КЛАССИКА COMPUTER SCIENCE

ВНУТРЕННЕЕ УСТРОЙСТВО WINDOWS

КЛЮЧЕВЫЕ КОМПОНЕНТЫ И ВОЗМОЖНОСТИ

СЕДЬМОЕ ИЗДАНИЕ

МАРК РУССИНОВИЧ, ДЭВИД СОЛОМОН
АЛЕКС ИОНЕСКУ, АНДРЕА АЛЛИЕВИ

Microsoft

Windows Internals
Seventh Edition
Part 2

Andrea Allievi
Alex Ionescu
Mark E. Russinovich
David A. Solomon

Внутреннее устройство Windows

Ключевые компоненты и возможности
Седьмое издание

Марк Руссинович
Дэвид Соломон
Алекс Ионеску
Андреа Аллиеви

ПИТЕР®

Санкт-Петербург • Москва • Минск

2025

ББК 32.973.2-018.2
УДК 004.451
В60

Руссинович Марк, Соломон Дэвид, Ионеску Алекс, Аллиеви Андреа

В60 Внутреннее устройство Windows. Ключевые компоненты и возможности. 7-е изд. — СПб.: Питер, 2025. — 992 с.: ил. — (Серия «Классика computer science»).

ISBN 978-5-4461-2015-4

Зная, что находится у операционной системы «под капотом», системные администраторы смогут быстро разобраться с поведением системы и решать задачи повышения производительности и диагностики сбоев. Специалистам по безопасности пригодится информация о борьбе с уязвимостями операционной системы.

Седьмое издание было полностью переработано под Windows 10/11 и Windows Server (2022, 2019 и 2016). Кроме этого в книгу были добавлены сведения по Hyper-V, полностью переработаны главы о процессе загрузки, новых технологиях хранения данных и механизмах управления Windows. Вы найдете уникальную информацию, основанную на исходном коде Microsoft, и практические эксперименты с применением новейших средств отладки, направленные на демонстрацию особенностей поведения внутренних компонентов Windows.

Новые элементы дизайна пользовательского интерфейса, появившиеся в Windows 11, основаны на знакомых по Windows 10 технологиях, поэтому читатели смогут сразу перейти на новый этап развития компьютерных технологий.

Для опытных программистов, архитекторов, администраторов и других специалистов по качеству, производительности, безопасности и сопровождению программного обеспечения.

Необходимо знание Windows на уровне опытного пользователя.

16+ (В соответствии с Федеральным законом от 29 декабря 2010 г. № 436-ФЗ.)

ББК 32.973.2-018.2
УДК 004.451

Информация, содержащаяся в данной книге, получена из источников, рассматриваемых издательством как надежные. Тем не менее, имея в виду возможные человеческие или технические ошибки, издательство не может гарантировать абсолютную точность и полноту приводимых сведений и не несет ответственности за возможные ошибки, связанные с использованием книги. В книге возможны упоминания организаций, деятельность которых запрещена на территории Российской Федерации, таких как Meta Platforms Inc., Facebook, Instagram и др. Издательство не несет ответственности за доступность материалов, ссылки на которые вы можете найти в этой книге. На момент подготовки книги к изданию все ссылки на интернет-ресурсы были действующими.

ISBN 978-0135462409 англ.
ISBN 978-5-4461-2015-4

Оглавление

Моим родителям Габриэле и Данило, а также моему брату Луке, которые всегда в меня верили и поддерживали мое стремление к мечте.

Андреа Аллиеви

Моим жене и дочери, которые всегда верили в меня и были неисчерпаемым источником любви и тепла. Моим родителям, вдохновившим меня следовать своей мечте, и их жертвам ради моих возможностей.

Алекс Ионеску

Об авторах этого издания

Андреа Аллиеви (Andrea Allievi) совмещает должности системного программиста и специалиста по безопасности уже более 15 лет. В 2010 году он получил диплом бакалавра по информатике в Миланском университете «Бикокка». В рамках дипломного проекта разработал 64-разрядный диспетчер главной загрузочной записи (MBR), способный обходить любые меры защиты ядра Windows 7, в частности PatchGuard и принудительную проверку подписания драйверов. Кроме того, Андреа имеет большой опыт в обратной разработке и специализируется на внутренних механизмах операционных систем, от ядра до пользовательского режима. Именно он создал первый буткит, обошедший защиту UEFI (который был разработан в научных целях и опубликован в 2012 году), множество хаков PatchGuard, был автором массы научных работ и статей. Его мастерству обязан своим существованием ряд системных утилит, предназначенных для удаления вредоносного программного обеспечения и постоянных серьезных угроз. Свою карьеру он строил во множестве IT-компаний: итальянской TgSoft, SaferBytes (теперь MalwareBytes), а также подразделении Talos корпорации Cisco Systems. Впервые он присоединился к Microsoft в 2016 году в роли специалиста по безопасности Центра защиты информации Microsoft (MSTIC). С января 2018-го Андреа числился старшим инженером ядра ОС подразделения защиты ядра в Microsoft, где в основном занимался инновациями (в частности, Retpoline и Speculation Mitigations) для NT и безопасного ядра.

Андреа остается активным участником сообщества по защите информации, периодически публикуя технические статьи о новинках в ядре Windows, в частности в блоге Microsoft Windows Internals, и выступая на множестве технологических конференций, таких как Recon и Microsoft BlueHat. Подпишитесь на Андреа в Twitter: @aall86.

Алекс Ионеску (Alex Ionescu) — вице-президент направления Endpoint в компании CrowdStrike Inc., где он начинал как сооснователь и главный архитектор. Алекс является специалистом мирового уровня по архитектуре безопасности и консультирующим экспертом в области низкоуровневого программного обеспечения, в разработке ядра, теории безопасности и обратной разработке. Более 20 лет его исследования помогали устранять десятки критических уязвимостей в защите ядра Windows и смежных с ним компонентов, а также множества поведенческих багов.

Ранее Алекс возглавлял разработку ядра ReactOS, клона Windows с открытым исходным кодом, созданного с нуля, где его перу принадлежит большинство Windows NT-подобных подсистем. Во время учебы Алекс работал в Apple, занимаясь там ядром iOS, загрузчиком и драйверами в составе оригинального костяка команды, стоявшей за iPhone, iPad и AppleTV. Наконец, Алекс Ионеску является основателем компании Winsider Seminars & Solutions, Inc., которая специализируется на низкоуровневом системном ПО, обратной разработке и тренингах по безопасности для организаций.

Алекс остается активным участником сообщества и выступает на множестве мероприятий по всему миру. На базе материалов книги «Внутреннее устройство Windows» он предоставляет услуги по обучению и поддержке как организациям, так и частным лицам. Подпишитесь на него в Twitter: @aionescu, а также на его блоги: www.alex-ionescu.com и www.windows--internals.com/blog.

Предисловие

Когда в 1993 году Microsoft выпустила чрезвычайно успешную ОС Windows NT 3.1, я, как человек, применявший ее и обозревавший ее внутреннее устройство, сразу же увидел, как сильно их новый продукт изменит мир. Дэвид Катлер (David Cutler), архитектор и глава разработки Windows NT, создал решение, которое отличалось безопасностью, надежностью и масштабируемостью, но при этом сохранило пользовательский интерфейс и совместимость с ПО, созданным под его старшего, более примитивного собрата. Книга Хелен Кастер (Helen Custer) «Внутри Windows NT» была фантастическим учебником по дизайну и архитектуре системы, но мне казалось, что назрел спрос на книгу еще более детальную. Книга «Внутреннее устройство и структуры данных VAX/VMS», известная как исчерпывающий гид по предыдущему детищу Дэвида Катлера, была настолько близка к исходному коду, насколько это возможно в прозе. Я решил, что напишу аналогичный учебник для Windows NT.

Дело шло медленно. Я был занят написанием диссертации и налаживанием карьеры в небольшой компании. Чтобы изучить Windows NT, я прочел документацию, дизассемблировал ее код и создал такие инструменты мониторинга системы, как Regmon и Filemon. Разработка системных утилит и их использование помогали мне лучше понять устройство Windows NT, так как позволяли обозревать процессы, происходящие «под капотом». В ходе своих изысканий я делился новыми знаниями в ежемесячной колонке «Внутреннее устройство NT» в журнале *Windows NT Magazine*, предназначенном для системных администраторов. Эти статьи позже легли в основу книги «Внутреннее устройство Windows», для создания которой меня наняло издательство IDG Press.

Сроки сдачи рукописи много раз менялись, так как у меня была основная работа и приходилось заниматься Sysinternals (тогда NTInternals) — бесплатным и коммерческим программным обеспечением в рамках моего стартапа Winternals Software. А в 1996 году Дэйв Соломон поразил меня вторым изданием «Внутри Windows NT». Книга показалась мне одновременно впечатляющей и деморализующей. В ней работа Хелен была полностью пересмотрена, устройство Windows NT описывалось гораздо глубже и масштабнее, как я и сам собирался сделать. В издание вошли уникальные исследования, где применялись встроенные инструменты и диагностические утилиты из наборов Windows NT Resource Kit и Device Driver Development Kit (DDK) и демонстрировались ключевые понятия и процессы. Дэйв поднял планку очень высоко. Мне стало ясно, что написание книги, которая могла бы соперничать с его работой по качеству и глубине, оказалось куда более монументальной задачей.

Как говорится в пословице, если не можешь прекратить процесс, возглавь его. Я знал Дэйва по совместному участию в конференциях по Windows. Не прошло

и двух недель с выхода его книги, как я отправил ему электронное письмо с предложением присоединиться к нему в качестве соавтора следующего издания, которое будет посвящено документированию системы, тогда называвшейся Windows NT 5, а позднее переименованной в Windows 2000. Моим вкладом стали бы новые главы, основанные на моей колонке NT Internals, на темы, которые Дэйв не упоминал, а также новые изыскания, для которых использовался пакет Sysinternals. Чтобы приукрасить предложение, я пообещал добавить последний на диск, прилагающийся к книге. Это вполне типичный способ распространения ПО.

Дэйв согласился. Правда, для начала ему нужно было получить разрешение от Microsoft. Я доставил им некоторые проблемы, раскрыв в публичном поле, что Windows NT Workstation и Windows NT Server — идентичные программы, меняющие свое поведение в зависимости от настроек в реестре. И если Дэйв имел полный доступ к исходному коду этих продуктов, я — нет и хотел, чтобы так и оставалось. Так не было бы риска вызвать проблемы с моей интеллектуальной собственностью, а именно с правами на ПО, созданное мною в рамках пакетов Sysinternals или Winternals, чей код опирался на недокументированный API. По удачному совпадению я уже давно восстанавливал отношения с ключевыми инженерами Windows, так что в Microsoft тактично согласились.

Писать вместе с Дэйвом «Внутри Windows 2000» было очень увлекательно. По невероятному совпадению он жил минутах в двадцати от меня (мы оба проживали в штате Коннектикут: я — в Данбери, а он — в Шермане). Мы ходили друг к другу в гости и проводили много времени вместе, копась в коде и исследуя глубины Windows. Было много профессиональных шуток, а попутно мы поднимали технические вопросы, при решении которых соперничали, кто разберется быстрее — он, читая исходный код, или я, вооружившийся отладчиком, дизассемблером и своим пакетом Sysinternals. (Если встретите его, не сыпьте соль на рану, но я всегда побеждал.)

Так я сделался соавтором самого исчерпывающего учебника по внутреннему устройству одной из самых коммерчески успешных операционных систем в истории. Для работы над пятым изданием, где шла речь о Windows XP и Windows Vista, мы пригласили в команду Алекса Ионеску. Алекс известен как один из сильнейших экспертов по операционным системам и обратной разработке в мире. Его участие обеспечило книге еще бóльшие глубину и кругозор, что позволило ей догнать и превзойти наши и без того высокие стандарты полноты изложения и достоверности. Количество охватываемых тем росло, да и сама система Windows увеличивалась и обрастала новыми возможностями и подсистемами. Наш труд превысил пределы, допустимые для одной книги, как получилось в пятом издании, поэтому пришлось разделить ее на два тома.

Когда начиналась работа над шестым изданием, я уже ушел в Azure. Команда была готова приступить к седьмому изданию, но возможности поучаствовать в создании книги у меня больше не было. Дэйв Соломон завершил карьеру, а задача обновления книги стала еще сложнее в силу того, что Windows перестали выпускать раз в несколько лет с новым номером основной версии, вместо этого ее назвали Windows 10 и стали добавлять новые функции и возможности постоянно. На помощь Алексу для работы над первым томом пришел Павел Йосифович (Pavel Yosifivitch). Но позже и он оказался слишком загружен другими проектами, поэтому

принять участие в написании второго тома не смог. Сам Алекс тоже был занят своим стартапом CrowdStrike, так что создание второго тома оказалось под угрозой.

К счастью, нам на помощь пришел Андреа. Вместе с Алексом они обновили немалую часть описания системы, включая процессы ее запуска и завершения работы, подсистему реестра и UWP. Не останавливаясь лишь на актуализации, они добавили три новые главы, где рассматривались Hyper-V, системы работы с файлами и кэшем, диагностика и журналирование. Наследие серии «Внутреннее устройство Windows» как самого технически детального и достоверного руководства о внутренних делах одного из самых важных программных продуктов в истории было спасено, и снова видеть свое имя среди участников — для меня предмет гордости.

Памятный момент в моей карьере наступил, когда мы попросили Дэвида Катлера написать предисловие для книги «Внутри Windows 2000». Я и Дэйв Соломон несколько раз приезжали в Microsoft, чтобы пообщаться с разработчиками Windows, среди которых был и Дэвид. Однако мы не могли сказать, согласится ли он, и были в восторге, когда это случилось. Очень необычно мне теперь ощущать себя с другой стороны, на месте Дэвида, и я почитаю за честь такую возможность. Я надеюсь, что похвалы, возданные мною в предисловии к этой книге, убедят вас в том, что она достоверна, ясна и понятна, как убедило предисловие Дэвида Катлера покупателей «Внутри Windows 2000».

Марк Руссинович (Mark Russinovich),
технический директор и стипендиат
проекта Azure, Microsoft
Март 2021 года, Белвью, Вашингтон

Введение

Книга «Внутреннее устройство Windows» (7-е изд., т. 2) предназначена для профессионалов в области информационных технологий — разработчиков, специалистов по защите информации и системных администраторов, которые желают понять, как устроены компоненты ядра операционных систем Windows 10 (вплоть до обновления, выпущенного в мае 2021 года и известного как 21H1) и Windows Server (с Server 2016 до Server 2022), включая множество тех, что присутствуют также в Windows 11X и операционной системе Xbox. Эти знания призваны помочь разработчикам лучше понимать обоснованность технических решений, принимаемых при построении приложений специально для платформы Windows, и лучше выбирать путь при создании мощных, масштабируемых и безопасных программных продуктов. Они помогут им отточить свои навыки отладки сложных проблем, корни которых тянутся глубоко в сердце системы. Попутно читатель освоит ряд утилит, которые будут очень полезны в дальнейшем.

Наконец, специалисты по защите информации смогут лучше понимать, когда приложения и операционная система могут вести себя ненормально, получилось это случайно или по чьему-то умыслу. Книга расскажет, какие функции безопасности, страхующие от подобных ситуаций, существуют в современных версиях Windows. Эксперты-криминалисты узнают, как пользоваться системными структурами данных и сервисами для поиска следов вторжения и как система сама их обнаруживает.

Кем бы ни был читатель этой книги, из нее он узнает очень многое о том, как работают системы семейства Windows и почему они ведут себя так, а не иначе.

ИСТОРИЯ СЕРИИ

Перед вами седьмое издание книги, которая изначально называлась «Внутри Windows NT» (Microsoft Press, 1992) и была написана Хелен Кастер (еще до первоначального релиза Microsoft Windows NT 3.1). Она стала самой первой книгой о Windows NT, где рассматривались ключевые вопросы архитектуры и дизайна данной системы.

Второе издание книги «Внутри Windows NT» (Microsoft Press, 1998) создано Дэвидом Соломоном. К материалу предшественника была добавлена информация о Windows NT 4.0, а заодно существенно детализирован разбор технической стороны.

Третье издание книги «Внутри Windows 2000» (Microsoft Press, 2000) было написано Дэвидом Соломоном и Марком Руссиновичем. Там затрагивалось много новых тем, таких как запуск и завершение работы системы, внутреннее устройство служб, реестр и драйверы файловой системы, а также взаимодействие с сетью. Рассматривались изменения функций ядра Windows 2000, таких как Windows Driver

Model (WDM), Plug and Play, управление питанием, инструментарий управления Windows (WMI), криптография, объекты заданий и службы терминалов.

В четвертом издании, названном «Внутреннее устройство Windows» (Microsoft Press, 2004), были описаны Windows XP и Windows Server 2003, появилось больше информации о том, как эффективнее пользоваться своими знаниями, в частности, о применении ключевых инструментов из пакета Windows SysInternals и анализе аварийных дампов.

Пятое издание книги «Внутреннее устройство Windows» (Microsoft Press, 2009) получило обновление до Windows Vista и Windows Server 2008. В числе авторов появился Алекс Ионеску, сменивший Марка Руссиновича, поступившего на работу в Microsoft, где он теперь является техническим директором Azure. Среди новых тем — загрузчик образов, отладчик пользовательского режима, продвинутый локальный вызов процедур (ALPR) и Hyper-V.

В шестом издании «Внутреннего устройства Windows» (Microsoft Press, 2012) были исчерпывающе изложены изменения ядра в Windows 7 и Windows Server 2008 R2 и добавлены примеры для отражения изменений в наших инструментах.

ИЗМЕНЕНИЯ В СЕДЬМОМ ИЗДАНИИ

Вышло так, что число страниц шестого издания оказалось больше, чем можно напечатать в современных условиях, поэтому оно стало первым, которое разделили на два тома. Заодно это позволило авторам опубликовать часть книги раньше (том 1 вышел в марте 2012-го, том 2 — в сентябре). Впрочем, в тот раз это решение было принято исключительно из-за количества страниц, так что порядок следования глав оставался таким же, как в предыдущих изданиях.

После выхода шестого издания Microsoft запустила процесс конвергенции ОС, в ходе которого сначала были объединены ядра Windows 8 и Windows Phone 8, а за этим последовали более современные среды выполнения приложений из Windows 8.1, Windows RT и Windows Phone 8.1. Венцом всему стала Windows 10, способная работать на настольных ПК, лэптопах, серверах, консолях XBox One, HoloLens и различных приборах, подключенных к Интернету вещей (IoT). Это великое воссоединение обозначало, что пришло время очередного издания, куда можно будет включить нововведения за без малого пять лет разработки.

В седьмом издании (Microsoft Press, 2017) авторы так и поступили, впервые пригласив Павла Йосифовича, заменившего Дэвида Соломона в роли «своего человека» в Microsoft и в целом ставшего менеджером проекта. Работая вместе с Алексом Ионеску, который, как и Марк, продолжил свой путь на новой основной работе в CrowdStrike (где он теперь вице-президент VP направления Endpoint), Павел принял решение переписать главы книги так, чтобы каждая часть была изложена последовательно и читателю не приходилось дожидаться второго тома, где завершалось бы разъяснение тем из первого. Это позволило тому 1 стать самостоятельным произведением, где читатель знакомится с ключевыми компонентами архитектуры системы Windows 10, управления процессами, планирования потоков, управления памятью, систем ввода/вывода, систем безопасности пользователей, данных и платформы. В этом томе рассматриваются

аспекты Windows 10 вплоть до версии 1073, обновления, выпущенного в мае 2017 года, а также Windows Server 2016.

Изменения в томе 2

После того как Алекс Ионеску и Марк Руссинович полностью посвятили себя основной работе, а Павел ушел на другие проекты, для тома 2 много лет не могли найти основного автора. Команда была очень благодарна Андреа Аллиеви, который решился взвалить на себя труд завершить серию. С Алексом в роли советника и наставника он имел полный доступ к базе исходных кодов Microsoft, как все соавторы до него. Впервые за историю серии Андреа, являясь полноправным участником команды разработки ядра Windows, привнес новое ви́дение и существенно переработал книгу.

Рассудив, что главы о взаимодействии с сетью и анализе аварийных дампов неактуальны для современного читателя, вместо них он добавил совершенно новое описание Hyper-V, что теперь является одной из ключевых частей стратегии платформы Windows как для клиентских систем, так и в рамках Azure. Вдобавок в книге появились полностью переписанные главы о процессе загрузки системы, новых технологиях хранения данных, таких как ReFS и DAX, и масса обновлений, относящихся к системным и административным механизмам. Практические примеры были полностью переработаны, чтобы продемонстрировать новые возможности отладчика и инструментальных средств.

Длительное ожидание выхода тома 2 позволило обеспечить соответствие книги последней публичной сборке Windows 10, версии 2103 (май 2021-го, 21H1), включая редакции Windows Server 2019 и 2022. Так читатели не отстанут после столь долгой паузы. Поскольку Windows 11 строится на основе ядра той операционной системы, они будут готовы к работе и с этой новой версией.

ПРАКТИЧЕСКИЕ ЭКСПЕРИМЕНТЫ

Даже не имея доступа к исходным кодам, вы сможете «увидеть» немалую часть внутреннего устройства Windows благодаря отладчику ядра, пакету SysInternals и утилитам, созданным специально для этой книги. Когда какой-то инструмент может быть использован для демонстрации некоторой особенности внутреннего поведения Windows, пошаговое рассмотрение того, как это сделать, будет размещаться во врезках, озаглавленных «ЭКСПЕРИМЕНТ». Они разбросаны по всей книге, и мы настоятельно рекомендуем по ходу чтения попробовать выполнить описанное. Живое доказательство того, что происходит внутри Windows, даст вам больше впечатлений, чем если вы просто прочтете об этом.

НЕЗАТРОНУТЫЕ ТЕМЫ

Windows — это большая и сложная операционная система. Данная книга не рассматривает темы, смежные с вопросами внутреннего устройства Windows, а фокусируется исключительно на базовых компонентах системы. К примеру, здесь не рассказывается о COM+ — распространяемой в составе Windows инфраструктуре для

объектно-ориентированного программирования, как и о .NET Framework — платформе для приложений с управляемым кодом. Эта книга не для пользователей, программистов или системных администраторов, она именно о внутреннем устройстве. Научить программировать, настраивать или использовать Windows в ее задачи не входит.

ПРЕДУПРЕЖДЕНИЕ И ПРЕДОСТЕРЕЖЕНИЕ

Описанное в книге поведение архитектуры и процессов операционной системы Windows, таких как внутренние функции и структуры ядра, публично не документировано. Это значит, что оно может произвольно изменяться между релизами продукта. Мы не утверждаем, что оно поменяется обязательно, но и на обратное рассчитывать нельзя. Любое программное обеспечение, опирающееся в своей работе на недокументированный внутренний функционал или конфиденциальные сведения об операционной системе, может перестать работать в следующей версии Windows. Хуже того, программы, работающие на уровне ядра (в частности, драйверы устройств) и применяющие недокументированные возможности, рискуют в новых версиях Windows вызвать отказ системы, ставя под серьезную угрозу данные тех, кто ими воспользуется.

Проще говоря, настоятельно не рекомендуется применять предназначенные для внутреннего использования Windows функции, разделы реестра, механики, API и другую недокументированную информацию из данной книги для разработки любого программного обеспечения для конечных пользователей. Да и с любой другой целью, кроме исследования и документирования. По всем вопросам следует обращаться к официальной документации в рамках MSDN (Microsoft Software Development Network).

ЧЕГО МЫ ОЖИДАЕМ ОТ ЧИТАТЕЛЯ

Мы предполагаем, что читатель работает с Windows на уровне опытного пользователя, знаком с основными понятиями из области операционных систем и аппаратного обеспечения, такими как регистры процессора, память, процессы и потоки. При чтении некоторых глав окажется полезным базовое знакомство с функциями, указателями и прочими понятиями языка программирования C.

СТРУКТУРА КНИГИ

Как и предыдущее, шестое издание, данная книга разделена на два тома, второй из которых сейчас перед вами.

- Глава 8 «Системные механизмы» рассказывает о важнейших внутренних механизмах, используемых операционной системой для предоставления драйверам устройств и приложениям ключевых служб, в частности ALPC, диспетчера объектов и процедур синхронизации. Кроме того, раскрываются детали аппаратной архитектуры в основе Windows, в том числе обработка системных прерываний, сегментация, уязвимости сторонних каналов и предосторожности, принимаемые для того, чтобы их избежать.

- Глава 9 «Технологии виртуализации» описывает, как операционная система Windows применяет технологии виртуализации, реализованные в современных процессорах, чтобы дать пользователям возможность запускать по несколько виртуальных машин в одной системе. Кроме того, виртуализация позволяет Windows выводить безопасность на новый уровень. Поэтому в главе активно обсуждаются технологии безопасного ядра и изолированного пользовательского режима.
- Глава 10 «Управление, диагностика и трассировка» разъясняет фундаментальные механизмы операционной системы, предназначенные для управления ею, а также ее настройки и диагностики. В частности, рассматриваются реестр Windows, службы, WMI и планировщик задач. Представлены диагностические инструменты, такие как служба событий и DTrace.
- Глава 11 «Кэширование и файловые системы» описывает, как самые важные компоненты системы хранения данных, такие как диспетчер кэша и драйверы файловой системы, взаимодействуют, чтобы позволить Windows эффективно и безотказно работать с файлами, папками и дисковыми устройствами. В главе рассмотрены поддерживаемые Windows файловые системы, особенно подробно — NTFS и ReFS.
- Глава 12 «Запуск и завершение работы системы» описывает порядок операций, выполняющихся при запуске и завершении работы, и компоненты операционной системы, задействованные в процессе загрузки. В главе также приводится анализ таких новых технологий, как Secure Boot, Measured Boot и Secure Launch.

ШРИФТОВЫЕ ВЫДЕЛЕНИЯ

В книге применяются следующие шрифтовые выделения.

- Рубленым шрифтом выделены элементы управления, названия клавиш и электронные адреса.
- *Курсивом* обозначены ранее не упоминавшиеся термины.
- Фрагменты кода и текст, который вам следует где-то ввести, обозначены моноширинным шрифтом.
- Первые буквы названий диалоговых окон и элементов управления в них пишутся прописными буквами.
- Сочетания горячих клавиш указываются со знаком плюс между названиями. К примеру, Ctrl+Alt+Delete означает требование одновременно нажать клавиши Ctrl, Alt и Delete.

БЛАГОДАРНОСТИ

Приводимые в книге сложные технические сведения и стоящую за ними логику часто нелегко описать и понять неспециалисту. На протяжении своей истории серия давала двойное преимущество. С одной стороны, составляющие ее книги освещали сторонние исследования по обратной разработке (reverse-engineering),

а с другой — сотрудники или подрядчики из Microsoft могли заполнить пробелы в них, имея доступ к огромной базе знаний, сокрытой в стенах Microsoft, и богатейшей истории разработки операционных систем Windows. В случае тома 2 седьмого издания это заслуга лично Андреа Аллиеви. Команда выражает ему благодарность за участие в роли основного автора и помощь в доведении до конца работы над большей частью книги и ее обновленным содержимым.

Вместе с тем эта книга не обладала бы технической глубиной и достоверностью без рецензирования, комментариев и поддержки от ключевых участников команды разработки Windows, других экспертов из Microsoft и иных коллег, друзей и экспертов в различных областях.

Отдельного упоминания заслуживает то, что переписанная глава 9 «Технологии виртуализации» не была бы столь полной и подробной без помощи Александра Греста (Alexander Grest) и Джона Ланге (Jon Lange), экспертов мирового класса, которые заслуживают особой благодарности за все те дни, когда они помогали Андреа разобраться в наиболее запутанных функциях гипервизора и безопасного ядра (Secure Kernel).

Алекс хотел бы выразить особую благодарность Аруну Кишану (Arun Kishan), Мехмету Игану (Mehmet Iyigun), Дэвиду Вестону (David Weston) и Энди Лурсу (Andy Luhrs), которые продолжают поддерживать книгу и обеспечивают ему возможность получать доступ к людям и информации внутри фирмы, чтобы сделать работу достоверной и законченной.

Мы хотим поблагодарить людей, которые предоставили технические рецензии и/или исходный материал для книги либо просто поддерживали авторов и помогали им: Саар Амар (Saar Amar), Крэйга Баркхауза (Craig Barkhouse), Мишель Бержерон (Michelle Bergeron), Джо Бялека (Joe Bialek), Кевина Броаса (Kevin Broas), Омара Кэри (Omar Carey), Нила Кристиансена (Neal Christiansen), Крис Ферналд (Chris Fernald), Стивена Финнигана (Stephen Finnigan), Элию Флорио (Elia Florio), Джеймса Форшоу (James Forshaw), Эндрю Харпера (Andrew Harper), Бена Хиллиса (Ben Hillis), Говарда Капустейна (Howard Kapustein), Сарухана Карадемира (Saruhan Karademir), Криса Клейнханса (Chris Kleynhans), Джона Ламберта (John Lambert), Аттилио Майнетти (Attilio Mainetti), Билла Мессмера (Bill Messmer), Мэта Миллера (Matt Miller), Джейка Ошинса (Jake Oshins), Саймона Поупа (Simon Pope), Джордана Рабета (Jordan Rabet), Лорен Робинсон (Loren Robinson), Арупа Роя (Arup Roy), Ярдена Шафира (Yarden Shafir), Андрея Шеделя (Andrey Shedel), Джейсона Ширка (Jason Shirk), Акселя Суше (Axel Souchet), Атула Талесару (Atul Talesara), Сатоси Танду (Satoshi Tanda), Педро Тейшейру (Pedro Teixeira), Габриэля Виалу (Gabrielle Viala), Нэйта Уорфилда (Nate Warfield), Мэтью Вулмана (Matthew Woolman) и Адама Заброцки (Adam Zabrocki).

Хочется еще раз поблагодарить Илфака Гуилфанова (Ilfak Guilfanov) из компании Hex-Rays (http://www.hex-rays.com) за предоставленные Алексу Ионеску лицензии на использование IDA Pro Advanced и Hex-Rays, в том числе за их недавнее пожизненное продление, что сыграло большую роль в ускорении обратной разработки ядра Windows. Команда Hex-Rays продолжает поддерживать Алекса, включая в каждый релиз полезные ему опции декомпилятора, что позволяет работать над такой книгой, как эта, не имея доступа к исходному коду системы.

Наконец, авторы выражают огромную признательность коллективу сотрудников Microsoft Press (Pearson), которые поспособствовали тому, что наша книга стала реальностью. Лоретта Ятес (Loretta Yates), Чарви Арора (Charvi Arora) и их коллеги — все заслуживают отдельного упоминания за их бесконечное терпение, начиная с подписания контракта в 2018-м и до выхода книги в свет через два с половиной года.

ОСТАЕМСЯ НА СВЯЗИ

Мы не прощаемся! Посетите наш Twitter: @MicrosoftPress.

ОТ ИЗДАТЕЛЬСТВА

Уважаемые читатели! Не удивляйтесь, что эта книга начинается с восьмой главы. Авторы разделили свой труд на две части. Первая часть книги вышла в издательстве «Питер» в 2019 году. В ней рассмотрены следующие темы.

Глава 1. Концепции и средства.

Глава 2. Архитектура системы.

Глава 3. Процессы и задания.

Глава 4. Потоки.

Глава 5. Управление памятью.

Глава 6. Подсистема ввода/вывода.

Глава 7. Безопасность.

К этому изданию прилагается набор утилит, которые помогут вам в выполнении экспериментов. Их можно скачать по адресу https://storage.piter.com/support_insale/files_list.php?code=978544612015&partner=3401.

Ваши замечания, предложения, вопросы отправляйте по адресу comp@piter.com (издательство «Питер», компьютерная редакция).

Мы будем рады узнать ваше мнение!

На веб-сайте издательства www.piter.com вы найдете подробную информацию о наших книгах.

ГЛАВА 8

Системные механизмы

Операционная система Windows предоставляет ряд основных механизмов, используемых такими компонентами режима ядра, как исполнительная система, ядро и драйверы устройств. В данной главе рассматриваются следующие системные механизмы и описывается порядок их использования.

- Модель выполнения кода процессором, включая понятие уровней колец, сегментация, состояния задач, диспетчеризация системных прерываний, отложенный вызов процедур (deferred procedure call, DPC), асинхронный вызов процедур (asynchronous procedure call, APC), таймеры, рабочие потоки системы, диспетчеризация исключений и диспетчеризация системных служб.
- Барьеры спекулятивного выполнения и другие средства защиты от уязвимостей по линии приложений.
- Диспетчер объектов.
- Синхронизация, включая спин-блокировки, объекты диспетчера ядра, порядок реализации ожиданий, а также примитивы синхронизации, относящиеся к пользовательскому режиму, такие как адресованные ожидания, условные переменные, тонкие блокировки «чтение — запись» (slim reader-writer, SRW).
- Продвинутый локальный вызов процедур (Advanced Local Procedure Call, ALPC).
- Средство уведомлений Windows (Windows Notification Facility, WNF).
- WoW64.
- Фреймворк отладки для пользовательского режима.

Дополнительно в главе подробно описываются универсальная платформа Windows (Universal Windows Platform, UWP) и набор служб как пользовательского режима, так и режима ядра, на работу которых она опирается, в частности:

- пакетных приложений и службы развертывания AppX;
- приложений Centennial и Windows Desktop Bridge;
- диспетчера состояний процессов (Process State Management, PSM) и диспетчера жизненного цикла процессов (Process Lifetime Manager, PLM);
- модератора активности хоста (Host Activity Moderator, HAM) и модератора фоновой активности (Background Activity Moderator, BAM).

МОДЕЛЬ ВЫПОЛНЕНИЯ КОДА ПРОЦЕССОРОМ

В этом разделе мы рассмотрим внутренние механизмы архитектуры процессоров i386 от Intel и ее расширения — архитектуры AMD64, фигурирующей в современных системах. Хотя обе компании создавали эти технологии самостоятельно, стоит отметить, что сегодня они вместе реализуют решения друг друга. Поэтому, несмотря на присутствие вышеупомянутых терминов в именах файлов и ключей реестра Windows, названия x86 (32 бита) и x64 (64 бита) сегодня употребляются более часто.

Речь пойдет о сегментации, задачах, уровнях привилегий, критических механизмах. Наконец, мы рассмотрим понятия диспетчеризации системных прерываний и системных вызовов.

Сегментация

Высокоуровневые языки программирования, такие как C/C++ или Rust, преобразуются компилятором в *машинный код*, часто называемый *ассемблером*. Этот низкоуровневый язык позволяет обращаться к регистрам процессора напрямую. Зачастую программам доступны три основных вида регистров (из тех, что можно увидеть в отладчике):

- программный счетчик (Program Counter, PC), в архитектурах x86/x64 называемый *указателем команд* (Instruction Pointer, IP), где он представлен регистрами EIP (для x86) и RIP (для x64). Его значение всегда указывает на команду в машинном коде, исполняемую в данный момент, кроме некоторых 32-разрядных архитектур класса ARM;
- указатель стека (Stack Pointer, SP), представленный регистрами ESP (для x86) и RSP (для x64). Его значение указывает на позицию в памяти, соответствующую текущей вершине стека;
- все остальные, известные как регистры общего назначения (General Purpose Registers, GPR), среди которых EAX/RAX, ECX/RCX, EDX/RDX, ESI/RSI, а также R8, R14 и множество других.

Хотя перечисленные регистры могут содержать значения адресов в памяти, существует еще несколько, используемых в условиях *сегментации в безопасном режиме*. Последняя предполагает проверки для различных *сегментных регистров*, также называемых *селекторами*.

- Все попытки доступа к программному счетчику предварительно оцениваются относительно регистра сегмента кода (Code Segment, CS).
- Все попытки доступа к указателю стека предварительно оцениваются относительно регистра сегмента стека (Stack Segment, SS).
- Обращения к прочим регистрам определяются *переопределением сегмента*, в ходе чего кодирование позволяло задать проверку сегмента данных (Data Segment, DS), дополнительных сегментов ES или FS.

Эти селекторы размещаются в 16-разрядных регистрах, а их значения хранятся в структуре данных, которая называется *глобальной таблицей дескрипторов* (Global

Descriptor Table, GDT). Чтобы обращаться к ней, процессор хранит ее адрес в еще одном особом регистре — GDTR. Структура такого селектора показана на рис. 8.1.

Рис. 8.1. Структура селектора сегмента в системе x86

Тем самым смещение, находящееся в селекторе сегмента, берется из GDT, если только не установлен флаг TI (Table Indicator). В последнем случае происходит обращение к другой структуре данных — *локальной таблице дескрипторов* (Local Descriptor Table, LDT). Ее адрес хранится в регистре LDTR, и она не используется в современных системах Windows. Результатом поиска становится запись о сегменте, а в случае неуспеха — некорректная запись, что приведет к общей ошибке безопасности (#GP) или исключению ошибки сегмента.

Эта запись, называемая в современных ОС *дескриптором сегмента*, служит для двух важнейших целей.

- Для сегментов кода дескриптор содержит *код уровня привилегий* (Code Privilege Level, CPL). Эти уровни привилегий часто описываются как *защитные кольца*, на которых код будет выполняться. Данный уровень (в диапазоне от 0 до 3) затем кэшируется в двух старших битах самого селектора (см. рис. 8.1). Операционные системы, такие как Windows, применяют уровень (кольцо) 0 для запуска компонентов ядра и драйверов, а уровень 3 — для приложений и служб.

  Кроме того, в системах типа x64 сегмент кода имеет признак того, используется он в *длинном режиме* (long mode) или в *режиме совместимости*. Первый применяется для прямого выполнения кода под x64, а второй обеспечивает совместимость с кодом под x86. Аналогичный механизм существует и в x86, где сегменты могут помечаться как 16- или 32-разрядные.
- Для прочих сегментов в дескрипторе указывается *уровень привилегий дескриптора* (Descriptor Privilege Level, DPL), который необходим для доступа к сегменту. Хотя в современных системах подобная проверка уже анахронизм, процессоры все еще требуют (а приложения все еще ожидают) ее правильной настройки.

Наконец, в системах типа x86 дескрипторы сегмента могут дополнительно хранить 32-разрядный *базовый адрес*, в случае чего любое значение, уже помещенное в регистр в результате переопределения сегмента, будет считаться смещением от этого адреса. Чтобы смещение не превысило некоторого предела, оно сверяется с соответствующим *пределом сегмента*. Поскольку в большинстве ОС базовый адрес приравнен к 0 (а предел — к 0xFFFFFFFF), в архитектуре x64 этот концепт реализован везде, за исключением селекторов FS и GS, поведение которых несколько отличается.

- Если сегмент кода работает в длинном режиме, базовый адрес для сегмента FS берется из моделезависимого регистра FS_BASE (Model Specific Register, MSR) — 0C0000100h. Для сегмента GS нужно загрузить GS_BASE MSR-0C0000101h или GS_SWAP MSR-0C0000102h в зависимости от *состояния свопа (замещения)*, задаваемого инструкцией swapgs.

Если в селекторе сегмента FS или GS установлен бит TI, тогда значение берется из записи в LDT с соответствующим смещением, которое ограничено только 32-разрядными базовыми адресами. Так сделано для обеспечения совместимости с некоторыми операционными системами, а сам предел игнорируется.

- Если сегмент кода работает в режиме совместимости, тогда базовый адрес обычным образом считывается из записи в GDT (или LDT, если установлен бит TI). Предел устанавливается и проверяется по смещению в регистре, заданному после переопределения сегмента.

Такое поведение сегментов FS и GS позволяет операционным системам вроде Windows реализовать *эффект локальности регистров в рамках потока*, где к отдельным структурам данных можно обращаться по базовому адресу сегмента, что дает возможность получить простой доступ к конкретным смещениям/полям внутри него. К примеру, согласно описанию, приведенному главе 3 тома 1, Windows хранит адрес блока окружения потока (Thread Environment Block, TEB) в сегментном регистре FS в системах x86 и в GS (замещенном) в системах x64. В таком случае при выполнении кода уровня ядра в x86 FS вручную переключается на другой сегмент, который соответствует адресу Kernel Processor Control Region (KPCR). В свою очередь, для архитектуры x64 этот адрес сразу помещается в регистр GS (не замещенный).

Таким образом, сегментация позволяет реализовать в Windows следующие эффекты: закодировать и установить уровень привилегий, который будет иметь исполняемый фрагмент кода на конкретном уровне процессора, обеспечить прямой доступ к структурам данных TEB и KPCR в пользовательском режиме и/или режиме уровня ядра соответственно. Заметим, что поскольку на глобальную таблицу дескрипторов ссылается регистр процессора GDTR, у каждого процессора она может быть своя. На самом деле именно этим Windows пользуется, чтобы обеспечить каждому процессору загрузку правильного KPCR для каждой GDT и то, что для текущего исполняемого потока на текущем процессоре его TEB находится в нужном сегменте.

ЭКСПЕРИМЕНТ. Просмотр GDT в x64

При проведении удаленной отладки или анализе аварийного дампа (в том числе при использовании LiveKD) вы сможете просмотреть содержимое GDT, включая состояние всех сегментов и их базовые адреса (где это применимо), с помощью команды отладчика dg. Она принимает на вход стартовый и конечный сегменты, в приведенном примере это 10h и 50h:

```
0: kd> dg 10 50

                                                    P Si Gr Pr Lo
Sel        Base              Limit          Type    l ze an es ng Flags
---- ----------------- ----------------- ---------- - -- -- -- -- --------
0010 00000000`00000000 00000000`00000000 Code RE Ac 0 Nb By P  Lo 0000029b
0018 00000000`00000000 00000000`00000000 Data RW Ac 0 Bg By P  Nl 00000493
0020 00000000`00000000 00000000`ffffffff Code RE Ac 3 Bg Pg P  Nl 00000cfb
0028 00000000`00000000 00000000`ffffffff Data RW Ac 3 Bg Pg P  Nl 00000cf3
0030 00000000`00000000 00000000`00000000 Code RE Ac 3 Nb By P  Lo 000002fb
0050 00000000`00000000 00000000`00003c00 Data RW Ac 3 Bg By P  Nl 000004f3
```

Ключевые сегменты здесь 10h, 18h, 20h, 28h, 30h и 50h. (Приведенный вывод несколько сокращен: удалена информация, не относящаяся к данной теме.)

В 10h (KGDT64_R0_CODE) вы можете наблюдать сегмент кода в длинном режиме на уровне нулевого кольца, что видно по `0` в столбце `Pl`, коду `Lo` в столбце `Long` и типу `Code RE`. Аналогично в 20h (KGDT64_R3_CMCODE) вы видите `Nl`-сегмент в третьем кольце (not long, то есть в режиме совместимости), где сейчас выполняется код для x86 в подсистеме WOW64. В 30h (KGDT64_R3_CODE), в свою очередь, находится аналогичный сегмент в длинном режиме. Затем обратите внимание на 18h (KGDT64_R0_DATA) и 28h (KGDT64_R3_DATA), которые соответствуют сегментам стека, данных и дополнительному сегменту.

Наконец, последний сегмент по 50h (KGDT_R3_CMTEB), как правило, имеет нулевой базовый адрес, если только вы не запускаете код под x86 в WOW64 при выводе дампа GDT. Как было сказано ранее, здесь будет храниться адрес TEB при активном режиме совместимости.

Чтобы посмотреть сегменты TEB и KPCR в 64-разрядном режиме, вам вместо этого потребуется выгрузить (dump) соответствующие MSR, что можно сделать следующими командами (актуально при удаленной и локальной отладке ядра, но не при анализе дампа):

```
lkd> rdmsr c0000101
msr[c0000101] = ffffb401`a3b80000

lkd> rdmsr c0000102
msr[c0000102] = 000000e5`6dbe9000
```

Вы можете сравнить эти значения с содержимым `@$pcr` и `@$teb`, где будут находиться те же данные:

```
lkd> dx -r0 @$pcr
@$pcr                 : 0xffffb401a3b80000 [Type: _KPCR *]
lkd> dx -r0 @$teb
@$teb                 : 0xe56dbe9000 [Type: _TEB *]
```

ЭКСПЕРИМЕНТ. Просмотр GDT в x86

В x86-системах GDT оформляется похожими сегментами, но с другими селекторами. Кроме того, из-за двойного назначения сегментного регистра FS вместо функциональности swapgs, а также из-за отсутствия длинного режима количество селекторов несколько отличается, что видно в примере:

```
kd> dg 8 38

                                 P Si Gr Pr Lo
Sel    Base     Limit     Type    l ze an es ng Flags
---- -------- -------- ---------- - -- -- -- -- --------
0008 00000000 ffffffff Code RE Ac 0 Bg Pg P  Nl 00000c9b
0010 00000000 ffffffff Data RW Ac 0 Bg Pg P  Nl 00000c93
```

```
0018 00000000 ffffffff Code RE    3 Bg Pg P  Nl 00000cfa
0020 00000000 ffffffff Data RW Ac 3 Bg Pg P  Nl 00000cf3
0030 80a9e000 00006020 Data RW Ac 3 Bg By P  Nl 00000493
0038 00000000 00000fff Data RW    3 Bg By P  Nl 000004f2
```

Ключевые сегменты здесь 8h, 10h, 18h, 20h, 30h и 38h. По смещению 08h (KGDT_R0_CODE) можно наблюдать сегмент кода на уровне нулевого кольца. Аналогично по смещению 18h (KGDT_R3_CODE) находится сегмент на уровне третьего кольца. Наконец, в 10h (KGDT_R0_DATA) и 20h (KGDT_R3_DATA) будут сегменты стека, данных и дополнительный сегмент.

В системах x86 в сегментах 30h (KGDT_R0_PCR) и 38h (KGDT_R3_TEB) будут базовые адреса KPCR и TEB текущего потока соответственно. В данной архитектуре никаких MSR для сегментации не используется.

Отложенная загрузка сегментов

С учетом описанного ранее поведения сегментов исследование значений сегментных регистров DS и ES в системах x86 и/или x64 может дать неожиданные результаты: они не всегда соответствуют значениям определенных для них уровней кольца. К примеру, какой-нибудь поток в архитектуре x86 в пользовательском режиме будет иметь следующие сегменты:

```
CS = 1Bh (18h | 3)
ES, DS = 23 (20h | 3)
FS = 3Bh (38h | 3)
```

Но во время системного вызова на уровне нулевого кольца обнаружатся другие сегменты:

```
CS = 08h (08h | 0)
ES, DS = 23 (20h | 3)
FS = 30h (30h | 0)
```

Похожим образом поток уровня ядра для x64 будет устанавливать свои сегментные регистры ES и DS в 2Bh (28h | 3). Это несоответствие сопутствует функциональности, называемой *отложенной загрузкой сегментов*, и отражает бессмысленность уровня привилегий дескриптора (DPL) для сегмента данных в ситуации, когда текущий уровень привилегий кода (CPL) равен 0 в условиях системы с плоской моделью памяти. Поскольку более высокий CPL всегда имеет доступ к данным более низких DPL (но не наоборот), установка DS и/или ES в «правильное» значение при входе в ядро потребовала бы их восстановления при возвращении в пользовательский режим.

Хотя инструкция MOV DS, 10h может показаться банальной, при встрече с ней микрокоду процессора потребуется выполнить несколько проверок на корректность селектора, что существенно повысит стоимость системных вызовов и обработки прерываний. Чтобы избежать этого, Windows всегда использует значения для регистров сегментов данных с уровня кольца 3.

Сегменты состояния задач

Кроме сегментных регистров кода и данных в системах архитектур x86 и x64, существует дополнительный особый регистр — регистр задач (Task Register, TR), который работает как еще один 16-битный селектор, где хранится смещение в пределах GDT. Однако в данном случае сегмент ассоциирован не с кодом или данными, а скорее с *задачей*. С точки зрения внутреннего состояния процессора он характеризует исполняемый фрагмент кода, называемый *Task State*, в случае Windows — текущий поток. Эти состояния задач, представленные сегментами (сегмент состояния задачи — Task State Segment, TSS), используются в современных операционных системах под x86 для решения ряда задач, связанных с критическими перехватами в процессоре (подробнее о них — в следующем разделе). Как минимум TSS характеризует каталог страниц (через регистр CR3), например PML4 в системах x64 (подробнее о страницах см. в главе 5 тома 1), сегмент кода, сегмент стека, указатель инструкций (IP) и до четырех указателей на стек (по одному на каждый уровень кольца). Подобные TSS используются в следующих целях.

- Они характеризуют текущее состояние исполнения кода в моменты, когда не происходит каких-либо системных прерываний. Впоследствии процессор пользуется этим для корректной обработки исключений и прерываний, загружая стек нулевого кольца из TSS, если перед этим он работал на уровне третьего кольца.
- Для разрешения ситуаций архитектурной гонки при работе с отказами отладки (#DB), для чего требуется выделенный TSS с отдельным обработчиком отказов отладки и стеком ядра.
- Характеризуют состояние исполнения кода, которое следует применить в случае срабатывания ловушки двойного отказа (#DF). С их помощью выполняется переход в обработчик двойного отказа с безопасным (запасным) стеком ядра вместо стека текущего потока, который и мог вызвать отказ.
- Характеризуют состояние исполнения кода, которое следует применить для немаскируемого прерывания (Non Maskable Interrupt, NMI). Используются для перехода в обработчик NMI с безопасным стеком ядра.
- Наконец, в похожих ситуациях при исключении машинных проверок (#MCE), которые по тем же причинам могут исполняться в условиях отдельного, безопасного стека ядра.

В системах x86 главный (текущий) TSS окажется в GDT по селектору 028h, что объясняет наличие в регистре TR аналогичного значения в ходе нормального выполнения кода в Windows. Кроме того, #DF TSS будет по 58h, NMI TSS — по 50h, а #MCE TSS — по 0A0h. Наконец, #DB TSS будет по 0A8h.

В системах x86 нет возможности содержать несколько TSS, так как эта функциональность была сведена в основном к единственной необходимости исполнять обработчики системных прерываний, выполняющиеся на выделенном стеке ядра. Таким образом, теперь используется лишь один TSS (в случае с Windows — по 040h), где содержится массив из восьми указателей стека, известный как таблица стеков прерываний (Interrupt Stack Table, IST). Каждое из предшествующих системных прерываний здесь ассоциируется с позицией в IST, а не с отдельным TSS. В следующем разделе, где мы разберем несколько записей IDT, вы сможете увидеть разницу в поведении систем x86 и x64 и в том, как они обрабатывают системные прерывания.

ЭКСПЕРИМЕНТ. Просмотр структур TSS в системах x86

В системе x86 мы можем заглянуть в общесистемный TSS по селектору 28h, воспользовавшись уже знакомой нам командой dg:

```
kd> dg 28 28
                                  P Si Gr Pr Lo
Sel    Base     Limit     Type    l ze an es ng Flags
---- -------- -------- ---------- - -- -- -- -- --------
0028 8116e400 000020ab TSS32 Busy 0 Nb By P  Nl 0000008b
```

На выходе появится виртуальный адрес структуры данных KTSS, которую можно выгрузить с помощью команд dx или dt:

```
kd> dx (nt!_KTSS*)0x8116e400
(nt!_KTSS*)0x8116e400                : 0x8116e400 [Type: _KTSS *]
    [+0x000] Backlink        : 0x0 [Type: unsigned short]
    [+0x002] Reserved0       : 0x0 [Type: unsigned short]
    [+0x004] Esp0            : 0x81174000 [Type: unsigned long]
    [+0x008] Ss0             : 0x10 [Type: unsigned short]
```

Важно отметить, что только двум полям, Esp0 и Ss0, присвоены значения. Это связано с тем, что Windows никогда не прибегает к аппаратному переключению задач, кроме случаев с системными прерываниями, описанных ранее. Таким образом, основным назначением данного конкретного TSS является загрузка подходящего стека ядра во время обработки аппаратных прерываний.

Как вы узнаете из раздела «Диспетчеризация системных прерываний», в системах, не подверженных архитектурной уязвимости Meltdown в процессоре, данный указатель стека будет ссылаться на стек ядра текущего потока (согласно структуре KTHREAD, описанной в главе 5 тома 1). В системах, где такая уязвимость есть, он будет указывать на переходный стек из зоны дескриптора процессора (Processor Descriptor Area, PDA). В то же время сегмент стека всегда в селекторе 10h, или KGDT_R0_DATA.

Как говорилось ранее, еще один TSS используется для исключения машинных проверок (#MC). На него можно посмотреть с помощью команды dg:

```
kd> dg a0 a0
                                  P Si Gr Pr Lo
Sel    Base     Limit     Type    l ze an es ng Flags
---- -------- -------- ---------- - -- -- -- -- --------
00A0 81170590 00000067 TSS32 Avl  0 Nb By P  Nl 00000089
```

Однако на этот раз вместо dx мы используем команду .tss, которая отформатирует некоторые поля из KTSS и отобразит задачу так, будто это поток, исполняемый в текущий момент. В данном случае входным параметром будет селектор задач (A0h):

```
kd> .tss a0

eax=00000000 ebx=00000000 ecx=00000000 edx=00000000 esi=00000000 edi=00000000
eip=81e1a718 esp=820f5470 ebp=00000000 iopl=0       nv up di  pl nz na po nc
cs=0008  ss=0010  ds=0023  es=0023  fs=0030  gs=0000             efl=00000000

hal!HalpMcaExceptionHandlerWrapper:
81e1a718 fa              cli
```

Обратите внимание на то, что сегментные регистры настраиваются так, как было описано в разделе про отложенную загрузку сегментов, а программный счетчик (EIP) ссылается на обработчик для #MC. Кроме того, стек настроен на безопасный стек из кода ядра, который не должен быть затронут повреждениями памяти. Наконец, пусть этого и не видно по команде `.tss`, CR3 указывает на системный каталог страниц. В разделе «Диспетчеризация системных прерываний» мы вернемся к данному TSS при рассмотрении команды `!idt`.

ЭКСПЕРИМЕНТ. Просмотр TSS и IST в системе x64

К сожалению, в системах x64 команда `dg` содержит баг, не позволяющий корректно отображать 64-битные базовые адреса сегментов, в силу чего для получения базового адреса сегмента TSS (по 40h) надо выгрузить как бы два сегмента и совместить старший, средний и младший байты:

```
0: kd> dg 40 48
                                                    P  Si  Gr  Pr  Lo
Sel        Base              Limit          Type    l ze an es ng Flags
---- ----------------- ----------------- ---------- - -- -- -- -- --------
0040 00000000`7074d000 00000000`00000067 TSS32 Busy 0 Nb By P  Nl 0000008b
0048 00000000`0000ffff 00000000`0000f802 <Reserved> 0 Nb By Np Nl 00000000
```

Таким образом, в приведенном примере KTSS64 оказывается по адресу `0xFFFFF8027074D000`. Есть и другой способ его получить: KPCR каждого процессора содержит поле `TssBase`, где также хранится указатель на KTSS64:

```
0: kd> dx @$pcr->TssBase
@$pcr->TssBase                : 0xfffff8027074d000 [Type: _KTSS64 *]
    [+0x000] Reserved0        : 0x0 [Type: unsigned long]
    [+0x004] Rsp0             : 0xfffff80270757c90 [Type: unsigned __int64]
```

Заметьте, что виртуальный адрес совпадает с тем, который можно увидеть в GDT. А также что все поля, кроме `RSP0`, обнулены. Заполненное поле, как и в системе x86, содержит либо адрес стека ядра для текущего потока (для систем без аппаратной уязвимости Meltdown), либо адрес переходного стека и зоны дескриптора процессора (PDA).

В системе, где проводился эксперимент, использовался процессор Intel десятого поколения, следовательно, `RSP0` указывает на текущий стек ядра:

```
0: kd> dx @$thread->Tcb.InitialStack
@$thread->Tcb.InitialStack : 0xfffff80270757c90 [Type: void *]
```

Наконец, взглянув на таблицу стеков прерываний, мы можем наблюдать различные стеки, связанные с системными прерываниями #DF, #MC, #DB и NMI.

В разделе «Диспетчеризация системных прерываний» мы рассмотрим, как таблица обработки прерываний (IDT) ссылается на них:

```
0: kd> dx @$pcr->TssBase->Ist
@$pcr->TssBase->Ist     [Type: unsigned __int64 [8]]
    [0]              : 0x0 [Type: unsigned __int64]
    [1]              : 0xfffff80270768000 [Type: unsigned __int64]
    [2]              : 0xfffff8027076c000 [Type: unsigned __int64]
    [3]              : 0xfffff8027076a000 [Type: unsigned __int64]
    [4]              : 0xfffff8027076e000 [Type: unsigned __int64]
```

Теперь, когда связь между уровнем кольца, исполнением кода и некоторыми ключевыми сегментами GDT прояснена, можно взглянуть на реальные переходы между различными сегментами кода и их уровнями кольца в разделе о диспетчеризации системных прерываний.

Прежде чем перейти к этой теме, проанализируем, как меняется конфигурация TSS в системах, уязвимых к атакам с аппаратной стороны типа Meltdown.

АППАРАТНЫЕ УЯЗВИМОСТИ К АТАКАМ ПО СТОРОННИМ КАНАЛАМ

Современные микропроцессоры способны вычислять и перемещать данные между своими внутренними регистрами очень быстро (счет идет на пикосекунды). Но сами эти регистры — очень ограниченный ресурс. В связи с этим операционной системе и приложениям приходится постоянно требовать от процессора перемещать данные из регистров в память и обратно. Существуют различные типы памяти, к которым он может обратиться. Память, находящаяся на самом процессоре и доступная непосредственно из блока исполнения, называется «кэш» и известна своей скоростью и дороговизной. Память, доступная по внешней шине, обычно называется RAM (Random Access Memory) и характеризуется низкой скоростью, малой стоимостью и большим объемом. Близость памяти к процессору помещает ее в некоторую иерархию, положение в которой зависит от скорости и физических размеров (чем ближе память к процессору, тем она быстрее и миниатюрнее). Как показано на рис. 8.2, в современных компьютерах процессоры часто имеют три уровня быстрой кэш-памяти, доступной каждому физическому ядру: L1, L2 и L3. Первые два ближе всех, и у каждого ядра процессора они свои. Кэш L3 дальше всех и всегда общий для всех ядер (заметим, что у встраиваемых процессоров кэш L3, как правило, отсутствует).

Одной из важнейших характеристик кэша является время доступа к нему, сравнимое с временем обращения к регистрам процессора (однако еще медленнее). Тем не менее обращение к основной памяти длится в сотни раз дольше. Это значит, что, если процессор будет исполнять все инструкции по порядку, он столкнется с большими паузами, вызванными попытками достучаться туда. Для решения этой проблемы в современных архитектурах прибегают к различным стратегиям. Исторически они привели к изобретению атак по сторонним каналам (также известных как спекулятивные атаки), очень эффективных при преодолении средств защиты пользовательских систем.

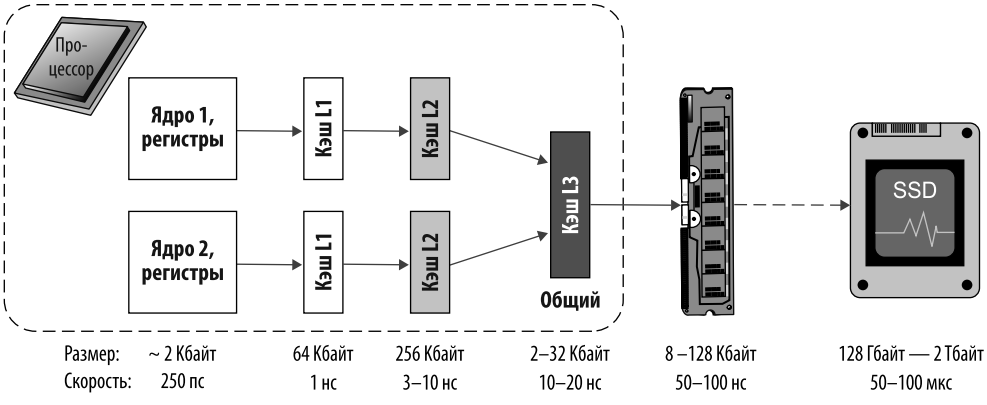


Рис. 8.2. Кэши и память современных процессоров, их средние размеры и время отклика

Чтобы правильно описать природу аппаратных атак по сторонним каналам и способы, которыми Windows борется с ними, следует рассмотреть несколько основных принципов работы внутренних механизмов процессора.

Внеочередное выполнение команд

Современный микропроцессор исполняет машинный код с помощью своего вычислительного конвейера. Тот включает в себя множество стадий, в числе которых — прочтение команды, декодирование, переименование и распределение регистров, переупорядочение, исполнение и фиксация результата выполнения. Когда процессору требуется обойти проблему медленной памяти, типичной стратегией является придание исполнительному устройству способности выполнять команды не по порядку, а в зависимости от готовности необходимых ресурсов. Таким образом, процессор обрабатывает код нелинейно, что позволяет максимально продуктивно задействовать все исполнительные устройства его ядра. Современный процессор способен исполнять сотни команд спекулятивно, пока однозначно не наступит момент, когда они потребуются, а их результат будет зафиксирован.

Одну из проблем описанного подхода создают команды условного перехода. Вычисление условия определяет два возможных пути дальнейшего продвижения по коду. Какой из них верен, зависит от команд, исполненных ранее. Когда этим командам требуется доступ к медленной памяти, могут возникать простои. В таком случае исполнитель ожидает, пока результат выполнения команд, определяющих условие, не будет зафиксирован (а именно, ожидание шины памяти, пока та обеспечивает доступ), чтобы потом продолжить внеочередное выполнение команд по корректному пути. Похожая проблема возникает при непрямом переходе. В этом случае исполнительное устройство не в курсе того, в какое место в коде произойдет переход (обычно это прыжок или вызов), поскольку целевой адрес еще нужно извлечь из основной памяти. В такой ситуации под *спекулятивным выполнением* подразумевается, что конвейер процессора декодирует и исполняет по нескольку команд параллельно, или, иными словами, вне очереди. Тем не менее результаты их исполнения не помещаются в регистры процессора, а запись данных в память откладывается по тех пор, пока условие перехода не будет окончательно вычислено.

Блок предсказания переходов

Откуда процессор узнает, код с какой ветви (перехода) должен исполняться, до того, как условие перехода будет полностью вычислено? (Та же проблема и с непрямыми переходами, целевой адрес которых заранее не известен.) Решение реализуется двумя компонентами в составе процессора: блоком предсказания переходов и блоком предсказания целей переходов.

Блок предсказания переходов (branch prediction) — это сложная цифровая схема в составе процессора, которая старается предугадать, какой вариант перехода будет выбран, до того, как это станет однозначно известно. Подобно ему, *блок предсказания цели переходов* — это устройство, задачей которого является предсказание адреса перехода прежде, чем тот станет известен. Хотя аппаратная реализация этих компонентов может серьезно различаться в зависимости от производителя процессора, оба они имеют доступ к служебному кэшу, который называется *буфером предсказания переходов* (Branch Target Buffer, BTB). В нем сохраняются адреса переходов или информация о результатах исходных переходов, выполненных ранее, с использованием хеш-меток, получаемых с помощью индексирующего алгоритма, подобно тому как кэш формирует метки (речь об этом пойдет в следующем разделе). Адрес перехода помещается в буфер после первого его выполнения. Как правило, в таком случае вычислительный конвейер приостанавливается, пока процессор дожидается, когда условие или адрес перехода будут получены из основной памяти. При следующем выполнении того же перехода адрес из буфера будет использован для помещения предсказанной цели в конвейер. На рис. 8.3 показана простая схема типового блока предсказания переходов.

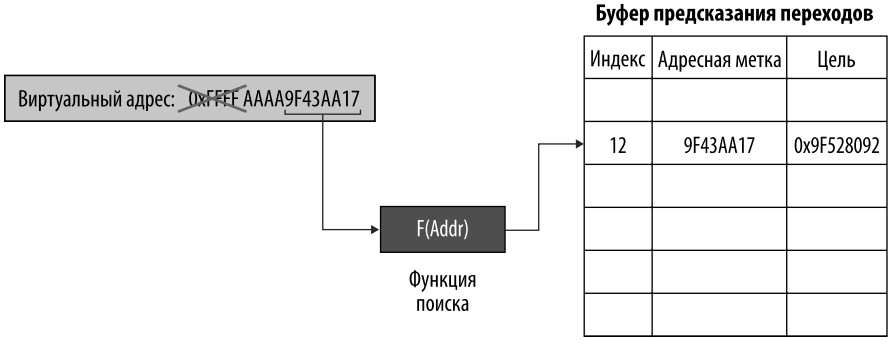


Рис. 8.3. Схема типового блока предсказания переходов процессора

В случае неправильного предсказания, когда неверный переход уже выполнен спекулятивно, вычислительный конвейер очищается, а результаты такого выполнения отменяются. В конвейер загружается код с альтернативного пути, а выполнение начинается с верного перехода. Такая ситуация называется *ошибочным предсказанием ветви* (branch misprediction). Общее число затраченных циклов процессора не превышает ожидания вычисления условия или получения адреса непрямого перехода при очередном выполнении. Однако спекулятивное выполнение все еще может страдать от различных побочных эффектов, таких как загрязнение строк кэша процессора.

К несчастью, некоторые побочные эффекты могут быть использованы со злым умыслом, что ставит под угрозу безопасность системы в целом.

Кэш процессора

Как говорилось ранее, кэшем процессора называется быстрая память, используемая для ускорения чтения/записи данных или кода. Информация перемещается между памятью и кэшем блоками фиксированной длины (обычно по 64 или 128 байт), которые называются *строками (блоками) кэша* (lines, cache blocks). Когда строка копируется из памяти в кэш, создается запись кэша. Она включает в себя скопированные данные и метку, идентифицирующую запрошенную область памяти. В отличие от буфера предсказания целей переходов кэш всегда индексируется по физическим адресам (в ином случае было бы сложно отслеживать множество типов отображения и изменения адресных пространств). С точки зрения кэша физический адрес делится на несколько частей. Старшие биты обычно служат меткой, а младшие идентифицируют саму строку и смещение до ее начала. Задачей метки является уникальная идентификация того, к какому блоку памяти принадлежит блок кэша, как показано на рис. 8.4.

Рис. 8.4. Типовой 48-битный одноканальный кэш процессора

Всякий раз, когда процессор выполняет чтение из какой-либо области памяти или запись в нее, в первую очередь проверяется соответствующая запись в кэше (в рамках всех строк кэша, где могут быть данные с целевого адреса; разные кэши действуют по-разному, о чем пойдет речь далее в этом разделе). Если процессор обнаружит содержимое запрошенной памяти в кэше, происходит *попадание в кэш*. В таком случае в найденной строке немедленно производится чтение/запись данных. Иначе же наблюдается *промах кэша*. В таком случае процессор создает в кэше новую запись и копирует туда данные из основной памяти, лишь затем давая к ним доступ.

На рис. 8.4 представлен одноканальный процессорный кэш, он предназначается максимум для 48 бит виртуального адресного пространства. В данном примере процессор считывает 48 байт по виртуальному адресу 0x19F566030. Содержимое памяти сначала копируется из основной памяти в блок кэша 0x60. Он заполнен под завязку, но запрошенные данные начинаются по смещению 0x30.

Данный типовой кэш содержит всего 256 блоков по 256 байт, так что блок 0x60 может соответствовать сразу нескольким физическим адресам. Метка (0x19F56) обеспечивает уникальную идентификацию физического расположения данных в основной памяти.

Когда от процессора требуется записать в память по указанному адресу что-то новое, он поступает похожим образом. Сначала обновляется строка (-и) кэша, к которой целевой адрес имеет отношение. В какой-то момент процессор отразит новые данные в оперативной памяти. Конкретное время будет зависеть от способа кэширования (обратная/сквозная запись, без кэша и т. д.), используемого для искомой страницы памяти. (Следует отметить, что это важнейшая проблема в работе многопроцессорных систем: чтобы избежать ситуаций, когда другой процессор пытается работать с данными, устаревшими в результате обновления блока кэша основным процессором, необходимо иметь некий протокол согласованности кэша. Существует ряд алгоритмов обеспечения согласованности кэша, но в книге они не рассматриваются.)

Когда случается промах кэша и требуется место для новой записи, процессору иногда приходится освобождать одну из ранее занятых. Алгоритм, который кэш использует для определения того, что освобождать (и выбора, где записать новые данные), называется *политикой размещения*. Если согласно этой политике всегда происходит замена только одного блока для конкретного виртуального адреса, такой кэш называется кэшем *прямого отображения* (кэш на рис. 8.4 является одноканальным прямого отображения). Если же кэш вправе выбрать любую запись (с тем же номером блока кэша) для записи новых данных, он называется *полностью ассоциативным*. Зачастую политика размещения в кэше подразумевает компромисс, при котором каждая запись в основной памяти может отразиться в любом из *N* мест в кэше, и тогда кэш называется *N*-канальным секторно-ассоциативным. На рис. 8.5 представлен четырехканальный секторно-ассоциативный кэш. В нем можно разместить до четырех разных физических адресов, индексируемых в один блок (с разными метками) в четырех разных секторах кэша.

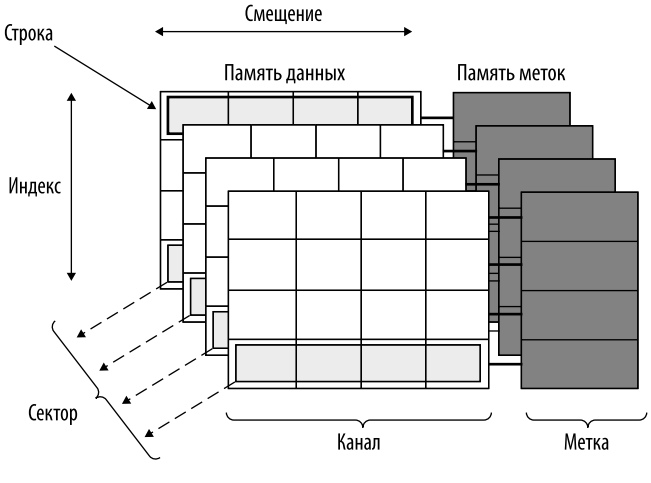


Рис. 8.5. Четырехканальный секторно-ассоциативный кэш

Атаки по сторонним каналам

Как говорилось ранее, исполнительные устройства современных процессоров не фиксируют результаты вычислений до тех пор, пока команда не считается окончательно выполненной. Это приводит к тому, что, несмотря на возможность внеочередного выполнения сразу множества инструкций без заметных архитектурных эффектов с точки зрения регистров процессора и памяти, проявляются побочные микроархитектурные эффекты, особенно для процессорного кэша. В конце 2017 года были выявлены новые способы атак на процессоры через их механизмы внеочередного исполнения и блоки предсказания переходов. Дело в том, что микроархитектурные побочные эффекты можно отследить, даже не имея возможности напрямую влиять на них из любого программного кода.

Два наиболее разрушительных и эффективных способа таких атак известны как Meltdown и Spectre.

Meltdown

Meltdown (впоследствии ставшая известной как Rogue Data Cache Load (загрузка в кэш мошеннических данных), RDCL) позволяла злонамеренному процессу пользовательского режима получать доступ ко всей памяти, даже к памяти ядра, что в обычной ситуации не позволяется. При атаке были использованы механизмы внеочередного выполнения в процессоре и разница во времени между доступом к памяти и проверкой привилегий при выполнении команд этого доступа.

В рамках реализации Meltdown процесс злоумышленника сначала вызывает полное очищение кэша (команды для этого доступны из пользовательского режима). Затем он делает нелегальный запрос доступа к памяти ядра, за которым сразу идет набор команд, заполняющих кэш предопределенным образом, задействуя *массив в роли зонда*. Доступа к памяти ядра процесс не получает, а процессор, в свою очередь, выдает исключение, которое отлавливается приложением. В иной ситуации все закончилось бы завершением процесса. Однако в условиях внеочередного выполнения процессор уже параллельно выполнил команды, следовавшие за отклоненным запросом к памяти ядра (но не зафиксировал, из-за чего никаких архитектурных эффектов не проявится ни в оперативной памяти, ни в регистрах процессора). Те команды, в свою очередь, уже заполнили кэш содержимым памяти ядра, к которому их не должны были допустить.

Затем приложение злоумышленника зондирует весь кэш, замеряя время, необходимое для доступа к каждой странице массива, использованного для заполнения блока кэша. Если время доступа не превышает некоторого предела, то данные присутствуют в строке кэша. Это позволяет злоумышленнику в точности воспроизвести байт, прочитанный из памяти ядра. Рисунок 8.6, заимствованный из оригинальной публикации о Meltdown (доступна по адресу https://meltdownattack.com/), демонстрирует время доступа с помощью массива-зонда размером 1 Мбайт (256 страниц по 4 Кбайт).

На рис. 8.6 видно, что время доступа ко всем страницам, кроме одной, примерно одинаковое. Если предположить, что засекреченные данные могут быть прочитаны по 1 байт за раз, а 1 байт может представить лишь 256 значений, то, зная точно, какая страница в массиве вызвала попадание в кэш, злоумышленник может определить байт, находящийся в памяти ядра.

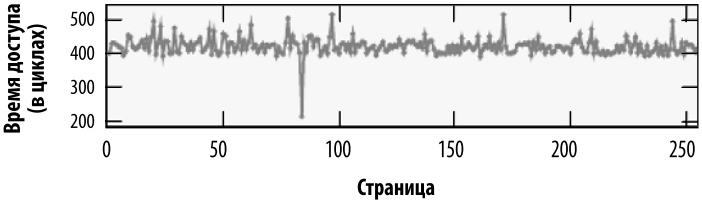


Рис. 8.6. Расход процессорного времени при доступе к массиву-зонду в 1 Мбайт

Spectre

Принцип атаки Spectre похож на принцип Meltdown и опирается на уязвимости внеочередного выполнения, описанные ранее. Однако в этом случае злоумышленник атакует другие важные компоненты процессора, а именно блок предсказания переходов и буфер предсказания целей переходов. Изначально были изучены два варианта Spectre. Оба можно свести к трем шагам.

1. В подготовительной фазе от лица низкопривилегированного процесса, который контролирует злоумышленник, исполняется множество повторяющихся операций с целью навязать блоку предсказания переходов неправильное обучение. В результате процессор приучается исполнять заданную ветку условного перехода либо непрямой переход по хорошо известному адресу.
2. Во второй фазе злоумышленник принуждает атакуемое привилегированное приложение (или такой же процесс) спекулятивно выполнить команды, находящиеся по адресу неверно предсказанного перехода. Зачастую эти команды переносят конфиденциальную информацию из контекста жертвы в микроархитектурный канал (обычно это процессорный кэш).
3. В финальной фазе снова от лица низкопривилегированного процесса злоумышленник воспроизводит оказавшиеся в кэше (микроархитектурном канале) чужие данные путем зондирования его тем же способом, что и при атаке Meltdown. Так в его руки попадает секретная информация из закрытого привилегиями жертвы адресного пространства.

Первый вариант атаки Spectre позволяет воспроизвести секреты, сокрытые в адресном пространстве процесса жертвы (которое может быть как раздельным, так и общим с адресным пространством процесса злоумышленника), принудив блок предсказания переходов дать процессору спекулятивно выполнить код по неверной ветви условного перехода. Обычно это оказывается часть функции, которая проверяет границы перед тем, как получить доступ к несекретным данным в буфере памяти. Если этот буфер находится рядом с секретными данными, а злоумышленник может контролировать смещение, передаваемое в условие перехода, ею можно натренировать блок предсказания корректными значениями, которые проходят проверку и указывают процессору верный путь.

Затем злоумышленник подготавливает кэш процессора нужным образом (в частности, так, чтобы размер буфера памяти, используемый для проверки, отсутствовал в кэше) и передает в функцию, выполняющую проверку границ,

неверное смещение. Блок предсказания переходов уже обучен всегда следовать одному верному выбору. Однако в этот раз путь будет неверен (нужно было идти по другому маршруту). Команды доступа к буферу будут выполнены спекулятивно, но результат прочитан за пределами допустимых границ, из секретных данных. Наконец, злоумышленник считывает их к себе путем полного зондирования кэша, как при атаке Meltdown.

Второй вариант Spectre злоупотребляет поведением буфера предсказания переходов процессора — злоумышленник может искажать непрямые переходы. Неверная цель непрямого перехода может быть использована для прочтения любой памяти процесса-жертвы (или ядра ОС) из зловредного контекста. Как показано на рис. 8.7, так злоумышленник провоцирует заполнение буфера целей переходов опасными адресами. В конечном счете процессор решается спекулятивно выполнить команды по переходу, нужному для атаки. В рамках адресного пространства жертвы адрес перехода будет указывать на гаджет. *Гаджет* — это группа команд, которая обращается к закрытым данным, отчего те предсказуемо попадают в кэш (злоумышленнику потребуется косвенно контролировать один или несколько регистров процессора в контексте жертвы, что зачастую возможно в тех случаях, когда API принимает непроверенные данные).

После того как блок предсказания переходов будет натренирован, злоумышленник очищает кэш процессора, а затем обращается к службе (интерфейсу), предоставляемой высокопривилегированной целью — процессом или ядром ОС. Код, в котором реализуется служба, должен иметь непрямые переходы, похожие на те, что показывал процесс злоумышленника. Буфер целей перехода провоцирует спекулятивное выполнение гаджета по подставному адресу. Тем самым, как и при первом способе или атаке Meltdown, появляются побочные микроархитектурные эффекты в кэше процессора, которые позволяют прочесть его из низкопривилегированного контекста.

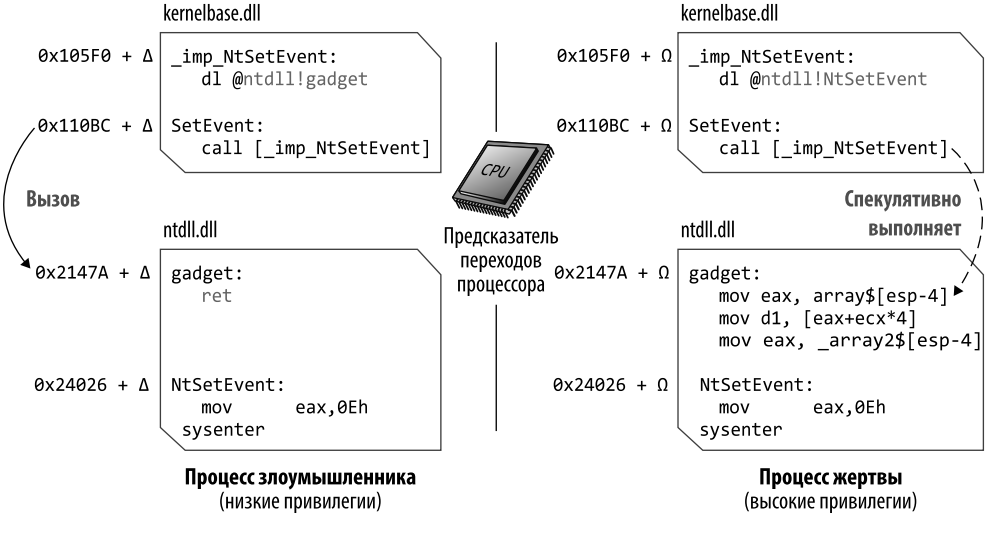


Рис. 8.7. Схема атаки Spectre, вариант 2

Другие атаки по сторонним каналам

После того как впервые были обнародованы методы атак Meltdown и Spectre, обнаружились еще несколько похожих техник. Пусть они были не настолько опасными и эффективными, важно как минимум понимать общую концепцию методов, используемых при подобных атаках.

Обход спекулятивного контекста (Speculative Store Bypass) возможен ввиду оптимизации со стороны процессора, позволяющей поместить команду, которую процессор признал независимой от прежнего контекста, на спекулятивное выполнение, пока тот контекст не обновился. Если предсказание перехода неверно, это может привести к загрузке устаревших данных, возможно секретных. Эти данные потом могут быть переданы другим спекулятивно выполняемым командам. Их действия могут обеспечивать доступ к памяти и создавать микроархитектурные побочные эффекты (обычно в кэше процессора). Злоумышленник может их оценить и завладеть данными.

Предзнаменование (Foreshadow, известно также как *L1TF*) — более серьезная атака, изначально проектировавшаяся для кражи данных из анклавов (изолированных областей данных, защищенных по технологии Intel SGX), но затем обобщенная для исполнения в рамках обычных программ без привилегий. Атакой Foreshadow использовались два аппаратных недостатка реализации спекулятивного исполнения в современных процессорах.

- Спекулятивное исполнение в отношении недоступной виртуальной памяти. В данной ситуации, когда процессор требует доступ к данным, находящимся по виртуальному адресу, указанному в записи таблицы страниц (Page Table Entry, PTE), но в котором сброшен бит присутствия (а значит, адрес невалиден), при корректной работе выдается исключение. Однако, если запись поддерживает преобразование действительного адреса, процессор может спекулятивно исполнить команды, зависящие от прочитанных данных. Как и в прочих подобных атаках, результат выполнения таких команд не зафиксирован процессором, но создает отслеживаемые побочные эффекты. В данном случае приложение пользовательского режима сможет прочесть защищенные данные из памяти ядра. Даже хуже, в особых случаях приложение сможет считать данные, принадлежащие другой виртуальной машине. Когда процессор выполняет преобразование гостевого физического адреса (Guest Physical Address, GPA) и встречает несуществующую запись в таблице преобразования адресов второго уровня (Second Level Adrress Translation Table, SLAT), могут возникать похожие побочные эффекты. (Больше информации о SLAT, GPA и механизмах преобразования сказано в главе 5 тома 1 и главе 9 данной книги).
- Спекулятивное исполнение на логических процессорах (с технологией Hyper-Threading) ядра. Современные процессоры могут иметь по несколько вычислительных конвейеров на одно физическое ядро, которые способны вне очереди выполнять множество потоков команд на общем исполнительном устройстве (это называется *симметричной многопоточностью* (Symmetric Multithreading) и рассматривается в главе 9). В таком ядре могут существовать два логических процессора, которые делят между собой общий кэш. Это значит, что пока один логический процессор выполняет какой-то код в привилегированном контексте,

его коллега может прочесть в кэше последствия такого выполнения. А это создает серьезные угрозы общей безопасности системы. Подобно первому варианту Foreshadow, логический процессор, исполняя вредоносный код в низкопривилегированном контексте, может украсть данные даже с другой, защищенной виртуальной машины, просто дожидаясь ее кода, который был назначен на исполнение родственным логическим процессором. Данный вариант Foreshadow относится к уязвимостям 4-й группы.

Микроархитектурные побочные эффекты не всегда задействуются для атак на кэш процессора. Модели от Intel применяют и другие промежуточные высокоскоростные буферы с целью ускорения доступа к кэшированной и некэшированной памяти, переупорядочивания микрокоманд (такие буферы в книге не рассматриваются). Атаки, использующие микроархитектурное зондирование данных (Microarchitectural Data Sampling, MDS), позволяют давать доступ к следующим структурам:

- **к буферам хранилища** — выполняя операции с контекстом, процессор записывает данные во внутреннюю временную микроархитектурную структуру, называемую буфером контекста, что позволяет ему выполнять команды еще до того, как данные будут отражены в кэше или в основной памяти (в случае доступа без кэша). Когда команда загрузки считывает данные с того же адреса в памяти, который применялся для прежнего контекста, процессор может выдать кому-то данные напрямую из буфера;
- **к буферам заполнения** — так называется внутренняя структура в процессоре, используемая для сбора или записи данных в случае промаха данных кэша первого уровня, а также при вводе/выводе и особых операциях с регистрами. Буферы заполнения служат промежуточным звеном между кэшем процессора и его механизмами внеочередного выполнения. В них могут оставаться данные от прошлых запросов к памяти, которые могут быть спекулятивно отданы команде загрузки;
- **к загрузочным портам** — к временным служебным структурам процессора, используемым для выполнения команд загрузки из памяти или портов ввода/вывода.

Микроархитектурные буферы обычно привязаны к конкретному ядру процессора и являются общими для SMT-потоков. Такой подход подразумевает, что, даже если атаки на подобные структуры сложно осуществить надежно, спекулятивное извлечение данных из них между SMT-потоками потенциально возможно, пусть и при особых условиях.

В конечном счете исходом всех рассмотренных аппаратных уязвимостей по сторонним каналам является то, что конфиденциальные данные могут быть похищены из адресного пространства жертвы. В системах Windows реализован ряд мер противостояния Spectre, Meltdown и почти всем подобным атакам.

МЕРЫ БОРЬБЫ С АТАКАМИ ПО СТОРОННИМ КАНАЛАМ В WINDOWS

В данном разделе мы разберемся, как в Windows реализуются различные меры противодействия атакам по сторонним каналам. Против некоторых техник они принимаются еще производителями процессоров посредством обновлений микрокода. Однако не все они активны постоянно, поэтому некоторые механизмы защиты должны быть включены программой (ядром Windows).

Затенение KVA

Затенение виртуальных адресов ядра (Kernel Virtual Address, KVA), известное также как затенение KVA (аналогично изоляции таблицы страниц ядра (Kernel Page Table Isolation, KPTI) в системах Linux), предотвращает исполнение атаки Meltdown введением четкого разделения между таблицами страниц ядра и пользовательского режима. Спекулятивное выполнение позволяет процессору исказить данные ядра при отсутствии должного уровня привилегий, но для этого необходимо, чтобы в таблице страниц, применяемой для преобразования целевой страницы ядра, находился валидный номер кадра страницы. Память ядра, которая выступает целью атаки Meltdown, обычно представлена допустимой записью в системной таблице страниц, что указывает на необходимость привилегий супервизора для доступа (таблицы страниц и преобразование виртуальных адресов рассматривались в главе 5 тома 1). Когда затенение KVA активно, система формирует и использует две высокоуровневые таблицы страниц для каждого процесса.

- Таблицы страниц ядра размечают адресное пространство целиком, включая системные и пользовательские страницы. В Windows последние отмечаются как неисполняемые, чтобы не позволить коду ядра выполнять содержимое памяти, выделенной в пользовательском режиме (эффект, схожий с тем, что реализуется аппаратной функцией SMEP).
- Пользовательские таблицы страниц (также называемые теневыми) размечают только пользовательские страницы и минимальный набор системных, не содержащих никаких системных данных. Последние применяются для предоставления минимального функционала с целью переключения таблиц страниц и стеков ядра, обработки системных прерываний, системных вызовов и пр. Этот набор системных страниц называется *транзитным* адресным пространством.

В транзитном адресном пространстве ядро NT обычно размещает структуру данных, входящую в блок управления процессором (Processor Region Control Block, PRCB) и известную как KPROCESSOR_DESCRIPTOR_AREA. В нее входят данные, которые должны присутствовать в пользовательских (теневых) и системных таблицах страниц, в частности TSS процессора, GDT и копия базового адреса сегмента GS режима ядра. Кроме того, в транзитном адресном пространстве находятся все обработчики системных прерываний из раздела .KVASCODE образа ядра NT.

Системы, где активно затенение KVA, запускают непривилегированные потоки пользовательского режима (то есть запускаемые без привилегий администратора) в рамках процессов, в которых не отображаются никакие страницы ядра, где могут быть секретные данные. Это делает атаку Meltdown бесполезной: системные страницы не считаются валидными с точки зрения пользовательской страницы, а значит, никакие спекуляции с процессором с целью доступа к ним невозможны. Когда пользовательский процесс делает системный вызов или срабатывает прерывание во время выполнения пользовательского кода, процессор строит в *транзитном стеке* кадр (или фрейм) ловушки (системного прерывания) (trap frame), представляющий собой структуру данных, которая, как отмечалось ранее, видна на страницах обоих типов. Затем процессор выполняет теневой код обработчика системного прерывания, который реагирует на вызов или прерывание. Этот код обычно переключается на системные таблицы страниц, копирует кадр ловушки

в стек ядра, а затем вызывает оригинальный обработчик. (Здесь подразумевается, что необходим корректно реализованный четкий алгоритм очистки TLB от устаревших записей. Подобный алгоритм описывается далее в этом разделе.) Оригинальный обработчик выполняется в условиях полной видимости всего адресного пространства.

Инициализация

Ядро NT определяет, уязвим ли процессор к атакам Meltdown, на старте фазы инициализации –1, после того как вычислены флаги функциональности процессора, в рамках процедуры `KiDetectKvaLeakage`. Последняя извлекает информацию о процессоре и устанавливает внутренний флаг `KiKvaLeakage` для всех процессоров Intel, кроме относящихся к категории Atom, которые не поддерживают внеочередное выполнение.

В таком случае затенение KVA активируется процедурой `KiEnableKvaShadowing`, которая готовит сегмент состояния задач (Task State Segment, TSS) для процессора и транзитные стеки размером по 512 байт. В этом сегменте стеки RSP0 (ядро) и IST настраивают на расположение правильных транзитных стеков. Те, в свою очередь, готовят путем добавления небольшой структуры данных под названием KIST_BASE_FRAME в начало стека. Она позволяет установить связь между транзитным стеком и его нетранзитным сородичем в ядре, который доступен, только когда таблицы страниц уже переключены (рис. 8.8). Однако эта структура не требуется для обычных, не IST-стеков ядра. Операционная система получает все необходимые данные из PRCB процессора. Каждый поток получает корректный стек ядра. Планировщик отмечает стек ядра как активный, связывая его с PRCB процессора, когда выбирает новый поток для выполнения. Это является ключевым отличием от IST-стеков, которые существуют по одному на процессор.

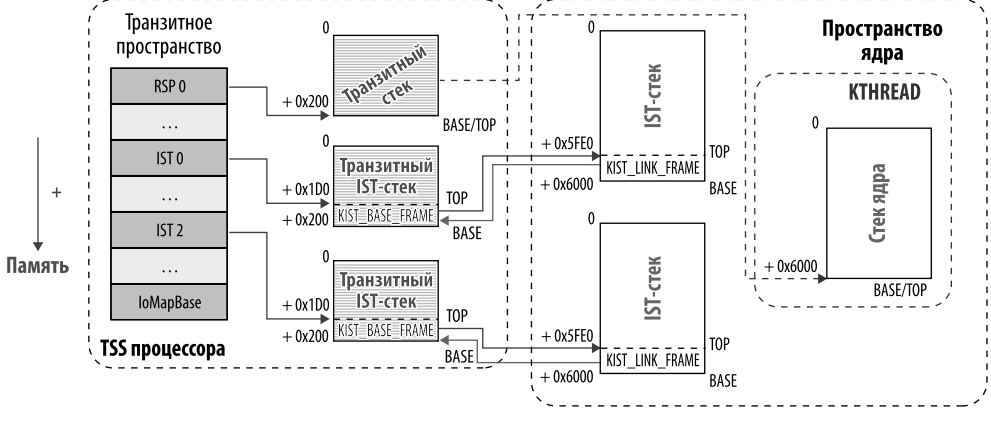


Рис. 8.8. Структура сегмента состояния задач процессора при активном затенении KVA

Процедура `KiEnableKvaShadowing` отвечает также за важный процесс — выбор алгоритма очистки TLB (рассматривается далее в этом разделе). Результат выбора (для глобальных записей или конкретных PCID) сохраняется в глобальной переменной `KiKvaShadowMode`. Наконец, для процессоров, не задействованных в загрузке системы, вызывается процедура `KiShadowProcessorAllocation`, которая для каждого

процессора отображает общие структуры данных для теневых таблиц страниц. Для BSP-процессора отображение выполняется позже, на фазе 1, после того как процесс SYSTEM и его теневые таблицы будут созданы, а IRQL переведен на пассивный уровень (Passive). Только в этом случае теневые обработчики системных прерываний отображаются в пользовательских таблицах страниц (они глобальны и не привязаны к отдельным процессорам).

Теневые таблицы страниц

Теневые (или пользовательские) таблицы страниц размещаются диспетчером памяти с помощью процедуры `MiAllocateProcessShadow` в тот момент, когда оформляется адресное пространство процесса. Изначально у нового процесса теневые таблицы страниц пусты. Затем диспетчер памяти копирует все высокоуровневые теневые таблицы страниц ядра из процесса SYSTEM в теневые таблицы нового процесса.

Это позволяет ОС быстро отобразить для нового процесса все транзитное адресное пространство, которое принадлежит ядру и применяется для всех процессов пользовательского режима. У процесса SYSTEM теневые таблицы остаются пустыми. Как описывалось в предыдущем разделе, они будут заполнены процедурой `KiShadowProcessorAllocation`, которая задействует функции диспетчера памяти, чтобы отобразить там отдельные блоки данных и воспроизвести всю иерархию страниц.

Теневые таблицы страниц обновляются диспетчером памяти лишь в особых случаях. Только ядру разрешается изменять таблицы страниц процесса для отображения или сокрытия отдельных блоков памяти. Когда имеет место запрос на выделение или отображение новой памяти в адресном пространстве пользовательского процесса, может оказаться, что верхняя запись таблицы страниц для отдельно взятого адреса будет отсутствовать. В таком случае диспетчер памяти размещает все страницы из иерархии таблиц страниц и вносит новую верхнюю запись в таблицы страниц ядра. Однако если активно затенение KVA, этого недостаточно — диспетчеру памяти нужно сделать подобное и с теневой таблицей страниц. В ином случае после того, как обработчик системного прерывания корректно переключит таблицы перед возвратом в пользовательский режим, этот адрес там будет недоступен.

В сравнении с таблицами страниц ядра адреса ядра по-другому отображаются в транзитном адресном пространстве. Во избежание ложного разделения близких к блоку памяти адресов, которые требуется отобразить, диспетчер памяти всегда пересоздает отражение иерархии записей таблиц страниц, которыми нужно поделиться. Иными словами, каждый раз, когда ядру требуется отразить какие-то новые страницы в транзитном адресном пространстве процесса, ему *нужно* воссоздать отображение всех теневых таблиц страниц процесса (внутренняя процедура `MiCopyTopLevelMappings` реализует именно это действие).

Алгоритм сброса TLB

В архитектуре x86 переключение таблиц страниц обычно приводит к сбросу текущего состояния буфера быстрого преобразования адреса (translation look-aside buffer, TLB). Этот буфер находится в специальном кэше процессора и служит для быстрого преобразования виртуальных адресов, использующихся при выполнении кода или получении доступа к данным. Валидная запись в TLB позволяет процессору

не сверяться с цепочкой таблиц страниц, отчего операции выполняются быстрее. В системах с затенением KVA записи из TLB, преобразующие адреса ядра, не обязательно именно сбрасывать — в Windows адресное пространство ядра зачастую уникально и имеет отношение ко всем процессам. Intel и AMD предложили различные средства, позволяющие избежать сброса записей о ядре при каждом переключении таблиц страниц, такие как флаг глобальности/локальности и идентификаторы процесса-контекста (Process-Context Identifiers, PCIDs). Методологии по работе с TLB и его сбросе детально описаны в архитектурных руководствах Intel и AMD и в данной книге не рассматриваются.

С помощью новых возможностей процессоров операционная система способна выборочно очищать только пользовательские записи и не терять в производительности. Это явно недопустимо в ситуациях с затенением KVA, когда поток обязан переключить таблицы страниц при входе в код ядра или выходе из него. Там, где это доступно, Windows задействует алгоритм, способный специально очищать либо только системные, либо только пользовательские записи в TLB по мере необходимости. Так решаются следующие две задачи.

- При выполнении пользовательского кода в потоке в TLB никогда не будет валидных записей о памяти ядра. В ином случае это может быть задействовано злоумышленником для спекуляций, подобных атаке Meltdown, что позволило бы похитить секретные данные.
- При переключении таблиц страниц очищается лишь минимум записей в TLB. Это позволяет поддерживать потери производительности из-за затенения KVA на приемлемом уровне.

Алгоритм сброса TLB реализуется при трех основных сценариях: переключении контекста, входе в ловушку и выходе из ловушки. Он может выполняться в системах, как поддерживающих только флаг глобальности, так и дополняющих это PCID. В первом случае, в отличие от конфигураций без затенения KVA, все страницы ядра отмечены как неглобальные, а транзитные и пользовательские страницы — наоборот. Глобальные страницы не очищаются при переключении таблиц страниц (система меняет значение регистра CR3). Системы с поддержкой PCID отмечают страницы ядра PCID 2, а пользовательские — PCID 1. Флаги глобальности/локальности в таком случае не применяются.

Когда выполняемый в данный момент поток завершает свой квант времени, запускается переключение контекста. Когда ядро планирует выполнение для потока из адресного пространства другого процесса, алгоритм TLB обеспечивает полное удаление оттуда всех пользовательских страниц. (Это означает, что в системах с флагом глобальности требуется полный сброс TLB. Опять же пользовательские страницы считаются глобальными.) При выходе из ловушки ядра (когда ядро заканчивает выполнять свой код и возвращается в пользовательский режим) алгоритм обеспечивает удаление (или объявляет недействительными) всех записей ядра из TLB. Достигается это легко: на процессорах с флагом глобальности сама по себе перезагрузка таблиц страниц вынуждает процессор отменить все неглобальные страницы. В свою очередь, в системах с поддержкой PCID пользовательские таблицы перезагружаются с пользовательским же

PCID, из-за чего все устаревшие записи ядра в TLB автоматически считаются невалидными.

Данная стратегия позволяет на входе в ловушки ядра, что может происходить, когда во время выполнения системой пользовательского кода срабатывает прерывание или какой-то поток делает системный вызов, ничего не удалять из TLB. Схема алгоритма очистки TLB представлена в табл. 8.1.

Таблица 8.1. Стратегии очистки TLB в условиях затенения KVA

Вид конфигурации	Пользовательские страницы	Страницы ядра	Транзитные страницы
Без затенения KVA	Неглобальны	Глобальны	Отсутствуют
Затенение KVA, стратегия PCID	PCID 1, неглобальны	PCID 2, неглобальны	PCID 1, неглобальны
Затенение KVA, стратегия флага глобальности	Глобальны	Неглобальны	Глобальны

Аппаратный контроль непрямых переходов (IBRS, IBPB, STIBP, SSBD)

Производители процессоров разработали аппаратные меры противодействия различным атакам по сторонним каналам. Они предназначены для использования совместно с программными решениями. В основном эти меры входят в следующие механизмы контроля непрямых переходов, взаимодействие с которыми обычно доступно через флаг в регистрах, зависящих от конкретной модели процессора (MSR).

- **Ограничение спекулятивного выполнения непрямых переходов** (Indirect Branch Restricted Speculation, IBRS) полностью блокирует блок предсказания переходов (и сбрасывает буфер целей переходов) при переключении на другой контекст безопасности (между уровнями пользователя/ядра и VM Root с VM non-root). Когда ОС активирует IBRS после перехода в более привилегированный режим, предсказанные цели непрямых переходов не могут быть подконтрольны коду, выполняемому с более низкими привилегиями. Кроме того, в этом режиме предсказанные цели непрямых переходов не могут быть получены другим логическим процессором. Обычно ОС устанавливает флаг IBRS и не сбрасывает его до тех пор, пока не вернется к менее привилегированному контексту безопасности.

  Реализация IBRS зависит от производителя процессора: какие-то модели при переходе в этот режим полностью отключают буферы предсказания переходов согласно принципу подавления, а другие просто сбрасывают эти буферы, реализуя принцип сброса. В подобных процессорах работа механизма контроля мер IBRS очень похожа на работу IBPB, так что обычно в процессоре присутствует только IBRS.
- **Ограничение предсказания непрямых переходов** (Indirect Branch Predictor Barrier, IBPB) подразумевает сброс содержимого предсказателей переходов при

активации, тем самым ограждая предсказанные цели переходов от влияния со стороны кода, недавно выполнявшегося тем же логическим процессором.

- **Однопоточное предсказание непрямых переходов** (Single Thread Indirect Branch Predictors, STIBP) запрещает общее предсказание между логическими процессорами в рамках одного физического ядра процессора. Установка флага STIBP на логическом процессоре предотвращает доступ к предсказанным для него целям непрямых переходов со стороны кода, выполняемого (или ранее выполнявшегося) на другом логическом процессоре в том же ядре.
- **Запрет спекулятивного доступа к хранилищам** (Speculative Store Bypass Disable, SSBD) запрещает процессору спекулятивно выполнять загрузку чего-либо, пока не станут известны адреса всех ранее выделенных хранилищ. Таким образом, никакая операция загрузки не приведет к спекулятивному доступу к устаревшим значениям данных из прежнего экземпляра хранилища для того же логического процессора, что дает защиту от атаки спекулятивного доступа к хранилищу (Speculative Store Bypass) (описано ранее в подразделе «Атаки по сторонним каналам»).

Ядро NT применяет сложный алгоритм для определения состояния рассмотренных средств контроля непрямых переходов, которое обычно меняется в тех же ситуациях, когда наблюдается затенение KVA: переключение контекста, вход в ловушку и выход из ловушки. На совместимых системах код ядра всегда выполняется с активным IBRS (за исключением случаев, когда активно Retpoline). Если IBRS недоступно, но IBPB и STIBP поддерживаются, ядро выполняется с активным STIBP, из-за чего буферы предсказания переходов (с IBPB) очищаются при каждом входе в ловушку (таким образом, блок предсказания не может оказаться под влиянием кода, выполняемого в пользовательском режиме или родственным потоком, действующим в другом контексте безопасности). Если процессор поддерживает SSBD, при выполнении кода ядра эта функция активна всегда.

По соображениям производительности потоки пользовательского режима обычно выполняются без применения аппаратных мер ограничения спекуляций или лишь с активным STIBP (зависит от доступности сопряжения STIBP, что рассматривается в следующем разделе). Защита от спекулятивного доступа к хранилищам должна быть активирована вручную глобально или в отношении конкретных процессов. В конечном счете все меры предотвращения спекуляций могут быть настроены через глобальную переменную реестра Windows по адресу HKLM\System\CurrentControlSet\Control\Session Manager\Memory Management\FeatureSettings. Значение представлено в виде 32-битной маски, где каждый бит отвечает за отдельную настройку. В табл. 8.2 поясняются каждая из них и ее назначение.

Таблица 8.2. Настройки функций и их значения

Название	Значение	Назначение
FEATURE_SETTINGS_DISABLE_IBRS_EXCEPT_HVROOT	0x1	Запретить IBRS, кроме невложенных корневых разделов (по умолчанию для серверных машин)
FEATURE_SETTINGS_DISABLE_KVA_SHADOW	0x2	Принудительный отказ от затенения KVA

Название	Значение	Назначение
FEATURE_SETTINGS_DISABLE_IBRS	0x4	Запрет IBRS независимо от конфигурации компьютера
FEATURE_SETTINGS_SET_SSBD_ALWAYS	0x8	Всегда активировать SSBD в режимах ядра и пользователя
FEATURE_SETTINGS_SET_SSBD_IN_KERNEL	0x10	Активировать SSBD только в режиме ядра (код пользовательского режима остается уязвим для атак SSB)
FEATURE_SETTINGS_USER_STIBP_ALWAYS	0x20	Всегда использовать STIBP для потоков пользовательского режима вне зависимости от сопряжения STIBP
FEATURE_SETTINGS_DISABLE_USER_TO_USER	0x40	Блокировать принятую по умолчанию стратегию предотвращения спекуляций (лишь для систем AMD) и разрешить только меры «пользователь — пользователь». Когда этот флаг установлен, при работе в режиме ядра все механизмы контроля спекуляций отключаются
FEATURE_SETTINGS_DISABLE_STIBP_PAIRING	0x80	Всегда блокировать сопряжение STIBP
FEATURE_SETTINGS_DISABLE_RETPOLINE	0x100	Всегда блокировать Retpoline
FEATURE_SETTINGS_FORCE_ENABLE_RETPOLINE	0x200	Разрешать Retpoline независимо от поддержки процессором IBPB или IBRS (чтобы добиться эффективной защиты от Spectre v2, Retpoline необходим как минимум IBPB)
FEATURE_SETTINGS_DISABLE_IMPORT_LINKING	0x20000	Блокировать оптимизацию импорта независимо от Retpoline

Retpoline и оптимизация импорта

Действия аппаратных средств защиты накладывают серьезные ограничения на производительность системы хотя бы потому, что предсказатель переходов процессора ограничен или отключен. Это оказалось неприемлемо для видеоигр и приложений с критическими задачами, которые столкнулись с существенным падением быстродействия. Наибольший ущерб производительности стали приносить IBRS или IBPB, которые служат защитой от атаки Spectre. Справиться с первым ее вариантом можно было без всяких аппаратных средств благодаря командам установки барьера памяти. Хорошим примером является LFENCE, доступная в архитектуре x86. Данные команды принуждают процессор не выполнять спекулятивно никаких новых команд, пока барьер не будет снят. Лишь после этого (и после того, как все команды перед барьером зафиксируются) вычислительный конвейер процессора возобновит выполнение новых команд, в том числе внеочередное. Вторая версия Spectre все еще требовала активности аппаратных средств защиты, что подразумевало все потери производительности, вызываемые IBRS и IBPB.

Для решения данной проблемы инженеры Google разработали уникальную технику бинарной модификации под названием Retpoline. Последовательность Retpoline (рис. 8.9) позволяет непрямым переходам оказываться защищенными от спекулятивного выполнения. Вместо выполнения небезопасного непрямого вызова процессор переходит на блок контроля, который динамически правит стек, отлавливает спекуляцию и затем по команде `Return` переходит на новое место в коде.

```
Trampoline:
    call SetupTarget    ; push address of CaptureSpec on the stack

CaptureSpec:
    int 3               ; Breakpoint to capture speculation
    jmp CaptureSpec     ; (similar to a LFENCE barrier)

SetupTarget:
    mov QWORD PTR [rsp], r10 ; Overwrite return address on the stack
    ret                 ; Return
```

Рис. 8.9. Последовательность кода Retpoline в x86-процессоре

В Windows техника Retpoline реализована в ядре NT, которое может применять Retpoline-последовательность к самому себе или к образам внешних драйверов динамически, используя таблицу динамического перемещения значений (Dynamic Value Relocation Table, DVRT). Когда ядро скомпилировано с помощью Retpoline (если компилятор позволяет), для каждого найденного в коде непрямого перехода в DVRT образа добавляется запись, описывающая его тип и адрес. Сам машинный код, выполняющий непрямой переход, остается неизменным, но дополняется заполнением переменного размера. Запись в DVRT содержит всю информацию, необходимую ядру NT, чтобы динамически модифицировать этот код. Данная архитектура обеспечивает способность внешних драйверов, скомпилированных с Retpoline, запускаться также на старых версиях ОС, где анализ DVRT попросту пропускается.

ПРИМЕЧАНИЕ Изначально DVRT разрабатывалась для поддержки рандомизации структуры адресного пространства (Address Space Layout Randomization, ASLR, см. главу 5 тома 1). Лишь позже таблица была расширена, в ней появились дескрипторы Retpoline. Система способна различать, какая ее версия задействована в образе.

На фазе инициализации –1 ядро определяет наличие у процессора уязвимости к Spectre и в случае совместимости системы и достаточности аппаратных средств разрешает использование Retpoline и применяет его к образу ядра и HAL. Процедура `RtlPerformRetpolineRelocationsOnImage` анализирует DVRT и заменяет каждый непрямой переход, отраженный в ее записях, прямым, ведущим к контрольной процедуре Retpoline и не подверженным спекулятивным атакам. Первоначальный адрес цели перехода сохраняется в регистре процессора (для AMD и Intel это регистр R10) вместе с одной командой, удаляющей оставленное компилятором заполнение. Код Retpoline хранится внутри образа ядра NT в разделе RETPOL. Страница с его копией отображается в конце образа каждого из драйверов.

Перед запуском загрузочные драйверы обрабатываются процедурой `MiReloadBootLoadedDrivers`, которая физически переносит их в память, а заодно модифицирует их образы, в том числе для Retpoline. Загрузчик Windows помещает все загрузочные драйверы, ядро NT и образы HAL в сопредельное виртуальное адресное пространство, они не связаны с какой-либо контрольной зоной, что делает их страничное индексирование невозможным. Это означает, что вся занятая ими память никогда не выгружается, а ядро NT может использовать все ту же функцию `RtlPerformRetpolineRelocationsOnImage` для модификации каждого непрямого перехода в коде напрямую. Если активна функция Hypervisor Code Integrity (HVCI), система должна вызывать безопасное ядро для применения Retpoline (через безопасный вызов `PERFORM_RETPOLINE_RELOCATIONS`). В итоге в такой ситуации память защищена исполняемым кодом драйверов от любых изменений в соответствии с принципом W^X, описанным в главе 9. Только безопасное ядро вправе проводить модификацию.

ПРИМЕЧАНИЕ Изменения, связанные с Retpoline и оптимизацией импорта, вносятся ядром в загрузочные драйверы до того, как PatchGuard, известный также как технология Kernel Patch Protection (защита ядра от модификации; см. главу 7 в томе 1, чтобы узнать о нем подробнее), инициализируется и возьмет какие-либо из них под свою защиту. Ни другие драйверы, ни само ядро NT не вправе изменять исполняемые разделы защищенных драйверов.

Рабочие драйверы, как описывалось в главе 5 тома 1, загружаются диспетчером памяти NT, который создает объект раздела, отражающий образ файла драйвера. Это означает, что создается контрольная зона, в том числе прототип массива PTE, призванная отслеживать страницы этого блока памяти. Для разделов драйверов некоторые физические страницы сначала загружаются в память лишь для проверки целостности кода, после чего отправляются в список отложенных. Когда впоследствии раздел потребуется отобразить и его страницы запросят впервые, отложенные физические страницы из списка (или из самого файла) материализуются по требованию обработчика ошибки поиска страницы. К общим страницам, на которые указывают прототипные PTE, Windows применяет Retpoline. Если то же раздел требуется отобразить для приложения пользовательского режима, диспетчер памяти создает новые *частные* страницы и копирует туда содержимое страниц *общих*, при этом отменяя изменения, внесенные обработчиком Retpoline.

ПРИМЕЧАНИЕ Надо заметить, что некоторые относительно новые процессоры от Intel также способны к спекулятивному выполнению по команде Return. Для таких моделей Retpoline активировать нельзя, потому что от Spectre v2 эта техника уже не поможет. Улучшенная IBRS (новое аппаратное средство защиты) решает проблемы производительности IBRS.

Битовая карта Retpoline

Одной из первоначальных задач проектирования (и ограничений) реализации Retpoline в Windows было поддержание неоднородной среды, где вместе работают драйверы, как совместимые с Retpoline, так и нет, в то же время сохраняя систему в целом защищенной от Spectre v2. Это подразумевает необходимость выполнять

код несовместимых драйверов при активном IBRS (или STIBP вместе с IBPB на входе в ядро, как рассматривалось ранее в подразделе «Аппаратный контроль непрямых переходов»), тогда как все прочие могут работать безо всякого аппаратного вмешательства (защиту обеспечивают Retpoline и барьеры памяти).

Чтобы добиться динамической совместимости со старыми драйверами, на фазе 0 инициализации ядро NT создает и заполняет динамическую битовую карту, на которой отслеживает все свое адресное пространство в разрезе блоков по 64 Кбайт каждый. В рамках этой модели значение бита 1 указывает, что на данном участке адресного пространства находится Retpoline-совместимый код, 0 означает обратное. Затем ядро NT помечает значением 1 все биты, относящиеся к адресным пространствам образов HAL и NT, которые всегда совместимы с Retpoline. Каждый раз, когда загружается образ в составе ядра NT, система пробует обработать его с помощью Retpoline. Если операция успешна, соответствующие биты на битовой карте получают значение 1.

Контрольная процедура Retpoline включает в себя проверку битовой карты: при каждом непрямом вызове система проверяет, находится ли оригинальный целевой адрес в Retpoline-совместимом модуле. В случае успеха проверки (и если соответствующий бит равен 1) система выполняет контроль Retpoline (как на рис. 8.9) и безопасно переходит по целевому адресу. В ином случае (когда на битовой карте попался 0) инициализируется процедура выхода из Retpoline. В PCRB текущего процессора устанавливается флаг `RUNNING_NON_RETPOLINE_CODE` (это требуется для переключения контекста), активируется IBRS (или STIBP в зависимости от аппаратной конфигурации), при необходимости используются IBPB и LFENCE и, наконец, создается событие ядра `SPEC_CONTROL`. После всего этого процессор переходит по целевому адресу уже в безопасности (защиту обеспечивают аппаратные средства).

Когда квант потока заканчивается и планировщик выбирает новый поток, последний сохраняет статус Retpoline, представленный наличием флага `RUNNING_NON_RETPOLINE_CODE`, у текущих процессов в структуре данных `KTHREAD` предыдущего потока. Таким образом, когда тот поток снова будет выбран на выполнение (или сработает системное прерывание ядра), система уже будет знать о необходимости перезапустить средства аппаратной защиты, чтобы защитить систему.

Оптимизация импорта

Кроме прочего, записи Retpoline в DVRT содержат непрямые переходы в импортированные функции. Запись в DVRT о передаче контроля импортированному коду описывает такой переход, используя индекс, указывающий на подходящую запись в IAT (таблица адресов импортированных образов (Image Import Address Table) — массив указателей на импортированные функции, формируемый загрузчиком). После того как загрузчик Windows скомпилировал IAT, дальнейшие изменения содержимого этой структуры маловероятны, не считая нескольких редких случаев. Как видно из рис. 8.10, непрямой переход в импортированную функцию не нуждается в превращении в Retpoline-вызов, так как ядро NT самостоятельно может обеспечить достаточную близость виртуальных адресов двух образов (вызывающего и вызываемого) для прямого перехода к целевой функции (менее 2 Гбайт).

```
StandardCall:
    call QWORD PTR [IAT+ExAllocatePoolOffset]  ; 7 bytes
    nop  DWORD PTR [RAX+RAX]     ; 5 bytes

ImportOptimizedCall:
    mov  R10, QWORD PTR [IAT+ExAllocatePoolOffset] ; 7 bytes
    call ExAllocatePool          ; Direct call (5 bytes)

RetpolineOnly:
    mov R10, QWORD PTR [IAT+ExAllocatePoolWithTagOffset]  ; 7 bytes
    call _retpoline_import_r10  ; Direct call (5 bytes)
```

Рис. 8.10. Различные непрямые переходы и функция ExAllocatePool

При оптимизации импорта используются динамические переносы Retpoline, чтобы трансформировать непрямые вызовы импортированных функций в прямые вызовы. Если для передачи управления в импортированную функцию применяется прямой вызов, задействовать Retpoline нет нужды, так как они не страдают от атак спекуляцией. Ядро NT оптимизирует импорт и трансформирует Retpoline одновременно, и хотя эти возможности могут быть настроены по отдельности, для корректной работы обе используют одну и ту же структуру DVRT. С помощью оптимизации импорта Windows удалось добиться прироста производительности даже в системах без уязвимости от Spectre v2 (прямой переход не требует дополнительного обращения к памяти).

Сопряжение STIBP

В системах с гиперпоточностью для защиты кода, выполняемого в пользовательском режиме, от уязвимостей Spectre v2, потоки с ним должны работать как минимум в условиях STIBP. В иных случаях в этом нет необходимости: защита против спекулятивного выполнения пользовательских потоков уже достигнута благодаря активности IBRS, имевшей место для только что выполнявшегося кода ядра. Когда же задействован Retpoline, нужный для этого режим IBPB запускается при первом выходе из ловушки ядра после межпроцессорного переключения потоков. Это гарантирует, что перед выполнением кода пользовательского потока буфер целей переходов процессора будет пуст.

То, что STIBP остается активным в гиперпоточной системе, грозит падением быстродействия, так что для потоков пользовательского режима данный режим отключен. Это оставляет их потенциально уязвимыми для спекулятивных атак из родственного SMT-потока. Конечный пользователь может вручную активировать STIBP для своих потоков с помощью настройки USER_STIBP_ALWAYS (см. раздел «Аппаратный контроль непрямых переходов» ранее в данной главе) или опции защиты процесса RESTRICT_INDIRECT_BRANCH_PREDICTION.

Описанный сценарий неидеален. Лучшее решение реализовано с помощью механизма сопряжения STIBP. Он активируется диспетчером ввода/вывода на фазе 1 инициализации ядра функцией `KeOptimizeSpecCtrlSettings` в строго определенных условиях. Система должна быть гиперпоточной, а процессор — поддерживать IBRS и STIBP. Более того, сопряжение STIBP совместимо либо лишь с невложенными

виртуализованными средами, либо когда заблокирован функционал Hyper-V (см. главу 9, чтобы узнать об этом подробнее).

В ситуациях, когда активно сопряжение STIBP, система присваивает каждому процессу идентификатор домена безопасности (сохраняется в структуре данных EPROCESS), представленный 64-разрядным числом. Идентификатор системного домена, равный 0, выдается только процессам, выполняемым под эгидой SYSTEM или с полными административными привилегиями. Несистемные домены безопасности назначаются при создании процесса (реализуется функцией `PspInitializeProcessSecurity`) по следующим правилам.

- Если новый процесс создается без прямого указания основного признака, он получает отметку того же домена безопасности, что и родительский процесс, его создавший.
- В случае, когда новому процессу специально назначается новый признак (к примеру, с помощью API-функций `CreateProcessAsUser` и `CreateProcessWithLogon`), для него создается новый пользовательский домен безопасности, начиная с внутреннего значения `PsNextSecurityDomain`. Последний инкрементируется каждый раз, когда создается новый домен, что исключает вероятность конфликта требований безопасности на протяжении всего сеанса работы системы.
- Заметим также, что новый признак безопасности может быть назначен API-функцией `NtSetInformationProcess` (с экземпляром информационного класса `ProcessAccessToken`) уже после создания процесса. Чтобы это сработало, процесс должен создаваться сразу приостановленным (в нем не будет выполняющихся потоков). На данном этапе у него все еще будет оригинальный незакрепленный признак. Новый домен безопасности назначается по тем же правилам, что и ранее.

Кроме того, домены безопасности можно вручную назначать различным процессам, относящимся к одной и той же группе. Приложение может заменить домен безопасности процесса таковым от другого процесса из той же группы с помощью API-функции `NtSetInformationProcess`, передав туда экземпляр класса `ProcessCombineSecurityDomainsInformation`. Функция принимает идентификаторы двух процессов и заменяет домен первого лишь в том случае, если у обоих токены безопасности закреплены и оба имеют друг к другу доступ с правами доступа `PROCESS_VM_WRITE` и `PROCESS_VM_OPERATION`.

Использование доменов безопасности позволяет механизму сопряжения STIBP функционировать. В рамках своей работы он устанавливает связь между логическим процессором (LP) с родственным ему, с которым они делят общее физическое ядро. (В данном разделе термины «процессор» и «логический процессор» взаимозаменяемы.) Два LP сопрягаются соответствующим алгоритмом STIBP, реализуемым системной функцией `KiUpdateStibpPairing`, лишь когда домен безопасности локального процессора равен домену безопасности удаленного либо когда один из LP ничем не занят. В подобных случаях оба LP могут работать без активного режима STIBP, все еще оставаясь защищенными от атак через спекулятивное выполнение (атака на родственный LP с тем же контекстом безопасности не дает преимуществ).

Алгоритм сопряжения STIBP реализуется в процедуре `KiUpdateStibpPairing`. Она вызывается обработчиком выхода из ловушки (он выполняется при выходе из кода ядра обратно в пользовательский поток), только когда статус сопряжения в PCRB процессора устарел. Статус сопряжения LP устаревает в основном по двум причинам.

- Планировщик NT принял решение выполнять на данном процессоре новый поток. Если тот имеет домен безопасности, отличный от предшественника, статус сопряжения в PCRB процессора считается устаревшим. Это дает алгоритму сопряжения STIBP возможность заново оценить статус их сопряжения.
- Когда родственный процессор выходит из состояния бездействия, он требует у удаленного процессора перепроверить статус своего сопряжения STIBP.

Заметим: когда логический процессор выполняет код в режиме STIBP, он находится под защитой от спекулятивного выполнения на родственном процессоре. Сопряжение STIBP задумывалось и с обратной целью: когда логический процессор выполняет что-то при STIBP, гарантируется, что родственный процессор защищен сам от себя. Это подразумевает, что, когда происходит переключение контекста в иной домен безопасности, родственный процессор не отвлекают, даже если там работает пользовательский поток без STIBP.

Описанная ситуация не произойдет, только когда планировщик выберет VP-переданный поток (он отражает процессор виртуальной машины, если активирован планировщик Root; см. главу 9, где об этом рассказано подробнее), принадлежащий процессу VMMEM. В таком случае система немедленно отправляет родственному потоку межпроцессорное прерывание (Interrupt Processor Interrupt, IPI) для обновления его статуса сопряжения STIBP. В конечном счете в VP-переданном потоке выполняется код из гостевой виртуальной машины, который сам всегда может отключить себе STIBP, лишая защиты родственный поток (а значит, оба работают без STIBP).

ЭКСПЕРИМЕНТ. Запрос статуса системных средств защиты от атак по сторонним каналам

Сведения о статусе средств защиты от атак по сторонним каналам в Windows можно получить от платформенно-зависимого API `NtQuerySystemInformation`, работая с экземплярами информационных классов `SystemSpeculationControlInformation` и `SystemSecureSpeculationControlInformation`. Существует несколько инструментов для взаимодействия с данным функционалом, способных показать конечному пользователю искомый статус.

- Скрипт SpeculationControl для Powershell, разработанный Мэтом Миллером (Matt Miller) и официально поддерживаемый Microsoft. Он имеет открытый исходный код, который доступен в репозитории GitHub по адресу https://github.com/microsoft/SpeculationControl.
- Инструмент SpecuCheck, разработанный Алексом Ионеску (одним из авторов данной книги), который имеет открытый исходный код и доступен в репозитории GitHub по адресу https://github.com/ionescu007/SpecuCheck.

- Утилита SkTool, разработанная Андреа Аллиеви (одним из авторов книги) и распространяемая (на момент написания этого текста) в новых релизах Windows Insider.

Все три представленные утилиты выдают примерно одинаковые результаты. Впрочем, лишь SkTool способна отобразить средства защиты от атак по сторонним каналам, реализуемые безопасным ядром (гипервизор и безопасное ядро подробно описываются в главе 9). В данном эксперименте вы узнаете, какие защитные технологии активны в вашей системе. Загрузите утилиту SpecuCheck и запустите через командную строку (можно ввести `CMD` в окне поиска Cortana). В итоге вы получите примерно следующий текст:

```
SpecuCheck v1.1.1 -- Copyright(c) 2018 Alex Ionescu
https://ionescu007.github.io/SpecuCheck/ -- @aionescu
--------------------------------------------------------

Mitigations for CVE-2017-5754 [rogue data cache load]
--------------------------------------------------------

[-] Kernel VA Shadowing Enabled:                      yes
     > Unnecessary due lack of CPU vulnerability:     no
     > With User Pages Marked Global:                 no
     > With PCID Support:                             yes
     > With PCID Flushing Optimization (INVPCID):     yes

Mitigations for CVE-2018-3620 [L1 terminal fault]
[-] L1TF Mitigation Enabled:                          yes
     > Unnecessary due lack of CPU vulnerability:     no
     > CPU Microcode Supports Data Cache Flush:       yes
     > With KVA Shadow and Invalid PTE Bit:           yes
```

(Вывод был обрезан из соображений экономии места.)

Чтобы попробовать SkTool, можете загрузить последний релиз Windows Insider. При запуске без параметров командной строки утилита по умолчанию отобразит статус гипервизора и безопасного ядра. Чтобы увидеть статус всех средств защиты, вам потребуется запустить утилиту, добавив параметр `/mitigations` в командную строку:

```
Hypervisor / Secure Kernel / Secure Mitigations Parser Tool 1.0

Querying Speculation Features... Success!
    This system supports Secure Speculation Controls.

System Speculation Features.
    Enabled: 1
    Hardware support: 1
    IBRS Present: 1
    STIBP Present: 1
    SMEP Enabled: 1
    Speculative Store Bypass Disable (SSBD) Available: 1
    Speculative Store Bypass Disable (SSBD) Supported by OS: 1
    Branch Predictor Buffer (BPB) flushed on Kernel/User transition: 1
    Retpoline Enabled: 1
    Import Optimization Enabled: 1
```

```
    SystemGuard (Secure Launch) Enabled: 0 (Capable: 0)
    SystemGuard SMM Protection (Intel PPAM / AMD SMI monitor) Enabled: 0

Secure system Speculation Features.
    KVA Shadow supported: 1
    KVA Shadow enabled: 1
    KVA Shadow TLB flushing strategy: PCIDs
    Minimum IBPB Hardware support: 0
    IBRS Present: 0 (Enhanced IBRS: 0)
    STIBP Present: 0
    SSBD Available: 0 (Required: 0)
    Branch Predictor Buffer (BPB) flushed on Kernel/User transition: 0
    Branch Predictor Buffer (BPB) flushed on User/Kernel and VTL 1 transition: 0
    L1TF mitigation: 0
    Microarchitectural Buffers clearing: 1
```

ДИСПЕТЧЕРИЗАЦИЯ СИСТЕМНЫХ ПРЕРЫВАНИЙ

Прерываниями и исключениями называют ситуации, когда операционная система отвлекает процессор от выполнения запущенных программ. Их могут создавать как программы, так и аппаратура. *Системным прерыванием* (иногда *ловушкой* — trap) называют механизм процессора, предназначенный для захвата выполняемого потока при возникновении исключения или прерывания и для передачи управления в определенное место в операционной системе. В Windows процессор передает управление обработчику системного прерывания, который является функцией, характерной для того или иного прерывания или исключения. Некоторые условия активации обработчиков системных прерываний представлены на рис. 8.11.

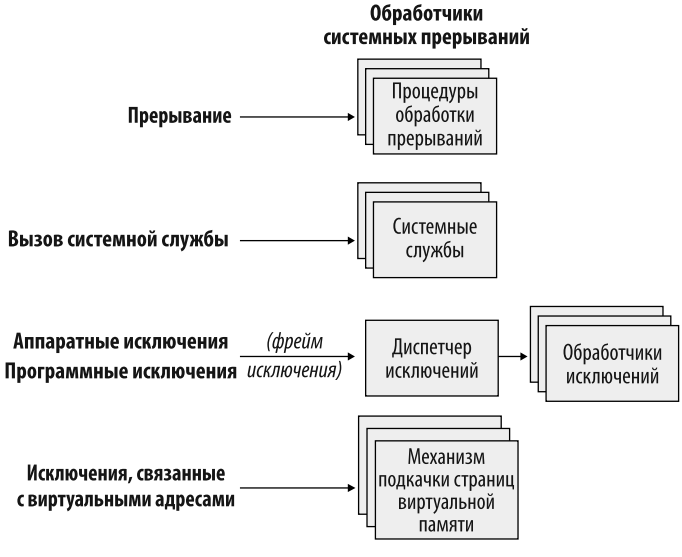


Рис. 8.11. Диспетчеризация системных прерываний

Ядро различает прерывания и исключения следующим образом. *Прерывание* (interrupt) является асинхронным событием (которое может произойти в любое время), не связанным с текущей задачей процессора. Прерывания генерируются главным образом устройствами ввода-вывода, процессорными часами или таймерами, и они могут быть разрешены (включены) или запрещены (выключены). *Исключение* (execution), напротив, является синхронным условием, которое обычно возникает в результате выполнения конкретной инструкции. *Аварийные завершения работы* (aborts), например, из-за машинного сбоя, обычно не связаны с выполнением инструкции. И исключения, и аварийные завершения порой называют *отказами* (faults), примеры — *отказ страницы* или *двойной отказ*. Повторение исключений может быть вызвано повторным запуском программы с теми же данными и при тех же условиях. В качестве примеров исключений можно привести нарушения доступа к памяти, определенные инструкции отладки и ошибки деления на нуль. Ядро также считает исключениями вызовы системных служб (хотя технически они являются системными прерываниями).

Исключения и прерывания могут генерироваться как оборудованием, так и программами. Например, причиной исключения, связанного с ошибкой шины, является оборудование, а исключение, связанное с делением на нуль, является результатом программной ошибки. Точно так же прерывание может генерироваться устройством ввода-вывода, или же программное прерывание может быть выдано самим ядром (прерывания APC или DPC будут рассмотрены в этой главе).

Когда происходит аппаратное исключение или прерывание, процессоры x86 и x64 в первую очередь проверяют, относится ли текущий сегмент кода (CS) к CPL 0 или ниже, то есть относится ли текущий поток к пользователю или системе. В случае если поток уже выполнялся на уровне кольца 0, процессор сохраняет *(выталкивает)* в текущий стек следующий блок информации, описывающий переход по коду из ядра в само же ядро:

- текущие флаги процессора (EFLAGS/RFLAGS);
- текущий сегмент кода (CS);
- текущий программный счетчик (EIP/RIP);
- не обязательно, но в некоторых ситуациях возможен *код ошибки*.

В случаях же, когда процессор выполняет код пользовательского режима на уровне кольца 3, он сначала оценивает текущее состояние TSS по регистру задачи (Task Register, TR) и переключается на SS0/ESP0 для x86 или просто RSP0 для x64, как описывалось в подразделе «Сегменты состояния задач» в данной главе. Далее, уже работая со стеком ядра, процессор сначала сохраняет предыдущие значения регистров SS (значение пользовательского режима) и ESP (стек пользовательского режима), а затем остальные данные аналогично переходу из ядра в ядро.

Сохранение этих данных дает сразу два преимущества. Во-первых, данных о состоянии выполнения в стеке ядра достаточно для того, чтобы потом вернуться в ту же точку потока и продолжить его выполнение как ни в чем не бывало. Во-вторых, так операционная система знает (исходя из сохраненного значения регистра

CS), откуда произошел переход на системное прерывание, к примеру, было выброшено исключение в коде пользовательского режима или ядра.

Процессор сохраняет столько данных, сколько необходимо для возобновления выполнения прерванного потока, в то время как прочие важные регистры, такие как EAX, EBX, ECX, EDI и т. д., записываются в кадр ловушки — структуру данных, в которой сохраняется состояние потока, выполнявшегося на момент прерывания. Эта информация позволяет ядру возобновить выполнение потока после обработки прерывания или исключения. Там оказывается полное состояние среды выполнения потока, включающее в себя контекст с дополнительной информацией. Вы сможете оценить ее определение, использовав команду `dt nt!_KTRAP_FRAME` в отладчике ядра либо, как вариант, загрузив Windows Driver Kit (WDK) и с его помощью рассмотрев заголовочный файл `NTDDK.H`, где также имеются дополнительные комментарии (подробности о контексте потока см. в главе 5 тома 1). Ядро может обрабатывать программные прерывания и как частный случай аппаратных, и в синхронном порядке, если поток обращается к функциям ядра, относящимся к программным прерываниям.

В большинстве случаев ядро обеспечивает вызов сначала передовых функций — обработчиков системных прерываний, выполняющих общие задачи их обработки до и после передачи управления другим функциям, выставившим системное прерывание. К примеру, если ситуация была вызвана прерыванием от какого-нибудь устройства, обработчик ядра передает управление *процедуре обработки прерывания* (interrupt service routine, ISR), предоставленной драйвером устройства, инициировавшего прерывание. Если ситуация создалась из-за вызова системной службы, общий обработчик системных прерываний, связанных с системными службами, передает управление конкретной функции системной службы в исполнительной системе.

В нестандартных ситуациях ядро может сталкиваться с активацией ловушек и прерываний, встреча или обработка которых не ожидалась. Иногда их называют *неожиданными* или *ложными* ловушками. Обработчики обычно вызывают системную функцию `KeBugCheckEx`. Если ядро обнаруживает проблемное или некорректное поведение, без присмотра способное привести к повреждению данных, эта функция приостанавливает работу компьютера. В следующих разделах более подробно рассматриваются прерывания, исключения и обращения к системным службам.

Диспетчеризация прерываний

Аппаратные прерывания, как правило, приходят с устройств ввода/вывода, которым требуется сообщить процессору о необходимости среагировать. Устройства, полагающиеся на прерывания, позволяют операционной системе использовать процессор максимально продуктивно, внедряя в основную работу операции ввода/вывода. Любой поток может запустить обмен данными с устройством и вернуться к основным задачам, пока не закончится передача. Когда это происходит, устройство снова обращается к процессору. Как правило, к таким устройствам относятся устройства указания координат, принтеры, клавиатуры, дисковые приводы и сетевые адаптеры.

Системное программное обеспечение тоже может инициировать прерывание. К примеру, ядро может создать программное прерывание для управления потоками и вмешаться в выполнение какого-то из них асинхронно. Ему также доступна возможность просто отключить прерывания, и тогда процессор не будет отвлекаться. Впрочем, такое происходит нечасто, а именно в критические моменты, к примеру, когда задается обработчик прерывания или идет обработка исключения.

Для ответа на прерывания, исходящие от устройств, ядро использует обработчики системных прерываний. Они могут передать управление либо внешней процедуре (ISR), обрабатывающей прерывание, либо внутренней процедуре ядра, реагирующей на прерывание. Драйверы устройств обеспечивают обработку прерываний своими ISR, а ядро предоставляет процедуры для обработки прочих типов прерываний.

В следующих разделах вы узнаете о том, как устройства оповещают процессор о своих прерываниях, видах прерываний, поддерживаемых ядром, о том, как драйверы устройств взаимодействуют с ядром (в рамках обработки прерываний), и о программных прерываниях, на которые ядро может реагировать (в том числе об объектах ядра, которые их реализуют).

Обработка аппаратных прерываний

На аппаратных платформах, поддерживаемых Windows, внешние прерывания ввода/вывода подаются на вход контроллера прерываний, к примеру улучшенного программируемого контроллера прерываний (I/O Advanced Programmable Interrupt Controller, IOAPIC). Тот, в свою очередь, прерывает локальный улучшенный программируемый контроллер прерываний (Local Advanced Programmable Interrupt Controllers, LAPIC) одного или нескольких процессоров, после чего происходит прерывание выполнения кода.

Когда процессор прерывают, он запрашивает у контроллера *глобальный вектор системного прерывания* (global system interrupt vector, GSIV), который иногда представлен номером *запроса прерывания* (interrupt request, IRQ). Контроллер прерываний преобразует GSIV в процессорный вектор прерывания, который затем используется для анализа структуры данных, называемой *таблицей диспетчеризации прерываний* (interrupt dispatch table, IDT). Эта таблица хранится в регистре IDT или IDTR, и из нее извлекается соответствующая вектору прерывания запись.

Исходя из содержания записи IDT, процессор может передать управление в соответствующий обработчик прерывания, выполняемый на уровне кольца 0 (сразу после действий, описанных в начале данного раздела). Более того, может быть даже загружен новый TSS и обновлен регистр задачи (TR), что происходит в рамках процедуры, называемой *шлюзом прерывания*. В случае Windows в ходе загрузки системы ядро заполняет IDT указателями как на специальные процедуры ядра и HAL для каждого исключения и обработки внутренних прерываний, так и на *перенаправляющие* процедуры ядра, называемые `KiIsrThunk` и предназначенные для обработки внешних прерываний, к которым могут привязаться драйверы устройств от сторонних разработчиков. В процессорных архитектурах x86 и x64 первые 32 записи IDT, связанные с векторами прерываний с 0-го по 31-й, считаются зарезервированными под системные прерывания процессора, перечисленные в табл. 8.3.

Таблица 8.3. Системные прерывания процессора

Номер (мнемоника)	Значение
0 (#DE)	Ошибка деления на 0 (Divide error)
1 (#DB)	Исключение отладки (Debug trap)
2 (NMI)	Немаскируемое внешнее прерывание (Nonmaskable interrupt)
3 (#BP)	Точка останова (Breakpoint trap)
4 (#OF)	Исключение по переполнению (Overflow fault)
5 (#BR)	Прерывание по выходу индекса за границы массива (Bound fault)
6 (#UD)	Недопустимый код операции (Undefined opcode fault)
7 (#NM)	Устройство недоступно (FPU error)
8 (#DF)	Двойной отказ (Double fault)
9 (#MF)	Сопроцессор недоступен (больше не используется) (Coprocessor fault)
10 (#TS)	Недопустимый сегмент состояния задачи (TSS fault)
11 (#NP)	Сегмент отсутствует (Segment fault)
12 (#SS)	Сбой в работе стека (Stack fault)
13 (#GP)	Общая ошибка защиты (General protection fault)
14 (#PF)	Отказ страницы (Page fault)
15	Зарезервировано
16 (#MF)	Ошибка операции с плавающей точкой (Floating point fault)
17 (#AC)	Контроль выравнивания (Alignment check fault)
18 (#MC)	Ошибка машинного контроля (Machine check abort)
19 (#XM)	Сбой SIMD (SIMD fault)
20 (#VE)	Исключение виртуализации (Virtualization exception)
21 (#CP)	Исключение контроля защиты (Control protection exception)
22–31	Зарезервировано

Остальные записи в IDT зависят от сочетания как жестко заданных значений (к примеру, векторы с 30-го по 34-й всегда применяются для связанных с Hyper-V прерываний VMBus), так и выбранных в результате компромисса между драйверами устройств, оборудованием, контроллерами прерываний и встроенного ПО вроде ACPI. Например, в одном экземпляре Windows контроллер клавиатуры может пользоваться вектором 82, а в другом — 67.

ЭКСПЕРИМЕНТ. Просмотр 64-разрядной IDT

Вы можете просмотреть содержимое IDT, в том числе информацию о том, какие обработчики Windows назначила на прерывания (в том числе исключения и IRQ), с помощью команды отладчика ядра `!idt`. Без параметров команда `!idt` выведет упрощенные сведения, где будут только зарегистрированные аппаратные прерывания (а на 64-разрядных системах еще и обработчики ловушек процессора).

Следующий пример демонстрирует результат выполнения команды `!idt` в системе x64:

```
0: kd> !idt

Dumping IDT: fffff8027074c000

00:    fffff8026e1bc700 nt!KiDivideErrorFault
01:    fffff8026e1bca00 nt!KiDebugTrapOrFault    Stack = 0xFFFFF8027076E000
02:    fffff8026e1bcec0 nt!KiNmiInterrupt        Stack = 0xFFFFF8027076A000
03:    fffff8026e1bd380 nt!KiBreakpointTrap
04:    fffff8026e1bd680 nt!KiOverflowTrap
05:    fffff8026e1bd980 nt!KiBoundFault
06:    fffff8026e1bde80 nt!KiInvalidOpcodeFault
07:    fffff8026e1be340 nt!KiNpxNotAvailableFault
08:    fffff8026e1be600 nt!KiDoubleFaultAbort    Stack = 0xFFFFF80270768000
09:    fffff8026e1be8c0 nt!KiNpxSegmentOverrunAbort
0a:    fffff8026e1beb80 nt!KiInvalidTssFault
0b:    fffff8026e1bee40 nt!KiSegmentNotPresentFault
0c:    fffff8026e1bf1c0 nt!KiStackFault
0d:    fffff8026e1bf500 nt!KiGeneralProtectionFault
0e:    fffff8026e1bf840 nt!KiPageFault
10:    fffff8026e1bfe80 nt!KiFloatingErrorFault
11:    fffff8026e1c0200 nt!KiAlignmentFault
12:    fffff8026e1c0500 nt!KiMcheckAbort         Stack = 0xFFFFF8027076C000
13:    fffff8026e1c0fc0 nt!KiXmmException
14:    fffff8026e1c1380 nt!KiVirtualizationException
15:    fffff8026e1c1840 nt!KiControlProtectionFault
1f:    fffff8026e1b5f50 nt!KiApcInterrupt
20:    fffff8026e1b7b00 nt!KiSwInterrupt
29:    fffff8026e1c1d00 nt!KiRaiseSecurityCheckFailure
2c:    fffff8026e1c2040 nt!KiRaiseAssertion
2d:    fffff8026e1c2380 nt!KiDebugServiceTrap
2f:    fffff8026e1b80a0 nt!KiDpcInterrupt
30:    fffff8026e1b64d0 nt!KiHvInterrupt
31:    fffff8026e1b67b0 nt!KiVmbusInterrupt0
32:    fffff8026e1b6a90 nt!KiVmbusInterrupt1
33:    fffff8026e1b6d70 nt!KiVmbusInterrupt2
34:    fffff8026e1b7050 nt!KiVmbusInterrupt3
35:    fffff8026e1b48b8 hal!HalpInterruptCmciService (KINTERRUPT fffff8026ea59fe0)
b0:    fffff8026e1b4c90 ACPI!ACPIInterruptServiceRoutine (KINTERRUPT ffffb88062898dc0)
ce:    fffff8026e1b4d80 hal!HalpIommuInterruptRoutine (KINTERRUPT fffff8026ea5a9e0)
d1:    fffff8026e1b4d98 hal!HalpTimerClockInterrupt (KINTERRUPT fffff8026ea5a7e0)
d2:    fffff8026e1b4da0 hal!HalpTimerClockIpiRoutine (KINTERRUPT fffff8026ea5a6e0)
d7:    fffff8026e1b4dc8 hal!HalpInterruptRebootService (KINTERRUPT fffff8026ea5a4e0)
d8:    fffff8026e1b4dd0 hal!HalpInterruptStubService (KINTERRUPT fffff8026ea5a2e0)
df:    fffff8026e1b4e08 hal!HalpInterruptSpuriousService (KINTERRUPT fffff8026ea5a1e0)
e1:    fffff8026e1b8570 nt!KiIpiInterrupt
e2:    fffff8026e1b4e20 hal!HalpInterruptLocalErrorService (KINTERRUPT fffff8026ea5a3e0)
e3:    fffff8026e1b4e28 hal!HalpInterruptDeferredRecoveryService
                        (KINTERRUPT fffff8026ea5a0e0)
fd:    fffff8026e1b4ef8 hal!HalpTimerProfileInterrupt (KINTERRUPT fffff8026ea5a8e0)
fe:    fffff8026e1b4f00 hal!HalpPerfInterrupt (KINTERRUPT fffff8026ea5a5e0)
```

На компьютере, где был получен данный текст, ISR-обработчик для ACPI SCI назначен на номер прерывания B0h. Вы также можете увидеть, что номер

прерывания 14 (0Eh) соответствует функции `KiPageFault`, которая, как описано ранее, относится к предопределенным ловушкам процессора.

Можно также заметить, что некоторые из прерываний, а именно 1, 2, 8 и 12, дополнены указателем стека. Это ловушки, описанные в подразделе «Сегменты состояния задач», которые требуют отдельных безопасных стеков ядра для обработки. Отладчик узнает эти указатели на стек, выгружая запись из IDT, что можете сделать и вы с помощью команды dx и сопоставления одного из векторов прерываний в IDT. Получить IDT можно или из регистра процессора IDTR, или другим способом, а именно из структуры данных ядра KPCR, где в поле `IdtBase` хранится аналогичный указатель:

```
0: kd> dx @$pcr->IdtBase[2].IstIndex
@$pcr->IdtBase[2].IstIndex : 0x3 [Type: unsigned short]

0: kd> dx @$pcr->IdtBase[0x12].IstIndex
@$pcr->IdtBase[0x12].IstIndex : 0x2 [Type: unsigned short]
```

Если вы сравните значения индексов IDT, выведенные здесь, с результатами приведенного ранее эксперимента по выгрузке TSS в системе x64, то увидите указатели на стек ядра, такие как в данном эксперименте.

Каждый процессор в системе имеет свою IDT (ее расположение указано в его регистре IDTR), так что они могут использовать свои наборы ISR-обработчиков, если необходимо. Например, в многопроцессорной системе все процессоры получают прерывание от встроенных часов, но только один из них реагирует. Однако остальные тоже его применяют для отслеживания истечения кванта времени активного потока, чтобы, когда это произойдет, запланировать другой. Подобным образом некоторые конфигурации оборудования могут требовать, чтобы конкретный процессор обрабатывал прерывания от определенного набора устройств.

Архитектура программируемого контроллера прерываний

Традиционно системы типа x86 опирались на программируемый контроллер прерываний (Programmable Interrupt Controller, PIC) i8259A, что соответствует стандарту, применявшемуся в оригинальных IBM PC. Данный PIC задействовался только в однопроцессорных системах и поддерживал лишь восемь конвейеров прерываний. Однако архитектура IBM PC определила возможность добавления второго PIC, известного как *вторичный*, прерывания которого мультиплексировались в один из конвейеров прерываний основного PIC. Это обеспечивало всего 15 конвейеров (семь на основном и восемь на вторичном, мультиплексируемых через восьмой конвейер основного PIC). Подобный способ управления более чем восемью устройствами сильно усложнял работу PIC, и даже 15 конвейеров однажды стали серьезным ограничением. Не облегчали ситуацию ограничения однопроцессорных систем, равно как и различные проблемы, связанные с электроникой (существовала вероятность возникновения ложных прерываний). В итоге подобные контроллеры

были изжиты из современных систем, а на смену им пришел улучшенный программируемый контроллер прерываний i82489 (Advanced Programmable Interrupt Controller, APIC).

Ввиду того что контроллеры APIC могли работать в мультипроцессорных системах, Intel и другие компании сформулировали *мультипроцессорную спецификацию* (Multiprocessor Specification, MPS) — стандарт проектирования мультипроцессорных x86-систем. Он был построен вокруг использования APIC и интеграции APIC ввода/вывода (IOAPIC), связанного с внешними устройствами, локальным APIC (LAPIC), связанным с ядром процессора. Со временем стандарт MPS стал частью усовершенствованного интерфейса управления конфигурацией и питанием (Advanced Configuration and Power Interface, ACPI). Так совпало, что новая аббревиатура оказалась анаграммой прежней. Чтобы обеспечить совместимость с однопроцессорными системами и загрузочным кодом, запускающим многопроцессорную систему в однопроцессорном режиме, контроллеры APIC поддерживают режим совместимости, где конвейеров прерываний всего 15 и все они обрабатываются только основным процессором. Архитектура APIC отражена на рис. 8.12.

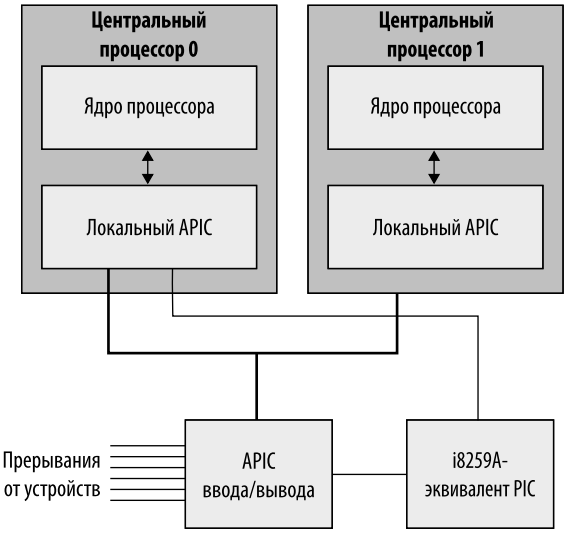


Рис. 8.12. Архитектура APIC

Как было упомянуто ранее, контроллер APIC состоит из нескольких компонентов: APIC ввода/вывода получает сигналы прерываний от устройств локальных APIC, которые получают эти прерывания по шине и останавливают ассоциированный с ними процессор; i8259A-совместимый контроллер прерываний преобразует сигналы APIC в аналогичные для PIC. Ввиду того что APIC ввода/вывода в системе тоже может быть несколько, материнские платы зачастую реализуют дополнительную низкоуровневую логику для управления взаимодействием между ними и процессорами. В ней реализуются алгоритмы маршрутизации прерываний, в чьи задачи входит как балансирование нагрузки

прерываний между процессорами, так и использование преимущества локальности, когда прерывания доставляются на тот же процессор, который только что обработал прерывание того же типа. Исполняемые программы могут перепрограммировать APIC ввода/вывода на фиксированный алгоритм маршрутизации, который обходит эту функцию чипсета. В большинстве случаев Windows записывает в них собственную логику, чтобы поддержать различные функции, такие как *перестраивание прерываний*, что, впрочем, не лишает права голоса драйверы устройств и код прошивок.

Поскольку архитектура x64 совместима с операционными системами под x86, x64-системы должны обеспечивать работу с теми же контроллерами прерываний. Однако существенное отличие заключается в том, что версии Windows под x64 применяли для работы с прерываниями только APIC, в силу чего отказывались работать в системах без него. Версии Windows под x86, в свою очередь, поддерживали оборудование и с PIC, и с APIC. Начиная с Windows 8 это изменилось, и новые системы работали только с оборудованием с APIC вне зависимости от архитектуры процессора. Еще одно различие: в системах x64 регистр приоритета задач (Task Priority Register, TPR) APIC теперь напрямую связан с контрольным регистром 8 (CR8) процессора. Современные операционные системы, включая Windows, используют этот регистр для хранения текущего уровня приоритета программных прерываний (в случае Windows называемого IRQL) и информирования IOAPIC при принятии им решений по маршрутизации. Больше сведений о работе с IRQL вы получите в дальнейшем.

ЭКСПЕРИМЕНТ. Просмотр PIC и APIC

Чтобы просмотреть конфигурацию контроллера PIC в однопроцессорной системе или текущего локального APIC в многопроцессорной, вы можете воспользоваться командами отладчика ядра `!pic` и `!apic` соответственно. В примере далее приведен вывод команды `!pic` в однопроцессорной системе. Обратите внимание на то, что даже в системе с APIC эта команда будет работать так, как везде, где есть APIC и соответствующий эквивалент PIC для эмуляции устаревшего оборудования:

```
lkd> !pic
----- IRQ Number ----- 00 01 02 03 04 05 06 07 08 09 0A 0B 0C 0D 0E 0F
Physically in service:  Y  .  .  .  .  .  .  .  .  Y  Y  Y  .  .  .  .
Physically masked:      Y  Y  Y  Y  Y  Y  Y  Y  Y  Y  Y  Y  Y  Y  Y  Y
Physically requested:   Y  .  .  .  .  .  .  .  .  Y  Y  Y  .  .  .  .
Level Triggered:        .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .
```

Далее приводится вывод команды `!apic` в системе, где разрешен Hyper-V, что заметно по наличию записей `SINTI`, ссылающихся на характерный для Hyper-V синтетический контроллер прерываний (Synthetic Interrupt Controller, SynIC), описываемый в главе 9. Обратите внимание на то, что при локальной отладке ядра данная команда покажет APIC, ассоциированный с текущим процессором. Иными словами, это будет процессор, на котором выполняется поток самого отладчика в момент ввода команды. Обозревая аварийный дамп или удаленную

систему, вы можете воспользоваться командой ~ (тильда) с номером процессора, чтобы переключиться на его локальный APIC. В обоих случаях под меткой `ID:` будет указано, о каком именно процессоре идет речь.

```
lkd> !apic
Apic (x2Apic mode) ID:1 (50014) LogDesc:00000002 TPR 00
TimeCnt: 00000000clk  SpurVec:df  FaultVec:e2  error:0
Ipi Cmd: 00000000`0004001f  Vec:1F  FixedDel    Dest=Self    edg high
Timer..: 00000000`000300d8  Vec:D8  FixedDel    Dest=Self    edg high    m
Linti0.: 00000000`000100d8  Vec:D8  FixedDel    Dest=Self    edg high    m
Linti1.: 00000000`00000400  Vec:00  NMI         Dest=Self    edg high
Sinti0.: 00000000`00020030  Vec:30  FixedDel    Dest=Self    edg high
Sinti1.: 00000000`00010000  Vec:00  FixedDel    Dest=Self    edg high    m
Sinti2.: 00000000`00010000  Vec:00  FixedDel    Dest=Self    edg high    m
Sinti3.: 00000000`000000d1  Vec:D1  FixedDel    Dest=Self    edg high
Sinti4.: 00000000`00020030  Vec:30  FixedDel    Dest=Self    edg high
Sinti5.: 00000000`00020031  Vec:31  FixedDel    Dest=Self    edg high
Sinti6.: 00000000`00020032  Vec:32  FixedDel    Dest=Self    edg high
Sinti7.: 00000000`00010000  Vec:00  FixedDel    Dest=Self    edg high    m
Sinti8.: 00000000`00010000  Vec:00  FixedDel    Dest=Self    edg high    m
Sinti9.: 00000000`00010000  Vec:00  FixedDel    Dest=Self    edg high    m
Sintia.: 00000000`00010000  Vec:00  FixedDel    Dest=Self    edg high    m
Sintib.: 00000000`00010000  Vec:00  FixedDel    Dest=Self    edg high    m
Sintic.: 00000000`00010000  Vec:00  FixedDel    Dest=Self    edg high    m
Sintid.: 00000000`00010000  Vec:00  FixedDel    Dest=Self    edg high    m
Sintie.: 00000000`00010000  Vec:00  FixedDel    Dest=Self    edg high    m
Sintif.: 00000000`00010000  Vec:00  FixedDel    Dest=Self    edg high    m
TMR: 95, A5, B0
IRR:
ISR:
```

Различные числа, следующие за пометками `Vec`, отображают соответствующие векторы из IDT с указанной командой. К примеру, в данном выводе номер прерывания `0x1F` связан с вектором прерывания процессора (Interrupt Processor Interrupt, IPI), а вектор прерывания `0xE2` обрабатывает ошибки при работе API. Возвращаясь к показаниям команды `!idt` из рассмотренных ранее экспериментов, можно заметить, что это прерывание APC ядра (признак того, что недавно через IPI от одного процессора к другому был передан APC), а `0xE2`, как и ожидалось, является обработчиком ошибок локального APIC в HAL.

Приведенный далее вывод команды `!ioapic` отображает конфигурацию APIC ввода/вывода, а именно компоненты контроллера прерываний, связанные с устройствами. К примеру, заметьте, как GSIV/IRQ 9, он же прерывание системного контроля (System Control Interrupt, SCI), ассоциирован с вектором `B0h`, который в показаниях `!idt` из прежних экспериментов был связан с ACPI.SYS:

```
0: kd> !ioapic
Controller at 0fffff7a8c0000898 I/O APIC at VA 0fffff7a8c0012000
IoApic @ FEC00000  ID:8 (11)  Arb:0
Inti00.: 00000000`000100ff  Vec:FF  FixedDel  Ph:00000000    edg high    m
Inti01.: 00000000`000100ff  Vec:FF  FixedDel  Ph:00000000    edg high    m
Inti02.: 00000000`000100ff  Vec:FF  FixedDel  Ph:00000000    edg high    m
Inti03.: 00000000`000100ff  Vec:FF  FixedDel  Ph:00000000    edg high    m
```

```
Inti04.: 00000000`000100ff  Vec:FF  FixedDel  Ph:00000000    edg high    m
Inti05.: 00000000`000100ff  Vec:FF  FixedDel  Ph:00000000    edg high    m
Inti06.: 00000000`000100ff  Vec:FF  FixedDel  Ph:00000000    edg high    m
Inti07.: 00000000`000100ff  Vec:FF  FixedDel  Ph:00000000    edg high    m
Inti08.: 00000000`000100ff  Vec:FF  FixedDel  Ph:00000000    edg high    m
Inti09.: ff000000`000089b0  Vec:B0  LowestDl  Lg:ff000000    lvl high
Inti0A.: 00000000`000100ff  Vec:FF  FixedDel  Ph:00000000    edg high    m
Inti0B.: 00000000`000100ff  Vec:FF  FixedDel  Ph:00000000    edg high    m
```

Уровни запросов программных прерываний

Хотя контроллеры прерываний управляют приоритетами прерываний, Windows устанавливает собственную схему их приоритизации, известную как *уровень запроса прерывания* (Interrupt Request Level, IRQL). Ядро представляет IRQL в виде чисел от 0 до 31 в системах x86 и от 0 до 15 — в системах x64 (как в ARM/ARM64). Чем больше число, тем выше приоритет. Хотя ядро определяет стандартный набор IRQL для программных прерываний, HAL отображает номера аппаратных прерываний на IRQL-уровни. На рис. 8.13 показаны IRQL, определенные для архитектур x86 и x64 (и ARM/ARM64).

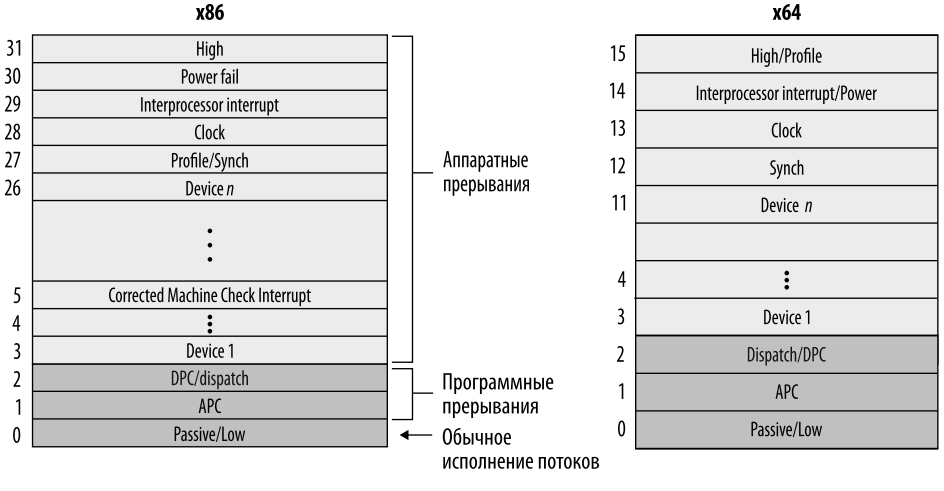


Рис. 8.13. Уровни запросов прерываний (IRQLs) для архитектур x86 и x64

Прерывания обслуживаются в порядке приоритета, и прерывание с высоким приоритетом обгоняет прерывание с низким. Когда срабатывает высокий приоритет, процессор сохраняет состояние прерванного потока и передает управление диспетчерам ловушек, ассоциированных с прерыванием. Диспетчер ловушки поднимает IRQL и вызывает процедуру, обслуживающую прерывание. После завершения ее выполнения диспетчер понижает IRQL процессора на уровень до срабатывания прерывания и загружает ранее сохраненное состояние. Прерванный поток возобновляет исполнение с прежнего места. Когда ядро понижает IRQL, могут реализоваться

ранее замаскированные низкоприоритетные прерывания. Если такое происходит, ядро повторяет процесс обработки прерывания.

Уровни приоритетов IRQL имеют совершенно иное значение, чем приоритеты при планировании потоков, описанные в главе 5 тома 1. Приоритет при планировании является атрибутом потока, в то время как IRQL имеет отношение к источнику прерывания, такому как клавиатура или мышь. Кроме того, каждый процессор имеет настройку IRQL, который меняется по ходу исполнения кода операционной системы. Как упоминалось ранее, в x64-системах IRQL хранится в регистре CR8, который отображается в TPR у APIC.

IRQL каждого процессора определяет, какие прерывания будут ему доставляться. Кроме того, IRQL применяются для синхронизации доступа к структурам данных уровня ядра (более подробно о них будет рассказано далее в главе). По ходу своего исполнения поток в режиме ядра повышает или понижает IRQL процессора либо напрямую, вызывая функции `KeRaiseIrql` и `KeLowerIrql`, либо чаще всего не напрямую, вызывая функции, обеспечивающие доступ к объектам синхронизации ядра. Как показано на рис. 8.14, прерывания с IRQL, превышающим текущий уровень, прерывают процессор, в то время как остальные, у которых IRQL ниже его или равен ему, маскируются до тех пор, пока исполняемый поток не понизит IRQL процессора.

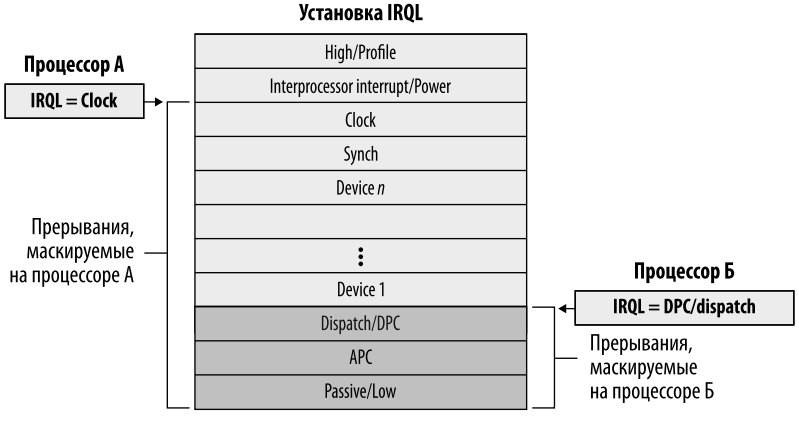


Рис. 8.14. Маскирование прерываний

Поток в режиме ядра повышает и понижает IRQL процессора, на котором он исполняется, в зависимости от стоящей перед ним задачи. К примеру, когда возникает прерывание, обработчик системного прерывания (или сам процессор в зависимости от его архитектуры) повышает IRQL процессора до IRQL, назначенного источнику прерывания. Это повышение маскирует все прерывания с приоритетом, равным ему или ниже его (только на этом процессоре), гарантируя, что процессор, обслуживающий прерывание, не будет отвлекаться на прерывания с таким IRQL. Замаскированные прерывания либо обрабатываются другим процессором, либо сохраняются, пока IRQL не снизится. Тем самым все компоненты системы, включая ядро и драйверы устройств, стремятся сохранять IRQL на *пассивном (passive)*

уровне (который иногда называют *низким (low)*). Это сделано для своевременной реакции драйверов устройств на аппаратные прерывания при условии, что IRQL не держится неоправданно высоким в течение продолжительных периодов времени. Таким образом, когда система не занята никакими задачами по аппаратным прерываниям (или не должна синхронизироваться с ними) либо не обрабатывает программные прерывания вроде DPC или APC, IRQL всегда равен нулю. Это касается любого кода пользовательского режима, поскольку позволение ему менять IRQL может существенно сказаться на работе всей системы. Фактически возврат в пользовательский режим с IRQL выше нуля приведет к критической ошибке *(bugcheck)* и указывает на существенный баг в драйвере.

Наконец, обратим внимание на то, что сами действия диспетчера, такие как переключение контекста между потоками для упреждения, работают на IRQL 2 (отсюда и название *«уровень dispatch»*), в силу чего процессор на этом уровне и выше действует однопоточно и кооперативно. К примеру, на этом уровне недопустимо ожидание объекта диспетчера (подробнее об этом — в разделе «Синхронизация» далее), потому что тогда переключение контекста на другой поток (или бездействующий поток) никогда не произойдет. Другое ограничение состоит в том, что при IRQL на уровне dispatch/DPC или выше доступ к невыгружаемой (nonpaged) памяти отсутствует.

Это правило на деле является побочным эффектом первого ограничения, потому что попытка доступа к нерезидентной памяти приведет к ошибке отказа страницы. Когда это происходит, диспетчер памяти начинает обращаться к диску, из-за чего должен дожидаться, пока драйвер файловой системы загрузит страницу оттуда. Ожидание, в свою очередь, потребует от планировщика переключения контекста (возможно, на бездействующий поток, если по плану нет пользовательских потоков), нарушая правило, по которому планировщик привлекать нельзя, поскольку во время чтения диска IRQL все еще на уровне DPC/dispatch или выше. Затем из-за того, что операция ввода/вывода обычно завершается на уровне APC, проблема усугубляется, ведь, даже если ожидания не потребуется, операция не завершится никогда, так как финальный APC никогда не запустится.

Если любое из этих двух ограничений нарушено, происходит отказ системы с кодом `IRQL_NOT_LESS_OR_EQUAL` или `DRIVER_IRQL_NOT_LESS_OR_EQUAL` (варианты отказа системы подробно обсуждаются в главе 10). Возникновение подобных ситуаций связано с типичными багами в драйверах устройств. Средство проверки драйверов Windows (Windows Driver Verifier) имеет режим, который может помочь в поиске подобных багов.

И наоборот, при работе на IRQL 1 (или *уровне APC*) упреждение все еще активно и переключение контекста может произойти. Из-за этого IRQL 1, в сущности, действует не как *зависимый от процессора*, а как *зависимый от потока* IRQL, коль скоро операции ожидания или упреждения будут активировать планировщик, который сохранит текущий IRQL в блоке управления потока (в структуре данных `KTHREAD`, как описано в главе 5) и восстановит IRQL процессора в блоке нового исполняемого потока. Это значит, что поток с пассивным уровнем (IRQL 0) все еще может быть вытеснен потоком на уровне APC (IRQL 1), потому что на уровнях ниже IRQL 2 планировщик сам решает, какому потоку передавать управление.

ЭКСПЕРИМЕНТ. Просмотр IRQL

Вы можете просмотреть сохраненный IRQL процессора с помощью команды отладчика `!irql`. Это будет уровень, имевший место непосредственно перед остановкой и переходом в отладчик, который повышает IRQL до фиксированного бессмысленного значения:

```
kd> !irql
Debugger saved IRQL for processor 0x0 -- 0 (LOW_LEVEL)
```

Обратите внимание на то, что значение IRQL сохраняется в двух местах. Первое — это зона контроля процессора (Processor Control Region, PCR), которая отражает текущий IRQL. Его расширение — блок зоны контроля процессора (Processor Region Control Block, PRCB) — в свою очередь, содержит сохраненный IRQL в поле `DebuggerSavedIRQL`. Этот трюк применяется, потому что, пока пользователь отлаживает систему, удаленный отладчик ядра повышает IRQL до `HIGH_LEVEL`, чтобы предотвратить любое асинхронное действие в процессоре, из-за чего показания команды `!irql` не будут иметь смысла. Это «сохраненное значение» и используется для отражения IRQL в момент подключения отладчика.

Каждый уровень прерываний применяется для определенных задач. К примеру, ядро вызывает *межпроцессорное прерывание* (Interprocessor Interrupt, IPI), чтобы потребовать от другого процессора выполнить какое-то действие, скажем передать некий поток на исполнение или обновить его данные в буфере быстрого преобразования адреса (TLB). Системные часы вызывают прерывания с регулярным интервалом, а ядро реагирует на это, обновляя текущее время и замеряя длительность исполнения потоков. HAL предоставляет уровни прерываний для устройств, управляемых через прерывания, точное число уровней зависит от процессора и настроек системы. Ядро использует программные прерывания (подробно о которых рассказывается далее в этой главе) для планировки потоков и асинхронного прерывания исполнения потока.

Распределение векторов прерываний по IRQL

В системах без APIC-архитектуры должно быть задано строгое отображение между GSIV/IRQ и IRQL. Во избежание ситуаций, когда контроллер прерываний может посчитать, что один вектор прерывания имеет больший приоритет, чем другой, в мире Windows уровни IRQL работали прямо противоположным образом. К счастью, при наличии APIC Windows может свободно отражать IRQL как часть TPR APIC, что, в свою очередь, может быть использовано контроллером APIC для более оптимальной маршрутизации. Более того, в системах с APIC приоритет каждого аппаратного прерывания привязан не к его GSIV/IRQ, а к вектору прерывания: верхние 4 бита вектора отражают приоритет. Поскольку IDT может содержать до 256 записей, это позволяет уместить 16 вариантов приоритета (к примеру, вектор `0x40` будет иметь приоритет 4) — на деле те же 16 чисел, умещающихся в TPR, которые отражаются на все те же 16 IRQL, реализуемые в Windows!

Таким образом, чтобы определить, какой IRQL назначить прерыванию, сначала Windows должна подобрать ему подходящий вектор прерывания и запрограммировать APIC ввода/вывода на его использование для соответствующего GSIV

устройства. И наоборот, если какому-то физическому устройству требуется конкретный IRQL, Windows следует подобрать ему вектор прерывания, соответствующий указанному приоритету. Данные решения принимаются диспетчером Plug and Play (PnP) сообща с особым драйвером, называемым *драйвером шины* (bus driver), который определяет наличие устройств (PCI, USB и т. д.) на его шине и то, какие прерывания могут быть к ним привязаны. Драйвер шины сообщает эти сведения диспетчеру Plug and Play, который, приняв во внимание допустимые назначения прерываний для всех остальных устройств, решает, какое прерывание будет назначено каждому устройству. После этого он обращается к арбитру Plug and Play, распределяющему прерывания по IRQL. Сам арбитр при этом находится в распоряжении HAL, который, в свою очередь, работает вместе с драйверами шин ACPI и PCI, чтобы сообща добиться подходящего распределения. В большинстве случаев окончательный номер вектора определяется по принципу циклического кругового перебора, так что вычислить его заранее невозможно. Однако в приведенном далее эксперименте демонстрируется, как отладчик может запросить данную информацию у арбитра прерываний.

Кроме векторов, назначаемых арбитром аппаратным прерываниям, в Windows зарезервировано несколько предопределенных векторов прерываний, которые всегда имеют одинаковый индекс в IDT (табл. 8.4).

Таблица 8.4. Предопределенные векторы прерываний

Вектор	Применение
0x1F	Прерывание APC
0x2F	Прерывание DPC
0x30	Прерывание гипервизора
0x31–0x34	Прерывание (-я) VMBus
0x35	Прерывание CMCI
0xCD	Прерывание термометра
0xCE	Прерывание IOMMU
0xCF	Прерывание DMA
0xD1	Прерывание от системного таймера
0xD2	Межпроцессорное прерывание от системного таймера (IPI)
0xD3	Постоянное прерывание от системного таймера
0xD7	Прерывание перезагрузки
0xD8	Прерывание-заглушка
0xD9	Тестовое прерывание
0xDF	Призрачное (ложное) прерывание
0xE1	Межпроцессорное прерывание (IPI)
0xE2	Прерывание ошибок LAPIC
0xE3	Прерывание DRS
0xF0	Прерывание Watchdog
0xFB	Прерывание HPET гипервизора
0xFD	Прерывание профилирования
0xFE	Прерывание оценки производительности

Здесь видно, что приоритет номера вектора (который, как вы помните, хранится в верхних четырех битах), как правило, соответствует IRQL на рис. 8.14. К примеру, у прерываний APC это 1, а у прерываний DPC — 2, в то время как прерывания IPI имеют уровень 14, а прерывания профилирования — 15. В рамках данной темы рассмотрим, каковы предопределенные IRQL в современной системе Windows.

Предопределенные IRQL

Давайте подробнее разберем, как используются предопределенные IRQL, начиная с наивысшего уровня, показанного на рис. 8.13.

- Как правило, ядро задействует уровень *высший (high)* уровень, только когда требуется остановить систему при исполнении `KeBugCheckEx` с маскированием всех прерываний или когда подключен удаленный отладчик ядра. Тем же значением на не-x86-системах пользуется уровень *профилирования*, на котором при активности данного функционала отсчитывается таймер профилирования. На том же уровне существует прерывание производительности, которое задействуется такими инструментами, как Intel Processor Trace (Intel PT), и прочими аппаратными средствами мониторинга производительности (performance monitoring unit, PMU).
- Уровень *межпроцессорных прерываний (interprocessor interrupt)* применяется для запроса к другому процессу, чтобы тот исполнил некое действие наподобие обновления кэша TLB процессора или перезаписи какого-то контрольного регистра во всех процессорах. Уровень *службы отложенного восстановления (Deferred Recovery Service, DRS)* использует такое же значение. Он применяется в x64-системах в рамках архитектуры аппаратных ошибок Windows (Windows Hardware Error Architecture, WHEA) для восстановления после некоторых исключений проверки машины (Machine Check Errors, MCE).
- Уровень *часов (clock)* предназначен для системных часов, ядро применяет его для отслеживания текущего времени, измерения и распределения процессорного времени между потоками.
- Уровень *синхронизации (synchronization)* нужен для внутреннего применения в коде диспетчера и планировщика потоков, чтобы обеспечить защиту от доступа к глобальному планированию и коду синхронизации/ожидания. Зачастую это наивысший уровень после IRQL для оборудования.
- IRQL для *оборудования (device)* используется для приоритизации прерываний от устройств (подробно о том, как уровни аппаратных прерываний преобразуются в IRQL, говорилось в предыдущем разделе).
- Уровень *устраненной ошибки аппаратного контроля (corrected machine check interrupt)* используется для оповещения операционной системы после существенной, но ликвидированной неисправности оборудования либо информирования об ошибке, переданной процессором или встроенным ПО через интерфейс ошибок аппаратного контроля (Machine Check Error, MCE).
- Уровни *DPC/dispatch* и *APC* служат для программных прерываний, вызываемых ядром и драйверами устройств (более подробно прерывания DPC и APC описаны далее в главе).

- Самый низкий IRQL, или *пассивный (passive)* уровень, по сути, существует лишь формально. Это состояние, в котором происходит нормальное исполнение потоков и могут возникать любые прерывания.

Объекты прерываний

Ядро предоставляет кросс-платформенный механизм — управляющий объект ядра, называемый *объектом прерывания* или *KINTERRUPT*, — который позволяет драйверам регистрировать ISR (служебную функцию прерывания) для своих устройств. Эта структура содержит всю информацию, нужную ядру для сопоставления ISR устройства с конкретным аппаратным прерыванием, включая адрес ISR, направленность и режим активации прерывания, IRQL для прерываний устройства, состояние общего доступа, GSIV и прочие данные контроллера прерывания, а также набор статистики по производительности.

Эти объекты прерываний размещаются в общей области памяти, и когда драйвер регистрирует прерывание, вызывая `IoConnectInterrupt` или `IoConnectInterruptEx`, тот инициализируется сразу со всей необходимой информацией. В зависимости от количества процессоров, которым позволено получать это прерывание (о чем драйвер устройства сообщает, *задавая соответствие для прерывания*), структура KINTERRUPT создается для каждого из них, что в типичном случае означает: для всех процессоров в системе. Затем, как только будет выбран вектор прерывания, массив данных в KPRCB (называемый *InterruptObject*) для каждого подходящего процессора обновляется, чтобы указывать на соответствующий каждому из них экземпляр KINTERRUPT.

После того как KINTERRUPT создан, проводится проверка на предмет того, является ли выбранный вектор прерывания совместно используемым. Если это так, то проверяется, занял ли его ранее другой KINTERRUPT. При положительном результате ядро обновляет в структуре данных KINTERRUPT, в поле `DispatchAddress`, так, чтобы оно указывало на функцию `KiChainedDispatch`, и добавляет этот KINTERRUPT в связный список (`InterruptListEntry`), хранящийся в первом KINTERRUPT, занявшем данный вектор. Если вектор прерывания эксклюзивный, в поле записывается указатель на функцию `KiInterruptDispatch`.

Объект прерывания также хранит связанный с ним IRQL, чтобы функции `KiInterruptDispatch` или `KiChainedDispatch` могли поднимать IRQL до правильного уровня прежде, чем вызвать ISR, а когда тот завершится, сбросить IRQL обратно. Данный двухфазный процесс необходим, так как невозможно передать в объект прерывания какой-либо указатель (да и в целом любой другой параметр) при первоначальном вызове, поскольку тот осуществляется самим оборудованием.

Когда срабатывает прерывание, по IDT вызывается одна из 256 копий функции `KiIsrThunk`, в каждой из которых содержится своя строка ассемблерного кода, помещающая вектор прерывания в стек ядра (потому что процессор не предоставляет своего), а после нее общая функция `KiIsrLinkage`, реализующая остальную обработку. Кроме прочего, последняя строит блок данных ловушки, суть которого описывалась ранее, и наконец передает управление по адресу из KINTERRUPT (одна из двух приведенных ранее функций). Нужный экземпляр KINTERRUPT находится путем обращения к массиву `InterruptObject` в текущем KPRCB

и разыменования соответствующего указателя с помощью вектора прерывания. В случае же, когда KINTERRUPT отсутствует, прерывание считается непредвиденным. В зависимости от значения `BugCheckUnexpectedInterrupts`, находящегося по адресу HKLM\SYSTEM\CurrentControlSet\Control\Session Manager\Kernel в реестре Windows, либо система может аварийно завершить работу с `KeBugCheckEx`, либо прерывание будет проигнорировано, а исполнение продолжится с исходной контрольной точки.

В x64-системах Windows ядро оптимизирует диспетчеризацию с помощью специальных процедур, которые экономят циклы процессора, пропуская ненужные процедуры, как, например, `KiInterruptDispatchNoLock`, применяемая для прерываний, у которых не назначено управляемой ядром спин-блокировки (как правило, используется драйверами, которым требуется синхронизация со своими ISR), `KiInterruptDispatchNoLockNoEtw` для прерываний, которым не требуется отслеживание производительности ETW, `KiSpuriousDispatchNoEOI` для прерываний, которым не нужно подавать сигнал о своем окончании, поскольку они ложные. Наконец, процедура `KiSpuriousDispatchNoEOI`, применяемая для прерываний, которые запрограммированы в APIC как *автозавершаемые* (Auto-EOI). Контроллер отправит сигнал EOI самостоятельно, а значит, ядру не потребуется выполнять для этого лишний код. К примеру, многие обработчики прерывания в составе HAL пользуются неблокирующим кодом перенаправления, поскольку HAL не нужна синхронизация своих ISR с ядром.

Еще один обработчик прерываний ядра — процедура `KiFloatingDispatch`, задействуемая для прерываний, которым требуется сохранение состояния плавающей точки, в отличие от кода режима ядра, которому зачастую не позволено выполнять операции с плавающей точкой (MMX, SSE, 3DNow!): поскольку регистры для них при переключении контекста не сохраняются, ISR они могут потребоваться, к примеру, когда обработчику прерывания от видеокарты требуется выполнить быструю графическую операцию. Регистрируя прерывание, драйверы могут передать в параметре `FloatingSave` значение `TRUE`, запросив у ядра применение соответствующей процедуры диспетчера, которая сохранит регистры для плавающей точки, что, однако, существенно увеличит задержку прерывания. Следует обратить внимание на то, что это поддерживается только в 32-разрядных системах.

Независимо от того, какая процедура диспетчера будет задействована, в конечном счете делается вызов по адресу в поле `ServiceRoutine` объекта KINTERRUPT, где находятся ISR-драйверы. Либо, если речь идет о *прерываниях, инициируемых сообщениями* (message signaled interrupts, MSI), подробнее о которых расскажем в дальнейшем, обработчиком прерывания будет процедура `KiInterruptMessageDispatch`, которая сделает вызов по адресу из поля `MessageServiceRoutine` объекта KINTERRUPT. Стоит заметить, что в некоторых случаях, например при взаимодействии с драйверами в рамках Kernel Mode Driver Framework (KMDF) либо с некоторыми драйверами мини-порта на основе NDIS или StorPort (подробнее о фреймворках для драйверов — в главе 6 тома 1), могут применяться процедуры диспетчера, специфичные для фреймворка, и/или драйверы порта, реализующие дополнительную обработку, прежде чем обратиться к конечному драйверу.

На рис. 8.15 представлен типичный порядок обработки прерываний, связанных с объектами прерываний.

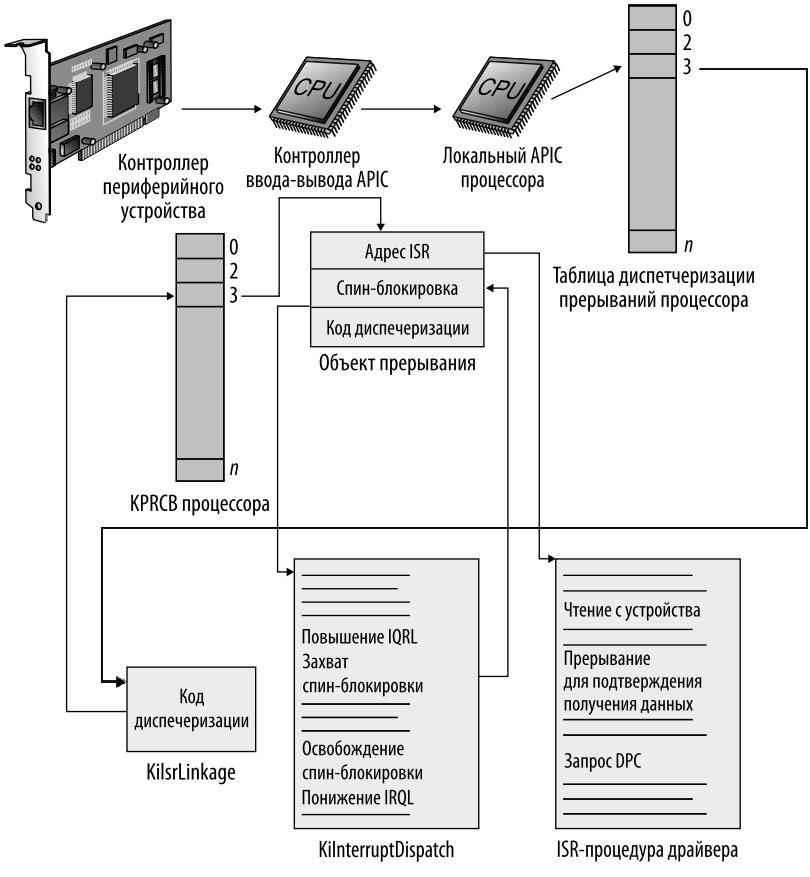


Рис. 8.15. Типовая схема управления прерываниями

Сопоставление ISR с определенным уровнем прерываний называется *подключением объекта прерывания*, а удаление связи между ISR и записью в IDT — его *отключением*. Данные операции, осуществляемые с помощью функций ядра `IoConnectInterruptEx` и `IoDisconnectInterruptEx`, позволяют драйверу устройства «включать» ISR, когда он загружен в систему, и «выключать», если требуется его выгрузить.

Как было показано ранее, использование объектов прерываний для регистрации ISR не позволяет драйверам напрямую обращаться к оборудованию прерываний, которое различно для разных архитектур процессоров, и освобождает от необходимости что-либо знать об IDT. Данный функционал ядра помогает в создании кроссплатформенных драйверов устройств, устраняя потребность в программировании на языке ассемблера или отражении в коде драйвера особенностей различных процессоров. Объекты прерываний тоже дают ряд преимуществ. С их помощью ядро способно синхронизировать исполнение ISR с другими компонентами драйвера устройства, которые могут делить данные с теми же ISR (более подробно о том, как драйверы устройств реагируют на прерывания, говорится в главе 6 тома 1).

Мы также описали принцип *связанной* (chained) обработки, который позволяет ядру легко вызывать более одной ISR на любом уровне прерываний. Если несколько драйверов устройств создадут объекты прерываний и присоединят их к одной записи в IDT, при срабатывании прерывания по соответствующему вектору процедура `KiChainedDispatch` вызовет ISR каждого из них. Эта возможность позволяет ядру свободно управлять *гирляндными* (daisy-chain) конфигурациями, где множество устройств могут делить между собой один вектор прерывания. Цепочка разрывается, когда один из ISR заявляет право на прерывание, вернув диспетчеру прерываний некий статус.

Если нескольким устройствам, использующим одно и то же прерывание, одновременно потребуется обработка, устройства, не признанные своими ISR, снова прервут систему, как только управляющая функция понизит IRQL. Цепочечная обработка допускается лишь тогда, когда все драйверы устройств, которым требуется занять одно и то же прерывание, сообщат ядру, что они готовы им делиться (определяется по полю `ShareVector` объекта KINTERRUPT). В ином случае диспетчер Plug and Play реорганизует назначение им прерываний так, чтобы удовлетворять требованиям каждого.

ЭКСПЕРИМЕНТ. Изучение внутреннего устройства прерываний

Используя отладчик ядра, вы сможете просматривать сведения из объекта прерывания: его IRQL, адрес ISR и специфический код для управления прерыванием. На первом шаге используйте команду отладчика `!idt` и посмотрите, получится ли у вас найти запись, ссылающуюся на `I8042KeyboardInterruptService`, ISR-обработчик для клавиатур по порту PS/2. Если такой у вас нет, можете поискать записи, указывающие на Stornvme.sys, Scsiport.sys или любой другой знакомый вам сторонний драйвер. Работая с виртуальной машиной Hyper-V, можете обойтись записью Acpi.sys. Далее приводится пример для системы с клавиатурой PS/2:

```
70:    fffff8045675a600 i8042prt!I8042KeyboardInterruptService
       (KINTERRUPT ffff8e01cbe3b280)
```

Чтобы просмотреть содержимое объекта, связанного с прерыванием, достаточно просто щелкнуть на ссылке, которую вам выдаст отладчик, в результате чего исполнится команда `dt`. С тем же успехом вы можете вручную ввести команду dx. Далее приводится содержимое KINTERRUPT с машины, на которой проводился эксперимент:

```
6: kd> dt nt!_KINTERRUPT ffff8e01cbe3b280
    +0x000 Type             : 0n22
    +0x002 Size             : 0n256
    +0x008 InterruptListEntry : _LIST_ENTRY
                                [ 0x00000000`00000000 - 0x00000000`00000000 ]
    +0x018 ServiceRoutine   : 0xfffff804`65e56820
                              unsigned char i8042prt!I8042KeyboardInterruptService
    +0x020 MessageServiceRoutine : (null)
    +0x028 MessageIndex     : 0
    +0x030 ServiceContext   : 0xffffe50f`9dfe9040 Void
    +0x038 SpinLock         : 0
```

```
    +0x040 TickCount       : 0
    +0x048 ActualLock      : 0xffffe50f`9dfe91a0 -> 0
    +0x050 DispatchAddress : 0xfffff804`565ca320  void nt!KiInterruptDispatch+0
    +0x058 Vector          : 0x70
    +0x05c Irql            : 0x7 ''
    +0x05d SynchronizeIrql : 0x7 ''
    +0x05e FloatingSave    : 0 ''
    +0x05f Connected       : 0x1 ''
    +0x060 Number          : 6
    +0x064 ShareVector     : 0 ''
    +0x065 EmulateActiveBoth : 0 ''
    +0x066 ActiveCount     : 0
    +0x068 InternalState   : 0n4
    +0x06c Mode            : 1 ( Latched )
    +0x070 Polarity        : 0 ( InterruptPolarityUnknown )
    +0x074 ServiceCount    : 0
    +0x078 DispatchCount   : 0
    +0x080 PassiveEvent    : (null)
    +0x088 TrapFrame       : (null)
    +0x090 DisconnectData  : (null)
    +0x098 ServiceThread   : (null)
    +0x0a0 ConnectionData  : 0xffffe50f`9db3bd90 _INTERRUPT_CONNECTION_DATA
    +0x0a8 IntTrackEntry   : 0xffffe50f`9d091d90 Void
    +0x0b0 IsrDpcStats     : _ISRDPCSTATS
    +0x0f0 RedirectObject  : (null)
    +0x0f8 Padding         : [8] ""
```

В данном примере Windows назначила этому прерыванию IRQL 7, что соответствует тому факту, что вектор прерывания равен `0x70` (и старшие 4 бита содержат 7). Далее в поле `DispatchAddress` видно, что здесь имеет место обычное прерывание с обработкой через `KiInterruptDispatch` без дополнительных оптимизаций или поддержки разделения.

Существует два способа узнать, какой номер запроса (IRQ) был связан с данным прерыванием. Первый заключается в следующем. В недавних версиях Windows IRQ хранятся в структуре `INTERRUPT_CONNECTION_DATA`, включенной в состав KINTERRUPT, в поле `ConnectionData`, что можно увидеть в приведенном ранее выводе. С помощью команды `dt` вы сможете выгрузить данные, к которым ведет этот указатель, следующим образом:

```
6: kd> dt 0xffffe50f`9db3bd90 _INTERRUPT_CONNECTION_DATA Vectors[0]..
nt!_INTERRUPT_CONNECTION_DATA
   +0x008 Vectors      : [0]
      +0x000 Type         : 0 ( InterruptTypeControllerInput )
      +0x004 Vector       : 0x70
      +0x008 Irql         : 0x7 ''
      +0x00c Polarity     : 1 ( InterruptActiveHigh )
      +0x010 Mode         : 1 ( Latched )
      +0x018 TargetProcessors :
      +0x000 Mask         : 0xff
      +0x008 Group        : 0
      +0x00a Reserved     : [3] 0
   +0x028 IntRemapInfo :
      +0x000 IrtIndex     : 0y00000000000000000000000000000000 (0)
      +0x000 FlagHalInternal : 0y0
```

```
      +0x000 FlagTranslated : 0y0
      +0x004 u              : <anonymous-tag>
   +0x038 ControllerInput :
      +0x000 Gsiv           : 1
```

По полю Type видно, что перед нами традиционный ввод конвейера/контроллера, а поля Vector и Irql подтверждают данные, которые мы ранее видели в KINTERRUPT. Затем в структуре ControllerInput вы можете увидеть, что GSIV равен 1 (то есть IRQ 1). Если же просматриваете прерывание иного типа, к примеру прерывание, инициированное сообщением (подробнее о которых мы поговорим далее), вам стоит разыменовать поле MessageRequest.

Для реализации другого способа определить связь между GSIV и векторами прерываний достаточно вспомнить, что Windows отслеживает ее, когда управляет ресурсами устройств через так называемых *арбитров*. Для каждого типа ресурсов существует арбитр, который поддерживает соответствие между использованием ресурсов виртуальных (таких как вектор прерывания) и физических (таких как конвейер прерывания). Таким образом, вы можете обратиться к арбитру IRQ ACPI и узнать соответствие. Для получения этих данных выполните команду !apciirqarb:

```
6: kd> !acpiirqarb

Processor 0 (0, 0):
Device Object: 0000000000000000
Current IDT Allocation:
...
000000070 - 00000070 D ffffe50f9959baf0 (i8042prt) A:ffffce0717950280 IRQ(GSIV):1
...
```

Обратите внимание на то, что GSIV клавиатуры — это IRQ 1, сохранившееся по сей день число, пришедшее к нам со времен PC/AT. Также вы можете дать команду !arbiter 4 (4 говорит отладчику показать только арбитров IRQ), чтобы увидеть непосредственно запись в арбитре IRQ ACPI:

```
6: kd> !arbiter 4

DEVNODE ffffe50f97445c70 (ACPI_HAL\PNP0C08\0)
    Interrupt Arbiter "ACPI_IRQ"" at fffff804575415a0
        Allocated ranges:
            0000000000000001 - 0000000000000001 ffffe50f9959baf0 (i8042prt)
```

В таком случае обратите внимание на то, что показанный диапазон представляет именно GSIV (IRQ), а не вектор прерывания. Кроме того, в обоих листингах отобразился владелец вектора в виде *объекта устройства* (в данном случае это 0xFFFFE50F9959BAF0). Зная это, вы сможете использовать команду !devobj, чтобы получить сведения о самом устройстве (в данном примере это i8042prt, который соответствует драйверу порта PS/2):

```
6: kd> !devobj 0xFFFFE50F9959BAF0
Device object (ffffe50f9959baf0) is for:
    00000049 \Driver\ACPI DriverObject ffffe50f974356f0
Current Irp 00000000 RefCount 1 Type 00000032 Flags 00001040
```

```
SecurityDescriptor ffffce0711ebf3e0 DevExt ffffe50f995573f0 DevObjExt
ffffe50f9959bc40
DevNode ffffe50f9959e670
ExtensionFlags (0x00000800) DOE_DEFAULT_SD_PRESENT
Characteristics (0x00000080) FILE_AUTOGENERATED_DEVICE_NAME
AttachedDevice (Upper) ffffe50f9dfe9040 \Driver\i8042prt
Device queue is not busy.
```

Объект устройства связан с *узлом оборудования*, где хранятся данные обо всех физических ресурсах устройства. Теперь вы сможете выгрузить их с помощью команды `!devnode`. Если при этом передать флаг `0xF`, будут выведены как сырые данные, так и оформленный список:

```
6: kd> !devnode ffffe50f9959e670 f
DevNode 0xffffe50f9959e670 for PDO 0xffffe50f9959baf0
    InstancePath is "ACPI\LEN0071\4&36899b7b&0"
    ServiceName is "i8042prt"
    TargetDeviceNotify List - f 0xffffce0717307b20 b 0xffffce0717307b20
    State = DeviceNodeStarted (0x308)
    Previous State = DeviceNodeEnumerateCompletion (0x30d)
    CmResourceList at 0xffffce0713518330 Version 1.1 Interface 0xf Bus #0
        Entry 0 - Port (0x1) Device Exclusive (0x1)
            Flags (PORT_MEMORY PORT_IO 16_BIT_DECODE
            Range starts at 0x60 for 0x1 bytes
        Entry 1 - Port (0x1) Device Exclusive (0x1)
            Flags (PORT_MEMORY PORT_IO 16_BIT_DECODE
            Range starts at 0x64 for 0x1 bytes
        Entry 2 - Interrupt (0x2) Device Exclusive (0x1)
            Flags (LATCHED
            Level 0x1, Vector 0x1, Group 0, Affinity 0xffffffff
...
    TranslatedResourceList at 0xffffce0713517bb0 Version 1.1 Interface 0xf Bus #0
        Entry 0 - Port (0x1) Device Exclusive (0x1)
            Flags (PORT_MEMORY PORT_IO 16_BIT_DECODE
            Range starts at 0x60 for 0x1 bytes
        Entry 1 - Port (0x1) Device Exclusive (0x1)
            Flags (PORT_MEMORY PORT_IO 16_BIT_DECODE
            Range starts at 0x64 for 0x1 bytes
        Entry 2 - Interrupt (0x2) Device Exclusive (0x1)
            Flags (LATCHED
            Level 0x7, Vector 0x70, Group 0, Affinity 0xff
```

Узел оборудования говорит о том, что у устройства есть список ресурсов из трех пунктов, один из которых является записью о прерывании, соответствующей IRQ 1 (номера *уровня* и *вектора* представляют GSIV, а не вектор прерывания). Далее виден оформленный список ресурсов, где уже указан IRQL 7 (это номер *уровня*) и вектор прерывания `0x70`.

В системах с ACPI эту информацию получить немного легче, прочитав расширенный вывод представленной ранее команды `!acpiirqarb`. Кроме прочего, она отображает таблицу соответствия IRQ и IDT:

```
Interrupt Controller (Inputs: 0x0-0x77):
     (01)Cur:IDT-70 Ref-1 Boot-0 edg hi     Pos:IDT-00 Ref-0 Boot-0 lev unk
     (02)Cur:IDT-80 Ref-1 Boot-1 edg hi     Pos:IDT-00 Ref-0 Boot-1 lev unk
```

```
(08)Cur:IDT-90 Ref-1 Boot-0 edg hi     Pos:IDT-00 Ref-0 Boot-0 lev unk
(09)Cur:IDT-b0 Ref-1 Boot-0 lev hi     Pos:IDT-00 Ref-0 Boot-0 lev unk
(0e)Cur:IDT-a0 Ref-1 Boot-0 lev low    Pos:IDT-00 Ref-0 Boot-0 lev unk
(10)Cur:IDT-b5 Ref-2 Boot-0 lev low    Pos:IDT-00 Ref-0 Boot-0 lev unk
(11)Cur:IDT-a5 Ref-1 Boot-0 lev low    Pos:IDT-00 Ref-0 Boot-0 lev unk
(12)Cur:IDT-95 Ref-1 Boot-0 lev low    Pos:IDT-00 Ref-0 Boot-0 lev unk
(14)Cur:IDT-64 Ref-2 Boot-0 lev low    Pos:IDT-00 Ref-0 Boot-0 lev unk
(17)Cur:IDT-54 Ref-1 Boot-0 lev low    Pos:IDT-00 Ref-0 Boot-0 lev unk
(1f)Cur:IDT-a6 Ref-1 Boot-0 lev low    Pos:IDT-00 Ref-0 Boot-0 lev unk
(41)Cur:IDT-96 Ref-1 Boot-0 edg hi     Pos:IDT-00 Ref-0 Boot-0 lev unk
```

Как и ожидалось, IRQ 1 связано с записью IDT 0x70. Более подробное описание объектов устройств, ресурсов и других связанных с ними структур см. в главе 6 тома 1.

Конвейерные и инициируемые сообщениями прерывания

Совместно используемые (shared) прерывания зачастую становятся причиной значительных задержек при обработке и даже могут нарушать стабильность. В большинстве случаев их наличие нежелательно и часто является побочным эффектом ограниченного количества конвейеров прерываний на компьютере. К примеру, если взять кардридер «4 в 1», способный работать с USB, Compact Flash, Sony Memory Stick, Secure Digital и другими форматами, все контроллеры, являясь частью одного физического устройства, как правило, будут присоединены к одному конвейеру прерываний, который затем будет настроен рядом различных драйверов устройств на общий вектор прерывания. Это приведет к дополнительным задержкам, поскольку каждый обработчик будет вызываться последовательно, пока не будет определено, какой именно контроллер вызвал прерывание для своего носителя.

Гораздо лучшим решением было бы дать каждому контроллеру устройства отдельное прерывание, а один драйвер работал бы с разными прерываниями, зная, от какого устройства они поступили. Однако если занять сразу четыре традиционных конвейера прерываний ради одного устройства, это быстро приведет к их исчерпанию. Кроме того, каждое устройство на шине PCI так или иначе подсоединяется к отдельному конвейеру, так что кардридер не смог бы использовать более одного IRQ в любом случае.

Другие проблемы со срабатыванием прерываний через конвейер IRQ возникают из-за того, что недочеты при обработке сигналов IRQ могут привести к наплыву прерываний или к другим зависаниям на машине, поскольку сигнал делается «высоким» или «низким», пока какой-то из ISR не признает его. (Кроме того, контроллеру прерываний надо дожидаться сигнала EOI.) Если что-то из этого происходит не из-за ошибки в программе, система может застрять в состоянии прерывания навсегда, все последующие прерывания могут оказаться замаскированными, а может, произойдет и то и другое. Наконец, конвейерные прерывания плохо масштабируются в многопроцессорной среде. Во множестве случаев последнее слово о том, какой процессор из набора, определенного диспетчером Plug and Play, будет выбран для данного прерывания, остается за оборудованием, и драйверы устройств тут мало что могут сделать.

Решение всех этих проблем впервые было представлено в стандарте PCI 2.2 под названием *«прерывания, инициируемые сообщениями»* (message-signaled interrupts, MSI). Хотя изначально оно было необязательным компонентом стандарта и в основном встречалось на клиентских компьютерах (а на серверах зачастую служило для повышения производительности сетевых адаптеров и контроллеров хранилищ данных), большинство современных систем благодаря стандартам, начиная с PCI Express 3.0, приняли эту модель во всей ее полноте. В реалиях MSI устройство отправляет своему драйверу сообщение, записывая его в особую область памяти через шину PCI. Фактически с точки зрения оборудования это считается операцией прямого доступа к памяти (Direct Memory Access, DMA). Данное действие вызывает срабатывание прерывания, после чего Windows вызывает ISR, передавая туда содержимое (значение) сообщения и адрес, по которому оно было доставлено. Кроме того, устройство может отправить несколько сообщений (до 32) на этот адрес, доставляя различную полезную нагрузку (payload) в зависимости от события.

Для систем, еще более чувствительных к задержкам и производительности, в стандарте PCI 3.0 было представлено расширение MSI, известное как MSI-X, где сообщения стали 32-разрядными (до этого были 16-разрядными), их предельное количество выросло до 2048 (вместо всего 32) и, что еще важнее, появилась возможность использовать различные адреса (их можно определять динамически) для каждого вида полезной нагрузки, приходящей с MSI. Применение этих адресов дает возможность помещать присланные данные в различные физические диапазоны адресов, принадлежащие разным процессорам или даже разным наборам процессоров, по сути позволяя доставлять прерывания с неравномерным доступом к памяти (nonuniform memory access, NUMA), где за обработку берется процессор, инициировавший запрос к соответствующему устройству. Отслеживание и загрузки, и ближайшего узла NUMA при обработке прерывания позволяет улучшить и задержки, и масштабирование.

В обеих моделях взаимодействие строится вокруг значения в памяти, а контент доставляется через прерывание. Из-за этого пропадает как нужда в конвейерах прерываний (что приравнивает системный предел MSI к числу векторов прерываний, а не конвейеров), так и необходимость для ISR-драйвера отдельно запрашивать у устройства данные, связанные с прерыванием, что снижает задержки. Ввиду большого количества прерываний устройств, обеспечиваемого такой моделью, разделение прерываний фактически становится бессмысленным. Доставка данных, связанных с прерыванием, напрямую заинтересованному ISR еще больше снижает задержки.

Это является одной из причин того, почему в данном тексте, как и в большинстве команд отладчика, чаще употребляется термин не IRQ, а GSIV, лучше подходящий на роль общего определения вектора MSI (который обозначается отрицательным числом), традиционного конвейера IRQ или даже контакта интерфейса ввода/вывода общего назначения (General Purpose Input Output, GPIO) на встроенном устройстве. Кроме того, в системах ARM и ARM64 не используется ни та ни другая модель, а вместо этого задействуется универсальный контроллер прерываний (Generic Interrupt Controller, GIC). На рис. 8.16 представлен диспетчер устройств на двух системах с традиционным назначением GSIV через IRQ- и MSI-значения соответственно.

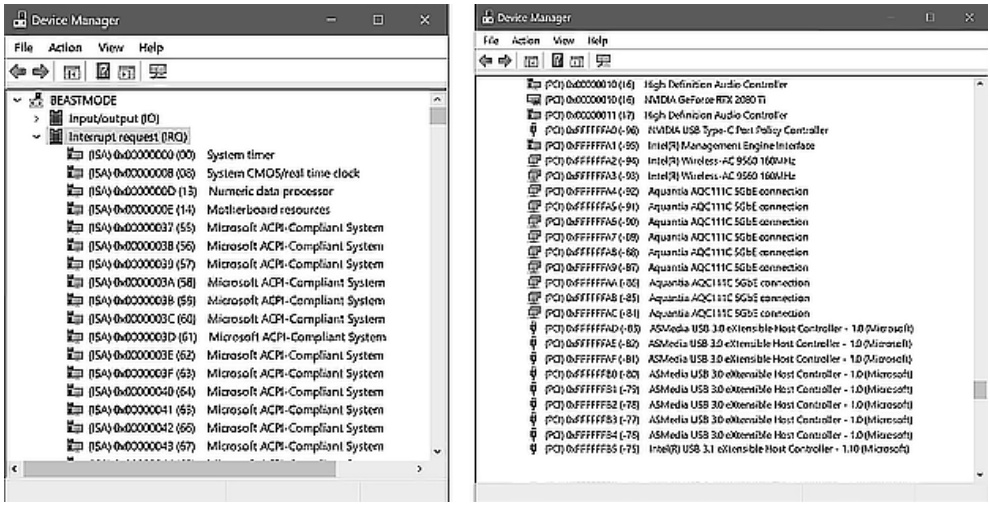


Рис. 8.16. Среды GSIV на основе IRQ и MSI

Перестраивание прерываний

В клиентских системах (не считая серверных SKU), работающих без виртуализации, где размер группы процессоров от 2 до 16, Windows задействует функциональность под названием *«перестраивание прерываний»* (interrupt steering), предназначенную для оптимизации задержек и энергопотребления в современных потребительских системах. С ее помощью нагрузка прерываниями может быть по необходимости распределена между процессорами, что позволяет избежать затора на отдельном ядре. Кроме того, механизм парковки ядер, описанный в главе 6 тома 1, способен перестраивать прерывания *вдали* от запаркованных ядер, не позволяя распределению держать слишком много процессоров активными одновременно.

Возможности перестраивания прерываний различаются в зависимости от контроллера прерываний. Например, на ARM-системах с GIC перестраиваться могут как зависящие от уровня, так и частотные прерывания, в то время как в системах с APIC (кроме тех, что работают внутри Hyper-V) это возможно только для зависящих от уровня прерываний. К сожалению, в силу того, что срабатывание MSI всегда зависит от уровня, преимущества данной технологии не так уж явны, поэтому в Windows реализована также дополнительная модель *перенаправления прерываний* специально для таких ситуаций.

Когда *перестраивание* разрешено, контроллер прерываний попросту перепрограммируется на доставку GSIV на LAPIC другого процессора (в терминах GIC из ARM — на эквивалент). Когда необходимо *перенаправление*, все процессоры становятся адресатами GSIV и всякий процессор, получивший прерывание, вручную отправляет IPI тому процессору, на который оно должно перестроиться.

Кроме использования перестраивания прерываний для нужд механизма парковки ядер, Windows предоставляет данную функциональность через класс сведений системной среды, управляемый функцией `KeIntSteerAssignCpuSetForGsiv` в рамках возможностей Real-Time Audio в Windows 10, и задание соответствия процессору,

описанное в разделе «Планирование потоков» главы 4 тома 1. Это позволяет определенным GSIV перестраиваться на конкретную группу процессоров, которая может быть выбрана приложением пользовательского режима, если оно обладает привилегией *увеличения базового приоритета* (Increase Base Priority), обычно присваиваемой только администраторам и локальным служебным учетным записям.

Приоритет и задание соответствия прерываний

Windows позволяет разработчикам драйверов и администраторам в некоторой степени контролировать задание соответствия процессорам (выбирать, какой процессор или группа процессоров получат прерывание). Более того, это дает примитивный инструмент приоритизации прерываний в зависимости от выбора IRQL. Политика задания соответствия определяется согласно табл. 8.5 и может быть перенастроена с помощью значения `InterruptPolicyValue` в реестре Windows по адресу Interrupt Management\Affinity Policy в разделе искомого устройства. Благодаря этому нет нужды что-то программировать — администратору достаточно добавить это значение в ключ нужного драйвера, тем самым повлияв на его поведение. Задание соответствия прерываний задокументировано в Microsoft Docs по адресу https://docs.microsoft.com/en-us/windows-hardware/drivers/kernel/interrupt-affinity-and-priority.

Таблица 8.5. Политики задания соответствия IRQ

Политика	Значение
IrqPolicyMachineDefault	Устройство не требует определенной политики задания соответствия. Windows использует политику, которая принята на данном компьютере по умолчанию и предполагает возможность выбрать любой доступный процессор (на компьютерах, где число логических процессоров менее 8)
IrqPolicyAllCloseProcessors	На компьютерах с NUMA диспетчер Plug and Play назначает прерывание всем процессорам вблизи устройства (на том же узле). На компьютере без NUMA — аналогично политике IrqPolicyAllProcessorsInMachine
IrqPolicyOneCloseProcessor	На компьютерах с NUMA диспетчер Plug and Play назначает прерывание одному процессору вблизи устройства (на том же узле). На компьютерах без NUMA процессор выбирается среди всех доступных в системе
IrqPolicyAllProcessorsInMachine	Прерывание может быть обработано любым доступным процессором в системе
IrqPolicySpecifiedProcessors	Прерывание может быть обработано только одним из процессоров, подпадающих под маску соответствия, указанную в значении AssignmentSetOverride в реестре Windows
IrqPolicySpreadMessagesAcrossAllProcessors	Различные прерывания, инициируемые сообщением, распределяются в рамках оптимального набора подходящих процессоров, по возможности отслеживаются нюансы топологии NUMA. И устройство, и платформа должны поддерживать MSI-X
IrqPolicyAllProcessorsInGroupWhenSteered	Прерывание может быть перестроено, в связи с чем оно будет внесено в IDT всех процессоров, один из которых будет выбираться для обработки динамически в зависимости от правил перестраивания

Кроме настройки политики сопоставления, можно еще одно значение в реестре использовать для задания приоритета прерывания (табл. 8.6).

Таблица 8.6. Приоритеты IRQ

Приоритет	Значение
IrqPriorityUndefined	Устройство не нуждается в конкретном приоритете. Ему будет назначен приоритет по умолчанию (IrqPriorityNormal)
IrqPriorityLow	Устройство не пострадает от больших задержек и будет иметь IRQL ниже, чем обычно (3 или 4)
IrqPriorityNormal	Устройство ожидает средних задержек. Оно получает IRQL, по умолчанию назначенный его вектору прерывания (от 5 до 11)
IrqPriorityHigh	Устройство требует минимальных задержек. Оно получает IRQL выше обычного (12)

Как обсуждалось ранее, важно учитывать, что Windows не является операционной системой реального времени, ввиду чего приведенные приоритеты IRQ — только подсказки для системы, контролирующей лишь связанный с прерыванием IRQL, и не дают никакого приоритета сверх того, что может дать механизм задания схем приоритета Windows. В силу того что приоритет IRQ еще и хранится в реестре, администраторы вправе свободно менять эти значения для драйверов, если вдруг возникнет потребность снизить задержки для драйвера, не пользующегося данной возможностью.

Программные прерывания

Хотя большинство прерываний генерируется оборудованием, для ряда задач ядро Windows использует программные прерывания, в частности:

- инициирование управления потоками;
- некритичную по времени обработку прерываний;
- истечение времени обработки;
- асинхронное исполнение процедуры в контексте конкретного потока;
- поддержку асинхронных операций ввода/вывода.

Эти задачи описываются в следующих разделах.

Прерывания отложенного вызова процедур

Как правило, *отложенным вызовом процедур* (deferred procedure call, DPC) называется связанная с прерыванием функция, которая выполняет какую-то задачу по обработке уже после того, как все прерывания устройства были обработаны. Такие функции называют отложенными, так как они не всегда исполняются сразу. Ядро использует DPC для обработки истечения таймеров (чтобы отпускать потоки, ожидающие таймер) и перепланирования работы процессора после истечения кванта у потока (следует заметить, что это происходит на IRQL для DPC, но не через обычное ядро DPC). Драйверы устройств применяют DPC для обработки прерываний и выполнения операций, недоступных на более высоких

IRQL. Чтобы вовремя обслуживать аппаратные прерывания, Windows совместно с драйверами с устройств пытается сохранять IRQL ниже уровня прерываний для устройств. Один из способов достижения этой цели — выполнение ISR-обработчиками драйверов необходимого минимума действий, чтобы признать свое устройство, сохранить состояние прерывания, а затем отложить передачу данных и прочие некритичные по времени задачи обработки прерывания на выполнение через DPC на уровне DPC/dispatch (система ввода/вывода подробно описана в главе 6 тома 1).

В случае, когда уровень IRQL задан как пассивный или APC, вызовы DPC будут исполняться немедленно и блокировать любую обработку, не касающуюся оборудования, в связи с чем они также часто используются, чтобы принудительно исполнять высокоприоритетный системный код. Это дает операционной системе возможность сгенерировать прерывание и исполнить системную функцию в режиме ядра. Например, когда какой-то поток не может продолжить исполнение, возможно, из-за необходимости его завершить или из-за того, что тот добровольно перешел в состояние ожидания, ядро напрямую вызывает диспетчер, чтобы немедленно выполнить переключение контекста. Однако иногда ядро обнаруживает, что перепланировку следует провести, когда поток ушел в глубь множества слоев кода. В таких ситуациях ядро запрашивает диспетчеризацию, но откладывает ее до тех пор, пока не закончит свою текущую задачу. Программные прерывания DPC — удобный инструмент для подобной отложенной обработки.

Когда ядру требуется синхронизировать доступ к своим структурам, связанным с планированием, оно всегда повышает IRQL процессора до уровня DPC или выше. Это запрещает дополнительные программные прерывания и управление потоками. Когда ядро обнаруживает, что диспетчеризация нужна, оно запрашивает прерывание уровня DPC, но, поскольку IRQL уже на этом уровне или выше, процессор его откладывает. Когда ядро заканчивает текущую работу, оно обеспечивает понижение IRQL ниже DPC и проверяет, не стоят ли в очереди какие-либо управляющие прерывания. Если они есть, IRQL сбрасывается до DPC и они обрабатываются. Активация диспетчера потоков через программное прерывание позволяет отложить управление до наступления подходящих условий. Прерывание DPC представлено в системе в виде *объекта DPC* — контрольной структуры данных ядра, которая не видна программам пользовательского режима, но доступна драйверам устройств и системному коду. Самая важная информация в ней — это адрес системной функции, которую ядро вызовет для обработки прерывания DPC. Пока такие функции ожидают выполнения, они записываются в обслуживаемые ядром очереди, по одной на процессор, которые называются *очередями DPC*. Чтобы запросить DPC, системный код обращается к ядру, которое инициализирует объект DPC и помещает его в очередь DPC.

По умолчанию ядро ставит объекты DPC в конец одной из двух очередей DPC, принадлежащих тому процессору, на котором их создание было запрошено (как правило, это процессор, исполнявший ISR). Однако драйвер устройства может изменить это поведение, указав у DPC приоритет (низкий, средний, умеренно высокий или высокий, по умолчанию — средний) и целевой процессор. Если DPC имеет целью конкретный процессор, его называют *направленным*. Если приоритет высокий, объект помещается в начало очереди, при всех других приоритетах он встает в ее конец.

Когда IRQL процессора вот-вот снизится с DPC или выше до более низкого уровня (APC или пассивный), ядро обрабатывает все DPC. Windows обеспечивает удержание IRQL на уровне DPC и достает объекты из очереди текущего процессора, пока та не опустеет (иными словами, ядро «опустошает» очередь), по порядку вызывая функции каждого DPC. Лишь когда очередь исчезнет, ядро позволит IRQL опуститься ниже уровня DPC, а обычным потокам — продолжить исполнение. Процесс работы с DPC приведен на рис. 8.17.

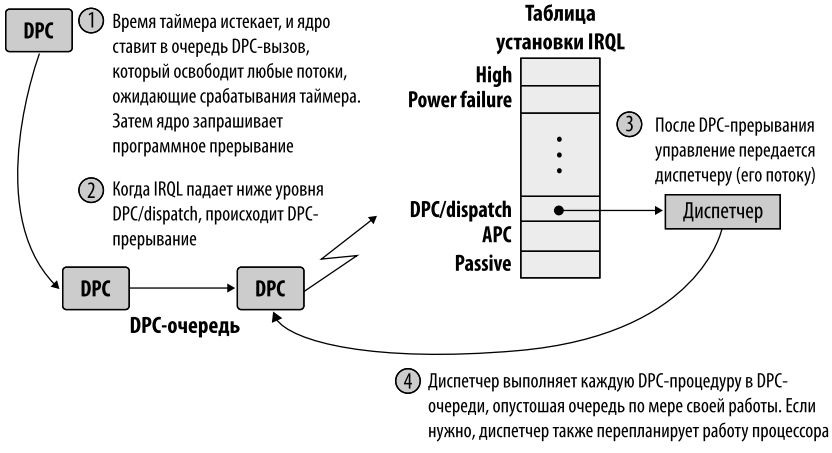


Рис. 8.17. Доставка DPC-вызова

Приоритеты DPC могут влиять на поведение системы и иначе. Обычно ядро запускает опустошение очереди через прерывание уровня DPC. Это происходит лишь тогда, когда вызов направлен на текущий процессор (тот, который исполняет IRS) и его приоритет выше низкого. При низком приоритете DPC ядро запрашивает прерывание, только если количество значимых запросов DPC (хранится в поле `DpcQueueDepth` в KPRCB) к этому процессору превышает определенный максимум (известный как `MaximumDpcQueueDepth` в KPRCB) либо число DPC в рамках некоторого временного окна низкое.

Если DPC направлен на процессор, отличный от того, где выполнялся ISR, а его приоритет умеренно высокий или высокий, ядро немедленно подает туда сигнал (посылая управляющее IPI) опустошить свою очередь DPC, но только если тот бездействует. Если же приоритет средний или низкий, число DPC в очереди на целевой процессор (опять же `DpcQueueDepth`) должно превысить лимит (`MaximumDpcQueueDepth`), иначе ядро не запустит прерывания. Поток бездействия системы также опустошает очередь DPC на тех процессорах, где он исполняется. Хотя уровни приоритетов DPC и их направленность гибкие, драйверам редко требуется изменять обычное поведение своих DPC-объектов. В табл. 8.7 приводятся основные ситуации, вызывающие опустошение очереди. Умеренно высокий и высокий приоритеты выглядят и являются практически аналогичными с точки зрения правил генерации прерывания. Разница лишь в том, что первый помещают в конец очереди, а второй — в начало.

Таблица 8.7. Правила генерации прерываний DPC

Приоритет DPC	DPC направлен на текущий процессор	DPC направлен на другой процессор
Низкий (Low)	Длина DPC-очереди превышает максимальную или частота DPC-запросов ниже минимума	Длина DPC-очереди выше максимума или система простаивает
Средний (Medium)	Всегда	Длина DPC-очереди выше максимума или система простаивает
Выше среднего (Medium-High)	Всегда	Целевой процессор простаивает
Высокий (High)	Всегда	Целевой процессор простаивает

Дополнительно в табл. 8.8 описываются различные переменные для настройки DPC и указывается, как их можно изменить через реестр. Также эти значения можно изменить, минуя реестр, воспользовавшись классом сведений о системе `SystemDpcBehaviorInformation`.

Таблица 8.8. Параметры генерации прерываний DPC

Переменная	Определение	Значение по умолчанию	Переназначение
KiMaximumDpcQueueDepth	Количество DPC в очереди, по превышении которого прерывание будет послано для DPC среднего приоритета или ниже	4	DpcQueueDepth
KiMinimumDpcRate	Количество DPC за такт часов, при превышении которого DPC низкого приоритета не вызовут прерывания	3	MinimumDpcRate
KiIdealDpcRate	Количество DPC за такт часов, прежде чем максимальная глубина очереди DPC будет уменьшена, если DPC находятся в ожидании, но прерывания не было	20	IdealDpcRate
KiAdjustDpcThreshold	Количество тактов часов, прежде чем максимальная глубина очереди DPC будет увеличена при отсутствии DPC в ожидании	20	AdjustDpcThreshold

Поскольку потоки пользовательского режима исполняются с низким IRQL, велика вероятность того, что такой поток будет прерван DPC. Обработчики DPC исполняются без оглядки на то, какой поток сейчас выполняется, в связи с чем их код не может предположить, адресное пространство какого процесса сейчас отображено. Они могут вызывать функции ядра, но не могут обращаться к системным службам, вызывать отказ

страницы, создавать или ожидать объекты диспетчера (подробнее — далее в данной главе). Однако они имеют доступ к адресам нестраничной системной памяти, поскольку адресное пространство системы всегда отображается независимо от текущего процесса.

Поскольку вся память пользовательского режима странична, а DPC обрабатываются в контексте неизвестного процесса, вызываемый ими код никогда и никоим образом не должен обращаться к памяти пользовательского режима. В системах, где поддерживаются предотвращение доступа к коду режима ядра к пользовательской части адресного пространства (Supervisor Mode Access Protection, SMAP) или механизм Privileged Access Never (PAN), Windows задействует эти инструменты на протяжении всей обработки очереди DPC и исполнения вызываемых ими подпрограмм, гарантируя, что любая попытка доступа к памяти пользовательского режима немедленно приведет к критической ошибке системы *(bugcheck)*.

Еще одним побочным эффектом прерываний DPC, вклинивающихся в работу потоков, является то, что они в итоге крадут их время: пока планировщик считает, что выполняется текущий поток, на самом деле исполняется обработчик DPC. В главе 4 тома 1 мы обсуждали механизмы, которые планировщик использует, чтобы возместить потерянное время, отсчитывая точное число циклов процессора, затраченных на исполнение потока, и вычитая время на ISR и DPC, если это возможно.

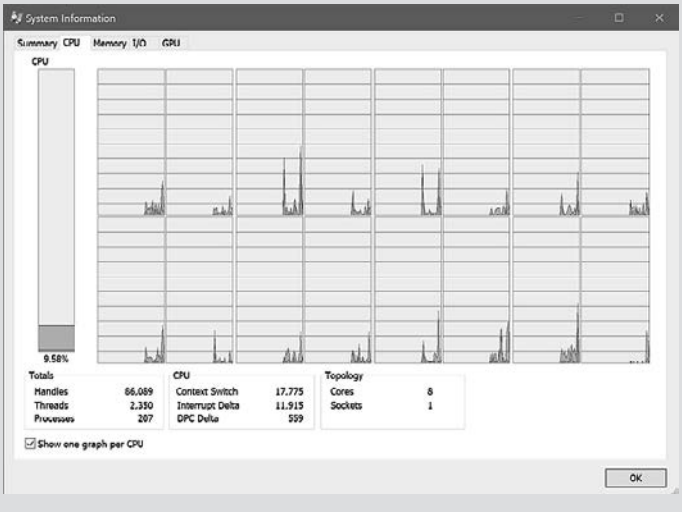
Пусть такие меры и гарантируют, что поток не понесет потерь в рамках своего кванта, это не отменяет того факта, что с точки зрения пользователя *фактическое время* (иными словами, *время*, прошедшее в реальной жизни) все еще тратится на что-то другое. Представьте себе ситуацию, когда пользователь слушает любимую песню в Интернете: если обработка DPC займет 2 с, на это время либо воспроизведение остановится, либо станет повторяться небольшой фрагмент. Подобное будет заметно при воспроизведении видео и даже при вводе с клавиатуры или с помощью мыши. Из-за этого DPC оказываются основной причиной неотзывчивости системы на внешнее воздействие на клиентских компьютерах или загруженности рабочих станций, ведь обработку DPC не сможет остановить даже поток с высочайшим приоритетом. Для работы с драйверами с медленными DPC Windows поддерживает *поточные DPC*. Поточные вызовы, как следует из их названия, опираются на исполнение своего обработчика на пассивном уровне, но с приоритетом реального времени. Это позволяет DPC обогнать большинство пользовательских потоков (поскольку чаще всего потоки приложений не применяются приоритетами реального времени), но дает возможность прерываниям, непоточным DPC, вызовам APC и прочим потокам с приоритетом 31 обогнать его самого.

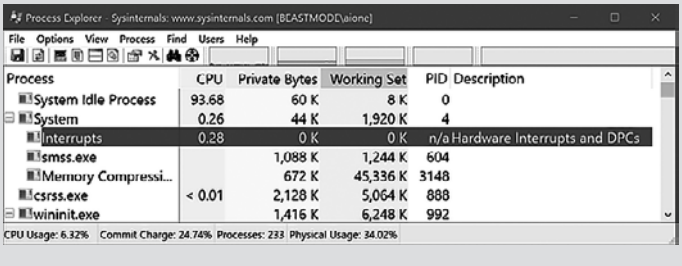
Механика поточных DPC по умолчанию активна, но ее можно отключить, добавив DWORD-параметр `ThreadDpcEnable` в раздел реестра (находится по адресу `HKEY_LOCAL_MACHINE\System\CurrentControlSet\Control\Session Manager\Kernel`) и записав туда `0`. Поточный вызов DPC должен быть инициализирован разработчиком с помощью API-функции `KeInitializeThreadedDpc`, которая установит его внутренний тип в `ThreadedDpcObject`. Поскольку поточные вызовы могут быть отключены, разработчики драйверов, использующие их, должны писать свои алгоритмы по тем же правилам, что и для непоточных вызовов, а поэтому не могут обращаться к страничной памяти, запрашивать у диспетчера ожидание или пытаться определить, на каком IRQL они исполняются. Им также нельзя пользоваться функциями `KeAcquire/ReleaseSpinLockAtDpcLevel`, так как те предполагают, что процессор работает на уровне DPC. Вместо этого поточным DPC следует применять `KeAcquire/ReleaseSpinLockForDpc`, которые сначала проверяют текущий IRQL, а уже потом принимают решение.

Хотя поточные вызовы DPC дают разработчикам драйверов отличный способ по возможности защитить ресурсы системы, их использование опционально как для разработчика, так и для системного администратора. Из-за этого подавляющее большинство DPC все еще работают непоточно и могут вызывать заметные задержки в системе. В Windows существует огромный арсенал средств отслеживания производительности для диагностики и решения проблем, связанных с DPC. В первую очередь это, конечно же, отслеживание времени исполнения DPC (и ISR) как через счетчики производительности, так и с помощью тщательной трассировки ETW.

ЭКСПЕРИМЕНТ. Отслеживание активности DPC

Чтобы наблюдать за активностью DPC, можно воспользоваться программой Process Explorer. Для этого нужно открыть окно Системная информация (System Information) и перейти на вкладку Процессор (CPU), где показывается количество прерываний и DPC выводится каждый раз, когда там обновляется изображение (по умолчанию один раз в секунду).





Вы также можете воспользоваться отладчиком ядра, чтобы рассмотреть различные поля KPRCB, начиная с DPC, такие как `DpcRequestRate`, `DpcLastCount`, `DpcTime` и `DpcData`, включающее в себя `DpcQueueDepth` и `DpcCount` для поточных и непоточных DPC. Кроме того, в новых версиях Windows туда добавлено поле

IsrDpcStats, которое содержит указатель на структуру _ISRDPCSTATS, взятую из файлов общедоступных символов. К примеру, приведенная далее команда отобразит общее число DPC в очереди на текущий KPRCB (поточных и непоточных) рядом с количеством уже *выполненных*:

```
lkd> dx new { QueuedDpcCount = @$prcb->DpcData[0].DpcCount +
    @$prcb->DpcData[1].DpcCount,
ExecutedDpcCount = ((nt!_ISRDPCSTATS*)@$prcb->IsrDpcStats)->DpcCount },d
    QueuedDpcCount   : 3370380
    ExecutedDpcCount : 1766914 [Type: unsigned __int64]
```

Несоответствие, показанное в примере, ожидаемо. Драйверы могли поставить в очередь вызов DPC, который уже там был, а с такими ситуациями Windows безопасно справляется. Кроме того, вызов DPC, изначально поставленный в очередь к конкретному процессору (без указания, к какому именно), иногда может быть исполнен другим процессором, в частности, если драйвер пользуется функцией KeSetTargetProcessorDpc, которая позволяет ему направить DPC на конкретный процессор.

Windows не просто ожидает, что пользователи будут сами расследовать задержки, вызванные проблемами с DPC, — она содержит встроенные инструменты на случай реализации нескольких типичных сценариев, способных причинять значительные неудобства. В первую очередь это средства *DPC Watchdog* и *DPC Timeout*, которые могут быть настроены через некоторые значения в реестре, находящиеся по адресу HKEY_LOCAL_MACHINE\SYSTEM\CurrentControlSet\Control\Session Manager\Kernel, такие как DPCTimeout, DpcWatchdogPeriod и DpcWatchdogProfileOffset.

Сторожевой таймер DPC Watchdog отвечает за мониторинг исполнения всего кода на уровне DISPATCH_LEVEL и выше на случай, если долго не будет снижения IRQL. Тайм-аут DPC Timeout, в свою очередь, отслеживает время исполнения отдельных DPC. По умолчанию тайм-аут для отдельного вызова DPC составляет 20 с, а для всего уровня DISPATCH_LEVEL (и выше) — 2 мин. Оба лимита могут быть перенастроены с помощью упомянутых ранее значений в реестре (DPCTimeout контролирует лимит времени для отдельных DPC, а DpcWatchdogPeriod отслеживает исполнение всего кода с высоким IRQL). Когда эти лимиты превышены, система либо выполнит проверку критической ошибки DPC_WATCHDOG_VIOLATION (с указанием, какая из ситуаций возникла), либо, если подключен отладчик ядра, инициирует подтверждение возможности продолжения.

Разработчики драйверов, желающие внести свою лепту в то, чтобы можно было избегать подобных ситуаций, могут воспользоваться функцией KeQueryDpcWatchdogInformation, чтобы узнать, какие в данный момент настроены параметры и сколько еще осталось времени. Более того, функция KeShouldYieldProcessor принимает эти значения во внимание (как и другие сведения о состоянии системы) и возвращает драйверу подсказку либо отложить обработку DPC на потом, либо сбросить IRQL обратно на PASSIVE_LEVEL (для случаев, когда DPC не исполняется, но драйверу требуется удержать блокировку или каким-то образом синхронизироваться с DPC).

В новейших сборках Windows 10 в каждом PRCB дополнительно помещается таблица истории исполнения DPC (DPC Runtime History Table, `DpcRuntimeHistoryHashTable`), где содержится хеш-таблица блоков отслеживания особых функций отзыва DPC, которые недавно исполнялись, и количества циклов процессора, ими потраченных. При анализе аварийного дампа памяти или удаленной системы это может сильно помочь в расследовании проблем с задержками без привлечения визуальных инструментов, но, что более важно, эти данные теперь тоже используются ядром.

Когда разработчик драйвера поместит очередь DPC с помощью функции `KeInsertQueueDpc`, API пролистает таблицу процессора и проверит, не замечено ли за этим DPC исполнение, превышающее определенный лимит (по умолчанию это 100 мс, но значение можно перенастроить с помощью параметра `LongDpcRuntimeThreshold`, находящегося в реестре по адресу HKEY_LOCAL_MACHINE\SYSTEM\CurrentControlSet\Control\Session Manager\Kernel). При положительном результате в вышеупомянутой структуре `DpcData` будет заполнено поле `LongDpcPresent`.

Для каждого потока бездействия (более подробное описание планирования потоков и потока бездействия вы найдете в главе 4 тома 1) ядро теперь создает поток делегирования DPC. Эти уникальные потоки, как и потоки бездействия, принадлежат процессу бездействия системы и по умолчанию никогда не участвуют в алгоритмах выбора потоков планировщиком. Ядро как бы собирает их для своих нужд. На рис. 8.18 приведен пример системы с 16 логическими процессорами, где теперь присутствуют 16 потоков бездействия и 16 потоков делегирования DPC. Обратите внимание на то, что эти потоки имеют реальные идентификаторы (Thread ID, TID) и наблюдать их следует в столбце Процессор (CPU).

Всякий раз, когда ядро диспетчеризует DPC, оно проверяет глубину очереди на предмет превышения лимита для так называемых *длительных DPC* (по умолчанию он равняется двум, но может быть перенастроен через то же значение в реестре, о котором упоминалось ранее). Если это так, решение по устранению возникшей проблемы принимается исходя из свойств текущего исполняемого потока. Бездействует ли? В режиме ли реального времени? Говорит ли его маска соответствия о том, что обычно этот поток исполняется на другом процессоре? В зависимости от этого ядро может вместо помещения в очередь фактически перенести вызов DPC с отягощенной позиции в отдельно запланированный делегируемый поток, который получит максимально возможный приоритет (но все еще на уровне `DISPATCH_LEVEL`). Таким образом предыдущий приостановленный поток (или любой другой в очереди ожидания) получит шанс быть перенаправленным на другой процессор.

Данный механизм напоминает описанные ранее поточные DPC, с некоторыми исключениями. Делегируемый поток все еще действует на уровне `DISPATCH_LEVEL`. После создания на фазе 1 инициализации ядра (подробно — в главе 12) он повышает свой IRQL до `DISPATCH_LEVEL`, сохраняет его в свой блок данных ядра в поле `WaitIrql`, а затем добровольно обращается к планировщику, чтобы тот переключил контекст на какой-нибудь другой готовый или ожидающий поток (вызывает процедуру `KiSwapThread`). Таким образом, делегируемые DPC позволяют системе автоматически обеспечивать баланс нагрузки, освобождая разработчика драйвера от необходимости самостоятельно принимать решение о повышении.

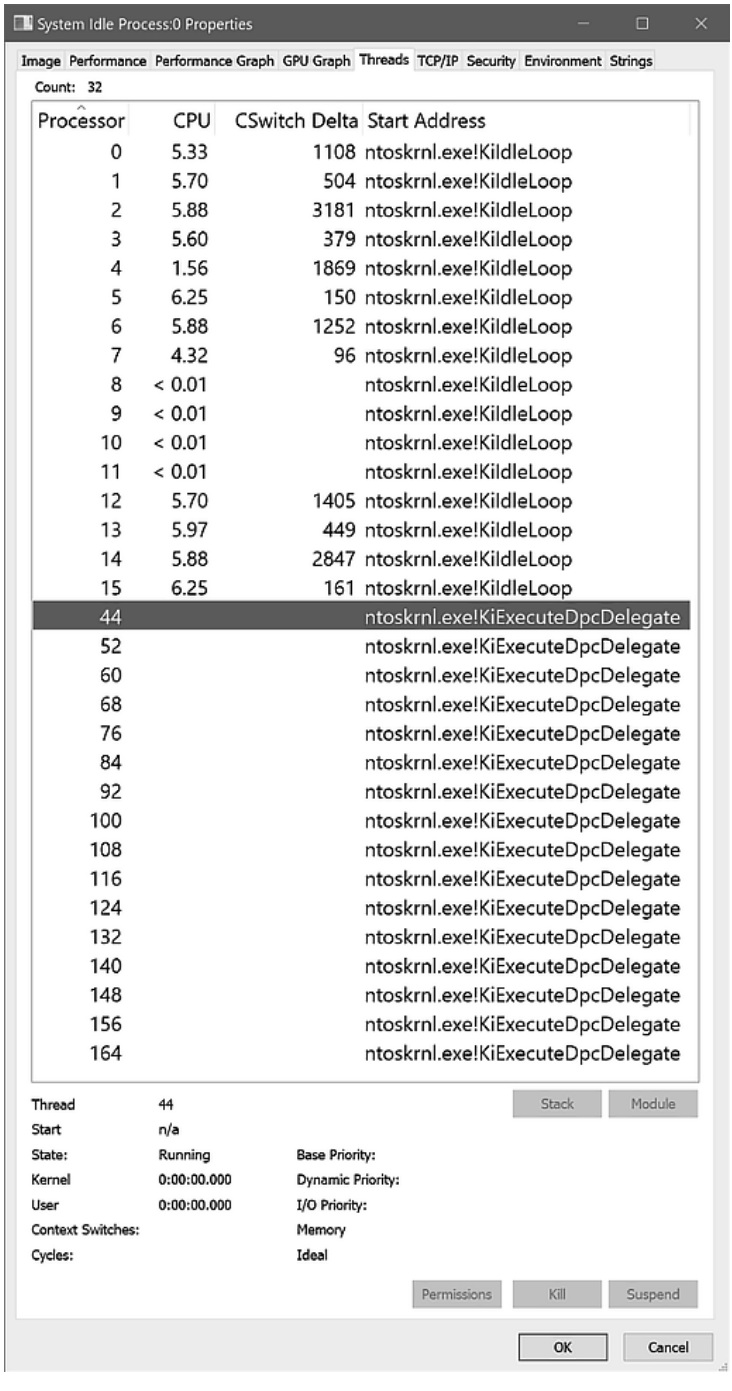


Рис. 8.18. Делегированные потоки DPC в системе с 16 процессорами

Если ваш экземпляр Windows 10 обновлен и имеет такую возможность, вы сможете самостоятельно узнать, как часто делегируемый поток привлекался к загрузке. Для этого потребуется выполнить следующую команду отладчика ядра, обращая внимание на число переключений контекста с момента запуска системы:

```
lkd> dx @$cursession.Processes[0].Threads.Where(t => t.KernelObject.ThreadName->
ToDisplayString().Contains("DPC Delegate Thread")).Select(t => t.KernelObject.Tcb.
ContextSwitches),d
    [44]                : 2138 [Type: unsigned long]
    [52]                : 4    [Type: unsigned long]
    [60]                : 11   [Type: unsigned long]
    [68]                : 6    [Type: unsigned long]
    [76]                : 13   [Type: unsigned long]
    [84]                : 3    [Type: unsigned long]
    [92]                : 16   [Type: unsigned long]
    [100]               : 19   [Type: unsigned long]
    [108]               : 2    [Type: unsigned long]
    [116]               : 1    [Type: unsigned long]
    [124]               : 2    [Type: unsigned long]
    [132]               : 2    [Type: unsigned long]
    [140]               : 3    [Type: unsigned long]
    [148]               : 2    [Type: unsigned long]
    [156]               : 1    [Type: unsigned long]
    [164]               : 1    [Type: unsigned long]
```

Прерывания асинхронных вызовов процедур

Асинхронные вызовы процедур (asynchronous procedure calls, APCs) дают пользовательским программам и системному коду возможность исполняться в контексте конкретного пользовательского потока, а значит, и в адресном пространстве конкретного процесса. Поскольку APC оказываются в очереди для подобного контекста, они подпадают под правила планирования потоков и не действуют в среде, отличной от DPC. То есть на практике они не работают на уровне `DISPATCH_LEVEL` и могут прерываться более приоритетными потоками, вставать в очередь ожидания блокировки и обращаться к страничной памяти.

Тем не менее APC все еще являются программными прерываниями, и им все еще требуется возможность как-то перехватывать контроль у основного кода потока, что отчасти достигается с помощью упомянутого ранее IRQL — `APC_LEVEL`. Это значит, что хотя они и не подвержены тем же ограничениям, что и DPC, но имеются определенные рамки, о которых разработчикам следует знать и которые мы вкратце рассмотрим далее.

APC описываются контрольными структурами ядра, которые называются *объектами APC*. В ожидании исполнения они пребывают в одной из двух управляемых ядром *очередей APC*. В отличие от очередей DPC, которые привязываются к процессорам (и делятся на поточные и непоточные), эти привязываются к потокам. Каждый поток имеет две очереди APC: одну для вызовов ядра и одну для пользовательских вызовов.

Когда ядро просят добавить APC в очередь, оно проверяет, в каком режиме *создан* вызов — пользовательском или ядра, а затем помещает его в соответствующую очередь того потока, где будет исполняться вызываемая процедура APC. Прежде

чем поговорить о том, как и когда APC сработает, рассмотрим различия, зависящие от режима его выполнения. При помещении вызова в очередь к потоку тот может пребывать в одном из следующих состояний:

- исполняется (возможно, даже является текущим на исполнении);
- ожидает;
- занят чем-то другим (готовится к исполнению, приостановлен и т. д.).

В первую очередь (см. главу 4 тома 1) можно припомнить, что всякий раз, вставая в очередь ожидания, поток попадает в оповещаемое (alertable) состояние. Для APC режима *ядра*, если поток может принимать вызовы, это состояние игнорируется — вызов всегда прерывает ожидание, последствия чего будут разъяснены в дальнейшем в этом разделе. Однако APC *пользовательского* режима прерывают поток, только либо если ожидание было оповещаемым и начато кодом пользовательского режима, либо если в очереди уже есть пользовательские APC, прерывающие ожидание (это происходит, когда множество процессоров пытаются поместить свои APC в один и тот же поток).

Пользовательские APC никогда не прерывают уже выполняющийся поток пользовательского режима — для этого он должен либо вызвать прерываемое ожидание, либо сменить привилегии или контекст, что приведет к проверке очереди APC. Вызовы же ядра вызывают прерывание на процессоре, где выполняется поток, повышая IRQL до `APC_LEVEL` и сообщая тому о необходимости пересмотреть очередь APC ядра для текущего потока. Если же поток был занят чем-либо другим, во всех случаях возникает необходимость завершить операции, переводящие поток в искомое состояние исполнения или ожидания. На практике это, в частности, приводит к тому, что остановленные потоки не будут выполнять APC из своей очереди.

До этого мы упомянули, что потоку может быть запрещено получать APC, не касаясь ранее описанных сценариев с оповещаемостью. Ядро или автор драйвера могут так сделать двумя способами. Первый — просто удерживать свой IRQL на `APC_LEVEL` или выше при выполнении какого-либо фрагмента кода. Так поток находится в состоянии исполнения, прерывание передается обычным образом, но, поскольку процессор уже на `APC_LEVEL` (или выше), оно маскируется. Таким образом, только когда IRQL будет сброшен обратно на `PASSIVE_LEVEL`, задержанное прерывание будет обработано, а APC-вызов — выполнен.

Второй способ намного более предпочтителен, поскольку позволяет не изменять состояние контроллера прерываний. Он сводится к использованию API-функций `KeEnterGuardedRegion` и `KeLeaveGuardedRegion` для включения и выключения доставки APC соответственно. Обе функции рекурсивны и могут вызываться повторно по принципу вложенности. При выполнении такого раздела кода переключение контекста на другой поток безопасно, поскольку нахождение в ней отражается в поле `SpecialApcDisable` объекта потока (KTHREAD) и не привязывается к процессору.

Подобным образом переключения контекста могут происходить на `APC_LEVEL`, хотя уже в разрезе процессоров. Диспетчер сохраняет IRQL в поле `WaitIrql` структуры KTHREAD, после чего устанавливает IRQL процессора равным `WaitIrql` нового входящего потока, который может оказаться `PASSIVE_LEVEL`. Так может получиться

занятная ситуация, когда поток PASSIVE_LEVEL технически может прервать поток APC_LEVEL. Подобные ситуации нередки и совершенно нормальны, это еще раз показывает, что в части исполнения потоков планировщик превалирует над правилами IRQL. Только начиная с уровня DISPATCH_LEVEL, где прерывание потоков не действует, IRQL главнее планировщика. Поскольку APC_LEVEL является единственным IRQL с таким поведением, его нередко называют внутрипоточным IRQL, что не совсем точно, но достаточно близко по смыслу для объяснения этого принципа.

Вне зависимости от того, как разработчик уровня ядра будет отключать APC, одно правило неизменно: коду нельзя ни вернуться в пользовательский режим через вызов APC на любом уровне выше PASSIVE_LEVEL, ни при этом поместить в поле SpecialApcDisable что-то, кроме 0. Данные действия немедленно вызывают критическую ошибку, типичным виновником которого является драйвер, не освободивший блокировку или не покинувший критический раздел APC.

В дополнение для каждого из двух режимов APC предусмотрены два типа APC — обычные и специальные, которые в зависимости от режима ведут себя по-разному. Рассмотрим каждое сочетание.

- **Специальные APC ядра.** Всегда помещается в очередь после существующих специальных APC ядра, но перед любыми обычными. Процедура ядра (kernel routine) получает указатели на параметры и на обычный вызов APC и будет исполняться на APC_LEVEL, где сможет принять решение поместить в очередь новый обычный APC.
- **Обычные APC ядра.** Всегда помещается в конец очереди, позволяя специальному APC ядра добавлять новые обычные вызовы, которые, как сказано ранее, однажды смогут выполниться. Подобные вызовы можно выключить не только упомянутыми способами, но и с помощью функции KeEnterCriticalRegion (вместе с KeLeaveCriticalRegion), которая вместо поля SpecialApcDisable в KTHREAD изменяет KernelApcDisable.

  Такие APC сначала исполняют свою процедуру ядра на APC_LEVEL, передавая туда указатели на параметры и на обычный вызов APC. Если он не был выдан, то IRQL понижается до PASSIVE_LEVEL и обычная процедура выполняется тоже, но с передачей параметров по значению. Когда та возвращает контроль, IRQL снова становится APC_LEVEL.
- **Обычные APC в пользовательском режиме.** При этой типичной комбинации APC помещается в конец очереди, а процедура ядра исполняется на APC_LEVEL, как и в предыдущих случаях. Если нормальная процедура все еще присутствует в памяти, APC готовится к передаче в пользовательском режиме (на PASSIVE_LEVEL) путем создания кадра ловушки и кадра исключения, которые в конечном счете передадут диспетчеру APC пользовательского режима в NtDLL.dll контроль над потоком после возврата в пользовательский режим, а затем сделают вызов по предоставленному пользовательскому указателю. Сразу после возврата APC в пользовательский режим диспетчер применит системный вызов NtContinue или NtContinueEx, чтобы вернуться к оригинальному кадру ловушки.

  Надо заметить, что если процедура ядра отменит обычную, то поток, если он в оповещаемом состоянии, теряет это состояние, и наоборот — если он не в оповещаемом состоянии, то переходит в него, и устанавливается флаг

«ожидается пользовательский APC», что потенциально приводит к доставке других пользовательских APC в ближайшее время. Это реализуется API-функцией `KeTestAlertThread`, в сущности обеспечивая поведение как у обычного вызова APC в пользовательском режиме, хотя процедура ядра и отменила обработку.

- **Специальные APC в пользовательском режиме.** Данная комбинация APC была добавлена в недавние сборки Windows 10, она обобщает специальную обработку, которая проводилась для APC, завершающих потоки, так, чтобы другие разработчики тоже могли ею пользоваться. Как скоро станет видно, завершение стороннего (не текущего) потока требует вызова APC, но это может произойти, только когда весь код режима ядра закончит работу. Реализация завершения потока в рамках пользовательского APC была бы хорошим решением, однако это значило бы, что разработчик в этом режиме смог бы спасти свой поток путем входа в неоповещаемое ожидание или заполнения очереди другими APC.

Во избежание такого сценария ядро долго содержало жестко заданную проверку того, является ли *процедура ядра* в пользовательском APC операцией `KiSchedulerApcTerminate`. В таком случае вызов считался специальным и попадал в начало очереди. Кроме того, статус самого потока не учитывался, ему сразу устанавливалось состояние ожидания пользовательского вызова *(user APC pending)*, после чего при следующем переключении контекста или смене уровня кольца APC выполнялся принудительно.

Однако данная функциональность, будучи предназначенной исключительно для кода завершения, приводила к тому, что те разработчики, которые похожим образом стремились гарантировать выполнение своих APC вне зависимости от состояния оповещаемости, были вынуждены пользоваться более сложными механизмами, такими как ручная смена контекста потока функцией `SetThreadContext`, в лучшем случае не дающей ошибок. В ответ на это была добавлена функция `QueueUserAPC2` с возможностью передачи флага `QUEUE_USER_APC_FLAGS_SPECIAL_USER_APC`, официально позволяющая разработчикам пользоваться аналогичной функциональностью. Такие APC всегда добавляются перед любыми другими пользовательскими (кроме вызовов на завершение работы, которые теперь исключительные) и игнорируют флаг оповещаемости, если поток в ожидании. Дополнительно этот APC будет в порядке исключения добавлен как специальный APC ядра, из-за чего процедура ядра будет исполнена почти мгновенно, а APC перерегистрирован как специальный пользовательский.

В табл. 8.9 подытоживаются порядок помещения в очередь и характер доставки каждого типа APC.

Таблица 8.9. Порядок помещения в очередь и доставка APC

Тип APC	Помещение в очередь	Порядок доставки
Специальный (ядро)	Сразу после последнего специального APC (перед всеми прочими обычными APC)	Процедура ядра доставлена на APC_LEVEL сразу после снижения IRQL и выхода потока из защищенного раздела. Она получает указатели на параметры при помещении вызова в очередь

Тип APC	Помещение в очередь	Порядок доставки
Обычный (ядро)	В конец очереди APC режима ядра	Процедура ядра стала APC_LEVEL сразу после снижения IRQL, когда поток находится вне критического (или защищенного) раздела. Она получает указатели на параметры при помещении вызова в очередь. Исполняет обычную процедуру, если есть, на PASSIVE_LEVEL после исполнения процедуры ядра. Она получает параметры, возвращаемые соответствующей процедурой ядра, которые могут оказаться как теми же, что были при добавлении вызова в очередь, так и новыми
Обычный (пользователь)	В конец очереди APC пользовательского режима	Процедура ядра стала APC_LEVEL сразу по снижении IRQL, а поток имеет установленный флаг Ожидание пользовательского APC (User APC pending) (что указывает на помещение APC в очередь, пока поток находился в оповещаемом состоянии). Она получает на вход параметры, указанные при помещении вызова в очередь. Исполняет обычную процедуру, если есть, на PASSIVE_LEVEL после исполнения процедуры ядра. Получает параметры, возвращаемые соответствующей процедурой ядра (которые могут оказаться как теми же, что были при добавлении вызова в очередь, так и новыми). Если обычная процедура была убрана процедурой ядра, та проводит проверку (test-alert) в отношении потока
Вызов завершения потока пользователя (KiSchedulerApcTerminate)	В начало очереди APC пользовательского режима	Немедленно устанавливает флаг Ожидание пользовательского APC (User APC pending) и далее действует так, как описано ранее, но по возвращении в пользовательский режим всегда переходит на уровень PASSIVE_LEVEL. На вход подаются параметры, возвращаемые специальным APC завершения потока
Специальный (пользователь)	В начало очереди APC пользовательского режима, но после того, как поток завершит APC, если таковой имелся	Аналогично описанному ранее, но параметры определяются тем, кто вызвал QueueUserAPC2 (NtQueueApcThreadEx2). Процедурой ядра является KeSpecialUserApcKernelRoutine, которая заново помещает APC в очередь, превращая первоначальный специальный вызов ядра в специальный вызов пользовательского режима

APC режима ядра задействуются для работы над задачами операционной системы, которые должны выполняться в адресном пространстве (в контексте) конкретного потока. Специальные APC ядра применяются для того, чтобы, к примеру, позволять потокам останавливать выполнение непрерываемых системных служб или сохранять результаты асинхронной операции ввода/вывода в адресное пространство потока. Подсистемы среды также задействуют их, чтобы заставить поток приостановиться/завершиться или получить/установить контекст его выполнения. Подсистема Windows для Linux (WSL) использует APC ядра для эмуляции доставки сигналов UNIX к подсистеме процессов приложений UNIX.

Другое важное применение APC режима ядра связано с приостановкой и завершением потоков. Поскольку эти действия могут быть инициированы посторонними потоками в отношении посторонних потоков, ядро использует эти вызовы как для входа в контекст потока, так и для его завершения. Драйверы устройств часто блокируют APC или находятся в защищенных или критических разделах, тем самым не позволяя применять эти действия к себе, пока они удерживают блокировку. В противном случае блокировки могут быть утрачены, а система зависнет.

Драйверы устройств также пользуются APC режима ядра. Например, если какой-то поток запрашивает операцию ввода/вывода и переходит в состояние ожидания, то можно запланировать работу другого потока в другом процессе. Когда устройство завершит передачу данных, подсистеме ввода/вывода необходимо как-то вернуться в контекст потока, их запросившего, чтобы тот мог скопировать результаты в буфер из адресного пространства своего процесса. Для этого применяется специальный APC режима ядра, кроме тех случаев, когда приложение воспользовалось функцией `SetFileIoOverlappedRange` или портами завершения. Тогда буфер либо будет расположен в памяти глобально, либо его копирование произойдет, лишь когда поток извлечет из порта результат завершения (применение APC системой ввода/вывода подробно рассматривалось в главе 6 тома 1).

Ряд функций Windows API, например `ReadFileEx`, `WriteFileEx` и `QueueUserAPC`, задействуют APC пользовательского режима. К примеру, `ReadFileEx` и `WriteFileEx` позволяют при их вызове обозначить процедуру, которая будет вызвана, когда операция ввода/вывода завершится. Завершение такой операции реализовано посредством помещения APC в очередь к потоку, ее запросившему. Однако вызов процедуры постобработчика не всегда происходит при помещении APC в очередь, так как в пользовательском режиме поток приступит к его обработке, только находясь в *оповещаемом состоянии ожидания*. Поток может войти в него и ожидать дескриптор какого-то объекта, указав наличие оповещаемости (с помощью функции `WaitForMultipleObjectsEx` в Windows) или напрямую проверив, есть ли в очереди APC (функцией `SleepEx`). В обоих случаях, если APC в очереди есть, ядро прерывает поток, передает контроль процедуре APC, а когда она отработает, возобновляет исполнение потока. В отличие от APC режима ядра, которые исполняются на `APC_LEVEL`, в пользовательском режиме вызовы работают на `PASSIVE_LEVEL`.

Доставка APC может менять порядок в очередях ожидания — списках того, какие потоки чего ожидают и в каком порядке они это делают (сериализация ожиданий описана в подразделе «Низкоуровневая IRQL-синхронизация» далее в главе). Если в момент доставки APC поток находился в состоянии ожидания того, когда его обработка закончится, ожидание возобновляется либо исполняется заново. Если коллизия

ожидания еще не разрешена, поток возвращается в состояние ожидания, но теперь в конец очереди ожидаемых объектов. Например, поскольку APC используют для приостановки выполнения потока, если тот ожидает какие-либо объекты, его очередь ожидания удаляется до возобновления исполнения, после чего он окажется в конце списка потоков, ожидающих тех же объектов. Поток, находящийся в оповещаемом ожидании в режиме ядра, также будет активирован по завершении, что позволит ему проверить, поступил сигнал из-за прекращения или по другой причине.

Обработка таймера

Интервальный таймер системных часов является, пожалуй, важнейшим устройством в компьютерах с Windows, на что указывают его высокое значение IRQL (`CLOCK_LEVEL`) и критический характер выполняемой им работы. Без прерывания от таймера Windows не сможет отслеживать время, что выразится в ошибочных результатах доступного времени и показаний часов и, что еще хуже, приведет к тому, что время таймеров больше не будет истекать, и потоки никогда больше не израсходуют свои кванты времени. Windows также перестанет быть вытесняющей операционной системой, и, пока текущий работающий поток не уступит центральный процессор, важные фоновые задачи и планировщики никогда не будут работать на этом процессоре.

Типы таймеров и интервалы

Традиционно Windows программировала системные часы так, чтобы те запускались через некоторый подходящий для компьютера интервал, впоследствии позволяя драйверам, приложениям и системным администраторам менять тактовый интервал для своих нужд. Эти часы подают сигнал со строгой периодичностью, что обеспечивается программируемым таймером прерываний (Programmable Interrupt Timer, PIT) — микросхемой, которая есть на всех компьютерах со времен PC/AT или часов реального времени (Real Time Clock, RTC). PIT работает на кристалле, настроенном на треть частоты цветовой поднесущей по стандарту NTSC (потому что изначально он применялся в ТВ-выходах первых видеоадаптеров CGA), а HAL пользуется рядом допустимых кратных величин, чтобы отмерять интервалы в миллисекундах, от 1 до 15. RTC, в свою очередь, действует на частоте 32 768 кГц, которая, будучи степенью двойки, легко настраивается на работу с различными интервалами, также являющимися степенью двойки. В системах с RTC мультипроцессорный HAL на базе APIC настраивал часы так, чтобы сигнал подавался каждые 15,6 мс, то есть примерно 64 раза в секунду.

PIT и RTC имеют массу недостатков: это медленные внешние устройства, подключаемые по устаревшим шинам, они имеют низкую детализацию, требуют от всех процессоров синхронизировать доступ к их аппаратным регистрам, очень сложны в эмуляции и все реже встречаются на встраиваемом оборудовании, таком как устройства Интернета вещей и мобильные устройства. В ответ на это производители оборудования создали новые типы таймеров, такие как таймер ACPI, иногда также называемый таймером управления питанием (Power Management, PM), и таймер APIC, который размещается непосредственно в процессоре. Таймер ACPI демонстрирует хорошую гибкость и совместимость с различными аппаратными платформами, но имеет проблемы из-за задержек и программных ошибок. Таймер

APIC, с другой стороны, очень эффективен, но часто уже занят под другие нужды платформы, в частности профилирование (хотя наиболее современные процессоры уже содержат отдельные таймеры для профилирования).

В свою очередь, Microsoft совместно с другими представителями отрасли разработали спецификацию, известную как высокопроизводительный таймер событий (High Performance Event Timer, HPET), который является значительно улучшенной версией RTC. В системах, где поддерживается HPET, он используется вместо RTC и PIC. Кроме того, в системах с архитектурой ARM64 имеется собственная архитектура таймера, известная как универсальный таймер прерываний (Generic Interrupt Timer, GIT). В конечном счете HAL задействует сложную иерархию поиска наиболее подходящего таймера для отдельно взятой системы в следующем порядке.

1. Попытаться обнаружить синтетический таймер гипервизора с целью избежать эмуляции на случай запуска в рамках виртуальной машины.
2. На физическом оборудовании попытаться обнаружить GIT. Ожидается в системах архитектуры ARM64.
3. Если возможно, найти какой-нибудь таймер на одном из процессоров, например еще не занятый локальный таймер APIC.
4. В ином случае найти HPET, начиная с MSI-совместимого HPET, устаревшего периодического HPET и заканчивая хоть каким-нибудь HPET.
5. Если не найдено никакого HPET, использовать RTC.
6. Если не найден RTC, попытаться отыскать какой-то другой таймер из таких, как PIT или SFI-таймер, в первую очередь отдавая предпочтение тем, которые поддерживают MSI-прерывания, если возможно.
7. Если ничего не нашлось, значит, в данной системе отсутствует поддерживаемый Windows таймер, чего не должно быть.

Таймеры HPET и LAPIC обладают *еще одним* преимуществом — кроме поддержки типичного *периодического режима*, описанного ранее, они также могут быть настроены на режим X. Эта способность позволяет поздним версиям Windows пользоваться *моделью динамических тактов*, о которой будет рассказано позже.

Детализация таймера

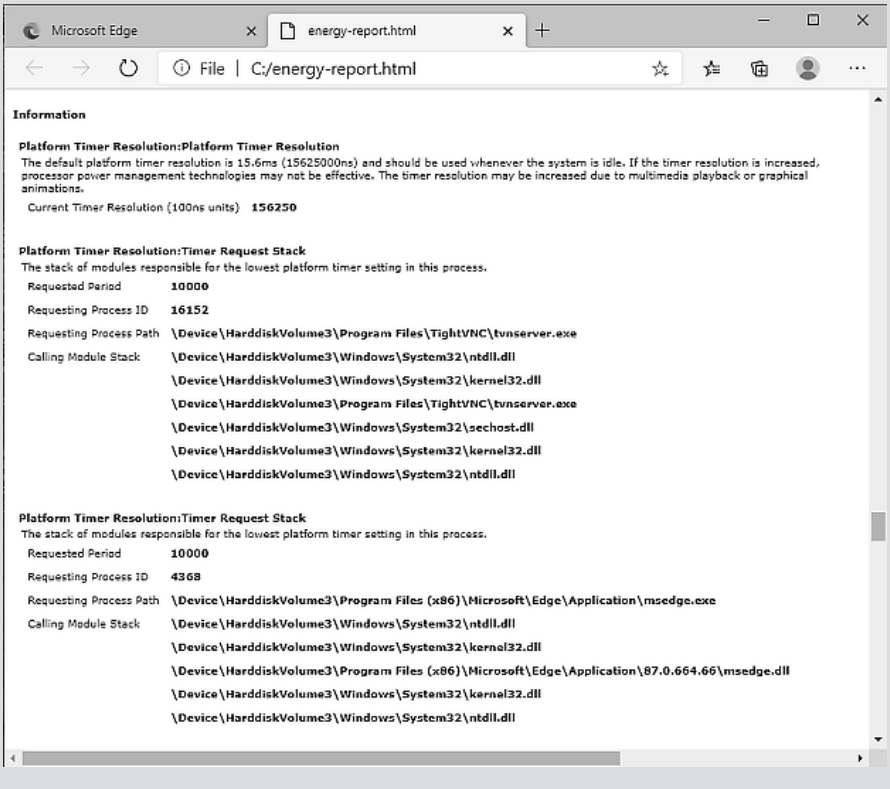
Некоторые категории приложений Windows, в частности мультимедийные приложения, могут требовать очень быстрого времени отклика. Фактически же некоторые мультимедийные задачи могут требовать частоты вплоть до 1 мс. По этой причине уже в ранних версиях Windows были реализованы механизмы и API, которые позволяли уменьшать интервал прерываний системных часов, что приводило к увеличению частоты их появления. Эта функциональность не устанавливала конкретную частоту отдельного взятого таймера (она была добавлена позднее в виде *улучшенных таймеров*, о которых более подробно будет рассказано в следующем разделе). Вместо этого в конечном счете изменялась частота всех таймеров в системе, в результате чего другие таймеры могли истекать чаще.

Тем не менее Windows делает все, что в ее силах, чтобы восстанавливать тактовую частоту часов до оригинального значения при любой возможности. Каждый раз, когда

процесс запрашивает изменение тактового интервала, Windows увеличивает внутренний механизм подсчета ссылок и ассоциирует его с этим процессом. Подобным образом драйверы (которые тоже могут изменять тактовую частоту) влияют на аналогичный глобальный подсчет. После того как все драйверы восстановили частоту, а все процессы, которые ее изменяли, либо вернули прежнее значение, либо завершились, Windows восстанавливает частоту до значения по умолчанию (в противном случае — до наибольшего значения из тех, что были запрошены процессом или драйвером).

ЭКСПЕРИМЕНТ. Отслеживание высокочастотных таймеров

Из-за проблем, которые способны вызывать высокочастотные таймеры, Windows использует службу трассировки событий (Event Tracing for Windows, ETW), чтобы отслеживать все процессы и драйверы, которые запрашивают изменение интервала системных часов, фиксируя при этом время запроса, запрошенный и текущий интервалы. Эти данные чрезвычайно полезны как разработчикам, так и системным администраторам, когда требуется определить причины быстрого расходования заряда батареи в обычно исправных системах либо снизить энергопотребление в больших системах. Чтобы получить их, достаточно запустить `powercfg/energy`. В результате откроется HTML-файл под названием `energy-report.html` наподобие приведенного далее.



Прокрутите ниже блока Разрешение таймера платформы (Platform Timer Resolution), и вы сможете увидеть все приложения, которые изменили детализацию таймера и до сих пор активны, а также стеки вызовов, приведших к этому. Разрешения таймеров отображаются в сотнях наносекунд, так что период 20 000 соответствует 2 мс. В приведенном примере два приложения, а именно Microsoft Edge и TightVNC remote desktop server, запросили повышенное разрешение.

Для получения этой информации можно воспользоваться отладчиком. Для каждого процесса в структуре EPROCESS предусмотрены следующие поля, которые позволяют определить изменение разрешения таймера:

```
+0x4a8 TimerResolutionLink : _LIST_ENTRY [ 0xfffffa80'05218fd8 - 0xfffffa80'059cd508 ]
+0x4b8 RequestedTimerResolution : 0
+0x4bc ActiveThreadsHighWatermark : 0x1d
+0x4c0 SmallestTimerResolution : 0x2710
+0x4c8 TimerResolutionStackRecord : 0xfffff8a0'0476ecd0 _PO_DIAG_STACK_RECORD
```

Обратим внимание на то, что отладчик показал нам дополнительную информацию — наименьший интервал таймера, когда-либо запрошенный данным процессом. В приведенном примере показан процесс, соответствующий PowerPoint 2010, который, как правило, запрашивает снижение интервала таймера для показа презентации, но во время редактирования слайдов так не делает. Поля структуры EPROCESS процесса PowerPoint, показанные в коде, это подтверждают, а стек можно проанализировать с помощью дампа структуры PO_DIAG_STACK_RECORD.

Наконец, поле TimerResolutionLink объединяет все процессы, которые внесли изменения в разрешение таймера, через поле ExpTimerResolutionListHead в двусвязный список. Разбор этого списка с помощью модели данных отладчика позволяет раскрыть все процессы в системе, которые вносили или внесли изменения в разрешение таймера, если команда X недоступна или требуется информация о процессах, делавших это ранее. В приведенном далее примере видно, что Edge в ряде моментов требовал установить интервал 1 мс, как и удаленный помощник Windows и Cortana. WinDbg Preview, в свою очередь, еще только запросил это и на момент написания данной команды все еще ждет ответа:

```
lkd> dx -g Debugger.Utility.Collections.FromListEntry(*(nt!_LIST_ENTRY*)
&nt!ExpTimerReso lutionListHead, "nt!_EPROCESS"", "TimerResolutionLink"").Select(p =>
new { Name = ((char*) p.ImageFileName).ToDisplayString("sb""),
Smallest = p.SmallestTimerResolution, Requested = p.RequestedTimerResolution}),d
=====================================================
=          = Name               = Smallest = Requested =
=====================================================
= [0]     - msedge.exe         - 10000    - 0         =
= [1]     - msedge.exe         - 10000    - 0         =
= [2]     - msedge.exe         - 10000    - 0         =
= [3]     - msedge.exe         - 10000    - 0         =
= [4]     - mstsc.exe          - 10000    - 0         =
= [5]     - msedge.exe         - 10000    - 0         =
```

```
= [6]     - msedge.exe        - 10000    - 0         =
= [7]     - msedge.exe        - 10000    - 0         =
= [8]     - DbgX.Shell.exe    - 10000    - 10000     =
= [9]     - msedge.exe        - 10000    - 0         =
= [10]    - msedge.exe        - 10000    - 0         =
= [11]    - msedge.exe        - 10000    - 0         =
= [12]    - msedge.exe        - 10000    - 0         =
= [13]    - msedge.exe        - 10000    - 0         =
= [14]    - msedge.exe        - 10000    - 0         =
= [15]    - msedge.exe        - 10000    - 0         =
= [16]    - msedge.exe        - 10000    - 0         =
= [17]    - msedge.exe        - 10000    - 0         =
= [18]    - msedge.exe        - 10000    - 0         =
= [19]    - SearchApp.exe     - 40000    - 0         =
=====================================================
```

Истечение времени таймеров

Как уже говорилось, одной из главных задач служебной фунции прерывания (ISR), связанных с прерыванием, генерируемым системными часами, является отсчет системного времени, в основном реализуемый процедурой `KeUpdateSystemTime`. Вторая ее задача — отслеживание логического времени исполнения, как то периоды исполнения процессов/потоков и системное время — внутреннее значение, задействуемое такими функциями, как `GetTickCount`, что используется разработчиками для подсчета времени в своих приложениях. Эта часть реализуется функцией `KeUpdateRunTime`. Однако, прежде чем выполнять всю эту работу, `KeUpdateRunTime` сначала проверяет, не истекло ли время ожидания каких-либо таймеров.

Таймеры Windows бывают *абсолютными*, то есть истекающими в фиксированный момент времени, и *относительными*, в которых в качестве момента истечения задано отрицательное значение, определяемое как смещение от текущего момента времени. Все таймеры считаются абсолютными в плане срока истечения, однако система отслеживает, какие из них являются действительно абсолютными, а какие преобразованы из относительных. Это различие бывает важно в некоторых ситуациях, например при переключении зимнего/летнего времени или даже при ручном изменении времени на часах. Абсолютной таймер в любом случае сработает в 20:00, даже если пользователь перевел время с 13:00 на 19:00, в то время как относительный таймер, скажем тот, который должен истечь через два часа, не ощутит на себе никакого эффекта перевода часов, потому что два часа на самом деле еще не прошли. Во время подобных событий изменения времени на часах ядро перепрограммирует абсолютное системное время, исходя из относительных таймеров, чтобы соответствовать новым настройкам.

Во времена, когда таймеры срабатывали только в периодическом режиме, поскольку отсчет велся в известных кратных величинах интервала, каждый кратный показатель системного времени, с которым мог быть связан таймер, назывался *исполнителем (hand)* и хранился в заголовке диспетчера объекта таймера. Windows использовала данный факт для того, чтобы организовать таймеры всех драйверов и приложений в связные списки на основе массива, где каждая запись соответствовала возможному кратному значению системного времени. Поскольку современные

версии Windows 10 уже не всегда работают в системах с периодическим тактом (ввиду функциональности динамических тактов), исполнитель был переопределен в верхние 46 бит времени, которые измерялись единицами по 100 нс. Таким образом, каждый исполнитель получает приблизительное «время» 28 мс. Кроме того, поскольку на отдельно взятом такте (особенно если срабатывание не происходит в фиксированный период времени) таймеры могли истекать по множеству исполнителей, Windows больше не могла проверять лишь текущий исполнитель. Вместо этого для отслеживания каждого исполнителя в таблице таймеров каждого процессора применяется битовая карта. Ожидающие исполнители обнаруживаются с ее помощью и проверяются на каждом прерывании от системного таймера.

Вне зависимости от используемого метода эти 256 связанных списков размещаются в структуре, называемой таблицей таймеров, которая, в свою очередь, входит в PRCB, что позволяет каждому процессору независимо обрабатывать истечение своих таймеров, не обращаясь к глобальной блокировке (рис. 8.19). Относительно новые сборки Windows 10 могут иметь до двух таблиц таймеров, то есть до 512 связанных списков.

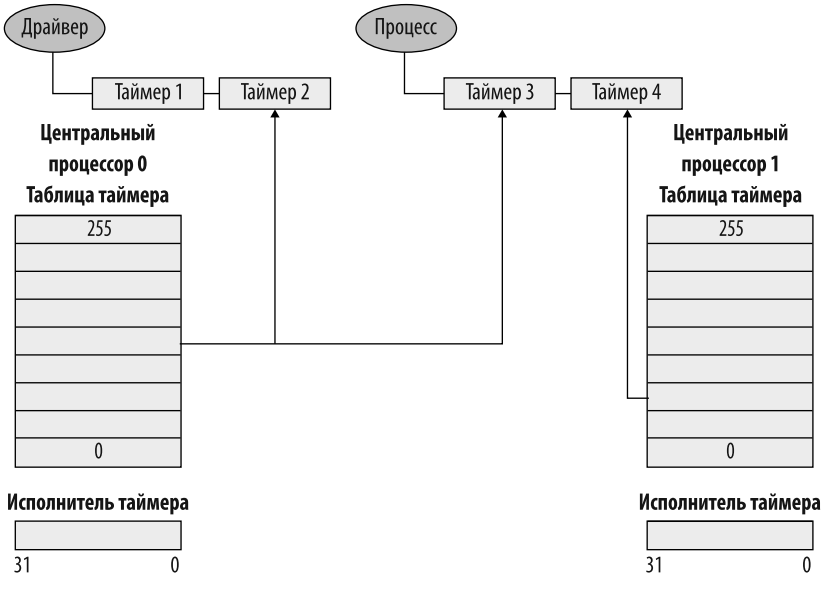


Рис. 8.19. Пример списков таймеров для каждого процессора

Далее вы поймете, как определяется, в таблицу таймеров какого логического процессора будет помещен новый таймер. Поскольку каждый процессор обладает собственной таблицей таймеров, они самостоятельно обрабатывают истечение своих таймеров. Для того чтобы избежать неоднозначного состояния, каждый процессор при инициализации заполняет свою таблицу абсолютными таймерами с бесконечным временем истечения. Таким образом, чтобы определить, прошел ли такт, достаточно проверить наличие таймеров в связанном списке, ассоциированном с текущим исполнителем.

Хотя операции обновления счетчиков и проверки связанного списка довольно быстрые, проход по каждому таймеру и изменение срока истечения его времени потенциально могут быть очень затратными — имейте в виду, что вся эта работа выполняется на CLOCK_LEVEL, исключительно высоком уровне IRQL. Подобно тому как программа обслуживания прерываний (ISR) драйвера помещает в очередь вызов DPC, чтобы отложить работу, ISR таймеров также запрашивает программное прерывание DPC, тем самым устанавливая флаг в PRCB и давая механизму опустошения очереди DPC знать, что таймеры должны истечь. Подобным образом во время обновления времени исполнения процесса или потока, если обработчик прерываний таймеров определяет, что какой-то поток исчерпал свой квант времени, он также запрашивает программное прерывание DPC и устанавливает другой флаг в PRCB. Поскольку в нормальной ситуации каждый процессор самостоятельно обновляет время исполнения и каждый процессор занят своим потоком и собственными задачами, эти флаги свои для каждой PRCB. В табл. 8.10 приведены различные поля, используемые при обработке и истечении таймеров.

Таблица 8.10. Поля KPRCB, используемые при обработке таймеров

Поле KPRCB	Тип	Описание
LastTimerHand	Индекс (до 256)	Последний исполнитель таймера, обработанный данным процессором. В недавних сборках является частью таблицы таймеров, поскольку теперь таблиц две
ClockOwner	Логический	Показывает, является ли текущий процессор владельцем этих часов
TimerTable	KTIMER_TABLE	Перечисляет заголовки списков таблицы таймеров (256 или 512 в поздних сборках)
DpcNormalTimerExpiration	Битовый	Показывает, было ли запущено прерывание DPC, чтобы запросить истечение таймера

Хотя DPC в основном предназначены для драйверов устройств, ядро тоже ими пользуется. Чаще всего оно делает это для обработки истечения квантов времени. На каждом такте системных часов срабатывает прерывание на уровне CLOCK_LEVEL. *Обработчик прерываний от системного таймера* (работающий на уровне IRQL системного таймера) обновляет системное время, а затем уменьшает значение счетчика, который отслеживает, как долго исполнялся текущий поток. Когда счетчик достигает нуля, квант текущего потока считается истекшим и ядру может потребоваться перепланировать работу процессора, что является менее приоритетной задачей, которую следует выполнять на DPC_LEVEL. Для этого обработчик помещает в очередь вызов DPC с целью запустить диспетчеризацию потоков, а затем завершает свою работу и понижает IRQL процессора. Поскольку прерывания DPC имеют более низкий приоритет, чем прерывания устройств, любое стоящее в очереди прерывание устройства, появившееся до того, как закончилась обработка прерывания от системных часов, обрабатывается до того, как происходит прерывание DPC.

Как только IRQL снизится до DISPATCH_LEVEL, в рамках обработки DPC, эти два флага будут приняты.

В главе 4 тома 1 рассматривались действия, связанные с планированием потоков и истечением квантов. Здесь же мы поговорим о действиях при истечении таймера. Поскольку таймеры связаны друг с другом исполнителем, код, реализующий истечение (исполняется в рамках DPC, связанного с PRCB в поле `TimerExpirationDpc`, — обычно `KiTimerExpirationDpc`), проверяет этот список от начала до конца. (Во время ввода первыми будут идти таймеры, находящиеся ближе всего к кратной величине тактового интервала, за ними пойдут таймеры, располагающиеся все ближе и ближе к следующему интервалу, но все еще в рамках данного исполнителя.) При истечении таймера выполняются две основные задачи.

- Таймер рассматривается как объект синхронизации диспетчера (потоки ждут таймера в рамках истечения лимита времени или непосредственно в рамках ожидания). К таймеру применяются алгоритмы проверки ожидания и удовлетворения ожидания. Более подробно эта работа описывается в разделе «Синхронизация» далее в данной главе. Именно таким образом таймер используется приложениями пользовательского режима и некоторыми драйверами.
- Таймер рассматривается как объект управления, связанный с процедурой обратного DPC-вызова, которая выполняется по истечении времени таймера. Этот метод зарезервирован только для драйверов и обеспечивает очень быструю реакцию на истечение времени таймера (метод ожидания/диспетчеризации требует задействовать всю дополнительную логику сигнализации ожидания). Кроме того, поскольку само истечение времени таймера выполняется на уровне `DISPATCH_LEVEL`, на котором также запускаются DPC-процедуры, для процедуры обратного вызова таймера складываются весьма благоприятные обстоятельства.

Поскольку каждый процессор активируется для обработки интервала системного таймера, чтобы обработать системное время и время исполнения, в итоге он обрабатывает истечение таймеров после некоторой задержки/ожидания, в ходе которого IRQL снижается с `CLOCK_LEVEL` до `DISPATCH_LEVEL`. На рис. 8.20 демонстрируется такое поведение двух процессоров: сплошные стрелки показывают срабатывания прерывания по таймеру, пунктирные — обработку истечений любого таймера, которая может происходить при наличии у процессора связанных таймеров.

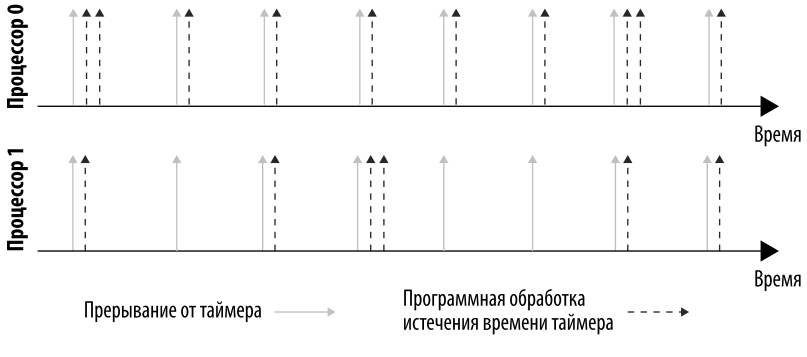


Рис. 8.20. Истечение времени таймера

Выбор процессора

Когда добавляется новый таймер, критически важным решением является подбор подходящей таблицы для использования — иными словами, нужно оптимально выбрать процессор. В первую очередь ядро проверяет, активна ли *сериализация таймеров*. Если это так, оно проверяет, имеется ли у таймера DPC, связанный с его истечением, и был ли этот DPC задан целевому процессору, в случае чего выбирается таблица этого самого процессора. Если DPC не ассоциирован с таймером или не был сопоставлен с процессором, ядро сканирует все процессоры в текущей процессорной группе, которые не были припаркованы (более подробное описание парковки ядер имеется в главе 4 тома 1). Если текущий процессор припаркован, ядро подбирает следующий ближайший по соседству непripаркованный процессор в том же режиме NUMA, в ином случае задействуется текущий процессор.

Такое поведение нацелено на то, чтобы повысить производительность и масштабируемость серверных систем, использующих технологию Hyper-V, хотя оно позволяет добиться похожего результата и на любой другой высоконагруженной системе. По мере того как системных таймеров становится все больше, потому что большинство драйверов не задают соответствие своих DPC процессорам, CPU 0 становится все более и более перегруженным исполнением кода истечения таймеров, что увеличивает задержки и даже может привести к длительным периодам ожидания, а то и к появлению необработанных вызовов DPC. Кроме того, истечения таймеров могут начать конкурировать с вызовами, обычно связанными с обработкой прерываний от драйверов, как код обработки сетевых пакетов, что может вызвать замедление системы в целом. Этот процесс еще более ощутим в ситуациях с Hyper-V, где CPU 0 должен обрабатывать таймеры и DPC, связанные с потенциально многочисленными виртуальными машинами, каждая — с собственными таймерами и устройствами.

При распределении таймеров между процессорами (рис. 8.21) нагрузка на каждый из них по обработке истечения таймеров полностью разделена между непripаркованными логическими процессорами. В 32-разрядных системах номер процессора, связанного с объектом таймера, хранится в управляющем заголовке, а в 64-разрядной системе — в самом объекте.

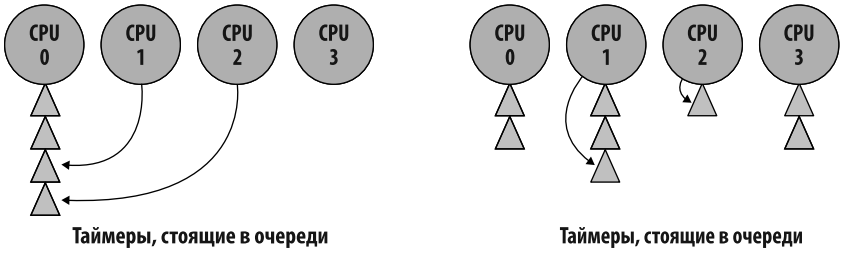


Рис. 8.21. Поведение таймеров, стоящих в очереди

Подобная стратегия, хотя и весьма выгодная для серверов, обычно не так важна для клиентских систем. Кроме того, она делает каждое событие истечения таймера (например, такт часов) более сложным, поскольку процессор мог уйти в бездействие,

все еще имея связанные с ним таймеры. Потенциально это могло бы вызвать для процессора, все еще получающего такты часов, необходимость заодно проверять чужие таблицы таймеров. Наконец, поскольку различные процессоры могут отменять и добавлять таймеры в одно и то же время, становится возможным принципиально асинхронное поведение в части истечения таймеров, что не всегда желательно. Эта сложность делает практически невозможной реализацию устойчивой фазы Modern Standby, потому что ни один конкретный процессор не может предназначаться специально для того, чтобы обрабатывать часы. Таким образом, в клиентских системах, где Modern Standby доступно, сериализация таймеров активна, что заставляет ядро в любом случае выбирать CPU 0. Это позволяет тому вести себя как владельцу *часов по умолчанию*, то есть как процессору, который всегда будет активен для того, чтобы обработать прерывание от часов (более подробно об этом — далее).

ПРИМЕЧАНИЕ Это поведение контролируется переменной ядра KiSerializeTimerExpiration, которая инициализируется на основании параметра в реестре, значение которого различно для серверной и клиентской конфигурации установки. Изменяя или пересоздавая параметр SerializeTimerExpiration в разделе HKLM\SYSTEM\CurrentControlSet\Control\Session Manager\Kernel и устанавливая для него любое значение, кроме 0 или 1, сериализацию можно заблокировать, позволяя таймерам распределяться между процессорами. Удаление данного параметра или приравнивание его к нулю позволяет ядру принимать решения, исходя из доступности технологий Modern Standby. Установка параметру значения 1 делает сериализацию всегда активной даже в системах без Modern Standby.

ЭКСПЕРИМЕНТ. Вывод списка системных таймеров

Вы можете воспользоваться отладчиком ядра, чтобы выгрузить все зарегистрированные на данный момент в системе таймеры, равно как и информацию о DPC, связанных с каждым из них (если они есть). Рассмотрим для примера следующий вывод:

```
0: kd> !timer
Dump system timers

Interrupt time: 250fdc0f 00000000 [12/21/2020 03:30:27.739]

PROCESSOR 0 (nt!_KTIMER_TABLE fffff8011bea6d80 - Type 0 - High precision)
List Timer             Interrupt Low/High Fire Time               DPC/thread

PROCESSOR 0 (nt!_KTIMER_TABLE fffff8011bea6d80 - Type 1 - Standard)
List Timer              Interrupt Low/High Fire Time DPC/thread
  1 ffffdb08d6b2f0b0   0807e1fb 80000000 [ NEVER ] thread ffffdb08d748f480
  4 ffffdb08d7837a20   6810de65 00000008 [12/21/2020 04:29:36.127]
  6 ffffdb08d2cfc6b0   4c18f0d1 00000000 [12/21/2020 03:31:33.230] netbt!TimerExpiry
                                                                 (DPC @ ffffdb08d2cfc670)
    fffff8011fd3d8a8 A fc19cdd1 00589a19 [ 1/ 1/2100 00:00:00.054]
                                  nt!ExpCenturyDpcRoutine (DPC @ fffff8011fd3d868)
  7 ffffdb08d8640440   3b22a3a3 00000000 [12/21/2020 03:31:04.772]
```

```
                                                              thread ffffdb08d85f2080
    ffffdb08d0fef300   7723f6b5 00000001 [12/21/2020 03:39:54.941]
                       FLTMGR!FltpIrpCtrlStackProfilerTimer (DPC @ ffffdb08d0fef340)
 11 fffff8011fcffe70   6c2d7643 00000000 [12/21/2020 03:32:27.052]
                                  nt!KdpTimeSlipDpcRoutine (DPC @ fffff8011fcffe30)
    ffffdb08d75f0180   c42fec8e 00000000 [12/21/2020 03:34:54.707]
                                                              thread ffffdb08d75f0080
 14 fffff80123475420   283baec0 00000000 [12/21/2020 03:30:33.060] tcpip!IppTimeout
                                                             (DPC @ fffff80123475460)
. . .
 58 ffffdb08d863e280 P 3fec06d0 00000000 [12/21/2020 03:31:12.803]
                                                              thread ffffdb08d8730080
    fffff8011fd3d948 A 90eb4dd1 00000887 [ 1/ 1/2021 00:00:00.054]
                                  nt!ExpNextYearDpcRoutine (DPC @ fffff8011fd3d908)
. . .
104 ffffdb08d27e6d78 P 25a25441 00000000 [12/21/2020 03:30:28.699]
                            tcpip!TcpPeriodicTimeoutHandler (DPC @ ffffdb08d27e6d38)
    ffffdb08d27e6f10 P 25a25441 00000000 [12/21/2020 03:30:28.699]
                            tcpip!TcpPeriodicTimeoutHandler (DPC @ ffffdb08d27e6ed0)
106 ffffdb08d29db048 P 251210d3 00000000 [12/21/2020 03:30:27.754]
                       CLASSPNP!ClasspCleanupPacketTimerDpc (DPC @ ffffdb08d29db088)
    fffff80122e9d110   258f6e00 00000000 [12/21/2020 03:30:28.575]
                               Ntfs!NtfsVolumeCheckpointDpc (DPC @ fffff80122e9d0d0)
108 fffff8011c6e6560   19b1caef 00000002 [12/21/2020 03:44:27.661]
                           tm!TmpCheckForProgressDpcRoutine (DPC @ fffff8011c6e65a0)
111 ffffdb08d27d5540 P 25920ab5 00000000 [12/21/2020 03:30:28.592]
                         storport!RaidUnitPendingDpcRoutine (DPC @ ffffdb08d27d5580)
    ffffdb08d27da540 P 25920ab5 00000000 [12/21/2020 03:30:28.592]
                         storport!RaidUnitPendingDpcRoutine (DPC @ ffffdb08d27da580)
. . .

Total Timers: 221, Maximum List: 8
Current Hand: 139
```

В этом примере, сокращенном для экономии места, есть несколько таймеров, которые скоро истекают и которые связаны с драйверами, в частности Netbt.sys и Tcpip.sys (оба связаны с сетевым взаимодействием), а также Ntfs — драйвер контроллера запоминающих устройств. Также представлены фоновые «хозяйственные» таймеры, скоро истекающие, в частности связанные с управлением питанием, трассировкой событий Windows (ETW), очисткой реестра и виртуализацией контроля учетных записей пользователей (Users Account Control, UAC). Кроме того, имеется около дюжины таймеров, с которыми не связаны никакие DPC, что с большой вероятностью перемещает вниз таймеры пользовательского режима или режима ядра, используемые для управления ожиданиями. Вы можете использовать команду `!thread` на указателях потока, чтобы это проверить.

Наконец, в системах Windows всегда присутствуют еще три интересных таймера. Это таймер, который отмечает изменения временных зон с целью перехода на летнее/зимнее время, таймер, отмечающий наступление следующего года, и таймер, который ожидает перехода в следующий век. Их довольно легко различить визуально благодаря обычно большому сроку истечения, если только вы не проводите этот эксперимент накануне одного из этих событий.

Интеллектуальное распределение тактов таймера

На рис. 8.20, где демонстрируются процессоры, обрабатывающие прерывания от часов и истекающие таймеры, видно, как процессор 1 просыпается несколько раз (сплошные стрелки), даже если с ним не связаны истекающие таймеры (пунктирные стрелки). Хотя подобное поведение необходимо, пока процессор 1 действует (чтобы обновлять время исполнения процессов/потоков и состояние планирования), что, если этот процессор бездействует и не имеет истекающих таймеров? Нужно ли ему все еще заниматься обработкой прерывания от часов? Поскольку единственная необходимая работа из описанных ранее — это в целом обновлять системное время и такты часов, было бы достаточно назначить всего лишь один процессор на роль хранителя времени (в данном случае процессор 0), а остальным позволить оставаться в состоянии сна. Когда они проснутся, все необходимые поправки времени можно выполнить, пересинхронизировавшись с процессором 0.

Фактически Windows реализует именно это (то, что называется *интеллектуальным распределением тактов таймера*), и рис. 8.22 демонстрирует состояния процессора в сценарии, когда процессор 1 спит (а не как раньше, когда мы предполагали, что он исполняет код). Как можно видеть, процессор 1 просыпается лишь пять раз для того, чтобы обработать свои истекающие таймеры, и неактивен в течение куда большего времени (период сна). Ядро использует переменную `KiPendingTimerBitmaps`, в которой содержится массив масок соответствия, где отображается, какой логический процессор должен получить интервал часов для текущего исполнителя таймера (интервал тактов часов). Это позволяет ему соответствующим образом запрограммировать контроллер прерываний, а заодно определить, каким процессорам тот будет посылать IPI для инициации обработки таймера.

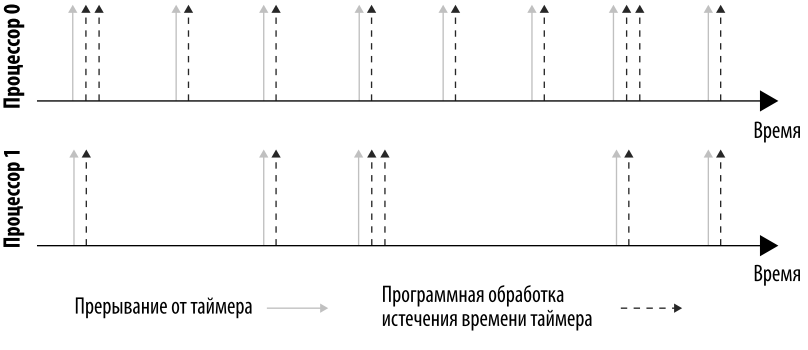


Рис. 8.22. Интеллектуальное распределение обработки таймерного такта применительно к процессору 1

Добиваться максимально возможных промежутков между активациями важно из-за специфики работы управления питанием процессора. По мере того как процессор обнаруживает, что нагрузка на него становится все меньше и меньше, он сокращает свое энергопотребление (P-состояние) до тех пор, пока не перейдет в состояние бездействия. Затем процессор может избирательно отключать собственные компоненты и входить во все более глубокое состояние бездействия/сна, в частности отключая кэши. Однако если ему снова потребуется проснуться, на его запуск уйдут энергия и время. По этой причине проектировщики процессора будут

рисковать входить в глубокое состояние бездействия/сна (C-состояние), только если время, проведенное в нем, перевешивает затраты времени и энергии на вход и выход. Очевидно, что нет никакого смысла тратить 10 мс на то, чтобы войти в состояние сна, которое продлится 1 мс. Не позволяя прерываниям от часов будить спящие процессоры до появления необходимости (таймеров), они могут входить в более глубокое C-состояние и дольше находиться в нем.

Объединение таймеров

Минимизация влияния прерываний от системного таймера на спящие процессоры в периоды, когда ни один таймер не истекает, значительно продлевает интервалы пребывания в C-состоянии при детализации таймера в 15 мс, но при этом на любом заданном исполнителе будет поставлено в очередь несколько таймеров, которые будут часто истекать, хоть и только на процессоре 0. Уменьшение нагрузки по обработке истечения программных таймеров поможет как снизить задержки (требуя меньше работы на уровне `DISPATCH_LEVEL`), так и еще дольше пребывать в состоянии сна другим процессорам. (Поскольку мы уже выяснили, что процессоры просыпаются только для обработки истечения таймеров, чем реже это происходит, тем дольше длится сон.) На практике на длительность сна влияет не столько количество истекающих таймеров (все еще дает задержку), сколько периодичность, с которой они истекают: шесть таймеров, истекающих на одном исполнителе, лучше, чем шесть таймеров, истекающих каждый на своем. Таким образом, чтобы полностью оптимизировать периоды бездействия, ядру потребуется механизм *объединения* (coalescing mechanism), который мог бы совместить разные исполнители таймера в единственный со множеством истечений.

Принцип объединения таймеров строится на предположении, что большинство драйверов и приложений пользовательского режима в целом не заботятся о том, с какой точно периодичностью срабатывают их таймеры (не считая, к примеру, мультимедийных приложений). Подобная некритическая погрешность растет вместе с длительностью периода таймера — приложению, которое просыпается каждые 30 с, скорее всего, не слишком повредит иногда делать это через 29 или 31 с, в то время как драйвер, каждую секунду что-то опрашивающий, сможет без особых проблем делать это раз в секунду ±50 мс. Важным для большинства периодических таймеров является то, чтобы периодичность их срабатывания оставалась всегда в конкретном диапазоне — к примеру, когда интервал таймера колеблется в пределах ±50 мс, он остается в нем всегда, не позволяя себе сработать через 2 с вместо 1 с, а на других фазах — через 0,5 с. Но даже при этом не все таймеры готовы к объединению в более грубые группы детализации, так что Windows включает этот механизм только для таймеров, которые сами себя пометили, как имеющие такую возможность либо через API-функцию ядра `KeSetCoalescableTimer` или ее аналог для пользовательского режима `SetWaitableTimerEx`.

С помощью этих API-функций разработчики драйверов и приложений имеют возможность предоставить ядру максимум допустимых отклонений (или приемлемых отсрочек), которые выдержит их таймер, что определяется как максимальное количество времени за пределами запрошенного периода, при котором таймер все еще будет правильно работать (в предыдущем примере односекундный таймер имел допустимое отклонение 50 мс). Рекомендуемое минимальное допустимое отклонение составляет

32 мс, что соответствует примерно двойному 15,6-миллисекундному такту часов, — любое меньшее значение не приведет к объединению, так как таймер с истечением срока невозможно даже будет переместить с одного такта часов на другой. Независимо от указанного допустимого отклонения Windows приводит таймер к одному из четырех наиболее предпочтительных интервалов объединения: 1 с, 250 мс, 100 мс или 50 мс.

Когда допустимая задержка устанавливается для периодического таймера, Windows использует процесс, называемый *смещением (shifting)*. Смещение заставляет таймер дрейфовать между периодами, пока он не будет выровнен по наиболее оптимальному кратному периоду интервала в пределах предпочтительного объединяющего интервала, связанного с заданным допустимым отклонением (который затем кодируется в заголовке диспетчера). Для абсолютных таймеров список предпочтительных объединительных интервалов сканируется и предпочтительное время истечения срока генерируется на основе ближайших допустимых объединительных интервалов к максимально допустимым отклонениям, указанным вызывающей программой. Такое поведение означает, что абсолютные таймеры всегда вытесняются как можно дальше от точки реального истечения времени, чтобы рассредоточить таймеры как можно дальше и создавать более продолжительные времена спячки процессоров.

Теперь, зная о возможности объединения таймеров, вернемся к рис. 8.20 и предположим, что для всех таймеров были указаны допустимые отклонения и тем самым указана допустимость объединения. По одному из сценариев Windows может принять решение по объединению таймеров, как показано на рис. 8.23. Обратите внимание на то, что теперь процессор 1 получает всего лишь три прерывания от часов, значительно увеличивая свои периоды сна и тем самым входя в более глубокое C-состояние. Более того, некоторым прерываниям часов на процессоре 0 тоже выпадает меньше работы. В частности, иногда задержка исчезает из-за необходимости опускаться на `DISPATCH_LEVEL` для каждого прерывания от часов.

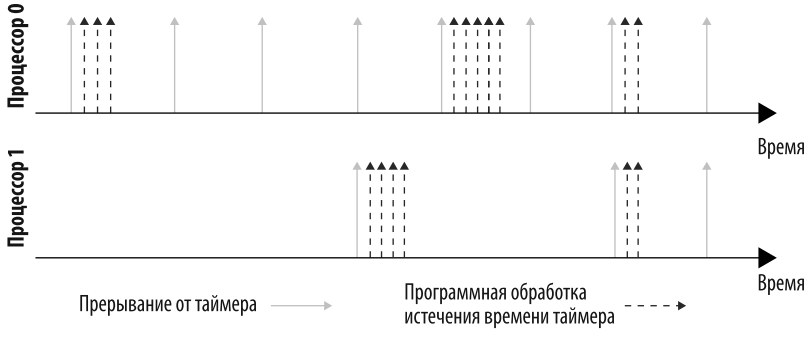


Рис. 8.23. Объединение таймеров

Улучшенные таймеры

Улучшенные таймеры были введены, чтобы удовлетворить длинный список требований, которым предыдущие улучшения системы таймеров все еще не отвечали. Во-первых, хотя объединение таймеров снижало энергопотребление, оно также делало время их истечения непостоянным, даже когда экономия не требовалась (другими словами, объединение таймеров следовало принципу «все или ничего»).

Во-вторых, единственный механизм Windows, который позволял драйверам и приложениям получить более редкий таймер, опирался на глобальное снижение частоты тактов, что, как мы уже видели, имело серьезные негативные последствия для системы в целом. Наконец, ирония ситуации заключается в том, что, хотя разрешение таймеров теперь было выше, их точность не всегда повышалась, поскольку регулярное истечение таймера могло случиться *до* такта часов вне зависимости от того, насколько детальными их делали.

Наконец, вспомним появление технологий Connected/Modern Standby, описанных в главе 6 тома 1, добавивших такие функции, как виртуализация таймеров и модератор активности Рабочего стола (Desktop Activity Moderator, DAM), которые активно задерживали истечение таймеров во время устойчивой фазы режима ожидания Modern Standby для того, чтобы стимулировать сон S3. Однако определенная активность ключевых системных таймеров все еще должна периодически иметь место даже в эту фазу.

Эти требования привели к созданию улучшенных таймеров, известных также под именем «объекты Timer2», к добавлению новых системных вызовов, таких как `NtCreateTimer2` и `NtSetTimer2`, а также функций для драйверов `ExAllocateTimer` и `ExSetTimer`. Улучшенные таймеры поддерживают четыре режима поведения, часть из которых взаимоисключающие.

- **Без пробуждения.** Этот тип таймера позволяет улучшить механизм объединения таймеров и задает допустимое ожидание, которое используется только в периоды сна.
- **Высокого разрешения.** Этот тип улучшенного таймера соответствует таймеру высокого разрешения с назначенной ему точной частотой. Часы должны будут ее поддерживать только накануне момента истечения таймера.
- **Устойчивый при бездействии.** Этот тип улучшенного таймера остается активным даже в режиме глубокого сна, в том числе в устойчивой фазе режима ожидания Modern Standby.
- **Конечный.** Этот тип таймера не имеет ни одного из описанных ранее свойств.

Таймеры высокого разрешения также могут быть устойчивы к простою. В то же время конечные таймеры не могут иметь ни одного из этих свойств. Тогда возникает вопрос: если конечные улучшенные таймеры не умеют ничего особенного, зачем их создавать? Дело в том, что, поскольку новая инфраструктура Timer2 была переработкой устаревшей логики таймеров, которая существовала с самого начала жизненного цикла ядра, она включает в себя несколько других преимуществ, не зависящих от какой-либо специальной функциональности.

- Вместо связанных списков, из которых состоят таблицы таймеров, используются самобалансирующиеся бинарные красно-черные деревья.
- Драйверам позволено разрешать или запрещать обратный вызов, не беспокоясь о необходимости вручную создавать DPC.
- Для каждой операции добавлены новые чистые записи для трассировки событий, что помогает в решении задач.
- Приняты дополнительные меры безопасности, в том числе ряд техник обфускации указателей и дополнительные проверки, усиливающие защиту от уязвимостей через повреждение данных.

Таким образом, разработчикам драйверов, которые ориентируются только на Windows 8.1 или более поздние версии, настоятельно рекомендуется пользоваться новой улучшенной инфраструктурой таймеров, даже если им не требуются ее новые возможности.

ПРИМЕЧАНИЕ Документированная функция ExAllocateTimer не позволяет драйверам создавать таймеры устойчивого бездействия. На практике попытка сделать это приведет к отказу системы. Только встроенные драйверы от Microsoft способны создавать подобные таймеры с помощью функции ExAllocateTimerInternal. Читателям не рекомендуется использовать ее, потому что в ядре содержится статический, неизменяемый список каждого допустимого вызывающего, отслеживаемого по уникальному идентификатору, который необходимо предоставить. Более того, системе известно, сколько таких таймеров каждый компонент может создать. Любые нарушения этих правил приведут к отказу системы (синему экрану смерти).

Кроме того, улучшенные таймеры подчиняются более сложному набору правил истечения времени, чем обычные, потому что у них оказывается два возможных *установленных срока*. Первый, называемый минимальным сроком, определяет самое раннее системное время, после которого таймеру разрешается сработать. Второй, максимальный срок — это самое позднее системное время, после которого таймеру нельзя срабатывать. Windows гарантирует, что таймер истечет где-то между этими двумя точками во времени, к примеру в ходе регулярного такта раз в интервал (например, 15 мс) или в результате отдельной проверки на истечение таймера (в частности, той, которую бездействующий поток выполняет при пробуждении из-за прерывания). Этот интервал вычисляют, взяв ожидаемое время истечения, предоставленное разработчиком, и адаптировав к возможному переданному «допустимому пробуждению». Если было указано, что толерантность к пробуждению неограниченна, тогда у таймера нет максимального времени истечения.

Таким образом, объект Timer2 живет в одном-двух узлах красно-черного дерева. Узел 0 предназначен для проверок минимального времени истечения, узел 2 — максимального. Таймеры без пробуждения и высокого разрешения живут в узлах 0, конечные и устойчивого бездействия — в узлах 1.

Раз мы упомянули, что некоторые из этих атрибутов могут сочетаться, как все это умещается всего в двух узлах? Очевидно, вместо одного красно-черного дерева система нуждается в нескольких. Они называются *коллекциями* (см. публичную структуру данных `KTIMER2_COLLECTION_INDEX`), по одной для каждого типа улучшенного таймера, который мы рассмотрели. Затем таймер можно внедрить в узел 0 или узел 1, либо в оба, либо ни в один из них, в зависимости от правил и комбинаций, приведенных в табл. 8.11.

Узел 1 можно считать зеркалом поведения по умолчанию устаревшего таймера — каждый такт часов проверяется, не истек ли таймер. Таким образом, таймер гарантированно истечет, пока он как минимум в узле 1, что подразумевает равенство минимального и максимального срока истечения. Если его допуск не ограничен, его не будет в узле 1, потому что технически этот таймер никогда не истечет, если процессор будет постоянно спать.

Таймеры высокого разрешения, напротив, проверяются точно тогда, когда им следует истечь, и никогда раньше, поэтому для них используется узел 0. Однако если их точное время истечения «рановато» для проверки по узлу 0, стоит проверить

и узел 1, в котором они также могут быть. В этом случае они обрабатываются как обычные (конечные) таймеры (таким образом они истекают чуть позже, чем ожидалось). Это может случиться, если вызвавшая сторона передала допуск, система бездействует и есть возможность приведения таймера.

Таблица 8.11. Типы таймеров и индексы коллекций узлов

Тип таймера	Индекс коллекции узла 1	Индекс коллекции узла 2
Без пробуждения (БП)	БП, если есть допуск	БП, если допуск ограничен или отсутствует
Конечный	Никогда не внедряется в этот узел	Конечный
Высокого разрешения (ВР)	ВР всегда	Конечный, если допуск ограничен или отсутствует
Устойчивый при бездействии	БП, если есть допуск	ВР, если допуск ограничен или отсутствует
Высокого разрешения и устойчивый при бездействии	ВР всегда	ВР, если допуск ограничен или отсутствует

Подобным образом таймер, способный работать в состоянии бездействия, если система не находится в устойчивой фазе, существует в коллекции «без пробуждения», если только не оказывается в коллекции «высокого разрешения» (состояние улучшенного таймера по умолчанию), либо живет в коллекции «высокого разрешения» в прочих случаях. Однако во время такта часов, который проверяет узел 1, он должен быть в особой коллекции «устойчивый при бездействии», чтобы обозначить необходимость его срабатывания, даже когда система в состоянии глубокого сна.

Хотя поначалу это может показаться сложным, такая комбинация состояний позволяет всем допустимым сочетаниям таймеров корректно работать при проверках как при системном такте (узел 1 — на максимальное время истечения), так и в ходе ближайшего пересчета оставшегося времени (узел 0 — на минимальное время истечения).

Каждый раз, когда таймер добавляется в соответствующую коллекцию (`KTIMER2_COLLECTION`) и связывается с узлом (узлами) красно-черного дерева, следующее время истечения в коллекции приравнивается к наиболее раннему времени истечения любого таймера в ней. В то же время глобальная переменная (`KiNextTimer2Due`) хранит в себе ближайшее время истечения любых таймеров во всех коллекциях.

ЭКСПЕРИМЕНТ. Вывод списка улучшенных системных таймеров

Вы можете использовать отладчик ядра, показанный ранее, чтобы отобразить улучшенные таймеры, которые будут видны в конце вывода:

```
KTIMER2s:
Address, Due time,                                Exp. Type   Callback, Attributes,
ffffa4840f6070b0 1825b8f1f4 [11/30/2020 20:50:16.089] (Interrupt) [None] NWF
                                              (1826ea1ef4 [11/30/2020 20:50:18.089])
ffffa483ff903e48 1825c45674 [11/30/2020 20:50:16.164] (Interrupt) [None] NW P
```

```
                                                                    (27ef6380)
ffffa483fd824960 1825dd19e8 [11/30/2020 20:50:16.326] (Interrupt) [None] NWF
                                          (1828d80a68 [11/30/2020 20:50:21.326])
ffffa48410c07eb8 1825e2d9c6 [11/30/2020 20:50:16.364] (Interrupt) [None] NW P
                                                                    (27ef6380)
ffffa483f75bde38 1825e6f8c4 [11/30/2020 20:50:16.391] (Interrupt) [None] NW P
                                                                    (27ef6380)
ffffa48407108e60 1825ec5ae8 [11/30/2020 20:50:16.426] (Interrupt) [None] NWF
                                          (1828e74b68 [11/30/2020 20:50:21.426])
ffffa483f7a194a0 1825fe1d10 [11/30/2020 20:50:16.543] (Interrupt) [None] NWF
                                          (18272f4a10 [11/30/2020 20:50:18.543])
ffffa483fd29a8f8 18261691e3 [11/30/2020 20:50:16.703] (Interrupt) [None] NW P
                                                                    (11e1a300)
ffffa483ffcc2660 18261707d3 [11/30/2020 20:50:16.706] (Interrupt) [None] NWF
                                          (18265bd903 [11/30/2020 20:50:17.157])
ffffa483f7a19e30 182619f439 [11/30/2020 20:50:16.725] (Interrupt) [None] NWF
                                          (182914e4b9 [11/30/2020 20:50:21.725])
ffffa483ff9cfe48 182745de01 [11/30/2020 20:50:18.691] (Interrupt) [None] NW P
                                                                    (11e1a300)
ffffa483f3cfe740 18276567a9 [11/30/2020 20:50:18.897] (Interrupt)
       Wdf01000!FxTimer::_FxTimerExtCallbackThunk (Context @ ffffa483f3db7360) NWF
                                (1827fdfe29 [11/30/2020 20:50:19.897]) P (02faf080)
ffffa48404c02938 18276c5890 [11/30/2020 20:50:18.943] (Interrupt) [None] NW P
                                                                    (27ef6380)
ffffa483fde8e300 1827a0f6b5 [11/30/2020 20:50:19.288] (Interrupt) [None] NWF
                                          (183091c835 [11/30/2020 20:50:34.288])
ffffa483fde88580 1827d4fcb5 [11/30/2020 20:50:19.628] (Interrupt) [None] NWF
                                          (18290629b5 [11/30/2020 20:50:21.628])
```

В приведенном примере вы можете наблюдать в основном улучшенные таймеры без пробуждения (NW) с показанным минимальным временем истечения. Некоторые из них являются периодическими и будут пересоздаваться при истечении. У некоторых также есть время максимального истечения, что означает наличие у них допустимых показателей, и вам видно самое позднее время для их истечения. Наконец, в одном из улучшенных примеров еще есть назначенный повторный вызов, принадлежащий фреймворку Windows Driver Foundation (WDF) (больше информации о драйверах WDF — в главе 6 тома 1).

Системные рабочие потоки

На фазе инициализации системы Windows создает несколько потоков в рамках системного процесса, называемых *системными рабочими потоками*, чьей единственной задачей является выполнение работы от лица других потоков. Существует множество ситуаций, когда потоки, выполняемые на уровне `DPC_LEVEL`, нуждаются в использовании функций, которые могут быть запущены только на более низком уровне. К примеру, функция DPC, запускаемая в контексте постороннего потока (потому что исполнение DPC может перехватить контроль у любого потока в системе), на уровне `DPC_LEVEL` может нуждаться в доступе к пулу подкачиваемой памяти или дожидаться объекта диспетчера, задействованного для синхронизации

выполнения с потоком из приложения. Поскольку процедуры DPC не могут снижать IRQL, они должны передать подобную работу потоку, который исполняется на уровне ниже `DPC_LEVEL`.

Некоторые драйверы устройств и управляющие компоненты создают собственные потоки, предназначенные для выполнения работы на пассивном уровне. Однако многие из них вместо этого используют системные рабочие потоки, что позволяет избежать ненужного планирования и затрат памяти, связанных с появлением в системе дополнительных потоков. Управляющий компонент запрашивает услуги системного рабочего потока, вызывая управляющие функции `ExQueueWorkItem` и `IoQueueWorkItem`. Драйверы устройств должны использовать только вторую функцию, потому что это даст возможность связать рабочую задачу с объектом устройства, позволяя лучше отслеживать и обрабатывать ситуации, в которых драйвер выгружается в момент, когда поставленная им задача еще активна. Эти функции помещают рабочую задачу в очередь объекта диспетчера, где эти потоки и ищут себе работу.

Процедуры `IoQueueWorkItemEx`, `IoSizeofWorkItem`, `IoInitializeWorkItem` и `IoUninitializeWorkItem` действуют подобным образом, но они создают ассоциацию с объектами драйвера или одним из объектов устройств.

Рабочие задачи включают в себя указатель на процедуру и параметр, который поток передаст ей при обработке. Реализует процедуру драйвер или управляющий компонент, которому требуется исполнение на пассивном уровне. К примеру, процедура DPC, которой нужно подождать управляющий объект, может инициировать задачу, где укажет на процедуру в драйвере, которая это сделает. На каком-то этапе системный рабочий поток удалит ее из очереди и выполнит процедуру драйвера. Когда она закончится, поток проверит, есть ли еще задачи в очереди. Если их больше нет, он заблокируется, пока туда не добавят задач. Процедура DPC может закончить или не закончить свою работу ко времени, когда системный рабочий поток выполняет ее задачу.

Существует много типов системных рабочих потоков.

- *Обычные рабочие потоки* исполняются с приоритетом 8, но в иных случаях ведут себя как отложенные.
- *Фоновые рабочие потоки* исполняются с приоритетом 7 и ведут себя так же, как обычные.
- *Отложенные рабочие потоки* исполняются с приоритетом 12 и занимаются некритическими по времени задачами.
- *Критические рабочие потоки* исполняются с приоритетом 13 и заняты критическими по времени задачами.
- *Суперкритические рабочие потоки* исполняются с приоритетом 14, в остальном совпадая с критическими.
- *Гиперкритические рабочие потоки* исполняются с приоритетом 15, в остальном совпадая с другими критическими.
- *Рабочие потоки реального времени* исполняются с приоритетом 18, который отмечает их особым статусом для планировщика (подробнее см. в главе 4 тома 1), после чего они не подпадают под настройку приоритета или разбивку по времени.

Поскольку названия этих рабочих очередей вводили в заблуждение, в новых версиях Windows добавили *пользовательские приоритетные рабочие потоки*, которые рекомендуются всем разработчикам драйверов и способны работать с приоритетом, нужным драйверу.

Особая функция ядра `ExpLegacyWorkerInitialization`, вызываемая в начале процесса загрузки, устанавливает исходное количество отложенных и критических рабочих потоков, которое можно конфигурировать через необязательные параметры в реестре. Возможно, вы даже могли найти сведения о ней в ранних изданиях данной книги. Однако следует иметь в виду, что эти переменные существуют лишь для совместимости с внешними инструментами и на деле никак и нигде в ядре Windows 10 не используются. Это связано с тем, что новые версии ядра задействуют новый управляющий объект ядра — *очередь приоритетов* (`KPRIQUEUE`), сочетая ее с полностью динамическим набором рабочих потоков ядра и еще больше разделив то, что когда-то было единой очередью рабочих потоков, на потоки, распределенные по узлам NUMA.

Начиная с Windows 10, ядро динамически создает, когда необходимо, новые рабочие потоки, максимальное количество которых по умолчанию ограничено 4096 (см. `ExpMaximumKernelWorkerThreads`) и которое можно перенастроить через реестр в диапазоне от 32 до 16 384. Параметр `ExpMaximumKernelWorkerThreads` можно найти по адресу `HKLM\SYSTEM\CurrentControlSet\Control\Session Manager\Executive`.

Каждый из объектов раздела, описанный в главе 5 тома 1, содержит раздел исполнения, являющийся частью объекта раздела, имеющего отношение к исполняемой программе — конкретно логикой системного рабочего потока. Он содержит структуру данных, отслеживающую диспетчер рабочей очереди для каждой части раздела, относящейся к некоему узлу NUMA. (Диспетчер очереди состоит из таймера обнаружения взаимных блокировок, сборщика элементов рабочей очереди и дескриптора потока, который и управляет очередью.) Там же находится массив указателей на каждую из восьми возможных рабочих очередей (`EX_WORK_QUEUE`). Эти очереди связаны с индивидуальным индексом и отслеживают минимальное (гарантированное) и максимальное число потоков, а также сколько задач уже было обработано.

В каждой системе по умолчанию есть две рабочие очереди: `ExPool` и `IoPool`. Первая используется драйверами и системными компонентами с помощью функции `ExQueueWorkItem`, а вторая — функции вроде `IoAllocateWorkItem`. Наконец, до шести дополнительных очередей определены для внутренних нужд системы, с которыми работают через внутреннюю (неэкспортируемую) функцию `ExQueueWorkItemToPrivatePool`, принимающую идентификатор пула от 0 до 5 (для очередей с номером от 2 до 7). На сегодня только диспетчер хранилища в составе диспетчера памяти (подробности вы найдете в главе 5 тома 1) пользуется данной функциональностью.

В рамках логики система пытается сопоставить количество критических рабочих потоков с изменяющейся по ходу работы нагрузкой. Каждый раз, когда задача для рабочего потока создается или обрабатывается, выполняется проверка на предмет потребности в новом потоке. Если он нужен, запускается событие, которое инициирует поток `ExpWorkQueueManagerThread` для связанного узла NUMA и раздела. Дополнительный рабочий поток создается в двух случаях:

- очередь обслуживает меньше потоков, чем ей по минимуму назначено;
- максимальное количество потоков еще не достигнуто, все рабочие потоки заняты, а в очереди ждут задачи либо в последний раз их не получилось туда добавить.

Дополнительно раз в секунду для каждого диспетчера рабочей очереди (то есть для каждого узла NUMA в каждом разделе) `ExpWorkQueueManagerThread` может попытаться определить, не случилась ли взаимоблокировка. Вывод об этом она делает, исходя из факта увеличения числа позиций в очереди на последнем интервале без соответствующего роста количества обрабатываемых задач. Если так произойдет, будет создан еще один рабочий поток сверх максимума в надежде разрешить образовавшийся затор. Затем эта проверка отключится до тех пор, пока в ней снова не возникнет необходимость (в частности, не будет достигнут максимум потоков). Поскольку топология процессоров может изменяться из-за *горячего добавления* новых процессоров, тот поток также в ответе за обновление соответствия процессорам и структурам данных для отслеживания новых процессоров.

Наконец, каждые два *тайм-аута рабочего потока* (по умолчанию по 10 мин, следовательно, 20 мин) этот поток решает, следует ли ему уничтожить какой-либо системный рабочий поток. По тому же адресу в реестре это значение может быть перенастроено в диапазоне от 2 до 120 мин с помощью параметра `WorkerThreadTimeoutInSeconds`. Эти действия называются *снятием (reaping)*, они призваны контролировать, чтобы число системных рабочих потоков не выходило из-под контроля. Если системный рабочий поток слишком долго ожидал (дольше тайм-аута рабочего потока) и в очереди нет позиций, ожидающих обработки (а значит, текущее количество потоков справляется с ними вовремя), его удаляют.

ЭКСПЕРИМЕНТ. Вывод списка системных рабочих потоков

К сожалению, в связи с переходом функциональности системных рабочих потоков на поразделную систему (она больше не делится по узлам NUMA и уже совсем не глобальна) команда отладчика ядра `!exqueue` больше не годится для их обзора и будет выдавать ошибку.

Поскольку структуры данных `EPARTITION`, `EX_PARTITION` и `EX_WORK_QUEUE` доступны в виде общедоступных символов, модель данных отладчика можно использовать для обзора потоков и их диспетчера. К примеру, вот как можно посмотреть диспетчер рабочих потоков на узле NUMA 0 главного (по умолчанию) системного раздела:

```
lkd> dx ((nt!_EX_PARTITION*)(*(nt!_EPARTITION**)&nt!PspSystemPartition)->
    ExPartition)->WorkQueueManagers[0]
((nt!_EX_PARTITION*)(*(nt!_EPARTITION**)&nt!PspSystemPartition)->ExPartition)->
    WorkQueueManagers[0]       : 0xffffa483edea99d0 [Type: _EX_WORK_QUEUE_MANAGER *]
     [+0x000] Partition        : 0xffffa483ede51090 [Type: _EX_PARTITION *]
     [+0x008] Node             : 0xfffff80467f24440 [Type: _ENODE *]
     [+0x010] Event            [Type: _KEVENT]
     [+0x028] DeadlockTimer    [Type: _KTIMER]
     [+0x068] ReaperEvent      [Type: _KEVENT]
     [+0x080] ReaperTimer      [Type: _KTIMER2]
     [+0x108] ThreadHandle     : 0xffffffff80000008 [Type: void *]
     [+0x110] ExitThread       : 0x0 [Type: unsigned long]
     [+0x114] ThreadSeed       : 0x1 [Type: unsigned short]
```

Как вариант, вот очередь ExPool узла NUMA 0, где сейчас действуют 15 потоков и где были обработаны уже почти 4 млн позиций:

```
lkd> dx ((nt!_EX_PARTITION*)(*(nt!_EPARTITION**)&nt!PspSystemPartition)->
    ExPartition)-> WorkQueues[0][0],d
((nt!_EX_PARTITION*)(*(nt!_EPARTITION**)&nt!PspSystemPartition)->ExPartition)->
    WorkQueues[0][0],d         : 0xffffa483ede4dc70 [Type: _EX_WORK_QUEUE *]
     [+0x000] WorkPriQueue     [Type: _KPRIQUEUE]
     [+0x2b0] Partition        : 0xffffa483ede51090 [Type: _EX_PARTITION *]
     [+0x2b8] Node             : 0xfffff80467f24440 [Type: _ENODE *]
     [+0x2c0] WorkItemsProcessed : 3942949 [Type: unsigned long]
     [+0x2c4] WorkItemsProcessedLastPass : 3931167 [Type: unsigned long]
     [+0x2c8] ThreadCount      : 15 [Type: long]
     [+0x2cc (30: 0)] MinThreads       : 0 [Type: long]
     [+0x2cc (31:31)] TryFailed        : 0 [Type: unsigned long]
     [+0x2d0] MaxThreads       : 4096 [Type: long]
     [+0x2d4] QueueIndex       : ExPoolUntrusted (0) [Type: _EXQUEUEINDEX]
     [+0x2d8] AllThreadsExitedEvent : 0x0 [Type: _KEVENT *]
```

После можно заглянуть в поле ThreadList структуры WorkPriQueue, чтобы вывести рабочие потоки, связанные с этой очередью:

```
lkd> dx -r0 @$queue = ((nt!_EX_PARTITION*)(*(nt!_EPARTITION**)&nt!PspSystemPartition)->
    ExPartition)->WorkQueues[0][0]
@$queue = ((nt!_EX_PARTITION*)(*(nt!_EPARTITION**)&nt!PspSystemPartition)->
    ExPartition)-> WorkQueues[0][0]   : 0xffffa483ede4dc70 [Type: _EX_WORK_QUEUE *]

lkd> dx Debugger.Utility.Collections.FromListEntry(@$queue->WorkPriQueue.
    ThreadListHead, "nt!_KTHREAD", "QueueListEntry")
Debugger.Utility.Collections.FromListEntry(@$queue->WorkPriQueue.ThreadListHead,
    "nt!_KTHREAD", "QueueListEntry")
     [0x0]           [Type: _KTHREAD]
     [0x1]           [Type: _KTHREAD]
     [0x2]           [Type: _KTHREAD]
     [0x3]           [Type: _KTHREAD]
     [0x4]           [Type: _KTHREAD]
     [0x5]           [Type: _KTHREAD]
     [0x6]           [Type: _KTHREAD]
     [0x7]           [Type: _KTHREAD]
     [0x8]           [Type: _KTHREAD]
     [0x9]           [Type: _KTHREAD]
     [0xa]           [Type: _KTHREAD]
     [0xb]           [Type: _KTHREAD]
     [0xc]           [Type: _KTHREAD]
     [0xd]           [Type: _KTHREAD]
     [0xe]           [Type: _KTHREAD]
     [0xf]           [Type: _KTHREAD]
```

И это была только ExPool. Как вы помните, есть еще очередь IoPool, которая идет следующим номером (1) на этом узле NUMA (0). Вы можете продолжить эксперимент, заглянув в частные пулы, например диспетчер хранилища:

```
lkd> dx ((nt!_EX_PARTITION*)(*(nt!_EPARTITION**)&nt!PspSystemPartition)->
    ExPartition)->WorkQueues[0][1],d
((nt!_EX_PARTITION*)(*(nt!_EPARTITION**)&nt!PspSystemPartition)->ExPartition)->
    WorkQueues[0][1],d         : 0xffffa483ede77c50 [Type: _EX_WORK_QUEUE *]
```

```
[+0x000] WorkPriQueue     [Type: _KPRIQUEUE]
[+0x2b0] Partition        : 0xffffa483ede51090 [Type: _EX_PARTITION *]
[+0x2b8] Node             : 0xfffff80467f24440 [Type: _ENODE *]
[+0x2c0] WorkItemsProcessed : 1844267 [Type: unsigned long]
[+0x2c4] WorkItemsProcessedLastPass : 1843485 [Type: unsigned long]
[+0x2c8] ThreadCount      : 5 [Type: long]
[+0x2cc (30: 0)] MinThreads       : 0 [Type: long]
[+0x2cc (31:31)] TryFailed        : 0 [Type: unsigned long]
[+0x2d0] MaxThreads       : 4096 [Type: long]
[+0x2d4] QueueIndex       : IoPoolUntrusted (1) [Type: _EXQUEUEINDEX]
[+0x2d8] AllThreadsExitedEvent : 0x0 [Type: _KEVENT *]
```

Управление исключениями

В отличие от прерываний, которые могут возникать в любой момент, исключения являются условиями, напрямую связанными с исполнением запущенной программы. В Windows используется средство, известное как *структурированная обработка исключений*, которое позволяет приложениям возвращать контроль при возникновении исключений. В таких случаях приложение может зафиксировать состояние и вернуться к тому месту, где возникло исключение, восстановить стек, тем самым прекратив исполнение процедуры, вызвавшей исключение, или вновь обратиться к системе, сообщив ей, что исключение не распознано и та должна продолжить поиск обработчика исключения, который сможет принять меры. Здесь мы предполагаем, что вы уже знакомы с базовыми принципами структурированной обработки исключений Windows. Если же нет, вам следует ознакомиться с документацией по Windows API либо прочесть главы 23–25 книги Джеффри Рихтера (Jeffrey Richter) и Кристофа Насарре (Christophe Nasarre) *Windows via C/C++* (Microsoft Press, 2007), прежде чем продолжить. Следует иметь в виду, что, хотя обработка исключений доступна через расширения языка программирования (к примеру, через конструкцию `__try` в Microsoft Visual C++), этот механизм является системным и поэтому от конкретного языка не зависит.

Для процессоров x86 и x64 все исключения имеют предопределенные номера прерываний, которые напрямую соответствуют записям в IDT, где хранятся указатели на обработчик ловушки для конкретного исключения. В табл. 8.12 показаны исключения, определенные для x86, и назначенные им номера прерываний. Поскольку первые записи в таблице IDT заняты для исключений, аппаратные прерывания описываются далее по списку.

Все исключения, кроме тех, что довольно просты для разрешения обработчиком ловушки, обслуживаются модулем ядра, называемым *диспетчером исключений*. Его задачей является поиск обработчика исключений, подходящего для того, чтобы от возникшего исключения избавиться. Примерами известных ядру архитектурно независимых исключений, кроме прочего, являются ошибки защиты памяти, целочисленное деление на ноль, целочисленное переполнение, ошибки операций с плавающей точкой и точки останова отладчика. Полный список архитектурно независимых исключений можно найти в документации по Windows SDK.

Таблица 8.12. Исключения и номера их прерываний

Номер прерывания	Исключение	Мнемоника
0	Divide Error (Ошибка деления)	#DE
1	Debug (Single Step) (Пошаговая отладка)	#DB
2	Non-Maskable Interrupt (NMI) (Немаскируемое прерывание)	—
3	Breakpoint (Контрольная точка)	#BP
4	Overflow (Переполнение)	#OF
5	Bounds Check (Проверка выхода за границы)	#BR
6	Invalid Opcode (Недопустимый код операции)	#UD
7	NPX Not Available (NPX недоступен)	#NM
8	Двойная ошибка (Double Fault)	#DF
9	NPX Segment Overrun (Выход за пределы сегмента NPX)	—
10	Invalid Task State Segment (TSS) (Неверный сегмент состояния задачи)	#TS
11	Segment Not Present (Сегмент отсутствует)	#NP
12	Stack Fault (Ошибка стека)	#SS
13	General Protection (Общее нарушение защиты)	#GP
14	Page Fault (Ошибка обращения к странице)	#PF
15	Intel Reserved (Зарезервировано компанией Intel)	-
16	Floating Point (Ошибка операции с плавающей точкой)	#MF
17	Alignment Check (Ошибка проверки выравнивания)	#AC
18	Machine Check (Ошибка проверки машины)	#MC
19	SIMD Floating Point (Ошибка операции с плавающей точкой в SIMD-архитектуре)	#XM или #XF
20	Virtualization Exception (Исключение виртуализации)	#VE
21	Control Protection (CET) (Защита управления)	#CP

Обработка ядром некоторых из этих исключений проходит прозрачно для пользовательской программы. К примеру, при встрече точки останова во время выполнения программы, находящейся в состоянии отладки, возникает исключение, которое ядро обрабатывает, обращаясь к отладчику. Некоторые другие исключения ядро обрабатывает, лишь возвращая вызвавшему компоненту код неуспешного результата.

Некоторым исключениям после проверки позволяется возвращаться в пользовательский режим без обработки. Например, некоторые случаи ошибок доступа к памяти или переполнение буфера вызывают возникновение исключения, которое операционная система не обрабатывает. Чтобы решать эту проблему самостоятельно, 32-разрядные приложения могут установить *кадрозависимый обработчик исключений*. Так называется обработчик, ассоциированный с вызовом конкретной процедуры. Когда процедура вызывается, в стек помещается соответствующий ей кадр стека (stack frame). У этого кадра может иметься один или несколько обработчиков

исключений, каждый из которых защищает определенный блок исходного кода программы. Когда возникает исключение, ядро разыскивает обработчик, связанный с текущим кадром стека. Если такового найти не удается, ядро пробует найти обработчик исключений, связанный с предыдущим кадром стека, и продолжает так до тех пор, пока не найдет кадрозависимый обработчик. Если и его не удалось найти, ядро вызывает собственный обработчик по умолчанию.

В 64-разрядных приложениях структурированная обработка исключений не задействует кадрозависимые обработчики (технология кадрозависимой обработки была признана уязвимой перед пользователями с недобрыми намерениями). Вместо этого для каждой функции таблица обработчиков строится и записывается в исполняемый образ на этапе компиляции. Ядро ищет обработчики, ассоциированные с каждой из функций, после чего обычно следует тому же алгоритму, что мы описали для 32-разрядного кода.

Структурированная обработка исключений часто применяется и внутри самого ядра, что позволяет ему безопасно проверять, можно ли, ничем не рискуя, обращаться по указателям из пользовательского режима с целью дать доступ для записи и чтения. Это же техникой могут пользоваться и драйверы, когда работают с указателями, пересылаемыми управляющими кодами ввода-вывода (IOCTL).

Еще один механизм обработки исключений называется *векторной обработкой исключений*. Его могут применять только приложения пользовательского режима. Больше информации о нем можно найти в документации к Windows SDK или на ресурсе Microsoft Docs по адресу https://docs.microsoft.com/en-us/windows/win32/debug/vectored-exception-handling.

Когда возникает исключение, то независимо от того, вызвано оно поведением программы или аппаратурой, в ядре начинается следующая цепочка событий. Аппаратные средства процессора передают контроль обработчику исключений ядра, который создает кадр ядра (подобное происходит и при срабатывании прерывания). Кадр ловушки позволяет системе продолжить работу с того места, где ее прервали, после того как исключение будет обработано. Обработчик ловушки также делает запись об исключении, которая содержит в себе причину исключения и прочую важную информацию.

Если исключение возникает в режиме ядра, диспетчер исключений попросту вызывает процедуру, которая будет искать кадрозависимый обработчик, чтобы тот им и занялся. Поскольку необработанные исключения уровня ядра считаются неустранимыми ошибками в работе системы, можно предполагать, что диспетчер всегда находит обработчик. Некоторые ловушки, однако, не ведут к обработчику исключений, потому что соответствующие ошибки ядро всегда считает неустранимыми. В их число входят те, которые могли бы быть вызваны исключительно серьезными багами во внутреннем коде ядра или значительными повреждениями в коде драйвера (они могут появиться только при сознательных низкоуровневых модификациях системы, за которые драйверы отвечать не должны). Такие критические ошибки будут приводить к сбою проверки кодом `UNEXPECTED_KERNEL_MODE_TRAP`.

Когда же исключение возникает в пользовательском режиме, диспетчер исключений идет более сложным путем. В подсистеме Windows имеются порт отладчика (который на самом деле объект отладчика, о чем будет рассказано позже) и порт исключений, позволяющий получать сообщения об исключениях пользовательского

режима в процессах Windows. (В данном случае под портом подразумевается объект порта ALPC, о котором пойдет разговор далее в данной главе.) Ядро задействует эти порты для обработки исключений по умолчанию, как показано на рис. 8.24.

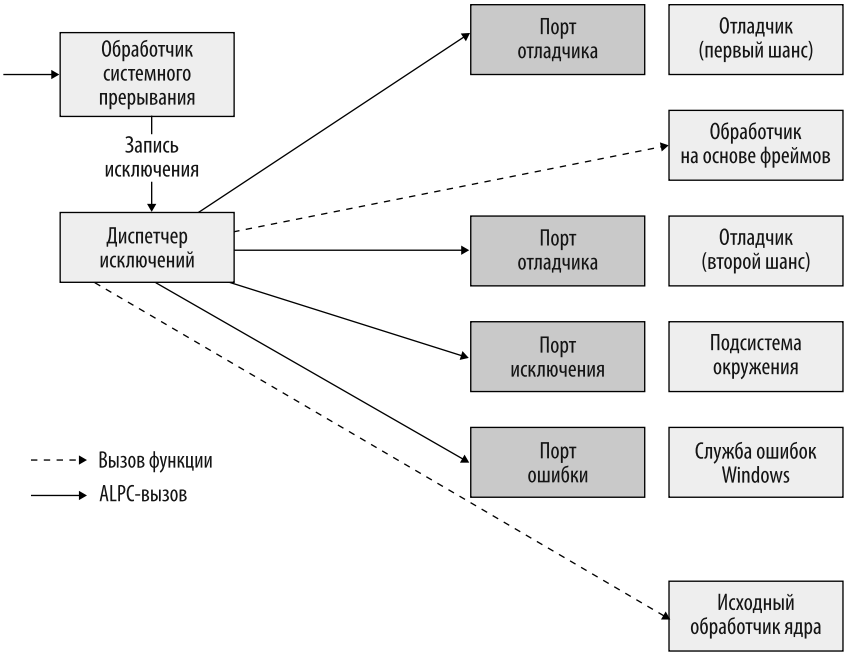


Рис. 8.24. Диспетчеризация исключения

Типичным источником исключений являются точки останова отладчика. Поэтому первым действием диспетчера исключений будет проверка того, имеет ли вызвавший исключение процесс ассоциированный с ним процесс отладчика. Если это так, диспетчер исключений отправляет сообщение через *ассоциированный с процессом объект отладчика* (система считает его портом для совместимости с программами, которые могут опираться на поведение Windows 2000, где вместо объектов отладки задействовались порты LPC). Если у процесса нет связанного с ним отладчика или отладчик не справляется с исключением, диспетчер исключений переходит в пользовательский режим, копирует кадр ловушки в стек пользователя, отформатировав его как структуру данных `CONTEXT` (документировано для Windows SDK), и вызывает процедуру поиска структурного или векторного обработчика исключений. Если не найдено ни того ни другого, а исключения обработать некому, диспетчер переключается обратно в режим ядра и вновь вызывает отладчик, чтобы позволить пользователю применить его (это называется *уведомлением второго шанса*).

Если отладчик не запущен и не найдено никаких обработчиков пользовательского режима, ядро отправляет сообщение через порт исключений, связанный с процессом исполняемого потока. Этот порт исключений, если он существует,

был зарегистрирован подсистемой окружения, контролирующей данный поток. Он дает подсистеме окружения, которое, предположительно, прослушивает этот порт, возможность перевести исключение в специфичные для данного окружения форму или сигнал. Однако, если ядро заходит в обработке исключения настолько далеко, а подсистема так и не справляется с этой задачей, ядро посылает сообщение через общесистемный *порт ошибок*, который CRSS (клиент-серверная подсистема исполнения) использует для службы ошибок Windows (WER) (см. главу 10) и исполняет обработчик исключений по умолчанию, попросту завершающий работу процесса, чей поток вызвал исключение.

Необработанные исключения

В Windows все потоки имеют обработчик, предназначенный для необработанных исключений. Он реализован в рамках внутренней функции *запуска потока* Windows. Она вызывается всякий раз, когда пользователь создает процесс либо любые дополнительные потоки. Эта функция вызывает процедуру запуска потока, предоставленную окружением и указанную в изначальной структуре контекста потока, которая, в свою очередь, вызывает пользовательскую процедуру запуска потока, указанную при вызове `CreateThread`.

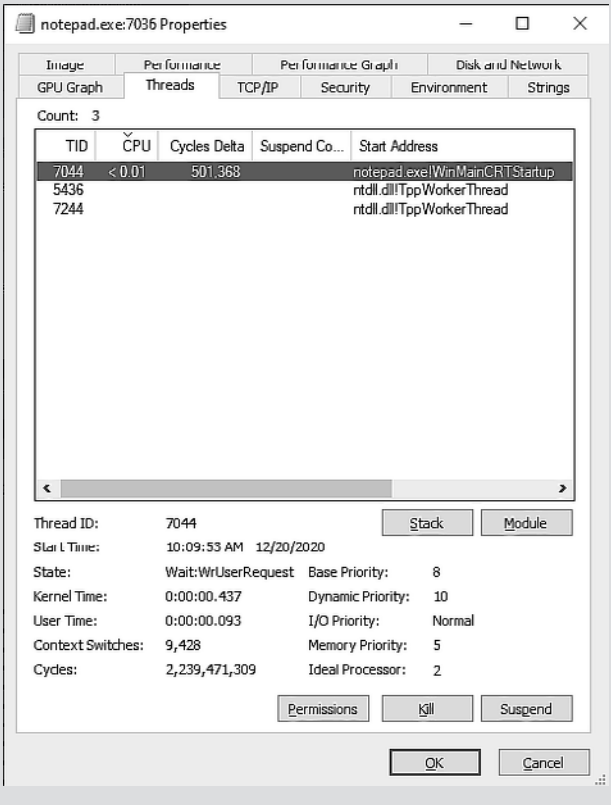
Далее представлен типичный код внутренней функции запуска потока:

```
VOID RtlUserThreadStart(VOID)
{
    LPVOID StartAddress = RCX; // Located in the initial thread context structure
    LPVOID Argument = RDX; // Located in the initial thread context structure
    LPVOID Win32StartAddr;
    if (Kernel32ThreadInitThunkFunction != NULL) {
        Win32StartAddr = Kernel32ThreadInitThunkFunction;
    } else {
        Win32StartAddr = StartAddress;
    }
    __try
    {
        DWORD ThreadExitCode = Win32StartAddr(Argument);
        RtlExitUserThread(ThreadExitCode);
    }
    __except(RtlpGetExceptionFilter(GetExceptionInformation()))
    {
        NtTerminateProcess(NtCurrentProcess(), GetExceptionCode());
    }
}
```

Обратите внимание на то, что в Windows фильтр необработанных исключений применяется, если поток вызвал исключение, которое не может обработать. Задача этой функции — реализовать определенное поведение для системы на случай, если исключение не обработано, а именно запустить процесс Werfault.exe. Однако в конфигурации по умолчанию служба сообщения об ошибках Windows, описанная в главе 10, обработает исключение, а значит, фильтр никогда не будет применен.

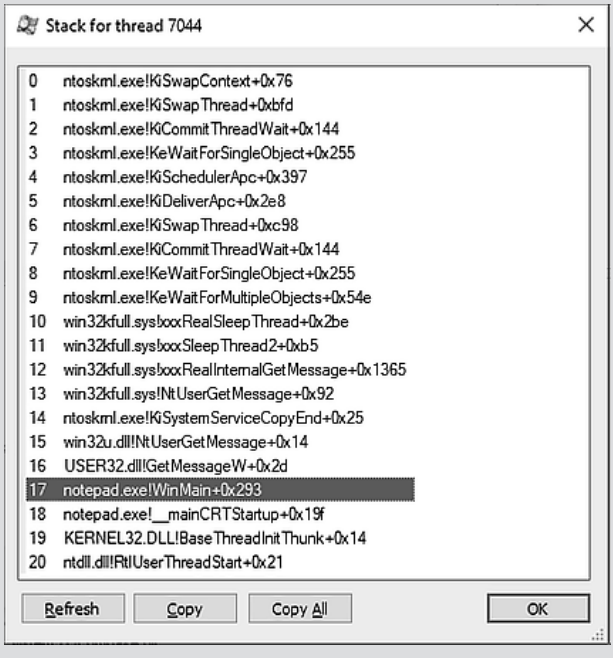
ЭКСПЕРИМЕНТ. Просмотр стартового адреса реального пользователя для потока Windows

Тот факт, что каждый поток Windows начинает свое исполнение с предоставленной системой (а не пользователем) функции, объясняет, почему стартовый адрес для потока 0 всегда одинаковый для всех процессов в системе и почему стартовые адреса для вторичных потоков точно такие же. Чтобы увидеть адрес функции, переданной пользователем, задействуйте Process Explorer или отладчик ядра.

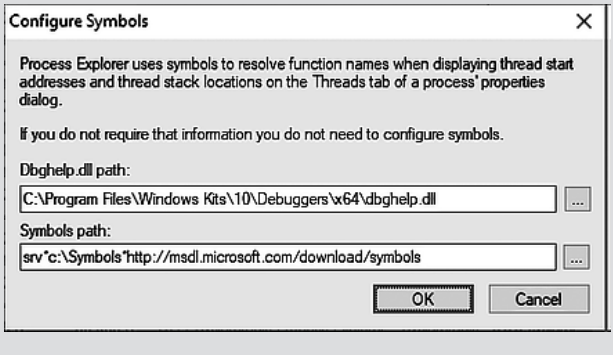


Когда Process Explorer отображает стек вызовов, он не выводит полную иерархию вызовов. Обратите внимание на следующий результат, когда нажмете кнопку Стек (Stack).

На снимке экрана далее строка 20 соответствует первому кадру стека — запуску внутренней оболочки потока. Второй кадр (строка 19) — это оболочка потока, предоставленная подсистемой окружения, в данном случае `kernel32`, поскольку мы имеем дело с приложением подсистемы Windows. Третий кадр (строка 18) и есть главная точка входа в процесс Notepad.exe.



Чтобы увидеть корректные имена функций, вам понадобится настроить Process Explorer с помощью подходящих отладочных символов. В первую очередь потребуется установить отладочные инструменты (Debugging Tools) в составе Windows SDK или WDK. Затем в меню Параметры (Options) выбрать пункт Настройка символов (Configure symbols). Путь к библиотеке `dbghelp.dll` должен указывать на файл, находящийся в папке с инструментами отладки, — обычно это C:\Program Files\Windows Kits\10\Debuggers. Обратите внимание на то, что `dbghelp.dll`, расположенная по адресу C:\Windows\System32, для этого не подойдет. Наконец, нужно правильно настроить путь к символам, чтобы скачать их из Microsoft Symbol Store в локальную папку, как показано далее.



Обработка действий системных служб

Как показано на рис. 8.24, обработчики ловушек ядра управляют прерываниями, исключениями и вызовами системных служб. В предыдущих разделах вы увидели, как работает обработка прерываний и исключений. В этом разделе поговорим о системных службах. Обработка системной службы (рис. 8.25) происходит в результате исполнения инструкции, связанной с диспетчеризацией системных служб. Инструкция, которую Windows использует для этого, зависит от модели действующего в системе центрального процессора, а также от того, активирована ли функция Hypervisor Code Integrity (HVCI), о которой мы скоро расскажем.

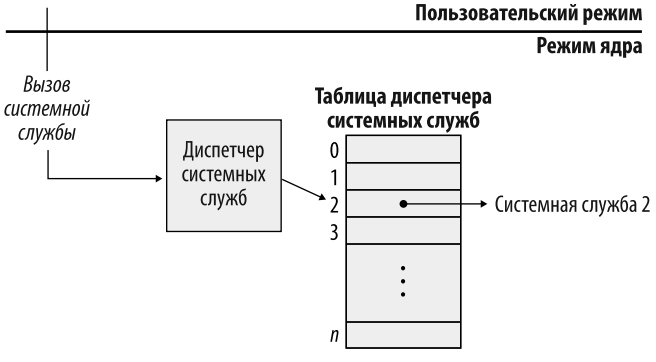


Рис. 8.25. Диспетчеризация системных служб

Зависимое от архитектуры управление системными службами

В большинстве 64-разрядных систем Windows использует инструкцию `syscall`, что приводит к изменению некоторых ключевых состояний процессора, о которых мы уже говорили в данной главе, в зависимости от определенных, заранее запрограммированных *моделезависимых регистров* (MSR):

- 0xC0000081, известного как STAR (регистр целевого адреса SYSCALL);
- 0xC0000082, известного как LSTAR (STAR длинного режима);
- 0xC0000084, известного как SFMASK (маска флагов SYSCALL).

При встрече с инструкцией `syscall` процессор ведет себя следующим образом.

- Сегмент кода (регистр CS) загружается из битов STAR с 32-го по 47-й, которые Windows приравнивает к 0x0010 (KGDT64_R0_CODE).
- Сегмент стека (регистр SS) загружается из битов STAR с 32-го по 47-й плюс 8, что в сумме дает 0x0018 (KGDT_R0_DATA).
- Указатель инструкций (RIP) сохраняется в RCX, а сам получает значение из LSTAR, которое Windows приравнивает к адресу `KiSystemCall64`, если не требуется защиты от уязвимости Meltdown (KVA Shadowing), либо к адресу `KiSystemCall64Shadow` (более подробно об уязвимости Meltdown рассказывалось ранее, в разделе «Атаки по сторонним каналам»).

- Текущее значение регистра флагов процессора RFLAGS сохраняется в R11, а затем меняется маской с SFMASK, который Windows приравнивает к 0x4700 (флаг ловушки, флаг направления, флаг прерывания и флаг вложенных задач).
- Указатель стека (RSP) и все другие сегментные регистры (DS, ES, FS, GS) сохраняют текущие значения в рамках пространства пользователя.

Таким образом, хотя эта инструкция исполняется каждые несколько циклов процессора, она не оставляет его в каком-либо нестабильном или небезопасном состоянии: указатель на стек пользовательского режима все еще загружен, GS все еще указывает на TEB, но уровень кольца, или CPL, теперь равен 0, что дает привилегии уровня ядра. Windows быстро возвращает процессор в стабильное рабочее окружение. Не считая операций, характерных для теневого копирования KVA, которое может происходить на старых процессорах, функция `KiSystemCall64` должна выполнять следующие точные операции.

С помощью инструкции `swapgs` GS переводится на адрес PCR, как сказано ранее.

Текущий указатель стека (RSP) сохраняется в поле `UserRsp` структуры PCR. Поскольку GS теперь корректно загружен, это можно сделать без применения какого-либо стека или регистра.

Новый указатель стека загружается из поля `RspBase` структуры PRCB (припомним, что эта структура хранится как часть структуры PCR).

Теперь, когда стек ядра загружен, функция создает кадр ловушки, используя формат, описанный ранее в данной главе. Сюда входят включение в кадр поля `SegSS` со значением `KGDT_R3_DATA` (0x2B), `Rsp` из поля `UserRsp` структуры PCR, `EFlags` из R11, поля `SegCs` со значением `KGDT_R3_CODE` (0x33) и сохранение там `Rip` из RCX. В обычной ситуации эти поля установила бы ловушка процессора, но Windows должна симулировать поведение на основе того, как работает функция `syscall`.

Что касается загрузки RCX из R10: в штатной ситуации ABI архитектуры x64 требует, чтобы первый аргумент любой функции, включая системный вызов, помещался в регистр RCX. Однако инструкция `syscall` переписывает этот регистр указателем инструкций вызывающего кода, как показано ранее. Windows об этом поведении известно, поэтому она копирует RCX в R10 и лишь затем вызывает `syscall`, что, как вы скоро увидите, восстанавливает данное значение.

Следующие шаги связаны с защитой процессора от атак, таких как запрещение доступа к режиму супервайзера (Supervisor Mode Access Prevention, SMAP), в частности вызов инструкции `stac`, и множеством других мер со стороны процессора против атак по сторонним каналам. В числе последних — очистка буферов трассировки переходов (BTB) и буфера хранения результатов (RCB). Кроме того, на процессорах с функцией технологии Control-flow Enforcement Technology (CET) теневой стек для потока тоже должен быть правильно синхронизирован. В дальнейшем сохраняются дополнительные элементы кадра ловушки, такие как различные неизменяемые регистры и регистры отладки, после чего начинается обработка неархитектурных регистров системного вызова, о чем вскоре мы поговорим более детально.

Однако не все процессоры имеют архитектуру x64, и будет нелишним заметить, что на процессорах x86, к примеру, используется другая инструкция, которая называется `sysenter`. Поскольку 32-разрядные процессоры встречаются все реже, мы

не станем тратить слишком много времени, разбираясь в ее работе, лишь заметим, что ее поведение довольно похоже. Какая-то доля состояний процессора требует загрузки из различных MSR, а ядро делает дополнительную работу, в частности подготавливает кадр ловушки. Больше деталей можно найти в соответствующих руководствах к процессорам семейства Intel. Подобным образом процессоры на основе ARM используют инструкцию `svc`, у которой есть собственные поведение и алгоритм управления на уровне ОС. Однако системы подобного рода составляют довольно малую часть от общего числа компьютеров с Windows.

Существует еще один неудобный момент, с которым Windows приходится считаться: процессоры без функции Mode Base Execution Controls (MBEC), но работающие в условиях активности механизма Hypervisor Code Integrity (HVCI), страдают от конструктивного недостатка, не позволяющего полностью обеспечивать функции HVCI. (HVCI и MBEC рассматриваются в главе 9.) То есть злоумышленник может создать в пользовательском пространстве исполняемый участок памяти, что HVCI не запрещает (обозначив соответствующую запись в SLAT исполняемой), после чего повредить PTE (которая не защищена от изменения ядром), тем самым представив виртуальный адрес как страницу ядра. Поскольку MMU сочтет ее собственностью ядра, система предотвращения исполнения в режиме супервайзера (Supervisor Mode Execution Prevention, SMEP) не запретит исполнять там код. В то же время, так как страница изначально была выделена как физическая и пользовательская, запись в SLAT тоже не станет поводом для запрета. Так злоумышленник только что получил возможность запускать посторонний код в режиме ядра, что прямо противоречит основной идее HVCI.

MBEC и родственные ей технологии (ограниченный пользовательский режим — Restricted User Mode) решают эту проблему путем дополнения структуры данных SLAT отдельными битами с признаком доступности пользователю или ядру, что позволяет гипервизору (или безопасному ядру за счет специфичных для VTL1 гипервызовов) помечать пользовательские страницы *не исполняемыми для ядра*, но исполняемыми для пользователя. К сожалению, на процессорах без такой возможности у гипервизора не остается иного выбора, кроме как перехватывать все попытки изменения привилегий по коду и вертеться между двумя наборами записей SLAT: один помечает все физические страницы пользователя неисполняемыми, а второй отмечает исключения из первого набора. Перехват осуществляется за счет представления таблицы IDT пустой (по сути, давая ей лимит в 0 записей) и разбора виновной инструкции, что является дорогой операцией. Однако, поскольку гипервизор способен перехватывать прерывания напрямую, избегая таких затрат, код диспетчера системных вызовов в пространстве пользователя, если опознает систему с активной функцией HVCI без опций MBEC, предпочитает вызывать прерывание. Бит `SystemCall` из структуры общих пользовательских данных (Shared User Data), описанной в главе 4 тома 1, играет ключевую роль в этой ситуации.

Поэтому, когда бит `SystemCall` переключен на 1, Windows x64 применяет инструкцию `int 0x2e`, что приводит к появлению перехватчика, а вместе с ним полностью готового кадра перехвата, которым не требуется участие ОС. Интересное совпадение — это оказалась та же самая инструкция, которая использовалась на

древних x86-процессорах еще до создания Pentium Pro и до сих пор поддерживается x86-системами для обратной совместимости с программами 30-летней давности, которым не повезло жестко зависеть от такой реализации. В свою очередь, в x64-системах инструкция `int 0x2e` применима только в таком сценарии, иначе ядро не заполнит IDT соответствующей записи.

Вне зависимости от того, какая инструкция в итоге была применена, системный вызов пользовательского режима всегда сохраняет *индекс системного вызова* в регистр — EAX на x86- и x64-системах, R12 на 32-разрядных ARM и X8 на ARM64. В дальнейшем он будет анализироваться не зависящим от архитектуры управляющим кодом системы, о котором мы поговорим далее. Наконец, чтобы упростить процесс, блок обработки стандартного вызова функции ABI (двоичный интерфейс приложения) поддерживается независимо от архитектуры: к примеру, для x86 аргументы помещаются в стек, а для x64 — в RCX (технически в RC10 из-за поведения `syscall`), RDX, R8, R9 и в стек для любых аргументов после первых четырех.

Как же процессор возвращается в свое прежнее состояние по завершении обработки? Для системных вызовов, опирающихся на перехват, прошедших через `int 0x2e`, инструкция `iret` восстанавливает состояние процессора, исходя из аппаратного кадра перехвата из стека. В свою очередь, для `syscall` и `sysenter` процессор в очередной раз обращается к MSR и жестко закодированным регистрам, которые мы видели при входе в вызов, с помощью специализированных инструкций — `sysret` и `sysexit` соответственно. Вот как ведет себя первая из них.

- Сегмент стека (регистр SS) загружается из битов STAR с 48-го по 63-й, который Windows приравнивает к 0x0023 (KGDT_R3_DATA).
- Сегмент кода (регистр CS) загружается из битов STAR с 48-го по 63-й плюс 0x10, в итоге давая 0x0033 (KGDT64_R3_CODE).
- Указатель инструкции (RIP) загружается из RCX.
- Регистр флагов (RFLAGS) загружается из R11.
- Указатель стека (RSP) и другие сегментные регистры (DS, ES, FS и GS) сохраняют текущие значения в соответствии с пространством ядра.

Таким образом, как и при входе в системный вызов, алгоритм выхода также должен выполнить очистку состояния процессора. А именно: RSP возвращается в поле `Rsp`, которое было сохранено в аппаратном кадре ловушки из кода входа, который мы анализировали. Аналогично всем другим сохраненным регистрам регистр RCX загружается из сохраненного `Rip`, R11 загружается из `EFlags`, затем сразу перед тем, как вызвать инструкцию `sysret`, вызывается инструкция `swapgs`. Поскольку DS, ES и FS никто не трогал, они сохраняют свои оригинальные значения из пространства пользователя. EDX и XMM0-XMM5 очищаются, и все другие неволатильные регистры восстанавливаются из кадра перехвата перед вызовом инструкции `sysret`. Эквивалентные меры принимаются для инструкций `sysexit` и `eret` (для ARM64). Дополнительно, если активно CET, теневой стек должен быть корректно синхронизирован при выходе, так же как и при входе.

ЭКСПЕРИМЕНТ. Обнаруживаем местонахождение диспетчера вызовов системных служб

Как уже упоминалось, системные вызовы в x64 происходят на основе ряда MSR, для обзора которых можно использовать команду отладчика `rdmsr`. В первую очередь обратите внимание на STAR, который показывает `KGDT_R0_CODE` (0x0010) и `KGDT64_R3_DATA` (0x0023):

```
lkd> rdmsr c0000081
msr[c0000081] = 00230010`00000000
```

Далее можно просмотреть LSTAR, а затем с помощью команды `ln` определить, указывает он на `KiSystemCall64` (для систем, где не требуется теневое копирование KVA) или `KiSystemCall64Shadow` (для тех, где требуется):

```
lkd> rdmsr c0000082
msr[c0000082] = fffff804`7ebd3740

lkd> ln fffff804`7ebd3740
(fffff804`7ebd3740) nt!KiSystemCall64
```

Наконец, взглянем на SFMASK, где должны оказаться значения, упомянутые ранее:

```
lkd> rdmsr c0000084
msr[c0000084] = 00000000`00004700
```

Системные вызовы в x86 делаются через `sysenter`, где используется иной набор MSR, в том числе 0x176, где хранится 32-разрядный обработчик системного вызова:

```
lkd> rdmsr 176
msr[176] = 00000000'8208c9c0

lkd> ln 00000000'8208c9c0
(8208c9c0) nt!KiFastCallEntry
```

Наконец, в системах x86 и x64, где нет MBEC, но есть HVCI, вы можете найти обработчик команды `int 0x2e` зарегистрированным в IDT, использовав команду отладчика `!idt 2e`:

```
lkd> !idt 2e

Dumping IDT: fffff8047af03000
2e:            fffff8047ebd3040 nt!KiSystemService
```

С помощью команды отладчика `u` вы можете дизассемблировать процедуру `KiSystemService` или `KiSystemCall64`, где со временем заметите:

```
nt!KiSystemService+0x227:
fffff804`7ebd3267 4883c408        add     rsp,8
fffff804`7ebd326b 0faee8          lfence
fffff804`7ebd326e 65c604255308000000 mov   byte ptr gs:[853h],0
fffff804`7ebd3277 e904070000      jmp     nt!KiSystemServiceUser
(fffff804`7ebd3980)
```

в то время как обработчик MSR попадет в:

```
nt!KiSystemCall64+0x227:
fffff804`7ebd3970 4883c408        add      rsp,8
fffff804`7ebd3974 0faee8          lfence
fffff804`7ebd3977 65c604255308000000 mov     byte ptr gs:[853h],0
nt!KiSystemServiceUser:
fffff804`7ebd3980 c645ab02        mov      byte ptr [rbp-55h],2
```

Так вы увидите, что обе ветви кода приводят к вызову `KiSystemServiceUser`, отвечающей за наиболее общие действия для всех процессоров, о которых пойдет речь в следующем разделе.

Независимое от архитектуры управление системными службами

Как показано на рис. 8.25, ядро использует нумерацию системных вызовов для поиска информации о службах в *таблице управления системными службами* (system service dispatch table). В системах x86 она подобна описанной ранее таблице управления прерываниями, с той лишь разницей, что в каждой записи находится указатель на системную службу, а не на процедуру обработчика прерывания. На других же платформах, в том числе 32-разрядной ARM и ARM64, таблица реализована немного иначе: вместо указателей на службы в ней хранятся смещения относительно нее самой. Подобный механизм адресации больше подходит для двоичного интерфейса приложений (application binary interface, ABI) x64 и ARM 64, их формата кодирования инструкций и характерных для ARM-решений систем команд типа RISC.

ПРИМЕЧАНИЕ Нумерация системных служб часто меняется между выпусками ОС. Это связано не только с решениями Microsoft об их удалении или добавлении — зачастую таблица специально заполняется случайными номерами, которые перемешиваются с целью защиты от попыток взлома, при которых номера жестко закодированы для скрытности.

Вне зависимости от архитектуры диспетчер системных служб на всех платформах выполняет несколько типичных задач:

- сохраняет в кадре ловушки дополнительные регистры, а именно отладочные и связанные с плавающей точкой;
- если текущий поток относится к процессу Pico, передает управление коду поставщику Pico системных вызовов (более подробно о поставщиках Pico — в главе 3 тома 1);
- если текущий поток запланирован через UMS, вызывает `KiUmsCallEntry` для синхронизации с основным (для знакомства с UMS см. главу 1 тома 1). UMS устанавливает на объектах основных потоков флаг `UmsPerformingSyscall`;
- сохраняет в поле `FirstArgument` объекта потока первый аргумент системного вызова, а в поле `SystemCallNumber` — номер системного вызова;

- вызывает пользовательский/системный разделяемый обработчик системных вызовов (`KiSystemServiceStart`), который помещает в поле `TrapFrame` объекта потока текущий указатель на стек, где он и хранится;
- разрешает доставку прерываний.

На этом этапе поток официально занят обработкой системного вызова, его состояние неизменно, а прерывание допустимо. На следующем шаге выбирается правильная таблица системных вызовов и в некоторых случаях этот поток превращается в GUI-поток, что будет отмечено в полях `GuiThread` и `RestrictedGuiThread` объекта потока, пока поле `GdiBatchCount` не обнулено.

Далее диспетчер системных вызовов должен скопировать все аргументы, не переданные через регистры (в зависимости от архитектуры процессора), из стека пользовательского режима в стек режима ядра. Это необходимо, чтобы освободить системные вызовы от ручного копирования (что потребовало бы дополнительного кода и обработки исключений) и обеспечить невозможность изменения аргументов пользователем, пока с ними работает ядро. Эта операция выполняется в рамках особого блока кода, который опознается обработчиками исключений как занятый копированием пользовательского стека, позволяя защитить ядро от аварийного завершения работы в случаях, когда злоумышленник или просто неверно написанная программа могут повредить стек пользователя. Поскольку системные вызовы могут принимать произвольное количество аргументов (ну почти), в следующем разделе рассматривается, откуда ядро знает, сколько их нужно скопировать.

Обратите внимание на то, что копирование аргументов неглубокое: если какой-либо из аргументов, переданных системной службе, является указателем на некий буфер в пространстве пользователя, его необходимо проверять *на безопасность доступа* прежде, чем код режима ядра сможет обращаться туда. Если ожидается *неоднократный* доступ к буферу, может потребоваться его захват, то есть копирование в локальный буфер ядра. За операции проверки и захвата буфера отвечает каждый системный вызов в отдельности, обработчик же этим не занимается. Однако ключевой операцией, которую должен выполнить обработчик вызова, является установка предыдущего режима искомого потока. Это значение может быть `KernelMode` или `UserMode`, и оно должно синхронизироваться всякий раз, когда текущий поток прерывают, с целью обозначить уровень привилегий поступившего исключения, ловушки или прерывания. Это позволит системному вызову задействовать функцию `ExGetPreviousMode`, чтобы корректно различать вызовы от пользователя или ядра.

Наконец, последние два шага реализованы в рамках кода диспетчера. На первом шаге, если настроен DTrace и активно отслеживание системных вызовов, вызываются соответствующие пред- и постобработчики для системного вызова. Иной вариант, если активно отслеживание ETW, но не DTrace, — до и после вызова журналируются соответствующие события ETW. Наконец, если неактивно ни то ни другое, системный вызов не обременяется никакой дополнительной логикой. Второй и последний шаг — увеличить счетчик `KeSystemCalls` в PRCB, который выводится в роли показателя производительности в системном мониторе Windows.

На этом обработка системного вызова завершается, после чего выполняются действия, обратные описанным, в рамках *выхода из системного вызова*. При этом

будет корректно скопировано и восстановлено состояние пользовательского режима, выполнена доставка APC, приняты меры по защите различных системных буферов от атак по сторонним каналам. В конечном счете работа вернется к одной из инструкций процессора, характерной для данной архитектуры.

Управление системными вызовами ядра

Поскольку системные вызовы могут осуществляться как из пользовательского режима, так и из ядра, все указатели, обработчики и прочая логика должны рассматриваться так, как если бы они происходили из пользовательского режима, что, очевидно, неправильно.

Чтобы разрешить эту проблему, ядро подгружает специализированные версии таких вызовов с префиксом `Zw`. Так, вместо `NtCreateFile` ядро приготовит `ZwCreateFile`. Кроме того, ввиду необходимости делать это вручную третьим лицам будут известны лишь те вызовы, которые Microsoft сочтет нужным показать. К примеру, `ZwCreateUserProcess` не подгружается по имени, так как драйверам уровня ядра не положено запускать пользовательские приложения. Эти экспортированные функции не являются псевдонимами или оболочками для обычных версий с префиксом `Nt`. На деле они служат трамплинами к соответствующим системным вызовам `Nt`, которые обрабатываются тем же самым диспетчером.

Подобно функции `KiSystemCall64`, они создают поддельный кадр аппаратной ловушки, помещая в стек данные, какие бы сформировал процессор после получения прерывания из режима ядра, и отключают прерывания, как сделал бы перехватчик. В системах x64 селектор `KGDT64_R0_CODE` (0x0010) сохраняется в роли регистра кода CS, а текущий стек ядра — как RSP. Каждый из трамплинов помещает номер системного вызова в соответствующий регистр (например, в x86 и x64 это EAX), после чего вызывает функцию `KiServiceInternal`, которая сохраняет в кадр ловушки дополнительные данные и записывает в поле предыдущего режима значение `KernelMode` (и это важное отличие).

Управление системными вызовами из пользовательского режима

Как отмечалось в главе 1 тома 1, инструкции по управлению системными службами для управляющих механизмов Windows располагаются в системной библиотеке `Ntdll.dll`. Библиотеки же подсистем обращаются в нее для реализации своих задокументированных функций. Исключением являются функции Windows из разряда USER и GDI, в том числе графическое ядро DirectX, в которых управление вызовами реализовано в `Win32u.dll`. Библиотека `Ntdll.dll` в этом не участвует. Оба случая представлены на рис. 8.26.

Как показано на схеме, функция Windows `WriteFile` из библиотеки `Kernel32.dll` импортирует одноименную функцию из `API-MS-Win-Core-File-L1-1-0.dll` — одной из перенаправляющих библиотек MinWin (подробнее о перенаправлении API рассказывается в главе 3 тома 1), которая, в свою очередь, вызывает `WriteFile` из `KernelBase.dll`, где и реализован искомый функционал. После нескольких специфичных для подсистемы проверок вызывается функция `NtWriteFile` из `Ntdll.dll`, в рамках которой выполняется инструкция, заставляющая сработать перехватчик системной службы, выдающий соответствующий номер системного вызова.

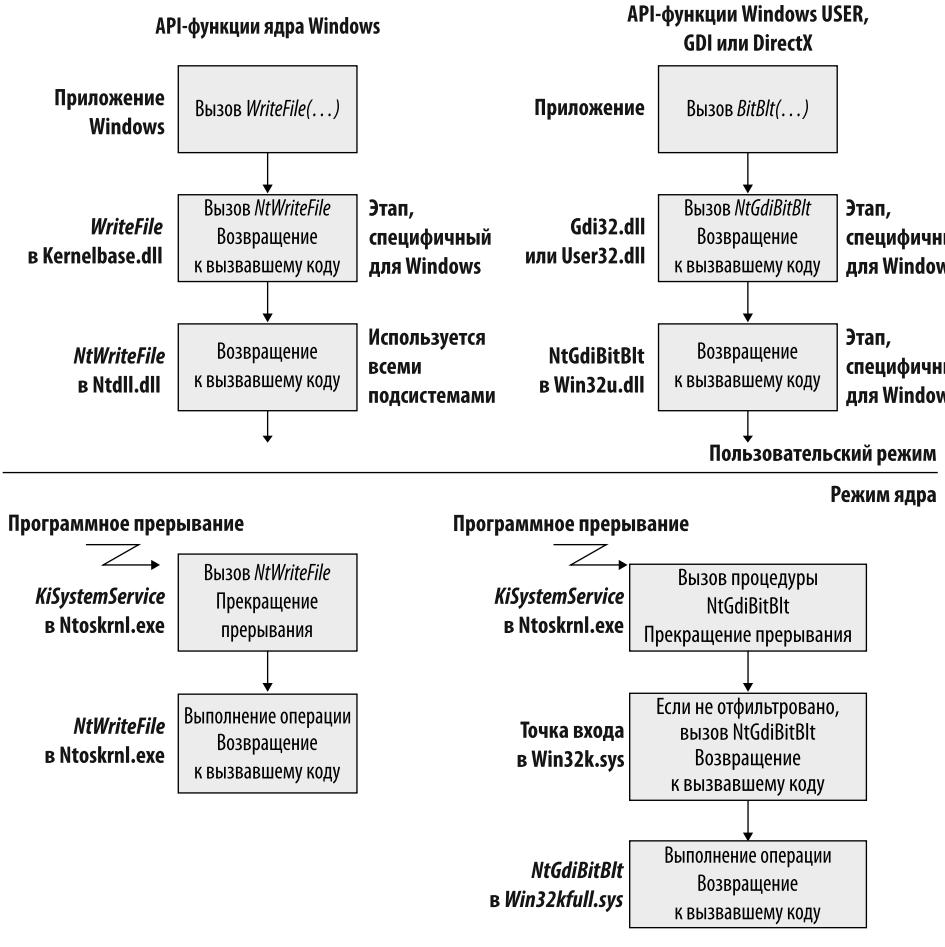


Рис. 8.26. Диспетчеризация системных служб

Далее диспетчер системных вызовов из `Ntoskrnl.exe` (в данном примере это `KiSystemService`) обращается к настоящему коду `NtWriteFile`, чтобы тот обработал запрос на операцию ввода/вывода. Для функций Windows из GDI, USER и графического ядра DirectX диспетчер вызовов обращается к функции из подгружаемой части кода режима ядра в рамках подсистемы Windows `Win32k.sys`, которая, в свою очередь, может отклонить системный вызов или перенаправить его в соответствующий модуль. В настольных системах это `Win32kbase.sys` либо `Win32kfull.sys`, в системах семейства `Win10X` — `Win32min.sys`, и, наконец, вызовы `DirectX` идут в `Dxgkrnl.sys`.

Безопасность системных вызовов

Поскольку ядро обладает механизмами, позволяющими корректно синхронизировать предыдущий режим для нужд системных вызовов, соответствующие службы могут доверять данному параметру в процессе работы. Мы уже упоминали, что при

этом необходимо *проверять* каждый аргумент с указателем на буфер в пользовательском пространстве, каким бы он ни был. Под *проверкой* подразумевается следующее.

1. Убедиться, что адрес находится ниже значения `MmUserProbeAddress`, которое на 64 Кбайт ниже максимального адреса для пользовательского режима (в случае 32 бит это 0x7FFF0000).
2. Убедиться, что адрес выровнен в границах, соответствующих тому, как вызывающий собирается работать с данными. К примеру, для работы с символами Юникод это 2 байта, для 64-битного указателя — 8 байт и т. д.
3. Если буфер предназначен для вывода, убедиться, что на момент запуска вызываемой системной функции он будет открыт для записи.

Обратите внимание на то, что буферы, предназначенные для вывода, в любой момент могут стать неактуальными или используемыми только для чтения, а поэтому системный вызов всегда должен обращаться к ним с помощью SEH, о которой уже говорилось в данной главе, чтобы не рисковать вызвать сбой ядра. По похожей причине, хотя буферы ввода не проверяют на возможность чтения, необходимо использовать SEH, чтобы обеспечить надежность доступа к ним. Увы, это не защитит от ошибок выравнивания или неопределенных указателей, так что первые два шага необходимы в любом случае.

Очевидно, что первая из упомянутых проверок провалится при любом вызове от лица ядра, и это первый шаг, на котором должен играть роль предыдущий режим. Для любого вызова не из пользовательского режима проверка не выполняется, а все буферы считаются актуальными и доступными для записи и/или чтения. Это не единственная предосторожность, которую должен предпринять системный вызов, ввиду вероятности возникновения еще нескольких опасных ситуаций.

- Вызвавшая сторона могла предоставить указатель на объект. Обычно при обращении к объектам ядро пропускает все проверки безопасности, имея при этом полный доступ ко всем подобным ссылкам у ядра (подробнее о которых — в разделе «Диспетчер объектов» далее в этой главе), в отличие от кода пользовательского режима. Значение предыдущего режима позволяет диспетчеру объектов понять, что запрос пришел из пространства пользователя, а значит, необходимо провести все проверки доступности.
- В более сложных случаях драйверу может оказаться необходимо воспользоваться флагом `OBJ_FORCE_ACCESS_CHECK`, чтобы, даже если он задействует функции с префиксом `Zw`, которые указывают предыдущий режим как `KernelMode`, диспетчеру объектов все равно нужно вести себя так, как если бы там был `UserMode`.
- Похожая ситуация возникает, когда вызвавший предоставил имя файла. Для системного вызова важно учитывать возможную необходимость установить флаг `IO_FORCE_ACCESS_CHECKING`, чтобы монитор безопасности ссылок проверил доступ к файловой системе, иначе вызов `ZwCreateFile` укажет в качестве предыдущего режима `KernelMode` и пропустит все проверки. Теоретически драйверу также может потребоваться данный шаг, если ему нужно создать файл от лица IRP в пространстве пользователя.
- Кроме того, сама файловая система создает риски, связанные с символическими ссылками и прочими атаками, построенными на перенаправлении, при которых

привилегированный код режима ядра может неверно применить различные специфичные для процесса/доступные пользователю точки разыменования.

- Наконец, в целом любая операция, приводящая к цепочке системных вызовов, среди которых какие-то имеют префикс `Zw`, должна реализовываться с учетом того, что они будут переводить значение предыдущего режима в `KernelMode`, и принятием соответствующих мер.

Таблица дескрипторов служб

Ранее мы упоминали, что прежде, чем исполнить системный вызов, трамплины пользовательского режима или ядра в первую очередь помещают номер системного вызова в регистр процессора вроде RAX, R12 или X8. Сам номер технически состоит из двух частей, что проиллюстрировано на рис. 8.27. Первая, занимающая нижние 12 бит, представляет *индекс* системного вызова. Вторая представлена еще 2 битами сверху и называется *идентификатором таблицы*. Как вы вскоре увидите, это позволяет ядру реализовать до четырех типов системных служб, каждая с таблицей, вмещающей до 4096 команд.

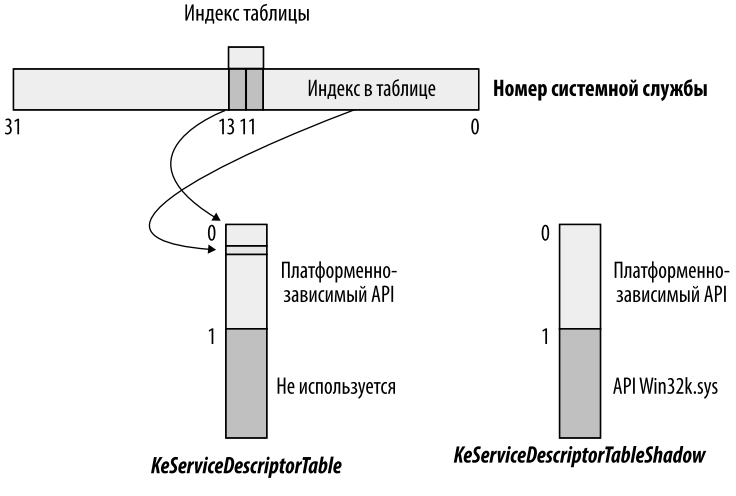


Рис. 8.27. Преобразование номера системной службы к системной службе

Для отслеживания таблиц системных служб ядро может использовать до трех возможных массивов: `KeServiceDescriptorTable`, `KeServiceDescriptorTableShadow` и `KeServiceDescriptorTableFilter`. Любой из них может содержать до двух элементов, в каждом из которых хранятся три параметра:

- указатель на список системных вызовов, реализованных в данной таблице;
- количество системных вызовов, присутствующих в данной таблице, называемое лимитом;
- указатель на массив байтов аргументов для каждого из системных вызовов, входящих в эту таблицу.

Первый массив всегда содержит лишь одну запись, которая указывает на `KiServiceTable` и `KiArgumentTable`, объемом чуть более 450 функций (точное количество зависит от вашей версии Windows). Все потоки по умолчанию делают вызовы, касающиеся только этой таблицы. В системах x86 это устанавливается с помощью указателя в поле `ServiceTable` объекта потока, на всех остальных платформах в диспетчере системных вызовов символ `KeServiceDescriptorTable` кодируется однозначно.

Когда поток впервые делает системный вызов сверх лимита, ядро использует функцию `PsConvertToGuiThread`, которая сообщает об этом потоке службам Windows USER и GDI в Win32k.sys и устанавливает флаг в полях `GuiThread` или `RestrictedGuiThread` объекта потока в случае успешной обработки. Какое из полей задействовать, определяется в зависимости от активности опции защиты процессов `EnableFilteredWin32kSystemCalls`, которая описывалась в разделе «Защитные меры уровня процессов» главы 7 тома 1. В системах x86 указатель из поля `ServiceTable` объекта потока теперь меняет значение на `KeServiceDescriptorTableShadow` или `KeServiceDescriptorTableFilter` в зависимости от конфигурации флагов. На других платформах это жестко закодированная ссылка, выбираемая отдельно при каждом системном вызове (пусть в ущерб производительности, это позволяет избавиться от очевидной для зловредных программ точки перехвата).

Как вы уже, вероятно, догадались, другие два массива содержат вторую запись, которая отображает службы Windows GDI и USER, реализованные в части подсистемы Windows для режима ядра Win32k.sys, и с некоторых пор — службы ядра подсистемы DirectX, реализованные в Dxgkrnl.sys (хотя и они поначалу проходят через Win32k.sys). Запись содержит указатели на `W32pServiceTable`/`W32pServiceTableFilter` и `W32pArgumentTable`/`W32pArgumentTableFilter` соответственно, где доступны примерно 1250 функций или более в зависимости от вашей версии Windows.

ПРИМЕЧАНИЕ Поскольку ядро не ссылается на Win32k.sys, оно экспортирует функцию KeAddSystemServiceTable, которая позволяет добавить еще одну запись в таблицы KeServiceDescriptorTableShadow и KeServiceDescriptorTableFilter, если они еще не заполнены. Если Win32k.sys уже вызывала эти API, функция не отрабатывает, но, так как ее уже вызвали, PatchGuard берет массивы под защиту, фактически предохраняя их от записи.

Единственное реальное отличие записей в таблице `Filter` от прочих состоит в том, что они указывают на системные вызовы в Win32k.sys с именами вроде `stub_UserGetThreadState`, в то время как настоящий массив указывает на `NtUserGetThreadState`. Вызовы с префиксом `stub` будут проверять, не активно ли фильтрование Win32k.sys для данного вызова, отчасти ориентируясь на набор фильтров, загруженный для данного процесса. В зависимости от результата проверки они либо откажут в вызове, вернув `STATUS_INVALID_SYSTEM_SERVICE`, если фильтры укажут на запрет, либо вызовут оригинальную функцию (в частности, `NtUserGetThreadState`), при активном аудите сохранив телеметрию.

Таблицы аргументов, в свою очередь, помогают ядру узнавать, сколько байтов необходимо скопировать из пользовательского стека в стек ядра, как объяснялось ранее в разделе об управлении. Каждая запись в них соответствует системному вызову с таким же индексом и хранит число байтов для копирования (до 255). Однако на платформах, кроме x86, ядро применяет механизм под названием *«сжатие*

таблицы системных вызовов», в рамках которого указатель системного вызова из таблицы вызовов и количество байтов из таблицы аргументов объединяются в одном значении. Это делается следующим образом.

1. Взять указатель на функцию системного вызова и рассчитать 32-битную разницу с адресом начала самой таблицы. Поскольку эти таблицы являются глобальными переменными в том же модуле, где реализованы функции, диапазона ±2 Гбайт должно быть более чем достаточно.
2. Взять количество байтов для стека из таблицы аргументов и разделить на 4, получив *количество аргументов* (некоторые функции имеют аргументы по 8 байт, но в таком случае они просто считаются парами аргументов).
3. Сдвинуть 32-битное смещение из шага 1 на 4 бита влево, сделав его 28-битным (опять же допустимо — в ядре нет компонентов больше 256 Мбайт), затем с помощью побитового сложения через «или» добавить количество аргументов из шага 2.
4. Заменить указатель на функцию системного вызова значением, полученным на шаге 3.

На первый взгляд данная оптимизация может показаться несерьезной, тем не менее у нее есть ряд преимуществ: ввиду отсутствия необходимости просматривать два массива при вызове снижается нагрузка на кэш, упрощается разыменовывание указателей. Кроме того, она представляет собой подобие слоя обфускации, отчего становится сложнее перехватить или изменить таблицу системных вызовов, облегчая работу PatchGuard при ее защите.

ЭКСПЕРИМЕНТ. Соответствие номеров системных вызовов функциям и аргументам

Вы можете продублировать поиск, осуществляемый ядром в ходе работы с ID системного вызова, чтобы определить, какая функция отвечает за его обработку и сколько аргументов она потребует. В системах x86 достаточно будет попросить отладчик выгрузить каждую таблицу вызовов, в том числе `KiServiceTable`, командой `dps`, которая сработает *как указатель для выгрузки*, и поиск будет сделан за вас. Затем можно выгрузить `KiArgumentTable` (или любую из Win32k.sys) командой `db` или просто выгрузить дамп побайтно.

Однако в свете описанного ранее кодирования более интересным упражнением будет выгрузить эти данные в системе x64 или ARM64. Сделать это поможет следующая инструкция.

1. Можете выгрузить конкретный системный вызов, проделав описанное ранее сжатие в обратном порядке. Нужно взять базовый адрес таблицы и прибавить к нему 28-битное смещение по искомому индексу. Как показано далее, системный вызов № 3 в служебной таблице ядра окажется функцией `NtMapUserPhysicalPagesScatter`:

```
lkd> ?? ((ULONG)(nt!KiServiceTable[3]) >> 4) + (int64)nt!KiServiceTable
unsigned int64 0xfffff803`1213e030
```

```
lkd> ln 0xfffff803`1213e030
(fffff803`1213e030) nt!NtMapUserPhysicalPagesScatter
```

2. Количество хранимых в стеке четырехбайтных аргументов можно увидеть, оценив четырехбитное значение:

```
lkd> dx (((int*)&(nt!KiServiceTable))[3] & 0xF)
(((int*)&(nt!KiServiceTable))[3] & 0xF) : 0
```

3. Обратите внимание на то, что это не говорит об отсутствии у вызова аргументов. Поскольку мы в системе x64, их может быть от 0 до 4 и все могут быть переданы через регистры (RCX, RDX, R8 и R9).
4. Можете также воспользоваться моделью данных отладчика, чтобы через проекцию создать предикат LINQ, выгрузив всю таблицу целиком, отталкиваясь от того факта, что переменная `KiServiceLimit` соответствует тому же полю лимита из служебной таблицы дескрипторов (аналогично `W32pServiceLimit` для записей Win32k.sys в теневой таблице дескрипторов). Получится следующий вывод:

```
lkd> dx @$table = &nt!KiServiceTable
@$table = &nt!KiServiceTable : 0xfffff8047ee24800 [Type: void *]

lkd> dx (((int(*)[90000])&(nt!KiServiceTable)))->Take(*(int*)&nt!KiServiceLimit)->
        Select(x => (x >> 4) + @$table)
(((int(*)[90000])&(nt!KiServiceTable)))->Take(*(int*)&nt!KiServiceLimit)->Select
         (x => (x >> 4) + @$table)
     [0]                : 0xfffff8047eb081d0 [Type: void *]
     [1]                : 0xfffff8047eb10940 [Type: void *]
     [2]                : 0xfffff8047f0b7800 [Type: void *]
     [3]                : 0xfffff8047f299f50 [Type: void *]
     [4]                : 0xfffff8047f012450 [Type: void *]
     [5]                : 0xfffff8047ebc5cc0 [Type: void *]
     [6]                : 0xfffff8047f003b20 [Type: void *]
```

5. Можно воспользоваться более сложным вариантом данной команды, который также позволит конвертировать указатели в символические ссылки, практически дублируя работу команды `dps` в Windows x86:

```
lkd> dx @$symPrint = (x => Debugger.Utility.Control.ExecuteCommand
     (".printf \"%y\\n\"," + ((unsigned __int64)x).ToDisplayString("x")).First())
@$symPrint = (x => Debugger.Utility.Control.ExecuteCommand(".printf \"%y\\n\"," +
((unsigned __int64)x).ToDisplayString("x")).First())

lkd> dx (((int(*)[90000])&(nt!KiServiceTable)))->Take(*(int*)
     &nt!KiServiceLimit)->Select (x => @$symPrint((x >> 4) + @$table))
(((int(*)[90000])&(nt!KiServiceTable)))->Take(*(int*)&nt!KiServiceLimit)->
     Select(x =>@$symPrint((x >> 4) + @$table))
     [0]            : nt!NtAccessCheck (fffff804`7eb081d0)
     [1]            : nt!NtWorkerFactoryWorkerReady (fffff804`7eb10940)
     [2]            : nt!NtAcceptConnectPort (fffff804`7f0b7800)
     [3]            : nt!NtMapUserPhysicalPagesScatter (fffff804`7f299f50)
     [4]            : nt!NtWaitForSingleObject (fffff804`7f012450)
     [5]            : nt!NtCallbackReturn (fffff804`7ebc5cc0)
```

6. Наконец, если вас интересует только служебная таблица ядра, без записей Win32k.sys, можно воспользоваться командой отладчика `!chksvctbl -v`,

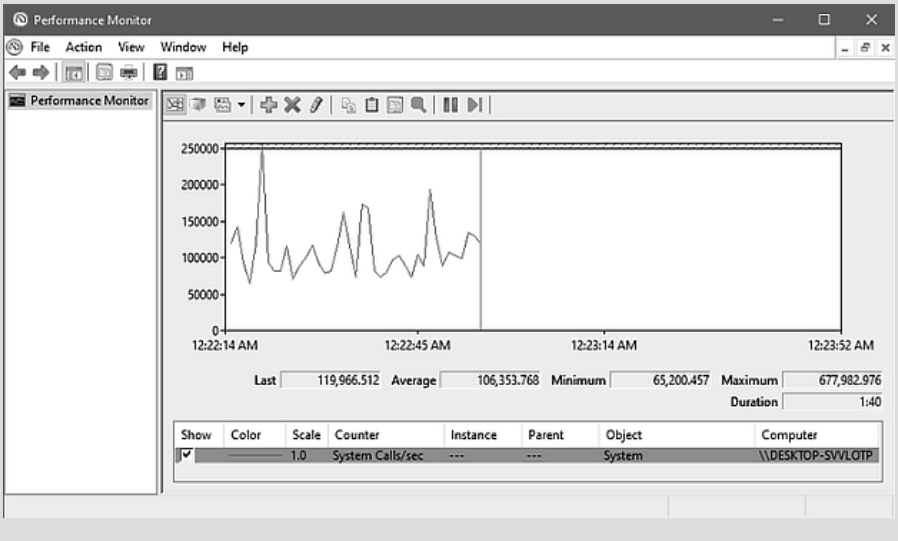
результат которой будет включать все те же данные, заодно с учетом перехватчиков, которые мог установить руткит:

```
lkd> !chksvctbl -v
#    ServiceTableEntry    DecodedEntryTarget(Address)              CompactedOffset
================================================================================
0    0xfffff8047ee24800   nt!NtAccessCheck(0xfffff8047eb081d0) 0n-52191996
1    0xfffff8047ee24804   nt!NtWorkerFactoryWorkerReady(0xfffff8047eb10940)
                                                                   0n-51637248
2    0xfffff8047ee24808   nt!NtAcceptConnectPort(0xfffff8047f0b7800) 0n43188226
3    0xfffff8047ee2480c   nt!NtMapUserPhysicalPagesScatter(0xfffff8047f299f50)
                                                                    0n74806528
4    0xfffff8047ee24810   nt!NtWaitForSingleObject(0xfffff8047f012450) 0n32359680
```

ЭКСПЕРИМЕНТ. Обзор активности системных служб

Вы можете отследить активность системных служб с помощью счетчика производительности системных вызовов. Запустите Системный монитор (Performance Monitor), откройте одноименный подраздел в разделе Средства наблюдения (Monitoring Tools) и нажмите кнопку Добавить (Add), чтобы добавить счетчик на график. Выберите объект Система (System), счетчик Системных вызовов/с (System Calls/Sec) и нажмите Добавить (Add), чтобы отобразить его на графике.

Не исключено, что вам потребуется существенно повысить максимальное значение шкалы, поскольку для системы нормально обрабатывать сотни тысяч системных вызовов в секунду, особенно с ростом числа процессоров. На рисунке далее показано, как выглядел график на компьютере автора.



WOW64 (WINDOWS-ON-WINDOWS)

WoW64 (эмуляция Win32 в среде 64-разрядной Windows) называется ПО, позволяющее запускать 32-разрядные приложения на 64-разрядных платформах (которые могут относиться к сторонней архитектуре). Изначально это был исследовательский проект на тему запуска кода под x86 в старой альфа- и MIPS-версии Windows NT 3.51. С тех пор (примерно с 1995 года) технология существенно развилась. Когда Microsoft в 2001 году выпустила 64-разрядную редакцию Windows XP, WoW64 была включена в ее состав для запуска старых 32-разрядных приложений в новой среде. В современных изданиях ОС технология была расширена до поддержки приложений ARM32 и приложений x86 в условиях ARM64.

Ядро WoW64 реализовано в виде набора DLL пользовательского режима с ограниченной поддержкой со стороны ядра в части создания соответствующих искомой архитектуре структур данных, которые в обычном случае являлись бы стандартными для 64 бит, как блок окружения процесса (process environment block, PEB) или блок окружения потока (thread environment block, TEB). Переключение контекста через функции `Get/SetThreadContext` также реализуется ядром. За WoW64 отвечают следующие основные DLL пользовательского режима.

- **Wow64.dll.** Реализует ядро WoW64 в пользовательском режиме. Создает тонкую программную оболочку, которая играет роль промежуточного ядра для 32-разрядных приложений и запускает симуляцию. Обрабатывает переключения контекста процессора и базовые системные вызовы, экспортируемые Ntoskrnl.exe. Кроме того, библиотека отвечает за перенаправление обращений к файловой системе и реестру Windows.
- **Wow64win.dll.** Реализует конверсию системных вызовов GUI из Win32k.sys. Эта и предыдущая библиотеки содержат код для конверсии, позволяющий вызывать функции с разными конвенциями вызова.

Несколько других модулей специфичны для архитектуры и применяются для преобразования машинного кода для других архитектур. В некоторых случаях (как с ARM64) машинный код приходится эмулировать или динамически компилировать. В данной книге этим термином мы называем технику компиляции кода на ходу, при которой небольшие блоки кода, называемые единицами компиляции, компилируются во время выполнения программы вместо эмуляции среды исполнения или исполнения инструкций по одной.

Далее приводятся DLL, отвечающие за преобразование, эмуляцию или динамическую компиляцию машинного кода, позволяющие последнему исполняться в целевой операционной системе.

- **Wow64cpu.dll.** Реализует симуляцию процессора для 32-разрядного кода под x86 в операционных системах на базе AMD64. Управляет 32-разрядным контекстом процессора для каждого потока в рамках WoW64 и предоставляет зависимый от архитектуры процессора функционал переключения между разрядностями и т. п.
- **Wowarmhw.dll.** Реализует симуляцию процессора для запуска приложений под ARM32 (AArch32) в системах под ARM64. Представляет собой эквивалент Wow64cpu.dll для x86 в среде ARM64.

- **Xtajit.dll.** Реализует эмулятор процессора для запуска 32-разрядных приложений под x86 в системах на базе ARM64. Включает в себя полноценную симуляцию x86, динамический компилятор кода и протокол коммуникации между динамическим компилятором и сервером кэша XTA. Динамический компилятор способен создавать блоки компиляции, в том числе кода под ARM64, преобразованного из данных образа под x86. Впоследствии они хранятся в локальном кэше.

Взаимодействие библиотек пользовательского режима WoW64 (в том числе других ключевых компонентов) показано на рис. 8.28.

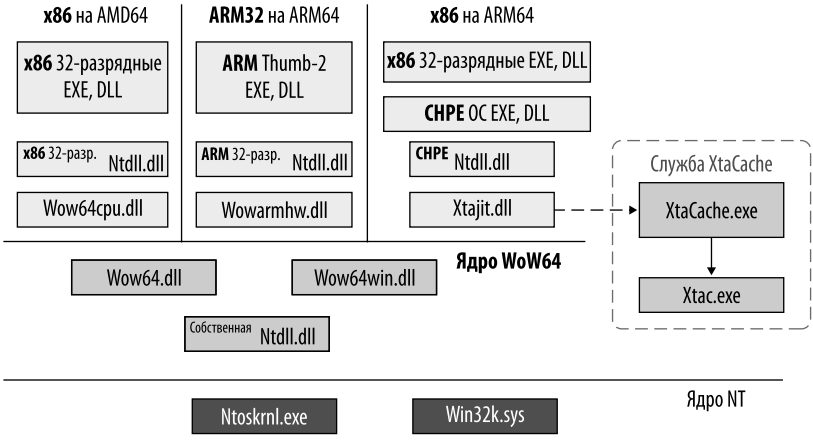


Рис. 8.28. Архитектура WoW64

ПРИМЕЧАНИЕ Старые версии Windows, предназначенные для работы с процессорами архитектуры Itanium, также имели в наличии полноценный эмулятор x86, встроенный в слой WoW64 в библиотеке Wowai32x.dll. Такие процессоры не могли самостоятельно исполнять 32-разрядные инструкции под x86 с должной эффективностью, так что это оказалось необходимо. Архитектура Itanium была официально снята с производства в январе 2019 года.

Существует новая инсайдерская версия Windows, которая также поддерживает исполнение 64-разрядного кода под x86 в системах ARM64. По этой причине был разработан отдельный динамический компилятор. Однако эмуляция кода под AMD64 в ARM-системах происходит без участия WoW64. Подробности архитектуры эмулятора AMD64 в этой книге мы рассматривать не будем.

Ядро WoW64

Как говорилось в предыдущем разделе, ядро WoW64 кросс-платформенно: оно лишь создает программный слой для исполнения 32-разрядного кода в 64-разрядной среде. Само же преобразование осуществляется другим компонентом — симулятором (также известен как двоичный транслятор), который уже зависит от платформы. В данном разделе мы рассмотрим роль ядра WoW64 и то, как оно

взаимодействует с симулятором. Хотя подавляющая часть ядра WoW64 реализована в пользовательском режиме (библиотека WoW64.dll), отдельные мелкие фрагменты работают в ядре NT.

Ядро WoW64 в ядре NT

Во время запуска системы (фаза 1) диспетчер ввода/вывода вызывает функцию `PsLocateSystemDlls`, которая отображает все поддерживаемые системные DLL (и сохраняет их базовые адреса в глобальном массиве) в адресном пространстве системного пользователя. В эту совокупность входят вариации Ntdll для WoW64, как показано в табл. 8.13. Во время запуска диспетчера процессов в фазе 2 вычисляются точки входа в некоторые из этих DLL, которые сохраняются во внутренних переменных ядра. Одна из экспортируемых функций, `LdrSystemDllInitBlock`, используется для переноса информации WoW64 и указателей на функции во вновь создаваемые процессы WoW64.

Таблица 8.13. Различные версии Ntdll

Путь	Внутреннее имя	Описание
c:\windows\system32\ntdll.dll	ntdll.dll	Системная Ntdll, отображаемая в любой пользовательский процесс (кроме минимальных). Единственная версия, считающаяся обязательной
c:\windows\SysWow64\ntdll.dll	ntdll32.dll	32-разрядная Ntdll под x86, отображаемая в процессах WoW64, запущенных в среде 64 бита под x86
c:\windows\SysArm32\ntdll.dll	ntdll32.dll	32-разрядная Ntdll под ARM, отображаемая в процессах WoW64, запущенных в среде 64 бита под ARM64
c:\windows\SyChpe32\ntdll.dll	ntdllwow.dll	32-разрядная Ntdll под CHPE x86, отображаемая в процессах WoW64, запущенных в среде 64 бита под x86

Когда процесс только создан, ядро определяет, нужен ли ему WoW64, с помощью алгоритма, который проверяет основной исполняемый PE-образ и разыскивает в памяти системы подходящую версию Ntdll. Если специальная среда необходима, то при выделении процессу адресного пространства ядро сразу отразит туда как основной вариант библиотеки, так и предназначенный для искомого варианта WoW64.

Как объяснялось в главе 3 тома 1, любой неминимальный процесс обладает структурой данных PEB (блок окружения процесса), доступной в пользовательском режиме. Для процессов под WoW64 ядро также размещает 32-разрядную версию PEB и сохраняет указатель на нее в небольшой структуре (`EWoW64PROCESS`), связанной с основным EPROCESS, представляющим вновь созданный процесс. Затем заполняется структура данных, описанная 32-разрядной версией символа `LdrSystemDllInitBlock`, включая указатели на функции из Ntdll для WoW64.

Когда в рамках процесса создается поток, ядро выполняет примерно те же действия: вслед за изначальным пользовательским стеком (его первоначальный размер указан в PE-заголовке главного образа) создается еще один для исполнения 32-разрядного кода. Последний называется WoW64-стеком потока. Когда TEB (блок окружения потока) построен, ядро выделяет столько памяти, сколько нужно для размещения 64-разрядного TEB и 32-разрядного вслед за ним.

Кроме того, внизу 64-разрядного стека размещается небольшая структура данных — сведения о процессорной области WoW64. Она включает в себя идентификатор компьютера, под который создавался исполняемый образ, зависящий от платформы 32-разрядный контекст для процессора (структуры `X86_NT5_CONTEXT` или `ARM_CONTEXT` в зависимости от целевой архитектуры) и указатель на специфические для потока общие данные процессора WoW64, которые могут быть использованы симулятором. Указатель на данную структуру хранится в слоте 1 TLS потока, чтобы двоичный транслятор мог быстро получить доступ к нему. На рис. 8.29 представлена итоговая конфигурация процесса WoW64 и его единственного основного потока.

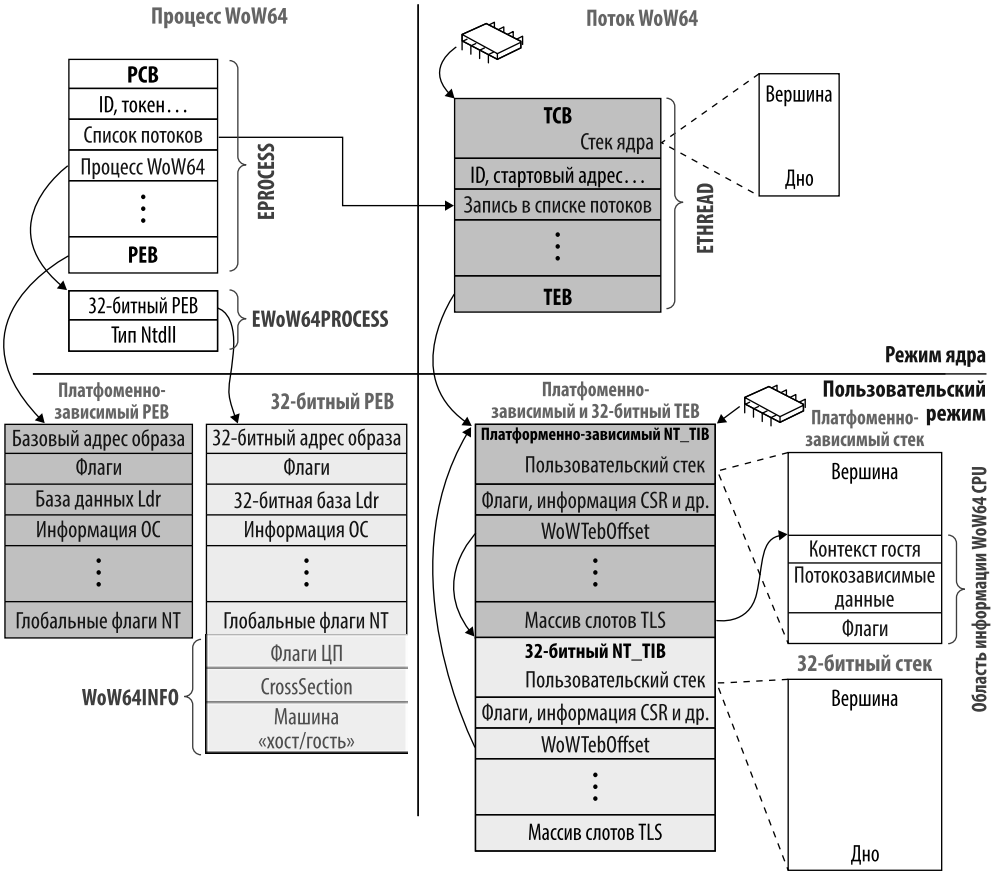


Рис. 8.29. Внутренняя конфигурация процесса WoW64 с одним-единственным потоком

Ядро WoW64 пользовательского режима

Не считая отличий, описанных в предыдущем разделе, инициализация процесса и его первого потока происходит так же, как и без участия WoW64, до тех пор, пока поток не начнет исполняться, вызвав функцию инициализации загрузчика `LdrpInitialize` из основного варианта Ntdll. Определив, что поток должен будет исполняться в контексте нового процесса первым, он вызывает функцию инициализации процесса `LdrpInitializeProcess`, которая, кроме множества других действий (подробнее — в подразделе «Ранняя стадия инициализации процесса» главы 3 тома 1), определяет, требуется ли WoW64. Признаком для проверки служит факт наличия 32-разрядного TEB, который следует сразу за 64-разрядным и связан с ним. При положительном результате основная Ntdll присваивает внутренней глобальной переменной `UseWoW64` значение `1`, строит путь к библиотеке ядра WoW64, wow64.dll, и отражает ее за пределами 4-гигабайтного виртуального адресного пространства (таким образом он сможет вмешиваться в симулируемое 32-разрядное адресное пространство процесса). Затем Ntdll находит адреса некоторых функций WoW64, связанных с приостановкой процесса/потока и обработкой APC и исключений, и сохраняет их в своих внутренних переменных.

По завершении функции инициализации процесса загрузчик Windows передает управление ядру WoW64, вызывая экспортированную функцию `Wow64LdrpInitialize`, из которой уже не будет возврата. С этого момента любой новый поток будет стартовать с этой точки входа вместо классической `RtlUserThreadStart`. Ядро WoW64 получает указатель на область процессора WoW64, сохраненный ранее ядром системы в слоте 1 TLS. Если поток является первым в рамках процесса, он запускает функцию инициализации процесса WoW64, которая выполняет следующие шаги.

1. Делает попытку загрузить библиотеку параллельного журналирования WoW64 (wow64log.dll). Она используется для отслеживания вызовов WoW64 и не входит в коммерческие издания Windows, так что этот шаг просто пропускается.
2. Находит базовый адрес Ntdll32 и указатели на функции с помощью `LdrSystemDllInitBlock`, заранее заполненной ядром NT.
3. Инициализирует перенаправление обращений к реестру и файловой системе. Происходит оно на слое Syscall ядра WoW64, который перехватывает 32-разрядные запросы и перестраивает пути в них, прежде чем передать управление обычным вызовам.
4. Подготавливает таблицы служб WoW64, где размещаются указатели на системные службы ядра NT и подсистемы GUI Win32k (подобно стандартным службам ядра). При этом дополнительно туда попадают системные службы консоли и NLS (системные вызовы WoW64 и перенаправление рассматриваются далее в этой главе).
5. Заполняет 32-разрядную версию PEB процесса, выделенную ядром, и загружает симулятор, подходящий для архитектуры основного исполняемого образа. Система обращается к реестру по стандартному пути HKLM\SOFTWARE\Microsoft\Wow64\<arch> (<arch> может быть x86 или ARM в зависимости от целевой архитектуры), где получает имя основной DLL симулятора. Последний загружается

и отражается в адресном пространстве процесса. Некоторые функции оттуда экспортируются и сохраняются во внутреннем массиве `BtFuncs`. Он играет ключевую роль в связи между зависящим от платформы двоичным транслятором и подсистемой WoW64: впредь она будет обращаться к нему только так. К примеру, функция `BtCpuProcessInit` реализует меры по инициализации процесса симулятором.

6. Инициализирует межпроцессный механизм преобразования путем выделения и отражения раздела в 16 Кбайт. Когда процесс WoW64 будет вызывать API из другого 32-разрядного процесса, там будет появляться синтезированная запись в очереди (таким образом преобразование работает между разными процессами).
7. Слой WoW64 оповещает симулятор (вызывая его функцию `BtCpuNotifyMapViewOfSection`), что основной модуль и 32-разрядная версия Ntdll были отражены в адресное пространство.
8. Наконец, ядро WoW64 сохраняет указатель на 32-разрядный диспетчер системных вызовов в экспортированную переменную `Wow64Transition` 32-разрядной Ntdll. Это обеспечивает работу основного диспетчера.

По окончании инициализации процесса поток готов приступить к симуляции процессора. Он запускает функцию симулятора по инициализации потока и готовит новый 32-разрядный контекст на основе 64-разрядного, ранее заполненного ядром. Наконец, на основе нового контекста готовится 32-разрядный стек для выполнения 32-разрядной же версии функции `LdrInitializeThunk`. Симуляция начинается с запуска экспортированной функции `BTCpuSimulate`, которая больше не вернет управление, если только в симуляторе не случится критических ошибок.

Перенаправление путей в файловой системе

Чтобы обеспечить совместимость приложений и облегчить их перенос из 32-разрядных Windows в 64-разрядные, пути к системным папкам были оставлены прежними. Таким образом, по пути `Windows\System32` стали находиться чисто 64-разрядные образы. WoW64, перехватывая все системные вызовы, преобразует все связанные с файлами функции API, заменяя пути к различным системным папкам эквивалентом, который зависит от архитектуры целевого процессора (табл. 8.14). В этой таблице предусмотрено также перенаправление по системным переменным окружения (к примеру, переменная `%ProgramFiles%` раскрывается как `\Program files (x86)` для 32-разрядных приложений и как `\Program Files` — для 64-разрядных).

Таблица 8.14. Перенаправление путей в WoW64

Путь	Архитектура	Путь перенаправления
c:\windows\system32	X86 на AMD64	C:\Windows\SysWow64
	X86 на AMD64	C:\Windows\SyChpe32 (или C:\Windows\SysWow64, если не найден путь к папке SyChpe32)
	ARM32	C:\Windows\SysArm32

Путь	Архитектура	Путь перенаправления
%ProgramFiles%	Родная	C:\Program Files
	X86	C:\Program Files (x86)
	ARM32	C:\Program Files (Arm)
%CommonProgramFiles%	Родная	C:\Program Files\Common Files
	X86	C:\Program Files (x86)
	ARM32	C:\Program Files (Arm)\Common Files
C:\Windows\regedit.exe	X86	C:\Windows\SysWow64\regedit.exe
	ARM32	C:\Windows\SysArm32\regedit.exe
C:\Windows\LastGood\System32	X86	C:\Windows\LastGood\SysWow64
	ARM32	C:\Windows\LastGood\SysArm32

По пути `\Windows\System32` существует несколько подпапок, к которым по соображениям совместимости и безопасности перенаправление не применяется, в частности, когда 32-разрядные приложения при попытке доступа обращаются по реальному пути. В их число входят:

- `%windir%\system32\catroot` и `%windir%\system32\catroot2`;
- `%windir%\system32\driverstore`;
- `%windir%\system32\drivers\etc`;
- `%windir%\system32\hostdriverstore`;
- `%windir%\system32\logfiles`;
- `%windir%\system32\spool`.

Наконец, в WoW64 имеется механизм контроля встроенного перенаправления до момента разделения на потоки с помощью функций `Wow64DisableWow64FsRedirection` и `Wow64RevertWow64FsRedirection`. Принцип его работы сводится к сохранению значения (да/нет) в позиции 8 в TLS, которое затем оценивается внутренней функцией WoW64 `RedirectPath`. Однако у этого подхода есть проблемы с отложенной загрузкой DLL, открытием файлов через стандартное диалоговое окно и даже интернационализацией. Если перенаправление выключить, система больше не задействует его при внутренних действиях по загрузке, из-за чего некоторые файлы, присущие только x64, могут быть не найдены. В типичном случае разработчикам было бы безопаснее использовать один из упомянутых ранее постоянных путей, таких как `%SYSTEMROOT%\Sysnative`.

ПРИМЕЧАНИЕ Поскольку у некоторых 32-разрядных приложений действительно порой есть потребность знать о 64-разрядных образах и взаимодействовать с ними, виртуальная папка %SYSTEMROOT%\Sysnative позволяет любым операциям ввода/вывода из 32-разрядного приложения не подвергаться перенаправлению в файловой системе. Данной папки на самом деле не существует, это виртуальное отражение содержимого реальной папки System32, даже если приложение контролируется WoW64.

Перенаправление в системном реестре

Приложения и их компоненты хранят настройки конфигурации в системном реестре. Зачастую эти данные помещаются туда в момент регистрации при установке. Если один и тот же компонент зарегистрировать как и 32-разрядный, и 64-разрядный исполняемый файл, последняя настройка перекроет первую из-за одинакового пути в реестре.

Чтобы помочь найти прозрачное решение данной проблемы без внесения существенных изменений в код 32-разрядных компонентов, реестр разделили на две части: родную и WoW64. По умолчанию 32-разрядные компоненты видят свое представление, а 64-разрядные — свое. Таким образом обеспечиваются безопасная среда исполнения для компонентов разных архитектур и разделение настроек 32-разрядного приложения и его 64-разрядного аналога, если тот существует.

WoW64 перехватывает все системные вызовы любого 32-разрядного процесса. Когда таким вызовом оказываются команды открытия или создания раздела в реестре, WoW64 переводит переданный путь так, чтобы он указывал на соответствующее представление (не считая случаев, когда запрашивается непосредственно 64-разрядное представление). Подсистема способна отслеживать перенаправленные разделы благодаря множеству древовидных структур данных, в которых сохраняются списки общих и раздельных разделов и подразделов (узловой элемент дерева указывает, в какую сторону системе следует выполнить перенаправление). WoW64 осуществляет перенаправление для следующих разделов:

- HKLM\SOFTWARE;
- HKEY_CLASSES_ROOT.

Не вся сеть расположенных ниже разделов разбивается. Подразделы могут быть сохранены в отдельном разделе реестра для WoW64 (в таком случае подраздел будет разделенным). Или же подраздел может использоваться совместно 64- и 32-разрядными приложениями (тогда он считается общим). В каждом из общих разделов WoW64 создает подраздел под названием `WoW6432Node` (для приложений x86) или `WoWAA32Node` (для приложений ARM32). В нем хранится 32-разрядная настроечная информация. Все остальная часть реестра используется совместно 32-разрядными и 64-разрядными приложениями (например, HKLM\SYSTEM).

В качестве дополнительной помощи: если 32-разрядное приложение x86 запишет `REG_SZ`- или `REG_EXPAND_CZ`-значение, которое начинается со строк `"%ProgramFiles%"` или `"%CommonProgramFiles%"`, в реестр, WoW64 будет заменять подобные фрагменты так, чтобы это соответствовало перенаправлению запросов к файловой системе по принципу, изложенному ранее. Тридцатидвухбитным приложениям в таком случае всегда следует пользоваться этими строками — любые другие данные будут проигнорированы и записаны как есть.

Для приложений, которым необходим доступ к разделам реестра из конкретного представления, получить его позволят следующие флаги для функций `RegOpenKeyEx`, `RegCreateKeyEx`, `RegOpenKeyTransacted`, `RegCreateKeyTransacted` и `RegDeleteKeyEx`:

- `KEY_WoW64_64KEY` — явно потребовать 64-разрядный раздел как из 32-разрядного, так и из 64-разрядного приложения и отказаться от перехвата вышеупомянутых значений `REG_SZ` или `REG_EXPAND_SZ`;
- `KEY_WoW64_32KEY` — явно потребовать 32-разрядный раздел как из 32-разрядного, так и из 64-разрядного приложения.

Симуляция X86 на платформах AMD64

Интерфейс симулятора x86 для платформ AMD64 (Wow64cpu.dll) довольно прост. Функция инициализации процесса в симуляторе разрешает использование быстрых системных вызовов в зависимости от наличия программного MBEC (подробнее о Mode Based Execute Control см. в главе 9). Когда ядро Wow64 начинает симуляцию, вызвав метод интерфейса симулятора `BtCpuSimulate`, тот строит кадр стека WoW64 (на основе 32-разрядного контекста процессора, заданного ядром WoW64), инициализирует массив скоростных (turbo) прокси-функций для управления быстрыми системными вызовами и готовит регистр FS, чтобы тот указывал на 32-разрядный TEB потока. Затем настраивается шлюз вызова, нацеленный на 32-разрядный сегмент (обычно это сегмент 0x20), переключаются стеки и выполняется длинный переход (far jump) к искомой 32-разрядной точке входа (при первом исполнении точка входа берется от 32-разрядной версии функции-загрузчика `LdrInitializeThunk`). Когда процессор обрабатывает этот длинный переход, он определяет, что шлюз вызова направлен на 32-разрядный сегмент, в связи с чем переводит свой режим исполнения в 32-разрядный. Этот режим прекращает действовать только в случае обработки прерывания или системного вызова. Больше деталей на тему шлюзов вызова можно найти в документации по разработке ПО для процессоров Intel и AMD.

ПРИМЕЧАНИЕ При первом переключении в 32-разрядный режим симулятор использует машинную команду IRET вместо обычного дальнего вызова. Так делается из-за необходимости инициализировать все 32-разрядные регистры, включая волатильные и EFLAGS.

Системные вызовы

С 32-разрядными приложениями слой WoW64 ведет себя примерно так же, как и ядро NT: особые версии Ntdll.dll, User32.dll и Gdi32.dll располагаются в папке `\Windows\Syswow64` (как и некоторые другие DLL, реализующие обмен данных между процессами, такие как Rpcrt4.dll). Когда 32-разрядному приложению требуется помощь от ОС, оно обращается к функциям, реализованным в этих 32-разрядных же системных библиотеках. Так же как и их 64-разрядные аналоги, те могут как выполняться напрямую в пользовательском режиме, так и запрашивать помощь ядра NT. Во втором случае они делают системные вызовы через функцию-заглушку наподобие тех, которые есть в обычной 64-разрядной ntdll.dll. Заглушка помещает индекс системного вызова в регистр, но вместо подачи обычной инструкции 32-разрядного системного вызова она обращается к диспетчеру системных вызовов WoW64 (через переменную `Wow64Transition`, зашитую в ядре WoW64).

Диспетчер системных вызовов WoW64 реализован в рамках соответствующего платформе симулятора (wow64cpu.dll). Тот переключается в родной 64-разрядный режим исполнения, покидая симуляцию. Двоичный транслятор переключает стек на 64-разрядный и сохраняет прежний контекст процессора. Затем он перехватывает параметры, связанные с системными вызовами, и конвертирует их. Процесс конвертации называют преобразованием (Thunking), он позволяет машинному коду в рамках 32-разрядного ABI взаимодействовать с 64-разрядным кодом. Данное соглашение о вызове, описываемое ABI, определяет, как структура данных, указатели и значения передаются через параметры каждой функции и предоставляются для использования машинным кодом.

Симулятор выполняет преобразование, руководствуясь двумя стратегиями. Для тех API, которым не требуется взаимодействовать со сложными структурами данных, переданных клиентом (но достаточно простых входных и выходных параметров), скоростные прокси-функции (небольшие функции-конвертеры, реализованные симулятором) исполняют конверсию и напрямую обращаются к родным 64-разрядным API. Другие же сложные API нуждаются в помощи функции `Wow64SystemServiceEx`, которая извлекает корректный номер из таблицы системных вызовов WoW64, после чего вызывает правильную функцию системного вызова WoW64. Системные вызовы WoW64 реализованы в библиотеке ядра и Wow64win.dll и имеют те же имена, что и родные вызовы, но с префиксом `wh-` (так, API WoW64 `NtCreateFile` будет называться `whNtCreateFile`).

После удачного завершения конверсии симулятор отправляет соответствующий родной 64-разрядный системный вызов. Когда тот вернет результат, WoW64 при необходимости конвертирует (или выполнит преобразование) все выходные параметры из 64-разрядного в 32-разрядный формат и перезапустит симуляцию.

Управление исключениями

Подобно системным вызовам, в WoW64 обработка исключений принудительно выводит процессор из симуляции. Когда возникает исключение, ядро NT определяет, был ли его виновником поток, исполнявший код пользовательского режима. Если это так, ядро строит в активном стеке расширенный кадр исключения и обрабатывает его, возвращаясь в функцию `KiUserExceptionDispatcher` в 64-разрядной Ntdll (более подробно об обработке исключений см. далее в пункте «Системные вызовы и обработка исключений»).

Стоит заметить, что 64-разрядный кадр исключения (где заодно сохранен захваченный контекст процессора) располагается в 32-разрядном стеке, который был активен в момент возникновения исключения. Следовательно, прежде, чем передавать его симулятору процессора, его нужно конвертировать. Именно эту задачу решает функция `Wow64PrepareForException` (экспортируемая из библиотеки ядра WoW64), которая выделяет пространство в родном 64-разрядном стеке и конвертирует как родное исключение, так и записи о контексте в их 32-разрядные аналоги, помещая результат обратно в 32-разрядный стек, заменяя этим 64-разрядный кадр исключения. На этом моменте ядро WoW64 может продолжить симуляцию с вызова 32-разрядной версии функции `KiUserExceptionDispatcher`, которая обрабатывает исключение тем же способом, что и родная 32-разрядная Ntdll.

Доставка 32-разрядных APC пользовательского режима реализована подобным образом. Обычный APC пользовательского режима доставляется через функцию `KiUserApcDispatcher` из родной Ntdll. Когда 64-разрядное ядро будет готово направить этот APC к процессу из WoW64, оно отразит 32-разрядный адрес APC в завышенный диапазон 64-разрядного адресного пространства. Затем 64-разрядная Ntdll обращается к функции `Wow64ApcRoutine`, экспортируемой из библиотеки ядра WoW64, которая перехватывает родной APC и данные контекста в пользовательском режиме, после чего возобновляет симуляцию процессора с 32-разрядной версии функции `KiUserApcDispatcher`, что, в свою очередь, обрабатывает APC тем же способом, что и 32-разрядная Ntdll.

ARM

ARM называют семейство архитектур ЭВМ с сокращенным набором команд (Reduced Instruction Set Computing), разработанное компанией ARM Holding. В отличие от Intel и AMD они лишь проектируют процессоры, после чего продают лицензии на производство другим фирмам, таким как Qualcomm и Samsung, где и производится конечный продукт. В итоге было выпущено множество релизов и версий архитектуры ARM, за несколько лет быстро развившейся из очень простых 32-разрядных процессоров, составлявших поколение ARMv3 в 1993 году, в последние ARMv8. Новейшие ARM64v8.2 из коробки поддерживают несколько режимов исполнения (или состояний), чаще всего это AArch32, Thumb-2 и AArch64.

- AArch32 — это классический режим исполнения, при котором процессор выполняет только 32-разрядный код и обменивается данными с основной памятью по 32-разрядной шине, используя 32-разрядные же регистры.
- Thumb-2 — это режим исполнения, ставший подвидом AArch32. Его набор инструкций был спроектирован с целью улучшения плотности кода во встраиваемых системах с низким энергопотреблением. В этом режиме процессор способен исполнять смесь из 16- и 32-разрядных команд, сохраняя при этом доступ к 32-разрядной памяти.
- AArch64 — это современный режим исполнения. В этом состоянии процессор имеет доступ к 64-разрядным регистрам общего назначения и способен обмениваться данными с основной памятью по 64-разрядной шине.

Windows 10 для систем с ARM64 способна работать в режимах исполнения AArch64 и Thumb-2 (AArch32 обычно не применяется). Thumb-2 был особо широко распространен в старых системах Windows RT. Текущее состояние процессора ARM64 также определяется уровнем исключений (exception level, EL), дающим различные уровни привилегий. На сегодняшний день ARM поддерживает три уровня исключений и два режима безопасности. Оба последних подробно рассматриваются в главе 9 и руководстве по архитектуре ARM (ARM Architecture Reference Manual).

Модели памяти

Ранее в данной главе, в разделе «Аппаратные уязвимости к атакам по сторонним каналам», мы описали концепцию протокола согласованности кэша, чья задача — гарантировать, что данные в кэше ядра процессора находятся под наблюдением при доступе туда множества процессоров (один из самых знаменитых протоколов согласованности кэша — MESI). Как и этот протокол, современные процессоры должны предоставлять *модель последовательности (упорядоченности) памяти*, чтобы решать еще одну проблему, порой возникающую в мультипроцессорной среде, — проблему сортировки памяти. Некоторые архитектуры (в частности, ARM64) действительно способны свободно пересортировывать порядок обработки запросов к памяти с целью повысить эффективность подсистемы управления памятью и распараллелить предоставление доступа к ней, обеспечивая повышенную производительность в условиях медленной шины памяти. Такой вариант архитектуры

использует слабую модель памяти, в отличие от AMD64, где модель сильная и предполагает, что команды доступа туда обычно исполняются в порядке следования в коде программы. Слабые модели позволяют процессору быть быстрее и работать с памятью более эффективно, но приносят с собой массу проблем синхронизации при разработке многопроцессорного ПО. Сильная же модель, наоборот, более интуитивна и стабильна, но значительно проигрывает в быстродействии.

Процессоры с возможностью пересортировки памяти, придерживающиеся слабой модели, поддерживают машинные инструкции, которые служат *барьерами памяти*. Барьер не позволяет процессору перестраивать очередь запросов к памяти до и после него, помогая решать проблемы синхронизации. Однако барьеры памяти довольно медлительны, а поэтому используются строго по потребностям мультипроцессорного кода в Windows, особенно в примитивах для синхронизации, а именно в спин-блокировках, мьютексах, push-блокировках и т. д.

Как будет подробнее рассмотрено далее, динамический компилятор ARM64 постоянно использует барьеры памяти для преобразования кода x86 в мультипроцессорной среде. На самом деле он не способен предсказать, будет ли поступивший код исполняться множеством потоков одновременно, а значит, будет рисковать проблемами с синхронизацией. В x86 используется сильная модель, и там проблемы с переупорядочением памяти нет, лишь очередной случай исполнения вне очереди, как объяснялось в предыдущем разделе.

ПРИМЕЧАНИЕ Кроме процессора, сортировка доступа может коснуться компилятора, который в ходе компиляции может изменить (и, возможно, сократить) последовательность обращений к памяти в исходном коде по соображениям скорости и эффективности. Подобные действия называются компиляторной сортировкой, в то время как в предыдущем разделе рассматривалась процессорная.

Симуляция ARM32 на платформах ARM64

Симуляция для приложений ARM32 в условиях ARM64 в своем исполнении весьма схожа с симуляцией x86 под AMD64. Как отмечалось в предыдущем разделе, процессор ARM64v8 способен динамически переключаться между режимами исполнения AArch64 и Thumb-2, в котором 32-разрядные инструкции исполняются аппаратно. Однако, в отличие от систем на базе AMD64, процессор не может поменять режим исполнения в пользовательском режиме какой-то особой командой, из-за чего слою WoW64 требуется обращаться к ядру NT для переключения. Чтобы это сделать, функция `BtCpuSimulate` из симулятора ARM на ARM64 (Wowarmhw.dll) сохраняет неволатильные регистры AArch64 в 64-разрядный стек, восстанавливает 32-разрядный контекст, оставленный в зоне процессора WoW64, после чего наконец отправляет корректный системный вызов, у которого некорректный номер syscall `-1`.

Обработчик исключений ядра NT (который в условиях ARM64 и есть обработчик syscall) определяет, что исключение пришло через системный вызов, проверяя его особый номер `-1`, и если это так, ядро NT понимает, что от WoW64 поступило требование изменить режим исполнения. В этом случае он вызывает функцию

`KiEnter32BitMode`, которая меняет состояние на более низкий уровень исключений, он же AArch32, гасит исключение и переходит в пользовательский режим.

Код продолжает исполняться в режиме AArch32. Как и в случае с симулятором x86 на AMD64, контроль над исполнением возвращается симулятору только при возникновении исключения или системном вызове. Обе ситуации обрабатываются так же, как и там.

Симуляция x86 на платформах ARM64

Симулятор процессора x86 под ARM64 (Xtajit.dll) отличается от других двоичных трансляторов, описанных далее, поскольку исполнять код аппаратно нет возможности. Процессор ARM64 попросту не может воспринимать никаких инструкций x86. По этой причине в симуляторе x86 под ARM64 реализованы полноценный эмулятор x86 и динамический компилятор, который способен переводить блоки машинного кода для x86 в код для ARM64 и уже их исполнять напрямую.

Когда функция инициализации процесса в симуляторе (`BtCpuProcessInit`) начинает работу над новым процессом в слое WoW64, она формирует для него главный раздел реестра динамического компилятора, комбинируя путь из раздела HKLM\SOFTWARE\Microsoft\Wow64\x86\xtajit с именем основного исполняемого образа процесса. Если этот раздел существует, симулятор запрашивает из него множество настроек конфигурации (чаще всего это касается мультипроцессорной совместимости и предельного размера блока динамического компилятора. Заметим также, что заодно с этим симулятор запрашивает настройки из базы данных совместимости приложений). Затем симулятор размещает и заполняет страницу системных вызовов (Syscall), которая, как следует из названия, используется для запуска страниц системных вызовов для x86 (а потом связывается с Ntdll с помощью переменной `Wow64Transition`). В этот момент симулятор определяет, может ли процесс задействовать кэш XTA.

Для хранения уже скомпилированных блоков кода симулятор использует два вида кэша: внутренний кэш распределяется в разрезе потоков и содержит блоки, полученные при компиляции кода x86, исполняемого конкретным потоком (эти блоки называют *динамически компилированными блоками*), внешний кэш XTA управляется службой XTACache и содержит все динамически компилированные блоки, созданные для образа x86 ею же. XTA-кэш для каждого образа находится во внешнем файле (более подробно — далее). Функция инициализации процесса дополнительно размещает битовую карту скомпилированного гибридного переносимого исполняемого файла (Compiled Hybrid Portable Executables, CHPE), которая полностью описывает адресное пространство в 4 Гбайт, теоретически занятое 32-разрядным процессом. Данная карта единичными битами отмечает страницы в памяти, содержащие код CHPE (более подробно о CHPE поговорим в дальнейшем).

Функция инициализации потока в симуляторе `BtCpuThreadInit` готовит к работе компилятор и распределяет состояние процессора для каждого потока в родном стеке. Это важная структура данных, где для каждого потока хранятся его контекст x86, состояние эмиттера кода x86, внутренний код кэша и конфигурация эмулируемого процессора x86 (сегментные регистры, состояние FPU, эмулируемые CPUID).

Оповещение загрузки образа для симулятора

В отличие от всех прочих двоичных трансляторов симулятор x86 под ARM64 необходимо информировать всякий раз, когда в памяти процесса отразился новый образ, в том числе CHPE для Ntdll. Эту задачу решает ядро WoW64, которое перехватывает вызов родной функции `NtMapViewOfSection` из 32-разрядного кода и сообщает об этом эмулятору Xtajit с помощью экспортированной функции `BTCpuNotifyMapViewOfSection`. Это оповещение важно, поскольку симулятору требуется обновить значимые внутренние данные компилятора, в том числе:

- битовую карту CHPE (ее надо обновлять, меняя значения битов на 1 и отмечая наличие страниц с кодом CHPE);
- состояние защиты потока управления (Control Flow Guard, CFG);
- состояние кэша XTA для этого образа.

В частности, всякий раз, когда загружается новый образ x86 или CHPE, симулятор определяет, следует ли ему использовать кэш XTA для данного модуля, посредством анализа данных в реестре и базе совместимости приложений. В случае успешной проверки симулятор обновляет состояние глобального кэша XTA в разрезе процессов и запрашивает у службы XTACache обновленный кэш для искомого образа. Если служба сможет опознать и открыть обновленный файл с кэшем, она возвращает симулятору объект раздела, который можно будет использовать для ускорения исполнения образа (в нем будут находиться заранее скомпилированные блоки кода под ARM64).

Гибридно скомпилированные исполняемые образы

Динамическая компиляция процесса x86 в средах ARM64 является сложной задачей, поскольку компилятору необходимо поддерживать производительность, достаточную для того, чтобы приложение не зависало. Одна из основных проблем связана с различием порядка доступа к памяти в двух архитектурах. Эмулятор x86 не в курсе, как был спроектирован оригинальный код, поэтому вынужден агрессивно применять барьеры памяти между любыми попытками доступа к ней в рамках образа. Как операция барьер работает медленно. В среднем более 40 % времени исполнения кода большинство приложений тратят на исполнение кода операционной системы. Это позволяет заключить, что неэмулируемые библиотеки имеют значительное преимущество по производительности приложения в целом.

Это и послужило мотивацией к созданию гибридно скомпилированных исполняемых образов (CHPE). Исполняемый файл CHPE — это особый гибридный образ, в котором содержится код, совместимый и с x86, и с ARM64, созданный с полным пониманием исходного кода (иными словами, компилятор точно знал, где использовать барьеры памяти). ARM64-совместимый машинный код называется гибридным (или CHPE): он все еще выполняется в режиме AArch64, но сформирован по правилам 32-разрядного ABI для лучшего взаимодействия с кодом x64.

Образы CHPE создаются как обычные исполняемые файлы под x86 (ID машины равен 014C, как и для x86), их основное отличие заключается в том, что в них содержится гибридный код, описываемый таблицей метаданных гибридного образа (она входит в каталог конфигурации загрузки образа). Когда CHPE загружается

в адресное пространство WoW64, симулятор обновляет битовую карту CHPE, устанавливая по одному биту для каждой страницы с гибридным кодом, исходя из гибридных метаданных. Когда динамический компилятор собирает блоки кода x86 и обнаруживает попытку вызова гибридной функции, он немедленно ее вызывает (через 32-разрядный стек), не тратя времени на ее компиляцию.

Обработанный динамическим компилятором код x86 исполняется после особого ABI, учитывающего наличие нестандартного соглашения о том, как использовать регистры ARM64 и как между функциями передаются параметры. Код CHPE не следует тем же соглашениям, которых придерживается код после динамического компилятора (хотя гибридный код все равно следует 32-разрядному ABI). Это значит, что напрямую вызвать код CHPE из динамически компилированного кода нет прямой возможности. Для решения этой проблемы образы CHPE дополнительно содержат три вида прокси-функций, позволяющих взаимодействовать с кодом для x86.

- Прокси-функция `pop` дает коду x86 возможность вызвать гибридную функцию путем конвертирования отправляемых (или возвращаемых) гостем (x86) аргументов в соответствии с соглашением CHPE и прямой передачи управления гибридному коду.
- Прокси-функция `push` позволяет CHPE вызвать функцию под x86 путем конвертирования отправляемых (или возвращаемых) гибридным кодом аргументов в соответствии с соглашением гостя (x86) и обращения к эмулятору, чтобы тот продолжил исполнение кода x86.
- Прокси-функция `export` обеспечивает совместимость с приложениями, которые модифицируют функции x86, экспортированные из модулей ОС, с целью изменения их функциональности. Функции, экспортированные из модулей CHPE, все еще содержат чуть-чуть кода x86 (обычно 8 байт), которые семантически не добавляют ничего, но позволяют модификации внешними приложениями.

Симулятор x86 под ARM64 прилагает максимум усилий для того, чтобы всегда загружать системные модули CHPE вместо стандартных x86, но порой это невозможно. Если CHPE-версия библиотеки отсутствует, симулятор использует обычную из папки SysWow64. В таком случае системная библиотека будет целиком перекомпилирована.

ЭКСПЕРИМЕНТ. Выгрузка таблицы диапазонов адресов гибридного кода

Инкрементальный компоновщик Microsoft (link.exe), входящий в состав Windows SDK и WDK, способен отобразить некоторую часть информации в рамках гибридных метаданных из каталога конфигурации загрузки образа CHPE. Более подробно этот инструмент и способ его установки описаны в главе 9.

В ходе следующего эксперимента вы сможете выгрузить гибридные метаданные из kernelbase.dll — системной библиотеки, которая также была скомпилирована с поддержкой CHPE. Можете поэкспериментировать и с другими такими

библиотеками. Установив SDK или WDK на компьютер с ARM64, откройте командную строку разработчика Visual Studio (или запустите скрипт LaunchBuildEnv.cmd, если используете образ ISO от EWDK). Переместите папку SHPE и выгрузите каталог конфигурации загрузки образа из kernelbase.dll следующими командами:

```
cd c:\Windows\SyChpe32
link /dump /loadconfig kernelbase.dll > kernelbase_loadconfig.txt
```

Заметим, что в данном примере результат выполнения команды был перенаправлен в текстовый файл `kernelbase_loadconfig.txt`, поскольку он был слишком велик для вывода в консоль. Откройте файл в Блокноте и прокрутите вниз, пока не увидите следующий текст:

```
Section contains the following hybrid metadata:

               4 Version
        102D900C Address of WowA64 exception handler function pointer
        102D9000 Address of WowA64 dispatch call function pointer
        102D9004 Address of WowA64 dispatch indirect call function pointer
        102D9008 Address of WowA64 dispatch indirect call function pointer
          (with CFG check)
        102D9010 Address of WowA64 dispatch return function pointer
        102D9014 Address of WowA64 dispatch leaf return function pointer
        102D9018 Address of WowA64 dispatch jump function pointer
        102DE000 Address of WowA64 auxiliary import address table pointer
        1011DAC8 Hybrid code address range table
               4 Hybrid code address range count


    Hybrid Code Address Range Table

               Address Range
          ----------------------
          x86   10001000 - 1000828F (00001000 - 0000828F)
          arm64 1011E2E0 - 1029E09E (0011E2E0 - 0029E09E)
          x86   102BA000 - 102BB865 (002BA000 - 002BB865)
          arm64 102BC000 - 102C0097 (002BC000 - 002C0097)
```

Кэш XTA

Как описывалось в предыдущих разделах, симулятор x86 под ARM64, кроме своего внутреннего кэша, в разрезе потоков использует некий глобальный кэш под названием XTA, управляемый защищенной службой XTACache, которая реализует отложенный динамический компилятор. Она запускается автоматически, после чего открывает или создает папку `C:\Windows\XtaCache`, которую защищает специальным ACL (только сама служба XTA и участники группы «Администраторы» имеют право туда заходить). Служба запускает собственный сервер ALPC через порт соединения `{BEC19D6F-D7B2-41A8-860C-8787BB964F2D}`. Затем, прежде чем отключиться, она регистрирует рабочие потоки для отложенного динамического компилятора и ALPC.

Рабочий поток ALPC отвечает за диспетчеризацию всех входящих запросов на сервере ALPC. В частности, если симулятор (клиент), запущенный в контексте процесса

из WoW64, присоединяется к службе XTACache, для его отслеживания создается новая структура данных, которая потом хранится во внутреннем списке вместе с отображенным в памяти разделом размером 128 Кбайт, доступным и клиенту, и XtaCache (оригинальная область памяти называется буфером отслеживания). Она используется симулятором для отправки подсказок о коде x86, который скомпилировали для исполнения, но которого еще нет ни в одном из кэшей, вместе с ID его модуля. Содержимое раздела обрабатывается кэшем XTA каждую секунду, а кроме того, при каждом заполнении буфера. В зависимости от количества корректных записей в списке служба может принять решение о прямом запуске отложенного динамического компилятора.

Когда в память процесса отражается новый образ, слой WoW64 информирует об этом симулятор, который обращается к XTACache для поиска существующих файлов кэша. Чтобы их найти, служба XTACache сначала должна открыть исполняемый образ, разгрузить и рассчитать его хеши. Хешей создается два: один — на основе пути к исполняемому файлу, второй — по бинарным данным внутри него. Оба они необходимы для того, чтобы избежать исполнения блоков, скомпилированных из устаревшей версии исполняемого образа. Имя файла кэша XTA формируется по следующей схеме: *<Имя_модуля>.<Хеш_заголовка>.<Хеш_расположения>.<много-/однопроц.>.<версия_кэш-файла>*.jc. Файл кэша содержит все ранее скомпилированные блоки кода, которые могут быть исполнены симулятором напрямую. В случаях, когда удается обнаружить подходящий файл кэша, XTACache создает отражение раздела из файла и внедряет его в клиентский процесс WoW64.

Отложенный динамический компилятор располагается в центре XTACache. Когда служба принимает решение вызвать его, создается и инициализируется новая версия файла кэша для искомого модуля x86. После этого динамический компилятор приступает к отложенной компиляции, запуская независимый компилятор XTA (xtac.exe). Последний исполняется в защищенной низкопривилегированной среде (в процессе AppContainer), которая имеет низкий приоритет. Единственная задача компилятора — транслировать код x86, исполненный симулятором. Новые блоки кода добавляются к находящимся в старой версии кэш-файла (если она существует) и сохраняются в новой версии.

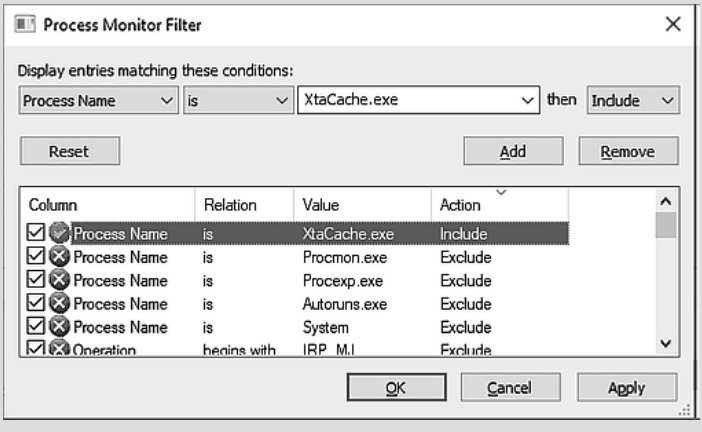
ЭКСПЕРИМЕНТ. Обзор кэша XTA

Новые версии Process Monitor способны без эмуляции работать в средах ARM64. Вы можете использовать их, чтобы посмотреть, как файлы кэша XTA создаются и применяются для работы процесса под x86.

Для данного эксперимента вам нужна будет система на базе ARM64 с установленной Windows 10 с обновлениями минимум от мая 2019 года (1903). Для начала потребуется приложение x86, которое еще никогда не запускалось на этой системе. В данном примере установим старую MPC-HC Media Player для x86, которую можно скачать по адресу https://sourceforge.net/projects/mpc-hc/files/latest/download. Впрочем, подойдет и любое другое приложение для x86.

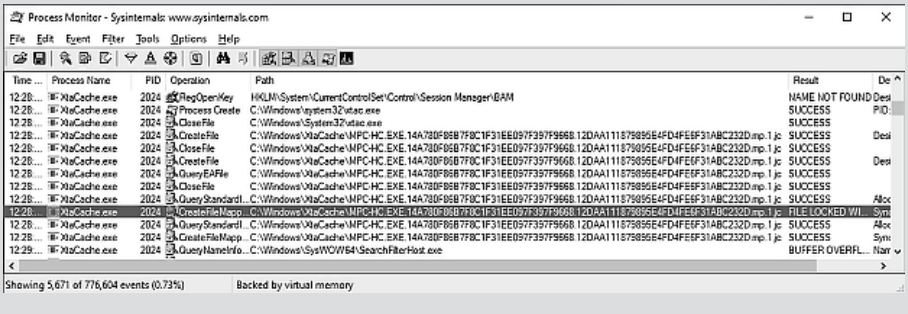
Установите MPC-HC (или собственное приложение x86), но прежде, чем запустите, включите Process Monitor и добавьте фильтр по имени процесса службы

XTACache (XTACache.exe, у службы свой процесс, она его ни с кем не делит). Настройки фильтра должны будут выглядеть следующим образом.



Если вы этого еще не сделали, начните захват событий, выбрав Захват событий (Capture Events) в меню Файл (File). Теперь запустите MPC-HC и попробуйте включить какое-нибудь видео. Выключите MPC-HC и остановите захват событий в Process Monitor. Количество последних в его окне будет велико. Вы можете проредить результат, отключив показ действий в реестре нажатием на соответствующий значок на панели инструментов (в данном эксперименте реестр нас не интересует).

Прокрутив список событий, вы обнаружите, что служба XTACache сначала попыталась открыть файл кэша MPC-HC, но не смогла, так как его нет. Это значит, что симулятор приступил к компиляции образа x86 самостоятельно, в ходе чего периодически отправлял информацию в XTACache. Позже отложенный динамический компилятор был запущен XTACache в рабочем потоке. Он создал новую версию файла кэша XTA и запустил компилятор Xtac, отразив раздел файла для себя и для него.



Перезапустив эксперимент, вы увидите в Process Monitor иные события: файл кэша будет немедленно отражен в процесс MPC-HC в WoW64. Таким образом,

симулятор может исполнять его напрямую. В результате быстродействие должно возрасти. Вы также можете попытаться удалить сформированный файл кэша XTA. Служба XTACache автоматически восстановит его, как только вы снова запустите приложение MPC-HC для x86.

Однако стоит помнить, что папка `%SystemRoot%\XtaCache` защищена точно определенным ACL, владеет которым лишь она сама. Чтобы попасть туда, вам потребуется открыть командную строку с правами администратора и выполнить следующие команды:

```
takeown /f c:\windows\XtaCache
icacls c:\Windows\XtaCache /grant Administrators:F
```

Динамическая компиляция и исполнение

Чтобы запустить гостевой процесс, симулятор x86 под ARM64 может лишь либо интерпретировать, либо динамически скомпилировать код x86. Интерпретирование гостевого кода сведется к переводу и исполнению по одной инструкции за раз, а это медленный процесс, поэтому эмулятор сразу прибегает к динамической компиляции: он динамически компилирует код x86 в ARM64 и сохраняет результат в гостевом блоке кода, пока не будут выполнены определенные условия:

- обнаружена точка останова на данных или в коде либо получена неверная команда;
- встретилась инструкция ветвления, направляющая в уже посещенный блок;
- размер блока превышает предопределенный лимит (512 байт).

Механизм симуляции начинает работу с проверки локального и XTA-кэшей на предмет того, существует ли уже искомый блок кода (индексируемый по своему RVA). Если блок уже присутствует в кэше, симулятор напрямую исполняет его с помощью функции *диспетчера*, которая компонует контекст ARM64, содержащий значения регистров хоста, и сохраняет их в 64-разрядный стек, потом переключается на 32-разрядный стек и подготавливает его для состояния гостевого потока x86. Кроме того, она готовит регистры ARM64 к исполнению скомпилированного кода x86, помещая туда контекст x86. Обратите внимание на наличие нестандартного соглашения вызова: диспетчер похож на прокси-функцию `pop`, которая используется для передачи управления из CHPE в контекст x86.

Когда исполнение блока кода заканчивается, диспетчер действует в обратном порядке — сохраняет новый контекст x86 в 32-разрядном стеке, переключается на 64-разрядный, а затем восстанавливает прежний контекст ARM64, содержащий состояние эмулятора. После завершения работы диспетчера симулятор знает точный виртуальный x86-адрес места, в котором исполнение прервалось. Позже он сможет возобновить эмуляцию по новому адресу. Подобно тому, как если бы он работал с записями кэша, симулятор проверяет, указывает ли целевой адрес на страницу с кодом CHPE (он может проверить это благодаря глобальной битовой карте CHPE). Если это так, симулятор находит прокси-функцию `pop` для искомой функции, добавляет ее адрес в локальный кэш потока, а потом напрямую исполняет ее.

Когда выполнено одно из двух приведенных ранее условий, симулятор может повести себя примерно так, как если бы работал с родными образами. Иначе ему бы пришлось обращаться к компилятору, чтобы тот собрал ему блок кода, преобразованного в машинный код. Процесс компиляции делится на три фазы.

1. На этапе **анализа** готовятся дескрипторы всех команд, которые могут потребоваться в блоке.
2. На этапе **оптимизации** последовательность инструкций оптимизируется.
3. Наконец, на этапе **генерации** в блок записывается готовый машинный код ARM64.

Затем сгенерированный блок кода помещается в локальный кэш каждого потока. Заметим, что симулятор не может добавить его в кэш XTA в основном по соображениям производительности и безопасности. В ином же случае злоумышленник сможет повредить кэш процесса с более высокими привилегиями, из-за чего в контексте такого процесса теоретически станет можно исполнить вредоносный код. Кроме того, у симулятора недостаточно процессорного времени на генерацию высокооптимизированного кода (даже при наличии фазы оптимизации) при поддержании стабильности работы приложения.

Однако информация о скомпилированных блоках x86 вместе с ID образов, к которым они принадлежали, вставляется в список, отражаемый в общем буфере отслеживания. Благодаря последнему отложенный динамический компилятор кэша в курсе, что ему необходимо скомпилировать код x86, недавно обработанный динамическим компилятором симулятора. В результате он создает оптимизированные блоки кода и добавляет их в XTA-кэш для модуля, который будет исполняться симулятором напрямую. Только самое первое исполнение процесса x86 будет в целом медленнее прочих.

Системные вызовы и обработка исключений

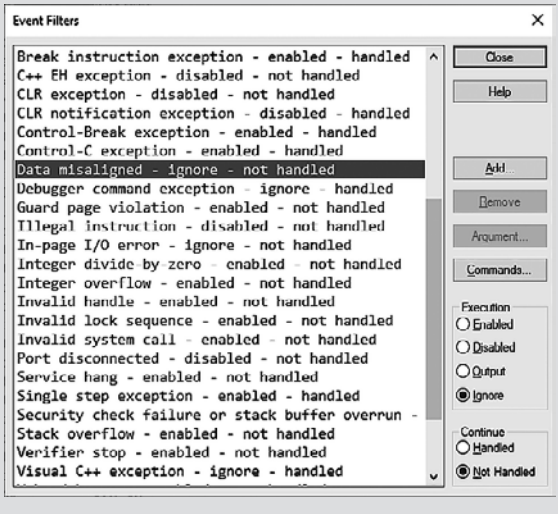
В симуляторе процессора x86 под ARM64, когда поток x86 совершает системный вызов, вызывается код со страницы системного вызова, созданной симулятором, который генерирует исключение 0x2E. Каждое исключение x86 вынуждает блок кода завершить работу. Диспетчер же при этом обрабатывает исключение через внутреннюю функцию, которая в итоге обращается либо к стандартному обработчику исключений WoW64, либо к диспетчеру системных вызовов — в зависимости от номера вектора исключения. Эти особенности уже упоминались в предыдущем разделе данной главы на тему эмуляции x86 на платформах AMD64.

ЭКСПЕРИМЕНТ. Отладка WoW64 в средах ARM64

Современные выпуски WinDbg (отладчика Windows) способны отлаживать машинный код в условиях любого симулятора. Это означает, что в системах с ARM64 вы сможете работать как с родными приложениями под ARM64, ARM Thumb-2, так и с кодом x86. В системах же с AMD64 отладка возможна для программ под 32- и 64-разрядную x86. Кроме того, отладчик способен легко переключаться между родными 32- и 64-разрядным стеками, что позволяет пользователю отлаживать как родной (включая слой WoW64 и сам эмулятор), так и гостевой код (а еще отладчик поддерживает CHPE).

В рамках следующего эксперимента вы откроете приложение для x86 на компьютере с ARM64 и переключитесь между тремя режимами исполнения: ARM64, Thumb-2 и x86. Для этого потребуется последняя версия инструментов отладки, которые можно найти в составе WDK или SDK. Установив один из этих пакетов, запустите версию WinDbg для ARM64 (доступна из меню Пуск).

Прежде чем начать сеанс отладки, вам нужно будет запретить исключения, создаваемые эмулятором XtaJit, такие как Data Misaligned (ошибка выравнивания данных) и ошибки ввода-вывода в рамках страницы (эти исключения эмулятор обрабатывает сам). В меню Отладка (Debug) выберите пункт Фильтры событий (Event Filters). Из списка выберите событие Data misaligned и отметьте его флажком Игнорировать (Ignore) из блока Выполнение (Execution). Повторите то же самое для ошибки In-page I/O. В итоге ваша конфигурация должна будет походить на следующий пример.



Нажмите Закрыть (Close), после чего в основном окне отладчика из меню Файл (File) выберите Открыть исполняемый (Open Executable). Выберите один из 32-разрядных исполняемых файлов под x86 из папки `%SystemRoot%\SysWOW64`. (В данном примере используется Блокнот, но подойдет любое приложение x86.) Откройте также окно дизассемблера из меню Вид (View). Если у вас правильно настроены отладочные символы (ищите инструкцию по настройке символов на сайте https://docs.microsoft.com/en-us/windows-hardware/drivers/debugger/symbol-path), вы увидите первую родную точку останова Ntdll, что можно проверить, выведя стек командой k:

```
0:000> k
# Child-SP          RetAddr           Call Site
00 00000000`001eec70 00007ffb`bd47de00 ntdll!LdrpDoDebuggerBreak+0x2c
01 00000000`001eec90 00007ffb`bd47133c ntdll!LdrpInitializeProcess+0x1da8
02 00000000`001ef580 00007ffb`bd428180 ntdll!_LdrpInitialize+0x491ac
03 00000000`001ef660 00007ffb`bd428134 ntdll!LdrpInitialize+0x38
04 00000000`001ef680 00000000`00000000 ntdll!LdrInitializeThunk+0x14
```

Симулятор к этому моменту все еще не загружен: Ntdll, платформенно-зависимая (native) и CHPE, были отображены в целевой образ ядром NT, в то время как исполняемые компоненты ядра WoW64 были загружены платформенно-зависимой Ntdll прямо перед той точкой останова с помощью функции `LdrpLoadWow64`. Это можно проверить, перечислив загруженные модули командой `lm` и перейдя в следующий кадр стека командой `.f+`. В окне дизассемблера вы увидите вызов функции `LdrpLoadWow64`:

```
00007ffb`bd47dde4 97fed31b bl         ntdll!LdrpLoadWow64 (00007ffb`bd432a50)
```

Теперь продолжим исполнение с помощью команды `g` (и клавиши F5). Вы увидите, как множество модулей подгружается в адресное пространство процесса, после чего срабатывает еще одна точка останова, теперь уже в контексте x86. Если вы снова выведете стек командой `k`, то заметите, что появилась новая колонка. Более того, теперь отладчик дополняет командную строку символами x86:

```
0:000:x86> k
 #     Arch ChildEBP RetAddr
00     x86 00acf7b8 77006fb8 ntdll_76ec0000!LdrpDoDebuggerBreak+0x2b
01    CHPE 00acf7c0 77006fb8 ntdll_76ec0000!#LdrpDoDebuggerBreak$push_thunk+0x48
02    CHPE 00acf820 76f44054 ntdll_76ec0000!#LdrpInitializeProcess+0x20ec
03    CHPE 00acfad0 76f43e9c ntdll_76ec0000!#_LdrpInitialize+0x1a4
04    CHPE 00acfb60 76f43e34 ntdll_76ec0000!#LdrpInitialize+0x3c
05    CHPE 00acfb80 76ffc3cc ntdll_76ec0000!LdrInitializeThunk+0x14
```

Сравнив старый и новый стеки, вы заметите, как сильно изменились адреса в нем, поскольку процесс теперь исполняется с 32-разрядным стеком. Обратите внимание на то, что у некоторых функций в начале имени появился символ #: Windbg обозначает им функции с кодом CHPE. На этот момент вы уже можете пошагово, со входом или без, отлаживать код x86, как если бы находились в его родной системе. Симулятор обеспечивает эмуляцию и скрывает все детали. Чтобы понаблюдать за его работой, вам понадобится перейти в 64-разрядный контекст по команде `.effmach`. Она может принимать разные параметры: `x86` для контекста x86, `arm64` или `amd64` для платформенно-зависимого 64-разрядного контекста (зависит от целевой платформы), `arm` для контекста ARM Thumb-2, `CHPE` для 32-разрядного контекста CHPE. Сейчас мы перейдем в 64-разрядный стек с помощью параметра `arm64`:

```
0:000:x86> .effmach arm64
Effective machine: ARM 64-bit (AArch64) (arm64)
0:000> k
 # Child-SP          RetAddr           Call Site
00 00000000`00a8df30 00007ffb`bd3572a8 wow64!Wow64pNotifyDebugger+0x18f54
01 00000000`00a8df60 00007ffb`bd3724a4 wow64!Wow64pDispatchException+0x108
02 00000000`00a8e2e0 00000000`76e1e9dc wow64!Wow64RaiseException+0x84
03 00000000`00a8e400 00000000`76e0ebd8 xtajit!BTCpuSuspendLocalThread+0x24c
04 00000000`00a8e4c0 00000000`76de04c8 xtajit!BTCpuResetFloatingPoint+0x4828
05 00000000`00a8e530 00000000`76dd4bf8 xtajit!BTCpuUseChpeFile+0x9088
06 00000000`00a8e640 00007ffb`bd3552c4 xtajit!BTCpuSimulate+0x98
07 00000000`00a8e6b0 00007ffb`bd353788 wow64!RunCpuSimulation+0x14
08 00000000`00a8e6c0 00007ffb`bd47de38 wow64!Wow64LdrpInitialize+0x138
09 00000000`00a8e980 00007ffb`bd47133c ntdll!LdrpInitializeProcess+0x1de0
0a 00000000`00a8f270 00007ffb`bd428180 ntdll!_LdrpInitialize+0x491ac
```

```
0b 00000000`00a8f350 00007ffb`bd428134 ntdll!LdrpInitialize+0x38
0c 00000000`00a8f370 00000000`00000000 ntdll!LdrInitializeThunk+0x14
```

Из этих двух стеков видно, что эмулятор исполнял код CHPE, после чего была вызвана прокси-функция `push`, которая возобновила симуляцию с x86-функции `LdrpDoDebuggerBreak`, где возникло исключение (обработанное платформенно-зависимой `Wow64RaiseException`), оповестившее отладчик через функцию `Wow64pNotifyDebugger`. С помощью команды отладчика `.effmach` вы сможете вести отладку фактически сразу в нескольких контекстах: платформенно-зависимого, CHPE- и x86-кода. С помощью команды `g @$exentry` вы можете переместиться к точке входа x86 в Блокноте и продолжить сеанс отладки кода x86 самого эмулятора. А можете повторить этот эксперимент в иной среде, попробовав отладить, к примеру, какое-то приложение из SysArm32.

ДИСПЕТЧЕР ОБЪЕКТОВ

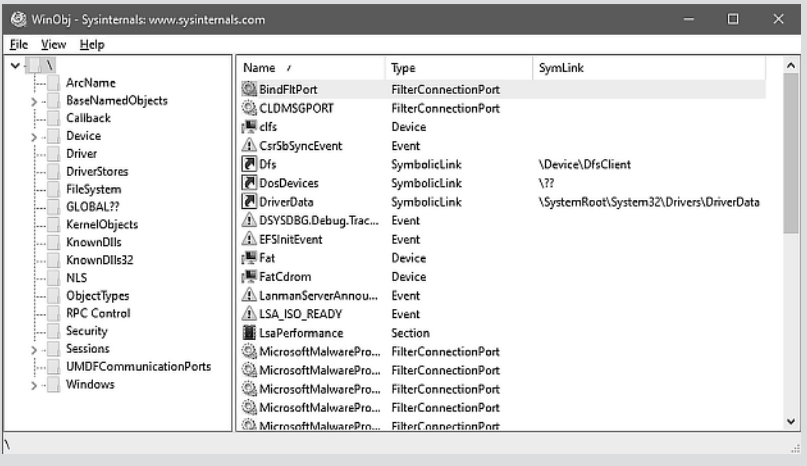
Как уже говорилось в главе 2 «Архитектура системы» тома 1, в Windows реализована объектная модель, обеспечивающая непрерывный и безопасный доступ к различным внутренним службам, реализованным в исполнительной системе. В этом разделе рассматривается *диспетчер объектов Windows*, компонент исполнительной системы, отвечающий за создание, удаление, защиту и отслеживание объектов. Диспетчер объектов централизует операции управления, которые в противном случае были бы разбросаны по всей операционной системе.

ЭКСПЕРИМЕНТ. Обзор диспетчера объектов

В данном разделе будут представлены эксперименты, показывающие способы изучения базы данных диспетчера объектов. В этих экспериментах используются следующие инструменты, с которыми вы познакомитесь, если они вам еще не известны.

- WinObj (доступно на сайте Sysinternals) показывает пространство имен диспетчера внутренних объектов и информацию об объектах (например, счетчик ссылок, количество открытых дескрипторов, дескрипторов безопасности и т. д. Инструмент WinObjEx64, доступный на GitHub, является более мощным аналогом с открытым исходным кодом, но он имеет подписи Microsoft.
- Process Explorer и Handle от Sysinternals, а также Монитор ресурсов (Resource Monitor) (представленный в главе 1) выводят открытые дескрипторы процесса. Кроме того, инструмент Process Hacker позволяет просматривать открытые дескрипторы и может показывать дополнительные детали для некоторых типов объектов.
- Команда отладчика ядра `!handle` отладчика ядра показывает открытые для процесса дескрипторы, равно как и структура данных `Io.Handles` в рамках Process, в частности `@$curprocess`.

WinObj и WinObjEx64 дает возможность обозревать пространство имен, поддерживаемое диспетчером объектов. (Чуть позже мы объясним, что имена имеются не у всех объектов.) Запустите WinObj и изучите показанную ниже схему.



Windows-команда Openfiles/query требует установки глобального флага Windows, который называется флагом обслуживания списка объектов — maintain objects list. (Для получения более подробной информации о глобальных флагах см. раздел «Глобальные флаги» главы 10.) Если набрать команду Openfiles/Local, можно получить информацию о том, установлен этот флаг или нет. Этот флаг можно установить с помощью команды Openfiles/Local ON. В любом случае, чтобы установка возымела эффект, нужно перезапустить систему. Process Explorer, Handle и Монитор ресурсов (Resource Monitor) не требуют включения отслеживания объектов, поскольку они запрашивают все системные дескрипторы и создает список объектов, принадлежащих каждому объекту.

Диспетчер объектов был разработан для достижения таких целей:

- обеспечение общего, унифицированного механизма для использования системных ресурсов;
- изоляция защиты объектов в одном месте операционной системы для обеспечения унификации и последовательности политики доступа к объектам;
- обеспечение механизма зарядки процессов на использование ими объектов, чтобы можно было установить лимиты на использование системных ресурсов;
- учреждение схемы названия объектов, в которую могут быть легко включены существующие объекты, например устройства, файлы и каталоги файловой системы или другие независимые коллекции объектов;
- поддержка требований различных окружений операционной системы, таких как возможность наследования процессом ресурсов от родительского процесса

(необходимых Windows и подсистеме Subsystem for UNIX Applications) и возможность создания имен файлов, чувствительных к регистру букв (необходимых для подсистемы Subsystem for UNIX Applications);

- установление унифицированных правил для сохранения объектов (object retention) (то есть для сохранения объектов доступными до тех пор, пока все процессы не завершат их использование);
- обеспечение возможности изоляции объектов для конкретного сеанса, чтобы учесть в пространстве имен как локальные, так и глобальные объекты;
- обеспечение возможности перенаправлять имена и местоположения объектов посредством символических ссылок и разрешить владельцам объектов, таким как файловая система, реализовывать свои механизмы перенаправления, в частности перекрестные точки в NTFS. Все вместе эти механизмы будут обеспечивать то, что называется *повторным анализом*.

У Windows есть три основных типа объектов: *объекты исполнительной системы*, *объекты ядра* и *объекты GDI/User*. Объекты исполнительной системы реализуются различными компонентами — диспетчером процессов, диспетчером памяти, подсистемой ввода/вывода и т. д. Объекты ядра представлены более простым набором, реализованным ядром Windows. Эти объекты не видимы коду пользовательского режима и создаются и используются только внутри исполнительной системы. Объекты ядра обеспечивают такие основные возможности, как синхронизация, на которых построены объекты исполнительной системы. Так, многие объекты исполнительной системы содержат (инкапсулируют) один или несколько объектов ядра, как видно на рис. 8.30.

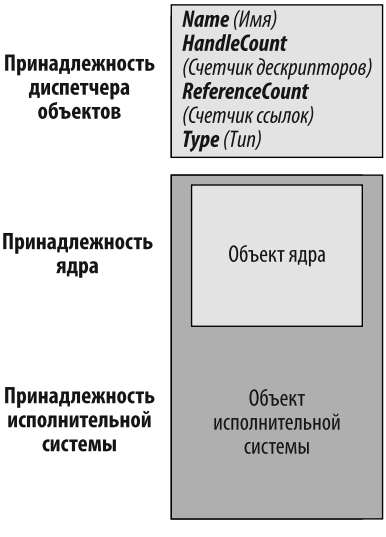


Рис. 8.30. Объекты исполнительной системы, содержащие объекты ядра

ПРИМЕЧАНИЕ Подавляющее большинство объектов GDI/User, в свою очередь, принадлежат подсистеме Windows (win32k.sys) и не взаимодействуют с ядром. По этой причине в сферу рассмотрения данной книги они не входят, но вы можете найти больше информации в Windows SDK. Пару исключений составляют объекты Desktop и Windows Station User, которые имеют оболочку управляющих объектов, равно как и большинство объектов DirectX (шейдеры, поверхности, композиции), у которых есть такая же оболочка.

Более подробно о структуре объектов ядра и о том, как они используются для реализации синхронизации, будет рассказано позже в этой главе. Оставшийся материал данного раздела будет сфокусирован на том, как работает диспетчер объектов, и на структуре объектов исполнительной системы, дескрипторах и таблицах дескрипторов. То, как объекты участвуют в реализации проверки безопасности доступа в Windows, мы упомянем кратко, подробно эта тема рассматривается в главе 7 тома 1.

Объекты исполнительной системы

Каждая подсистема среды окружения Windows создает для своих приложений свой1 образ операционной системы. Объекты исполнительной системы и службы объектов являются элементами, используемыми подсистемами среды окружения для создания их собственных версий объектов и других ресурсов.

Как правило, объекты исполнительной системы создаются либо подсистемой среды окружения от имени пользовательского приложения, либо различными компонентами операционной системы в рамках их обычного функционирования. Например, чтобы создать файл, приложение Windows вызывает функцию `CreateFileW`, реализованную в библиотеке подсистемы Windows Kernelbase.dll. После проверки и инициализации `CreateFileW`, в свою очередь, вызывает платформенно-зависимую системную функцию Windows `NtCreateFile`, чтобы та создала объект файла исполнительной системы.

Набор объектов подсистемы среды окружения, предоставляемый ее приложениям, может быть больше или меньше набора, предоставляемого исполнительной системой. Подсистема Windows использует объекты исполнительной системы для экспорта своего собственного набора объектов, многие из которых напрямую связаны с объектами исполнительной системы. Например, мьютексы и семафоры Windows непосредственно основаны на объектах исполнительной системы (которые, в свою очередь, основаны на соответствующих объектах ядра). Кроме того, подсистема Windows предоставляет именованные каналы и почтовые слоты, ресурсы, основанные на объектах файлов исполнительной системы. Поддерживая подсистему Windows для Linux (WSL), ее драйвер (Lxcore.sys) использует объекты исполнительной системы и службы в качестве основы для представления процессов, каналов и других ресурсов в стиле UNIX для своих приложений.

В табл. 8.15 перечислены первичные объекты, предоставляемые исполнительной системой, и дано краткое описание того, что они представляют. Более подробную информацию об объектах исполнительной системы можно найти в главах, описывающих связанные с ними компоненты этой системы, а для объектов исполнительной системы, непосредственно экспортируемых в Windows, — в справочной документации по Windows API). Полный список типов объектов можно получить после запуска Winobj с привилегированными правами и перехода в каталог ObjectTypes.

ПРИМЕЧАНИЕ В исполнительной системе реализовано около 69 типов объектов (в зависимости от версии Windows). Многие из этих объектов предназначены только для использования теми компонентами исполнительной системы, которые их определили, и получить непосредственный доступ к ним из функций Windows API невозможно. Примерами таких объектов являются Driver, CallBack и Adapter.

Таблица 8.15. Объекты исполнительной системы, видимые функциям Windows API

Тип объекта	Что он представляет
Process (Процесс)	Виртуальное адресное пространство и управляющую информацию, необходимую для выполнения набора объектов типа «поток»
Thread (Поток)	Исполняемая категория внутри процесса
Job (Задание)	Коллекция процессов, управляемых как единое целое, в рамках задания

Тип объекта	Назначение
Section (Раздел)	Область общей памяти (известная в Windows как проекция файла)
File (Файл)	Экземпляр открытого файла либо устройство ввода-вывода, например именованный канал или сокет
Token (Токен)	Профиль безопасности (ID безопасности, права пользователя и т. д.) процесса или потока
Event, KeyedEvent (Событие)	Объект с постоянным состоянием (о котором поступил или не поступил сигнал), который может использоваться для синхронизации или уведомления
Semaphore (Семафор)	Счетчик, ограничивающий доступ к ресурсу путем разрешения доступа к этому ресурсу, защищенному семафором, вполне определенному максимальному количеству потоков
Mutex (Мьютекс)	Механизм синхронизации, используемый для последовательного доступа к ресурсу
Timer, IRTimer (Таймер)	Механизм для оповещения потока о том, что прошел конкретный период времени. Второй вариант объектов, называемый также Idle Resilient Timers, задействуется UWP-приложениями и некоторыми службами, чтобы создавать таймеры, которые не зависят от состояния Connected Standby
IoCompletion, IoCompletion-Reserve	Метод для потоков по постановке в очередь и извлечении из нее уведомлений о завершении операций ввода-вывода (известен в Windows API как порт завершения ввода/вывода). Второй вариант предполагает также предварительное размещение порта в целях борьбы с ситуациями нехватки памяти
Key (Раздел реестра)	Механизм ссылки на данные в реестре. Хотя разделы появляются в пространстве имен диспетчера объектов, они управляются диспетчером конфигурации точно так же, как файловые объекты управляются драйверами файловой системы
Directory (Каталог)	Виртуальный каталог в пространстве имен диспетчера объектов, отвечающий за содержание в других объектов или каталогов объектов
SymbolicLink (Символическая ссылка)	Виртуальная ссылка — перенаправление наименования между объектом в пространстве имен и другим объектом. Например, C:, который является символической ссылкой на \Device\HarddiskVolumeN
TpWorkerFactory	Коллекция потоков, назначенных для выполнения конкретного набора задач. Ядро может управлять количеством рабочих элементов, которые будут выполняться по очереди, тем, сколько именно потоков будут отвечать за работу, а также динамическим созданием и завершением рабочих потоков, исходя из конкретных ограничений, устанавливаемых вызывающей программой
TmRm (Диспетчер ресурсов), TmTx (Транзакция), TmTm (Диспетчер транзакций), TmEn (Включение в список)	Объекты, используемые диспетчером транзакций ядра (Kernel Transaction Manager, KTM) для различных транзакций и/или включений в списки в качестве части диспетчера ресурсов или диспетчера транзакций. Объекты могут создаваться с помощью API-функций CreateTransactionManager, CreateResourceManager, CreateTransaction и CreateEnlistment

Продолжение ⇩

Таблица 8.15 *(продолжение)*

Тип объекта	Назначение
RegistryTransaction (Транзакция реестра)	Объекты, используемые низкоуровневыми ресурсоемкими функциями транзакций в реестре, которые не задействуют полностью возможностей KTM, но все еще позволяют транзакциям получать простой доступ к разделам реестра
WindowStation (Станция окна)	Объект, содержащий буфер обмена, таблицу atom и группу объектов типа Рабочий стол
Desktop (Рабочий стол)	Объект, содержащийся внутри объекта станции окна. Рабочий стол имеет логическую поверхность дисплея и содержит окна, меню и связи
PowerRequest	Объект, связанный с выполняемым потоком, который помимо прочего является вызовом функции SetThreadExecutionState для запроса заданного изменения режима электропитания, такого как блокировка перехода в спящий режим (например, при воспроизведении какого-нибудь фильма)
EtwConsumer	Представляет собой подключенный ETW-потребитель режима реального времени, зарегистрированный с помощью API-функции StartTrace (и способный вызвать функцию ProcessTrace для получения событий в очереди объектов)
CoverageSampler	Создан ETW в ходе разрешения отслеживания покрытия кода в рамках указанного сеанса ETW
EtwRegistration	Представляет собой объект регистрации, связанный с ETW-провайдером пользовательского режима (или режима ядра), который зарегистрирован с помощью API-функции EventRegister
ActivationObject	Представляет объект, отслеживающий состояние переднего плана для дескрипторов окон, которые управляются диспетчером необработанных входных данных (Raw Input Manager) в Win32k.sys
ActivityReference	Отслеживает процессы, находящиеся под контролем диспетчера жизненного цикла процессов (Process Lifetime Manager, PLM), которые должны оставаться бодрствующими во время сценариев Connected Standby
ALPC Port	В основном используется библиотекой удаленного вызова процедур (RPC) для обеспечения функциональности Local RPC (LRPC) с помощью протокола передачи ncalrpc. Доступна внутренним службам как базовый механизм IPC между процессами и/или ядром
Composition, DxgkCompositionObject, DxgkCurrentDxgProcessObject, DxgkDisplayManagerObject, DxgkSharedBundleObject, DxgkSharedKeyedMutexObject, DxgkShartedProtectedSessionObject, DgxkSharedResource, DxgkSwapChainObject, DxgkSharedSyncObject	Применяется функциями DirectX 12 в пространстве пользователя как часть продвинутых возможностей шейдеров и GDGPU. Эти объекты управления являются оболочкой для нижележащих дескрипторов DirectX

Тип объекта	Назначение
CoreMessaging	Представляет объект CoreMessaging IPC, который служит оболочкой порта ALPC с собственным пространством имен и возможностями. В основном используется современным диспетчером ввода, но также виден любому компоненту MinUser в системах WCOS
EnergyTracker	Доступен службе User Mode Power (UMPO), позволяет ей отслеживать и подсчитывать расход энергии в рамках определенного набора оборудования и ассоциировать его в разрезе приложений
FilterCommunicationPort, FilterConnectionPort	Подлежащие (являющиеся основой) объекты, поддерживающие интерфейс на основе IRP, предоставляемый API менеджера фильтров, обеспечивающий передачу информации между службами пользовательского режима и приложениями, и мини-фильтры, управляемые менеджером фильтров, как при использовании функции FilterSendMessage
Partition (Раздел)	Позволяет диспетчерам памяти и кэша и управляющим программам рассматривать область физической памяти как уникальную с точки зрения управления относительно остальной памяти системы, давая ей собственный экземпляр потоков управления, функциональные возможности, обмен страницами, кэширование и т. д. Используется среди прочих Hyper-V и игровым режимом
Profile (Профиль)	Используется интерфейсом профилирования, позволяющим захватывать контролируемые по времени области памяти, которые отслеживают все, от указателя инструкций (IP) до самой низкоуровневой информации о кэшировании в процессоре, хранимой в счетчиках PMU
RawInputManager	Представляет собой объект, привязанный к устройству HID, такому как мышь, клавиатура или планшет, и позволяет считывать данные ввода, которые он получает, и управлять ими в диспетчере окон. Задействуется современным кодом управления пользовательским интерфейсом, например при использовании службы Core Messaging
Session (Сеанс)	Объект, который представляет собой интерактивный пользовательский сеанс с точки зрения диспетчера памяти. Отслеживает уведомления диспетчера ввода-вывода по поводу подключения/отключения/входа в систему/выхода из системы для нужд сторонних драйверов
Terminal (Терминал)	Доступен, только если активен Terminal Thermal Manager (TTM), который представляет собой пользовательский терминал на устройстве, управляемом диспетчером питания пользовательского режима (UMPO)
TerminalEventQueue	Доступен только в системах TTM наподобие предыдущего объекта, представляет собой события, поступающие на терминал устройства, о котором UMPO общается с диспетчером питания ядра
UserApcReserve	Подобно IoCompletionReserve, который позволяет заранее создавать структуру данных, чтобы заново использовать ее при нехватке памяти, этот объект инкапсулирует объект APC ядра (KAPC) как объект управления
WaitCompletionPacket	Применяемый в рамках новых возможностей асинхронного ожидания, добавленных в API пула потоков пользовательского режима, объект является оболочкой над завершением диспетчером ожидания того, когда пакет ввода/вывода будет доставлен в порт завершения ввода/вывода
WmiGuid	Задействуется API инструментария управления Windows (WMI) при открытии блоков данных WMI по GUID как из пользовательского режима, так и из режима ядра, как в случае IoWMIOpenBlock

ПРИМЕЧАНИЕ Поскольку Windows NT изначально проектировалась совместимой с операционной системой OS/2, мьютекс должен был быть совместим с уже существующей конструкцией объектов взаимного исключения, то есть иметь конструкцию, от которой требовалось, чтобы поток мог отказаться от объекта, оставив его недоступным. Поскольку подобное поведение для такого объекта считалось необычным, был создан еще один объект ядра — мутант. Со временем от поддержки OS/2 отказались, и объект стал использоваться подсистемой Windows 32 под названием мьютекс (но при этом он сохранил внутреннее имя «мутант»).

Структура объектов

Как показано на рис. 8.31, у каждого объекта есть заголовок, тело и опционально — футер. Диспетчер объектов управляет заголовками и футерами, а телами объектов управляют компоненты исполнительной системы, являющиеся владельцами тех типов объектов, которые ими созданы. В каждом заголовке объекта есть индекс другого специального объекта — объекта типа, в котором содержится информация, общая для всех экземпляров объекта. Вдобавок к этому существует еще до восьми подзаголовков: информационные заголовки названия, квоты, процесса, дескриптора, аудита, заполнения (padding), расширенный и создателя. Наличие расширенного информационного заголовка говорит о том, что у объекта есть футер, на который в нем будет указатель.

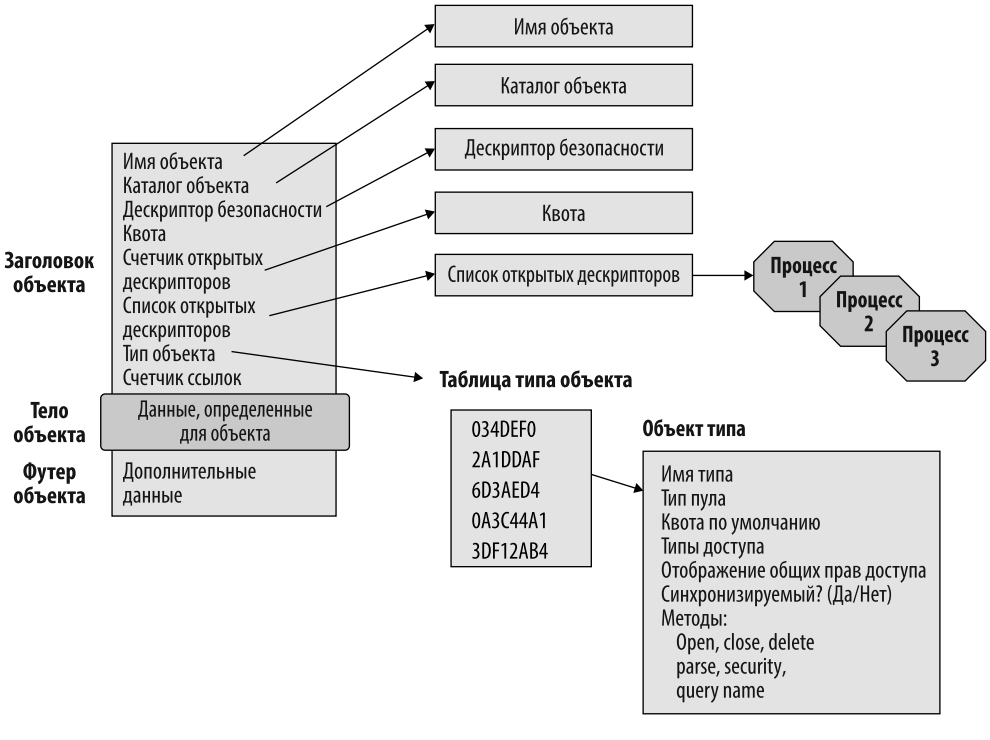


Рис. 8.31. Структура объекта

Заголовки и тела объектов

Диспетчер объектов использует данные из заголовков, чтобы управлять объектами, независимо от их типа. В табл. 8.16 кратко описаны поля заголовка объекта, а в табл. 8.17 — поля из дополнительных подзаголовков.

Таблица 8.16. Поля заголовка объекта

Поле	Назначение
Счетчик дескрипторов	Обеспечивает подсчет количества открытых в данный момент дескрипторов к объекту
Счетчик указателей	Обеспечивает подсчет количества ссылок на объект (в том числе по ссылке для каждого дескриптора) и количества ссылок использования каждого дескриптора (до 32 в 32-разрядных системах и до 32 768 — в 64-разрядных). Компоненты режима ядра могут ссылаться на объект по указателю без использования дескрипторов
Дескриптор безопасности	Определяет, кто может использовать данный объект и что можно с ним делать. Надо заметить, что неименованные объекты не могут обладать безопасностью по определению
Индекс типа объекта	Хранит индекс типа объекта, который содержит атрибуты, общие для всех объектов данного типа. Таблица, в которой хранятся все объекты типа, называется ObTypeIndexTable. Из соображений безопасности данный индекс делится по XOR на динамически сгенерированное сторожевое значение, которое хранится в ObHeaderCookie, и последние восемь бит адреса заголовка самого объекта
Инфомаска (InfoMask)	Битовая маска, которая описывает, какие из дополнительных структур подзаголовков, перечисленных в табл. 8.17, присутствуют, за исключением подзаголовка с информацией о создателе, который, если имеется, всегда предшествует объекту. С помощью таблицы ObpInfoMaskToOffset эта битовая маска конвертируется в отрицательное смещение, где каждый подзаголовок связан с однобайтовым индексом, который определяет его позицию относительно остальных подзаголовков
Флаги	Характеристики и атрибуты объекта (см. табл. 8.20, где приведен список всех флагов объекта)
Блокировка	Персональная блокировка объекта, используемая при изменении полей, принадлежащих заголовку этого объекта или любому из его подзаголовков
Флаги отслеживания	Дополнительные флаги, специально предназначенные для инструментов отслеживания и отладки, также описаны в табл. 8.20
Информация о создании	Кратковременные сведения о создании объекта, которые хранятся до тех пор, пока объект не будет полностью размещен в пространстве имен. После создания это поле конвертируется в указатель на квоту

Кроме заголовка объекта, хранящего информацию, относящуюся к любым объектам, в подзаголовках содержится дополнительная информация, относящаяся к специфическим деталям объекта. Учтите, что эти структуры находятся по переменному смещению от заголовка объекта, значение которого зависит от количества подзаголовков, связанных с основным заголовком объекта. (За исключением, как

ранее указывалось, информации о создателе.) Для каждого подзаголовка с целью отображения его существования обновляется поле `InfoMask`. Когда диспетчер объектов проводит проверку заданного подзаголовка, он проверяет, выставлен ли соответствующий бит в `InfoMask`, а затем использует оставшиеся биты для выбора правильного смещения в таблице `ObpInfoMaskToOffset`, где он находит смещение подзаголовка с начала заголовка объекта.

Эти смещения существуют для всех возможных комбинаций имеющихся подзаголовков, но, поскольку подзаголовки, если они есть, всегда размещаются в фиксированном, постоянном порядке, для заданного заголовка количество возможных мест будет равно максимальному количеству предшествующих ему подзаголовков. Например, поскольку первым всегда следует подзаголовок с информацией об имени, у него есть только одно возможное смещение. С другой стороны, подзаголовок с информацией о дескрипторах (который размещается третьим) имеет три возможных места, поскольку он может размещаться после подзаголовка квоты, а может и не размещаться после этого подзаголовка, который, в свою очередь, мог бы располагаться после информации об имени. В табл. 8.17 приведены описания всех дополнительных подзаголовков объекта и их мест. Что же касается информации о создателе, наличине этого подзагловка определяется значением флагов заголовка объекта (см. табл. 8.20).

Таблица 8.17. Необязательные подзаголовки объекта

Имя	Назначение	Бит	Смещение
Информация о создателе	Связывает объект со списком всех объектов того же типа и записывает процесс, который создал объект, вместе с обратным отслеживанием	0 (0x1)	ObpInfoMaskToOffset[0])
Информация об имени	Содержит имя объекта, ответственного за обеспечение видимости объекта другим процессам для его совместного использования, и указатель на каталог объекта, предоставляющий иерархическую структуру, в которой хранятся имена объектов	1 (0x2)	ObpInfoMaskToOffset [InfoMask & 0x3]
Информация о дескрипторах	Содержит базу данных, состоящую из записей (или из одной записи) для процесса, у которого имеется открытый дескриптор объекта (наряду со счетчиком дескрипторов для каждого процесса)	2 (0x4)	ObpInfoMaskToOffset [InfoMask & 0x7]
Информация о квоте	Регистрирует расход ресурсов, забираемых у процесса, когда тот открывает дескриптор объекта	3 (0x8)	ObpInfoMaskToOffset [InfoMask & 0xF]
Информация о процессе	Содержит указатель на владельца процесса, если объект относится к эксклюзивному. Далее в этой главе будет дана дополнительная информация об эксклюзивных объектах	4 (0x10)	ObpInfoMaskToOffset [InfoMask & 0x1F]

Имя	Назначение	Бит	Смещение
Информация об аудите	Содержит указатель на изначальный дескриптор безопасности, использованный при первом создании данного объекта. Применяется для объектов-файлов, когда аудит активен, чтобы гарантировать целостность	5 (0x20)	ObpInfoMaskToOffset [InfoMask & 0x3F]
Расширенная информация	Хранит указатель на футер объекта для объектов, которым это требуется, в частности для файловых объектов и объектов контекста хранилища	6 (0x40)	ObpInfoMaskToOffset [InfoMask & 0x7F]
Информация об отступах	Хранит пустое место, но используется, если требуется, для выравнивания тела объекта в границах кэша	7 (0x80)	ObpInfoMaskToOffset [InfoMask & 0xFF]

Каждый из этих подзаголовков является дополнительным и присутствует лишь при определенных обстоятельствах либо во время загрузки системы, либо при создании самого объекта. В табл. 8.18 дается описание этих обстоятельств.

Таблица 8.18. Условия, требуемые для наличия подзаголовков объектов

Имя	Условие
Информация о создателе	У объекта должен быть установлен флаг поддержки списка типа (type list). У объектов драйверов (Driver) этот флаг установлен, если включена проверка Driver Verifier. Но установка глобального флага поддержки списка типа (рассмотренного ранее) создаст это условие для всех объектов, а для Type-объектов этот флаг установлен всегда
Информация об имени	Объект должен быть создан с именем
Информация о дескрипторах	У типа объекта должен быть установлен флаг поддержки счетчика дескрипторов (handle count). Для файловых объектов, ALPC-объектов, объектов WindowStation и Desktop-объектов этот флаг установлен в их структуре типа объекта
Информация о квотах	Объект не должен быть создан исходным системным процессом (или процессом простоя)
Информация о процессе	Объект должен быть создан с эксклюзивным (exclusive) флагом объекта (описание флагов объектов см. в табл. 8.20)
Информация об аудите	Объект должен быть файлом. Должен быть активен аудит событий с файловыми объектами
Расширенная информация	У объекта должен быть футер для размещения либо информации об изъятии дескриптора (применяется файловыми объектами и объектами разделов реестра), либо дополнительного контекста пользователя (для объектов контекста хранилища)
Информация об отступах	У типа объекта должен быть установлен флаг выравнивания в кэше (Cache aligned). Это характерно для объектов процессов и потоков

Как можно заметить, если имеется подзаголовок с расширенной информацией, в хвосте тела объекта размещается футер. В отличие от подзаголовков футер — это структура статического размера, память выделяется сразу для всех ее типов. Этих типов два, и они описаны в табл. 8.19.

Таблица 8.19. Условия наличия футера объекта

Имя	Условие
Информация об изъятии дескриптора	Объект должен быть создан функцией ObCreateObjectEx с флагом AllowHandleRevocation в структуре OB_EXTENDED_CREATION_INFO. Так создаются файловые объекты и объекты разделов реестра
Расширенная информация о пользователе	Объект должен быть создан функцией ObCreateObjectEx с флагом AllowExtendedUserInfo в структуре OB_EXTENDED_CREATION_INFO. Так создаются объекты типа Silo context

И наконец, ряд атрибутов и (или) флагов определяют поведение объекта во время создания или во время определенных операций. Диспетчер объектов получает эти флаги при каждом создании нового объекта в структуре под названием *атрибуты объекта*. Эта структура определяет имя объекта, корневой каталог объекта, в который он должен быть вставлен, дескриптор безопасности объекта и *флаги атрибутов объекта*. В табл. 8.20 перечислены различные флаги, которые могут быть связаны с объектом.

ПРИМЕЧАНИЕ Когда объект создается с помощью API-функций в подсистеме Windows (таких как CreateEvent или CreateFile), вызывающая программа не определяет каких-либо атрибутов объекта — DLL-библиотека подсистемы делает все закулисно. Поэтому все именованные объекты, созданные с помощью Win32, помещаются в каталог BaseNamedObjects, либо в его глобальный экземпляр, либо в экземпляр, созданный для конкретного сеанса, поскольку он является корневым каталогом объектов, определяемых Kernelbase.dll в качестве части структуры атрибутов объекта. Дополнительную информацию о BaseNamedObjects и о его связи с пространством имен, создаваемым для каждого сеанса, можно будет найти далее в этой главе.

Таблица 8.20. Флаги объектов

Флаг атрибута	Бит флага заголовка	Назначение
OBJ_INHERIT	Хранится в записи таблицы дескрипторов	Определяет, будет ли дескриптор унаследован дочерними процессами и сможет ли процесс использовать функцию DuplicateHandle для создания копии
OBJ_PERMANENT	PermanentObject	Определяет способность объекта сохранять поведение, зависящее от рассматриваемых далее подсчетов ссылок
OBJ_EXCLUSIVE	ExclusiveObject	Указывает на то, что объект может использоваться только тем процессом, который его создал

Флаг атрибута	Бит флага заголовка	Назначение
OBJ_CASE_INSENSITIVE	Не сохраняется, используется во время выполнения	Указывает на то, что поиск этого объекта в пространстве имен должен вестись без учета регистра символов. Может быть отменен установкой флага *case insensitive* в типе объекта
OBJ_OPENIF	Не сохраняется, используется во время выполнения	Указывает на то, что операция создания объекта с таким же именем должна привести к открытию уже существующего объекта, а не к выдаче ошибки
OBJ_OPENLINK	Не сохраняется, используется во время выполнения	Указывает на то, что диспетчер объектов должен открыть дескриптор символической ссылки, а не целевого объекта
OBJ_KERNEL_HANDLE	KernelObject	Указывает на то, что дескриптор объекта должен быть *дескриптором ядра* (подробности будут даны чуть позже)
OBJ_FORCE_ACCESS_CHECK	Не сохраняется, используется во время выполнения	Указывает на то, что даже при открытии объекта из режима ядра требуется полная проверка прав доступа
OBJ_KERNEL_EXCLUSIVE	KernelOnlyAccess	Не дает процессу пользовательского режима открыть дескриптор объекта; используется для защиты раздела объектов \Device\PhysicalMemory и \Win32kSessionGlobals
OBJ_IGNORE_IMPERSONATED_DEVICEMAP	Не сохраняется, используется во время выполнения	Отмечает, что, если токен не настоящий, карта устройств DOS исходного пользователя не должна использоваться, а вместо нее для поиска объекта нужно использовать карту устройств процессора. Это сделано для защиты от некоторых атак, основанных на файловых перенаправлениях
OBJ_DONT_REPARSE	Не сохраняется, используется во время выполнения	Запрещает любые ситуации, включающие повторный анализ (символические ссылки, точки повторного анализа NTFS, перенаправление разделов реестра), и возвращает STATUS_REPARSE_POINT_ENCOUNTERED в случае их возникновения. Это сделано для защиты от некоторых типов атак через перенаправление
Нет	DefaultSecurityQuota	Указывает, что дескриптор безопасности объекта использует исходную квоту 2 Кбайт
Нет	SingleHandleEntry	Указывает на то, что подзаголовок информации о дескрипторах содержит только одну запись и не является базой данных

Продолжение ⇩

Таблица 8.20 *(продолжение)*

Флаг атрибута	Бит флага заголовка	Назначение
Нет	NewObject	Указывает на то, что объект был создан, но еще не вставлен в пространство имен объектов
Нет	DeletedInline	Указывает на то, что объект удаляется через отложенное удаление рабочего потока

Кроме заголовка, у каждого объекта есть тело, чей формат и содержимое уникальны для типа этого объекта; все объекты одного и того же типа используют один и тот же формат тела. Создавая тип объекта и предоставляя для него соответствующие службы, компонент исполнительной системы может контролировать работу с данными в телах всех объектов такого типа. Поскольку заголовок объекта имеет статический, заранее известный размер, диспетчер объектов может легко найти заголовок объекта путем простого вычитания размера заголовка из указателя на объект. Как уже ранее говорилось, для получения доступа к подзаголовку диспетчер объектов вычитает еще одно известное значение из указателя на заголовок объекта. Для получения указателя на футер используются данные из заголовка расширенной информации.

Благодаря нормированным структурам заголовка объекта и подзаголовков диспетчер объектов может предоставить небольшой набор служб, способный работать с атрибутами, сохраненными в любом заголовке объекта, и применимый к объектам любого типа (хотя некоторые общие службы не имеют для определенных объектов никакого смысла). Эти общие службы, часть из которых подсистема Windows делает доступными Windows-приложениям, перечислены в табл. 8.21.

Таблица 8.21. Общие службы объекта

Служба	Назначение
Close (Закрытия)	Закрывает дескриптор объекта, если позволено
Duplicate (Дублирования)	Обеспечивает совместное использование объекта путем создания дубликата его дескриптора и предоставления его другому процессу
Inheritance (Наследования)	Если дескриптор отмечен как наследуемый и был создан дочерний процесс, которому разрешено наследовать дескрипторы, он действует так же, как при дублировании дескриптора
Make permanent/temporary (Превращения в постоянный или временный)	Изменяет сохранность объекта (рассматриваемую далее)
Query object (Запроса объекта)	Получает информацию о стандартных атрибутах объекта и других деталях, управляемых на уровне диспетчера объектов
Query security (Запроса безопасности)	Получает дескриптор безопасности объекта

Служба	Назначение
Set security (Установки безопасности)	Изменяет степень защиты объекта
Wait for a single object (Ожидания одиночного объекта)	Синхронизирует выполнение потока с одним объектом, который позже может синхронизировать выполнение потока или быть связанным с портом завершения ввода-вывода посредством пакета завершения ожидания
Signal an object and wait for another (Сообщения об объекте и ожидания другого объекта)	Сообщает об объекте (таком, как событие) и синхронизирует выполнение потока с другим объектом. Операция пробуждения ожидания производится атомарно с точки зрения планировщика
Wait for multiple objects (Ожидания нескольких объектов)	Синхронизирует выполнение потока с несколькими объектами (максимум до 64), которые после могут синхронизировать исполнение потока или быть связанными с портом завершения ввода-вывода с помощью пакета завершения ожидания

Хотя эти общие службы объектов поддерживаются для объектов всех типов, у каждого объекта есть свои службы создания (create), открытия (open) и запроса (query). Например, система ввода-вывода реализует службу создания файла для своих файловых объектов, а диспетчер процессов реализует службу создания процесса для своих объектов процесса.

Однако некоторые объекты могут не предоставлять таких служб напрямую, а создаваться в рамках работы внутренних процедур как результат какой-то пользовательской операции. Например, при открытии блока данных WMI из пользовательского режима создается объект WmiGuid, но приложению не предоставляется никакого дескриптора со службами запроса или закрытия. Ключевой момент здесь в том, что не существует никакой обобщенной процедуры создания объекта.

Подобная процедура была бы слишком сложной, поскольку набор параметров, необходимых для инициализации, к примеру, файлового объекта, существенно бы отличался от того набора, который понадобился бы для инициализации объекта типа «процесс». Кроме того, диспетчер объектов нес бы дополнительную нагрузку при каждом вызове потоком службы объекта на определение типа объекта, на который ссылается дескриптор, и на вызов соответствующей версии службы.

Объекты типа

Заголовки объектов содержат данные, общие для всех объектов, которые могут принимать различные значения для каждого экземпляра объекта. Например, у каждого объекта есть уникальное имя и может быть уникальный дескриптор безопасности. Но объекты также содержат некоторые данные, остающиеся постоянными для всех объектов конкретного типа. Исполнительная система предоставляет для объектов типа «поток», кроме всех прочих, доступ к завершению и приостановке, а для объектов типа «файл» — доступ для чтения, записи, добавления и удаления. Еще одним примером атрибута, присущего определенному типу объектов, является синхронизация, которая вскоре будет рассмотрена.

В целях экономии памяти диспетчер объектов сохраняет эти статические атрибуты, присущие объектам определенного типа, единожды при создании нового объекта типа. Для записи этих данных он использует свой собственный объект, так называемый объект типа. Как показано на рис. 8.32, если установлен флаг отладки для отслеживания объектов (описывается в разделе «Глобальные флаги» главы 10), объект типа также связывает вместе все объекты одного и того же типа (в данном случае типа «процесс»), позволяя диспетчеру объектов находить эти объекты и вести их подсчет, если это необходимо. Эта функция использует возможность ранее рассмотренного подзаголовка с информацией о создателе.

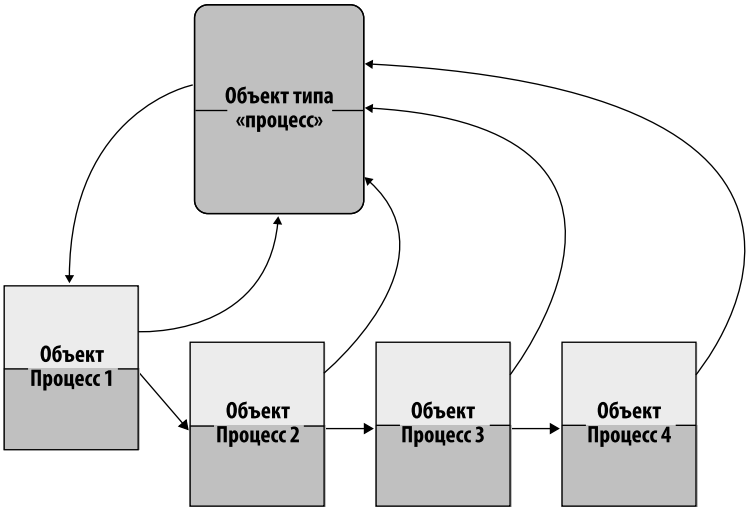


Рис. 8.32. Объекты процессов и объект типа «процесс»

ЭКСПЕРИМЕНТ. Просмотр заголовков объектов и объектов типа

Вы можете просмотреть структуру данных объекта типа «процесс» в отладчике ядра, предварительно идентифицировав этот объект с помощью команды модели данных отладчика `dx@$cursession.Processes` какого-нибудь объекта процесса:

```
lkd> dx -r0 &@$cursession.Processes[4].KernelObject
&@$cursession.Processes[4].KernelObject           : 0xffff898f0327d300
                                                      [Type: _EPROCESS *]
```

Затем выполните команду `!object` с адресом объекта процесса в качестве аргумента:

```
lkd> !object 0xffff898f0327d300
Object: ffff898f0327d300 Type: (ffff898f032954e0) Process
    ObjectHeader: ffff898f0327d2d0 (new version)
    HandleCount: 6 PointerCount: 215645
```

Учтите, что в 32-разрядной версии Windows заголовок объекта начинается с 0x18 (32 в десятичном формате) байт, предшествующих телу объекта, а в 64-разрядной версии — с 0x30 (48 в десятичном формате) байт, предшествующих телу, то есть с размера самого заголовка объекта. Просмотреть заголовок объекта можно с помощью следующей команды:

```
lkd> dx (nt!_OBJECT_HEADER*)0xffff898f0327d2d0
(nt!_OBJECT_HEADER*)0xffff898f0327d2d0         : 0xffff898f0327d2d0 [Type: _
OBJECT_HEADER *]
     [+0x000] PointerCount     : 214943 [Type: __int64]
     [+0x008] HandleCount      : 6 [Type: __int64]
     [+0x008] NextToFree       : 0x6 [Type: void *]
     [+0x010] Lock             [Type: _EX_PUSH_LOCK]
     [+0x018] TypeIndex        : 0x93 [Type: unsigned char]
     [+0x019] TraceFlags       : 0x0 [Type: unsigned char]
     [+0x019 ( 0: 0)] DbgRefTrace      : 0x0 [Type: unsigned char]
     [+0x019 ( 1: 1)] DbgTracePermanent      : 0x0 [Type: unsigned char]
     [+0x01a] InfoMask         : 0x80 [Type: unsigned char]
     [+0x01b] Flags            : 0x2 [Type: unsigned char]
     [+0x01b ( 0: 0)] NewObject            : 0x0 [Type: unsigned char]
     [+0x01b ( 1: 1)] KernelObject         : 0x1 [Type: unsigned char]
     [+0x01b ( 2: 2)] KernelOnlyAccess     : 0x0 [Type: unsigned char]
     [+0x01b ( 3: 3)] ExclusiveObject      : 0x0 [Type: unsigned char]
     [+0x01b ( 4: 4)] PermanentObject      : 0x0 [Type: unsigned char]
     [+0x01b ( 5: 5)] DefaultSecurityQuota      : 0x0 [Type: unsigned char]
     [+0x01b ( 6: 6)] SingleHandleEntry    : 0x0 [Type: unsigned char]
     [+0x01b ( 7: 7)] DeletedInline        : 0x0 [Type: unsigned char]
     [+0x01c] Reserved          : 0xffff898f [Type: unsigned long]
     [+0x020] ObjectCreateInfo  : 0xfffff8047ee6d500
                                          [Type: _OBJECT_CREATE_INFORMATION *]
     [+0x020] QuotaBlockCharged   : 0xfffff8047ee6d500 [Type: void *]
     [+0x028] SecurityDescriptor  : 0xffffc704ade03b6a [Type: void *]
     [+0x030] Body              [Type: _QUAD]
    ObjectType        : Process
    UnderlyingObject  [Type: _EPROCESS]
```

Теперь посмотрим на структуру данных объекта типа, скопировав указатель, ранее выданный командой `!object`:

```
lkd> dx (nt!_OBJECT_TYPE*)0xffff898f032954e0
(nt!_OBJECT_TYPE*)0xffff898f032954e0            : 0xffff898f032954e0 [Type: _
OBJECT_TYPE *]
     [+0x000] TypeList         [Type: _LIST_ENTRY]
     [+0x010] Name             : "Process" [Type: _UNICODE_STRING]
     [+0x020] DefaultObject    : 0x0 [Type: void *]
     [+0x028] Index                : 0x7 [Type: unsigned char]
     [+0x02c] TotalNumberOfObjects  : 0x2e9 [Type: unsigned long]
     [+0x030] TotalNumberOfHandles  : 0x15a1 [Type: unsigned long]
     [+0x034] HighWaterNumberOfObjects  : 0x2f9 [Type: unsigned long]
     [+0x038] HighWaterNumberOfHandles  : 0x170d [Type: unsigned long]
     [+0x040] TypeInfo         [Type: _OBJECT_TYPE_INITIALIZER]
     [+0x0b8] TypeLock         [Type: _EX_PUSH_LOCK]
     [+0x0c0] Key              : 0x636f7250 [Type: unsigned long]
     [+0x0c8] CallbackList     [Type: _LIST_ENTRY]
```

Из полученного вывода видно, что структура объекта типа включает имя объекта типа и пиковое количество дескрипторов и объектов данного типа. Кроме того, в поле `CallbackList` также отслеживается фильтрация обратных вызовов диспетчера объектов, связанная с этим объектом типа. В поле TypeInfo хранится указатель на структуру данных, в которой хранятся общие для всех объектов этого типа атрибуты, а также указатели на методы объекта типа:

```
lkd> dx ((nt!_OBJECT_TYPE*)0xffff898f032954e0)->TypeInfo
((nt!_OBJECT_TYPE*)0xffff898f032954e0)->TypeInfo   [Type: _OBJECT_TYPE_INITIALIZER]
    [+0x000] Length           : 0x78 [Type: unsigned short]
    [+0x002] ObjectTypeFlags  : 0xca [Type: unsigned short]
    [+0x002 ( 0: 0)] CaseInsensitive  : 0x0 [Type: unsigned char]
    [+0x002 ( 1: 1)] UnnamedObjectsOnly  : 0x1 [Type: unsigned char]
    [+0x002 ( 2: 2)] UseDefaultObject  : 0x0 [Type: unsigned char]
    [+0x002 ( 3: 3)] SecurityRequired  : 0x1 [Type: unsigned char]
    [+0x002 ( 4: 4)] MaintainHandleCount  : 0x0 [Type: unsigned char]
    [+0x002 ( 5: 5)] MaintainTypeList  : 0x0 [Type: unsigned char]
    [+0x002 ( 6: 6)] SupportsObjectCallbacks  : 0x1 [Type: unsigned char]
    [+0x002 ( 7: 7)] CacheAligned      : 0x1 [Type: unsigned char]
    [+0x003 ( 0: 0)] UseExtendedParameters  : 0x0 [Type: unsigned char]
    [+0x003 ( 7: 1)] Reserved          : 0x0 [Type: unsigned char]
    [+0x004] ObjectTypeCode     : 0x20 [Type: unsigned long]
    [+0x008] InvalidAttributes  : 0xb0 [Type: unsigned long]
    [+0x00c] GenericMapping     [Type: _GENERIC_MAPPING]
    [+0x01c] ValidAccessMask    : 0x1fffff [Type: unsigned long]
    [+0x020] RetainAccess       : 0x101000 [Type: unsigned long]
    [+0x024] PoolType           : NonPagedPoolNx (512) [Type: _POOL_TYPE]
    [+0x028] DefaultPagedPoolCharge  : 0x1000 [Type: unsigned long]
    [+0x02c] DefaultNonPagedPoolCharge  : 0x8d8 [Type: unsigned long]
    [+0x030] DumpProcedure      : 0x0 [Type: void (__cdecl*)
                                                (void *,_OBJECT_DUMP_CONTROL *)]
    [+0x038] OpenProcedure      : 0xfffff8047f062f40 [Type: long (__cdecl*)
           (_OB_OPEN_REASON,char,_EPROCESS *,void *,unsigned long *,unsigned long)]
    [+0x040] CloseProcedure     : 0xfffff8047F087a90 [Type: void (__cdecl*)
                            (_EPROCESS *,void *,unsigned __int64,unsigned __int64)]
    [+0x048] DeleteProcedure    : 0xfffff8047f02f030 [Type: void (__cdecl*)(void *)]
    [+0x050] ParseProcedure    : 0x0 [Type: long (__cdecl*)(void *,void *,
     _ACCESS_STATE *, char,unsigned long,_UNICODE_STRING *,_UNICODE_STRING *,void *,
                                        _SECURITY_QUALITY_OF_SERVICE *,void * *)]
    [+0x050] ParseProcedureEx   : 0x0 [Type: long (__cdecl*)(void *,void *,
     _ACCESS_STATE *, char,unsigned long,_UNICODE_STRING *,_UNICODE_STRING *,void *,
         _SECURITY_QUALITY_OF_SERVICE *,_OB_EXTENDED_PARSE_PARAMETERS *,void * *)]
    [+0x058] SecurityProcedure  : 0xfffff8047eff57b0 [Type: long (__cdecl*)
          (void *,_SECURITY_OPERATION_CODE,unsigned long *,void *,unsigned long *,
                                      void * *,_POOL_TYPE,_GENERIC_MAPPING *,char)]
    [+0x060] QueryNameProcedure  : 0x0 [Type: long (__cdecl*)(void *,unsigned char,_
                    OBJECT_NAME_INFORMATION *,unsigned long,unsigned long *,char)]
    [+0x068] OkayToCloseProcedure  : 0x0 [Type: unsigned char (__cdecl*)(_EPROCESS *,
                                                          void *,void *,char)]
    [+0x070] WaitObjectFlagMask  : 0x0 [Type: unsigned long]
    [+0x074] WaitObjectFlagOffset  : 0x0 [Type: unsigned short]
    [+0x076] WaitObjectPointerOffset  : 0x0 [Type: unsigned short]
```

Из пользовательского режима работать с объектами типа нельзя, поскольку диспетчер объектов не предоставляет для них никаких служб. Тем не менее некоторые, определяемые ими атрибуты, видимы с помощью некоторых

платформенно-зависимых служб операционной системы и процедур Windows API. Информация, хранящаяся в инициализаторах типа, описана в табл. 8.22.

Таблица 8.22. Поля инициализаторов типа

Атрибут	Назначение
Type name (Имя типа)	Имя объектов данного типа («процесс», «событие», «порт APLC» и т. д.)
Pool type (Тип пула)	Показывает, может ли объектам данного типа выделяться выгружаемая или невыгружаемая память
Default quota charges (Квота по умолчанию)	Значения пулов выгружаемой и невыгружаемой памяти, составляющие по умолчанию квоту процесса
Valid access mask (Действующая маска доступа)	Виды доступа, которые поток может запросить при открытии дескриптора объекта данного типа («чтение», «запись», «завершение», «приостановка» и т. д.)
Generic access rights mapping (Отображение общих прав доступа)	Отображение четырех общих прав доступа (для чтения, записи, выполнения и всех прав) на права доступа, присущие конкретному типу
Retain access (Сохранение доступа)	Права доступа, которые никаким сторонним обратным вызовам диспетчера объектов нельзя удалять (появляются в списке обратных вызовов, описанном ранее)
Flags (Флаги)	Указывают на то, что объекты не должны иметь имен (например, в случае с объектами типа «процесс»), что в их именах учитывается регистр символов, что они требуют наличие дескриптора безопасности, что они поддерживают обратные вызовы, фильтруемые объектами, и должна ли поддерживаться база данных дескрипторов (подзаголовок информации о дескрипторах) и (или) взаимозависимость списка типов (подзаголовок информации о создателе). Флаг Use default object также определяет поведение показанного далее в этой таблице поля Default object
Object type code (Код объекта типа)	Используется для описания того, что из себя представляет тип объекта (в отличие от сравнения с известным значением имени). Для файловых объектов значение этого поля устанавливается в 1, для объектов синхронизации — в 2 и для объектов потоков — в 4. Это поле также используется ALPC для хранения информации об атрибуте дескриптора, связанной с сообщением
Invalid attributes (Недопустимые атрибуты)	Задает флаги атрибутов объекта (описаны в табл. 8.20), недопустимые для этого типа объекта
Default object (Объект по умолчанию)	Определяет внутреннее событие диспетчера объектов, которое должно использоваться при ожидании данного объекта, если этого требует создатель объекта типа. Следует учесть, что такие объекты, как File и ALPC-порт, уже содержат свой встроенный диспетчер объектов; в таком случае это поле является смещением в теле объекта. Например, событие внутри структуры FILE_OBJECT встроено в поле под названием Event

Продолжение ⇩

Таблица 8.22 *(продолжение)*

Атрибут	Назначение
Wait object flags, pointer, offset (флаги ожидания объекта, указатель, смещение)	Позволяют диспетчеру объектов базовым способом обнаружить нижележащий объект — диспетчер ядра, который должен быть использован для синхронизации, если какая-то из базовых служб ожидания, показанных ранее (WaitForSingleObject и т. п.), была применена к данному объекту
Methods (Методы)	Одна или несколько процедур, вызываемых диспетчером объектов автоматически в определенные моменты жизни объекта

Одним из видимых приложениям Windows атрибутов является *синхронизация*, которая указывает на способность потоков синхронизировать их выполнение, ожидая, пока какой-либо объект не перейдет из одного состояния в другое. Поток может синхронизироваться с объектами выполняемого задания, процесса, потока, файла, семафора, мьютекса и таймера. Все остальные объекты исполнительной системы синхронизацию не поддерживают. Способность объекта поддерживать синхронизацию основана на трех возможностях.

- Объект исполнительной системы является оболочкой для диспетчера объектов и содержит заголовок диспетчера, основную структуру, о которой рассказывается в разделе «Низкоуровневая IRQL-синхронизация» далее в этой главе.
- Создатель объекта типа требует объект по умолчанию, и диспетчер объектов предоставляет такой объект.
- Объект исполнительной системы имеет встроенный диспетчер объектов, такой как событие, где-нибудь внутри тела объекта, и владелец объекта предоставляет смещение на него диспетчеру объектов при регистрации объекта типа.

Методы объекта

Атрибут «методы» (см. табл. 8.22) содержит набор внутренних процедур, аналогичных конструкторам и деструкторам C++, то есть процедурам, которые автоматически вызываются при создании или удалении объекта. Диспетчер объектов расширяет этот замысел, вызывая метод объекта и в других ситуациях, например когда кто-нибудь открывает или закрывает дескриптор объекта или когда кто-нибудь пытается изменить защиту объекта. Некоторые типы объектов указывают методы, а некоторые их не указывают, в зависимости от того, как должен использоваться объект того или иного типа.

Когда компонент исполнительной системы создает новый объект типа, он с помощью диспетчера объектов может зарегистрировать один или несколько методов. После этого диспетчер объектов вызывает методы во вполне определенные моменты жизнедеятельности объектов данного типа, обычно при создании объекта, его удалении или каком-нибудь изменении. Методы, поддерживаемые диспетчером объектов, перечислены в табл. 8.23.

Причиной обращения к этим методам объектов служит факт той или иной операции над объектом (закрытия, дублирования, изменения степени безопасности и т. д.). Чтобы полностью обобщить эти методы, разработчикам диспетчера объектов пришлось предвидеть объекты любого типа. Это крайне усложнило бы работу ядра,

вдобавок функции, отвечающие за создание объектов, на деле экспортируются ядром же. Поскольку это позволяет внешним компонентам ядра создавать собственные типы объектов, ядро не смогло бы предвидеть возможные варианты несогласованного поведения. Хотя подобная функциональность не фигурирует в документации для разработчиков драйверов, она используется в Pcw.sys, DxgKrnl.sys, Win32k.sys, FltMgr.sys и др., чтобы определять WindowsStation, Desktop, PcwObject, Dxgk*, FilterCommunication/ConnectionPort, NdisCmState и другие объекты. Благодаря возможности расширения методов объектов эти драйверы определяют свои процедуры для проведения таких операций, как создание и запрос.

Таблица 8.23. Методы объекта

Метод	Когда вызывается
Open (Открыть)	При создании, открытии, дублировании или наследовании дескриптора объекта
Close (Закрыть)	При закрытии дескриптора объекта
Delete (Удалить)	Перед удалением объекта диспетчером объектов
Query name (Запросить имя)	Когда поток запрашивает имя объекта, такого как файл, которое существует в пространстве имен производного объекта
Parse (Провести анализ)	При поиске диспетчером объектов имени объекта, которое существует в пространстве имен производного объекта
Dump (Вывести дамп)	Не используется
Okay to close (Подтвердить закрытие)	Когда диспетчер объектов получает указание закрыть дескриптор
Security (Определить степень безопасности)	Когда процесс читает состояние защиты или изменяет защиту такого объекта, как файл, то есть имеет дело со сведениями, существующими в пространстве имен производного объекта

Другая причина существования таких методов сводится к тому, чтобы для управления жизненным циклом объекта иметь подобие виртуальных конструктора и деструктора. Это позволит механизмам более низкого уровня выполнять дополнительные действия во время создания и закрытия дескрипторов, а также при уничтожении объекта. Они дают возможность даже запрещать создание или закрытие дескриптора, когда это нежелательно. К примеру, механизм защищенных процессов, описанный в главе 3 тома 1, пользуется собственным методом создания дескрипторов, который запрещает менее защищенному процессу открывать дескрипторы для более защищенного. Кроме того, эти методы обеспечивают видимость некоторых внутренних функций диспетчера объектов, таких как дублирование и наследование, которые можно получить с помощью базовых служб.

Наконец, поскольку эти методы замещают функционал анализа и запроса имени, их можно применять для реализации дополнительного пространства имен вне досягаемости диспетчера объектов. Фактически именно так работают объекты файлов и разделов реестра — их пространство имен управляется драйвером файловой системы и диспетчером конфигурации, а диспетчер объектов видит лишь объекты `\REGISTRY` и `\Device\HarddiskVolumeN`. Несколько позже мы разберем детали и примеры для каждого из этих методов.

Диспетчер объектов вызывает методы, только если в инициализаторе типа их указатель не был задан как NULL. Единственным исключением из этого правила является процедура безопасности, которая, если не указано иное, выполняет по умолчанию `SeDefaultObjectMethod`. Этой процедуре не нужно знать внутреннюю структуру объекта, так как она имеет дело только с дескриптором безопасности объекта, а вы уже видели, что указатель на дескриптор безопасности хранится в общем заголовке объекта, а не внутри тела объекта. Но если объект требует собственной дополнительной проверки безопасности, он может определить свою процедуру безопасности, что опять же играет роль для файловых объектов и объектов разделов реестра, которые хранят свои данные по защите так, как напрямую решает файловая система или диспетчер конфигурации.

Диспетчер объектов вызывает метод *open* всякий раз, когда создает дескриптор объекта, что происходит при его создании, открытии, дублировании или наследовании. Объекты WindowStation и Desktop предоставляют метод *open*; например, объект типа WindowStation требует такой метод *open*, чтобы Win32k.sys мог совместно использовать часть памяти с процессом, который служит в качестве пула памяти, связанного с рабочим столом.

Пример использования метода *close* встречается в системе ввода/вывода. Диспетчер ввода/вывода регистрирует метод *close* для объекта типа «файл», а диспетчер объектов вызывает метод *close* при каждом закрытии дескриптора файлового объекта. Этот метод *close* проверяет, не владеет ли процесс, закрывающий дескриптор файла, какими-либо незавершенными блокировками, касающимися этого файла, и если такие блокировки имеются, он их удаляет. Проверка наличия блокировок, связанных с файлами, не является обязанностью, которую должен или может выполнять сам диспетчер объектов.

Перед удалением временного объекта из памяти диспетчер объектов вызывает метод *delete*, если такой метод был зарегистрирован. Диспетчер памяти, к примеру, регистрирует метод *delete* для объекта типа «раздел», и этот метод освобождает физические страницы, использовавшиеся в разделе. Он также проверяет перед удалением объекта типа «раздел» факт удаления любых внутренних структур данных, выделенных разделу диспетчером памяти. И, опять же, диспетчер объектов не может справиться с этой работой, поскольку ему ничего не известно о внутренней работе диспетчера памяти. Методы удаления для других типов объектов выполняют аналогичные функции.

Метод *parse* (так же, как и метод запроса имени *query name*) позволяет диспетчеру объектов уступать управление поиском объекта производному диспетчеру объектов, если он найдет объект, существующий за пределами пространства имен диспетчера объектов. Когда диспетчер объектов ищет имя объекта, он приостанавливает свой поиск, если обнаруживает объект в пути, который он связывает с методом *parse*. Диспетчер объектов вызывает метод *parse*, передавая ему оставшуюся часть от имени объекта, поиском которого он занимается. Кроме пространства имен диспетчера объектов, в Windows есть еще два пространства имен: пространство имен реестра, реализуемое диспетчером конфигурации, и пространство имен файловой системы, реализуемое диспетчером ввода-вывода с помощью драйверов файловой системы. (Более подробно диспетчер конфигурации рассматривается в главе 10, а диспетчер ввода/вывода и драйверы файловой системы — в главе 6 тома 1.)

Например, когда процесс открывает дескриптор объекта `\Device\HarddiskVolume1\docs\resume.doc`, проходит по дереву его имени до тех пор, пока не дойдет до имени

объекта устройства `HarddiskVolume1`. Заметив, что с этим объектом связан метод *parse*, он его вызывает, передавая как аргумент оставшуюся часть искомого имени — в данном случае строку `docs\resume.doc`. Метод *parse* для объектов устройств является процедурой ввода-вывода, поскольку объекты типа «устройство» определялись диспетчером ввода-вывода и им же регистрировался для них метод *parse*. Процедура *parse*, задаваемая диспетчером ввода-вывода, получает строку имени и передает ее соответствующей файловой системе, которая ищет файл на диске и открывает его.

Метод безопасности *security*, который также использует система ввода-вывода, аналогичен методу *parse*. Он вызывается в том случае, если поток пытается запросить или изменить информацию безопасности, защищающую файл. Эта информация для файлов отличается от такой же информации для других объектов, поскольку хранится в самом файле, а не в памяти. Поэтому для поиска информации о безопасности и ее считывании или изменении должна быть вызвана система ввода-вывода.

И наконец, метод подтверждения закрытия *okay-to-close* используется в качестве дополнительного уровня защиты от злонамеренного или ошибочного закрытия дескрипторов, используемых для системных целей. Например, у каждого процесса есть дескриптор на объект Desktop или объекты, с которыми у его потока или потоков имеются видимые окна. В соответствии со стандартной моделью безопасности такие потоки могут закрывать свои дескрипторы на их рабочие столы, поскольку у процесса есть полный контроль над его собственными объектами. По этому сценарию потоки остаются без связанного с ними рабочего стола, что является нарушением многооконной модели. Для предотвращения подобного поведения Win32k.sys регистрирует процедуру *okay-to-close* для объектов рабочего стола Desktop и объектов WindowStation.

Дескрипторы объекта и таблица дескрипторов процесса

Когда какой-либо процесс создает или открывает объект по имени, он получает дескриптор, дающий ему доступ к этому объекту. Ссылаться на объект по его дескриптору быстрее, чем использовать его имя, поскольку диспетчер объектов может не заниматься поиском по имени и находить объект напрямую. Как было упомянуто ранее, процессы также могут получать дескрипторы объектов, наследуя их во время создания процесса (если создатель передаст флаг наследования при вызове `CreateProcess` и дескриптор был помечен как наследуемый, и при создании, и при вызове функции Windows `SetHandleInformation`) либо при получении дубликата дескриптора от другого процесса (см. функцию Windows `DuplicateHandle`).

Все процессы пользовательского режима должны иметь дескриптор объекта, прежде чем их потоки смогут использовать объект. Идея применения дескрипторов для управления системными ресурсами не нова. Например, библиотеки C и C++, возвращают дескрипторы открытых файлов. Дескрипторы служат косвенными указателями на системные ресурсы, эта косвенность не позволяет приложениям напрямую манипулировать структурами данных системы.

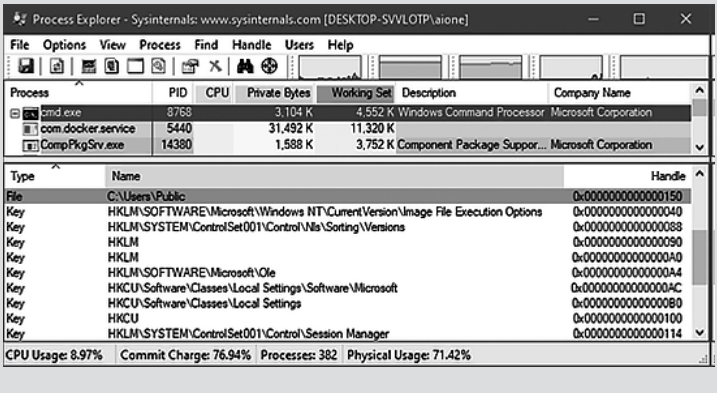
Дескрипторы объектов дают и другие преимущества. Во-первых, за исключением того, к чему они относятся, нет разницы между дескриптором файла, события или процесса. Эта схожесть обеспечивает единый интерфейс для ссылки на объекты вне зависимости от их типа. Во-вторых, диспетчер объектов имеет исключительное право на создание дескрипторов и на определение местоположения объекта, который относится к дескриптору. Это дает ему возможность проверять каждое действие

в пользовательском режиме, влияющее на объект, чтобы убедиться, что профиль безопасности вызывающей программы разрешает проводить запрашиваемую операцию над объектом.

ПРИМЕЧАНИЕ Компоненты исполнительной системы и драйверы устройств могут обращаться к объектам напрямую, так как они выполняются в режиме ядра, а значит, имеют доступ к структурам объекта в системной памяти. Однако и они должны объявить о своем использовании объекта, увеличив значение счетчика ссылок, чтобы объект нельзя было удалить из памяти, пока он еще используется. (Подробности см. далее в подразделе «Сохранение объектов».) Но для успешного использования объекта драйверы устройств должны знать определение внутренней структуры объекта, а для многих объектов она не предоставляется. Взамен драйверам устройств рекомендуется использовать соответствующие API-функции ядра для изменения или чтения информации из объекта. Например, хотя драйверы устройств могут получить указатель на объект типа «процесс» (EPROCESS), его структура им не известна, и должны быть использованы API-функции вида Ps*. Для других объектов непрозрачен сам тип (это касается большинства объектов исполнительной системы, в которые заключен диспетчер объектов, в качестве примеров можно привести события или мьютексы). Для таких объектов драйверы должны использовать такие же системные вызовы, которые в конечном счете используются приложениями пользовательского режима (такие как ZwCreateEvent), и задействовать дескрипторы, а не указатели на объекты.

ЭКСПЕРИМЕНТ. Просмотр открытых дескрипторов

Запустите Process Explorer и убедитесь, что нижняя панель включена и настроена на показ дескрипторов (откройте меню Вид (View), щелкните на пункте Вид нижней панели (Lower Pane View), а потом выберите Дескрипторы (Handles)). Откройте командную строку и просмотрите таблицу дескрипторов для нового процесса cmd.exe. Вы должны увидеть открытый дескриптор для текущего каталога. Например, предположим, что это каталог `C:\Users\Public`, и тогда Process Explorer покажет следующее.

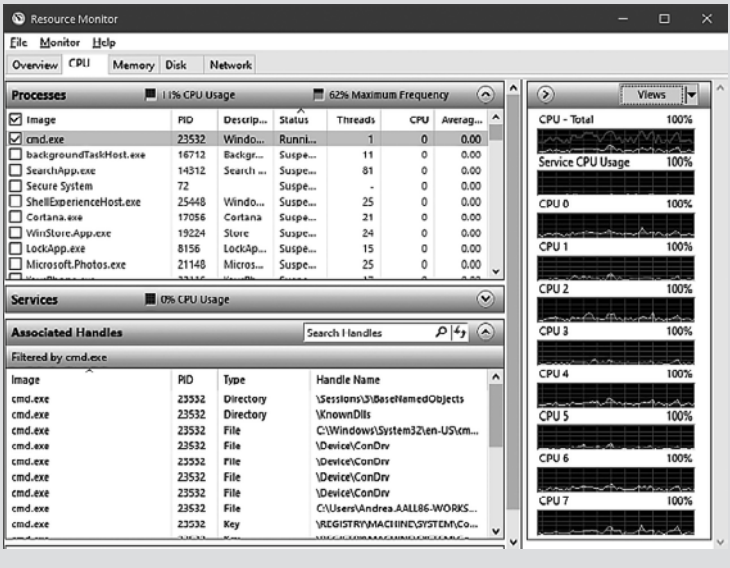


После этого остановите Process Explorer, нажав клавишу пробела или выбрав в меню Вид (View), Обновление скорости (Update Speed) пункт Пауза (Pause).

Теперь смените текущий каталог командой `cd` и нажмите клавишу F5, чтобы обновить изображение. В Process Explorer вы увидите, что дескриптор предыдущего каталога закрыт, а для текущего открыт новый. Старый дескриптор помечен красным цветом, а новый — зеленым.

Возможность цветовой подсветки различий в Process Explorer позволяет легко отслеживать изменения в таблице дескрипторов. Например, если у процесса возникает утечка дескрипторов, Process Explorer сможет очень быстро показать, какой или какие дескрипторы открываются, но не закрываются (зачастую вы увидите длинный список дескрипторов одного и того же объекта). Эта информация сможет помочь программисту отыскать место утечки.

Монитор ресурсов тоже показывает открытые (именованные) дескрипторы для процессов, выбранных путем установки флажков напротив их имен. Вот как выглядят дескрипторы открытого окна командной строки.



Таблицу открытых дескрипторов можно также вывести, используя средство командной строки Handle из серии программных продуктов Sysinternals. Посмотрите, к примеру, на следующий, частично показанный вывод, полученный с помощью средства Handle при изучении дескрипторов файловых объектов, находящихся в таблице дескрипторов для процесса Cmd.exe до и после изменения каталога. По умолчанию Handle отфильтровывает нефайловые дескрипторы, пока не будет использован ключ –a, который приводит к выводу всех дескрипторов в процессе, аналогично Process Explorer:

```
C:\Users\aione>\sysint\handle.exe -p 8768 -a users
Nthandle v4.22 - Handle viewer
Copyright (C) 1997-2019 Mark Russinovich
Sysinternals - www.sysinternals.com
cmd.exe            pid: 8768 type: File            150: C:\Users\Public
```

Дескриптор объекта — это индекс в *таблице дескрипторов*, относящейся к конкретному процессу. Он указывается исполнительным блоком процесса (EPROCESS) (описан в главе 3 тома 1). Этот индекс умножается на 4 (сдвиг на 2 бита), чтобы оставить место для битов под дескриптор, которые управляют поведением некоторых API — например, запрещают оповещения от портов завершения ввода/вывода или изменяют ход отладки процесса. Таким образом, первый индекс дескриптора имеет значение 4, второй — 8 и т. д. Использование дескрипторов 5, 6 или 7 попросту перенаправляется к объекту с дескриптором 4, а значения 9, 10 и 11 будут восприняты как ссылка на объект с дескриптором 8.

Таблица дескрипторов процесса содержит указатели на все объекты, для которых процесс прямо сейчас имеет открытые дескрипторы, а сами они экономятся, из-за чего новым индексом дескриптора по возможности будет уже существующий индекс закрытого дескриптора. Таблицы дескрипторов (рис. 8.33) реализованы по древовидной схеме, подобной той, которую реализует блок управления памятью x86 для перевода виртуальных адресов в физические, но из соображений совместимости — с ограничением в 24 бита, тем самым оставляя возможность создать максимум 16 777 215 (2^{24}–1) дескрипторов на процесс. На рис. 8.34 изображена схема таблицы дескрипторов в Windows. Чтобы сэкономить память ядра, при создании процесса размещают только самые низкоуровневые дескрипторы, на остальных уровнях их создают по мере надобности. Субтаблица дескрипторов содержит столько записей, сколько поместится на одной странице, минус одна для нужд аудита дескрипторов. Например, в 64-разрядных системах объем страницы 4096 байт нужно разделить на размер записи таблицы дескрипторов (16 байт), выходит 256 – 1 байт, в итоге у нас 255 записей для дескрипторов самого низкого уровня. В таблице среднего уровня хранится целая страница указателей на субтаблицы дескрипторов, так что количество последних зависит от размера страницы и разрядности указателей на конкретной платформе. Опять же, взяв в качестве примера 64-разрядные системы, мы получаем 4096/8, или 512 записей. Ввиду ограничения в 24 бита в таблице указателей высшего уровня можно разместить только 128 записей.

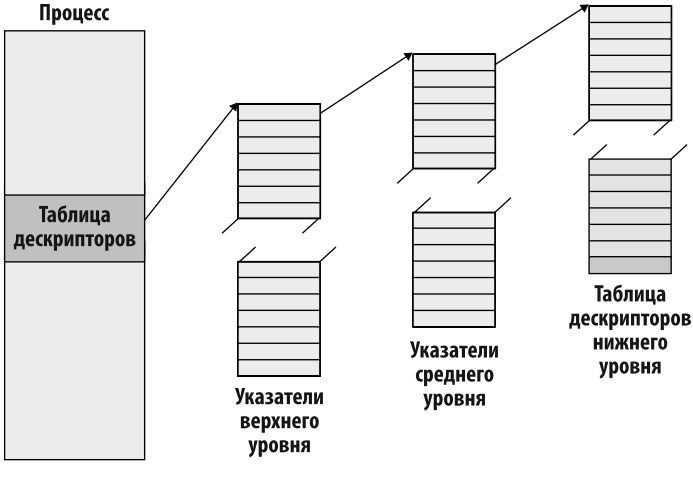


Рис. 8.33. Архитектура таблицы дескрипторов процесса Windows

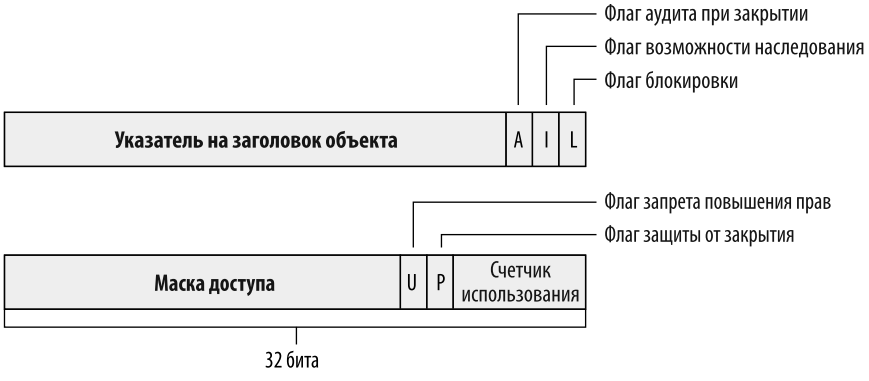


Рис. 8.34. Структура записи в таблице дескрипторов

ЭКСПЕРИМЕНТ. Создание максимального количества дескрипторов

Тестовая программа TestLimit от SysInternals дает возможность открывать дескрипторы объекта до тех пор, пока не будет получен отказ. Этим можно воспользоваться, чтобы узнать, сколько дескрипторов может открыть один процесс в вашей системе. Поскольку таблицы дескрипторов размещаются в пуле подкачиваемой памяти, вы можете выйти за его рамки прежде, чем достигнете искомого максимума дескрипторов. Чтобы узнать, сколько же все-таки их в вашей системе можно создать, выполните следующие шаги.

1. Скачайте исполняемый файл TestLimit, соответствующий 32/64-разрядной Windows, с сайта https://docs.microsoft.com/en-us/sysinternals/downloads/testlimit.
2. Запустите Process Explorer, откройте меню Вид (View), а затем щелкните на пункте Системная информация (System Information). Перейдите на вкладку Память (Memory). Обратите внимание на максимальный размер пула подкачиваемой памяти. (Чтобы показать максимальные размеры пула памяти, Process Explorer должен быть корректно настроен для доступа к символам образа ядра, Ntoskrnl.exe.) Оставьте окно с системной информацией открытым, чтобы видеть уровень использования пула, когда программа TestLimit будет запущена.
3. Откройте командную строку.
4. Запустите программу TestLimit с аргументом `-h` (просто наберите `testlimit -h`). Когда TestLimit будет отказано в открытии нового дескриптора, она сообщит, сколько всего смогла сделать. Если это число меньше примерно 1 млн, вы, скорее всего, исчерпали пул подкачиваемой памяти прежде, чем достигли теоретического максимума дескрипторов для одного процесса.
5. Закройте командную строку. Это приведет к завершению процесса TestLimit, а значит, все открытые дескрипторы закроются.

Как видно на рис. 8.34, в 32-разрядных системах каждая запись дескриптора состоит из структуры с двумя 32-разрядными элементами: указателем на объект (где младшие 3 бита заняты флагами, ввиду того что все объекты выровнены кратно 8 байтам, и эти биты можно считать 0) и маской предоставленного доступа (где из 32 нужны только 25, так как базовые права никогда не хранятся в записи о дескрипторе), а также двух флагов и *счетчика использования ссылок*, речь о котором пойдет чуть позже.

В 64-разрядных системах присутствуют те же базовые блоки данных, но иначе закодированные. Например, поскольку объекты выровнены по 16 байт и старшие 4 бита можно принять за 0, здесь для кодировки указателя на объект требуются 44 бита (при условии, что у процессора четырехуровневая структура страниц и 48-разрядное управление виртуальной памятью). Это уже позволяет кодировать флаг `Protect from close` (Защита от закрытия) как часть первоначальной тройки флагов, которых в итоге получается четыре. Еще одно изменение заключается в том, что *счетчик использования* размещается в последних 16 битах рядом с указателем, а не с маской доступа. Наконец, флаг `No rights upgrade` (Запретить повышение прав) все еще находится рядом с маской доступа, но остающиеся 6 бит свободны, как и 34 бита от выравнивания, итого 16 байт. В системах же с LA57 с пятиуровневой структурой страниц изменения идут еще дальше и указатель требует уже 53 бита, оставляя для счетчика использования всего 7 бит.

Поскольку мы упомянули ряд флагов, рассмотрим их назначение. Первым флагом является бит блокировки, показывающий, что запись в настоящее время используется. Технически он называется `unlocked` (Разблокировано), в связи с чем вам следует ожидать, что старший бит будет установлен. Вторым флагом является указатель на возможность наследования, то есть он показывает, получат ли процессы, созданные данным процессом, копию этого дескриптора в свои таблицы дескрипторов. Как уже говорилось, наследование дескриптора может быть определено как при его создании, так и позже с помощью функции `SetHandleInformation`. Третий флаг показывает, должно ли закрытие объекта генерировать контрольное сообщение (флаг не показывается в Windows, диспетчер объектов использует его для внутренних нужд). Следующий бит, `Protect from close` (Защита от закрытия), показывает, разрешено ли вызывающей программе закрыть этот дескриптор. (Этот флаг тоже можно изменить функцией `SetHandleInformation`.) Наконец, бит `No rights upgrade` (Запретить повышение прав) определяет, следует ли расширять права доступа при дублирования данного дескриптора в процесс с более высокими привилегиями.

Эти последние четыре флага будут доступны драйверам при использовании структуры `OBJECT_HANDLE_INFORMATION`, передаваемой таким функциям, как `ObReferenceObjectByHandle`, и отражающейся в `OBJ_INHERIT` (0x2), `OBJ_AUDIT_OBJECT_CLOSE` (0x4), `OBJ_PROTECT_CLOSE` (0x1) и `OBJ_NO_RIGHTS_UPGRADE` (0x8), в точности совпадающих с «дырами» в ранее упомянутых определениях атрибутов `OBJ_`, которые можно установить при создании объекта. Таким образом, во время исполнения атрибуты объекта могут сыграть роль в кодировании особенностей поведения как объекта, так и дескриптора к нему.

Наконец, мы упомянули о наличии *счетчика использования ссылок* сразу и в счетчике ссылок в заголовке объекта, и в записи таблицы дескрипторов. Этот

полезный функционал кодирует кэшированное количество (на основе количества доступных битов) уже существующих ссылок как часть каждого дескриптора, а затем суммирует счетчики использования всех процессов, у которых есть дескриптор объекта, в счетчик указателей — в заголовке объекта. В итоге этот счетчик включает в себя количество дескрипторов, ссылок ядра через `ObReferenceObject` и количество кэшированных ссылок для каждого дескриптора.

Каждый раз, когда процесс заканчивает пользоваться объектом, разыменовывая один из его дескрипторов — в сущности, вызывая *любую* функцию Windows API, которая принимает на вход дескриптор и в конечном счете превращает его в объект, — кэшированное количество ссылок сбрасывается. Иными словами, счетчик использования уменьшается на 1, пока не достигнет нуля, после чего его отслеживание прекращается. Это позволяет точно определить, сколько раз искомый объект подвергался бы использованию/доступу/управлению через дескриптор конкретного процесса.

Команда отладчика `!truedef`, запущенная с аргументом `-v`, пользуется данной возможностью, чтобы показать каждый дескриптор, ссылающийся на объект, и то, сколько раз он был применен (если вместе с этим считать число поглощенных/отброшенных счетчиков). В одном из дальнейших экспериментов вы воспользуетесь ею, чтобы детальнее узнать о работе с объектом.

Системным компонентам и драйверам устройств нередко требуется открывать дескрипторы к объектам, к которым не должно быть доступа со стороны приложений пользовательского режима или же не требуется привязка к какому-либо процессу в принципе. Это достигается путем создания дескрипторов в *таблице дескрипторов ядра* для внутреннего применения, известной как `ObpKernelHandleTable`, которая принадлежит процессу SYSTEM. Эти дескрипторы доступны только из режима ядра в контексте любого процесса. Таким образом, функции режима ядра могут обращаться к такому дескриптору из любого процесса, не влияя на производительность.

Диспетчер объектов опознает дескрипторы из таблицы ядра по установленному старшему биту, иными словами, когда значение ссылки на них превышает 0x80000000 на 32-разрядных системах или 0xFFFFFFFF80000000 на 64-разрядных (поскольку с точки зрения типов данных дескриптор является указателем, компилятор принудительно выполняет расширение знака).

Кроме того, таблица ядра служит для хранения дескрипторов системного и минимальных процессов, таким образом, все дескрипторы, созданные SYSTEM (в частности, при исполнении кода в системных потоках), фактически относятся к ядру, поскольку в структуре `EPROCESS` таких процессов символ `ObpKernelHandleTable` указан как `ObjectTable`. Теоретически это значит, что процесс пользовательского режима с достаточно высокими привилегиями мог бы задействовать функцию `DuplicateHandle`, чтобы извлечь дескриптор ядра для использования в пользовательском режиме. Но возможность данной атаки минимизирована, начиная с Windows Vista, где было введено понятие защищенных процессов, подробно описанное в томе 1.

Более того, по соображениям безопасности любой дескриптор, созданный драйвером уровня ядра, у которого прежний режим указан как ядро, в недавних версиях Windows автоматически становится дескриптором ядра, что позволяет предотвратить его утечку в пользовательские приложения.

ЭКСПЕРИМЕНТ. Просмотр таблицы дескрипторов с помощью отладчика ядра

Команда `!handle` отладчика ядра имеет три аргумента:

```
!handle <индекс дескриптора> <флаги> <идентификатор процесса>
```

Индекс дескриптора идентифицирует искомую запись в таблице (0 будет значить «показать все дескрипторы»). Индекс первого дескриптора имеет значение 4, второго — 8 и т. д. Например, по команде `!handle 4` будет показан первый дескриптор для текущего процесса.

Доступные для указания флаги можно задать в виде битовой маски, где бит 0 значит «показать информацию только из записи о дескрипторе», бит 1 — «показать свободные дескрипторы (не только используемые)», а бит 2 — «показать информацию об объекте, на который ссылается этот дескриптор». Следующая команда выводит полную информацию о таблице дескрипторов процесса с ID 0x1540:

```
lkd> !handle 0 7 1540

PROCESS ffff898f239ac440
    SessionId: 0 Cid: 1540 Peb: 1ae33d000 ParentCid: 03c0
    DirBase: 211e1d000 ObjectTable: ffffc704b46dbd40 HandleCount: 641.
    Image: com.docker.service

Handle table at ffffc704b46dbd40 with 641 entries in use

0004: Object: ffff898f239589e0 GrantedAccess: 001f0003 (Protected) (Inherit) Entry:
ffffc704b45ff010
Object: ffff898f239589e0 Type: (ffff898f032e2560) Event
    ObjectHeader: ffff898f239589b0 (new version)
        HandleCount: 1 PointerCount: 32766

0008: Object: ffff898f23869770 GrantedAccess: 00000804 (Audit) Entry:
ffffc704b45ff020
Object: ffff898f23869770 Type: (ffff898f033f7220) EtwRegistration
    ObjectHeader: ffff898f23869740 (new version)
        HandleCount: 1 PointerCount: 32764
```

Вместо того чтобы запоминать значение каждого из этих битов и переводить ID процесса в шестнадцатеричную систему, можете воспользоваться моделью данных отладчика и обращаться к дескрипторам посредством пространства имен процесса Io.Handles. Например, команда `dx@$curprocess.Io.Handles[4]` покажет первый дескриптор текущего процесса, включая права доступа и имя. А приведенная далее команда отображает детальное описание дескрипторов процесса и ID 5440 (он же 0x1540):

```
lkd> dx -r2 @$cursession.Processes[5440].Io.Handles
@$cursession.Processes[5440].Io.Handles
     [0x4]
        Handle           : 0x4
        Type             : Event
        GrantedAccess    : Delete | ReadControl | WriteDac | WriteOwner | Synch |
QueryState | ModifyState
```

```
      Object           [Type: _OBJECT_HEADER]
   [0x8]
      Handle           : 0x8
      Type             : EtwRegistration
      GrantedAccess
      Object           [Type: _OBJECT_HEADER]
   [0xc]
      Handle           : 0xc
      Type             : Event
      GrantedAccess    : Delete | ReadControl | WriteDac | WriteOwner | Synch |
                         QueryState | ModifyState
      Object           [Type: _OBJECT_HEADER]
```

Кроме того, воспользовавшись моделью данных отладчика с предикатом LINQ, можно выполнить поиск по более интересным критериям, например найти отображения именованных объектов разделов с доступом на чтение/запись:

```
lkd> dx @$cursession.Processes[5440].Io.Handles.Where(h => (h.Type == "Section") &&
(h.GrantedAccess.MapWrite) && (h.GrantedAccess.MapRead)).Select(h => h.ObjectName)
@$cursession.Processes[5440].Io.Handles.Where(h => (h.Type == "Section") &&
(h.GrantedAccess.MapWrite) && (h.GrantedAccess.MapRead)).Select(h => h.ObjectName)
    [0x16c]          : "Cor_Private_IPCBlock_v4_5440"
    [0x170]          : "Cor_SxSPublic_IPCBlock"
    [0x354]          : "windows_shell_global_counters"
    [0x3b8]          : "UrlZonesSM_DESKTOP-SVVLOTP$"
    [0x680]          : "NLS_CodePage_1252_3_2_0_0"
```

ЭКСПЕРИМЕНТ. Поиск открытых файлов с помощью отладчика ядра

Хотя для поиска дескрипторов открытых файлов можно воспользоваться такими средствами, как Process Explorer, Handle и OpenFiles.exe, они недоступны при просмотре аварийного дампа или удаленном анализе системы. Вместо них для поиска дескрипторов, открытых для файлов в томе, можно воспользоваться командой `!devhandles` (подробнее об устройствах, файлах и томах — в главе 11).

1. В первую очередь необходимо выбрать букву интересующего вас диска и получить указатель на его объект Device. Для этого воспользуйтесь командой `!object`, как показано далее:

   ```
   lkd> !object \Global??\C:
   Object: ffffc704ae684970 Type: (ffff898f03295a60) SymbolicLink
       ObjectHeader: ffffc704ae684940 (new version)
       HandleCount: 0 PointerCount: 1
       Directory Object: ffffc704ade04ca0 Name: C:
       Flags: 00000000 ( Local )
       Target String is '\Device\HarddiskVolume3'
       Drive Letter Index is 3 (C:)
   ```

2. Затем снова выполните команду `!object`, чтобы получить объект Device для нужного имени тома:

   ```
   1: kd> !object \Device\HarddiskVolume1
   Object: FFFF898F0820D8F0 Type: (fffffa8000ca0750) Device
   ```

3. Наконец, передайте полученный указатель на объект Device команде !devhandles. Каждый показанный объект будет указывать на файл:

```
lkd> !devhandles 0xFFFF898F0820D8F0

Checking handle table for process 0xffff898f0327d300
Kernel handle table at ffffc704ade05580 with 7047 entries in use

PROCESS ffff898f0327d300
    SessionId: none Cid: 0004 Peb: 00000000 ParentCid: 0000
    DirBase: 001ad000 ObjectTable: ffffc704ade05580 HandleCount: 7023.
    Image: System

019c: Object: ffff898F080836a0 GrantedAccess: 0012019f (Protected) (Inherit)
(Audit) Entry: ffffc704ade28670
Object: ffff898F080836a0 Type: (ffff898f032f9820) File
    ObjectHeader: ffff898F08083670 (new version)
        HandleCount: 1 PointerCount: 32767
        Directory Object: 00000000 Name: \$Extend\$RmMetadata\$TxfLog\
                                         $TxfLog.blf {HarddiskVolume4}
```

Хотя этот инструмент справляется с задачей, вы наверняка заметили, что до того, как покажутся первые несколько дескрипторов, приходится подождать от 30 с до 1 мин. Вы можете добиться аналогичного результата, воспользовавшись моделью данных отладчика с предикатом LINQ, и начать получать результаты немедленно:

```
lkd> dx -r2 @$cursession.Processes.Select(p => p.Io.Handles.Where(h =>
    h.Type == "File").Where(f => f.Object.UnderlyingObject.DeviceObject ==
     (nt!_DEVICE_OBJECT*)0xFFFF898F0820D8F0).Select(f =>
    f.Object.UnderlyingObject.FileName))
@$cursession.Processes.Select(p => p.Io.Handles.Where(h => h.Type == "File").
Where(f => f.Object.UnderlyingObject.DeviceObject == (nt!_DEVICE_OBJECT*)
0xFFFF898F0820D8F0).Select(f => f.Object.UnderlyingObject.FileName))
     [0x0]
     [0x19c]    : "\$Extend\$RmMetadata\$TxfLog\$TxfLog.blf" [Type: _UNICODE_STRING]
     [0x2dc]    : "\$Extend\$RmMetadata\$Txf:$I30:$INDEX_ALLOCATION"
                  [Type: _UNICODE_STRING]
     [0x2e0]    : "\$Extend\$RmMetadata\$TxfLog\$TxfLogContainer00000000000000000002"
                  [Type: _UNICODE_STRING]
```

Резервные объекты

Поскольку с помощью объектов представлено все, от событий до файлов и до сообщений, которыми процессы обмениваются между собой, способность приложений и кода ядра создавать объекты является основой для нормального и желаемого поведения в процессе выполнения любого фрагмента кода Windows. Если назначить объект не удается, это обычно служит причиной чего угодно, от утраты функционирования (процесс не может открыть файл) до утраты данных или попадания в аварийное состояние (процесс не может назначить объект синхронизации). Хуже того, в определенных ситуациях создание отчетов об ошибках, связанных с отказами в создании объектов, может само требовать назначения новых объектов. Для разрешения подобных ситуаций в Windows реализуются два специальных резервных объекта: резервный объект асинхронного вызова процедуры в пользовательском режиме — User

APC reserve object и резервный объект пакета завершения ввода-вывода — I/O Completion packet reserve object. Следует заметить, что сам механизм резервных объектов полностью открыт для расширения и в будущих версиях Windows могут быть добавлены и другие типы резервных объектов. По большому счету, резервный объект является механизмом, позволяющим любой структуре данных режима ядра быть заключенной в оболочку объекта (с соответствующим дескриптором, именем и мерами обеспечения безопасности) для дальнейшего использования.

Как уже говорилось ранее в разделе, посвященном APC-вызовам, эти вызовы используются для таких операций, как приостановка, завершение работы и завершение ввода-вывода, а также для связи между приложениями пользовательского режима, которым нужно обеспечивать асинхронные обратные вызовы. Когда приложение пользовательского режима запрашивает User APC, нацеленный на другой поток, оно использует API-функцию `QueueUserApc` из Kernelbase.dll, которая, в свою очередь, осуществляет системный вызов `NtQueueApcThread`. В ядре этот системный вызов пытается выделить часть выгружаемого пула для хранения структуры управляющего объекта KAPC, связанной с APC-вызовом. Когда испытывается дефицит памяти, эта операция терпит неудачу, мешая доставке APC, что, в зависимости от того, какой APC-вызов для какой цели использовался, может привести к потере данных или к нарушению функционирования системы.

Чтобы предотвратить такую возможность, приложение пользовательского режима может в самом начале своей работы воспользоваться системным вызовом `NtAllocateReserveObject`, чтобы запросить у ядра заблаговременное выделение памяти под KAPC-структуру. Затем приложение использует другой системный вызов, `NtQueueUserApcThreadEx`, содержащий дополнительный параметр, предназначенный для хранения дескриптора резервного объекта. Вместо размещения новой структуры ядро пытается получить резервный объект (путем установки его бита использования `InUse` в `true`) и воспользоваться им, пока у KAPC-объекта не минует в нем надобность, в случае чего резервный объект будет возвращен системе. В данное время, в целях недопущения сторонними разработчиками неправильного управления системными ресурсами, API-функции, работающие с резервными объектами, доступны только внутри системы, через системные вызовы компонентов операционной системы. Например, RPC-библиотека использует резервные APC-объекты, чтобы гарантировать, что асинхронные обратные вызовы все же получат возможность возврата в случае дефицита памяти.

Похожий сценарий реализуется, когда приложению нужна гарантированная доставка сообщения или пакета порта завершения ввода-вывода. Как правило, пакеты отправляются с помощью функции `PostQueuedCompletionStatus` из Kernelbase.dll, которая вызывает функцию `NtSetIoCompletion`. Так же, как и в случае с пользовательскими APC-вызовами, ядро должно разместить структуру диспетчера ввода-вывода, чтобы в ней содержалась информация пакета завершения, и если размещение потерпит неудачу, пакет не сможет быть создан. При наличии резервных объектов приложение может воспользоваться в начале своей работы API-функцией `NtAllocateReserveObject`, чтобы получить заранее размещенный в ядре пакет завершения ввода-вывода, а для предоставления дескриптора этому резервному объекту, гарантирующего успешное прохождение, может быть использован системный вызов `NtSetIoCompletionEx`. как и в случае с резервными объектами User APC, эта

функциональная возможность зарезервирована для системных компонентов и используется как RPC-библиотекой, так и Windows-службой Peer-To-Peer BranchCache для гарантированного завершения асинхронных операций ввода-вывода.

Безопасность объектов

При открытии файла нужно указать, что с ним собираются делать: читать или записывать данные. При попытке записи в файл, открытый для доступа только по чтению, будет получена ошибка. Точно так же в исполнительной системе, когда процесс создает объект или открывает дескриптор уже существующего объекта, процесс должен указать набор желаемых прав доступа, то есть что он хочет делать с этим объектом. Он может запросить либо набор стандартных прав доступа (например, по чтению, записи и выполнению), применимый ко всем объектам типа, либо конкретные права доступа, изменяющиеся в зависимости от типа объекта. Например, процесс может запросить доступ к файловому объекту для удаления или для добавления. Аналогично этому, он может потребовать возможность приостановки или завершения работы объекта потока.

Когда процесс открывает дескриптор объекта, диспетчер объектов вызывает *монитор безопасности ссылок*, ту часть системы безопасности, которая работает в режиме ядра, передавая ему от процесса набор желаемых прав доступа. Монитор безопасности ссылок проверяет, разрешает ли дескриптор безопасности объекта тот вид доступа, который запрашивается процессом. Если да, монитор возвращает набор выделенных прав доступа, которым разрешено пользоваться процессу, а диспетчер объектов сохраняет права доступа в создаваемом им дескрипторе объекта. Как система безопасности определяет, кто получает права и к каким объектам, рассматривалось в главе 7 тома 1.

Впоследствии, когда потоки процесса используют дескриптор через системный вызов, диспетчер объектов может быстро проверить, сохранен ли в дескрипторе объекта набор выделенных прав доступа, соответствующий использованию применительно к той службе объекта, которая была вызвана потоком. Например, если вызывающая программа запрашивает доступ по чтению к объекту раздела, но затем вызывает службу записи в этот объект, служба ответит отказом.

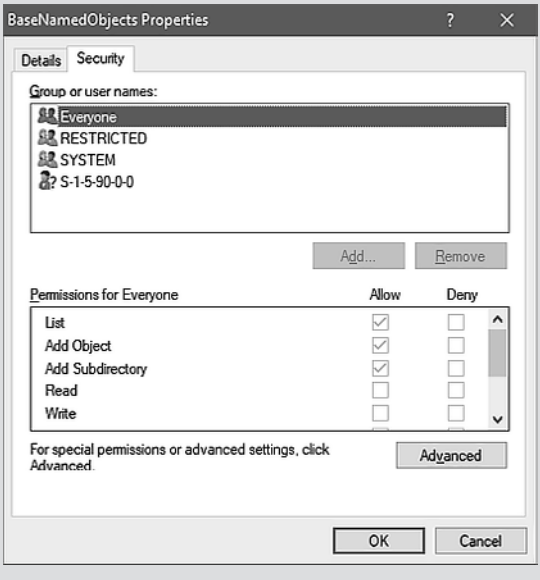
ЭКСПЕРИМЕНТ. Просмотр безопасности объекта

Для просмотра разных разрешений для какого-либо объекта вы можете воспользоваться Process Hacker, Process Explorer, WinObj, WinObjEx64 или AccessChk — инструментами из набора Sysinternals или из числа утилит с открытым исходным кодом, доступных на GitHub. Рассмотрим различные способы вывода списка контроля доступа объекта (access control list, ACL).

- Для перехода к любому объекту системы, включая каталоги объектов, вы можете воспользоваться утилитами WinObj или WinObjEx64. Щелкните правой кнопкой мыши на объекте и выберите Свойства (Properties). Например, выберите каталог `BaseNamedObjects`, нажмите Свойства (Properties) и перейдите на вкладку Безопасность (Security). Отобразится диалоговое

окно, подобное приведенному далее. Поскольку WinObjEx64 поддерживает больше типов объектов, это окно доступно для более широкого набора системных ресурсов.

Изучая настройки в этом окне, вы можете заметить, что группа Все (Everyone) не имеет права удалять каталог, но пользователь SYSTEM, например, имеет (поскольку там будут хранить свои объекты службы сеанса 0 с правами SYSTEM).



- Вместо использования WinObj или WinObjEx64 вы можете просмотреть таблицу дескрипторов процесса с помощью Process Explorer, как было показано в эксперименте «Просмотр открытых дескрипторов» ранее в главе, или же с помощью Process Hacker, отображающего примерно то же самое. Рассмотрим таблицу дескрипторов процесса Explorer.exe. Вы наверняка заметите там дескриптор объекта-каталога с адресом `\Sessions\n\BaseNamedObjects`, где *n* — номер сеанса, определяемый на этапе загрузки. (О пространстве имен в рамках сеанса будет рассказано чуть позже.) Вы можете дважды щелкнуть на дескрипторе объекта и, перейдя на вкладку Безопасность (Security), увидеть похожее диалоговое окно, где еще больше пользователей и выданных прав.
- Наконец, вы можете воспользоваться AccessChk, чтобы запросить информацию о параметрах безопасности объекта, добавив аргумент `-o`, как показано в консольном выводе далее (подробнее об уровнях целостности и мониторе защиты говорится в главе 7 тома 1):

```
C:\sysint>accesschk -o \Sessions\1\BaseNamedObjects

Accesschk v6.13 - Reports effective permissions for securable objects
Copyright (C) 2006-2020 Mark Russinovich
Sysinternals - www.sysinternals.com
```

```
\Sessions\1\BaseNamedObjects
  Type: Directory
  RW Window Manager\DWM-1
  RW NT AUTHORITY\SYSTEM
  RW DESKTOP-SVVLOTP\aione
  RW DESKTOP-SVVLOTP\aione-S-1-5-5-0-841005
  RW BUILTIN\Administrators
  R  Everyone
     NT AUTHORITY\RESTRICTED
```

Windows также поддерживает расширенные Ex-версии (от слова Extended) API-функций — `CreateEventEx`, `CreateMutexEx`, `CreateSemaphoreEx`, — которые добавляют еще один аргумент для указания маски доступа. Это дает возможность приложениям правильно использовать списки управления избирательным доступом (discretionary access control lists, DACL) для защиты их объектов, не нарушая при этом их способности использовать API-функции создания объектов для открытия их дескрипторов. А почему бы клиентскому приложению просто не воспользоваться функцией OpenEvent, поддерживающей желаемый аргумент доступа? Использование API-функций открытия объекта приводит к присущим этим функциям состояниям гонки, когда речь идет об отказе в вызове открытия, то есть когда клиентское приложение попыталось открыть событие еще до того, как оно было создано. Во многих приложениях такого типа в случае отказа API-функция открытия следует за API-функцией создания. К сожалению, гарантированного способа сделать эту операцию создания атомарной, иными словами, происходящей только один раз, не существует. В действительности, API-функции создания могут выполняться сразу несколькими потоками и (или) процессами параллельно, и все они могут пытаться создать событие в одно и то же время. Эти соревновательные условия и дополнительные сложности, требуемые для того, чтобы попытаться справиться с проблемой, делают использование API-функций открытия объекта неприемлемым решением проблемы, именно поэтому вместо них должны использоваться API-функции Ex.

Сохранение объектов

Существует два вида объектов: временные и постоянные. Большинство объектов — временные. Иными словами, они существуют, пока ими пользуются, и уничтожаются, когда в них больше нет нужды. Постоянные объекты существуют до тех пор, пока не будет дана явная команда избавиться от них. Поскольку большинство объектов временные, далее в этом разделе речь пойдет о том, как диспетчер объектов реализует сохранение объектов, а именно — поддерживает существование объектов не дольше, чем они нужны, с последующим их удалением.

Поскольку всем процессам пользовательского режима для доступа к какому-либо объекту сначала нужно открыть для него дескриптор, диспетчеру объектов несложно определить, сколько из них и какие именно этим объектом пользуются. Отслеживание таких дескрипторов составляет первую часть реализации сохранения. Диспетчер объектов реализует сохранение в два этапа. Первый называется сохранением имени и управляется с помощью количества открытых дескрипторов

существующего объекта. Каждый раз, когда процесс открывает дескриптор для объекта, диспетчер объектов увеличивает счетчик открытых дескрипторов в заголовке объекта. Когда процессы заканчивают пользоваться объектом и закрывают свои дескрипторы к нему, диспетчер объектов этот счетчик уменьшает. Когда тот опускается до 0, диспетчер объектов удаляет имя объекта из своего глобального пространства имен. Это удаление не дает процессам открыть дескриптор объекта.

Второй этап сохранения объектов состоит в прекращении сохранения самих объектов (иными словами, их удалении), когда они больше не используются. Поскольку код операционной системы обычно обращается к объектам через указатели, а не дескрипторы, диспетчер объектов должен также следить за тем, сколько указателей на объект он раздал системным процессам. Как мы уже видели, выдавая *указатель на объект*, он увеличивает *счетчик указателей*. Когда компоненты уровня ядра заканчивают пользоваться указателем, они обращаются к диспетчеру объектов, чтобы тот уменьшил счетчик ссылок у объекта. Поскольку дескриптор тоже является ссылкой на объект и требует отслеживания, увеличение счетчика дескрипторов увеличивает счетчик ссылок, а уменьшение — уменьшает.

Наконец, мы упомянули *счетчик использования ссылок*, который суммирует *кэшированные ссылки* со счетчиком указателей и уменьшается каждый раз, когда процесс применяет дескриптор. Счетчик использования ссылок был добавлен в Windows 8 из соображений производительности. Когда от ядра требуют получить указатель на объект, исходя из его дескриптора, оно может это сделать, не нуждаясь в блокировке глобальной таблицы дескрипторов. Это значит, что в новых версиях Windows запись в таблице дескрипторов, описанная в разделе «Дескрипторы объектов и таблица дескрипторов процесса» ранее в данной главе, содержит счетчик использования ссылок, который инициализируется при первом использовании дескриптора соответствующего объекта приложением или драйвером уровня ядра. Заметим, что в данном контексте под «использованием» подразумевается разыменование указателя на объект исходя из дескриптора — операция, выполняемая ядром с помощью таких функций, как `ObReferenceObjectByHandle`.

Рассмотрим работу всех трех счетчиков на примере, подобном показанному на рис. 8.35. На нем изображены два объекта событий, используемых в условиях 64-разрядной системы. Процесс А создает первое событие и получает к нему дескриптор. У события есть имя, что подразумевает помещение его диспетчером объектов в корректный объект-каталог (например, `\BaseNamedObjects`) и присвоение счетчику ссылок и счетчику дескрипторов начальных значений 2 и 1 соответственно. После завершения инициализации процесс А переходит к ожиданию первого события. Эта операция позволяет ядру задействовать его дескриптор (или ссылаться на него), из-за чего счетчик ссылок использования устанавливается в 32 767 (0x7FFF в шестнадцатеричной системе, первые 15 бит равны 1). Это значение прибавляется к счетчику ссылок объекта первого события, который тоже увеличивается на 1, приходя к окончательному значению 32 770.

Процесс Б инициализируется, создает второе именованное событие и сигнализирует о нем. Последняя операция использует второе событие (то есть ссылается на него), тоже доводя его счетчик ссылок до значения 32 770. Затем процесс Б открывает первое событие, размещенное процессом А. Эта операция позволяет ядру создать новый дескриптор (действительный только в адресном пространстве процесса Б),

который увеличивает для первого события счетчики дескрипторов и ссылок, приводя их к значениям 2 и 32 771 соответственно. (Не забываем, что новая запись в таблице дескрипторов все еще не инициализировала свой счетчик ссылок использования.) Процесс Б прежде, чем сигнализировать о первом событии, трижды использует его дескриптор. Первая операция инициализирует *счетчик ссылок использования* значением 32 767. Оно суммируется со счетчиком ссылок объекта, который увеличивается еще на 1 и достигает общего значения 65 539. Последующие операции над дескриптором просто *понижают* счетчик ссылок использования, не трогая счетчик ссылок объекта. Когда ядро заканчивает использовать объект, оно, однако, всегда уничтожает ссылку на его указатель, — действие, уменьшающее счетчик ссылок на объект. Таким образом, после четырех использований (включая операцию сигнализации) первый объект имеет счетчик дескрипторов 2 и счетчик ссылок 65 535. Кроме того, на первое событие ссылается некая структура режима ядра, что в итоге доводит счетчик ссылок до значения 65 536.

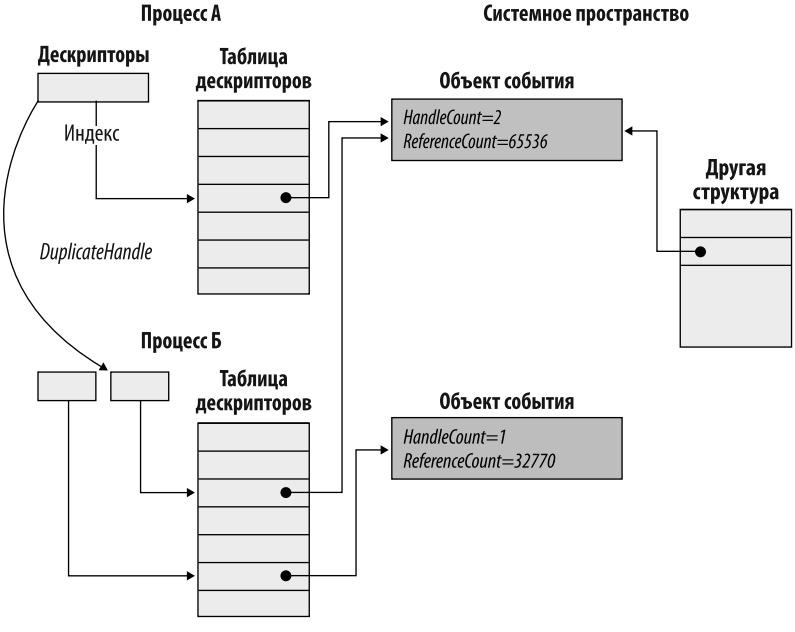


Рис. 8.35. Счетчики дескрипторов и ссылок

Когда процесс закрывает дескриптор от объекта (эта операция приводит к вызову ядром функции `NtClose`), диспетчер объектов знает, что ему нужно вычесть счетчик ссылок использования дескриптора из счетчика ссылок объекта. Это позволяет корректно отозвать дескриптор. В приведенном примере, даже если процессы А и Б закроют свои дескрипторы к первому объекту, тот продолжит существовать, поскольку его счетчик ссылок станет равным 1 (при этом счетчик дескрипторов будет 0). Однако, когда процесс Б закроет свой дескриптор от объекта второго события, этот объект будет уничтожен, поскольку его счетчик ссылок достигнет значения 0.

Подобное поведение означает, что, даже когда счетчик открытых дескрипторов объекта достигает 0, счетчик ссылок на объект может оставаться положительным, то есть система все еще как-то его использует. В конечном счете диспетчер объектов удаляет объект из памяти, только когда его счетчик ссылок достигает 0. Это удаление должно следовать определенным правилам, а в некоторых случаях требует дополнительного участия вызвавшего процесса. Например, поскольку объекты могут присутствовать как в подкачиваемом, так и в статичном пуле памяти (в зависимости от настроек, указанных в характеристиках типа объекта), если отзыв происходит на уровне IRQL `DISPATCH_LEVEL` или выше и приводит к снижению счетчика ссылок до 0, при попытке немедленно освободить память из-под объекта в подкачиваемом пуле происходит отказ системы (напомним, что подобная операция нелегальна, так как ошибку доступа к странице некому будет обработать). При таком сценарии диспетчер объектов выполняет *операцию отложенного удаления*, помещая ее в рабочий поток, действующий на пассивном уровне (IRQL 0). Мы опишем системные рабочие потоки более подробно далее в этой главе.

Другой сценарий, требующий отложенного удаления, возникает в ходе работы с объектами диспетчера транзакций ядра (Kernel Transaction Manager, KTM). В некоторых ситуациях определенные драйверы могут удерживать блокировку, связанную с этим объектом, и попытка его удалить приведет к тому, что система попытается эту блокировку получить. Однако сам драйвер может и вовсе не получить шанса ее снять, что приведет к взаимной блокировке. Работая с объектами KTM, разработчики драйверов должны использовать функцию `ObDereferenceObjectDeferDelete`, чтобы принудительно применять отложенное удаление вне зависимости от уровня IRQL. Наконец, диспетчер ввода/вывода тоже задействует этот механизм в целях оптимизации, тем самым позволяя некоторым операциям завершаться быстрее, вместо того чтобы дожидаться, пока диспетчер объектов удалит объект.

Из-за особенностей того, как работает сохранение объектов, приложение может обеспечить наличие объекта и его имени в памяти, просто поддерживая его дескриптор открытым. Программистам, которые пишут приложения, состоящие из двух и более взаимодействующих между собой процессов, не нужно беспокоиться о том, что один процесс может удалить какой-то объект прежде, чем другой закончит им пользоваться. Кроме того, закрытие дескрипторов объекта, принадлежащих приложению, не приведет к удалению самого объекта, если операционная система еще не закончила с ним работать. Например, один процесс может создать второй для выполнения какой-нибудь программы в фоне, а затем немедленно закрыть свой дескриптор от него. Поскольку операционной системе второй процесс нужен для исполнения программы, она сохраняет ссылку на объект ее процесса. Лишь когда фоновая программа закончит исполняться, диспетчер объектов понизит второму процессу счетчик ссылок и удалит его.

Утечки памяти, связанные с объектами, могут представлять опасность для системы, вызывая утечки пула памяти ядра и в итоге вызывая истощение общесистемной памяти (может прекратить работу приложений с трудно выявляемыми причинами). Поэтому в Windows предусмотрен ряд отладочных механизмов, которые могут быть активированы для отслеживания, анализа и отладки проблем с дескрипторами и объектами. Кроме того, для WinDbg существует два расширения, которые используются в целях графического анализа. Они описаны в табл. 8.24.

Таблица 8.24. Механизмы отладки для дескрипторов объектов

Механизм	Чем активируется	Расширение отладчика ядра
Handle Tracing Database (База данных отслеживания дескрипторов)	Общесистемным средством отслеживания стека ядра — Kernel Stack Trace и (или) установкой для каждого процесса с помощью средства Gflags.exe флага отслеживания пользовательского стека — User Stack Trace	!htrace <значение дескриптора> < ID процесса>
Object Reference Tracing (Отслеживание ссылок на объект)	Установка в режиме отслеживания ссылок на объект — Object Reference Tracing, флагов для имени конкретного процесса (или имен конкретных процессов) или флагов тега (тегов) пула конкретного типа объекта с помощью Gflags.exe»	!obtrace <указатель на объект>
Object Reference Tagging (Тегирование ссылок на объект)	Соответствующие API-функции должны вызываться драйверами	Не определено

Включение базы данных отслеживания дескрипторов пригодится при попытке разобраться с использованием каждого описателя внутри контекста приложения или системы. Расширение отладчика `!handle` может показать моментальный снимок трассировки стека на время открытия конкретного дескриптора. После обнаружения утечки памяти, связанной с дескриптором, трассировка стека может точно определить код, создающий дескриптор, который можно проанализировать на предмет пропущенного вызова такой функции, как `CloseHandle`.

Расширение `!objtrace`, позволяющее отслеживать ссылки на объект, дает возможность вести более широкое наблюдение, показывая трассировку стека для каждого вновь созданного дескриптора, а также при каждой ссылке на дескриптор из ядра (а также при каждом ее открытии, дублировании или наследовании) и при каждом удалении ссылки. При каждом анализе подобных схем появляется возможность более легкой отладки кода при неправильном использовании объекта на системном уровне. Кроме того, такое отслеживание ссылок дает возможность разобраться в поведении системы при работе с конкретными объектами. К примеру, при отслеживании процессов выводятся ссылки от всех, имеющихся в системе драйверов, которые зарегистрировали уведомительные функции обратного вызова (как, например, Process Monitor). Это помогает выявить неконтролируемые или некачественные драйверы сторонних производителей, которые могут ссылаться на дескрипторы в режиме ядра, никогда не удаляя ссылок на них.

ПРИМЕЧАНИЕ При включении отслеживания ссылок на объект для конкретного типа объекта вы можете узнать имя его тега пула по полю key структуры данных OBJECT_TYPE, воспользовавшись командой dx. У каждого объекта типа в системе есть глобальная переменная со ссылкой на эту структуру, например PsProcessType. Как альтернативу можно применять команду !object, которая вернет указатель на структуру.

В отличие от предыдущих двух механизмов, тегирование ссылок на объект не является функцией отладки, которую можно включить с помощью глобальных флагов или отладчика, а представляет собой набор API-функций, которые разработчики драйверов устройств должны использовать для установки и удаления ссылок на объекты, включающий такие функции, как `ObReferenceObjectWithTag` и `ObDereferenceObjectWithTag`. Аналогично тегированию пулов (более подробно об этом рассказано в главе 5 тома 1) эти API-функции позволяют разработчикам предоставлять каждой паре установки-удаления ссылки для ее идентификации тег, состоящий из четырех символов. При использовании только что рассмотренного расширения `!objtrace` показывается также тег для каждой операции установки и удаления ссылки, что позволяет использовать в качестве механизма идентификации мест утечки памяти или подчиненных ссылок не только стек вызовов; это особенно важно, когда заданный вызов осуществляется драйвером тысячи раз.

Учет ресурсов

Учет ресурсов, как и сохранение объектов, тесно связан с использованием дескрипторов объектов. Положительный счетчик открытых дескрипторов указывает на то, что какой-то процесс пользуется данным ресурсом. Кроме того, это значит, что процесс платит квотой за память, занимаемую этим объектом. Когда счетчики дескрипторов и ссылок объекта достигают 0, процесс, который им пользовался, освобождается от связанных с этим обязательств.

Множество операционных систем ограничивают доступ процессов к ресурсам системы с помощью системы квот. Однако типы квот, накладываемые на процессы, порой разнообразны и сложны, а код для их отслеживания разбросан по всей операционной системе. Например, в некоторых ОС компонент ввода/вывода может отслеживать и ограничивать количество файлов, которые один процесс может открыть, когда компонент, ответственный за память, может устанавливать предельный объем памяти, выделяемый себе потоками в рамках процесса. Компонент, управляющий процессами, может ограничивать пользователя в количестве новых процессов, которые он может запустить, и максимуме потоков в рамках процесса. Каждое из этих ограничений отслеживается и используется в различных частях операционной системы.

В отличие от этого, диспетчер объектов Windows предоставляет централизованный учет ресурсов. В каждом заголовке объекта есть атрибут под названием *квота* (quota charges), в котором записывается, сколько диспетчером объектов вычитается ресурсов из выделенной процессу квоты пулов выгружаемой и (или) невыгружаемой памяти, когда поток в процессе открывает дескриптор объекта.

Каждый процесс в Windows указывает на структуру квот, где записаны лимиты и текущие значения используемых невыгружаемого пула, выгружаемого пула и виртуальной памяти. По умолчанию эти квоты имеют значение 0 (без ограничений), но могут быть указаны путем изменения значений реестра. (Для этого нужно дополнить или отредактировать параметры `NonPagedPoolQuota`, `PagedPoolQuota` и `PagingFileQuota` по адресу `HKLM\SYSTEM\CurrentControlSet\Control\Session Manager\Memory Management`.) Следует заметить, что все процессы в интерактивном сеансе

работы используют общий блок квот, и документированного способа создания процессов со своими собственными блоками квот не существует.

Именование объектов

Важной проблемой, о которой следует помнить при создании множества объектов, является необходимость организовать эффективную систему их отслеживания. Чтобы помочь вам сделать это, диспетчеру объектов потребуется следующая информация:

- способ, позволяющий отличить объекты друг от друга;
- метод для поиска и получения конкретного объекта.

Первое требование обеспечивается возможностью давать объектам имена. Почти в любой операционной системе есть подобная функциональность — возможность именовать выбранные ресурсы, файлы, каналы или, к примеру, разделяемый блок памяти. Исполнительная система, в свою очередь, позволяет дать имя любому ресурсу, который можно представить с помощью объекта. Второе требование — поиск и получение объекта — обеспечивается также с помощью имен объектов. Храня объекты именованными, диспетчер может искать и по имени.

Кроме того, именование объектов позволяет обеспечить и третье требование, а именно возможность совместного доступа к объекту разных процессов. Пространство имен объектов исполнительной системы является глобальным и видно всем процессам в системе. Один процесс может создать объект и поместить его имя в глобальном пространстве имен, а другой — открыть дескриптор для этого объекта, назвав его по имени. Если же публиковать объект подобным образом не требуется, создатель может не давать ему имя.

Из соображений эффективности диспетчер объектов не выполняет поиск объекта по имени каждый раз, когда кто-то его использует. Поиск происходит только в двух случаях. Первый имеет место, когда процесс создает именованный объект: диспетчер объектов выполняет поиск по имени, чтобы проверить, что его еще не существует, прежде чем добавлять новый элемент в глобальное пространство имен. Второй случай наступает, когда процесс открывает дескриптор для именованного объекта: диспетчер объектов ищет объект по имени, находит его и возвращает вызвавшей стороне дескриптор, а тот в дальнейшем обращается к объекту с помощью дескриптора. Диспетчер объектов позволяет вызвавшему выбирать как зависимый, так и независимый от регистра поиск. Эта возможность поддерживает подсистему Windows для Linux (WSL) и другие среды, где используются зависимые от регистра имена файлов.

Каталоги объектов

Объект каталога объектов — это инструмент, с помощью которого диспетчер объектов обеспечивает иерархическую структуру именования. Такой объект аналогичен каталогу в файловой системе и содержит имена других объектов, в том числе даже объектов-каталогов. Сам он хранит достаточно информации для того, чтобы преобразовать эти имена в указатели на заголовки самих объектов. Эти указатели диспетчер объектов применяет для дескрипторов объектов, которые он возвращает

клиентам из пользовательского режима. Код как режима ядра (в том числе компоненты исполнительной системы и драйверы устройств), так и пользовательского режима (в частности, подсистемы) может создавать каталоги объектов и хранить там другие объекты.

Объекты могут помещаться в любом месте пространства имен, но для некоторых типов они всегда появляются в определенных каталогах из-за того, что создаются специализированным компонентом в особом порядке. Например, диспетчер ввода/вывода создает каталог объектов по имени `\Driver`, в котором содержатся имена объектов, представляющих драйверы режима ядра, не относящиеся к файловой системе. Поскольку этот диспетчер является единственным компонентом, способным создавать объекты драйверов (с помощью функции `IoCreateDriver`), объектов другого типа там находиться не должно.

В табл. 8.25 стандартные каталоги объектов, находящиеся во всех системах Windows, и те типы объектов, чьи имена в них хранятся. Из всего перечисленного только `\AppContainerNamedObjects`, `\BaseNamedObjects` и `\Global??` доступны для стандартных приложений Win32 и UWP, опирающихся на документированные API (более подробно — в пункте «Пространство имен сеанса» далее).

Таблица 8.25. Каталоги стандартных объектов

Каталог	Типы хранящихся в нем объектов
\AppContainerNamedObjects	Существует только в рамках каталога объектов \Session для интерактивных сеансов (не нулевых), содержит именованные объекты ядра, созданные функциями API Win32 или UWP внутри процессов, работающих в AppContainer
\ArcName	Символические ссылки, отображающие пути в ARC-стиле на пути в NT-стиле
\BaseNamedObjects	Объекты глобальных мьютексов, событий, семафоров, таймеров ожидания, заданий, ALPC-портов, символических ссылок и разделов
\Callback	Объекты обратных вызовов (создавать их могут только драйверы)
\Device	Объекты устройств, принадлежащих большинству драйверов, кроме файловой системы и устройств диспетчера фильтров. Вдобавок к этому событие VolumesSafeForWriteAccess и отдельные символические ссылки, такие как SystemPartition и BootPartition. Кроме того, содержит объект раздела PhysicalMemory, дающий компонентам ядра прямой доступ к оперативной памяти. Наконец, здесь же располагаются некоторые каталоги объектов, такие как Http, используемый драйвером-ускорителем Http.sys, и каталоги вида HarddiskN для каждого физического диска
\Driver	Объекты драйверов, чей тип не «драйвер файловой системы» (SERVICE_FILE_SYSTEM_DRIVER) или «определитель файловой системы» (SERVICE_RECOGNIZER_DRIVER)
\DriverStore(s)	Символические ссылки для местоположений, откуда можно устанавливать драйверы и управлять ими. В типичных случаях там как минимум есть SYSTEM, указывающая на \SystemRoot, но на устройствах с Windows 10X могут быть и другие

Продолжение ⇩

Таблица 8.25 *(продолжение)*

Каталог	Типы хранящихся в нем объектов
\FileSystem	Объекты драйверов файловой системы (SERVICE_FILE_SYSTEM_DRIVER) и определителей файловой системы (SERVICE_RECOGNIZER_DRIVER), а также объекты устройства. Кроме того, диспетчер фильтров создает свои объекты устройства в каталоге объектов Filters
\GLOBAL??	Объект — символическая ссылка, представляющая имена устройств MS-DOS. (Каталоги вида \Sessions\0\DosDevices\<LUID>\Global являются символическими ссылками на этот каталог)
\KernelObjects	Содержит объекты событий, сигнализирующих о состоянии ресурсов пула ядра, завершении определенных задач операционной системы, а также объекты сеанса (по крайней мере Session0), представляющие каждый интерактивный сеанс, и объекты раздела (по крайней мере MemoryPartition0) для каждого раздела памяти. Также хранит мьютекс, используемый для синхронизации доступа к базе данных конфигурации загрузки (BC). Наконец, содержит динамические символические ссылки, использующие пользовательский обратный вызов для ссылки на правильный раздел для физической памяти и условий фиксации ресурсов, а также для обнаружения ошибок памяти
\KnownDlls	Объекты разделов для известных DLL, размеченных SMSS при запуске системы, и символическая ссылка на путь, где эти DLL хранятся
\KnownDlls32	Если установлена 64-разрядная Windows, \KnownDLLs содержит платформенно-зависимые 64-разрядные образы, а данный каталог — 32-разрядные версии DLL для применения в WoW64
\NLS	Объекты разделов для отображения таблиц поддержки национальных языков
\ObjectTypes	Объекты типа для всех типов объектов, созданных с помощью ObCreateObjectTypeEx
\RPC Control	Порты ALPC, созданные для представления конечных точек удаленного вызова процедур (RPC) в условиях применения локальных RPC (ncalrpc). Сюда входят как однозначно именованные точки, так и автоматически сгенерированные имена портов COM (OLEXXXXX) и неименованных портов (LRPC-XXXX, где XXXX — случайно сгенерированное шестнадцатеричное значение)
\Security	Порты ALPC и события, используемые специфическими объектами в рамках системы безопасности
\Sessions	Каталог пространств имен сеансов (см. следующий пункт)
\Silo	Если был создан хотя бы один Windows Server Container с помощью Docker для Windows или не-VM-контейнеров, содержит каталоги объектов для каждого ID хранилища (ID корневого задания в этом контейнере), в которых располагаются пространства имен объектов в рамках соответствующих хранилищ
\UMDFCommunicationPorts	Порты ALPC, применяемые платформой для разработки драйверов пользовательского режима (User-Mode Driver Framework, UMDF)

Каталог	Типы хранящихся в нем объектов
\VmSharedMemory	Объекты раздела, используемые виртуализованными экземплярами Win32k.sys и других компонентов диспетчера окон на устройствах с Windows 10X при запуске устаревших приложений Win32. Кроме того, содержит каталог объектов Host, представляющий вторую сторону соединения
\Windows	Порты ALPC подсистемы Windows, совместно используемые разделы и объекты типа Windows Station, хранимые в каталоге WindowsStations. Кроме того, диспетчер окон Рабочего стола (DWM) хранит здесь свои порты ALPC, события и совместно используемые разделы для сеансов из числа ненулевых. Наконец, здесь находится объект раздела для службы тем Рабочего стола

Имена объектов являются глобальными в рамках отдельно взятого компьютера (или для всех процессоров в мультипроцессорном компьютере), но невидимы через сеть. Однако метод анализа диспетчера объектов делает возможным доступ к именованным объектам, существующим на других компьютерах. Например, диспетчер ввода/вывода, который предоставляет службы для работы с файловыми объектами, расширяет возможности диспетчера объектов на файлы с удаленных ресурсов. При попытке доступа к объекту удаленного файла диспетчер вызывает метод анализа, который позволяет диспетчеру ввода/вывода перехватить запрос и доставить его средству сетевого перенаправления — драйверу, который обращается к файлам через сеть. Серверный код на удаленной системе Windows вызывает диспетчеры объектов и ввода/вывода на этой системе, чтобы найти файловый объект и вернуть нужную информацию отправителю.

Поскольку все объекты ядра, созданные процессами вне Appcontainer посредством API Win32 или UWP, такие как мьютексы, события, семафоры, таймеры ожидания и разделы, хранятся в одном каталоге объектов, никакие два из них не могут иметь одно и то же имя, даже если они относятся к объектам разных типов. Из-за этого ограничения нужно осторожно выбирать имена, чтобы избежать конфликтов с другими именами. Например, вы можете снабжать имена префиксом в виде GUID и/или комбинировать имя с пользовательским индикатором безопасности (SID), но даже тогда это поможет лишь в условиях запуска одного экземпляра приложения на одного пользователя.

Проблема конфликта имен может показаться безобидной, но главным нюансом безопасности, о котором не следует забывать, является возможность *сквоттинга*, то есть *незаконного присвоения имени объекта*. Хотя объекты в различных сеансах защищены друг от друга, внутри пространства имен текущего сеанса не существует стандартной защиты, которая могла бы быть установлена с помощью стандартных API-функций Windows. Это дает возможность непривилегированным приложениям запускаться в одном сеансе с привилегированными приложениями для доступа к их объектам, как было рассмотрено ранее в подразделе, посвященном безопасности объектов. К сожалению, даже если создатель объекта использует для защиты объекта правильный DACL-список, это не спасает от сквоттинг-атаки, при которой непривилегированное приложение создает объект до привилегированного приложения, отказывая таким образом в доступе законному приложению.

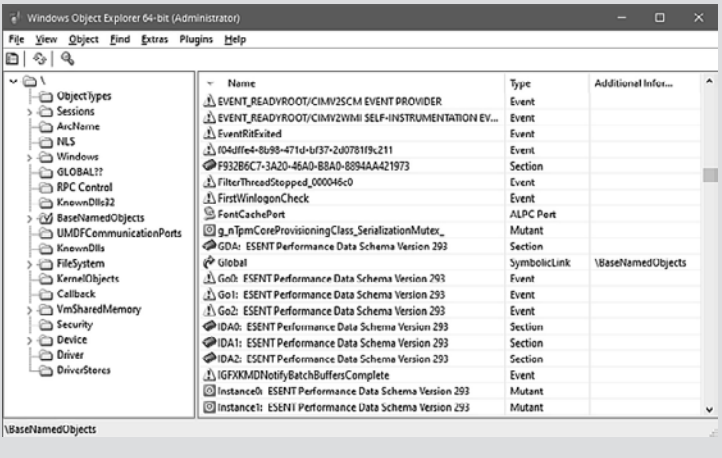
Уменьшить эту проблему позволяет существующая в Windows концепция *защищенного пространства имен*. Она позволяет приложениям пользовательского режима

создавать каталоги объектов с помощью API-функции `CreatePrivateNamespace` и связывать эти каталоги с дескрипторами границ, создаваемыми с помощью функции `CreateBoundaryDescriptor` и являющимися специальными структурами данных, защищающими каталоги. Таким образом, привилегированное приложение может быть уверено, что непривилегированные приложения не смогут проводить атаку отказа от обслуживания в отношении этих объектов, что не остановит привилегированные приложения от совершения таких же действий (но это спорный момент). Кроме того, дескриптор границы может также содержать уровень целостности, основанный на уровне целостности процесса, защищая объекты, которые, возможно, принадлежат той же самой учетной записи пользователя, что и приложение (подробное описание уровней целостности см. в главе 7 тома 1).

Одним из факторов, делающих дескрипторы границ эффективной защитой от скваттинга, является то, что, в отличие от объектов, создатель такого дескриптора должен иметь доступ к нему (через SID или уровень целостности). Это значит, что приложение без привилегий сможет создать лишь дескриптор без привилегий. Аналогично, когда приложение желает открыть объект в защищенном пространстве имен, оно должно открыть это пространство с помощью такого дескриптора границ, с которым то было создано. В свою очередь, привилегированное приложение будет обладать привилегированным дескриптором, который не совпадет с тем, который создало приложение без привилегий.

ЭКСПЕРИМЕНТ. Просмотр основных именованных объектов и закрытых объектов

Вы можете просмотреть список основных объектов, имеющих имена, с помощью инструмента WinObj из набора Sysinternals или WinObjEx64. В этом эксперименте мы пользуемся вторым, потому что он поддерживает дополнительные типы объектов и может показывать закрытые пространства имен. Запустите Winobjex64.exe и щелкните на узле дерева BaseNamedObjects, как показано далее.

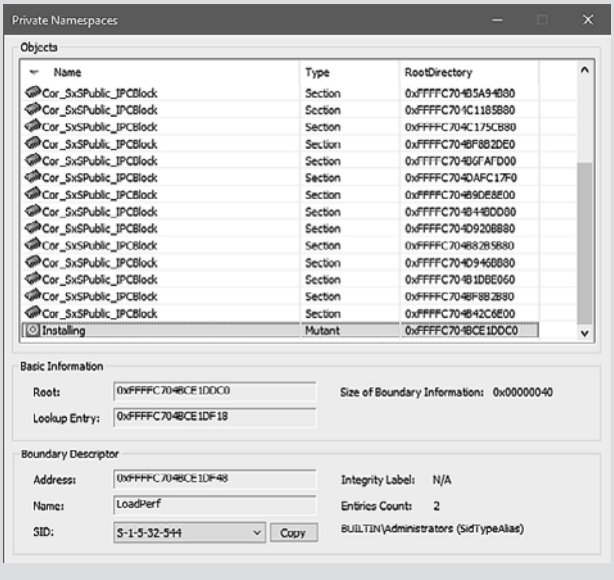


Именованные объекты перечисляются в правой части окна. Значки показывают тип объекта:

- мьютексы обозначены стоп-сигналом;
- разделы (объекты Windows для просмотра файлов) показаны как микросхемы памяти;
- события отмечены восклицательным знаком;
- семафоры показаны значком, похожим на светофор;
- символические ссылки обозначены изогнутой стрелкой;
- объекты-каталоги показаны в виде папки;
- порты ALPC отмечены значком в виде электрической вилки;
- таймеры показаны часами.

Прочие значки используются для остальных типов объектов.

Теперь откройте меню Дополнительно (Extras) и выберите Закрытые пространства имен (Private Namespaces). Отобразится список наподобие приведенного далее.

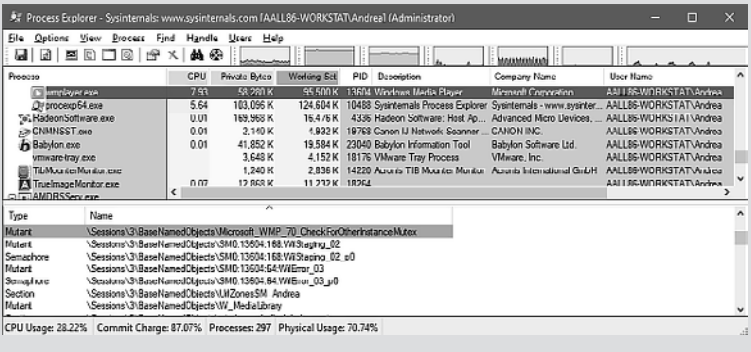


Для каждого объекта вы сможете увидеть имя дескриптора границы (например, мьютекс Installing ограничен дескриптором LoadPerf) и связанный с ним уровень целостности (в данном случае конкретной целостности не назначено, а SID входит в группу «Администраторы»). Обратите внимание на то, что для работы с данной функциональностью на компьютере должна быть активна отладка ядра, поскольку WinObjEx64 использует драйвер локальной отладки ядра WinDbg для чтения памяти ядра.

ЭКСПЕРИМЕНТ. ВМЕШАТЕЛЬСТВО В ПРОЦЕСС ИСПОЛЬЗОВАНИЯ ЕДИНСТВЕННОГО ЭКЗЕМПЛЯРА

Приложения наподобие Проигрывателя Windows Media и те, что входят в Microsoft Office, могут послужить ярким примером принудительного использования единственного экземпляра посредством именованных объектов. Обратите внимание на то, что при запуске исполняемого файла Wmplayer.exe сам Проигрыватель Windows Media появляется лишь единожды — все остальные запуски будут лишь возвращать окну фокус. С помощью Process Explorer вы можете изменить список дескрипторов так, чтобы превратить компьютер в медиамикшер.

1. Запустите Проигрыватель Windows Media и Process Explorer, в котором пройдите к таблице дескрипторов (в меню Вид (View) выберите Вид нижней панели (Lower Pane View), а потом — Дескрипторы (Handles)). Там вы увидите дескриптор Microsoft_WMP_70_CheckForOtherInstanceMute:



2. Щелкните на дескрипторе правой кнопкой мыши и выберите Закрыть дескриптор (Close handle). Подтвердите действие после предупреждения. Обратите внимание: чтобы иметь возможность закрывать дескрипторы других процессов, Process Explorer следует запустить с правами администратора.
3. Снова запустите Проигрыватель Windows Media. Обратите внимание на то, что теперь запустился второй процесс.
4. Проиграйте в каждом экземпляре разные песни. Вы также можете использовать микшер из области уведомлений (щелкните на значке Громкость), чтобы выбрать, какой из процессов будет звучать громче, фактически создав микшер.

Вместо закрытия дескриптора именованного объекта некое приложение могло бы запуститься до Проигрывателя Windows Media и создать объект с таким же именем. В этом случае Проигрыватель не запустится никогда, поскольку решит, что его экземпляр уже работает в системе.

Символические ссылки

В некоторых файловых системах (к примеру, NTFS, Linux, MacOS) символические ссылки позволяют пользователю создать имя файла или папки таким образом, что при обращении к ним операционная система преобразует их в другое имя файла или

папки. Применение символических ссылок — это простой способ позволить пользователям косвенно иметь общий доступ к файлам или содержимому каталога, создавая перекрестную ссылку между каталогами в рамках обычно иерархической структуры.

Диспетчер объектов реализует объект, который называется *объектом символической ссылки* и выполняет аналогичные функции для имен объектов в своем пространстве имен объектов. Символическая ссылка может встретиться в любом месте строки с именем объекта. Когда вызывающий обращается по имени объекта символической ссылки, диспетчер просматривает свое пространство имен, пока не обнаружит искомый объект. Внутри него он находит строку, которая заменит имя символической ссылки. После этого поиск по имени начинается заново.

Одним из случаев, когда исполнительная система задействует объекты символических ссылок, является преобразование имен устройств в стиле MS-DOS во внутренние имена устройств Windows. Пользователь Windows обычно обращается к жестким дискам по именам C:, D: и т. д., а к последовательным — COM1, COM2 и т. п. Подсистема Windows создает эти объекты символических ссылок и помещает их в пространство имен диспетчера объектов в каталог `\Global??`, что можно проделать и с дополнительными буквами для дисков, воспользовавшись функцией `DefineDosDevice`.

В некоторых случаях цель, стоящая за символической ссылкой, нестатична и может зависеть от контекста обратившегося. Например, в старых версиях Windows в каталоге `\KernelObjects` имелось событие под названием `LowMemoryCondition`, но после появления разделов памяти (описание см. в главе 5 тома 1) условие срабатывания события стало зависеть от того, в каком разделе обратившийся работает и какие из них может видеть. Таким образом, теперь для каждого раздела памяти есть свое событие `LowMemoryCondition`, а его клиенты должны быть перенаправлены на экземпляр, соответствующий их разделу. Это достигается благодаря особому флагу на объекте, отсутствию строки с целью и наличию обратного вызова символической ссылки, связанная с которым функция вызывается каждый раз, когда ссылку анализирует диспетчер объектов. С помощью `WinObjEx64` вы сможете увидеть зарегистрированный обратный вызов, как показано на рис. 8.36. (Также можете воспользоваться отладчиком, запустив команду `!object\KernelObjects\LowMemoryCondition`, а затем выгрузив структуру `_OBJECT_SYMBOLIC_LINK` с помощью команды `dx`.)

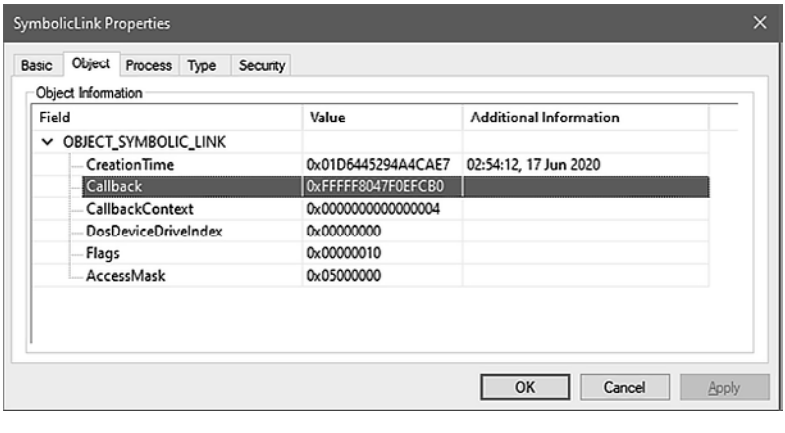


Рис. 8.36. Перенаправляющий обратный вызов символической ссылки LowMemoryCondition

Пространство имен сеанса

Службы имеют полный доступ к *глобальному* пространству имен, которое служит в качестве первого экземпляра пространства имен. Обычные пользовательские приложения имеют к нему доступ на чтение и запись, но не на удаление (за некоторыми исключениями, о которых чуть позже). Интерактивные пользовательские сеансы, в свою очередь, получают закрытое для сеанса отображение пространства имен, известное как *локальное* пространство имен. Оно предоставляет полный доступ на чтение и запись к базовым именованным объектам для всех приложений, запущенных в рамках данного сеанса, а также применяется для изоляции некоторых специфических объектов подсистемы Windows, все еще обладающих привилегиями. В число участков пространства имен, локальный экземпляр которых создается для каждого сеанса, входят `\DosDevices`, `\Windows`, `\BaseNamedObjects` и `\AppContainerNamedObjects`.

Создание отдельных копий одних и тех же участков пространства имен называется *инстанцированием* пространства имен. Применение его к `\DosDevices` позволяет каждому пользователю иметь собственный набор букв для сетевых дисков и объектов Windows наподобие последовательных портов. В Windows глобальный каталог `\DosDevices` называется `\Global??`, и это каталог, на который он ссылается, при этом локальные `\DosDevices` различаются по ID сеанса авторизации пользователя.

Каталог `\Windows` служит местом, куда Win32k.sys помещает интерактивный объект Windows Station, создаваемый Winlogon, — `\WinSta0`. Среда служб терминалов способна поддерживать по несколько интерактивных пользователей, но каждый из них будет нуждаться в индивидуальной версии `WinSta0`, чтобы сохранить иллюзию, будто он обращается к предопределенному экземпляру интерактивного Windows Station. Наконец, обычные приложения Win32 создают разделяемые объекты в `\BaseNamedObjects`, в том числе события, мьютексы и разделы в памяти. Если два пользователя запустят приложение, создающее именованный объект, сеансу каждого из них потребуется закрытая версия этого объекта, чтобы два экземпляра приложения не вмешивались в работу друг друга, пытаясь обращаться к одному и тому же объекту. Однако, если приложение Win32 работает в AppContainer или же это приложение UWP, механизмы поддержания «песочницы» не позволят ему получить доступ к `\BaseNamedObjects`. Вместо этого запрос будет перенаправлен каталогу `\AppContainerNamedObjects`, где находятся подкаталоги, имена которых соответствуют SID пакета AppContainer (подробности об AppContainer и модели «песочницы» Windows см. в главе 7 тома 1).

Диспетчер объектов реализует локальное пространство имен, создавая закрытые версии четырех упомянутых ранее каталогов по адресу, соответствующему сеансу пользователя в `\Sessions\`*`n`* (где *n* — идентификатор сеанса). Например, когда приложение Windows из удаленного сеанса 2 создает именованное событие, подсистема Win32 как участник функции `BaseGetNamedObjectDirectory` из Kernelbase.dll неявно перенаправляет имя объекта с `\BaseNamedObjects` на `\Sessions\2\BaseNamedObjects`, а в случае с AppContainer — на `\Sessions\2\AppContainerNamedObjects\<PackageSID>\`.

Еще одним способом доступа к именованным объектам является использование защитной функции под названием *«изоляция базовых именованных объектов»* (Base Named Object, BNO). Процесс-родитель может запустить потомка с атрибутом

ProcThreadAttribute BnoIsolation (подробности об атрибутах запуска процесса см. в главе 3 тома 1), передав там персональный префикс для каталога объектов. Это приведет к тому, что Kernelbase.dll создаст этот каталог и начальный набор объектов (например, символические ссылки) для его поддержки, после чего с помощью функции NtCreateUserProcess установит префикс и связанные с этим начальные дескрипторы в объекте-токене процесса-потомка (конкретно — в поле BnoIsolationHandlesEntry) через данные в платформенно-зависимой версии атрибута процесса.

В дальнейшем функция BaseGetNamedObjectDirectory запрашивает объект-токен, чтобы проверить, действует ли изоляция BNO, и если это так, добавляет префикс оттуда во все операции с именованными объектами. Например, \Sessions\2\BaseNamedObjects станет \Sessions\2\BaseNamedObjects\IsolationExample. Это позволяет создать подобие «песочницы» для процесса, не используя функциональность AppContainer.

Все функции диспетчера объектов, связанные с управлением пространством имен, в курсе наличия инстанцированных каталогов и принимают участие в поддержании иллюзии, будто бы все сеансы делят одно пространство имен. DLL подсистемы Windows дополняют имена в каталоге \DosDevices, передаваемые от приложений Windows, префиксами вида \?? (например, C:\Windows становится \??\C:\Windows). Когда диспетчер объектов сталкивается с особым префиксом \??, предпринимаемые им шаги зависят от версии Windows, но в любом случае он опирается на поле DeviceMap в исполнительном объекте процесса (EPROCESS, более подробно описан в главе 3 тома 1), которое указывает на структуру данных, разделяемую с прочими процессами в том же сеансе.

Поле DosDevicesDirectory структуры DeviceMap указывает на каталог диспетчера объектов, представляющий локальную для процесса копию \DosDevices. Когда диспетчер встречает ссылку на \??, он обнаруживает локальный для процесса каталог \DosDevices, используя поле DosDevicesDirectory из DeviceMap. Если же диспетчер не находит в этом каталоге исходного объекта, он проверяет поле DeviceMap объекта каталога. Если оно валидно, он ищет нужный объект в каталоге по ссылке из поля GlobalDosDevicesDirectory структуры DeviceMap, которое всегда указывает на \Global\??.

При определенных обстоятельствах чувствительные к сеансу приложения требуют доступа к объектам из глобального сеанса, даже если они сами работают в другом сеансе. Приложение может нуждаться в этом для синхронизации с другими своими экземплярами, запущенными в других удаленных сеансах или из консольного (то есть нулевого) сеанса. Для таких случаев диспетчер объектов предоставляет особое перенаправление \Global, которым приложение может предварить имя любого объекта, чтобы обеспечить доступ в глобальное пространство имен. Например, приложение из сеанса 2, открывая объект \Global\ApplicationInitialize, перенаправляется в \BaseNamedObjects\ApplicationInitialized вместо \Sessions\2\BaseNamedObjects\ApplicationInitialized.

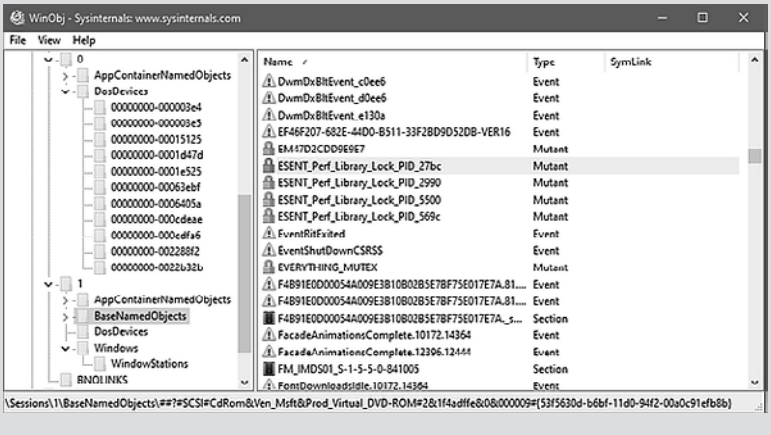
Любое приложение, желающее получить доступ к глобальному каталогу \DosDevices, не нуждается в префиксе \Global, покуда объект отсутствует в его локальном каталоге \DosDevices. Так происходит потому, что диспетчер объектов, не найдя искомого объекта в локальном каталоге, автоматически продолжает поиск в глобальном. Тем не менее приложение может принудительно потребовать поиска в глобальном каталоге, обратившись по \GLOBALROOT.

Каталоги сеансов изолированы друг от друга, но, как упоминалось ранее, обычные пользовательские приложения могут создавать глобальные объекты с помощью префикса `\Global`. Однако существует важное защитное ограничение: объекты сеансов и символических ссылок не могут быть созданы глобально, если только вызывающий не работает в нулевом сеансе или не обладает особой привилегией под названием *«создание глобальных объектов»* (create global object), исключая случаи, когда имя этого объекта входит в авторизованный список небезопасных имен, хранящийся в реестре по адресу HKLM\SYSTEM\CurrentControlSet\Control\Session Manager\kernel в параметре `ObUnsecureGlobalNames`. По умолчанию к этим именам обычно относятся такие строки:

- `netfxcustomperfcounters.1.0`;
- `SharedPerfIPCBlock`;
- `Cor_Private_IPCBlock`;
- `Cor_Public_IPCBlock_`.

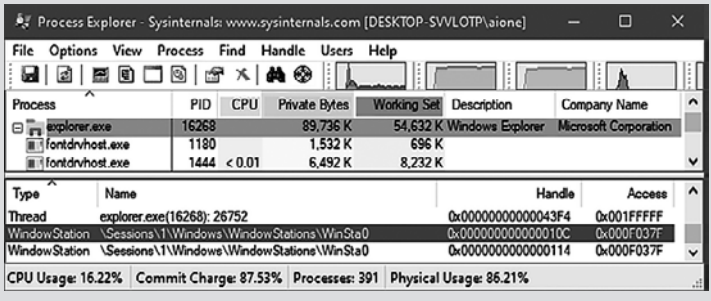
ЭКСПЕРИМЕНТ. Просмотр экземпляров пространства имен

Вы можете наблюдать разделение между пространством имен нулевого сеанса и таковыми у других сеансов сразу после входа в систему. Причина этого в том, что первый консольный пользователь авторизуется в сеанс 1 (службы работают в сеансе 0). Запустите Winobj.exe с правами администратора и щелкните на каталоге `\Sessions`. Отобразятся подкаталоги с числовыми именами для каждого активного сеанса. Открыв один из них, вы увидите подкаталоги с названиями `DosDevices`, `Windows`, `AppContainerNamedObjects` и `BaseNamedObjects`, которые относятся к локальному пространству имен этого сеанса. Далее приведен пример локального пространства имен.



Затем запустите Process Explorer и выберите процесс из своего сеанса (например, Explorer.exe), а после просмотрите его таблицу дескрипторов, выбрав в меню

Вид (View) пункт Вид нижней панели (Lower Pane View), а потом — Дескрипторы (Handles)). Вы заметите, что дескриптор для `\Windows\WindowStations\WinSta0` оказался под `\Sessions\`*n*, где *n* — ID сеанса.



Фильтрация объектов

Windows включает в диспетчер объектов модель фильтрации наподобие модели мини-фильтров в файловой системе и обратных вызовов реестра, упоминаемых в главе 10. Одним из основных преимуществ данной модели фильтрации является возможность использования понятия *высоты* (altitude), которое применяется в данных существующих технологиях фильтрования, это означает, что несколько драйверов могут фильтровать события диспетчера объектов в соответствующих местах стека фильтрации. Кроме того, драйверам разрешается перехватывать вызовы таких функций, как `NtOpenThread` и `NtOpenProcess`, и даже изменять маски доступа, запрашиваемые диспетчером процессов. Это позволяет защититься от конкретных операций над открытым дескриптором — не позволить программе злоумышленника завершить благонадежный защитный процесс или же не дать ворующему пароли приложению получить разрешение на чтение памяти процесса LSA. Однако следует заметить, что операцию открытия нельзя заблокировать полностью из соображений совместимости, так как это, например, отняло бы у диспетчера задач способность запрашивать командную строку запуска или имя образа процесса.

Более того, драйверы получают возможность воспользоваться как предшествующими, так и последующими обратными вызовами, что позволит им приготовиться к конкретной операции до ее осуществления, а также отреагировать на информацию или придать ей окончательную форму после завершения этой операции. Эти обратные вызовы могут быть указаны для каждой операции (на данный момент поддерживаются только операции создания, открытия и создания дубликата) и могут быть указаны для каждого типа объекта (поддерживаются только объект процесса и объект потока). Для каждого обратного вызова драйверы могут указать свое собственное внутреннее значение контекста, который может возвращаться при всех вызовах драйвера или при вызовах пары предваряющего и завершающего обратных вызовов. Эти обратные вызовы могут быть зарегистрированы с помощью API-функции `ObRegisterCallbacks`, а их регистрация может быть отменена с помощью API-функции `ObUnregisterCallbacks`, причем отмена регистрации возлагается на драйвер.

Использовать эти функции позволено исполняемым образам с определенными характеристиками.

- Образ должен иметь подпись даже на 32-разрядных компьютерах согласно правилам, которые сформулированы в политике подписи кода режима ядра (Kernel Mode Code Signing, KMCS). Образ должен быть скомпилирован с установленным флагом компоновщика `/integritycheck`, который приводит к установке значения `IMAGE_DLLCHARACTERISTICS_FORCE_INTEGRITY` в PE-заголовке. Это сообщает диспетчеру памяти о необходимости проверять подпись образа вне зависимости от любых умолчаний, обычно не требующих проверки.
- Образ должен быть подписан с каталогом, содержащим криптографические хеши для каждой страницы исполняемого кода. Это позволяет системе обнаружить изменения образа после его загрузки в память.

Прежде чем выполнить обратный вызов, диспетчер объектов вызывает функцию `MmVerifyCallbackFunction` с указателем на функцию обратного вызова в роли аргумента, которая, в свою очередь, определяет местонахождение таблицы данных загрузчика, связанной с модулем — владельцем этого адреса, и проверяет, установлен ли флаг `LDRP_IMAGE_INTEGRITY_FORCED`.

СИНХРОНИЗАЦИЯ

При разработке операционных систем важную роль играет понятие взаимного исключения. Оно относится к обеспечению того, что в каждый момент времени доступ к конкретному ресурсу может иметь один и только один поток. Взаимное исключение необходимо, когда ресурс не предполагает совместного использования или когда такое использование может привести к непредсказуемому исходу. Например, если два потока одновременно копируют файл на порт принтера, их вывод может быть перепутан. Аналогично этому, если один поток производит чтение из какого-нибудь места памяти, в то время как другой поток ведет запись в это же место, первый поток получит непредсказуемые данные. В общем, ресурсы, которые могут подвергаться записи, не могут совместно использоваться без ограничений, а ресурсы, которые не подвергаются изменениям, могут использоваться совместно. На рис. 8.37 показано, что произойдет, когда и тот и другой из двух потоков, запущенные на разных процессорах, записывают данные в круговую очередь.

Поскольку второй поток получил значение указателя на хвост очереди раньше, чем первый поток закончил его обновлять, второй поток вставил свои данные туда же, куда свои поместил первый, перезаписав данные и оставив один из элементов очереди пустым. Хотя рис. 8.37 показывает, что могло бы произойти в мультипроцессорной системе, та же ошибка может возникнуть и при одном процессоре, если операционная система переключит контекст на второй поток прежде, чем первый обновит указатель на хвост очереди.

Разделы кода, обращающиеся к необобщаемому ресурсу, называются *критическими разделами*. Для обеспечения корректности кода только один поток может выполнять критический раздел в отдельно взятый момент времени. Пока один поток ведет запись в файл, обновляя базу данных или изменяя значение общей переменной, никакой другой поток не должен получать разрешение на доступ к

тому же самому ресурсу. Псевдокод на рис. 8.37 является критическим разделом, который, вопреки правилам, без взаимного исключения получил доступ к совместно используемой структуре данных.

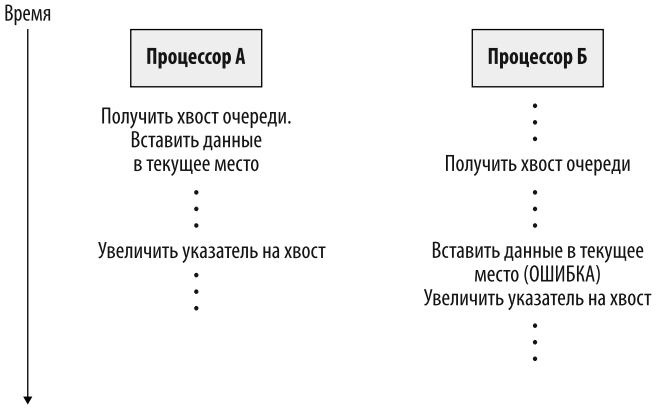


Рис. 8.37. Неправильное совместное использование памяти

Вопрос взаимного исключения важен для всех операционных систем, но особенно критичен (и сложен) для *тесно связанных, симметричных многопроцессорных* (symmetric multiprocessing, SMP) операционных систем, таких как Windows, где один и тот же системный код одновременно выполняется на более чем одном процессоре, совместно используя определенные структуры данных, хранящиеся в глобальной памяти. В обязанности ядра Windows входит обеспечение механизмов, которые системный код может использовать для предотвращения ситуаций, когда два потока одновременно изменяют одни и те же данные. Для этого оно предоставляет примитивы взаимного исключения, которые оно само и другие компоненты исполнительной системы применяют для синхронизации своего доступа к глобальным структурам данных.

Поскольку планировщик синхронизирует доступ к своим структурам данных на IRQL-уровне DPC/dispatch, ядро и исполнительная система не могут полагаться на механизмы синхронизации, способные окончиться ошибкой страницы или операцией перепланировки, при синхронизации доступа к структурам данных при IRQL-уровне DPC/dispatch или выше (уровни, также известные как *повышенные* или *высокие*). В следующих разделах вы узнаете, как ядро и исполнительная система используют взаимное исключение для защиты своих глобальных структур данных в условиях высоких IRQL и какие механизмы взаимного исключения и синхронизации применяются ими же в условиях *низких* IRQL (ниже уровня DPC/dispatch).

Высокоуровневая IRQL-синхронизация

На различных стадиях своей работы ядро должно гарантировать, что один и только один процессор за раз исполняет код в критическом разделе. Критическими разделами ядра являются участки кода, которые вносят изменения в глобальные структуры данных, такие как база данных диспетчеров ядра или его очередь DPC.

Чтобы операционная система корректно работала, ядро должно гарантировать, что потоки обращаются к этим структурам данных в порядке взаимного исключения.

Самой проблемной темой являются прерывания. Например, ядро может вносить изменения в глобальную структуру данных в момент, когда срабатывает прерывание и его обработчик тоже вносит туда изменения. Простые однопроцессорные операционные системы иногда предотвращают такой сценарий, запрещая все прерывания при каждом обращении к глобальным данным, но в ядре Windows предусмотрено не такое простое решение. Прежде чем воспользоваться глобальным ресурсом, ядро временно маскирует прерывания, чьи обработчики задействуют тот же ресурс. Это достигается поднятием IQL процессора до высшего уровня из всех, какие могут использовать потенциальные источники прерываний, обращающиеся к глобальным данным. Например, прерывание на уровне DPC/dispatch заставляет диспетчер, использующий базу данных диспетчеров, запуститься. Таким образом, любая другая часть ядра, которая задействует базу данных диспетчеров, повышает IRQL до уровня DPC/dispatch, маскируя все прерывания этого уровня, прежде чем туда обратиться.

Данная стратегия подходит для однопроцессорной системы, но не работает в многопроцессорных конфигурациях. Повышение IRQL на одном процессоре не мешает прерываниям срабатывать на других. Кроме того, ядро должно гарантировать взаимоисключающий доступ для нескольких процессоров.

Взаимоблокируемые операции

Эта простейшая форма механизмов синхронизации опирается на аппаратную поддержку безопасных для мультипроцессорной среды манипуляций с целочисленными значениями и операций сравнения. К ним относятся такие функции, как `InterlockedIncrement`, `InterlockedDecrement`, `InterlockedExchange` и `InterlockedCompareExchange`.

Например, функция `InterlockedDecrement` использует инструкции с поддерживаемым в x86 и x64 префиксом `lock` (например, `lock xadd`) для блокировки мультипроцессорной шины при операциях вычитания. Это делается для того, чтобы другой процессор, который также изменяет то же самое место в памяти, подвергающееся уменьшению значения, не мог его изменить в тот период, когда процессор читает исходное значение и записывает обратно уменьшенное значение. Эта форма базовой синхронизации используется ядром и драйверами. В современном инструментарии компиляторов Microsoft такие функции называются *встроенными* (intrinsic), поскольку код для них генерируется встроенным ассемблером непосредственно в фазе компиляции, вместо того чтобы проходить через вызов функции. Вполне вероятно, что помещение параметров в стек, вызов функции, копирование параметров в регистры, а затем извлечение параметров из стека и возвращение управления вызывающему коду будут более затратными операциями, чем та работа, которой функция должна заниматься в первую очередь.

Спин-блокировки

Механизм, который ядро использует для достижения мультипроцессорной взаимной блокировки, называется *спин-блокировкой* или *спинлоком*. Спин-блокировка — это блокирующий примитив, связанный с глобальной структурой данных наподобие очереди DPC, показанной на рис. 8.38.

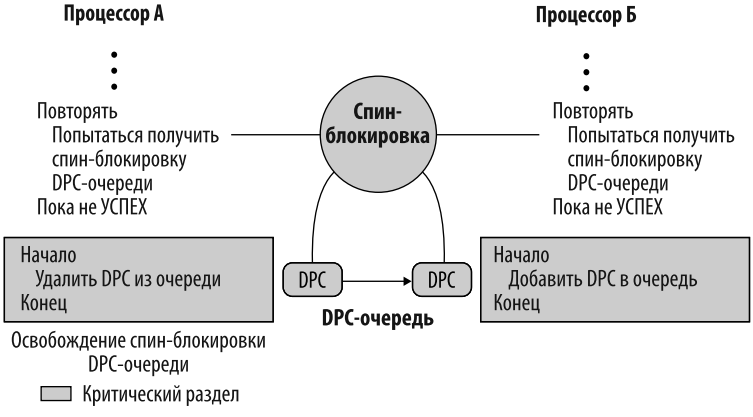


Рис. 8.38. Использование спин-блокировки

Прежде чем войти в ту или иную критический раздел из изображенных на рис. 8.38, ядро должно получить спин-блокировку, связанную с защищенной очередью DPC. Если спин-блокировка занята, ядро будет продолжать попытки до тех пор, пока не добьется успеха. Спин-блокировка (циклическая блокировка) называется так потому, что ядро ожидает «в цикле», пока не получит блокировку.

Спин-блокировки, как и защищаемые ими структуры данных, располагаются в пуле неподкачиваемой памяти, отображаемой на системное адресное пространство. Код для получения и освобождения спин-блокировки написан на языке ассемблера из соображений скорости и возможности эксплуатировать любой механизм блокировки, доступный в архитектуре реально используемого процессора. Во множестве архитектур спин-блокировки реализованы с помощью аппаратно поддерживаемой операции вида test-and-set, которая проверяет значение переменной блокировки по логике «И» и получает блокировку в рамках одной атомарной операции. Проверка и получение блокировки в рамках одной инструкции предотвращает захват вторым потоком блокировки в период между тем, как первый поток проверит переменную, и тем, когда он получит блокировку. Кроме того, аппаратная инструкция, такая как упомянутая ранее `lock`, также может быть использована в рамках операции test-and-set, что приводит к появлению комбинированной команды `lock bts` на процессорах x86 и x64, которая также блокирует шину мультипроцессора. В ином случае стало бы возможно выполнить эту операцию атомарно более чем на одном процессоре. (Без блокировки данная операция будет гарантированно атомарной только на текущем процессоре.) На процессорах ARM инструкции вроде `ldrex` и `strex` могут быть использованы подобным образом.

В системе Windows все спин-блокировки режима ядра имеют ассоциированный с ними уровень IRQL, который всегда DPC/dispatch или выше. Таким образом, когда поток пытается получить доступ к спин-блокировке, любые другие действия на уровне IRQL этой спин-блокировки или ниже прекращаются. Поскольку диспетчеризация потоков происходит на уровне DPC/dispatch, поток, который удерживает спин-блокировку, никогда не прерывается, потому что данный IRQL маскирует механизмы диспетчеризации. Маскировка позволяет коду, выполняющемуся в критическом разделе, защищенном спин-блокировкой, продолжить работу, чтобы

в итоге быстро освободить блокировку. Ядро использует спин-блокировки очень осторожно, минимизируя количество инструкций, исполняемых во время удержания спин-блокировки. Любой процессор при попытке получить спин-блокировку оказывается фактически занят, бесконечно ожидая, потребляя электроэнергию (занятость при ожидании аналогична 100%-ной нагрузке) и не выполняя никакой реальной работы.

На процессорах x86 и x64 в циклы ожидания в режиме занятости может быть вставлена специальная ассемблерная инструкция `pause`, в то время как на процессорах ARM аналогичное преимущество дает инструкция `yield`. Данная инструкция подсказывает процессору, что инструкции цикла, которые он сейчас обрабатывает, являются частью цикла на получение спин-блокировки или иной подобной конструкции. Инструкция `pause` обеспечивает три преимущества.

- Она существенно сокращает потребление энергии за счет небольшой задержки в работе ядра вместо постоянного выполнения цикла.
- В случае с ядрами SMT позволяет процессору понять, что «работа», выполняемая в рамках цикла на этом логическом ядре, не слишком важна, что дает больше процессорного времени второму логическому ядру.
- Поскольку цикл занятого ожидания приводит к большому потоку запросов на чтение, приходящих с шины от ожидающего потока (они могут возникать беспорядочно), процессор пытается скорректировать нарушение порядка использования памяти, как только замечает операцию записи, иными словами, когда поток-владелец отпускает блокировку. Таким образом, как только спин-блокировка освобождается, процессор заново сортирует все ожидающие операции чтения, чтобы обеспечить правильный порядок работы с памятью. Дополнительная сортировка приводит к серьезному снижению производительности системы и может быть предотвращена с помощью инструкции `pause`.

  Если ядро замечает, что оно работает в рамках гипервизора, совместимого с Hyper-V, который поддерживает паравиртуализацию (описывается в главе 9), механизм спин-блокировки может использовать библиотечную функцию `HvlNotifyLongSpinWait`, когда обнаруживает, что спин-блокировка сейчас принадлежит другому процессору, вместо того чтобы оставаться в цикле и использовать инструкцию `pause`.

Ядро делает спин-блокировки доступными для других частей исполнительной системы с помощью набора своих функций, в числе которых `KeAcquireSpinLock` и `KeReleaseSpinLock`. Например, драйверам устройств спин-блокировки необходимы для того, чтобы гарантировать, что регистры устройства и другие глобальные структуры данных в определенный момент предоставлялись для доступа только одной части драйвера устройства и только на одном процессоре. Спин-блокировки не предназначены для пользовательских программ — те должны задействовать объекты, которые описываются в следующем разделе. Кроме того, драйверам устройств требуется защищать собственные структуры данных от прерываний, с ними же связанных. API для спин-блокировок, как правило, повышают IRQL только до DPC/dispatch, но этого мало для защиты от прерываний. По этой причине ядро также

экспортирует функции `KeAcquireInterruptSpinLock` и `KeReleaseInterruptSpinLock`, принимающие в качестве параметра объект `KINTERRUPT`, обсуждавшийся в начале данной главы. Система анализирует объект прерывания на предмет связанного с ним уровня DIRQL и повышает IRQL в достаточной степени, чтобы обеспечить корректный доступ к структурам при совместном доступе с ISR.

Кроме того, устройства могут воспользоваться API `KeSynchronizeExecution`, чтобы синхронизировать функцию с ISR целиком, вместо просто критического раздела. В любом случае код, защищенный прерыванием спин-блокировки, должен выполняться максимально быстро — любое промедление вызывает ненормальную задержку прерывания, что значительно ухудшает производительность.

Спин-блокировки ядра предполагают ряд ограничений для использующего их кода. Поскольку спин-блокировки всегда имеют IRQL-уровень DPC/dispatch или выше, как упоминалось ранее, код, удерживающий спин-блокировку, приведет к отказу системы, если попытается спровоцировать планировщик исполнить операцию диспетчеризации или наткнется на промах при доступе к странице.

Спин-блокировки с очередью

Улучшить масштабируемость спин-блокировок можно, если задействовать *спин-блокировку с очередью*, которая часто используется вместо обычной, особенно когда ожидается конфликт и требуется его справедливое разрешение.

Спин-блокировка с очередью работает следующим образом: когда процессор находит необходимым получить такую спин-блокировку, которая кем-то занята, он помещает свой идентификатор в связанную с ним очередь. Когда процессор, удерживающий спин-блокировку, освобождает ее, блокировка передается следующему процессору, записанному в очередь.

То, что спин-блокировки с очередью приводят к циклической обработке флагов для каждого процессора, в отличие от глобальных спин-блокировок, имеет два следствия. Первое состоит в том, что мультипроцессорная шина не слишком сильно загружена данными для межпроцессорной синхронизации, а расположение бита в памяти не сводится к отдельному узлу NUMA, который в итоге нужно разыскивать через кэши каждого из логических процессоров. Второе следствие таково: вместо того чтобы отдать спин-блокировку случайному процессору из группы ожидания, спин-блокировка с очередью внедряет обслуживание по принципу «первым пришел — первым ушел» (FIFO). Хотя снижение загруженности шины и повышение справедливости распределения в целом оказываются значительными преимуществами, спин-блокировки с очередью требуют дополнительного контроля, в том числе лишних блокирующих операций, увеличивающих стоимость. Разработчикам необходимо с осторожностью делать выбор между дополнительным управлением и преимуществами, чтобы решать, стоит ли спин-блокировка с очередью в их случае затрачиваемых усилий.

В Windows используются два типа спин-блокировок с очередью. Первый предназначен для внутреннего применения исключительно ядром, второй доступен также внешним и созданным сторонними разработчиками драйверам. Сначала Windows определяет некоторое количество глобальных спин-блокировок с очередью,

сохраняя указатели на них в массиве, расположенном в контрольном пространстве (PCR) каждого процессора. Например, в системах x64 они находятся в поле `LockArray` структуры данных KPCR.

Обратиться к глобальной спин-блокировке можно с помощью функции `KeAcquireQueuedSpinLock`, передав ей индекс в массиве, где хранится указатель на спин-блокировку. Изначально число глобальных спин-блокировок возрастало с каждым новым релизом операционной системы, однако со временем стали использоваться более эффективные иерархии блокировки, которые уже не требовали глобальных блокировок в разрезе процессора. Вы можете просмотреть таблицу с определением индексов для этих блокировок в заголовочном файле `Wdm.h` в структуре `KSPIN_LOCK_QUEUE_NUMBER`, но следует иметь в виду, что обращение к одной из этих спин-блокировок с очередью со стороны драйвера устройства не поддерживается и настоятельно не рекомендуется. Как уже отмечалось, данные блокировки зарезервированы для внутреннего применения ядром.

ЭКСПЕРИМЕНТ. Просмотр глобальных спин-блокировок с очередью

Вы можете просмотреть состояние глобальных спин-блокировок с очередью (указатели на них хранятся в массиве спин-блокировок с очередью в PCR каждого процессора), воспользовавшись командой отладчика ядра `!qlocks`. Обратите внимание на то, что в приведенном далее примере ни один из процессоров не обращается ни к одной из этих блокировок и это стандартная ситуация в рамках локальной системы в состоянии отладки в реальном времени:

```
lkd> !qlocks
Key: O = Owner, 1-n = Wait order, blank = not owned/waiting, C = Corrupt

                      Processor Number
   Lock Name         0  1  2  3  4  5  6  7

KE - Unused Spare
MM - Unused Spare
MM - Unused Spare
MM - Unused Spare
CC - Vacb
CC - Master
EX - NonPagedPool
IO - Cancel
CC - Unused Spare
```

Внутристековые спин-блокировки с очередью

Драйверы устройств могут пользоваться динамически размещаемыми спин-блокировками с очередью посредством функций `KeAcquireInStackQueuedSpinLock` и `KeReleaseInStackQueuedSpinLock`. Ряд компонентов, включая диспетчер кэша, диспетчер пула исполнительной системы и NTFS, пользуются для своих нужд этими блокировками вместо того, чтобы обращаться к глобальным спин-блокировкам с очередью.

Функция KeAcquireInStackQueuedSpinLock принимает указатель на структуру данных спин-блокировки и дескриптор очереди спин-блокировки. На деле этот дескриптор является структурой данных, где ядро хранит информацию о состоянии блокировки, включая владельца и очередь процессоров, которые могут ожидать ее доступности. По этой причине данный дескриптор не следует делать глобальной переменной. Обычно она размещается в стеке, что гарантирует ее *локальность* с точки зрения вызывающего потока и отвечает за присутствие в названии функции и самой блокировки фрагмента InStack.

Спин-блокировки на чтение/запись

Хотя использование спин-блокировок с очередью значительно повышает производительность в конкурентных ситуациях, в Windows поддерживается дополнительный вид спин-блокировок, способных обеспечить даже более значительные преимущества за счет потенциального исключения коллизий как таковых во множестве ситуаций. Спин-блокировка на множество чтений/единственную запись, также известная как *исполнительная спин-блокировка*, является усовершенствованием обычных спин-блокировок, доступных с помощью функций ExAcquireSpinLockExclusive, ExAcquireSpinLockShared и соответствующих им функций вида ExReleaseXxx. Кроме того, для продвинутых вариантов применения существуют функции ExTryAcquireSpinLockSharedAtDpcLevel и ExTryConvertSharedSpinLockToExclusive.

Как следует из названия, блокировки такого типа позволяют получать совместный бесконфликтный доступ к спин-блокировке в тех случаях, когда нет заявки на запись. Когда происходит обращение к блокировке с целью записи, все обратившиеся за чтением должны отозвать ее, после чего запросы на чтение приниматься не будут, пока запись активна (как и другие запросы на запись). Например, если разработчик драйвера часто прибегает к просмотру связанного списка, лишь изредка добавляя или удаляя элементы, благодаря исполнительным блокировкам в большинстве случаев можно избежать коллизий доступа, не сталкиваясь со сложностями применения спин-блокировок с очередью.

Взаимоблокируемые операции исполнительной системы

Для более сложных операций, таких как добавление или удаление элементов в одно- и двусвязных списках, ядро предоставляет несколько простых функций синхронизации, построенных на спин-блокировках. В их числе ExInterlockedPopEntryList и ExInterlockedPushEntryList для односвязных списков, ExInterlockedInsertHeadList и ExInterlockedRemoveHeadList — для двусвязных. Кроме них, доступны такие функции, как ExInterlockedAddUlong и ExInterlockedAddLargeInteger. Все они требуют в качестве аргумента стандартной спин-блокировки и широко применяются в коде ядра и драйверов устройств.

Вместо того чтобы полагаться на стандартный API для получения и отзыва спин-блокировки из параметра, эти функции встраивают нужный код напрямую и, кроме того, пользуются иной схемой очередности. Функции для спин-блокировок вида Ke сначала проверяют и устанавливают бит с целью узнать, освободилась ли

блокировка, а затем для фактического получения блокировки проводят ее в рамках атомарной операции по принципу «проверки и установки». Эти процедуры запрещают прерывания процессора и сразу же пытаются провести атомарную операцию «проверки и установки». Если начальная попытка будет неудачной, прерывания снова разрешаются и продолжается выполнение стандартного алгоритма ожидания в режиме занятости до тех пор, пока операция «проверки и установки» не вернет 0, в случае чего снова перезапускается вся функция. Из-за этих неочевидных различий спин-блокировка, используемая для взаимоблокируемых функций, не должна использоваться с ранее рассмотренными стандартными API-функциями ядра. Конечно, операции со списками, не использующие взаимную блокировку, не должны смешиваться со взаимоблокируемыми операциями.

ПРИМЕЧАНИЕ Некоторые взаимоблокируемые операции исполнительной системы по возможности молча игнорируют спин-блокировку. Например, функции ExInterlocked-IncrementLong и ExInterlockedCompareExchange пользуются тем же префиксом lock, что и стандартные блокирующие и встроенные функции. Они были полезны в старых системах (или системах не-x86), где операция lock не годилась или не была доступна. Поэтому теперь вместо этих нерекомендуемых вызовов предпочтение отдается встроенным функциям.

Низкоуровневая IRQL-синхронизация

Исполняемое ПО, не входящее в состав ядра, тоже нуждается в синхронизации доступа к глобальным структурам данных в условиях многопроцессорной среды. Например, у диспетчера памяти есть только одна база данных страничных блоков, к которой он обращается как к глобальной структуре данных, а драйверам устройств нужно обеспечить возможность получения исключающего доступа к их устройствам. С помощью функций ядра исполнительная система может создать спин-блокировку, получить ее и освободить.

Однако спин-блокировки лишь обеспечивают потребность исполнительной системы в механизмах синхронизации. Поскольку их ожидание буквально приостанавливает работу процессора, использовать их можно лишь при строгом соблюдении следующих ограничений.

- Доступ к защищенному ресурсу должен осуществляться быстро и без сложных взаимодействий с остальным кодом.
- Критический раздел кода не должен попадать в страницы, выгружаемые из памяти, не должен ссылаться на данные, находящиеся в выгружаемой памяти, не должен вызывать внешние процедуры (включая системные службы) и не должен генерировать прерывания или исключения.

Эти ограничения являются исчерпывающими и не могут соблюдаться при всех возможных обстоятельствах. Кроме того, исполняющей системе нужно выполнять другие виды синхронизации в добавок к взаимному исключению, и она должна также предоставлять механизмы синхронизации пользовательскому режиму.

Для тех случаев, когда спин-блокировки не годятся, предусмотрено несколько дополнительных инструментов:

- объекты диспетчера ядра (мьютексы, семафоры, события и таймеры);
- быстрые мьютексы и защищенные мьютексы;
- push-блокировки;
- ресурсы исполнительной системы;
- инициализация Run-once (InitOnce).

Кроме того, код пользовательского режима, также действующий на низких IRQL, должен иметь собственные примитивы для блокировок. В Windows их предусмотрено несколько:

- системные вызовы, обращающиеся к управляющим объектам ядра (мьютексы, семафоры, события и таймеры);
- условные переменные (CondVars);
- тонкие блокировки чтения/записи (блокировки SRW);
- ожидание на основе адреса;
- инициализация Run-Once (InitOnce);
- критические разделы.

Примитивы для пользовательского режима и то, как работает их поддержка ядром, мы рассмотрим позже, а пока что сфокусируемся на объектах режима ядра. В табл. 8.26 сравниваются возможности данных механизмов и их взаимодействие с работой APC режима ядра.

Таблица 8.26. Механизмы синхронизации режима ядра

Механизм синхронизации	Доступны драйверам устройств	Запрет нормальных APC режима ядра	Запрос особых APC режима ядра	Поддержка рекурсивного доступа	Поддержка совместного и эксклюзивного доступа
Мьютексы диспетчера ядра	Да	Да	Нет	Да	Нет
Семафоры, события, таймеры диспетчера ядра	Да	Нет	Нет	Нет	Нет
Быстрые мьютексы	Да	Нет	Нет	Нет	Нет
Защищенные мьютексы	Да	Да	Да	Нет	Нет
Push-блокировки	Да	Да	Да	Нет	Да
Ресурсы исполнительной системы	Да	Нет	Нет	Да	Да
Защита с обратным отсчетом	Да	Нет	Нет	Да	Нет

Объекты диспетчера ядра

Ядро предоставляет исполнительной системе дополнительные механизмы синхронизации в форме особых объектов, известных как *объекты диспетчера ядра*. На них основываются возможности объектов для синхронизации, видимых Windows API. Всякий объект, доступный в Windows API и поддерживающий синхронизацию, инкапсулирует хотя бы один объект диспетчера ядра. Семантика синхронизации исполнительной системы в Windows доступна разработчику посредством функций `WaitForSingleObject` и `WaitForMultipleObjects`, которые подсистема Windows реализует, обращаясь к системным службам, предоставляемым диспетчером объектов. Поток в рамках приложения Windows способен синхронизироваться с разнообразными объектами, такими как процесс, другой поток, событие, семафор, мьютекс, таймер для ожидания, порт финализации ввода/вывода, порт ALPC, раздел в реестре или файл. Фактически дожидаться можно почти любых объектов, предоставляемых ядром. Какие-то из них сами являются объектами диспетчера, тогда как другие, будучи более крупными структурами (например, порты, разделы или файлы), содержат в своем составе объект диспетчера. В табл. 8.27 (см. далее в пункте «Что переводит объект в сигнальное состояние») приводятся объекты диспетчера, которые реализуют примитивы, зачастую присутствующие в составе любого объекта, которого Windows API позволяет дожидаться.

Другие два типа механизмов синхронизации в исполнительной системе, заслуживающие внимания, — это *ресурс исполнительной системы* и *push-блокировка*. Они могут обеспечивать как эксклюзивный доступ (подобно мьютексам), так и совместный на чтение (несколько читателей делят доступ на чтение к структуре). Однако они доступны только коду режима ядра, а поэтому недоступны в Windows API. Кроме того, они не являются в полной мере объектами — у них есть API, предоставляемый через чистые указатели и функции вида `Ex`, а диспетчер объектов с его системой дескрипторов в их работе не участвует. В следующих разделах описываются детали реализации ожидания объектов диспетчера.

Ожидание объектов диспетчера

Традиционным для потока способом синхронизироваться с объектом диспетчера является ожидание его дескриптора или, для отдельных типов объектов, прямое ожидание по его указателю. Класс функций вида `NtWaitForXxx` (также доступный в пользовательском режиме) работает с дескрипторами, в то время как функции вида `KeWaitForXxx` обращаются непосредственно к объектам диспетчера.

Поскольку API-функции `Nt` взаимодействуют с диспетчером объектов (функциями вида `ObWaitForXxx`), они проходят через абстракции, описанные в разделе о типах объектов ранее в данной главе. Например, API в части `Nt` позволяет передать дескриптор файловому объекту, поскольку диспетчер объектов использует информацию из типа объекта, чтобы перенаправить ожидание на поле `Event` внутри `FILE_OBJECT`. API в части `Ke`, в свою очередь, работает только с самими объектами диспетчера — иными словами, с теми, у которых сначала идет структура `DISPATCHER_HEADER`. Вне зависимости от подхода подобные вызовы так или иначе ведут к переводу ядром потока в состояние ожидания.

Совершенно иной, более современный подход к ожиданию объектов диспетчера базируется на *асинхронном ожидании*. Он предполагает ассоциирование существующего порта финализации ввода/вывода с поддерживающей его очередью ядра с помощью объекта-посредника, называемого *пакетом окончания ожидания*. Благодаря этому механизму поток, по сути, лишь регистрирует ожидание, не блокируя объект напрямую и не входя в состояние ожидания. Вместо этого, когда приходит срок, порт финализации ввода/вывода добавит пакет окончания ожидания, оповещая всех, кто извлекает элементы из этого порта или дожидается сигнала от него. Это позволяет одному или нескольким потокам зарегистрировать отметки ожидания на различных объектах, следить за которыми сможет отдельный поток или пул потоков. Как вы уже, вероятно, догадались, данный механизм является стержнем функциональности API пула потоков, поддерживающей ожидание обратных вызовов для таких функций, как `CreateThreadPoolWait` и `SetThreadPoolWait`.

Наконец, в относительно новых версиях Windows 10 механизм асинхронного ожидания был расширен посредством функции *события ожидания DPC*, которая сегодня зарезервирована Hyper-V (хотя API уже экспортирован, документации пока нет). Отсюда вытекает окончательный подход к ожиданиям диспетчера, предназначенный для драйверов режима ядра, при котором отложенный вызов процедуры (DPC, описан ранее в данной главе) может быть ассоциирован с объектом диспетчера, а не с потоком или портом финализации ввода/вывода. Подобно механизму, описанному ранее, DPC регистрируется с помощью объекта, и когда срок истекает, он попадает в очередь к текущему процессору, как если бы драйвер только что вызвал `KeInsertQueueDpc`. Как только блокировка диспетчера будет снята, а IRQL опустится ниже уровня DISPATCH, DPC исполнится на текущем процессоре, сыграв роль оставленного драйвером обратного вызова, который теперь сможет отреагировать на сигнальное состояние объекта.

Вне зависимости от механизма ожидания ожидающие объекты синхронизации могут находиться в одном из двух состояний: *сигнальном* и *несигнальном*. Поток не сможет продолжить исполнение, пока срок ожидания не подойдет, то есть не возникнет ситуация, когда объект диспетчера, чей дескриптор ожидает поток, сам изменит состояние с несигнального на сигнальное (когда другой поток запустит событие, например).

Для синхронизации с объектом поток обращается к одной из системных функций ожидания, предоставляемых диспетчером объектов, передав дескриптор от искомого объекта. Поток может ожидать один или несколько объектов, а также указать на необходимость прервать ожидание по истечении определенного промежутка времени. Всякий раз, когда ядро переводит объект в сигнальное состояние, один из сигнальных алгоритмов ядра проверяет наличие потоков, ожидающих искомый объект, но не ожидающих сигнала от других. В случае подтверждения ядро отпускает один или несколько потоков из состояния ожидания, чтобы те могли продолжить исполняться.

Чтобы получить асинхронное оповещение о сигнале с объекта, поток создает порт финализации ввода/вывода, после чего вызывает функцию `NtCreateWaitCompletionPacket`, чтобы создать объект пакета окончания ожидания и получить назад его дескриптор. Затем вызывается `NtAssociateWaitCompletionPacket`, куда

передаются дескрипторы и от порта, и от только что созданного пакета вместе с дескриптором объекта, о котором следует получить оповещение. Когда ядро переводит объект в сигнальное состояние, сигнальный алгоритм обнаруживает, что никакие потоки искомый объект не ждут, и проверяет наличие порта финализации ввода/вывода, связанного с ожиданием. В случае положительного результата он подает сигнал на объект очереди, связанный с портом, из-за чего все потоки, ожидающие его в данный момент, просыпаются и получают пакет окончания ожидания (или очередь просто переходит в сигнальное состояние). Другой вариант: если порта финализации ввода/вывода не нашлось, проверяется наличие DPC, который, если существует, будет добавлен в очередь к текущему процессору. Последняя часть касается описанного ранее механизма ожидания события DPC, предназначенного только для ядра.

Синхронизация влияет на диспетчеризацию потоков следующим образом.

1. Поток пользовательского режима ожидает дескриптор объекта события.
2. Ядро изменяет состояние диспетчеризации потока на ожидание и включает его в список потоков, ожидающих события.
3. Другой поток запускает событие.
4. Ядро проходится по списку потоков, ожидающих события. Если условия окончания ожидания потока удовлетворены (см. примечание), ядро выводит поток из состояния ожидания. Если это поток изменяемого приоритета, ядро может также поднять его приоритет выполнения (более подробно о планировании потоков рассказано в главе 4 тома 1).

ПРИМЕЧАНИЕ Некоторые потоки могут ожидать более одного объекта, поэтому они продолжают ожидание, если только ими не указано ожидание любого из объектов — WaitAny, что их разбудит, как только в сигнальное состояние перейдет один объект (а не все объекты).

Что переводит объект в сигнальное состояние

Сигнальное состояние для разных объектов определяется по-разному. Объект потока имеет несигнальное состояние в течение всего своего жизненного цикла и устанавливается в сигнальное состояние ядром, когда выполнение потока завершается. Подобным образом ядро устанавливает объект процесса в сигнальное состояние, когда завершается выполнение последнего потока этого процесса. В отличие от этих объектов объект таймера, подобно будильнику, устанавливается в положение «состоявшегося» в определенное время. Когда его время истекает, ядро устанавливает объект таймера в сигнальное состояние.

При выборе механизма синхронизации разработчик должен принимать во внимание правила поведения различных объектов синхронизации. Порядок завершения ожидания потока при переводе объекта в сигнальное состояние зависит от типа того объекта, который ожидается потоком (табл. 8.27).

Когда объект устанавливается в сигнальное состояние, ожидающие потоки в основном тут же освобождаются от своих ожидающих состояний. Например, объект уведомительного события (называемый в Windows API событием с ручным сбросом) используется для объявления о наступлении какогонибудь события.

Когда объект события переводится в сигнальное состояние, освобождаются все потоки, ожидавшие это событие. Исключение составляет любой поток, одновременно ожидающий более одного объекта, такому потоку может понадобиться продолжить ожидание, пока еще один объект не достигнет сигнального состояния.

Таблица 8.27. Определение сигнальных состояний

Тип объекта	Условие сигнального состояния	Реакция ожидающих потоков
Процесс	Завершается последний поток	Освобождаются все потоки
Поток	Завершается поток	Освобождаются все потоки
Событие (уведомительного типа)	Поток устанавливает событие	Освобождаются все потоки
Событие (синхронизирующего типа)	Поток устанавливает событие	Один поток освобождается и может получить повышение приоритета; объект события сбрасывается
Шлюз (блокирующего типа)	Поток сообщает о шлюзе	Освобождается и получает более высокий приоритет первый ожидающий поток
Шлюз (сигнализирующего типа)	Поток сообщает о типе	Освобождается первый ожидающий поток
Событие с ключом	Поток устанавливает событие с помощью ключа	Освобождается поток, ожидающий ключа и относящийся к тому же процессу, что и сигнализирующий поток
Семафор	Счетчик семафора уменьшается на 1	Освобождается один поток
Таймер (уведомительного типа)	Настало установленное время, или истек интервал времени	Освобождаются все потоки
Таймер (синхронизирующего типа)	Настало установленное время, или истек интервал времени	Освобождается один поток
Мьютекс	Поток освобождает мьютекс	Освобождается один поток, который становится владельцем мьютекса
Очередь	Элемент помещается в очередь	Освобождается один поток

В отличие от объекта события объект мьютекса имеет связанного с ним владельца (если только он не был приобретен во время DPC-вызова). Он используется для получения взаимно исключающего доступа к ресурсу, и только один поток в одно и то же время может удерживать мьютекс. Когда объект мьютекса становится свободным, ядро устанавливает его в сигнальное состояние, а затем выбирает для выполнения, а также для наследования любого примененного повышения приоритета один из ожидающих потоков (подробнее о повышении приоритетов говорится в главе 4 тома 1). Поток, выбранный ядром, получает объект мьютекса, а все остальные потоки продолжают ожидание.

Объект мьютекса может быть также заброшен: это случается, когда завершается выполнение того потока, который им обладает в данный момент. Когда завершается выполнение потока, ядро определяет общее количество мьютексов, которыми владел поток, и переводит их в заброшенное состояние (в понятиях логики сигнализации оно рассматривается как сигнальное состояние, в котором владение мьютексом передается ожидающему потоку).

Этот краткий обзор не предназначался для перечисления всех причин и особенностей использования различных объектов исполнительной системы. Здесь перечислены их основные функциональные возможности и поведение, связанное с синхронизацией. Для получения информации о том, как воспользоваться этими объектами в программах Windows, обратитесь к справочной документации Windows по объектам синхронизации или к книге авторов Джеффри Рихтера и Кристофера Назара «Windows via C/C++. Программирование на языке Visual C++» (*Windows via C/C++*, Microsoft Press).

Безобъектное ожидание (поточные оповещения)

Хотя возможность ожидать и получать оповещения от объекта, перешедшего в сигнальное состояние, — весьма мощный инструмент, а у программиста имеется богатый выбор разных объектов диспетчера, иногда требуется значительно более простой подход. Одному потоку потребовалось *ожидать* выполнения какого-то условия, а второму необходимо о таком выполнении *просигнализировать*. Добиться этого можно, связав искомое условие с событием, но для этого потребуются ресурсы (для начала — дескрипторы и память), а получение их может оказаться непростым, долгим и не обязательно успешным процессом. Ядро Windows обеспечивает два механизма синхронизации, не привязываемых к объектам:

- оповещения потоков;
- оповещения потоков по ID.

Несмотря на схожесть названий, принципы действия этих двух механизмов различаются. Рассмотрим, как работают оповещения потока. Сначала поток, желающий синхронизироваться, переходит в состояние прерываемого сна с помощью функции `SleepEx`, так или иначе вызывающей `NtDelayExecutionThread`. Поток режима ядра может сделать выбор также в пользу `KeDelayExecutionThread`. Ранее, в разделе о программных прерываниях и APC, мы уже обсуждали принципы оповещаемости. В данном случае поток может задать для себя интервал времени или сделать свой сон бесконечным. Другая сторона, в свою очередь, использует функцию `NtAlertThread` или `KeAlertThread`, чтобы оповестить первый поток, выводя его из состояния сна и возвращая код статуса `STATUS_ALERTED`. Для полноты картины стоит также отметить, что поток может принять решение не впадать в сон, а просто в какой-то момент времени в дальнейшем вызвать функцию `NtTestAlert` или `KeTestAlertThread`. Наконец, поток может обойтись без состояния оповещаемого ожидания, вместо этого приостановив сам себя (`NtSuspendThread` или `KeSuspendThread`). В таком случае другая сторона может применить `NtAlertResumeThread`, чтобы и оповестить, и активировать первый поток.

Несмотря на свою простоту и элегантность, данный механизм имеет недостатки, начиная с того, что не существует способа определить, было ли оповещение связано

с причиной ожидания, — иными словами, первый поток мог получить сигнал от любого другого потока, суть которого различить не получится. Кроме того, API таких оповещений официально не документирован, в связи с чем внутренним компонентам ядра и службам пользовательского режима так делать можно, но от сторонних разработчиков этого не ждут. И как только поток активируется, все ожидающие в очереди APC тоже возвращаются к исполнению, например APC пользовательского режима в случаях, когда функции оповещения применяются в приложении. Наконец, функция `NtAlertThread` все еще требует открытия дескриптора для искомого потока — операции, технически считающейся запросом на ресурсы, который может быть не удовлетворен. Теоретически вызывающие стороны могут открыть себе дескрипторы заранее, чтобы гарантировать успешное оповещение, но это все равно вносит свою лепту в общую стоимость из-за привлечения дескрипторов.

Для решения этих проблем, начиная с Windows 8, ядро получило более современный механизм оповещений потоков по ID. Хотя системные функции `NtAlertThreadByThreadId` и `NtWaitForAlertByThreadId`, стоящие за ним, не документированы, инструменты ожидания Win32 для пользовательского режима публичны. Эти вызовы очень просты и не требуют никаких ресурсов, принимая в роли аргумента лишь ID потока. Ввиду отсутствия дескриптора, что может навредить безопасности, единственным недостатком этих функций является то, что они позволяют синхронизироваться с потоками лишь в рамках текущего процесса.

Объясняется поведение данного механизма довольно просто: поток блокируется функцией `NtWaitForAlertByThreadId`, при необходимости передавая тайм-аут. Так он переходит в полное ожидание, не нуждаясь в оповещениях. Фактически, несмотря на название, какой-то сигнализации здесь не предусмотрено технически. Затем другой поток вызывает функцию `NtAlertThreadByThreadId`, которая заставляет ядро выполнить поиск по ID потока, убедиться, что тот принадлежит вызывающему процессу, а потом проверить, что действительно имеется остановка из-за вызова `NtWaitForAlertByThreadId`. Если поток в этом состоянии, он просто активируется. Этот простой и элегантный механизм играет ключевую роль в ряде примитивов синхронизации в условиях пользовательского режима, о которых будет сказано далее в этой главе, и может быть задействован в реализации чего угодно, от барьеров до более сложных методик синхронизации.

Структуры данных

Для отслеживания, *кто* находится в состоянии ожидания, *как* осуществляется это ожидание, *что именно* ожидается и *в каком состоянии* находится вся операция ожидания, есть три ключевые структуры данных: заголовок диспетчера, блок ожидания и регистр состояния ожидания. Первые две структуры определены в заголовочном файле WDK `Wdm.h`, а третья не задокументирована, но видна среди публичных символов с типом `KWAIT_STATUS_REGISTER`, где поле `Flags` принимает значения из последовательности `KWAIT_STATE`.

Заголовок диспетчера является упакованной структурой, поскольку в нем нужно хранить большой объем информации в структуре фиксированного размера

(подробнее о структуре данных заголовка диспетчера — далее во врезке «Эксперимент. Просмотр очередей ожидания»). Одной из основных тонкостей является определение взаимоисключающих флагов в структуре, в одном и том же месте памяти (с одинаковым смещением), что в теории программирования известно как *объединение*. Из поля `Type` ядро узнает, какое из этих полей применяется в данном случае. Например, мьютекс может быть заброшенным, а таймер может быть абсолютным или относительным. Аналогичным образом, таймер может быть вставлен в список таймеров, а поле активности отладчика Debug Active имеет смысл только для процессов. С другой стороны, заголовок диспетчера действительно содержит информацию, общую для любого объекта диспетчера: тип объекта, сигнальное состояние и список потоков, ожидающих этот объект.

Блок ожидания представляет поток, ожидающий объект. У каждого потока, находящегося в состоянии ожидания, есть массив блоков ожидания, вмещающий до 64 блоков, представляющих объекты, которых данный поток дожидается, в том числе потенциально блок ожидания, указывающий на внутренний таймер потока и таймер для реализации тайм-аута, предложенного вызывающей стороной. Альтернативно, если используются примитивы `alert-by-ID`, существует единственный блок со специальным обозначением, что данное ожидание реализуется без диспетчера. Поле `Object` заменяется подсказкой (`Hint`), которую указывает код, вызывающий функцию `NtWaitForAlertByThreadId`. У этого массива блоков две цели.

- Когда поток завершается, все объекты, которых он ожидал, должны быть освобождены от ссылок на них, а сами блоки ожидания удалены и от этих объектов отвязаны.
- Когда поток активизируется по сигналу одного из объектов, которых он ожидал, все остальные объекты, которых он мог дожидаться, должны быть освобождены от ссылок, а блоки ожидания удалены.

Подобно тому как у потока есть массивом с данными об объектах, им ожидаемых, каждый объект диспетчера тоже имеет список связанных с ним блоков ожидания. Он нужен для того, чтобы, когда объект перейдет в сигнальное состояние, ядро смогло быстро определить, кто его ожидает (или с каким портом финализации ввода/вывода он связан), и применить логику завершения ожидания, о которой расскажем в дальнейшем.

Наконец, поскольку поток *диспетчера настройки баланса*, действующего на каждом из процессоров (подробности о диспетчере настройки баланса см. в главе 5 тома 1), должен анализировать время, затрачиваемое каждым потоком на ожидание (и решить, подкачать ли стек ядра), каждый PRCB обладает списком подходящих ожидающих потоков, недавно исполнявшихся на этом ядре. Для этого вновь задействуется поле `Ready List` из структуры `KTHREAD`, поскольку поток не может пребывать в состоянии готовности и ожидания одновременно. Подходящие потоки должны отвечать трем условиям.

- Ожидание должно быть запущено в режиме, соответствующем пользовательскому режиму (ожидания режима ядра считаются чувствительными к времени и не стоящими затрат на своппинг стека).

- У потока должен быть установлен флаг `EnableStackSwap` (драйверы уровня ядра могут отключить эту возможность, вызвав функцию `KeSetKernelStackSwapEnable`).
- Приоритет потока должен быть ниже стартового значения диапазона приоритетов реального времени Win32 или равен ему (24 — значение по умолчанию для нормального потока в рамках процесса с приоритетом реального времени).

Структура блока ожидания всегда фиксированная, но некоторые ее поля могут использоваться по-разному в зависимости от типа ожидания. Например, обычно блок ожидания содержит указатель на ожидаемый объект, но в случае применения примитива `alert-by-ID` никакого объекта нет, а значит, нам нужна подсказка, указанная при вызове.

Однако во всех случаях остаются два поля, а именно *тип ожидания* и *состояние блока ожидания*. Кроме того, может присутствовать *ключ ожидания*. Тип ожидания особенно важен в момент его завершения, так как от него зависит, в каком из пяти режимов оно будет происходить. В режиме `wait any` ядро не обращает внимания на состояние остальных объектов в списке, поскольку уже как минимум один из них (а именно, текущий!) подал сигнал. В то же время в режиме `wait all` ядро может пробудить поток только в том случае, когда все ожидаемые объекты оказались в состоянии сигнализации одновременно, из-за чего приходится пролистывать все блоки ожидания и связанные с ними объекты.

В режиме `wait dequeue` обрабатывается особый случай, когда объект диспетчера на самом деле является очередью (порт финализации ввода/вывода) и существует поток, ожидающий появления оттуда пакета финализации (с помощью вызова `KeRemoveQueue(Ex)` или `(Nt)IoRemoveIoCompletion`). Блоки ожидания для очередей активизируются в порядке LIFO — подобно стеку (в отличие от принципа FIFO — подобно очереди, действующего для других объектов диспетчера), так что, когда очередь сигнализируется, появляется возможность принять верные меры (стоит помнить, что поток мог ожидать нескольких объектов в режиме `wait all` или `wait any`, за которыми тоже нужно постоянно следить).

В режиме `wait notification` ядро знает, что никаких потоков с данным объектом не связано, и наблюдается асинхронное ожидание с привязкой к порту финализации ввода/вывода, чья очередь и будет просигнализирована (поскольку очередь и есть объект диспетчера, это приводит к вторичному завершению ожидания для очереди и всех потоков, которые могли бы ее ожидать).

Наконец, `wait DPC`, самый новый режим из представленных, дает ядру понять, что с ожиданием не связаны ни поток, ни порт финализации ввода/вывода, но связан объект DPC. В таком случае указатель ссылается на инициализированную структуру KDPC, которую ядро помещает в очередь к текущему процессору для исполнения в ближайшее время, как только будет снята блокировка диспетчера.

Блок ожидания также содержит изменчивое поле `wait block state` (`KWAIT_BLOCK_STATE`), определяющее текущее состояние данного блока в рамках операции ожидания транзакции, в которой он участвует. Различные состояния, их значение и влияние на поведение кода логики ожидания описаны в табл. 8.28.

Таблица 8.28. Состояние блоков ожидания

Состояние	Значение	Эффект
WaitBlockActive (4)	Данный блок ожидания напрямую связан с объектом как частью потока в режиме ожидания	При удовлетворении ожидания блок будет исключен из списка
WaitBlockInactive (5)	Поток, связанный с данным блоком, был удовлетворен (или срок, если таковой на это отводится, уже истек)	При удовлетворении ожидания блок не будет исключен из списка, так как это уже произошло, когда блок находился в активном состоянии
WaitBlockSuspended (6)	По отношению к потоку, связанному с данным блоком ожидания, будет выполнена операция легкой приостановки	Фактически то же поведение, что и в состоянии WaitBlockActive, но случается только при выходе из приостановки. При обычном удовлетворении ожидания игнорируется (не должно случаться никогда, так как приостановленный поток ничего ожидать не может!)
WaitBlockBypassStart (0)	Сигнал доставляется потоку, но он еще не перешел в режим ожидания	При удовлетворении ожидания (которое может произойти немедленно, даже раньше, чем поток по-настоящему перейдет в состояние ожидания) ожидающий поток должен синхронизироваться с источником сигнала из-за риска того, что ожидаемый объект может быть в стеке. То, что этот блок будет помечен как неактивный, вынудит ожидающего разбирать стек, в то время как источник сигнала все еще может им пользоваться
WaitBlockBypassComplete (1)	Поток, ассоциированный с данным блоком ожидания, был корректно синхронизирован (ожидание удовлетворено), и сценарий обхода завершен	Данный блок ожидания теперь фактически считается неактивным (игнорируется)
WaitBlockSuspendBypassStart (2)	Потоку доставляется сигнал, хотя легкая приостановка еще не зафиксирована	Блок ожидания фактически считается находящимся в состоянии WaitBlockBypassStart
WaitBlockSuspendBypassComplete (3)	Легкая приостановка, связанная с данным блоком ожидания, теперь корректно синхронизирована	Блок ожидания теперь ведет себя так, как будто находится в состоянии WaitBlockSuspended

Наконец, мы упоминали регистр состояния ожидания. С тех пор как в Windows 7 была упразднена блокировка глобального диспетчера ядра, общее состояние потока (или любого из объектов, требуемого для начала ожидания) может измениться в то время, пока операции ожидания все еще подготавливаются. Поскольку никакой глобальной синхронизации состояний больше нет, ничто не мешает другому потоку, выполняемому на другом логическом процессоре, попытаться ввести в сигнальное состояние один из ожидаемых объектов, или, может быть, оповестить поток, или даже отправить APC-вызов. В связи с этим диспетчер ядра нуждается в отслеживании двух дополнительных фрагментов данных для каждого ожидающего потока: текущего четко определяемого состояния ожидания потока (`KWAIT_STATE`, не путать с состоянием блока ожидания), а также любых изменений подвешенного состояния, способных изменить результат попытки провести операцию ожидания. Эти два блока данных и составляют регистр состояния ожидания (`KWAIT_STATUS_REGISTER`).

Когда потоку дают команду ожидать какой-то объект (например, вызовом `WaitForSingleObject`), он сначала пытается войти в действующее состояние ожидания (`WaitInProgress`). Эта операция достигает успеха, если на данный момент времени в адрес потока нет отложенных оповещений. Здесь все основано на способности получения оповещений в режиме ожидания и текущем режиме ожидания процессора, которые определяют, может ли оповещение получить приоритет над ожиданием. Если оповещение имеется, вход в режим ожидания вообще не происходит, и вызывающая программа получает соответствующий код состояния, в ином случае поток входит в состояние `WaitInProgress`, и в этот момент основное состояние его меняется на `Waiting` (ожидающий). При этом записываются причина и время ожидания, кроме того, должны быть зарегистрированы какие-нибудь сроки истечения ожидания.

После входа в режим ожидания поток может нужным образом инициализировать блоки ожидания (и в процессе пометить их как `WaitBlockActive`), а затем приступить к отслеживанию всех объектов, являющихся частями этого ожидания. Поскольку у каждого объекта есть своя собственная блокировка, важно, чтобы ядро могло обеспечить последовательную схему очередности блокировки, при которой несколько процессоров могли бы анализировать цепочку ожиданий, состоящую из многих объектов (выстроенную с помощью вызова функции `WaitForMultipleObjects`). Для этого ядром используется технология, известная как *сортировка адресов* (address ordering). Поскольку у каждого объекта есть отличный от других и статичный адрес режима ядра, объекты могут быть выстроены в монотонно-возрастающем порядке адресов, гарантирующем, что блокировки будут всегда получаться и освобождаться в одном и том же порядке всеми вызывающими программами. Это означает, что соответствующим образом может быть продублирован и отсортирован массив объектов, предоставляемый вызывающей программой.

Следующим этапом будет проверка возможности немедленного удовлетворения ожидания, например, когда потоку предписывается ожидание мьютекса, который уже освободился, или события, о наступлении которого уже был получен сигнал. В таких случаях ожидание удовлетворяется немедленно, что включает в себя отсоединение связанных с этим блоков ожидания (в данном случае никакие блоки ожидания еще и не были вставлены) и выполнение выхода из ожидания (обработка любых, отложенных операций планировщика, помеченных в регистре состояния

ожидания). Если этот кратчайший путь пройти не удается, ядро пытается проверить, не истекло ли указанное для ожидания время (если таковое имелось). В этом случае ожидание не «удовлетворяется», а просто считается «просроченным», что приводит к немного более быстрой обработке кода выхода, хотя и с таким же результатом.

Если ни один из этих кратчайших путей не был пройден, блок ожидания вставляется в список ожидания процесса и теперь поток пытается совершить свое ожидание. Тем временем блокировка или блокировки объекта освобождается, позволяя другим процессорам изменить состояние любого из объектов, на котором, как теперь предполагается, поток пытается построить свое ожидание. Если предположить развитие событий по сценарию, в котором отсутствуют моменты соперничества, где другие процессоры не интересуются этим потоком или его объектами ожидания, ожидание переключается в состояние совершения ожидания, если нет незавершенных изменений, отмеченных в регистре состояния ожидания. Операция совершения ожидания связывает ожидающий поток со списком PRCB, активирует, если это нужно, дополнительный поток очереди ожидания, и вставляет таймер, связанный с подсчетом времени истечения ожидания, если таковое имеется. Поскольку потенциально к этому моменту завершится довольно много циклов, появится новая возможность истечения лимита времени. При таком сценарии вставка таймера приведет к немедленной отправке сигнала потоку, и, таким образом, к удовлетворению ожидания по таймеру, и к общему истечению времени ожидания. В противном случае, по наиболее возможному сценарию, теперь контекст центрального процессора переключится на следующий поток, готовый к выполнению (подробности о планировании см. в главе 4 тома 1).

На многопроцессорных компьютерах в ветвях кода с высокой конкуренцией возможно и вероятно, что поток может претерпеть изменения в момент, когда его попытка зафиксировать свое ожидание только оформляется. Один из возможных сценариев возникает, когда один из ожидавшихся объектов только что перешел в сигнальное состояние. Как уже упоминалось, это переводит блок ожидания в состояние `WaitBlockBypassStart`, а регистр состояния ожидания потока теперь показывает состояние ожидания `WaitAborted`. Другой возможный сценарий заключается в том, что ожидающему потоку было выдано оповещение или APC-вызов. Тогда состояние потока не меняется — `WaitAborted`, но в регистре состояния устанавливается один из соответствующих битов. Поскольку вызовы APC могут прекращать ожидание (в зависимости от типа APC, режима ожидания и возможности принимать оповещения), такой вызов доставляется, а ожидание прекращается. Другие операции будут изменять регистр состояния ожидания без генерирования полного цикла прекращения, включая изменения, вносимые в состояния приоритетности или родственности потока, которые будут обработаны при выходе из режима ожидания по причине ошибки перехода в этот режим, как уже упоминалось в предыдущих примерах.

Как кратко было сказано ранее, а также в разделе о планировщике из главы 4 тома 1, в недавно появившихся версиях Windows реализован механизм легкой приостановки, где используются функции `SuspendThread` и `ResumeThread`, более не полагающиеся во всех случаях на APC, который должен получить событие приостановки в составе объекта потока. Вместо этого ожидание может быть преобразовано в состояние приостановки при соблюдении следующих условий.

- Параметр `KiDisableLightWeightSuspend` равен `0` (администраторы могут отключить эту оптимизацию, изменив параметр в разделе реестра HKLM\SYSTEM\CurrentControlSet\Session Manager\Kernel).

- Состояние потока — `Waiting`, иными словами, поток уже в состоянии ожидания.
- Регистр состояния ожидания установлен в `WaitCommitted`, иными словами, поток полностью перешел к ожиданию.
- Поток не является основным для UMS или планируемым (подробности о планировании пользовательского режима см. в главе 4 тома 1), поскольку таким требуется дополнительная логика, реализованная в APC приостановки планировщика.
- Поток вызвал ожидание на IRQL 0 (пассивный уровень), поскольку ожидания на `APC_LEVEL` требуют особой обработки, которую может выполнить только APC-вызов приостановки планировщика.
- У потока не запрещены APC и на данный момент никаких APC не выполняется, потому что иные ситуации требуют дополнительной синхронизации, которой позволяет достичь доставка APC планировщика.
- Поток на данный момент не привязан к другому процессу посредством вызова `KeStackAttachProcess`, поскольку в ином случае потребуется особая обработка, как в пункте ранее.
- Если первый блок ожидания, связанный с ожиданием потока, не в состоянии `WaitBlockInactive`, тип ожидания должен быть `WaitAll`, в ином случае это значит, что имеется минимум один блок `WaitAny`.

Как следует из этого списка, преобразование сводится к тому, что каждый активный на данный момент блок ожидания переводится в состояние `WaitBlockSuspended`. Если в блоке есть ссылка на какой-то объект, тот исключается из списка ожидания заголовка диспетчера, так что сигнализация данного объекта не приведет к активизации потока. Если с потоком был связан таймер, он отменяется и удаляется из массива блоков ожидания потока, по факту чего устанавливается флаг. Наконец, изначальный режим ожидания (пользовательский или ядра) также сохраняется в виде флага.

Поскольку настоящий объект ожидания больше не используется, механизм потребовал трех дополнительных состояний блока ожидания, показанных в табл. 8.28, как четырех новых состояний ожидания: `WaitSuspendInProgress`, `WaitSuspended`, `WaitResumeInProgress` и `WaitResumeAborted`. Эти новые состояния схожи в поведении со своими обычными версиями, но учитывают конкурентные обстоятельства, описанные ранее, во время операции легкой приостановки.

Например, когда поток возобновляется, ядро определяет, был ли он помещен в состояние легкой приостановки, и, в сущности, обращает операцию вспять, устанавливая регистр ожидания в `WaitResumeInProgress`. Затем каждый блок ожидания проверяется, все находившиеся в состоянии `WaitBlockSuspended` переводятся в `WaitBlockActive` и вновь связываются со списком блоков ожидания заголовка диспетчера своего объекта. Однако, если тот объект в то время был сигнализирован, состояние меняется на `WaitBlockInactive`, как и при обычной активизации. Наконец, если с потоком был связан тайм-аут, чье ожидание было отменено, таймер потока возвращается в таблицу таймеров с сохранением оригинального срока истечения.

На рис. 8.39 показана связь объектов диспетчера с блоками ожидания, через них — с потоками и далее с PRCB. В данном примере процессор 0 имеет два ожидающих (зафиксировано) потока: поток 1 ожидает объект Б, а поток 2 — объекты А и Б. Если объект А сигнализируется, ядро заключает, что раз поток 2 ожидает еще один объект,

значит, не может быть подготовлен к исполнению. В то же время, если сигнализировать объект Б, ядро сможет приготовить поток 1 к исполнению сразу же, так как никаких других объектов тот не ждет (как вариант, если бы поток 1 ждал другие объекты, но ожидание началось в режиме `WaitAny`, ядро смогло бы его пробудить).

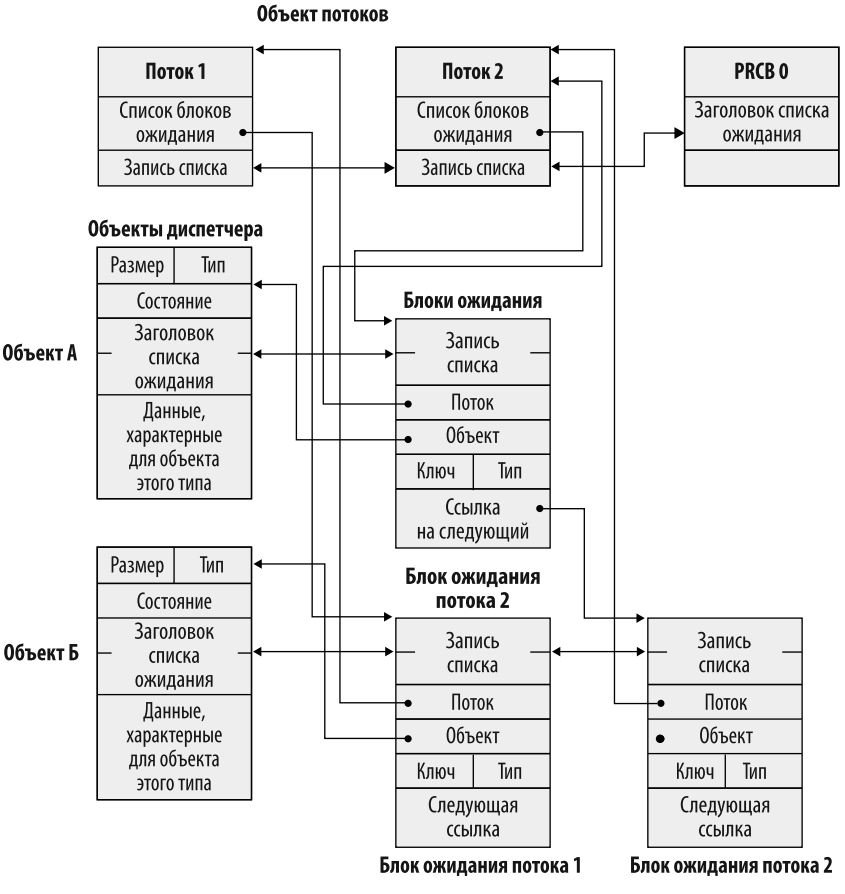


Рис. 8.39. Структуры данных, необходимые для обеспечения ожидания

ЭКСПЕРИМЕНТ. Просмотр очередей ожидания

Вы можете просмотреть список объектов, ожидаемых потоком, с помощью команды отладчика ядра `!thread`.

Например, следующий фрагмент вывода команды `!process` показывает, что поток ожидает объект события:

```
lkd> !process 0 4 explorer.exe

    THREAD ffff898f2b345080 Cid 27bc.137c Teb: 00000000006ba000
    Win32Thread: 0000000000000000 WAIT: (UserRequest) UserMode Non-Alertable
            ffff898f2b64ba60 SynchronizationEvent
```

Используйте команду dx, чтобы интерпретировать заголовок диспетчера объекта следующим образом:

```
lkd> dx (nt!_DISPATCHER_HEADER*)0xffff898f2b64ba60
(nt!_DISPATCHER_HEADER*)0xffff898f2b64ba60: 0xffff898f2b64ba60 [Type: _DISPATCHER_
HEADER*]
    [+0x000] Lock             : 393217 [Type: long]
    [+0x000] LockNV           : 393217 [Type: long]
    [+0x000] Type             : 0x1 [Type: unsigned char]
    [+0x001] Signalling       : 0x0 [Type: unsigned char]
    [+0x002] Size             : 0x6 [Type: unsigned char]
    [+0x003] Reserved1        : 0x0 [Type: unsigned char]
    [+0x000] TimerType        : 0x1 [Type: unsigned char]
    [+0x001] TimerControlFlags : 0x0 [Type: unsigned char]
    [+0x001 ( 0: 0)] Absolute         : 0x0 [Type: unsigned char]
    [+0x001 ( 1: 1)] Wake             : 0x0 [Type: unsigned char]
    [+0x001 ( 7: 2)] EncodedTolerableDelay : 0x0 [Type: unsigned char]
    [+0x002] Hand             : 0x6 [Type: unsigned char]
    [+0x003] TimerMiscFlags   : 0x0 [Type: unsigned char]
    [+0x003 ( 5: 0)] Index            : 0x0 [Type: unsigned char]
    [+0x003 ( 6: 6)] Inserted         : 0x0 [Type: unsigned char]
    [+0x003 ( 7: 7)] Expired          : 0x0 [Type: unsigned char]
    [+0x000] Timer2Type       : 0x1 [Type: unsigned char]
    [+0x001] Timer2Flags      : 0x0 [Type: unsigned char]
    [+0x001 ( 0: 0)] Timer2Inserted   : 0x0 [Type: unsigned char]
    [+0x001 ( 1: 1)] Timer2Expiring   : 0x0 [Type: unsigned char]
    [+0x001 ( 2: 2)] Timer2CancelPending : 0x0 [Type: unsigned char]
    [+0x001 ( 3: 3)] Timer2SetPending : 0x0 [Type: unsigned char]
    [+0x001 ( 4: 4)] Timer2Running    : 0x0 [Type: unsigned char]
    [+0x001 ( 5: 5)] Timer2Disabled   : 0x0 [Type: unsigned char]
    [+0x001 ( 7: 6)] Timer2ReservedFlags : 0x0 [Type: unsigned char]
    [+0x002] Timer2ComponentId : 0x6 [Type: unsigned char]
    [+0x003] Timer2RelativeId : 0x0 [Type: unsigned char]
    [+0x000] QueueType        : 0x1 [Type: unsigned char]
    [+0x001] QueueControlFlags : 0x0 [Type: unsigned char]
    [+0x001 ( 0: 0)] Abandoned        : 0x0 [Type: unsigned char]
    [+0x001 ( 1: 1)] DisableIncrement : 0x0 [Type: unsigned char]
    [+0x001 ( 7: 2)] QueueReservedControlFlags : 0x0 [Type: unsigned char]
    [+0x002] QueueSize        : 0x6 [Type: unsigned char]
    [+0x003] QueueReserved    : 0x0 [Type: unsigned char]
    [+0x000] ThreadType       : 0x1 [Type: unsigned char]
    [+0x001] ThreadReserved   : 0x0 [Type: unsigned char]
    [+0x002] ThreadControlFlags : 0x6 [Type: unsigned char]
    [+0x002 ( 0: 0)] CycleProfiling   : 0x0 [Type: unsigned char]
    [+0x002 ( 1: 1)] CounterProfiling : 0x1 [Type: unsigned char]
    [+0x002 ( 2: 2)] GroupScheduling  : 0x1 [Type: unsigned char]
    [+0x002 ( 3: 3)] AffinitySet      : 0x0 [Type: unsigned char]
    [+0x002 ( 4: 4)] Tagged           : 0x0 [Type: unsigned char]
    [+0x002 ( 5: 5)] EnergyProfiling  : 0x0 [Type: unsigned char]
    [+0x002 ( 6: 6)] SchedulerAssist  : 0x0 [Type: unsigned char]
    [+0x002 ( 7: 7)] ThreadReservedControlFlags : 0x0 [Type: unsigned char]
    [+0x003] DebugActive      : 0x0 [Type: unsigned char]
    [+0x003 ( 0: 0)] ActiveDR7        : 0x0 [Type: unsigned char]
    [+0x003 ( 1: 1)] Instrumented     : 0x0 [Type: unsigned char]
    [+0x003 ( 2: 2)] Minimal          : 0x0 [Type: unsigned char]
    [+0x003 ( 5: 3)] Reserved4        : 0x0 [Type: unsigned char]
    [+0x003 ( 6: 6)] UmsScheduled     : 0x0 [Type: unsigned char]
    [+0x003 ( 7: 7)] UmsPrimary       : 0x0 [Type: unsigned char]
    [+0x000] MutantType       : 0x1 [Type: unsigned char]
```

```
[+0x001] MutantSize       : 0x0 [Type: unsigned char]
[+0x002] DpcActive        : 0x6 [Type: unsigned char]
[+0x003] MutantReserved   : 0x0 [Type: unsigned char]
[+0x004] SignalState      : 0 [Type: long]
[+0x008] WaitListHead     [Type: _LIST_ENTRY]
    [+0x000] Flink            : 0xffff898f2b3451c0 [Type: _LIST_ENTRY *]
    [+0x008] Blink            : 0xffff898f2b3451c0 [Type: _LIST_ENTRY *]
```

Поскольку данная структура является объединением, следует игнорировать любые значения, не соответствующие типу искомого объекта, так как они не имеют к нему отношения. К сожалению, понять, какие поля относятся к какому типу, непросто, если только не сверяться с исходным кодом ядра Windows или с комментариями в заголовочных файлах WDK. Для удобства в табл. 8.29 приводятся флаги заголовка диспетчера и объекты, к которым они относятся.

Таблица 8.29. Использование и предназначение флагов заголовка диспетчера

Флаг	Область применения	Предназначение
Type	Все объекты диспетчера	Значение из последовательности KOBJECTS, определяющее тип объекта диспетчера
Lock	Все объекты	Используется для блокировки объекта во время операций ожидания, которым нужно обновить его состояние или связи. На самом деле соответствует 7-му биту (0x80) поля Тип
Signaling	Шлюзы	При переходе шлюза в сигнальное состояние нужно повысить уровень приоритета рабочего потока
Size	События, семафоры, шлюзы, процессы	Размер объекта делится на 4, чтобы поместиться в сигнальный байт
Timer2Type	Устойчивые бездействующие таймеры	Отражение поля Тип
Timer2Inserted	Устойчивые бездействующие таймеры	Установлен, когда таймер вставлен в таблицу дескрипторов таймеров
Timer2Expiring	Устойчивые бездействующие таймеры	Установлен, пока таймер истекает
Timer2CancelPending	Устойчивые бездействующие таймеры	Установлен, пока таймер отменяется
Timer2SetPending	Устойчивые бездействующие таймеры	Установлен, пока таймер регистрируется
Timer2Running	Устойчивые бездействующие таймеры	Установлен, пока активна функция отклика на таймер
Timer2Disabled	Устойчивые бездействующие таймеры	Установлен, когда таймер заблокирован

Флаг	Область применения	Предназначение
Timer2ComponentId	Устойчивые бездействующие таймеры	Идентифицирует широко известный компонент, связанный с таймером
Timer2RelativeId	Устойчивые бездействующие таймеры	В рамках ID предыдущего компонента показывает, который это из его таймеров
TimerType	Таймеры	Отражение поля Тип
Absolute	Таймеры	Срок истечения таймера абсолютный, а не относительный
Wake	Таймеры	Это пробуждаемый таймер, при сигнализации ему нужно выходить из состояния сна
EncodedTolerableDelay	Таймеры	Максимальная погрешность (представленная в степени двойки), которую таймер может допускать при работе вне ожидаемой периодичности
Hand	Таймеры	Индекс в таблице дескрипторов таймеров
Index	Таймеры	Индекс в таблице истечения времени таймеров
Inserted	Таймеры	Установлен, если таймер был добавлен в таблицу дескрипторов таймеров
Expired	Таймеры	Установлен, если таймер уже истек
ThreadType	Потоки	Отражение поля Тип
ThreadReserved	Потоки	Не используется
CycleProfiling	Потоки	Для данного потока задействовано профилирование циклов процессора
CounterProfiling	Потоки	Для данного потока задействовано отслеживание/профилирование аппаратного счетчика производительности процессора
GroupScheduling	Потоки	Для данного потока задействованы группы планирования, например, в режиме DFSS (Distributed Fair-Share Scheduler) или с объектом задания, где используется ускорение процессора
AffinitySet	Потоки	Поток связан с набором процессоров

Продолжение ⇩

Таблица 8.29 *(продолжение)*

Флаг	Область применения	Предназначение
Tagged	Потоки	Потоку присвоен тег свойства
EnergyProfiling	Потоки	Для процесса, которому принадлежит данный поток, задействована оценка энергопотребления
SchedulerAssist	Потоки	Задействован Hyper-V XTS (eXTended Scheduler), а сам этот поток принадлежит потоку виртуального процессора (VP) в рамках минимального процесса VM
ActiveDR7	Потоки	Используются аппаратные точки останова, а значит, DR7 активен и требует очистки при операциях с контекстом. Иногда этот флаг называют DebugActive
Minimal	Потоки	Поток принадлежит минимальному процессу
AltSyscall	Потоки	Для процесса — владельца данного потока зарегистрирован альтернативный обработчик системных вызовов, таких как поставщик Pico или Windows CE PAL
UmsScheduled	Потоки	Это рабочий поток UMS (планируемый)
UmsPrimary	Потоки	Это планировочный поток UMS (основной)
MutantType	Мутанты	Отражение поля Тип
MutantSize	Мутанты	Не используется
DpcActive	Мутанты	Этот мутант был получен в ходе DPC
MutantReserved	Мутанты	Не используется
QueueType	Очереди	Отражение поля Тип
Abandoned	Очереди	Не осталось потоков, ожидающих данную очередь
DisableIncrement	Очереди	Для потоков, активизируемых с целью обработки пакета из этой очереди, не будет повышен приоритет

Наконец, заголовок диспетчера имеет также поля `SignalState` и `WaitListHead`, описанные ранее. Следует иметь в виду, что если указатели на заголовок списка

ожиданий идентичны, то либо объект не ожидает ни один поток, либо его ожидает один поток. Разницу можно заметить, если идентичный указатель оказался в списке адресов — это признак того, что ожидающий поток отсутствует. В примере ранее `0XFFFF898F2B3451C0` не было в списке адресов, поэтому мы можем выгрузить блок ожидания следующим образом:

```
lkd> dx (nt!_KWAIT_BLOCK*)0xffff898f2b3451c0
(nt!_KWAIT_BLOCK*)0xffff898f2b3451c0     : 0xffff898f2b3451c0 [Type: _KWAIT_BLOCK *]
    [+0x000] WaitListEntry    [Type: _LIST_ENTRY]
    [+0x010] WaitType         : 0x1 [Type: unsigned char]
    [+0x011] BlockState       : 0x4 [Type: unsigned char]
    [+0x012] WaitKey          : 0x0 [Type: unsigned short]
    [+0x014] SpareLong        : 6066 [Type: long]
    [+0x018] Thread           : 0xffff898f2b345080 [Type: _KTHREAD *]
    [+0x018] NotificationQueue : 0xffff898f2b345080 [Type: _KQUEUE *]
    [+0x020] Object           : 0xffff898f2b64ba60 [Type: void *]
    [+0x028] SparePtr         : 0x0 [Type: void *]
```

В данном случае тип ожидания показан как `WaitAny`, так что имеется поток, блокирующий это событие, указатель на который мы и видим. Мы также видим, что блок ожидания активен.

Далее можем исследовать несколько связанных с ожиданием полей в структуре данных потока:

```
lkd> dt nt!_KTHREAD 0xffff898f2b345080 WaitRegister.State WaitIrql WaitMode
  WaitBlockCount WaitReason WaitTime
   +0x070 WaitRegister       :
   +0x000 State              : 0y001
   +0x186 WaitIrql           : 0 ''
   +0x187 WaitMode           : 1 ''
   +0x1b4 WaitTime           : 0x39b38f8
   +0x24b WaitBlockCount     : 0x1 ''
   +0x283 WaitReason         : 0x6 ''
```

Эти данные говорят о том, что перед нами зафиксированное ожидание, которое было запущено на IRQL 0 (Passive) с режимом ожидания в пользовательском режиме 15 мс с момента загрузки системы, а причиной ожидания выступает запрос со стороны приложения пользовательского режима. Мы также можем увидеть, что это единственный блок ожидания у данного потока, а значит, никаких других объектов он не ждет.

Если в списке ожидания более одной записи, можно выполнить ту же самую команду для значения второго указателя в поле `WaitListEntry` каждого блока ожидания (путем выполнения команды `!thread` в отношении указателя потока в блоке ожидания), чтобы пройти по списку и посмотреть, какие еще потоки ожидают объект. Если эти потоки дожидались более чем одного объекта, вам придется посмотреть на их поле `WaitBlockCount`, чтобы увидеть, сколько еще блоков ожидания там в наличии, и просто увеличивать указатель на `sizeof(KWAIT_BLOCK)`.

События с ключом

Объект синхронизации, называемый *событием с ключом*, заслуживает особого упоминания ввиду роли, которую он играет в примитивах синхронизации в пользовательском режиме и разработке примитива `alert-by-ID`, который, как вы скоро заметите, является реализованным для Windows аналогом *фьютекса* из операционной системы Linux (широко известный в информатике концепт). События с ключом изначально были реализованы, чтобы помочь процессам справляться с ситуацией дефицита памяти при использовании критических разделов, и они представляют собой объекты синхронизации пользовательского режима. Событие с ключом, которое не задокументировано, позволяет потоку указать «ключ», который он ожидает, и поток возобновляет выполнение, когда другой поток того же процесса сигнализирует о событии с таким же ключом. Как мы уже отмечали, если это похоже на механизм оповещений, то это потому, что события с ключом легли в его основу.

Если возникает соперничество, функция `EnterCriticalSection` динамически размещает объект события, и поток, ожидающий получения критического раздела, ждет, пока тот поток, который владеет критическим разделом, не выдаст ему сигнал, вызвав функцию `LeaveCriticalSection`. В условиях дефицита памяти появляется очевидная проблема: получение критического раздела может не состояться из-за того, что системе не удалось разместить требуемый объект события. В крайнем случае нехватка памяти сама по себе может быть вызвана приложением, пытающимся получить критический раздел, и тогда в системе случится взаимная блокировка. Дефицит памяти не единственный сценарий, который может привести к подобному отказу: менее вероятным сценарием может стать дефицит дескрипторов. Если процесс достигнет своего лимита дескрипторов, в новом дескрипторе для события может быть отказано.

Может показаться, что решить проблему позволит предварительное выделение глобального стандартного *объекта события*. Однако, поскольку процесс может иметь несколько критических разделов, каждый со своим состоянием блокировки, это может потребовать неизвестно сколько заранее созданных объектов события, поэтому данное решение не подходит. Основным преимуществом событий с ключом было то, что одно и то же событие могло быть использовано несколькими различными потоками при условии, что каждый предоставит ключ, чтобы себя уникально обозначить. Предоставление напрямую самому критическому разделу в качестве ключа виртуального адреса позволяет нескольким критическим разделам (а значит, и тем, кто их ожидает) пользоваться одним и тем же дескриптором события с ключом, который можно заранее разместить при запуске процесса.

Когда поток посылает сигнал о событии с ключом или ждет этого события, он использует уникальный идентификатор, называемый *ключом*, который идентифицирует экземпляр события с ключом (ассоциируя событие с ключом с конкретным критическим разделом). Когда поток-владелец освобождает событие с ключом, переводя его в сигнальное состояние, возобновляется выполнение только один поток, ожидавший этот ключ (такое же поведение, как у *событий синхронизации*, в отличие от *событий оповещений*). В нашем варианте, где критические разделы используют свой адрес в роли ключа, это будет подразумевать, что каждый процесс все

еще нуждается в собственном событии с ключом, поскольку виртуальные адреса, очевидно, уникальны для адресного пространства конкретного процесса. Однако так вышло, что ядро способно активизировать ожидающих только в текущем процессе, а значит, этот ключ изолирован даже между процессами, позволяя обойтись лишь одним событием с ключом на всю систему.

Таким образом, когда функция `EnterCriticalSection` вызывает `NtWaitForKeyedEvent`, чтобы выполнить ожидание события с ключом, она может дать в качестве параметра для события с ключом дескриптор `NULL`, сообщая ядру, что не может создать такое событие. Ядро распознает такое поведение и использует глобальное событие с ключом `ExpCritSecOutOfMemoryEvent`. Основное преимущество здесь в том, что процессам больше не нужно тратить дескриптор на именованное событие с ключом, поскольку отслеживанием этого объекта и ссылок на него занимается ядро.

Однако события с ключом служат не только альтернативными объектами, применяемыми при дефиците памяти. Когда один и тот же ключ ожидается сразу несколькими потоками, которые нуждаются в возобновлении своего выполнения, о ключе, фактически, сообщается многократно, что требует от объекта вести список всех ожидающих потоков, чтобы в отношении каждого из них можно было провести операцию «пробуждения» (вспомним, что выдача сигнала о событии с ключом приводит к тому же, что и выдача сигнала о событии синхронизации). Но поток может отправить сигнал о событии с ключом и без потоков в списке ожиданий. В таком случае поток, выдающий сигнал, сам ждет это событие.

Без такой альтернативы поток, выдающий сигнал, может просигналить о событии с ключом в то самое время, когда код пользовательского режима увидел событие в несигнальном состоянии и попытался перейти в режим ожидания. Ожидание, возможно, наступило после того, как сигнализирующий поток ввел событие с ключом в сигнальное состояние, в результате чего пропустил импульс, следовательно, ожидающий поток попадает в тупиковую ситуацию. Если заставить сигнализирующий поток ждать по этому сценарию, он, фактически, выдаст сигнал о событии с ключом только тогда, когда кто-то ищет этот сигнал (ждет его). Это поведение сделало такие события похожими на *фьютексы* из Linux, но не идентичными им, позволив задействовать их в ряде примитивов пользовательского режима, таких как SRW-блокировки (Slim Read Writer Locks).

ПРИМЕЧАНИЕ Когда код, использующий ожидание событий с ключом, сам нуждается в ожидании, он задействует встроенный семафор, расположенный в объекте потока режима ядра (ETHREAD), под названием KeyedWaitSemaphore. (Он фактически делит свое местоположение с семафором ожидания ALPC.) Подробности об объектах потоков см. в главе 4 тома 1.

Тем не менее в реализации критических разделов события с ключом не заменили стандартных объектов событий. Изначально причина состояла в том, что во времена Windows XP события с ключом не предлагали масштабируемой производительности в часто используемых сценариях. Напомним, что все описанные ранее алгоритмы предназначались ны только для критических сценариев, реализуемых в условиях нехватки ресурсов, когда производительность и масштабируемость

были не так важны. Попытка заменить ими стандартные объекты возложила бы на события с ключом нагрузку, для обработки которой они не предназначались. Основным узким местом в вопросах производительности было то, что события с ключом вели список ожидающих потоков, который реализовывался в виде списка с двойной связью. Списки такого рода имеют *низкую скорость прохода* по элементам, то есть на перебор элементов списка требуется много времени. В данном случае это время зависит от количества ожидающих потоков. Поскольку объект является глобальным, в списке могут быть десятки потоков, требуя большого времени прохода при каждой установке ключа или ожидании события с ключом.

ПРИМЕЧАНИЕ Начало списка хранится в самом объекте события с ключом, в то время как потоки привязаны через поле KeyedWaitChain (оно делит место с временем выхода из потока, которое хранится как LARGE_INTEGER, размером как раз с элемент двунаправленного списка) в объекте потока режима ядра (ETHREAD). Подробности об этом объекте см. в главе 4 тома 1.

В Windows Vista производительность событий с ключом была улучшена за счет применения для хранения ожидающих потоков хеш-таблицы взамен связанного списка. Именно эта оптимизация в итоге позволила Windows обзавестись тремя новыми облегченными примитивами синхронизации пользовательского режима (о них чуть позже), которые полагались на эти события. Однако критические разделы продолжили пользоваться стандартными объектами событий в основном для совместимости приложений и отладки, поскольку они и их внутреннее устройство широко известны и документированы, а события с ключом непрозрачны и недоступны для Win32 API.

Однако с введением в Windows 8 новых возможностей за счет оповещения по ID потока все снова изменилось и события с ключом перестали применяться в масштабах системы, не считая одной ситуации при синхронизации Init Once (об этом немного позже). Со временем сами критические разделы перестали пользоваться обычными объектами событий и тоже перешли на эту новую функцию (с опцией режима совместимости приложений, позволяющей вернуться к применению оригинальных объектов событий).

Быстрые и защищенные мьютексы

Быстрые мьютексы, также известные как мьютексы исполнительной системы, обычно обеспечивают лучшую производительность, чем их обычные аналоги, поскольку, хотя они все еще построены на основе объектах события диспетчера, они реализуют ожидание через диспетчер, только если вступают в состязание, в отличие от стандартного мьютекса, который всегда пытается получить доступ через диспетчер. Это дает быстрому мьютексу особенно высокую производительность в конкурентных средах. Быстрые мьютексы широко применяются в драйверах устройств.

Однако подобная эффективность имеет свою цену. Быстрые мьютексы подходят для работы, лишь когда доставку всех APC-вызовов режима ядра (описаны ранее в данной главе) можно заблокировать, в отличие от обычных, которые блокируют

только *обычные* APC. С учетом этого в исполнительной подсистеме определены две функции для их получения: `ExAcquireFastMutex` и `ExAcquireFastMutexUnsafe`. Первая из них блокирует доставку всех APC, поднимая уровень IRQL этого процессора до APC. Вторая же, небезопасная функция ожидает, что при ее вызове доставка всех APC уровня ядра уже заблокирована, что можно сделать, повысив уровень IRQL до `APC_LEVEL`. Функция `ExTryToAcquireFastMutex` действует подобно первой функции, но на самом деле, если мьютекс уже занят, даже не начинает ожидания, вместо этого возвращая `FALSE`. Другое ограничение быстрых мьютексов — невозможность получать их рекурсивно, тогда как для обычных мьютексов это позволено.

Начиная с Windows 8, защищенные мьютексы идентичны быстрым, но доступ к ним получают с помощью вызова `KeAcquireGuardedMutex` и `KeAcquireGuardedMutexUnsafe`. Как и у аналога, у них есть метод `KeTryToAcquireGuardedMutex`.

До Windows 8 эти функции блокировали APC не путем поднятия IRQL до уровня `APC_LEVEL`, а с помощью входа в особую защищенную область, где в структуре объекта потока устанавливались особые счетчики, блокировавшие их доставку, пока функции не покинут эту область. В устаревших системах с PIC (о котором рассказывалось ранее в главе) такое решение работало быстрее, чем изменение IRQL. Кроме того, защищенные мьютексы пользовались объектом диспетчера шлюза, что было чуть быстрее, чем событие, — еще одно более не актуальное отличие.

Дополнительной проблемой, связанной с защищенным мьютексом, была функция ядра `KeAreApcsDisabled`. До Windows Server 2003 она отображала, заблокированы ли нормальные APC, проверяя, не исполнялся ли в тот момент код в критическом разделе. После выхода этой системы функция была изменена и стала сообщать, выполняется ли код в критическом разделе или защищенной области, возвращая `TRUE`, если особые APC ядра тоже отключены.

Во время запрета особых APC ядра некоторые операции драйверам выполнять не следует, поэтому имело смысл определять, находится ли IRQL на уровне `APC_LEVEL`, с помощью функции `KeGetCurrentIrql`. Это был единственный способ выяснить, заблокированы ли особые APC ядра. Однако с появлением защищенных мьютексов и областей, которые постоянно использовались даже диспетчером памяти, проверка перестала работать, так как защищенные мьютексы не повышали IRQL. Тогда драйверам приходилось полагаться на функцию `KeAreAllApcsDisabled`, которая также проверяла, запрещены ли особые APC ядра, путем входа в защищенную область. Подобные нестыковки в сочетании с ненадежными проверками, выполняемыми средством проверки драйверов (Driver Verifier), выдававшим ложные срабатывания, в итоге привели к решению просто отказаться от защищенных мьютексов в пользу прежних быстрых.

Ресурсы исполнительной системы

Ресурсы исполнительной системы — это механизм синхронизации, который поддерживает общий и эксклюзивный доступ. Как и быстрые мьютексы, они требуют, чтобы доставка APC режима ядра была приостановлена прежде, чем станет можно их получить. Они также построены на объектах диспетчера, которые используются только при наличии конкуренции. Ресурсы исполнительной системы используются

по всей ОС, особенно в драйверах файловой системы, поскольку такие драйверы склонны к продолжительным периодам ожидания, во время которых должны быть по-прежнему до определенной степени разрешены операции ввода-вывода (например, для чтения).

Потоки, ожидающие получения ресурса исполняющей системы для совместного доступа, ждут семафора, связанного с ресурсом, а потоки, ожидающие получения ресурса исполняющей системы для исключающего доступа, ждут события. Семафор с неограниченным счетчиком используется для потоков, ожидающих совместного доступа, поскольку все они могут быть разбужены и им может быть предоставлено право доступа к ресурсу, когда исключающий держатель ресурса освобождает этот ресурс, путем простой выдачи сигнала семафора. Когда поток ждет исключающего доступа к ресурсу, который в настоящий момент находится в чьем-нибудь владении, он находится в ожидании объекта события синхронизации, поскольку только один из ожидающих потоков будет разбужен, когда поступит сигнал о событии. В разделе, посвященном событиям синхронизации, упоминалось, что некоторые операции, не ожидающие события, могут стать причиной повышения уровня приоритета: этот сценарий разыгрывается, когда используются ресурсы исполняющей системы, что является одной из причин, почему они также, подобно мьютексам, отслеживают принадлежность (подробности о повышении приоритета исполнительного ресурса см. в главе 4 тома 1).

Благодаря той гибкости, которая предоставляется совместным и исключающим доступом, существует несколько функций для получения ресурсов: `ExAcquireResourceSharedLite`, `ExAcquireResourceExclusiveLite`, `ExAcquireSharedStarveExclusive` и `ExAcquireShareWaitForExclusive`. Они задокументированы в WDK.

В последних версиях Windows также появились быстрые ресурсы исполнительной системы с такими же именами с добавлением слова fast: `ExAcquireFastResourceExclusive`, `ExReleaseFastResource` и т. д. Они должны были заменить прежние функции, обеспечивая ускорение за счет нового способа управления блокировками, однако их не использует ни один компонент, кроме файловой системы ReFS (Resilient File System). Во время периодов высокой конкуренции доступа к файловой системе ReFS дает производительность чуть выше, чем у NTFS, отчасти за счет более быстрых блокировок.

ЭКСПЕРИМЕНТ. Вывод ресурсов исполнительной системы

Команда отладчика ядра `!locks` обращается к связному списку ядра с ресурсами исполнительной системы и выгружает их состояние. По умолчанию команда выводит только ресурсы, в данный момент имеющие владельца, но в документации упоминается флаг `-d`, позволяющий показать их все. К сожалению, это больше не так. Однако вы все еще можете использовать вместо него флаг `-v`, чтобы выгрузить подробную информацию обо всех ресурсах. Далее приводится фрагмент вывода такой команды:

```
lkd> !locks -v
**** DUMP OF ALL RESOURCE OBJECTS ****

Resource @ nt!ExpFirmwareTableResource (0xfffff8047ee34440)    Available
Resource @ nt!PsLoadedModuleResource (0xfffff8047ee48120)    Available
```

```
    Contention Count = 2
Resource @ nt!SepRmDbLock (0xfffff8047ef06350)    Available
    Contention Count = 93
Resource @ nt!SepRmDbLock (0xfffff8047ef063b8)    Available
Resource @ nt!SepRmDbLock (0xfffff8047ef06420)    Available
Resource @ nt!SepRmDbLock (0xfffff8047ef06488)    Available
Resource @ nt!SepRmGlobalSaclLock (0xfffff8047ef062b0)    Available
Resource @ nt!SepLsaAuditQueueInfo (0xfffff8047ee6e010)    Available
Resource @ nt!SepLsaDeletedLogonQueueInfo (0xfffff8047ee6ded0)    Available
Resource @ 0xffff898f032a8550    Available
Resource @ nt!PnpRegistryDeviceResource (0xfffff8047ee62b00)    Available
    Contention Count = 27385
Resource @ nt!PopPolicyLock (0xfffff8047ee458c0)    Available
    Contention Count = 14
Resource @ 0xffff898f032a8950    Available
Resource @ 0xffff898f032a82d0    Available
```

Отметим, что счетчик конкуренции (contention count), который извлекается из структуры ресурса, отображает количество раз, когда потоки пытались этот ресурс получить, но вынуждены были ждать, потому что тот был занят. Попытавшись вклиниться отладчиком в работающую систему, вы при определенном везении сможете лишь заметить несколько занятых ресурсов, как видно из следующего вывода:

```
2: kd> !locks
**** DUMP OF ALL RESOURCE OBJECTS ****
KD: Scanning for held locks.....

Resource @ 0xffffde07a33d6a28 Shared 1 owning threads
    Contention Count = 28
        Threads: ffffde07a9374080-01<*>
KD: Scanning for held locks....

Resource @ 0xffffde07a2bfb350 Shared 1 owning threads
    Contention Count = 2
        Threads: ffffde07a9374080-01<*>
KD: Scanning for held locks.............................................

Resource @ 0xffffde07a8070c00 Shared 1 owning threads
        Threads: ffffde07aa3f1083-01<*> *** Actual Thread ffffde07aa3f1080
KD: Scanning for held locks.............................................

Resource @ 0xffffde07a8995900 Exclusively owned
        Threads: ffffde07a9374080-01<*>
KD: Scanning for held locks.............................................
    9706 total locks, 4 locks currently held
```

Вы можете рассмотреть детали конкретного объекта ресурса, включая информацию о потоке, который им владеет, и всех потоках, ожидающих его, указав флаг `-v` и адрес ресурса, если сможете отловить его занятым (имеющим владельца). Например, далее приводится занятый общий ресурс, который, по всей видимости, связан с NTFS, в то время как какой-то поток делает попытку чтения из файловой системы:

```
2: kd> !locks -v 0xffffde07a33d6a28
```

```
Resource @ 0xffffde07a33d6a28 Shared 1 owning threads
    Contention Count = 28
        Threads: ffffde07a9374080-01<*>

        THREAD ffffde07a9374080 Cid 0544.1494 Teb: 000000ed8de12000
        Win32Thread: 0000000000000000 WAIT: (Executive) KernelMode Non-Alertable
            ffff8287943a87b8 NotificationEvent
        IRP List:
            ffffde07a936da20: (0006,0478) Flags: 00020043 Mdl: ffffde07a8a75950
            ffffde07a894fa20: (0006,0478) Flags: 00000884 Mdl: 00000000
        Not impersonating
        DeviceMap                 ffff8786fce35840
        Owning Process            ffffde07a7f990c0      Image:         svchost.exe
        Attached Process          N/A            Image:       N/A
        Wait Start TickCount      3649           Ticks: 0
        Context Switch Count      31             IdealProcessor: 1
        UserTime                  00:00:00.015
        KernelTime                00:00:00.000
        Win32 Start Address 0x00007ff926812390
        Stack Init ffff8287943aa650 Current ffff8287943a8030
        Base ffff8287943ab000 Limit ffff8287943a4000 Call 0000000000000000
        Priority 7 BasePriority 6 PriorityDecrement 0 IoPriority 0 PagePriority 1
        Child-SP          RetAddr           Call Site
        ffff8287`943a8070 fffff801`104a423a nt!KiSwapContext+0x76
        ffff8287`943a81b0 fffff801`104a5d53 nt!KiSwapThread+0x5ba
        ffff8287`943a8270 fffff801`104a6579 nt!KiCommitThreadWait+0x153
        ffff8287`943a8310 fffff801`1263e962 nt!KeWaitForSingleObject+0x239
        ffff8287`943a8400 fffff801`1263d682 Ntfs!NtfsNonCachedIo+0xa52
        ffff8287`943a86b0 fffff801`1263b756 Ntfs!NtfsCommonRead+0x1d52
        ffff8287`943a8850 fffff801`1049a725 Ntfs!NtfsFsdRead+0x396
        ffff8287`943a8920 fffff801`11826591 nt!IofCallDriver+0x55
```

Push-блокировки

Push-блокировки — это еще один оптимизированный механизм синхронизации, построенный на объектах событий: как и в случае с быстрыми и защищенными мьютексами, ожидание события начинается только при наличии конкуренции. Однако в сравнении с последними у них есть преимущество: их тоже можно получать в общий или эксклюзивный доступ, подобно ресурсам исполнительной системы. Но push-блокировки выгодно выделяются своим размером: объект ресурса занимает 104 байт, в то время как push-блокировка имеет размер указателя. Из-за этого им не требуются размещение и инициализация, а их работоспособность в условиях нехватки памяти гарантирована. Множество компонентов в рамках ядра перешли с ресурсов исполнительной системы на push-блокировки. Современные драйверы от сторонних разработчиков тоже ими пользуются.

Существует четыре типа push-блокировок: обычные (normal), осведомленные о кэш-памяти (cache-aware), авторасширяемые (auto-expand) и адресно базированные (address-based). Обычным push-блокировкам нужна память размером с указатель (4 байта на 32-разрядных и 8 байт — на 64-разрядных системах). Когда поток получает обычную push-блокировку, которая еще не занята, ответственный код помечает ее как имеющую владельца. Если же та уже находится в эксклюзивном

владении или поток претендует на эксклюзивный доступ к push-блокировке, а она занята в совместном использовании, то поток размещает в стеке потока блок ожидания, в котором инициализирует объект события, а ее саму помещает в список ожидания, связанный с этой push-блокировкой. Когда какой-то поток освободит push-блокировку, он разбудит ожидающий поток, если он есть, переведя в сигнальное состояние объект события, имеющийся в блоке ожидания потока.

Поскольку push-блокировка имеет размер указателя, она фактически содержит различные биты, описывающие ее состояние. Значения этих битов изменяются в зависимости от перехода push-блокировки из конкурирующего в неконкурирующее состояние. В своем исходном состоянии push-блокировка содержит следующую структуру:

- один бит блокировки, установленный в 1, если блокировка получена;
- один бит ожидания, установленный в 1, если за блокировку возник спор и кто-то ее дожидается;
- один бит пробуждения, установленный в 1, если блокировка была предоставлена какому-то потоку и список ожидающих требуется оптимизировать;
- один бит множественного доступа, установленный в 1, если push-блокировка в общем доступе и на данный момент получена более чем одним потоком;
- 28 (в 32-разрядных Windows) или 60 (в 64-разрядных Windows) бит счетчика доступа, в котором содержится число потоков, получивших эту push-блокировку.

Как говорилось ранее, когда поток претендует на эксклюзивное владение push-блокировкой, которая уже занята совместным чтением или кем-то одним для записи, ядро размещает блок ожидания push-блокировки. При этом изменяется сама структура значения push-блокировки. Биты счетчика общего доступа теперь становятся указателем на блок ожидания. Поскольку этот блок находится в стеке, а заголовочные файлы содержат особую директиву выравнивания, заставляя его выровняться по 16-битному формату, последние 4 бита структуры любой push-блокировки с блоком ожидания будут всегда содержать нули. Таким образом, эти биты игнорируются из соображений разыменования указателя, вместо этого 4 бита, упомянутые ранее, комбинируются со значением указателя. Поскольку в результате этого выравнивания удаляются биты счетчика совместного использования, теперь этот счетчик хранится в блоке ожидания.

Push-блокировка, *осведомленная о кэш-памяти*, добавляет к обычной (основной) push-блокировке уровни путем размещения push-блокировки для каждого процессора в системе и связывания ее с push-блокировкой, осведомленной о кэш-памяти. Когда потоку нужно получить кэш-блокироку, осведомленную о кэш-памяти, для совместного доступа, он просто получает push-блокировку, размещенную для его текущего процессора в режиме совместного использования; для получения эксклюзивной push-блокировки, осведомленной о кэш-памяти, поток запрашивает эксклюзивный доступ для ее аналога у каждого процессора.

Однако, как можете представить, в современных условиях, когда Windows поддерживает системы с числом процессоров до 2560, количество потенциальных выровненных в кэше слотов для push-блокировки, осведомленной о кэш-памяти, потребовало бы огромных фиксированных размещений, даже когда процессоров

немного. Поддержка динамического добавления процессоров при работающей системе без ее остановки еще больше усложняет задачу, поскольку технически потребуется заранее выделить место под все 2560 слотов, создавая структуры для блокировок на много килобайтов. Для ее решения в современных версиях Windows реализована авторасширяемая push-блокировка. Как следует из названия, push-блокировки этого типа способны по мере необходимости динамически наращивать число слотов кэша, исходя как из наличия конкуренции, так и из числа процессоров и при этом гарантируя дальнейший прогресс, поскольку пользуется разместителем слотов в рамках исполнительной системы, который заранее резервирует пространство в подкачиваемом или неподкачиваемом пуле (в зависимости от флагов, переданных при размещении авторасширяемой push-блокировки).

К сожалению сторонних разработчиков, push-блокировки, осведомленные о кэш-памяти (и более новые, авторасширяемые), официально не документированы для использования, даже притом что некоторые структуры данных, такие как заголовки FCB в Windows 10 21H1 и далее, незаметно ими пользуются (подробности о структуре FCB см. в главе 11). К внутренним компонентам ядра, применяющим авторасширяемые push-блокировки, относится диспетчер памяти, в котором те используются для защиты структур данных интерфейса Address Windowing Extension (AWE).

Наконец, еще одним типом недокументированной, но экспортированной push-блокировки является адресно базированная push-блокировка, реализация которой сводится к механизму, похожему на адресно базированное ожидание, которое мы скоро рассмотрим в пользовательском режиме. Кроме отнесения к отдельному виду, такая push-блокировка больше отличается способом ее применения. С одной стороны, вызывающий поток задействует `ExBlockOnAddressPushLock`, передавая push-блокировку, виртуальный адрес какой-то интересующей переменной, ее размер (до 8 байт) и адрес для сравнения, по которому находится ожидаемое или желаемое значение этой переменной. Если ее значение на данный момент не равно искомому, с помощью вызова `ExTimedWaitForUnblockPushLock` инициализируется ожидание. Это поведение похоже на конкурентное получение push-блокировки, с той разницей, что может быть задано значение тайм-аута. В то же время кто-то вызывает `ExUnblockOnAddressPushLockEx` после внесения изменений по наблюдаемому адресу, чтобы подать сигнал ожидающему об обновлении значения. Данная техника особенно полезна, когда имеются изменения в данных, защищенных блокировкой, или в рамках блокирующей операции, что позволяет участвующим в гонке читателям получить от записавшего уведомление о завершении изменений извне блокировки. Наряду с намного меньшими потребностями в памяти одним из крупных преимуществ push-блокировок перед ресурсами исполнительной системы является то, что при отсутствии конкуренции им не требуются учет использования и целочисленные вычисления при получении и освобождении. Из-за их размера, равного указателю, ядро может выполнять эти задачи с помощью атомарных инструкций процессора (например, на процессорах x86 и x64 используется `lock cmpxchg`, которая атомарно сравнивает старую блокировку с новой). Если атомарное сравнение-обмен не срабатывает, блокировка содержит значение, которого вызывающий не ожидает (обычно предполагают, что блокировка будет не занята или находится в общем доступе), в силу чего происходит обращение к более сложному конкурентному варианту.

Чтобы еще значительнее улучшить производительность, ядро предоставляет функциональность push-блокировок в виде вложенных функций, тем самым при отсутствии конкуренции на доступ вообще обходясь без формирования вызовов — ассемблерный код вставляется в каждую функцию напрямую. Размер кода от этого слегка увеличивается, зато удается избежать затрат времени на вызов процедуры. Наконец, в push-блокировках используются несколько алгоритмических трюков, позволяющих избежать очередей на блокировке — ситуаций, когда несколько потоков с одинаковым приоритетом вместе ждут какую-то блокировку, почти ничего не делая по существу. Кроме того, в них предусмотрена самооптимизация: список потоков, ожидающих push-блокировку, будет периодически переупорядочиваться для обеспечения более справедливого поведения при освобождении.

Еще одной оптимизацией производительности, применимой к получению push-блокировок (в том числе адресно базированных), является оппортунистическое поведение при конкуренции, сходное с поведением спин-блокировки, наблюдающееся перед тем, как объект события диспетчера станет ожидать блок ожидания push-блокировки. Если в системе есть хоть один еще не припаркованный процессор (подробности о парковке ядер см. в главе 4 тома 1), ядро входит в круговой цикл на `ExpSpinCycleCount` коротких фаз, подобно спин-блокировке, но без повышения IRQL, выполняя инструкцию-передышку (например, `pause` для x86/x64) для каждой итерации. Если в ходе любой из итераций окажется, что push-блокировка освободилась, выполняется блокирующая операция ее получения.

Если тайм-аут фаз цикла заканчивается, блокирующая операция получает отказ (из-за гонки) или в системе нет хотя бы одного лишнего неприпаркованного процессора, к объекту-событию в блоке ожидания push-блокировки применяется `KeWaitForSingleObject`. Значение `ExpSpinCycleCount` приравнивается к 10 240 циклам на любом компьютере с более чем одним логическим процессором, и его нельзя настроить. Для систем с процессором AMD, в котором реализована спецификация MWAITT (MWAIT Timer), вместо спин-цикла используются инструкции `monitorx` и `mwaitx`. Эти аппаратные функции позволяют на уровне процессора ожидать изменения значения по какому-либо адресу без необходимости входить в цикл, но при этом принимают параметр с тайм-аутом, чтобы ожидание не получалось бесконечным (ядро определяет его по значению `ExpSpinCycleCount`).

И напоследок: с появлением технологии AutoBoost (описывается в главе 4 тома 1) push-блокировки по умолчанию пользуются ее возможностями, кроме случаев, когда вызываются новые функции вида `ExXxxPushLockXxxEx`, принимающие флаг `EX_PUSH_LOCK_FLAG_DISABLE_AUTOBOOST`, который эту функциональность отключает (не документировано официально). По умолчанию все функции без `Ex` теперь вызывают новые версии с `Ex`, но без этого флага.

Адресно базированные ожидания

Уроки, полученные из практики событий с ключом, обусловили то, что основным примитивом синхронизации, предоставляемым Windows в пользовательском режиме, теперь является системный вызов `alert-by-ID` (и его обратный вариант `wait-on-alert-by-ID`). С помощью этих двух простых функций, которым не требуется ни выделения памяти, ни дескрипторов, можно установить любое количество синхронизаций в рамках процесса, к которым относится механизм адресно

базированного ожидания. На нем основаны другие примитивы, например критические разделы и SRW-блокировки.

Адресно базированное ожидание основывается на трех документированных функциях Win32 API: `WaitOnAddress`, `WakeByAddressSingle` и `WakeByAddressAll`. Они размещены в KernelBase.dll, но на деле являются лишь оболочками реальных реализаций из Ntdll.dll, где те размещены с похожими именами с префиксом `Rtl`, что значит Run Time Library (библиотека среды выполнения). Функция `Wait` принимает на вход адрес, указывающий на интересующее значение, размер значения (до 8 байт), адрес нежелательного значения и тайм-аут. Функция `Wake` принимает только адрес.

В первую очередь `RtlWaitOnAddress` готовит локальный блок ожидания адреса, отслеживающий ID потока и адрес, и помещает его в персональную для процесса хеш-таблицу, расположенную в блоке окружения процесса (Process Environment Block, PEB). Эти действия аналогичны поведению `ExBlockOnAddressPushLock`, которое мы обсуждали ранее, за исключением того, что там хеш-таблица была не нужна, поскольку вызывающему нужно было где-то хранить указатель push-блокировки. Затем, подобно API ядра, `RtlWaitOnAddress` проверяет, не содержится ли по искомому адресу значение, отличное от нежелательного, и, если это так, удаляет блок ожидания адреса, возвращая `FALSE`. В противном случае применяется внутренняя функция блокировки.

Если доступен более чем один неприпаркованный процессор, блокирующая функция сначала попытается избежать перехода в ядро, выполняя циклический просмотр в пользовательском режиме, чтобы проверить значение бита блока ожидания адреса, отвечающего за доступность. Количество фаз цикла определяется значением `RtlpWaitOnAddressSpinCount`, на уровне кода приравниваемым к 1024, если в системе больше одного процессора. Если блок ожидания все еще показывает конкуренцию, в отношении ядра выполняется системный вызов функцией `NtWaitForAlertByThreadId`, куда передается искомый адрес в качестве параметра-подсказки вместе с тайм-аутом.

Если функция возвращает управление по причине тайм-аута, то в блоке ожидания адреса устанавливается флаг, блокировка снимается, а возвращаемым результатом оказывается `STATUS_TIMEOUT`. Однако существует незначительная вероятность состояния гонки, при котором вызвавшая сторона могла воспользоваться функцией `Wake` за несколько циклов до истечения тайм-аута ожидания. Поскольку флаг блока ожидания изменяется инструкцией сравнения-обмена, код может заметить это и на деле выполнить вызов `NtWaitForAlertByThreadId` второй раз, теперь без тайм-аута. Данный вызов гарантированно вернется, так как коду известно, что происходит активизация. Заметим, что в случаях без тайм-аута не нужно удалять блок ожидания, поскольку код пробуждения уже это сделал.

На стороне записывающего функции `RtlWakeOnAddressSingle` и `RtlWakeOnAddressAll` полагаются на одну и ту же вспомогательную функцию, которая хеширует входящий адрес и выполняет поиск по хеш-таблице в PEB, описанной выше. Осторожно синхронизируясь посредством инструкций сравнения-обмена, она удаляет блок ожидания адреса из хеш-таблицы и, если ей требуется пробудить кого-то из ожидающих, проходится по всем подходящим блокам ожидания того же адреса, вызывая `NtAlertThreadByThreadId` для каждого из них в версии `All` или только первого из них в версии `Single`.

Благодаря такой реализации мы фактически имеем отдельную версию событий с ключом для пользовательского режима, которая не полагается ни на один объект

или дескриптор ядра, даже на глобальные, полностью исключая любые отказы в условиях нехватки памяти. Единственной обязанностью ядра остается перевод потока в состояние ожидания и вывод его из этого состояния.

В следующих нескольких разделах мы рассмотрим различные примитивы, которые обеспечивают синхронизацию в условиях конкуренции, полагаясь на эту функциональность.

Критические разделы

Критические разделы являются одним из главных примитивов синхронизации, предоставляемых Windows разработчикам приложений пользовательского режима на основе примитивов синхронизации режима ядра. Критические разделы и другие примитивы, как позже станет понятно, имеют одно существенное преимущество перед своими аналогами из ядра — позволяют избежать переключения в режим ядра и обратно в случаях, когда за блокировку нет конкуренции (доля которых обычно 99 % или более). Но в случае конкуренции по-прежнему требуется обращение к ядру, поскольку это единственный компонент системы, способный реализовать сложную логику активизации и диспетчеризации, необходимую для работы этих объектов.

Критические разделы могут оставаться в пользовательском режиме путем использования локального бита, обеспечивающего логику эксклюзивной блокировки, во многом похожей на push-блокировку. Если этот бит содержит 0, критический раздел можно получить, и новый владелец помещает туда 1. Эта операция не требует вызова ядра, но использует рассмотренные ранее операции взаимоблокировки центрального процессора. Освобождение критического раздела осуществляется аналогичным образом, когда с помощью операции взаимоблокировки бит состояния изменяется с 1 на 0. С другой стороны, как вы уже могли догадаться, когда бит уже имеет значение 1 и другой вызывающий код пытается получить критический раздел, должно быть вызвано ядро, чтобы поместить поток в состояние ожидания.

Подобно push-блокировкам и адресно базированным ожиданиям, критические разделы прибегают к дальнейшей оптимизации, чтобы избежать обращений к ядру: циклической проверке, во многом подобно спин-блокировке (правда, на пассивном уровне IRQL 0) для бита блокировки в надежде, что он будет сброшен достаточно скоро для того, чтобы избежать блокирующего ожидания. По умолчанию установлено ограничение в 2000 циклов, но его можно перенастроить с помощью функций `InitializeCriticalSectionAndSpinCount` во время создания или `SetCriticalSectionSpinCount` позднее.

Когда возникает потребность пойти напрямую по пути конкуренции, при первом обращении критические разделы пытаются инициализировать свое поле `LockSemaphore`. В современных версиях Windows это происходит, только если установлен флаг `RtlpForceCSToUseEvents`, то есть когда для текущего процесса, исходя из базы совместимости приложений, установлен флаг `KACF_ALLOCDEBUGINFOFORCRITSECTIONS` (0x400000). Но если это так, будет создан объект события диспетчера (даже если поле ссылается на семафор, объект все равно событие). Затем исходя из того, что событие было создано, выполняется `WaitForSingleObject`, чтобы установить блокировку на критическом разделе (как правило, с индивидуальным для процесса настраиваемым тайм-аутом, полезным при отладке взаимных блокировок, после которого попытка ожидания повторяется).

В случаях, когда слой совместимости приложению не нужен или же нехватка памяти такова, что его потребовали, но создать событие не удалось, критические разделы прекращают пользоваться им, как и любым другим функционалом событий с ключом, описанным ранее. Вместо этого они напрямую полагаются на вышеупомянутый адресно базированный механизм (с тем же самым механизмом определения взаимных блокировок, что и в предыдущем разделе). Адрес локального бита предоставляется при вызове `WaitOnAddress`, и как только критический раздел освобождается через `LeaveCriticalSection`, последняя использует либо `SetEvent`, либо `WakeAddressSingle` в отношении локального бита.

ПРИМЕЧАНИЕ Хотя мы и упоминали функции по их именам в Win32 API, на самом деле критические разделы реализованы в Ntdll.dll, а KernelBase.dll просто передает управление идентичным функциям с префиксом Rtl, поскольку они входят в библиотеку среды выполнения. Таким образом, RtlLeaveCriticalSection вызывает NtSetEvent, RtlWakeAddressSingle и т. д.

Наконец, поскольку критические разделы не являются объектами ядра, на них накладываются определенные ограничения. Основным из них является то, что для критического раздела нельзя получить дескриптор ядра, из-за этого защита, именование и прочие функции диспетчера объектов к ним неприменимы. Два процесса не могут использовать один критический раздел для координации своих действий, равно как и нельзя применять дублирование или наследование.

Ресурсы пользовательского режима

Ресурсы пользовательского режима тоже обеспечивают более тонкие механизмы блокировки, чем примитивы ядра. Ресурс может быть получен для режима совместного использования или для эксклюзивного режима, что позволяет ему работать в режиме блокировки для нескольких читающих потоков (при совместном использовании), в режиме блокировки для единственного записывающего потока (при эксклюзивном использовании) для таких структур данных, как базы данных. Когда происходит получение ресурса в режиме совместного использования и другие потоки пытаются получить его же, не требуется никаких переходов в режим ядра, поскольку ни один из потоков не будет ожидающим. Это потребуется, только когда поток попытается получить ресурс для эксклюзивного доступа или когда ресурс уже заблокирован исключающим владельцем.

Чтобы воспользоваться тем же механизмом диспетчеризации и синхронизации, который мы видели в ядре, ресурсы задействуют существующие там примитивы. В структуре данных ресурса (`RTL_RESOURCE`) содержатся дескрипторы от двух объектов семафоров ядра. Когда ресурс запрашивается в эксклюзивном режиме более чем одним потоком, он пользуется эксклюзивным семафором с единичным счетчиком освобождения, поскольку владелец может быть только один. Когда ресурс запрошен в совместном режиме более чем одним потоком, он предоставляет общий семафор со счетчиком освобождения, соответствующим количеству владельцев. Детали подобного уровня обычно скрыты от программиста, а эти внутренние объекты ни в коем случае нельзя использовать напрямую.

Изначально ресурсы были реализованы для поддержки диспетчера учетных записей безопасности (Security Account Manager, SAM) (подробности см. в главе 7

тома 1) и не были доступны в Windows API для стандартных приложений. SRW-блокировки (Slim Reader-Writer Locks), о которых вскоре поговорим, были введены позже с целью предоставить похожий, но высокооптимизированный примитив блокировки через документированный API, хотя некоторые системные компоненты до сих пор пользуются механикой ресурсов.

Условные переменные

Условные переменные обеспечивают присущую Windows реализацию синхронизации набора потоков, ожидающих конкретного результата проверки условия. Хотя эта операция была возможна и с применением других методов синхронизации в пользовательском режиме, атомарного механизма для проверки результата проверки условия и для начала ожидания изменения этого результата не существовало. Это потребовало использования в отношении таких фрагментов кода этой дополнительной синхронизации.

Поток пользовательского режима инициализирует условную переменную, вызывая `InitializeConditionVariable`, чтобы задать исходное состояние. Когда требуется начать ожидание этой переменной, он может вызвать функцию `SleepConditionVariableCS`, которая использует критический раздел (поток должен его инициализировать), чтобы дожидаться изменений. Даже лучше, можно использовать `SleepConditionVariableSRW`, где применяется SRW-блокировка, давая вызвавшему потоку преимущества запроса совместного или эксклюзивного доступа.

Тем временем устанавливающий поток должен использовать функцию `WakeConditionVariable` (или `WakeAllConditionVariable`) после внесения изменений в переменную. В результате этого вызова освобождается критический раздел или SRW-блокировка либо одного, либо всех ожидающих потоков, в зависимости от того, которая из этих функция была использована. Если это похоже на адресно базированное ожидание, то лишь потому, что так и есть — с дополнительной гарантией применения атомарной операции сравнения-ожидания. Кроме того, условные переменные были реализованы раньше адресно базированного ожидания (а значит, и раньше `alert-by-ID`) и поэтому должны были полагаться на события с ключом, которые демонстрировали поведение, лишь приближенное к желаемому.

До появления условных переменных было обычным делом применение события уведомления или синхронизации (напомним, что в Windows API они известны как `auto-reset` или `manual-reset`), чтобы сигнализировать об изменении какой-то переменной, например состояния рабочей очереди. Ожидание изменения требовало получения, а затем освобождения критического раздела, после чего выполнялось ожидание события. После ожидания нужно было снова получить критический раздел. В ходе всех этих получений и освобождений поток мог переключать контексты, но возникали проблемы, если какой-то из потоков вызывал `PulseEvent` (похожую проблему события с ключом решали, принуждая сигнализирующий поток дожидаться, пока не появится ожидающий). С помощью условных переменных получение критического раздела или SRW-блокировки может сохраняться приложением, пока вызвана `SleepConditionVariableCS/SRW`, а освобождение возможно только после завершения основной работы. Это делает написание кода, связанного с рабочими очередями и другими подобными реализациями, более простым и предсказуемым.

Тем не менее с переходом и критических разделов, и SRW-блокировок на адресно базированные примитивы условные переменные могут напрямую полагаться на `NtWaitForAlertByThreadId` и напрямую подавать сигнал потоку, в то же время пользуясь блоком ожидания условной переменной, структурно похожим на ранее описанный блок ожидания адреса. Таким образом, потребность в событиях с ключом полностью отпала, а сохраняются они лишь ради обратной совместимости.

Тонкие блокировки чтения/записи

Хотя условные переменные и являются механизмом синхронизации, полностью примитивными блокировками они не являются, так как их блокировочное поведение сопряжено со скрытыми сравнениями значений и полагается на предоставленные со стороны абстракции более высокого уровня — собственно блокировку. В свою очередь, адресно базированное ожидание является примитивной операцией, но обеспечивает лишь базовый примитив синхронизации, а не настоящую блокировку. Посреди этих двух вариантов Windows обзавелась настоящим блокирующим примитивом, который почти идентичен push-блокировке, — тонкой блокировкой чтения/записи (Slim Reader-Writer Locks), или SRW-блокировкой.

Как и их аналог из ядра, SRW-блокировки умещаются в размер указателя, используют атомарные операции для получения и высвобождения, переупорядочивают список ожидающих, защищаются от очередей на блокировке и могут быть запрошены как в эксклюзивном, так и в совместном режиме. Как и push-блокировки, SRW-блокировки могут быть улучшены или конвертированы из общего доступа в эксклюзивный и обратно. Единственное реальное различие в том, что первые доступны только коду режима ядра, а вторые — только коду пользовательского режима. Они не могут использоваться совместно с другой стороной или быть там видимыми. Поскольку SRW-блокировки тоже задействуют примитив `NtWaitForAlertByThreadId`, они не требуют выделения памяти и гарантированно безотказны (не считая случаев некорректного применения).

SRW-блокировки не только способны полностью заменить собой в коде приложения критические разделы, снижая потребность в размещении больших структур `CRITICAL_SECTION` (которые раньше требовали создания объекта события), но и предоставляют функциональность множественного чтения и одиночной записи. В первую очередь SRW-блокировки следует инициализировать с помощью либо функции `InitializeSRWLock`, либо контрольного значения, после чего их можно будет запрашивать и высвобождать в эксклюзивном или совместном режиме с помощью соответствующих функций: `AcquireSRWLockExclusive`, `ReleaseSRWLockExclusive`, `AcquireSRWLockShared` и `ReleaseSRWLockShared`.

ПРИМЕЧАНИЕ В отличие от большинства других API в Windows, функции SRW-блокировки не возвращают результата, вместо этого в случае провала они генерируют исключения. При этом становится очевидным тот факт, что получить блокировку не удалось, поэтому выполнение кода, который пытается это сделать, будет завершено, вместо того чтобы дать ему потенциальную возможность повредить пользовательские данные. Поскольку SRW-блокировки не получают отказ из-за нехватки ресурсов, единственным возможным исключением остается STATUS_RESOURCE_NOT_OWNED, возникающее при некорректной попытке высвободить несовместную блокировку в совместном режиме.

SRW-блокировки Windows не отдают предпочтения ни коду, ведущему чтение, ни коду, ведущему запись, так что производительность в обоих случаях должна быть одинаковой. Это делает их прекрасной заменой критическим разделам, являющимся механизмами синхронизации только для записи или *эксклюзивными*, и оптимизированной альтернативой ресурсам. Будь они оптимизированы для чтения, из них бы вышли плохие блокировки для эксклюзивного доступа, но это не тот случай. Вот почему мы ранее упомянули, что условные переменные тоже могут использовать SRW-блокировки с помощью функции `SleepConditionVariableSRW`. Вместе с тем, поскольку события с ключом больше не используются в одном механизме (SRW), но все еще применяются в другом (CS), адресно базированное ожидание не проявляет большинство преимуществ, кроме размера кода и возможности использовать блокировку для чтения или записи. Тем не менее код, предназначенный для более старых версий Windows, должен задействовать SRW-блокировки, чтобы гарантировать их преимущества в условиях ядра, где все еще применяются события с ключом.

Однократная инициализация

Возможность гарантировать *атомарное* исполнение участка кода, ответственного за какую-либо задачу по инициализации, как то: выделение памяти, инициализация каких-то переменных или даже создание объектов по требованию, является типичной проблемой в многопоточном программировании. На участке кода, который может быть вызван одновременно несколькими потоками (хорошим примером будет подпрограмма DLLMain, инициализирующая DLL), существует несколько способов обеспечить корректное, атомарное и уникальное выполнение задач инициализации.

Для таких случаев Windows реализует то, что называется *init once*, или *one-time initialization*, то есть проводит *однократную инициализацию*. Этот API существует и в версии для Win32, он переводит вызов на библиотеку режима выполнения Ntdll.dll (`Rtl`), как делают все ранее описанные механизмы и документированный набор API `Rtl`, которые, в свою очередь, видны разработчикам уровня ядра в Ntoskrnl.exe (очевидно, что разработчики пользовательского режима тоже могут обходить Win32 и применять функции `Rtl` из Ntdll.dll, но это не рекомендуется). Единственное различие между этими двумя реализациями в том, что ядро в итоге задействует для синхронизации объект события, а в пользовательском режиме применяются события с ключом (фактически в параметр дескриптора помещается `NULL`, чтобы можно было воспользоваться событием с ключом для условий дефицита памяти, которое раньше использовалось критическими разделами).

ПРИМЕЧАНИЕ Поскольку недавние версии Windows теперь реализуют адресно базированную push-блокировку для режима ядра и адресно зависимый примитив в пользовательском режиме, библиотека Rtl тоже могла бы быть обновлена для применения функций RtlWakeAddressSingle и ExBlockOnAddressPushLock. И на самом деле в будущих версиях Windows могут пойти на это — событие с ключом всего лишь предоставляло интерфейс, более похожий на объект события диспетчера из старых версий Windows. Как всегда, не полагайтесь на детали, изложенные в данной книге, потому что они могут поменяться.

Механизм `init once` позволяет как синхронное (в таком случае другие потоки должны подождать окончания инициализации) исполнение конкретного участка кода, так и асинхронное (при котором другие потоки могут сделать попытку самостоятельно наперегонки провести инициализацию). Логику, стоящую за асинхронным исполнением, мы рассмотрим позже, а пока остановимся на синхронном.

При синхронном выполнении разработчик пишет фрагмент кода, который обычно будет выполняться после двойной проверки глобальной переменной в специальной функции. Любая информация, которая нужна этой процедуре, может быть передана через переменную параметра `parameter`, которую принимает процедура однократной инициализации. Любая выходная информация возвращается через переменную контекста `context`. Состояние самой инициализации возвращается как булево значение. Все, что должен сделать разработчик для обеспечения надлежащего выполнения, — вызвать функцию `InitOnceExecuteOnce` с `parameter`, `context` и указателем на функцию однократного запуска после инициализации объекта `INIT_ONCE` с помощью API-функции `InitOnceInitialize`. Обо всем остальном позаботится система.

Для приложений, которым вместо этого нужно использовать асинхронную модель, потоки вызывают функцию `InitOnceBeginInitialize` и получают булево значение отложенного состояния — `pending status` — и описанное ранее значение `context`. Если значение `pending status` равно `FALSE`, значит, инициализация уже состоялась и поток использует в качестве результата значение `context`. Вернуть `FALSE` может и сама функция, что будет означать неудачный исход инициализации. Но если в `pending status` возвращается `TRUE`, поток должен вступить в соревнование, чтобы первым создать объект. Последующий код выполняет все, что требуется от задачи инициализации, например, занимается созданием объектов или выделением памяти. Когда эта работа будет сделана, поток вызывает функцию `InitOnceComplete` с результатом работы в виде значения `context` и получает значение `status`. Если оно равно `TRUE`, значит, поток выиграл соревнование, и объект, который создан или распределен, станет глобальным объектом. Теперь поток может сохранить этот объект или вернуть его вызывавшему модулю, в зависимости от того, что применимо.

При более сложном сценарии, если `status` имеет значение `FALSE`, это означает, что поток проиграл соревнование. Поток должен произвести откат всей проделанной работы, например отменить удаление объектов или освобождение памяти, а затем снова вызвать функцию `InitOnceBeginInitialize`. Но вместо запроса на старт соревнований, как он это первоначально сделал, он использует флаг проведения всего лишь проверки возможности однократной инициализации `INIT_ONCE_CHECK_ONLY`, зная о своих потерях, и запрашивает вместо прежних данных значение `context` победителя (например, тех созданных победителем объектов или той выделенной им памяти). Здесь уже возвращается другое значение `status`, которое может быть `TRUE` и означать, что `context` имеет действующее значение, которое может быть использовано или возвращено вызывавшему модулю, или может быть `FALSE`, и означать, что инициализация прошла неудачно и никто в данный момент не сможет выполнить работу в условиях дефицита памяти.

В обоих случаях механизм однократной инициализации подобен механизму условных переменных и SRW-блокировок. Структура `init once` имеет размер указателя, а при отсутствии конкуренции используются вложенные ассемблерные версии кода получения/высвобождения SRW, в то время как события с ключом применяются, когда конкуренция есть (что происходит, когда механизм используется в синхронном

режиме) и другие потоки должны дожидаться инициализации. В асинхронном случае блокировки используются в совместном режиме, позволяя нескольким потокам выполнять инициализацию одновременно. Хотя это не так эффективно, как примитив `alert-by-ID`, применение событий с ключом все еще гарантирует, что механизм `init once` будет работать даже в большинстве случаев исчерпания памяти.

ПРОДВИНУТЫЙ ЛОКАЛЬНЫЙ ВЫЗОВ ПРОЦЕДУР

Всем современным операционным системам необходим механизм для безопасной и эффективной передачи данных между одним или несколькими процессами в пользовательском режиме, равно как и между службой в ядре и ее клиентами в пользовательском режиме. Обычно для переносимости используются такие UNIX-механизмы, как почтовые слоты, файлы, именованные каналы и сокеты, но есть и такие разработчики, которые для графических приложений используют оконные сообщения, применяемые в графических приложениях Win32. В Windows реализуется внутренний IPC-механизм — *продвинутый (или асинхронный) локальный вызов процедур* (Advanced Local Procedure Call, ALPC), обладающий высокой скоростью работы, масштабируемостью и средством обеспечения передачи сообщений произвольного размера.

ПРИМЕЧАНИЕ ALPC заменил устаревший IPC-механизм, поставлявшийся с самой первой версией ядра Windows NT и известный как LPC. Поэтому даже сегодня некоторые переменные, поля и функции могут ссылаться на LPC. Следует помнить, что механизм LPC теперь эмулируется на верхнем уровне ALPC для совместимости и удален из ядра, но существуют еще устаревшие системные вызовы, которые заключаются в оболочку ALPC-вызовов).

Хотя этот механизм является внутренним и поэтому недоступен сторонним разработчикам, ALPC широко используется в различных частях Windows.

- Приложения Windows, использующие вызов удаленных процедур (remote procedure call, RPC), относящийся к документированному API, опосредованно задействуют ALPC, когда определяют локальный RPC-вызов через `ncalrpc`-транспортировку. Эта форма RPC применяется для связи между процессами в рамках одной системы. На данный момент это протокол передачи по умолчанию почти для всех клиентов RPC. Кроме того, когда драйверы Windows полагаются на RPC режима ядра, это приводит к неявному использованию ALPC как единственного разрешенного варианта передачи.
- Всякий раз, когда в Windows запускается процесс или поток, равно как и во время любой операции подсистемы Windows, ALPC используется для связи с процессом этой подсистемы (CSRSS). Все подсистемы связываются с диспетчером сеанса (SMSS) через ALPC.
- Winlogon задействует ALPC для связи с процессом проверки подлинности локальной системы безопасности (LSASS).
- Монитор безопасности ссылок (компонент исполнительной системы, описанный в главе 7 тома 1) использует ALPC для связи с процессом LSASS.
- Диспетчер электропитания пользовательского режима и монитор электропитания связываются с диспетчером электропитания режима ядра через ALPC всякий раз, когда на дисплее меняют яркость.

- User-Mode Driver Framework (UMDF) позволяет драйверам пользовательского режима обмениваться данными с драйвером-ответчиком режима ядра через ALPC.
- Новый механизм Core Messaging, применяемый в CoreUI и современных UWP UI-компонентах, использует ALPC как для регистрации в Core Messaging Registrar, так и для отправки объектов сериализованных сообщений, которые пришли на замену устаревшей модели оконных сообщений Win32.
- Изолированный процесс LSASS в условиях активности технологии Credential Guard общается с LSASS через ALPC. Подобным образом безопасное ядро передает информацию из аварийного дампа доверия в WER через ALPC.

Модель подключения

Обычно ALPC-вызовы используются между процессом-сервером и одним или несколькими процессами-клиентами этого сервера. APLC-подключение может быть установлено между двумя или более процессами пользовательского режима, или между компонентом режима ядра и одним или более процессами пользовательского режима, или даже между двумя компонентами режима ядра (хотя это было бы не самым эффективным способом связи). Для поддержания необходимого для связи состояния ALPC предоставляет один исполняемый объект, называемый *объектом порта*. Хотя это всего лишь один объект, на самом деле существует несколько видов ALPC-портов, которые он может представлять.

- **Порт подключения к серверу.** Именованный порт, указываемый в запросе на подключение к серверу. Клиенты могут подключаться к серверу, подключаясь к этому порту.
- **Серверный порт связи.** Безымянный порт, который сервер использует для связи с конкретным клиентом. Для каждого активного клиента у сервера есть один такой порт.
- **Клиентский порт связи.** Безымянный порт, который каждый клиент использует для связи со своим сервером.
- **Неподключенный порт связт.** Безымянный порт, который клиент может использовать для локальной связи с самим собой. Данная модель была упразднена во время перехода с LPC на ALPC, но устаревший LPC эмулируется по соображениям совместимости.

ALPC придерживается моделей подключения и связи, в чем-то напоминающих программирование сокетов под BSD. Сначала сервер создает порт подключения к серверу (`NtAlpcCreatePort`), а клиент предпринимает попытку подключиться к этому порту (`NtAlpcConnectPort`). Если сервер находился в состоянии прослушивания, используя `NtAlpcSendWaitReceivePort`, он получает сообщение о запросе подключения и может его принять (`NtAlpcAcceptConnectPort`). В таком случае создаются как клиентский, так и серверный порты связи, а каждый соответствующий конечный процесс получает дескриптор своего порта связи. Впоследствии через этот дескриптор пересылаются сообщения (также с помощью `NtAlpcSendWaitReceivePort`), которые сервер продолжает получать, используя тот же API. Таким образом, в наиболее

простом сценарии один из потоков сервера находится в цикле, вызывая функцию `NtAlpcSendWaitReceivePort` и с ее же помощью получая запросы на подключение, которые он принимает, или сообщения, которые обрабатывает и, возможно, на которые отвечает. Сервер может различать сообщения, читая структуру `PORT_HEADER`, которая находится в начале любого сообщения и хранит его тип. Различные типы сообщений приведены в табл. 8.30.

Таблица 8.30. Типы сообщений ALPC

Тип	Назначение
LPC_REQUEST	Обычное сообщение ALPC с возможностью синхронного ответа
LPC_REPLY	Сообщение-датаграмма ALPC, отправленное в качестве асинхронного ответа на предыдущую датаграмму
LPC_DATAGRAM	Сообщение-датаграмма ALPC, которое немедленно освобождается и не может послужить основанием синхронного ответа
LPC_LOST_REPLY	Устарело, применяется в Legacy LPC Reply API
LPC_PORT_CLOSED	Отправляется всякий раз, когда последний дескриптор порта ALPC оказывается закрыт, сообщая клиентам и серверам, что другая сторона недоступна
LPC_CLIENT_DIED	Отправляется диспетчером процессов (PspExitThread) с помощью Legacy LPC на зарегистрированные порты завершения потока и зарегистрированный порт исключений процесса
LPC_EXCEPTION	Отправляется платформой отладки пользовательского режима (DbgkForwardException) в порт исключений посредством Legacy PLC
LPC_DEBUG_EVENT	Устарело, применялось устаревшими службами отладки пользовательского режима во времена, когда они еще были частью подсистемы Windows
LPC_ERROR_EVENT	Отправляется каждый раз, когда в пользовательском режиме генерируется аппаратный сбой (NtRaiseHardError). Отправляется с помощью Legacy LCP в порт исключений целевого потока, если таковой есть, а если нет — в порт ошибок, как правило принадлежащий CSRSS
LPC_CONNECTION_REQUEST	Сообщение ALPC, означающее попытку клиента подключиться к порту подключения к серверу
LPC_CONNECTION_REPLY	Внутреннее сообщение, отправляемое сервером, когда он вызывает функцию, чтобы принять запрос на подключение от клиента
LPC_CANCELED	Ответ, полученный от клиента или сервера, ожидавшего сообщение, которое оказалось отменено
LPC_UNREGISTER_PROCESS	Отправляется диспетчером процессов, когда порт исключения текущего процесса заменяется другим, позволяя владельцу (как правило, CSRSS) отменить регистрацию его структур данных для потока, переключая его порт на другой

Сервер может отклонить подключение как по соображениям безопасности, так и просто из-за проблем с протоколом или версиями. Поскольку клиенты могут отправить с запросом на подключение какую-нибудь свою полезную нагрузку, это обычно используется различными службами, чтобы убедиться, что с сервером связывается надлежащий или только единственный клиент. Если будут обнаружены

какие-нибудь аномалии, сервер может отклонить подключение и дополнительно вернуть нагрузку, содержащую информацию о том, почему клиент был отклонен. Эта информация позволяет клиенту предпринять корректирующее действие или воспользоваться этой информацией в целях отладки программы.

Когда подключение установлено, в информационной структуре этого подключения (на самом деле это двоичный объект, о чем мы поговорим позже) сохраняются все связи между всеми портами, как показано на рис. 8.40.

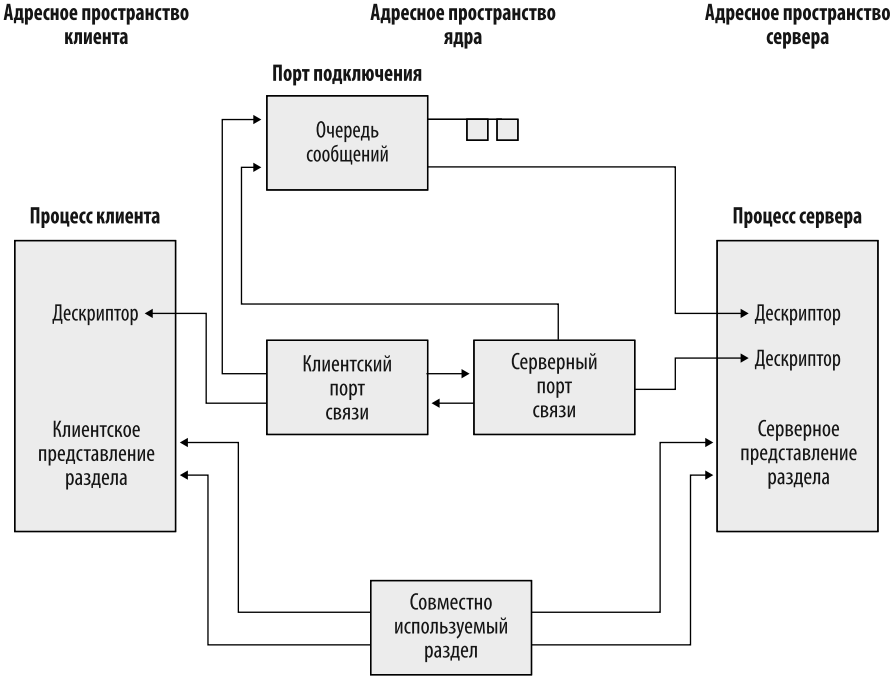


Рис. 8.40. Использование портов ALPC

Модель сообщений

Используя ALPC, клиент и поток посредством блокирующих сообщений по очереди выполняют цикл вокруг системного вызова `NtAlpcSendWaitReceivePort`, в котором одна из сторон отправляет запрос и ждет ответ, в то время как другая сторона делает противоположное. Поскольку ALPC поддерживает асинхронные сообщения, у любой из сторон есть возможность не заниматься блокировкой, а выполнить вместо этого какую-нибудь другую задачу времени выполнения и проверить наличие сообщения чуть позже (некоторые из этих методов вскоре будут описаны). ALPC поддерживает следующие три метода обмена полезной нагрузкой, отправляемой вместе с сообщением.

- Сообщение может быть отправлено другому процессу через стандартный механизм двойного буферирования, в котором ядро сохраняет копию сообщения (копируя его из процесса-источника), переключается на целевой процесс и копирует данные из буфера ядра. С целью поддержки совместимости, если используется устаревший LPC, этим способом могут быть отправлены только сообщения длиной до 256 байт, хотя у ALPC есть возможность выделить *расширенный буфер* для сообщений объемом примерно 64 Кбайт.
- Сообщение может быть сохранено в объекте ALPC-раздела, откуда клиент и сервер выстраивают свои видения.

Важным побочным эффектом возможности отправлять асинхронные сообщения является то, что сообщение может быть отменено, например, в ситуациях, когда запрос занимает слишком много времени или пользователь изъявил желание отменить операцию, связанную с этим сообщением. ALPC поддерживает эту возможность с помощью системного вызова `NtAlpcCancelMessage`.

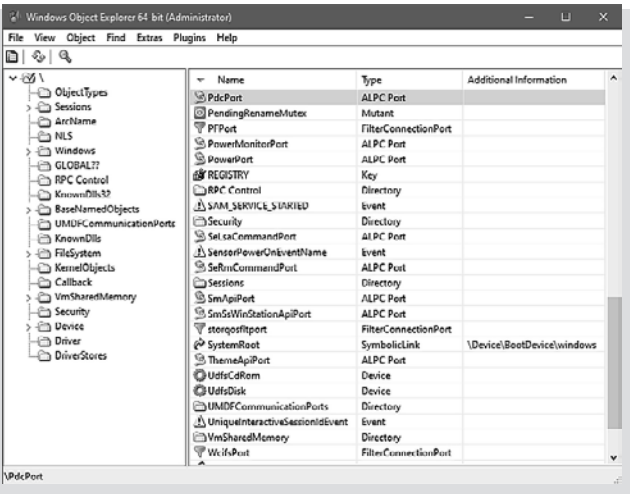
Сообщение ALPC может находиться в одной из пяти очередей, реализуемых объектом порта ALPC.

- **Главная очередь.** Сообщение было отправлено, и клиент его обрабатывает.
- **Очередь ожидания.** Сообщение было отправлено, и отправитель ожидает ответа, который еще не был отправлен.
- **Большая очередь сообщений.** Сообщение было отправлено, но буфер вызывающего кода оказался слишком мал, чтобы его вместить. Вызывающему коду дается еще один шанс разместить буфер побольше и повторно запросить полезную нагрузку сообщения.
- **Очередь аннулированных сообщений.** Сообщение было отправлено в порт, но потом отменено.
- **Прямая очередь.** Сообщение было отправлено привязанным к прямому событию.

Обратите внимание на то, что шестая очередь, названная *очередью ожидания*, не связывает сообщения вместе, вместо этого она связывает все потоки, ожидающие сообщение.

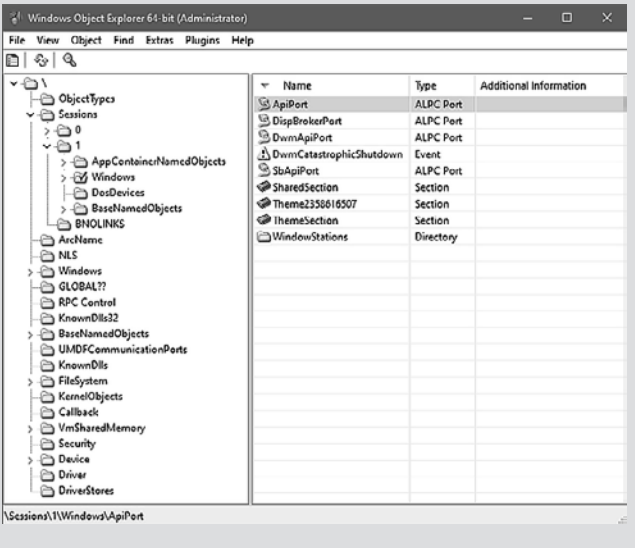
ЭКСПЕРИМЕНТ. Просмотр объектов ALPC-порта подсистемы

Вы можете просмотреть именованные объекты портов ALPC с помощью инструмента WinObj от SysInternals или WinObj64 с GitHub. Запустите один из этих двух инструментов с правами администратора и выберите корневую папку. В WinObj объекты портов отмечены значком с шестеренками, а в WinObj64 — значком в виде электрической вилки, как показано на рисунке (можете щелкнуть на поле Тип (Type), чтобы быстро отсортировать объекты по типу).



Вы должны увидеть порты ALPC, используемые диспетчером электропитания, диспетчером безопасности и другими внутренними службами Windows. Если хотите увидеть объекты порта ALPC, задействуемые RPC, можете выбрать папку \RPC Control.

Одним из основных пользователей ALPC, помимо локального RPC, является подсистема Windows, которая использует его для связи со своими DLL-библиотеками, доступными для всех Windows-процессов. Поскольку CSRSS загружается лишь раз за сеанс, вы найдете ее объекты портов ALPC в папке \Sessions\X\Windows, как показано на снимке экрана.



Асинхронные операции

Синхронная модель ALPC связана с исходной LPC-архитектурой в ранней конструкции NT и похожа на другие блокирующие IPC-механизмы, например на Mach-порты. Хотя блокирующий IPC-алгоритм прост по конструкции, в нем имеется множество возможностей для взаимных блокировок, и обходы подобных сценариев приводят к созданию сложного кода, требующего поддержки более гибкой асинхронной (не блокирующей) модели. Поэтому ALPC изначально проектировался в первую очередь для поддержки также асинхронной деятельности, что требуется для масштабируемых RPC, равно как и для других применений, таких как отложенный ввод/вывод у драйверов пользовательского режима. Одной из базовых возможностей ALPC, отсутствовавшей в LPC, является возможность назначать блокирующим вызовам параметр тайм-аута. Это позволяет старым приложениям избегать определенных сценариев взаимной блокировки.

Однако ALPC оптимизирован для асинхронных сообщений и обеспечивает три различные модели асинхронных уведомлений. Первая на самом деле не уведомляет клиент или сервер, а просто копирует данные полезной нагрузки при сообщении. В рамках этой модели выбор надежного метода синхронизации зависит от автора реализации. Например, клиент и сервер могут делить между собой объект события уведомления или же клиент может отслеживать прибытие данных. Структура данных, используемая этой моделью, называется *списком завершения* ALPC (не путать с портом завершения ввода-вывода Windows). Список завершения ALPC — это эффективная неблокирующая структура данных, которая позволяет выполнять атомарную передачу данных между клиентами, ее внутреннее устройство будет более подробно описано в подразделе «Производительность».

Следующей моделью уведомлений является ожидающая модель, которая задействует механизм портов завершения Windows (поверх списка завершения ALPC). Это позволяет потоку получать по несколько блоков полезной нагрузки сразу, контролировать максимальное количество одновременных запросов и пользоваться преимуществами функциональных возможностей, присущих порту завершения. Реализация пула потоков пользовательского режима обеспечивает внутренние функции API, с помощью которых процессы управляют сообщениями ALPC в рамках той же инфраструктуры, что и рабочие потоки, реализованные с помощью этой модели. Система RPC в Windows в случае применения Local RPC (через `ncalrpc`) тоже пользуется данной функциональностью для обеспечения эффективной доставки сообщений за счет преимущества в виде поддержки со стороны ядра, как поступает исполняемый образ RPC режима ядра Msrpc.sys.

Наконец, поскольку драйверы могут также использовать асинхронные ALPC, но обычно не поддерживают порты завершения на таком высоком уровне, ALPC также предоставляет механизм для более основательного оповещения, основанного на использовании ядра с использованием объектов обратного вызова исполнительной системы. Драйвер может зарегистрировать собственную функцию обратного вызова и контекст с помощью функции `NtSetInformationAlpcPort`, после чего она будет вызвана при получении сообщения. Например, Power Dependency Coordinator (Pdc.sys) в рамках ядра использует данный механизм для связи со своими клиентами. Стоит заметить, что применение объекта обратного вызова исполнительной системы дает потенциальные преимущества в плане производительности, как, впрочем, и риски

в сфере безопасности. Поскольку обратные вызовы выполняются в блокирующей манере (по сигналу) и вложены в сигнализирующий код, они всегда будут запускаться в контексте отправителя сообщения ALPC, а именно вложенными в поток пользовательского режима, вызвавший `NtAlpcSendWaitReceivePort`. Это означает, что компонент ядра может получить шанс оценить состояние своего клиента, не волнуясь о стоимости переключения контекста, и теоретически получить присланную полезную нагрузку в контексте отправителя.

Невозможно дать абсолютные гарантии (и уберечь от риска, если разработчик не в курсе) потому, что множество клиентов могут послать сообщение на конкретный порт в одно и то же время и сообщение может быть отправлено клиентом до того, как сервер зарегистрирует свой объект обратного вызова исполнительной системы. Также возможна ситуация, когда еще один клиент прислал сообщение, пока сервер занят обработкой первого сообщения от другого клиента. Серверу следует замечать такие ситуации (поскольку Client ID отправителя закодирован в `PORT_HEADER` сообщения) и привязывать/анализировать состояние корректного отправителя (в результате чего, возможно, придется расходовать усилия на переключение контекста).

Просмотры, области и разделы

Вместо обмена буферами с сообщениями между двумя соответствующими процессами сервер и клиент могут выбрать более эффективный механизм передачи данных, который лежит в основе диспетчера памяти Windows, — *объект раздела* (подробности см. в главе 5 тома 1). Он позволяет выделить часть памяти для совместного использования и открыть как для клиента, так и для сервера последовательный и одинаковый просмотр этой памяти. При таком подходе можно передавать столько данных, сколько туда поместится, а сами они просто копируются в один диапазон адресов и немедленно становятся доступны в другом. К сожалению, связь с использованием общей памяти, такая, которую традиционно предоставляет LPC-вызов, имеет свою долю недостатков, особенно если брать в расчет вопросы безопасности. Поскольку и клиент, и сервер должны иметь доступ к этой памяти, непривилегированный клиент может в первую очередь воспользоваться этим, чтобы неавторизованно изменить совместно используемую память сервера или даже создать там выполняемую полезную нагрузку с потенциально вредоносным кодом. Кроме того, поскольку клиент знает, где находятся серверные данные, он может воспользоваться этой информацией для обхода ASLR-защиты (подробности см. в главе 5 тома 1).

ALPC обеспечивает собственную защиту вдобавок к той, что обеспечивается объектами раздела. При использовании ALPC указанный объект раздела должен быть создан с помощью соответствующей API-функции `NtAlpcCreatePortSection`, которая создает корректные ссылки на порт и позволяет применить автоматический сбор мусора в разделе (существуют также API-функции для ручного удаления мусора). Как только владелец объекта раздела ALPC приступает к его использованию, из выделенных участков памяти составляются области ALPC, которые представляют диапазон используемых внутренних адресов, и добавляют дополнительную ссылку на сообщение. И наконец, внутри диапазона совместно используемой памяти клиенты получают возможности просмотра этой памяти, представляющие собой локальные отображения внутри их адресного пространства.

Области также поддерживают несколько вариантов обеспечения безопасности. Прежде всего области могут быть отображены либо с использование режима безопасности, либо в небезопасном режиме. При использовании режима безопасности для раздела допускаются только два просмотра (отображения). Это обычно используется в том случае, когда серверу нужно иметь совместно используемые данные, имеющие закрытый характер, только с одним клиентским процессом. Кроме того, только одна область для заданного диапазона совместно используемой памяти может быть открыта в рамках применения данного порта. И наконец, области могут также быть помечены защищенными от любого доступа, кроме как доступа по чтению, что разрешает только одному контексту процесса (серверу) иметь доступ по записи к просматриваемой области (это делается с использованием функции `MmSecureVirtualMemoryAgainstWrites`). Тем временем все остальные клиенты будут иметь доступ только по чтению. Такие настройки подавляют многие атаки, связанные с повышением уровня привилегий, которые могут произойти во время атак на совместно используемую память, и они делают ALPC-вызовы более устойчивыми по сравнению с обычными IPC-механизмами.

Атрибуты

Механизм ALPC предоставляет не только простую передачу сообщений, он также позволяет определенной контекстной информации быть добавленной к каждому сообщению и допускает со стороны ядра отслеживание достоверности, срока службы и осуществление реализации этой информации. Пользователи ALPC имеют также возможность назначения своей собственной пользовательской контекстной информации. Как при системном, так и при пользовательском управлении в ALPC эти данные называются *атрибутами*. Ядро управляет семью из них:

- атрибутом безопасности, в котором содержится ключевая информация, позволяющая заимствовать права клиента, а также получать расширенные функциональные возможности безопасности ALPC (о них будет рассказано позже);
- атрибутом просмотра данных, отвечающими за управление различными просмотрами, которые связаны с областями ALPC-раздела. Он используется также для установки флагов, таких как флаг автоосвобождения, и ручного отключения просмотра при ответе;
- атрибутом контекста, позволяющим добавлять характерный для пользователя указатель контекста, который может быть связан с сообщением, отсылаемым через этот порт. Кроме того, здесь хранятся и управляются ядром номер в последовательности, идентификатор сообщения и идентификатор обратного вызова, что позволяет пользователям ALPC обеспечивать уникальность, хеширование на основе сообщений и сортировку;
- атрибутом дескриптора, который содержит информацию о том, какие дескрипторы нужно связывать с сообщением (более подробно об этом — в разделе «Передача дескрипторов» далее);
- атрибутом токена, который может быть использован для получения ID токена, ID аутентификации и ID отправителя сообщения без применения полноценного атрибута безопасности (при этом данный атрибут не позволяет определить клиента);

- прямым атрибутом, используемым для отправки прямых сообщений, с которыми ассоциирован объект синхронизации (подробнее об этом рассказывается в разделе «Атрибут прямого события ALPC» далее);
- атрибутом «делегирования задания», используемым для кодирования рабочего задания, применяемого для оптимизации принятия решений в части управления электропитанием и ресурсами (см. подраздел «Управление электропитанием» далее).

Некоторые их этих атрибутов изначально предоставляются сервером или клиентом при отправке сообщения и конвертируются в специфичное для ядра внутреннее представление ALPC. Если пользователь ALPC потребует эти данные обратно, они будут безопасно раскрыты ему. В некоторых случаях сервер или клиент может запросить атрибут (например, атрибуты контекста или токена), потому что именно ALPC внутренне связывает его с сообщением и всегда делает его доступным. Реализуя такую модель и сочетая ее с собственной внутренней таблицей дескрипторов, о которой поговорим в дальнейшем, ALPC может сохранять критические данные непрозрачными между клиентами и серверами, при этом все еще имея настоящие указатели в режиме ядра.

Для правильного определения атрибутов внутренние ALPC-потребители располагают различными API-функциями, среди которых `AlpcInitializeMessageAttribute` и `AlpcGetMessageAttribute`.

Двоичные объекты, дескрипторы и ресурсы

Хотя подсистема ALPC публикует только один тип объектов диспетчера объектов — порт, она должен управлять рядом структур данных, которые позволяют ей выполнять выполнять задачи, требуемые ее механизмам. Например, ALPC требуется размещать и отслеживать сообщения, связанные с каждым портом, равно как и атрибуты этих сообщений, причем он должен отслеживать их все время, пока они существуют. Вместо того чтобы управлять данными с помощью функций диспетчера объектов, ALPC реализует собственные нересурсоемкие объекты, называемые *двоичными объектами*, или *блобами* (**b**inary **l**inked **ob**ject — «объект двоичной компоновки»). Как и обычные объекты, двоичные могут быть автоматически размещены в памяти или собраны вместе с мусором, отслежены по ссылкам и заблокированы с помощью синхронизации. Кроме того, они могут иметь специализированные функции обратного вызова, занимающиеся их размещением и освобождением занимаемой ими памяти, которые позволяют их владельцам управлять дополнительной информацией, которая может понадобиться для отслеживания каждого двоичного объекта. ALPC также использует имеющуюся в исполнительной системе реализацию таблицы дескрипторов (она необходима для объектов и идентификаторов процессов и идентификатором потоков), чтобы получить таблицу дескрипторов, специализированную под ALPC, которая позволяет ей генерировать для двоичных объектов закрытые дескрипторы вместо применения указателей.

Например, в рамках модели ALPC сообщения — это двоичные объекты, и их конструктор генерирует идентификатор сообщения, который сам же служит дескриптором в таблице дескрипторов ALPC. К другим видам двоичных объектов ALPC относятся:

- двоичный объект подключения, в котором хранятся порты связи клиента и сервера, а также порт подключения к серверу и таблица дескрипторов ALPC;
- двоичный объект безопасности, в котором хранятся данные, необходимые для того, чтобы представлять клиента. Там же хранится атрибут безопасности;
- двоичные объекты раздела, области и просмотра, которые описывают модель разделяемой памяти ALPC. Двоичный объект просмотра отвечает за хранение атрибута просмотра данных;
- резервный двоичный объект, реализующий поддержку резервных объектов ALPC (см. пункт «Резервные объекты» ранее в данной главе);
- двоичный объект дескриптора данных, где хранится информация, позволяющая поддерживать атрибут дескрипторов ALPC.

Поскольку двоичные объекты выделяются из выгружаемой памяти, их необходимо осторожно отслеживать для своевременного удаления. Для некоторых типов двоичных объектов это просто: например, когда сообщение ALPC освобождается, вместе с ним удаляется и двоичный объект, который его содержит. Однако некоторые двоичные объекты могут представлять ряд атрибутов, привязанных к отдельному сообщению ALPC, и тогда корректно управлять их временем жизни придется ядру. Например, поскольку сообщение может иметь несколько связанных с ним просмотров (когда доступ к одной и той же совместно используемой памяти есть сразу у нескольких клиентов), просмотры должны отслеживаться наряду со ссылающимися на них сообщениями. ALPC реализует данную функциональность, используя концепцию *ресурсов*. С каждым сообщением связан список ресурсов, и всякий раз, когда в памяти размещается новый двоичный объект (это не простой указатель), связанный с каким-то сообщением, он добавляется в качестве ресурса сообщения. В свою очередь, библиотека ALPC предоставляет функциональность для поиска, сброса и удаления связанных ресурсов. Двоичные объекты безопасности, резервные двоичные объекты, двоичные объекты просмотра хранятся в качестве ресурсов.

Передача дескрипторов

Ключевой особенностью технологий Unix Domain Sockets и Mach ports — наиболее сложных и востребованных механизмов IPC в Linux и Mac OS соответственно — является способность отправить сообщение, где закодирован *дескриптор файла*, который будет продублирован процессом-получателем, что даст ему доступ к файлу в стиле UNIX (наподобие именованного канала, сокета или реального местоположения в файловой системе). Благодаря ALPC и предоставляемому им атрибуту дескриптора Windows теперь тоже может пользоваться преимуществами данной модели. Этот атрибут позволяет отправителю закодировать тип объекта, некоторую информацию о том, как продублировать дескриптор, и индекс этого дескриптора в таблице отправителя. Если индекс дескриптора подходит типу объекта, заявленному отправителем при отправке, создается дубликат дескриптора на текущий момент в таблице дескрипторов системы (ядра). Первая часть гарантирует, что отправитель действительно прислал то, что заявил, и что на текущий момент любая операция, которую он мог бы выполнить с дескриптором или объектом, с ним связанным, не повредит его легитимности.

Затем получатель запрашивает атрибут дескриптора, указав тип объекта, который он ожидает. В случае совпадения дескриптор ядра дублируется еще раз, на этот раз как дескриптор пользовательского режима в таблице получателя (при этом копия, находящаяся под управлением ядра, закрывается). Передача дескриптора завершена, и получателю дана гарантия, что в его распоряжении находится дескриптор точно того объекта, на который отправитель ссылался, и именно того типа, которого получатель ожидает. Более того, поскольку дублирование выполняется ядром, это значит, что привилегированный сервер может послать сообщение непривилегированному клиенту, не требуя от последнего никакого типа доступа к процессу-отправителю.

Первая реализация данного механизма передачи дескрипторов в основном использовалась подсистемой Windows (CSRSS), которой требовалось знать обо всех дочерних процессах, созданных существующими процессами Windows, чтобы те могли успешно подключиться к CSRSS, когда придет их черед выполняться, и CSRSS уже знала об их создании родителем. Однако там имелся ряд проблем, в частности неспособность передать более одного дескриптора (и уж точно не более одного типа объекта). Кроме того, получателям приходилось всегда принимать любой дескриптор, связанный с сообщением этого порта, не зная заранее, должно ли это сообщение для начала работы вовсе быть связано с каким-либо дескриптором.

Для решения этих проблем начиная с Windows 8 был реализован механизм передачи *косвенных дескрипторов*, который позволил отправлять по несколько дескрипторов разных типов одновременно, а получателям — вручную забирать их по одному за сообщение. Если порт принимает и допускает подобные косвенные дескрипторы (ALPC-серверы, не основанные на RPC, как правило, не пользуются косвенными дескрипторами), обычные дескрипторы больше не будут автоматически дублироваться в зависимости от атрибута дескриптора, полученного вместе с новым сообщением с помощью `NtAlpcSendWaitReceivePort`. Вместо этого клиенты и серверы ALPC должны будут самостоятельно запрашивать, сколько дескрипторов содержится в конкретном сообщении, размещать подходящие структуры данных для получения значений этих дескрипторов и их типов, а затем уже запрашивать дублирование их всех, отбирая те, которые подходят под ожидаемый тип (при этом закрывая и отбрасывая неожиданные типы), с помощью `NtAlpcQueryInformationMessage` и пропуская дальше полученное сообщение.

Это новое поведение дает преимущество в области безопасности — вместо того чтобы автоматически дублировать дескриптор, как только вызывающий код покажет атрибут дескриптора с подходящим типом, теперь они дублируются только тогда, когда запрашиваются от сообщения к сообщению. Поскольку сервер может ожидать дескриптор для сообщения А, но не обязательно для всех других сообщений, некосвенные дескрипторы могут создать проблемы, если только сервер не решит закрывать любой возможный дескриптор даже при обработке сообщений Б или В. Пользуясь косвенными дескрипторами, сервер попросту не станет вызывать `NtAlpcQueryInformationMessage` для таких сообщений, а дескрипторы оттуда никогда не будут дублироваться или требовать их закрытия.

Благодаря этим улучшениям механизм передачи дескрипторов ALPC теперь доступен не только в рамках описанных приемов использования, но и интегрирован в библиотеке выполнения RPC и компиляторе IDL. Теперь можно при помощи синтаксиса `system_handle(sh_type)` обозначить более 20 различных типов

дескрипторов, которые библиотека RPC может маршалировать с клиента на сервер и обратно. Более того, хотя ALPC предоставляет проверку типа с точки зрения ядра, как говорилось ранее, компоненты RPC сами по себе тоже проводят дополнительные проверки. Например, хотя именованные каналы, сокеты и настоящие файлы имеют тип объекта «Файл», функции могут выполнять проверки при маршаллинге и анмаршаллинге специально, чтобы выяснить, передается ли дескриптор сокета, когда файл IDL показывает `system_handle(sh_pipe)`, например (это делается с помощью таких функций API, как `GetFileAttribute`, `GetDeviceType` и т. д.).

Безопасность

ALPC реализует несколько механизмов безопасности, полноценные границы безопасности и смягчение условий для предотвращения атак в случае общих ошибок анализа IPC. На базовом уровне объекты порта ALPC управляются теми же интерфейсами диспетчера объектов, который управляет объектами в безопасном режиме, препятствуя непривилегированным приложениям в получении дескрипторов на порты сервера с ACL. Вдобавок к этому ALPC предоставляет модель доверия на основе SID, унаследованную из оригинального дизайна LPC. Она позволяет клиентам проверить достоверность сервера, к которому они подключаются, полагаясь не только на имя порта. Получив доступ к порту, процесс клиента передает ядру SID процесса сервера, который ожидает встретить в конечной точке. Во время подключения ядро проверяет достоверность того, что клиент действительно подключился к ожидаемому серверу, снижая тем самым возможность атаки с незаконным проникновением в пространство имен, когда непроверенный сервер создает порт для имитации настоящего сервера.

ALPC также позволяет и клиентам, и серверам в рамках атомарной операции однозначно идентифицировать поток и процесс, ответственный за каждое сообщение. Этот механизм полностью поддерживает имеющуюся в Windows модель заимствования прав, используя для этого API-функцию `NtAlpcImpersonateClientThread`. Другие API-функции дают серверу ALPC возможность запросить SID, связанные со всеми подключенными клиентами, и запросить LUID (locally unique identifier) токена безопасности клиента (о чем подробнее рассказано в главе 7 тома 1).

Владение портами ALPC

Концепция владения портами важна для ALPC, поскольку эта технология предлагает ряд гарантий безопасности заинтересованным клиентам и серверам. В первую очередь только владелец порта соединения ALPC вправе принимать по нему подключения. Это гарантирует, что, каким бы образом дескриптор порта ни был дублирован или унаследован другим процессом, там не получится нечестным образом разрешать входящие подключения. Кроме того, когда атрибуты дескрипторов используются (косвенно или нет), они всегда дублируются в контексте процесса владельца порта вне зависимости от того, кто на данный момент анализирует сообщение.

Эти проверки имеют большое значение, когда компонент ядра может общаться с клиентом через ALPC, — компонент ядра может быть сейчас привязан к совершенно другому процессу (или даже действовать в роли части системного процесса,

передавая сообщения с порта ALPC системному потоку), и наличие информации о владельце порта позволяет ALPC избежать неоправданного доверия к текущему процессу.

Однако иногда компоненту ядра может быть выгодно не глядя принимать входящие подключения на порт вне зависимости от текущего процесса. Одним из ярких примеров является ситуация, когда объект обратного вызова исполнительной системы используется для доставки сообщения. В данном сценарии, поскольку обратный вызов происходит синхронно в контексте одного или нескольких процессов отправителей, тогда как порт подключения ядра был, скорее всего, создан во время исполнения, в системном контексте, таком как `DriverEntry`, может возникнуть несоответствие между текущим процессом и процессом — владельцем порта во время подключения. ALPC предоставляет особый флаг атрибута порта, которым могут пользоваться только вызывающие из ядра. Он помечает порт подключения как системный, в таком случае проверки владельца порта игнорируются.

Другим важным случаем применения владения портом является ситуация, когда происходит проверка SID сервера по запросу клиента, как описано в подразделе «Безопасность». Данная проверка всегда проводится в отношении токена владельца порта подключения вне зависимости от того, кто прямо сейчас может отслеживать поступление на него сообщений.

Производительность

ALPC для повышения производительности использует несколько стратегий, главным образом путем поддержки списков завершения, о которых мы кратко поговорили ранее. На уровне ядра список завершения, по сути, является пользовательской таблицей описания памяти (Memory Descriptor List, MDL), которая была опробирована и заблокирована, а затем сопоставлена с адресом (подробности об MDL см. в главе 5 тома 1). Поскольку этот механизм связан с MDL-таблицей (отслеживающей физические страницы), когда клиент отправляет сообщение на сервер, полезная нагрузка может копироваться непосредственно на физическом уровне, и не нужно запрашивать у ядра двойное буферирование сообщения, как это принято в других IPC-механизмах.

Сам по себе список завершения реализован в виде 64-разрядной очереди записей о завершениях, и для вставки и удаления записей из очереди ее потребители могут воспользоваться заблокированной операцией сравнения-обмена, как в пользовательском режиме, так и в режиме ядра. Кроме того, чтобы упростить размещение, сразу после инициализации MDL используется битовая карта для идентификации доступных областей памяти, которые могут использоваться для хранения новых сообщений, еще находящихся в очереди. Алгоритм битовой карты также использует простые инструкции блокировки процессора для атомарного выделения и освобождения областей физической памяти, которые могут использоваться списками завершения. Списки завершения могут быть настроены с помощью `NtAlpcSetInformationPort`.

И заключительная оптимизация, о которой стоит упомянуть, заключается в том, что вместо копирования данных сразу же после их отправки ядро устанавливает полезную нагрузку для отложенной копии, забирает только нужную информацию, но без какого-либо копирования. Данные сообщения копируются только тогда, когда

получатель востребует это сообщение. Вполне очевидно, что при использовании зоны сообщений или совместно используемой памяти этот метод не дает никакого преимущества, но при передаче асинхронных, буферируемых в ядре сообщений его можно использовать для оптимизации отмен и для работы сценариев с напряженным трафиком.

Управление электропитанием

Как мы видели раньше, оказавшись в среде с дефицитом электроэнергии, например на мобильных устройствах, Windows использует ряд техник, чтобы лучше распоряжаться потреблением электроэнергии и доступностью процессоров, например гетерогенную обработку на архитектурах, на которых она поддерживается (таких как архитектура ARM64 big.LITTLE), и Connected Standby как способ еще больше снизить энергопотребление в пользовательских системах без существенной нагрузки. Чтобы не мешать этим механизмам работать, ALPC реализует две дополнительные возможности: способность клиентов ALPC отправлять список удержания в канал удержания своего сервера ALPC и введение атрибута «делегирования задания». Последний является атрибутом, который отправитель может решить ассоциировать с сообщением, когда ему нужно связать свой запрос с текущим рабочим заданием, с которым он ассоциирован, или создать новое рабочее задание с описанием потока отправителя.

Такие рабочие задания используются, к примеру, когда отправитель является частью объекта задания (либо в качестве Silo/Windows-контейнера, либо в качестве части системы гетерогенного планирования и/или системы Connected StandBy), и их связь с потоком может заставить различные части системы передавать циклы процессора, пакеты запросов ввода/вывода, доли пропускной способности диска/сети и энергетические квоты «делегированному» потоку, а не действующему.

Кроме того, предоставление приоритета переднего плана и другие вмешательства в планирование предпринимаются, чтобы избежать проблем с инверсией приоритетов в архитектуре big.LITTLE, когда поток RPC «застревает» на малом ядре просто из-за того, что является фоновой службой. С рабочим заданием поток принудительно планируется на большое ядро и получает приоритет переднего плана.

Наконец, список удержания используется для обхода ситуаций с риском взаимной блокировки, когда система переходит в *режим ожидания с подключением* (Connected StandBy, в Windows 10 он же текущий режим ожидания, Modern Standby), описанный в главе 6 тома 1, либо когда готовится приостановка UWP-приложения. С его помощью можно зафиксировать время жизни процесса-владельца ALPC-порта, предотвратив принудительные операции приостановки/«замораживания» на случай, если их предпримет диспетчер жизненного цикла процессов (Process Lifetime Manager, PLM, он же диспетчер электропитания, Power Manager) даже для приложений Win32. Как только сообщения будут доставлены и обработаны, запись в списке удержания удаляется, позволяя при необходимости приостановить процесс. (Напомним, что операция завершения не проблема, так как отправка сообщения в завершенный процесс/закрытый порт немедленно активизирует отправителя особым ответом `PORT_CLOSED`, избегая блокировки из-за ответа, который не придет.)

Атрибут прямого события ALPC

Напомним, что ALPC обеспечивает клиентам и серверам два способа связи: *запросы*, которые можно отправлять в обе стороны и которые требуют ответа, и *датаграммы*, однонаправленные и не предусматривающие синхронизированного ответа. Здесь мог бы оказаться полезен промежуточный вариант — сообщение-датаграмма, на которое нельзя дать ответ, но получение которого может быть подтверждено таким образом, чтобы отправитель узнал о реакции на него, не усложняя работу обработкой ответа. Именно это фактически и реализуется с помощью *атрибута прямого события*.

Позволяя отправителю ассоциировать дескриптор с объектом события ядра (через `CreateEvent`) с сообщением ALPC, атрибут прямого события фиксирует связанную с последним структуру `KEVENT` и добавляет ссылку на него к структуре `KALPC_MESSAGE`. Затем, когда это сообщение попадет к процессу-получателю, тот сможет выдать этот атрибут и сигнализировать о нем. При этом клиент может иметь связанный с портом завершения ввода/ввода пакет окончания ожидания, также может наблюдаться синхронизированный вызов на ожидание, такой как `WaitForSingleObject` в отношении дескриптора события, который теперь получит оповещение и/или подтверждение ожидания, сообщая об успешной доставке сообщения.

Прежде такая функциональность достигалась вручную средствами RPC, которые позволяли клиентам при вызове `RpcAsyncInitializeHandle` передать параметр `RpcNotificationTypeEvent`, после чего связать `HANDLE` с событием с помощью асинхронного сообщения RPC. Вместо того чтобы принудить средства RPC с другой стороны отвечать на сообщение-запрос, тем самым позволив RPC отправителя сигнализировать о завершении события локально, ALPC фиксирует его в атрибуте Direct Event, а само сообщение помещается в отдельную очередь прямых сообщений (Direct Message Queue) вместо обычной очереди сообщений. Так подсистема ALPC подаст сигнал сразу в момент доставки сообщения, не покидая режима ядра и избегая дополнительного перехода и переключения контекста.

Отладка и трассировка

В проверенных версиях ядра возможно ведение журнала сообщений ALPC. Все атрибуты, двоичные объекты, зоны сообщений и транзакции диспетчера можно записывать по отдельности, а сами журналы — вывести в виде дампа с помощью отсутствующих в документации команд `!alpc` в WinDbg. В системах для конечного пользования системные администраторы и операторы могут добавить события ALPC в журналы ядра NT, чтобы отслеживать сообщения в них (служба трассировки событий Windows (Event Tracing, ETW) рассматривается в главе 10). События ETW не включают полезную нагрузку, но в них содержится информация о подключениях, отключениях, а также об отправке-получении и об ожидании-разблокировании. Наконец, даже в таких системах некоторые команды `!alpc` способны получать информацию о ALPC-портах и сообщениях.

ЭКСПЕРИМЕНТ. Вывод дампа порта подключения

В данном эксперименте будет использоваться порт CSRSS API для процессов Windows, запущенных в сеансе 1, который обычно является интерактивным сеансом для пользователя консоли. Всякий раз, когда запускается приложение Windows, оно подключается к API-порту CSRSS в соответствующем сеансе.

1. Сначала получите указатель на порт подключения с помощью команды !object:

```
lkd> !object \Sessions\1\Windows\ApiPort
Object: ffff898f172b2df0 Type: (ffff898f032f9da0) ALPC Port
    ObjectHeader: ffff898f172b2dc0 (new version)
    HandleCount: 1 PointerCount: 7898
    Directory Object: ffffc704b10d9ce0 Name: ApiPort
```

2. Теперь выведите информационный дамп самого объекта порта с помощью команды !alpc /p. Это, к примеру, подтвердит, что владельцем является CSRSS:

```
lkd> !alpc /P ffff898f172b2df0
Port ffff898f172b2df0
    Type : ALPC_CONNECTION_PORT
    CommunicationInfo           : ffffc704adf5d410
        ConnectionPort          : ffff898f172b2df0 (ApiPort), Connections
        ClientCommunicationPort : 0000000000000000
        ServerCommunicationPort : 0000000000000000
    OwnerProcess                : ffff898f17481140 (csrss.exe), Connections
    SequenceNo                  : 0x0023BE45 (2342469)
    CompletionPort              : 0000000000000000
    CompletionList              : 0000000000000000
    ConnectionPending           : No
    ConnectionRefused           : No
    Disconnected                : No
    Closed                      : No
    FlushOnClose                : Yes
    ReturnExtendedInfo          : No
    Waitable                    : No
    Security                    : Static
    Wow64CompletionList         : No

    5 thread(s) are waiting on the port:

        THREAD ffff898f3353b080 Cid 0288.2538 Teb: 00000090bce88000
        Win32Thread: ffff898f340cde60 WAIT
        THREAD ffff898f313aa080 Cid 0288.19ac Teb: 00000090bcf0e000
        Win32Thread: ffff898f35584e40 WAIT
        THREAD ffff898f191c3080 Cid 0288.060c Teb: 00000090bcff1000
        Win32Thread: ffff898f17c5f570 WAIT
        THREAD ffff898f174130c0 Cid 0288.0298 Teb: 00000090bcfd7000
        Win32Thread: ffff898f173f6ef0 WAIT
        THREAD ffff898f1b5e2080 Cid 0288.0590 Teb: 00000090bcfe9000
        Win32Thread: ffff898f173f82a0 WAIT
        THREAD ffff898f3353b080 Cid 0288.2538 Teb: 00000090bce88000
        Win32Thread: ffff898f340cde60 WAIT
        Main queue is empty.
        Direct message queue is empty.
        Large message queue is empty.
        Pending queue is empty.
        Canceled queue is empty.
```

3. Чтобы просмотреть, какие клиенты подключены к порту, в том числе все процессы Windows, действующие в рамках сеанса, можно использовать отсутствующую в документации команду `!alpc /lpc`. А в более новых версиях WinDbg достаточно щелкнуть на ссылке рядом с названием ApiPort. Кроме того, отобразятся порты связи сервера и клиента для каждого подключения и любые ожидающие сообщения в любой из очередей:

```
lkd> !alpc /lpc ffff898f082cbdf0

ffff898f082cbdf0('ApiPort') 0, 131 connections
        ffff898f0b971940 0 ->ffff898F0868a680 0 ffff898f17479080('wininit.exe')
        ffff898f1741fdd0 0 ->ffff898f1742add0 0 ffff898f174ec240('services.exe')
        ffff898f1740cdd0 0 ->ffff898f17417dd0 0 ffff898f174da200('lsass.exe')
        ffff898f08272900 0 ->ffff898f08272dc0 0 ffff898f1753b400('svchost.exe')
        ffff898f08a702d0 0 ->ffff898f084d5980 0 ffff898f1753e3c0('svchost.exe')
        ffff898f081a3dc0 0 ->ffff898f08a70070 0 ffff898f175402c0('fontdrvhost.ex')
        ffff898F086dcde0 0 ->ffff898f17502de0 0 ffff898f17588440('svchost.exe')
        ffff898f1757abe0 0 ->ffff898f1757b980 0 ffff898f17c1a400('svchost.exe')
```

4. Учтите, что при наличии других сеансов этот эксперимент можно повторить и для них (в том числе для нулевого системного сеанса). В конечном счете вы получите список всех процессов Windows на вашем компьютере.

СРЕДСТВО УВЕДОМЛЕНИЙ WINDOWS

Средство уведомлений Windows (Windows Notification Facility, WNF) лежит в основе современного механизма публикации и чтения без регистрации, добавленного в Windows 8 в качестве ответа на ряд архитектурных проблем, связанных с оповещением заинтересованных сторон о наличии какого-либо действия, события или состояния и с возможностью предоставить нужные данные в случае его изменения.

Проиллюстрируем это следующим сценарием: служба А желает оповестить потенциальных клиентов Б, В и Г о том, что диск был просканирован и готов к безопасной записи, при этом какое-то количество поврежденных секторов (если нашлись) были удалены. Нет никакой гарантии, что Б, В или Г запущены после А — весьма вероятно, что они стартовали раньше. В таком случае им было бы небезопасно продолжать выполнение и пришлось бы дождаться, пока А запустится и доложит о готовности диска к записи. Но если последняя еще даже не стартовала, как за ней следить?

Типичным решением для Б было бы создать событие «МОЖНО_МНЕ_ЖДАТЬ_А», затем А должна создать событие «А_ДИСК_БЕЗОПАСЕН». После этого она просигналит событие «МОЖНО_МНЕ_ЖДАТЬ_А», дав Б понять, что теперь можно безопасно дожидаться события «А_ДИСК_БЕЗОПАСЕН». В случае, когда клиент всего один, это приемлемо, но все оказывается намного сложнее, если учесть наличие В и Г, которые могут действовать по той же логике,

наперегонки создавая событие «МОЖНО_МНЕ_ЖДАТЬ_А». На тот момент они на деле получат уже существующее событие (в данном примере созданное Б) и будут ждать по нему сигнала. Хотя это и возможно, есть ли гарантия, что событие действительно создано Б? Здесь возникают проблемы, связанные с вредоносными атаками сквоттинга и отказа в обслуживании. В конечном счете можно спроектировать безопасный протокол, но это повлечет за собой значительные сложности для разработчиков А, Б, В и Г — и это мы еще не обсуждали, как посчитать поврежденные секторы.

Возможности WNF

Сценарий, описанный в предыдущем разделе, — типичная проблема проектирования операционной системы, и корректный способ ее разрешения явно не следует отдавать на откуп индивидуальным разработчикам. В обязанности операционной системы входит обеспечение простых, масштабируемых и быстродействующих решений для подобных архитектурных проблем. Решением для современных платформ Windows и является WNF. Вот его функции.

- Возможность определить *именованное состояние*, которое могут публиковать или добавлять в подписку сторонние процессы, обеспеченное стандартным дескриптором защиты Windows, включая DACL или SACL.
- Возможность ассоциировать это состояние с полезной нагрузкой до 4 Кбайт, которую можно получить по подписке на изменения состояния или опубликовать в момент изменения.
- Возможность иметь *известные имена состояний*, которые предоставляются операционной системой и не должны создаваться издателем, которые будет потенциально соревноваться с потребителями, — таким образом, потребители будут блокироваться по уведомлениию об изменении состояния, даже если издатель еще не начал работу.
- Возможность сохранять *данные состояния* даже между перезагрузками, чтобы потребители смогли увидеть их опубликованными еще до своего запуска.
- Возможность сохранять пометку о времени изменения каждого именованного состояния, чтобы потребители даже между перезагрузками могли знать, публиковались ли новые данные в какой-то момент до их запуска и следует ли реагировать на данные с прошлого раза.
- Возможность задавать *область видимости* именованному состоянию, позволяя последнему существовать во множестве экземпляров, ассоциированных с ID интерактивного сеанса, серверным хранилищем (контейнером), пользовательским ID/токеном или даже отдельно взятым процессом.
- Наконец, возможность публиковать и получать изменения именованных состояний WNF через границы режима ядра и пользовательского режима, позволив взаимодействовать компонентам, находящимся по обе ее стороны.

Пользователи WNF

Как вы уже могли заметить, реализация всей этой семантики позволит большому множеству служб компонентов ядра применять WNF, чтобы доставлять оповещения и сигналы изменения состояния сотням клиентов, от специфических API в различных системных библиотеках до крупномасштабных процессов. Сегодня WNF используется несколькими ключевыми системными компонентами и инфраструктурой.

- Диспетчер электропитания и ряд связанных с ним компонентов используют WNF, чтобы сигнализировать об открытии и закрытии крышки ноутбука, состоянии заряда батареи, включении и выключении монитора, нахождении пользователя за компьютером и т. д.
- Оболочка и ее компоненты задействуют WNF, чтобы отслеживать запуск приложений, действия пользователя, поведение меню Пуск и экрана блокировки, применение Cortana.
- Брокер системных событий (System Events Broker, SEB) является целой инфраструктурой, применяемой приложениями и брокерами UWP для получения оповещений о системных событиях наподобие ввода/вывода аудио.
- Диспетчер процессов использует временные именованные состояния WNF в разрезе процессов для реализации канала активизации, который задействуется диспетчером жизненного цикла процессов для частичного обеспечения работы механизма, позволяющего отдельным событиям принудительно сохранять активными процессы, помеченные на приостановку (глубокую заморозку).

Перечисление всех пользователей WNF могло бы занять всю эту книгу: существует более 6000 общеизвестных именованных состояний, не считая тех, что создаются временно, в том числе каналы активизации в разрезе процессов. Однако далее приводится эксперимент, где демонстрируется использование утилиты wnfdump из набора инструментов к этой книге, которая позволит читателю обозревать и активировать все события WNF в системе вместе с их данными. Инструментарий отладки Windows имеет расширение `!wnf`, тоже представленное в эксперименте и пригодное для той же цели. Кроме того, в табл. 8.31 приводятся некоторые из ключевых состояний WNF и указывается их применение. По мере раскрытия глубины и важности механизма WNF вам встретится множество компонентов и кодовых имен Windows на протяжении длинного ряда конфигураций, от Windows Phone до XBOX.

Таблица 8.31. Префиксы именованных состояний WNF

Префикс	Количество имен	Применение
9P	2	Перенаправление по протоколу Plan 9
A2A	1	Взаимодействие между приложениями
AAD	2	Azure Active Directory
AA	3	Ограниченный доступ
ACC	1	Доступность

Префикс	Количество имен	Применение
ACHK	1	Проверка целостности загрузочного диска (Autochk)
ACT	1	Активность
AFD	1	Вспомогательный функциональный драйвер (Winsock)
AI	9	Установка приложений
AOW	1	Подсистема Windows для Android (устаревшая)
ATP	1	Защита от сложных угроз (ATP) Защитника Microsoft
AUDC	15	Захват аудио
AVA	1	Голосовая активация
AVLC	3	Изменение предела громкости
BCST	1	Служба трансляции приложений
BI	16	Инфраструктура брокера
BLTH	14	Bluetooth
BMP	2	Media Player в фоновом режиме
BOOT	3	Загрузчик
BRI	1	Яркость
BSC	1	Конфигурация браузера (для прежних версий IE, устаревшая)
CAM	66	Диспетчер доступа к возможностям
CAPS	1	Централизованные политики доступа
CCTL	1	Брокер управления вызовами
CDP	17	Платформа подключенных устройств (проект Rome/передача приложений)
CELL	78	Сотовые сервисы
CERT	2	Кэш сертификатов
CFCL	3	Изменения настроек клиента в режиме полета
CI	4	Целостность кода
CLIP	6	Буфер обмена
CMFC	1	Управление конфигурацией и конфигурация свойств
CMPT	1	Совместимость
CNET	10	Сотовые сети (данные)
CONT	1	Контейнеры
CSC	1	Кэширование со стороны клиента
CSHL	1	Адаптивная оболочка
CSH	1	Хост пользовательской оболочки
CXH	6	Приложение Cloud Experience Host
DBA	1	Доступ брокера устройств
DCSP	1	Диагностический журнал поставщика служб шифрования (CSP)

Продолжение ⇩

Таблица 8.31 *(продолжение)*

Префикс	Количество имен	Применение
DEP	2	Развертывание (установка Windows)
DEVM	3	Управление устройствами
DICT	1	Словарь
DISK	1	Диск
DISP	2	Дисплей
DMF	4	Платформа переноса данных
DNS	1	Сервер доменных имен (DNS)
DO	2	Оптимизация доставки
DSM	2	Диспетчер состояний устройств
DUMP	2	Аварийный дамп
DUSM	2	Управление подписками использования данных
DWM	9	Диспетчер окон рабочего стола
DXGK	2	Ядро DirectX
DX	24	DirectX
EAP	1	Расширяемый протокол проверки подлинности
EDGE	4	Браузер Edge
EDP	15	Защита корпоративных данных
EDU	1	Образование
EFS	2	Служба зашифрованных файлов
EMS	1	Службы аварийного управления
ENTR	86	Корпоративные групповые политики
EOA	8	Специальные возможности
ETW	1	Трассировка событий для Windows
EXEC	6	Компоненты исполнения (контроль температуры)
FCON	1	Конфигурация свойств
FDBK	1	Обратная связь
FLTN	1	Уведомления о размещении рекламы
FLT	2	Диспетчер фильтрации
FLYT	1	Идентификатор полета
FOD	1	Функции по запросу
FSRL	2	Среда выполнения файловой системы (FsRtl)
FVE	15	Шифрование всего тома
GC	9	Игровое ядро
GIP	1	Графика
GLOB	3	Глобализация
GPOL	2	Групповая политика

Префикс	Количество имен	Применение
HAM	1	Диспетчер активности хоста
HAS	1	Служба аттестации хоста
HOLO	32	Службы голографии
HPM	1	Диспетчер присутствия человека
HVL	1	Библиотека гипервизора (Hvl)
HYPV	2	Hyper-V
IME	4	Редактор метода ввода
IMSN	7	Уведомления интерактивной оболочки
IMS	1	Разрешения
INPUT	5	Ввод
IOT	2	Интернет вещей
ISM	4	Диспетчер состояния ввода
IUIS	1	Масштабирование интерактивного интерфейса пользователя
KSR	2	Мягкая перезагрузка ядра
KSV	5	Потоковая передача на уровне ядра
LANG	2	Языковые свойства
LED	1	Светодиодная сигнализация
LFS	12	Платформа службы определения местоположения
LIC	9	Лицензирование
LM	7	Диспетчер лицензий
LOC	3	Геолокация
LOGN	8	Вход в систему
MAPS	3	Карты
MBAE	1	Мобильные широкополосные подключения (MBAE)
MM	3	Диспетчер памяти
MON	1	Устройства мониторинга
MRT	5	Диспетчер ресурсов Microsoft
MSA	7	Учетная запись Microsoft
MSHL	1	Минимальная оболочка
MUR	2	Запрос медийного интерефса пользователя
MU	1	Неизвестно
NASV	5	Служба физической аутентификации
NCB	1	Брокер сетевого соединения
NDIS	2	Спецификация интерфейсов сетевых драйверов (NDIS) ядра
NFC	1	Службы коммуникации ближнего поля (NFC)
NGC	12	Криптография нового поколения

Продолжение ⇩

Таблица 8.31 *(продолжение)*

Префикс	Количество имен	Применение
NLA	2	Сведения о подключенных сетях
NLM	6	Диспетчер сетевого расположения
NLS	4	Службы языковой локализации
NPSM	1	Диспетчер сеансов просмотра текущего списка воспроизведения
NSI	1	Служба интерфейса сохранения сети
OLIC	4	Лицензирование ОС
OOBE	4	Запуск при первом включении компьютера
OSWN	8	Хранилище ОС
OS	2	Базовая ОС
OVRD	1	Переопределение окон
PAY	1	Брокер платежей
PDM	2	Диспетчер устройства печати
PFG	2	Жест для первоочередного применения пера
PHNL	1	Телефонная линия
PHNP	3	Частный телефон
PHN	2	Телефон
PMEM	1	Постоянная память
PNPA-D	13	Диспетчер устройств Plug-and-Play
PO	54	Диспетчер электропитания
PROV	6	Контроль времени выполнения
PS	1	Диспетчер процессов ядра
PTI	1	Служба удаленной установки приложений
RDR	1	Перенаправитель блока серверных решений (SMB) ядра
RM	3	Диспетчер ресурсов игрового режима
RPCF	1	Диспетчер брандмауэра удаленного вызова процедур (RPC)
RTDS	2	Массив данных триггера времени исполнения
RTSC	2	Рекомендуемый клиент устранения неполадок
SBS	1	Состояние защищенной загрузки
SCH	3	Защищенный канал (SChannel)
SCM	1	Диспетчер служб
SDO	1	Изменение ориентации простых устройств
SEB	61	Брокер системных событий
SFA	1	Двухступенчатая аутентификация
SHEL	138	Оболочка
SHR	3	Общий доступ к Интернет-подключению (ICS)
SIDX	1	Индексатор поиска

Префикс	Количество имен	Применение
SIO	2	Опции входа в систему
SYKD	2	SkyDrive (Microsoft OneDrive)
SMSR	3	Маршрутизатор SMS
SMSS	1	Диспетчер сеансов
SMS	1	Сообщения SMS
SPAC	2	Дисковые пространства
SPCH	4	Речь
SPI	1	Информация о параметрах системы
SPLT	4	Обслуживание
SRC	1	Изменение системного радио
SRP	1	Системная репликация
SRT	1	Восстановление системы (Windows Recovery Environment)
SRUM	1	Исследование спящего режима
SRV	2	Блок серверных сообщений (SMB/CIFS)
STOR	3	Хранилище
SUPP	1	Поддержка
SYNC	1	Синхронизация телефона
SYS	1	Система
TB	1	Брокер времени
TEAM	4	Платформа TeamOS
TEL	5	Телеметрия ATP Защитника Microsoft
TETH	2	Связывание
THME	1	Темы
TKBN	24	Брокер сенсорной клавиатуры
TKBR	3	Брокер токена
TMCN	1	Уведомление контроля планшетного режима
TOPE	1	Событие касания
TPM	9	Доверенный платформенный модуль (TPM)
TZ	6	Часовой пояс
UBPM	4	Диспетчер электропитания пользовательского режима
UDA	1	Доступ к данным пользователя
UDM	1	Диспетчер устройств пользователя
UMDF	2	Платформа драйверов пользовательского режима
UMGR	9	Диспетчер пользователей
USB	8	Стек универсальной последовательной шины (USB)
USO	16	Оркестратор обновления

Продолжение ⇩

Таблица 8.31 *(продолжение)*

Префикс	Количество имен	Применение
UTS	2	Доверенные сигналы пользователя
UUS	1	Неизвестно
UWF	4	Объединенный фильтр записи
VAN	1	Виртуальные сети
VPN	1	Виртуальные частные сети
VTSV	2	Служба хранения учетных данных
WAAS	2	Доставка обновлений Windows
WBIO	1	Биометрия Windows
WCDS	1	Беспроводные ЛВС
WCM	6	Диспетчер подключений Windows
WDAG	2	Защита приложений Защитника Windows
WDSC	1	Настройки безопасности Защитника Windows
WEBA	2	Сетевая аутентификация
WER	3	Отчеты об ошибках Windows
WFAS	1	Служба приложения брандмауэра Windows
WFDN	3	Подключение дисплея WiFi (MiraCast)
WFS	5	Семейная безопасность Windows
WHTP	2	Библиотека протокола HTTP Windows
WIFI	15	Стек беспроводной сети (WiFi) Windows
WIL	20	Библиотека инструментария Windows
WNS	1	Служба уведомлений Windows
WOF	1	Фильтр наложения Windows
WOSC	9	Конфигурация настройки Windows
WPN	5	Push-уведомления Windows
WSC	1	Центр безопасности Windows
WSL	1	Подсистема Windows для Linux
WSQM	1	Метрики качества ПО (SQM) для Windows Software Quality Metrics
WUA	6	Обновление Windows
WWAN	5	Служба беспроводной глобальной сети (WWAN)
XBOX	116	Службы XBOX

Именованные состояния WNF и хранение

Именованные состояния WNF представлены случайными на вид 64-разрядными идентификаторами наподобие 0xAC41491908517835, которые затем скрываются макросами препроцессора C под дружественными именами, похожими на

`WNF_AUDC_CAPTURE_ACTIVE`. На деле же в их номерах кодируются номер версии (1), время жизни (постоянный или временный), область видимости (процесс, контейнер, пользователь, сеанс, компьютер), постоянный флаг данных и, для общеизвестных состояний, префикс, указывающий на владельца состояния, уникальный номер в последовательности. На рис. 8.41 приводится структура формата.

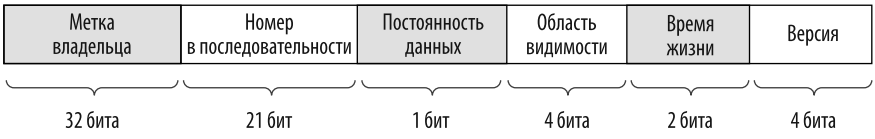


Рис. 8.41. Формат именованного состояния WNF

Как было упомянуто ранее, именованные состояния могут быть общеизвестными, а значит, они созданы заранее для общего пользования. WNF обеспечивает это, используя реестр в качестве запасного хранилища, где кодируются дескриптор защиты, максимальный размер данных и ID типа (если есть), по адресу HKLM\SYSTEM\CurrentControlSet\Control\Notifications. Для каждого состояния информация сохраняется в виде значения, соответствующего 64-разрядному коду именованного состояния WNF.

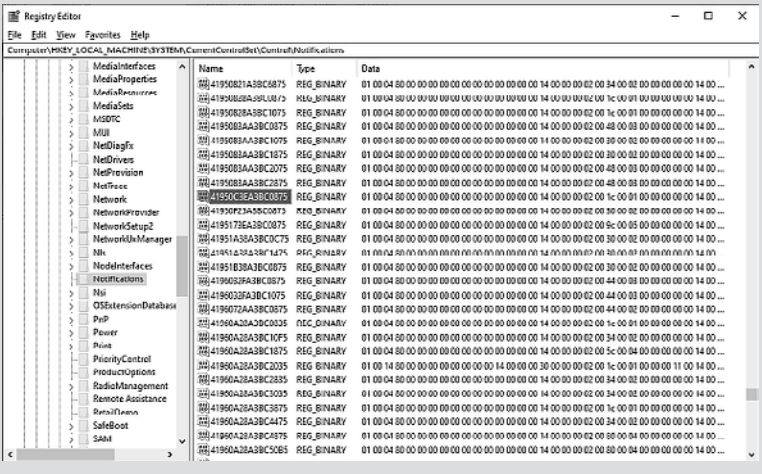
Дополнительно состояния WNF могут быть зарегистрированы как *сохраняемые (persistent)*, что означает: они будут оставаться актуальными на протяжении всего периода активности системы вне зависимости от жизненного цикла процессора-автора. Здесь повторяется поведение, упомянутое в разделе «Диспетчер объектов» данной главы, из-за чего, как и там, для регистрации таких состояний потребуется привилегия `SeCreatePermanentPrivilege`. Сохраняемые состояния тоже содержатся в реестре, в их случае по адресу HKLM\SOFTWARE\Microsoft\Windows NT\CurrentVersion\VolatileNotifications, и используют не сохраняющийся при выключении питания флаг реестра, что позволяет им исчезнуть сразу после перезагрузки компьютера. Применение не сохраняющихся при выключении питания разделов реестра для сохраняемых данных WNF может показаться странным, но стоит помнить, что, как показано ранее, фиксируются они лишь в рамках сеанса работы с включенным компьютером (в отличие от привязки к жизни процесса, что в WNF и названо временным, о чем расскажем позже).

Более того, именованное состояние WNF может быть зарегистрировано как *постоянное*, что позволяет ему сохраняться и между перезагрузками. Примерно о такой сохраняемости говорилось абзацем выше. Здесь используется еще один раздел реестра, на этот раз без флага волатильности, находящийся по адресу HKLM\SOFTWARE\Microsoft\Windows NT\CurrentVersion\Notifications. Соответственно, для такого уровня фиксации тоже потребуется привилегия `SeCreatePermanentPrivilege`. Для таких состояний WNF используется дополнительный раздел реестра, располагающийся ниже по иерархии, под названием `Data`, где для каждого 64-разрядного идентификатора хранятся время последнего изменения и бинарные данные. Обратим внимание на то, что, если состояние WNF на данном компьютере еще не менялось, последнего блока не будет.

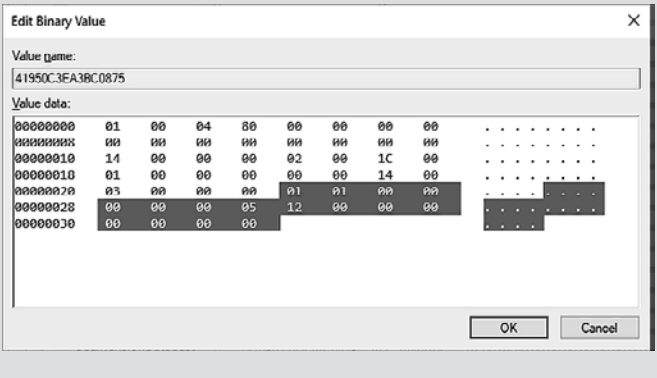
ЭКСПЕРИМЕНТ. Просмотр именованных состояний WNF и их данных через реестр

В этом эксперименте используется редактор реестра, чтобы просмотреть общеизвестные имена NWF, а также несколько примеров постоянных и сохраняемых среди них. В бинарных данных в реестре вы сможете разглядеть информацию о дескрипторе защиты и данных.

Запустите редактор реестра и проследуйте в раздел HKEY_LOCAL_MACHINE\SYSTEM\CurrentControlSet\Control\Notifications. Взгляните на отобразившиеся значения, выглядящие примерно так, как на снимке экрана ниже.

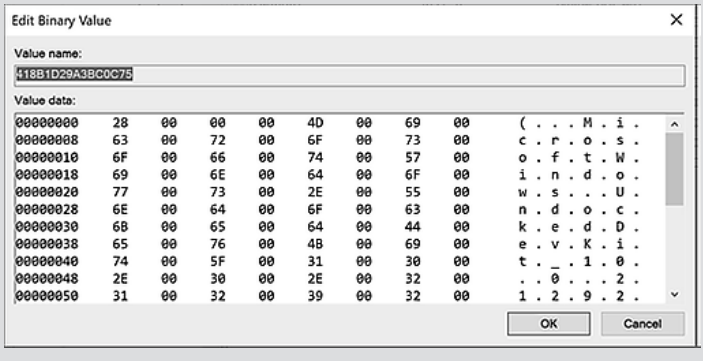


Дважды щелкните на значении 41950C3EA3BC0875 (WNF_SBS_UPDATE_AVAILABLE), чтобы открылся редактор бинарных данных реестра. Обратите внимание на то, что на следующем снимке экрана виден дескриптор защиты (выделенные бинарные значения, включающие в себя SID S-1-5-18), а также максимальный размер данных (0 байт).



Воздержитесь от изменения любых показанных значений, поскольку из-за этого система может стать уязвимой к атаке или потерять работоспособность.

Наконец, чтобы увидеть некоторые примеры постоянного состояния WNF, перейдете в редакторе реестра по адресу HKEY_LOCAL_MACHINE\SOFTWARE\Microsoft\Windows NT\CurrentVersion\Notifications\Data и откройте значение `418B1D29A3BC0C75 (WNF_DSM_DSMAPPINSTALLED)`. На следующем рисунке видно, какое последнее приложение было установлено в данной системе (`MicrosoftWindows.UndockedDevKit`).



Случайное именованное состояние может быть зарегистрировано и как *временное*. У таких есть несколько отличий от упомянутых ранее. Во-первых, поскольку их имена не известны заранее, потребителю и производителю требуется какой-то способ передать друг другу этот идентификатор. В обычном случае тот, кто первый попытается опубликовать или прочесть данные состояния, вместо этого сам создает/использует с помощью внутренних средств соответствующий раздел реестра для сохранения данных. Однако с временными состояниями WNF такое невозможно, поскольку имя основано на монотонно возрастающем порядковом номере.

Во-вторых, как следует из первого, для кодировки временных именованных состояний разделы реестра не используются — они привязаны к процессу, зарегистрировавшему их экземпляр, а все данные хранятся только в пуле ядра. Такие задействуются, к примеру, в роли каналов активизации для процессов, описанных ранее. Другие варианты применения включают в себя оповещения диспетчера электропитания и триггеры прямого обслуживания, используемые в SCM.

Модель публикаций и подписок WNF

Когда для публикации используется WNF, производитель придерживается стандартной последовательности из регистрации именованного состояния (в случае не общеизвестных состояний) и публикации каких-либо данных, которые требуется показать. Можно и не показывать последние, а просто передать буфер нулевой длины, который включает состояния и дает сигнал подписчику, даже если данных не было.

Потребители, в свою очередь, используют регистрационные возможности WNF, чтобы связать свою функцию-реакцию с определенным именованным состоянием. Всякий раз при публикации изменения вызывающая сторона должна обратиться к соответствующему WNF API для получения данных по искомому состоянию (чтобы вызвавший при необходимости мог выделить область памяти или стека, предоставляется размер буфера). В пользовательском режиме, в свою очередь, низкоуровневый компонент оповещения WNF из NtDll.dll самостоятельно размещает буфер в куче и передает указатель на данные напрямую в функцию-реакцию подписчика.

В обоих случаях также предоставляется метка изменения — уникальный номер в монотонной последовательности, который можно использовать, чтобы обнаружить пропущенные публикации данных в случае, если подписчик был неактивен, а публикации продолжились. Наконец, с реакцией можно ассоциировать персональный контекст, что полезно для ситуаций в C++, когда статический указатель на функцию привязывается к ее классу.

ПРИМЕЧАНИЕ В WNF есть API, позволяющий запросить, зарегистрировано ли в системе то или иное именованное состояние (позволяющее потребителю реализовать особое поведение, если производитель пока не работает), равно как и API для проверки наличия активных подписок на состояния (позволяя производителю реализовать особое поведение, к примеру воздержаться от дополнительных публикаций, которые перепишут старые данные).

WNF может управлять тысячами подписок посредством привязки структуры данных к каждой подписке пользовательского режима или режима ядра, объединяя их вместе для каждого заданного именованния состояния WNF. Таким образом, когда состояние публикуется, анализируется список подписок, а в пользовательском режиме в связанный список добавляется блок полезной нагрузки, после чего каждому процессу сигнализируется о событии уведомления. Это предписывает доставочному коду WNF из NtDll.dll начать вызывать API, чтобы тот потребил эту полезную нагрузку, как и все прочие блоки полезных данных в связанном списке. В режиме ядра механика проще — обратный вызов синхронно выполняется в контексте публикующего.

Обратите внимание на то, что есть еще возможность подписываться на оповещение в двух режимах: `data-notification` и `meta-notification`. Первый предполагает типичное поведение: вызвать реакцию, когда с состоянием WNF оказались связаны новые данные. Второй интереснее, так как может давать оповещения о том, стал активным новый подписчик или ушел старый, равно как и сообщать, что издатель завершил работу (в случае с волатильными состояниями, когда такое возможно).

Наконец, стоит добавить, что подписки пользовательского режима имеют еще одну особенность. Поскольку Ntdll.dll управляет оповещениями WNF в рамках процесса в целом, сразу несколько компонентов (например, динамических библиотек/DLL) могут запросить собственные реакции для одного именованного состояния, но по разным причинам и в разных контекстах. В таком случае Ntdll.dll вынуждена связать с каждым модулем свои контексты регистрации, чтобы направленную в процесс полезную нагрузку можно было ретранслировать на вход нужной реакции и доставить, только если запрошенный режим доставки подходит типу оповещения для подписчика.

ЭКСПЕРИМЕНТ. Использование утилиты WnfDump для вывода дампа именованных состояний WNF

В данном эксперименте вы воспользуетесь одним из инструментов, прилагаемых к книге (WnfDump), чтобы подписаться на именованные состояния `WNF_SHEL_DESKTOP_APPLICATION_STARTED` и `WNF_AUDC_RENDER`.

Запустите `WnfDump` из командной строки со следующими аргументами:

```
-i WNF_SHEL_DESKTOP_APPLICATION_STARTED -v
```

Утилита показывает сведения об именованном состоянии и читает его данные, что можно увидеть в следующем выводе:

```
C:\>wnfdump.exe -i WNF_SHEL_DESKTOP_APPLICATION_STARTED -v
WNF State Name                                  | S | L | P | AC | N | CurSize |
MaxSize
-------------------------------------------------------------------------------
WNF_SHEL_DESKTOP_APPLICATION_STARTED            | S | W | N | RW | I |    28   |
   512
65 00 3A 00 6E 00 6F 00-74 00 65 00 70 00 61 00  e.:.n.o.t.e.p.a.
64 00 2E 00 65 00 78 00-65 00 00 00              d...e.x.e...
```

Поскольку данное событие связано с Проводником (оболочкой), запускающим приложения, вы увидите одно из последних приложений, которые открывали двойным щелчком кнопкой мыши или из меню Пуск, в том числе командой Выполнить. В общем, любое, в отношении которого использовался API `ShellExecute`. Метка изменения тоже отображена, по сути, в виде счетчика приложений, которые запускали таким образом с момента загрузки данного экземпляра Windows (поскольку это *сохраняемое*, а не *постоянное событие*).

Откройте еще какое-нибудь приложение, например Paint, из меню Пуск и снова попробуйте запустить `wnfdump`. Вы увидите, как поменялась метка изменений и отобразились другие бинарные данные.

Агрегация событий WNF

Хотя подсистема WNF сама по себе обеспечивает клиентов и серверы мощным инструментом обмена информацией и оповещения о статусах друг друга, встречаются ситуации, когда какой-либо подписчик/клиент заинтересован более чем в одном именованном состоянии. Например, есть одно состояние WNF, которое публикуется при отключении подсветки экрана, второе, отвечающее за момент отключения беспроводного адаптера, и третье, сообщающее, что пользователь ушел от компьютера. Одному подписчику нужно будет, чтобы его оповещали обо *всех* этих событиях, тогда как другому — только о первых двух или лишь о третьем.

К сожалению, системные вызовы и инфраструктура WNF, предусмотренные в Ntdll.dll для клиентов пользовательского режима (равно как и для видимого API от ядра), работают с именованными состояниями только поодиночке. Таким

образом, в данных случаях потребуется самостоятельная обработка по принципу конечного автомата, которую придется реализовать каждому подписчику.

Чтобы упростить это часто встречающееся требование, существует специальный компонент, облегчающий задачу предоставлением простого API, — общий агрегатор событий (Common Event Aggregator, CEA) для обращений в режиме ядра, реализованный в CEA.SYS, и EventAggregation.dll для пользовательского режима. Обе библиотеки экспортируют набор функций (таких как `EaCreateAggregatedEvent` и `EaSignalAggregatedEvent`), которые разрешают поведение наподобие прерываний (с реакцией на «старт», когда состояние WNF активизируется, и реакцией на «стоп», когда завершается), в том числе с возможностью указания наборов условий с помощью операторов AND, OR и NOT.

В число пользователей CEA входят стек USB, а вместе с ним платформа драйверов Windows, которая демонстрирует реакцию на изменение именованных состояний WNF. Кроме того, служба Power Delivery Coordinator (Pdc.sys) использует CEA, чтобы собирать именования состояния питания компьютеров, как в примере, приведенном в начале этого раздела. Универсальный диспетчер фоновых процессов (Unified Background Process Manager, UBPM), рассматриваемый в главе 9, также полагается на CEA при реализации таких возможностей, как запуск и остановка служб в зависимости от нехватки питания или бездействия.

Наконец, WNF имеет ключевое значение для службы, называемой брокером системных событий (System Event Broker, SEB) и реализованной в SystemEventBroker.dll в паре с клиентской библиотекой SystemEventBrokerClient.dll. Брокер экспортирует такие API, как `SebRegisterPrivateEvent`, `SebQueryEventData` и `SebSignalEvent`, которые перенаправляются к самой службе через RPC. В пользовательском режиме SEB лежит в основе универсальной платформы Windows (UWP) и других API для опроса состояния системы, а также служб, которые реагируют в зависимости от определенных изменений в состояниях, предоставляемых WNF. Наконец, наиболее мощно SEB показывает себя на одноядерных системах, таких как Windows Phone и XBOX (где, как отмечалось ранее, есть несколько сотен общеизвестных именованных состояний WNF), там он заменил устаревший диспетчер окон, работавший через такие сообщения, как `WM_DEVICEARRIVAL`, `WM_SESSIONENDCHANGE`, `WM_POWER` и т. д.

SEB имеет каналы с инфраструктурой брокеров, используемой приложениями UWP, которая позволяет приложениям даже в рамках `AppContainer` иметь доступ к событиям, дающим выход на состояние системы в целом. В свою очередь, для приложений WinRT пространство имен `Windows.ApplicationModel.Background` публикует класс `SystemTrigger`, реализующий интерфейс `IBackgroundTrigger`, который выводит на RPC-сервисы SEB и его API для C++, а тот фактически преобразует некоторые общеизвестные системные события в именованные состояния `WNF_SEB_XXX`. Все это служит превосходным примером того, как что-то столь глубинное и недокументированное, как WNF, может лежать в основе столь детально описанного и современного API для разработки приложений UWP. SEB является лишь одним из множества брокеров, которые предоставляет UWP, и в конце главы мы подробно рассмотрим фоновые задачи и инфраструктуру брокеров.

ОТЛАДКА В ПОЛЬЗОВАТЕЛЬСКОМ РЕЖИМЕ

Поддержка отладки в пользовательском режиме разделена на три модуля. Первый входит непосредственно в исполнительную систему и имеет префикс `Dbgk`, что значит *Debugging Framework*. Он обеспечивает необходимые внутренние функции для регистрации и отслеживания событий отладки, управления объектом отладки и упаковки информации с целью ее потребления его партнерами, работающими в пользовательском режиме. Компоненты пользовательского режима, напрямую связанные с `Dbgk`, расположены в платформенно-зависимой системной библиотеке Ntdll.dll, и представлены там набором API-функций с префиксом `DbgUI`. Они играют роль оболочки для непубличной реализации объекта отладки и позволяют всем приложениям подсистемы использовать отладку, чтобы оборачивать уже свои API вокруг реализации `DbgUi`. Наконец, третий компонент отладки пользовательского режима входит в DLL подсистемы. Это открытый документированный API для подсистемы Windows, входящий в KernelBase.dll и поддерживаемый каждой подсистемой для отладки прочих приложений.

Поддержка со стороны ядра

Ядро поддерживает отладку пользовательского режима через упомянутую ранее структуру — *объект отладки*. Оно обеспечивает ряд системных вызовов, большинство из которых являются прямым отражением отладочного API Windows, как правило доступного в первую очередь через уровень `DbgUi`. Сам по себе объект отладки является довольно простой конструкцией, состоящей из серии флагов, определяющих состояние, события для уведомления любого ожидающего потока о присутствии событий отладчика, списка с двойной связью с событиями отладки, ожидающими обработки, и быстрого мьютекса, используемого для блокировки объекта. Это вся информация, которая требуется ядру для успешного получения и отправки событий отладки, и каждый подвергаемый отладке процесс имеет в своей структуре элемент порта отладки, указывающий на этот объект отладки.

После того как у процесса появится связанный с ним порт отладки, события, указанные в табл. 8.32, могут стать причиной вставки события отладки в список событий.

Таблица 8.32. События отладки режима ядра

Идентификатор события	Его значение	Причина возникновения
DbgKmExceptionApi	Произошло исключение	Вызов KiDispatchException в ходе исключения, произошедшего в пользовательском режиме
DbgKmCreateThreadApi	Создан новый поток	Запуск потока пользовательского режима
DbgKmCreateProcessApi	Создан новый процесс	Запуск потока пользовательского режима, являющегося первым потоком процесса, если только флаг CreateReported еще не установлен в EPROCESS

Продолжение ⇩

Таблица 8.32 *(продолжение)*

Идентификатор события	Его значение	Причина возникновения
DbgKmExitThreadApi	Произошел выход из потока	Завершение потока пользовательского режима в случае, когда флаг ThreadInserted установлен в ETHREAD
DbgKmExitProcessApi	Произошел выход из процесса	Завершение потока пользовательского режима, оказавшегося последним в текущем процессе, в случае, когда флаг ThreadInserted установлен в ETHREAD
DbgKmLoadDllApi	Загружена DLL-библиотека	Вызов NtMapViewOfSection, если раздел является файлом образа (может быть и с расширением EXE), когда флаг SuppressDebugMsg из TEB не установлен
DbgKmUnloadDllApi	Выгружена DLL-библиотека	Вызов NtUnmapViewOfSection, если раздел является файлом образа (может быть и с расширением EXE), когда флаг SuppressDebugMsg из TEB не установлен
DbgKmErrorReportApi	Исключение должно быть направлено в Windows Error Reporting (WER)	Когда DbgKmExceptionApi возвращает DBG_EXCEPTION_NOT_HANDLED, это особое сообщение отправляется не через объект отладки, а через ALPC, чтобы WER мог продолжить обработку исключения

Кроме случаев из таблицы, существует еще пара особых ситуаций, не входящих в типичный сценарий, которые происходят в то время, когда подвергаемый отладке объект становится связанным с процессом. Когда отладчик активен, сначала первые сообщения *create process* и *create thread* будут отправлены принудительно для самого процесса и его главного потока, затем сообщение *create thread* будет выдано всем остальным потокам данного процесса. Наконец, будут отправлены события *load dll* для исполняемой программы, подвергаемой отладке (с помощью Ntdll.dll), и остальных загружаемых DLL данного процесса. Подобным образом, если отладчик подключен, но создается клон процесса (форк), такие события будут отправлены для первого потока в нем, поскольку в клонированном адресном пространстве уже есть не только Ntdll.dll, но и все остальные загруженные DLL.

Также предусмотрен особый флаг `hide from debugger` (Скрыть от отладчика), который можно установить на потоке как при его создании, так и при исполнении. Когда этот флаг установлен, что отражается на флаге `HideFromDebugger` из TEB, все действия текущего потока даже при наличии порта отладки не будут производить отладочных сообщений.

Когда объект отладчика устанавливает связь с процессом, тот входит в состояние *глубокой заморозки* (deep freeze), используемое и для приложений UWP. Напомним, что также это приостанавливает все потоки и исключает удаленное создание новых потоков. С этого момента ответственность за запуск запросов на отправку событий отладки возлагается на отладчик. Отладчики запрашивают обратную отправку событий отладки в пользовательский режим, выполняя ожидание на объекте отладки. Это приводит к циклическому перебору списка событий отладки. По мере того как

каждый запрос удаляется из списка, его содержимое переводится из внутренней dbgk-структуры в структуру, привычную и понятную для следующего вышестоящего уровня. Как вы увидите, эта структура также отличается и от структуры Win32, и должен быть проведен еще один уровень преобразования. Даже после обработки отладчиком всех отложенных сообщений отладки ядро не возобновляет выполнение процесса в автоматическом режиме. Чтобы возобновить выполнение, отладчик должен вызвать функцию `ContinueDebugEvent`.

Не считая некоторых других, более сложных ситуаций в условиях многопоточности, базовая модель для данного фреймворка сводится к работе *продюсеров* — кода ядра, создающего отладочные события, приведенные в табл. 8.32, и *потребителей* — отладчика, ожидающего данные события и подтверждающего их получение.

Платформенно-зависимая поддержка

Хотя базовый протокол отладки в пользовательском режиме довольно прост, приложения Windows не могут задействовать его напрямую — вместо этого он заключается в оболочку `DbgUi`-функциями из библиотеки Ntdll.dll. Такая абстракция нужна, чтобы позволить пользоваться этими процедурами как платформенно-зависимым приложениям, так и различным подсистемам, поскольку код в составе Ntdll.dll не имеет зависимостей. Функции, которые предоставляет данный компонент, в основном аналогичны таковым в рамках Windows API и связанным с ними системным вызовам. Внутренний же код обеспечивает функциональность, необходимую для создания объекта отладки, связанного с конкретным потоком. Для этого объекта создается дескриптор, но он никогда не раскрывается. Вместо этого система сохраняет его в блок окружения потока (TEB) от потока отладчика, который и выполняет привязку. (Подробности о TEB см. в главе 4 тома 1.) Само значение помещается в поле `DbgSsReserved[1]`.

Когда отладчик присоединяется к процессу, ожидается, что процесс будет разбит на части, то есть должна произойти операция `int 3` (установка контрольной точки), сгенерированная потоком, вставленным в процесс. Если этого не случится, отладчик никогда не сможет взять контроль над процессом и будет просто наблюдать за пролетающими мимо событиями отладки. За создание этого потока и его внедрение в нужный процесс отвечает библиотека Ntdll.dll. Надо заметить, что поток создается с особым флагом, который ядро заносит в TEB, а это приводит к установке флага `SkipThreadAttach`, позволяя избежать оповещений `DLL_THREAD_ATTACH` и занятия слота TLS, что могло бы вызывать нежелательные побочные эффекты всякий раз, когда отладчик вмешивается в процесс.

Наконец, Ntdll.dll также предоставляет API для преобразования платформенно-зависимой структуры событий отладки в структуру, понятную Windows API-функциям. Это происходит согласно правилам, приведенным в табл. 8.33.

Таблица 8.33. Правила конвертации платформенно-зависимого формата в Win32

Изменение платформенно-зависимого состояния	Изменение состояния Win32	Примечания
DbgCreateThreadStateChange	CREATE_THREAD_DEBUG_EVENT	lpImageName всегда NULL, а fUnicode всегда TRUE

Продолжение ⇩

Таблица 8.33 *(продолжение)*

Изменение платформенно-зависимого состояния	Изменение состояния Win32	Примечания
DbgCreateProcessStateChange	CREATE_PROCESS_DEBUG_EVENT	
DbgExitThreadStateChange	EXIT_THREAD_DEBUG_EVENT	
DbgExitProcessStateChange	EXIT_PROCESS_DEBUG_EVENT	
DbgExceptionStateChange, DbgBreakpointStateChange, DbgSingleStepStateChange	OUTPUT_DEBUG_STRING_EVENT, RIP_EVENT или EXCEPTION_ DEBUG_EVENT	Конкретное решение зависит от кода исключения — это может быть DBG_PRINTEXCEPTION_C/ DBG_PRINTEXCEPTION_ WIDE_C, DBG_RIPEXCEPTION или что-то иное
DbgLoadDllStateChange	LOAD_DLL_DEBUG_EVENT	fUnicode всегда TRUE
DbgUnloadDllStateChange	UNLOAD_DLL_DEBUG_EVENT	

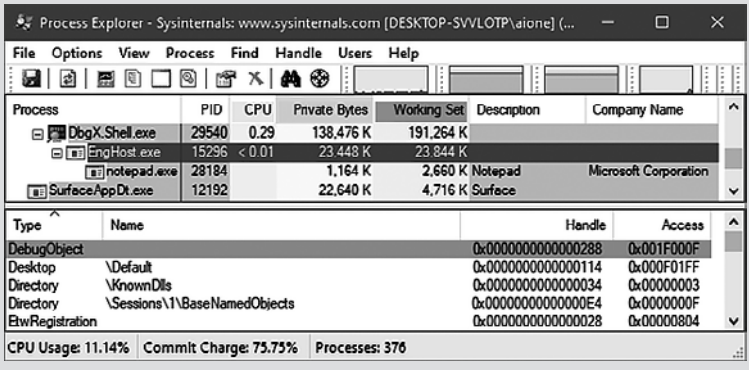
ЭКСПЕРИМЕНТ. Просмотр объектов отладчика

Ранее вы использовали утилиту WinDbg для отладки в режиме ядра, но ее можно применять и для отладки программ пользовательского режима. Для начала запустите Блокнот — Notepad.exe, подключив отладчик следующим образом.

1. Запустите WinDbg, затем в меню Файл (File) выберите Открыть исполняемый (Open Executable).
2. Зайдите в каталог `\Windows\System32\` и выберите Notepad.exe.
3. Мы не собираемся ничего отлаживать, так что проигнорируйте вопросы отладчика. Чтобы указать WinDbg продолжить выполнение Notepad.exe, можно ввести команду g.

Затем запустите Process Explorer и включите нижнюю панель, настроив ее на отображение открытых дескрипторов (в меню Вид (View) выберите Вид нижней панели (Lower Pane View), а потом Дескрипторы (Handles)). Вам также нужно видеть безымянные дескрипторы, поэтому в меню Вид (View) выберите Показать безымянные дескрипторы и отображения (Show Unnamed Handles And Mappings).

Далее щелкните на процессе WinDbg.exe (или EngHost.exe, если пользуетесь WinDbg Preview) и взгляните на его таблицу дескрипторов. Там должен отобразиться открытый безымянный дескриптор объекта отладки. (Чтобы быстрее отыскать нужную запись, эту таблицу можно отсортировать по типу, щелкнув на заголовке Тип (Type).) Вы должны увидеть нечто подобное.



Можете попробовать щелкнуть правой кнопкой мыши на этом дескрипторе и закрыть его. Тогда окно Блокнота пропадет, а в WinDbg появится такое сообщение:

```
ERROR: WaitForEvent failed, NTSTATUS 0xC0000354
This usually indicates that the debuggee has been
killed out from underneath the debugger.
You can use .tlist to see if the debuggee still exists.
```

Фактически если вы заглянете в описание показанного кода NTSTATUS, то найдете там следующий текст: «Попытка выполнить операцию над портом отладки не удалась, так как указанный порт находится в процессе удаления». Это в точности описывает то, что вы сделали, закрыв дескриптор.

Как видите, платформенно-зависимый `DbgUi` мало что делает для поддержки фреймворка, кроме этой абстракции. Самая сложная его задача — это конвертация структур данных отладчика из платформенно-зависимого формата в Win32. Это подразумевает несколько дополнительных изменений в них.

Поддержка подсистемы Windows

Последний компонент, обеспечивающий таким отладчикам, как Microsoft Visual Studio или WinDbg, возможность вести отладку приложений пользовательского режима, находится в KernelBase.dll. В нем реализована соответствующая часть документированного Windows API. Помимо простого преобразования одного имени функции в другое, эта сторона инфраструктуры отладки отвечает еще за одну важную управленческую задачу: управление продублированными дескрипторами файла и потока.

Напомним, что каждый раз, когда отправляется событие `load DLL`, дескриптор файла образа дублируется ядром и передается через объект события, аналогично поступая с дескриптором исполняемого файла процесса для события `create process`. Во время каждого вызова `wait` KernelBase.dll проверяет, приводит ли это событие к созданию новых дублированных дескрипторов процесса и/или потока ядром (два

события `create`). Если это так, библиотека размещает структуру, в которой сохраняет ID процесса, ID потока и дескриптор процесса и/или потока, связанного с этим событием. Затем она привязывается к первому элементу массива `DbgSsReserved` в TEB, где, как упоминалось ранее, хранится дескриптор объекта отладки. Подобным образом KernelBase.dll проверяет события `exit`. При обнаружении таких событий она помечает дескрипторы в этой структуре данных.

Как только отладчик закончит пользоваться дескрипторами и выполнит вызов *continue*, KernelBase.dll проанализирует эти структуры, определит дескрипторы потоков, сделавших *exit*, и закроет их для отладчика. В ином случае такие потоки и процессы никогда не завершатся, потому что из-за присутствия отладчика их дескрипторы не будут закрываться.

ПАКЕТНЫЕ ПРИЛОЖЕНИЯ

Начиная с Windows 8, появилась потребность в некоторых API, работающих на различных типах устройств, от мобильных телефонов до XBox и полноценных персональных компьютеров. Действительно, тогда Windows всерьез стала проектироваться даже для новых видов устройств, которые используют сторонние платформы и процессорные архитектуры (хорошим примером будет ARM). В Windows 8 была представлена новая независимая от платформы архитектура приложений Windows Runtime, также известная как WinRT. Она поддерживала разработку на C++, JavaScript и языках с управляемым кодом (C#, VB.Net и т. д.), базировалась на технологии COM и в неуправляемом коде позволяла работу на процессорах x86, AMD64 и ARM. Универсальная платформа Windows (Universal Windows Platform, UWP) стала продолжением WinRT. Она была спроектирована с целью избавиться от некоторых ограничений WinRT и построена на ее основе. Приложениям UWP больше не нужно сообщать, на какой версии ОС их разрабатывали, через файлы `manifest`, вместо этого они предназначены сразу для целого семейства устройств, и не одного.

UWP предоставляет API, называемый универсальным семейством устройств (Universal Device Family), который гарантированно присутствует во всех семействах устройств и зависимых от устройства API-расширениях. Разработчик может специализироваться на одном типе устройства, добавив SDK-расширения в `manifest`, и даже больше, имеет возможность при необходимости протестировать наличие какого-либо API во время исполнения, соответствующим образом адаптируя поведение своего приложения. Таким образом, приложение UWP, работая на смартфоне, может начать вести себя так, как если бы его запустили на ПК, если устройство подключат к стационарному компьютеру или подходящей док-станции.

UWP предоставляет своим приложениям ряд возможностей.

- Адаптивные элементы управления и ввод. Графические элементы реагируют на размер и разрешение экрана, изменяя свои масштаб и раскладку. Более того, обработка ввода отделена от самого приложения. Это значит, что приложение UWP сможет надежно работать на разных экранах с разными типами устройств ввода, такими как тачпад, перо, мышь, клавиатура или джойстик от XBox.
- Централизованный магазин для всех приложений UWP, который обеспечивает комфортный процесс установки, обновления или удаления.

- Унифицированная система дизайна, известная как Fluent (интегрирована в Visual Studio).
- Среда-«песочница» под названием AppContainer.

Технология AppContainer изначально создавалась для WinRT и по сей день используется приложениями UWP. Аспекты безопасности сред AppContainer мы описывали в главе 7 тома 1.

Чтобы правильно запускать и администрировать приложения UWP, в Windows была выстроена новая модель приложений, известная как AppModel (полное название — современная модель приложений, Modern Application Model). С каждым новым релизом ОС современная модель приложений развивалась и менялась. В данной книге мы рассматриваем эту технологию применительно к Windows 10. В новую модель вошло множество компонентов, которые действуют сообща, чтобы корректно управлять состояниями пакетного приложения и его фоновой деятельностью в энергосберегающей манере.

- **Диспетчер активности хоста** (Host Activity Manager, HAM). Это новый компонент, добавленный в Windows 10, который замещает и интегрирует ряд старых компонентов, контролировавших жизненный цикл (и состояния) приложения UWP (диспетчер жизненного цикла процессов, диспетчер переднего плана, политики ресурсов и диспетчер ресурсов). Диспетчер активности хоста действует в рамках службы фоновой инфраструктуры задач (BrokerInfrastructure) — не путайте с компонентом фоновой инфраструктуры брокеров — и тесно взаимодействует с диспетчером состояний процессов. Он реализован в двух библиотеках, одна из которых представляет интерфейс клиента (Rmclient.dll), а вторая — интерфейс сервера (PsmServiceExtHost.dll).
- **Диспетчер состояний процессов** (Process State Manager, PSM). PSM был частично заменен HAM и считается частью последнего (на самом деле PSM стал клиентом HAM). Он обслуживает и хранит состояние каждого хоста пакетного приложения. Реализован диспетчер в рамках той же службы, что и HAM (BrokerInfrastructure), но в другой DLL — Psmsrv.dll.
- **Диспетчер активации приложений** (Application Activation Manager, AAM). Компонент AAM отвечает за различные типы и виды активации пакетного приложения. Он реализован в библиотеке ActivationManager.dll, которая действует в рамках службы, называемой диспетчером пользователей. Диспетчер активации приложений является клиентом HAM.
- **Диспетчер отображения** (View Manager, VM). VM определяет и администрирует деятельность и события пользовательского интерфейса UWP, взаимодействует с HAM, поддерживая UI-приложения на переднем плане и в неприостановленном состоянии. Кроме того, VM помогает HAM определять, когда приложение UWP переходит в фоновое состояние. Диспетчер отображения реализован в управляемой .Net библиотеке CoreUiComponents.dll, которая зависит от клиентского интерфейса модернизированного диспетчера исполнения (ExecModelClient.dll), что позволяет корректно регистрироваться для работы с HAM. Обе библиотеки действуют в рамках диспетчера пользователей, который работает в составе процесса SiHost (он нужен службе для корректного управления UI-событиями).

- **Фоновая инфраструктура брокеров** (Background Broker Infrastructure, BI). BI управляет фоновыми задачами приложений, их политиками исполнения и событиями. Основной сервер, по большей части реализованный в библиотеке bisrv.dll, управляет событиями, которые генерируют брокеры, и оценивает политики для принятия решений по запуску фоновой задачи. Фоновая инфраструктура брокеров работает в рамках службы BrokerInfrastructure и на момент написания данной книги не используется в приложениях Centennial.

Чтобы обеспечить возможность запускать даже стандартные приложения Win32 в защищенных системах вроде Windows 10 S и позволить конвертировать устаревшие приложения под новую модель, в Microsoft была разработана технология Desktop Bridge (внутреннее название Centennial). Разработчикам она доступна в таких средствах, как Visual Studio или Desktop App Converter. Запускать приложения Win32 в AppContainer возможно, но не рекомендуется просто потому, что стандартные программы под Win32 спроектированы для использования более широкого набора системных API, доступ к которым в средах AppContainer существенно ограничен.

Приложения UWP

Мы уже описывали появление приложений UWP и защищенную среду, в которой они работают, в главе 7 тома 1. Чтобы лучше понять концепции, изложенные в данной главе, полезно будет определить некоторые базовые особенности современных приложений UWP. В Windows 8 процессы получили такие новые важные свойства, как:

- идентификация пакета;
- идентификация приложения;
- AppContainer;
- модернизированный UI.

Ранее мы уже подробно разбирали среду AppContainer (см. главу 7 тома 1). Когда пользователь скачивает современное приложение UWP, обычно оно доставляется оформленным в рамках пакета AppX. В пакете могут находиться по несколько приложений, опубликованных одним автором и связанных друг с другом. Идентификация пакета — это логический конструкт, обеспечивающий его уникальность. В него входят пять параметров: имя, версия, архитектура, ID ресурса и издатель. Идентификация пакета может быть представлена двумя вариантами: полным именем пакета, а именно строкой из всех фрагментов идентификации, разделенных символом подчеркивания, либо именем семейства пакетов — другой строкой, содержащей название пакета и издателя. В обоих случаях издатель представлен строкой со своим полным именем, закодированным по алгоритму Base32. В терминологии UWP термины «ID пакета» и «полное имя пакета» эквивалентны. Например, пакет с программой Adobe Photoshop распространяется под полным именем `AdobeSystemsIncorporated.AdobePhotoshopExpress_2.6.235.0_neutral_split.scale-125_ynb6jyjzte8ga`, где:

- `AdobeSystemsIncorporated.AdobePhotoshopExpress` — это имя пакета;
- `2.6.235.0` — версия;

- `neutral` — целевая архитектура;
- `split_scale` — ID ресурса;
- `ynb6jyjzte8ga` — имя издателя, закодированное по алгоритму Base32 (вариация по Крокфорду, которая исключает буквы i, l, u и o, чтобы не путать их с цифрами).

Имя семейства пакетов будет проще, а именно `AdobeSystemsIncorporated.AdobePhotoshopExpress_ynb6jyjzte8ga`.

Каждое приложение в составе пакета представлено идентификацией приложения. Этот конструкт обеспечивает уникальное обозначение набора окон, процессов, ярлыков, значков и функциональности в рамках отдельного пользовательского приложения вне зависимости от его реальной реализации (таким образом, в методологии UWP отдельное приложение может состоять из нескольких процессов, все еще входящих в состав одной идентификации приложения). Идентификация приложения представлена простой строкой, в методологии UWP известной как ID приложения в рамках пакета (Package Relative Application ID, часто сокращенно PRAID). Та, в свою очередь, часто комбинируется с именем семейства пакетов для формирования ID пользовательской модели приложения (Application User Model ID, AUMID). Например, приложение современного меню Пуск в Windows имеет следующий AUMID: `Microsoft.Windows.ShellExperienceHost_cw5n1h2txyewy!App`, где `App` — это PRAID.

Полное имя пакета вместе с идентификацией приложения располагаются в атрибуте безопасности `WIN://SYSAPPID` в рамках токена, описывающего контекст безопасности современного приложения. Подробное описание защищенной среды, в которой работают приложения UWP, приведено в главе 7 тома 1.

Приложения Centennial

Начиная с Windows 10, новая модель приложений стала совместима со стандартными приложениями Win32. Единственное, что требуется от разработчика, — это запустить программу установки приложения с помощью специального инструмента от Microsoft под названием Desktop App Converter. Эта программа запускает установщик в среде-«песочнице» в рамках сервера Silo (внутреннее название Argon Container), где перехватывается весь ввод и вывод относительно файловой системы и реестра, а все файлы помещаются в защищенные папки VFS (виртуализованная файловая система). Подробное описание приложения Desktop App Converter выходит за рамки тематики данной книги.

В отличие от приложений UWP, исполнительная система Centennial не создает для своих процессов «песочницу», а лишь оборачивает их в легкий слой виртуализации. В результате в сравнении со стандартными программами Win32 приложения Centennial не страдают от ослабленной защиты или необходимости запускаться с токеном низкого уровня целостности. Приложение Centennial можно запускать даже от имени администратора. Программы подобного рода работают в рамках капсул для приложений (внутреннее название Helium Container), которые с целью реализовать разделение состояний, не теряя совместимости, предоставляют две формы «изолятора»: перенаправление реестра и виртуальную файловую систему (Virtual File System, VFS). На рис. 8.42 приведен пример приложения Centennial — Kali Linux.

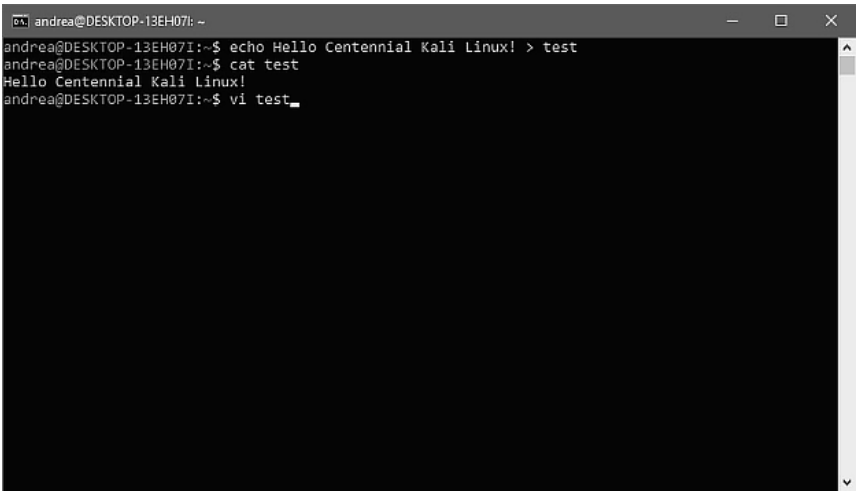


Рис. 8.42. Kali Linux, распространяемая через Windows Store, является типичным примером приложения Centennial

При активации пакета система применяет к приложению перенаправление реестра и выполняет слияние основных системных хранилищ с хранилищами реестра приложения Centennial. При установке на пользовательскую рабочую станцию каждое приложение Centennial может включать в себя три хранилища реестра: `registry.dat`, `user.dat` и (не обязательно) `userclasses.dat`. Файлы реестра, сформированные Desktop Convert, представляют собой неизменяемые хранилища. При запуске приложения исполнительная система Centennial проводит слияние неизменяемых хранилищ с настоящими хранилищами системного реестра (на самом деле Centennial выполняет процедуру детокенизации, поскольку каждое значение в хранилище — это относительное значение).

Возможности по слиянию реестра и виртуализации предоставляются драйвером фильтрации пространства имен виртуального реестра (WscVReg), который интегрирован в ядро NT (диспетчер конфигурации). В момент активации пакета служба пользовательского режима AppInfo взаимодействует с устройством VRegDriver с целью обеспечить слияние и перенаправление действий в реестре от имени приложений Centennial. В рамках данной модели, когда приложение пытается прочесть значение из реестра, присутствующее в виртуализованных хранилищах, операции ввода/вывода на деле перенаправляются в хранилища пакета. Операция записи в такие значения недопустима. Если искомое значение в виртуализованном хранилище отсутствует, оно создается в реальном без всякого перенаправления. Иной способ перенаправления используется для всего корневого раздела `HKEY_CURRENT_USER`. В нем все новые подразделы или значения хранятся *исключительно* в хранилище пакета, которое находится по адресу `C:\ProgramData\Packages\<PackageName>\<UserSid>\SystemAppData\Helium\Cache`.

В табл. 8.34 приведены правила виртуализации реестра для приложений Centennial.

Таблица 8.34. Виртуализация реестра для приложений Centennial

Операция	Результат
Чтение или перечисление в HKEY_LOCAL_MACHINE\Software	Операция возвращает динамическое слияние хранилищ пакета с их локальными системными аналогами. Разделы и значения реестра, существующие в хранилищах пакета, ВСЕГДА имеют приоритет по отношению к разделам и значениям, уже имеющимся в локальной системе
Любые операции записи в HKEY_CURRENT_USER	Перенаправление в виртуализованное хранилище пакета Centennial
Все операции записи в пределах пакета	Если значение реестра присутствует в одном из хранилищ пакета, запись в раздел HKEY_LOCAL_MACHINE\Software не разрешается
Все операции записи за пределами пакета	Если значения реестра еще нет ни в одном из хранилищ пакета, запись в раздел HKEY_LOCAL_MACHINE\Software разрешается

Когда исполнительная система Centennial готовит контейнер приложения Silo, она проходится по всем файлам и каталогам в папке VFS нужного пакета. Эта процедура является частью конфигурации виртуальной файловой системы Centennial, предоставляемой при активации пакета. Также система включает в себя список соответствий для каждой папки в каталоге VFS (табл. 8.35).

Таблица 8.35. Список системных папок, виртуализуемых для приложений Centennial

Имя папки	Адрес перенаправления	Архитектура
SystemX86	C:\Windows\SysWOW64	32 бита/64 бита
System	C:\Windows\System32	32 бита/64 бита
SystemX64	C:\Windows\System32	Только 64 бита
ProgramFilesX86	C:\Program Files (x86)	32 бита/64 бита
ProgramFilesX64	C:\Program Files	Только 64 бита
ProgramFilesCommonX86	C:\Program Files (x86)\Common Files	32 бита/64 бита
ProgramFilesCommonX64	C:\Program Files\Common Files	Только 64 бита
Windows	C:\Windows	Нейтральная
CommonAppData	C:\ProgramData	Нейтральная

Виртуализация файловой системы обеспечивается тремя драйверами, постоянно используемыми средами Argon Container.

- **Драйвер мини-фильтра Windows Bind** (Windows Bind minifilter driver, BindFlt). Управляет перенаправлением для файлов приложения Centennial. Это значит, что, если приложению потребуется прочесть или изменить один из своих уже существующих виртуализованных файлов, ввод/вывод будет перенаправлен в оригинальное местоположение искомого файла. Если же приложение создает еще не существующий файл в одной из виртуализованных папок (например, в `C:\Windows`), операция разрешается (если, конечно, пользователь имеет необходимые привилегии) и перенаправление не применяется.

- **Драйвер мини-фильтра изоляции Windows Container** (Windows Container Isolation minifilter driver, Wcifs). Отвечает за слияние контента различных виртуализованных папок, называемых слоями, и создание уникального представления. Приложения Centennial задействуют этот драйвер для слияния содержимого папки `Application Data` локального пользователя (обычно по адресу `C:\Users\<UserName>\AppData`) с папкой локального кэша приложения по адресу `C:\User\<UserName>\Appdata\Local\Packages\<Package Full Name\LocalCache`. Драйвер способен справиться даже со слиянием для нескольких пакетов, позволяя каждому из них работать со своим персональным представлением файловой системы. Для обеспечения такой возможности он сохраняет ID слоя для каждого пакета в точке `Reparse` искомой папки. Таким образом у него получается сформировать в памяти карту слоев и позволить себе работать в различных частных зонах (внутреннее название — scratch-зоны). На момент написания книги эта продвинутая функция разрешена только для связанного набора, о котором будет рассказано далее в данной главе.
- **Драйвер мини-фильтра виртуализации имен Windows Container** (Windows Container Name Virtualization minifilter driver, Wcnfs). Если драйвер Wcifs выполняет слияние нескольких папок, Wcnfs применяется в Centennial для организации перенаправления имен для локальной пользовательской папки `Application Data`. В отличие от предыдущего случая, когда приложение создает новый файл или папку в виртуализованной папке `Application Data`, результат оказывается в папке кэша приложения, притом всегда в виртуальной, независимо от того, существовал ли он ранее.

Важный момент, о котором не следует забывать, состоит в том, что фильтр BindFlt работает с отдельными файлами, тогда как драйверы Wcnfs и Wcifs работают с папками. Centennial использует порты коммуникации с мини-фильтрами для корректной организации инфраструктуры виртуализованной файловой системы. Этот процесс выполняется с помощью системы коммуникации через сообщения, в рамках которой Centennial отправляет сообщение мини-фильтру и ожидает от него ответа. В табл. 8.36 приведены правила виртуализации файловой системы для приложений Centennial.

Таблица 8.36. Виртуализация файловой системы для приложений Centennial

Операция	Результат
Чтение или просмотр в общеизвестной папке Windows	Операция возвращает динамическое слияние с соответствующей папкой WFS и ее локальным аналогом из системы. Файлы, которые уже находятся в папке VFS, ВСЕГДА имеют приоритет по отношению к имеющимся в локальных папках системы
Запись в папку Application Data	Все операции записи в папку Application Data перенаправляются в локальный кэш приложения Centennial
Любая запись в папки Centennial внутри пакета	Недопустимо, только чтение
Любая запись в папки снаружи пакета	Допустимо, если у пользователя есть разрешение

Диспетчер активности хоста

В Windows 10 были объединены различные компоненты, взаимодействовавшие с состоянием пакетного приложения независимо. В результате совсем новый компонент, известный как диспетчер активности хоста (HAM), стал играть центральную роль в администрировании состояний пакетных приложений, чем может заниматься только он сам, при этом предоставляя унифицированный набор API для всех своих клиентов.

В отличие от предшественников диспетчер активности хоста раскрывает своим клиентам интерфейсы на базе деятельности. Хостом называется объект, обеспечивающий минимальный элемент изоляции, воспринимаемый моделью приложений. Ресурсы, состояния приостановки/возобновления и заморозки управляются как один элемент, что обычно соответствует объекту задания Windows, представляющего пакетное приложение. В этом объекте может содержаться только один процесс для простых приложений, но может оказаться и сразу несколько, если приложение задает по несколько фоновых задач (к примеру, так делают проигрыватели мультимедиа).

В новой модернизированной модели приложений бывает три типа заданий:

- **смешанное** (mixed). Смесь деятельности переднего плана и фоновой, но зачастую связанной именно с передним планом приложения. Когда требуется выполнение фоновых задач, таких как проигрывание музыки или вывод на печать, используется этот тип;
- **чистое** (pure). Применяется сугубо для фоновых задач;
- **системное** (system). Хост исполняет код Windows от лица приложения (например, для фонового скачивания).

Активность всегда принадлежит хосту и представляет собой общий интерфейс для клиентских концепций, таких как окна, фоновые задачи, завершения задач и т. д. Хост считается «активным», если его задание не заморожено и у него есть хотя бы одна запущенная активность. Клиенты HAM — это компоненты, которые взаимодействуют с жизненным циклом активностей и контролируют его. Несколько компонентов являются клиентами HAM: View Manager, Broker Infrastructure, различные компоненты Shell (например, Shell Experience Host), AudioSrv, завершения задач и даже диспетчер служб Windows

Жизненный цикл приложения в такой среде включает в себя четыре состояния: «выполняется», «приостанавливается», «завершение приостановки» и «приостановлено» (состояния и взаимодействия между ними приведены на рис. 8.43).

- **Выполняется** (`Running`). Состояние, при котором в приложении исполняется часть кода, не связанная с приостановкой. Приложение может быть в данном состоянии не только когда оно на переднем плане, но и при выполнении фоновых задач — проигрывании музыки, печати и множестве прочих активностей подобного рода.
- **Приостанавливается** (`Suspending`). Это состояние представляет собой ограниченный по времени период, который наступает, когда HAM просит приложение приостановиться. Для этого могут быть различные причины: приложение теряет

фокус переднего плана, у системы ограничены ресурсы или происходит переход в режим экономии заряда батареи либо просто приложение само ожидает какого-то UI-события. Когда это случается, у приложения есть ограниченное время, чтобы приостановиться (обычно максимум 5 с), иначе оно будет принудительно закрыто.

- **Завершение приостановки** (`SuspendComplete`). Данное состояние указывает, что приложение закончило готовиться к приостановке и сообщает об этом системе. В этом случае процедура приостановки считается завершенной.
- **Приостановлено** (`Suspend`). Как только приложение заканчивает приостановку и сообщает об этом системе, та замораживает его объект задания посредством вызова функции `NtSetInformationJobObject` (через экземпляр информационного класса `JobObjectFreezeInformation`), в результате чего никакой код приложения не может быть подан на исполнение.

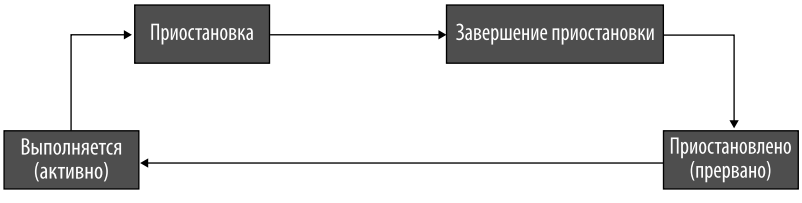


Рис. 8.43. Схема жизненного цикла пакетного приложения

С целью сохранения производительности и ресурсов системы диспетчер активности хоста будет по умолчанию всегда требовать от приложения приостановки. Клиентам HAM необходимо требовать от него сохранять приложение активным. Компонентом, ответственным за активность приложений переднего плана, является диспетчер отображения. То же самое касается и фоновых задач: компонент «инфраструктура брокеров» отвечает за определение того, как процесс, являющийся хостом фоновой задачи, должен оставаться активным (и потребовать у HAM не приостанавливать приложение).

Для пакетных приложений не предусмотрено состояния «завершено» (`Terminated`). Это значит, что у них нет реального представления о переходе на выход или завершение и им не следует пытаться самостоятельно прекращать работу. На деле модель подразумевает, что для завершения пакетного приложения сначала его приостанавливают, а затем HAM, если необходимо, вызывает функцию `NtTerminateJobObject` в отношении задания приложения. HAM автоматически управляет жизненным циклом приложения и уничтожает процесс только по потребности (хорошие примеры — диспетчер отображения или диспетчер активации приложений). Пакетное приложение не способно определить, приостановлено оно или завершено. Это позволяет Windows автоматически восстановить прежнее состояние приложения, даже если его завершили или система перезагрузилась. В итоге модель пакетных приложений оказывается совершенно не такой, как стандартная модель приложений Win32.

Чтобы корректно приостанавливать или возобновлять пакетные приложения, диспетчер активности хоста использует новые функции ядра `PsFreezeProcess` и `PsThawProcess`. Операции над процессом `Freeze` и `Thaw` похожи на приостановку и возобновление, но имеют следующие существенные отличия.

- Новый поток, созданный или внедренный в контексте глубоко замороженного процесса, не запустится, даже если при создании не был установлен флаг `CREATE_SUSPENDED` или запустить поток пытаются с помощью функции `NtResumeProcess`.
- В структурах данных `EPROCESS` был реализован новый счетчик `Freeze`. Это значит, что процесс может быть заморожен многократно. Чтобы процесс возобновил работу, число запросов на разморозку должно быть равно числу запросов на заморозку. Только в этом случае неприостановленные потоки смогут продолжить работу.

Репозиторий состояний

Модернизированная модель приложений представила новый способ хранения настроек, зависимостей между пакетами и прочих данных приложения. Репозиторий состояний является новым централизованным хранилищем, где содержатся самые разные данные. Он реализует важнейшее ключевое правило управления всеми современными приложениями: всякий раз, когда приложение скачивается из магазина, устанавливается, активируется или удаляется, в репозитории считываются или записываются новые данные. Классическим примером использования репозитория состояний является щелчок кнопкой мыши на элементе в меню Пуск. Компонент меню определяет полный путь к файлу активации приложения (который, как уже упоминалось в главе 7 тома 1, может быть типа EXE или DLL), читая данные из репозитория. (На самом деле это упрощение, поскольку процесс `ShellExecutionHost` перечисляет все современные приложения на этапе инициализации.)

Репозиторий состояний по большей части реализован в двух библиотеках: Windows.StateRepository.dll и Windows.StateRepositoryCore.dll. Хотя служба репозитория состояний отвечает за серверную часть компонента, приложения UWP общаются с ним через библиотеку Windows.StateRepositoryClient.dll. (Весь API репозитория пользуется безусловным доверием, из-за клиента WinRT для корректного общения с сервером необходим прокси. Это правило еще одной DLL под названием Windows.StateRepositoryPs.dll.) Корень местоположения репозитория состояний находится в значении реестра HKLM\SOFTWARE\Microsoft\Windows\CurrentVersion\Appx\PackageRepositoryRoot, которое обычно ссылается на путь `C:\ProgramData\Microsoft\Windows\AppRepository`.

Репозиторий состояний реализован поверх нескольких баз данных — разделов. Таблицы в такой базе называются сущностями. Разделы имеют разные ограничения по доступу и жизненному циклу.

- **Машинный** (machine). Эта база данных содержит определения пакетов, данные и идентификации приложения, а также основной и вторичный элементы, применяемые в меню Пуск. Эти данные постоянно считываются различными компонентами, такими как библиотека TileDataRepository, используемая Проводником

и меню Пуск для управления элементами, но запись их в основном выполняется при развертывании AppX (изредка некоторыми другими незначительными компонентами). Машинный раздел обычно хранится в файле `StateRepository-Machine.srd`, расположенном в корневой папке репозитория состояний.

- **Развертывание** (deployment). Хранит общесистемные данные, по большей части используемые только службой развертывания (AppxSvc), когда новый пакет регистрируется в системе или удаляется из нее. Это касается списков файлов приложений и копий файла `manifest` для каждого приложения. Раздел развертывания обычно хранится в файле `StateRepository-Deployment.srd`.

Все разделы хранятся в базах данных формата SQLite. В Windows имеется особая сборка SQLite, скомпилированная в библиотеку StateRepository.Core.dll. Последняя публикует API слоя доступа к данным репозитория состояний (также известный как DAL), которые в основном являются оболочкой внутреннего ядра СУБД и вызываются службой репозитория состояний.

Иногда различным компонентам требуется знать, когда какие-то отдельно взятые данные в репозитории состояний записываются или изменяются. В годовом обновлении Windows 10 в репозиторий состояний была добавлена поддержка отслеживания изменений и событий. Она позволяет разбираться со следующими ситуациями.

- Компоненту требуется подписаться на изменение данных в какой-то конкретной сущности. Компонент получает отклик, когда данные изменены и зафиксированы SQL-транзакцией. В операцию развертывания может входить по несколько SQL-транзакций. В конце каждой из них репозиторий состояний проверяет, закончена ли операция развертывания, и, если это так, обращается к каждому из зарегистрированных слушателей.
- Процесс запускается или возобновляется из состояния приостановки, и ему требуется узнать, изменились ли данные после последнего оповещения или обращения. Репозиторий состояний может удовлетворить такой запрос, используя поле `ChangeId`, которое в таблицах, где эта функция поддерживается, является уникальным хронологическим идентификатором записи.
- Процесс запрашивает данные из репозитория состояний, и ему требуется знать, менялись ли они после последней проверки. Изменения данных всегда записываются в совместимых сущностях с помощью таблицы `Changelog`. Последняя всегда записывает время, ID изменения из события, создавшего данные, и, если применимо, ID изменения из события, удалившего данные.

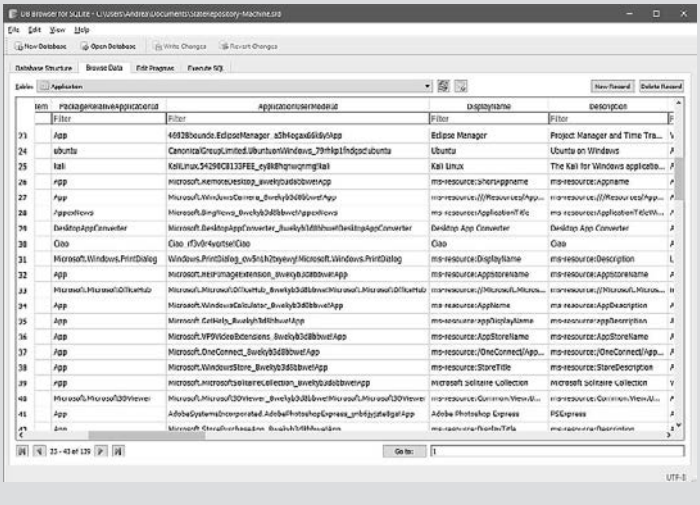
Современное меню Пуск для корректной работы использует возможность отслеживания изменений и событий репозитория состояний. Когда запускается процесс `ShellExperienceHost`, он обращается к репозиторию состояний, чтобы тот оповещал его контроллер (NotificationController.dll) каждый раз, когда элемент меню изменяется, создается или удаляется. Когда пользователь устанавливает или удаляет современное приложение через магазин, сервер развертывания приложений выполняет в базе данных транзакцию для добавления или удаления элемента. Под конец транзакции репозиторий состояний сигналит событие, которое активизирует контроллер. Таким образом меню Пуск способно менять свой облик почти в реальном времени.

ПРИМЕЧАНИЕ Подобным образом современное меню Пуск способно автоматически добавлять или удалять вхождения для каждого вновь установленного стандартного приложения Win32. Программа установки приложения обычно создает один или несколько ярлыков в одном из местоположений папки классического меню Пуск (общесистемный путь: C:\ProgramData\Microsoft\Windows\Start Menu, пользовательский путь: C:\Users\<UserName>\AppData\Roaming\Microsoft\Windows\Start Menu). Современное меню Пуск пользуется библиотекой AppResolver, чтобы подписаться на все файловые события папок меню Пуск (с помощью функции ReadDirectoryChangesW для Win32). Таким образом, всякий раз, когда в отслеживаемых папках появляется новый ярлык, библиотека может получить обратный вызов и дать меню Пуск сигнал перерисоваться.

ЭКСПЕРИМЕНТ. Наблюдение за репозиторием состояний

Вы можете довольно просто открыть каждый из разделов репозитория состояний с помощью любого приятного вам просмотрщика SQLite. Для данного эксперимента вам потребуется скачать какой-нибудь из них вроде DB Browser (с открытым кодом) для SQLite, который можно получить с сайта `http://sqlitebrowser.org/`. Путь к репозиторию состояний недоступен стандартным пользователям. Более того, файл любого из разделов может оказаться занятым как раз в тот момент, когда вы пытаетесь получить к нему доступ. Так что потребуется скопировать файл с базой данных, прежде чем пытаться открыть его обозревателем SQLite. Откройте командную строку администратора (для этого наберите в поисковой строке Cortana команду `cmd` и, щелкнув правой кнопкой мыши на значке командной строки, выберите Запустить от имени администратора (Run As Administrator)) и вставьте следующие команды:

```
C:\WINDOWS\system32>cd "C:\ProgramData\Microsoft\Windows\AppRepository"
C:\ProgramData\Microsoft\Windows\AppRepository>copy StateRepository-Machine.srd
"%USERPROFILE%\Documents"
```

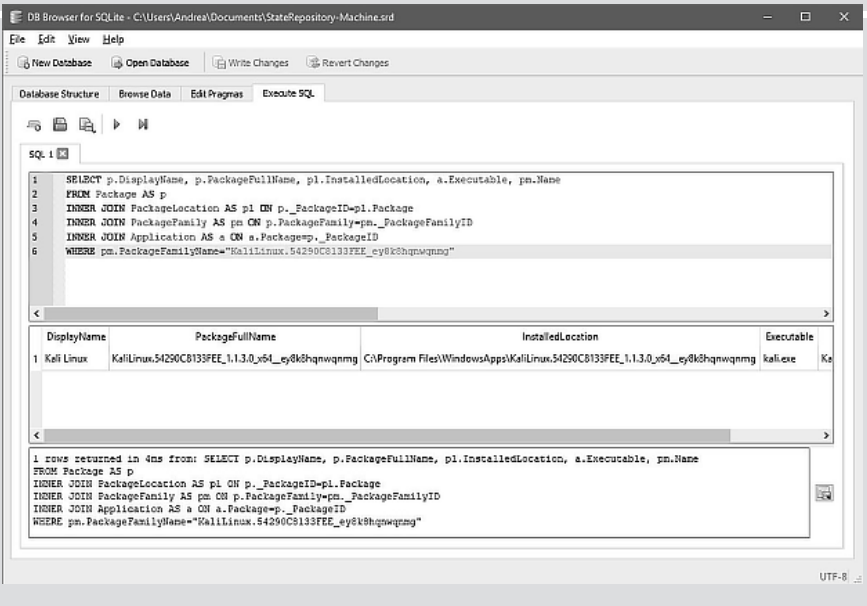


Так вы скопировали машинный раздел репозитория состояний к себе в папку Документы. Следующим шагом будет его открытие. Запустите DB Browser for SQLite, используя ссылку, появившуюся в меню Пуск или в поисковой строке Cortana, и нажмите кнопку Открыть базу данных (Open Database). Перейдите в папку Документы, в раскрывающемся списке Тип файла (File Type) выберите Все файлы (*) (All Files (*)) (базы данных репозитория состояний не пользуются стандартным расширением файлов SQLite) и откройте скопированный файл `StateRepository-machine.srd`. Основным режимом DB Browser for SQLite является просмотр структуры базы. Для данного эксперимента вам потребуется перейти на вкладку Обзор данных (Browse Data) и пройти через таблицы `Package`, `Application`, `PackageLocation` и `PrimaryTile`.

Диспетчер активации приложений и многие другие компоненты модернизированной модели приложений применяют стандартные SQL-запросы для извлечения из репозитория состояний необходимых данных. Например, для получения местонахождения пакета и имени исполняемого файла современного приложения может быть использован SQL-запрос, подобный следующему:

```
SELECT p.DisplayName, p.PackageFullName, pl.InstalledLocation, a.Executable, pm.Name
FROM Package AS p
INNER JOIN PackageLocation AS pl ON p._PackageID=pl.Package
INNER JOIN PackageFamily AS pm ON p.PackageFamily=pm._PackageFamilyID
INNER JOIN Application AS a ON a.Package=p._PackageID
WHERE pm.PackageFamilyName="<Package Family Name>"
```

DAL (слой доступа к данным) задействует похожие запросы для предоставления данных своим клиентам.



Вы можете зафиксировать общее количество записей в таблице, а затем установить из магазина новое приложение. Если после завершения процесса развертывания вы снова скопируете файл базы данных, то обнаружите, что число записей изменилось! Это происходит в нескольких таблицах. А если новое приложение добавляет новый элемент, то даже в таблице `PrimaryTile` появляется запись о новом элементе для показа в меню Пуск.

Мини-репозиторий зависимостей

Открытие базы данных SQLite и извлечение оттуда нужной информации посредством SQL-запроса может быть довольно дорогой операцией. Кроме того, текущая архитектура требует некоторого объема общения между процессами средствами RPC. Эти два ограничения порой чересчур строги, чтобы пытаться им соответствовать. Классическим примером может служить ситуация, когда пользователь запускает новое приложение (или псевдоним выполнения) через консоль командной строки. Проверка репозитория состояний каждый раз, когда система создает процесс, вызывает серьезные проблемы с производительностью. Чтобы их решить, в модель приложений было добавлено еще одно хранилище поменьше для хранения информации о современных приложениях — мини-репозиторий зависимостей (Dependency Mini Repository, DMR).

В отличие от репозитория состояний мини-репозиторий зависимостей никаких баз данных не использует, а хранит свои данные в проприетарном бинарном формате Microsoft, который можно прочесть через любую файловую систему в любом контексте безопасности (даже драйвер режима ядра мог бы поразбирать данные DMR). Каталог метаданных системы, представленный папкой `Packages`, которая находится в корневом каталоге репозитория состояний, содержит список подкаталогов — по одному для каждого установленного пакета. Мини-репозиторий зависимостей представлен файлом с расширением `.pckgdep` и именем по SID пользователя. Файл DMR создается службой развертывания, когда у пользователя регистрируется пакет (подробности см. в подразделе «Регистрация пакетов» далее).

Мини-репозиторий зависимостей часто применяется, когда система создает процесс, принадлежащий пакетному приложению (в расширении AppX PreCreateProcess). Поэтому он полностью реализован в рамках Win32-библиотеки kernelbase.dll (с несколькими функциями-заглушками в kernel.appcore.dll). Когда при создании процесса открывается файл DMR, он прочитывается, анализируется и отображается в памяти процесса-родителя. Когда дочерний процесс создан, кодзагрузчик отражает эти данные и в его память. Файл DMR содержит различную информацию, в том числе:

- информацию о пакете — ID, полное имя, полный путь и издателя;
- информацию о приложении — ID пользовательской модели приложения и относительный ID, описание, отображаемое имя и графические логотипы;
- контекст безопасности — SID и возможности среды AppContainer;
- целевую платформу и граф зависимостей пакета (применяется в случаях, когда пакет зависит от одного или нескольких других пакетов).

Файл DMR спроектирован с перспективой добавления дополнительных данных в будущих версиях Windows (если потребуется). Использование мини-репозитория зависимостей делает создание процесса довольно быстрым и не требует запросов к репозиторию состояний. Стоит заметить, что после того, как процесс создан, файл DMR закрывается. Это позволяет переписать файл `.pckgdep` для установки нового пакета даже во время работы современного приложения. Таким образом, пользователь может добавить функции в свое приложение, даже не перезапуская его. Некоторые маленькие фрагменты (чаще всего полное имя пакета и путь к нему) дублируются в различных разделах реестра в роли кэша для ускорения доступа. Такой кэш часто применяется для простейших операций, например, чтобы проверить, существует ли пакет.

Фоновые задачи и инфраструктура брокеров

Обычно приложениям UWP требуется способ запускать часть своего кода в фоновом режиме. Такому коду нет необходимости взаимодействовать с основным процессом на переднем плане. UWP поддерживает фоновые задачи, которые позволяют обеспечить приложению функциональность, доступную, даже когда основной процесс приостановлен или не выполняется. Существует масса причин, почему программе могут потребоваться фоновые задачи: общение в реальном времени, почта, IM, музыка и мультимедиа, видеопроигрыватель и т. д. Фоновая задача может быть привязана с помощью триггеров и условий. Триггером называется глобальное системное асинхронное событие, которое при срабатывании сигнализирует о запуске фоновой задачи. В этот момент запуск может и не произойти, что зависит от существующих условий. Например, фоновая задача в рамках некоего приложения для IM может быть запущена, только когда пользователь авторизовался (триггер системного события) и только если есть доступ в Интернет (условие).

В Windows 10 существует два вида фоновых задач:

- **фоновая задача внутри процесса** (in-process background task). Код приложения и его фоновой задачи выполняются в рамках одного процесса. С точки зрения разработчика, такое проще реализовать, но возникает серьезная проблема: если в коде окажется баг, произойдет отказ приложения. Фоновые задачи внутри процесса поддерживают меньше триггеров, чем задачи вне процесса;
- **фоновая задача вне процесса** (out-of-process background task). Код приложения и его фоновой задачи выполняются в разных процессах (процесс мог бы выполняться еще и отдельно в объекте задания). Этот тип задач более устойчив, действует в рамках хоста процесса backgroundtaskhost.exe и может использовать любые триггеры и условия. Никакой баг, имеющийся в фоне, никогда не воздействует на приложение целиком. Основной же недостаток обусловлен затратами производительности на код RPC, которые необходимы для коммуникации между двумя разными процессорами.

Чтобы обеспечить пользователю максимальный комфорт, все фоновые задачи имеют лимит времени на выполнение, в целом равный 30 с. Через 25 с служба инфраструктуры брокеров фонового режима обращается к дескриптору отмены фоновой задачи — в WinRT это называется событием `OnCanceled`. Когда это событие

происходит, у фоновой задачи есть всего 5 с на то, чтобы полностью очиститься и выйти. В ином случае процесс, содержащий код фоновых задач (в случае задач вне процесса это мог бы быть BackgroundTaskHost.exe, в ином случае — процесс самого приложения), может оказаться завершенным. Разработчики персональных и бизнес-приложений UWP могут не придерживаться этого ограничения, однако тогда их работы нельзя будет опубликовать в официальном Microsoft Store.

Фоновая инфраструктура брокеров (BI) — это центральный компонент сферы управления всеми фоновыми задачами. В основном он реализуется в bisrv.dll (сторона сервера), которая запускается в составе службы инфраструктуры брокеров. Возможностями фоновой инфраструктуры брокеров могут воспользоваться два типа клиентов: стандартные приложения и службы Win32 могут импортировать клиентскую библиотеку bi.dll, приложения для WinRT всегда подключают библиотеку biwinrt.dll, которая предоставляет API для современных приложений. Фоновая инфраструктура брокеров не может существовать без брокеров. Брокерами называются компоненты, генерирующие событие, которые потом принимаются фоновым сервером брокеров. Существует несколько их разновидностей. Наиболее важными считаются:

- **брокер системных событий** (System Event Broker). Предоставляет триггеры для таких системных событий, как изменение статуса подключения к сети, вход и выход пользователей, изменение состояния системной батареи и т. д.;
- **брокер времени** (Time Broker). Обеспечивает поддержку циклического или однократного таймера;
- **брокер подключения к сети** (Network Connection Broker). Обеспечивает UWP-возможность получать события для успешной установки соединений по определенным портам;
- **брокер устройств** (Device Services Broker). Поддерживает триггеры появления устройств, когда пользователь подключает или извлекает устройство. Работает, прослушивая события `Pnp`, приходящие из ядра;
- **брокер широкополосной мобильной связи** (Mobile Broad Band Experience Broker). Предоставляет все критически важные триггеры для работы с телефонами и SIM-картами.

Серверная часть брокера представляет собой службу Windows. Все брокеры реализованы по-разному. Большинство используют для работы подписку на статусы WNF (см. раздел «Средство оповещений Windows» ранее в данной главе), публикуемые ядром Windows; прочие, такие как брокер времени, выстроены на основе стандартных функций Win32. Особенности реализации каждого типа брокеров выходят за рамки тематики данной книги. Брокер может попросту пересылать события, сгенерированные где-то еще (например, ядром Windows), либо создавать их самостоятельно, основываясь на каких-то других условиях и статусах. Брокеры пересылают сообщения, которые они обработали с помощью WNF: каждый из них создает именованное состояние WNF, на которое подписывается фоновая инфраструктура. Таким образом, когда брокер публикует новые данные о состоянии, последняя, будучи подписчиком, просыпается и пересылает событие своим клиентам.

В состав всех видов брокеров входит и клиентская инфраструктура — библиотеки для WinRT и Win32. Фоновая инфраструктура брокеров и они сами предоставляют своим клиентам три вида функций:

- **недоверительные** (non-trust APIs). Обычно используются компонентами WinRT, работающими в AppContainer или в среде-«песочнице». Проводятся дополнительные проверки безопасности. Те, кто вызывает такие функции, а именно `BiRtCreateEventForApp`, не вправе указывать имя постороннего пакета или действовать от лица другого пользователя;
- **полудоверительные** (partial-trust APIs). Применяются компонентами Win32, работающими в среде Medium-IL. Те, кто вызывает такие функции, а именно `BiPtCreateEventForApp`, могут указать полное имя пакета современного приложения, но не вправе действовать от лица другого пользователя;
- **доверительные** (full-trust API). Применяются только высокопривилегированными системными и административными службами Win32. Те, кто вызывает такие функции, а именно `BiCreateEventForPackageName`, вправе действовать от лица другого пользователя с любым именем пакета.

Клиенты брокеров вправе решать, подписаться им напрямую на событие, выдаваемое конкретным брокером, или на фоновую инфраструктуру брокеров. В WinRT всегда предпочитали второй вариант. На рис. 8.44 приведен пример инициализации таймера-триггера для задач современного приложения.

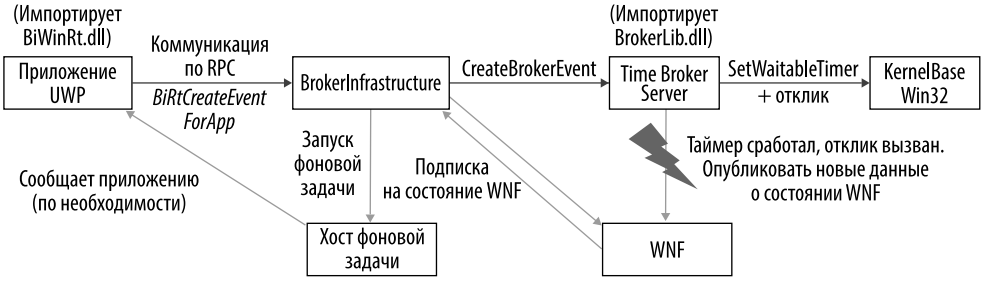


Рис. 8.44. Архитектура брокера времени

Еще одна важная возможность, которую фоновая инфраструктура предоставляет своим брокерам, — сохранение фоновых заданий. Это значит, что, когда пользователь перезагрузит или включит/выключит систему, все зарегистрированные фоновые задачи восстанавливаются и перепланируются так, как было до перезагрузки, чтобы достичь этого без ошибок, когда система загружается и диспетчер контроля служб (подробнее о диспетчере контроля служб будет рассказано в главе 10) запускает службу фоновой инфраструктуры, последняя, как часть своей инициализации, размещает в памяти GUID корневого хранилища закрытую копию хранилища реестра фонового брокера и, используя платформенно-зависимую функцию `NtLoadKeyEx`, загружает ее. Служба сообщает ядру NT о потребности загрузить закрытую копию хранилища с помощью специального флага `REG_APP_HIVE`. Хранилище BI находится в файле `C:\Windows\System32\Config\BBI`. Корневой раздел хранилища монтируется как `\Registry\A\<Root Storage GUID>` и доступен только процессу службы

инфраструктуры брокеров (в данном случае это svchost.exe, инфраструктура брокеров работает в общем хосте для служб). Хранилище инфраструктуры хранит список событий и рабочих записей, которые сортируются и идентифицируются с помощью кодов GUID.

- **Событие, представляющее триггер для фоновой задачи.** Относится к ID брокера (представляет брокера, дающего событию тип), полным именем пакета, пользователем приложения UWP, с которым тот связан, и некоторыми другими параметрами.
- **Рабочая запись представляет запланированную фоновую задачу.** Содержит имя, список условий, точку входа в задачу и GUID связанного с задачей события триггера.

Служба BI перечисляет каждый подраздел. При этом очищаются осиротевшие события (те, что не связаны с рабочими записями). Затем наконец публикуется наименование состояния для WNF. Таким образом, все брокеры могут проснуться и завершить свою инициализацию.

Фоновая инфраструктура брокеров играет важнейшую роль в работе приложений UWP. Даже обычные приложения и сервисы Win32 могут воспользоваться BI и брокерами с помощью соответствующих клиентских библиотек. В число заметных примеров входят служба планировщика задач, интеллектуальная фоновая служба передачи, служба Push-уведомлений Windows и AppReadiness.

Установка и запуск пакетных приложений

Жизненный цикл пакетных приложений отличается от такового у стандартных приложений Win32. В мире последних процедура установки может варьироваться от простого копирования исполняемого файла до комплексных процедур настройки. Даже если запуск приложения сводится к запуску исполняемого файла, загрузчик Windows возьмет на себя всю работу. Установка современного приложения, в свою очередь, сводится к четко определенной процедуре, которая в основном происходит с участием Windows Store. В режиме разработчика администратор даже может установить программу из внешнего файла AppX, хотя последний и необходимо заверить цифровой подписью. Процедура регистрации пакета сложна и затрагивает множество компонентов.

Прежде чем углубляться в тему регистрации пакетов, важно разобраться в другом ключевом аспекте современных приложений — активации пакетов. Это процесс запуска приложения, которое может отобразить пользовательский интерфейс или не может сделать этого. Для различных типов приложений активация проходит по-разному и затрагивает ряд системных компонентов.

Активация пакетов

Пользователь не сможет запустить приложение UWP, просто запустив его исполняемый файл (не считая случаев для недавно введенных псевдонимов AppExecution, созданных специально для этого. Их описание приведено далее в данной главе). Чтобы корректно активировать современное приложение, пользователю следует щелкнуть на элементе в меню Пуск, задействовать специальный файл-ссылку,

который Проводник сможет проанализировать, или же обратиться к иным точкам активации (например, дважды щелкнуть на связанном с приложением документе, пройти по специальному URL и т. д.). Какой вариант активации использовать в зависимости от типа приложения, решает процесс ShellExperienceHost.

Приложения UWP

Основным компонентом, управляющим активацией, является диспетчер активации, реализованный в ActivationManager.dll. Он работает под крылом службы sihost.exe, поскольку ему требуется взаимодействовать с Рабочим столом. Диспетчер активации тесно связан с диспетчером отображения. Меню Пуск обращается к диспетчеру активации посредством RPC. Последний начинает процедуру активации, схема которой приведена на рис. 8.45.

1. Получить SID пользователя, запросившего активацию, ID семейства пакетов и PRAID пакета. Таким образом можно проверить, зарегистрирован ли пакет в системе (с помощью мини-фильтра репозитория зависимостей и его кэша из реестра).
2. Если проверка установила, что пакет необходимо зарегистрировать, вызывается клиент развертывания AppX и начинается регистрация. Иногда пакет не нуждается в регистрации в силу действия политики «регистрация по требованию», а именно, когда приложение скачалось, но не было до конца установлено (это экономит время, особенно в условиях предприятия) или когда требуется выполнить обновление. Диспетчер активации опознает один из таких случаев при помощи репозитория состояний.
3. Приложение регистрируется в HAM, создается HAM-хост для нового пакета и его первоначальной деятельности.
4. Диспетчер активации обращается к диспетчеру отображения (через RPC) с целью инициализировать активацию пользовательского интерфейса для нового сеанса (диспетчер отображения должен быть оповещен даже при фоновой активации).
5. Активация продолжается в службе DcomLaunch, поскольку на данном этапе диспетчер активации задействует класс WinRT для запуска низкоуровневого создания процесса.
6. Служба DcomLaunch отвечает за запуск серверов COM, DCOM и WinRT в ответ на запрос об активации объекта. Она реализована в библиотеке rpcss.dll. DcomLaunch перехватывает запрос на активацию и готовится вызвать Win32-функцию `CreateProcessAsUser`. Прежде чем это сделать, службе потребуется установить для процесса корректные атрибуты (например, полное имя пакета), определить, что у пользователя есть актуальная лицензия на запуск приложения, продублировать токен пользователя, заменить низкое значение уровня целостности новым и отметить процесс нужными атрибутами безопасности. (Надо заметить, что служба DcomLaunch выполняется от лица системной учетной записи, которая имеет привилегии на TCB. Их наличие необходимо для подобных манипуляций с токенами. Подробности см. в главе 7 тома 1.) Затем DcomLaunch вызывает `CreateProcessAsUser`, передавая на вход полное имя пакета с помощью одного из атрибутов процесса. Это приведет к созданию приостановленного процесса.

7. Далее активация продолжается в KernelBase.dll. Токен, созданный DcomLaunch, все еще не является AppContainer, но уже содержит атрибуты безопасности UWP. Особый код из функции `CreateProcessInternal` использует реестровый кэш мини-репозитория зависимостей для сбора следующей информации о пакетном приложении: корневой каталог, статус пакета, SID пакета AppContainer и список возможностей приложения. Затем проверяется целостность лицензии (очень популярная функция у игр). К этому моменту файл мини-репозитория зависимостей уже отражен в память родительского процесса, а у приложения UWP определен альтернативный путь загрузки DLL.
8. Токен AppContainer, пространство имен его объекта и символические ссылки создаются функцией `BasepCreateLowBox`, которая выполняет в пользовательском режиме бóльшую часть работы, не считая собственно создания токена AppContainer, за что отвечает функция ядра `NtCreateLowBoxToken`. Тема токенов AppContainer уже рассматривалась в главе 7 тома 1.
9. Объект процесса уровня ядра, как обычно, создается с помощью функции ядра `NtCreateUserProcess`.
10. После того как подсистема CSRSS уведомлена, функция `BasepPostSuccessAppXExtension` отражает мини-репозиторий зависимостей в PEB дочернего процесса и убирает это отражение из памяти родительского процесса. Новый процесс наконец можно запустить, возобновив его основной поток.

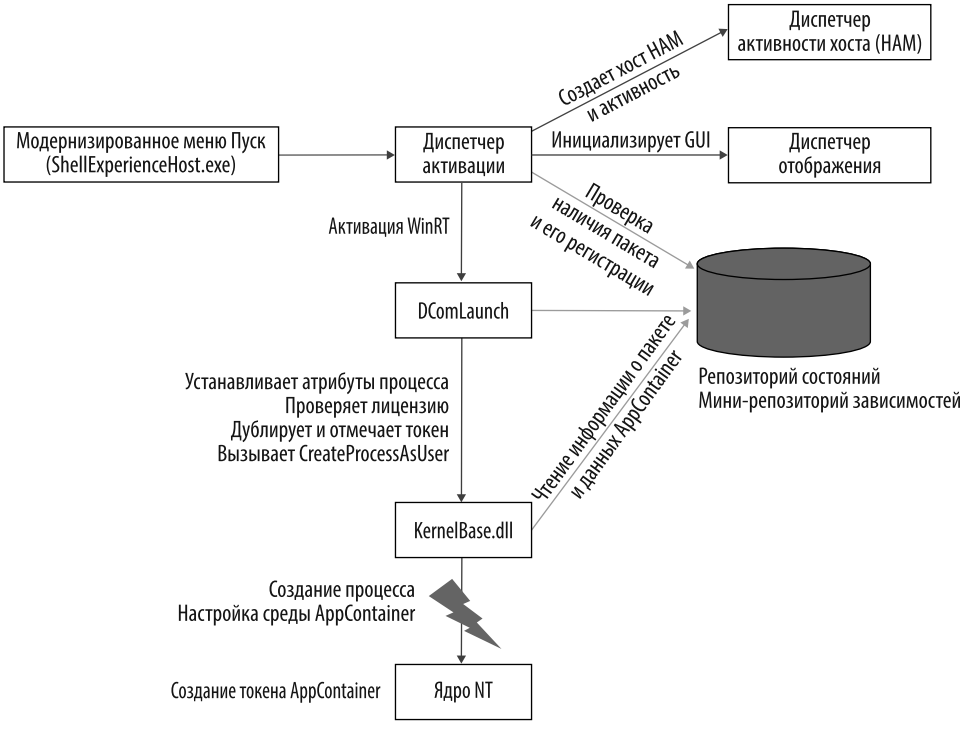


Рис. 8.45. Схема активации современного приложения UWP

Приложения Centennial

Процесс активации приложений Centennial подобен таковому для приложений UWP, но реализован совершенно иначе. При этом модернизированное меню, ShellExperienceHost, всегда обращается к Explorer.exe (Проводник). В активации Centennial участвует множество библиотек, отраженных в Проводнике, например Daxexec.dll, Twinui.dll и Windows.Storage.dll. Когда Проводник получает запрос на активацию, он получает полное имя пакета и ID приложения, после чего через RPC подхватывает путь к основному исполняемому файлу и свойства пакета из репозитория состояний. Затем выполняются те же шаги, что для UWP (шаги со 2-го по 4-й). Основное отличие заключается в том, что вместо службы DComLaunch активация Centennial на данном этапе запускает процесс с помощью функции `ShellExecute` из библиотеки Shell32. Код последней был обновлен так, чтобы узнавать приложение Centennial и использовать особую процедуру из Windows.Storage.dll (через COM). Эта библиотека применяет RPC для вызова функции `RAiLaunchProcessWithIdentity` из службы AppInfo. Последняя задействует репозиторий состояний для проверки лицензии на приложение и целостности его файлов, после чего обращается к токену процесса. Затем токен помечается необходимыми атрибутами безопасности и наконец создает процесс в приостановленном состоянии. AppInfo передает полное имя пакета в функцию `CreateProcessAsUser` посредством атрибута процесса `PROC_THREAD_ATTRIBUTE_PACKAGE_FULL_NAME`.

В отличие от активации UWP, никакого AppContainer не создается, а AppInfo вызывает функцию `PostCreateProcessDesktopAppXActivation` из DaxExec.dll для инициализации виртуализационного слоя приложений Centennial (реестр и файловая система). Подробнее об этом рассказывалось в подразделе «Приложения Centennial» ранее в главе.

ЭКСПЕРИМЕНТ. Активация современных приложений через командную строку

Проведя данный эксперимент, вы лучше поймете различия между UWP и Centennial и узнаете, какова мотивация выбора функции ShellExecute для активации последних. Вам потребуется установить хотя бы одно приложение Centennial. На момент написания книги простой метод выяснить, что такой тип приложения существует, сводится к обращению в Windows Store. В магазине после выбора искомого предложения надо прокрутить страницу вниз до раздела Дополнительная информация (Additional Information). Если увидите фразу: «Приложению разрешается: использовать все системные ресурсы», которая обычно расположена под табличкой Поддерживаемые языки (Supported languages), — значит, перед вами приложение Centennial.

Для данного эксперимента мы воспользуемся Notepad++. Найдите в Windows Store приложение (unofficial) Notepad++ и установите его. Затем откройте

приложения Camera и Notepad++. Откройте командную строку администратора (можно сделать это, набрав команду `cmd` в поисковой строке Cortana, правой кнопкой щелкните на значке командной строки и выберите Запустить от имени администратора (Run As Administrator)). Вам потребуется найти полные пути к обоим запущенным пакетным приложениям с помощью следующих команд:

```
wmic process where "name='WindowsCamera.exe'" get ExecutablePath
wmic process where "name='notepad++.exe'" get ExecutablePath
```

Теперь вы сможете создать пару ссылок на исполняемые файлы этих приложений с помощью команд:

```
mklink "%USERPROFILE%\Desktop\notepad.exe" "<Notepad++ executable Full Path>"
mklink "%USERPROFILE%\Desktop\camera.exe" "<WindowsCamera executable full path>"
```

Замените то, что находится между < и >, реальными путями к исполняемым файлам, которые вы нашли с помощью двух предыдущих команд.

Теперь можно закрыть командную строку и оба приложения. На вашем Рабочем столе должны появиться две новые ссылки. В отличие от ссылки на Notepad.exe, если вы попытаетесь запустить приложение Camera с Рабочего стола, активация не сработает, а Windows ответит примерно таким диалоговым окном.

(«Windows не удалось получить доступ к указанному устройству, пути или файлу. Возможно, у вас нет разрешений, достаточных для доступа к данному элементу».)

Это происходит потому, что Проводник использует для активации ссылок на исполняемые файлы библиотеку Shell32. В случае с UWP она не имеет понятия, какое приложение сейчас запустит данный образ, поэтому обращается к функции `CreateProcessAsUser`, не указав идентификации пакета. В ином случае Shell32 способна опознать приложение Centennial, тогда, как в данном случае, процесс активации проходит до конца, а приложение корректно запускается. Если вы попробуете запустить обе ссылки через командную строку, ни одна не запустит приложение как надо. Это объясняется тем, что командная строка не использует Shell32. Вместо этого она применяет функцию `CreateProcess` напрямую из собственного кода. Таковы различия в активации разных типов пакетных приложений.

ПРИМЕЧАНИЕ Начиная с обновления Windows 10 Creators Update (RS2), модернизированная модель приложений поддерживает концепцию опциональных пакетов (внутреннее название RelatedSet). Такие пакеты широко применяются в играх, где основная часть поддерживает добавление DLC (или дополнения), и в пакетах, представляющих собой комплекты (хороший пример — Microsoft Office). Пользователь может загрузить и установить Word и вместе с ним неявно получить пакет с фреймворком, который содержит весь общий код Office. Если пользователю потребуется дополнительно загрузить Excel, операция развертывания сможет пропустить загрузку основного фреймворка, так как Word уже является его опциональным пакетом.

Опциональные пакеты связаны с главными через свои файлы manifest. Такой файл содержит декларацию о зависимости от главного пакета (с помощью AMUID). Подробное описание принципов работы опциональных пакетов выходит за рамки тематики данной книги.

Псевдонимы AppExecution

Как уже говорилось, пакетные приложения нельзя активировать напрямую через их исполняемые файлы. Это накладывает значительное ограничение, особенно на новые современные консольные приложения. Начиная с обновления Windows 10 Fall Creators Update (build 1709), модернизированная модель приложений получила поддержку псевдонимов AppExecution, чтобы иметь возможность запускать современные приложения (как Centennial, так и UWP) из командной строки. Теперь пользователь может запускать Edge и любые другие современные приложения прямо из консоли командной строки. Псевдоним AppExecution в простейшем понимании представляет собой исполняемый файл длиной 0 байт, расположенный по адресу `C:\Users\<UserName>\AppData\Local\Microsoft\WindowsApps` (рис. 8.46). Данное местоположение добавлено в системный список путей поиска исполняемых файлов (через переменную окружения PATH). В итоге, чтобы запустить современное предложение, пользователь может указать имя любого исполняемого файла из этой папки, не указывая полного пути (как в диалоговом окне Выполнить (Run) или в командной строке консоли).

Как же система исполнит файл в 0 байт? Ответ кроется в малоизвестной функции файловой системы — точке повторного анализа. Обычно эти функции используются для создания символических ссылок, однако могут хранить не только информацию о них, но и в целом любые данные. Модифицированная модель приложений задействует эту функцию для сохранения данных об активации пакетных приложений (имя семейства пакетов, ID пользовательской модели приложения, путь к приложению) напрямую в точке повторного анализа.

Когда пользователь запускает исполняемый файл псевдонима AppExecution, то, как и раньше, задействуется функция `CreateProcess`. Системный вызов `NtCreateUserProcess`, используемый для управления созданием процесса в режиме ядра (подробности см. в разделе «Порядок работы CreateProcess» главы 3 тома 1), не срабатывает, так как файл пуст. Файловая система, придерживаясь штатного порядка создания процесса, открывает искомый файл

(с помощью функции `IoCreateFileEx`), наталкивается на данные точки повторного анализа (в ходе анализа последнего узла в пути) и возвращает вызвавшему код `STATUS_REPARSE`. `NtCreateUserProcess` интерпретирует этот код в ошибку `STATUS_IO_REPARSE_TAG_NOT_HANDLED` и завершает работу. Функция `CreateProcess` теперь знает, что создать процесс не удалось из-за некорректной точки повторного анализа, поэтому загружает библиотеку ApiSetHost.AppExecutionAlias.dll и обращается к ней. Та содержит код, который может интерпретировать содержимое точек повторного анализа для современных приложений.

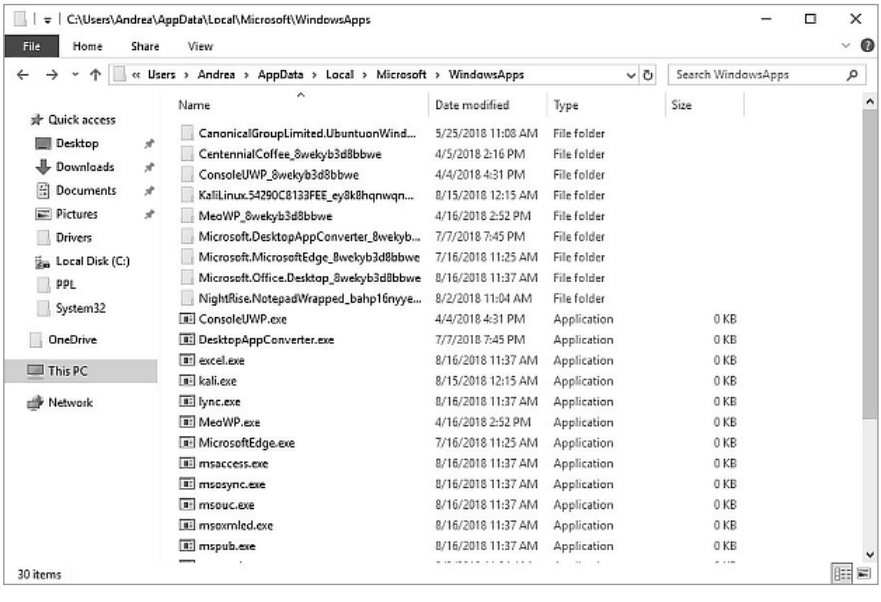


Рис. 8.46. Основная папка с псевдонимами AppExecution

Код из этой библиотеки разбирает содержимое точки повторного анализа, извлекает оттуда данные для активации пакетного приложения и обращается к службе AppInfo с целью корректно пометить токен нужными атрибутами безопасности. AppInfo проверяет наличие у пользователя корректной лицензии на запуск пакетного приложения и целостность его файлов (с помощью репозитория состояний). Сам процесс создается вызвавшим процессом. Функция `CreateProcess` определяет ошибку повторного анализа и перезапускается с правильным путем к исполняемому файлу пакета (обычно находится по адресу `C:\Program Files\WindowsApps\`). На этот раз она корректно создает процесс и токен AppContainer для него либо в случае приложений Centennial инициализирует слой виртуализации (на самом деле вновь происходит обращение к AppInfo по RPC). Кроме того, там же создаются хост HAM и его действия, необходимые приложению. В этот момент активация закончена.

ЭКСПЕРИМЕНТ. Чтение данных псевдонима AppExecution

В данном эксперименте вы извлечете данные псевдонима AppExecution из исполняемого файла размером 0 байт. Можете воспользоваться утилитой FsReparser (входит в состав загружаемых инструментов к данной книге) для проверки как расширенных атрибутов файловой системы NTFS, так и точек повторного анализа. Запустите утилиту в окне командной строки, указав параметр `READ`:

```
C:\Users\Andrea\AppData\Local\Microsoft\WindowsApps>fsreparser read MicrosoftEdge.exe

File System Reparse Point / Extended Attributes Parser 0.1
Copyright 2018 by Andrea Allievi (AaLl86)

Reading UWP attributes...
Source file: MicrosoftEdge.exe.

The source file does not contain any Extended Attributes.

The file contains a valid UWP Reparse point (version 3).
Package family name: Microsoft.MicrosoftEdge_8wekyb3d8bbwe
Application User Model Id: Microsoft.MicrosoftEdge_8wekyb3d8bbwe!MicrosoftEdge
UWP App Target full path: C:\Windows\System32\SystemUWPLauncher.exe
Alias Type: UWP Single Instance
```

Как видно из выведенного утилитой текста, функция `CreateProcess` может извлечь всю информацию, нужную ей для корректной активации современного приложения. Это объясняет, почему можно запускать Edge из командной строки.

Регистрация пакетов

Когда пользователю требуется установить современное приложение, обычно открывают AppStore, находят нужный продукт и нажимают кнопку Get. Начинается скачивание архива, в котором содержится ряд файлов: файл `manifest` для пакета, цифровая подпись приложения и блочная карта, обозначающая цепочку доверенных сертификатов, задействованных в цифровой подписи. Изначально архив помещается в папку, находящуюся по адресу `C:\Windows\SoftwareDistribution\Download`. Процесс AppStore (WinStore.App.exe) взаимодействует со службой Центра обновления Windows (wuaueng.dll), которая работает с запросами на скачивание.

Скачанные файлы являются манифестами, где содержатся список всех файлов современного приложения, зависимости приложения, данные о лицензии и шаги, необходимые для корректной регистрации пакета. Служба Центра обновления Windows определяет, что запрос на скачивание поступил от современного приложения, проверяет токен обратившегося процесса (это должен быть AppContainer) и, пользуясь функциями библиотеки AppXDeploymentClient.dll, выясняет, не был ли искомый пакет установлен ранее. Затем она создает запрос на развертывание для AppX и посредством RPC отправляет его на сервер развертывания AppX. Последний действует на правах службы PPL в рамках процесса общего хоста для служб,

под чьим крылом работает также служба клиентских лицензий, пользуясь тем же уровнем защиты. Запрос на развертывание помещается в очередь, находящуюся под асинхронным управлением. Когда сервер развертывания AppX замечает запрос, то выносит его из очереди и порождает поток, который и займется развертыванием современного приложения.

ПРИМЕЧАНИЕ Начиная с Windows 8.1, средства развертывания UWP стали поддерживать концепцию комплектов. Под ними подразумеваются пакеты, где может быть по несколько ресурсов, таких как разные языки или функции, спроектированные для использования только в отдельно взятых регионах. Средства развертывания реализуют логику применимости, по которой скачивается лишь нужная часть сжатого комплекта, исходя из профиля пользователя и системных настроек.

Процесс развертывания современного приложения предусматривает сложную цепочку событий. Мы вкратце изложим здесь процесс развертывания в виде трех основных фаз.

Фаза 1. Размещение пакета

После того как Windows Update закончит скачивание манифеста приложения, сервер развертывания AppX проверяет, все ли зависимости удовлетворены, подходят ли системные требования, такие как целевое семейство устройств (телефон, настольный компьютер, консоль XBox и т. д.), и годится ли файловая система тома, в который будет производиться установка. Все требования, необходимые приложению, изложены в файле `manifest` со всеми зависимостями. Если все проверки пройдены, процедура размещения создает корневой каталог пакета (обычно по адресу `C:\Program Files\WindowsApps\<PackageFullName>`) и его подкаталоги. Кроме того, каталоги пакета будут защищены посредством установки на каждый из них соответствующих ACL. Если современное приложение относится к типу Centennial, оно загружает библиотеку daxexec.dll и создает точки повторного анализа VFS, необходимые драйверу мини-фильтра изоляции Windows Container (см. подраздел «Приложения Centennial» ранее в главе), с целью корректной виртуализации папки с данными приложения. Наконец, путь к корневому каталогу пакета сохраняется в реестре в разделе HKLM\SOFTWARE\Classes\LocalSettings\Software\Microsoft\Windows\CurrentVersion\AppModel\PackageRepository\Packages\<PackageFullName>, в параметре `Path`.

Процедура размещения заранее выделяет на диске место для файлов приложения, рассчитывает окончательный объем данных для скачивания и извлекает URL сервера, где размещены все файлы пакетов (в сжатом виде с расширением AppX). Наконец, с удаленного сервера загружается последний файл AppX, опять же с помощью службы Центра обновления Windows.

Фаза 2. Размещение данных пользователя

Эта фаза что-то значит только тогда, когда пользователь обновляет приложение. В ходе нее пользовательские данные просто берутся из старого пакета и переносятся по новому пути к приложению.

Фаза 3. Регистрация пакета

Самая важная фаза развертывания — это регистрация пакета. На этом сложном этапе используются возможности библиотеки AppXDeploymentExtensions.onecore.dll (и AppXDeploymentExtensions.desktop.dll для развертывания частей, специфичных для Рабочего стола). Мы называем это установкой ядра пакета. На данном этапе основной задачей сервера развертывания AppX остается обновление репозитория состояний. Он создает новые записи для пакета, для одного или нескольких приложений в его составе, новые элементы, разрешения для пакета, лицензию на приложение и т. д. Чтобы это сделать, сервер развертывания AppX использует транзакции в базе данных, которые фиксирует, только если не возникло никаких ошибок (в противном случае все они будут отменены). Когда все транзакции в базах данных для проведения развертывания в репозитории состояний будут зафиксированы, репозиторий состояний сможет обратиться к зарегистрированным слушателям, чтобы передать оповещение каждому запросившему его клиенту. (Более подробно о возможности отслеживания изменений и событий в репозитории состояний рассказано в подразделе «Репозиторий состояний» ранее в главе.)

Последними шагами при регистрации пакета будут создание файла минирепозитория зависимостей и обновление реестра системы для отражения изменений в репозитории состояний. На этом процесс развертывания завершается. Новое приложение готово к активации и запуску.

ПРИМЕЧАНИЕ Чтобы материал было легче читать, процесс развертывания был значительно упрощен. Например, на стадии размещения мы пропустили несколько предварительных фаз, таких как фаза индексирования, которая анализирует файл manifest; фаза диспетчера зависимостей, на которой составляется рабочий план и анализируются зависимости пакета; фаза проверки занятости пакета, целью которой является взаимодействие с PLM для выяснения того, что пакет еще не установлен в системе и не занят.

Кроме того, при провале какой-либо операции средства развертывания должны быть способны откатить все изменения. Иные фазы отката в разделе не упоминались.

ЗАКЛЮЧЕНИЕ

В данной главе мы рассмотрели ключевые базовые механизмы, на которых построена исполнительная система Windows. В следующей главе познакомим вас с технологиями виртуализации, поддерживаемыми Windows для повышения уровня безопасности в системе в целом и обеспечения производительной среды исполнения для виртуальных машин, изолированных контейнеров и защищенных анклавов.

ГЛАВА 9

Технологии виртуализации

Одной из важнейших технологий, применяемых для запуска нескольких операционных систем на одном физическом компьютере, является виртуализация. На момент написания данной книги существует множество типов развивающихся технологий виртуализации от различных производителей оборудования. Технологии виртуализации используются не только для запуска нескольких операционных систем на одном компьютере — они стали основой для таких важных средств защиты, как Virtual Secure Mode (VSM) и Hypervisor-Enforced Code Integrity (HVCI), которые не могут работать без гипервизора.

В этой главе вы познакомитесь с технологическим решением Windows по виртуализации — системой Hyper-V. Она состоит из гипервизора, компонента, управляющего зависимым от платформы оборудованием для виртуализации, и стека виртуализации. Мы разберем внутреннюю архитектуру Hyper-V и кратко опишем его составляющие: диспетчер памяти, виртуальные процессоры, перехваты, планировщик и т. д. Стек виртуализации строится поверх гипервизора и предоставляет различные службы для корневого и гостевых разделов. Мы рассмотрим все его компоненты (процесс VM Worker, службу управления виртуальными машинами, драйвер VID, шина VMBus и т. д.) и ряд поддерживаемых аппаратных средств эмуляции.

В заключительной части главы поговорим о технологиях, основанных на виртуализации, таких как VSM и HVCI. Вам будут представлены все защищенные сервисы, которые эти технологии позволяют обеспечить в системе.

ГИПЕРВИЗОР WINDOWS

Гипервизор Hyper-V, также известный как гипервизор Windows, — это гипервизор первого типа (низкоуровневый, или «голое железо»): операционная система в миниатюре, работающая непосредственно на оборудовании хоста и управляющая одной корневой и одной или несколькими гостевыми операционными системами. В отличие от гипервизоров второго типа (хостовых), которые работают на базе традиционной ОС, подобно обычному приложению, гипервизор Windows абстрагируется от корневой ОС, которая знает о его существовании и взаимодействует с ним для обеспечения выполнения одной или нескольких гостевых виртуальных машин. Поскольку гипервизор является частью операционной системы, управление гостями внутри нее, равно как и взаимодействие с ними, полностью интегрированы

в нее посредством стандартных механизмов управления, таких как WMI и службы. В таком случае корневая ОС содержит несколько компонентов паравиртуализации (в терминологии Microsoft enlightenment — букв. «просветление». — *Примеч. пер.*). *Компонентами паравиртуализации* называются особые средства оптимизации в ядре, а иногда и в драйверах устройств, которые могут обнаружить, что исполняющийся код виртуализован гипервизором, из-за чего некоторые задачи они выполняют иначе и более эффективно в зависимости от среды.

На рис. 9.1 приведена базовая архитектура стека виртуализации Windows, подробное описание которой приводится далее.

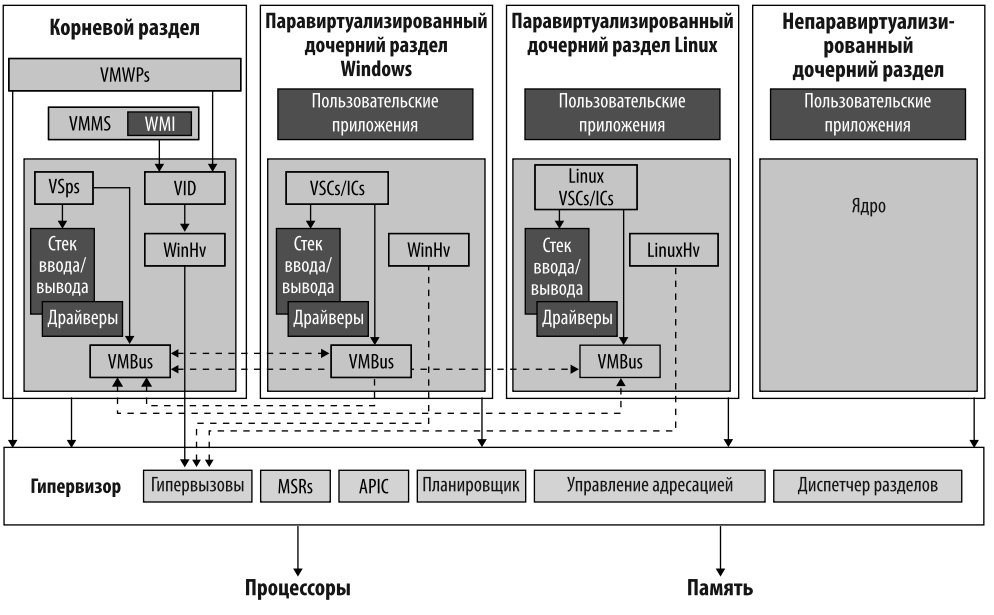


Рис. 9.1. Архитектурный стек Hyper-V (гипервизор и стек виртуализации)

На нижнем уровне архитектуры расположен гипервизор, который запускается очень рано, прямо перед началом загрузки ОС, и предоставляет свои сервисы в пользование стеку виртуализации посредством обращения к интерфейсу гипервызовов. Ранняя инициализация гипервизора будет описана в главе 12. Запуск гипервизора инициируется загрузчиком Windows, от которого зависит, запускать ли гипервизор и безопасное ядро. Если запущены и гипервизор, и безопасное ядро, гипервизор использует сервисы HvLoader.dll, чтобы корректно определить аппаратную платформу для загрузки и запуска своей подходящей версии. Поскольку в процессорах от Intel и AMD (и ARM64) имеются различные элементы реализации аппаратной поддержки виртуализации, то и гипервизоры существуют разные. Правильный выбирается в момент начала загрузки, когда процессор уже обработал очередь инструкций CPUID. В системах Intel загружается образ Hvix64.exe, в системах AMD используется образ Hvax64.exe. Начиная с обновления Windows от 10 мая 2019 г. (19H1), версия Windows под ARM64 стала поддерживать собственный гипервизор, который реализован в образе Hvaa64.exe.

На высоком уровне гипервизор использует расширение для аппаратной виртуализации, которое является тонким слоем между ядром ОС и процессором. Этот слой, который перехватывает и безопасно эмулирует критичные операции, исполняемые ОС, пользуется более высоким уровнем привилегий, чем само ядро ОС. (Intel называют этот режим VMXROOT. Большинство учебников и книг определяют зону безопасности VMXROOT как кольцо 1.) Когда операция, исполняемая нижестоящей ОС, перехватывается, процессор прекращает исполнять код ОС и передает управление гипервизору на повышенном уровне привилегий. Данную операцию обычно называют событием VMEXIT. Аналогично, когда гипервизор закончил обрабатывать перехваченную операцию, ему нужно как-то позволить физическому процессору возобновить исполнение кода операционной системы. Аппаратное расширение виртуализации добавляет новые команды, что позволяет случиться событию VMEnter — процессор возобновляет исполнение кода ОС на прежнем своем уровне привилегий.

Разделы, процессы и потоки

Одним из ключевых компонентов, стоящих за гипервизором Windows, является раздел. Это, в сущности, главная единица изоляции, экземпляр установленной операционной системы, который по традиции называют либо хостом, либо гостем. В рамках модели гипервизора Windows эти два термина не используются, вместо этого мы говорим либо о корневом разделе, либо о дочернем разделе соответственно. Раздел состоит из некоторого объема физической памяти и одного или нескольких виртуальных процессоров (VP) со своими локальными виртуальными APIC и таймерами (в глобальном смысле раздел включает в себя также виртуальную материнскую плату и несколько виртуальных периферийных устройств. Все это — понятия стека виртуализации, которые не имеют отношения к гипервизору).

Как минимум система с Hyper-V имеет корневой раздел (тот, в котором работает операционная система, контролирующая компьютер), стек виртуализации и связанные с ним компоненты. Каждая операционная система, работающая в виртуализованной среде, представляет собой дочерний раздел, который может содержать некоторые дополнительные инструменты для оптимизации доступа к оборудованию или обеспечения возможности управлять этой ОС. Разделы организованы иерархически. Корневой раздел контролирует каждый из дочерних разделов и иногда получает оповещения (перехваты) об определенных видах происходящих там событий. Большинство обращений к физическому оборудованию, происходящих в корневом разделе, гипервизором пропускаются, таким образом, родительский раздел может общаться с устройствами напрямую (за некоторыми исключениями). Со своей стороны дочерние разделы обычно не вправе напрямую взаимодействовать с оборудованием физического компьютера (опять же за некоторыми исключениями, о чем более подробно говорится в разделе «Стек виртуализации» далее в данной главе). Все операции ввода/вывода перехватываются гипервизором и при необходимости перенаправляются в корневой раздел.

Одной из главных целей проектирования гипервизора было сделать его настолько маленьким и модульным, насколько возможно, почти как микроядро, — не нужно поддерживать никакой *драйвер гипервизора* или создавать монолитный, полнофункциональный модуль. Это значит, что большая часть работы по виртуализации

на деле выполняется отдельным стеком виртуализации (см. рис. 9.1). Гипервизор пользуется уже существующей архитектурой драйверов Windows и общается с настоящими драйверами устройств. Подобный подход образует несколько компонентов, которые обеспечивают и регулируют это поведение и коллективно носят имя *стека виртуализации*. Хотя сам гипервизор считывается с загрузочного диска и подается загрузчиком Windows на выполнение еще до начала существования корневой ОС (и корневого раздела), именно корневой раздел ответствен за предоставление всего этого стека. Поскольку все это компоненты Microsoft, только машина с Windows может находиться в корневом разделе. ОС Windows из корневого раздела отвечает за драйверы устройств в системе, равно как и за весь стек виртуализации. Она же является пунктом управления всеми дочерними разделами. Основные компоненты, обеспечиваемые корневым разделом, приведены на рис. 9.2.

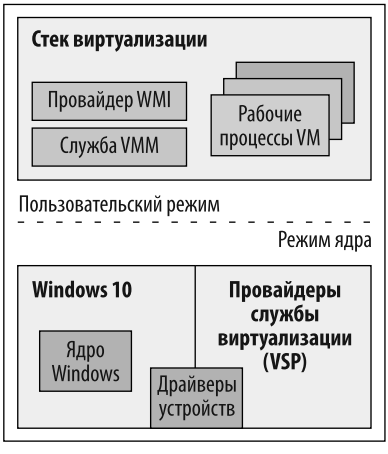


Рис. 9.2. Компоненты корневого раздела

Дочерние разделы

Дочерний раздел — это экземпляр любой операционной системы, исполняющийся одновременно с корневым разделом. (Поскольку вы можете сохранить или приостановить состояние любого дочернего раздела, возможно, что и не исполняющийся.) В отличие от корневого раздела, у которого есть полный доступ к APIC, портам ввода/вывода и физической памяти, кроме той, что принадлежит гипервизору и безопасному ядру, дочерние разделы из соображений безопасности и контроля ограничены собственным представлением адресного пространства (гостевая физическая адресация, или GPA, — пространство, управляемое гипервизором) и не имеют прямого доступа к оборудованию (тем не менее способны напрямую обращаться к некоторым видам устройств, подробности см. в разделе «Стек виртуализации»). С точки зрения гипервизора дочерний раздел также в основном ограничен в оповещениях и изменениях состояния. К примеру, он не может контролировать другие разделы и создавать новые.

В дочерних разделах намного меньше компонентов виртуализации, так как они не отвечают за управление стеком виртуализации — только за взаимодействие с ним. Кроме того, эти компоненты можно считать необязательными, поскольку они улучшают производительность среды, но не критичны для ее работы. На рис. 9.3 приведены компоненты, присутствующие в типичном дочернем разделе с Windows.

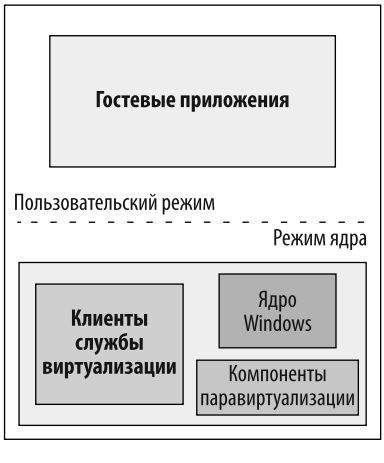


Рис. 9.3. Компоненты дочернего раздела

Процессы и потоки

Гипервизор Windows представляет собой виртуальную машину со структурой данных на основе разделов. Раздел, как сказано ранее, состоит из некоторого объема памяти и одного или нескольких виртуальных процессоров (VP). Внутри же гипервизора каждый виртуальный процессор — это диспетчеризуемый объект, а сам гипервизор, как и стандартное ядро NT, оснащен планировщиком. Планировщик распределяет исполнение виртуальных процессоров, принадлежащих разным разделам, по всем физическим процессорам. (Мы обсудим различные типы планировщиков для гипервизора в подразделе «Планировщики Hyper-V» далее.) Поток гипервизора (структура данных `TH_THREAD`) служит связующим элементом между виртуальным процессором и его объектом планирования. На рис. 9.4 приводится структура данных, отражающая текущий контекст реального исполнения. В ней содержатся стек исполнения потока, данные о планировании, указатель на виртуальный процессор потока, точка входа в цикл диспетчеризации потока (подробности далее) и, что важнее всего, указатель на процесс гипервизора, которому данный поток принадлежит.

Для каждого создаваемого виртуального процессора гипервизор создает структуру данных виртуального процессора (`VM_VP`) и привязывает ее к новообразованному потоку.

В рамках гипервизора процесс (структура данных `TH_PROCESS`), показанный на рис. 9.5, представляет раздел и является контейнером для своего физического (и виртуального) адресного пространства. В эту структуру входят список потоков, поддерживаемых виртуальными процессорами, данные о планировании (задание процессу соответствия физическим процессам, на которых ему позволено работать) и указатель на базовые структуры данных о памяти раздела (секция памяти, зарезервированные страницы, корень каталога страниц и т. д.). Обычно процесс создается, когда гипервизор оформляет раздел (структура данных `VM_PARTITION`), который будет представлять новую виртуальную машину.

Рис. 9.4. Структура данных потока гипервизора

Рис. 9.5. Структура данных процесса в гипервизоре

Компоненты паравиртуализации

Компоненты паравиртуализации, — это одно из ключевых средств оптимизации, которыми пользуются средства виртуализации в Windows. Они являются прямыми модификациями стандартного ядра Windows, способными определять нахождение

операционной системы в дочернем разделе и принимать решение выполнять свою работу иначе. Зачастую такие средства оптимизации сильно зависят от конкретного оборудования и сводятся к гипервызову с целью оповещения гипервизора.

Примером является оповещение гипервизора о длительном цикле спин-ожидания освобождения. Гипервизор может запоминать какое-то состояние спин-ожидания и принимать решение запланировать другой VP (virtual processor), пока ожидание не закончилось. Вход в состояние прерывания, выход из него и доступ к APIC можно координировать гипервизором, который может быть паравиртуализирован во избежание захвата реального доступа, а затем виртуализирован.

Другой пример относится к управлению памятью, в частности, к очистке буфера быстрого преобразования адресов (Translation Look-aside Buffer, TLB) (подробности см. в главе 5 тома 1). Обычно, когда операционная система исполняет инструкцию процессора очистить одну или несколько устаревших записей TLB, это касается только одного процессора. В многопроцессорных системах запись в TLB зачастую нужно очищать в кэше каждого активного процессора (для этого система отправляет межпроцессорное прерывание каждому активному процессору). Но поскольку один дочерний раздел может делить физические процессоры с другими дочерними разделами, часть из которых может на момент начала очистки TLB обслуживать процессор другой виртуальной машины, данная операция может привести к очистке TLB и для них. Более того, виртуальный процессор будет перепланирован на исполнение только очистки TLB, что приведет к заметному снижению производительности. Если же Windows действует в рамках гипервизора, вместо этого тот получает гипервызов для очистки только определенных данных отдельного дочернего раздела.

Привилегии и свойства раздела, особенности версий

В момент создания раздела (как правило, драйвером VID) к нему еще не привязано ни одного виртуального процессора (VP). На этой фазе драйвер VID может свободно добавлять или удалять привилегии раздела. Гипервизор, в свою очередь, при создании раздела выдает ему некоторый набор привилегий по умолчанию в зависимости от его типа.

Привилегия раздела отражает, какое действие — обычно представленное гипервызовами или синтетическими, зависящими от модели регистрами (Model Specific Registers, MSR) — паравиртуализированная ОС, действующая в рамках раздела, вправе исполнить от лица этого самого раздела. Например, привилегия «доступ к корневому планировщику» позволяет дочернему разделу оповестить корневой о том, что было просигнализировано о событии и гостевой VP может быть перепланирован, что обычно повышает приоритет потока гостевого VP. Привилегия «доступ к VSM», в свою очередь, позволяет разделу разблокировать VTL 1 и доступ к его свойствам и конфигурации, обычно предоставляемым через синтетические регистры. В табл. 9.1 приведены все привилегии, назначаемые гипервизором по умолчанию.

Привилегии раздела могут быть заданы только перед тем, как тот создаст или запустит какие-либо VP: гипервизор будет отказывать в запросах на установку привилегий, если хотя бы один VP в разделе уже запущен. Свойства раздела похожи на привилегии, но подобным ограничением не обладают: они могут быть установлены или запрошены в любой момент. Существуют различные группы свойств, которые могут быть запрошены или установлены для раздела (табл. 9.2).

Таблица 9.1. Привилегии разделов

Тип раздела	Привилегии по умолчанию
Корневой и дочерний разделы	Чтение/запись счетчика времени исполнения VP. Чтение эталонного времени текущего раздела. Доступ к таймерам и регистрам SynIC. Запрос/установка для VP вспомогательной страницы виртуального APIC. Чтение/запись MSR гипервызова. Запрос записи VP IDLE. Отображение или сокрытие зоны кода гипервызова. Чтение эмулированного TSC (счетчик временных меток) и его частоты у VP. Контроль TSC раздела и эмуляция повторной паравиртуализации. Чтение/запись синтетических регистров VSM. Чтение/запись VTL-зависимых регистров PV. Запуск виртуального процессора AP. Разрешение поддержки разделом быстрых гипервызовов
Только корневой раздел	Создание дочернего раздела. Поиск и обращение к разделу по ID. Запрос/отдача памяти в распоряжение раздела. Отправка сообщений в порт подключения. Сигнализация события в разделе порта подключения. Создание, удаление или считывание свойств порта подключения раздела. Соединение/разъединение с портом подключения раздела. Отображение/сокрытие страницы статистики гипервизора, где описываются VP, LP, раздел или гипервизор. Планирование VP дочернего раздела и доступ к синтетическим MSR SyncIC. Вызов перезагрузки паравиртуализированной системы. Чтение настроек отладчика гипервизора для раздела
Только дочерний раздел	Генерирование расширенного перехвата гипервызова в корневом разделе Оповещение потока VP корневого планировщика о сигнале события
Раздел EXO	Нет

Таблица 9.2. Свойства разделов

Группа свойств	Описание
Свойства планирования	Устанавливают/запрашивают свойства классического планировщика и планировщика ядер, как то: предел, вес и резерв
Свойства времени	Позволяют разделу быть приостановленным/возобновленным
Свойства отладки	Изменяют конфигурацию отладчика гипервизора в ходе выполнения
Свойства ресурсов	Запрашивают специфичные для платформы свойства виртуального оборудования для раздела, такие как размер TLB, поддержка SGX и т. д.
Свойства совместимости	Запрашивают специфичные для платформы свойства виртуального оборудования, связанные с изначальными функциями совместимости

Когда раздел создан, инфраструктура сообщает гипервизору уровень совместимости, указанный в конфигурационном файле виртуальной машины. В зависимости от него гипервизор разрешает или запрещает конкретные функции оборудования, которые могут быть предоставлены VP его операционной системе. Существует множество функций для настройки поведения VP в зависимости от уровня совместимости виртуальной машины. Хорошим примером может служить аппаратная таблица атрибутов страниц (Page Attribute Table, PAT), которая представляет настраиваемый способ кэширования виртуальной памяти. До выхода Windows Anniversary Update 10 (RS1) гостевые виртуальные машины (VM) не могли использовать PAT, так что, даже если уровень совместимости VM указывает Windows 10 RS1, гипервизор не позволит гостевой ОС под ним увидеть регистры PAT. Но если уровень совместимости выше, чем Windows 10 RS1, гипервизор предоставит поддержку PAT ОС из гостевой VM. Когда корневой раздел исходно создан во время загрузки, гипервизор активирует наивысший уровень совместимости. Таким образом, корневая ОС получает возможность пользоваться всеми функциями, поддерживаемыми физическими устройствами.

Запуск гипервизора

В главе 12 мы проанализируем метод, которым загружается рабочая станция на основе UEFI, и все компоненты, задействованные в загрузке и запуске исполняемого образа гипервизора. В данном подразделе вкратце обсудим, что происходит на компьютере после того, как модуль HvLoader переключил исполнение на гипервизор, передав ему управление в первый раз.

HvLoader загружает корректную версию двоичного образа гипервизора (она зависит от производителя процессора) и создает блок загрузки гипервизора. Тот захватывает минимальный контекст процессора, который нужен гипервизору для запуска первого виртуального процессора. Затем HvLoader переключается на новое, только что созданное адресное пространство и передает управление в образ гипервизора посредством обращения к его точке входа `KiSystemStartup`, которая готовит процессор к запуску гипервизора и инициализирует структуру данных `CPU_PLS`. Последняя отражает физический процессор и выполняет роль структуры PRCB из ядра NT, гипервизор, в свою очередь, получает быстрый доступ к ней (через сегмент GS). В отличие от ядра NT, `KiSystemStartup` вызывается только для процессора-загрузчика (алгоритм запуска процессоров для приложений рассматривается в подразделе «Запуск процессоров приложений» далее в этой главе), в итоге оставляя реальную инициализацию другой функции, `BmpInitBootProcessor`.

Функция `BmpInitBootProcessor` запускает сложную последовательность инициализации. Она оценивает систему и запрашивает все функции виртуализации, поддерживаемые процессором (EPT и VPID; запрашиваемые возможности зависят от платформы и различаются для версий гипервизора для Intel, AMD или ARM). Затем функция определяет тип планировщика, который будет управлять тем, как гипервизор планирует работу виртуальных процессоров. Для серверных систем на основе Intel и AMD по умолчанию используется планировщик ядер, в то время как корневой планировщик по умолчанию действует во всех клиентских системах, в том

числе ARM64. Тип планировщика может быть переопределен вручную с помощью опции BCD `hypervisorschedulertype` (более подробно разные планировщики гипервизора описываются далее в данной главе).

Затем инициализируются вложенные компоненты паравиртуализации. Они позволяют гипервизору исполняться во вложенных конфигурациях, где корневой гипервизор (L0) управляет реальным оборудованием, а другой гипервизор (L1) исполняется на виртуальной машине. После этого этапа функция `BmpInitBootProcessor` выполняет инициализацию следующих компонентов.

- Диспетчера памяти (инициализирует базу данных PFN и корневое пространство).
- Слоя аппартных абстракций гипервизора (HAL).
- Процесса и подсистемы потоков гипервизора, которая зависит от выбранного типа планировщика. Создаются системный процесс и его первоначальный поток. Это особенный процесс, он не привязан ни к каким разделам или потокам хоста, где исполняется код гипервизора.
- Слоя абстракции виртуализации VMX (VAL). Назначение VAL заключается в абстрагировании различий между всеми поддерживаемыми аппаратными расширениями виртуализации (Intel, AMD и ARM64). В него входит код, работающий со специфичными для платформы возможностями технологии виртуализации на машине, используемой гипервизором (в частности, на платформе Intel слой VAL управляет поддержкой «неограниченного гостя», EPT, SGX, MBEC и т. д.).
- Синтетического контроллера прерываний (Interrupt Controller, SynIC) и модуля управления вводом/выводом памяти (I/O Memory Management Unit, IOMMU).
- Диспетчера адресации (Address Manager, AM) — компонента, отвечающего за управление физической памятью, назначенной разделу (известный как гостевая физическая память или GPA), и преобразование в общую физическую память (известную как системная физическая память). Хотя первая реализация Hyper-V поддерживала теневые таблицы страниц (программная техника преобразования адресов), начиная с Windows 8.1, диспетчер адресации использует зависящий от платформы код для настройки преобразования адресов для гипервизора, предоставляемый оборудованием (расширенные таблицы страниц для Intel, вложенные таблицы страниц для AMD). В контексте гипервизора это физическое адресное пространство называется *адресным доменом*. Такое не зависящее от платформы преобразование физического адресного пространства обычно называется вложенным пейджингом (Second Layer Address Translation, SLAT). Под этим подразумеваются EPT у Intel, NPT у AMD или механизм двухэтапного преобразования адресов ARM.

Теперь гипервизор может продолжить оформление структуры данных `CPU_PLS`, связанной с процессором-загрузчиком, размещая первоначальные, зависящие от оборудования, контрольные структуры виртуальной машины (VMCS для Intel, VMCB для AMD) и активируя виртуализацию с помощью первой операции VMXON. Наконец, происходит инициализация структур данных для управления прерываниями в разрезе процессоров.

ЭКСПЕРИМЕНТ. Подключение отладчика гипервизора

В данном эксперименте вы подключите отладчик гипервизора для анализа последовательности запуска гипервизора, описанной в предыдущем разделе. Отладчик гипервизора поддерживается только через сетевое или последовательное соединение. Для отладки гипервизора могут быть использованы только физические компьютеры либо виртуальные машины, на которых активна функция вложенной виртуализации (см. подраздел «Вложенная виртуализация» далее). В последнем случае для гипервизора L1 возможна только отладка по последовательному подключению.

Для данного эксперимента вам потребуется отдельный физический компьютер, на котором поддерживаются расширения виртуализации и установлена и активна роль Hyper-V. Вы будете использовать ее в качестве отлаживаемой системы, привязанной к системе-хосту (она играет роль отладчика), где будут запущены инструменты отладки. В качестве альтернативы можете настроить вложенную VM, как показано в эксперименте «Активация вложенной виртуализации на Hyper-V» далее в данной главе (в таком случае отдельная физическая машина не потребуется).

В качестве первого шага нужно скачать и установить в системе-хосте пакет Debugging Tools for Windows, который имеется в составе Windows SDK (или WDK) и может быть загружен с сайта https://developer.microsoft.com/en-us/windows/downloads/windows-10-sdk. В качестве альтернативы можете воспользоваться утилитой WinDbgX, которую на момент написания книги можно получить в Windows Store по запросу WinDbg Preview.

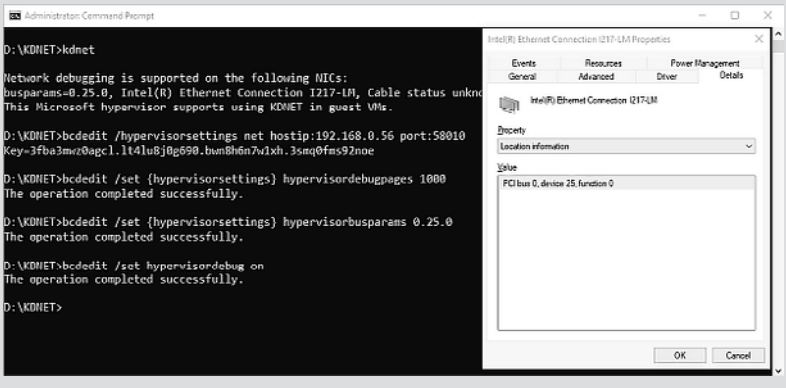
Данный эксперимент требует, чтобы в отлаживаемой системе была отключена функция `Secure Boot`. Отладка гипервизора с этой технологией несовместима. По поводу отключения `Secure Boot` обратитесь к руководству пользователя для вашей рабочей станции (обычно настройки `Secure Boot` можно найти в UEFI Bios). Для активации отладчика гипервизора в отлаживаемой системе вам потребуется открыть командную строку администратора, набрав в поисковой строке Cortana слово `cmd` и выбрав Запустить с правами администратора (Run As Administrator).

Если вы собираетесь отлаживать гипервизор через сетевую карту, наберите приведенные далее команды, заменив вхождения `<HostIp>` IP-адресом системы-хоста, `<HostPort>` — соответствующим номером порта на хосте (начиная с 49152) и `<NetCardBusParams>` — параметрами сетевой карты отлаживаемой системы, указанными в формате XX.YY.ZZ, где XX — номер шины, YY — номер устройства, а ZZ — номер функции:

```
bcdedit /hypervisorsettings net hostip:<HostIp> port:<HostPort>
bcdedit /set {hypervisorsettings} hypervisordebugpages 1000
bcdedit /set {hypervisorsettings} hypervisorbusparams <NetCardBusParams>
bcdedit /set hypervisordebug on
```

Узнать параметры шины можно через диспетчер устройств или с помощью утилиты KDNet.exe, доступной в Windows SDK.

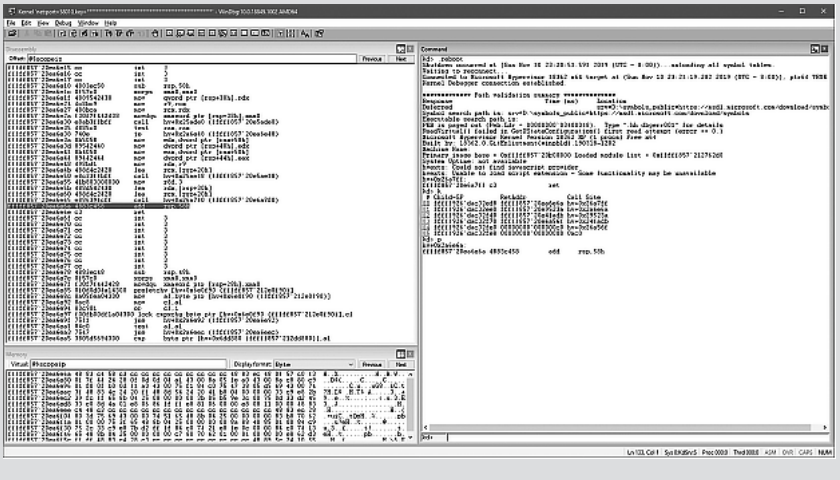
На рисунке далее приводится пример системы, где сетевой интерфейс, используемый для отладки гипервизора, имеет параметры шины 0.25.0, а отладчик нацелен на систему-хост, настроенную под IP-адрес 192.168.0.56 и порт 58010.



Обратите внимание на возвращенный ключ отладки. После перезагрузки отлаживаемой системы нужно будет запустить на хосте WinDbg следующей командой:

```
windbg.exe -d -k net:port=<HostPort>,key=<DebuggingKey>
```

Вы сможете отлаживать гипервизор, отслеживая ход его запуска, даже несмотря на то, что Microsoft может и не публиковать отладочные символы для главного модуля гипервизора.



В случае VM с активной вложенной виртуализацией вы сможете включить отладчик гипервизора L1 только через последовательный порт, выполнив в отлаживаемой системе следующую команду:

```
bcdedit /hypervisorsettings SERIAL DEBUGPORT:1 BAUDRATE:115200
```

Создание корневого раздела и загрузочного виртуального процессора

Первыми шагами, которые полностью инициализированный гипервизор должен предпринять, является создание корневого раздела и первого виртуального процессора, используемого для запуска системы (известного как BSP VP). Корневой раздел создается почти по тем же правилам, что и дочерние: множество слоев раздела инициализируются один за другим. А если конкретно, то происходит следующее.

1. Слой VM инициализирует максимально допустимое число уровней VTL и устанавливает привилегии раздела в зависимости от его типа (подробности см. в предыдущем подразделе). Кроме того, слой VM определяет допустимые функции в зависимости от уровня совместимости указанного раздела. Корневой раздел поддерживает максимум допустимых функций.
2. Слой VP инициализирует виртуализованные данные CPUID, используемые всеми виртуальными процессорами, когда гостевая операционная система запрашивает у них CPUID. Слой VP создает процесс в рамках гипервизора, который и обеспечивает раздел.
3. Диспетчер адресации (AM) строит изначальное адресное пространство раздела, используя зависмый от платформы машинный код, который выстраивает EPT для Intel, NPT — для AMD. Построенное физическое адресное пространство зависит от типа раздела. Корневой раздел использует идентичное отображение, в силу чего вся физическая память гостя соответствует системной физической памяти. (подробности приводятся в пункте «Физическое адресное пространство раздела» далее).

Наконец, после того, как SynIC, IOMMU и общие страницы перехватов были корректно настроены для данного раздела, гипервизор создает и запускает виртуальный процессор BSP для корневого раздела, который уникален и используется для перезапуска процесса загрузки.

Виртуальный процессор гипервизора (VP) представлен большой структурой данных `VM_VP` (рис. 9.6). Она содержит все данные, используемые для отслеживания состояния виртуального процессора: состояние и данные зависимых от платформы регистров (общего назначения, отладочные, зона XSAVE и стек), закрытое адресное пространство VP и массив структур данных `VM_VPLC`, в которых отслеживается состояние каждого виртуального уровня доверия (Virtual Trust Level, VTL) виртуального процессора. Кроме того, в состав `VM_VP` входят указатель на поддерживающий VP поток и указатель на физический процессор, на котором VP на данный момент выполняется.

Как и в случае с разделами, создание виртуального процессора BSP похоже на создание обычных виртуальных процессоров. Функция `VmAllocateVp` отвечает за размещение и инициализацию необходимой памяти из секции памяти, используемой для хранения структуры данных `VM_VP`, ее зависящей от платформы части и массива `VM_VPLC` (одного для каждого поддерживаемого VTL). Гипервизор копирует изначальный контекст процессора, определенный HvLoader во время загрузки, в структуру `VM_VP`, после чего создает закрытое адресное пространство VP и присоединяется к нему (только если допущена изоляция адресного пространства).

Наконец, он создает поток, обеспечивающий VP. Это важный шаг: далее построение виртуального процессора продолжается в контексте его собственного потока. Главный системный поток гипервизора ожидает, пока новый BSP VP не инициализируется полностью. Это побуждает планировщик гипервизора выбрать только что созданный поток, где выполняется функция `ObConstructVp`, которая строит VP в контексте нового потока.

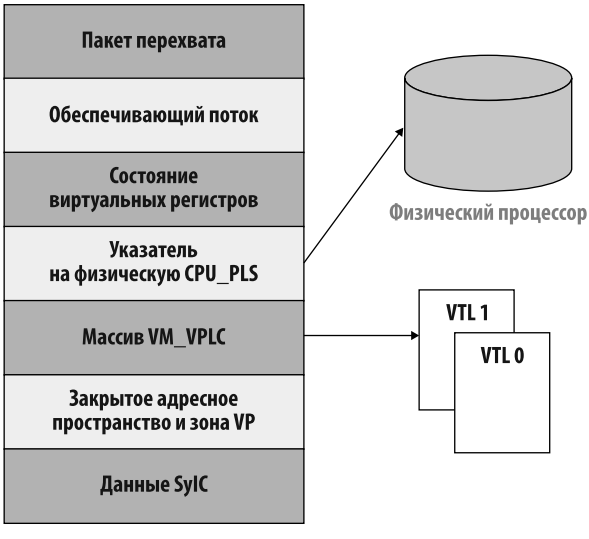


Рис. 9.6. Структура данных VM_VP, представляющая виртуальный процессор

Подобно тому как происходит с разделами, `ObConstructVp` выстраивает и инициализирует каждый слой виртуального процессора.

1. Слой диспетчера виртуализации связывает структуру данных физического процессора (`CPU_PLS`) с VP и устанавливает VTL 0 в качестве активного.
2. Слой VAL инициализирует зависящие от платформы фрагменты VP — его регистры, зону XSAVE, стек и данные для отладки. Кроме того, для каждого поддерживаемого VTL он размещает и инициализирует структуру данных VMCS (VMCBB для систем AMD), которая используется оборудованием для отслеживания состояния виртуальной машины, и таблицы страниц SLAT для каждого VTL. Последнее позволяет изолировать все VTL друг от друга (подробнее о VTL — в подразделе «Виртуальные уровни доверия и виртуальный безопасный режим»). Наконец, слой VAL разблокирует и активирует VTL 0. Зависящая от платформы VMCS (или VMCB для систем AMD) полностью скомпилирована, таблица SLAT для VTL 0 активирована, и эмулятор реального режима инициализирован. Часть VMCS, касающаяся состояния хоста, нацеливается на цикл диспетчеризации VAL гипервизора. Этот цикл — самая важная часть гипервизора, поскольку он управляет всеми событиями VMEXIT, генерируемыми каждым из гостей.

3. Слой VP размещает страницу гипервызовов VP и для каждого гипервызова — страницы перехватов и поддержки. Они используются гипервизором для того, чтобы делиться кодом и/или данными с гостевой операционной системой.

Когда `ObConstructVp` завершает работу, поток диспетчеризации VP активирует виртуальный процессор и его синтетический контроллер прерываний (SynIC). Если этот VP — первый в корневом разделе, поток диспетчеризации восстанавливает изначальный контекст VP, сохраненный в структуре данных `VM_VP`, посредством записи каждого из захваченных регистров в зависимую от платформы VMCS (или VMCB) область процессора (этот контекст был определен HvLoader в процессе загрузки). Наконец, поток диспетчеризации сигналит об окончании инициализации VP (в результате чего главный системный поток входит в цикл бездействия) и входит в зависимый от платформы цикл диспетчеризации VAL. Этот цикл обнаруживает, что данный VP новый, готовит его к первому исполнению и запускает новую виртуальную машину, выполняя инструкцию VMLAUNCH. Новая VM возобновляет работу ровно с той же точки, в которой HvLoader передал управление в гипервизор. Загрузка продолжается обычным путем, но уже в контексте нового раздела гипервизора.

Диспетчер памяти гипервизора

Диспетчер памяти гипервизора довольно прост в сравнении с диспетчером памяти безопасного ядра или ядра NT. Процесс, управляющий набором физической памяти, известен как *секция памяти гипервизора*. Прежде чем начнется запуск гипервизора, его загрузчик (HvLoader.dll) размещает блок загрузки гипервизора и заранее подсчитывает максимальное количество физических страниц, которые гипервизор будет использовать для корректного запуска и создания корневого раздела. Это количество зависит от страниц, используемых для инициализации IOMMU, чтобы хранить структуры диапазонов памяти, системную базу данных PFN, таблицы страниц SLAT и VA пространство HAL. Загрузчик гипервизора заранее выделяет рассчитанное количество физических страниц, отмечает их как зарезервированные и присоединяет к блоку загрузки массив со списком страниц. Позже, когда гипервизор запустится, он создаст секцию корневого раздела, используя список страниц, размещенный ранее его загрузчиком.

На рис. 9.7 показано строение структуры данных секции памяти. Она отслеживает общее количество страниц, внесенных в секцию, которые можно где-нибудь выделить или освободить. Секция запоминает свои физические страницы в разных списках, упорядочиваемых узлом NUMA. Сама она помнит лишь заголовок каждого списка. Состояние каждой физической страницы и ссылка на нее в списке NUMA поддерживаются за счет записей в базе данных PFN. Также секция отслеживает отношение страниц к корневому разделу. Новая секция может быть создана за счет физических страниц, принадлежащих родителю (корню). Когда секцию удаляют, все оставшиеся в нем физические страницы возвращаются родителям.

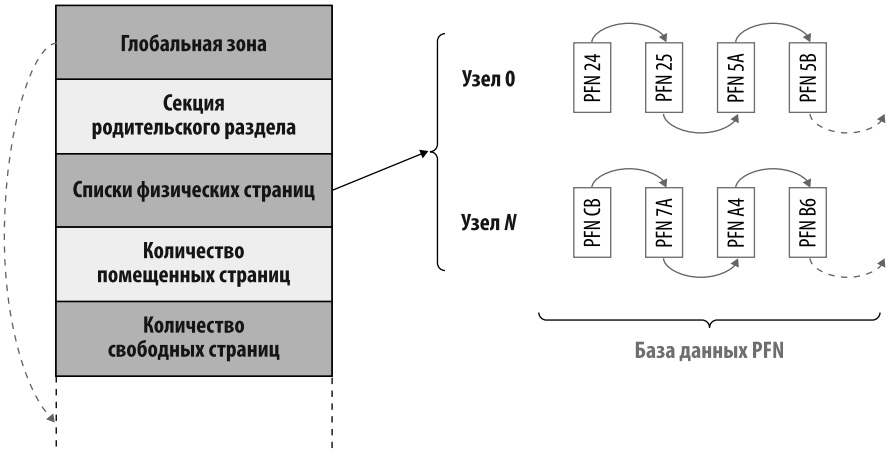


Рис. 9.7. Секция памяти гипервизора. Виртуальное адресное пространство для глобальной зоны резервируется с конца структуры данных секции

Когда гипервизору требуется сколько-то физической памяти для неважно какой задачи, он выделяет ее из активной секции (в зависимости от раздела). Это значит, что выделить может и не получиться. В случае отказа возможны два сценария.

- Если выделение было запрошено для нужд внутренней службы гипервизора (в основном от лица корневого раздела), отказ считается недопустимым и происходит фатальный отказ системы. (Это объясняет, почему изначальный расчет общего количества страниц для назначения корневому разделу должен быть точным.)
- Если выделение было запрошено от лица дочернего раздела (обычно через гипервызов), гипервизор откажет в запросе, присвоив статус `INSUFFICIENT_MEMORY`. Корневой раздел, обнаружив эту ошибку, выделяет несколько физических страниц (подробнее в разделе «Стек виртуализации» далее), которые будут добавлены в секцию дочернего раздела посредством гипервызова `HvDepositMemory`. Наконец, операцию можно начать заново (обычно с успехом).

Физические страницы, выделенные из секции, обычно отображаются гипервизором с помощью виртуального адреса. Когда создается секция, виртуальный диапазон адресов (4 или 8 Гбайт в зависимости от того, для корневого или дочернего раздела оно предназначено) выделяется с целью отображения новой секции, его битовой карты PDE и глобальной зоны.

Зона гипервизора инкапсулирует закрытый диапазон VA, который не делится остальным адресным пространством гипервизора (см. пункт «Изоляция адресного пространства» далее в данной главе). Гипервизор работает с одной корневой таблицей страниц, в отличие от ядра NT, где применяется теневое копирование KVA. Две записи в корневой таблице страниц зарезервированы для динамического переключения между каждой из зон и адресными пространствами виртуальных процессоров.

Физическое адресное пространство раздела

Как говорилось ранее, когда новый раздел создается, гипервизор выделяет ему физическое адресное пространство. Оно содержит все структуры данных, нужных оборудованию для преобразования гостевых физических адресов (GPA) в системные физические адреса (SPA). Данная аппаратная функция позволяет осуществлять преобразование, зачастую называемое преобразованием адресов второго уровня (SLAT). Термин SLAT не привязан к платформе, разные производители используют свои названия. Intel называет это расширенными таблицами страниц (extended page tables, EPT), AMD использует термин «вложенные таблицы страниц» (nested page tables, NPT), в ARM просто говорят о преобразовании адресов второго уровня.

Обычно реализация SLAT подобна реализации таблиц страниц в x64, где используются четыре уровня преобразования (преобразование виртуальных адресов x64 детально разбиралось в главе 5 тома 1). ОС, действующая внутри этого раздела, использует тот же метод преобразования виртуальных адресов, который применялся бы на «железном» оборудовании. Однако в первом случае физический процессор на деле выполняет преобразование на двух уровнях: одном — для виртуальных адресов, втором — для физических. На рис. 9.8 показана настройка SLAT для гостевого раздела. В дочерних разделах GPA, как правило, преобразуется в отличный от него SPA. В корневом разделе это не так.

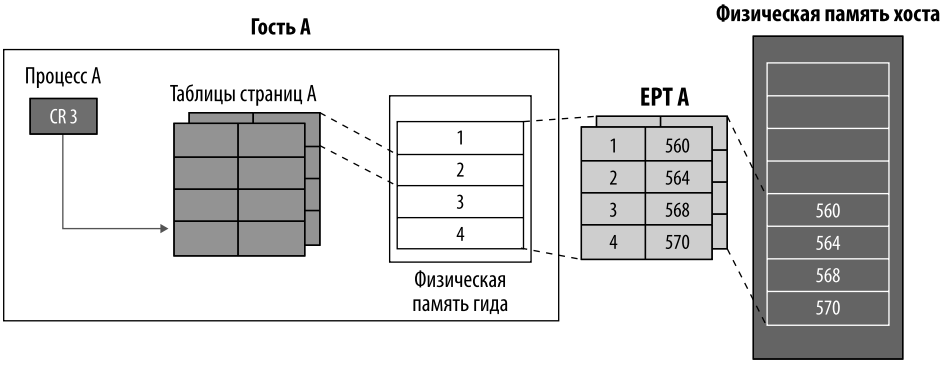


Рис. 9.8. Преобразование адресов для гостевого раздела

Когда гипервизор создает корневой раздел, он строит его изначальное физическое адресное пространство, используя сопоставление идентификации. В данной модели каждый GPA соответствует такому же SPA (например, гостевой кадр 0x1000 в корневом разделе отображается как физически существующий кадр 0x1000). Гипервизор заранее выделяет память, необходимую для отображения всего физического адресного пространства машины (о котором загрузчик Windows должен был узнать с помощью сервисов UEFI, подробности см. в главе 12) на все допустимые виртуальные уровни доверия (VTL) корневого раздела. Уровень защиты, применяемый для физического кадра каждого раздела, позволяет создавать различные домены безопасности (VTL), которые могут быть изолированы друг от друга. Более подробно VTL рассматриваются в разделе «Безопасное ядро» далее в данной главе.

Страницы самого гипервизора помечаются как зарезервированные аппаратно и не отражаются в таблице SLAT раздела (на самом деле они отражаются с помощью заведомо неверной точки входа в несуществующую PFN).

ПРИМЕЧАНИЕ Из соображений производительности, когда гипервизор строит отображение физической памяти, он способен обнаруживать крупные непрерывные фрагменты физической памяти и по аналогии с памятью виртуальной может отображать их с помощью больших страниц. Если ОС, действующая в данном разделе, по какой-то причине решит применить более детальный подход к защите физической страницы, гипервизор воспользуется резервной памятью, чтобы разбить большую страницу в таблице SLAT.

Ранние версии гипервизора поддерживали и еще одну технику отображения физического адресного пространства раздела — теневые страницы. Она использовалась на машинах без поддержки SLAT. Эта техника очень сильно влияла на производительность, поэтому больше не поддерживается. (На машине должна поддерживаться SLAT, в ином случае гипервизор откажется запускаться.)

Таблица SLAT корневого раздела строится во время его создания, но у гостевых разделов это не совсем так. Когда создается дочерний раздел, гипервизор формирует для него первоначальное адресное пространство, но размещает только корневую таблицу страниц (PML4) для каждого из его VTL. Перед запуском новой VM драйвер VID (часть стека виртуализации) резервирует необходимые для нее физические страницы (их точное число зависит от объема памяти VM), выделяя из корневого раздела. (Напомним, что речь идет о физической памяти — только драйвер может выделять физические страницы.) Драйвер VID обслуживает список физических страниц, которые анализируются и делятся на большие страницы и затем отправляются гипервизору с помощью повторного гипервызова `HvMapGpaPages`.

Прежде чем послать запрос на отображение, драйвер VID обращается к гипервизору, чтобы тот создал необходимые таблицы страниц SLAT и внутренние структуры данных о пространстве физической памяти. Каждая иерархия таблиц страниц SLAT размещается для каждого VTL, доступного в этом разделе (операция называется *pre-commit*). Эта операция может закончиться отказом, в частности, когда секция памяти нового раздела не может уместить достаточно физических страниц. В таком случае, как уже говорилось, драйвер VID выделяет дополнительную память из секции корневого раздела и отдает ее дочернему. После этого драйвер VID сможет свободно отобразить все физические страницы дочернего раздела. Гипервизор компонует и компилирует все необходимые таблицы страниц SLAT, задавая различные уровни защиты в зависимости от уровня VTL. (Большим страницам нужно на один уровень перенаправления меньше.) На этом создание физического адресного пространства дочернего раздела завершается.

Изоляция адресного пространства

Уязвимости спекулятивного исполнения, обнаруженные в современных процессорах (также известные как Meltdown, Spectre и Foreshadow), позволяли злоумышленнику считывать секретные данные, расположенные в более привилегированном контексте,

с помощью спекулятивного чтения устаревших данных из кэша процессора. Это значит, что программа, исполняемая на гостевой VM, была способна прочесть собственную память гипервизора или более привилегированного корневого раздела. Внутренние детали Spectre, Meltdown и всех прочих уязвимостей по сторонним каналам и меры, которые Windows предпринимает против них, подробно рассматривались в главе 8.

Гипервизор был способен отражать большинство видов таких атак, реализуя технику HyperClear. Эта мера опирается на три ключевых компонента изоляции VM: планировщик ядер, изоляцию адресного пространства виртуального процессора и очистку важных данных. В современных многоядерных процессорах различные потоки SMT часто делят между собой один и тот же кэш. (Подробности о планировщике ядер и симметричной многопоточности см. в подразделе «Планировщики Hyper-V».) В виртуализованной среде потоки SMT на одном ядре способны независимо входить в контекст гипервизора и выходить из него в зависимости от своей активности. Например, такие события, как прерывания, могут заставить поток SMT переключиться с исполнения на виртуальном процессоре на исполнение в контексте гипервизора. Это может происходить независимо для каждого потока SMT, в силу чего один такой поток может работать в контексте гипервизора, пока другой родственный ему поток все еще действует в контексте виртуального процессора гостевой VM. Злоумышленник, запустивший код в контексте виртуального процессора гостевой VM с меньшим уровнем доверия на одном потоке SMT, может воспользоваться уязвимостью стороннего канала для получения потенциальной возможности обозревать важные данные и контекст гипервизора, где действует родственный поток SMT.

Гипервизор обеспечивает мощную изоляцию данных для защиты от вредоносной гостевой VM путем предоставления отдельных виртуальных диапазонов адресов для каждого гостевого потока SMT, которые обеспечивают виртуальные процессоры. Когда отдельно взятый поток SMT входит в контекст гипервизора, никакие секретные данные адресовать нельзя. В кэш процессора могут быть занесены только данные, ассоциированные с текущим гостевым виртуальным процессором либо представляющие общий ресурс в рамках гипервизора. Как показано на рис. 9.9, когда VP, действующий в потоке SMT, входит в гипервизор, корневой планировщик принудительно делает так, чтобы родственный поток SMT работал на другом VP в рамках той же VM. Более того, никакие общие секретные данные в гипервизоре не отражаются. В случаях, когда гипервизору требуется доступ к секретным данным, он удостоверяется, что для родственного потока SMT не запланировано других VP.

В отличие от ядра NT, гипервизор всегда работает с одной корневой таблицей страниц, что создает единое глобальное адресное пространство. В его работе имеет место понятие закрытого адресного пространства, но это название не совсем правильное. На деле гипервизор резервирует две записи в глобальной корневой таблице страниц (записи PML4, создающие диапазон виртуальных адресов 1-TB) для отображения и скрытия закрытого адресного пространства. При изначальном создании VP гипервизор выделяет две корневые записи в закрытой таблице страниц. Они используются для отображения секретных данных VP, таких как стек и содержащие закрытые данные структуры данных. Переключение адресного пространства означает создание двух записей в глобальной корневой таблице страниц

(что объясняет неверность термина *«закрытое адресное пространство»* — на самом деле это закрытый *диапазон* адресов. Гипервизор переключается между закрытыми адресными пространствами только в двух случаях: когда создан новый виртуальный процессор и во время переключения между потоками. (Напомним: потоки обеспечиваются VP. Планировщик ядер проверяет, чтобы ни один родственный поток SMT не выполнял VP из другого раздела.) Во время исполнения поток гипервизора отображает лишь закрытые данные собственного VP — никакие другие секретные данные этому потоку не доступны.

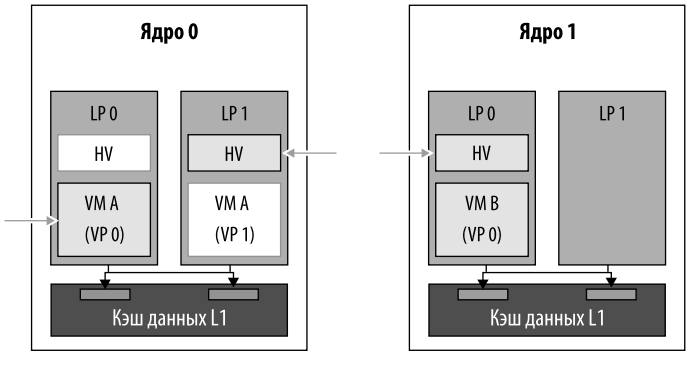


Рис. 9.9. Техника HyperClear

Секретные данные отображаются в закрытом адресном пространстве с помощью зоны в памяти, представленной структурой данных `MM_ZONE`. Эта зона инкапсулирует закрытый поддиапазон VA закрытого адресного пространства, где гипервизор обычно хранит секретные данные отдельно взятого VP.

Зона в памяти похожа на закрытое адресное пространство. Но вместо отображения записей корневой таблицы страниц в корне глобальной таблицы страниц она отображает закрытые каталоги страниц в двух корневых записях, используемых закрытым адресным пространством. В рамках структуры зоны имеются массив каталогов страниц, которые будут отображены или скрыты в закрытом адресном пространстве, и битовая карта для отслеживания используемых таблиц страниц. На рис. 9.10 показаны связи между закрытым адресным пространством и зоной памяти. Эти зоны могут быть по требованию отображены или скрыты (в закрытом адресном пространстве), но обычно это происходит только при создании VP. На самом деле гипервизору не нужно переключать их при смене потока — закрытое адресное пространство уже включает в себя диапазон адресов, доступный через зону памяти.

На рис. 9.10 структуры таблицы страниц, связанные с закрытым адресным пространством, оформлены по-разному: те, что относятся к зоне памяти, написаны серым шрифтом, а общие, принадлежащие гипервизору, заштрихованы. Переключение между адресными пространствами — это довольно малозатратная операция, которая требует изменений в двух записях PML4 в корне таблицы страниц гипервизора. Привязка или отсоединение зоны памяти с закрытым адресным пространством требует лишь одного изменения в PDPTE этой зоны (размер VA зоны непостоянен, PDPTE всегда расположены подряд).

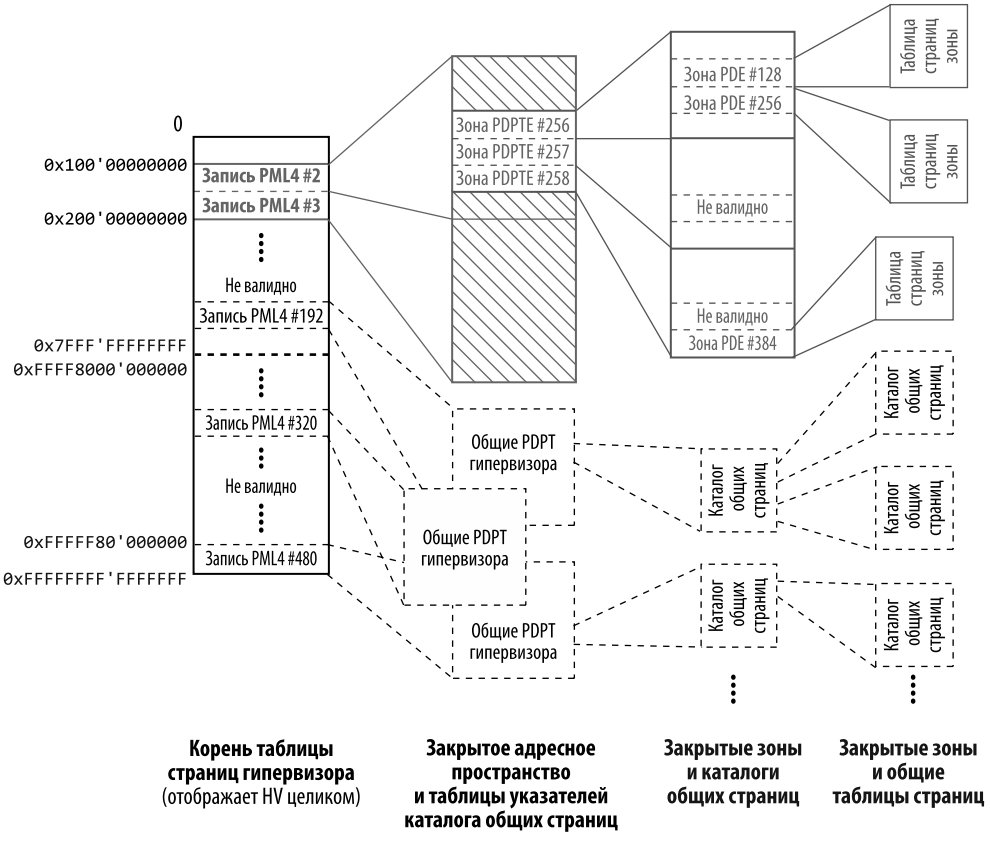


Рис. 9.10. Закрытые адресные пространства и закрытые зоны памяти в рамках гипервизора

Динамическая память

Виртуальные машины могут использовать разный процент выделенной им физической памяти. Какие-то из них задействуют лишь малую долю назначенной им гостевой физической памяти, а бóльшую часть оставляют свободной или обнуленной. Производительность же других виртуальных машин может страдать в условиях высокой нагрузки на память, когда файл подкачки используется слишком часто, так как назначенной гостевой физической памяти слишком мало. Во избежание такого сценария гипервизор и стек виртуализации поддерживают технологию динамической памяти. *Динамической памятью* называют физическую память виртуальной машины, способную динамически же увеличивать или уменьшать свой объем. Данную функцию обеспечивают несколько компонентов:

- диспетчер памяти ядра NT, который поддерживает добавление или удаление физической памяти на ходу, в том числе в «пустой» системе;
- гипервизор с помощью SLAT (управляется диспетчером адресации);

- процесс VMWorker, который использует модуль контроллера динамической памяти Vmdynmem.dll для установки связи с драйвером динамической памяти VSC для VMBus (Dmvsc.sys), работающим в дочернем разделе.

Чтобы корректно описать динамическую память, следует вкратце ознакомиться с тем, как база данных номеров фреймов страницы (Page Frame Number, PFN) создается ядром NT. База данных PFN используется Windows для отслеживания физической памяти. Детально она рассматривалась в главе 5 тома 1. Для создания такой базы ядро NT сначала рассчитывает гипотетический объем, который потребуется, чтобы отобразить максимально возможный физический адрес (256 Тбайт для стандартных 64-разрядных систем), после чего отмечает виртуальное адресное пространство, необходимое для того, чтобы полностью отобразить его как зарезервированное (базовый адрес хранится в глобальной переменной `MmPfnDatabase`). Обратите внимание на то, что это зарезервированное пространство все еще не имеет выделенных таблиц страниц. Ядро NT в цикле обходит каждый дескриптор физической памяти, найденный диспетчером загрузки (с помощью сервисов UEFI), соединяет их в максимально длинные диапазоны и для каждого из них формирует записи в базе данных PFN, используя большие страницы. Здесь есть важный нюанс: как показано на рис. 9.11, в базе данных поместится максимальный из возможных объем физической памяти, но лишь маленький ее фрагмент отображает реальные физические страницы. Она называется *разреженной памятью* (sparse memory).

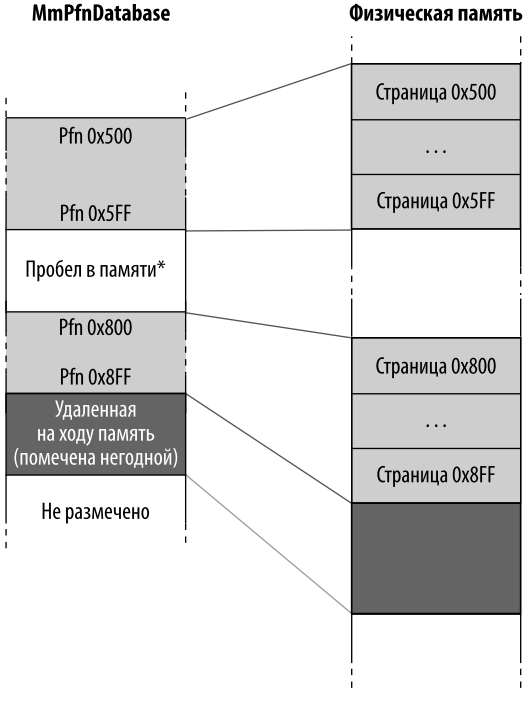


Рис. 9.11. Пример базы данных PFN, откуда была удалена часть физической памяти

Благодаря этому принципу работает горячее добавление и удаление памяти на ходу. Когда в систему добавляется физическая память, драйвер памяти Plug and Play (Pnpmem.sys) обнаруживает это и вызывает функцию `MmAddPhysicalMemory`, которая экспортируется из ядра NT. Она проводит сложную процедуру, в ходе которой вычисляются точное число страниц в новом диапазоне и узел NUMA, им владеющий, после чего новые записи заносятся в базу данных PNF путем создания необходимых таблиц страниц в зарезервированном виртуальном адресном пространстве. Эти новые физические страницы добавляются в список свободных (подробности см. в главе 5 тома 1).

Когда часть физической памяти на ходу удалена, система выполняет обратную процедуру. Она проверяет, относятся ли страницы к корректному списку физических страниц, обновляет внутренние счетчики памяти (например, общего количества физических страниц) и, наконец, освобождает соответствующие записи в PFN, после чего те будут помечены как «плохие». Физические страницы, которые ими описывались, больше никогда не будут задействоваться диспетчером памяти. Из базы данных не удаляется никакое виртуальное пространство. Физическая память, описывавшаяся освобожденными записями PFN, снова может быть добавлена в будущем.

Когда запускается паравиртуализированная VM, драйвер динамической памяти (Dmvsc.sys) определяет, поддерживает ли дочерняя VM функцию горячего добавления. Если да, то он создает рабочий поток, который согласует протокол и устанавливает связь с каналом VMBus из VSP. (См. раздел «Стек виртуализации» далее в данной главе.) Канал подключения VMBus обеспечивает связь между драйвером динамической памяти, работающим в дочернем разделе, и модулем контроллера динамической памяти (Vmdynmem.dll), отображенным в процессе VM Worker в корневом разделе. Запускается протокол обмена сообщениями. Каждую секунду дочерний раздел получает отчет о нагрузке на память и запрашивает различные счетчики производительности, публикуемые диспетчером памяти (глобальное использование файла подкачки, количество доступных, зафиксированных и грязных страниц, количество ошибок страниц в секунду, количество страниц в списке освобожденных или очищенных страниц). Затем отчет отправляется корневому разделу.

Чтобы сделать вычисления, необходимые для определения возможности выполнить операцию горячего добавления, процесс VM Worker из корневого раздела пользуется сервисами, предоставляемыми балансировщиком VMMS — компонентом в составе службы VmCompute. Если статус памяти в корневом разделе позволяет выполнить горячее добавление, балансировщик VMMS вычисляет правильное количество страниц для добавления в дочерний раздел и отправляет обратный вызов процессу VM Worker (через COM), который при поддержке драйвера VID начинает операцию горячего добавления.

1. Резервирует подходящий объем памяти в корневом разделе.
2. Обращается к гипервизору с целью отразить системные физические страницы, зарезервированные корневым разделом на несколько физических страниц гостя, размещенных в дочерней VM, с должной защитой.
3. Отправляет сообщение драйверу динамической памяти с указанием приступить к операции горячего добавления на нескольких гостевых физических страницах, недавно отраженных гипервизором.

Драйвер динамической памяти в дочернем разделе использует функцию `MmAddPhysicalMemory`, публикуемую ядром NT, чтобы выполнить горячее добавление. Та, в свою очередь, размещает записи, описывающие новую гостевую физическую память в базе данных PFN, при необходимости добавляя туда новые страницы.

Подобным образом, когда балансировщик VMMS замечает, что у гостевой VM много доступных физических страниц, он может запросить у дочернего раздела (все так же через процесс VM Worker) выполнить горячее удаление какого-то их количества. Драйвер динамической памяти использует функцию `MmRemovePhysicalMemory` для выполнения операции горячего удаления. Ядро NT проверяет, что каждая страница из диапазона, указанного балансировщиком, находится в списке свободных или очищенных либо относится к стеку, который можно безопасно удалить. Если все условия выполнены, драйвер динамической памяти отправляет диапазон страниц для горячего удаления процессу VM Worker, который воспользуется сервисами драйвера VID, чтобы убрать отображение физических страниц из дочернего раздела и освободить их, чтобы они поступили в распоряжение ядра NT.

ПРИМЕЧАНИЕ Динамическая память несовместима с режимом вложенной виртуализации.

Планировщики Hyper-V

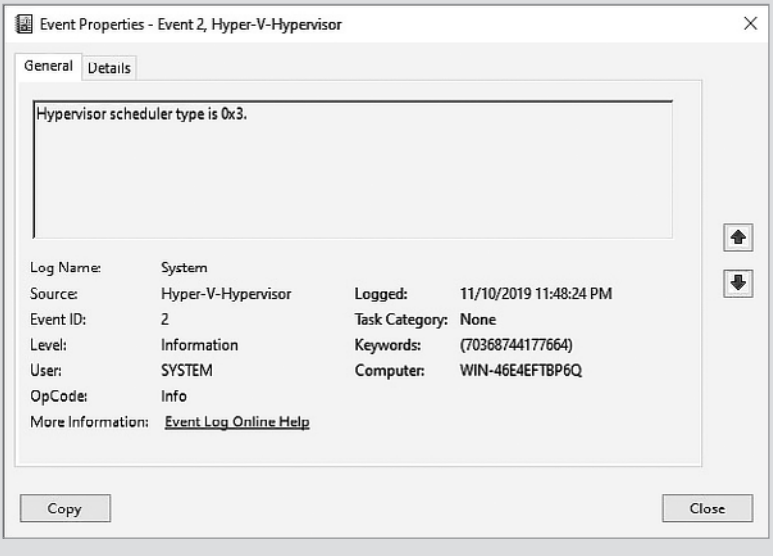
Гипервизор чем-то сам похож на операционную микросистему, действующую на уровень ниже ОС корневого раздела (Windows). Это значит, что она должна быть способна решать, какой поток (поддерживающий виртуальный процессор) на каком физическом процессоре исполняется. Это особенно жизненно, когда система поддерживает множество виртуальных машин с общим числом виртуальных процессоров бóльшим, чем установлено на рабочей станции физических. Задача планировщика гипервизора состоит в том, чтобы выбирать следующий поток для выполнения на физическом процессоре после того, как закончится отведенный отрезок времени или выполняемый на данный момент поток. Hyper-V может использовать три разных планировщика. Чтобы правильно управлять всеми тремя, гипервизор публикует *API планировщика* — набор функций, позволяющий к нему обратиться. И единственное назначение сводится к перенаправлению вызовов API к конкретной реализации планировщика.

ЭКСПЕРИМЕНТ. Выбор типа планировщика для гипервизора

Хотя клиентские версии Windows по умолчанию запускаются с корневым планировщиком, Windows Server 2019 работает с планировщиком ядер. В ходе данного эксперимента вы определите, какой планировщик гипервизора активен в вашей системе, и узнаете, как переключиться на другой после следующей перезагрузки системы.

Определив, какой планировщик выбрать, гипервизор Windows делает запись об этом в журнале системных событий. Вы можете поискать его с помощью

инструмента Просмотр журналов событий (Event Viewer), который можно запустить, набрав в поисковом окне Cortana слово `eventvwr`. Когда апплет запустится, раскройте узел Журналы Windows (Windows Logs) и выберите Система (System log). Вам следует поискать события с ID 2-источником события Hyper-V-Hypervisor. Вы можете сделать это, нажав кнопку Фильтр текущего журнала (Filter Current Log) в правой части окна либо на заголовке графы Код события (Event ID), после чего события отсортируются по убыванию ID (имейте в виду, это действие может затянуться). Дважды щелкнув на найденном событии, вы увидите окно наподобие следующего.



Событие запуска с кодом 2 указывает на действительный тип планировщика, где:

1 — классический, протокол SMT отключен;

2 — классический;

3 — планировщик ядер;

4 — корневой.

Здесь приведен снимок экрана системы Windows Server, которая по умолчанию использует планировщик ядер. Чтобы сменить тип планировщика на классический или корневой, вам потребуется открыть командную строку администратора, набрав в поисковом окне Cortana слово `cmd` и выбрав Запустить с правами администратора (Run As Administrator), и ввести там следующую команду:

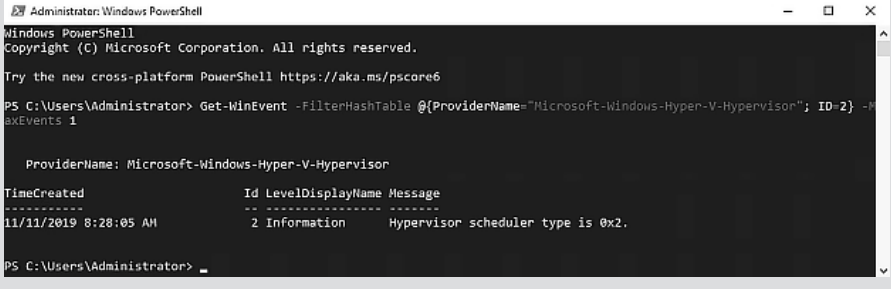
```
bcdedit /set hypervisorschedulertype <Type>
```

где `<Type>` — это `Classic` для классического, `Core` для планировщика ядер или `Root` для корневого. Теперь можно перезагрузить систему и проверить новую запись события от Hyper-V-Hypervisor с кодом 2. Вы также можете узнать текущий

планировщик гипервизора с помощью консоли PowerShell администратора и следующей команды:

```
Get-WinEvent -FilterHashTable @{ProviderName="Microsoft-Windows-Hyper-V-Hypervisor";
ID=2}
-MaxEvents 1
```

Она извлекает последнее событие с кодом 2 из журнала событий Система (System).



Классический планировщик

Классический планировщик использовался по умолчанию во всех версиях Hyper-V с самого первого релиза. В стандартной конфигурации он реализует простую циклическую политику, в которой любой виртуальный процессор в текущем состоянии исполнения (состояние исполнения зависит от общего количества VM, действующих в системе) имеет равные шансы запуститься. Кроме того, классический планировщик поддерживает установку привязки виртуального процессора и принимает решение о планировании с учетом узла NUMA физического процессора. Классический планировщик не знает, что именно сейчас исполняет гостевой VP. Единственное исключение связано с паравиртуализацией спин-блокировки. Когда ядро Windows, поддерживающее раздел, сочтет необходимым войти в активное ожидание методом спин-блокировки, оно сделает гипервызов, чтобы проинформировать гипервизор (механизмы синхронизации высокого IRQL описывались в главе 8). Классический планировщик может приостановить исполнение виртуального процесса, который еще не отработал отведенный ему отрезок времени или не закончил работу, и запланировать для работы другой. Таким образом он экономит спин-циклы активного процессора.

При конфигурации по умолчанию классический планировщик выделяет каждому VP одинаковые отрезки времени. Это значит, что в сильно нагруженных системах с избытком подписчиков, где множество виртуальных процессоров пытаются исполняться, а все физические процессоры очень заняты, производительность может стремительно снижаться. Для решения этой проблемы классический планировщик поддерживает ряд опций тонкой настройки (рис. 9.12), способных изменить принятие решений о планировании.

- **Резервирование VP.** Пользователь может заранее зарезервировать ресурсы процессора от лица гостевой машины. Резервирование представляет собой

указание процента от предельной загрузки физического процессора, который должен быть доступен гостевой машине на момент ее запланированного запуска. В итоге Hyper-V планирует этот VP только тогда, когда указанный минимум ресурсов процессора свободен, а значит, выделенный отрезок времени гарантируется.

- **Лимиты VP.** Аналогично резервированию VP пользователь может ограничить нагрузку VP на физический процессор. Это значит, что в ситуации с высокой нагрузкой отрезки времени для VP будут уменьшены.
- **Веса VP.** Позволяет контролировать вероятность того, что VP будет запланирован, когда зарезервированный ресурс доступен. В конфигурациях по умолчанию все VP имеют равные шансы продолжить работу. Если пользователь определит вес VP какой-то виртуальной машины, решения о планировании будут принимать, исходя из относительного коэффициента этого веса. Предположим, что в системе с четырьмя процессорами работают одновременно три виртуальные машины. У первой VM коэффициент веса 100, у второй — 200, а у третьей — 300. Полагая, что все физические процессоры в системе работают с одинаковым количеством VP, вероятность запуститься у VP из первой VM будет 17 %, у VP из второй — 33 %, а у VP из третьей — 50 %.

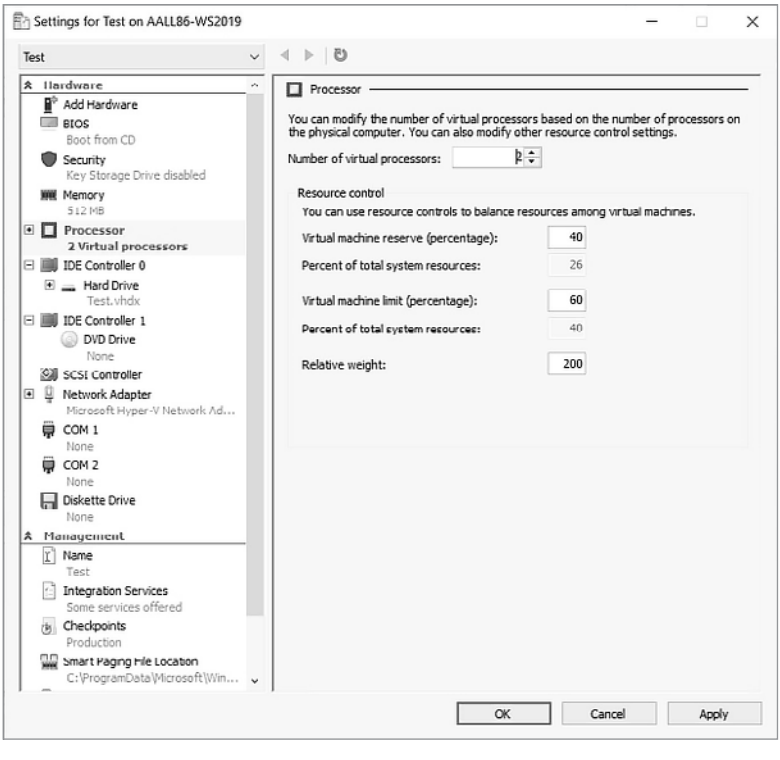


Рис. 9.12. Окно тонкой настройки дополнительных параметров, которое доступно, только когда активен классический планировщик

Планировщик ядер

В обычном случае классическое ядро процессора имеет один конвейер исполнения, на котором потоки инструкций исполняются по очереди. Инструкция попадает на конвейер, проходит несколько этапов (загрузить данные, вычислить, сохранить данные, например) и уходит с конвейера. Различные типы инструкций пользуются разными частями ядра процессора. Современное ядро процессора часто способно исполнять по несколько последовательных инструкций из потока вне очереди (с учетом порядка, в котором они зашли на конвейер). Современные процессоры, поддерживающие исполнение вне очереди, часто реализуют принцип, называемый симметричной многопоточностью (symmetric multithreading, SMT): ядро процессора имеет два конвейера исполнения и представляет более одного логического процессора в системе, таким образом, два разных потока инструкций могут исполняться бок о бок на одном общем устройстве (ресурсы ядра, как и его кэши, общие). Эти два конвейера исполнения представляются программе как отдельные независимые процессоры. С этого момента термином *«логический процессор»* (LP) мы будем называть конвейер исполнения некоего SMT-ядра, представленный Windows как независимый процессор (SMT рассматривались в главах 2 и 4 тома 1).

Эта аппаратная реализация привела к множеству проблем с безопасностью: инструкция, выполняемая общим процессором, может вмешаться и повлиять на инструкцию, выполняемую родственным LP. Более того, кэш физического процессора тоже общий, так что какой-нибудь LP может изменить его содержимое. Другой родственный процессор теоретически может заглянуть в данные, хранящиеся в кэше, соизмеряя время, затрачиваемое процессором на доступ к одному и тому же каналу кэша, и тем самым раскрывая секретные данные, используемые другим логическим процессором, как описано в разделе «Аппаратные уязвимости к атакам по сторонним каналам» главы 8. Классический планировщик в обычном случае сможет выбрать два потока из разных VM и выполнить на двух LP в составе одного процессорного ядра. Это явно недопустимо, потому что в таком контексте одна виртуальная машина могла бы потенциально прочитать данные, принадлежащие другой.

Чтобы решить эту проблему и иметь возможность запускать VM с предсказуемой производительностью, в Windows Server 2016 был представлен планировщик ядер. Он опирается на SMT, чтобы обеспечить изоляцию и мощные защитные ограничения для гостевых VP. Когда он активен, Hyper-V распределяет виртуальные ядра по физическим ядрам. Более того, планировщик стремится никогда не позволять VP из разных VM распределяться на родственные SMT потоки в рамках физического ядра. Планировщик ядер позволяет виртуальной машине самой пользоваться SMT. VP, предоставленные VM, могут быть частью набора SMT. ОС и приложения, запущенные на гостевой виртуальной машине, могут пользоваться программными интерфейсами (API) и поведением SMT, чтобы контролировать и распределять работу между потоками SMT, как если бы они работали в невиртуализованной среде.

На рис. 9.13 приводится пример системы с SMT с четырьмя логическими процессами, распределенными между двумя физическими ядрами процессора. В этом

примере запущены три VM. Первая и вторая VM имеют четыре VP, объединенных в группы по два, а третья — только один назначенный ей VP. Группы VP в этих VM обозначены буквами от A до E. Отдельные VP в рамках группы не заняты (нет кода для исполнения), они отмечены черным цветом.

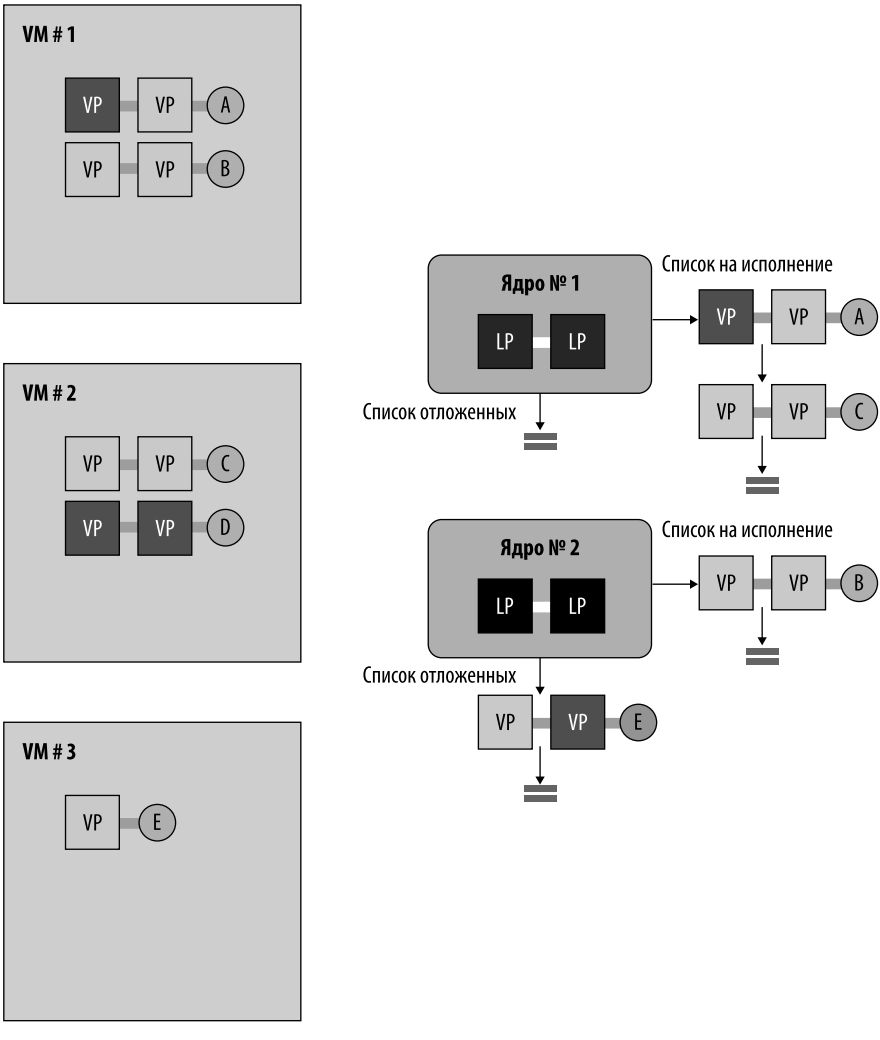


Рис. 9.13. Пример системы SMT с двумя процессорными ядрами и тремя запущенными VM

Каждое ядро имеет список на исполнение, куда входят группы VP, готовые к исполнению, и список отложенных групп, которые тоже готовы, но в список ядра на исполнение пока не добавлены. Группы VP исполняются на физических ядрах. Если все VP в группе окажутся бездействующими, она исключается из плана и ни в одном списке на исполнение не фигурирует (на рис. 9.13 в таком положении оказалась

группа D). Единственный VP группы E недавно вышел из состояния бездействия. Его назначили на процессорное ядро № 2. На рисунке рядом с ним показан фиктивный родственный VP. Это потому, что LP ядра № 2 никогда не возьмется за другие VP, пока на его родственном LP исполняется VP, принадлежащий VM 3. Таким же образом никакие другие VP не будут распределены на физическом ядре, если один VP из группы стал бездействующим, когда остальные все еще заняты (например, как в группе A). Каждое ядро обрабатывает группу VP на вершине своего списка на исполнение. Если групп на исполнение не осталось, ядро бездействует, пока новую группу VP не добавят в его список отложенных. Когда это происходит, ядро возобновляет работу и очищает свой отложенный список, перенося его содержимое в список на исполнение.

Планировщик ядер реализован рядом различных компонентов (рис. 9.14), которые соблюдают между собой строгое разграничение по слоям. Сердцем планировщика является *планировочный модуль*, соответствующий виртуальному ядру группы VP SMT (для VM без SMT он представляет отдельный VP). В зависимости от типа виртуальной машины планировочный модуль содержит один или два привязанных к ней потока. Процесс гипервизора имеет список планировочных модулей, владеющих потоками, обеспечивающими VP из этой VM. Планировочный модуль является для планировщика отдельной единицей планирования, к которой во время исполнения применяются такие характеристики, как резервирование, вес и предел. Планировочный модуль остается активным в ходе выделенного отрезка времени, может быть заблокирован и разблокирован и способен мигрировать между разными ядрами физического процессора. Важным концептом является то, что планировочный модуль аналогичен потоку в классическом планировщике, но не имеет стека или контекста VP, чтоб там работать. Этим владеет один из потоков, привязанных к данному модулю и исполняемых на физическом ядре. *Планировщик групп потоков* является арбитром для каждого планировочного модуля. Эта сущность принимает решение, какой поток из активного планировочного модуля будет работать на каком LP в рамках физического ядра процессора. Она задает соответствие потоков ядрам, применяет политики планирования потоков и обновляет связанные с потоками счетчики.

Каждый LP в рамках физического ядра процессора имеет связанный с ним экземпляр диспетчера логического процессора. *Диспетчер логического процессора* отвечает за переключение потоков, обслуживание таймеров и очистку VMCS (или VMCV в зависимости от архитектуры) для текущего потока. Диспетчеры логического процессора принадлежат *диспетчеру ядра*, который представляет отдельное физическое процессорное ядро и имеет ровно два LP SMT. Диспетчер ядра управляет текущим (активным) планировочным модулем. Диспетчер модулей, привязанный к собственному диспетчеру ядра, решает, какой планировочный модуль должен быть запущен следующим на физическом процессорном ядре, которому тот принадлежит. Последним важным компонентом планировщика ядер является *диспетчер планировщика*, который владеет всеми планировщиками модулей в системе и имеет единое представление обо всех их состояниях. Он обеспечивает балансирование нагрузки и идеальный сервис по назначению ядер для планировщика модулей.

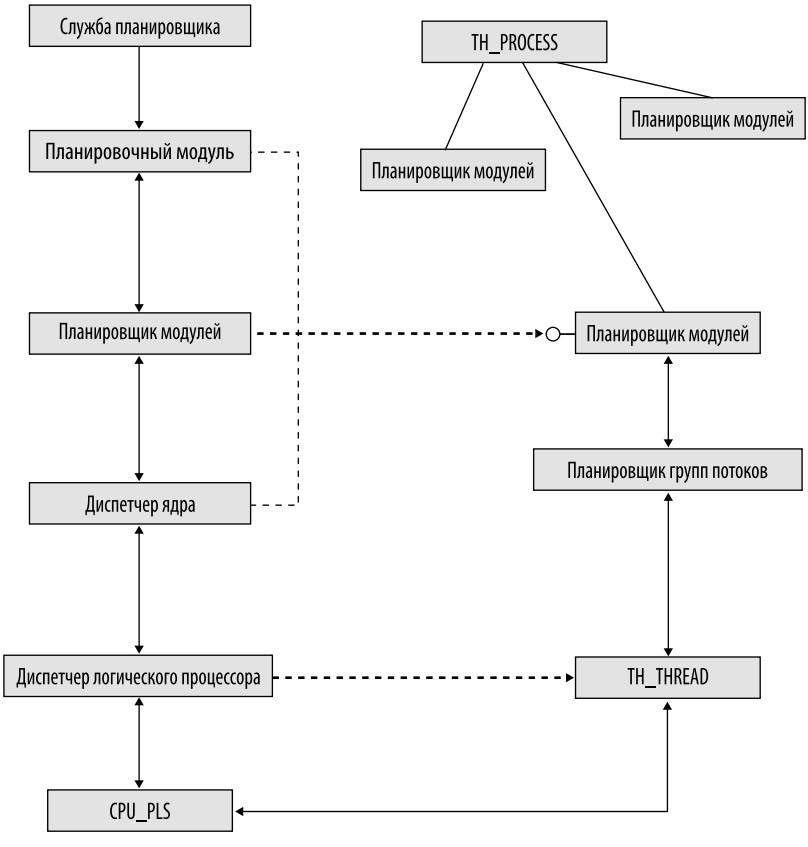


Рис. 9.14. Компоненты планировщика ядер

Корневой планировщик

Корневой планировщик, известный также как интегрированный планировщик, был представлен в обновлении Windows 10 от 18 апреля 2018 года (RS4) с целью позволить корневому разделу планировать работу виртуальных процессоров (VP), принадлежащих гостевым разделам. Корневой планировщик был спроектирован для поддержки легковесных контейнеров, используемых Application Guard Защитника Windows. Эти типы контейнеров, известные как контейнеры типа Krypton и Barcelona, должны управляться корневым разделом, потреблять мало памяти и занимать немного пространства на жестком диске. (Детальное описание контейнеров Krypton выходит за рамки тем данной книги. Познакомиться с серверными контейнерами вы можете в главе 3 тома 1.) В придачу к этому планировщик корневой ОС может оперативно собирать показатели, относящиеся к применению нагрузки процессора внутри контейнера, и использовать эти данные в качестве входных для тех же политик планирования, которые применяются ко всем прочим нагрузкам в системе.

Планировщик NT из экземпляра ОС в корневом разделе управляет всеми аспектами планирования работы системных LP. Чтобы этого достичь, корневой компонент

интегрированного планировщика внутри драйвера VID создает поток диспетчеризации VP в рамках корневого раздела (в контексте нового процесса VMMEM) для каждого гостевого VP (виртуальные машины с поддержкой VA рассматриваются далее в данной главе). Планировщик NT планирует потоки диспетчеризации VP как обычные потоки, для которых применяются дополнительные специфичные для VM/VP политики планирования и компоненты паравиртуализации. Каждый поток диспетчеризации VP выполняет цикл диспетчеризации VP до тех пор, пока драйвер VID не завершит работу соответствующего VP.

Поток диспетчеризации VP создается драйвером VID после того, как процесс VM Worker, подробнее о котором — в разделе «Стек виртуализации» далее в данной главе, запросит создание раздела и VP через `SETUP_PARTITION IOCTL`. Драйвер VID общается с драйвером WinHvr, который, в свою очередь, инициализирует создание гипервизором гостевого раздела посредством гипервызова `HvCreatePartition`. В случаях, когда созданный раздел представляет VM с поддержкой VA или в системе активен корневой планировщик, драйвер VID обращается к ядру NT (через расширение ядра) с целью создать минималистичный процесс VMMEM, ассоциированный с вновь созданным разделом. Драйвер VID также создает поток диспетчеризации VP для каждого VP, принадлежащего этому разделу. Поток диспетчеризации VP исполняется в контексте драйвера VID (и WinHvr). Как показано на рис. 9.15, каждый поток диспетчеризации VP выполняет цикл диспетчеризации VP до тех пор, пока VID не уничтожит соответствующий VP или гостевой раздел и не сгенерирует перехват.

В рамках цикла диспетчеризации VP поток диспетчеризации VP ответствен за следующие действия.

1. Обратиться к новому интерфейсу гипервизора `HvDispatchVp` через гипервызов с целью диспетчеризации VP на текущем процессоре. При каждом обращении по `HvDispatchVp` гипервизор пытается переключить контекст с текущего корневого VP на указанный гостевой VP и позволить ему запустить гостевой код. Одной из важнейших характеристик этого гипервызова является то, что код, который его производит, должен работать на уровне IRQL `PASSIVE_LEVEL`. Гипервизор позволяет гостевому коду выполняться, пока VP не приостановится добровольно, VP не сгенерирует перехват для корня или не произойдет прерывания в адрес корневого VP. Прерывания системных часов все еще обрабатываются корневыми разделами. Когда гостевой VP исчерпает своей отрезок времени, обеспечивающий его поток приостанавливается планировщиком NT. При любом из этих трех событий гипервизор переключается обратно на корневой VP и завершает гипервызов `HvDispatchVo`. Затем он возвращается в корневой раздел.
2. Блокироваться по событию VP-dispatch, если соответствующий VP в гипервизоре заблокирован. Всякий раз, когда гостевой VP блокируется добровольно, поток диспетчеризации VP блокируется по событию VP-Dispatch, пока гипервизор не разблокирует соответствующий гостевой VP и не уведомит об этом драйвер VID. Тот, в свою очередь, сигналит событие VP-Dispatch, и планировщик NT разблокирует поток диспетчеризации VP, который может сделать другой гипервызов `HvDispatchUp`.
3. Обработать все перехваты, о которых гипервизор сообщает при возврате гипервызова. Если гостевой VP генерирует перехват для корня, поток диспетчеризации VP обрабатывает запрос на перехват, полученный в ответ на прошлый

HvDispatchVp, и делает еще один гипервызов HvDispatchVp, после того как VID завершает обработку перехвата. Каждый перехват обрабатывается по-своему. Если он требует обработки процессом VMWP в пользовательском режиме, драйвер WinHvr выходит из цикла и возвращается к VID, который сигнализирует о событии потоку под VMWP и ожидает, пока процесс VMWP не обработает сообщение перехвата, чтобы потом возобновить цикл.

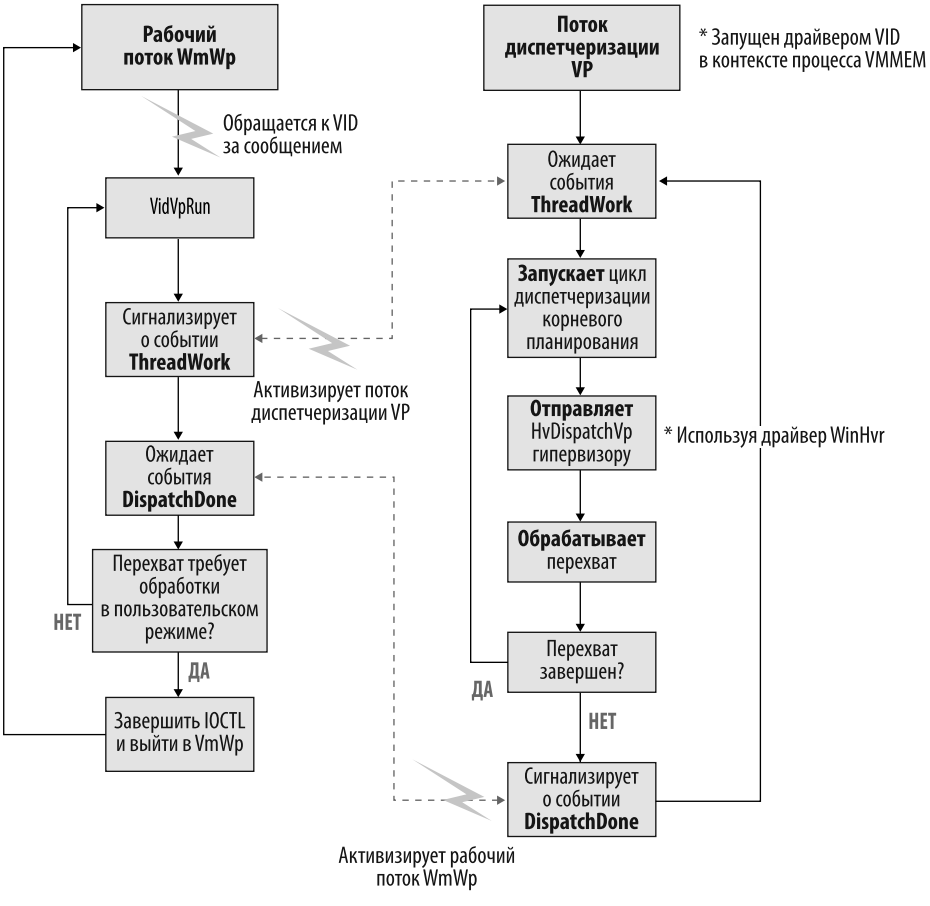


Рис. 9.15. Поток диспетчеризации VP в рамках корневого диспетчера и ассоциированный рабочий поток VMWP, который обрабатывает сообщения гипервизора

Чтобы корректно доставлять сигналы потокам диспетчеризации VP из гипервизора в корень, интегрированный планировщик обеспечивает механизм обмена сообщениями планировщика. Гипервизор отправляет сообщения планировщика корневому разделу через общую страницу. Когда новое сообщение готово к отправке, гипервизор внедряет в корень прерывание SINT, и корень доставляет его соответствующему обработчику ISR в драйвере WinHvr, который направляет сообщение в функцию VID (VidInterceptIsrCallback). Функция отклика на перехват

пытается передать сообщение напрямую в драйвер VID. В случае невозможности передачи сигналится событие синхронизации, которое позволяет циклу диспетчеризации VP прекратиться, а одному из рабочих потоков VmWp — отправить перехват в пользовательском режиме.

Переключения контекста в корневом планировщике более затратны, чем в других реализациях планировщика. Например, когда система переключается между двумя гостевыми VP, ей всегда требуется сгенерировать два выхода в корневые разделы. Интегрированный планировщик воспринимает гостевые и корневые потоки VP очень по-разному, хотя они представлены одинаковыми структурами данных `TH_THREAD`.

- Только корневой поток VP может добавить гостевой поток VP в очередь к физическому процессору. Корневой поток VP имеет повышенный приоритет в сравнении с любыми гостевыми VP, которые на данный момент запущены и диспетчеризуются. Если корневой VP не заблокирован, то интегрированный планировщик прилагает все усилия, чтобы переключить контекст корневого потока VP.
- Гостевой VP имеет два набора состояний: *внутренние* и *корневые состояния потока*. Корневые состояния потока отражают состояния потока диспетчеризации VP, о которых гипервизор общается с корневым разделом. Интегрированный планировщик поддерживает эти состояния для каждого гостевого VP, чтобы знать, как отправить сигнал активации соответствующему потоку диспетчеризации VP в корень.

Только корневой VP может инициировать отправку гостевого VP на его процессор. Он может так сделать или из-за гипервызовов `HvDispatchVp` (в данной ситуации мы говорим, что гипервизор занят внешней работой), или из-за любого другого гипервызова, который требует синхронного запроса в адрес целевого VP (это называется внутренней работой). Если гостевой VP в последний раз запускался на текущем физическом процессоре, планировщик может запустить гостевой поток VP немедленно. В ином случае ему потребуется отправить процессору, на котором гостевой VP работал в последний раз, запрос на очистку и дожидаться, пока другой процессор не очистит контекст VP. Последний случай называется миграцией и создает ситуацию, которую гипервизор вынужден отслеживать через внутренние и корневые состояния потока, которые здесь не описываются.

ЭКСПЕРИМЕНТ. Фокусы с корневым планировщиком

Планировщик NT решает, когда выбрать и запустить виртуальный процессор из какой-то VM и как надолго. Этот эксперимент демонстрирует то, что мы обсуждали ранее: все потоки диспетчеризации VP исполняются в контексте процесса VMMEM, созданного драйвером VID. Для этого эксперимента вам потребуется рабочая станция как минимум с установленным обновлением Windows 10 от 18 апреля 2018 года (RS4), активной ролью Hyper-V и VM с любой операционной системой, готовой к использованию. Процедура создания виртуальной машины детально описана здесь: https://docs.microsoft.com/en-us/virtualization/hyper-v-on-windows/quick-start/quick-create-virtual-machine.

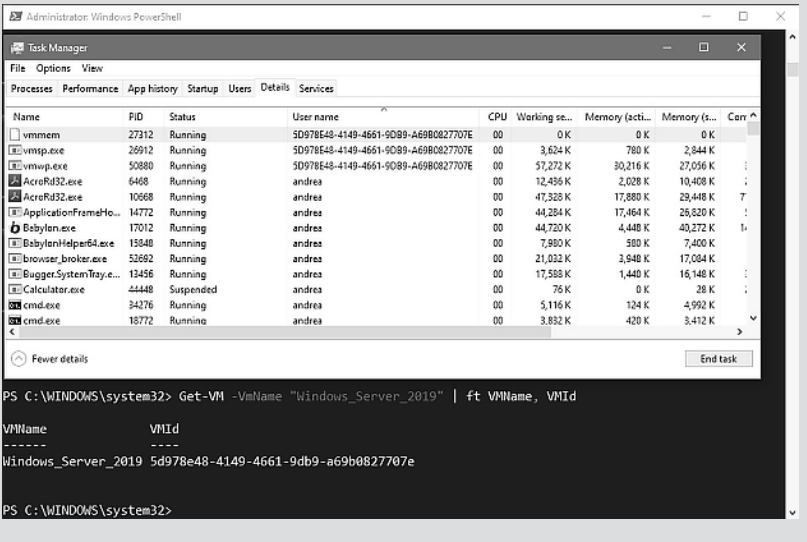
В первую очередь вам следует убедиться, что корневой планировщик активен. Детали данной процедуры можно найти в эксперименте «Выбор типа планировщика для гипервызова» ранее в данной главе. Используемая в эксперименте VM должна быть выключена.

Откройте диспетчер задач, щелкнув правой кнопкой мыши на панели задач и выбрав Диспетчер задач (Task Manager), перейдите на вкладку Подробно (Details) и проверьте, сколько процессов VMMEM сейчас активны. Если не запущено никаких VM, то их там быть не должно, если же установлена роль Application Guard Защитника Windows (WDAG), может иметься экземпляр процесса VMMEM, который содержит предварительно загруженный экземпляр контейнера WDAG. (Этот тип виртуальных машин описывается далее в подразделе «Виртуальные машины с поддержкой VA».) В случае существования экземпляра процесс VMMEM вам следует записать его ID процесса (PID).

Откройте Hyper-V Manager, набрав в поисковом окне Cortana `Hyper-V Manager`, и запустите свою виртуальную машину. После того как VM запустилась и гостевая операционная система успешно загружена, вернитесь в диспетчер задач и поищите новые процессы VMMEM. Если вы щелкнете на новом процессе VMMEM и откроете графу Пользователь (User Name), то увидите, что этот процесс был ассоциирован с токеном, принадлежащим пользователю по имени VM's GUID. Вы можете получить свой VM's GUID, выполнив в консоли PowerShell-администратора следующую команду (замените вхождение `<VmName>` именем своей VM):

```
Get-VM -VmName "<название_VM>" | ft VMName, VmId
```

ID VM и имя пользователя процесса VMMEM должны быть одинаковыми, как показано на следующем рисунке.



Установите Process Explorer (скачав его с сайта https://docs.microsoft.com/en-us/sysinternals/downloads/process-explorer) и запустите с правами администратора. Отыщите процесс VMMEM, найденный на предыдущем шаге (в примере это 27 312), щелкните на нем правой кнопкой мыши и выберите Приостановить (Suspend). Теперь графа Процессор (CPU) процесса VMMEM должна показывать вместо корректного времени ЦП слово Приостановлен (Suspended).

Если вы вернетесь в VM, то заметите, что она не отвечает и полностью зависла. Так случилось, потому что вы приостановили процесс, содержащий потоки диспетчеризации всех виртуальных процессоров, принадлежавших этой VM. Это не позволило планировщику NT их распределить, что помешало драйверу WinHvr отправить нужный гипервызов `HvDispatchVp`, который возобновил бы исполнение VP.

Если щелкнете правой кнопкой мыши на приостановленном VMMEM и выберете Возобновить (Resume), VM возобновит исполнение и продолжит корректно работать.

Гипервызовы и TLFS гипервизора

Гипервызовы обеспечивают механизм, позволяющий операционным системам из корневого и дочернего разделов запрашивать сервисы у гипервизора. Гипервызовы имеют хорошо проработанный набор входных и выходных параметров. Спецификация высокоуровневого функционала (Top Level Functional Specification, TLFS) гипервизора доступна онлайн (https://docs.microsoft.com/en-us/virtualization/hyper-v-on-windows/reference/tlfs), там определены различные соглашения о вызовах, используемые при указании этих параметров. Кроме того, в ней перечислены все общедоступные функции гипервизора, свойства разделов, гипервизор и интерфейсы VSM.

Гипервызовы доступны благодаря зависящему от платформы операционному коду (VMCALL — для систем Intel, VMMCALL — для систем AMD и HVC — для ARM64), который при исполнении всегда вызывает VM_EXIT в гипервизор. WM_EXIT — это события, которые заставляют гипервизор возобновить работу и исполнять собственный код на уровне привилегий гипервизора, который выше, чем у любого другого ПО, запущенного в системе (не считая контекста SMM-прошивок), пока VM приостановлена. События VM_EXIT могут быть сгенерированы по ряду причин. В зависимой от платформы VMCS (или VMCB) непрозрачная структура данных содержит индекс, описывающий причину выхода через VM_EXIT. Гипервизор этот индекс получает и, если выход произошел из-за гипервызова, считывает его входное значение, указанное отправителем (обычно из регистра общего назначения ЦП RCX для систем AMD или 64-разрядных — для Intel). Входное значение гипервызова (рис. 9.16) — это 64-битный параметр, который содержит код гипервызова, его свойства и использованное соглашение о вызове. Доступно три типа соглашений о вызовах.

- **Стандартные гипервызовы.** Входные и выходные параметры хранятся по выровненным по 8 байт гостевым физическим адресам (GPA). ОС передает адреса

через регистры общего назначения (RDX и R8 в системах Intel и 64-разрядных AMD).

- **Быстрые гипервызовы.** Обычно не допускают выходных параметров и используют два регистра общего назначения как стандартные гипервызовы для передачи гипервизору только входных параметров (размером до 16 байт).
- **Расширенные быстрые гипервизоры (или быстрые гипервызовы XMM).** Подобны быстрым гипервызовам, но используют дополнительные регистры с плавающей точкой, чтобы позволять отправителю посылать входные параметры размером до 112 байт.

Рис. 9.16. Входное значение гипервызова (из TLFS по гипервызову)

Существуют два класса гипервызовов: простые и повторяющиеся. *Простой* гипервызов выполняет одиночную операцию и имеет набор входных и выходных параметров фиксированного общего размера. Повторяющийся гипервызов ведет себя как серия простых гипервызовов. Когда отправитель только начинает повторяющийся гипервызов, он указывает количество повторов, где сообщает число элементов в списке входных или выходных параметров. Также отправитель может указать индекс начала повторения, чтобы показать на следующий входной или выходной параметр, который следует использовать.

Все гипервызовы возвращают еще одно 64-битное значение, известное как *результирующее значение гипервызова* (рис. 9.17). Как правило, оно описывает, чем закончилась операция, и в случае повторения гипервызов указывает общее число законченных повторов.

Рис. 9.17. Результирующее значение гипервызова (из TLFS по гипервызову)

Для выполнения гипервызовам может потребоваться некоторое время. Присутствие процессора, который не получает прерываний, может быть опасно для системы-хоста. Например, в Windows существует механизм, который проверяет процессор на предмет неполучения им своего прерывания такта системных часов за период дольше 16 мс. Если это условие будет выполнено, система мгновенно остановится и покажет BSOD. Поэтому при некоторых гипервызовах гипервизор полагается на *механизм продолжения гипервызова*, в том числе все формы гипервызовов rep. Если гипервызов не может завершиться за отведенное ему время

(обычно 50 мкс), управление передается обратно отправителю (через операцию `VM_ENTRY`), но его указатель управления не двигается дальше команды, инициировавшей гипервызов. Это позволяет ожидающим прерываниям обработаться, а другим виртуальным процессорам — запланироваться. Когда оригинальный поток возобновит работу, он выполнит команду гипервызова заново и продвинется в сторону завершения операции.

Как правило, драйвер никогда не делает гипервызов напрямую через зависящий от платформы операционный код. Вместо этого он пользуется сервисами драйвера интерфейса гипервизора Windows, который доступен в двух версиях:

- **WinHvr.sys.** Загружается при запуске системы, если ОС действует в корневом разделе, и предоставляет гипервызовы, доступные как для корневого, так и для дочернего разделов;
- **WinHv.sys.** Загружается, только если ОС действует в дочернем разделе. Он позволяет использовать только гипервызовы, доступные в дочерних разделах.

Функции и структуры данных, экспортируемые драйвером интерфейса гипервизора Windows, широко используются стеком виртуализации, особенно драйвером VID, который, как мы уже рассказывали, играет ключевую роль в функциональности всей платформы Hyper-V.

Перехваты

Корневой раздел должен быть способен создать виртуальную среду, которая позволяла бы немодифицированной гостевой ОС, написанной для работы на физическом оборудовании, действовать в гостевом разделе гипервизора. Подобные устаревшие гостевые системы могут пытаться получить доступ к физическим устройствам, которых не существует в разделе гипервизора (к примеру, путем обращения к определенным портам ввода/вывода или записи в специфические MSR). Для таких случаев гипервизор предоставляет инструмент *перехватов хоста*: когда VP или гостевая ОС выполняет определенные инструкции, авторизованный корневой раздел может перехватить это событие и изменить результат действия перехваченной инструкции так, чтобы с точки зрения дочернего раздела он повторял ожидавшееся поведение физического устройства.

Когда в дочернем разделе происходит событие перехвата, его VP приостанавливается, а гипервизор отправляет в корневой раздел сообщение перехвата с помощью синтетического контроллера прерываний (SynIC, подробности — в следующем подразделе). Сообщение доходит благодаря синтетическому обработчику прерываний (Interrupt Service Routine, ISR) гипервизора, который ядро NT устанавливает на фазе 0 своего запуска только в случаях, когда паравиртуализированная система работает в среде гипервизора (подробнее — в главе 12). Синтетический ISR гипервизора (`KiHvInterrupt`) обычно установлен на вектор 0x30 и передает управление внешней функции обратного вызова, которую драйвер VID зарегистрировал во время своего запуска (через публичную функцию ядра NT `HvlRegisterInterruptCallback`).

Драйвер VID — это драйвер перехвата, который может регистрировать перехваты хоста в гипервизоре, тем самым получая все перехваты, происходящие в дочерних разделах. После инициализации раздела процесс VM Worker регистрирует

перехваты для различных компонентов стека виртуализации. (Например, виртуальная материнская плата регистрирует перехваты ввода/вывода для каждого последовательного порта виртуальной машины.) Он отправляет драйверу IOCTL, который использует гипервызов `HvInstallIntercept`, чтобы установить перехват в дочернем разделе. Когда дочерний раздел вызовет перехват, гипервизор приостановит VP и внедряет прерывание в корневой раздел, где им займется ISR (`KiHvInterrupt`). Последний передаст управление зарегистрированному обратному вызову перехвата VID, который обработает событие и возобновит работу VP, очистив синтетический регистр приостановки из-за перехвата в приостановленном VP.

Гипервизор поддерживает перехват следующих событий в дочернем разделе:

- доступа к портам ввода/вывода (чтение или запись);
- доступа к MSR VP (чтение или запись);
- выполнения инструкции `CPU_ID`;
- исключений;
- доступа к регистрам общего назначения;
- гипервызовов.

Синтетический контроллер прерываний

Гипервизор виртуализует прерывания и исключения как для корневых, так и для дочерних разделов посредством синтетического контроллера прерываний (SynIC), являющегося расширением для виртуализованного APIC (подробности об APIC см. в руководстве для разработчиков ПО для Intel и AMD). SynIC отвечает за доставку виртуальных прерываний виртуальным процессорам. Прерывания, доставленные в раздел, делятся на две категории: *внешние* и *синтетические* (также известные как внутренние или просто виртуальные прерывания). Внешние прерывания приходят от других разделов или устройств, синтетические прерывания приходят от гипервизора и направляются VP раздела.

Когда в разделе создается VP, гипервизор создает и инициализирует SynIC для каждого VTL. Затем он начинает с SynIC VTL 0, тем самым разрешая виртуализацию APIC физического процессора в аппаратной структуре данных VMCS (или VMCB). Гипервизор поддерживает три варианта виртуализации APIC.

- В стандартной конфигурации APIC виртуализуется через аппаратную поддержку внедрения событий. Это значит, что каждый раз, когда раздел обращается к регистрам локального APIC своего VP, портам ввода/вывода или MSR (в случае x2APIC), он вызывает VMEXIT, заставляя код гипервизора отправлять прерывание через SynIC, который в конечном счете внедряет событие нужному гостевому VP путем управления непрозрачными полями в VMCS/VMCB (после того как пройдет через логику, подобную физическому APIC, чтобы определить, возможно ли это прерывание доставить).
- Режим эмуляции APIC работает подобно аналогу из стандартной конфигурации. Каждое физическое прерывание, отправленное оборудованием (обычно через IOAPIC), все еще вызывает VMEXIT, но гипервизору ничего внедрять не нужно. Вместо этого он изменяет *страницу виртуального APIC*, используемую

процессором для виртуализации некоторых вариантов доступа к регистрам APIC. Когда гипервизору требуется внедрить событие, он просто управляет некоторыми виртуальными регистрами, отображаемыми на странице виртуального APIC. Когда происходит VMENTRY, событие доставляется аппаратно. В то же время, когда гостевой VP управляет определенными частями своего локального APIC, никакого VMEXIT не происходит, но изменения будут сохранены на странице виртуального APIC.

- Отложенные прерывания позволяют некоторым видам внешних прерываний доставляться напрямую в гостевой раздел, не вызывая никаких VMEXIT. Это дает возможность устройствам прямого доступа отображаться прямо в дочерний раздел, не вызывая никаких потерь производительности из-за VMEXIT. Физический процессор обрабатывает виртуальные прерывания напрямую, записывая их как отложенные на странице виртуального APIC (подробности см. в руководствах для разработчиков ПО для Intel и AMD).

Когда гипервизор запускает процессор, он обычно инициализирует модуль синтетического контроллера прерываний для физического процессора, представленного структурой данных `CPU_PLS`. Модуль SynIC физического процессора является массивом из дескрипторов прерываний, который обеспечивает связь между физическим и виртуальным прерываниями. Дескриптор прерывания гипервизора (запись в IDT) содержит данные, необходимые SynIC, чтобы корректно отправить прерывание, в частности объект, в который прерывание доставляется (раздел, гипервизор, ложное прерывание), целевой процессор (корневой, дочерний разделы, несколько VP или синтетическое прерывание), вектор прерывания, целевой VTL и ряд других характеристик прерывания (рис. 9.18).

Рис. 9.18. Дескриптор физического прерывания гипервизора

В конфигурациях по умолчанию все прерывания доставляют в корневой раздел на VTL 0 либо самому гипервизору (во втором случае запись о прерывании — это «Зарезервировано гипервизором»). Внешние прерывания могут быть доставлены в дочерний раздел, только если туда отображено устройство прямого доступа (хорошим примером могут послужить устройства NVMe).

Каждый раз, когда обеспечивающий VP поток выбирается для исполнения, гипервизор проверяет, требуется ли доставить одно или несколько синтетических прерываний. Как обсуждалось ранее, синтетические прерывания от оборудования не приходят — обычно их генерирует сам гипервизор (при определенных условиях), и ими все равно управляет SynIC, который способен внедрять синтетические прерывания в правильный VP. Хотя синтетические прерывания активно используются ядром NT (паравиртуализированный таймер часов тому хороший пример), они имеют фундаментальное значение для виртуального безопасного режима (Virtual Secure Mode, VSM). Мы обсуждаем их в разделе «Безопасное ядро» далее в этой главе.

Корневой раздел может отправить настраиваемое виртуальное прерывание в дочерний раздел, используя гипервызов `HvAssertVirtualInterrupt` (описано в TFLS).

Коммуникация между разделами

Синтетический контроллер прерываний также играет важную роль в обеспечении виртуальных машин средствами коммуникации между разделами. Гипервизор предоставляет два принципиальных механизма для коммуникации одного раздела с другим: сообщения и события. В обоих случаях оповещения доставляются целевому VP с помощью синтетических прерываний. Сообщения и события отправляются из исходного раздела в целевой через заранее выделенное *соединение*, ассоциированное с *портом* назначения.

Одним из важнейших компонентов, использующих предоставляемые SynIC сервисы коммуникации между разделами, является VMBus. (Архитектура VMBus рассматривается в разделе «Стек виртуализации» далее в данной главе.) Корневой драйвер VMBus (VMBusr.sys) выделяет в корневом разделе идентификатор порта (порты идентифицируются по 32-битным идентификатором) и создает порт в дочернем разделе, выполняя гипервызов `HvCreatePort` с помощью сервисов, предоставляемых драйвером WinHv.

Порт выделяется в гипервизоре за счет пула памяти получателя. Когда порт создан, гипервизор выделяет 16 буферов сообщений в памяти порта. Буферы сообщений поддерживаются очередью, ассоциированной с источником синтетического прерывания (synthetic interrupt source, SINT) в SynIC виртуального процессора. Гипервизор публикует 16 источников прерывания, что позволяет драйверу VMBus управлять максимум 16 очередями сообщений. Синтетическое сообщение имеет фиксированный размер 256 байт и может вместить только 240 байт (16 байт используются в роли заголовка). Отправитель гипервызова `HvCreatePort` указывает, в какой виртуальный процессор SINT целится.

Чтобы корректно получать сообщения, драйвер WinHv выделяет страницу сообщений синтетических прерываний (synthetic interrupt message page, SIMP), которой затем делится с гипервизором. Когда сообщение встает в очередь на отправку в целевой раздел, гипервизор копирует сообщение из своей внутренней очереди в слот SIMP, соответствующий корректному SINT. Затем корневой драйвер VMBus создает *подключение*, с которым ассоциируется порт, открытый на дочерней VM для родительского раздела, с помощью гипервызова `HvConnectPort`. После того как дочерний раздел разрешил получение синтетических прерываний в корректный слот SINT, коммуникация может начинаться: отправитель может оставить сообщение клиенту, указав ID целевого порта и инициировав гипервызов `HvPostMessage`. Гипервизор внедряет синтетическое прерывание целевому VP, который сможет прочесть содержимое сообщения со страницы сообщений (SIMP).

Гипервизор поддерживает порты и соединения трех типов.

- **Порты сообщений.** Передают 240-байтные *сообщения* в раздел и из раздела. Порт сообщений ассоциируется с отдельным SINT в родительском и дочернем разделах. Сообщения доставляются по порядку через отдельную очередь порта сообщений. Эта характеристика делает сообщения идеальными для подключения и разрыва канала VMBus (подробности описываются в разделе «Стек виртуализации» далее в данной главе).
- **Порты событий.** Получают простые прерывания, связанные с набором флагов, устанавливаемых, когда другая сторона делает гипервызов `HvSignalEvent`. Этот

вид портов обычно используется в качестве механизма синхронизации. Например, VMBus задействует порт событий, чтобы оповестить о том, что в кольцевой буфер, описанный определенным каналом, было помещено сообщение. Когда прерывание события доставляется в целевой раздел, получатель с помощью флага, ассоциированного с событием, точно знает, в какой канал направили прерывание.

- **Порты мониторинга.** Оптимизация портов событий. Вызывать VMEXIT-переключение контекста VM для каждого гипервызова `HvSignalEvent` — затратная операция. Порты мониторинга настраиваются посредством выделения общей страницы (между гипервизором и разделом), где располагается структура данных, описывающая, какой порт событий конкретно ассоциирован с отслеживанием флага оповещения (1 бит на странице). Таким образом, если исходный раздел пожелает отправить прерывание синхронизации, он может просто установить соответствующий флаг на общей странице. Рано или поздно гипервизор заметит на общей странице установленный бит и активирует прерывание в порт событий.

API платформы гипервизора Windows и разделы EXO

Windows все чаще использует гипервизор Hyper-V для обеспечения функциональности, не всегда связанной с запуском традиционных VM. В частности, во второй половине главы мы будем обсуждать VSM — важный компонент защиты современных версий Windows, который использует гипервизор, чтобы обеспечивать более высокий уровень изоляции для функций, предоставляющих критические системные сервисы или обращение с секретными данными, например паролями. Их активация требует, чтобы гипервизор выполнялся на машине по умолчанию.

Сторонние продукты виртуализации, такие как VMWare, Qemu, VirtualBox, Android Emulator и множество подобных, используют виртуализующие расширения, обеспечиваемые оборудованием, чтобы создавать собственные гипервизоры, необходимые им для корректной работы. Это явно несовместимо с Hyper-V, который запускает свой гипервизор раньше, чем ядро Windows запустится в корневом разделе (гипервизор Windows является низкоуровневым, на «голом железе» гипервизором).

Как и Hyper-V, сторонние решения для виртуализации состоят из гипервизора, предоставляющего базовые низкоуровневые абстракции для исполнения процессора и управления памятью виртуальной машины, и стека виртуализации, под которым подразумеваются компоненты решения, обеспечивающие эмуляцию среды для VM (материнская плата, встроенное ПО, контроллеры запоминающих устройств, сами устройства и т. д.).

Главная цель API платформы гипервизора Windows, документированного по адресу https://docs.microsoft.com/en-us/virtualization/api/, — позволить запускать сторонние решения для виртуализации в гипервизоре Windows. В частности, сторонний виртуализующий продукт должен иметь возможность создавать, удалять, запускать и останавливать VM с характеристиками (встроенное ПО, эмулируемые устройства, контроллеры запоминающих устройств), определяемыми их собственным стеком виртуализации. Сторонний стек виртуализации при своих интерфейсах управления продолжает работать в Windows в корневом разделе, что позволяет их клиентам пользоваться его VM прежним способом.

Как показано на рис. 9.19, все API платформы гипервизора Windows работают в пользовательском режиме и реализованы поверх драйверов VID и WinHvr в двух библиотеках: WinHvPlatform.dll и WinHvEmulation.dll (в последней реализован эмулятор инструкций для MMIO).

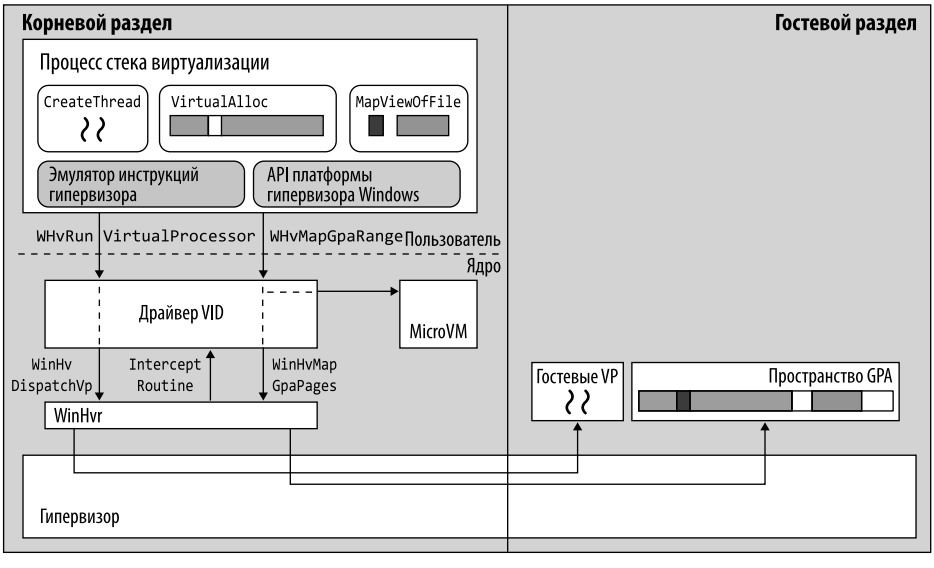


Рис. 9.19. Архитектура API платформы гипервизора Windows

Приложение пользовательского режима, которому требуется создать VM и связанные с ней процессоры, обычно должно сделать следующее.

1. Создать раздел в библиотеке VID (vid.dll) с помощью API `WHvCreatePartition`.
2. Настроить различные внутренние свойства раздела — число виртуальных процессоров, режим эмуляции APIC, тип запрашиваемых VMEXIT и т. д. Все это с помощью API `WHvSetPartitionProperty`.
3. Создать раздел в драйвере VID с помощью API `WHvSetupPartition`. (Этот тип разделов в гипервизоре называется разделами EXO, о них чуть позже.) Данный API также создаст для раздела виртуальные процессоры, которые изначально будут находиться в приостановленном состоянии.
4. Создать соответствующий виртуальный процессор (или несколько) в библиотеке VID с помощью API `WHvCreateVirtualProcessor`. Этот шаг важен, поскольку здесь будет настроен и отображен буфер сообщений для приложения пользовательского режима, используемый для асинхронной коммуникации с гипервизором и потоком, в котором работают виртуальные VP.
5. Выделить адресное пространство для раздела путем резервирования большого диапазона виртуальной памяти классической функцией `VirtualAlloc` (подробности см. в главе 5 тома 1) и отразить его в гипервизоре функцией `WHvMapGpaRange`. Более точечную защиту физической памяти гостевой машины можно обеспечить

при выделении ее в гостевом адресном пространстве, фиксируя различные диапазоны резервированной виртуальной памяти.

6. Создать таблицы страниц и копировать код исходной прошивки в переданную память.
7. Установить содержимое регистров исходного VP с помощью API `WHvSetVirtualProcessorRegisters`.
8. Запустить этот виртуальный процессор, вызвав блокирующую функцию `WHvRunVirtualProcessor`. Управление от него возвращается, только когда в гостевом коде попадается операция, требующая обработки через стек виртуализации (требуется управление VMEXIT в гипервизоре сторонним стеком виртуализации), или в случае внешнего запроса (например, при удалении данного виртуального процессора).

Средства API платформы гипервизора Windows обычно способны обращаться к сервисам гипервизора, отправляя различные IOCTL в объект устройства Device\VidExo, который создается драйвером VID в период инициализации, при условии, что значение параметра реестра `HKLM\System\CurrentControlSet\Services\Vid\Parameters\ExoDeviceEnabled` равняется 1. В ином случае система не предоставляет никакой поддержки API гипервизора.

Некоторые чувствительные к производительности средства API платформы гипервизора (хорошим примером служит `WHvRunVirtualProcessor`) могут напрямую обращаться к гипервизору из пользовательского режима благодаря *сигнальной странице* — особой изначально недействительной гостевой странице, доступ к которой всегда вызывает VMEXIT. В свою очередь, API платформы гипервизора Windows узнает адрес сигнальной страницы от драйвера VID. Он вносит записи в сигнальную страницу при каждом гипервызове из пользовательского режима. Гипервизор определяет отказ и обрабатывает его иным образом благодаря тому, что этот физический адрес отмечен в таблице страниц SLAT как особенный. Гипервизор считывает код и параметры гипервызова из регистров VP так же, как для обычных гипервызовов, и в конечном счете передает управление программе обработки гипервызова. Когда та завершает свою работу, гипервизор наконец выполняет VMENTRY, переходя к инструкции, следующей за ошибочной. Это позволяет потоку, обеспечивающему данный VP, сэкономить много циклов, поскольку ему больше не нужно переходить в ядро, чтобы отправить гипервызов. Более того, операционный код VMCALL и ему подобные всегда требуют привилегий ядра для выполнения.

Виртуальные процессоры новых сторонних виртуальных машин диспетчеризуются корневым планировщиком. Если в данной системе он отключен, все функции API платформы гипервизора будут недоступны. Созданный в гипервизоре раздел имеет тип EXO. Разделы EXO представляют собой минимальные разделы, которые не содержат никакой синтетической функциональности и обладают определенными свойствами, идеальными для сторонних виртуальных машин.

- Они всегда относятся к типам с поддержкой VA (подробности о виртуальных машинах с поддержкой VA или микровиртуальных машинах приведены далее в разделе «Стек виртуализации»). Процесс, владеющий памятью раздела, является не экземпляром процесса VMMEM, а приложением пользовательского режима, создавшим виртуальную машину.

- У них нет никаких привилегий раздела или поддержки каких-либо VTL (виртуальный уровень доверия), кроме нулевого. Все классические привилегии раздела относятся к синтетической функциональности, что обычно предоставляется гипервизором в рамках стека виртуализации Hyper-V. Разделы EXO используются для сторонних стеков. Функциональность, предоставляемая любыми классическими привилегиями раздела, им не нужна.
- Они сами управляют временной синхронизацией. Гипервизор не предоставляет разделам EXO никаких источников прерываний виртуальных часов. Эту ответственность должен взять на себя сторонний стек виртуализации. Это значит, что каждая попытка прочесть счетчик меток времени виртуального процессора будет заставлять гипервизор осуществить VMEXIT, после чего тот перенаправит перехват в поток пользовательского режима, реализующий этот виртуальный процессор.

ПРИМЕЧАНИЕ В сравнении с классическими разделами гипервизора разделы EXO имеют еще ряд мелких отличий. Однако к данному обсуждению они отношения не имеют, а поэтому здесь не приводятся.

Вложенная виртуализация

Иногда крупным серверам и облачным провайдерам требуется возможность запускать контейнеры или дополнительные виртуальные машины в рамках гостевого раздела. Данный сценарий проиллюстрирован на рис. 9.20: гипервизор, работающий поверх «голого» оборудования, обозначаемый L0 (уровень 0), пользуется предоставляемыми оборудованием расширениями виртуализации, чтобы создать гостевую виртуальную машину. Кроме того, гипервизор L0 эмулирует расширения виртуализации от процессора и предоставляет их гостю (возможность предоставлять расширения виртуализации называется *вложенной виртуализацией*). Гостевая виртуальная машина может принять решение запустить еще один экземпляр гипервизора, который в таком случае обозначается L1 (уровень 1), используя эмулированные расширения виртуализации от гипервизора L0. Гипервизор L1 создает вложенный корневой раздел, где запускает корневую операционную систему L2. По тому же принципу корень L2 может взаимодействовать с гипервизором L1, чтобы запустить вложенную гостевую виртуальную машину. Последний гость в данной конфигурации будет называться гостем L2.

Вложенная виртуализация является программной моделью: гипервизор должен быть способен эмулировать расширения виртуализации и управлять ими. Каждая касающаяся виртуализации инструкция, будучи выполненной на гостевой машине L1, вызывает VMEXIT в гипервизоре L0, который через свой эмулятор реконструирует ее и выполняет действия, необходимые для эмуляции. На момент написания данной книги поддерживаются только аппаратные расширения Intel и AMD. Возможность вложенной виртуализации обязательно должна быть явным образом активирована на виртуальной машине L1, в ином случае, если гостевая операционная система выполнит инструкцию виртуализации, гипервизор L0 внедрит в VM исключение общей ошибки защиты.

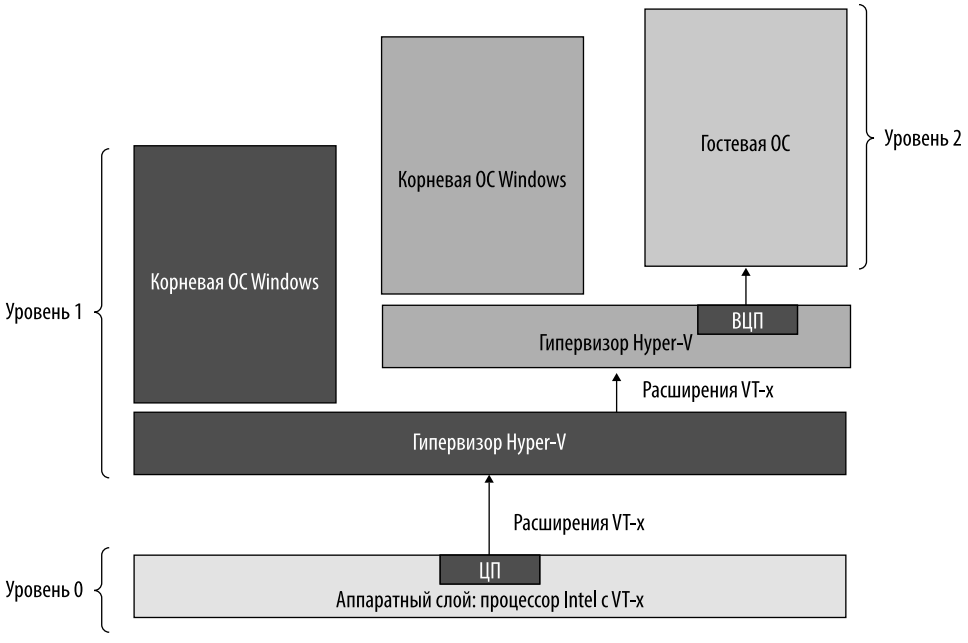


Рис. 9.20. Схема вложенной виртуализации

На оборудовании Intel Hyper-V позволяет вложенную виртуализацию благодаря двум главным принципам, таким как:

- эмуляция расширений виртуализации VT-x;
- преобразование вложенных адресов.

Как уже обсуждалось в данном подразделе, для оборудования Intel основной структурой данных при описании виртуальной машины является контрольная структура виртуальной машины (machine control structure, VMCS). Когда гипервизор L0 создает VP для раздела с поддержкой вложенной виртуализации, кроме стандартной физической VMCS, представляющей VM L1, он размещает несколько вложенных структур данных VMCS (не следует путать с виртуальными VMCS — это часть другой концепции). Вложенная VMCS — это программный дескриптор, содержащий всю информацию, необходимую гипервизору L0 для запуска и работы вложенного VP для раздела L2. Как вкратце упоминалось в подразделе «Запуск гипервизора», когда гипервизор L1 загружается, он определяет, является ли его среда виртуализованной, и, если это так, активирует ряд вложенных компонентов паравиртуализации, как то: паравиртуализированная VMCS или прямая виртуальная очистка (обсуждается далее).

Как показано на рис. 9.21, для каждой вложенной VMCS гипервизор L0 дополнительно выделяет виртуальную VMCS и физическую аппаратную VMCS — две похожие структуры данных, представляющие VP, на котором действует виртуальная машина L2. Виртуальная VMCS важна потому, что играет ключевую роль в обслуживании вложенных виртуализованных данных. Физическая VMCS, в свою

очередь, загружается гипервизором L0 при запуске виртуальной машины L2. Это происходит, когда гипервизор L0 перехватывает инструкцию VMLAUNCH, запущенную гипервизором L1.

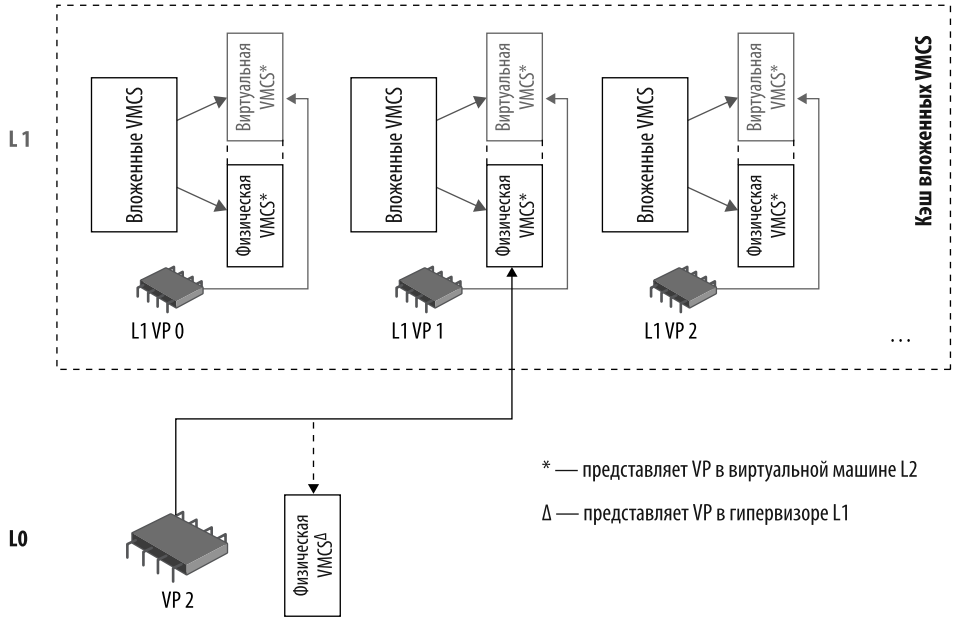


Рис. 9.21. Гипервизор L0 запускает VM L2 на виртуальном процессоре 2

В приведенном примере гипервизор L0 запланировал VP 2 для запуска виртуальной машины L2 под управлением гипервизора L1 (посредством вложенного виртуального процессора 1). Гипервизор L1 может работать только с данными виртуализации, дублированными в виртуальной VMCS.

Эмуляция расширений виртуализации VT-x

На оборудовании Intel гипервизор L0 поддерживает гипервизоры L1 как с паравиртуализацией, так и без. Впрочем, единственной официально поддерживаемой конфигурацией является работа Hyper-V поверх Hyper-V.

В гипервизоре без паравиртуализации все инструкции VT-x, выполняемые в госте L1, вызывают VMEXIT. Разместив гостевую физическую VMCS для описания новой виртуальной машины L2, гипервизор L1 обычно отмечает ее как активную (с помощью аппаратной инструкции Intel VMPTRLD). Гипервизор L0 перехватывает эту операцию и устанавливает связь между выделенной вложенной VMCS и гостевой физической VMCS, указанной гипервизором L1. Далее он заполняет начальные значения для виртуальной VMCS и отмечает вложенные VMCS как активные для текущего VP. (Однако переключения физической VMCS не происходит — контекстом исполнения должен оставаться гипервизор L1.) Все последующие операции

чтения или записи физической VMCS от лица гипервизора L1 всегда перехватываются гипервизором L0 и перенаправляются на виртуальную VMCS (см. рис. 9.21).

Когда гипервизор L1 запускает виртуальную машину (выполняя операцию под названием VMENTRY), он дает на выполнение специальную аппаратную инструкцию (для устройств Intel это VMLAUNCH), которая перехватывается гипервизором L0. В сценариях без паравиртуализации гипервизор L0 копирует все гостевые поля из виртуальной VMCS в другую физическую VMCS, представляющую VM L2, записывает поля хоста, делая их указателями на точки входа гипервизора L0, и помечает ее как активную, используя на платформах Intel инструкцию VMPTRLD. Если гипервизор L1 применяет преобразование адресов второго уровня (на устройствах Intel это EPT), гипервизор L0 выполняет теневое копирование расширенных таблиц страниц текущего L1 (подробнее — в следующем разделе). Наконец, он проделывает действительный VMENTRY с помощью особой аппаратной инструкции. В итоге код с виртуальной машины L2 исполняется аппаратно.

При выполнении в рамках VM L2 каждая операция, вызывающая VMEXIT, переключает контекст исполнения обратно на гипервизор L0 вместо L1. В ответ на это гипервизор L0 проделает VMENTRY снова, уже с оригинальной физической VMCS, представляющей контекст гипервизора L1, внедряя синтетическое событие VMEXIT. Гипервизор L1 возобновляет исполнение и обрабатывает перехваченное событие как обычное невложенное VMEXIT. Закончив внутреннюю обработку синтетического события VMEXIT, он выполняет операцию VMRESUME, которая будет вновь перехвачена гипервизором L0 и обработана подобно упомянутой ранее изначальной VMENTRY.

Генерация VMEXIT каждый раз, когда гипервизор L1 выполняет инструкцию виртуализации, — это затратная операция, которая определенно может внести вклад в общее замедление виртуальной машины L2. Для обхода данной проблемы гипервизор Hyper-V поддерживает паравиртуализированную VMCS. Активация этого средства оптимизации позволяет гипервизору L1 загружать, считывать и записывать данные виртуализации на странице памяти, общей для гипервизоров L1 и L0, вместо физической VMCS. Эта страница и называется паравиртуализированной VMCS. Когда гипервизор VMCS работает с данными виртуализации от VM L2, вместо использования аппаратных инструкций, из-за которых происходит VMEXIT в гипервизоре L0, он считывает и записывает данные в паравиртуализированной VMCS. Данная мера существенно повышает производительность виртуальной машины L2.

В сценариях с паравиртуализацией гипервизор L0 перехватывает только операции VMENTRY и VMEXIT (и некоторые другие, которые здесь не обсуждаются). Гипервизор L0 обрабатывает VMENTRY в порядке, похожем на случай без паравиртуализации, но прежде, чем приступить к действиям, описанным ранее, копирует данные паравиртуализированной VMCS, расположенные на общей странице, в виртуальную VMCS, представляющую виртуальную машину L2.

ПРИМЕЧАНИЕ Стоит упомянуть, что в сценариях без паравиртуализации гипервизор L0 поддерживает еще одну технику предотвращения VMEXIT при работе с данными виртуализации, известную как теневая VMCS. Так называется аппаратная оптимизация, очень похожая на паравиртуализированную VMCS.

Преобразование вложенных адресов

Как обсуждалось в подразделе «Физическое адресное пространство раздела», гипервизор пользуется SLAT, чтобы обеспечить изолированное гостевое адресное пространство для виртуальной машины и преобразовать GPA в реальные SPA. Вложенным виртуальным машинам потребуется еще один аппаратный уровень преобразования поверх двух существующих. Для поддержки вложенной виртуализации этот новый слой должен быть способен преобразовать GPA L2 в GPA L1. Из-за роста сложности электроники, необходимой для создания MMU процессора, управляющего тремя уровнями преобразований, гипервизор Hyper-V избрал иную стратегию обеспечения дополнительного слоя преобразования адресов — *теневые вложенные таблицы страниц*. Они используют технику, схожую с применяемой теневыми страницами (см. предыдущий подраздел), для прямого преобразования GPA L2 в SPA.

Когда создается раздел, поддерживающий вложенную виртуализацию, гипервизор L0 создает и инициализирует домен теневого копирования вложенных таблиц страниц. Эта структура данных используется для хранения списка теневых вложенных таблиц страниц, связанных с различными виртуальными машинами L2, созданными в этом разделе. Кроме того, там хранятся номер генерации активного домена текущего раздела (обсуждается далее в данном подразделе) и статистика вложенной памяти.

Когда гипервизор L0 выполняет начальную операцию VMENTRY для запуска виртуальной машины L2, он размещает теневую вложенную таблицу страниц, связанную с этой машиной, и инициализирует ее пустыми значениями (в итоге получается пустое физическое адресное пространство). Когда VM L2 начинает исполнять код, это немедленно вызывает VMEXIT в гипервизоре L0 из-за отказа вложенной страницы (на платформах Intel это нарушение EPT). Гипервизор L0 вместо внедрения отказа в L1 выполняет обход вложенных таблиц страниц гостя, построенных гипервизором L1. В случае нахождения корректной записи для указанного GPA L2 он считывает соответствующее GPA L1, преобразует в SPA и создает необходимую иерархию в теневой вложенной таблице страниц, чтобы отразить это в VM L2. Затем он делает запись в таблице корректной SPA (для отражения теневых вложенных страниц гипервизор использует большие страницы) и возобновляет исполнение VM L2 напрямую, помечая описывающие вложенные VMCS как активные.

Чтобы преобразование вложенных адресов проходило корректно, гипервизор L0 должен знать обо всех изменениях, происходящих во вложенных таблицах страниц L1, в ином случае виртуальная машина L2 может остаться работать с устаревшими записями. Реализация этого зависит от платформы: обычно гипервизоры защищают вложенную таблицу страниц L2, позволяя только ее чтение. Таким образом, когда гипервизор L1 вносит туда изменения, им это становится известно. Гипервизор Hyper-V следует иной хитрой стратегии. Согласно ей гарантируется, что теневая вложенная таблица страниц, описывающая VM L2, всегда обновляется в силу следующих условий.

- Когда гипервизор L1 добавляет новые записи во вложенную таблицу страниц L2, он не выполняет более никаких действий для вложенной VM (в гипервизоре L0 не появляется никаких перехватов). Запись в теневой вложенной таблице страниц появляется, только когда отказ на вложенной странице приводит к VMEXIT в гипервизоре L0, как описано в сценарии ранее.

- Что до невложенных VM, то, когда запись во вложенной таблице страниц изменяют или удаляют, гипервизор всегда должен вызывать очистку TLB, чтобы корректно аннулировать аппаратный TLB. В случае вложенной виртуализации, когда гипервизор L1 запускает очистку TLB, L0 перехватывает этот запрос и полностью аннулирует теневую вложенную таблицу страниц. Гипервизор L0 обслуживает виртуальный TLB с помощью идентификаторов генерации, хранящихся как в теневой VMCS, так и в домене теневого копирования вложенных таблиц страниц. (Описание архитектуры виртуального TLB выходит за рамки тематик данной книги.)

Полное аннулирование теневой вложенной таблицы страниц из-за смены одного адреса может показаться избыточным, но это продиктовано аппаратной поддержкой. (Аппаратная инструкция IVEPT для Intel не позволяет указать, какое конкретное GPA удалить из TLB.) В классических виртуальных машинах это не проблема, так как изменения физического адресного пространства случаются нечасто. Когда такая машина запускается, вся ее память уже выделена. (В разделе «Стек виртуализации» это рассматривается более подробно.) Впрочем, для виртуальных машин с поддержкой VA и VSM это не так.

Для улучшения производительности в процессе работы с неклассическими VM и VSM (см. подробности в следующем разделе) гипервизор поддерживает компонент виртуализации «прямая виртуальная очистка», которая предоставляет гипервизору L1 два гипервызова, чтобы аннулировать TLB напрямую. В частности, гипервызов `HvFlushGuestPhysicalAddressList` (документирован в TLFS) позволяет гипервызову L1 аннулировать конкретную запись в теневой вложенной таблице страниц, избавляясь от потерь производительности, связанных с очисткой всей теневой вложенной таблицы страниц и необходимостью нескольких операций VMEXIT, чтобы ее реконструировать.

ЭКСПЕРИМЕНТ. Активация вложенной виртуализации в Hyper-V

Как объяснялось в данном разделе, чтобы запустить виртуальную машину на машине L1 в Hyper-V, сначала вы должны включить вложенную виртуализацию в системе-хосте. Для данного эксперимента нужен будет процессор Intel или AMD и ОС Windows 10 или Windows Server 2019 (как минимум с ежегодным обновлением RS1). Вам потребуется создать виртуальную машину Type-2 с минимум 4 Гбайт оперативной памяти, используя Hyper-V Manager или Windows PowerShell. В данном эксперименте вы создадите вложенную виртуальную машину L2 в первой виртуальной машине, поэтому нужно назначить достаточно памяти.

После первого запуска виртуальной машины и первоначальной настройки ее нужно будет выключить и открыть консоль администратора PowerShell (в поисковом окне Cortana наберите `Windows PowerShell`, щелкните правой кнопкой мыши на значке PowerShell и выберите Запуск от имени администратора (Run As Administrator)). Наберите там следующую команду, заменяя вхождения "<*название_VM*>" именем своей виртуальной машины:

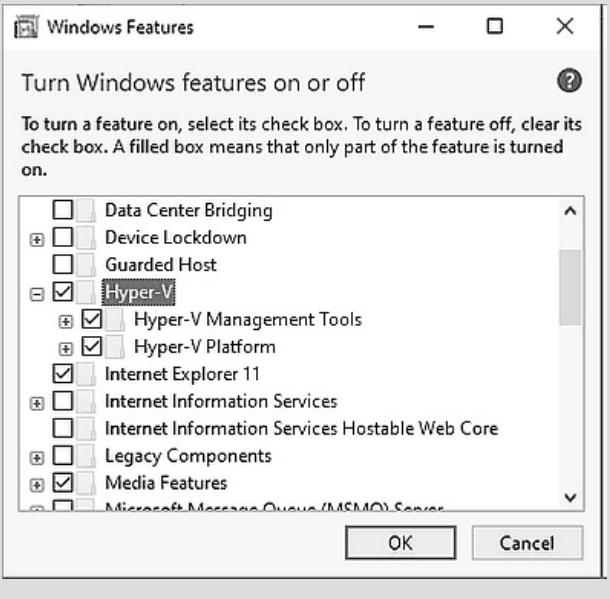
```
Set-VMProcessor -VMName "<название_VM>" -ExposeVirtualizationExtension $true
```

Чтобы точно проверить, активна ли у вас возможность вложенной виртуализации, команда:

```
$(Get-VMProcessor -VMName "<название_VM>").ExposeVirtualizationExtensions
```

должна вернуть `True`.

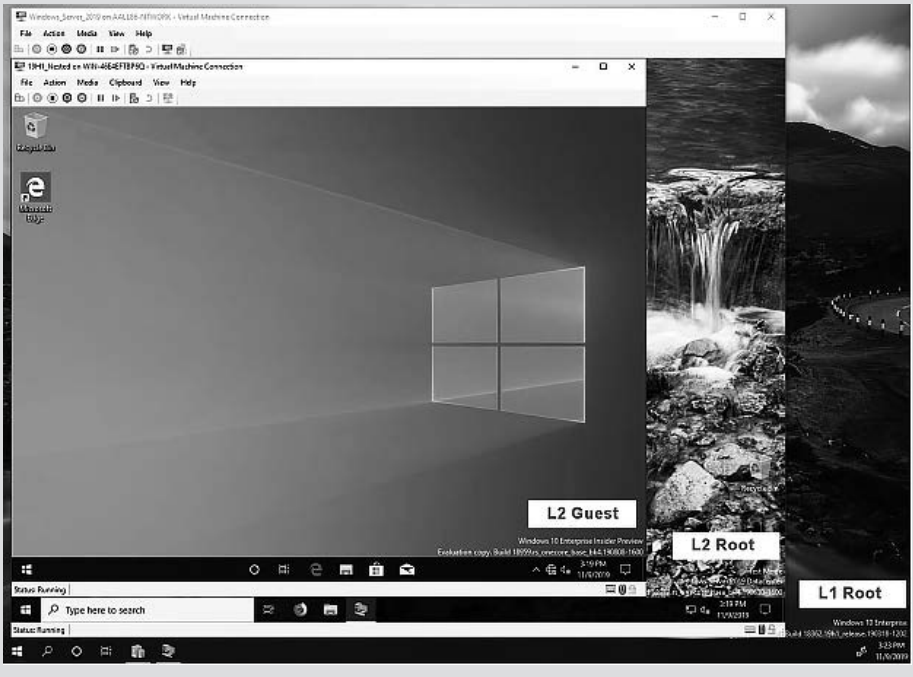
После того как возможность вложенной виртуализации будет активирована, можете перезапустить свою виртуальную машину. Прежде чем сделать возможным запуск гипервизора L1 на этой машине, на панель управления надо добавить необходимый компонент. В поисковом окне Cortana найдите Панель управления (Control Panel), откройте Программы (Programs) и выберите Включение и отключение компонентов Windows (Turn Windows Features On Or Off). Отметьте все дерево Hyper-V, как показано на следующем рисунке.



Нажмите OK. После завершения процедуры нажмите Перезагрузка (Restart) для перезапуска виртуальной машины (этот шаг обязателен). Когда машина будет готова к работе, вы сможете проверить присутствие гипервизора L1 с помощью приложения «Сведения о системе». (В поисковом окне Cortana наберите `msinfo32`. Для получения подробностей см. эксперимент «Определение VBS и предоставляемых им сервисов» далее в данной главе.) Если по каким-то причинам гипервизор не был запущен, можете запустить его принудительно, открыв консоль администратора на виртуальной машине (наберите `cmd` в поисковом окне Cortana и выберите Запуск от имени администратора (Run As Administrator)) и вставив следующую команду:

```
bcdedit /set {current} hypervisorlaunchtype Auto
```

На этом этапе вы уже можете использовать Hyper-V Manager или Windows PowerShell для создания гостевой машины L2 напрямую на своей виртуальной машине. Результат будет чем-то похож на следующую иллюстрацию.



Из корневого раздела L2 вы также сможете активировать отладчик гипервизора L1 способом, объяснявшимся в эксперименте «Подключение отладчика гипервизора» ранее в данной главе. На момент написания книги ограничением было отсутствие во вложенных конфигурациях возможности отладки через сеть — единственный поддерживаемый способ отладки гипервизора L1 предполагает использование последовательного порта. Это значит, что в системе-хосте вам нужно будет активировать в виртуальной машине L1 два виртуальных последовательных порта (один для гипервизора, второй для корневого раздела L2) и привязать их к именованным каналам. Для виртуальных машин Type-2 нужно будет использовать такие команды PowerShell, чтобы настроить два последовательных порта в VM L1 (как и прежде, вы должны поменять вхождение "<*название_VM*>" на имя своей виртуальной машины):

```
Set-VMComPort -VMName "<название_VM>" -Number 1 -Path \\.\pipe\HV_dbg
Set-VMComPort -VMName "<название_VM>" -Number 2 -Path \\.\pipe\NT_dbg
```

После этого вам потребуется настроить отладчик гипервизора на привязку к последовательному порту COM1, тогда как отладчик ядра NT надо привязать к COM2 (подробности см. в предыдущем эксперименте).

Гипервизор Windows на ARM64

В отличие от архитектур x86 и AMD64, где аппаратная поддержка виртуализации была добавлена значительно позже появления их исходной разработки, архитектура ARM64 изначально разрабатывалась с аппаратной поддержкой виртуализации. В частности, как показано на рис. 9.22, среда исполнения ARM64 была разделена на три домена безопасности, известных как *уровни исключений* (exception levels, EL). Каждый EL определяет уровень привилегий: чем выше EL, тем больше их у исполняемого кода. Хотя все приложения пользовательского режима исполняются на EL0, ядро NT и драйверы режима ядра обычно работают на EL1. Как правило, одна программа исполняется только на одном уровне исключений. Уровень привилегий EL2 разработан для запуска гипервизора (в ARM64 известен как *диспетчер виртуальных машин* — Virtual Machine Manager) и является исключением из данного правила. Гипервизор предоставляет сервисы виртуализации и может работать в небезопасной среде сразу на EL2 и EL1. (В безопасной среде EL2 не существует. ARM TrustZone будет обсуждаться далее в данном разделе.)

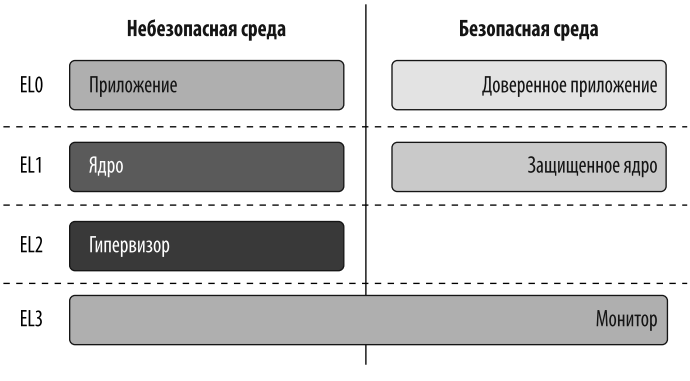


Рис. 9.22. Среда исполнения в ARM64

В отличие от архитектуры AMD64, где ЦП входит в корневой режим (исполнительный домен, где действует гипервизор) только из контекста ядра и при определенных условиях, когда загружается стандартное устройство AMD64, встроенное ПО UEFI и диспетчер загрузки начинают свое исполнение на EL2. На таких устройствах загрузчик гипервизора (или безопасное средство запуска (Secure Launcher) — в зависимости от последовательности загрузки) имеет возможность запустить гипервизор напрямую и через какое-то время сбросить уровень исключений до EL1, отправив инструкцию возврата исключения, известную как ERET.

Вдобавок к уровням исключений технология TrustZone позволяет разделить систему между двумя состояниями защиты исполнения — безопасным и небезопасным. Безопасное ПО, как правило, может получать доступ и к защищенным, и к незащищенным памяти и ресурсам, тогда как обычное может работать только с незащищенными памятью и ресурсами. Небезопасное состояние называют также нормальной средой. Это позволяет какой-нибудь ОС работать параллельно с доверенной ОС на одном оборудовании и обеспечивает защиту от некоторых

программных и аппаратных атак. Безопасное состояние, также известное как безопасная среда, обычно характерно для безопасных устройств (их встроенного ПО и диапазонов IOMMU) и в целом всего, что требует от процессора находиться в безопасном состоянии.

Чтобы корректно сообщаться с безопасной средой, небезопасная ОС отправляет *безопасные вызовы методов* (secure method calls, SMC), которые обеспечивает механизм, похожий на стандартные системные вызовы ОС. Обычно TrustZone реализует разделение на нормальную и безопасную среды через тонкий слой защиты памяти, состоящий из строго определенных модулей защиты памяти (в Quallcom их называют XPU). XPU настраиваются прошивкой так, чтобы только определенные среды исполнения имели доступ к конкретным зонам памяти. (Память безопасной среды недоступна программам из нормальной среды.)

На серверных машинах ARM64 Windows имеет возможность запускать гипервизор напрямую. Клиентские же машины часто не оснащены XPU, даже если функции TrustZone активны. (Большинство клиентских устройств ARM64, на которых работает Windows, произведены QualComm.) На этих клиентских устройствах разделение между нормальной и безопасной средами обеспечивается проприетарным гипервизором под названием QHEE, реализующим изоляцию памяти с помощью преобразования памяти второго уровня (этот слой аналогичен слою SLAT, применяемому в гипервизоре Windows). QHEE перехватывает все SMC, отправляемые действующей ОС: он может перенаправить SMC напрямик в TrustZone (после проверки необходимых прав доступа) или же проделать какую-то работу за него. В таких устройствах TrustZone также несет на себе большую ответственность в части загрузки и авторизации встроенного ПО компьютера и координации с QHEE для корректного исполнения последовательности загрузки по методу Secure Launch.

Хотя в Windows безопасная среда в основном не используется (грань между двумя средами уже обеспечена гипервизором с помощью уровней VTL), гипервизор Hyper-V все же работает на уровне EL2. Это несовместимо с гипервизором QHEE, который действует там же. Чтобы решить эту проблему, Windows при загрузке придерживается особой стратегии: безопасный запуск (Secure Launch) происходит при участии QHEE. Когда он завершается, гипервизор QHEE выгружается из памяти и передает управление гипервизору Windows, загрузившемуся в составе Secure Launch. На более поздних этапах загрузки, после запуска безопасного ядра, в момент, когда SMSS уже создает первый сеанс пользовательского режима, формируется новый специальный доверительный процесс (или «трастлет») (в QualComm его назвали QcExt). Он исполняет обязанности оригинального гипервизора ARM64: перехватывает все запросы SMC, проверяет их целостность, обеспечивает требуемую изоляцию зон памяти (через сервисы, предоставляемые безопасным ядром) и способен получать и отправлять команды из монитора безопасности в EL3.

СТЕК ВИРТУАЛИЗАЦИИ

Хотя гипервизор предоставляет изоляцию и низкоуровневые сервисы для управления виртуализационным оборудованием, высокоуровневая реализация виртуальных машин обеспечивается стеком виртуализцаии. Стек виртуализации управляет состояниями виртуальных машин, предоставляет им память и виртуализует

оборудование, обеспечивая наличие виртуальной материнской платы, встроенного системного ПО и множества видов виртуальных устройств — эмулированных, синтетических и с прямым доступом. Туда же входит VMBus — важный компонент, обеспечивающий скоростной канал коммуникации между гостевой виртуальной машиной и корневым разделом и доступный через слой абстракции за счет клиентской библиотеки режима ядра (kernel mode client library, KMCL).

В данном разделе мы поговорим о некоторых важных сервисах, предоставляемых стеком виртуализации, и проанализируем его компоненты. На рис. 9.23 приведены главные компоненты стека виртуализации.

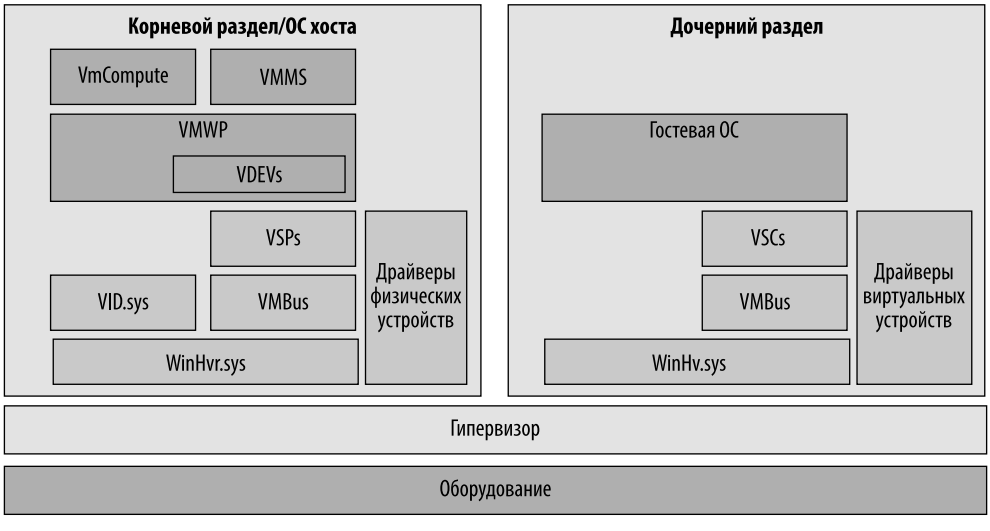


Рис. 9.23. Компоненты стека виртуализации

Служба диспетчера виртуальных машин и рабочие процессы

Служба диспетчера виртуальных машин (Vmms.exe) отвечает за обеспечение корневого раздела интерфейсом инструментария управления Windows (Windows Management Instrumentation, WMI), что позволяет управлять дочерними разделами через плагин консоли управления MMC (Microsoft Management Console) для PowerShell. Служба VMMS обрабатывает запросы, полученные через интерфейс WMI от лица виртуальной машины (обозначаемой с помощью GUID), такие как запуск, выключение, пауза, возобновление, перезагрузка и т. д. Она же управляется настройками в части того, какие устройства видны дочерним разделам и как каждому разделу выделяются память и процессор. VMMS управляет процессом добавления или удаления устройств. Когда виртуальная машина запускается, служба VMM играет критическую роль в создании рабочего процесса виртуальной машины (Virtual Machine Worker Process, VMWP.exe). VMMS управляет снимками состояния VM, перенаправляя запросы на них в процесс VMWP, когда машина запущена, а в противном случае делая снимки самостоятельно.

VMWP выполняет различные работы по виртуализации, относящиеся к обязанностям обычного монолитного гипервизора (подобно деятельности программных решений виртуализации). Сюда входят управление состоянием машины в отдельно взятом дочернем разделе (чтобы поддержать такие функции, как снимки состояния и переход между состояниями), реагирование на всяческие оповещения со стороны гипервизора, эмуляция определенных устройств, видимых в дочернем разделе (они называются эмулируемыми устройствами), и взаимодействие со службой VM и компонентом конфигурации. Рабочий процесс играет важную роль в запуске виртуальной материнской платы и поддержке состояния каждого виртуального устройства в рамках виртуальной машины. Туда же входят компоненты, отвечающие за удаленное управление стеком виртуализации, в том числе компонентом RDP, который позволяет использовать клиент удаленного Рабочего стола для подключения к любому гостевому разделу, удаленно же видеть его пользовательский интерфейс и взаимодействовать с ним. Рабочий процесс VM поддерживает COM-объекты, предоставляющие интерфейс, используемый VMMS (и службой VmCompute) для взаимодействия с экземпляром VMWP, представляющим конкретную виртуальную машину.

Служба вычислений хоста VM, реализованная в образах VmCompute.exe и Vmcompute.dll, — это еще один важный компонент, на который ложатся большинство операций с интенсивными вычислениями, не реализованных в рамках службы диспетчера VM. Операции наподобие анализа отчета о памяти VM (для динамической памяти), управления файлами VHD и VHDX и создания базовых слоев для контейнеров реализуются службой вычисления хоста VM. Рабочие процессы и сама VMMS могут взаимодействовать с ней через предоставляемый ей COM-объект.

Служба диспетчера виртуальных машин, рабочий процесс и служба вычислений VM способны открывать и интерпретировать ряд конфигурационных файлов, отражающих список всех виртуальных машин, созданных в данной системе, и конфигурацию каждой из них. В частности:

- конфигурационный репозиторий хранит список установленных в системе виртуальных машин, их имена, файлы конфигурации и GUID в файле `data.vmcx`, расположенном по адресу `C:\ProgramData\Microsoft\Windows Hyper-V`;
- репозиторий хранилищ данных VM (часть службы вычислений VM) способен открывать, читать и переписывать файл конфигурации (обычно с расширением `.vmcx`) виртуальной машины, содержащий список виртуальных устройств и конфигурацию виртуального оборудования.

Репозиторий хранилищ данных VM также используется для чтения и записи файла сохранения состояния VM. Этот файл формируется во время приостановки VM на паузу и содержит сохраненное состояние запущенной машины (состояние раздела, содержимое памяти машины, состояние каждого виртуального устройства), которое впоследствии можно будет восстановить. Сами файлы конфигурации отформатированы в нотации XML в виде пар «ключ — значение». Текстовое представление данных XML хранится в сжатом виде с использованием проприетарного двоичного формата, дополнительно реализующего логику журнализации, чтобы обеспечить защиту от перебоев с энергоснабжением. Документирование данного формата не входит в задачи книги.

Драйвер VID и диспетчер памяти стека виртуализации

Драйвер виртуальной инфраструктуры (Virtual Infrastructure Driver, VID.sys), вероятно, является одним из важнейших компонентов стека виртуализации. Он предоставляет сервисы управления разделами, памятью и процессорами виртуальных машин, действующих в дочернем разделе, и делает их видимыми для рабочего процесса VM, который существует в корневом разделе. Рабочий процесс и службы VMMS используют драйвер VID для взаимодействия с гипервизором благодаря интерфейсам, реализованным в драйвере интерфейса гипервизора Windows (WinHv.sys и WinHvr.sys), которые драйвер VID импортирует. Они включают в себя весь код для поддержки управления гипервызовами гипервизора и позволяют операционной системе или просто драйверам режима ядра получать доступ к гипервизору с помощью вызова стандартных API Windows вместо гипервызовов.

В драйвер VID входит также диспетчер памяти стека виртуализации. В предыдущем разделе мы описывали диспетчер памяти гипервизора, который управляет физической и виртуальной памятью самого гипервизора. Гостевая физическая память виртуальной машины выделяется и управляется диспетчером памяти стека виртуализации. При запуске VM вновь созданный рабочий процесс VM (VMWP.exe) обращается к сервисам диспетчера памяти, определенным в виде COM-интерфейса `IMemoryManager`, для построения оперативной памяти гостя. Выделение памяти для виртуальной машины проходит в два этапа.

1. Рабочий процесс VM получает отчет о глобальном состоянии системной памяти (с помощью сервисов балансировщика памяти из процесса VMMS) и, исходя из доступной системной памяти, определяет размер физических блоков памяти, чтобы запросить их у драйвера VID (это делается с помощью `IOCTL VID_RESERVE`, размер блока варьируется от 64 Мбайт до 4 Гбайт). Драйвер VID выделяет блоки с помощью функций управления MDL, в частности `MmAllocatePartitionNodePagesForMdlEx`. Из соображений производительности и во избежание фрагментации памяти драйвер VID реализует алгоритм, стремящийся в первую очередь выделить громадные и большие физические страницы (1 Гбайт и 2 Мбайт), не полагаясь на стандартные маленькие страницы. После того как блоки памяти будут выделены, их страницы помещаются во внутренний резервный сегмент, обслуживаемый драйвером VID. Тот содержит списки страниц, отсортированные в массиве в зависимости от уровня качества обслуживания (quality of service, QOS). Этот уровень определяется в зависимости от типа страницы (громадные, большие и малые) и узла NUMA, которому она принадлежит. В номенклатуре VID этот процесс называется резервированием физической памяти (не путайте с термином «резервирование виртуальной памяти», который относится к диспетчеру памяти NT).
2. С точки зрения стека виртуализации *фиксация* физической памяти — это процесс очистки резервных страниц в сегменте и перенос их в блок памяти VID (структура данных `VSMM_MEMORY_BLOCK`), который создается рабочим процессом VM и остается в его распоряжении с помощью сервисов драйвера VM. В ходе создания блока памяти драйвер VID сначала помещает в гипервизор дополнительные физические страницы посредством драйвера Winhvr и гипервызова

`HvDepositMemory`. Эти страницы необходимы для формирования иерархии страниц для VM в таблице SLAT. Затем драйвер VID просит у гипервизора отразить физические страницы, описывающие всю оперативную память гостевого раздела. Гипервизор добавляет корректные записи в таблицу SLAT и назначает им соответствующие разрешения. Создается гостевое адресное пространство раздела. Этот диапазон GPA добавляется в список, принадлежащий разделу VID. Блок памяти VID принадлежит рабочему процессу VM. Кроме того, он используется для отслеживания гостевой памяти и задействуется в блоках памяти, относящихся к файлам DAX (подробности о томах DAX и PMEM см. в главе 11). Впоследствии рабочий процесс VM сможет использовать этот блок памяти для множества целей, например для доступа к некоторым страницам в ходе управления эмулируемыми устройствами.

Работа виртуальной машины

Запуск виртуальной машины в основном управляется VMMS и процессом VMWP. Когда служба VMMS получает запрос (через PowerShell или визуальное приложение Hyper-V Manager) на запуск виртуальной машины, идентифицируемой по GUID, она начинает с чтения конфигурации VM из репозитория хранилищ данных, где также хранятся GUID VM и список виртуальных устройств (VDEV), из которых состоит ее виртуальное оборудование. Затем проверяется, что путь, содержащий VHD (или VHDX), представляющий виртуальный жесткий диск виртуальной машины, имеет правильный список управления доступом (ACL, более подробная информация будет предоставлена далее). Если ACL окажется некорректным, то, если конфигурация VM позволяет, служба VMMS, действующая от лица пользователя SYSTEM, перезаписывает его новым, совместимым с новым экземпляром процесса VMWP. Для создания нового процесса VMWP служба VMMS связывается со службой вычислений хоста посредством служб COM.

Служба вычислений хоста определяет путь к рабочему процессу VM, запрашивая его регистрационные данные из реестра Windows с помощью раздела `HKCU\CLSID\{f33463e0-7d59-11d9-9916-0008744f51f3}`. Затем она создает новый процесс, используя надежный токен доступа, построенный с помощью SID виртуальной машины в роли владельца. В действительности учетная запись NT Authority в модели безопасности Windows определяет общепринятое значение уполномоченного центра (subauthority) (83) для идентификации виртуальных машин (больше информации о защитных компонентах системы имеется в главе 7 тома 1). Служба вычислений хоста дожидается, пока процесс WMWP не закончит свою инициализацию, в результате чего его COM-интерфейсы окажутся доступными. Управление возвращается в службу VMMS, которая наконец-то может запросить у процесса VMWP запуск виртуальной машины (через его COM-интерфейс `IVirtualMachine`).

Как показано на рис. 9.24, рабочий процесс VM реализует для виртуальной машины переход в состояние «холодный старт». В рамках рабочего процесса VM вся виртуальная машина управляется через сервисы, предоставляемые виртуальной материнской платой. Для виртуальных машин поколения 1 виртуальная материнская плата эмулирует плату Intel I440BX, а для машин поколения 2 — проприетарную материнскую плату. Она обслуживает и контролирует список виртуальных

устройств и выполняет переходы между состояниями для каждого из них. В следующем разделе будет более подробно рассмотрено, что каждое виртуальное устройство реализовано в виде COM-объекта (у которого открыт интерфейс `IVirtualDevice`) в DLL. Виртуальная материнская плата перечисляет все устройства из конфигурации виртуальной машины и загружает соответствующие COM-объекты для их представления.

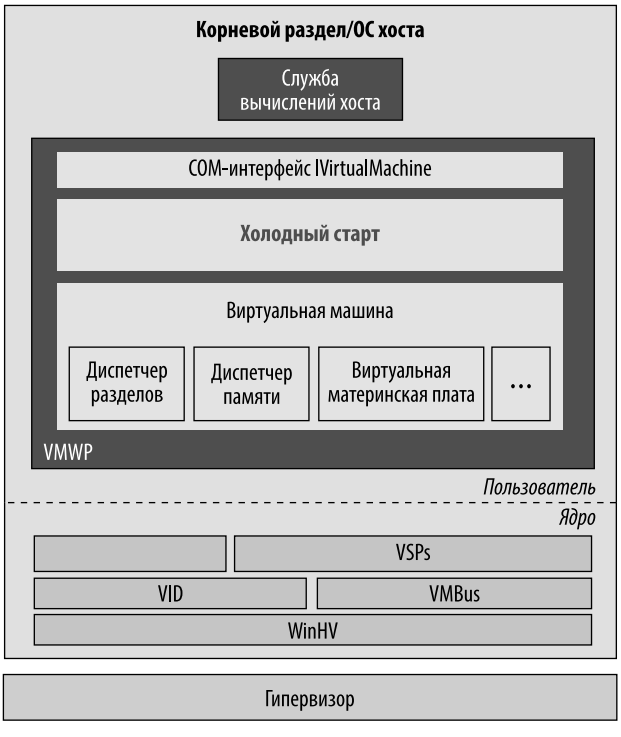


Рис. 9.24. Рабочий процесс VM и его интерфейс для холодного старта виртуальной машины

Рабочий процесс VM начинает процедуру запуска с резервирования ресурсов, необходимых для каждого из виртуальных устройств. Затем он формирует гостевое физическое адресное пространство VM (виртуальную оперативную память), выделяя физическую память из корневого раздела через драйвер VID. На данном этапе процесс уже может включить виртуальную материнскую память, которая циклически пройдется по всем VDEV и включит их. Процедура включения для каждого устройства своя: например, синтетические устройства для первоначальной настройки обычно взаимодействуют с собственным поставщиком служб виртуализации (Virtualization Service Provider, VSP).

Более глубокого обсуждения заслуживает виртуальное устройство, представляющее виртуальный BIOS (оно реализовано в библиотеке Vmchipset.dll). Методика его включения позволяет виртуальной машине дополнительно запустить встроенное ПО после запуска загрузочного VP. Виртуальное устройство BIOS извлекает

корректную для данной VM прошивку (для машин поколения 1 это обратно совместимый BIOS, в ином случае — UEFI) из раздела ресурсов собственной библиотеки, формирует переменную часть встроенного ПО (таблицы ACPI и SRAT), которую потом вводит в нужный участок гостевой памяти с помощью драйвера VID. Драйвер VID действительно способен отражать диапазоны памяти, описанные в блоке памяти VID, в память пользовательского режима, доступную рабочему процессу VM (внутреннее название данной процедуры — создание проема в память).

После того как все виртуальные устройства будут успешно запущены, рабочий процесс VM может запустить загрузочный VP виртуальной машины, отправив соответствующий IOCTL драйверу VID, который запустит этот виртуальный процессор и его канал сообщений (используется для обмена сообщениями между рабочим процессом VM и драйвером VID).

ЭКСПЕРИМЕНТ. Разбор защиты рабочего процесса VM и файлов виртуальных жестких дисков

В предыдущем разделе мы обсуждали, как рабочий процесс VM запускается службой вычислений хоста (Vmcompute.exe) после того, как в процесс VMMS поступил запрос о запуске виртуальной машины (через WMI). Прежде чем начать коммуникацию со службой вычислений, VMMS формирует токен безопасности для нового экземпляра рабочего процесса.

Для корректной поддержки виртуальных машин в модель защиты Windows были добавлены три новых элемента (модель безопасности Windows подробно рассматривалась в главе 7 тома 1):

- группа безопасности «Виртуальные машины» с идентификатором безопасности S-1-5-83-0;
- идентификатор безопасности виртуальной машины (SID), основанный на ее уникальном идентификаторе (GUID). SID VM становится владельцем токена безопасности, созданного для рабочего процесса VM;
- способность обеспечивать безопасность рабочего процесса VM используется, чтобы дать приложениям, действующим в рамках AppContainer, доступ к сервисам Hyper-V, требующимся рабочему процессу VM.

В ходе данного эксперимента вы с помощью Hyper-V Manager создадите виртуальную машину по адресу, который будет доступен только текущему пользователю и членам группы администраторов, после чего проверите, как изменится защита файлов виртуальной машины и самого рабочего процесса VM.

Первым делом откройте командную строку администратора и в одном из томов рабочей станции создайте папку (это будет `C:\TestVM`), используя следующую команду:

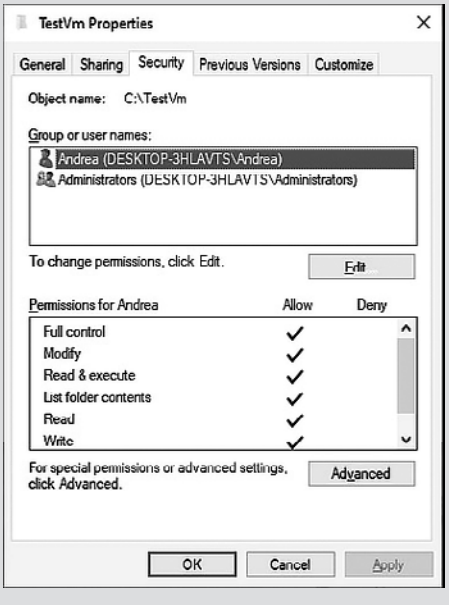
```
md c:\TestVm
```

Затем нужно будет снять с нее все унаследованные ACE (access control entries — записи контроля доступа; подробности см. в главе 7 тома 1) и задать возможность

получения полного доступа к ним для текущего пользователя и группы «Администраторы». Следующие команды выполнят описанные действия (при этом нужно будет заменить `C:\TestVM` путем к вашему каталогу, а вхождение *<имя_пользователя>* — именем текущего пользователя):

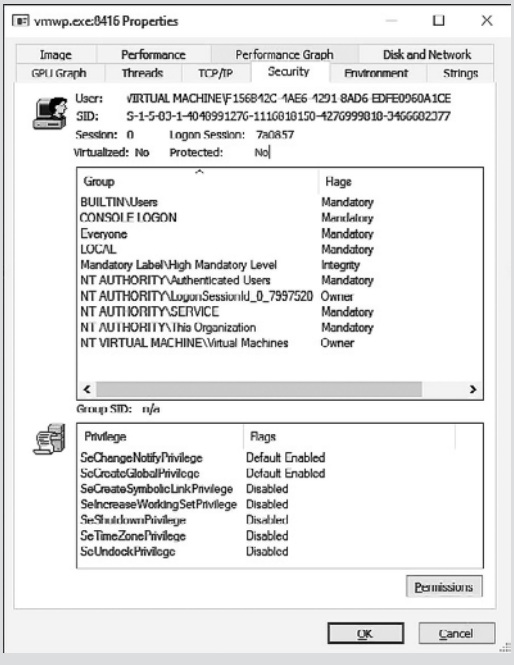
```
icacls c:\TestVm /inheritance:r
icacls c:\TestVm /grant Administrators:(CI)(OI)F
icacls c:\TestVm /grant <имя_пользователя>:(CI)(OI)F
```

Чтобы убедиться в правильности ACL у этой папки, откройте Проводник (нажмите на клавиатуре Win+E), щелкните на папке правой кнопкой мыши, выберите Свойства (Properties) и перейдите на вкладку Безопасность (Security). Отобразится окно наподобие приведенного на рисунке.



Откройте Hyper-V, создайте виртуальную машину (и виртуальный диск для нее) и поместите ее в созданную папку (процедура расписана по адресу: `https://docs.microsoft.com/en-us/virtualization/hyper-v-on-windows/quick-start/create-virtual-machine`). Для данного эксперимента нет необходимости ставить на нее операционную систему. Закончив общение с Мастером создания виртуальной машины, запустите VM (в данном примере она называется VM1).

Запустите Process Exporer с правами администратора и отыщите процесс vmwp.exe. Щелкните на нем правой кнопкой мыши и выберите Свойства (Properties). Как и ожидалось, его родительским процессом оказался vmcompute.exe (служба вычислений хоста). Перейдя на вкладку Безопасность (Security), вы должны будете увидеть, что владельцем процесса назначен SID VM, а сам токен принадлежит к группе Virtual Machines.



SID сформирован на основе GUID виртуальной машины. В приведенном примере это {F156B42C-4AE6-4291-8AD6-EDFE0960A1CE}. (Можете проверить это с помощью PowerShell, как демонстрировалось в эксперименте «Фокусы с корневым планировщиком» ранее в данной главе.) GUID представляет собой последовательность из 16 байт, разделенную на 32-битное целое значение (4 байта), два 16-битных (2 байта) целых значения и финальные 8 байт. GUID в этом примере организован так:

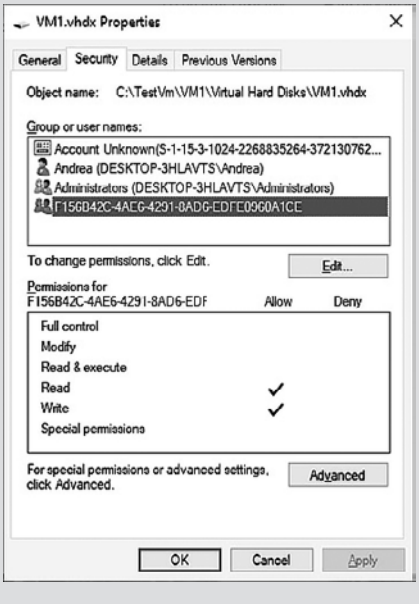
- 0xF156B42C — первое 32-битное целое, в десятичной системе — 4 048 991 276;
- 0x4AE6 и 0x4291 — два 16-битных целых, которые при объединении в одно 32-битное значение станут 0x42914AE6, или в десятичной системе 1 116 818 150 (напомним, что в этой системе формат данных Little Endian, а значит, менее значимый байт находится по меньшему адресу);
- финальная последовательность байтов — 0x8A, 0xD6, 0xED, 0xFE, 0x09, 0x60, 0xA1 и 0xCE (третья часть показанного читабельного GUID, 8AD6 — это уже последовательность байтов, а не 16-битное значение), которая, если представить ее двумя 32-битными значениями, будет 0xFEEDD68A и 0xCEA16009, или 4 276 999 818 и 3 466 682 377 в десятичной системе.

Если совместить все вычисленные десятичные значения с общим идентификатором SID, выданным NT Authority (S-1-5) и базовым RID виртуальной машины, у вас должен получиться тот же SID, который отобразился в Process Explorer, — в данном примере это S-1-5-83-4048991276-1116818150-4276999818-3466682377.

Как вы можете увидеть в Process Explorer, токен безопасности процесса VMWP не относится к группе администраторов и не был создан от лица текущего пользователя. Как же тогда рабочий процесс VM сможет работать с виртуальным жестким диском и файлами конфигурации виртуальной машины?

Ответ кроется в процессе VMMS, который в момент создания VM сканирует каждый компонент ее местонахождения и изменяет DACL необходимых файлов и папок. В частности, корневой каталог виртуальной машины (он имеет то же имя, что и сама VM, таким вы его и найдете в виде подкаталога в созданной вами папке) доступен благодаря добавлению ACE от группы виртуальных машин. А файл виртуального жесткого диска доступен за счет ACE, разрешающего доступ для SID этой виртуальной машины.

Можете проверить это с помощью Проводника: откройте папку с виртуальным жестким диском машины (находится в корневом каталоге VM и называется Виртуальные жесткие диски (Virtual Hard Disks)), щелкните правой кнопкой мыши на файле VHDX (или VHD), выберите Свойства (Properties) и перейдите на вкладку Безопасность (Security). В придачу к первоначальной вы найдете там две новые ACE: одну для виртуальной машины, другую — для обеспечения возможностей процесса VmWorker для Appcontainer.



Если вы приостановите виртуальную машину и попытаетесь удалить из файла ее ACE, то увидите, что она больше не способна запуститься. Чтобы восстановить на виртуальном жестком диске правильный ACL, можете воспользоваться скриптом для PowerShell, доступным по адресу https://gallery.technet.microsoft.com/Hyper-V-Restore-ACL-e64dee58.

VMBus

VMBus — это механизм, предоставляемый стеком виртуализации Hyper-V для обеспечения коммуникации между разделами виртуальных машин. Это устройство — виртуальная шина, которая устанавливает каналы связи между гостем и хостом. Эти каналы позволяют разделам делить общие данные и организовывать паравиртуализованные (известные также как синтетические) устройства.

Корневой раздел содержит поставщики служб виртуализации (Virtualization Service Providers, VSPs), которые общаются через VMBus с целью обработки запросов к устройствам со стороны дочерних разделов. Со своей стороны дочерние разделы (или гости) используют потребители служб виртуализации (Virtualization Service Consumers, VSCs), чтобы перенаправлять запросы к устройствам к VSP по VMBus. Дочерним разделам требуются драйверы VMBus и VSC для использования стеков паравиртуализированных устройств (подробности о поддержке виртуального оборудования приводятся в подразделе «Поддержка виртуального оборудования» далее). Каналы VMBus позволяют различным VCS и VSP передавать данные в основном через два кольцевых буфера: наверх (Upstream) и вниз (Downstream). Эти кольцевые буферы отображаются в обоих разделах благодаря гипервизору, который, как обсуждалось ранее, также предоставляет возможность связи между разделами через SynIC.

Одним из первых виртуальных устройств (VDEV), которые рабочий процесс запускает при включении виртуальной машины, является VDEV VMBus (реализовано в Vmbusr.dll). Его исполняемая при включении питания процедура подключает рабочий процесс виртуальной машины к корневому драйверу VMBus (Vmbusr.sys), отправляя `IOCTL VMBUS_VDEV_SETUP` корневому устройству VMBus (по имени `\Device\RootVmBus`). Корневой драйвер VMBus реализует родительскую конечную точку в рамках двунаправленной коммуникации с дочерней виртуальной машиной. Его процедура инициализации, запускаемая до того, как целевая VM будет включена, играет важную роль в создании структуры данных XPartition, которая используется для представления экземпляра для данной дочерней VM и подсоединения необходимых синтетических источников прерываний SynIC (также известных как SINT; подробности см. в подразделе «Синтетический контроллер прерываний» ранее в данной главе). В корневом разделе VMBus используют два синтетических источника прерываний: один для первоначального квитирующего сообщения (что происходит до создания канала), а другой — для синтетических событий, сигнализируемых кольцевыми буферами. Дочерние же разделы используют только один SINT. Процедура настройки выделяет главный порт сообщений в дочерней виртуальной машине и соответствующее соединение в корне, а для каждого виртуального процессора в рамках данной VM — порт событий и соединение с ним (используется для получения синтетических событий из дочерней VM).

Эти два синтетических источника прерываний отображаются с помощью двух функций ISR — `KiVmbusInterrupt0` и `KiVmbusInterrupt1`. Благодаря им корневой раздел готов принимать синтетические прерывания и сообщения от дочерней виртуальной машины. Когда получено сообщение (или событие), ISR помещает в очередь отложенный вызов процедур (DPC), в котором проверяется валидность сообщения. Если оно валидно, то в очередь добавляется элемент задания, который

позже будет обработан системой на пассивном уровне IRQL, что в дальнейшем дополнительно повлияет на очередь сообщений.

Когда VMBus в корневом разделе готов к работе, каждый драйвер VSP может воспользоваться клиентской библиотекой VMBus режима ядра (kernel mode client library, KMCL), чтобы выделить канал VMBus и предложить его дочерней VM. Эта библиотека представляет канал VMBus через непрозрачную структуру данных `KMODE_CLIENT_CONTEXT`, которая размещается и инициализируется во время создания канала, когда VSP вызывает функцию `VmbChannelAllocate`. Корневой VSP после этого, как правило, предлагает данный канал дочерней VM, вызывая функцию `VmbChannelEnabled` (она устанавливает в госте реальное подключение к корню путем открытия канала). KMCL реализована в виде двух драйверов: один действует в корневом разделе (Vmbkmclr.sys), а другой загружается в дочерних разделах (Vmbkmcl.sys).

В рамках корневого раздела предложение канала является довольно сложной операцией, включающей следующие шаги.

1. Драйвер KMCL обращается к драйверу VMBus корня через файловый объект, создаваемый в рамках процедуры включения VDEV. Драйвер VMBus находит для искомого дочернего раздела представляющую его структуру XPartition и начинает процесс предложения канала.
2. Низкоуровневые сервисы драйвера VMBus размещают и инициализируют структуру данных `LOCAL_OFFER`, представляющую отдельное предложение канала. Также они предварительно размещают несколько предопределенных сообщений SynIC. Затем VMBus создает в корневом разделе синтетический порт событий, через который дочерний сможет подключиться к сигнальным событиям после записи данных в кольцевой буфер. Структура данных `LOCAL_OFFER`, представляющая предложенный канал, добавляется во внутренний список каналов сервера.
3. Создав канал, VMBus делает попытку отправить в дочерний раздел сообщение `OfferChannel` с целью оповестить о новом канале. Однако на данном этапе она не даст результата, поскольку другая сторона (дочерняя VM) еще не готова и не начинала первоначальный обмен сообщениями.

После того как все VSP закончат готовить предложение канала, а все VDEV будут включены (подробности см. в предыдущем разделе), процесс VM Worker запускает виртуальную машину. Чтобы полностью завершить инициализацию каналов, а соответствующие им подключения начали работать, гостевому разделу потребуется загрузить и запустить дочерний драйвер VMBus (VMBus.sys).

Первичный обмен квитирующими сообщениями VMBus

В среде Windows драйвер VMBus дочернего раздела является шинным драйвером WDF, который определяется и запускается диспетчером PNP и находится в ведении корневого перечислителя ACPI. (Существует и отдельная версия дочернего драйвера VMBus для Linux. Впрочем, ее описание в тематику данной книги не входит.) Когда на дочерней VM запускается ядро NT, драйвер VMBus начинает работу с инициализации собственного внутреннего состояния (а именно, размещения необходимых структур данных и заданий) и создания корневого функционального объекта устройства `\Device\VMBus` (functional device object, FDO). Затем диспетчер PNP

запускает у VMBus процедуру обработчика распределения ресурсов. Та, в свою очередь, устанавливает корректный источник SINT, выполняя с помощью драйвера WinHv гипервызов `HvSetVpRegisters` на одном из регистров `HvRegisterSint`, и подключает его к ISR `KiVMBusInterrupt2`. Кроме того, она получает страницу SIMP, используемую для отправки синтетических сообщений в корневой раздел и их получения из него (см. подробности в подразделе «Синтетический контроллер прерываний» ранее в данной главе), и создает структуру данных XPartition, представляющую родительский (корневой) раздел.

Когда от диспетчера Pnp поступает запрос на запуск FDO от VMBus, драйвер последнего приступает к первичному обмену квитирующими сообщениями. На данном этапе каждое сообщение отправляется посредством гипервызова `HvPostMessage` (с помощью драйвера WinHv), что позволяет гипервизору внедрить синтетическое прерывание в целевой раздел (в данном случае раздел и есть цель). Получатель может прочесть сообщение, просто обратившись к странице SIMP. О том, что сообщение из очереди было прочитано, он сигнализирует, устанавливая новый тип сообщения MessageTypeNone (подробнее — в документации к гипервизору в TLFS). Читатель может представить себе первичный обмен квитирующими сообщениями, приведенный на рис. 9.25, как процесс, разделяемый на две фазы.

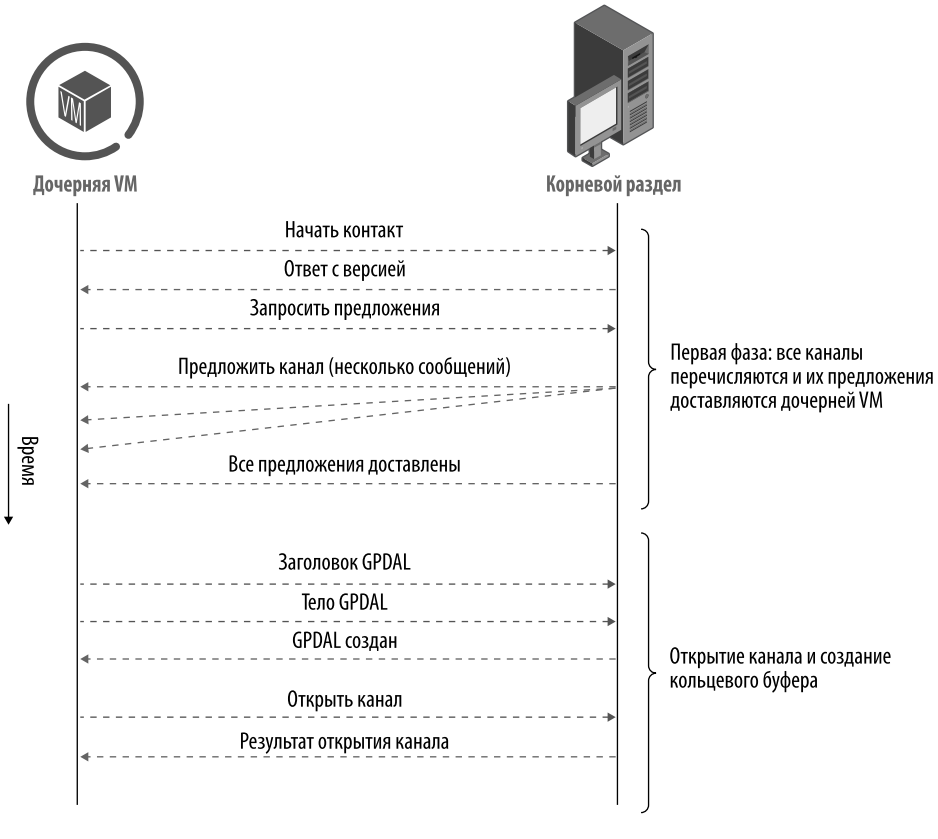


Рис. 9.25. Первичный обмен квитирующими сообщениями VMBus

Первая фаза обозначается сообщением `Initiate Contact` («Начать контакт»), доставляемым единожды за весь сеанс работы виртуальной машины. Оно отправляется с дочерней виртуальной машины в корень с целью договориться о версии протокола VMBus, поддерживаемой обеими сторонами. На время написания данного текста существует пять основных версий протокола VMBus, имеющих ряд незначительных вариаций. Корневой раздел анализирует сообщение, просит гипервизор отобразить контрольные страницы, размещенные клиентом (если поддерживается данным протоколом), и отвечает принятием предложенной версии протокола. Заметим, что в другой ситуации (когда версия Windows из корневого раздела старше, чем действующая на дочерней виртуальной машине) дочерняя виртуальная машина перезапускает процесс со снижением версии протокола VMBUS до тех пор, пока не будет установлена версия, поддерживаемая обеими сторонами. В этот момент раздел готов отправить сообщение `Request Offers`, что приводит к тому, что в корневой раздел отправляется список всех каналов, уже предложенных VSP. Это позволяет дочернему разделу в дальнейшем открыть каналы в рамках протокола первоначального обмена.

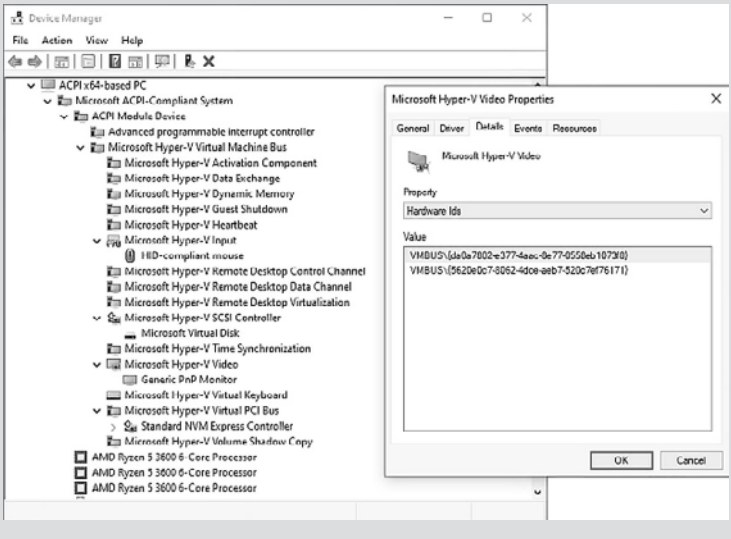
На рис. 9.25 приводится ряд различных синтетических сообщений для настройки канала или каналов, расположенных в списке каналов сервера (структура данных `LOCAL_OFFER`, упомянутая ранее), и для каждого из них отправляет сообщение о предложении канала в дочернюю виртуальную машину. Это то же самое сообщение, которое отправляется на финальном этапе протокола предложения канала, о чем говорилось в подразделе «VMBus». Таким образом, притом что первая фаза первоначального обмена квитирующими сообщениями происходит лишь один раз за все время работы виртуальной машины, вторая фаза может начаться в любой момент после предложения канала. Сообщение `Offer Channel` несет в себе важные данные, применяемые для его уникальной идентификации, в частности GUID экземпляра и тип канала. Для каналов VDEV эти два GUID используются диспетчером Pnp для корректной идентификации связанного виртуального устройства.

Дочерний раздел отвечает на это сообщение размещением клиентской структуры данных `LOCAL_OFFER`, представляющей канал и связанный с ним объект XInterrupt, и определением того, требуется ли каналу создание объекта физического устройства, что для каналов VDEV обычно верно. В таком случае драйвер VMBus создает экземпляр объекта физического устройства, представляющий новый канал. Созданное устройство защищено посредством дескриптора защиты, который делает его доступным исключительно для учетных записей системы и администраторов. Стандартный интерфейс устройства VMBus, который связали с новым объектом, сохраняет связь между новым каналом VMBus через структуру данных `LOCAL_OFFER` и объектом устройства. После создания физического объекта диспетчер Pnp получает возможность идентифицировать и загрузить правильный драйвер за счет GUID типа VDEV и экземпляра в составе сообщения `Offer Channel`. Эти интерфейсы становятся частью нового PDO и отображаются в диспетчере устройств. Подробности приведены в следующем эксперименте. Когда после этого загрузится драйвер VSC, он обычно вызывает функцию `VmbEnableChannel` (публикуемую KMCL, как обсуждалось ранее), чтобы открыть канал и создать конечный кольцевой буфер.

ЭКСПЕРИМЕНТ. Вывод списка виртуальных устройств (VDEV), отображаемых из VMBus

Каждый канал VMBus идентифицируется через GUID экземпляра и типа. Для каналов, принадлежащих VDEV, GUID экземпляра и типа также идентифицирует отображаемое устройство. Когда дочерний драйвер VMBus создает физические объекты экземпляра, он добавляет GUID экземпляра в ряд свойств устройства, такие как путь к экземпляру, идентификаторы оборудования и совместимости. В данном эксперименте демонстрируется, как перечислить все VDEV, сформированные на основе VMBus.

Для данного эксперимента вам потребуется создать и запустить виртуальную машину с операционной системой Windows 10 посредством Hyper-V Manager. Когда она запущена и работает, откройте диспетчер устройств, например набрав его имя в поисковом окне Cortana. В Диспетчере устройств щелкните на пункте меню Вид (View) и выберите Устройства по подключению (Device by Connection). Драйвер VMBus указан и запущен через перечислитель acpi, поэтому вы сможете раскрыть корневой узел ACPI для x64, а затем узел модульного устройства в дочернем узле Microsoft, как показано на следующем рисунке.



Открыв модульное устройство ACPI, вы должны будете увидеть другой узел, а именно шину виртуальной машины Microsoft Hyper-V, который представляет корневой объект физического устройства VMBus. Под этим узлом Диспетчер устройств отобразит все экземпляры устройств, созданные PDO VMBus после того, как соответствующие им каналы VMBus были предложены корневым разделам.

Теперь правой кнопкой мыши щелкните на одном из устройств Hyper-V, таком как устройство видео Microsoft Hyper-V, и выберите Свойства (Properties).

Чтобы отобразить GUID экземпляра и типа канала VMBus, обеспечивающего это виртуальное устройство, откройте вкладку Сведения (Details) окна Свойства (Properties). Среди трех показанных устройств будут GUID экземпляра и типа канала, показанные в различных форматах: путь к экземпляру устройства, идентификаторы оборудования и совместимости. Хотя в идентификаторе оборудования находится только GUID типа канала VMBus (на рисунке это {da0a7802-e377-4aac-8e77-0558eb1073f8}), идентификатор оборудования и путь к экземпляру устройства содержат GUID и экземпляра, и типа.

Открытие канала VMBus и создание кольцевого буфера

Для корректного начала коммуникации между разделами и созданием кольцевого буфера необходимо открыть канал. Обычно VSC после размещения клиентской стороны канала (все еще через размещение `VmbChannelAllocate`) вызывают функцию `VmbChannelEnable`, публикуемую драйвером KMCL. Как говорилось в предыдущем разделе, в дочерних разделах данная функция открывает канал VMBus, который уже был предложен корнем. Драйвер KMCL взаимодействует с драйвером VMBus, получает параметры канала, такие как тип канала, GUID экземпляра и используемое пространство MMIO, и формирует из полученных пакетов рабочую запись. После этого он размещает кольцевой буфер (рис. 9.26). Размер кольцевого буфера обычно определяется со стороны VSC посредством вызова функции `VmbClientChannelInitSetRingBufferPageCount`, публикуемой KMCL.

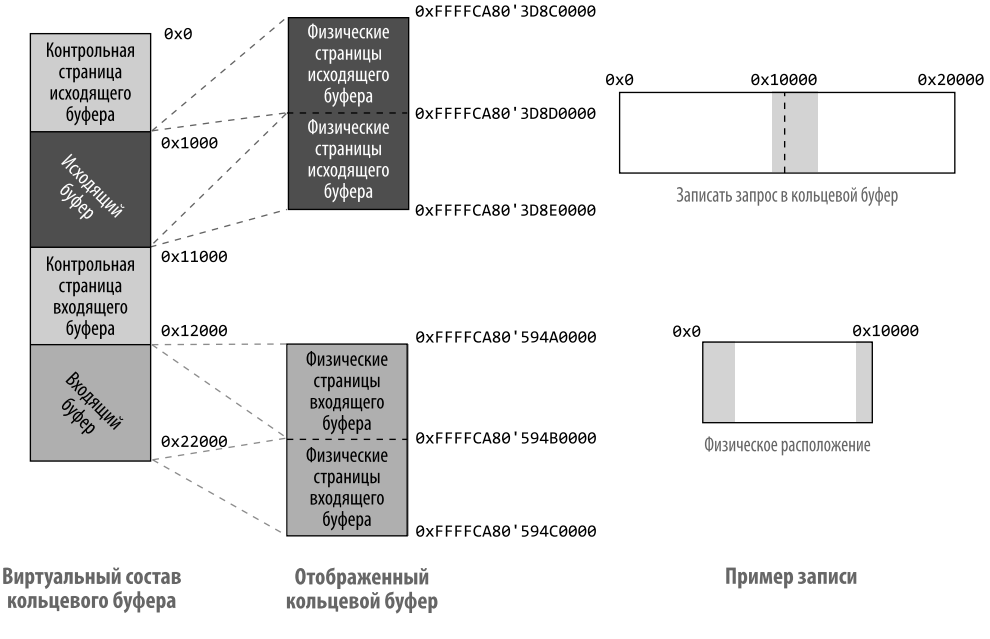


Рис. 9.26. Пример 16-страничного кольцевого буфера, размещенного в дочернем разделе

Кольцевой буфер размещается со стороны неподкачиваемого пула дочерней виртуальной машины и отображается через список дескрипторов памяти (MDL) с использованием техники, известной как *двойное отображение* (double mapping) (MDL описаны в главе 5 тома 1). Согласно этой технике список размещенных дескрипторов описывает двойное количество физических страниц входящего или исходящего буфера. Массив PFN в рамках MDL заполняется посредством включения физических страниц из буфера дважды: первый раз — в первую половину массива, второй раз — во вторую. Так создается кольцевой буфер.

Например, на рис. 9.26 входящий и исходящий буферы имеют по 16 страниц (0x10). Входящий буфер отображается по адресу 0xFFFFCA803D8C0000. Если отправитель посылает по VMBus пакет размером 1 Кбайт, то на позиции, близкой к концу буфера, скажем, по смещению 0x9FF00, запись проходит успешно (не возникает ошибки защиты), но данные будут записаны частично в конце буфера, а частично — в начале. На рис. 9.26 только 256 байт записаны в конец буфера, остальные 768 байт — в начале.

Как входящие, так и исходящие буферы заключены в контрольную страницу. Она разделена между двумя конечными точками и представляет собой контрольный блок кольца виртуальной машины. Эта структура данных используется для отслеживания положения последнего пакета, записанного в кольцевой буфер. Кроме того, она содержит несколько битов для контроля того, следует ли направить прерывание, когда пакет необходимо доставить.

После того как кольцевой буфер создан, драйвер KMCL отправляет IOCTL к VMBus, тем самым запрашивая создание списка дескрипторов GPA (GPADL). GPADL — это структура данных, очень похожая на MDL, она применяется для описания фрагмента физической памяти. Отличие заключается в том, что GPADL содержит массив гостевых физических адресов (GPA, которые всегда представлены 64-битными значениями, в отличие от используемых в MDL PFN). Драйвер VMBus отправляет в корневой раздел различные сообщения для переноса GPADL, формируя описание как входящего, так и исходящего кольцевых буферов. (Максимальный размер синтетического сообщения 240 байт, как говорилось ранее.) Корневой раздел полностью реконструирует GPADL и сохраняет его в своем внутреннем списке. Отображение в корневой раздел происходит, когда дочерний раздел отправляет завершающее сообщение `Open Channel`. Корневой драйвер VMBus анализирует полученный GPADL и отображает его в собственное адресное пространство, используя возможности, предоставляемые драйвером VID (он обслуживает список диапазонов блоков памяти, из которых состоит физическое адресное пространство виртуальной машины).

На данном этапе канал уже готов: корневой и дочерний разделы могут взаимодействовать, просто считывая или записывая данные в кольцевом буфере. Когда отправитель заканчивает запись данных, он вызывает функцию `VmbChannelSendSynchronousRequest`, публикуемую драйвером KMCL. Та обращается к сервисам VMBus, чтобы сигнализировать события на контрольной странице объекта XInterrupt, связанного с каналом (устаревшие версии протокола VMBus использовали страницу прерывания, которая содержала бит, отвечающий за каждый канал). В качестве альтернативы VMBus может сигнализировать события напрямую через событийный порт канала, что зависит лишь от времени задержки.

VMBus задействуется не только VSC, но и другими компонентами для реализации высокоуровневых интерфейсов. Хорошим примером являются каналы VMBus, которые реализованы в двух библиотеках режима ядра — VMBuspipe.dll и VMBuspiper.dl, опирающихся на сервисы, предоставляемые драйвером VMBus (через IOCTL). Сокеты Hyper-V, также известные как HvSockets, обеспечивают высокоскоростное взаимодействие с помощью стандартных сетевых интерфейсов (сокетов). Клиент подключает Socket типа `AF_HYPERV` к искомой виртуальной машине, указывая цель с помощью ее GUID и GUID регистрации в службе сокетов Hyper-V (для использования HvSockets обе конечные точки должны быть зарегистрированы в реестре по адресу `HKLM\SOFTWARE\Microsoft\Windows NT\CurrentVersion\Virtualization\GuestCommunicationServices`), вместо того чтобы задействовать для этого IP-адрес и порт. Сокеты Hyper-V реализованы в нескольких драйверах: HvSocket.sys — это драйвер транспорта, предоставляющий низкоуровневые службы, используемые инфраструктурой сокетов; HvSocketControl.sys — это драйвер управления поставщиком, используемый для загрузки поставщика HvSocket в случае, если интерфейс VMBus отсутствует в системе; HvSocket.dll — это библиотека, предоставляющая дополнительные интерфейсы сокетов (привязанные к сокетам Hyper-V), вызываемые из приложений пользовательского режима. Описание внутренней инфраструктуры как каналов VMBus, так и сокетов Hyper-V не относится к темам данной книги, но обе технологии документированы в Microsoft Docs.

Поддержка виртуального оборудования

Для правильной работы виртуальных машин стек виртуализации должен поддерживать виртуализованные устройства. Hyper-V поддерживает различные типы виртуальных устройств, реализуемые в ряде компонентов стека виртуализации. Ввод на виртуальные устройства и вывод с них организуется в основном в корневой ОС. В ее ведении находятся запоминающие устройства, сеть, клавиатура, мышь, последовательные порты и графический процессор (graphics processing unit, GPU). Стек виртуализации предоставляет гостевым виртуальным машинам три типа устройств:

- эмулируемые устройства, также известные (в ставшей отраслевым стандартом форме) как полностью виртуализованные устройства;
- синтетические, или паравиртуализованные, устройства;
- устройства с аппаратным ускорением, называемые также устройствами прямого доступа.

Для выполнения ввода на физические устройства и вывода с них процессор обычно считывает данные из портов вывода и записывает на порты ввода, принадлежащие устройству (в целом — порты ввода-вывода). Получить доступ к портам ввода-вывода он может двумя способами:

- через отдельное адресное пространство ввода-вывода, которое отличается от физического адресного пространства и на платформах AMD64 состоит из 64 тыс. индивидуально адресуемых портов ввода-вывода. Этот метод старый и обычно используется для устаревших устройств;

- через отображенный в памяти ввод-вывод. Доступ к устройствам, которые реагируют как компоненты памяти, можно получить через физическое адресное пространство процессора. Это означает, что процессор обеспечивает доступ к памяти посредством стандартных инструкций — базовая физическая память отображается на устройство.

На рис. 9.27 приведен пример эмулируемого устройства (виртуального контроллера IDE, используемого в виртуальных машинах первого поколения), которое применяет ввод-вывод с отражением в памяти для передачи данных в виртуальный процессор и получения их обратно.

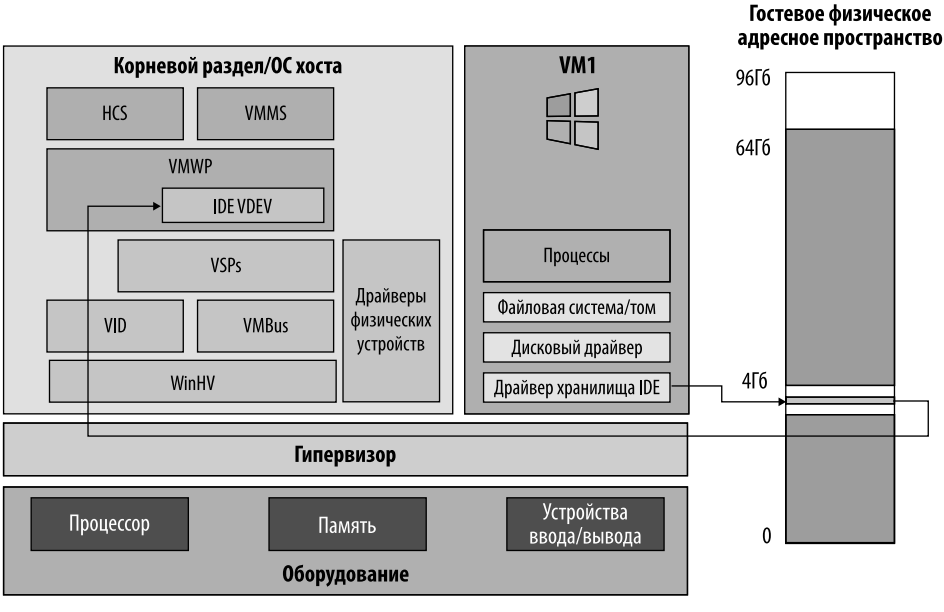


Рис. 9.27. Виртуальный контроллер IDE, который использует эмулированный ввод-вывод для передачи данных

В этой модели каждый раз, когда виртуальный процессор читает из пространства MMIO устройства, или записывает в него, или выдает инструкции для доступа к портам ввода-вывода, он вызывает VMEXIT для гипервизора. Гипервизор вызывает соответствующую процедуру перехвата, которая передается драйверу VID. Тот создает сообщение VID и помещает его во внутреннюю очередь. Потребителем этой очереди является внутренний поток VMWP, который ожидает и отправляет сообщения виртуального процесса, полученные от драйвера VID. Этот поток называется *потоком подкачки сообщений* (message pump) и принадлежит внутреннему пулу потоков, инициализированному во время создания VMWP. Рабочий процесс VM определяет физический адрес, где вызывается событие VMEXIT, связанное с соответствующим виртуальным устройством (VDEV), и выполняет один из обратных вызовов VDEV (обычно чтения или записи). Код VDEV использует сервисы, предоставляемые эмулятором инструкций, для выполнения сбойной инструкции с целью правильной эмуляции виртуального устройства (в примере — IDE-контроллера).

ПРИМЕЧАНИЕ Эмулятор полных инструкций, расположенный в рабочем процессе VM, используется и для других целей, например для ускорения выполнения кода с интенсивным перехватом в дочернем разделе. Эмулятор в этом случае позволяет контексту выполнения оставаться в рабочем процессе между перехватами, поскольку VMEXIT приводит к серьезным издержкам производительности. Более старые версии расширений аппаратной виртуализации запрещают выполнение кода реального режима на виртуальной машине, в этих случаях стек виртуализации задействовал эмулятор для выполнения кода реального режима на виртуальной машине.

Паравиртуализованные устройства

Эмулируемые устройства всегда создают события VMEXIT и работают довольно медленно, а мы рассмотрим пример синтетического, или паравиртуализованного, устройства — синтетического адаптера хранилища (рис. 9.28). Синтетические устройства умеют работать в виртуализованной среде — это снижает сложность виртуального устройства и позволяет ему достичь более высокой производительности. Некоторые синтетические виртуальные устройства существуют только в виртуальной форме и не эмулируют какое-либо реальное физическое оборудование, пример — синтетический RDP.

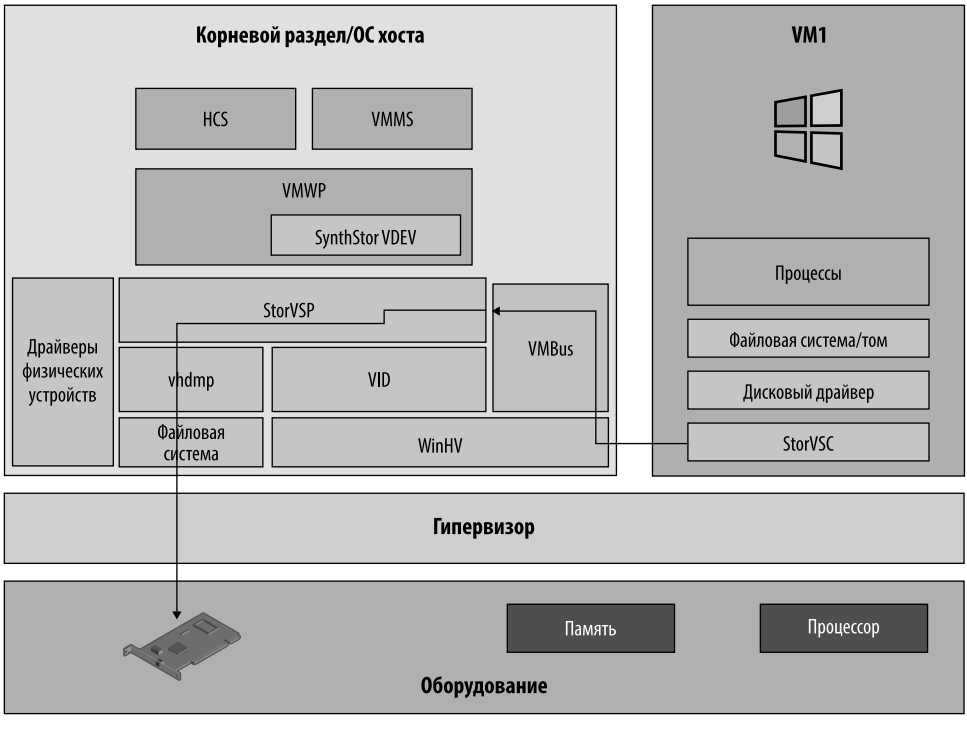


Рис. 9.28. Паравиртуализованное устройство контроллера хранилища

Паравиртуализованным устройствам обычно требуются три основных компонента.

- Драйвер поставщика служб виртуализации (VSP) действует в корневом разделе и дает гостю доступ к специфичным интерфейсам виртуализации благодаря сервисам, предоставляемым VMBus (подробнее об VMBus — в подразделе ранее).
- Синтетический VDEV отображается в рабочем процессе VM и обычно участвует только в запуске, отсоединении, сохранении и восстановлении виртуального устройства. Как правило, в штатной работе устройства это не используется. Синтетический VDEV инициализирует и распределяет ресурсы для конкретного устройства (в примере SynthStor VDEV инициализирует адаптер виртуального хранилища), но, что наиболее важно, позволяет VSP предлагать канал связи VMBus гостевому VSC. Канал будет применяться для связи с корнем и сигнализации уведомлений, специфичных для устройства, через гипервизор.
- Драйвер потребителя службы виртуализации (VSC) действует в дочернем разделе. Он способен работать со специфичными для виртуализации интерфейсами, предоставляемыми VSP, и читает сообщения и уведомления из общей памяти, доступной VSP через VMBus, а также записывает в нее. Это позволяет виртуальному устройству работать на дочерней виртуальной машине быстрее, чем эмулируемому устройству.

Устройства с аппаратным ускорением

На серверных SKU *устройства с аппаратным ускорением* (известны также как устройства прямого доступа) позволяют переназначать физические устройства в гостевом разделе благодаря службам инфраструктуры VPCI. Если физическое устройство поддерживает такие технологии, как однокорневая виртуализация ввода-вывода (SR IOV) или дискретное назначение устройств (DDA), оно может быть отражено в гостевом разделе. Этот раздел может напрямую обращаться к пространству MMIO, связанному с устройством, и осуществлять прямой доступ (DMA) к гостевой памяти без перехвата со стороны гипервизора. IOMMU обеспечивает необходимую безопасность и гарантирует, что устройство может инициировать выполнение DMA только в физической памяти, принадлежащей виртуальной машине.

На рис. 9.29 показаны компоненты, отвечающие за управление устройствами с аппаратным ускорением.

- VPci VDEV (Vpcievdev.dll) запускается в рабочем процессе виртуальной машины. Его обязанности — извлечь список устройств с аппаратным ускорением из файла конфигурации виртуальной машины, настроить виртуальную шину VPCI и назначить устройство VSP.
- Прокси-драйвер PCI (Pcip.sys) отвечает за отсоединение и подсоединение DDA-совместимого физического устройства из корневого раздела. Кроме того, он играет ключевую роль в получении списка ресурсов, используемых устройством (через протокол SR-IOV), таких как пространство MMIO и прерывания. Драйвер обеспечивает доступ к пространству физической конфигурации устройства и делает несмонтированное устройство недоступным для ОС хоста.

- Поставщик виртуальных служб VPCI (Vpcivsp.sys) создает и поддерживает объект виртуальной шины, который связан с одним или несколькими устройствами с аппаратным ускорением (в VPCI VSP они называются *виртуальными устройствами*). Виртуальные устройства доступны гостевой виртуальной машине через канал VMBus, созданный VSP и предлагаемый VSC в гостевом разделе.
- Клиент виртуальной службы VPCI (Vpci.sys) — это шинный драйвер WDF, который работает на гостевой виртуальной машине. Он подключается к каналу VMBus, предоставляемому VSP, получает список устройств прямого доступа, доступных виртуальной машине, и их ресурсов и создает PDO (объект физического устройства) для каждого из них. Далее драйвер устройства сможет подключаться к созданным объектам PDO так же, как и в невиртуализованных средах.

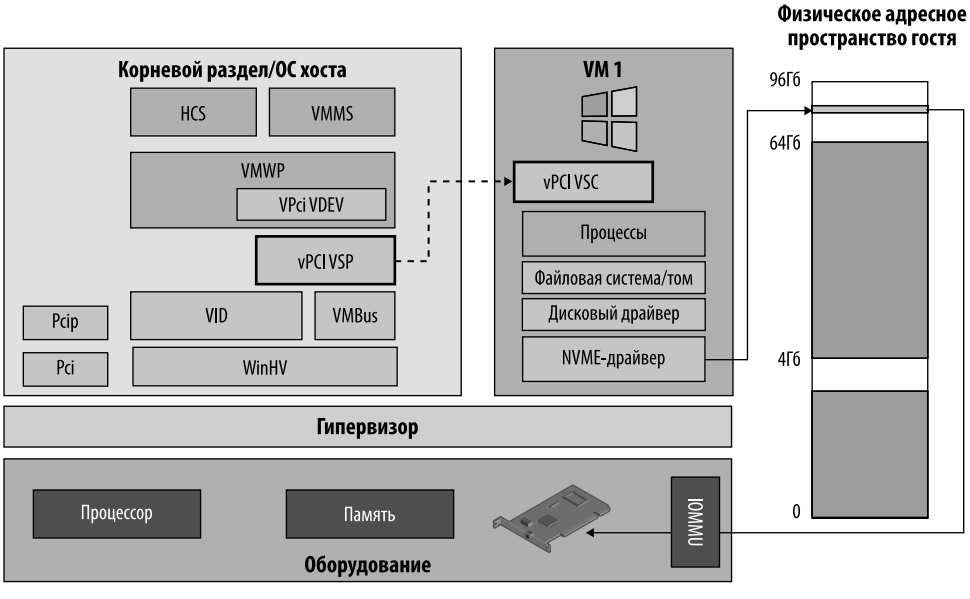


Рис. 9.29. Устройство с аппаратным ускорением

Когда пользователь хочет отобразить устройство с аппаратным ускорением в виртуальной машине, он применяет некоторые команды PowerShell (подробности — в следующем эксперименте), которые начинаются с отсоединения устройства от корневого раздела. Это действие заставляет службу VMMS взаимодействовать со стандартным драйвером PCI (через публикуемое им устройство, называемое *PciControl*). Служба VMMS отправляет IOCTL `PCIDRIVE_ADD_VMPROXYPATH` драйверу PCI, предоставляя дескриптор устройства в форме идентификаторов шины, устройства и функции. Драйвер PCI проверяет дескриптор и, если проверка прошла успешно, добавляет его в значение реестра `HKLM\System\CurrentControlSet\Control\PnP\Pci\VmProxy`. Затем VMMS запускает (иногда повторно) перечисление устройств PNP, используя службы, предоставляемые диспетчером PNP. На этапе перечисления драйвер PCI находит новое прокси-устройство и загружает прокси-драйвер PCI (Pcip.sys), который

помечает устройство как зарезервированное для стека виртуализации и делает его невидимым для операционной системы хоста.

Второй шаг требует назначения устройства виртуальной машине. В этом случае VMMS записывает дескриптор устройства в файл конфигурации виртуальной машины. При запуске виртуальной машины VPCI VDEV (vpcievdev.dll) считывает дескриптор устройства прямого доступа из конфигурации виртуальной машины и запускает сложную фазу настройки, которая управляется в основном VPCI VSP (Vpcivsp.sys). Действительно, во время обратного вызова при включении питания VPCI VDEV отправляет различные IOCTL на VPCI VSP, который работает в корневом разделе, с целью создания виртуальной шины и назначения устройств с аппаратным ускорением для гостевой виртуальной машины.

Виртуальная шина — это структура данных, используемая инфраструктурой VPCI в качестве связующего средства для поддержания соединения между корневым разделом, гостевой виртуальной машиной и назначенными ей устройствами прямого доступа. VPCI VSP выделяет и запускает канал VMBus, предлагаемый гостевой виртуальной машине, и инкапсулирует его в виртуальную шину. Кроме того, виртуальная шина включает в себя указатели на важные структуры данных, такие как некоторые выделенные пакеты VMBus, используемые для двунаправленной связи, определения состояния питания гостя и т. д. После создания виртуальной шины VPCI VSP выполняет назначение устройства.

Устройство с аппаратным ускорением идентифицируется LUID и представлено объектом виртуального устройства, который выделяется VPCI VSP. На основании LUID устройства VPCI VSP находит подходящий прокси-драйвер, который также известен как драйвер Mux (обычно это Pcip.sys). VPCI VSP запрашивает интерфейсы SR-IOV или DDA у прокси-драйвера и использует их для получения информации Plug and Play (дескриптора оборудования) устройства прямого доступа и сбора требований к ресурсам (пространство MMIO, регистры BAR и каналы DMA). На этом этапе устройство готово к подключению к гостевой виртуальной машине: VPCI VSP использует сервисы, предоставляемые драйвером WinHvr, для отправки гипервызова `HvAttachDevice` гипервизору, который перенастраивает системный IOMMU для отображения адресного пространства устройства в гостевом разделе.

Гостевая виртуальная машина знает о подключенном устройстве благодаря VPCI VSC (Vpci.sys). VPCI VSC — это шинный драйвер WDF, перечисляемый и запускаемый драйвером шины VMBus, расположенным в гостевой виртуальной машине. Он состоит из двух основных компонентов: FDO (объект функционального устройства), созданного во время загрузки виртуальной машины, и одного или нескольких PDO (объектов физических устройств), представляющих физические устройства прямого доступа, переназначенные в гостевой виртуальной машине. Когда драйвер шины VPCI VSC выполняется в гостевой виртуальной машине, он создает и запускает клиентскую часть канала VMBus, задействуемого для обмена сообщениями с VSP. Первым сообщением, отправляемым VPCI VSC по каналу VMBus, является `Send bus relations` (Отправить связи шины). VSP в корневом разделе отвечает на это, отправляя список идентификаторов оборудования, описывающих устройства с аппаратным ускорением, подключенные в данный момент к виртуальной машине. Когда диспетчер PNP требует связи новых устройств с VPCI VSC, последний создает новый PDO для каждого обнаруженного устройства прямого доступа. Драйвер VSC отправляет VSP еще одно сообщение с целью запроса ресурсов, используемых PDO.

После завершения первоначальной настройки VSC и VSP редко участвуют в управлении устройством. Драйвер конкретного устройства с аппаратным ускорением в гостевой виртуальной машине подключается к соответствующему PDO и управляет периферийным устройством так, как если бы оно было установлено на физической машине.

ЭКСПЕРИМЕНТ. Присоединение диска NVMe с аппаратным ускорением к виртуальной машине

Как объяснялось в предыдущем разделе, физические устройства, поддерживающие технологии SR-IOV и DDE, могут быть напрямую сопоставлены с гостевой виртуальной машиной, работающей на хосте Windows Server 2019. В этом эксперименте мы сопоставляем диск NVMe, который подключен к системе через шину PCI-Ex и поддерживает DDE, с виртуальной машиной Windows 10. (Windows Server 2019 поддерживает также прямое назначение видеокарты, но это выходит за рамки данного эксперимента.)

Как объясняется на https://docs.microsoft.com/en-us/virtualization/community/team-blog/2015/20151120-discrete-device-assignment-machines-and-devices, чтобы иметь возможность переназначения, устройство должно обладать определенными характеристиками, такими как поддержка прерываний, сигнализируемых сообщениями, и операций ввода-вывода, отображаемых в памяти. Более того, компьютер, на котором работает гипервизор, должен поддерживать SR-IOV и иметь соответствующий MMU ввода-вывода. В этом эксперименте следует начать с проверки того, что стандарт SR-IOV включен в BIOS системы (здесь не объясняется — процедура зависит от производителя вашего компьютера).

Следующий шаг — загрузка сценария PowerShell, который проверяет, совместим ли ваш контроллер NVMe с дискретным назначением устройств. Следует загрузить сценарий PowerShell под название Survey-dda.ps1 с https://github.com/MicrosoftDocs/Virtualization-Documentation/tree/master/hyperv-samples/benarm-powershell/DDA. Откройте административное окно PowerShell, введя `PowerShell` в поле поиска Cortana и выбрав Запуск от имени администратора (Run As Administrator), и проверьте, настроена ли политика реализации сценариев PowerShell на неограниченное применение, выполнив команду Get-ExecutionPolicy. Если она дает результат, отличный от Неограниченный (Unrestricted), следует ввести `Set-ExecutionPolicy -Scope LocalMachine -ExecutionPolicy Unrestricted`, нажать Enter и подтвердить нажатием Y.

Если выполняется загруженный сценарий Survey-dda.ps1, его выходные данные должны указать, можно ли переназначить ваше устройство NVMe гостевой виртуальной машине. Далее приводится корректный пример вывода:

```
Standard NVM Express Controller
Express Endpoint -- more secure.
    And its interrupts are message-based, assignment can work.
PCIROOT(0)#PCI(0302)#PCI(0000)
```

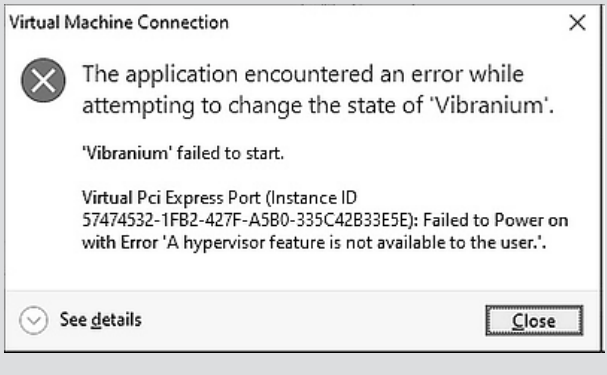
Обратите внимание на путь к местоположению (в примере — строка `PCIROOT(0)#PCI(0302)#PCI(0000)`). Теперь мы установим действие автоматической остановки для целевой виртуальной машины как выключенное (обязательный шаг для DDA) и отключим устройство. В нашем примере виртуальная машина называется Vibranium. Напишите в окне PowerShell следующие команды, заменив примерное имя виртуальной машины и местоположение устройства собственными:

```
Set-VM -Name "Vibranium" -AutomaticStopAction TurnOff
Dismount-VMHostAssignableDevice -LocationPath "PCIROOT(0)#PCI(0302)#PCI(0000)"
```

Если последняя команда выдает ошибку сбоя операции, вполне вероятно, что вы не отключили устройство. Откройте диспетчер устройств, найдите свой контроллер NVMe (в данном примере это стандартный контроллер NVMe Express), щелкните на нем правой кнопкой мыши и выберите Отключить устройство (Disable Device). Затем можете снова ввести последнюю команду. На этот раз все должно получиться. Затем назначьте устройство своей виртуальной машине, набрав следующее:

```
Add-VMAssignableDevice -LocationPath "PCIROOT(0)#PCI(0302)#PCI(0000)" -VMName
"Vibranium"
```

Последняя команда должна была полностью удалить контроллер NVMe с хоста. Вы должны убедиться в этом, проверив диспетчер устройств в хост-системе. Теперь пришло время включить виртуальную машину. Можете использовать инструмент Hyper-V Manager или PowerShell. Если вы запускаете виртуальную машину и получаете ошибку, подобную приведенной на рисунке, ваш BIOS неправильно настроен для предоставления SR-IOV или MMU ввода-вывода не имеет нужных характеристик (скорее всего, не поддерживает переназначения ввода-вывода).



В противном случае виртуальная машина должна просто загрузиться, как и ожидалось. В этом случае вы сможете увидеть как контроллер NVMe, так и диск NVMe, указанные в апплете диспетчера устройств дочерней виртуальной

машины. Можете использовать инструмент управления дисками для создания разделов в дочерней виртуальной машине так же, как делаете это в основной ОС. Диск NVMe будет работать на полной скорости без каких-либо потерь производительности (можете убедиться в этом с помощью любого инструмента оценки производительности диска).

Чтобы правильно удалить устройство из виртуальной машины и переключить его в ОС хоста, нужно сначала выключить виртуальную машину, а затем использовать следующие команды (не забывайте менять имя виртуальной машины и расположение контроллера NVMe):

```
Remove-VMAssignableDevice -LocationPath "PCIROOT(0)#PCI(0302)#PCI(0000)" -VMName
"Vibranium"
Mount-VMHostAssignableDevice -LocationPath "PCIROOT(0)#PCI(0302)#PCI(0000)"
```

После последней команды контроллер NVMe должен снова появиться в списке диспетчера устройств хостовой ОС. Вам просто нужно снова включить его для перезапуска, чтобы использовать диск NVMe на хосте.

Виртуальные машины с поддержкой VA

Виртуальные машины применяются для различных целей. Одна из них заключается в возможности корректно запускать традиционное программное обеспечение в изолированных средах, называемых *контейнерами*. (Бункеры серверов и приложений, которые представляют собой два типа контейнеров, описаны в главе 3 тома 1.) Полностью изолированные контейнеры (внутренние названия — Xenon и Krypton) требуют функций быстрого запуска, низких накладных расходов и возможности получить минимальный объем памяти. Гостевая физическая память этой разновидности виртуальных машин обычно распределяется между несколькими контейнерами. Хорошими примерами могут послужить Application Guard в Защитнике Windows, где контейнер используется для обеспечения полной изоляции браузера, и Windows Sandbox, с помощью контейнеров обеспечивающая полностью изолированную виртуальную среду. Обычно в контейнерах задействуются одна общая прошивка виртуальной машины, операционная система и зачастую некоторые приложения, работающие в них (общие компоненты составляют базовый уровень контейнера). Запуск каждого контейнера в персональном гостевом пространстве физической памяти был бы невозможен и привел бы к большим потерям физической памяти.

Чтобы решить эту проблему, стек виртуализации обеспечивает работу виртуальных машин с поддержкой VA (virtual appliance — предварительно конфигурированный образ виртуальной машины). У них используется диспетчер памяти операционной системы хоста для предоставления расширенных функций физической памяти гостевого раздела, таких как дедупликация памяти, обрезка памяти, прямое отображение, клонирование памяти и, что наиболее важно, подкачка памяти (все эти концепции подробно описаны в главе 5 тома 1). В традиционных виртуальных

машинах гостевая память назначается драйвером VID путем статического выделения физических страниц системы с хоста и сопоставления их в пространстве GPA виртуальной машины прежде, чем какой-либо виртуальный процессор сможет выполнить код, но для VM с поддержкой VA задействуется новый слой-посредник между пространством GPA и пространством SPA. Вместо отражения страниц SPA непосредственно в пространство GPA VID выделяет изначально пустое пространство GPA, создает минимальный процесс пользовательского режима (называемый VMMEM) в роли владельца адресов VA и настраивает сопоставление GPA с VA с помощью MicroVM. MicroVM — это новый компонент ядра NT, тесно интегрированный с диспетчером памяти NT, который в итоге отвечает за управление сопоставлением GPA с SPA путем объединения задач сопоставления GPA с VA (обслуживаются драйвером VID) и VA с SPA (обслуживаются диспетчером памяти NT).

Новый слой-посредник позволяет виртуальным машинам с поддержкой VA применять большинство функций управления памятью, доступных процессам Windows. Как обсуждалось в предыдущем подразделе, рабочий процесс VM при запуске виртуальной машины просит драйвер VID создать блок памяти раздела. Если это виртуальная машина с VA, она создает битовую карту сопоставления диапазона блоков памяти GPA, которая используется для отслеживания выделенных виртуальных страниц, составляющих оперативную память новой виртуальной машины. Затем создается оперативная память раздела, поддерживаемая обширным диапазоном виртуального пространства. Пространство VA обычно равно выделенному объему оперативной памяти виртуальной машины (обратите внимание на то, что это необязательное условие, — разные диапазоны VA могут быть сопоставлены как разные диапазоны GPA) и резервируется в контексте процесса VMMEM с помощью нативной функции `NtAllocateVirtualMemory`.

Если оптимизационная функция отложенной фиксации неактивна (подробности в следующем разделе), драйвер VID выполняет еще один вызов API `NtAllocateVirtualMemory` с целью фиксации всего диапазона VA. Как обсуждалось в главе 5 тома 1, при выделении памяти используется системный лимит выделения, но по-прежнему не выделяется ни одной физической страницы (все записи PTE, описывающие весь диапазон, — это недопустимые PTE с нулевым требованием). Драйвер VID на этом этапе использует Winhvr, чтобы попросить гипервизор сопоставить пространство GPA всего раздела со специальным недопустимым SPA, задействуя гипервызов `HvMapGpaPages`, применяемый для стандартных разделов. Когда гостевой раздел обращается к гостевой физической памяти, представленной в таблице SLAT с помощью специального недопустимого SPA, у гипервизора происходит VMEXIT и он понимает специальное значение этого и выполняет перехват памяти в корневой раздел.

Наконец, драйвер VID уведомляет MicroVM о новом диапазоне GPA, поддерживаемом VA, вызывая процедуру `VmCreateMemoryRange` (службы MicroVM предоставляются ядром NT драйверу VID через расширение ядра). MicroVM выделяет и инициализирует структуру данных `VM_PROCESS_CONTEXT`, которая содержит два важных красно-черных дерева (RB): одно описывает выделенные диапазоны GPA в виртуальной машине, другое — соответствующие диапазоны системных виртуальных адресов (SVA) в корневом разделе. Указатель на выделенную структуру данных затем сохраняется в EPROCESS экземпляра VMMEM.

Когда рабочий процесс VM хочет выполнить запись в память виртуальной машины с поддержкой VA или генерируется перехват памяти из-за недопустимого преобразования GPA в SPA, драйвер VID вызывает обработчик ошибки страницы MicroVM (`VmAccessFault`). Обработчик выполняет две важные операции: сначала устраняет ошибку, вставляя допустимый PTE в таблицу страниц, описывающую сбойную виртуальную страницу (подробнее см. в главе 5 тома 1), а затем обновляет таблицу SLAT дочерней виртуальной машины путем вызова драйвера WinHvr, который генерирует еще один гипервызов `HvMapGpaPages`. После этого гостевые физические страницы виртуальной машины могут быть выгружены просто потому, что собственную память процесса обычно разрешается выгрузить из памяти. Это имеет важное значение: от большинства функций MicroVM требуется возможность работать на пассивном IRQL.

Для виртуальных машин с поддержкой VA можно использовать несколько служб диспетчера памяти NT. В частности, *шаблоны клонирования* позволяют быстро клонировать память двух разных виртуальных машин с поддержкой VA, *прямое отображение* (direct map) дает возможность разделяемым образам или файлам данных сопоставлять объекты разделов с процессом VMMEM и диапазоном GPA, указывающим на этот регион VA. Базовые физические страницы могут совместно задействоваться различными виртуальными машинами и хост-процессами, что приводит к повышению плотности памяти.

Средства оптимизации виртуальных машин с поддержкой VA

Как рассказывалось в предыдущем подразделе, затратность гостевого доступа к динамически резервируемой памяти, которая в данный момент не выделена или не предоставляет требуемых разрешений, может быть довольно высокой: при попытке гостевого доступа к недоступной памяти происходит VMEXIT, что требует от гипервизора приостановить гостевую виртуальную машину, запланировать виртуальную машину корневого раздела и передать ей сообщение о перехвате памяти. Обработчик отклика на перехват VID вызывается при высоком IRQL, но обработка запроса и вызов MicroVM требуют запуска на `PASSIVE_LEVEL`. Таким образом, DPC ставится в очередь. Процедура DPC устанавливает событие, которое активизирует соответствующий поток, отвечающий за обработку перехвата. После того как обработчик ошибки страницы MicroVM устранит проблему и вызовет гипервизор для обновления записи SLAT (через другой гипервызов, который создает еще один VMEXIT), тот возобновит гостевую VP.

Большое количество перехватов памяти, генерируемых во время выполнения, приводит к значительному снижению производительности. Чтобы избежать этого, было реализовано несколько средств оптимизации в виде гостевых компонентов паравиртуализации (или простых конфигураций):

- компоненты паравиртуализации для обнуления памяти;
- подсказки по доступу к памяти;
- ошибка паравиртуализированной страницы;
- отложенная фиксация и др.

Компоненты паравиртуализации для обнуления памяти

Чтобы избежать раскрытия виртуальной машине информации об артефактах памяти, ранее использовавшихся корневым разделом или другой виртуальной машиной, перед отображением для доступа гостя выделенная гостевая оперативная память обнуляется. Обычно во время загрузки операционная система обнуляет всю физическую память, поскольку ее содержимое недетерминированно. Для виртуальной машины это означает, что память может быть обнулена дважды: один раз хостом виртуализации, второй раз гостевой операционной системой. Для виртуальных машин *с физической поддержкой* это в лучшем случае пустая трата ресурсов ЦП. Для виртуальных машин обнуление гостевой ОС приводит к затратным перехватам памяти. Чтобы избежать ненужных перехватов, гипервизор предоставляет компоненты паравиртуализации для обнуления памяти.

Когда загрузчик Windows загружает основную операционную систему, он пользуется сервисами UEFI, чтобы получить карту физической памяти компьютера. Когда гипервизор запускает виртуальную машину с поддержкой VA, он предоставляет гипервызов `HvGetBootZeroedMemory`, который загрузчик Windows может использовать для запроса списка диапазонов физической памяти, которые фактически уже обнулены. Прежде чем передать выполнение ядру NT, загрузчик Windows включает полученные обнуленные диапазоны в список дескрипторов физической памяти, предоставленный сервисом EFI и хранящийся в блоке Loader (подробности о механизмах запуска см. в главе 12). Ядро NT вставляет объединенный дескриптор непосредственно в список обнуленных страниц, пропуская начальное обнуление памяти.

Аналогичным образом гипервизор поддерживает функцию обнуления памяти при горячем добавлении с помощью упрощенной реализации: когда драйвер VSC динамической памяти (dmvsc.sys) инициирует запрос на добавление физической памяти в ядро NT, он устанавливает флаг `MM_ADD_PHYSICAL_MEMORY_ALREADY_ZEROED`, который дает диспетчеру памяти (MM) указание добавить новые страницы непосредственно в список обнуленных страниц.

Подсказки по доступу к памяти

Для виртуальных машин *с физической поддержкой* корневой раздел владеет очень ограниченной информацией о том, как гостевой диспетчер памяти собирается использовать свои физические страницы. Для таких VM это по большей части не имеет значения, поскольку почти все сопоставления памяти и GPA создаются при запуске виртуальной машины и остаются статически сопоставленными. Для виртуальных машин с VA такие сведения могут оказаться очень полезными, поскольку диспетчер памяти хоста управляет рабочим набором минимального процесса, владеющего памятью VM (VMMEM).

Горячая подсказка позволяет гостю сообщить, что какой-то набор физических страниц должен быть сопоставлен с гостем, поскольку к ним будут скоро или часто обращаться. Это означает, что страницы следует добавить в рабочий набор минимального процесса. VID обрабатывает подсказку, приказывая MicroVM

немедленно вызвать ошибку на физических страницах и не удалять их из рабочего набора процесса VMMEM.

Подобным образом холодная подсказка позволяет гостю указать, что набор физических страниц должен быть отключен от гостя, поскольку он не будет использоваться в ближайшее время. Драйвер VID обрабатывает подсказку, пересылая ее в MicroVM, которая немедленно удаляет страницы из рабочего набора. Обычно гость применяет холодную подсказку для страниц, которые были обнулены местным фоновым потоком (подробности см. в главе 5 части 1).

Гостевой раздел с поддержкой VA задает подсказку памяти для страницы с помощью гипервызова `HvMemoryHeatHint`.

Ошибка паравиртуализированной страницы

Обработка ошибок страниц через паравиртуализацию (enlightened page fault, EPF) — это функция, которая позволяет гостевому разделу с поддержкой VA перепланировать потоки на виртуальном процессоре, которые вызвали перехват памяти для страницы GPA с поддержкой VA. Обычно перехват памяти для такой страницы обрабатывается путем синхронного разрешения ошибки доступа в корневом разделе и возобновления работы ВП после обработки ошибки доступа. Когда функция EPF активна и происходит перехват памяти для страницы GPA с поддержкой VA, драйвер VID в корневом разделе создает фоновый рабочий поток, который вызывает обработчик ошибки страницы MicroVM и доставляет асинхронное исключение (не путать с асинхронным прерыванием) на виртуальный процессор гостя с целью сообщить ему, что текущий поток вызвал перехват памяти.

Гость перепланирует поток, тем временем хост обрабатывает ошибку доступа. После устранения ошибки доступа драйвер VID добавит исходный GPA, вызвавший ошибку, в очередь завершения и доставит гостю асинхронное прерывание. Оно заставляет гостя проверить очередь завершения и разблокировать все потоки, ожидающие завершения EPF.

Отложенная фиксация и другие средства оптимизации

Отложенная фиксация — это средство оптимизации, которое, если оно активно, заставляет драйвер VID не фиксировать каждую зарезервированную страницу до момента первого доступа. Это позволяет запускать больше виртуальных машин одновременно без увеличения файла подкачки, но, поскольку выделенное пространство VM еще только зарезервировано, а не зафиксировано, виртуальные машины могут аварийно завершиться во время выполнения из-за достижения предела фиксации в корневом разделе. В этом случае останется больше свободной памяти.

Доступны и другие средства оптимизации для установки размера страниц, которые будут выделены обработчиком ошибок страниц MicroVM (маленькие или большие), а также для *закрепления* резервных страниц при первом доступе. Это предотвращает устаревание и обрезку памяти, что в целом стабилизирует производительность, но повышает расход памяти и снижает ее плотность.

Процесс VMMEM

Процесс VMMEM существует в основном по двум причинам.

- Он размещает цикл потоков диспетчеризации VP, когда включен корневой планировщик, который представляет планируемый блок гостевого VP.
- Он размещает пространство виртуального устройства для VM, поддерживаемых виртуальным устройством.

Процесс VMMEM создается драйвером VID при формировании раздела виртуальной машины. Что касается обычных разделов (подробности — в предыдущем разделе), то рабочий процесс VM инициализирует настройку виртуальной машины через библиотеку VID.dll, которая вызывает VID через IOCTL. Если драйвер VID обнаруживает, что новый раздел имеет поддержку VA, он вызывает MicroVM (с помощью функции `VsmmNtSlatMemoryProcessCreate`), чтобы создать минимальный процесс. MicroVM использует функцию `PsCreateMinimalProcess`, которая выделяет процесс, создает его адресное пространство и вставляет процесс в список процессов. Затем она резервирует нижние 4 Гбайт адресного пространства, чтобы гарантировать, что туда не попадут битовые карты с прямым сопоставлением (это может снизить энтропию и безопасность для гостя). Драйвер VID задействует для нового процесса VMMEM особый дескриптор защиты, только система и рабочий процесс VM могут получить к нему доступ. (Рабочий процесс VM запускается с особым токеном, владельцу которого присваивается SID, сгенерированный на основе уникального GUID виртуальной машины.) Это важно, поскольку в противном случае виртуальное адресное пространство процесса VMMEM могло бы быть доступно кому угодно. Читая виртуальную память процесса, злоумышленник может прочитать закрытую физическую память виртуальной машины.

БЕЗОПАСНОСТЬ НА ОСНОВЕ ВИРТУАЛИЗАЦИИ (VBS)

Как сказано в предыдущем разделе, Hyper-V предоставляет сервисы, необходимые для управления виртуальными машинами и их запуска в системах Windows. Гипервизор гарантирует необходимую изоляцию для каждого раздела. Таким образом, одна виртуальная машина не может вмешиваться в работу другой. В данном разделе опишем еще один важный компонент инфраструктуры виртуализации Windows — безопасное ядро, которое предоставляет базовые службы безопасности на основе виртуализации (VBS).

Сначала мы перечислим сервисы, предоставляемые безопасным ядром, и его требования, а затем опишем его архитектуру и основные компоненты. Далее представим некоторые из его основных внутренних структур данных. Затем обсудим метод запуска безопасного ядра и компонента Virtual Secure Mode, описав его высокую зависимость от гипервизора. В заключение проанализируем компоненты, построенные на основе безопасного ядра: изолированный пользовательский режим, механизм Hypervisor Enforced Code Integrity, анклавы безопасного программного обеспечения, безопасные устройства, а также службы микрокода и горячих исправлений ядра Windows.

Виртуальные уровни доверия и виртуальный безопасный режим

Как обсуждалось в предыдущем разделе, гипервизор применяет SLAT для облсуживания каждого раздела в собственном пространстве памяти. Операционная система, работающая в разделе, обращается к памяти стандартным способом (гостевые виртуальные адреса преобразуются в гостевые физические адреса с помощью таблиц страниц). «Под капотом» аппаратное обеспечение преобразует все GPA разделов в настоящие SPA, а затем получает реальный доступ к памяти. Этот последний уровень преобразования поддерживается гипервизором, который использует отдельную таблицу SLAT для каждого раздела. Аналогичным образом гипервизор может применять SLAT для создания разных доменов безопасности в одном разделе. Благодаря этой функции Microsoft смогла разработать безопасное ядро (Secure Kernel) — основу виртуального безопасного режима.

Традиционно операционная система имела единое физическое адресное пространство, и программное обеспечение, работающее на кольце 0 (то есть в режиме ядра), могло получить доступ к любому адресу физической памяти. Таким образом, если какая-то программа, работающая в привилегированном режиме супервизора (ядро, драйверы и т. д.), будет скомпрометирована, то вся система также окажется под угрозой. Виртуальный безопасный режим применяет гипервизор, чтобы обеспечить новые границы доверия для системного программного обеспечения. С помощью VSM можно установить границы безопасности (они описываются гипервизором с помощью SLAT), ограничивающие ресурсы, к которым может получить доступ код режима супервизора. Таким образом, при использовании VSM, даже если код режима супервизора скомпрометирован, вся система скомпрометирована не будет.

VSM обеспечивает эти границы с помощью концепции виртуальных уровней доверия (VTL). По сути, VTL представляет собой набор средств защиты доступа к физической памяти. Каждый VTL может иметь свой набор средств защиты доступа. Таким образом, VTL можно применять для изоляции памяти. Защиту доступа к памяти VTL можно настроить так, чтобы ограничить объем физической памяти, к которой на нем можно получить доступ. При использовании VSM виртуальный процессор всегда работает на определенном VTL и может получить доступ только к физической памяти, помеченной как доступная через SLAT гипервизора. Например, если процессор работает на VTL 0, он может получить доступ к памяти только в соответствии с защитой доступа к памяти, назначенной VTL 0. Это принудительное управление доступом к памяти происходит на уровне преобразования гостевой физической памяти и, следовательно, не может быть изменено кодом режима супервизора в данном разделе.

VTL организованы в виде иерархии. У более высоких уровней больше привилегий, чем у более низких. Кроме того, высокие уровни могут регулировать защиту доступа к памяти для более низких уровней. Таким образом, программное обеспечение, работающее на VTL 1, может настроить защиту доступа к памяти VTL 0, чтобы

ограничить его. Это позволяет программному обеспечению на VTL 1 скрывать (изолировать) память от VTL 0. Это важная концепция, лежащая в основе VSM. На данный момент гипервизор поддерживает только два VTL: VTL 0 представляет собой обычную среду выполнения ОС, с которой взаимодействует пользователь; VTL 1 представляет безопасный режим, в котором работают безопасное ядро и изолированный пользовательский режим (IUM). Поскольку уровень VTL 0 — это среда, в которой действуют стандартная операционная система и приложения, его часто называют нормальным режимом.

ПРИМЕЧАНИЕ Архитектура VSM изначально была разработана для поддержки максимум 16 VTL. На момент написания книги гипервизор поддерживает только два VTL. Вполне возможно, в будущем Microsoft добавит один или несколько VTL. Например, последние версии Windows Server, работающие в Azure, дополнительно обеспечивают работу конфиденциальных виртуальных машин, которые запускают свой уровень совместимости хостов (HCL) в двух VTL.

С каждым VTL связаны следующие характеристики.

- **Защита доступа к памяти.** Как уже обсуждалось, каждый уровень виртуального доверия имеет набор гостевых средств защиты доступа к физической памяти, которые определяют, как программное обеспечение может получить доступ к памяти.
- **Состояние виртуального процессора.** Виртуальный процессор в гипервизоре использует ряд регистров совместно с каждым VTL, тогда как некоторые другие регистры являются частными для каждого VTL. Частное состояние виртуального процессора для VTL недоступно для программного обеспечения, работающего на более низком уровне VTL. Это позволяет изолировать состояние процессора между VTL.
- **Подсистема прерываний.** Каждый VTL имеет уникальную подсистему прерываний, управляемую синтетическим контроллером прерываний гипервизора. Подсистема прерываний VTL не может быть доступна программному обеспечению, работающему на более низком уровне VTL. Это позволяет безопасно управлять прерываниями на определенном VTL, не рискуя тем, что более низкий VTL создаст или замаскирует неожиданные прерывания.

На рис. 9.30 показана схема защиты памяти, обеспечиваемая гипервизором в виртуальном безопасном режиме. Гипервизор представляет каждый VTL виртуального процессора через другую структуру данных, VMCS (подробнее — в предыдущем разделе), которая включает отдельную таблицу SLAT. Таким образом, программное обеспечение, работающее на определенном VTL, может получить доступ только к страницам физической памяти, назначенным этому уровню. Важная концепция заключается в том, что защита SLAT применяется к *физическим* страницам, но не к *виртуальным*, так как те защищены стандартными таблицами страниц.

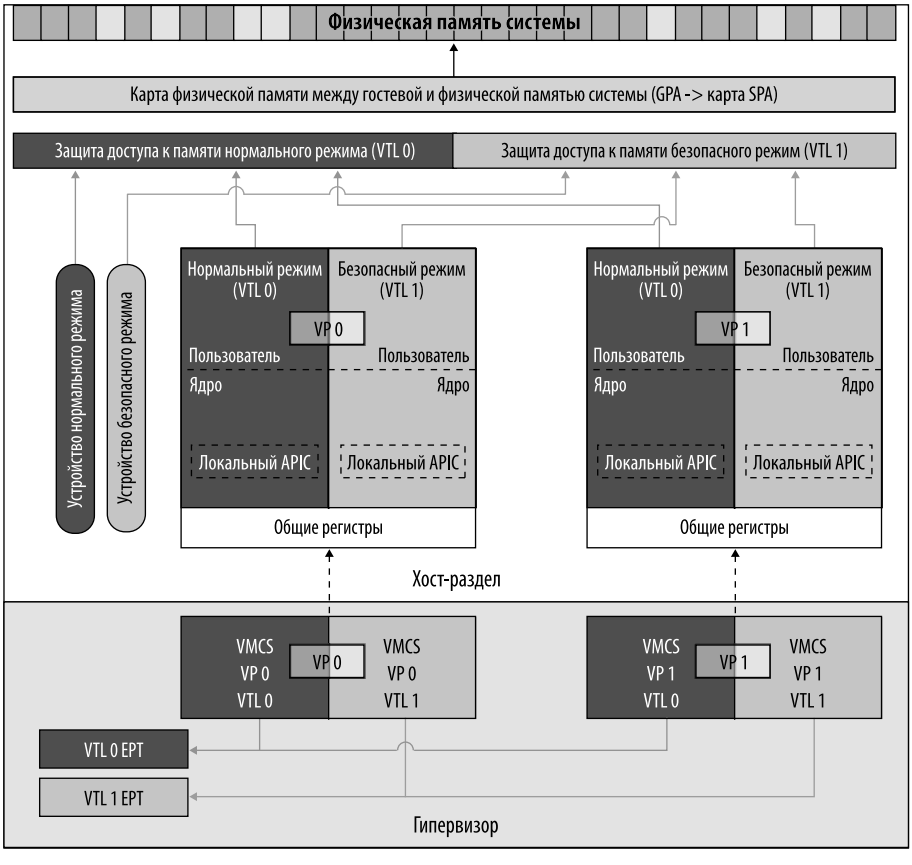


Рис. 9.30. Схема архитектуры защиты памяти, предоставляемой гипервизором VSM

Сервисы, предоставляемые VSM, и требования

Виртуальный безопасный режим, построенный на основе гипервизора, предоставляет экосистеме Windows следующие сервисы.

- **Изоляция.** Изолированный пользовательский режим IUM обеспечивает аппаратную изолированную среду для каждого вида программного обеспечения, работающего в VTL 1. Защищенные устройства, управляемые безопасным ядром, изолированы от остальной системы и работают в пользовательском режиме VTL 1. Программное обеспечение, работающее в VTL 1, обычно хранит секреты, которые невозможно перехватить или раскрыть в VTL 0. Эта служба активно применяется Credential Guard — функцией, которая сохраняет все системные учетные данные в адресном пространстве памяти трастлета LsaIso, работающего в пользовательском режиме VTL 1.
- **Контроль над VTL 0.** Функция Hypervisor Code Integrity (HVCI) проверяет целостность и подпись каждого модуля, который загружает и запускает обычная

ОС. Проверка целостности выполняется в VTL 1, имеющем доступ ко всей физической памяти VTL 0. Никакое программное обеспечение VTL 0 не может помешать проверке подписи. Более того, HVCI гарантирует, что все страницы памяти нормального режима, содержащие исполняемый код, помечены как недоступные для записи (эта функция называется W^X. И HVCI, и W^X обсуждались в главе 7 тома 1).

- **Безопасные перехваты.** VSM предоставляет механизм, позволяющий более высокому VTL блокировать критически важные системные ресурсы и предотвращать доступ к ним более низких VTL. Безопасные перехваты широко используются HyperGuard, который обеспечивает еще один уровень защиты ядра VTL 0, предотвращая вредоносные модификации критических компонентов операционных систем.
- **Анклавы на базе VBS.** *Анклав безопасности* — это изолированная область памяти в адресном пространстве процесса пользовательского режима. Область памяти анклава недоступна даже для более высоких уровней привилегий. Первоначальная реализация этой технологии заключалась в применении аппаратных средств для правильного шифрования памяти, принадлежащей процессу. Анклав на базе VBS — это безопасный анклав, гарантии изоляции которого предоставляются с помощью VSM.
- **Защита потока управления ядра.** Когда HVCI активно, VSM обеспечивает защиту потока управления (Control Flow Guard, CFG) каждому модулю ядра, загружаемому в обычном мире (и самому ядру NT). Программное обеспечение режима ядра, работающее в обычной среде, имеет доступ к битовой карте только для чтения, поэтому эксплойт потенциально не может его изменить. По этой причине конфигурация ядра в Windows также известна как Secure Kernel CFG (SKCFG).

ПРИМЕЧАНИЕ CFG — это реализация целостности потока управления Microsoft (Control Flow Integrity) — метода, который предотвращает перенаправление потока выполнения программы самыми разными злонамеренными атаками. И пользовательский режим, и режим ядра CFG подробно обсуждались в главе 7 тома 1.

- **Безопасные устройства.** Это новый тип устройств, которые полностью отображаются и управляются безопасным ядром в VTL 1. Драйверы для них работают целиком в пользовательском режиме VTL 1 и задействуют сервисы безопасного ядра для отображения пространства ввода-вывода для устройства.

VSM предъявляет к оборудованию ряд требований, обеспечивающих его правильное включение и корректную работу. Хост-система должна поддерживать расширения виртуализации (Intel VT-x, AMD SVM или ARM TrustZone) и SLAT. VSM не будет работать, если в системном процессоре отсутствует одна из перечисленных аппаратных функций. Некоторые другие аппаратные функции не являются строго необходимыми, но при их отсутствии часть условий безопасности VSM не будут гарантированы.

- Функция IOMMU необходима для защиты от атак DMA физического устройства. Если системные процессоры не имеют IOMMU, VSM все равно сможет работать, но будет уязвим для атак на физические устройства.

- Функция UEFI BIOS с включенной безопасной загрузкой необходима для защиты цепочки загрузки, которая приводит к запуску гипервизора и безопасного ядра. Если безопасная загрузка не включена, система уязвима для атак при загрузке, которые могут изменить целостность этих компонентов до того, как они смогут запуститься.

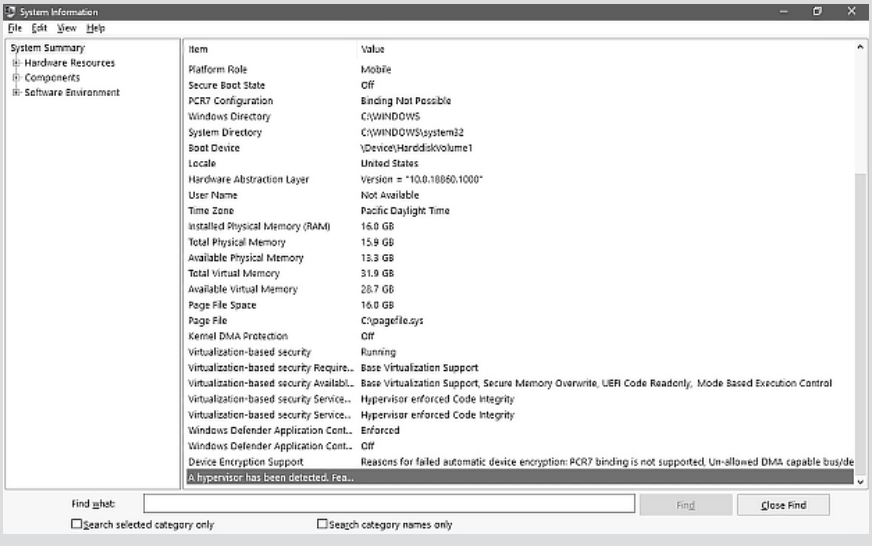
Некоторые другие компоненты необязательны, но их наличие повышает общую безопасность и скорость реагирования системы. Хорошим примером может послужить присутствие TPM. Он используется безопасным ядром для хранения главного ключа шифрования и выполнения безопасного запуска, известного также как DRTM (подробности см. в главе 12). Еще одним аппаратным компонентом, который может улучшить скорость реагирования VSM, является аппаратная поддержка контроля исполнения процессора на основании режима (Mode-Based Execute Control, MBEC), который применяется, когда HVCI включен, для защиты состояния выполнения страниц пользовательского режима в режиме ядра. С помощью аппаратного MBEC гипервизор может установить исполняемое состояние страницы физической памяти на основе домена CPL (ядра или пользователя) конкретного VTL. Таким образом, память, принадлежащая приложению пользовательского режима, может быть физически помечена как исполняемая только с помощью кода пользовательского режима (эксплойты ядра больше не смогут выполнять собственный код, расположенный в памяти приложения пользовательского режима). При отсутствии аппаратного MBEC гипервизор должен будет эмулировать его, задействуя две разные таблицы SLAT для VTL 0 и переключая их, когда выполнение кода меняет домен безопасности CPL (в таком случае переход из пользовательского режима в режим ядра и наоборот будет вызывать VMEXIT). Более подробно HVCI обсуждалась в главе 7 тома 1.

ЭКСПЕРИМЕНТ. Обнаружение VBS и предоставляемых им функций

В главе 12 мы обсудим политику запуска VSM и проинструктируем относительно ручного включения или отключения безопасности на основе виртуализации. В этом же эксперименте определяем состояние различных функций, предоставляемых гипервизором и безопасным ядром. VBS — это технология, которая не видна пользователю. Инструмент System Information (Информация о системе), поставляемый вместе с базовой установкой Windows, может отображать подробную информацию о безопасном ядре и связанных с ним технологиях. Вы можете запустить его, набрав `msinfo32` в поле поиска Cortana. Обязательно запустите его от имени администратора, поскольку для некоторых деталей потребуется учетная запись пользователя с полными привилегиями.

На рисунке VBS включен, вместе с ним HVCI (а именно, целостность кода, обеспечиваемая гипервизором), виртуализация среды выполнения UEFI (а именно, UEFI только для чтения) и MBEC (как управление выполнением на основе режима). Однако система, описанная в примере, не подразумевает активной безопасной загрузки и не имеет работающего IOMMU, указанного как Защита DMA (DMA

Protection) в строке Доступные свойства безопасности на основе виртуализации (Virtualization-Based Security Available Security Properties).



Более подробную информацию о том, как включать, отключать и блокировать конфигурацию VBS, можно найти в эксперименте «Знакомство с политикой VSM» в главе 12.

БЕЗОПАСНОЕ ЯДРО (SECURE KERNEL)

Основная часть безопасного ядра реализована в файле `securekernel.exe` и запускается Загрузчиком Windows после успешного запуска гипервизора. Как показано на рис. 9.31, безопасное ядро — это минимизированная ОС, которая работает строго с обычным ядром, находящимся в VTL 0. Как и любая обычная ОС, безопасное ядро работает в CPL 0 (известно также как кольцо 0 или режим ядра) VTL 1 и предоставляет сервисы (большинство из них посредством системных вызовов) для изолированного пользовательского режима (IUM), который находится в CPL 3 (известно также как кольцо 3 или пользовательский режим) VTL 1. Безопасное ядро было разработано с упором на минимальный размер с целью уменьшения поверхности внешней атаки. Его нельзя расширить с помощью драйверов внешних устройств, как обычное ядро. Единственные модули этого ядра с возможностью расширения их функциональности загружаются загрузчиком Windows перед запуском VSM и импортируются из `securekernel.exe`.

- **Skci.dll** реализует часть безопасного ядра, обеспечивающую защиту гипервизором целостности кода.
- **Cng.sys** предоставляет криптографический механизм для безопасного ядра.

- **Vmsvcext.dll** обеспечивает поддержку аттестации компонентов безопасного ядра в средах Intel TXT (безопасную загрузку) (дополнительная информация о безопасной загрузке есть в главе 12).

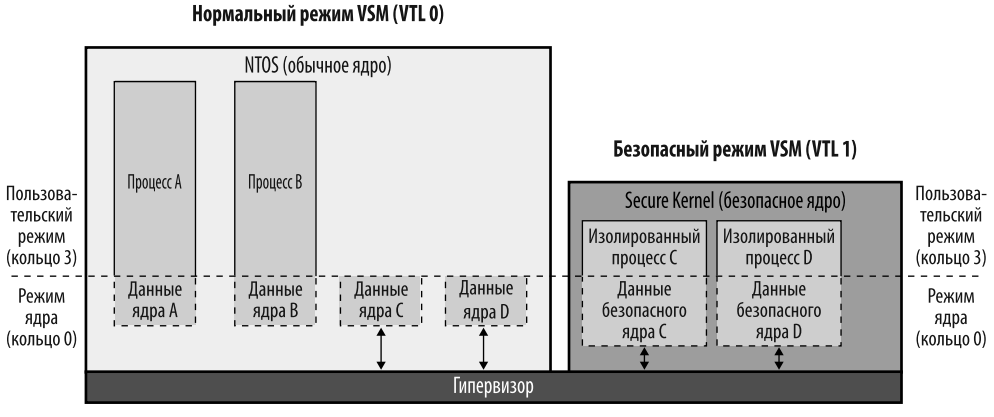


Рис. 9.31. Схема архитектуры виртуального безопасного режима, построенная на основе гипервизора

Хотя безопасное ядро не может быть расширено, изолированный пользовательский режим включает в себя специализированные процессы, называемые *трастлетами*. Они изолированы друг от друга и имеют особые требования к цифровой подписи. Трастлеты могут связываться с безопасным ядром посредством системных вызовов и с обычным миром через почтовые ящики и ALPC. Изолированный пользовательский режим обсуждается далее в данной главе.

Виртуальные прерывания

Когда гипервизор настраивает нижестоящие виртуальные разделы, он требует, чтобы физические процессоры выполняли VMEXIT каждый раз, когда внешнее прерывание вызывается их физическим APIC (расширенный программируемый контроллер прерываний). Расширения виртуальной машины оборудования позволяют гипервизору внедрять виртуальные прерывания в гостевые разделы (более подробную информацию можно найти в руководствах пользователя Intel, AMD и ARM). Благодаря этим двум факторам в гипервизоре реализована концепция синтетического контроллера прерываний (SynIC). SynIC может управлять двумя видами прерываний. Виртуальные прерывания — это прерывания, доставляемые виртуальному APIC гостевого раздела. Виртуальное прерывание может представлять собой физическое аппаратное прерывание, генерируемое реальным оборудованием, и быть связано с ним. В противном случае виртуальное прерывание может быть синтетическим прерыванием, которое генерируется самим гипервизором в ответ на определенные виды событий. SynIC может отображать физические прерывания на виртуальные. VTL имеет SynIC, связанный с каждым виртуальным процессором, в котором работает VTL. На момент написания данного текста гипервизор был спроектирован для поддержки 16 различных синтетических векторов прерываний (хотя на самом деле применяются только два).

При запуске системы (фаза 1 инициализации ядра NT) драйвер ACPI сопоставляет каждое прерывание с корректным вектором, используя службы, предоставляемые HAL. NT HAL паравиртуализирован и знает, работает ли он под управлением VSM. В этом случае он вызывает гипервизор для сопоставления каждого физического прерывания с собственным VTL. Даже безопасное ядро может сделать то же самое. Однако на момент написания книги с безопасным ядром не были связаны никакие физические прерывания (это может измениться в будущем — гипервизор уже поддерживает эту функцию). Вместо этого безопасное ядро обращается к гипервизору за получением только следующих виртуальных прерываний: безопасных таймеров, помощника для уведомления о виртуальных прерываниях (VINA) и безопасных перехватов.

ПРИМЕЧАНИЕ Важно понимать: гипервизор требует, чтобы низкоуровневое оборудование выполняло VMEXIT при управлении прерываниями только внешнего типа. Исключения по-прежнему управляются на том же VTL, на котором выполняется процессор (VMEXIT не генерируется). Если инструкция вызывает исключение, последнее по-прежнему управляется кодом структурированной обработки исключений (SEH), расположенным в текущем VTL.

Чтобы понять три типа виртуальных прерываний, мы должны сначала представить, как гипервизор управляет прерываниями.

В гипервизоре каждый VTL спроектирован так, чтобы безопасно получать прерывания от устройств, связанных с его собственным VTL, иметь безопасный таймер, которому не могут помешать менее безопасные VTL, и иметь возможность предотвращать прерывания, направленные на более низкие VTL при выполнении кода с более высоким VTL. Более того, VTL должен иметь возможность отправлять прерывания IPI другим процессорам. Эта модель создает следующие сценарии.

- В ходе работы на определенном VTL прием прерываний, нацеленных на текущий VTL, приводит к их стандартной обработке, как определяется виртуальным контроллером APIC VP.
- Получение прерывания, нацеленного на более высокий VTL, вызывает переключение на более высокий VTL, которому оно предназначено, если его значение IRQL позволяет искомое прерывание представить. Если же значение IRQL более высокого VTL не позволяет доставить прерывание, оно ставится в очередь без переключения текущего VTL. Такое поведение позволяет более высокому VTL выборочно маскировать прерывания при возврате к более низкому VTL. Это может быть полезно, если более высокий VTL обрабатывает прерывание и ему необходимо вернуться к более низкому VTL, чтобы ускорить процесс.
- Когда получено прерывание, нацеленное на более низкий VTL, чем выполняющийся в данный момент VTL виртуального процессора, прерывание ставится в очередь для будущей доставки на более низкий VTL. Прерывание, нацеленное на более низкий VTL, никогда не будет прерывать выполнение текущего VTL. Вместо этого оно отображается, когда виртуальный процессор скоро перейдет на целевой VTL.

Предотвращение прерываний, направленных на более низкие VTL, — не всегда хорошее решение. Во многих случаях это может замедлить нормальную работу ОС, особенно в критически важных или игровых средах. Чтобы лучше управлять этими условиями, был добавлен помощник для уведомления о виртуальных прерываниях

(VINA). В рамках обычного цикла диспетчеризации событий гипервизор проверяет, есть ли ожидающие прерывания, поставленные в очередь на более низкий VTL. Если это так, он внедряет прерывание VINA в текущий выполняющийся VTL. У безопасного ядра есть обработчик, зарегистрированный для вектора VINA в его виртуальном IDT. Обработчик (функция `ShvlVinaHandler`) выполняет обычный вызов (`NORMALKERNEL_VINA`) к VTL 0 (обычные и безопасные вызовы обсуждаются далее в данной главе). Этот вызов заставляет гипервизор переключиться на обычное ядро (VTL 0). Пока VTL переключен, все прерывания в очереди будут корректно отправлены. Обычное ядро повторно войдет в VTL 1, отправив безопасный вызов `SECUREKERNEL_RESUMETHREAD`.

Безопасные IRQL

Обработчик VINA не всегда будет выполняться в VTL 1. Как и в ядре NT, это зависит от фактического IRQL выполняемого кода. IRQL текущего исполняемого кода маскирует все прерывания, связанные с IRQL, который меньше его или равен ему. Сопоставление между вектором прерывания и IRQL поддерживается регистром приоритета задач (TPR) виртуального APIC, как и в случае реальных физических APIC (дополнительную информацию см. в руководстве по архитектуре Intel). Как показано на рис. 9.32, безопасное ядро поддерживает разные уровни IRQL, в отличие от обычного. Эти IRQL называются безопасными.

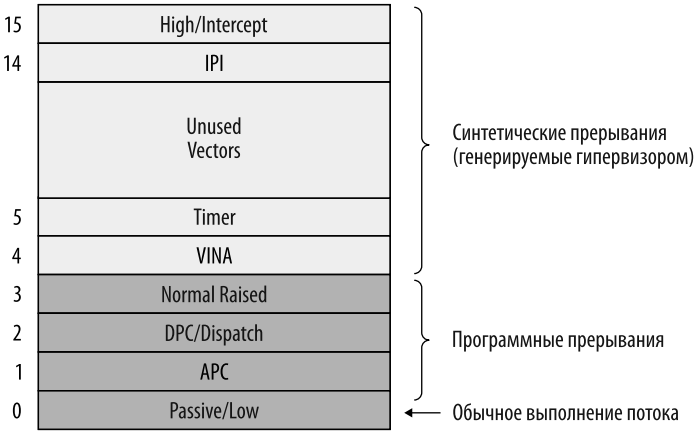


Рис. 9.32. Уровни запросов прерываний безопасного ядра (IRQL)

Первыми тремя безопасными IRQL управляет безопасное ядро, как и в обычной среде. Обычные APC и DPC, нацеленные на VTL 0, по-прежнему не могут вытеснять код, выполняющийся в VTL 1 через гипервизор, но прерывание VINA по-прежнему доставляется в безопасное ядро. (Операционная система управляет тремя программными прерываниями, записывая данные в регистр приоритета задач APIC целевого процессора, что вызывает VMEXIT для гипервизора. Дополнительные сведения об APIC TPR см. в руководствах Intel, AMD или ARM.) Это означает, что, если DPC в обычном режиме нацелен на процессор, пока тот выполняет код VTL 1 (с совместимым безопасным IRQL, который должен быть

меньше, чем Dispatch), прерывание VINA будет доставлено и переключит контекст выполнения на VTL 0. Фактически при этом выполняется DPC в обычной среде, и на некоторое время IRQL обычного ядра повышается до уровня Dispatch. Когда очередь DPC опустошается, IRQL обычного ядра сбрасывается. Поток выполнения возвращается в безопасное ядро благодаря коду цикла связи VSM, который находится в процедуре `VslpEnterIumSecureMode`. Цикл обрабатывает каждый обычный вызов, исходящий от безопасного ядра.

Безопасное ядро сопоставляет первые три безопасных IRQL с одним и тем же IRQL обычной среды. Когда безопасный вызов выполняется из кода, запущенного с определенным IRQL (все еще меньше Dispatch или равным ему) в обычной среде, безопасное ядро переключает собственный безопасный IRQL на тот же уровень. И наоборот, когда безопасное ядро выполняет обычный вызов для входа в ядро NT, оно переключает IRQL обычного ядра на тот же уровень, что и его собственный. Это работает только для первых трех уровней.

Нормальный повышенный уровень используется, когда ядро NT входит в безопасную среду с IRQL выше уровня DPC. В этих случаях безопасное ядро отображает все обычные IRQL, находящиеся выше DPC, на их обычный повышенный уровень безопасности. Код безопасного ядра, выполняющийся на этом уровне, не может получать VINA для каких-либо программных IRQL в обычном ядре, но может получать VINA для аппаратных прерываний. Каждый раз, когда ядро NT входит в безопасную среду с нормальным IRQL выше DPC, безопасное ядро повышает свой безопасный IRQL до нормального уровня.

Безопасные IRQL, равные или превышающие VINA, никогда не могут быть вытеснены каким-либо кодом в обычном мире. Это объясняет, почему безопасное ядро поддерживает концепцию безопасных, невытесняемых таймеров и безопасных перехватов. Безопасные таймеры генерируются процедурой обработчика прерываний часов гипервизора (ISR). Этот ISR, прежде чем внедрять синтетическое прерывание часов в ядро NT, проверяет, есть ли один или несколько безопасных таймеров, срок действия которых истек. Если есть, он внедряет синтетическое безопасное прерывание таймера в VTL 1. Затем переходит к пересылке прерывания по тактовому сигналу в обычный VTL.

Безопасный перехват

В некоторых случаях безопасному ядру может потребоваться запретить ядру NT, работающему с более низким VTL, доступ к определенным критическим системным ресурсам. Например, запись в MSR некоторых процессоров потенциально может быть использована для организации атаки, которая приведет к отключению гипервизора или нарушению некоторых его средств защиты. VSM предоставляет механизм, позволяющий более высокому VTL блокировать критические системные ресурсы и предотвращать доступ к ним со стороны более низких VTL. Этот механизм называется *безопасным перехватом*.

Безопасные перехваты реализуются в безопасном ядре путем регистрации синтетического прерывания, которое предоставляется гипервизором (переназначается в безопасном ядре на вектор 0xF0). Когда определенные события вызывают VMEXIT, гипервизор внедряет синтетическое прерывание в более высокий VTL на виртуальном процессоре, который инициировал перехват. На момент написания

данного текста безопасное ядро регистрируется в гипервизоре для следующих типов перехваченных событий:

- записи в некоторые важные MSR процессора (Star, Lstar, Cstar, Efer, Sysenter, Ia32Misc и базу APIC на архитектурах AMD64) и специальные регистры (GDT, IDT, LDT);
- записи в определенные регистры управления (CR0, CR4 и XCR0);
- записи в некоторые порты ввода-вывода (хорошими примерами являются порты 0xCF8 и 0xCFC — перехват управляет реконфигурацией устройств PCI);
- недопустимый доступ к защищенной гостевой физической памяти.

Когда программное обеспечение VTL 0 выполняет перехват, который будет вызван в VTL 1, безопасному ядру необходимо распознать тип перехвата из своей процедуры обработки прерываний. Для этой цели оно использует очередь сообщений, выделенную SynIC для источника синтетического прерывания «Перехват» (дополнительную информацию о SynIC и SINT см. в подразделе «Коммуникация между разделами» ранее). Безопасное ядро может обнаруживать и отображать страницу физической памяти, проверяя синтетический MSR SIMP, который виртуализуется гипервизором. Отображение физической страницы выполняется во время инициализации безопасного ядра в VTL 1. Запуск безопасного ядра описан далее в этой главе.

Перехваты широко используются HyperGuard для защиты критичных частей обычного ядра NT. Если вредоносный руткит, установленный в ядре NT, пытается изменить систему, записывая определенное значение в защищенный регистр, например в обработчики системных вызовов CSTAR и LSTAR или специфичные для модели регистры, то обработчик перехвата безопасного ядра (`ShvlpInterceptHandler`) фильтрует новое значение регистра и, обнаружив, что значение неприемлемо, вводит немаскируемое исключение общей ошибки защиты (General Protection Fault, GPF) в ядро NT в VLT 0. Это вызывает немедленную проверку ошибок, приводящую к остановке системы. Если значение приемлемо, безопасное ядро записывает новое значение регистра с помощью гипервизора через гипервызов `HvSetVpRegisters` (в этом случае безопасное ядро служит прокси для доступа к регистру).

Контроль над гипервызовами

Последний тип перехвата, который безопасное ядро регистрирует в гипервизоре, — это перехват гипервызова. Обработчик перехвата гипервызова убеждается, что гипервызов, отправленный кодом VTL 0 гипервизору, допустим и исходит из самой операционной системы, а не поступил через какие-то внешние модули. Каждый раз, когда в любом VTL генерируется гипервызов, происходит VMEXIT в гипервизоре (по замыслу). Гипервызовы — это базовый сервис, используемый компонентами ядра каждого VTL для запроса к сервисам друг друга (и самого гипервизора). Гипервизор внедряет синтетическое прерывание перехвата в VTL более высокого уровня только для гипервызовов для запросов сервисов непосредственно к нему, пропуская все гипервызовы, применяемые для безопасных и обычных вызовов к Secure Kernel и от него.

Если гипервызов не признан действительным, он не будет выполнен, в этом случае безопасное ядро обновит младшие регистры VTL, чтобы сигнализировать об ошибке гипервызова. Система избежит аварийного завершения (хотя в будущем такое поведение может измениться), вызывающий код сможет решить, как справиться с ошибкой.

Системные вызовы VSM

Как говорилось в предыдущих разделах, VSM использует гипервызовы для запроса сервисов к безопасному ядру и от него. Гипервызовы изначально разрабатывались как способ запроса сервисов к гипервизору, но в VSM модель была расширена для поддержания новых типов системных вызовов.

- Безопасные вызовы выполняются обычным ядром NT в VTL 0, чтобы потребовать сервисы безопасного ядра.
- Обычные вызовы запрашиваются безопасным ядром в VTL 1, когда ему нужны сервисы, предоставляемые ядром NT, которое работает в VTL 0. Кроме того, некоторые из них используются защищенными процессами (трастлетами), работающими в изолированном пользовательском режиме (IUM), для запроса сервиса из безопасного ядра или обычного ядра NT.

Подобные системные вызовы реализованы в гипервизоре, безопасном ядре и обычном ядре NT. Гипервизор определяет два гипервызова для переключения между разными VTL: `HvVtlCall` и `HvVtlReturn`. Ядро Secure Kernel и ядро NT определяют цикл диспетчеризации, используемый для отправки безопасных и обычных вызовов.

Кроме того, безопасное ядро реализует еще один тип системных вызовов — безопасные системные вызовы. Они дают возможность только защиты процессов (трастлетов), выполняющихся в IUM. Эти системные вызовы недоступны обычному ядру NT. Гипервизор не участвует в обработке безопасных системных вызовов.

Состояние виртуального процессора

Прежде чем углубляться в архитектуру безопасных (Secure) и обычных (Normal) вызовов, необходимо проанализировать, как виртуальный процессор управляет переходами между VTL. Безопасные VTL всегда работают в длинном режиме (это модель выполнения для процессоров AMD64, в которой ЦП обращается только к 64-битным инструкциям и регистрам) с включенной подкачкой. Любая другая модель выполнения не поддерживается. Это упрощает запуск безопасных VTL и управление ими, а также обеспечивает дополнительный уровень защиты кода, работающего в безопасном режиме. (Некоторые другие важные выводы обсуждаются далее в этой главе.)

Для повышения эффективности виртуальный процессор имеет ряд регистров, используемых совместно несколькими VTL, и ряд других регистров, частных для каждого VTL. Состояние общих регистров не меняется при переключении между VTL. Это позволяет быстро передавать небольшой объем информации между уровнями доверия и снижает затраты на переключение контекста при переходе. Каждый VTL имеет собственный экземпляр частных регистров, доступ к которым может быть получен только им. Гипервизор занимается сохранением и восстановлением содержимого частных регистров при переключении между VTL. Таким образом, при входе на какой-то VTL на виртуальном процессоре состояние частных регистров имеет те же параметры, что и при последнем запуске этого VTL на данном виртуальном процессоре.

Большая часть состояния регистра виртуального процессора разделяется между VTL. В частности, регистры общего назначения, векторные регистры и регистры с плавающей запятой используются всеми VTL, за некоторыми исключениями, такими как регистры RIP и RSP. Частными могут быть ряд регистров управления, некоторые архитектурные регистры и виртуальные MSR гипервизора. Механизм безопасного перехвата (подробнее см. в предыдущем подразделе) применяется, чтобы позволить безопасной среде контролировать, к какому MSR может получить доступ среда обычного режима. В табл. 9.3 показано, какие регистры являются общими для всех VTL, а какие — частными для каждого VTL.

Таблица 9.3. Состояния регистров виртуального процессора для каждого VTL

Тип	Основные регистры	MSR
Общие	Rax, Rbx, Rcx, Rdx, Rsi, Rdi, Rbp CR2 R8–R15 DR0–DR5 Состояние X87 с плавающей точкой XMM-регистры AVX-регистры XCR0 (КСЭМ) DR6 (зависит от процессора)	HV_X64_MSR_TSC_FREQUENCY HV_X64_MSR_VP_INDEX HV_X64_MSR_VP_RUNTIME HV_X64_MSR_RESET HV_X64_MSR_TIME_REF_COUNT HV_X64_MSR_GUEST_IDLE HV_X64_MSR_DEBUG_DEVICE_OPTIONS HV_X64_MSR_BELOW_1MB_PAGE HV_X64_MSR_STATS_PARTITION_RETAIL_PAGE HV_X64_MSR_STATS_VP_RETAIL_PAGE MTRR и PAT MCG_CAP MCG_STATUS
Частные	RIP, RSP RFLAGS CR0, CR3, CR4 DR7 IDTR, GDTR CS, DS, ES, FS, GS, SS, TR, LDTR TCK DR6 (зависит от процессора)	SYSENTER_CS, SYSENTER_ESP, SYSENTER_EIP, STAR, LSTAR, CSTAR, SFMASK, EFER, KERNEL_GSBASE, FS.BASE, GS.BASE HV_X64_MSR_HYPERCALL HV_X64_MSR_GUEST_OS_ID HV_X64_MSR_REFERENCE_TSC HV_X64_MSR_APIC_FREQUENCY HV_X64_MSR_EOI HV_X64_MSR_ICR HV_X64_MSR_TPR HV_X64_MSR_APIC_ASSIST_PAGE HV_X64_MSR_NPIEP_CONFIG HV_X64_MSR_SIRBP HV_X64_MSR_SCONTROL HV_X64_MSR_SVERSION HV_X64_MSR_SIEFP HV_X64_MSR_SIMP HV_X64_MSR_EOM HV_X64_MSR_SINT0–HV_X64_MSR_SINT15 HV_X64_MSR_STIMER0_CONFIG–HV_X64_MSR_STIMER3_CONFIG HV_X64_MSR_STIMER0_COUNT–HV_X64_MSR_STIMER3_COUNT Локальные регистры APIC, включая CR8/TPR

Безопасные вызовы

Когда ядру NT требуются сервисы, предоставляемые безопасным ядром, оно использует специальную функцию `VslpEnterIumSecureMode`. Процедура принимает 104-байтную структуру данных, называемую *SKCALL*, которая применяется для описания типа операции (вызов службы, очистка TB, возобновление потока или вызов анклава), номера защищенного вызова и максимум 12 восьмибайтных параметров. Функция при необходимости повышает IRQL процессора и определяет значение cookie безопасного потока (Secure Thread). Это значение сообщает безопасному ядру, какой безопасный поток будет обрабатывать запрос. Затем она (пере)запускает цикл диспетчеризации безопасных вызовов. Состояние исполняемости каждого VTL — это конечный автомат, который зависит от другого VTL.

Цикл, описываемый функцией `VslpEnterIumSecureMode`, управляет всеми операциями, показанными в левой части рис. 9.33, в VTL 0 (за исключением случая безопасных прерываний). Ядро NT может решить войти в безопасное ядро или в обычное ядро NT. Цикл начинается с входа в безопасное ядро через процедуру `HvlSwitchToVsmVtl1`, определяющую операцию, запрошенную вызывающей стороной. Последняя функция, которая возвращает результат только в том случае, если безопасное ядро запрашивает переключение VTL, сохраняет все общие регистры и копирует всю структуру данных SKCALL в некоторые общеизвестные регистры процессора: RBX и регистры SSE с XMM10 по XMM15. Наконец, она отправляет гипервызов `HvVtlCall` гипервизору. Последний переключается на целевой VTL, загрузив сохраненную VMCS для каждого VTL, и записывает причину входа безопасного вызова VTL на страницу управления VTL. Действительно, чтобы иметь возможность определить, почему произошел вход на безопасный VTL, гипервизор поддерживает информационную страницу памяти, общую для всех защищенных VTL. Эта страница используется для двунаправленной связи между гипервизором и кодом, выполняющимся в безопасном VTL на виртуальном процессоре.

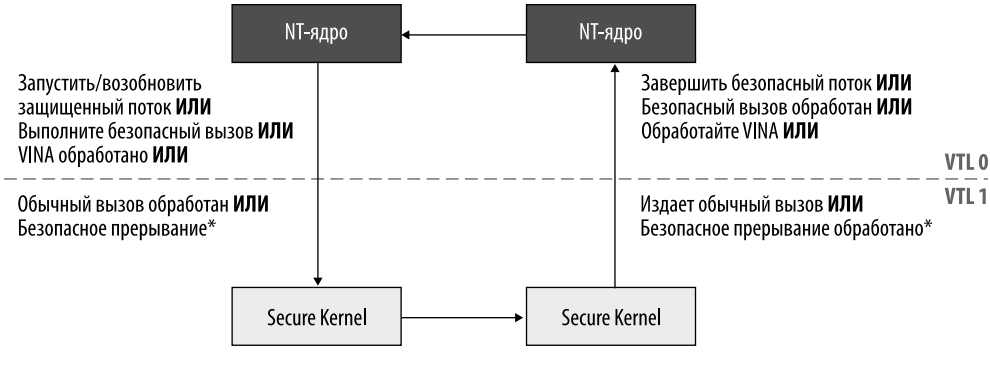


Рис. 9.33. Цикл диспетчеризации VSM

Виртуальный процессор перезапускает выполнение в контексте VTL 1 в функции `SkCallNormalMode` безопасного ядра. Код считывает причину входа в VTL: если это небезопасное прерывание, оно загружает текущий процессор SKPRCB (блок

управления процессором безопасного ядра), выбирает поток для запуска (начиная с файла cookie безопасного потока), копирует содержимое структуры данных SKCALL из общего ЦП и регистрируется в буфере памяти. Наконец, он вызывает процедуру диспетчера `IumInvokeSecureService`, которая будет обрабатывать запрошенный безопасный вызов, отправляя вызов правильной функции, и реализует часть цикла диспетчеризации в VTL 1.

Важно понимать, что безопасное ядро может отображать память VTL 0 и получать к ней доступ, поэтому нет необходимости маршалировать и копировать любую возможную структуру данных, на которую указывает один или несколько параметров, в память VTL 1. Эту концепцию нельзя применять к обычному вызову, о чем поговорим в следующем разделе.

Как мы видели в предыдущем разделе, безопасные прерывания (и перехваты) отправляются гипервизором, который вытесняет любой код, выполняющийся в VTL 0. В этом случае, начиная выполнение, код VTL 1 отправляет прерывание в подходящий ISR. После завершения ISR безопасное ядро немедленно отправляет гипервызов `HvVtlReturn`. В результате код в VTL 0 перезапускает выполнение в той точке, в которой оно было ранее прервано и которая не находится в цикле диспетчеризации безопасных вызовов. Следовательно, безопасные прерывания не являются частью цикла диспетчеризации, даже если они все еще выполняют переключение VTL.

Обычные вызовы

Обычными вызовами управляют аналогично безопасным вызовам (с аналогичным циклом диспетчеризации, действующим в VTL 1 и называемым *циклом обычных вызовов*), но с некоторыми важными отличиями.

- Все общие регистры VTL надежно очищаются безопасным ядром перед отправкой `HvVtlReturn` гипервизору для переключения VTL. Это предотвращает утечку любых защищенных данных в обычный режим.
- Обычное ядро NT не может читать защищенную память VTL 1. Для правильной передачи параметров системного вызова и структур данных, необходимых для обычного вызова, требуется буфер памяти, который могут совместно использовать как безопасное, так и обычное ядро. Ядро безопасности выделяет этот общий буфер с помощью обычного вызова `ALLOCATE_VM`, который не требует передачи какого-либо указателя в качестве параметра. Последний передается функции `MmAllocateVirtualMemory` в обычном ядре NT. Выделенная память переназначается в безопасное ядро по тому же виртуальному адресу и становится частью общего пула памяти безопасного процесса.
- Как мы обсудим позже в этой главе, изолированный пользовательский режим (IUM) изначально проектировался для того, чтобы иметь возможность запускать специальные исполняемые файлы Win32, которые должны были работать независимо как в обычной, так и в безопасной среде. Стандартные неизмененные библиотеки Ntdll.dll и KernelBase.dll отображаются даже в IUM. Этот факт имеет важное последствие: требуется, чтобы почти все собственные API NT, от которых зависят Kernel32.dll и многие другие библиотеки пользовательского режима, были проксированы безопасным ядром.

Для корректного решения описанных проблем в безопасное ядро включен маршализатор, идентифицирующий и корректно копирующий в общий буфер структуры данных, на которые указывают параметры NT API. Маршализатор также может определить размер общего буфера, который будет выделен из пула памяти безопасного процесса. Безопасное ядро определяет три типа обычных вызовов.

- **Запрещенный обычный вызов** не реализован в безопасном ядре, и если он делается из IUM, то просто завершается с ошибкой с кодом завершения `STATUS_INVALID_SYSTEM_SERVICE`. Этот тип вызова не может быть задействован непосредственно самим безопасным ядром.
- **Разрешенный обычный вызов** реализован только в ядре NT и может быть сделан из IUM в исходной версии Nt или Zw (через Ntdll.dll). Даже безопасное ядро может запросить разрешенный обычный вызов — но только с помощью небольшого кода-заглушки, который загружает номер обычного вызова, — установить старший бит в номере и вызвать диспетчер обычных вызовов (процедура `IumGenericSyscall`). Самый старший бит идентифицирует обычный вызов, требуемый самим безопасным ядром, а не модулем Ntdll.dll, загруженным в IUM.
- **Специальный обычный вызов** частично или полностью реализован в безопасном ядре (VTL 1), которое может фильтровать результаты исходной функции или полностью перепроектировать ее код.

Разрешенные и специальные обычные вызовы могут быть помечены как `KernelOnly`. В последнем случае обычный вызов можно запросить только у самого безопасного ядра, а не у защищенных процессов. В главе 3 тома 1, в разделе «Системные вызовы, доступные через трастлет», мы представили список разрешенных и специальных обычных вызовов, которые можно делать из программного обеспечения, работающего в VSM.

На рис. 9.34 показан пример специального обычного вызова. В нем трастлет LsaIso вызвал собственный API `NtQueryInformationProcess` для запроса информации о конкретном процессе. Ntdll.dll, отображенный в IUM, подготавливает номер системного вызова и выполняет инструкцию SYSCALL, которая передает поток выполнения глобальному диспетчеру системных вызовов `KiSystemServiceStart`, находящемуся в безопасном ядре (VTL 1). Диспетчер глобальных системных вызовов распознает, что номер системного вызова принадлежит обычному вызову, и использует его для доступа к массиву `IumSyscallDispatchTable`, который представляет таблицу диспетчеризации обычных вызовов.

Таблица диспетчеризации обычных вызовов включает массив сжатых записей, которые генерируются в фазе 0 запуска безопасного ядра (обсуждается далее в этой главе). Каждая запись содержит смещение целевой функции, рассчитанное относительно самой таблицы, и количество ее аргументов с некоторыми флагами. Все смещения в таблице изначально продумываются так, чтобы указывать на обычную процедуру диспетчера вызовов (`IumGenericSyscall`). После первого цикла инициализации процедура запуска безопасного ядра исправляет каждую запись, представляющую специальный вызов. Новое смещение указывает на часть кода, которая реализует обычный вызов в безопасном ядре.

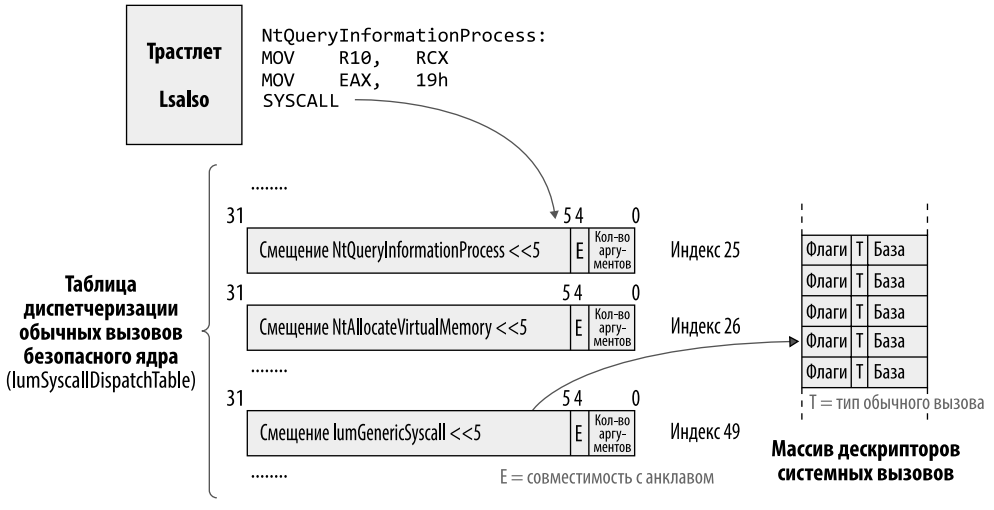


Рис. 9.34. Трастлет, выполняющий специальный обычный вызов API NtQueryInformationProcess

В результате на рис. 9.34 диспетчер глобальных системных вызовов передает выполнение части функции `NtQueryInformationProcess`, реализованной в безопасном ядре. Последнее проверяет, является ли запрошенный класс информации одним из небольших подмножеств, предоставляемых безопасному ядру, и если да, то использует небольшой код-заглушку для вызова обычной процедуры диспетчера вызовов (`IumGenericSyscall`).

На рис. 9.35 показан номер селектора системного вызова для API `NtQueryInformationProcess`. Обратите внимание на то, что заглушка устанавливает старший бит (бит N) номера системного вызова, чтобы указать: обычный вызов запрашивается безопасным ядром. Обычный диспетчер вызовов проверяет параметры и вызывает маршализатор, который может маршалировать каждый аргумент и копировать его в правильное смещение общего буфера. В селекторе есть еще один бит, который дополнительно различает обычный вызов и безопасный системный вызов (обсуждается позже в этой главе).

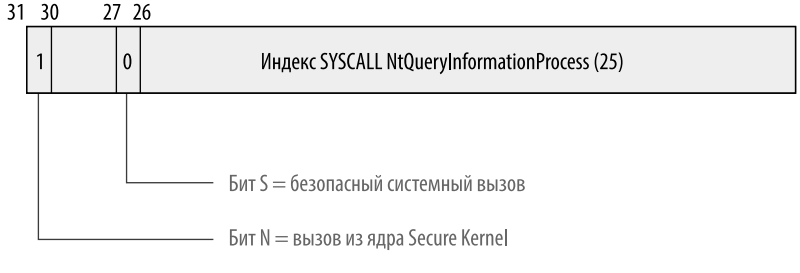


Рис. 9.35. Номер селектора системного вызова безопасного ядра

Маршализатор работает благодаря двум важным массивам, описывающим каждый обычный вызов: массиву дескрипторов (показан в правой части рис. 9.34)

и массиву дескрипторов аргументов. Из них маршализатор может получить всю необходимую информацию: тип обычного вызова, индекс функции маршалинга, тип аргумента, размер и тип данных, на которые указывает (если аргумент является указателем).

После того как общий буфер правильно заполнен маршализатором, безопасное ядро компилирует структуру данных `SKCALL` и входит в обычный цикл диспетчера вызовов (`SkCallNormalMode`). Эта часть цикла сохраняет и очищает все общие регистры виртуального процессора, отключает прерывания и перемещает контекст потока в поток PRCB (подробнее о планировании потоков — далее в этой главе). Затем копирует содержимое структуры данных `SKCALL` в некоторый общий регистр. На заключительном этапе вызывает гипервизор посредством гипервызова `HvVtlReturn`.

Затем выполнение кода возобновляется в цикле диспетчеризации безопасных вызовов в VTL 0. Если в очереди есть ожидающие прерывания, они обрабатываются как обычно (если IRQL это позволяет). Цикл распознает запрос на нормальную операцию вызова и вызывает функцию `NtQueryInformationProcess`, реализованную в VTL 0. После того как функция завершила обработку, цикл перезапускается и снова входит в безопасное ядро (как и для безопасных вызовов), все еще через процедуру `HvlSwitchToVsmVtl1`, но с запросом другой операции — `Resume thread` (Возобновить поток). Это, как следует из названия, позволяет безопасному ядру переключиться на исходный защищенный поток и продолжить выполнение, прерванное для обычного вызова.

Реализация разрешенных обычных вызовов такая же, за исключением того, что они имеют записи в таблице диспетчеризации обычных вызовов, указывающие непосредственно на процедуру диспетчера обычных вызовов `IumGenericSyscall`. Таким образом, код будет передан непосредственно обработчику, пропуская любой код реализации API в безопасном ядре.

Безопасные системные вызовы

Последний тип системных вызовов, доступных в безопасном ядре, аналогичен стандартным системным вызовам, предоставляемым ядром NT для программного обеспечения пользовательского режима VTL 0. Безопасные системные вызовы применяются для предоставления сервисов только защищенным процессам (трастлетам). Программное обеспечение VTL 0 никоим образом не может выполнять безопасные системные вызовы. Как мы обсудим в разделе «Изолированный пользовательский режим» далее в этой главе, каждый трастлет сопоставляет Dll собственного уровня IUM (Iumdll.dll) в своем адресном пространстве. Iumdll.dll выполняет ту же задачу, что и его аналог в VTL 0, Ntdll.dll, — реализует встроенные функции-заглушки системных вызовов для приложения пользовательского режима. Заглушка копирует номер системного вызова в регистр и выдает инструкцию SYSCALL (в зависимости от платформы она задействует разные коды операций).

В номерах безопасных системных вызовов 28-й бит всегда установлен в 1 (в архитектурах AMD64, тогда как ARM64 использует 16-й бит). Таким образом, глобальный диспетчер системных вызовов (`KiSystemServiceStart`) распознает, что номер

системного вызова принадлежит защищенному системному вызову, а не обычному вызову и переключается на `SkiSecureServiceTable`, который представляет таблицу диспетчеризации безопасных системных вызовов. Как и в случае с обычными вызовами, глобальный диспетчер проверяет соответствие номера вызова лимиту, выделяет место в стеке для аргументов (при необходимости), вычисляет конечный адрес системного вызова и передает ему выполнение кода.

В целом выполнение кода остается в VTL 1, но текущий уровень привилегий виртуального процессора повышается с 3 (пользовательский режим) до 0 (режим ядра). Таблица диспетчеризации безопасных системных вызовов сжимается — аналогично таблице диспетчеризации обычных вызовов — в фазе 0 запуска безопасного ядра. Однако все записи в ней действительны и указывают на функции, реализованные в безопасном ядре.

Защищенные потоки и планирование

Как мы объясним в разделе «Изолированный пользовательский режим», исполнительными блоками в VSM являются защищенные потоки, которые существуют в адресном пространстве, описываемом защищенным процессом. Защищенные потоки могут быть потоками режима ядра или пользовательского режима. VSM поддерживает строгое соответствие между каждым защищенным потоком пользовательского режима и обычным потоком, живущим в VTL 0.

Действительно, планирование потоков безопасного ядра полностью зависит от обычного ядра NT — безопасное не включает собственный планировщик (проектировалось, что поверхность атаки безопасного ядра должна быть небольшой). В главе 3 тома 1 мы описали, как ядро NT создает процесс и относительный начальный поток. В пункте, описывающем этап 4, «Создание исходного потока, его стека и контекста», объясняется, что создание потока состоит из двух этапов.

- Создается объект потока исполнительной системы — выделяются его ядро и пользовательский стек. Процедура `KeInitThread` вызывается для настройки начального контекста потока для потоков пользовательского режима. `KiStartUserThread` — это первая процедура, которая будет выполнена в контексте нового потока, она снизит IRQL потока и вызовет `PspUserThreadStartup`.
- Затем управление возвращается в функцию `NtCreateUserProcess`, которая на более позднем этапе вызывает `PspInsertThread` для завершения инициализации потока и вставки его в пространство имен диспетчера объектов.

В ходе своей работы функция `PspInsertThread`, обнаружив, что поток принадлежит защищенному процессу, вызывает `VslCreateSecureThread`, которая, как следует из названия, использует безопасный вызов службы Create Thread, чтобы попросить безопасное ядро создать связанный защищенный поток. Безопасное ядро проверяет параметры и получает структуру данных безопасного образа процесса (подробнее об этом позже в этой главе). Затем оно выделяет объект защищенного потока и его TEB, создает начальный контекст потока (первая выполняемая процедура — `SkpUserThreadStartup`) и, наконец, делает поток планируемым. Более того,

обработчик безопасной службы в VTL 1 после маркировки потока как готового к выполнению возвращает особое значение конкретного потока, которое хранится в структуре данных ETHREAD.

Новый защищенный поток по-прежнему запускается в VTL 0. Как описано в разделе «Этап 7» главы 3 тома 1, функция `PspUserThreadStartup` выполняет окончательную инициализацию пользовательского потока в новом контексте. Если она определяет, что процесс-владелец потока является трастлетом, то запускает функцию `VslStartSecureThread`, которая вызывает цикл диспетчеризации безопасных вызовов через процедуру `VslpEnterIumSecureMode` в VTL 0, передавая файл cookie безопасного потока, возвращенный обработчиком безопасной службы Create Thread. Первая операция, которую цикл диспетчеризации запрашивает у безопасного ядра, — это возобновление выполнения защищенного потока (все еще через гипервызов `HvVtlCall`).

Безопасное ядро до перехода на VTL 0 выполняло код в обычном цикле диспетчера вызовов (`SkCallNormalMode`). Гипервызов, реализуемый обычным ядром, перезапускает выполнение той же процедуры цикла. Цикл диспетчера VTL 1 распознает запрос на возобновление нового потока — переключает контекст выполнения на новый защищенный поток, присоединяется к его адресному пространству и делает его работоспособным. В ходе переключения контекста выбирается новый стек, который был предварительно инициализирован безопасным вызовом Create Thread. Последний содержит адрес первой системной функции защищенного потока, `SkpUserThreadStartup`, которая, как и в случае обычных потоков NT, устанавливает начальный контекст преобразования для запуска процедуры инициализации загрузчика образа (`LdrInitializeThunk` в Ntdll.dll).

После запуска новый защищенный поток может вернуться в нормальный режим по двум основным причинам: он генерирует обычный вызов, который необходимо обработать в VTL 0, или прерывания VINA останавливают выполнение кода. Несмотря на то что эти два случая обрабатываются немного по-разному, оба они приводят к выполнению обычного цикла диспетчера вызовов (`SkCallNormalMode`).

Как обсуждалось ранее в главе 4 тома 1, планировщик NT работает благодаря тактовой частоте процессора, которая генерирует прерывание каждый раз, когда срабатывают системные часы (обычно каждые 15,6 мс). Процедура обслуживания прерывания часов обновляет время процессора и вычисляет, истекает ли квант потока. Прерывание адресовано VTL 0, поэтому, когда виртуальный процессор выполняет код в VTL 1, гипервизор внедряет прерывание VINA в безопасное ядро (рис. 9.36). Прерывание VINA вытесняет текущий исполняемый код, понижает IRQL до значения IRQL предыдущего вытесненного кода и генерирует обычный вызов VINA для ввода VTL 0.

В качестве стандартного процесса обычной диспетчеризации вызовов безопасное ядро перед тем, как выдать гипервызов `HvVtlReturn`, отменяет выбор текущего потока выполнения из PRCB виртуального процессора. Это важно: VP в VTL 1 больше не привязан ни к какому контексту потока, и в следующем цикле безопасное ядро может переключиться на другой поток или принять решение о переносе выполнения текущего.

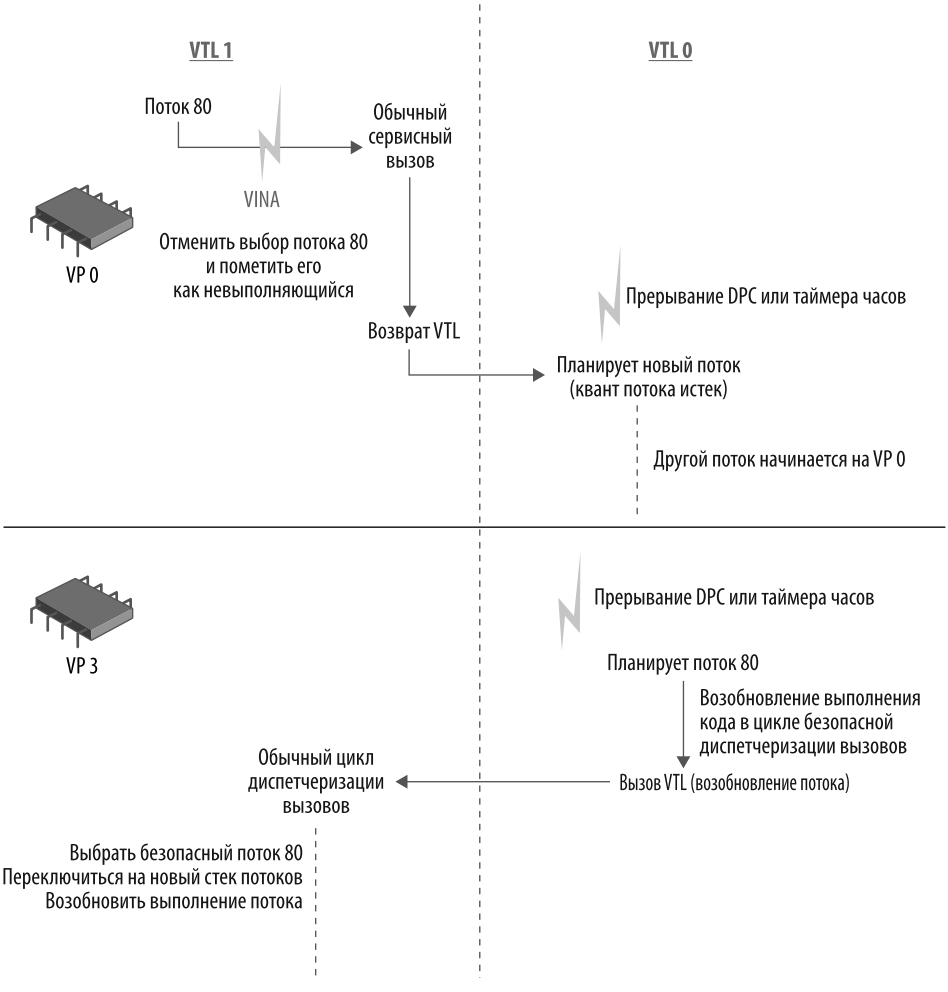


Рис. 9.36. Схема планирования защищенных потоков

После переключения VTL ядро NT возобновляет выполнение в цикле диспетчеризации безопасных вызовов, все еще в контексте нового потока. Прежде чем у него появится шанс выполнить какой-либо код, он вытесняется процедурой обработчика прерываний таймера, которая может вычислить новое значение такта и, если оно истекло, переключить выполнение другого потока. Когда происходит переключение контекста и другой поток входит в VTL 1, обычный цикл диспетчеризации вызовов планирует другой защищенный поток в зависимости от значения особого значения защищенного потока:

- безопасный поток из пула защищенных потоков, если обычное ядро NT ввело VTL 1 для отправки безопасного вызова (в этом случае особое значение защищенного потока равно 0);

- вновь созданный безопасный поток, если поток был перепланирован для выполнения (особое значение защищенного потока является допустимым значением). Как показано на рис. 9.36, новый поток может быть перепланирован также другим виртуальным процессором (в примере VP 3).

В описанной схеме все решения по планированию выполняются только в VTL 0. Цикл безопасного вызова и цикл обычного вызова взаимодействуют для правильного переключения контекста защищенного потока в VTL 1. Все защищенные потоки имеют связанный поток в обычном ядре. А обратное неверно: если обычный поток в VTL 0 решает выполнить безопасный вызов, безопасное ядро отправляет запрос, используя произвольный контекст потока из пула потоков.

Hypervisor-Enforced Code Integrity (HVCI)

Hypervisor-Enforced Code Integrity (защищенная гипервизором целостность кода) — это функция, которая обеспечивает работу Device Guard и обеспечивает характеристику W^X памяти ядра VTL 0. Ядро NT не может отображать и выполнять какие-либо исполняемые файлы в режиме ядра без помощи безопасного ядра. Последнее позволяет запускать в ядре машины только правильные драйверы с цифровой подписью. Как мы обсудим в следующем разделе, безопасное ядро отслеживает каждую виртуальную страницу, выделенную в обычном ядре NT. Страницы памяти, помеченные в ядре NT как исполняемые, считаются *привилегированными*. Только безопасное ядро может писать в них после того, как модуль SKCI корректно проверит их содержимое.

Подробнее о HVCI можно прочитать в главе 7 тома 1, в подразделах «Device Guard» и «Охранник учетных данных».

Виртуализация UEFI во время выполнения

Еще один сервис, предоставляемый безопасным ядром (при включенной HVCI), — это возможность виртуализации и защиты служб времени выполнения UEFI. Как мы обсудим в главе 12, службы прошивки UEFI в основном реализуются с использованием большой таблицы указателей функций. Часть таблицы будет удалена из памяти после того, как ОС возьмет на себя управление и вызовет функцию `ExitBootServices`, но другая часть таблицы, представляющая службы времени выполнения, останется сопоставленной даже после того, как ОС взяла на себя полный контроль над машиной. Действительно, это необходимо, поскольку иногда ОС требуется взаимодействовать с конфигурацией и службами UEFI.

Каждый поставщик оборудования реализует собственную прошивку UEFI. При использовании HVCI ожидается, что встроенное ПО будет сотрудничать, чтобы обеспечить недоступное для записи состояние каждой из своих исполняемых страниц памяти (ни одна страница встроенного ПО не может быть отображена в VTL 0 с состояниями чтения, записи и выполнения). Диапазон памяти, в котором находится прошивка UEFI, описывается несколькими структурами данных `MEMORY_DESCRIPTOR`, расположенными в отраженной памяти EFI. Загрузчик Windows анализирует эти данные, чтобы обеспечить надлежащую защиту памяти прошивки UEFI. К сожалению, в исходной реализации UEFI код и данные хранились

смешанными в одном разделе (или нескольких разделах) и описывались относительными дескрипторами памяти. Более того, некоторые драйверы устройств считывают данные конфигурации непосредственно из областей памяти UEFI или записывают в них. Это явно несовместимо с HVCI.

Для решения этой проблемы безопасное ядро использует такие две стратегии.

- Новые версии прошивки UEFI, соответствующие спецификациям UEFI 2.6 и выше, поддерживают новую таблицу конфигурации, связанную с таблицей служб загрузки и называемую таблицей атрибутов памяти (MAT). MAT определяет детализированные разделы области памяти UEFI, которые являются подразделами дескрипторов памяти, определенных картой памяти EFI. Каждый раздел никогда не имеет атрибута защиты одновременно исполняемого и записываемого файлов.
- Для старой прошивки безопасное ядро отображает в VTL 0 всю физическую память региона прошивки UEFI с правом доступа только для чтения.

При реализации первой стратегии во время загрузки загрузчик Windows объединяет информацию, найденную как в отраженной памяти EFI, так и в MAT, создавая массив дескрипторов памяти, которые точно описывают всю область прошивки. Затем он копирует их в зарезервированный буфер, расположенный в VTL 1 (используемый в пути гибернации), и проверяет то, что каждый раздел прошивки не нарушает предположения W^X. Если это так, то при запуске безопасное ядро использует надлежащую защиту SLAT для каждой страницы, принадлежащей базовой области прошивки UEFI. Физические страницы защищены SLAT, но их виртуальное адресное пространство в VTL 0 по-прежнему полностью помечено как RWX. Сохранение защиты RWX виртуальной памяти важно, поскольку безопасное ядро должно поддерживать выход из режима гибернации в сценарии, когда защита, примененная в записях MAT, может измениться. Более того, это обеспечивает совместимость со старыми драйверами, которые читают непосредственно из области памяти UEFI или записывают в нее, при условии, что запись выполняется в правильных разделах. (Кроме того, код UEFI должен иметь возможность записи в собственную память, которая отображается в VTL 0.) Эта стратегия позволяет безопасному ядру избежать отображения любого кода прошивки в VTL 1, а единственная часть прошивки, которая остается в VTL 1, — это сама таблица функций времени выполнения. Сохранение таблицы в VTL 1 позволяет коду возобновления работы из режима гибернации напрямую обновлять указатель функции служб среды выполнения UEFI.

Вторая стратегия не является оптимальной и применяется только для того, чтобы позволить старым системам работать с включенным HVCI. Когда безопасное ядро не находит MAT в прошивке, у него нет другого выхода, кроме как сопоставить весь код служб времени выполнения UEFI в VTL 1. Исторически в коде прошивки UEFI, особенно в SMM, обнаруживалось множество ошибок. Использовать прошивку в VTL 1 может быть опасно, но это единственное решение, совместимое с HVCI. (Новые системы, как уже говорилось, никогда не отображают какой-либо код прошивки UEFI в VTL 1.) Во время запуска NT HAL обнаруживает, что HVCI включен и прошивка полностью отображается в VTL 1. Поэтому он переключает свою внутреннюю службу EFI (указатель таблицы) на новую таблицу, называемую

таблицей-оболочкой UEFI. Записи таблицы-оболочки содержат подпрограммы-заглушки, которые используют безопасный вызов `INVOKE_EFI_RUNTIME_SERVICE` для входа в VTL 1. Безопасное ядро маршалирует параметры, выполняет вызов прошивки и выдает результаты в VTL 0. В этом случае вся физическая память, описывающая прошивку UEFI, по-прежнему отображается в режиме только для чтения в VTL 0. Цель состоит в том, чтобы позволить драйверам правильно считывать информацию из области памяти прошивки UEFI, например таблицы ACPI. Старые драйверы, которые напрямую записывают данные в области памяти UEFI, в этом сценарии несовместимы с HVCI.

Когда безопасное ядро выходит из режима гибернации, оно обновляет таблицу служб UEFI в памяти, чтобы указать расположение новых служб. Кроме того, в системах с новой прошивкой UEFI безопасное ядро повторно применяет защиту SLAT к каждой области памяти, отображенной в VTL 0 (загрузчик Windows может при необходимости изменять виртуальные адреса областей).

Запуск VSM

Хотя мы описываем весь механизм запуска и завершения работы Windows в главе 12, в этом разделе рассматривается способ запуска безопасного ядра и всей инфраструктуры VSM. Правильный запуск безопасного ядра зависит от гипервизора, загрузчика Windows и ядра NT. В главе 12 мы обсудим загрузчик Windows, загрузчик гипервизора и предварительные этапы, на которых безопасное ядро инициализируется в VTL 0 этими двумя модулями. Здесь же сосредоточимся на методе запуска VSM, который реализован в исполняемом файле Securekernel.exe.

Первый код, выполняемый двоичным файлом Securekernel.exe, все еще действует в VTL 0 — гипервизор уже запущен, и таблицы страниц, используемые для VTL 1, построены. Безопасное ядро инициализирует следующие компоненты в VTL 0:

- функцию инициализации диспетчера памяти, которая сохраняет PFN структуры уровня страницы корневого уровня VTL 0 и данные целостности кода, включает HVCI, MBEC (управление выполнением на основе режима), конфигурацию ядра и горячее исправление;
- компоненты ЦП, специфичные для общей архитектуры, такие как GDT и IDT;
- таблицы диспетчеризации обычных вызовов и защищенных системных вызовов (инициализация и сжатие);
- загрузочный процессор. Процесс запуска загрузочного процессора требует, чтобы безопасное ядро выделило свое ядро и стеки прерываний, инициализировало компоненты, зависящие от архитектуры, которые не могут использоваться разными процессорами (например, TSS), и, наконец, выделило SKPRCB процессора. Последняя представляет собой важную структуру данных, которая, как и структура данных PRCB в VTL 0, применяется для хранения важной информации, связанной с каждым процессором.

Код инициализации безопасного ядра готов к первому переходу в VTL 1. Функция инициализации подсистемы гипервизора (процедура `ShvlInitSystem`) подключается к гипервизору (через классы CPUID гипервизора, подробности

в предыдущем разделе) и проверяет поддерживаемые сведения. Затем он сохраняет таблицу страниц VTL 1, ранее созданную загрузчиком Windows, и выделенные страницы гипервызова, используемые для хранения параметров гипервызова. Наконец, он инициализируется и входит в VTL 1 следующим образом.

1. Включает VTL 1 для текущего раздела гипервизора посредством гипервызова `HvEnablePartitionVtl`. Гипервизор копирует существующую таблицу SLAT обычного VTL в VTL 1 и активирует MBEC и новый VTL 1 для данного раздела.
2. Включает VTL 1 для загрузочного процессора посредством гипервызова `HvEnableVpVtl`. Гипервизор инициализирует новую структуру данных VMCS для каждого уровня, компилирует ее и устанавливает таблицу SLAT.
3. Запрашивает у гипервизора расположение зависящего от платформы кода гипервызова `VtlCall` и `VtlReturn`. Коды операций ЦП, необходимые для выполнения вызовов VSM, скрыты от реализации безопасного ядра. Это позволяет большей части кода безопасного ядра быть независимым от платформы. Наконец, безопасное ядро выполняет переход к VTL 1 посредством гипервызова `HvVtlCall`. Гипервизор загружает VMCS для нового VTL и переключается на него (активизирует его). По сути, это делает новый VTL работоспособным.

Безопасное ядро запускает сложную процедуру инициализации в VTL 1, которая по-прежнему зависит от загрузчика Windows и ядра NT. Стоит отметить, что на этом этапе память VTL 1 все еще сопоставлена с идентификаторами в VTL 0, а безопасное ядро и зависимые от него модули по-прежнему доступны обычному миру. После переключения на VTL 1 безопасное ядро инициализирует загрузочный процессор.

1. Получает виртуальный адрес общей страницы контроллера синтетических прерываний, страницы поддержки TSC и VP, которые предоставляются гипервизором для обмена данными между гипервизором и кодом VTL 1. Сопоставляет в VTL 1 страницу гипервызовов.
2. Блокирует возможность запуска других виртуальных процессоров системы с помощью более низкого VTL и запрашивает обнуление памяти при перезагрузке гипервизора.
3. Инициализирует и заполняет таблицу дескрипторов прерываний загрузочного процессора (IDT). Настраивает IPI, обратные вызовы и обработчики прерываний безопасного таймера и устанавливает текущий защищенный поток в качестве потока SKPRCB по умолчанию.
4. Запускает диспетчер защищенной памяти VTL 1, который создает сопоставление таблицы загрузки и отображает память загрузчика в VTL 1, формирует защищенную базу данных PFN и системное гиперпространство, инициализирует поддержку пула защищенной памяти и считывает блок загрузчика VTL 0, чтобы скопировать дескрипторы модулей импортированных образов безопасного ядра (Skci.dll, Cnf.sys и Vmsvcext.sys). Наконец, он просматривает список загруженных модулей NT, чтобы установить состояние каждого драйвера, создавая структуру данных NAR (нормальный диапазон адресов) для них и компилируя запись нормальной таблицы (NTE) для каждой страницы, составляющей разделы

загрузочного драйвера. Кроме того, функция инициализации диспетчера защищенной памяти применяет правильную защиту SLAT VTL 0 к каждому разделу драйвера.

5. Инициализирует HAL, пул безопасных потоков, подсистему процессов, синтетический APIC, Secure PNP и Secure PCI.
6. Применяет защиту SLAT VTL 0 только для чтения в отношении страниц безопасного ядра, настраивает MBEC и включает виртуальное прерывание VINA на загрузочном процессоре.

Когда эта часть инициализации завершается, безопасное ядро прекращает отображение загрузочной памяти. Диспетчер защищенной памяти, как мы обсудим в следующем разделе, зависит от диспетчера памяти VTL 0, поскольку он способен выделять и освобождать память VTL 1. VTL 1 не имеет физической памяти — на этом этапе он использует некоторые ранее выделенные загрузчиком Windows физические страницы для удовлетворения запросов на выделение памяти. При последующем запуске ядра NT безопасное ядро выполняет обычные вызовы для запроса служб памяти к диспетчеру памяти VTL 0. В результате некоторые фазы инициализации безопасного ядра после запуска ядра NT необходимо отложить. Поток выполнения возвращается к загрузчику Windows в VTL 0. Последний загружает и запускает ядро NT. Завершающая часть инициализации безопасного ядра происходит в фазах 0 и 1 инициализации ядра NT (более подробную информацию см. в главе 12).

В фазе 0 инициализации ядра NT все еще нет доступных служб памяти, но это последний момент, когда безопасное ядро полностью доверяет нормальной среде. Загруженные драйверы еще не инициализированы, а первоначальный процесс загрузки уже должен быть защищен с помощью Secure Boot. Обработчик безопасных вызовов PHASE3_INIT изменяет защиту SLAT всех физических страниц, принадлежащих безопасному ядру и зависимым от него модулям, делая их недоступными для VTL 0. Кроме того, он применяет защиту только для чтения к битовым картам CFG ядра. На этом этапе безопасное ядро обеспечивает поддержку целостности файла подкачки, создает начальный системный процесс и его адресное пространство, а также сохраняет все доверенные значения общих регистров ЦП, например IDT, GDT, Syscall MSR и т. д. Структуры данных, на которые указывают общие регистры, проверяются благодаря базе данных NTE. Наконец, запускается пул защищенных потоков и инициализируются диспетчер объектов, модуль целостности безопасного кода (Skci.dll) и HyperGuard (более подробную информацию о HyperGuard можно найти в главе 7 тома 1).

Когда поток выполнения возвращается в VTL 0, ядро NT может запустить все остальные процессоры приложений (AP). Когда безопасное ядро активно, инициализация точки доступа происходит немного по-другому (мы обсудим ее в следующем разделе).

В фазе 1 инициализации ядра NT система запускает диспетчер ввода-вывода. Как обсуждалось в главе 6 тома 1, он является ядром системы ввода-вывода и определяет модель, в рамках которой запросы ввода-вывода доставляются драйверам устройств. Одна из обязанностей диспетчера ввода-вывода — инициализация и запуск драйверов, загружаемых при запуске системы ELAM. Перед созданием специальных разделов для сопоставления системных DLL пользовательского режима

функция инициализации диспетчера ввода-вывода генерирует безопасный вызов `PHASE4_INIT`, чтобы начать последнюю фазу инициализации безопасного ядра. В этой фазе безопасное ядро больше не доверяет VTL 0, но может задействовать службы, предоставляемые диспетчером памяти NT. Безопасное ядро инициализирует содержимое страницы данных Secure User Shared, которая отображается как в пользовательском режиме VTL 1, так и в режиме ядра, и завершает инициализацию исполнительной подсистемы. Оно освобождает все ресурсы, зарезервированные во время загрузки, вызывает каждую из собственных зависимых точек входа модуля, в частности cng.sys и vmsvcext.sys, которые запускаются перед любыми обычными загрузочными драйверами. Затем выделяет необходимые ресурсы для шифрования файла гибернации, аварийного дампа, файлов подкачки и обеспечения целостности страниц памяти. Наконец, оно считывает и отображает файл схемы набора API в памяти VTL 1. На этом этапе VSM полностью инициализируется.

Запуск процессоров приложений

Одна из функций безопасности, обеспечиваемых безопасным ядром, — это запуск процессоров приложений (AP), которые не используются для загрузки системы. При запуске системы спецификации Intel и AMD для архитектур x86 и AMD64 определяют точный алгоритм выбора процессора начальной загрузки (BSP) в многопроцессорных системах. Загрузочный процессор всегда запускается в 16-разрядном реальном режиме (где он имеет доступ только к 1 Мбайт физической памяти) и обычно выполняет код прошивки машины, в большинстве случаев UEFI, который должен быть расположен в определенном месте физической памяти. Это местоположение называется вектором сброса. Загрузочный процессор выполняет почти всю инициализацию ОС, гипервизора и безопасного ядра. Для запуска других незагрузочных процессоров системе необходимо отправить специальный IPI (межпроцессорное прерывание) локальным APIC, принадлежащим каждому процессору. Вектор запуска IPI (SIPI) содержит адрес физической памяти стартового блока процессора — блока кода, включающего в себя инструкции по выполнению следующих основных операций.

1. Загрузить GDT и переключиться из 16-разрядного реального режима в 32-разрядный защищенный режим (без включенной подкачки).
2. Установить базовую таблицу страниц, включить подкачку и войти в 64-разрядный длинный режим.
3. Загрузить 64-разрядные IDT и GDT, установить соответствующие регистры процессора и перейти к функции запуска ОС (`KiSystemStartup`).

Этот процесс уязвим для вредоносных атак. Код запуска процессора может быть изменен внешними объектами во время выполнения на процессоре AP (ядро NT на этом этапе ничего не контролирует). В этом случае все перспективы безопасности, предлагаемые VSM, можно было бы легко нарушить. Когда гипервизор и безопасное ядро включены, процессоры приложений по-прежнему запускаются ядром NT, но с использованием гипервизора.

`KeStartAllProcessors` — функция, вызываемая в фазе 1 инициализации ядра NT (подробнее см. в главе 12) с целью запуска всех точек доступа, создает общий IDT и перечисляет все доступные процессоры, сверяясь с таблицей ACPI описания

множественных APIC (Multiple APIC Description Table, MADT). Каждому обнаруженному процессору она выделяет память для PRCB и всех частных структур данных ЦП для ядра и стека DPC. Если VSM включен, то затем запускается точка доступа, отправляя безопасный вызов `START_PROCESSOR` в безопасное ядро. Последнее проверяет правильность всех структур данных, выделенных и заполненных для нового процессора, включая начальные значения регистров процессора и процедуры запуска (`KiSystemStartup`), а также гарантирует, что запуск точек доступа происходит последовательно и только один раз для каждого процессора. Затем оно инициализирует структуры данных VTL 1, необходимые для нового процессора приложений, в частности SKPRCB. Поток PRCB, который используется для отправки безопасных вызовов в контексте нового процессора, запускается, а структуры данных ЦП VTL 0 защищаются с помощью SLAT. Наконец, безопасное ядро включает VTL 1 для нового процессора приложения и запускает его с помощью гипервызова `HvStartVirtualProcessor`. Гипервизор запускает точку доступа способом, аналогичным описанному в начале этого подраздела, — отправкой IPI запуска. Однако в этом случае точка доступа начинает выполнение в контексте гипервизора, переключается на выполнение 64-разрядного длинного режима и возвращается в VTL 1.

Первая функция, выполняемая процессором приложения, находится в VTL 1. Процедура инициализации ЦП безопасного ядра сопоставляет страницу помощи VP для каждого процессора и страницу управления SynIC, настраивает MBEC и включает VINA. Затем он возвращается в VTL 0 через гипервызов `HvVtlReturn`. Первой процедурой, выполняемой в VTL 0, является `KiSystemStartup`, которая инициализирует структуры данных, необходимые ядру NT для управления точкой доступа, инициализирует HAL и входит в цикл ожидания (подробнее см. в главе 12). Цикл диспетчеризации безопасного вызова инициализируется позже обычным ядром NT, когда выполняется первый безопасный вызов.

В этом случае злоумышленник не сможет изменить блок запуска процессора или какое-либо начальное значение регистров и структур данных ЦП. При описанном безопасном запуске точки доступа безопасным ядром будет обнаружено любое изменение и проделана системная проверка на предмет критических ошибок, чтобы отразить любую попытку атаки.

Диспетчер памяти безопасного ядра

Диспетчер памяти безопасного ядра во многом зависит от диспетчера памяти NT (и от диспетчера памяти загрузчика Windows в его коде запуска). Полное описание диспетчера памяти безопасного ядра выходит за рамки этой книги. Здесь мы обсуждаем только наиболее важные концепции и структуры данных, используемые безопасным ядром.

Как упоминалось в предыдущем подразделе, инициализация диспетчера памяти безопасного ядра разделена на три фазы (этапа). В фазе 1, самом важном, диспетчер памяти выполняет следующее.

1. Отображает список дескрипторов памяти встроенного ПО загрузчика в VTL 1, сканирует список и определяет первую физическую страницу, которую он может использовать для выделения памяти, необходимой для первоначального запуска (этот тип памяти называется SLAB). Сопоставляет таблицы страниц

VTL 0 с виртуальным адресом, который расположен ровно на 512 Гбайт раньше таблицы страниц VTL 1. Это позволяет безопасному ядру выполнять быстрое преобразование между виртуальным адресом NT и адресом безопасного ядра.

2. Инициализирует структуры данных диапазона PTE. Этот диапазон содержит битовую карту, которая описывает каждый фрагмент выделенного диапазона виртуальных адресов и помогает ядру безопасности выделять PTE для своего адресного пространства.
3. Создает базу данных Secure PFN и инициализирует пул памяти.
4. Инициализирует разреженную таблицу адресов NT. Для каждого загружаемого драйвера он создает и заполняет NAR, проверяет целостность двоичного файла, заполняет информацию о горячем обновлении и, если HVCI активен, защищает каждый исполняемый раздел драйвера с помощью SLAT. Затем он циклически переключается между всеми PTE образа памяти и записывает таблицу адресов NT (NTE) в таблицу адресов NT.
5. Инициализирует пакеты страниц.

Безопасное ядро отслеживает память, которую использует обычное ядро NT. Диспетчер памяти безопасного ядра задействует структуру данных `NAR` для описания диапазона виртуальных адресов ядра, содержащего исполняемый код. NAR хранит некоторую информацию о диапазоне (например, его базовый адрес и размер) и указатель на структуру данных `SECURE_IMAGE`, которая используется для описания драйверов среды выполнения (как правило, образы, проверенные с помощью безопасного HVCI, включая образы пользовательского режима, применяемые для трастлетов), загруженных в VTL 0. Загруженные при запуске системы драйверы не используют структуру данных `SECURE_IMAGE`, поскольку диспетчер памяти NT обрабатывает их как частные страницы, содержащие исполняемый код. Последняя структура данных содержит информацию о загруженном образе в ядре NT (она проверяется SKCI), например адрес его точки входа, копию его таблиц перемещения (используется также для работы с Retpoline и оптимизацией импорта), указатель на его общие прототипы PTE, информацию о горячем обновлении и структуре данных, которая определяет авторизованное применение его страниц памяти. Структура данных `SECURE_IMAGE` очень важна, поскольку она используется безопасным ядром для отслеживания и проверки страниц общей памяти, с которыми работают драйверы среды выполнения.

Для отслеживания частных страниц VTL 0 безопасное ядро задействует структуру данных NTE. Она существует для каждой виртуальной страницы в адресном пространстве VTL 0, которая требует контроля со стороны безопасного ядра — он часто выполняется для частных страниц. NTE отслеживает PTE виртуальной страницы VTL 0 и сохраняет ее состояние и защиту. Когда HVCI включен, таблица NTE делит все виртуальные страницы на привилегированные и непривилегированные. Привилегированная страница представляет собой страницу памяти, к которой ядро NT не может обратиться самостоятельно, поскольку она защищена с помощью SLAT и обычно соответствует исполняемой странице или странице CFG ядра, доступной только для чтения. К непривилегированным страницам относятся все другие типы страниц памяти, над которыми ядро NT имеет полный контроль. Безопасное ядро использует недопустимые NTE для представления непривилегированных страниц. Когда HVCI выключен, все частные страницы являются непривилегированными (ядро NT действительно полностью контролирует все свои страницы).

В системах с поддержкой HVCI диспетчер памяти NT не может изменять защищенные страницы. В противном случае в гипервизоре возникнет исключение нарушения EPT, что приведет к сбою системы. После того как эти системы завершат этап загрузки, оказывается, что безопасное ядро уже обработало все неисполняемые физические страницы, защищая их SLAT только для доступа на чтение и запись. В этом сценарии новые исполняемые страницы могут быть выделены, только если целевой код проверен безопасным HVCI.

Когда система, приложение или диспетчер Plug and Play требуют загрузки нового драйвера среды выполнения, запускается сложная процедура, в которой участвуют NT и диспетчер памяти безопасного ядра.

1. Диспетчер памяти NT создает объект раздела, выделяет и заполняет новую область управления (более подробную информацию о диспетчере памяти NT можно найти в главе 5 тома 1), считывает первую страницу двоичного файла и вызывает безопасное ядро, чтобы создать относительно безопасный образ, описывающий новый загружаемый модуль.
2. Безопасное ядро создает структуру данных `SECURE_IMAGE`, анализирует все разделы двоичного файла и заполняет массив PTE защищенного прототипа.
3. Ядро NT считывает весь двоичный файл в неисполняемую разделяемую память, на которую указывают прототипные PTE области управления. Вызывает безопасное ядро, которое с помощью HVCI циклически переключается между всеми разделами двоичного образа и вычисляет его окончательный хеш.
4. Если вычисленный хеш файла совпадает с хранящимся в цифровой подписи, диспетчер памяти NT проходит весь образ и для каждой страницы вызывает безопасное ядро, которое проверяет страницу (хеш каждой страницы уже рассчитан на предыдущем этапе), применяет необходимые перемещения (ASLR, Retpoline и оптимизация импорта) и новую защиту SLAT, позволяя странице быть исполняемой, но недоступной для записи.
5. Объект «Раздел» создан. Диспетчеру памяти NT необходимо отобразить драйвер в своем адресном пространстве. Он вызывает безопасное ядро для выделения необходимых привилегированных PTE для описания диапазона виртуальных адресов драйвера. Безопасное ядро создает структуру данных NAR. Затем оно сопоставляет предварительно проверенные физические страницы драйвера с помощью процедуры `MiMapSystemImage`.

ПРИМЕЧАНИЕ Когда NAR инициализируется для драйвера среды выполнения, часть таблицы NTE заполняется для описания нового адресного пространства драйвера. NTE не используются для отслеживания диапазона виртуальных адресов драйвера среды выполнения (его виртуальные страницы общие, а не частные), поэтому относительная часть таблицы адресов NT заполняется недействительными зарезервированными NTE.

В то время как диапазоны виртуальных адресов ядра VTL 0 представлены с помощью структуры данных NAR, безопасное ядро применяет безопасные VAD (дескрипторы виртуальных адресов) для отслеживания виртуальных адресов пользовательского режима в VTL 1. Безопасные VAD создаются каждый раз, когда создается новое частное виртуальное распределение. Двоичный образ отображается

в адресное пространство трастлета (защищенного процесса), а при создании VBS-анклава или модуля — в его адресное пространство. Безопасный VAD аналогичен VAD ядра NT и содержит дескриптор диапазона VA, счетчик ссылок, некоторые флаги и указатель на раздел Secure, созданный SKCI. (Указатель безопасного раздела устанавливается в 0 в случае безопасных VAD, описывающих частные виртуальные распределения.) Более подробная информация о трастлетах и анклавах на основе VBS будет обсуждаться позже в этой главе.

Идентификация страницы и безопасная база данных PFN

Когда драйвер загружен и правильно отображен в память VTL 0, диспетчер памяти NT должен иметь возможность управлять своими страницами памяти, например, при выгрузке страниц выгружаемого раздела драйвера, создании частных страниц, применении частных исправлений и т. д. (подробнее см. в главе 5 тома 1). Каждый раз, когда диспетчер памяти NT работает с защищенной памятью, ему требуется помощь безопасного ядра. Диспетчеру памяти NT для работы с привилегированной памятью предлагаются два основных вида безопасных сервисов: копирование и удаление защищенных страниц.

Структура данных `PAGE_IDENTITY` — это связующее звено, позволяющее ядру безопасности отслеживать все типы страниц. Структура данных состоит из двух полей: контекста адреса и виртуального адреса. Каждый раз, когда ядро NT вызывает безопасное ядро для работы с привилегированными страницами, ему необходимо указать номер физической страницы вместе с допустимой структурой данных `PAGE_IDENTITY`, описывающей, для чего задействуется физическая страница. С помощью этой структуры данных безопасное ядро может проверить использование запрошенной страницы и решить, нужно ли разрешить запрос.

В табл. 9.4 показаны структура данных `PAGE_IDENTITY` (вторая и третья графы), а также все типы проверки, выполняемые безопасным ядром на разных страницах памяти.

- Если безопасное ядро получает запрос на копирование или освобождение общей исполняемой страницы драйвера среды выполнения, оно проверяет дескриптор защищенного образа, указанный вызывающей стороной, и получает его относительную структуру данных (`SECURE_IMAGE`). Затем использует относительный виртуальный адрес (RVA) в качестве индекса в массиве защищенных прототипов для получения фрейма физической страницы (physical page frame, PFN) драйвера. Если найденный PFN равен указанному вызывающей стороной, безопасное ядро разрешает запрос, в противном случае блокирует его.
- Аналогичным образом, если ядро NT запрашивает работу на трастлете или странице анклава (более подробная информация о трастлетах и безопасных анклавах представлена далее в этой главе), безопасное ядро использует указанный виртуальный адрес вызывающей стороны для проверки того, что безопасный PTE в таблице страниц защищенного процесса содержит правильный PFN.
- Как говорилось ранее в данном подразделе, для частных страниц безопасное ядро находит NTE, начиная с указанного виртуального адреса вызывающей стороны, и проверяет, содержит ли он действительный PFN, который должен совпадать с именем вызывающего абонента.

- Страницы-заполнители — это незанятые страницы, защищенные SLAT. Безопасное ядро проверяет состояние страницы-заполнителя, используя базу данных PFN.

Таблица 9.4. Идентификаторы страниц, управляемые безопасным ядром

Тип страницы	Контекст адреса	Виртуальный адрес	Структура проверки
Ядро общее	Дескриптор защиты образа	RVA страницы	Безопасный прототип PTE
Трастлет/анклав	Дескриптор защиты процесса	Виртуальный адрес безопасного процесса	Безопасное PTE
Ядро частное	0	Виртуальный адрес ядра страницы	Запись таблицы адресов NT (NTE)
Заполнитель	0	0	Запись PFN

Диспетчер памяти безопасного ядра поддерживает базу данных PFN для представления состояния каждой физической страницы. Запись PFN в Secure Kernel намного меньше по сравнению с ее эквивалентом в NT и в основном содержит состояние страницы и счетчик общего доступа. Физическая страница с точки зрения безопасного ядра может находиться в одном из следующих состояний: недопустимое, свободное, общее, ввода-вывода, защищенное или образ (защищенный частный NT).

Защищенное состояние используется для физических страниц, которые являются частными для безопасного ядра (ядро NT никогда не может претендовать на них), или для физических страниц, выделенных ядром NT и позже защищенных SLAT с помощью безопасного ядра для хранения исполняемого кода, проверенного HVCI. Только защищенные общие физические страницы имеют идентификатор страницы.

Когда ядро NT собирается выгрузить защищенную страницу, оно запрашивает у безопасного ядра операцию удаления страницы. Безопасное ядро анализирует указанный идентификатор страницы и проверяет его (как объяснялось ранее). Если идентификатор страницы относится к анклаву или странице трастлета, безопасное ядро шифрует содержимое страницы, прежде чем передать его ядру NT, которое затем сохранит страницу в файле подкачки. Таким образом, ядро NT по-прежнему не имеет возможности перехватить реальное содержимое частной памяти.

Безопасное распределение памяти

Как обсуждалось в предыдущих разделах, при первом запуске безопасное ядро анализирует списки дескрипторов памяти встроенного ПО с целью выделения физической памяти для собственного применения. В фазе 1 своей инициализации безопасное ядро не может использовать службы памяти, предоставляемые ядром NT (ядро NT действительно еще не инициализировано), поэтому оно задействует свободные записи списков дескрипторов памяти встроенного ПО для резервирования двухмегабайтной SLAB — непрерывной физической памяти размером 2 Мбайт, которая отображается как одна запись каталога вложенной таблицы страниц в гипервизоре. Все страницы SLAB имеют одинаковую защиту SLAT. SLAB

были разработаны с учетом требований производительности. При сопоставлении фрагмента физической памяти размером 2 Мбайт с использованием одной вложенной записи страницы в гипервизоре преобразование адресов дополнительной аппаратной памяти происходит быстрее и приводит к меньшему количеству промахов в кэше таблицы SLAT.

Первый пакет страниц безопасного ядра заполняется 1 Мбайт выделенной памяти SLAB. Пакет страниц — это структура данных (рис. 9.37), которая содержит список смежных номеров свободных фреймов физических страниц (PFN). Когда безопасное ядро нуждается в памяти для своих целей, оно выделяет физические страницы из пакета страниц, удаляя один или несколько свободных фреймов страниц из хвоста массива PFN пакета. В этом случае безопасному ядру не нужно проверять список дескрипторов памяти прошивки до тех пор, пока пакет не будет полностью израсходован. Когда фаза 3 инициализации безопасного ядра завершена, службы памяти ядра NT становятся доступными, поэтому безопасное ядро освобождает все списки дескрипторов загрузочной памяти, сохраняя страницы физической памяти, ранее расположенные в пакетах.

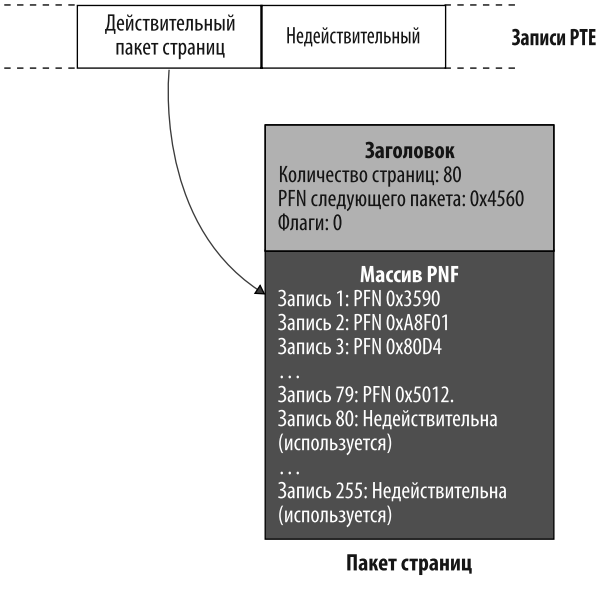


Рис. 9.37. Пакет защищенных страниц с 80 доступными страницами. Он состоит из заголовка и массива свободных PFN

В будущем для безопасного распределения памяти станут использоваться обычные вызовы, предоставляемые ядром NT. Пакеты страниц были разработаны для минимизации количества обычных вызовов, необходимых для выделения памяти. Когда пакет полностью выделен, он не содержит страниц (все его страницы в настоящее время назначены) и новая страница будет сгенерирована путем запроса у ядра NT 1 Мбайт смежных физических страниц (через обычный вызов `ALLOC_PHYSICAL_PAGES`). Физическая память будет выделена ядром NT из соответствующей SLAB.

Таким же образом каждый раз, когда безопасное ядро освобождает часть своей собственной памяти, оно сохраняет соответствующие физические страницы в правильном пакете, увеличивая свой массив PFN до предела в 256 свободных страниц. Когда массив заполняется целиком и пакет освобождается, в очередь ставится новый рабочий элемент. Он обнулит все страницы и выдаст обычный вызов `FREE_PHYSICAL_PAGES`, который заканчивается выполнением функции `MmFreePagesFromMdl` диспетчера памяти NT.

Каждый раз, когда достаточное количество страниц перемещается в пакет или из него, они полностью защищаются в VTL 0 с помощью SLAT (эта процедура называется *защитой пакета*). Безопасное ядро поддерживает три типа пакетов, каждый из которых распределяет память из разных SLAB: «Нет доступа», «Только чтение» и «Чтение-выполнение».

Горячее исправление

Несколько лет назад 32-разрядные версии Windows поддерживали горячее исправление компонентов операционной системы. Исправляемые функции содержали избыточный двухбайтовый код операции в прологе и дополнительно несколько байтов, расположенных перед самой функцией. Это позволяло ядру NT динамически заменять исходный код операции косвенным переходом, который задействует свободное пространство, предоставленное дополнением, для перенаправления кода на исправленную функцию, находящуюся в другом модуле. Эта функция активно использовалась Центром обновления Windows, что позволяло обновлять систем, не перезагружая компьютер немедленно. При переходе на 64-разрядные архитектуры это уже было невозможно из-за различных проблем. Хорошим примером является защита исправлений ядра — больше не существовало надежного способа изменить двоичный файл защищенного режима ядра и разрешить обновление PatchGuard, не раскрывая некоторые из его частных интерфейсов, а открытые интерфейсы PatchGuard мог легко использовать злоумышленник, чтобы обойти защиту.

Безопасное ядро решило все проблемы, связанные с 64-разрядными архитектурами, и вернуло в ОС возможность горячего исправления двоичных файлов ядра. Когда безопасное ядро активно, могут быть исправлены в горячем режиме следующие типы исполняемых образов:

- модули пользовательского режима VTL 0 (исполняемые файлы и библиотеки);
- драйверы режима ядра, HAL и двоичный файл ядра NT независимо от того, защищены ли они с помощью PatchGuard;
- двоичный файл безопасного ядра и зависимые от него модули, работающие в режиме ядра VTL 1;
- гипервизор (Intel, AMD и версия ARM).

Двоичные файлы исправлений, созданные для программного обеспечения, работающего в VTL 0, называются обычными исправлениями, а остальные — безопасными. Если безопасное ядро неактивно, можно исправлять только приложения пользовательского режима.

Образ горячего исправления — это стандартный двоичный файл переносимого исполняемого файла (PE), который включает таблицу горячего исправления — структуру данных, применяемую для отслеживания функций исправления.

Таблица горячих исправлений связана в двоичном виде через каталог данных конфигурации загрузки образа. Он содержит один или несколько дескрипторов, описывающих каждый изменяемый базовый образ, который идентифицируется по его контрольной сумме и отметке даты и времени. (Таким образом, горячий патч совместим только с правильными базовыми образами. Система не может применить исправление к неправильному образу.) Таблица горячих исправлений (патчей) также включает в себя список функций или глобальных фрагментов данных, которые необходимо обновить в базе или в образе исправления; мы кратко опишем механизм. Каждая запись в этом списке содержит смещения функций в базовых и исправленных вариантах образов, а также исходные байты базовых функций, которые будут заменены.

К базовому образу можно применить несколько исправлений, но применение исправления само по себе идемпотентно. Можно задействовать одно и то же исправление несколько раз или разные исправления — последовательно. В любом случае последний задействованный патч будет активным для базового образа. Когда системе необходимо применить горячее исправление, она использует системный вызов `NtManageHotPatch` для установки и удаления горячих исправлений или управления ими. (Системный вызов поддерживает различные классы информации об исправлениях для описания всех возможных операций.) Горячее исправление может быть установлено глобально для всей системы или, если оно предназначено для кода пользовательского режима (VTL 0), для всех процессов, которые принадлежат определенному сеансу пользователя.

Когда система запрашивает применение исправления, ядро NT находит таблицу горячих исправлений в двоичном файле исправления и проверяет ее. Затем оно использует безопасный вызов `DTERMINE_HOT_PATCH_TYPE` для безопасного определения типа исправления. В случае безопасного исправления его может применить только безопасное ядро, поэтому задействуется вызов `APPLY_HOT_PATCH`, никакой другой обработки ядром NT не требуется. Во всех остальных случаях ядро NT сначала пытается применить исправление к драйверу ядра. Он циклически переключается между всеми загруженными модулями ядра в поисках базового образа, имеющего контрольную сумму, которая описана в одном из дескрипторов горячего исправления образа исправления.

Горячее исправление разрешено, только если значение реестра `HKEY_LOCAL_MACHINE\SYSTEM\CurrentControlSet\Control\Session Manager\Memory Management\HotPatchTableSize` кратно стандартному размеру страницы памяти (4096). Так или иначе, когда разрешено горячее исправление, каждому образу, отображаемому в виртуальном адресном пространстве, необходимо зарезервировать определенный объем этого адресного пространства сразу после самого образа. Зарезервированное пространство используется для таблицы адресов горячего исправления образа (HPAT, не путать с таблицей горячего исправления). HPAT задействуется для минимизации количества дополнений, необходимых для каждой исправляемой функции, путем сохранения адреса новой функции в исправленном образе. При исправлении функции местоположение HPAT будет использоваться для выполнения косвенного перехода от исходной функции в базовом образе к исправленной функции в образе исправления (обратите внимание на то, что для совместимости с Retpoline вместо непрямого перехода применяется другой вид процедуры Retpoline).

Когда ядро NT находит подходящий для исправления драйвер режима ядра, оно загружает и отображает двоичный файл исправления в адресном пространстве ядра и создает соответствующую запись в таблице данных загрузчика (более подробно см. в главе 12). Затем оно сканирует каждую страницу памяти как базового образа, так и образа исправления и блокирует в памяти те, которые задействованы в горячем исправлении (это важно — страницы не могут быть выгружены на диск, пока применение исправления не закончено). Наконец, оно отправляет безопасный вызов `APPLY_HOT_PATCH`.

Настоящий процесс применения исправлений начинается в безопасном ядре. Последнее захватывает и проверяет таблицу горячих исправлений образа исправления (посредством переназначения образа исправления также в VTL 1) и находит NAR базового образа (более подробную информацию о NAR см. в предыдущем подразделе «Диспетчер памяти безопасного ядра»), который также сообщает безопасному ядру, защищен ли образ PatchGuard. Затем безопасное ядро проверяет, достаточно ли зарезервировано места в образе HPAT. Если это так, оно выделяет одну или несколько свободных физических страниц (получая их из защищенного пакета или используя обычный вызов `ALLOC_PHYSICAL_PAGES`), которые будут отображаться в зарезервированном пространстве. На этом этапе, если базовый образ защищен, безопасное ядро запускает сложный процесс, который обновляет внутреннее состояние PatchGuard для нового исправленного образа и, наконец, вызывает механизм исправлений.

Механизм исправлений ядра выполняет следующие высокоуровневые операции, которые описываются разными типами записей в таблице горячих исправлений.

1. Исправляет все вызовы измененных функций в образе исправления с целью перехода к соответствующим функциям в базовом образе. Это гарантирует, что весь неисправленный код всегда будет выполняться в исходном базовом образе. Например, если функция А вызывает Б в базовом образе, а исправление изменяет функцию А, но не функцию Б, то механизм исправлений обновит функцию Б в исправлении, чтобы перейти к функции Б в базовом образе.
2. Исправляет необходимые ссылки на глобальные переменные в исправленных функциях, чтобы они указывали на соответствующие глобальные переменные в базовом образе.
3. Исправляет необходимые ссылки на таблицу адресов импорта (IAT) в образе исправления, копируя соответствующие записи IAT из базового образа.
4. Атомарно исправляет необходимые функции в базовом образе для перехода к соответствующей функции в образе исправления. Как только это будет сделано для данной функции в базовом образе, все ее последующие вызовы будут выполнять ее новый исправленный код в образе исправления. Когда исправленная функция будет возвращена, она вернется к вызывающей стороне исходной функции в базовом образе.

Поскольку указатели новых функций имеют разрядность 64 бита (8 байт), механизм исправлений вставляет каждый указатель в HPAT, который находится в конце двоичного файла. Таким образом, для размещения косвенного перехода в пространстве заполнения, расположенном в начале каждой функции, требуется

всего 5 байт. (Процесс был упрощен. Для горячих исправлений, совместимых с Retpoline, требуется совместимый Retpoline. Кроме того, HPAT разделен на страницы кода и данных.)

Как показано на рис. 9.38, механизм исправлений совместим с различными типами двоичных файлов. Если ядро NT не обнаружило ни одного исправляемого модуля режима ядра, оно перезапускает поиск по всем процессам пользовательского режима и применяет процедуру, аналогичную правильному горячему исправлению совместимого исполняемого файла или библиотеки пользовательского режима.

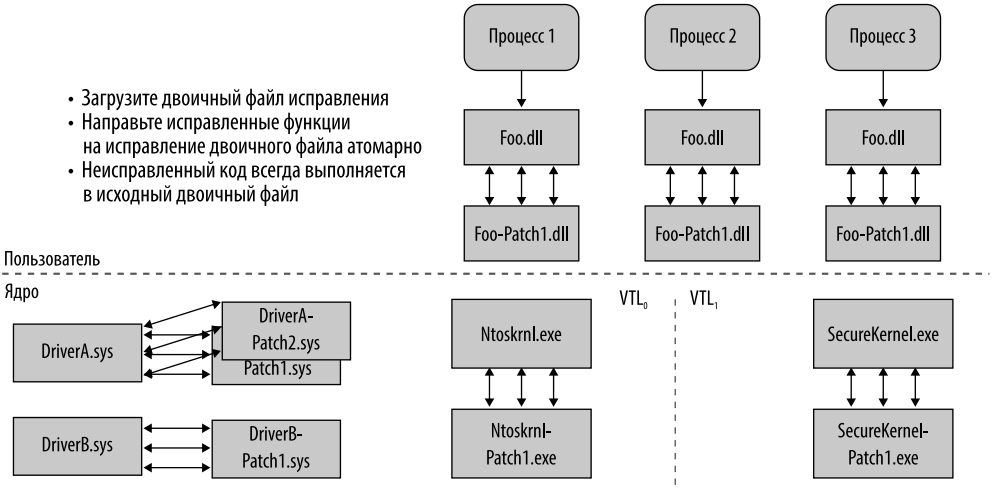


Рис. 9.38. Схема механизма горячего исправления, работающего на различных типах двоичных файлов

ИЗОЛИРОВАННЫЙ ПОЛЬЗОВАТЕЛЬСКИЙ РЕЖИМ

Изолированный пользовательский режим (IUM), сервисы, предоставляемые ядром безопасности его безопасным процессам (трастлетам), а также общая архитектура трастлетов описаны в главе 3 тома 1. В этом разделе мы продолжим обсуждение, начиная с того момента, и перейдем к описанию некоторых сервисов, предоставляемых изолированным пользовательским режимом, таких как защищенные устройства и анклавы VBS.

Как было сказано в главе 3 тома 1, когда трастлет создается в VTL 1, он обычно отображает в своем адресном пространстве следующие библиотеки.

- **Iumdll.dll.** Библиотека платформенно-ориентированного слоя IUM реализует заглушку защищенного системного вызова. Это эквивалент Ntdll.dll VTL 0.
- **Iumbase.dll.** Библиотека DLL базового слоя IUM реализует большинство безопасных API, которые могут использоваться исключительно программным обеспечением VTL 1. Он предоставляет всем безопасным процессам различные сервисы, такие как безопасная идентификация, связь, криптография и безопасное управление памятью. Трастлеты обычно не делают безопасные системные вызовы напрямую, а проходят через Iumbase.dll, которая является эквивалентом kernelbase.dll в VTL 0.

- **IumCrypt.dll.** Предоставляет функции шифрования с открытым/закрытым ключом, применяемые для подписи и проверки целостности. Большинство криптофункций, предоставляемых VTL 1, реализованы в Iumbase.dll, а в IumCrypt — лишь небольшое количество специализированных процедур шифрования. LsaIso — основной потребитель сервисов, предоставляемых IumCrypt, который не загружается во многие другие трастлеты.
- **Ntdll.dll**, **Kernelbase.dll** и **Kernel32.dll.** Трастлет может быть предназначен для работы как в VTL 1, так и в VTL 0. В этом случае он должен использовать только процедуры, реализованные в стандартной поверхности API VTL 0. Не все службы, доступные для VTL 0, реализованы также в VTL 1. Например, трастлет никогда не может выполнять какие-либо операции ввода-вывода с реестром и какие-либо файловые операции, но может применять процедуры синхронизации, ALPC, API потоков и структурированную обработку исключений, а также управление виртуальной памятью и объектами разделов. Почти все службы, предлагаемые библиотеками kernelbase и kernel32, выполняют системные вызовы через Ntdll.dll. В VTL 1 подобные системные вызовы преобразуются в обычные вызовы и перенаправляются в ядро VTL 0. (Мы подробно обсуждали обычные вызовы ранее в этой главе.) Обычные вызовы часто используются функциями IUM и самим безопасным ядром. Это объясняет, почему ntdll.dll всегда отображается в каждом трастлете.
- **Vertdll.dll.** Библиотека времени выполнения анклава VSM — это библиотека DLL, которая управляет временем существования анклава VBS. Программное обеспечение, выполняющееся в защищенном анклаве, предоставляет лишь ограниченное количество сервисов. Эта библиотека реализует все сервисы анклава, доступные программному анклаву, и обычно не загружается для стандартных процессов VTL 1.

Учитывая эти сведения, рассмотрим участников процесса создания трастлета, начиная с API `CreateProcess` в VTL 0, для которого поток выполнения подробно описан в главе 3.

Создание трастлетов

Как неоднократно обсуждалось в предыдущих разделах, при выполнении ряда операций безопасное ядро полагается на ядро NT. Создание трастлета тоже сюда входит: оно обеспечивается как безопасным ядром, так и ядром NT. В главе 3 тома 1 были представлены структура трастлета и требования к его подписи, рассмотрены связанные с ним важные метаданные политики. Кроме того, мы подробно описали последовательность действий функции `CreateProcess`, которая по-прежнему является отправной точкой для создания трастлета.

Чтобы правильно создать трастлет, приложение при вызове функции `CreateProcess` должно указать флаг создания `CREATE_SECURE_PROCESS`. Он преобразуется в атрибут `PS_CP_SECURE_PROCESS` NT и передается в платформенно-зависимый интерфейс `NtCreateUserProcess`. После успешного открытия исполняемого образа тот создает для него объект-секцию, указывая специальный флаг, сообщающий диспетчеру памяти, что нужно использовать Secure HVCI для проверки его содержимого. Это позволяет безопасному ядру создать структуру данных `SECURE_IMAGE`, применяемую для описания образа PE, проверенного с помощью Secure HVCI.

Как и в обычных процессах, ядро NT создает необходимые структуры данных и исходное адресное пространство VTL 0 (каталоги страниц, гиперпространство и рабочий набор). Если новый процесс является трастлетом, ядро генерирует безопасный вызов `CREATE_PROCESS`. Безопасное ядро поддерживает это, создавая объект защищенного процесса и соответствующую структуру данных, называемую `SEPROCESS`. Оно связывает обычный объект-процесс (`EPROCESS`) с новым защищенным объектом и создает исходное безопасное адресное пространство, выделяя таблицу защищенных страниц и дублируя корневые записи, описывающие часть ядра безопасного адресного пространства, в верхней его половине.

Ядро NT завершает настройку пустого адресного пространства процесса и отображает туда библиотеку Ntdll (подробности см. в главе 3 тома 1). При этом для защищенных процессов ядро NT делает безопасный вызов `INITIALIZE_PROCESS` для завершения настройки в VTL 1. Безопасное ядро копирует идентификатор и атрибуты трастлета, указанные во время создания процесса, в новый защищенный процесс, создает таблицу защищенных идентификаторов и сопоставляет защищенную общую страницу с адресным пространством.

Последним шагом, необходимым для защищенного процесса, будет создание защищенного потока. Исходный объект потока создается так же, как и для обычных процессов в ядре NT: когда интерфейс `NtCreateUserProcess` вызывает функцию `PspInsertThread`, последняя уже выделила стек ядра потока и вставила необходимые данные для запуска из функции ядра `KiStartUserThread` (подробности см. в главе 3 тома 1). Если процесс является трастлетом, ядро NT генерирует безопасный вызов `CREATE_THREAD` для завершения создания защищенного потока. Безопасное ядро подключается к адресному пространству нового процесса, выделяет и инициализирует структуру данных защищенного потока, безопасные TEB потока и стек ядра. Затем оно заполняет стек ядра потока, добавляя туда начальную процедуру ядра потока `SkpUserThreadStart`. Далее оно инициализирует машинно-зависимый аппаратный контекст для защищенного потока, определяющий фактический начальный адрес образа и адрес первой процедуры пользовательского режима. Наконец, безопасное ядро связывает обычный объект потока с вновь созданным безопасным объектом, заносит поток в список защищенных и помечает его как работоспособный.

Когда объект обычного потока выбран для запуска планировщиком ядра NT, выполнение по-прежнему начинается в рамках функции KiStartUserThread в VTL 0. Последняя снижает IRQL потока и вызывает системную процедуру первоначального запуска потока (`PspUserThreadStartup`). Выполнение продолжается как для обычных потоков, пока ядро NT не установит начальный контекст преобразователя функций. Но вместо последнего оно запускает цикл диспетчеризации безопасного ядра, для чего вызывает процедуру `VslpEnterIumSecureMode` и указывает безопасный вызов `RESUMETHREAD`. Данный цикл завершится не раньше, чем сам поток. Первоначальный безопасный вызов обрабатывается обычным циклом диспетчера вызовов в VTL 1, где определяется причина возобновления потока в VTL 1, происходит подключение к адресному пространству нового процесса и переключение на новый стек защищенных потоков. В таком случае безопасное ядро не вызывает функцию диспетчера `IumInvokeSecureService`, поскольку знает, что начальная функция потока находится в стеке а поэтому просто возвращает управление по адресу, расположенному в стеке, и указывающему на безопасную начальную процедуру `SkpUserThreadStart` из VTL 1.

Процедура `SkpUserThreadStart`, подобно аналогам из стандартных потоков VTL 0, настраивает начальный контекст преобразователя функций для запуска процедуры инициализации загрузчика образа (`LdrInitializeThunk` в Ntdll.dll), а также общесистемную заглушку запуска потока (`RtlUserThreadStart` в Ntdll.dll). Эти шаги выполняются путем редактирования контекста текущего потока с последующим инициированием выхода из служебной системной операции, в ходе которой загружается специально созданный пользовательский контекст и происходит возврат в пользовательский режим. Инициализация нового защищенного потока проводится так же, как и для обычных потоков VTL 0: процедура `LdrInitializeThunk` инициализирует загрузчик и необходимые ему структуры данных. По возвращении функция управления `NtContinue` восстанавливает новый пользовательский контекст. Здесь действительно начинается выполнение потока: `RtlUserThreadStart` принимает адрес фактической точки входа исполняемого образа и параметры запуска и передает управление в точку входа самого приложения.

ПРИМЕЧАНИЕ Внимательный читатель мог заметить, что безопасное ядро не делает ничего для защиты двоичного образа нового трастлета. Это связано с тем, что общая память, описывающая базовый двоичный образ трастлета, изначально спроектирована так, чтобы по-прежнему быть доступной для VTL 0.

Предположим, что трастлет хочет записать частные данные, расположенные в области глобальных данных образа. PTE, отображающие раздел данных, доступных для записи в рамках области глобальных данных образа, помечены как копируемые при записи. Таким образом, процессор будет генерировать ошибку отказа в доступе. Та относится к диапазону адресов пользовательского режима (помните, что NAR не задействуются для отслеживания общих страниц). Обработчик ошибок страниц безопасного ядра через обычный вызов передает управление ядру NT, которое выделит новую страницу, скопирует в нее содержимое старой и защитит ее через SLAT с помощью операции защищенного копирования (дополнительную информацию см. в подразделе «Диспетчер памяти безопасного ядра» ранее в этой главе).

ЭКСПЕРИМЕНТ. Отладка трастлета

Отладка трастлета средствами пользовательского режима возможна, только когда трастлет однозначно позволяет это в соответствии с метаданными своей политики (хранится в разделе .tPolicy). В рамках данного эксперимента мы попробуем отладить трастлет через отладчик ядра. Для этого потребуется подключить отладчик (локальный тоже подойдет) к тестовой системе, где должна быть активна опция VBS. Присутствие HVCI необязательно.

Первым делом найдем трастлет Lsaiso.exe:

```
lkd> !process 0 0 lsaiso.exe
PROCESS ffff8904dfdaa080
    SessionId: 0 Cid: 02e8 Peb: 8074164000 ParentCid: 0250
    DirBase: 3e590002 ObjectTable: ffffb00d0f4dab00 HandleCount: 42.
    Image: LsaIso.exe
```

Анализ PEB процесса показывает, что часть информации обнулена или нечитаема:

```
lkd> .process /P ffff8904dfdaa080
lkd> !peb 8074164000
PEB at 0000008074164000
    InheritedAddressSpace:    No
    ReadImageFileExecOptions: No
    BeingDebugged:            No
    ImageBaseAddress:         00007ff708750000
    NtGlobalFlag:             0
    NtGlobalFlag2:            0
    Ldr                       0000000000000000
    *** unable to read Ldr table at 0000000000000000
    SubSystemData:     0000000000000000
    ProcessHeap:       0000000000000000
    ProcessParameters: 0000026b55a10000
    CurrentDirectory:  'C:\Windows\system32\'
    WindowTitle:  '< Name not readable >'
    ImageFile:    '\??\C:\Windows\system32\lsaiso.exe'
    CommandLine:  '\??\C:\Windows\system32\lsaiso.exe'
    DllPath:      '< Name not readable >'lkd
```

Чтение по базовому адресу образа процесса может быть успешным, но это зависит от того, осуществлялся ли уже ранее доступ к отображенному в адресном пространстве VTL 0 образу Lsalso. Это обычно справедливо для первой страницы (не забываем, что разделяемая память основного образа доступна в VTL 0). В нашей системе первая страница отображена и действительна, притом что третья недействительна:

```
lkd> db 0x7ff708750000 l20
00007ff7`08750000 4d 5a 90 00 03 00 00 00-04 00 00 00 ff 00 00  MZ..............
00007ff7`08750010 b8 00 00 00 00 00 00 00-40 00 00 00 00 00 00 00  ........@.......
lkd> db (0x7ff708750000 + 2000) l20
00007ff7`08752000 ?? ?? ?? ?? ?? ?? ?? ??-?? ?? ?? ?? ?? ?? ?? ??  ????????????????
00007ff7`08752010 ?? ?? ?? ?? ?? ?? ?? ??-?? ?? ?? ?? ?? ?? ?? ??  ????????????????
lkd> !pte (0x7ff708750000 + 2000)
1: kd>  !pte (0x7ff708750000 + 2000)
                                           VA 00007ff708752000
PXE at FFFFD5EAF57AB7F8    PPE at FFFFD5EAF56FFEE0    PDE at FFFFD5EADFFDC218
contains 0A0000003E58D867  contains 0A0000003E58E867  contains 0A0000003E58F867
pfn 3e58d     ---DA--UWEV  pfn 3e58e     ---DA--UWEV  pfn 3e58f     ---DA--UWEV

PTE at FFFFD5BFFB843A90
contains 00000000000000
not valid
```

Выгрузка потоков процесса раскрывает важную информацию, подтверждающую то, что обсуждалось в разделах ранее:

```
!process ffff8904dfdaa080 2
PROCESS ffff8904dfdaa080
    SessionId: 0 Cid: 02e8 Peb: 8074164000 ParentCid: 0250
    DirBase: 3e590002 ObjectTable: ffffb00d0f4dab00 HandleCount: 42.
    Image: LsaIso.exe
        THREAD ffff8904dfdd9080 Cid 02e8.02f8 Teb: 0000008074165000
```

```
        Win32Thread: 0000000000000000 WAIT: (UserRequest) UserMode Non-Alertable
            ffff8904dfdc5ca0 NotificationEvent

        THREAD ffff8904e12ac040 Cid 02e8.0b84 Teb: 0000008074167000
        Win32Thread: 0000000000000000 WAIT: (WrQueue) UserMode Alertable
            ffff8904dfdd7440 QueueObject

lkd> .thread /p ffff8904e12ac040
Implicit thread is now ffff8904`e12ac040
Implicit process is now ffff8904`dfdaa080
.cache forcedecodeuser done
lkd> k
  *** Stack trace for last set context - .thread/.cxr resets it
 # Child-SP RetAddr Call Site
00 ffffe009`1216c140 fffff801`27564e17 nt!KiSwapContext+0x76
01 ffffe009`1216c280 fffff801`27564989 nt!KiSwapThread+0x297
02 ffffe009`1216c340 fffff801`275681f9 nt!KiCommitThreadWait+0x549
03 ffffe009`1216c3e0 fffff801`27567369 nt!KeRemoveQueueEx+0xb59
04 ffffe009`1216c480 fffff801`27568e2a nt!IoRemoveIoCompletion+0x99
05 ffffe009`1216c5b0 fffff801`2764d504 nt!NtWaitForWorkViaWorkerFactory+0x99a
06 ffffe009`1216c7e0 fffff801`276db75f nt!VslpDispatchIumSyscall+0x34
07 ffffe009`1216c860 fffff801`27bab7e4 nt!VslpEnterIumSecureMode+0x12098b
08 ffffe009`1216c8d0 fffff801`276586cc nt!PspUserThreadStartup+0x178704
09 ffffe009`1216c9c0 fffff801`27658640 nt!KiStartUserThread+0x1c
0a ffffe009`1216cb00 00007fff`d06f7ab0 nt!KiStartUserThreadReturn
0b 00000080`7427fe18 00000000`00000000 ntdll!RtlUserThreadStart
```

По данному стеку ясно видно, что исполнение начинается в процедуре `KiStartUserThread`. В свою очередь, `PspUserThreadStartup` вызвала цикл диспетчера вызовов, который никогда не завершался, но был приостановлен операцией ожидания (`wait`). Отладчик ядра никогда не сможет отобразить ни структуру данных безопасного ядра, ни частные данные трастлета.

Защищенные устройства

VBS предоставляет драйверам возможность запускать часть своего кода в безопасной среде. Само безопасное ядро не может быть расширено для поддержки драйверов ядра — его поверхность атаки станет слишком большой. Более того, Microsoft не позволит сторонним компаниям вносить возможные ошибки в компонент, применяемый в первую очередь в целях безопасности.

Платформа драйверов пользовательского режима (UMDF) решает данную проблему, представляя концепцию сопутствующих драйверов-компаньонов, которые могут работать как в пользовательском режиме VTL 0, так и в VTL 1. В этом случае их так и называют — *безопасными компаньонами*. Безопасный компаньон берет подмножество кода драйвера, которое должно работать в другом режиме (в данном случае IUM), и загружает его как расширение, оно же компаньон, основного драйвера KMDF. Тем не менее стандартные драйверы WDM тоже поддерживаются. Основной драйвер, который по-прежнему работает в режиме ядра VTL 0, продолжает управлять устройством в части PnP и состояния питания, но

ему необходима возможность обращаться к своим компаньонам по поводу задач, которые необходимо выполнять в IUM.

Платформа Secure Driver Framework (SDF), упомянутая в главе 3, устарела, однако на рис. 9.39 приведена архитектура новой модели безопасных компаньонов UMDF, построенная на основе все той же базовой структуры UMDF (Wudfx02000.dll), задействуемой в пользовательском режиме VTL 0. Последний пользуется хостом безопасных компаньонов UMDF (WUDFCompanionHost.exe) для загрузки драйвера-компаньона, представленного в виде библиотеки DLL, и управления им. Хост безопасных компаньонов UMDF управляет временем жизни безопасного компаньона и инкапсулирует многие функции UMDF, непосредственно относящихся к среде IUM.

Безопасный компаньон обычно связан с основным драйвером, действующим в ядре VTL 0. Он должен быть правильно подписан (в том числе IUM EKU в подписи, как и для каждого трастлета) и декларировать свои возможности в собственном разделе метаданных. Безопасный компаньон полностью владеет управляемым устройством (это объясняет, почему такое устройство часто называют *защищенным*). Защищенный контроллер устройства в рамках компаньона поддерживает следующие функции.

- **Защищенный DMA.** Драйвер может дать указание устройству передать DMA непосредственно в защищенную память VTL 1, недоступную в условиях VTL 0. Безопасный компаньон может обработать данные, отправленные или полученные подобным образом, а затем передать часть их драйверу VTL 0 через стандартный интерфейс связи KMDF (ALPC). Безопасные системные вызовы `IumGetDmaEnabler` и `IumDmaMapMemory`, предоставляемые через Iumbase.dll, позволяют безопасному компаньону сопоставлять диапазоны физической памяти для DMA непосредственно в пользовательском режиме VTL 1.
- **Порты ввода-вывода с отображением в памяти** (memory mapped IO, MMIO). Безопасный компаньон может обратиться к устройству, чтобы отразить его доступный диапазон MMIO в VTL 1 (пользовательский режим). После этого он может получить доступ к отображенным в памяти регистрам устройства непосредственно в IUM. Эта возможность предоставляется функциями `MapSecureIo` и `ProtectSecureIo`.
- **Защищенные разделы.** Компаньон может создавать (функцией `CreateSecureSection`) и отображать безопасные разделы, которые представляют собой память, доступную для совместного использования трастлетами и основным драйвером, работающим в VTL 0. Кроме того, безопасный компаньон может задействовать иной тип защиты SLAT, если происходит доступ к памяти через защищенное устройство — DMA или MMIO.

Безопасный компаньон не может напрямую реагировать на прерывания устройства, так как за их отображение и обработку отвечает соответствующий драйвер режима ядра в VTL 0. Драйвер режима ядра по-прежнему должен действовать как высокоуровневый интерфейс для системы и приложений пользовательского режима, обрабатывая все входящие IOCTL. Основной драйвер взаимодействует со своим безопасным компаньоном, отправляя задачи WDF через объект очереди задач UMDF, внутренняя реализация которого базируется на функционале ALPC, предоставляемом платформой WDF.

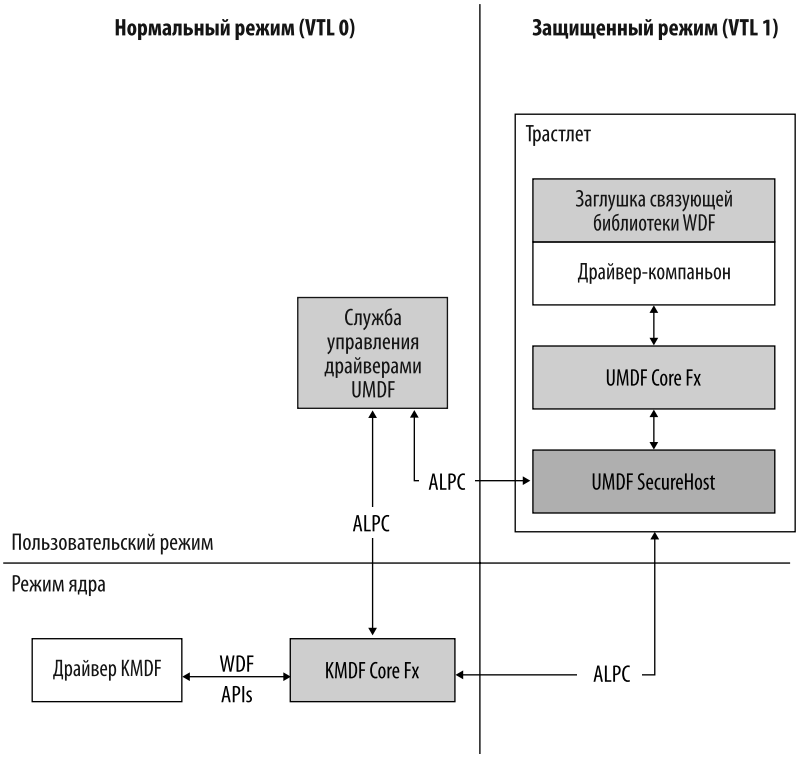


Рис. 9.39. Архитектура безопасных компаньонов для драйверов WDF

Типичный драйвер KMDF регистрирует своего компаньона с помощью директив INF. Последнего WDF запускает автоматически в контексте вызова драйвером функции `WdfDeviceCreate`, что для драйверов Plug and Play обычно происходит в ответном вызове `AddDevice`. Для этого сообщение ALPC отправляется в службу диспетчера драйверов UMDF, которая с помощью платформенно-зависимого интерфейса `NtCreateUserProcess` создает новый трастлет хоста безопасных компаньонов UMDF WUDFCompanionHost.exe. Тот, в свою очередь, загружает DLL компаньона в свое адресное пространство. Затем уже диспетчер драйверов UMDF отправляет еще одно сообщение ALPC в адрес WUDFCompanionHost для фактического запуска безопасного компаньона. В рамках последнего процедура `DriverEntry` выполняет безопасную инициализацию драйвера и создает объект `WDFDRIVER` с помощью классической функции `WdfDriverCreate`.

Далее платформа вызывает процедуру обратного вызова `AddDevice` компаньона в VTL 1, обычно создающую устройство компаньона с помощью новой функции UMDF `WdfDeviceCompanionCreate`. Та через безопасный системный вызов `IumCreateSecureDevice` передает управление безопасному ядру, которое создает новое защищенное устройство. С этого момента безопасный компаньон полностью владеет отданным под его управление устройством. Обычно первое, что компаньон делает после создания защищенного устройства, — создает объект

очереди (WDFTASKQUEUE), используемый для обработки любых входящих задач, доставляемых связанным с ним драйвером VTL 0. Управление возвращается драйверу режима ядра, который теперь сможет отправлять новые задачи своему безопасному компаньону.

Эта модель поддерживается также драйверами WDM. Они могут задействовать режим мини-порта KMDF для взаимодействия со специальным драйвером фильтра WdmCompanionFilter.sys, который подключается на нижнем уровне стека устройства. Так, фильтр Wdm Companion позволяет драйверам WDM использовать объект очереди задач для взаимодействия с безопасным компаньоном.

Анклавы на основе VBS

В главе 5 тома 1 мы обсуждали расширение Software Guard Extension (SGX) — аппаратную технологию, позволяющую создавать защищенные анклавы памяти, представляющие собой безопасные зоны в адресном пространстве процесса, где код и данные последнего защищаются (шифруются) аппаратным обеспечением от влияния кода, действующего за пределами анклава. Впервые представленное в процессорах Intel Core шестого поколения (Skylake), данное решение столкнулось с рядом проблем, которые помешали его широкому распространению. (Кроме того, AMD выпустила еще одну технологию под названием Secure Encrypted Virtualization, несовместимую с SGX.)

Для преодоления этих сложностей Microsoft представила анклавы на основе VBS — безопасные анклавы, гарантии изоляции которых обеспечиваются с помощью инфраструктуры VSM. Код и данные внутри такого анклава видны только ему самому и безопасному ядру VSM, но недоступны для ядра NT, процессов VTL 0 и безопасных трастлетов, работающих в системе.

Защищенный анклав на основе VBS создается посредством формирования единого диапазона виртуальных адресов в рамках обычного процесса. Затем в него загружаются код и данные, после чего первый вход в анклав осуществляется путем передачи управления на точку входа через безопасное ядро. Оно, в свою очередь, сначала проверяет, что весь код и данные подлинные и разрешены для запуска внутри анклава, выполняя проверку подписи образа в образе анклава. Если проверки подписи пройдены, то управление передается на точку входа в анклав, где будет обеспечен доступ ко всему его коду и данным. По умолчанию система поддерживает выполнение только правильно подписанных анклавов. Это исключает вероятность того, что неподписанное вредоносное ПО сможет выполняться в системе вне поля зрения антивирусного программного обеспечения, не способного обозревать содержимое никаких анклавов.

Во время выполнения управление может передаваться туда и обратно между анклавом и содержащим его процессом. Код, выполняющийся внутри анклава, имеет доступ ко всем данным в диапазоне его виртуальных адресов. Кроме того, он имеет доступ для чтения и записи в незащищенном адресном пространстве процесса. В свою очередь, процесс-владелец анклава не сможет обращаться к памяти в его диапазоне виртуальных адресов. Если же в рамках одного хост-процесса существует несколько анклавов, каждый из них сможет получить доступ только к собственной памяти и памяти, доступной его владельцу.

Если код выполняется в аппаратном анклаве, он может получить запечатанный отчет анклава, который может быть использован сторонним объектом для оценки на предмет обеспечения гарантий изоляции анклава VBS, а впоследствии — для проверки конкретной версии выполняемого кода. В отчет входит информация о хост-системе, самом анклаве и всех библиотеках DLL, которые могли быть в него загружены. Кроме того, там отражаются сведения о том, работает ли анклав при активных инструментах отладки.

Анклав на основе VBS распространяется в виде библиотеки DLL с рядом определенных особенностей.

- Имеет подпись аутентификации, а конечный сертификат включает действительный EKU, разрешающий запускать данный образ как анклав. Корневым центром сертификации, выдавшим цифровой сертификат, должны быть либо Microsoft, либо сторонний удостоверяющий центр, указанный в манифесте сертификата, подписанном Microsoft. Это означает, что сторонние компании могут подписывать и запускать собственные анклавы. Допустимыми EKU цифровой подписи являются IUM EKU (1.3.6.1.4.1.311.10.3.37) для внутренних анклавов, подписанных для Windows, и анклавные EKU (1.3.6.1.4.1.311.10.3.42) для всех сторонних анклавов.
- Включает раздел конфигурации анклава (представлен структурой данных `IMAGE_ENCLAVE_CONFIG`), в котором приводится информация об анклаве и который привязан к каталогу данных конфигурации загрузки его образа.
- Включает в себя соответствующий инструментарий Control Flow Guard (CFG).

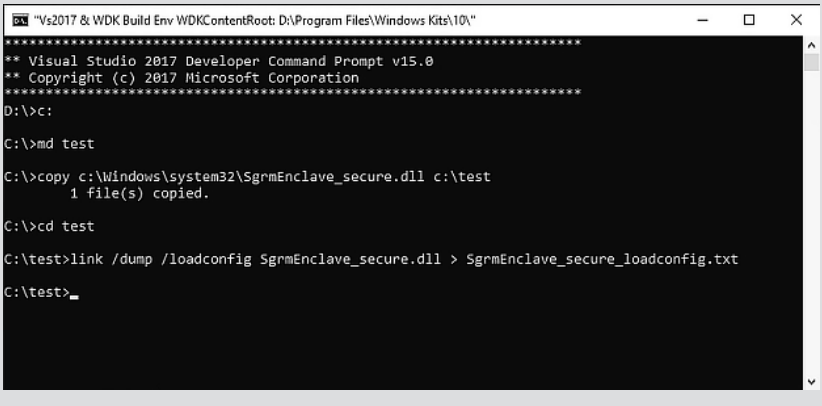
Роль раздела конфигурации анклава заключается в том, что он содержит важную информацию, необходимую для правильной работы и запечатывания анклава: уникальные идентификаторы семейства и образа, указываемые автором и идентифицирующие двоичный образ, номер безопасной версии и информацию о политиках — ожидаемый виртуальный размер, максимальное число выполняемых потоков, возможность отладки анклава. Кроме того, в раздел конфигурации анклава входит список образов, которые могут быть им импортированы, в том числе их идентификационные данные. Импортированный модуль анклава может быть идентифицирован по комбинации идентификаторов семейства и образа или же по комбинации сгенерированного уникального идентификатора, рассчитываемого на основе хеша двоичного файла, и идентификатора автора, получаемого из сертификата, использованного при подписании анклава. (Это значение удостоверяет личность того, кто построил анклав.) Дескриптор импортированного модуля должен включать также номер минимальной безопасной версии.

Безопасное ядро предоставляет анклавам некоторые базовые системные сервисы через Vertdll.dll — библиотеку времени выполнения для анклавов VBS, которая отражается в адресное пространство анклава. В их рамках предоставляются: ограниченный функционал стандартной библиотеки времени выполнения C, возможность выделять или освобождать защищенную память в пределах диапазона адресов анклава, возможности синхронизации, поддержка структурированной обработки исключений, базовые криптографические функции и возможность запечатывания данных.

ЭКСПЕРИМЕНТ. Выгрузка конфигурации анклава

В рамках данного эксперимента мы воспользуемся инкрементным компоновщиком Microsoft (link.exe) в составе Windows SDK и WDK, чтобы выгрузить данные о конфигурации программного анклава. Оба пакета доступны для скачивания в Интернете. Можете также воспользоваться EWDK, содержащим все необходимые инструменты и не требующим установки. Его можно скачать на сайте https://docs.microsoft.com/en-us/windows-hardware/drivers/download-the-wdk.

Откройте консоль Visual Studio Developer Command Prompt через окно поиска Cortana или запустив скрипт LaunchBuildEnv.cmd из ISO-образа EWDK. Мы будем анализировать данные конфигурации анклава регулярной аттестации защиты системы (System Guard Routine Attestation) (он присутствует на рис. 9.40 и рассматривает далее в данной главе) с помощью команды `link.exe/dump/loadconfig`.



Листинг результата выполнения этой команды весьма велик. Поэтому на приведенном снимке экрана мы перенаправили его в файл `SgrmEnclave_secure_loadconfig.txt`. Открыв его, вы увидите, что двоичный образ включает в себя таблицу CFG и действительный указатель конфигурации анклава, по которому находятся следующие данные:

```
Enclave Configuration

            00000050 size
            0000004C minimum required config size
            00000000 policy flags
            00000003 number of enclave import descriptors
            0004FA04 RVA to enclave import descriptors
            00000050 size of an enclave import descriptor
            00000001 image version
            00000001 security version
    0000000010000000 enclave size
            00000008 number of threads
            00000001 enclave flags
            family ID : B1 35 7C 2B 69 9F 47 F9 BB C9 4F 44 F2 54 DB 9D
```

```
            image ID : 24 56 46 36 CD 4A D8 86 A2 F4 EC 25 A9 72 02

   ucrtbase_enclave.dll

            0 minimum security version
            0 reserved

              match type : image ID
               family ID : 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00
                image ID : F0 3C CD A7 E8 7B 46 EB AA E7 1F 13 D5 CD DE 5D
        unique/author ID : 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00
                           00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00

        bcrypt.dll

            0 minimum security version
            0 reserved

              match type : image ID
               family ID : 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00
                image ID : 20 27 BD 68 75 59 49 B7 BE 06 34 50 E2 16 D7 ED
        unique/author ID : 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00
                           00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00
...
```

Раздел конфигурации содержит данные двоичного образа анклава (идентификаторы семейства и образа, номер версии защиты) и массив дескрипторов импорта, согласно которому безопасное ядро определяет, на какую библиотеку анклав может безопасно полагаться. Вы можете повторить этот эксперимент с библиотекой Vertdll.dll, равно как и с прочими образами, импортируемыми из анклава System Guard Routine Attestation.

Жизненный цикл анклава

В главе 5 тома 1 мы обсуждали жизненный цикл аппаратного анклава (на базе SGX). Жизненный цикл анклава на основе VBS аналогичен — корпорация Microsoft усовершенствовала уже доступные API-интерфейсы анклавов для поддержки их нового типа на базе VBS.

Шаг 1. Создание. Приложение создает анклав на основе VBS, передав флаг `ENCLAVE_TYPE_VBS` в функцию `CreateEnclave`. Вызывающая сторона должна указать владельца анклава, предоставив его идентификатор. Код создания анклава, как и для аппаратных анклавов, заканчивается вызовом функции ядра `NtCreateEnclave`. Последняя проверяет параметры, копирует переданные структуры и присоединяется к целевому процессу на случай, если анклав должен быть создан в постороннем для вызывающего процессе. Функция `MiCreateEnclave` размещает VAD анклавного типа, где описывается диапазон виртуальной памяти анклава, и выбирает базовый виртуальный адрес, если тот не указан вызывающей стороной. Ядро выделяет структуру данных анклава VBS для диспетчера памяти и хеш-таблицу анклава для всех процессов, используемую для быстрого поиска анклава, начиная с его номера. Если анклав создается для процесса первым, система создает также пустой защищенный процесс (он играет роль контейнера для анклавов) в VTL 1

с помощью безопасного вызова CREATE _PROCESS (подробности см. в подразделе «Создание трастлетов» ранее).

Обработчик безопасного вызова CREATE_ENCLAVE из VTL 1 выполняет фактическую работу по созданию анклава: выделяет структуру данных защищенного ключа анклава (SKMI_ENCLAVE), устанавливает ссылку на защищенный процесс контейнера (только что созданный ядром NT) и создает защищенный VAD с описанием всего виртуального адресного пространства анклава (такой VAD содержит информацию, аналогичную имеющейся в его коллеге из VTL 0). Последний вставляется в дерево VAD процесса-владельца, а не в сам анклав. Пустое виртуальное адресное пространство для анклава создается так же, как и для содержащего его процесса: корень таблицы страниц заполняется только системными записями.

Шаг 2. Загрузка модулей в анклав. В отличие от случая аппаратных анклавов родительский процесс не может загружать в анклав произвольные данные, ограничиваясь только модулями. Это приводит к тому, что каждая страница образа будет скопирована в адресное пространство VTL 1. Каждая страница образа в анклаве VTL 1 будет *частной копией*. В анклав необходимо загрузить хотя бы один модуль, исполняющий роль основного образа анклава, так как в противном случае он не сможет быть инициализирован. После создания анклава VBS приложение вызывает интерфейс LoadEnclaveImage, передавая ему базовый адрес анклава и имя модуля, который необходимо в него загрузить. Код загрузчика Windows (в Ntdll.dll) выполняет поиск по указанному имени DLL, открывает и проверяет ее двоичный файл, после чего создает объект раздела, который отображается в вызывающем процессе с правом доступа только для чтения.

Отобразив этот раздел, загрузчик анализирует таблицу адресов импорта образа с целью создания списка зависимых модулей — импортированных, с отложенной загрузкой и пересылаемых. Он проверяет, достаточно ли в анклаве места для отображения каждого найденного модуля, и вычисляет корректный базовый адрес образа. Как показано на рис. 9.40, где представлен анклав System Guard Runtime Attestation, модули в анклаве отображаются по принципу «сверху вниз». Это означает, что основной образ отображается по максимально возможному виртуальному адресу, а все зависимые — по младшим адресам рядом друг с другом. На этом этапе для каждого модуля загрузчик Windows вызывает интерфейс ядра NtLoadEnclaveData.

Для загрузки искомого образа в анклав VBS ядро запускает сложный процесс, который позволяет копировать общие страницы его объекта раздела в частные страницы анклава в VTL 1. Функция MiMapImageForEnclaveUse получает область управления объектом раздела и проверяет его с помощью SKCI. Если проверка не удалась, процесс прерывается, а вызывающей стороне возвращается ошибка. (Как обсуждалось ранее, все модули анклава должны иметь корректную подпись.) В противном случае система подключается к защищенному системному процессу и отображает объект раздела этого образа в его адресное пространство в VTL 0. Общие страницы самого модуля в это время могут быть действительными или недействительными (подробности см. в главе 5 тома 1). Затем она фиксирует виртуальное адресное пространство модуля в содержащем его процессе. В итоге создаются частные структуры данных подкачки VTL 0 для PTE с принудительным обнулением, которые позже будут заполнены безопасным ядром в ходе загрузки образа в VTL 1.

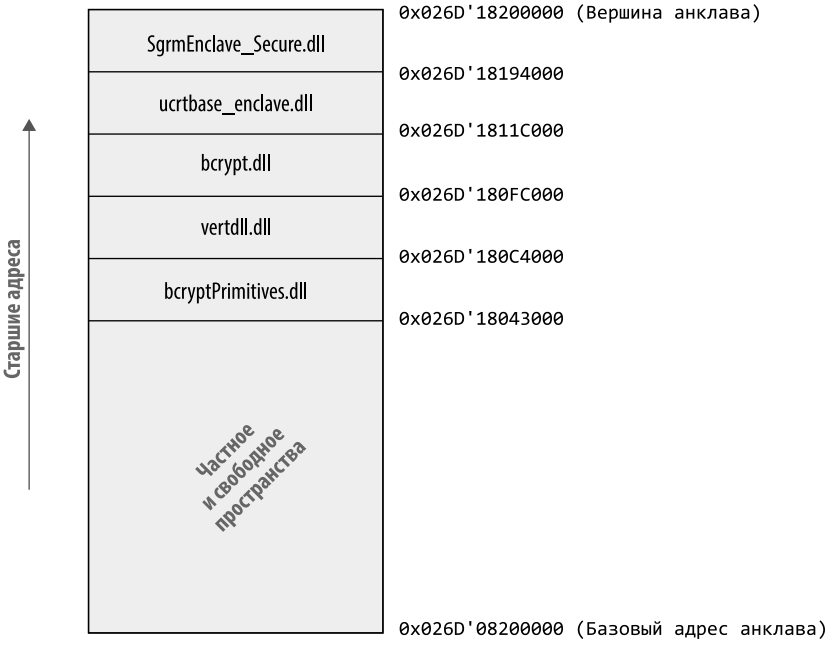


Рис. 9.40. Защищенный анклав System Guard Runtime Attestation (обратите внимание на пустое пространство у базы анклава)

Обработчик защищенного вызова `LOAD_ENCLAVE_MODULE` в VTL 1 получает `SECURE_IMAGE` нового модуля, созданного SKCI, и проверяет, подходит ли образ для использования в анклаве на базе VBS, путем проверки характеристик цифровой подписи. Затем он подключается к защищенному системному процессу в VTL 1 и отображает защищенный образ по тому же виртуальному адресу, который ранее был сопоставлен ядром NT. Это позволяет совместно применять прототипные PTE из VTL 0. Затем безопасное ядро создает безопасный VAD, описывающий модуль, и вставляет его в адресное пространство VTL 1 анклава. Наконец, производится циклический обход прототипов PTE секций каждого модуля. Для каждого отсутствующего прототипа PTE он подключается к защищенному системному процессу и использует обычный вызов `GET_PHYSICAL_PAGE` для вызова обработчика ошибок страницы NT (`MmAccessFault`), который помещает общую страницу в память. Безопасное ядро выполняет аналогичный процесс для частных страниц анклава, которые ранее были зафиксированы ядром NT в VTL 0 для PTE с принудительным обнулением. В таком случае обработчик ошибок страниц NT выделяет обнуленные страницы. Безопасное ядро копирует содержимое каждой общей физической страницы в каждую новую личную страницу, при необходимости выполняя требуемые частные перемещения.

Загрузка модуля в анклав на основе VBS завершена. Безопасное ядро устанавливает защиту частных страниц анклава посредством SLAT (ядро NT не имеет доступа к коду образа и данным в анклаве), прекращает отражение общего раздела в защищенном системном процессе и передает управление ядру NT. Теперь загрузчик может приступить к следующему модулю.

Шаг 3. Инициализация анклава. После загрузки всех модулей в анклав приложение инициализирует его с помощью интерфейса `InitializeEnclave` и указывает максимальное количество потоков, поддерживаемое им (сам анклав, в свою очередь, будет привязан к потокам, способным выполнять вызовы анклава в рамках процесса-владельца). Обработчик безопасного вызова `INITIALIZE_ENCLAVE` безопасного ядра проверяет, что политики, указанные при создании анклава, совместимы с политиками, изложенными в данных конфигурации основного образа, убеждается в том, что библиотека платформы анклава (Vertdll.dll) загружена, вычисляет окончательный 256-битный хеш анклава (используется для создания запечатанного отчета анклава) и создает все защищенные потоки анклава. Когда управление возвращается коду загрузчика Windows в VTL 0, система инициирует первый вызов анклава, в рамках которого выполняется код инициализации DLL платформы.

Шаг 4. Вызовы анклава (входящие и исходящие). После корректной инициализации анклава приложение сможет выполнять в его адрес произвольное количество вызовов. Все вызываемые функции в анклаве должны быть экспортированы. Приложение сможет вызывать стандартный интерфейс `GetProcAddress` для получения адреса функции анклава, а затем использовать процедуру `CallEnclave` для передачи управления в сам защищенный анклав. В этом сценарии, описывающем *входящий вызов*, процедура ядра `NtCallEnclave` выполняет алгоритм выбора потока, который связывает вызывающий поток VTL 0 с потоком анклава в соответствии со следующими правилами.

- Если обычный поток ранее не вызывался анклавом (анклавы поддерживают вложенные вызовы), то для выполнения выбирается его произвольный простаивающий поток. Если свободных потоков анклава нет, вызов блокируется до тех пор, пока они не появятся (если это указано вызывающей стороной, в противном случае вызов просто закончится отказом).
- Если анклав ранее вызывал обычный поток, то вызов анклава выполняется в том же потоке анклава, который отправлял хосту предыдущий вызов.

Список дескрипторов потоков анклава обслуживается как ядром NT, так и безопасным ядром. Когда обычный поток привязан к потоку анклава, последний вставляется в другой список, который называется *списком связанных потоков*. Отслеживаемые им потоки анклава считаются на данный момент занятыми и более недоступны.

После успешного завершения алгоритма выбора потока ядро NT отправляет безопасный вызов `CALLENCLAVE`. Безопасное ядро создает новый кадр стека для анклава и возвращается в пользовательский режим. Первой функцией пользовательского режима, выполняемой в контексте анклава, является `RtlEnclaveCallDispatcher`. Если вызов анклава был первым когда-либо инициированным, она передает управление процедуре инициализации библиотеки времени выполнения анклава VSM (Vertdll.dll), которая инициализирует CRT, загрузчик и все сервисы, предоставляемые анклаву. Наконец, она вызывает функцию `DllMain` основного модуля анклава и всех зависимых от него образов, передав код причины `DLL_PROCESS_ATTACH`.

В обычных ситуациях, когда DLL платформы анклава уже инициализирована, диспетчер анклава вызывает `DllMain` каждого модуля, передавая код причины `DLL_THREAD_ATTACH`, проверяет, действителен ли указанный адрес искомой функции анклава, и, если да, наконец вызывает ее. Функция, завершив свое выполнение, возвращается к VTL 0 посредством ответного вызова в адрес процесса-владельца. Для

этого по-прежнему используется DLL платформы анклава, которая снова вызывает процедуру ядра `NtCallEnclave`. Хотя она реализована в безопасном ядру несколько иначе, но задействует аналогичную стратегию для возврата к VTL 0. Сам анклав может генерировать свои вызовы для выполнения каких-либо функций в контексте незащищенного процесса-владельца. В этом сценарии, описывающем исходящий вызов, код анклава использует процедуру `CallEnclave`, куда передает адрес функции, экспортируемой в главном модуле содержащего ее процесса.

Шаг 5. Завершение и уничтожение. Когда через функцию `TerminateEnclave` запрашивается полное завершение работы анклава, все потоки, выполняющиеся внутри него, будут вынуждены вернуться в VTL 0. При этом все дальнейшие вызовы анклава будут заканчиваться отказом. По мере завершения потоков их состояние из VTL 1 (включая стеки потоков) уничтожается. Как только все потоки остановятся, анклав можно уничтожить. Когда это происходит, все его данные, оставшиеся в VTL 1, также уничтожаются (включая все адресное пространство анклава), а все страницы в VTL 0 освобождаются. Наконец, VAD анклава удаляется, равно как и вся выделенная ему память. Уничтожение запускается, когда процесс-владелец вызывает `VirtualFree`, передав базовый диапазон адресов анклава. Уничтожение невозможно, если анклав не завершился или никогда не инициализировался.

ПРИМЕЧАНИЕ Как обсуждалось ранее, все страницы памяти, отражаемые в адресное пространство анклава, являются частными. Это имеет ряд последствий. Впрочем, никакие страницы памяти, относящиеся к VTL 0 процесса-владельца, не отображаются в адресном пространстве анклава, равно как и VAD, описывающие расположение памяти процесса-владельца. Так как же анклав может получить доступ ко всем страницам памяти содержащего его процесса?

Ответ находится в обработчике ошибок страниц безопасного ядра (SkmmAccessFault). В своем коде обработчик ошибок проверяет, является ли процесс, вызвавший ошибку, анклавом. Если это так, делается следующая проверка на предмет того, произошла ли ошибка из-за попытки анклава выполнить какой-то код за пределами своей области памяти. В таком случае выдается ошибка нарушения доступа. Если ошибка вызвана доступом для чтения или записи за пределами адресного пространства анклава, обработчик ошибок защищенной страницы обращается к обычному сервису GET_PHYSICAL_PAGE, в результате чего вызывается обработчик ошибок доступа VTL 0. Обработчик VTL 0 проверяет VAD-дерево процесса-владельца, получает PFN страницы из ее PTE (при необходимости возвращая ее в память) и возвращает его в VTL 1. На этом этапе безопасное ядро создает необходимые структуры подкачки для отражения физической страницы по тому же виртуальному адресу, который гарантированно доступен благодаря свойству самого анклава, после чего возобновляет выполнение. Теперь эта страница действительна в контексте безопасного анклава.

Запечатывание и аттестация

Анклавы на базе VBS, аналогично аппаратным анклавам, поддерживают как запечатывание, так и аттестацию данных. Термин *«запечатывание»* относится к шифрованию произвольных данных с использованием одного или нескольких ключей шифрования, которые не видны коду анклава, но управляются безопасным ядром и завязаны на идентификацию анклава и машины. Сами анклавы никогда не будут иметь доступа к этим ключам, вместо этого безопасное ядро предлагает возможности по запечатыванию и распечатыванию произвольного содержимого (с помощью `EnclaveSealData` и `EnclaveUnsealData`) с использованием соответствующего ключа,

назначенного анклавом. Во время запечатывания данных предоставляется набор параметров, которые контролируют, каким анклавам разрешено распечатывать данные. Поддерживаются следующие политики.

- **Номер версии защиты (SVN) безопасного ядра и основного образа.** Ни один анклав не сможет распечатать любые данные, запечатанные более поздней версией анклава или безопасного ядра.
- **Точный код.** Данные могут быть распечатаны только анклавом, который отображает те же самые модули, что и анклав, который их запечатал. Безопасное ядро проверяет хеш уникального идентификатора каждого образа, отраженного в анклаве, чтобы обеспечить корректное распечатывание.
- **Те же образ, семейство или автор.** Данные могут быть распечатаны только анклавом, имеющим тот же идентификатор автора, семьи и/или образа.
- **Политика времени выполнения.** Данные могут быть распечатаны только в том случае, если распечатывающий анклав имеет ту же политику отладки, что и запечатавший (отладка разрешена или не разрешена).

Каждый анклав может подтвердить любой третьей стороне, что он работает как анклав VBS со всеми средствами защиты, предлагаемыми архитектурой анклавов VBS. Отчет об аттестации анклава предоставляет доказательство того, что конкретный анклав работает под контролем безопасного ядра. Отчет об аттестации содержит идентификационные данные всего кода, загруженного в анклав, равно как и все политики, контролирующие исполнение в нем.

Описание внутренних деталей операций запечатывания и аттестации выходит за рамки этой книги. Анклав может создать отчет об аттестации с помощью интерфейса `EnclaveGetAttestationReport`. Буфер памяти, возвращаемый интерфейсом, может быть передан в другой анклав, который сможет подтвердить целостность среды, в которой работал исходный анклав, проверив отчет об аттестации с помощью функции `EnclaveVerifyAttestationReport`.

Аттестация среды выполнения System Guard

Аттестация среды выполнения System Guard (System Guard Runtime Attestation, SGRA) — это компонент поддержания целостности ОС, который вместе с компонентом службы удаленной аттестации помогает вышеупомянутым анклавам VBS, обеспечивая надежные гарантии в отношении своей среды исполнения. Эта среда используется для подтверждения критичных свойств системы во время выполнения и позволяет проверяющей стороне наблюдать за нарушениями требований безопасности, которые система обязуется обеспечить. Первая реализация этой новой технологии была представлена в обновлении Windows от 10 апреля 2018 г. (RS4).

SGRA позволяет приложению просматривать заявление о состоянии безопасности устройства. Это заявление состоит из трех частей:

- отчета о сеансе, куда входит уровень безопасности, описывающий проверяемые свойства устройства в момент загрузки системы;
- отчета времени выполнения, описывающего состояние устройства во время выполнения;

- подписанного сертификата сеанса, который можно использовать для проверки отчетов.

Служба SGRA, SgrmBroker.exe, содержит компонент (SgrmEnclave_secure.dll), который работает в VTL 1 как анклав VBS и постоянно проверяет систему на предмет нарушения функций безопасности во время выполнения. Эти проверки отображаются в отчете времени выполнения, который может быть проверен на серверной стороне проверяющей частью. Поскольку проверки выполняются в отдельном домене доверия, непосредственная атака на содержимое отчета времени выполнения становится затруднительной.

Внутреннее устройство SGRA

На рис. 9.41 показан общий обзор архитектуры аттестации среды выполнения System Guard Защитника Windows, в которую входят следующие клиентские компоненты:

- SgrmEnclave_secure.dll — механизм проверок VTL 1;
- SgrmAgent.sys — агент режима ядра VTL 0;
- SgrmBroker.exe — защищенный брокерский процесс VTL 0 WinTCB, в рамках которого действует механизм проверок;
- SgrmLpac.exe — процесс VTL 0 LPAC, используемый процессом брокера WinTCBPP для взаимодействия с сетевым стеком.

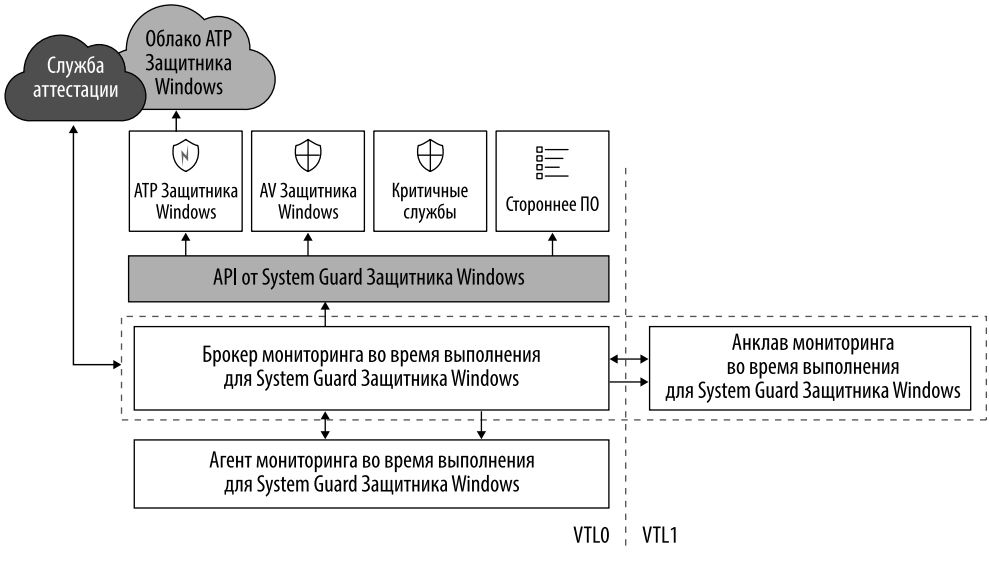


Рис. 9.41. Архитектура аттестации среды выполнения в рамках System Guard Защитника Windows

Чтобы иметь возможность быстро реагировать на угрозы, SGRA содержит средство динамического исполнения скриптов (на языке Lua), что служит основой механизма проверок, который действует в анклаве VTL 1. Данный подход позволяет часто обновлять логику проверок.

Из-за изоляции, обеспечиваемой анклавом VBS, потоки, выполняющиеся в VTL 1, ограничены в части возможности доступа к API-интерфейсам NT в условиях VTL 0. Таким образом, чтобы компонент времени выполнения SGRA мог проделывать значимую работу, необходим способ обхода ограничений анклава VBS на доступ к API.

Чтобы предоставить средства VTL 0 логике, работающей в VTL 1, реализован принцип агента: эти средства называются *помощниками* (assists) и обслуживаются компонентом пользовательского режима SgrmBroker или же драйвером агента, работающим в режиме ядра VTL 0 (SgrmAgent.sys). Логика VTL 1, выполняющаяся в анклаве, может обращаться к этим компонентам из VTL 0 с запросами о помощи в части ряда возможностей, в том числе примитивов синхронизации ядра NT, функции отражения страниц и т. д.

В качестве примера работы этого механизма можно привести подсистему SGRA, позволяющую механизму проверок из VTL 1 напрямую читать физические страницы, принадлежащие VTL 0. Анклав запрашивает отражение произвольной страницы через помощник. Искомая страница будет заблокирована и отражена в адресное пространство SgrmBroker в VTL 0, что сделает ее резидентной. Поскольку анклавы VBS имеют прямой доступ к адресному пространству хост-процесса, защищенная логика может считывать данные непосредственно по отраженным виртуальным адресам. Эти операции чтения должны быть синхронизированы непосредственно с ядром VTL 0. Резидентный агент-брокер уровня VTL 0 (драйвер SgrmAgent.sys) также используется для выполнения синхронизации.

Логика проверок

Как упоминалось ранее, SGRA оценивает показатели безопасности системы во время выполнения. Проверки проводит соответствующий механизм, действующий в анклаве на базе VBS. Подписанный байт-код Lua, описывающий логику проверок, передается последнему во время запуска.

Проверки выполняются периодически. Когда обнаруживается нарушение заявленного свойства (то есть когда утверждение, лежащее в основе проверки, не работает), сбой записывается и сохраняется в анклаве. Он будет доступен проверяющей стороне в отчете среды выполнения, который создается и подписывается (с сертификатом сеанса) внутри анклава.

Примером возможностей оценки, предоставляемых SGRA, являются проверки, окружающие различные атрибуты объекта исполнительного процесса, например периодическое перечисление запущенных процессов и оценка состояния битов защиты процесса, управляющих политиками защищенных процессов. Последовательность действий механизма проверок, выполняющего эту работу, можно свести к следующим шагам.

1. Механизм проверок, работающий в VTL 1, обращается к своему хост-процессу в VTL 0 (SgrmBroker) с запросом, чтобы ядро ссылалось на объект исполнительного процесса.
2. Процесс-брокер пересылает этот запрос агенту режима ядра (SgrmAgent), который обслуживает его, получая ссылку на запрошенный объект исполнительного процесса.

3. Агент уведомляет брокера о том, что запрос обслужен, и передает ему все необходимые метаданные.
4. Брокер пересылает этот ответ запрашивающей логике проверки в VTL 1.
5. Затем логика может выбрать, чтобы физическая страница, обеспечивающая указанный объект исполнительного процесса, была заблокирована и отображена в доступное ей адресное пространство. Это делается путем вызова из анклава, аналогичного шагам 1–4.
6. После отражения страницы механизм из VTL 1 сможет прочитать ее напрямую и проверить бит защиты объекта исполнительного процесса на соответствие внутреннему контексту.
7. Логика VTL 1 снова обращается к VTL 0, чтобы прекратить отражение страниц и сбросить ссылку ядра на объект.

Отчеты и установление доверия

Чтобы позволить проверяющим сторонам получить сертификат сеанса SGRA, а также подписанные отчеты о сеансе и времени выполнения, им предоставляется API на основе WinRT. Он не является общедоступным и предоставляется в соответствии с соглашением о неразглашении поставщикам, являющимся участниками Microsoft Virus Initiative. (Обратите внимание на то, что Microsoft Defender Advanced Threat Protection (улучшенная защита от угроз Защитника Windows) сейчас является единственным компонентом «из коробки», напрямую взаимодействующим с SGRA через этот API.)

Порядок получения заявления о доверии от SGRA следующий.

1. Между проверяющей стороной и SGRA создается сеанс. Для его установления требуется подключение к сети. Механизм проверок SgrmEnclave, работающий в VTL-1, генерирует ключевую пару типа «открытый/закрытый», а защищенный процесс SgrmBroker получает журнал TCG и отчет об аттестации VBS, которые отправляет в службу аттестации Microsoft System Guard вместе с публичной частью ключа, сгенерированного чуть ранее.
2. Служба аттестации проверяет журнал TCG (из TPM) и отчет о аттестации VBS (как доказательство того, что логика работает в анклаве VBS) и создает отчет о сеансе, описывающий аттестованные свойства времени загрузки устройства. Она подписывает открытый ключ промежуточным ключом службы аттестации SGRA для создания сертификата, который будет использоваться для проверки отчетов во время выполнения.
3. Отчет о сеансе и сертификат возвращаются проверяющей стороне. С этого момента она сможет проверять достоверность отчета о сеансе и сертификата времени выполнения.
4. Периодически проверяющая сторона может запрашивать отчет времени выполнения у SGRA, используя установленный сеанс: механизм проверок SgrmEnclave генерирует отчет времени выполнения, описывающий состояние проведенных проверок. Отчет будет подписан с применением парного закрытого ключа, сгенерированного во время создания сеанса, и возвращен проверяющей стороне (закрытый ключ никогда не покидает анклав).

5. Проверяющая сторона может оценить достоверность отчета времени выполнения по сертификату времени выполнения, полученному ранее, и принять решение в части политик на основе содержимого как отчета о сеансе (подтвержденное состояние во время загрузки), так и отчета о времени выполнения (проверяемое состояние).

SGRA предоставляет API, который проверяющие стороны могут использовать для подтверждения состояния устройства в определенный момент. API возвращает отчет времени выполнения, в котором подробно описаны заявления о состоянии безопасности системы, сделанные аттестацией среды выполнения System Guard в Защитнике Windows. Эти утверждения подразумевают проверки, которые представляют собой измерения конфиденциальных свойств системы во время выполнения. Например, приложение может запросить у System Guard Защитника Windows оценку безопасности системы из аппаратного анклава и вернуть отчет. Приводимые в нем подробности приложение может использовать, чтобы решить, можно ли ему выполнить конфиденциальную финансовую транзакцию или отобразить личные данные.

Как обсуждалось в предыдущем разделе, анклав на базе VBS также может предоставлять отчет об аттестации анклава, подписанный с особым ключом подписи VBS. Если System Guard Защитника Windows сможет получить подтверждение того, что хост-система работает с активным VSM, это вместе с подписанным отчетом о сеансе подойдет в качестве доказательства того, что конкретный анклав запущен и работает. Таким образом, для установления доверия, необходимого, чтобы гарантировать подлинность отчета времени выполнения, требуется следующее.

1. Подтверждение состояния загрузки машины. Двоичные файлы ОС, гипервизора и безопасного ядра должны быть подписаны Microsoft и настроены в соответствии с политикой безопасности.
2. Привязка доверия между TPM и работоспособностью гипервизора, чтобы обеспечить доверие к журналу измерений загрузки.
3. Извлечение необходимого ключа (VSM IDK) из журнала измерений загрузки и использование его для проверки подписи анклава VBS (подробности см. в главе 12).
4. Подписание открытого компонента пары эфемерных ключей, созданной в анклаве, с доверенным центром сертификации для выдачи сертификата сеанса.
5. Подписание отчета времени выполнения эфемерным закрытым ключом.

Сетевые вызовы между анклавом и службой аттестации System Guard Защитника Windows выполняются из VTL 0. Однако модель протокола аттестации гарантирует его устойчивость к несанкционированному вмешательству даже в условиях ненадежных транспортных механизмов.

Прежде чем описанная ранее цепочка доверия сможет быть в достаточной степени установлена, требуется применить ряд технологий нижнего уровня. Чтобы информировать проверяющую сторону об уровне доверия к отчету среды выполнения, которого она может ожидать в любой конкретной конфигурации, каждому отчету о сеансе, подписанному службой аттестации System Guard Защитника Windows, назначается уровень безопасности. Последний отражает базовые технологии, действующие на платформе, и определяет уровень доверия, основанный на ее возможностях.

Microsoft сопоставляет наличие различных технологий безопасности с уровнями безопасности и делится результатами, когда API публикуется для использования третьими лицами. Самый высокий уровень доверия, вероятно, потребует как минимум следующих функций:

- аппаратного обеспечения с поддержкой VBS и OEM-конфигурации;
- динамических измерений корневого доверия при загрузке;
- безопасной загрузки для проверки образов гипервизора, NT и SK;
- политики безопасности, обеспечивающей создаваемую гипервизором целостность кода (HVCI) и целостность кода режима ядра (KMCI), отключение тестовой подписи и отладки ядра;
- наличия драйвера ELAM.

ЗАКЛЮЧЕНИЕ

Windows может управлять несколькими виртуальными машинами и запускать их благодаря гипервизору Hyper-V и его стеку виртуализации, которые в совокупности поддерживают различные операционные системы, работающие на виртуальной машине. С годами эти два компонента развивались, чтобы обеспечить больше возможностей оптимизации и расширенных функций для виртуальных машин, таких как вложенная виртуализация, вариации планировщика для виртуальных процессоров, различные типы поддержки виртуального оборудования, VMBus, виртуальные машины с поддержкой VA и т. д.

Безопасность на основе виртуализации обеспечивает корневой операционной системе новый уровень защиты от вредоносных программ и скрытых руткитов, которые больше не способны похищать частную и конфиденциальную информацию из памяти корневой операционной системы. Безопасное ядро использует сервисы, предоставляемые гипервизором Windows, для создания новой защищенной среды выполнения (VTL 1), недоступной для программного обеспечения, действующего в основной ОС. Кроме того, безопасное ядро привносит в экосистему Windows ряд сервисов, которые помогают поддерживать более безопасную среду.

Безопасное ядро также вводит понятие изолированного пользовательского режима, позволяющего выполнять код пользовательского режима в новой защищенной среде через трастлеты, защищенные устройства и анклавы. В конце главы был изучен механизм аттестации времени выполнения System Guard — компонент, который задействует сервисы, предоставляемые безопасным ядром, для оценки среды выполнения рабочей станции и предоставления надежных гарантий ее целостности.

В следующей главе мы рассмотрим компоненты управления и диагностики Windows, а также обсудим важные механизмы, связанные с их инфраструктурой: реестр, службы, планировщик задач, инструментарий управления Windows (WMI), журналирование событий ядра и т. д.

ГЛАВА 10

Управление, диагностика и трассировка

В этой главе описываются фундаментальные механизмы операционной системы Microsoft Windows, имеющие критическое значение для ее конфигурирования и управления ею. В частности, рассмотрим реестр Windows, службы, унифицированный диспетчер фоновых процессов и инструментарий управления Windows (WMI). В главе также будут представлены некоторые фундаментальные компоненты, используемые для диагностики и отслеживания, такие как отслеживание событий для Windows (ETW), средство уведомлений Windows (WNF) и отчеты об ошибках Windows (WER). Глава завершается обсуждением глобальных флагов Windows и кратким введением в ядро и механизм User Shim Engine.

РЕЕСТР

Реестр играет ключевую роль в настройке систем Windows и управлении ими. Это хранилище как общесистемных, так и индивидуальных для пользователя настроек. Большинство людей думают о реестре как о статических данных, хранящихся на жестком диске, но, как вы увидите в данном разделе, он является также окном в различные структуры памяти, поддерживаемые исполнительной системой Windows и ядром.

Мы начнем с обзора структуры реестра, обсуждения типов данных, которые он поддерживает, и краткого описания ключевых сведений, хранимых в реестре Windows. Затем рассмотрим содержимое диспетчера конфигурации — исполнительного компонента, отвечающего за реализацию базы данных реестра. Среди тем, которых мы коснемся, будут внутренняя структура реестра на диске, способы получения Windows информации о конфигурации, когда приложение ее запрашивает, и меры, принимаемые для защиты этой критически важной системной базы данных.

Просмотр и изменение реестра

Как правило, у вас не должно быть повода редактировать реестр напрямую. Параметры приложений и системы, хранящиеся там и требующие изменения, должны иметь соответствующий пользовательский интерфейс для управления их изменениями. Однако, как неоднократно упоминалось в данной книге, некоторые расширенные

и отладочные настройки не имеют пользовательского интерфейса для редактирования. Таким образом, в состав Windows входят как инструменты с графическим интерфейсом пользователя (GUI), так и инструменты для командной строки, позволяющие просматривать и изменять реестр.

Windows поставляется с одним основным инструментом с графическим интерфейсом для редактирования реестра — Regedit.exe — и несколькими инструментами, запускаемыми из командной строки. Например, Reg.exe позволяет импортировать и экспортировать разделы, делать резервную копию и восстанавливать их, а также сравнивать, изменять и удалять разделы и параметры. Он также может устанавливать или запрашивать флаги, используемые при виртуализации UAC. Вдобавок Regini.exe позволяет импортировать данные реестра на основе текстовых файлов, содержащих данные конфигурации в кодировке ASCII или Юникод.

В комплект драйверов Windows (WDK) входит распространяемый компонент Offregs.dll, в котором содержится библиотека автономного реестра. Она позволяет загружать файлы кустов реестра (описаны в разделе «Кусты» далее в этой главе) в двоичном формате и выполнять операции над самими файлами, минуя обычную логическую загрузку и сопоставление, которые Windows требует для операций с реестром. Ее назначение в первую очередь включает в себя поддержку автономного доступа к реестру, например для проверки целостности и валидности. Библиотека также может обеспечить выигрыш в производительности в случаях, когда низкоуровневые данные не должны быть видны системе, поскольку вместо системных вызовов в отношении реестра доступ осуществляется через локальный файловый ввод-вывод.

Использование реестра

Существуют четыре основные ситуации, когда считываются данные конфигурации.

- В начале загрузчик считывает данные конфигурации и список драйверов загрузочных устройств для загрузки в память перед инициализацией ядра. Поскольку база данных конфигурации загрузки (BCD) на самом деле хранится в кусте реестра, можно утверждать, что доступ к реестру происходит даже раньше, когда диспетчер загрузки отображает список операционных систем.
- В процессе загрузки ядро считывает настройки, которые определяют, какие драйверы устройств следует загружать и как различные элементы системы, такие как диспетчер памяти и диспетчер процессов, настраивают себя и поведение системы.
- Во время входа в систему Проводник и другие компоненты Windows считывают из реестра настройки каждого пользователя, включая сопоставление букв сетевых дисков, обои Рабочего стола, заставку, поведение меню, размещение значков и, что, возможно, наиболее важно, какие программы запускать при старте и какие файлы недавно открывались.
- Во время запуска приложения считывают общесистемные настройки, такие как список дополнительно установленных компонентов и данные лицензирования. Кроме того, они считывают индивидуальные настройки пользователей, куда могут входить способы размещения меню и панелей инструментов, а также список документов, к которым обращались в прошлый раз.

Реестр можно читать и в других случаях, например в ответ на изменение параметра или раздела реестра. Хотя реестр позволяет выполнять асинхронные обратные вызовы, являющиеся предпочтительным способом получения уведомлений об изменениях, некоторые приложения отслеживают свои параметры конфигурации в реестре постоянно путем опроса и автоматически учитывают обновленные параметры. Однако в простаивающей системе в целом не должно быть активного взаимодействия с реестром, а такие приложения отступают от рекомендуемых практик. (Process Monitor из пакета Sysinternals — отличный инструмент для отслеживания подобной активности и виновных в ней приложений.)

Изменения в реестр обычно вносятся в следующих случаях.

- Хотя это и не является модификацией, первоначальная структура реестра и многие параметры по умолчанию определяются прототипной версией реестра, поставляемой на установочном носителе Windows и копируемой в новый экземпляр инсталляции.
- Установщики приложений создают параметры приложения по умолчанию и записывают настройки, отражающие выбор варианта конфигурации при установке.
- Во время установки драйвера устройства система Plug and Play создает в реестре параметры, которые сообщают диспетчеру ввода-вывода, как запускать драйвер, а также другие параметры его работы. (Дополнительную информацию о процессе установки драйверов устройств см. в главе 6 тома 1.)
- Когда вы меняете настройки приложения или системы через пользовательский интерфейс, изменения зачастую сохраняются в реестре.

Типы данных реестра

Реестр представляет собой базу данных, структура которой похожа на структуру дискового тома. Реестр содержит *разделы*, аналогичные каталогам диска, и *параметры*, сравнимые, по сути, с файлами на диске. Раздел — это контейнер, который может состоять из других разделов (подразделов) или параметров. Последние, в свою очередь, хранят данные. Разделы верхнего уровня являются корневыми. Мы будем использовать слова *«подраздел»* и *«раздел»* как взаимозаменяемые.

И разделы, и параметры заимствуют соглашения об именах из файловой системы. Таким образом, вы можете однозначно идентифицировать параметр `mark`, который хранится в разделе `trade`, по имени `trade\mark`. Единственное исключение из этой схемы именования — безымянный параметр каждого раздела. Regedit отображает безымянный параметр как `(По умолчанию)`.

В параметрах хранятся различные виды данных, которые могут относиться к одному из 12 типов, перечисленных в табл. 10.1. Для большинства параметров реестра это `REG_DWORD`, `REG_BINARY` или `REG_SZ`. Параметры типа `REG_DWORD` могут хранить числа или логические значения (значения `true`/`false`), параметры REG_BINARY — числа размером более 32 бит или необработанные данные, например зашифрованные пароли. Параметры `REG_SZ` хранят строки (разумеется, в кодировке Юникод), которые могут представлять такие элементы, как наименования, имена файлов, пути и типы.

Таблица 10.1. Типы параметров реестра

Тип параметра	Описание
REG_NONE	Нет типа параметра
REG_SZ	Строка Юникод фиксированной длины
REG_EXPAND_SZ	Строка Юникод переменной длины, которая может содержать вложенные переменные среды
REG_BINARY	Двоичные данные произвольной длины
REG_DWORD	32-битное число
REG_DWORD_BIG_ENDIAN	32-битное число, начиная со старшего байта
REG_LINK	Символическая ссылка Юникод
REG_MULTI_SZ	Массив строк Юникод, завершающихся NULL
REG_RESOURCE_LIST	Описание аппаратного ресурса
REG_FULL_RESOURCE_ DESCRIPTOR	Описание аппаратного ресурса
REG_RESOURCE_ REQUIREMENTS_LIST	Требования к ресурсам
REG_QWORD	64-битное число

Тип `REG_LINK` особенно интересен, поскольку позволяет разделу прямо указывать на другой раздел. Когда вы перемещаетесь по реестру по ссылке, поиск пути продолжается до цели ссылки. Например, если `\Root1\Link` имеет параметр `\Root2\RegKey` типа `REG_LINK`, а `RegKey` содержит параметр `RegValue`, то последний идентифицируют два пути: `\Root1\Link\RegValue` и `\Root2\RegKey\RegValue`. Как объясняется в следующем разделе, Windows активно использует ссылки реестра: три из шести корневых разделов реестра — это просто ссылки на подразделы в трех других, нессылочных корневых разделах.

Логическая структура реестра

Вы можете составить схему организации реестра с помощью хранящихся в нем данных. Существуют девять корневых разделов, в которых хранится информация (табл. 10.2), добавлять новые корневые разделы или удалять существующие нельзя.

Таблица 10.2. Девять корневых разделов

Корневой раздел	Описание
HKEY_CURRENT_USER	Хранит данные, связанные с пользователем, вошедшим в систему в данный момент
HKEY_CURRENT_USER_LOCAL_ SETTINGS	Хранит данные, связанные с пользователем, вошедшим в систему в данный момент, которые являются локальными для компьютера и исключены из перемещаемого профиля пользователя

Продолжение ⇩

Таблица 10.2 *(продолжение)*

Корневой раздел	Описание
HKEY_USERS	Хранит информацию обо всех учетных записях на компьютере
HKEY_CLASSES_ROOT	Хранит информацию об ассоциации файлов и регистрации объектов модели COM (Component Object Model)
HKEY_LOCAL_MACHINE	Хранит системную информацию
HKEY_PERFORMANCE_DATA	Хранит данные о производительности
HKEY_PERFORMANCE_NLSTEXT	Хранит текстовые строки, описывающие счетчики производительности на языке региона, в котором работает данная компьютерная система
HKEY_PERFORMANCE_TEXT	Хранит текстовые строки, описывающие счетчики производительности на американском английском языке
HKEY_CURRENT_CONFIG	Хранит некоторые данные о текущем профиле оборудования (устарел)

Почему имена корневых разделов начинаются с буквы «H»? Потому что они представляют собой дескрипторы Windows (H — handle) для разделов (KEY). Как упоминалось в главе 1 тома 1, HKLM — это аббревиатура, используемая для HKEY_LOCAL_MACHINE. В табл. 10.3 перечислены все корневые разделы и их сокращения. В последующих разделах подробно объясняются содержание и назначение каждого из этих корневых разделов.

Таблица 10.3. Корневые разделы реестра

Корневой раздел	Сокращение	Описание	Ссылка
HKEY_CURRENT_USER	HKCU	Указывает на профиль пользователя, вошедшего в систему в данный момент	Подраздел в разделе HKEY_USERS, соответствующий текущему пользователю, вошедшему в систему
HKEY_CURRENT_USER_LOCAL_SETTINGS	HKCULS	Указывает на локальные настройки текущего пользователя, вошедшего в систему	Ссылка на HKCU\Software\Classes\Local Settings
HKEY_USERS	HKU	Содержит подразделы для всех загруженных профилей пользователей	Не ссылка
HKEY_CLASSES_ROOT	HKCR	Содержит информацию об ассоциациях типов файлов и регистрации COM	Не прямая ссылка, а скорее объединенное представление HKLM\SOFTWARE\Classes и HKEY_USERS\<SID>\SOFTWARE\Classes
HKEY_LOCAL_MACHINE	HKLM	Глобальные настройки компьютера	Не ссылка

Корневой раздел	Сокращение	Описание	Ссылка
HKEY_CURRENT_CONFIG	HKCC	Текущий профиль оборудования	HKLM\SYSTEM\CurrentControlSet\Hardware Profiles\Current
HKEY_PERFORMANCE_DATA	HKPD	Счетчики производительности	Не ссылка
HKEY_PERFORMANCE_NLSTEXT	HKPNT	Текстовые строки счетчиков производительности	Не ссылка
HKEY_PERFORMANCE_TEXT	HKPT	Текстовые строки счетчиков производительности на американском английском	Не ссылка

HKEY_CURRENT_USER

Корневой раздел HKCU содержит данные о настройках и конфигурации программного обеспечения локально вошедшего в систему пользователя. Он указывает на его профиль, расположенный на жестком диске в папке \Users\<имя пользователя>\Ntuser.dat. (См. раздел «Внутреннее устройство реестра» далее в данной главе, чтобы узнать, как корневые разделы сопоставляются с файлами на жестком диске.) Всякий раз, когда загружается профиль пользователя, например во время входа в систему или когда процесс службы запускается в контексте определенной учетной записи пользователя, HKCU создается для отображения раздела пользователя в разделе HKEY_USERS (таким образом, если в систему вошли несколько пользователей, каждый из них увидит свой HKCU). В табл. 10.4 перечислены некоторые подразделы HKCU.

Таблица 10.4. Подразделы HKEY_CURRENT_USER

Подраздел	Описание
AppEvents	Ассоциации звуков/событий
Console	Настройки окна команд (например, ширина, высота и цвета)
Control Panel	Заставка, тема Рабочего стола, настройки клавиатуры и мыши, специальных возможностей, а также региональные настройки
Environment	Определения переменных среды
EUDC	Информация о символах, определяемых конечным пользователем
Keyboard Layout	Настройки раскладки клавиатуры
Network	Отображение сетевых дисков и их монтирование
Printers	Настройки подключения принтера
Software	Пользовательские настройки программного обеспечения
Volatile Environment	Определения переменных нестабильной среды

HKEY_USERS

HKU содержит подраздел для каждого загруженного профиля пользователя и базу данных регистрации классов пользователей в системе. Там же находится подраздел HKU\.DEFAULT, связанный с профилем системы (задействуется процессами,

запущенными под учетной записью локальной системы, более подробно описан в разделе «Службы» далее в данной главе). В частности, этот профиль применяется Winlogon, поэтому изменения настроек фона Рабочего стола в нем отразятся на экране входа в систему. Когда пользователь входит в систему впервые, а его учетная запись не зависит от перемещаемого профиля домена (то есть профиль пользователя приходит из центрального сетевого расположения по указанию контроллера домена), система создает профиль для его учетной записи на основе профиля, хранящегося в `%SystemDrive%\Users\Default`.

Место, где система хранит профили, определяется параметром реестра HKLM\Software\Microsoft\Windows NT\CurrentVersion\ProfileList\ProfilesDirectory, для которого по умолчанию установлено значение `%SystemDrive%\Users`. Раздел `ProfileList` также хранит список профилей, присутствующих в системе. Информация для каждого профиля находится в подразделе, имя которого отражает идентификатор безопасности (SID) учетной записи, которой соответствует профиль. (Дополнительную информацию о SID см. в главе 7 тома 1.) Данные, хранящиеся в разделе профиля, включают время последней загрузки профиля в параметре `LocalProfileLoadTimeLow`, двоичное представление SID учетной записи — в параметре `Sid`, параметр и путь к кусту профиля на диске (файл Ntuser.dat, описанный далее в разделе «Кусты») — в каталоге, заданном параметром `ProfileImagePath`. Windows отображает профили, хранящиеся в системе, в диалоговом окне управления профилями пользователей (рис. 10.1), доступ к которому можно получить, щелкнув на Настройка дополнительных свойств профиля пользователя (Configure Advanced User Profile Properties) в апплете Учетные записи пользователей (User Accounts) панели управления.

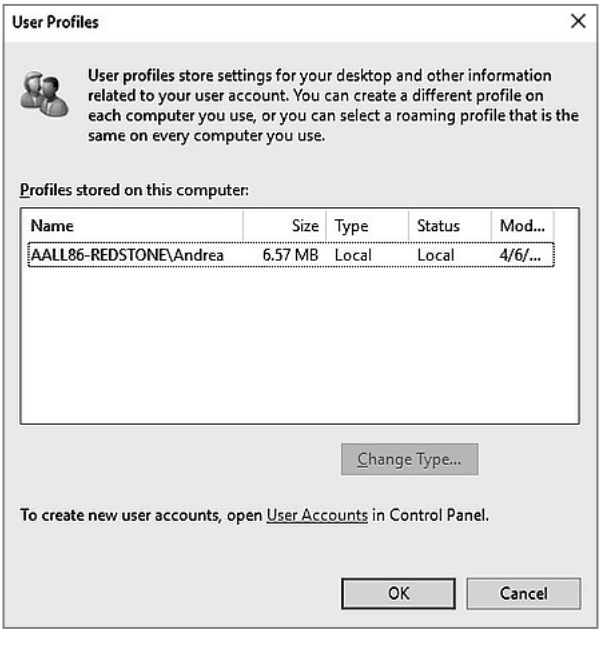
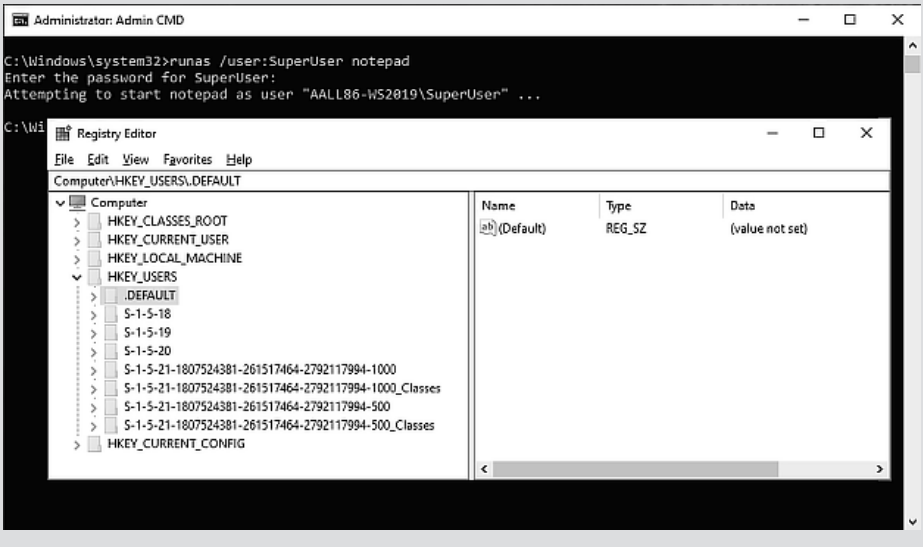


Рис. 10.1. Диалоговое окно управления профилями пользователей

ЭКСПЕРИМЕНТ. Наблюдение за загрузкой и выгрузкой профиля

Вы можете наблюдать за загрузкой профиля в реестр, а затем его выгрузкой с помощью команды Runas, запуская процесс в учетной записи, которая в данный момент не авторизована на этом компьютере. Пока новый процесс исполняется, запустите Regedit и обратите внимание на загруженный раздел профиля в разделе HKEY_USERS. После завершения процесса обновите Regedit, нажав клавишу F5, и профиль оттуда исчезнет.



HKEY_CLASSES_ROOT

HKCR содержит информацию трех типов: ассоциации расширений файлов, регистрацию классов COM и виртуальный корень реестра для контроля учетных записей пользователей (UAC). (Дополнительную информацию о UAC см. в главе 7 тома 1.) Для каждого зарегистрированного расширения имени файла существует раздел. Большинство разделов содержат параметр `REG_SZ`, указывающий на другой раздел в HKCR, хранящий информацию об ассоциации для класса файлов, который представляет данное расширение.

Например, HKCR\.xls будет указывать на информацию о файлах Microsoft Office Excel. В частности, параметр по умолчанию содержит строку `Excel.Sheet.8`, которая используется для создания экземпляра COM-объекта Excel. Другие разделы содержат сведения о конфигурации для всех COM-объектов, зарегистрированных в системе. Виртуализованный реестр UAC расположен в разделе `VirtualStore`, который не связан с другими видами данных, хранящихся в HKCR.

Данные, находящиеся в разделе HKEY_CLASSES_ROOT, поступают из двух источников:

- регистрационные данные классов в разрезе пользователей в HKCU\SOFTWARE\Classes сопоставляются с файлом на жестком диске \Users\<имя пользователя>\AppData\Local\Microsoft\Windows\Usrclass.dat;
- общесистемные регистрационные данные классов в HKLM\SOFTWARE\Classes.

Регистрационные данные каждого пользователя отделены от общесистемных, чтобы перемещаемые профили могли содержать персональные настройки. Непривилегированные пользователи и приложения могут читать общесистемные данные и добавлять новые разделы и параметры (которые отразятся в их пользовательских данных), но способны изменять только существующие разделы и параметры в своих данных. Это также закрывает дыру в безопасности: непривилегированный пользователь не сможет изменить или удалить разделы в общесистемной версии HKEY_CLASSES_ROOT и таким образом повлиять на работу приложений в системе.

HKEY_LOCAL_MACHINE

HKLM — это корневой раздел, который содержит все подразделы общесистемной конфигурации: BCD00000000, COMPONENTS (загружается динамически по мере необходимости), HARDWARE, SAM, SECURITY, SOFTWARE и SYSTEM.

Подраздел HKLM\BCD00000000 содержит информацию из базы данных конфигурации загрузки (Boot Configuration Database, BCD), загруженную в виде куста реестра. Эта база данных заменяет файл Boot.ini, который использовался до Windows Vista, и обеспечивает большую гибкость и изоляцию данных конфигурации загрузки для каждой инсталляции. Подраздел BCD00000000 поддерживается скрытым файлом BCD, который в системах UEFI находится в \EFI\Microsoft\Boot. (Дополнительную информацию о BCD см. в главе 12.)

Каждая запись в BCD, как то установка Windows или параметры командной строки для установки, хранится в подразделе Objects либо как объект, на который ссылается GUID (в случае загрузочной записи), либо как числовой подраздел, называемый *элементом*. Большинство этих необработанных элементов описаны в справочнике BCD в Microsoft Docs и определяют различные параметры командной строки или загрузки. Параметр, связанный с каждым подразделом элемента, соответствует значению флага командной строки или параметра загрузки.

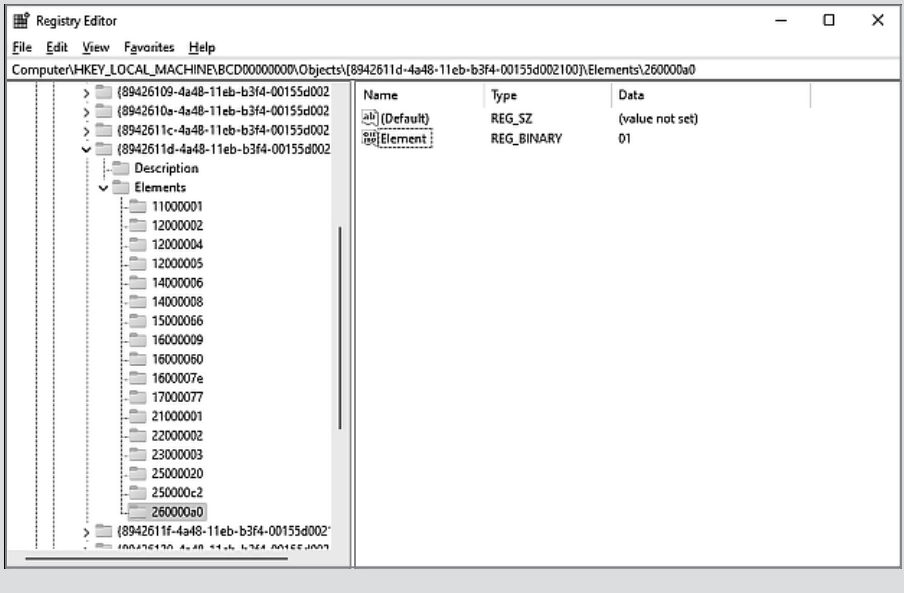
Утилита командной строки BCDEdit позволяет изменять BCD, используя символические имена для элементов и объектов. Она также предоставляет обширную справку по всем доступным вариантам загрузки. Куст реестра можно открыть удаленно, а также импортировать из файла куста: вы можете изменить или прочитать BCD на удаленном компьютере с помощью редактора реестра. В следующем эксперименте показано, как с помощью редактора реестра активировать отладку ядра.

ЭКСПЕРИМЕНТ. Удаленное редактирование BCD

Хотя можно изменить автономные хранилища BCD с помощью команды `bcdedit/store`, в рамках данного эксперимента вы включите отладку, отредактировав хранилище BCD внутри реестра. В данном случае редактируется локальная копия BCD, но смысл этого метода в том, что его можно использовать

в отношении куста BCD на любом другом компьютере. Выполните следующие действия, чтобы добавить флаг командной строки `/DEBUG`.

1. Откройте редактор реестра и перейдите в раздел HKLM\BCD00000000. Разверните каждый подраздел так, чтобы числовые идентификаторы каждого раздела Elements были видны полностью.
2. Определите загрузочную запись для своей установки Windows, найдя раздел Description с параметром `Type`, равным 0x10200003, а затем в дереве `Elements` выберите раздел 12000004. В параметре Element этого подраздела вы должны найти название своей версии Windows, например Windows 10. В недавно разработанных системах у вас может быть более одной установки Windows или различных загрузочных приложений, таких как среда восстановления Windows или приложение возобновления Windows. В таких случаях может потребоваться проверить подраздел 22000002 раздела Elements, где будет содержаться путь, например `\Windows`.
3. Теперь, когда вы нашли правильный GUID для своей установки Windows, создайте новый подраздел в подразделе `Elements` для этого GUID и назовите его 0x260000a0. Если этот подраздел уже существует, просто перейдите к нему. Найденный GUID должен соответствовать параметру идентификатора в разделе Загрузчик Windows (Windows Boot Loader), показанному командой `bcdedit/v` (можете использовать параметр командной строки `/store`, чтобы просмотреть файл автономного хранилища).
4. Если вы создавали подраздел, теперь добавьте в него двоичный параметр с именем `Element`.
5. Отредактируйте параметр и установите его равным 1. Это включит отладку в режиме ядра. Вот как должны выглядеть эти изменения.



ПРИМЕЧАНИЕ Идентификатор 0x12000004 соответствует аргументу BcdLibraryString_ApplicationPath, тогда как идентификатор 0x22000002 — BcdOSLoaderString_SystemRoot. Наконец, добавленный вами идентификатор 0x260000a0 соответствует BcdOSLoaderBoolean_KernelDebuggerEnabled. Эти значения описаны в справочнике BCD в Microsoft Docs.

Подраздел HKLM\COMPONENTS содержит информацию, относящуюся к стеку обслуживания на основе компонентов (Component Based Servicing, CBS). Этот стек содержит различные файлы и ресурсы, являющиеся частью установочного образа Windows (задействуется пакетом автоматической установки или пакетом предварительной установки OEM) либо активной инсталляции. API CBS, предназначенные для обслуживания, используют информацию, расположенную в данном разделе, для идентификации установленных компонентов и информации об их конфигурации. Эти данные используются всякий раз, когда компоненты устанавливаются, обновляются или удаляются индивидуально (называются *модулями*) или группами *(пакетами)*. Поскольку этот раздел может стать довольно большим, из соображений оптимизации системных ресурсов он загружается и выгружается динамически и только по мере необходимости, когда стек CBS обслуживает запрос. Этот раздел поддерживается файлом куста COMPONENTS, расположенным в \Windows\system32\config.

Подраздел HKLM\HARDWARE содержит описания устаревшего оборудования системы и некоторые сопоставления аппаратных устройств с драйверами. В современной системе здесь, скорее всего, можно найти лишь несколько периферийных устройств, таких как клавиатура, мышь и данные ACPI BIOS. Инструмент Диспетчер устройств позволяет просматривать из реестра информацию об оборудовании, которую он получает, просто считывая параметры по разделу HARDWARE (хотя чаще всего используется дерево HKLM\SYSTEM\CurrentControlSet\Enum).

HKLM\SAM содержит информацию о локальных учетных записях и группах — пароли пользователей, определения групп и ассоциации доменов. Системы Windows Server, действующие в роли контроллеров домена, хранят учетные записи и группы домена в Active Directory — базе данных, где находятся общие для домена настройки и информация. (Active Directory не описывается в этой книге.) По умолчанию дескриптор защиты раздела SAM настроен так, что даже учетная запись администратора не имеет к нему доступа.

В HKLM\SECURITY хранятся общесистемные политики безопасности и назначения прав пользователей. HKLM\SAM ссылается на подраздел в рамках SECURITY по адресу HKLM\SECURITY\SAM. По умолчанию вы не сможете просматривать содержимое HKLM\SECURITY или HKLM\SAM, поскольку настройки безопасности этих разделов разрешают доступ только для системной учетной записи. (Системные учетные записи обсуждаются более подробно далее в данной главе.) Но можно изменить дескриптор безопасности так, чтобы позволить администраторам доступ на чтение. Также можно использовать PsExec для запуска Regedit от лица локальной системной учетной записи, что тоже позволит заглянуть внутрь. Однако такой осмотр будет не очень информативным, поскольку данные не документированы, а пароли зашифрованы с помощью одностороннего отображения, то есть вы не можете определить пароль по его зашифрованной форме. Подразделы SAM и SECURITY поддерживаются файлами кустов SAM и SECURITY, расположенными по пути \Windows\system32\config загрузочного раздела.

HKLM\SOFTWARE — это место, где Windows хранит общесистемную информацию о конфигурации, не используемую при загрузке системы. Кроме того, сторонние приложения хранят здесь свои общесистемные настройки, такие как пути к файлам и каталогам, а также информацию о лицензировании и сроке действия.

HKLM\SYSTEM содержит общесистемную информацию о конфигурации, необходимую для загрузки системы, например, какие драйверы устройств загружать и какие службы запускать. Раздел поддерживается файлом куста SYSTEM, расположенным по адресу \Windows\system32\config. Загрузчик Windows использует службы реестра, предоставляемые загрузочной библиотекой, для чтения и навигации по кусту SYSTEM.

HKEY_CURRENT_CONFIG

HKEY_CURRENT_CONFIG — это просто ссылка на текущий профиль оборудования, хранящийся в разделе HKLM\SYSTEM\CurrentControlSet\Hardware Profiles\Current. Профили оборудования больше не поддерживаются в Windows, но раздел все еще существует для поддержки устаревших приложений, которые могут полагаться на его наличие.

HKEY_PERFORMANCE_DATA и HKEY_PERFORMANCE_TEXT

Реестр — это механизм, используемый для доступа к значениям счетчиков производительности в Windows независимо от того, получены они от компонентов операционной системы или от серверных приложений. Одним из побочных преимуществ предоставления доступа к счетчикам производительности через реестр является то, что удаленный мониторинг производительности работает «бесплатно», поскольку удаленный доступ к реестру легко получить через его обычные API.

Вы можете получить доступ к информации счетчика производительности реестра напрямую, открыв специальный раздел HKEY_PERFORMANCE_DATA и запросив параметры под ним. Вы не найдете этот раздел, заглянув в редактор реестра, так как получить к нему доступ можно только программным путем с помощью функций реестра Windows, таких как `RegQueryValueEx`. На самом деле данные о производительности в реестре не хранятся — функции реестра перенаправляют доступ по этому разделу к оперативной информации о производительности, получаемой от поставщиков данных о производительности.

HKEY_PERFORMANCE_TEXT — это еще один особый раздел, используемый для получения информации о счетчике производительности — обычно имени и описания. Имя любого счетчика производительности можно получить, запросив данные из специального параметра реестра `Counter`. Специальный параметр реестра `Help` вместо этого дает описание всех счетчиков. Информация, возвращаемая этим специальным разделом, формулируется на американском английском языке. В свою очередь, раздел HKEY_PERFORMANCE_NLSTEXT извлекает имена и описания счетчиков производительности на языке, который используется в ОС.

Вы также можете получить доступ к информации счетчика производительности с помощью функций Performance Data Helper (PDH), доступных через соответствующий API (Pdh.dll). На рис. 10.2 показаны компоненты, участвующие в получении информации от счетчиков производительности.

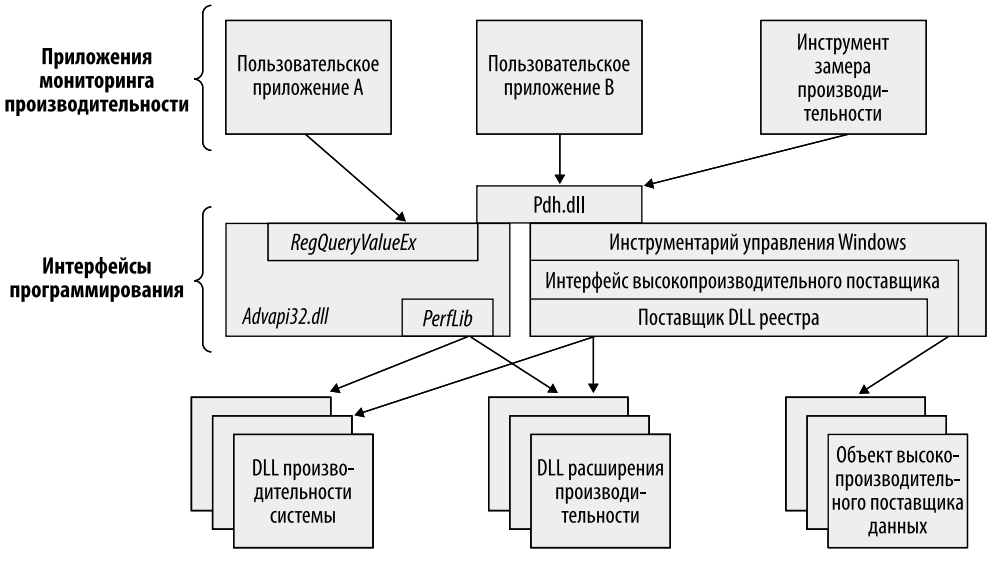


Рис. 10.2. Архитектура счетчиков производительности в реестре

Как показано на рис. 10.2, данный раздел реестра абстрагирован библиотекой производительности (Perflib), которая статически скомпонована с Advapi32.dll. Ядро Windows не знает о существовании раздела HKEY_PERFORMANCE_DATA, что объясняет невозможность найти его через редактор реестра.

Кусты приложений

Обычно приложения вправе читать данные из глобального реестра и записывать в него. Когда какое-то из них открывает раздел реестра, ядро Windows проверяет доступ по токену доступа его процесса или потока, если тот олицетворяющий (более подробную информацию см. в главе 7 тома 1), и ACL конкретного раздела. Приложение также может загружать и сохранять кусты реестра с помощью функций `RegSaveKeyEx` и `RegLoadKeyEx`. В этих сценариях приложение работает с данными, в которые могут вмешаться другие процессы, выполняющиеся с более высоким или тем же уровнем привилегий. Кроме того, для загрузки и сохранения кустов приложению необходимо подключить привилегии резервного копирования и восстановления. Последние предоставляются только процессам, запущенным от имени учетной записи администратора.

Очевидно, это стало ограничением для большинства приложений, которым требуется доступ к частному репозиторию для хранения собственных настроек. В Windows 7 была представлена концепция кустов приложений. Куст приложения — это стандартный файл куста (он связан с соответствующими файлами журнала), который можно смонтировать и сделать видимым только для запросившего его приложения. Разработчик может создать базовый файл куста с помощью функции `RegSaveKeyEx` (экспортирует содержимое обычного раздела реестра в файл куста). Затем приложение может смонтировать куст в частном порядке с помощью функции

RegLoadAppKey (передача флага REG_PROCESS_APPKEY предотвращает доступ других приложений к тому же кусту). Внутренне же эта функция выполняет следующие операции.

1. Создает случайный идентификатор GUID и присваивает его частному пространству имен в форме \Registry\A\<случайный Guid>. (\Registry образует пространство имен реестра ядра NT, описанное в разделе «Пространство имен и функционирование реестра» далее в этой главе.)
2. Преобразует путь DOS к указанному имени файла куста в формат NT и вызывает нативную функцию NtLoadKeyEx с соответствующим набором параметров.

Функция NtLoadKeyEx выполняет обычные обратные вызовы реестра. Но когда она обнаруживает, что куст относится к приложению, используется функция CmLoadAppKey для загрузки его (и связанных с ним файлов журнала) в частное пространство имен, которое не обнаружимо никаким другим приложением и привязано ко времени жизни вызывающего процесса. (Однако файлы куста и журнала по-прежнему отображаются в процессе реестра. Процесс реестра будет описан в разделе «Процесс реестра» далее в этой главе.) Приложение может применять стандартные API реестра для чтения и записи собственных настроек, которые будут храниться в кусте приложения. Куст будет автоматически выгружен при выходе из приложения или при закрытии последнего дескриптора раздела.

Кусты приложений используются различными компонентами Windows, такими как агент телеметрии совместимости приложений (CompatTelRunner.exe) и модель современных приложений (Modern Application Model). Приложения универсальной платформы Windows (UWP) задействуют кусты приложений для хранения информации о классах WinRT, которые могут быть созданы и являются частными для приложения. Этот куст хранится в файле ActivationStore.dat и применяется главным образом диспетчером активации при запуске приложения или, точнее, активации. Компонент фоновой инфраструктуры модели современных приложений использует данные в кусте для хранения информации о фоновых задачах. Таким образом, по истечении таймера фоновой задачи он точно знает, в какой библиотеке приложения находится код задачи, а также тип активации и модель потоков.

Кроме того, стек современных приложений предоставляет разработчикам UWP концепцию контейнеров данных приложения, которые можно использовать для хранения настроек, которые могут быть локальными для устройства, на котором выполняется приложение (в этом случае контейнер данных называется локальным), либо автоматически распространяться между всеми устройствами пользователя, на которых установлено приложение. Оба типа контейнеров реализованы в библиотеке WinRT Windows.Storage.ApplicationData.dll, использующей куст приложения, локальный для приложения (обеспечивается файлом settings.dat), для хранения параметров, созданных приложением UWP.

Файлы кустов settings.dat и ActivationStore.dat создаются в процессе развертывания модели современных приложений (во время установки приложения), подробно описанном в главе 8 (с общим обсуждением пакетных приложений). Документация о контейнерах данных приложения доступна по адресу https://docs.microsoft.com/en-us/windows/uwp/get-started/settings-learning-track.

Расширение для работы с реестром в режиме транзакций — Transactional Registry (TxR)

С помощью диспетчера транзакций ядра (Kernel Transaction Manager, KTM; подробности см. в разделе о KTM главы 8) разработчики получают доступ к простому API, который позволяет им реализовать надежные возможности восстановления после ошибок при выполнении операций с реестром, что может быть связано с операциями, не касающимися реестра, например с файлами или базами данных.

Транзакционное изменение реестра поддерживается тремя API: `RegCreateKeyTransacted`, `RegOpenKeyTransacted` и `RegDeleteKeyTransacted`. Эти новые процедуры принимают те же параметры, что и их нетранзакционные аналоги, за исключением того, что добавляется новый параметр дескриптора транзакции. Разработчик предоставляет этот дескриптор после вызова функции KTM `CreateTransaction`.

После транзакционной операции создания или открытия все последующие операции реестра, как то создание, удаление или изменение параметров внутри раздела, также будут выполняться через транзакции. Однако операции с подразделами транзакционного раздела не будут запускаться автоматически, на случай чего и предусмотрен третий API, `RegDeleteKeyTransacted`. Процедура позволяет выполнять транзакционное удаление подразделов, чего в рамках `RegDeleteKeyEx` обычно не делается.

Данные для этих транзакционных операций записываются в файлы журналов с использованием служб общей файловой системы журналирования (common logging file system, CLFS) аналогично другим операциям KTM. До тех пор пока транзакция не будет зафиксирована или отменена (и то и другое может произойти согласно программе или в результате сбоя питания или аварийного завершения системы в зависимости от состояния транзакции), изменения разделов, параметров и другие операции в реестре, выполняемые с помощью дескриптора транзакции, не будут видны внешним приложениям через нетранзакционные API. Транзакции также изолированы друг от друга: изменения, сделанные изнутри одной транзакции, не будут видны внутри других транзакций или за ее пределами до тех пор, пока транзакция не будет зафиксирована.

ПРИМЕЧАНИЕ Нетранзакционные средства записи прервут транзакцию в случае конфликта. Например, если параметр был создан внутри транзакции и позже, пока она еще активна, такой модуль попробует создать параметр под тем же разделом. Нетранзакционная операция завершится успешно, а все операции конфликтующей транзакции будут прерваны.

Уровень изоляции («I» в ACID) реализован диспетчерами ресурсов TxR по принципу read-commit, а это означает, что изменения становятся доступными для других читателей (с транзакциями или без них) сразу после фиксации. Этот механизм важен для людей, знакомых с транзакциями в базах данных, где уровень изоляции — предсказуемое чтение (или стабильность курсора, как это называют в литературе по базам данных). При уровне изоляции предсказуемого чтения повторный доступ к параметру внутри транзакции будет возвращать те же данные. Read-commit не дает такой гарантии. Одним из последствий оказывается то, что

транзакции реестра нельзя использовать для атомарных операций увеличения/уменьшения параметра реестра.

Чтобы внести в реестр постоянные изменения, приложение, задействующее дескриптор транзакции, должно вызвать функцию KTM `CommitTransaction`. (Если приложение решит отменить изменения, например во время обработки сбоя, оно может вызвать функцию `RollbackTransaction`.) Затем изменения все так же будут видны через обычные API реестра.

ПРИМЕЧАНИЕ Если дескриптор транзакции, созданный с помощью CreateTransaction, закрывается до фиксации транзакции и нет других дескрипторов, открытых для этой транзакции, система откатывает ее.

Помимо использования поддержки CLFS, предоставляемой KTM, TxR также хранит свои внутренние файлы журналов в папке `%SystemRoot%\System32\Config\Txr` на системном томе. Эти файлы имеют расширение `.regtrans-ms` и по умолчанию скрыты. За обслуживание всех кустов, смонтированных во время загрузки системы, отвечает диспетчер ресурсов глобального реестра (RM). Для каждого куста, который монтируется явно, создается диспетчер ресурсов. Для приложений, использующих транзакции реестра, создание RM прозрачно, поскольку KTM гарантирует, что все они, участвующие в одной и той же транзакции, скоординированы в двухфазном протоколе фиксации/прерывания. Для RM глобального реестра файлы журнала CLFS хранятся, как упоминалось ранее, внутри `System32\Config\Txr`. Для других кустов они хранятся рядом с ними, в том же каталоге. Эти файлы скрыты, соответствуют тому же соглашению об именах — имеют расширение `.regtrans-ms`. Имена файлов журнала начинаются с имени куста, которому они соответствуют.

Мониторинг активности реестра

Поскольку система и приложения в значительной степени зависят от параметров конфигурации, определяющих их поведение, сбои системы и приложений могут быть результатом изменения данных или настроек безопасности реестра. Когда системе или приложениям не удается прочитать настройки, к которым, как предполагается, они всегда смогут получить доступ, они могут работать неправильно, отображать сообщения об ошибках, скрывающие основную причину, или даже аварийно завершать работу. Практически невозможно узнать, какие разделы или параметры реестра настроены неправильно, не понимая, как система или приложение, в которых произошел сбой, получают доступ к реестру. В таких ситуациях ответ может дать утилита Process Monitor из набора Windows Sysinternals (https://docs.microsoft.com/en-us/sysinternals/).

Process Monitor позволяет отслеживать активность в реестре по мере ее возникновения. Для каждого обращения к реестру программа может показывать процесс, получивший доступ, время, тип и результат доступа, стек потока в момент доступа. Эта информация полезна для просмотра того, как приложения и система используют реестр, определения того, где они хранят параметры конфигурации, а также устранения неполадок, связанных с приложениями, у которых не хватает разделов или параметров реестра. Process Monitor включает в себя расширенную фильтрацию и выделение,

позволяющие более точно отследить операции, связанные как с конкретными разделами или параметрами, так и с активностью определенных процессов в целом.

Внутреннее устройство Process Monitor

Process Monitor полагается на драйвер устройства, который извлекает из своего исполняемого образа во время выполнения перед его запуском. Его первое выполнение требует, чтобы учетная запись, под которой он запущен, имела привилегию загрузки драйвера, а также привилегию отладки, последующие выполнения в том же сеансе загрузки требуют только привилегии отладки, поскольку после загрузки драйвер остается резидентным.

ЭКСПЕРИМЕНТ. Отслеживание активности реестра в условиях бездействия системы

Поскольку в реестре реализована функция `RegNotifyChangeKey`, которую приложения могут использовать для запроса уведомлений об изменениях реестра без постоянной проверки появления изменений, при запуске Process Monitor в простаивающей системе вы не должны видеть повторяющиеся обращения к одним и тем же разделам или параметрам реестра. Любая такая деятельность выявляет плохо написанное приложение, которое неоправданно снижает общую производительность системы.

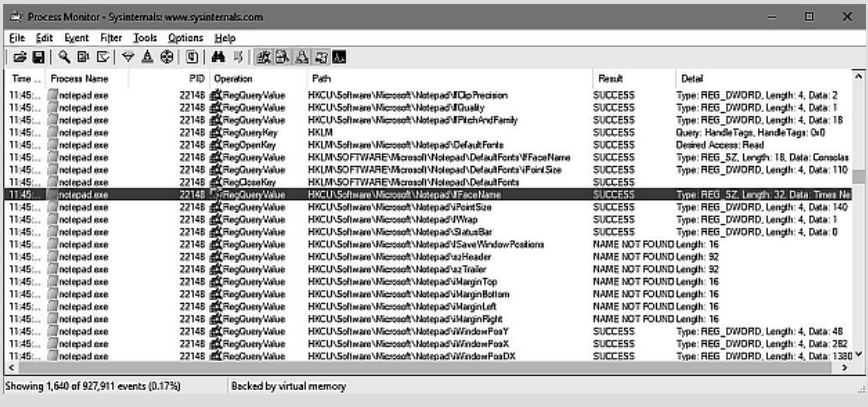
Запустите Process Monitor, убедитесь, что на панели инструментов включен только значок Показать активность реестра (Show Registry Activity) (с целью удаления шума, создаваемого файловой системой, сетью, процессами или потоками), и через несколько секунд просмотрите журнал вывода, чтобы увидеть, происходит ли опрос измененных данных. Щелкните правой кнопкой мыши на строке вывода, связанной с опросом, а затем выберите в контекстном меню Свойства процесса (Process Properties), чтобы просмотреть подробную информацию о процессе, выполняющем действие.

ЭКСПЕРИМЕНТ. Использование Process Monitor для поиска настроек реестра приложения

В некоторых сценариях устранения неполадок вам может потребоваться определить, где в реестре система или приложение хранит конкретные параметры. В этом эксперименте воспользуйтесь Process Monitor, чтобы определить расположение настроек Блокнота. Как и большинство приложений Windows, Блокнот хранит пользовательские настройки, такие как режим переноса слов, шрифт, размер шрифта, а также положение окна между запусками. Заставив Process Monitor отслеживать, когда Блокнот читает или записывает свои настройки, вы можете определить раздел реестра, в котором хранятся параметры. Последовательность шагов будет такой.

1. Укажите Блокноту сохранить настройку, которую вы сможете легко найти в трассировке Process Monitor. Для этого откройте Блокнот, установите шрифт Times New Roman, а затем закройте приложение.

2. Запустите Process Monitor. Откройте диалоговое окно фильтра и фильтр Имя процесса (Process Name) и введите `notepad.exe` в качестве фильтра. Подтвердите действие, нажав на кнопку Добавить (Add). Этот шаг указывает, что Process Monitor будет регистрировать только активность процесса notepad.exe.
3. Запустите Блокнот еще раз и затем остановите захват событий Process Monitor, переключив Захват событий (Capture Events) в меню Файл (File) Process Monitor.
4. Прокрутите до верхней строки результирующего журнала и выберите ее.
5. Нажмите Ctrl+F, чтобы открыть диалоговое окно Найти (Find), и выполните поиск строки `times new`. Process Monitor должен выделить строку, примерно как на следующем рисунке. Она представляет собой чтение Блокнотом значения шрифта из реестра. Другие операции в непосредственной близости должны относиться к другим настройкам Блокнота.



6. Щелкните правой кнопкой мыши на выделенной строке и выберите Перейти к (Jump To). Process Monitor запустит Regedit (если он еще не запущен) и заставит его перейти к параметру реестра, на который ссылался Блокнот, и выбрать его.

Внутреннее устройство реестра

В данном разделе описывается, как диспетчер конфигурации — исполнительная подсистема, реализующая реестр, организует файлы реестра на диске. Мы рассмотрим, как диспетчер управляет реестром, когда приложения и другие компоненты операционной системы читают и изменяют разделы и параметры реестра. Также обсудим механизмы, с помощью которых диспетчер пытается гарантировать, что реестр всегда находится в восстанавливаемом состоянии, даже если система выйдет из строя во время изменения реестра.

Кусты

На диске реестр представляет собой не просто один большой файл, а скорее набор файлов, называемых кустами[1]. Каждый куст содержит дерево реестра, у которого

[1] Также их называют ульями. — *Примеч. ред.*

есть раздел, служащий корнем или начальной точкой дерева. Подразделы и их параметры находятся под корнем. Может показаться, что корневые разделы, отображаемые редактором реестра, соответствуют корневым разделам в кустах, но это не так. В табл. 10.5 перечислены кусты реестра и имена их файлов на диске. Имена путей ко всем кустам, за исключением профилей пользователей, закодированы в диспетчере конфигурации. Когда тот загружает кусты, включая системные профили, он отмечает путь каждого из них в параметрах подраздела HKLM\SYSTEM\CurrentControlSet\Control\Hivelist, а если куст выгружен — удаляет путь. Он формирует корневые разделы, связывая эти кусты вместе для создания знакомой вам структуры реестра, которую отображает редактор реестра.

Таблица 10.5. Файлы на диске, соответствующие путям в реестре

Путь к кусту реестра	Путь к файлу куста
HKEY_LOCAL_MACHINE\BCD00000000	\EFI\Microsoft\Boot
HKEY_LOCAL_MACHINE\COMPONENTS	%SystemRoot%\System32\Config\Components
HKEY_LOCAL_MACHINE\SYSTEM	%SystemRoot%\System32\Config\System
HKEY_LOCAL_MACHINE\SAM	%SystemRoot%\System32\Config\Sam
HKEY_LOCAL_MACHINE\SECURITY	%SystemRoot%\System32\Config\Security
HKEY_LOCAL_MACHINE\SOFTWARE	%SystemRoot%\System32\Config\Software
HKEY_LOCAL_MACHINE\HARDWARE	Непостоянный куст
\HKEY_LOCAL_MACHINE\WindowsAppLockerCache	%SystemRoot%\System32\AppLocker\AppCache.dat
HKEY_LOCAL_MACHINE\ELAM	%SystemRoot%\System32\Config\Elam
HKEY_USERS\<SID локальной учетной записи службы>	%SystemRoot%\ServiceProfiles\LocalService\Ntuser.dat
HKEY_USERS\<SID учетной записи сетевой службы>	%SystemRoot%\ServiceProfiles\NetworkService\NtUser.dat
HKEY_USERS\<SID имени пользователя>	\Users\<имя пользователя>\Ntuser.dat
HKEY_USERS\<SID имени пользователя>_Classes	\Users\<имя пользователя>\AppData\Local\Microsoft\Windows\Usrclass.dat
HKEY_USERS\.DEFAULT	%SystemRoot%\System32\Config\Default
Виртуализированный HKEY_LOCAL_MACHINE\SOFTWARE	Пути варьируются. Обычно это \ProgramData\Packages\<PackageFullName>\<UserSid>\SystemAppData\Helium\Cache\<RandomName>.dat для Centennial
Виртуализированный HKEY_CURRENT_USER	Пути варьируются. Обычно это \ProgramData\Packages\<PackageFullName>\<UserSid>\SystemAppData\Helium\User.dat для Centennial
Виртуализированный HKEY_LOCAL_MACHINE\SOFTWARE\Classes	Пути варьируются. Обычно это \ProgramData\Packages\<PackageFullName>\<UserSid>\SystemAppData\Helium\UserClasses.dat для Centennial

Вы заметите, что некоторые из кустов, перечисленных в табл. 10.5, не являются постоянными и не имеют связанных файлов. Система создает их и управляет ими полностью в памяти, сами же они носят временный характер. Непостоянные кусты формируются заново при каждой загрузке. Примером такого куста является HKLM\HARDWARE, в котором хранится информация о физических устройствах и назначенных им ресурсах. Назначение ресурсов и обнаружение оборудования происходят каждый раз при загрузке системы, поэтому логично не хранить эти данные на диске. Обратите внимание также на последние три записи в таблице, которые представляют собой виртуализированные кусты. Начиная с Windows 10 Anniversary Update, ядро NT поддерживает виртуальный реестр (VReg), позволяя использовать пакетные приложения Centennial, которые работают в контейнере Helium. Каждый раз, когда пользователь запускает приложение подобного типа (например, современный Skype), система монтирует необходимые кусты пакетов. Приложения Centennial и Modern Application Model подробно обсуждались в главе 8.

ЭКСПЕРИМЕНТ. Ручная загрузка и выгрузка кустов

Regedit способен загружать кусты, доступ к чему можно получить через меню Файл (File). Эта возможность может оказаться полезной при устранении неполадок, когда нужно просмотреть или отредактировать куст в системе, которая не грузится, или с носителя резервной копии. В этом эксперименте используйте Regedit для загрузки версии куста HKLM\SYSTEM, которую программа установки Windows создает в процессе инсталляции.

1. Кусты можно загружать только под HKLM или HKU, поэтому откройте Regedit, выберите HKLM, а затем Загрузить куст (Load Hive) в меню Файл (File) Regedit.
2. Перейдите в каталог %SystemRoot%\System32\Config\RegBack в диалоговом окне Загрузка куста (Load Hive), выберите Система (System) и откройте его. В некоторых новых системах в папке RegBack может отсутствовать какой-либо файл. В этом случае можете попробовать тот же эксперимент, открыв куст ELAM, расположенный в папке Config. При появлении запроса введите `Test` в качестве имени раздела, под которым он будет загружаться.
3. Откройте вновь созданный раздел HKLM\Test и изучите содержимое куста.
4. Откройте HKLM\SYSTEM\CurrentControlSet\Control\Hivelist и найдите запись \Registry\Machine\Test, которая продемонстрирует, как диспетчер конфигурации перечисляет загруженные кусты в разделе Hivelist.
5. Выберите HKLM\Test, а затем Выгрузить куст (Unload Hive) в меню Файл (File) Regedit, чтобы выгрузить куст.

Ограничения на размер куста

В некоторых случаях размеры кустов ограниченны. Например, в Windows существует ограничение на размер куста HKLM\SYSTEM. Это происходит потому, что Winload считывает весь это куст в физическую память перед началом процесса загрузки, когда подкачка виртуальной памяти не включена. Также в физическую память загружаются ядро Ntoskrnl и драйверы загрузочных устройств. Ввиду этого объем физической памяти, назначенной HKLM\SYSTEM, следует ограничивать. (Подробности о роли Winload в процессе запуска см. в главе 12.) В 32-разрядных системах Winload позволяет кусту иметь размер до 400 Мбайт или половину объема физической памяти в системе в зависимости от того, что меньше. В системах x64 нижняя граница составляет 2 Гбайт.

Процесс реестра

До Windows 8.1 ядро NT использовало подкачиваемый (выгружаемый) пул памяти для хранения содержимого каждого загруженного файла куста. Большинство кустов, загруженных в систему, оставались в памяти до завершения ее работы (хорошим примером является куст SOFTWARE, который загружается диспетчером сеансов после завершения фазы 1 запуска системы и иногда может иметь размер в несколько сотен мегабайт). Память оттуда может быть выгружена диспетчером балансировки диспетчера памяти, если в течение определенного времени к ней не осуществлялся доступ (подробности см. в главе 5 «Управление памятью» тома 1). Это означает, что неиспользуемые части куста не остаются в рабочем наборе надолго. Выделенная виртуальная память поддерживается файлом подкачки и требует увеличения резервирования системной виртуальной памяти. Это уменьшает общий объем виртуальной памяти, доступной для других целей.

Чтобы решить эту проблему, в обновлении Windows 10 April 2018 Update (RS4) появилась поддержка обслуживания реестра в разрезе разделов. На этапе 1 инициализации ядра NT процедура запуска диспетчера конфигурации инициализирует несколько компонентов реестра: кэш, рабочие потоки, транзакции, поддержку обратных вызовов и т. д. Затем она создает тип объекта раздела и перед загрузкой необходимых кустов создает процесс реестра. Процесс реестра — это полностью защищенный (такой же уровень защиты, как и у процесса SYSTEM, — уровень WinSystem) минимальный процесс, который диспетчер конфигурации использует для выполнения большинства операций ввода-вывода в открытых кустах реестра. Во время инициализации в процессе реестра он размещает предварительно загруженные кусты. Однако те (кусты SYSTEM и ELAM) остаются в невыгружаемой памяти, которая размещается по адресации уровня ядра. Позже в процессе загрузки диспетчер сеансов загружает куст SOFTWARE, делая системный вызов `NtInitializeRegistry`.

Создается объект раздела, поддерживаемый файлом куста SOFTWARE: диспетчер конфигурации делит файл на фрагменты по 2 Мбайт и создает зарезервированное сопоставление в адресном пространстве пользовательского режима процесса реестра для каждого из них с помощью собственного API `NtMapViewOfSection`. Зарезервированные сопоставления отслеживаются действительными VAD, но фактически страницы не выделяются (подробности см. в главе 5 тома 1). Каждое представление

размером 2 Мбайт доступно только для чтения. Когда диспетчер конфигурации хочет прочитать какие-либо данные из куста, он обращается к страницам представления и выдает ошибку доступа, в результате чего разделяемые страницы переносятся в память диспетчером памяти. В это время увеличивается фактическое использование виртуальной памяти, но не ее общее резервирование (страницы поддерживаются самим файлом куста, а не файлом подкачки).

Во время инициализации диспетчер конфигурации устанавливает для процесса реестра жесткое ограничение на уровне 64 Мбайт. Это означает, что в сценариях с большой нехваткой памяти гарантировано, что реестр будет использовать не более 64 Мбайт рабочего набора. Каждый раз, когда приложение или система задействует API-интерфейсы для доступа к реестру, диспетчер конфигурации подключается к адресному пространству процесса реестра, выполняет необходимую работу и возвращает результаты. Диспетчеру конфигурации не всегда требуется переключать адресные пространства: когда приложению требуется доступ к разделу реестра, который уже находится в кэше (блок управления разделами уже существует), диспетчер конфигурации пропускает присоединение процесса и возвращает кэшированные данные. Процесс реестра в основном используется для выполнения ввода-вывода в файле куста низкого уровня.

Когда система записывает или изменяет разделы и параметры реестра, хранящиеся в кусте, она выполняет операцию копирования — записи, сначала меняя защиту памяти 2 Мбайт представления на `PAGE_WRITECOPY`. Запись в память, помеченная как `PAGE_WRITECOPY`, создает новые частные страницы и увеличивает размер выделенной виртуальной памяти в системе. Когда запрашивается обновление реестра, система немедленно добавляет новые записи в журнал куста, но запись фактических страниц, принадлежащих основному файлу, откладывается. «Грязные» страницы куста, как и любую обычную страницу памяти, можно выгрузить на диск. Эти страницы записываются в основной файл куста при выгрузке или с помощью Reconciler — одного из потоков отложенной записи диспетчера конфигурации, который запускается по умолчанию один раз в час (период времени настраивается установкой параметра реестра HKLM\SYSTEM\CurrentControlSet\Control\Session Manager\Configuration Manager\RegistryLazyReconcileInterval).

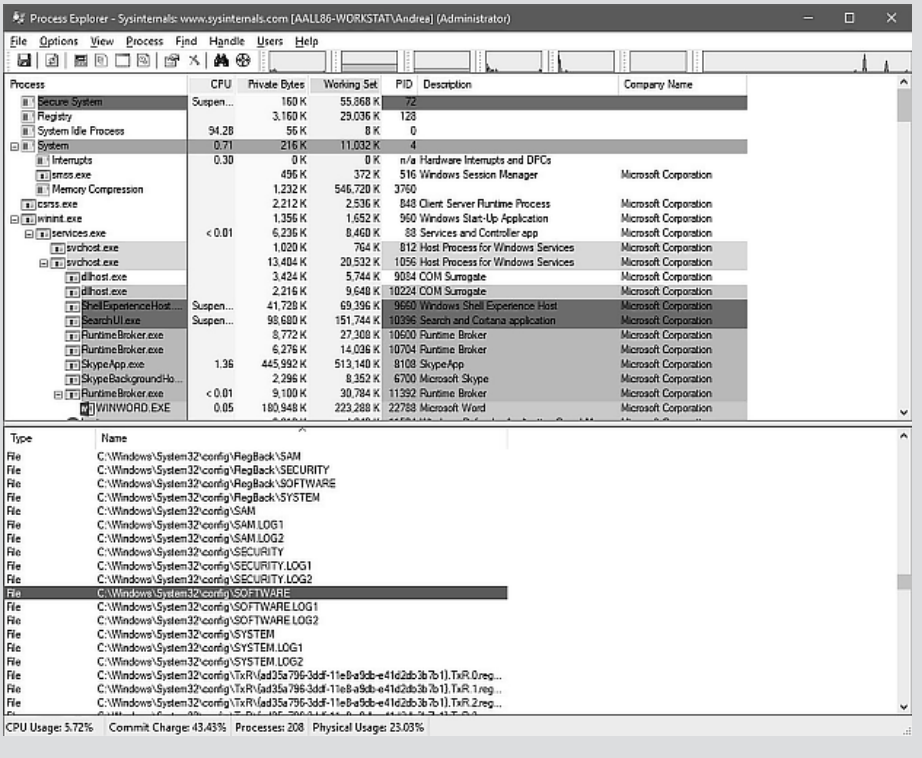
Поток Reconciler и инкрементальное ведение журнала обсуждаются в разделе «Инкрементальное ведение журнала» далее в этой главе.

Символические ссылки реестра

Специальный тип раздела, известный как символическая ссылка реестра, позволяет диспетчеру конфигурации связывать разделы для организации реестра. Символическая ссылка — это раздел, который перенаправляет диспетчера на другой раздел. Например, раздел HKLM\SAM представляет собой символическую ссылку на раздел в корне куста SAM. Символические ссылки создаются указанием параметра `REG_CREATE_LINK` при вызове `RegCreateKey` или `RegCreateKeyEx`. При обработке запроса диспетчер конфигурации создаст параметр `REG_LINK` с именем `SymbolicLinkValue`, который содержит путь к целевому разделу. Поскольку этот параметр представляет собой тип `REG_LINK`, а не `REG_SZ`, он не будет отображаться в Regedit, однако является частью куста реестра на диске.

ЭКСПЕРИМЕНТ. Обзор дескрипторов кустов

Диспетчер конфигурации открывает кусты, применяя таблицу дескрипторов ядра, описанную в главе 8, чтобы иметь доступ к ним из контекста любого процесса. Участие данной таблицы — эффективная альтернатива подходам, предполагающим использование драйверов или исполнительных компонентов для доступа из системного процесса только к дескрипторам, которые должны быть защищены от пользовательских процессов. Вы можете запустить Process Explorer от имени администратора, чтобы увидеть дескрипторы кустов, которые будут отображаться как открытые в системном процессе. Выберите процесс Система (System), затем щелкните на пункте Вид нижней панели (Lower Pane View) меню Вид (View) и выберите Дескрипторы (Handles). Отсортируйте список по типу дескриптора и прокручивайте, пока не увидите файлы куста, как показано на следующем рисунке.



Структура куста

Диспетчер конфигурации логически разделяет куст на области, называемые блоками, примерно так же, как файловая система делит диск на кластеры. По определению размер блока реестра составляет 4096 байт (4 Кбайт). Когда новые данные расширяют куст, он всегда расширяется кратно размеру блока. Первый блок куста называется базовым.

Базовый блок включает в себя глобальную информацию о кусте, в том числе подпись (regf), которая идентифицирует файл как куст, два обновленных порядковых номера, отметку времени, показывающую, когда в последний раз была инициирована операция записи в кусте, информацию об исправлении ошибок или восстановлении реестра, выполнявшемся Winload, номер версии формата куста, контрольную сумму и внутреннее имя файла куста, например \Device\HarddiskVolume1\WINDOWS\SYSTEM32\CONFIG\SAM. Мы поясним значение двух обновленных порядковых номеров и отметки времени, когда будем говорить о том, как данные записываются в файл куста.

Номер версии формата куста определяет формат данных внутри. Диспетчер конфигурации использует формат куста версии 1.5, который поддерживает большие параметры (свыше 1 Мбайт) и улучшенный поиск (вместо кэширования первых четырех символов имени во избежание коллизий используется хеш всего имени). Кроме того, диспетчер конфигурации поддерживает разностные кусты, добавленные для поддержки контейнеров. Для них применяется формат версии 1.6.

Windows организует данные реестра, которые хранит куст, в контейнерах, называемых *ячейками*. Ячейка может содержать раздел, параметр, дескриптор безопасности, список подразделов или список параметров разделов. Четырехбайтовый символьный тег в начале данных ячейки подписывает ее тип данных. В табл. 10.6 подробно описывается каждый тип данных ячейки. Заголовок ячейки — это поле, которое определяет размер ячейки как дополнение до 1 (отсутствует в структурах `CM_`). Когда ячейка присоединяется к кусту и он должен расшириться, чтобы вместить ячейку, система создает единицу распределения, называемую *пакетом*.

Пакет — это размер новой ячейки, округленный до границы следующего блока или страницы в зависимости от того, что больше. Система считает любое пространство между концом ячейки и концом пакета свободным пространством, которое она может выделить другим ячейкам. У ячеек также есть заголовки, которые содержат подпись hbin и поле, в котором записаны смещение в файле куста пакета и его размер.

Используя для отслеживания активных частей реестра пакеты вместо ячеек, Windows сводит к минимуму ряд задач по управлению ими. Например, система обычно выделяет и освобождает пакеты ячеек реже, чем сами ячейки, позволяя диспетчеру конфигурации более эффективно работать с памятью. Когда диспетчер конфигурации считывает куст реестра в память, он читает его весь, включая пустые пакеты, но может отказаться от них позже. Когда система добавляет ячейки в куст и удаляет их из него, пустые пакеты в нем могут перемежаться с активными. Эта ситуация аналогична фрагментации диска, которая возникает, когда система создает файлы на диске и удаляет их. Когда пакет пустеет, диспетчер конфигурации объединяет с ним все соседние пустые пакеты, чтобы сформировать как можно больший непрерывный пустой пакет. Диспетчер конфигурации также объединяет соседние удаленные ячейки, образуя свободные ячейки большего размера. (Диспетчер конфигурации сжимает куст только тогда, когда ячейки в конце него освобождаются. Вы можете сжать реестр, создав его резервную копию и восстановив его с помощью функций Windows `RegSaveKey` и `RegReplaceKey`, которые использует утилита Windows Backup. Кроме того, система уплотняет контейнеры во время инициализации куста с помощью алгоритма реорганизации, о котором поговорим позже.)

Таблица 10.6. Типы данных ячеек

Тип данных	Тип структуры	Описание
Ячейка раздела	CM_KEY_NODE	Ячейка, содержащая раздел реестра, также называемая узлом раздела. Ячейка раздела содержит подпись (kn для раздела, kl для узла связи), отметку времени последнего обновления раздела, индекс ячейки родительской ячейки раздела, индекс ячейки списка подразделов, который идентифицирует подразделы раздела, индекс ячейки для ячейки дескриптора безопасности раздела, индекс ячейки для строкового раздела, который определяет имя класса раздела, и имя раздела (например, CurrentControlSet). Он также сохраняет кэшированную информацию, такую как количество подразделов под разделом, и размер самого большого раздела, имя параметра, данные параметра и имя класса подразделов этого раздела
Ячейка параметра	CM_KEY_VALUE	Ячейка, содержащая информацию о параметре раздела. Она включает подпись (kv), тип параметра (например, REG_DWORD или REG_BINARY) и имя параметра (например, Boot-Execute). Ячейка параметра также содержит индекс ячейки, содержащей данные параметра
Ячейка большого параметра	CM_BIG_DATA	Ячейка, представляющая параметр реестра размером более 16 Кбайт. Содержимое ячейки этого типа представляет собой массив индексов ячеек, каждый из которых указывает на ячейку размером 16 Кбайт, которая содержит фрагмент параметра реестра
Ячейка списка подразделов	CM_KEY_INDEX	Ячейка, состоящая из списка индексов ячеек разделов, которые являются подразделами общего родительского раздела
Ячейка списка параметров	CM_KEY_INDEX	Ячейка, состоящая из списка индексов ячеек параметров, которые являются параметрами общего родительского раздела
Ячейка дескриптора безопасности	CM_KEY_SECURITY	Ячейка, содержащая дескриптор безопасности. Ячейки дескриптора безопасности включают подпись (ks) в заголовке ячейки и счетчик ссылок, который записывает количество узлов разделов, совместно использующих дескриптор безопасности. Несколько ячеек разделов могут совместно использовать ячейки дескриптора безопасности

Ссылки, создающие структуру куста, называются индексами ячеек. Индекс ячейки — это смещение ячейки в файле куста за вычетом размера базового блока. Таким образом, индекс ячейки похож на указатель из одной ячейки на другую, который диспетчер конфигурации интерпретирует относительно начала куста. Например, как вы видели в табл. 10.6, ячейка, описывающая раздел, содержит поле,

определяющее индекс ячейки ее родительского раздела, а индекс ячейки для подраздела указывает ячейку, описывающую подразделы, подчиненные указанному подразделу. Ячейка списка подразделов содержит список индексов ячеек, которые ссылаются на ячейки разделов подраздела. Следовательно, если хотите найти, например, ячейку раздела подраздела A, родителем которой является раздел B, вы должны сначала найти ячейку, содержащую список подразделов раздела B, с помощью индекса ячейки списка подразделов в ячейке раздела B. Затем находите каждую из ячеек подраздела раздела B, используя список индексов ячеек в ячейке списка подразделов. Для каждой ячейки подраздела проверяете, соответствует ли имя подраздела, хранящееся в ячейке раздела, тому, которое вы хотите найти, — в данном случае подразделу A.

Различия между ячейками, пакетами и блоками могут сбивать с толку, поэтому давайте рассмотрим пример простой структуры куста реестра, чтобы прояснить различия. На рис. 10.3 мы видим базовый блок и два пакета. Первый пакет пуст, а во втором имеется несколько ячеек. Логически куст имеет только два раздела: корневой раздел Root и подраздел корня Sub Key. У Root два значения: Val 1 и Val 2. Ячейка списка подразделов определяет подраздел корневого раздела, а ячейка списка значений определяет значения корневого раздела. Свободные места во втором пакете представляют собой пустые ячейки. На рис. 10.3 не показаны ячейки безопасности для двух разделов, которые должны присутствовать в кусте.

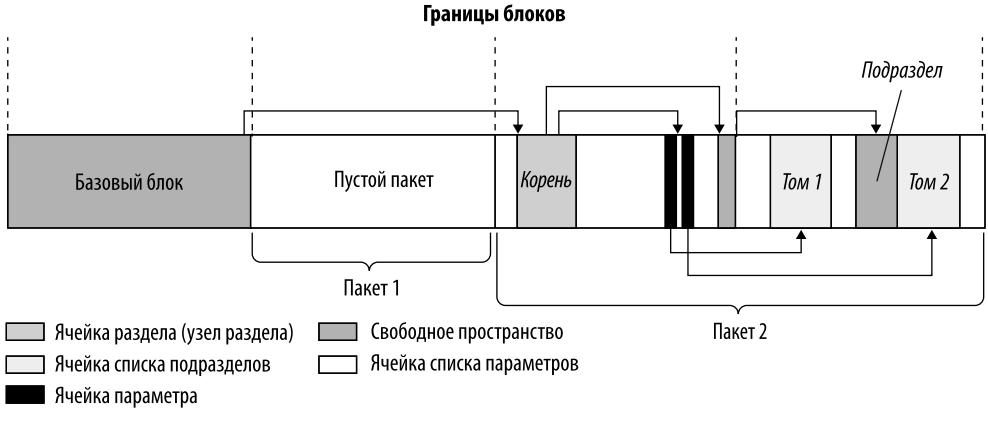


Рис. 10.3. Внутренняя структура куста реестра

Чтобы оптимизировать поиск параметров и подразделов, диспетчер конфигурации сортирует ячейки списка подразделов в алфавитном порядке. Затем диспетчер конфигурации может выполнить двоичный поиск, когда ищет подраздел в этом списке. Диспетчер конфигурации проверяет подраздел в середине списка, и если имя подраздела, который ищет диспетчер конфигурации, предшествует имени среднего подраздела, диспетчер конфигурации знает, что этот подраздел находится в первой половине списка подразделов. В противном случае подраздел находится во второй половине списка подразделов. Процесс разделения продолжается до

тех пор, пока диспетчер конфигурации не найдет подраздел или не обнаружит отсутствие совпадений. Однако ячейки списка параметров не сортируются, поэтому новые параметры всегда добавляются в конец списка.

Карты ячеек

Если бы кусты никогда не разрастались, диспетчер конфигурации мог бы выполнять все операции по управлению реестром в версии куста, находящейся в памяти, как если бы куст был файлом. Зная индекс ячейки, диспетчер конфигурации может вычислить ее местоположение в памяти, просто добавив индекс ячейки, который представляет собой смещение файла куста, к базовому образу куста в памяти. В начале загрузки системы этот процесс аналогичен тому, что Winload делает с кустом SYSTEM: Winload считывает его целиком в память как куст, доступный только для чтения, и добавляет индексы ячеек к базовому адресу образа куста в памяти для поиска ячеек. К сожалению, кусты разрастаются по мере приема новых разделов и параметров, а это означает, что система должна выделять новые зарезервированные представления и расширять файл куста для хранения новых пакетов ячеек, содержащих добавленные разделы и параметры. Зарезервированные представления, в которых хранятся данные реестра в памяти, не обязательно являются смежными.

Чтобы работать с несмежными адресами памяти, ссылающимися на данные куста в памяти, диспетчер конфигурации применяет стратегию, аналогичную той, которую использует диспетчер памяти Windows для сопоставления адресов виртуальной памяти с адресами физической памяти. Хотя индекс ячейки — это всего лишь смещение в файле куста, задействуется двухуровневая схема (рис. 10.4), когда куст представляется с помощью сопоставленных отображений в процессе реестра. Схема принимает в качестве входных данных индекс ячейки (то есть смещение в файле куста) и возвращает в качестве выходных данных адреса в памяти блока, в которых находятся как индекс ячейки, так и ячейка. Напомним: пакет может содержать один или несколько блоков, а кусты прирастают пакетами, поэтому Windows всегда представляет их непрерывными областями памяти. Таким образом, все блоки в интервале встречаются в одном и том же отображенном представлении куста размером 2 Мбайт.

Чтобы реализовать сопоставление, диспетчер конфигурации логически разделяет индекс ячейки на поля точно так же, как диспетчер памяти делит виртуальный адрес. Windows интерпретирует первое поле индекса ячейки как индекс каталога карты ячеек куста. Каталог карты ячеек содержит 1024 записи, каждая из которых ссылается на таблицу карты ячеек, хранящую 512 записей карты. Запись в таблице карты ячеек определяется вторым полем индекса ячейки. Эта запись определяет адреса ячеек и блоков памяти ячейки.

На последнем этапе трансляции диспетчер конфигурации интерпретирует последнее поле индекса ячейки как смещение в идентифицированном блоке, чтобы точно определить местоположение ячейки в памяти. При инициализации куста диспетчер конфигурации динамически создает таблицы сопоставления, назначая запись сопоставления каждому блоку в кусте, а также добавляет таблицы в каталог ячеек и удаляет их оттуда в зависимости от изменения размера куста.

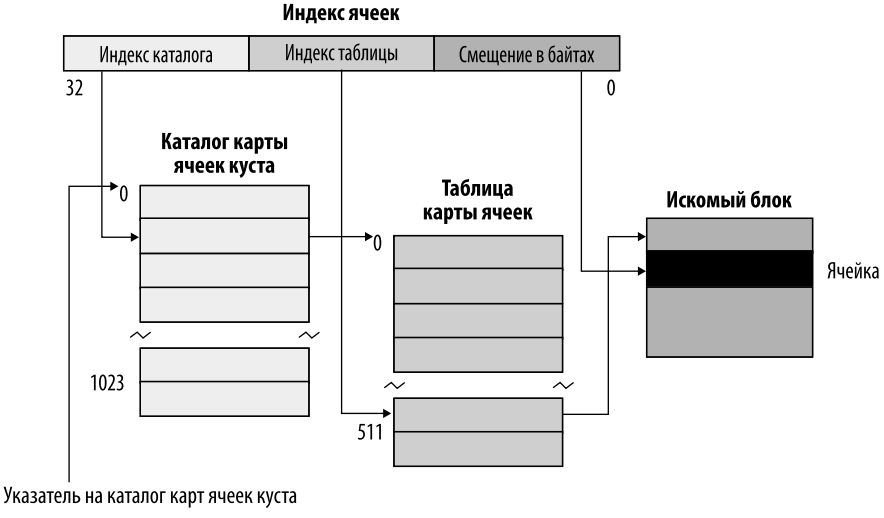


Рис. 10.4. Структура индекса ячеек

Реорганизация кустов

Как и реальные файловые системы, кусты реестра страдают от проблем фрагментации, когда ячейки в пакете освобождаются и невозможно объединить смежные. Так в отдельно взятых пакетах возникают небольшие фрагментированные участки свободного пространства. Если для новых ячеек недостаточно доступного непрерывного пространства, новые пакеты добавляются в конец файла куста, тогда как фрагментированные редко используются повторно. Чтобы решить эту проблему, начиная с Windows 8.1, каждый раз, когда диспетчер конфигурации монтирует файл куста, он проверяет, необходимо ли реорганизовать его. Время последней реорганизации записано в базовом блоке куста. Если куст имеет действительные файлы журналов, не является непостоянным и после предыдущей реорганизации прошло свыше семи дней, операция запускается. Реорганизация преследует две основные цели: сжать файл куста и оптимизировать его. Она начинается с создания нового пустого куста, идентичного исходному, но без ячеек. Созданный клон используется для копирования корневого раздела исходного куста со всеми его параметрами, но без подразделов. Сложный алгоритм анализирует все дочерние разделы, в свою очередь, во время своей обычной деятельности диспетчер конфигурации записывает, осуществляется ли доступ к определенному разделу, и, если да, сохраняет индекс, представляющий текущую фазу выполнения операционной системы (загрузочную или обычную) в его ячейке раздела.

Алгоритм реорганизации сначала копирует разделы, доступ к которым был получен во время нормального выполнения ОС, затем те, которые использовались на этапе загрузки, и, наконец, разделы, к которым не обращались с момента последней реорганизации. Таким образом разные разделы группируются в соседних ячейках файла куста, а операция копирования по определению создает нефрагментированный

файл куста (каждая ячейка сохраняется в пакете последовательно, а новая всегда добавляется в конец файла). Кроме того, у нового куста есть особенность — он содержит горячие и холодные классы разделов, хранящиеся в больших непрерывных фрагментах. Это значительно ускоряет этап загрузки и выполнения операционной системы при чтении данных из реестра.

Алгоритм реорганизации сбрасывает состояние доступа всех новых скопированных ячеек. Таким образом, система может отслеживать использование разделов в кусте, перезапустившись из нейтрального состояния. Новая статистика их применения будет задействована при следующей реорганизации, которая начнется через семь дней. Диспетчер конфигурации сохраняет результаты цикла реорганизации в ключе реестра HKLM\SYSTEM\CurrentControlSet\Control\Session Manager\Configuration Manager\Defrag (рис. 10.5). На снимке экрана видно, что последняя реорганизация проводилась 10 апреля 2019 года и сохранила 10 Мбайт фрагментированного пространства куста.

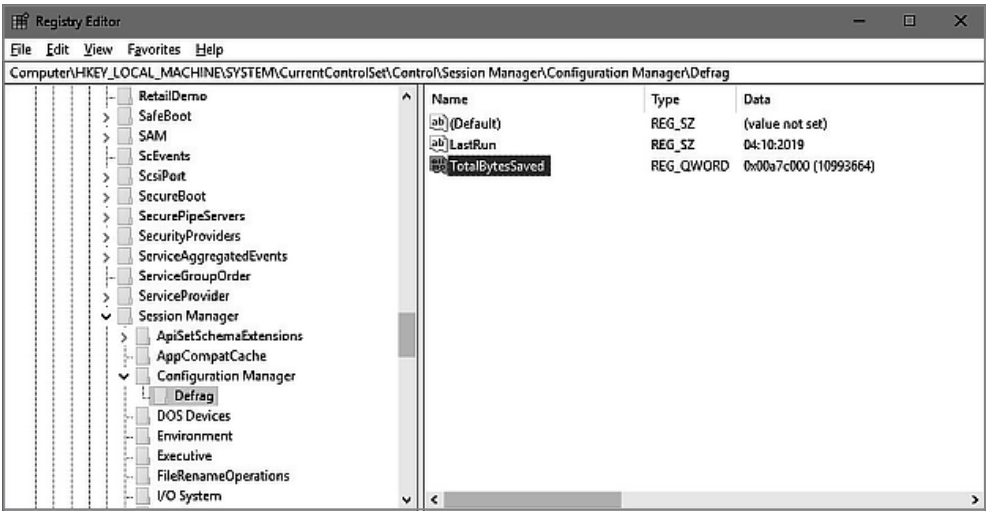


Рис. 10.5. Данные реорганизации реестра

Пространство имен и работа реестра

Диспетчер конфигурации определяет тип объекта раздела для интеграции пространства имен реестра с общим пространством имен ядра. Диспетчер конфигурации вставляет объект раздела с именем Registry в корень пространства имен Windows, который служит точкой входа в реестр. Regedit отображает имена разделов в форме HKEY_LOCAL_MACHINE\SYSTEM\CurrentControlSet, но подсистема Windows преобразует такие имена в форму пространства имен объектов, например \Registry\Machine\System\CurrentControlSet. Когда диспетчер объектов Windows анализирует это имя, он сначала находит объект раздела по имени Registry и передает остальную часть имени диспетчеру конфигурации. Тот берет на себя анализ имени, просматривая свое внутреннее дерево кустов в поисках нужного раздела или параметра. Прежде чем описать рабочий процесс управления типичным функционированием реестра,

нам необходимо обсудить объекты разделов и блоки управления разделами. Всякий раз, когда приложение открывает или создает раздел реестра, диспетчер объектов предоставляет приложению дескриптор, с помощью которого можно ссылаться на раздел. Дескриптор соответствует объекту раздела, который диспетчер конфигурации выделяет с помощью диспетчера объектов. Поддерживаемый диспетчером объектов диспетчер конфигурации использует преимущества безопасности и функций подсчета ссылок, которые предоставляет диспетчер объектов.

Для каждого открытого раздела реестра диспетчер конфигурации выделяет *блок управления разделами*. Такой блок хранит имя раздела, включает индекс ячейки узла раздела, на который ссылается блок управления, и содержит флаг, который отмечает, нужно ли диспетчеру удалить ячейку раздела, на которую ссылается блок управления разделом, когда последний дескриптор раздела закрывается. Windows помещает все блоки управления разделами в хеш-таблицу, чтобы обеспечить быстрый поиск существующих блоков по имени. Объект раздела указывает на соответствующий ему блок управления разделом, поэтому, если два приложения открывают один и тот же раздел реестра, каждое из них получает объект раздела и оба объекта указывают на общий управляющий блок.

Когда приложение открывает существующий раздел реестра, процесс начинается с указания приложением имени раздела в API реестра, что вызывает процедуру анализа имен диспетчера объектов. Диспетчер объектов, обнаружив в пространстве имен объект раздела реестра диспетчера конфигурации, передает тому имя пути. Диспетчер конфигурации выполняет поиск в хеш-таблице блока управления разделами. Если соответствующий блок управления разделами найден, дальнейшая работа не требуется (присоединять процесс реестра не нужно), в противном случае поиск предоставляет диспетчеру конфигурации ближайший к искомому разделу блок управления разделами, и поиск продолжается путем подключения к процессу реестра и использования структур данных куста в памяти для поиска по разделам и подразделам. Если диспетчер конфигурации находит ячейку раздела, он просматривает дерево блоков управления разделами, чтобы определить, открыт ли раздел (тем же приложением или другим). Процедура поиска оптимизирована так, чтобы всегда начинаться с ближайшего предка с уже открытым блоком управления. Например, если приложение открывает \Registry\Machine\Key1\Subkey2, а \Registry\Machine уже открыт, процедура анализа использует блок управления разделами \Registry\Machine в качестве отправной точки. Если раздел открыт, диспетчер конфигурации увеличивает счетчик ссылок существующего блока управления разделом. Если раздел не открыт, диспетчер конфигурации выделяет новый блок управления разделом и вставляет его в дерево. Затем диспетчер конфигурации выделяет объект раздела, нацеливает его на блок управления разделами, отделяется от процесса реестра и передает управление диспетчеру объектов, который возвращает приложению дескриптор.

Когда приложение создает новый раздел реестра, диспетчер конфигурации сначала находит ячейку раздела для родительского элемента нового раздела. Затем диспетчер конфигурации просматривает список свободных ячеек куста, в котором будет находиться новый раздел, чтобы определить, существуют ли ячейки, достаточно большие для хранения новой ячейки раздела. Если свободные ячейки недостаточно большого размера, диспетчер конфигурации выделяет новый пакет и использует его для ячейки, помещая ее в конец пакета в списке свободных ячеек. Новая ячейка раздела заполняется соответствующей информацией, включая имя

раздела, и диспетчер конфигурации добавляет ячейку раздела в список подразделов ячейки списка подразделов родительского раздела. Наконец, система сохраняет индекс родительской ячейки в ячейке раздела нового подраздела.

Диспетчер конфигурации использует счетчик ссылок блока управления разделами, чтобы определить, когда следует удалить этот блок. Когда все дескрипторы, которые ссылаются на раздел в блоке управления, закрываются, счетчик ссылок становится равным 0, что означает: блок управления клавишами больше не нужен. Если приложение, которое вызывает API для удаления раздела, устанавливает флаг удаления, диспетчер конфигурации может удалить связанный раздел из куста разделов, поскольку знает, что ни одно приложение не держит раздел открытым.

ЭКСПЕРИМЕНТ. Обзор контрольных блоков разделов

Вы можете использовать отладчик ядра для получения списка всех блоков управления разделами, выделенных в системе, с помощью команды `!reg openkeys`. В качестве альтернативы, если хотите просмотреть блок управления разделами для определенного открытого раздела, задействуйте `!reg querykey`:

```
0: kd> !reg querykey \Registry\machine\software\microsoft

Found KCB = ffffae08c156ae60 :: \REGISTRY\MACHINE\SOFTWARE\MICROSOFT

Hive         ffffae08c03b0000
KeyNode      00000225e8c3475c

[SubKeyAddr]         [SubKeyName]
225e8d23e64          .NETFramework
225e8d24074          AccountsControl
225e8d240d4          Active Setup
225ec530f54          ActiveSync
225e8d241d4          Ads
225e8d2422c          Advanced INF Setup
225e8d24294          ALG
225e8d242ec          AllUserInstallAgent
225e8d24354          AMSI
225e8d243f4          Analog
225e8d2448c          AppServiceProtocols
225ec661f4c          AppV
225e8d2451c          Assistance
225e8d2458c          AuthHost
...
```

Затем можете проверить сообщенный блок управления разделами с помощью команды `!reg kcb`:

```
kd> !reg kcb ffffae08c156ae60

Key              : \REGISTRY\MACHINE\SOFTWARE\MICROSOFT
RefCount         : 1f
Flags            : CompressedName, Stable
ExtFlags         :
Parent           : 0xe1997368
KeyHive          : 0xe1c8a768
```

```
KeyCell          : 0x64e598 [cell index]
TotalLevels      : 4
DelayedCloseIndex: 2048
MaxNameLen       : 0x3c
MaxValueNameLen  : 0x0
MaxValueDataLen  : 0x0
LastWriteTime    : 0x1c42501:0x7eb6d470
KeyBodyListHead  : 0xe1034d70 0xe1034d70
SubKeyCount      : 137
ValueCache.Count : 0
KCBLock          : 0xe1034d40
KeyLock          : 0xe1034d40
```

Поле `Flags` указывает, что имя хранится в сжатом виде, а поле `SubKeyCount` — что раздел имеет 137 подразделов.

Обеспечение надежного хранения

Чтобы гарантировать, что постоянный куст реестра (у которого файл на диске) всегда находится в восстанавливаемом состоянии, диспетчер конфигурации использует кусты журналов. С каждым постоянным кустом связан куст журнала, который представляет собой скрытый файл с тем же базовым именем, что и у куста, и расширением `logN`. Чтобы обеспечить прогресс, диспетчер конфигурации задействует схему двойного журналирования. Потенциально существуют два файла журнала: `.log1` и `.log2`. Если по какой-либо причине файл `.log1` был записан, но при записи измененных данных в основной файл журнала произошел сбой, при следующей очистке произойдет переключение на `.log2` с накопленными измененными данными. Если и это не удается, совокупные «грязные» данные (данные в `.log1` и данные, которые были тем временем изменены) сохраняются в `.log2`. Как следствие, в следующий раз снова будет использоваться `.log1`, пока не будет выполнена успешная операция записи в основной файл журнала. Если сбоя не происходит, применятся только `.log1`.

Например, если вы заглянете в каталог %SystemRoot%\System32\Config (и у вас выбрана опция Показать скрытые файлы и папки (Show Hidden Files And Folders), а флажок Скрыть защищенные системные файлы (Hide Protected Operating System File) снят), то увидите файлы `System.log1`, `Sam.log1` и другие файлы `.log1` и `.log2`. При инициализации куста диспетчер конфигурации выделяет битовый массив, в котором каждый бит представляет 512-байтовую часть (сектор) куста. Этот массив называется массивом «грязных» секторов, поскольку бит, установленный в массиве, означает, что система изменила соответствующий сектор в кусте в памяти и должна записать этот сектор обратно в файл куста. (Если бит не установлен, то сектор соответствует содержимому куста в памяти.)

Когда создаются новые раздел или параметр или изменяются существующие раздел или параметр, диспетчер конфигурации отмечает обновленные сектора основного куста и записывает их в массив «грязных» секторов куста в памяти. Затем он планирует операцию отложенной очистки или синхронизацию журналов. Системный поток отложенной записи куста просыпается через минуту после

запроса на синхронизацию. Он генерирует новые записи журнала из секторов куста в памяти, на которые ссылаются действительные биты массива «грязных» секторов, и записывает их в файлы журнала куста на диске. В то же время система записывает на диск все изменения реестра, произошедшие между моментом запроса синхронизации куста и моментом ее выполнения. Ленивая запись использует операции ввода-вывода с низким приоритетом и записывает «грязные» сектора в файл журнала на диске (а не в основной куст). Когда завершается синхронизация куста, следующая сможет начаться не раньше чем через 1 мин.

Если при ленивой записи все «грязные» сектора куста просто попадут в файл куста и система выйдет из строя в середине работы, тот окажется в противоречивом (поврежденном) и не поддающемся восстановлению состоянии. Чтобы предотвратить подобное, ленивая запись сначала сбрасывает массив «грязных» секторов куста и все «грязные» сектора в файл журнала куста, увеличивая при необходимости размер файла журнала. Базовый блок куста содержит два порядковых номера. После первой операции сброса (а не при последующих сбросах) диспетчер конфигурации обновляет один из порядковых номеров, который становится больше второго. Таким образом, если система выйдет из строя во время операций записи в куст, при следующей перезагрузке диспетчер конфигурации заметит, что два порядковых номера в базовом блоке куста не совпадают. Тогда он сможет обновить куст, используя измененные сектора в файле журнала куста, чтобы актуализировать его. Куст снова будет актуальным и последовательным.

После сохранения записей в журнал куста ленивая очистка очищает соответствующие допустимые биты в массиве «грязных» секторов, но вставляет эти биты в другой важный вектор — несогласованный массив. Последний используется диспетчером конфигурации, чтобы понять, какие записи журнала следует вносить в основной куст. Благодаря новой функции инкрементального ведения журнала (обсудим ее позже) основной файл куста редко переписывается во время выполнения операционной системы. Протокол синхронизации куста (не путать с синхронизацией журнала) — это алгоритм, применяемый для записи всех изменений реестра в памяти и журнале в основной файл куста и для установки там двух порядковых номеров. Это действительно дорогостоящая многоэтапная операция, о которой мы еще поговорим далее.

Reconciler (Согласователь), который представляет собой еще один тип системного потока отложенной записи, просыпается раз в час, замораживает журнал и записывает все ненужные записи журнала в основной файл куста. Алгоритм согласования знает, какие части куста в памяти следует записать в основной файл, благодаря как «грязным» секторам, так и несогласованному массиву. Однако бывает это редко. В случае сбоя системы у диспетчера конфигурации есть вся информация, необходимая для восстановления, благодаря записям журнала, которые уже внесены в файлы журнала. Выполнение согласования реестра только один раз в час (или когда размер журнала выходит за пороговое значение, которое зависит от размера тома, на котором находится куст) значительно повышает производительность. Единственный временной интервал, в течение которого может произойти некоторая потеря данных, — это период между очистками журнала.

Обратите внимание на то, что сверка по-прежнему не обновляет второй порядковый номер в основном файле куста. Два порядковых номера будут обновлены

равным значением только на этапе проверки (еще одна форма очистки куста), которая происходит только во время выгрузки куста (при этом приложение вызывает API `RegUnloadKey`), когда система выключается, или при первой загрузке. Это означает, что большую часть времени существования операционной системы главный куст реестра находится в «грязном» состоянии и его правильное прочтение потребует обращения к файлу журнала.

Загрузчик Windows также содержит код, связанный с надежностью реестра. Например, он может анализировать файл `System.log` перед загрузкой ядра и выполнять исправления для согласованности. Кроме того, в некоторых случаях повреждения куста (например, если базовый блок, пакет или ячейка содержит данные, которые не прошли проверку согласованности) диспетчер конфигурации может повторно инициализировать поврежденные структуры данных, возможно, в процессе этого удалив подразделы, и продолжить нормальную работу. Если ему необходимо прибегнуть к операции самовосстановления, появится диалоговое окно с системной ошибкой, уведомляющее пользователя.

Инкрементальное ведение журнала

Как упоминалось в предыдущем разделе, в Windows 8.1 значительно улучшена производительность алгоритма синхронизации кустов благодаря инкрементному ведению журнала. Обычно ячейки в файле куста могут быть в четырех разных состояниях.

- **Чистое.** Данные ячейки находятся в основном файле куста и не были изменены.
- **Грязное.** Данные ячейки были изменены, но находятся только в памяти.
- **Несогласованное.** Данные ячейки были изменены и правильно записаны в файл журнала, но еще не в основном файле.
- **Грязное и несогласованное.** После записи ячейки в файл журнала она снова изменяется. В файле журнала сохраняется только первая модификация, тогда как последняя хранится только в памяти.

Исходный алгоритм синхронизации до Windows 8.1 выполнялся через 5 с после изменения одной или нескольких ячеек. Алгоритм можно свести к четырем шагам.

1. Диспетчер конфигурации записывает все измененные ячейки, отмеченные «грязным» вектором, в одну запись в файле журнала.
2. Он делает недействительным базовый блок куста, устанавливая только один порядковый номер, бóльший, чем другой.
3. Он записывает все измененные данные в файл основного куста.
4. Он выполняет проверку основного куста (при проверке устанавливаются два одинаковых порядковых номера в основном файле куста).

Чтобы поддерживать целостность и возможность восстановления куста, алгоритм должен направлять запрос на операцию записи изменений на диск драйверу файловой системы после каждого этапа, в противном случае может произойти повреждение данных. Операции записи изменений данных произвольного доступа могут быть очень дорогими, особенно на обычных жестких дисках.

Инкрементное ведение журнала решило проблему производительности. В устаревшем алгоритме делалась одна запись журнала, содержащая все «грязные» данные между несколькими проверками куста. Инкрементальная модель отошла от этого принципа. Новый алгоритм синхронизации заносит одну запись в журнал каждый раз, когда выполняется ленивая очистка, что, как обсуждалось ранее, делает недействительным базовый блок основного куста только при первом запуске. Последующие записи изменений продолжают вносить новые записи в журнал, не затрагивая основного файла куста. Каждый час, или если место в журнале заканчивается, алгоритм согласования записывает все данные, хранящиеся в записях журнала, в файл основного куста, не выполняя этап проверки. Таким образом, место в файле журнала освобождается, при этом сохраняется возможность восстановления куста. Если на этом этапе система выйдет из строя, журнал будет содержать исходные записи, которые будут повторно использованы во время загрузки куста. В противном случае новые записи используются повторно в начале журнала, а при последующем сбое системы во время загрузки куста используются только новые записи из журнала.

На рис. 10.6 показаны возможные ситуации сбоя и то, как они управляются с помощью схемы инкрементного журналирования. В случае А система записала новые данные в куст в памяти, а ленивая запись изменений на диск сделала соответствующие записи в журнал (но сверки не произошло). При перезапуске системы процедура восстановления применяет все записи журнала к основному кусту и снова проверяет файл куста. В случае Б средство согласования уже записало данные, хранящиеся в записях журнала, в основной куст до сбоя (проверка куста не проводилась). При перезагрузке системы процедура восстановления повторно применяет существующие записи журнала, но никаких изменений в основной файл куста не вносится. Случай В аналогичен ситуации Б, но в журнале после сверки была сделана новая запись. В этом случае процедура восстановления записывает только последнюю модификацию, которой нет в основном файле.

Проверка куста выполняется только в определенных редких случаях. Когда тот выгружается, система выполняет сверку, а затем проверяет основной файл куста. В конце проверки она устанавливает два идентичных порядковых номера основного файла куста и выдает последний запрос на запись изменений файловой системы на диск перед выгрузкой куста из памяти. При перезапуске системы код загрузки куста обнаруживает, что основной куст в порядке (по двум одинаковым порядковым номерам), и не запускает каких-либо процедур восстановления. Благодаря новому протоколу инкрементальной синхронизации операционная система больше не страдает от снижения производительности, вызванного старым протоколом ведения журнала.

ПРИМЕЧАНИЕ Загрузка куста, созданного в Windows 8.1 или более новой операционной системе, на старых машинах проблематична, если основной файл куста находится в нечистом состоянии. Старая ОС (например, Windows 7) понятия не имеет, как обрабатывать новые файлы журналов. По этой причине Microsoft создала драйвер мини-фильтра RegHiveRecovery, который распространяется через комплект Windows Assessment and Deployment Kit (ADK). Драйвер RegHiveRecovery использует обратные вызовы реестра, которые перехватывают запросы загрузки куста от системы и определяют, нуждается ли в восстановлении основной файл куста, и задействует инкрементальные журналы. Если это так, он выполняет восстановление и исправляет основной файл куста до того, как система сможет его прочитать.

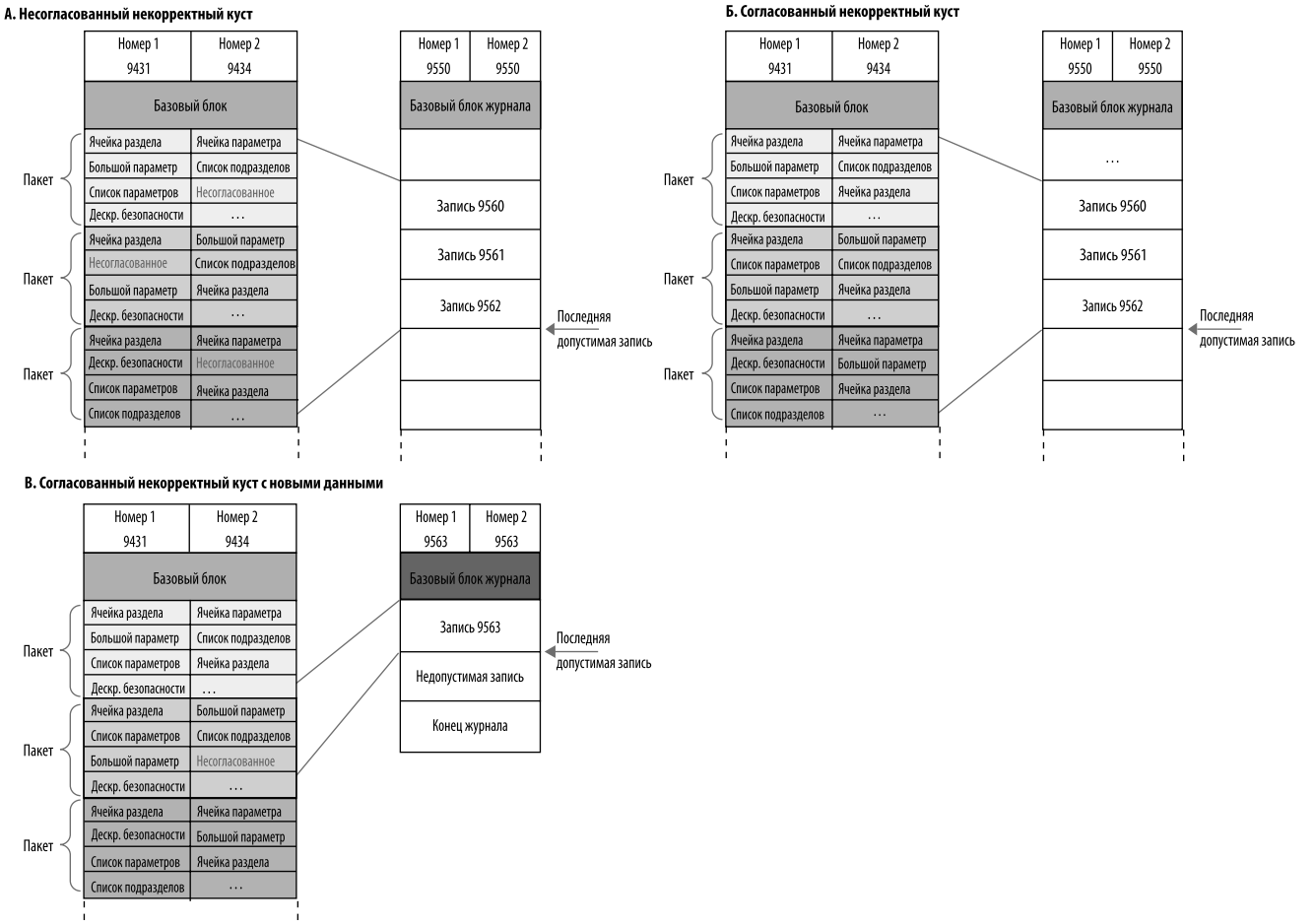


Рис. 10.6. Последствия возможных сбоев системы в различные моменты

Фильтрация реестра

Диспетчер конфигурации в ядре Windows реализует мощную модель фильтрации реестра, которая позволяет отслеживать активность реестра с помощью таких инструментов, как Process Monitor. Когда драйвер использует механизм обратного вызова, он регистрирует соответствующую функцию в диспетчере конфигурации. Тот выполняет функцию обратного вызова драйвера до и после выполнения системных служб реестра, чтобы драйвер имел полную видимость и контроль над доступом к реестру. Антивирусные продукты, которые просматривают данные реестра на наличие вирусов или предотвращают изменение реестра несанкционированными процессами, тоже являются пользователями данного механизма.

Обратные вызовы реестра также связаны с концепцией высот. Это способ, с помощью которого разные поставщики могут зарегистрировать высоту в стеке фильтрации реестра, чтобы порядок, в котором система вызывает каждую процедуру обратного вызова, мог быть детерминированным и правильным. Это позволяет избежать сценария, в котором антивирусный продукт будет просматривать зашифрованные разделы до того, как продукт шифрования запустит собственный обратный вызов для их расшифровки. В модели обратного вызова реестра Windows обоим типам инструментов назначается базовая высота, соответствующая типу выполняемой ими фильтрации — в данном случае шифрованию в противоположность просмотру. Кроме того, компании, создающие подобные инструменты, должны зарегистрироваться в Microsoft, чтобы внутри своей группы они не сталкивались с аналогичными или конкурирующими продуктами.

Модель фильтрации также включает в себя возможность либо полностью взять на себя обработку операции реестра (в обход диспетчера конфигурации и не позволяя ему обрабатывать запрос), либо перенаправить операцию на другую операцию (например, перенаправить реестр в WoW64). Кроме того, можно изменить выходные параметры и возвращаемое значение операции реестра.

Наконец, драйверы могут назначать и помечать информацию, определенную ими в разрезе разделов или операций, для собственных целей. Драйвер может создавать и назначать эти контекстные данные во время операции создания или открытия, которые диспетчер конфигурации запоминает и возвращает во время каждой последующей операции с разделом.

Виртуализация реестра

В обновлении Windows 10 Anniversary Update (RS1) была представлена виртуализация реестра для контейнеров Argon и Helium, а также возможность загружать разностные кусты, соответствующие новой версии куста 1.6. Виртуализация реестра обеспечивается как диспетчером конфигурации, так и драйвером VReg (интегрирован в ядро Windows). Эти два компонента предоставляют следующие возможности.

- **Перенаправление пространства имен.** Приложение может перенаправить содержимое виртуального раздела на реальный раздел на хосте. Приложение также может перенаправить виртуальный раздел на раздел, принадлежащий разностному кусту, который объединяется с корневым разделом на узле.
- **Слияние реестров.** Разностные кусты интерпретируются как набор отличий от базового куста. Базовый куст представляет базовый уровень, который содержит

представление неизменяемого реестра. Разделы в разностном кусте могут быть добавлением к основному или тем, что из него вычитается. Последние называются «мертвыми разделами».

Диспетчер конфигурации на этапе 1 инициализации ОС создает объект устройства `VRegDriver` (с соответствующим дескриптором безопасности, который разрешает доступ только системе и администратору) и тип объекта `VRegConfigurationContext`, который представляет контекст хранилища, используемый для отслеживания перенаправления пространства имен и куста, и слияние, которое принадлежит контейнеру. Серверные хранилища рассматривались в главе 3 тома 1.

Перенаправление пространства имен

Перенаправление пространства имен реестра можно включить только в изолированном хранилище как сервера, так и приложения. После создания хранилища, но перед его запуском приложение отправляет IOCTL инициализации объекту устройства VReg, передавая дескриптор хранилища. Драйвер VReg создает пустой контекст конфигурации и присоединяет его к объекту хранилища. Затем он создает один узел пространства имен, который переназначает корневой раздел `\Registry\WC` контейнера на раздел хоста, поскольку все контейнеры имеют одно и то же его представление. Корневой раздел `\Registry\WC` создается для монтирования всех кустов, виртуализированных для разных хранилищ.

Драйвер VReg — это драйвер фильтра реестра, который использует механизм обратных вызовов реестра для правильной реализации перенаправления пространства имен. Когда приложение впервые инициализирует перенаправление пространства имен, драйвер VReg регистрирует свою основную процедуру уведомления `RegistryCallback` через внутренний API, аналогичный `CmRegisterCallbackEx`. Чтобы правильно добавить перенаправление пространства имен к корневому разделу, приложение отправляет IOCTL Create Namespace Node на устройство VReg и указывает путь виртуального раздела (он будет виден контейнеру), реальный путь раздела хоста и дескриптор задания контейнера. В ответ драйвер VReg создает новый узел пространства имен (небольшую структуру данных, содержащую данные раздела и некоторые флаги) и добавляет его в контекст конфигурации хранилища.

После того как приложение завершило настройку всех перенаправлений реестра для контейнера, оно присоединяет собственный процесс (или новый порожденный процесс) к изолированному объекту, используя `AssignProcessToJobObject` (дополнительные сведения см. в главе 3 тома 1). С этого момента каждый ввод-вывод реестра, выполняемый контейнерным процессом, будет перехватываться мини-фильтром реестра VReg. Давайте проиллюстрируем на примере, как работает перенаправление пространства имен.

Предположим, что современная платформа приложений установила несколько перенаправлений пространства имен реестра для приложения Centennial. В частности, один из узлов перенаправления перенаправляет разделы из HKCU на раздел `\Registry\WC\a20834ea-8f46-c05f-46e2-a1b71f9f2f9cuser_sid` хоста. В определенный момент приложение Centennial хочет создать новый раздел с именем AppA в родительском разделе `HKCU\Software\Microsoft`. Когда процесс вызывает API `RegCreateKeyEx`, обратный вызов реестра VReg перехватывает запрос и получает контекст конфигурации задания. Затем он ищет в контексте ближайший узел пространства имен к пути раздела,

указанному вызывающей стороной. Если ничего не находит, то возвращает ошибку отсутствующего объекта: «Для контейнера не разрешена работа по невиртуализированным путям». Предполагая, что узел пространства имен, описывающий корневой раздел HKCU, существует в контексте и является родительским для подраздела HKCU\Software\Microsoft, драйвер VReg заменяет относительный путь исходного виртуального раздела именем родительского раздела хоста и пересылает его запрос к диспетчеру конфигурации. Итак, в этом случае диспетчер конфигурации действительно видит запрос на создание \Registry\WC\a20834ea-8f46-c05f-46e2-a1b71f9f2f9cuser_sid\Software\Microsoft\AppA и выполняет его. Контейнерное приложение не обнаруживает никакой разницы. Со стороны приложения раздел реестра находится в хосте HKCU.

Разностные (дифференциальные) кусты

Хотя перенаправление пространства имен реализовано в драйвере VReg и доступно только в контейнерных средах, объединение реестра также может работать глобально и реализуется в основном в самом диспетчере конфигурации. (Однако драйвер VReg по-прежнему используется в качестве точки входа, позволяя монтировать разностные кусты к базовым разделам.) Как указано в предыдущем разделе, разностные кусты применяют формат версии 1.6, который очень похож на версию 1.5, но поддерживает метаданные для разностных разделов. Увеличение версии куста также предотвращает возможность монтирования куста в системах, не поддерживающих виртуализацию реестра.

Приложение может создать разностный куст и смонтировать его глобально в системе или изолированном контейнере, отправив IOCTL на устройство VReg. Однако необходимы привилегии резервного копирования и восстановления, поэтому только административные приложения могут управлять разностными кустами. Чтобы смонтировать разностный куст, приложение заполняет структуру данных именем базового раздела (называемого базовым уровнем; базовый уровень — это корневой раздел, на который накладываются все подразделы и параметры, содержащиеся в разностном кусте), путем к разностному кусту и точкой монтирования. Затем он отправляет структуру данных драйверу VReg через управляющий код `VR_LOAD_DIFFERENCING_HIVE`. Точка монтирования содержит объединение данных, хранящихся в разностном кусте, и данных на базовом уровне.

Драйвер VReg поддерживает список всех загруженных разностных кустов в хеш-таблице. Это позволяет драйверу VReg монтировать разностный куст в нескольких точках монтирования. Как было сказано ранее, модель современных приложений использует случайные GUID в корневом разделе \Registry\WC с целью монтирования независимых разностных кустов приложений Centennial. После создания записи в хеш-таблице драйвер VReg просто пересылает запрос функции внутреннего диспетчера конфигурации `CmLoadDifferencingKey`. Последний выполняет бóльшую часть работы. Он инициирует обратные вызовы реестра и загружает разностный куст. Создание этого куста происходит так же, как и для обычного куста. После создания куста на нижнем уровне диспетчера конфигурации появляется также структура данных блока управления разделами. Новый блок управления разделами связан с блоком управления разделами базового уровня.

Когда запрос направлен на открытие или чтение параметров, расположенных в разделе, используемом в качестве точки монтирования, либо в его дочернем элементе, диспетчер конфигурации знает, что соответствующий блок управления

разделами представляет собой разностный куст. Поэтому процедура синтаксического анализа начинается с разностного куста. Если диспетчер конфигурации обнаруживает там подраздел, он останавливает процедуру анализа и возвращает разделы и данные, хранящиеся в разностном кусте. Если же данные в разностном кусте не найдены, диспетчер конфигурации перезапускает процедуру анализа с базового куста. Также проверяется, найден ли мертвый раздел в разностном кусте: диспетчер конфигурации скрывает искомый раздел и не возвращает никаких данных (или ошибки). Мертвые разделы действительно используются для пометки раздела как удаленного из базового куста.

Система поддерживает три типа разностных кустов.

- **Изменяемые кусты** можно записывать и обновлять. Все запросы на запись, направленные к точке монтирования или ее дочерним разделам, сохраняются в разностном кусте.
- **Неизменяемые кусты** не могут быть изменены. Это означает, что все изменения, запрошенные для раздела, расположенного в разностном кусте, завершатся ошибкой.
- **Изменяемые кусты сквозной записи** нельзя изменять, но запросы на запись в точке монтирования и ее дочерние разделы сразу идут на базовый уровень, который теперь не неизменяемый.

Ядро и приложения NT также могут монтировать разностный куст, а затем использовать перенаправление пространства имен поверх его точки монтирования, что позволяет реализовать сложные виртуализированные конфигурации, подобные той, которая используется для приложений Centennial (рис. 10.7). Модель современных приложений и архитектура приложений Centennial описаны в главе 8.

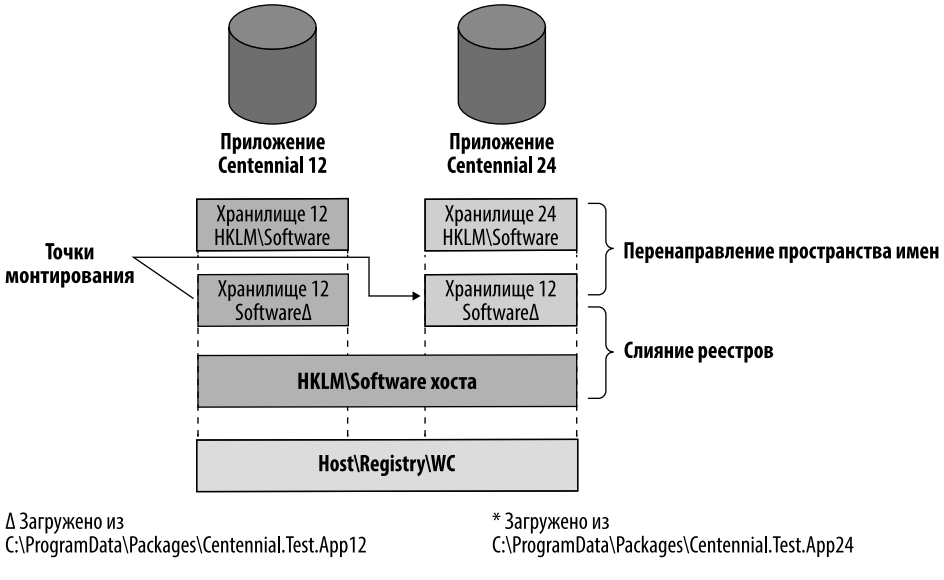


Рис. 10.7. Виртуализация реестра программного куста в модели современных приложений для приложений Centennial

Оптимизация реестра

Диспетчер конфигурации делает несколько примечательных оптимизаций производительности. Во-первых, практически каждый раздел реестра имеет дескриптор безопасности, ограничивающий доступ к нему. Однако хранить уникальную копию дескриптора безопасности для каждого раздела в кусте было бы крайне неэффективно, поскольку одни и те же параметры безопасности часто применяются ко всем поддеревьям реестра. Когда система применяет меры защиты к разделу, диспетчер конфигурации проверяет пул уникальных дескрипторов безопасности, используемых в том же кусте, что и раздел, к которому применяется новая защита, и совместно использует любой существующий дескриптор раздела, гарантируя, что существует не более одной копии каждого уникального дескриптора безопасности в кусте.

Диспетчер конфигурации также оптимизирует способ хранения имен разделов и параметров в кусте. Хотя реестр полностью поддерживает Юникод и указывает все имена, следуя правилам этой кодировки, но если имя содержит только символы ASCII, диспетчер конфигурации сохраняет его в кусте в кодировке ASCII. Когда диспетчер конфигурации считывает имя (например, при поиске имени), он преобразует его в кодировку Юникод в памяти. Сохранение имени в форме ASCII может значительно уменьшить размер куста.

Чтобы минимизировать использование памяти, блоки управления разделами не хранят полные имена путей разделов реестра. Вместо этого они ссылаются только на имя раздела. Например, блок управления, который указывает на `\Registry\System\Control`, будет ссылаться на имя `Control`, а не на полный путь. Дополнительная оптимизация памяти заключается в том, что диспетчер конфигурации использует блоки управления именами разделов для хранения имен разделов, а все блоки управления разделами применяют для разделов с одинаковым именем один и тот же блок управления именем раздела. Для оптимизации производительности диспетчер конфигурации сохраняет имена блоков управления разделами в хеш-таблице для быстрого поиска.

Чтобы обеспечить быстрый доступ к блокам управления разделами, диспетчер конфигурации сохраняет часто используемые блоки из них в таблице кэша, которая настроена как хеш-таблица. Когда диспетчеру конфигурации необходимо найти блок управления разделами, он сначала проверяет таблицу кэша. Наконец, у диспетчера конфигурации есть еще один кэш — таблица отложенного закрытия, в которой хранятся блоки управления разделами, которые приложения закрывают, чтобы приложение могло быстро повторно открыть недавно закрытый раздел. Оптимизировать поиск помогает то, что эти таблицы кэша хранятся для каждого куста. Диспетчер конфигурации удаляет самые старые блоки управления разделами из таблицы отложенного закрытия, чтобы добавить в нее самые последние закрытые блоки.

СЛУЖБЫ WINDOWS

Почти каждая операционная система имеет механизм запуска процессов во время загрузки системы, не привязанный к интерактивному пользователю. В Windows такие процессы называются службами или службами Windows. Службы аналогичны процессам-демонам UNIX и часто реализуют серверную часть клиент-серверных приложений. Примером службы Windows может быть веб-сервер, поскольку он

должен работать независимо от того, вошел ли кто-либо в систему на компьютере, и запускаться при загрузке системы, чтобы администратору не приходилось помнить о необходимости запуска или вообще присутствовать для него.

Службы Windows состоят из трех компонентов: приложения служб, программы управления службами (service control program, SCP) и диспетчера управления службами (Service Control Manager, SCM). Для начала поговорим о приложениях служб, учетных записях служб, пользовательских и пакетных службах, а также обо всех операциях SCM. Затем объясним, как запускаются службы автозапуска во время загрузки системы. Мы также рассмотрим действия, которые SCM предпринимает в случае сбоя службы во время ее запуска, и способы, которыми он отключает службы. А закончим описанием процесса разделяемой службы и того, как система управляет защищенными службами.

Приложения служб

Приложения служб, такие как веб-серверы, состоят как минимум из одного исполняемого файла, который работает как служба Windows. Пользователь, который хочет запустить, остановить или настроить службу, задействует SCP. Хотя Windows предоставляет встроенные SCP (наиболее распространенными являются инструмент командной строки sc.exe и пользовательский интерфейс, предоставляемый оснасткой MMC Services.msc), которые обеспечивают общие функции запуска, остановки, паузы и продолжения, некоторые приложения служб включают в себя собственный SCP, который позволяет администраторам указывать параметры конфигурации, присущие только службе, которой они управляют.

Непосредственно же службы — это просто исполняемые файлы Windows (с графическим или консольным интерфейсом) с дополнительным кодом для получения команд от SCM, а также для передачи статуса приложения обратно в SCM. Поскольку большинство служб не имеют пользовательского интерфейса, они созданы как консольные программы.

Когда вы устанавливаете приложение, содержащее службу, программа установки приложения, которая обычно действует так же, как SCP, должна зарегистрировать службу в системе. Чтобы сделать это, программа установки вызывает `CreateService` — функцию Windows, связанную со службами и экспортированную в Advapi32.dll (%SystemRoot%\System32\Advapi32.dll). Advapi32 — библиотека Advanced API DLL — реализует лишь небольшую часть клиентских API SCM. Все наиболее важные клиентские API SCM реализованы в другой DLL — Sechost.dll, которая является хост-библиотекой для клиентских API SCM и LSA. Все API SCM, не реализованные в Advapi32.dll, просто перенаправляются в Sechost.dll. Большинство клиентских API SCM взаимодействуют с диспетчером управления службами через RPC. SCM реализован в исполняемом файле Services.exe. Более подробная информация приведена далее в разделе «Диспетчер управления службами».

Когда программа установки регистрирует службу, вызывая `CreateService`, выполняется вызов RPC к экземпляру SCM, работающему на целевом компьютере. Затем SCM создает раздел реестра для службы в разделе HKLM\SYSTEM\CurrentControlSet\Services. Раздел Services — это постоянное представление базы данных SCM. Отдельные разделы для каждой службы определяют путь к исполняемому образу, содержащему службу, а также параметры конфигурации.

После создания службы приложение установки или управления может запустить службу с помощью функции `StartService`. Поскольку для работы некоторые приложения служб должны инициализироваться во время процесса загрузки, программа установки нередко регистрирует службу для автозапуска, просит пользователя перезагрузить систему для завершения установки и позволяет SCM запустить службу при загрузке системы.

Когда программа вызывает `CreateService`, она должна указать ряд параметров, описывающих характеристики службы. Они включают тип службы (если это служба, которая работает в собственном процессе, а не использующая общий процесс с другими), расположение исполняемого образа, необязательное отображаемое имя, необязательное имя учетной записи и пароль, применяемый для запуска службы в контексте безопасности конкретной учетной записи. Кроме того, туда входят тип запуска, указывающий, запускается ли служба автоматически при загрузке системы или вручную под руководством SCP, код ошибки, указывающий, как система должна реагировать, если служба выдает ошибку при запуске и дополнительная информация, если есть автозапуск, указывающая, когда запускаться относительно других служб. Хотя службы с отложенной загрузкой поддерживаются, начиная с Windows Vista, в Windows 7 появилась поддержка триггерных служб, которые запускаются или останавливаются при подтверждении наступления одного или нескольких определенных событий. SCP может указать информацию о событии триггера с помощью API `ChangeServiceConfig2`.

Приложение службы выполняется в процессе службы. В таком процессе может размещаться одно или несколько приложений служб. Когда SCM запускает процесс службы, тот должен немедленно вызвать функцию `StartServiceCtrlDispatcher` (до истечения четко определенного времени ожидания, подробности см. в разделе «Вход службы в систему»). `StartServiceCtrlDispatcher` принимает список точек входа в службы, по одной для каждой службы в процессе. Каждая точка входа идентифицируется именем службы, которой она соответствует. После установления локального коммуникационного соединения RPC (ALPC) с SCM (который действует как канал) `StartServiceCtrlDispatcher` в цикле ожидает поступления команд по каналу от SCM. Обратите внимание на то, что дескриптор соединения сохраняется SCM во внутреннем списке, который используется для отправки служебных команд нужному процессу и их получения от него. SCM отправляет команду запуска службы каждый раз, когда запускает службу, принадлежащую процессу. Для каждой полученной команды запуска функция `StartServiceCtrlDispatcher` создает поток, называемый потоком службы, для вызова точки входа стартовой службы (`ServiceMain`) и реализации командного цикла для службы. `StartServiceCtrlDispatcher` бесконечно ожидает команд от SCM и возвращает управление основной функции процесса только тогда, когда все службы процесса остановлены, что позволяет процессу службы очистить ресурсы перед выходом.

Первым действием точки входа в службу (`ServiceMain`) является вызов функции `RegisterServiceCtrlHandler`. Та получает и сохраняет указатель на функцию, называемую обработчиком управления, которую служба реализует для обработки различных команд, получаемых от SCM. `RegisterServiceCtrlHandler` не взаимодействует с SCM, но сохраняет функцию в локальной памяти процесса для функции `StartServiceCtrlDispatcher`. Точка входа службы продолжает инициализацию

службы, которая может включать в себя выделение памяти, создание точек связи и чтение частных данных конфигурации из реестра. Как объяснялось ранее, соглашение, которому следуют большинство служб, заключается в хранении своих параметров в подразделе Parameters раздела реестра службы.

Пока точка входа инициализирует службу, она должна периодически отправлять в SCM сообщения о состоянии с помощью функции `SetServiceStatus`, указывая, как продвигается запуск службы. После того как точка входа завершает инициализацию (служба сообщает об этом SCM посредством статуса `SERVICE_RUNNING`), поток службы обычно находится в цикле, ожидая запросов от клиентских приложений. Например, веб-сервер инициализирует сокет прослушивания TCP и ожидает входящих запросов HTTP-соединения.

Основной поток процесса службы, который выполняется в функции `StartServiceCtrlDispatcher`, получает команды SCM, направляемые службам в ходе работы, и вызывает функцию обработчика управления целевой службы, хранящуюся в `RegisterServiceCtrlHandler`. Команды SCM включают команды остановки, паузы, возобновления, опроса и выключения, а также команды, определяемые приложением. На рис. 10.8 показаны внутренняя организация процесса службы, основной поток и поток службы, составляющие процесс, в котором размещается одна служба.

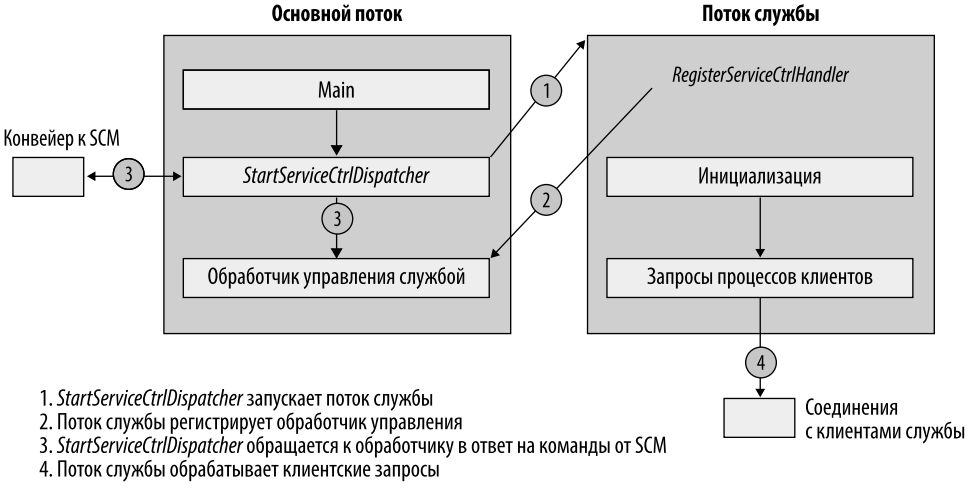


Рис. 10.8. Внутреннее устройство процесса службы

Характеристики службы

SCM сохраняет каждую характеристику как параметр в разделе реестра службы. Пример раздела реестра службы показан на рис. 10.9.

В табл. 10.7 перечислены все характеристики службы, многие из которых можно применять и к драйверам устройств. (Не каждая характеристика применима к каждому типу службы или драйверу устройства.)

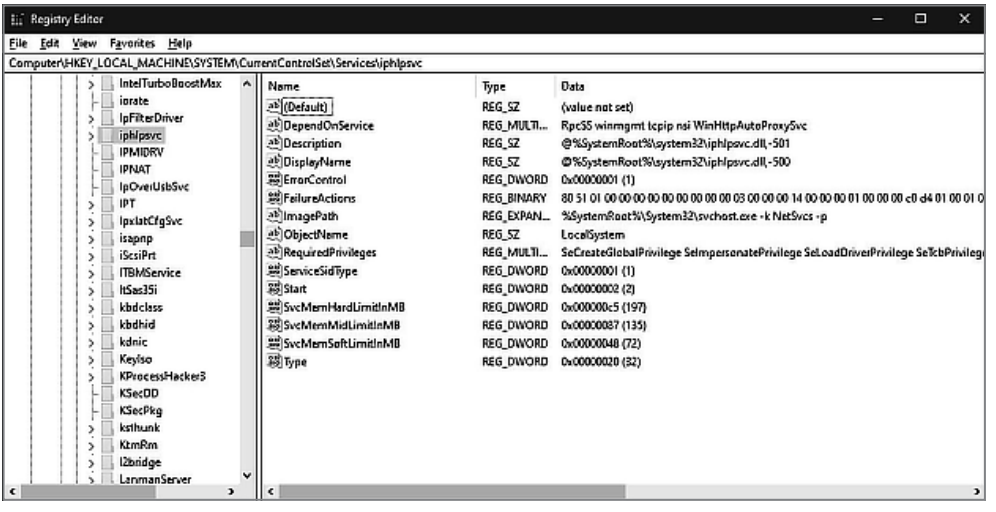


Рис. 10.9. Пример раздела службы в реестре

ПРИМЕЧАНИЕ SCM не имеет доступа к подразделу Parameters службы до тех пор, пока служба не будет удалена, после чего SCM удаляет весь раздел службы, включая такие подразделы, как Parameters.

Таблица. 10.7. Параметры реестра служб и драйверов

Установка параметра	Название параметра	Описание настройки
Start	SERVICE_BOOT_START (0x0)	Winload предварительно загружает драйвер, чтобы он находился в памяти во время загрузки. Эти драйверы инициализируются непосредственно перед драйверами SERVICE_SYSTEM_START
	SERVICE_SYSTEM_START (0x1)	Драйвер загружается и инициализируется во время инициализации ядра после инициализации драйверов SERVICE_BOOT_START
	SERVICE_AUTO_START (0x2)	SCM запускает драйвер или службу после запуска процесса SCM Services.exe
	SERVICE_DEMAND_START (0x3)	SCM запускает драйвер или службу по требованию (он запускается триггером, когда клиент вызывает для него StartService или когда от него зависит другая запускаемая служба)
	SERVICE_DISABLED (0x4)	Невозможно загрузить или инициализировать драйвер или службу

Установка параметра	Название параметра	Описание настройки
ErrorControl	SERVICE_ERROR_IGNORE (0x0)	Любая ошибка, возвращаемая драйвером или службой, игнорируется, и никакие предупреждения не регистрируются и не отображаются
	SERVICE_ERROR_NORMAL (0x1)	Если драйвер или служба сообщает об ошибке, записывается сообщение в журнал событий
	SERVICE_ERROR_SEVERE (0x2)	Если драйвер или служба возвращает ошибку и последнее известное исправное состояние системы не используется, перезагрузитесь в последнее известное исправное состояние, в ином случае зарегистрируйте сообщение о событии
	SERVICE_ERROR_CRITICAL (0x3)	Если драйвер или служба возвращает ошибку и последнее известное исправное состояние не используется, перезагрузитесь в последнее известное исправное состояние, в ином случае зарегистрируйте сообщение о событии
Type	SERVICE_KERNEL_DRIVER (0x1)	Драйвер устройства
	SERVICE_FILE_SYSTEM_ DRIVER (0x2)	Драйвер файловой системы режима ядра
	SERVICE_ADAPTER (0x4)	Устарело
	SERVICE_RECOGNIZER_ DRIVER (0x8)	Драйвер распознавателя файловой системы
	SERVICE_WIN32_OWN_PROCESS (0x10)	Служба выполняется в процессе, в котором размещается только одна служба
	SERVICE_WIN32_SHARE_ PROCESS (0x20)	Служба выполняется в процессе, в котором размещены несколько служб
	SERVICE_USER_OWN_PROCESS (0x50)	Служба запускается с токеном безопасности вошедшего в систему пользователя в собственном процессе
	SERVICE_USER_SHARE_PROCESS (0x60)	Служба запускается с токеном безопасности вошедшего в систему пользователя в процессе, в котором размещаются несколько служб
	SERVICE_INTERACTIVE_ PROCESS (0x100)	Службе разрешено отображать окна на консоли и получать пользовательский ввод, но только в сеансе консоли (0), чтобы предотвратить взаимодействие с пользовательскими/консольными приложениями в других сеансах. Параметр устарел

Продолжение ⇩

Таблица 10.7 *(продолжение)*

Установка параметра	Название параметра	Описание настройки
Groupname	Имя группы	Драйвер или служба инициализируется при инициализации своей группы
Tag	Номер тега	Указанное местоположение в порядке инициализации группы. Этот параметр не применяется к службам
ImagePath	Путь к исполняемому файлу службы или драйвера	Если ImagePath не указан, диспетчер ввода-вывода ищет драйверы в %SystemRoot%\System32\Drivers. Требуется для служб Windows
DependOnGroup	Имя группы	Драйвер или служба не будет загружаться, пока не загрузится драйвер или служба из указанной группы
DependOnService	Имя службы	Служба не будет загружаться до тех пор, пока не загрузится указанная служба. Этот параметр не применяется к драйверам устройств или службам с типом запуска, отличным от SERVICE_AUTO_START или SERVICE_DEMAND_START
ObjectName	Обычно LocalSystem, но это может быть имя учетной записи, например .\ Администратор	Указывает учетную запись, под которой будет запускаться служба. Если ObjectName не указано, используется учетная запись LocalSystem. Этот параметр не применяется к драйверам устройств
DisplayName	Имя службы	Приложение службы отображает службы под этим именем. Если имя не указано, им становится имя раздела реестра службы
DeleteFlag	0 или 1 (TRUE или FALSE)	Временный флаг, устанавливаемый SCM, когда служба помечена для удаления
Описание	Описание службы	До 32 767 байтов описания службы
FailActions	Описание действий, которые должен предпринять SCM в случае неожиданного завершения процесса обслуживания	Действия при сбое включают перезапуск процесса службы, перезагрузку системы и запуск указанной программы. Этот параметр не распространяется на драйверы
FailCommand	Командная строка программы	SCM считывает этот параметр только в том случае, если в поле FailureActions указано, что программа должна выполняться в случае сбоя службы. Этот параметр не распространяется на драйверы
DelayedAutoStart	0 или 1 (TRUE или FALSE)	Указывает SCM запустить эту службу по истечении определенной задержки с момента запуска SCM. Это уменьшает количество служб, запускаемых одновременно во время запуска

Установка параметра	Название параметра	Описание настройки
PreshutdownTimeout	Тайм-аут в миллисекундах	Этот параметр позволяет службам переопределить тайм-аут уведомления перед завершением работы по умолчанию, равный 180 с. По истечении этого времени SCM выполняет действия по завершении работы службы, если она еще не ответила
ServiceSidType	SERVICE_SID_TYPE_NONE (0x0)	Параметр для обратной совместимости
	SERVICE_SID_TYPE_UNRESTRICTED (0x1)	SCM добавляет SID службы в качестве владельца группы к токену процесса службы при его создании
	SERVICE_SID_TYPE_RESTRICTED (0x3)	SCM запускает службу с токеном с ограниченной записью, добавляя SID службы в список SID с ограниченным доступом процесса службы вместе с глобальными, времени входа и ограниченными для записи SID
Alias	Строка	Имя псевдонима службы
RequiredPrivileges	Список привилегий	Этот параметр содержит список привилегий, необходимых службе для работы. SCM вычисляет их объединение при создании токена для общего процесса, связанного с этой службой, если таковой имеется
Security	Дескриптор безопасности	Этот параметр содержит необязательный дескриптор безопасности, который определяет, кто имеет какой доступ к объекту службы, созданному внутри SCM. Если этот параметр опущен, SCM применяет дескриптор безопасности по умолчанию
LaunchProtected	SERVICE_LAUNCH_PROTECTED_NONE (0x0)	SCM запускает службу незащищенной (значение по умолчанию)
	SERVICE_LAUNCH_PROTECTED_WINDOWS (0x1)	SCM запускает службу в процессе, защищенном Windows
	SERVICE_LAUNCH_PROTECTED_WINDOWS_LIGHT (0x2)	SCM запускает службу в свете защищенного процесса Windows
	SERVICE_LAUNCH_PROTECTED_ANTIMALWARE_LIGHT (0x3)	SCM запускает службу в свете процесса, защищенного от вредоносного ПО
	SERVICE_LAUNCH_PROTECTED_APP_LIGHT (0x4)	SCM запускает службу в индикаторе процесса, защищенного приложением (только для внутреннего применения)

Продолжение ⇩

Таблица 10.7 *(продолжение)*

Настраиваемое значение	Название значения	Описание настройки
UserServiceFlags	USER_SERVICE_FLAG_DSMA_ ALLOW (0x1)	Разрешить пользователю по умолчанию запускать службу пользователя
	USER_SERVICE_FLAG_NONDSMA_ ALLOW (0x2)	Не разрешать пользователю по умолчанию запускать службу
SvcHostSplitDisable	0 или 1 (TRUE или FALSE)	Если установлено значение, 1 запрещает SCM включать разделение Svchost. Этот параметр применяется только к разделяемым службам
PackageFullName	String	Полное имя пакета пакетной службы
AppUserModelId	String	Идентификатор модели пользователя приложения (AUMID) пакетной службы
PackageOrigin	PACKAGE_ORIGIN_UNSIGNED (0x1), PACKAGE_ORIGIN_INBOX (0x2), PACKAGE_ORIGIN_STORE (0x3), PACKAGE_ORIGIN_DEVELOPER_ UNSIGNED (0x4), PACKAGE_ORIGIN_ DEVELOPER_SIGNED (0x5)	Этот параметр определяют происхождение пакета AppX (субъект, который его создал)

Обратите внимание на то, что среди параметров типа есть применяемые к драйверам устройств: драйвер устройства, драйвер файловой системы и распознаватель файловой системы. Они используются драйверами устройств Windows, которые также сохраняют свои параметры в виде данных реестра в разделе реестра Services. SCM отвечает за запуск драйверов, не поддерживающих PNP, с параметром `Start`, содержащим `SERVICE_AUTO_START` или `SERVICE_DEMAND_START`, поэтому для базы данных SCM естественно включать драйверы. Службы используют другие типы — взаимоисключающие `SERVICE_WIN32_OWN_PROCESS` и `SERVICE_WIN32_SHARE_PROCESS`.

Исполняемый файл, в котором размещается только одна служба, задействует тип `SERVICE_WIN32_OWN_PROCESS`. Аналогично исполняемый файл, в котором размещены несколько служб, указывает `SERVICE_WIN32_SHARE_PROCESS`. Размещение нескольких служб в одном процессе экономит системные ресурсы, которые в противном случае стали бы накладными расходами при запуске нескольких процессов служб. Потенциальный недостаток заключается в следующем: если одна из служб коллекции, запущенной в том же процессе, вызывает ошибку, завершающую процесс, все службы этого процесса завершаются. Кроме того, еще одним ограничением является то, что все службы должны работать под одной и той же учетной записью (но если служба использует механизмы усиления безопасности службы, она может стать менее подверженной вредоносным атакам). Флаг `SERVICE_USER_SERVICE` добавляется для обозначения пользовательской службы, которая представляет собой тип службы, запускаемой с идентификатором текущего пользователя, вошедшего в систему.

Информация о триггере обычно хранится в SCM, в подразделе с именем `TriggerInfo`. Каждое событие триггера сохраняется в дочернем разделе, называемом индексом события и начинающемся с 0 (например, третье событие триггера хранится

в подразделе `TriggerInfo\2`). В табл. 10.8 перечислены все возможные параметры реестра, составляющие информацию о триггере.

Таблица 10.8. Параметры реестра триггерных служб

Установка параметра	Название параметра	Описание установки параметра
Action	SERVICE_TRIGGER_ACTION_SERVICE_START (0x1)	Запустить службу при возникновении триггерного события
	SERVICE_TRIGGER_ACTION_SERVICE_STOP (0x2)	Остановить службу при возникновении триггерного события
Type	SERVICE_TRIGGER_TYPE_DEVICE_INTERFACE_ARRIVAL (0x1)	Определяет событие, инициируемое, когда устройство указанного класса интерфейса устройства прибывает или присутствует при запуске системы
	SERVICE_TRIGGER_TYPE_IP_ADDRESS_AVAILABILITY (0x2)	Указывает событие, инициируемое, когда IP-адрес становится доступным или недоступным в сетевом стеке
	SERVICE_TRIGGER_TYPE_DOMAIN_JOIN (0x3)	Указывает событие, инициируемое, когда компьютер присоединяется к домену или покидает его
	SERVICE_TRIGGER_TYPE_FIREWALL_PORT_EVENT (0x4)	Указывает событие, инициируемое при открытии или закрытии порта брандмауэра
	SERVICE_TRIGGER_TYPE_GROUP_POLICY (0x5)	Указывает событие, инициируемое при изменении политики компьютера или пользователя
	SERVICE_TRIGGER_TYPE_NETWORK_ENDPOINT (0x6)	Указывает событие, инициируемое, когда пакет или запрос поступает по определенному сетевому протоколу
	SERVICE_TRIGGER_TYPE_CUSTOM (0x14)	Указывает пользовательское событие, созданное поставщиком ETW
Guid	GUID подтипа триггера	GUID, идентифицирующий подтип события триггера. GUID зависит от типа триггера
Data[Index]	Данные, специфичные для триггера	Данные, специфичные для триггера для события триггера службы. Этот параметр зависит от типа события-триггера
DataType[Index]	SERVICE_TRIGGER_DATA_TYPE_BINARY (0x1)	Данные, относящиеся к триггеру, имеют двоичный формат
	SERVICE_TRIGGER_DATA_TYPE_STRING (0x2)	Данные, специфичные для триггера, имеют строковый формат
	SERVICE_TRIGGER_DATA_TYPE_LEVEL (0x3)	Данные, специфичные для триггера, представляют собой байтовое значение
	SERVICE_TRIGGER_DATA_TYPE_KEYWORD_ANY (0x4)	Данные, специфичные для триггера, представляют собой 64-битное (8 байтов) целое число без знака
	SERVICE_TRIGGER_DATA_TYPE_KEYWORD_ALL (0x5)	Данные, специфичные для триггера, представляют собой 64-битное (8 байтов) целое число без знака

Учетные записи служб

Контекст безопасности службы — важный фактор для разработчиков служб и системных администраторов, поскольку он определяет, к какому ресурсу может получить доступ процесс. Большинство встроенных служб работают в контексте безопасности соответствующей учетной записи службы, которая имеет ограниченные права доступа, как описано в следующих подразделах. Когда программа установки службы или системный администратор создает службу, они обычно указывают контекст безопасности локальной системной учетной записи (иногда отображается как SYSTEM, а иногда как LocalSystem), что очень важно. Две другие встроенные учетные записи — это учетные записи сетевых и локальных служб. У них меньше возможностей, чем у учетной записи локальной системы с точки зрения безопасности. В следующих разделах описаны особые характеристики всех учетных записей служб.

Учетная запись локальной системы

Учетная запись локальной системы — это та же учетная запись, под которой запускаются основные компоненты операционной системы Windows в пользовательском режиме, включая диспетчер сеансов (%SystemRoot%\System32\Smss.exe), процесс подсистемы Windows (Csrss.exe), процесс локального центра безопасности (%SystemRoot%\System32\Lsass.exe) и процесс входа в систему (%SystemRoot%\System32\Winlogon.exe). Дополнительную информацию об этих процессах см. в главе 7 тома 1.

С точки зрения безопасности учетная запись локальной системы чрезвычайно мощная — более мощная, чем любая локальная или доменная учетная запись, когда речь идет о возможностях безопасности в локальной системе. Эта учетная запись имеет следующие характеристики.

- Является членом локальной группы администраторов. В табл. 10.9 показаны группы, к которым принадлежит учетная запись локальной системы. (Информацию о том, как членство в группе используется при проверке доступа к объектам, см. в главе 7 тома 1.)
- Имеет право активировать все привилегии, даже те, которые обычно не предоставляются учетной записи локального администратора, например, создание токенов безопасности. В табл. 10.10 приведен список привилегий, назначенных учетной записи локальной системы. (Глава 7 тома 1 описывает использование каждой привилегии.)
- Большинство файлов и разделов реестра предоставляют полный доступ к учетной записи локальной системы. Даже если они не дают полного доступа, процесс, запущенный под учетной записью локальной системы, для получения доступа может воспользоваться привилегией получения владения.
- Процессы, запущенные под учетной записью локальной системы, запускаются с профилем пользователя по умолчанию (HKU\.DEFAULT). Таким образом, они не могут напрямую получить доступ к информации о конфигурации, хранящейся в профилях пользователей других учетных записей, если явно не задействуют API `LoadUserProfile`.

Таблица 10.9. Членство учетных записей служб в группах (и уровень целостности)

Локальная система	Сетевая служба	Локальная служба	Учетная запись службы
Администраторы	Все	Все	Все
Все	Пользователи	Пользователи	Пользователи
Проверенные пользователи	Проверенные пользователи Локальная Сетевая служба Вход в консоль	Проверенные пользователи Локальная Локальная служба Вход в консоль Группы возможностей UWP	Проверенные пользователи Локальная Локальная служба Все службы Ограничение на запись Вход в консоль
Системный уровень целостности	Системный уровень целостности	Системный уровень целостности	Высокий уровень целостности

Таблица 10.10. Привилегии учетных записей служб

Локальная система	Локальная/сетевая служба	Учетная запись службы
SeAssignPrimaryTokenPrivilege SeAuditPrivilege SeBackupPrivilege SeChangeNotifyPrivilege SeCreateGlobalPrivilege SeCreatePagefilePrivilege SeCreatePermanentPrivilege SeCreateSymbolicLinkPrivilege SeCreateTokenPrivilege SeDebugPrivilege SeDelegateSessionUserImpersonatePrivilege SeImpersonatePrivilege SeIncreaseBasePriorityPrivilege SeIncreaseQuotaPrivilege SeIncreaseWorkingSetPrivilege SeLoadDriverPrivilege SeLockMemoryPrivilege SeManageVolumePrivilege SeProfileSingleProcessPrivilege SeRestorePrivilege SeSecurityPrivilege SeShutdownPrivilege SeSystemEnvironmentPrivilege SeSystemProfilePrivilege SeSystemtimePrivilege SeTakeOwnershipPrivilege SeTcbPrivilege SeTimeZonePrivilege SeTrustedCredManAccessPrivilege SeRelabelPrivilege SeUndockPrivilege (только клиент)	SeAssignPrimaryTokenPrivilege SeAuditPrivilege SeChangeNotifyPrivilege SeCreateGlobalPrivilege SeImpersonatePrivilege SeIncreaseQuotaPrivilege SeIncreaseWorkingSetPrivilege SeShutdownPrivilege SeSystemtimePrivilege SeTimeZonePrivilege SeUndockPrivilege (только клиент)	SeChangeNotifyPrivilege SeCreateGlobalPrivilege SeImpersonatePrivilege SeIncreaseWorkingSetPrivilege SeShutdownPrivilege SeTimeZonePrivilege SeUndockPrivilege

- Если система является членом домена Windows, учетная запись локальной системы включает в себя идентификатор безопасности компьютера (SID) для компьютера, на котором запущен служебный процесс. Таким образом, служба, запущенная под учетной записью локальной системы, будет автоматически аутентифицироваться на других компьютерах в том же лесу с использованием своей учетной записи компьютера. (*Лес* — это группа доменов.)
- Если учетной записи компьютера специально не предоставлен доступ к ресурсам, таким как общие сетевые ресурсы, именованные конвейеры и т. д., процесс может получить доступ к сетевым ресурсам, которые допускают нулевые сеансы, то есть соединения, не требующие учетных данных. Вы можете указать общие ресурсы и конвейеры на конкретном компьютере, которые разрешают нулевые сеансы, в параметрах `NullSessionPipes` и `NullSessionShares` в разделе реестра HKLM\SYSTEM\CurrentControlSet\Services\LanmanServer\Parameters.

Учетная запись сетевой службы

Учетная запись сетевой службы предназначена для использования службами, которым требуется проходить аутентификацию на других компьютерах в сети с помощью учетной записи компьютера, подобно локальной учетной записи системы, но которые не нуждаются в ее многочисленных привилегиях или членстве в группе администраторов. Отсутствие полномочий администратора приводит к тому, что службы, действующие под сетевой учетной записью, по умолчанию имеют доступ к гораздо меньшему количеству разделов реестра, папок файловой системы и файлов, чем работающие от имени локальной системы. Кроме того, малое количество привилегий ограничивает влияние скомпрометированного процесса сетевой службы. Например, будучи запущенным под учетной записью сетевой службы, тот не может загрузить драйвер устройства или открыть произвольные процессы.

Еще одно различие между учетными записями сетевой службы и локальной системы заключается в том, что процессы, запущенные от лица сетевой службы, используют соответствующий профиль учетной записи. Компонент реестра профиля сетевой службы загружается в раздел HKU\S-1-5-20, а файлы и каталоги, составляющие его, находятся в папке %SystemRoot%\ServiceProfiles\NetworkService.

Примером службы, действующей под учетной записью сетевой службы, является DNS-клиент, который отвечает за разрешение DNS-имен и поиск контроллеров домена.

Учетная запись локальной службы

Учетная запись локальной службы практически идентична учетной записи сетевой службы, с той важной разницей, что она может получить доступ только к сетевым ресурсам, которые разрешают анонимный доступ. В табл. 10.10 показано, что учетная запись сетевой службы имеет те же привилегии, что и учетная запись локальной службы, а из табл. 10.9 видно, что она принадлежит к тем же группам, за исключением того, что она относится к группе локальных служб, а не сетевых. Профиль, используемый процессами, работающими в локальной службе, загружается в HKU\S-1-5-19 и сохраняется в %SystemRoot%\ServiceProfiles\LocalService.

К примерам служб, запускаемых под учетной записью локальной службы, относятся служба удаленного реестра, обеспечивающая удаленный доступ к реестру локальной системы, и служба LmHosts, которая выполняет разрешение имен NetBIOS.

Запуск служб в альтернативных учетных записях

Из-за только что названных ограничений некоторые службы необходимо запускать с настройками безопасности учетной записи пользователя. Вы можете настроить службу для запуска под альтернативной учетной записью при ее создании или указав учетную запись и пароль, под которыми службе следует работать, с помощью оснастки MMC Службы Windows (Windows Services). В оснастке Службы (Services) щелкните правой кнопкой мыши на службе, выберите Свойства (Properties), перейдите на вкладку Вход в систему (Log On) и выберите параметр С учетной записью (This Account) (рис. 10.10).

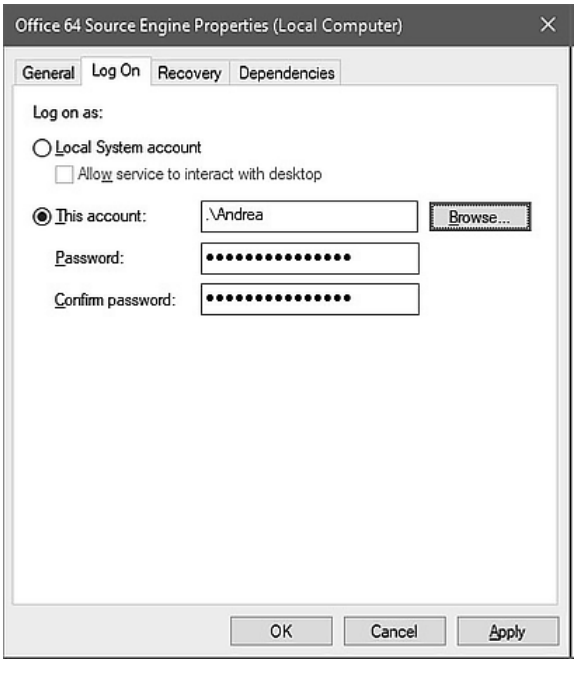


Рис. 10.10. Настройки учетной записи службы

Обратите внимание: при необходимости загрузки служба, работающая с альтернативной учетной записью, всегда запускается с помощью учетных данных этой записи, даже если вход в систему от ее лица в этот момент не выполнен. Это означает, что профиль пользователя загружается, даже если тот не вошел в систему. Пользовательские службы, которые описаны далее в этой главе (в разделе «Пользовательские службы»), также были разработаны для решения этой проблемы. Они загружаются только при входе пользователя в систему.

Запуск с наименьшими привилегиями

Процесс службы обычно подчиняется модели «все или ничего», а поэтому все привилегии, доступные учетной записи, под которой он запущен, доступны работающей в нем службе, возможно, нуждающейся только в подмножестве этих привилегий.

Чтобы лучше соответствовать принципу наименьших привилегий, согласно которому Windows назначает службам только необходимые полномочия, разработчики могут указать привилегии, необходимые их службе, а SCM создает токен безопасности, описывающий только эти привилегии.

Разработчики служб используют API `ChangeServiceConfig2` (с указанием уровня информации `SERVICE_CONFIG_REQUIRED_PRIVILEGES_INFO`), чтобы привести список желаемых привилегий. API сохраняет эту информацию в реестре в параметре `RequiredPrivileges` корневого раздела службы (см. табл. 10.7). Когда служба запускается, SCM считывает раздел и добавляет эти привилегии к токену процесса, в котором работает служба.

Если существует параметр реестра `RequiredPrivileges` и служба автономна (выполняется как выделенный процесс), SCM создает токен, описывающий только те привилегии, которые ей необходимы. Для служб, работающих в рамках общего хост-процесса (подобно подмножеству служб, которые являются частью Windows) и определяющих необходимые привилегии, SCM вычисляет объединение этих привилегий, тем самым комбинируя их и добавляя в токен хост-процесса службы. Другими словами, будут удалены только привилегии, не указанные ни одной из служб, размещенных в одном и том же процессе службы. Когда параметра реестра не существует, у SCM нет иного выбора, кроме как предположить, что служба либо несовместима с минимальными привилегиями, либо требует для работы всех привилегий. В этом случае создается полный токен, содержащий все привилегии, и эта модель не обеспечивает никакой дополнительной безопасности. Чтобы лишиться почти всех привилегий, службы могут указать только привилегию «Уведомление об изменении».

ПРИМЕЧАНИЕ Привилегии, указанные службой, должны быть подмножеством тех, которые доступны учетной записи службы, под которой она работает.

ЭКСПЕРИМЕНТ. Просмотр привилегий, запрашиваемых службами

Вы можете просмотреть привилегии, необходимые службе, с помощью утилиты управления службами sc.exe и параметра `qprivs`. Кроме того, Process Explorer может отображать информацию о токене безопасности любого процесса служб в системе, поэтому вы можете сравнить информацию, возвращаемую sc.exe, с привилегиями в рамках его токена. Рассмотрим, как это сделать для некоторых из наиболее ограниченных служб в системе.

1. Используйте sc.exe, чтобы просмотреть необходимые привилегии, указанные `CryptSvc`, введя в командную строку следующее:

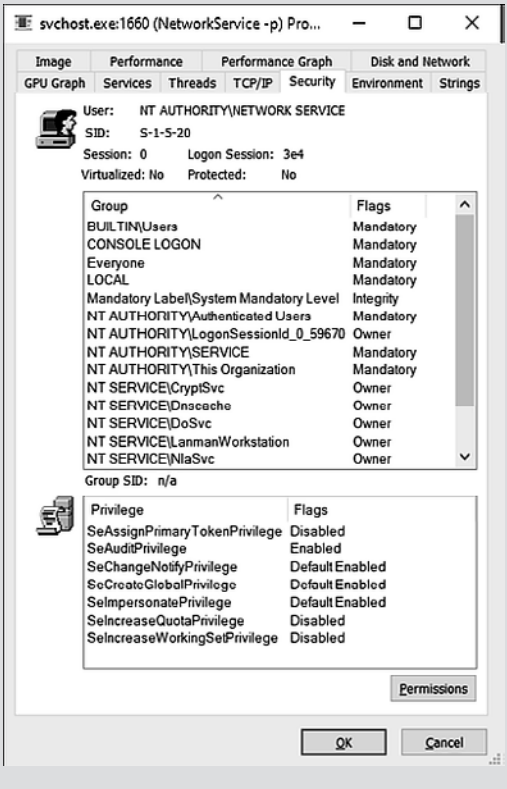
   ```
   sc qprivs cryptsvc
   ```

   Вы должны будете увидеть три запрашиваемые привилегии: `SeChangeNotifyPrivilege`, `SeCreateGlobalPrivilege` и `SeImpersonatePrivilege`.

2. Запустите Process Explorer от имени администратора и просмотрите список процессов.

Вы должны будете увидеть несколько процессов Svchost.exe, на которых размещаются службы на вашем компьютере (если включено разделение Svchost, количество экземпляров Svchost будет еще больше). Process Explorer выделяет их розовым цветом.

3. CryptSvc — это служба, работающая в рамках общего хост-процесса. В Windows 10 легко найти правильный экземпляр процесса с помощью диспетчера задач. Вам не нужно знать имя служебной DLL, указанное в разделе реестра HKLM\SYSTEM\CurrentControlSet\Services\CryptSvc\Parameters.
4. Откройте диспетчер задач и просмотрите вкладку Службы (Services). Вы легко сможете найти PID хост-процесса CryptSvc.
5. Вернитесь в Process Explorer и дважды щелкните на процессе Svchost.exe, который имеет тот же PID, который вы нашли в диспетчере задач, чтобы открыть диалоговое окно Свойства (Properties).
6. Повторно убедитесь, что на вкладке Службы (Services) есть служба CryptSvc. Если разделение служб включено, процесс должен содержать только одну службу, в противном случае их будет несколько. Затем перейдите на вкладку Безопасность (Security). Вы должны увидеть информацию о безопасности, примерно как на следующем снимке экрана.



Обратите внимание: хотя служба работает от лица учетной записи локальной службы, список назначенных ей привилегий Windows намного короче, чем указано в списке привилегий, доступных для этой учетной записи (см. табл. 10.10).

Для хост-процесса службы часть токена, относящаяся к привилегиям, представляет собой объединение привилегий, запрошенных всеми службами, работающими в его рамках, из чего можно сделать вывод, что такие службы, как `DnsCache` и `LanmanWorkstation`, не запрашивали иных привилегий, кроме тех, которые показаны Process Explorer. Вы можете убедиться в этом, запустив инструмент Sc.exe для этих двух служб (доступно, только если разделение `Svchost` по службам отключено).

Изоляция служб

Хотя ограничение привилегий, к которым имеет доступ служба, помогает уменьшить способность скомпрометированного процесса службы компрометировать другие процессы, это не помогает изолировать службу от ресурсов, к которым учетная запись, в которой она работает, имеет доступ в нормальных условиях. Как упоминалось ранее, учетная запись локальной системы имеет полный доступ к критически важным системным файлам, разделам реестра и другим защищаемым объектам в системе, поскольку списки управления доступом (ACL) предоставляют такие разрешения этой учетной записи.

Иногда доступ к некоторым из этих ресурсов имеет решающее значение для работы службы, тогда как другие объекты должны быть защищены от нее. Раньше, чтобы избежать применения локальной системной учетной записи для получения доступа к необходимым ресурсам, служба запускалась под стандартной учетной записью пользователя, а у системных объектов дополнялись списки ACL, что значительно увеличивало риск атаки на систему со стороны вредоносного кода. Другим решением было создание выделенных учетных записей служб и установка определенных списков ACL для каждой учетной записи, связанной с службой, но этот подход легко превратился в административную проблему.

Сегодня Windows объединяет эти два подхода в гораздо более управляемое решение: система позволяет службам запускаться под непривилегированной учетной записью, но при этом иметь доступ к определенным привилегированным ресурсам без снижения безопасности этих объектов. Действительно, списки ACL объекта теперь могут устанавливать разрешения непосредственно для службы, не требуя специальной учетной записи. Вместо этого Windows генерирует специальный SID (идентификатор безопасности) для представления службы, и этот SID можно использовать для установки разрешений в отношении таких ресурсов, как разделы реестра и файлы.

Диспетчер управления службами задействует SID службы различными способами. Если служба настроена на запуск с помощью учетной записи виртуальной службы (в домене NT SERVICE\), генерируется SID службы и назначается основным

пользователем токена новой службы. Этот токен будет входить также в группу NT SERVICE\ALL SERVICES. Данная группа используется системой, чтобы обеспечить любой службе доступ к защищаемому объекту. В случае разделяемых служб SCM создает хост-процессы служб (процессы, содержащие более одной службы) с токеном, содержащим идентификаторы SID всех служб, входящих в группы служб, связанных с процессом, включая те, что еще не запущены (невозможно добавить новые SID после создания токена). Ограниченные и неограниченные службы (описаны далее в этом разделе) всегда имеют SID службы в токене хост-процесса.

ЭКСПЕРИМЕНТ. Разбор SID службы

В главе 9 мы представили эксперимент «Разбор защиты рабочего процесса VM и файлов виртуальных жестких дисков», в котором показали, как система генерирует SID виртуальной машины для различных рабочих процессов VM. Подобно рабочему процессу виртуальной машины, система генерирует SID для служб с помощью четко определенного алгоритма. В этом эксперименте используется Process Explorer для отображения идентификаторов SID служб и объясняется, как система их создает.

Прежде всего вам нужно выбрать службу, которая работает с учетной записью виртуальной службы или токеном ограниченного/неограниченного доступа. Откройте редактор реестра (введя `regedit` в поле поиска Cortana) и перейдите к разделу реестра HKLM\SYSTEM\CurrentControlSet\Services. Затем в меню Правка (Edit) выберите Найти (Find). Как обсуждалось ранее в этом разделе, учетная запись службы хранится в параметре реестра ObjectName. К сожалению, вы не найдете много служб, работающих под учетной записью виртуальной службы (эти учетные записи начинаются с виртуального домена NT SERVICE\), поэтому лучше взглянуть на ограниченный токен (неограниченные токены тоже подойдут). Введите `ServiceSidType` (его параметр сохраняется независимо от того, должна ли служба работать с ограниченным или неограниченным токеном) и нажмите кнопку Найти далее (Find Next).

Для этого эксперимента вам потребуется учетная запись службы с ограниченным доступом (ее параметр `ServiceSidType` установлен на 3), но и службы с неограниченным тоже хорошо подойдут (параметр установлен на 1). Если нужный параметр не совпадает, можете нажать клавишу F3, чтобы найти следующую службу. В этом эксперименте используйте службу BFE.

Откройте Process Explorer, найдите хост-процесс BFE (вернитесь к предыдущему эксперименту, чтобы понять, как найти правильный) и дважды щелкните на нем. Выберите вкладку Безопасность (Security) и щелкните на группе NT SERVICE\BFE (удобочитаемое обозначение SID службы) или SID службы, если выбрали другой. Обратите внимание на расширенный идентификатор безопасности группы, который отображается под списком групп. Если служба работает под

учетной записью виртуальной службы, вместо этого идентификатор безопасности службы отображается в Process Explorer, во второй строке вкладки Безопасность (Security):

```
S-1-5-80-1383147646-27650227-2710666058-1662982300-1023958487
```

Управляющая система NT (идентификатор 5) отвечает за идентификаторы SID службы, генерируемые с использованием базового RID службы (80) и хеша SHA-1 строки Юникода UTF-16 с именем службы в верхнем регистре. SHA-1 — это алгоритм, который создает 160-битное (20-байтовое) значение. В мире безопасности Windows это означает, что SID будет иметь пять 4-байтовых значений управляющих подсистем. Хеш SHA-1 имени службы BFE в Юникоде (UTF-16):

```
7e 28 71 52 b3 e8 a5 01 4a 7b 91 a1 9c 18 1f 63 d7 5d 08 3d
```

Если разделить полученный хеш на пять групп по восемь шестнадцатеричных цифр, вы обнаружите следующее:

- 0x5271287E (первое значение DWORD), которое равно 1 383 147 646 в десятичном формате (помните, что Windows — это ОС с порядком байтов Little Endian — прямым порядком);
- 0x01A5E8B3 (второе значение DWORD), которое равно 27 650 227 в десятичном формате;
- 0xA1917B4A (третье значение DWORD), которое равно 2 710 666 058 в десятичном формате;
- 0x631F189C (четвертое значение DWORD), которое равно 1 662 982 300 в десятичном формате;
- 0x3D085DD7 (пятое значение DWORD), которое равно 1 023 958 487 в десятичном формате.

Если объединить числа и добавить значение полномочий SID службы и первый RID (S-1-5-80), вы создадите тот же SID, который отображается в Process Explorer. Это демонстрирует, как система генерирует идентификаторы безопасности служб.

Полезность наличия SID для каждой службы выходит за рамки простой возможности добавлять записи ACL и разрешения для различных объектов в системе как способа детального контроля над их доступом. В нашем обсуждении первоначально рассматривался случай, когда определенные объекты в системе, доступные для данной учетной записи, должны быть защищены от службы, работающей под этой учетной записью. Как мы описывали ранее, идентификаторы SID службы предотвращают эту проблему, требуя лишь, чтобы записи Deny (Запрет), связанные с SID службы, размещались на каждом объекте, который необходимо защитить, что явно является неуправляемым подходом.

Чтобы избежать необходимости требовать записей управления доступом Deny (ACE) как способа запретить службам доступ к ресурсам, к которым имеет доступ учетная запись пользователя, в которой они работают, применяют два типа SID

для служб: ограниченный SID службы SERVICE_SID_TYPE_RESTRICTED и неограниченный SID службы SERVICE_SID_TYPE_UNRESTRICTED. Последний используется по умолчанию и задействовался в случаях, которые мы рассматривали до сих пор. В данном случае названия немного вводят в заблуждение. SID службы всегда генерируется одинаково (см. предыдущий эксперимент). Токен хост-процесса генерируется особым способом.

Неограниченные SID службы создаются как по умолчанию активные SID владельца группы, а токену процесса также присваивается новый ACE, который предоставляет полный доступ в отношении SID службы, что позволяет ей продолжать взаимодействовать с SCM. (Основное применение данного решения — включение или отключение SID службы внутри процесса во время запуска или завершения работы службы.) Служба, работающая с учетной записью SYSTEM, запущенной с неограниченным токеном, имеет даже больше полномочий, чем стандартная служба SYSTEM.

В то же время SID ограниченной службы превращает токен хост-процесса службы в токен с ограничением записи. Токены с ограниченным доступом (дополнительную информацию о токенах см. в главе 7 тома 1) обычно требуют, чтобы система выполняла две проверки при доступе к защищаемым объектам: одну с использованием стандартного списка активных SID группы, а другую с применением списка SID с ограниченным доступом. Для стандартного ограниченного токена доступ предоставляется только в том случае, если обе проверки доступа разрешают запрошенные права доступа. В свою очередь, токены с ограничением записи, которые обычно создаются путем указания флага `WRITE_RESTRICTED` в API `CreateRestrictedToken`, выполняют двойные проверки доступа только для запросов на запись — запросы доступа только для чтения вызывают лишь одну проверку доступа для активных групповых SID токена, подобно обычным токенам.

Хост-процесс службы, работающий с токеном с ограничением записи, может записывать только в объекты, предоставляющие SID службы явный доступ на запись (и следующим трем дополнительным SID, добавленным для совместимости) независимо от учетной записи, под которой он работает. По этой причине все службы, работающие внутри этого процесса (входящие в одну и ту же группу служб), должны иметь ограниченный SID, в противном случае службы с ограниченным SID не запустятся. Как только токен становится токеном с ограничением записи, в целях совместимости добавляются еще три SID.

- Общемировой SID добавлен, чтобы разрешить доступ для записи к объектам, которые обычно доступны любому, в первую очередь определенным DLL в пути загрузки.
- SID входа в службу добавляется, чтобы позволить службе взаимодействовать с SCM.
- SID с ограничением записи добавляется, чтобы позволить объектам явно разрешать доступ к ним любой службе с ограничением записи. Например, ETW использует этот SID в своих объектах, чтобы позволить любой службе с ограничением на запись генерировать события.

На рис. 10.11 показан пример хост-процесса, содержащего службы, помеченные как имеющие ограниченные идентификаторы SID. Например, механизм базовой фильтрации (BFE), который отвечает за применение правил фильтрации

брандмауэра Windows, входит в данный хост-процесс, поскольку эти правила хранятся в разделах реестра, которые должны быть защищены от злонамеренного доступа для записи в случае компрометации службы. (Это может позволить использовать уязвимость службы и отключить правила брандмауэра для исходящего трафика, например, включив двустороннюю связь со злоумышленником.)

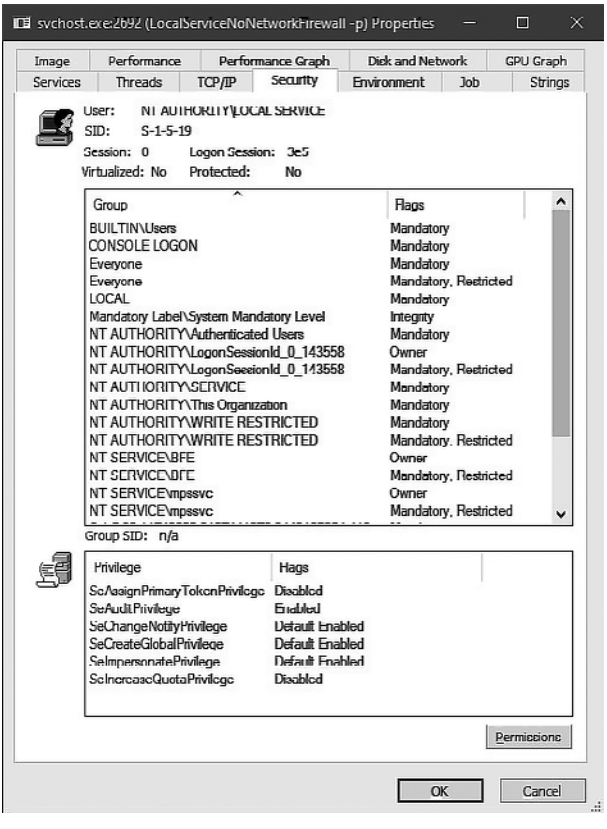


Рис. 10.11. Служба с ограниченными SID

Блокируя доступ для записи к объектам, которые в противном случае были бы доступны для записи службой (путем наследования разрешений учетной записи, под которой она работает), ограниченные SID службы решают другую часть изначально обозначенной проблемы, поскольку пользователям не нужно ничего делать, чтобы запретить службе, работающей под привилегированной учетной записью, иметь доступ на запись к критическим системным файлам, разделам реестра или другим объектам, ограничивая уязвимость к атакам любой такой службы, которая могла быть скомпрометирована.

Windows также допускает использование правил брандмауэра, которые ссылаются на SID служб, связанных с одним из трех вариантов поведения, описанных в табл. 10.11.

Таблица 10.11. Правила ограничения сети

Сценарий	Пример	Ограничения
Доступ к сети заблокирован	Служба обнаружения оборудования оболочки (ShellHWDetection)	Заблокированы все сетевые соединения, как входящие, так и исходящие
Доступ к сети статически ограничен портами	Служба RPC (Rpcss) работает на порте 135 (TCP и UDP)	Сетевые коммуникации ограничены определенными портами TCP или UDP
Доступ к сети динамически ограничен по портам	Служба DNS (Dns) прослушивает переменные порты (UDP)	Сетевые соединения ограничены настраиваемыми портами TCP или UDP

Учетная запись виртуальной службы

Как сказано в предыдущем разделе, SID службы можно установить в качестве владельца токена службы, работающей в контексте учетной записи виртуальной службы. Служба, работающая с учетной записью виртуальной службы, имеет меньше привилегий, чем типы служб `LocalService` или `NetworkService` (список привилегий см. в табл. 10.10), и у нее нет учетных данных для аутентификации через сеть. SID службы является владельцем токена, а токен входит в группы «Все», «Пользователи», «Прошедшие проверку» и «Все службы». Это означает, что служба может читать (или записывать, если только она не применяет ограниченный тип SID) объекты, принадлежащие обычным пользователям, но не пользователям с высоким уровнем привилегий, принадлежащим к группе «Администратор» или «Система». В отличие от других типов, служба, работающая под учетной записью виртуальной службы, имеет частный профиль, который загружается службой `ProfSvc` (Profsvc.dll) во время входа в нее, аналогично обычным службам (подробнее см. раздел «Вход службы в систему»). Профиль изначально создается во время первого входа в службу с использованием папки с тем же именем, что и у службы, расположенной по пути %SystemRoot%\ServiceProfiles. Когда профиль службы загружается, ее куст реестра монтируется в корневой раздел HKEY_USERS под разделом, именуемым удобочитаемым SID учетной записи виртуальной службы (начиная с S-1-5-80, как описано в эксперименте «Разбор SID службы»).

Пользователи могут легко назначить службе учетную запись виртуальной службы, установив для учетной записи входа значение NT SERVICE\<имя_службы>. Во время входа в систему диспетчер управления службами распознает, что учетная запись для входа является учетной записью виртуальной службы (благодаря поставщику входа в систему NT SERVICE) и проверяет, соответствует ли имя учетной записи имени службы. Службу нельзя запустить с использованием учетной записи виртуальной службы, принадлежащей другой, и это обеспечивается SCM (через внутреннюю функцию `ScIsValidAccountName`). Службы, задействующие общий хост-процесс, не могут работать с учетной записью виртуальной службы.

В ходе работы с защищаемыми объектами пользователи могут добавлять в список ACL объекта с помощью учетной записи входа в службу (в виде NT SERVICE\<имя_службы>), ACE, который разрешает или запрещает доступ к виртуальной службе.

Как показано на рис. 10.12, система способна преобразовать имя учетной записи виртуальной службы в соответствующий SID, тем самым устанавливая детальный контроль доступа к объекту со стороны службы. (Это верно и для обычных служб, работающих под несистемной учетной записью, как описано в предыдущем разделе.)

Интерактивные службы и изоляция нулевой сессии

Одним из ограничений для служб, работающих под соответствующей учетной записью службы, учетными записями локальной системы, локальной или сетевой службы, которые всегда присутствовали в Windows, является то, что эти службы не могли отображать диалоговые окна или окна на интерактивном рабочем столе пользователя. Это ограничение было не прямым результатом работы под этими учетными записями, а скорее следствием того, как подсистема Windows назначает служебные процессы оконным станциям. Это ограничение еще больше усиливается из-за использования сеансов в рамках модели под названием «Изоляция нулевой сессии», в результате чего службы не могут напрямую взаимодействовать с рабочим столом пользователя.

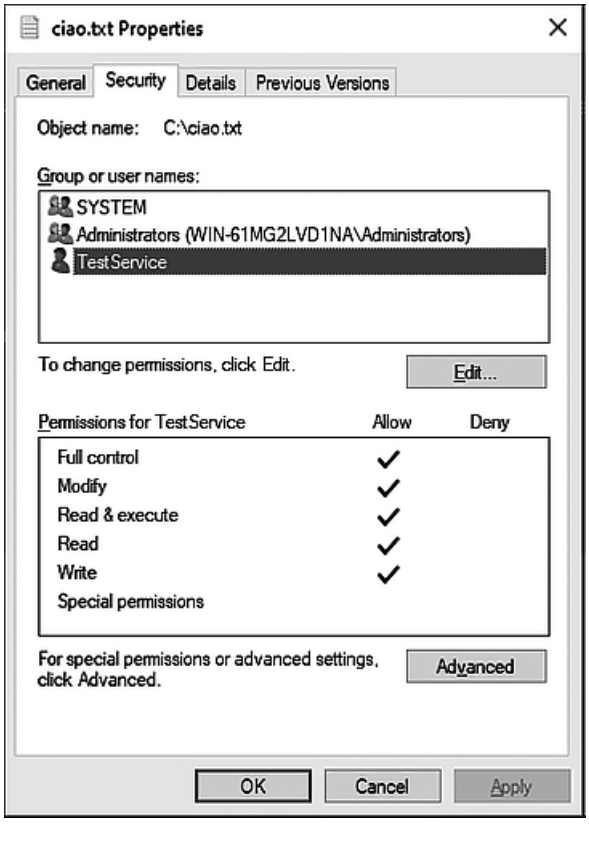


Рис. 10.12. Файл (защищаемый объект) с ACE, обеспечивающим полный доступ к TestService

Подсистема Windows связывает каждый процесс Windows с оконной станцией. Последняя включает в себя рабочие столы, а рабочие столы содержат окна. Одновременно только одна оконная станция может быть видимой и принимать данные, вводимые пользователем с помощью мыши и клавиатуры. Во время работы служб терминалов видна одна оконная станция за сеанс, но все службы выполняются как часть скрытой нулевой сессии. Windows называет видимую оконную станцию WinSta0, и все интерактивные процессы получают доступ к ней.

Если не указано иное, подсистема Windows связывает службы, работающие под соответствующей учетной записью службы или учетной записью локальной системы, с невидимой оконной станцией с именем Service-0x0-3e7$, которую используют все неинтерактивные службы. Число 3e7 в имени представляет идентификатор сеанса входа в систему, назначаемый процессом локальной управляющей системы безопасности (LSASS) сеансу входа в систему, который SCM применяет для неинтерактивных служб, работающих под учетной записью локальной системы. Аналогичным образом службы, работающие в учетной записи локальной службы, связаны с оконной станцией, созданной сеансом входа в систему 3e5, тогда как службы, работающие в учетной записи сетевой службы, связаны с оконной станцией, созданной сеансом входа в систему 3e4.

Службы, настроенные для запуска под учетной записью пользователя (то есть не под учетной записью локальной системы), запускаются на другой невидимой оконной станции, названной с помощью идентификатора входа LSASS, назначенного для сеанса входа в службу. На рис. 10.13 показан пример экрана утилиты Sysinternals WinObj, показывающий каталог диспетчера объектов, в котором Windows размещает объекты оконной станции. Видны интерактивная оконная станция (WinSta0) и три оконные станции неинтерактивных служб.

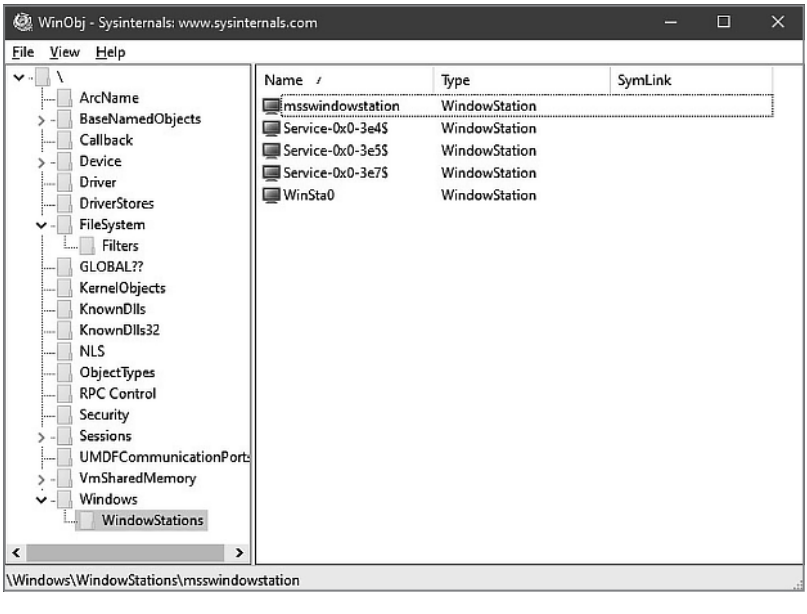


Рис. 10.13. Список оконных станций

Независимо от того, запущены ли службы под учетной записью пользователя, учетной записью локальной системы или учетной записью локальной или сетевой службы, те из них, которые не работают на станции с видимым окном, не могут получать входные данные от пользователя или отображать видимые окна. Если бы такая служба открыла модальное диалоговое окно, то оказалась бы по сути зависшей, потому что ни один пользователь не смог бы увидеть это окно, а это, естественно, помешало бы ему ввести данные с клавиатуры или с помощью мыши, чтобы то закрылось и позволило службе продолжить работу.

У службы может быть веская причина взаимодействовать с пользователем через диалоговые или обычные окна. Службы, настроенные с помощью флага `SERVICE_INTERACTIVE_PROCESS` в параметре Type раздела реестра службы, запускаются с хост-процессом, подключенным к интерактивной оконной станции WinSta0. (Обратите внимание на то, что службы, настроенные для запуска под учетной записью пользователя, не могут быть помечены как интерактивные.) Если бы пользовательские процессы запускались в одном сеансе со службами, подключение к WinSta0 позволило бы службе отображать диалоговые окна и окна и включать их, чтобы реагировать на ввод пользователя, поскольку они будут задействовать оконную станцию совместно с интерактивными службами. Однако в нулевой сессии выполняются только процессы, принадлежащие системе и службам Windows, а все остальные сеансы входа в систему, включая сеансы пользователей консоли, выполняются в разных сессиях. Таким образом, любое окно, отображаемое процессами в нулевой сессии, не видно пользователю.

Это дополнительное ограничение помогает предотвратить разрушительные атаки, при которых менее привилегированное приложение отправляет сообщения в окно, видимое на той же оконной станции, чтобы использовать ошибку в более привилегированном процессе, владеющем окном, что позволяет ему выполнять код в этом процессе. Раньше в Windows была служба обнаружения интерактивных служб (UI0Detect), которая уведомляла пользователей, когда служба отображала окно на главном рабочем столе оконной станции WinSta0 нулевой сессии. Это позволяло пользователю переключиться на окно сеанса 0 станции, обеспечивая правильную работу интерактивных служб. В целях безопасности эта функция сначала была отключена, а начиная с обновления Windows от 10 апреля 2018 года (RS4), удалена.

В результате, хотя интерактивные службы по-прежнему поддерживаются диспетчером управления службами (только путем установки для параметра реестра `HKLM\SYSTEM\CurrentControlSet\Control\Windows\NoInteractiveServices` значения `0`), доступ к нулевой сессии теперь невозможен. Ни одна служба больше не может отображать какое-либо окно (по крайней мере, без какого-либо недокументированного взлома).

Диспетчер управления службами

Исполняемый файл SCM — `%SystemRoot%\System32\Services.exe`, и, подобно большинству процессов служб, он запускается как консольная программа Windows. Процесс Wininit запускает SCM на ранней стадии загрузки системы. (Подробную информацию о процессе загрузки см. в главе 12.) Функция запуска SCM `SvcCtrlMain` организует запуск служб, настроенных для автоматического запуска.

`SvcCtrlMain` сначала выполняет собственную инициализацию, устанавливая средства защиты своего процесса и фильтр необработанных исключений,

а также создавая в памяти представление общеизвестных идентификаторов SID. Затем она создает два события синхронизации: SvcctrlStartEvent_A3752DX и SC_AutoStartComplete. Оба инициализируются как несигнализированные. Первое событие сигнализируется SCM после завершения всех шагов, необходимых для получения команд от SCP. Второе сигнализируется по завершении инициализации SCM. Событие используется для предотвращения запуска системой или другими пользователями другого экземпляра диспетчера управления службами. Функция, которую SCP применяет для установления диалога с SCM, — это OpenSCManager. Она предотвращает попытки SCP связаться с SCM до его инициализации, ожидая получения сигнала события SvcctrlStartEvent_A3752DX.

Затем SvcCtrlMain приступает к делу: создает соответствующий дескриптор безопасности и вызывает ScGenerateServiceDB — функцию, которая создает внутреннюю базу данных служб SCM. ScGenerateServiceDB считывает и сохраняет содержимое HKLM\SYSTEM\CurrentControlSet\Control\ServiceGroupOrder\List — параметр REG_MULTI_SZ, в котором перечислены имена и порядок определенных групп служб. Раздел реестра службы содержит необязательный параметр группы, если этой службе или драйверу устройства необходимо контролировать порядок запуска по отношению к службам из других групп. Например, сетевой стек Windows построен снизу вверх, поэтому сетевые службы должны указывать параметры группы, которые помещают их в последовательность запуска позже, чем драйверы сетевых устройств. SCM внутренне создает список групп, в котором сохраняется порядок групп, считываемых из реестра. Эти группы включают (но на том не ограничиваются) NDIS, TDI, основной диск, порт клавиатуры, класс клавиатуры, фильтры и т. д. Дополнительные и сторонние приложения могут даже определять собственные группы и добавлять их в этот список. Например, Microsoft Transaction Server добавляет группу с именем MS Transactions.

Затем ScGenerateServiceDB просматривает содержимое HKLM\SYSTEM\CurrentControlSet\Services, создавая запись (она называется служебной записью) в базе данных службы для каждого встреченного раздела. Запись базы данных включает в себя все соответствующие параметры, определенные для службы, а также поля, которые отслеживают состояние службы. SCM добавляет записи и для драйверов устройств, и для служб, поскольку он запускает службы и драйверы, помеченные для автозапуска, и обнаруживает сбои при запуске для драйверов, помеченных для запуска, при загрузке или запуске системы. Он также дает приложениям возможность запрашивать состояние драйверов. Диспетчер ввода-вывода загружает драйверы, помеченные для запуска при загрузке и при старте системы, до выполнения любых процессов пользовательского режима, поэтому любые драйверы, имеющие эти типы запуска, загружаются до запуска SCM.

ScGenerateServiceDB считывает параметр группы службы, чтобы определить ее членство в группе, и связывает его с записью группы в созданном ранее списке групп. Функция также считывает и записывает в базу данных группу службы и зависимости службы, запрашивая параметры реестра DependOnGroup и DependOnService. На рис. 10.14 показано, как SCM организует списки записей служб и порядка групп. Обратите внимание на то, что список служб отсортирован в алфавитном порядке. Причина этого заключается в том, что SCM создает список из раздела реестра Services, а Windows перечисляет разделы реестра в алфавитном порядке.

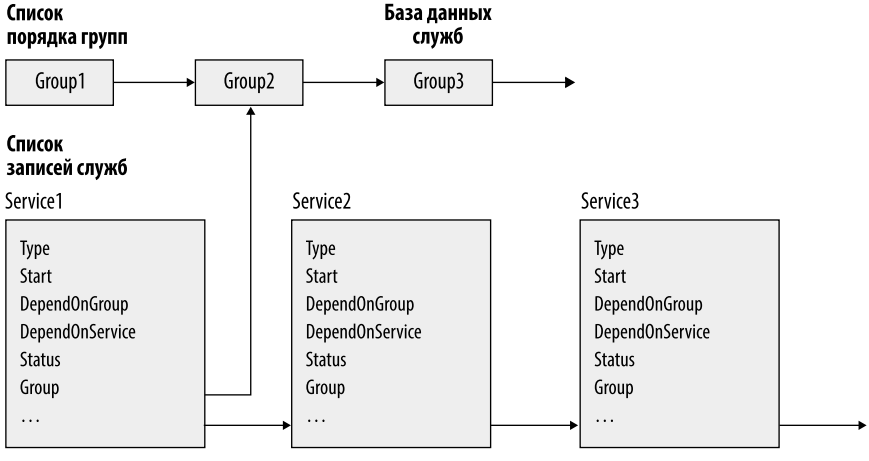


Рис. 10.14. Организация базы данных служб

Во время запуска службы SCM вызывает LSASS (например, для входа в систему службы под нелокальной системной учетной записью), а потом ожидает, пока LSASS сигнализирует о событии синхронизации `LSA_RPC_SERVER_ACTIVE`, что тот и делает после завершения инициализации. Wininit также отвечает за запуск процесса LSASS, поэтому инициализация LSASS происходит одновременно с инициализацией SCM, а порядок завершения инициализации LSASS и SCM может различаться. SCM вычищает (из реестра, вдобавок к базе данных) все службы, которые были помечены как удаленные (с помощью параметра реестра `DeleteFlag`), и создает список зависимостей для каждой записи службы в базе данных. Это позволяет SCM узнать, какая служба зависит от конкретной записи службы, что является информацией о зависимости, противоположной хранящейся в реестре.

Затем SCM запрашивает, запущена ли система в безопасном режиме (из параметра реестра HKLM\System\CurrentControlSet\Control\Safeboot\Option\OptionValue). Эта проверка необходима для того, чтобы в дальнейшем определить, следует ли запускать службу (подробности описаны в разделе «Запуск автозапускаемых служб» далее в этой главе). Далее он создает именованный конвейер удаленного вызова процедур (RPC) \Pipe\Ntsvcs, а затем RPC запускает поток для прослушивания конвейера на предмет входящих сообщений от SCP. SCM сигнализирует о событии завершения инициализации `SvcctrlStartEvent_A3752DX`. Регистрация обработчика событий завершения работы консольного приложения и регистрация в процессе подсистемы Windows через `RegisterServiceProcess` подготавливают SCM к завершению работы системы.

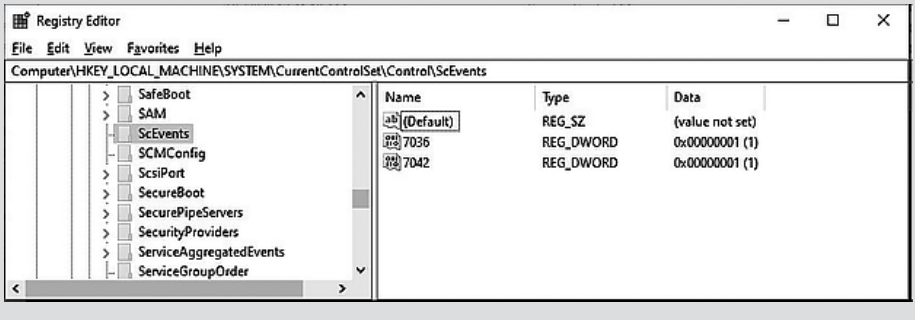
Перед запуском автозапускаемых служб SCM выполняет еще несколько шагов. Он инициализирует диспетчер драйверов UMDF, который отвечает за управление драйверами UMDF. Начиная с Windows 10 Fall Creators Update (RS3), он является частью диспетчера управления службами и ожидает полной инициализации известных DLL, ожидая события `\KnownDlls\SmKnownDllsInitialized`, о котором сигнализирует диспетчер сеансов.

ЭКСПЕРИМЕНТ. Активация журналирования служб

Диспетчер управления службами обычно регистрирует события ETW только при обнаружении аномальных ошибок, например при невозможности запуска службы или изменения ее конфигурации. Это поведение можно переопределить, вручную включив или отключив различные типы событий SCM. В этом эксперименте вы включите два типа событий, которые особенно полезны для отладки изменения состояния службы. События 7036 и 7042 возникают, когда меняется состояние службы или в службу отправляется запрос управления STOP.

Эти два события включены по умолчанию в серверных SKU, но не в клиентских выпусках Windows 10. На компьютере с Windows 10 вам следует открыть редактор реестра, введя `regedit.exe` в поле поиска Cortana, и перейти к разделу реестра HKLM\SYSTEM\CurrentControlSet\Control\ScEvents. Если последнего подраздела не существует, его нужно создать, щелкнув правой кнопкой мыши на подразделе Control и выбрав пункт Раздел (Key) в контекстном меню Создать (New).

Теперь следует создать два параметра DWORD и назвать их 7036 и 7042. Установите для обоих значение 1. (Можете установить для них значение 0, чтобы получить противоположный эффект и предотвратить создание этих событий даже на серверах SKU.) Вы должны получить состояние реестра, подобное следующему.

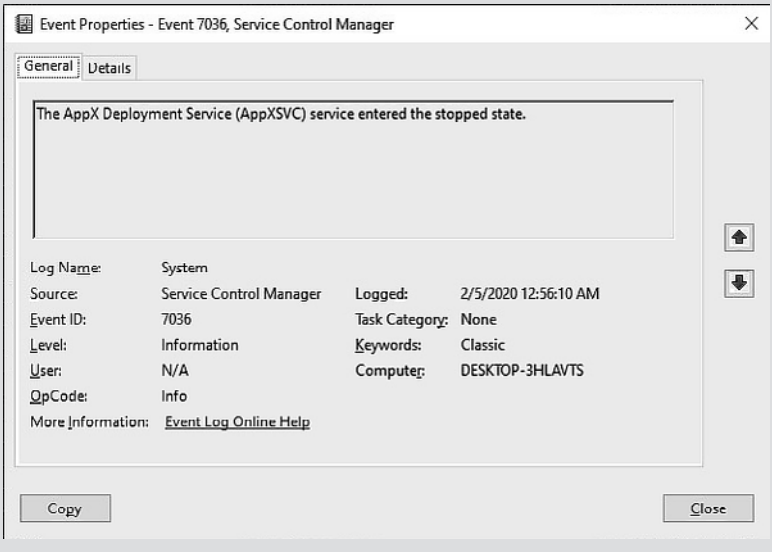


Перезагрузите рабочую станцию, а затем запустите и остановите какую-либо службу, например AppXSvc, с помощью инструмента sc.exe, открыв административную командную строку и введя следующие команды:

```
sc stop AppXSvc
sc start AppXSvc
```

Откройте Просмотр событий, введя `eventvwr` в поле поиска Cortana, и перейдите в Журналы Windows (Windows Logs), а затем в Система (System). Вы должны заметить различные события из диспетчера управления службами с идентификаторами событий 7036 и 7042. В верхних из них нужно найти событие

остановки, сгенерированное службой AppXSvc, как показано на следующем рисунке.



Обратите внимание на то, что диспетчер управления службами по умолчанию регистрирует все события, генерируемые службами, запускаемыми автоматически при запуске системы. Это может привести к нежелательному количеству событий, заполняющих журнал системных событий. Чтобы устранить эту проблему, можете отключить события автозапуска SCM, создав параметр реестра с именем `EnableAutostartEvents` в разделе HKLM\System\CurrentControlSet\Control и установив для него значение `0` (неявное значение по умолчанию — `1` как в клиентских, так и в серверных SKU). В результате будут записываться только события, генерируемые приложениями служб при запуске, приостановке или остановке целевой службы.

Буквы сетевых дисков

Помимо своей роли интерфейса для служб, SCM имеет еще одну совершенно не связанную с ней зону ответственности: он уведомляет приложения с графическим интерфейсом в системе всякий раз, когда система создает или удаляет сетевое соединение с буквой диска. SCM ожидает, пока маршрутизатор с несколькими поставщиками (MPR) сигнализирует об именованном событии `\BaseNamedObjects\ScNetDrvMsg`, о котором MPR сообщает всякий раз, когда приложение назначает букву диска удаленному сетевому ресурсу или отменяет назначение буквы диска удаленному общему ресурсу. Когда MPR сигнализирует об этом событии, SCM вызывает функцию Windows `GetDriveType`, чтобы запросить список букв подключенных сетевых дисков. Если список изменяется в сигнале события, SCM отправляет широковещательное сообщение Windows типа `WM_DEVICECHANGE`. SCM использует в качестве

подтипа сообщения либо DBT_DEVICEREMOVECOMPLETE, либо DBT_DEVICEARRIVAL. Это сообщение в первую очередь предназначено для проводника Windows, чтобы он мог обновлять любые открытые окна компьютера, показывая наличие или отсутствие буквы сетевого диска.

Программы управления службами

Как описано в разделе «Приложения служб», программы управления службами (SCP) — это стандартные приложения Windows, которые используют функции управления службами SCM, включая CreateService, OpenService, StartService, ControlService, QueryServiceStatus и DeleteService. Чтобы использовать функции SCM, SCP сначала должен открыть канал связи с SCM, вызвав функцию OpenSCManager, чтобы указать, какие типы действий он хочет выполнить. Например, если SCP просто хочет перечислить и отобразить службы, присутствующие в базе данных SCM, он запрашивает доступ к перечислению служб при вызове OpenSCManager. Во время своей инициализации SCM создает внутренний объект, представляющий базу данных SCM, и использует функции безопасности Windows для защиты объекта с помощью дескриптора безопасности, который определяет, какие учетные записи могут открывать объект и с какими разрешениями доступа. Например, дескриптор безопасности указывает, что группа пользователей «Прошедшие проверку» может открыть объект SCM с доступом к функции перечисления служб. Однако открыть объект с доступом, необходимым для создания или удаления службы, могут только администраторы.

Как и в случае с базой данных SCM, SCM обеспечивает безопасность самих служб. Когда SCP создает службу с помощью функции CreateService, он указывает дескриптор безопасности, который SCM неявно связывает с записью службы в базе данных службы. SCM сохраняет дескриптор безопасности в разделе реестра службы как значение Security и считывает его при сканировании раздела Services реестра во время инициализации, чтобы настройки безопасности сохранялись при перезагрузках. Точно так же, как SCP должен указать, какие типы доступа к базе данных SCM ему нужны при вызове OpenSCManager, SCP должен сообщить SCM, какой доступ к службе ему требуется при вызове OpenService. Виды доступа, которые SCP может запросить, включают возможность запрашивать состояние службы, а также настраивать, останавливать и запускать ее.

SCP, с которым вы, вероятно, знакомы лучше всего, — это оснастка MMC «Службы», входящая в состав Windows и расположенная в %SystemRoot%\System32\Filemgmt.dll. Windows также включает Sc.exe (инструмент Service Controller) — программу управления службами из командной строки, о которой мы неоднократно упоминали.

SCP иногда накладывают политику служб поверх того, что реализует SCM. Хорошим примером является тайм-аут, который реализует оснастка «Службы» MMC, когда служба запускается вручную. Оснастка предоставляет индикатор выполнения, отображающий ход запуска службы. Службы косвенно взаимодействуют с SCP, устанавливая статус их конфигурации, чтобы отражать их прогресс при ответе на команды SCM, такие как команда запуска. SCP запрашивают статус с помощью функции QueryServiceStatus. Они могут определить, когда служба активно обновляет статус и когда она зависает, и SCM может предпринять соответствующие действия для уведомления пользователя о том, что делает служба.

Запуск служб

`SvcCtrlMain` вызывает функцию SCM `ScAutoStartServices` для запуска всех служб, имеющих значение Start, обозначающее автозапуск (кроме отложенного автозапуска и пользовательских служб). `ScAutoStartServices` также запускает автозапускаемые драйверы. Во избежание путаницы следует исходить из того, что термин «службы» означает службы и драйверы, если не указано иное. `ScAutoStartServices` начинает с запуска двух важных базовых служб: Plug and Play (реализованной в библиотеке Umpnpmgr.dll) и Power (реализованной в библиотеке Umpo.dll), которые необходимы системе для управления оборудованием Plug-and-Play и интерфейсами питания. Затем SCM регистрирует свое состояние автозапуска WNF, которое используется для указания текущей фазы автозапуска для Power и других служб.

Прежде чем начнется запуск других служб, процедура `ScAutoStartService` вызывает `ScGetBootAndSystemDriverState` для просмотра базы данных служб в поисках записей драйверов устройств при загрузке и запуске системы. `ScGetBootAndSystemDriverState` определяет, успешно ли запустился драйвер с типом запуска «Запуск при загрузке» или «Запуск при старте системы», путем поиска его имени в каталоге пространства имен диспетчера объектов с именем `\Driver`. Когда драйвер устройства успешно загружается, диспетчер ввода-вывода вставляет объект драйвера в пространство имен этого каталога, поэтому, если его имя отсутствует, он не загрузился. На рис. 10.15 показано, как WinObj отображает содержимое каталога Driver. `ScGetBootAndSystemDriverState` вносит имена драйверов, которые не запустились и являются частью текущего профиля, в список `ScStoppedDrivers`. Он будет использоваться позже, в конце инициализации SCM, для регистрации события в журнале системных событий (ID 7036), содержащем список загрузочных драйверов, которые не удалось запустить.

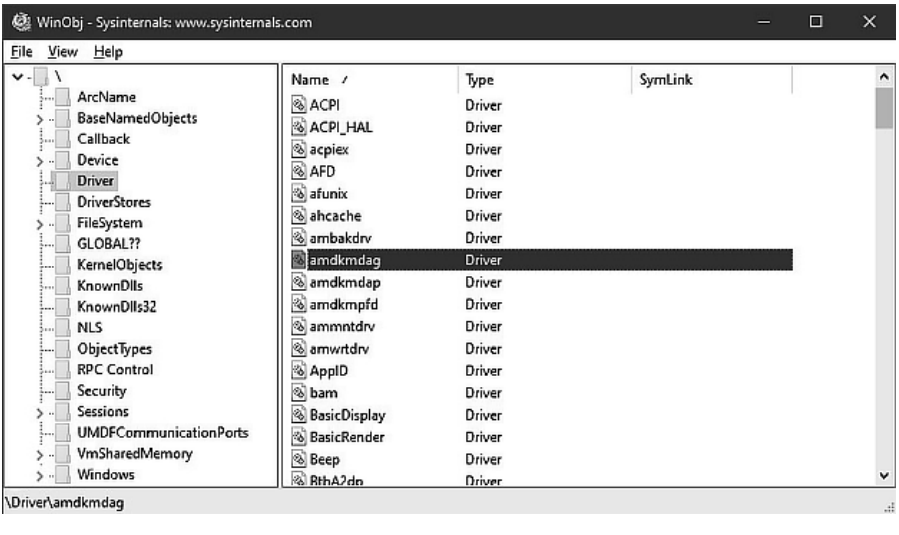


Рис. 10.15. Список объектов драйверов

Алгоритм `ScAutoStartServices` для запуска служб в правильном порядке выполняется поэтапно, при этом этап соответствует группе, а этапы выполняются

в последовательности, определяемой порядком групп, хранящимся в параметре реестра HKLM\SYSTEM\CurrentControlSet\Control\ServiceGroupOrder\List. Параметр `List`, показанный на рис. 10.16, включает имена групп в том порядке, в котором SCM должен их запускать. Таким образом, назначение службы группе не имеет никакого эффекта, кроме тонкой настройки ее запуска по отношению к другим службам, принадлежащим к другим группам.

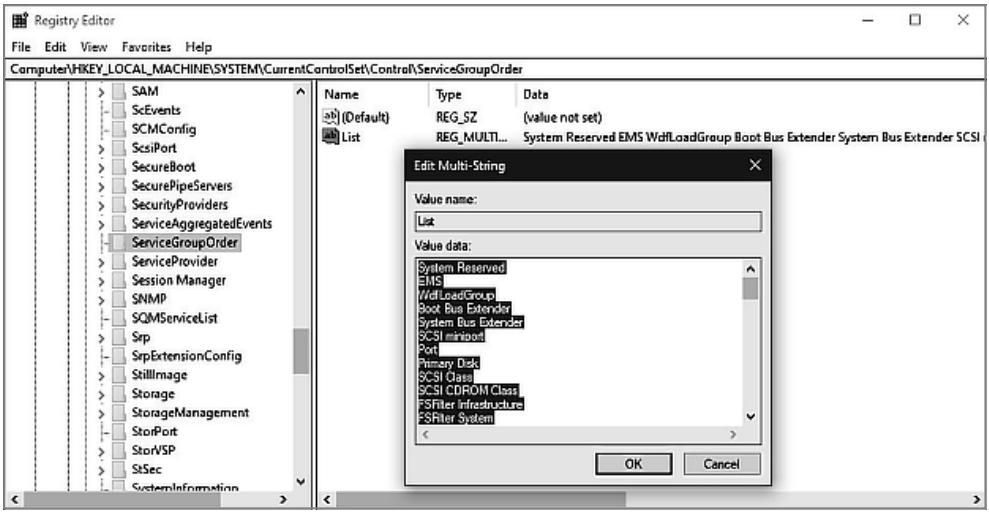


Рис. 10.16. Раздел реестра ServiceGroupOrder

При запуске фазы `ScAutoStartServices` помечает все записи служб, принадлежащие группе фазы, для запуска. Затем `ScAutoStartServices` просматривает отмеченные службы, чтобы проверить, может ли она запустить каждую из них. Сюда входит проверка того, помечена служба как служба отложенного автозапуска или как служба с шаблоном пользователя. В обоих случаях SCM запустит ее на более позднем этапе. (Службы отложенного автозапуска также необходимо вывести из групп. Пользовательские службы обсуждаются позже в одноименном разделе.) Другая часть выполняемой проверки состоит в определении того, зависит ли служба от другой группы, как указано в существовании параметра `DependOnGroup` в разделе реестра службы. Если зависимость существует, группа, от которой зависит служба, должна быть уже инициализирована и по крайней мере одна служба этой группы — успешно запущена. Если служба зависит от группы, которая запускается позже группы службы в последовательности запуска группы, SCM отмечает ошибку циклической зависимости для службы. Если `ScAutoStartServices` рассматривает службу Windows или драйвер устройства автозапуска, он затем проверяет, зависит ли служба от одной или нескольких других служб, и если да — определяет, запущены ли уже эти службы. Зависимости службы обозначаются параметром реестра `DependOnService` в разделе реестра службы. Если служба зависит от других служб, принадлежащих к группам, которые указаны позже в `ServiceGroupOrder\List`, SCM также генерирует ошибку циклической зависимости и не запускает службу. Если служба зависит от каких-либо еще не запущенных служб из той же группы, она пропускается.

Когда зависимости службы удовлетворены, `ScAutoStartServices` перед запуском службы выполняет окончательную проверку того, является ли служба частью текущей конфигурации загрузки. Когда система загружается в безопасном режиме, SCM гарантирует, что служба идентифицирована по имени или по группе в соответствующем разделе реестра безопасной загрузки. В разделе HKLM\SYSTEM\CurrentControlSet\Control\SafeBoot есть два раздела безопасной загрузки — Minimal и Network, и тот, который проверяет SCM, зависит от того, в каком из безопасных режимов загрузился пользователь. Если он выбрал безопасный режим или безопасный режим с поддержкой командной строки в современном или устаревшем меню загрузки, SCM ссылается на раздел Minimal, если безопасный режим с поддержкой сети — SCM относится к Network. Наличие строкового параметра с именем `Option` под разделом SafeBoot указывает не только на то, что система загрузилась в безопасном режиме, но и на тип безопасного режима, выбранный пользователем. Дополнительную информацию о безопасной загрузке см. в разделе «Безопасная загрузка» главы 12.

Запуск службы

Как только SCM решает запустить службу, он вызывает `StartInternal`, который выполняет для служб шаги, отличные от таковых для драйверов устройств. Когда `StartInternal` запускает службу Windows, он сначала определяет имя файла, запускающего процесс службы, считывая параметр `ImagePath` из раздела реестра службы. Если служебный файл соответствует LSASS.exe, SCM инициализирует конвейер управления, подключается к уже запущенному процессу LSASS и ожидает его ответа. Когда конвейер готов, процесс LSASS подключается к SCM, вызывая классическую процедуру `StartServiceCtrlDispatcher`. Как показано на рис. 10.17, некоторые службы, такие как Credential Manager или Encrypting File System, должны взаимодействовать со службой подсистемы локальной управляющей безопасности (LSASS) — обычно для выполнения операций шифрования для локальных системных политик, например паролей, привилегий и аудита безопасности (подробнее см. в главе 7 тома 1).

Затем SCM определяет, является служба критической (анализируя параметры реестра FailureAction) или работает под управлением WoW64. (Если служба 32-разрядная, SCM должен применить перенаправление файловой системы. Дополнительные сведения см. в разделе «WoW64» главы 8.) Он также проверяет параметр `Type` службы. SCM инициирует поиск во внутренней базе данных записей образов, если применяются следующие условия:

- параметр `Type` службы содержит значение `SERVICE_WINDOWS_SHARE_PROCESS` (0x20);
- служба не была перезапущена после ошибки;
- разделение службы Svchost не разрешено для службы (дополнительную информацию см. в разделе «Разделение Svchost по службам» далее в этой главе).

Запись образа — это структура данных, которая представляет запущенный процесс, где размещена хотя бы одна служба. Если предыдущие условия применимы, SCM ищет запись образа, имеющую то же имя исполняемого файла процесса, что и новый параметр `ImagePath` службы.

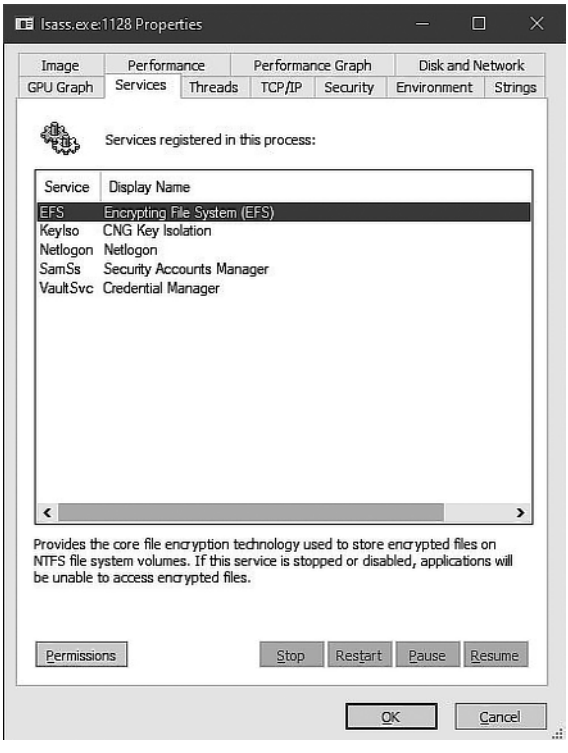


Рис. 10.17. Службы, размещенные в процессе LSASS

Если SCM обнаружит существующую запись базы данных образов с соответствующими данными `ImagePath`, службу можно использовать совместно и один из хост-процессов уже запущен. SCM проверяет, что вход в систему найденного хост-процесса совершен с помощью той же учетной записи, которая указана для запускаемой службы. (Это необходимо для того, чтобы служба была настроена не с неправильной учетной записью, допустим `LocalService`, а с путем к образу, указывающим на работающий Svchost, например netsvcs, который работает как `LocalSystem`.) Параметр реестра `ObjectName` службы хранит учетную запись пользователя, под которой должна работать служба. Служба без `ObjectName` или `ObjectName` с параметром `LocalSystem` запускается под учетной записью локальной системы. Процесс может войти только под одной учетной записью, поэтому SCM сообщает об ошибке, когда служба указывает имя учетной записи, отличное от имени другой службы, которая уже запущена в том же процессе.

Если запись образа существует, то перед запуском новой службы необходимо провести еще одну окончательную проверку: SCM открывает токен выполняющегося хост-процесса и проверяет, находится ли в токене необходимый SID службы (и включены ли все необходимые привилегии). Даже в этом случае SCM сообщает об ошибке, если условие не выполнено. Обратите внимание: как мы описываем в разделе «Вход службы в систему», для общих служб все идентификаторы SID размещенных служб добавляются во время создания токена. Ни один компонент пользовательского режима не может добавлять идентификаторы SID групп в токен после того, как он создан.

Если в базе данных образов нет записи для нового параметра ImagePath службы, SCM создает ее. Когда SCM создает новую запись, он сохраняет имя учетной записи, используемое для службы, и данные из параметра ImagePath службы. SCM требует, чтобы службы имели параметр ImagePath. Если у службы нет параметра ImagePath, SCM сообщает об ошибке, говоря, что не смог найти путь к службе и не может ее запустить. После того как SCM создает запись образа, он входит в учетную запись службы и запускает новый хост-процесс. (Процедура описана далее, в разделе «Вход службы в систему».)

После входа в систему и правильного запуска хост-процесса SCM ожидает от службы начального сообщения о подключении. Служба подключается к SCM благодаря конвейеру SCM RPC (\Pipe\Ntsvcs, как описано в разделе «Диспетчер управления службами») и структуре данных контекста канала, созданной процедурой LogonAndStartImage. Когда SCM получает первое сообщение, он приступает к запуску службы, отправляя управляющее сообщение SERVICE_CONTROL_START процессу службы. Обратите внимание на то, что в описанном протоколе связи всегда указана служба, которая подключается к SCM.

Приложение службы может обрабатывать сообщение благодаря циклу обработки сообщений, расположенному в API StartServiceCtrlDispatcher (более подробную информацию см. в разделе «Приложения служб» ранее в этой главе). Приложение службы включает SID группы служб в своем токене (при необходимости) и создает новый поток службы, который будет выполнять функцию Main службы. Затем он выполняет обратный вызов SCM для создания дескриптора новой службы и сохранения его во внутренней структуре данных (INTERNAL_DISPATCH_TABLE), аналогичной таблице службы, указанной в качестве входных данных для API StartServiceCtrlDispatcher. Структура данных используется для отслеживания активных служб в хост-процессе. Если служба не отвечает положительно на команду запуска в течение периода ожидания, SCM сдается и отмечает в системном журнале событий ошибку, которая указывает на то, что службу не удалось запустить своевременно.

Если служба, которую SCM запускает с вызова StartInternal, имеет параметр реестра Type со значением SERVICE_KERNEL_DRIVER или SERVICE_FILE_SYSTEM_DRIVER, служба на самом деле является драйвером устройства, поэтому StartInternal включает привилегию безопасности загрузки драйвера для процесса SCM, а затем вызывает службу ядра NtLoadDriver, передавая в данных параметра ImagePath раздела реестра драйвера. В отличие от служб, драйверам не нужно указывать параметр ImagePath, и если он отсутствует, SCM создает путь к образу, добавляя имя драйвера к строке %SystemRoot%\System32\Drivers\.

ПРИМЕЧАНИЕ Драйвер устройства со стартовым значением SERVICE_AUTO_START или SERVICE_DEMAND_START запускается SCM в качестве драйвера времени выполнения, что означает: загруженный образ использует общие страницы и имеет контрольную область, которая их описывает. Это отличает его от драйверов со стартовым значением SERVICE_BOOT_START или SERVICE_SYSTEM_START, которые загружаются загрузчиком Windows и запускаются диспетчером ввода-вывода. Все эти драйверы задействуют частные страницы, не могут использоваться совместно и не имеют связанной области управления. (Более подробную информацию можно найти в главе 5 тома 1.)

ScAutoStartServices продолжает циклически перебирать службы, принадлежащие группе, до тех пор пока все службы не запустятся или не сгенерируют ошибки

зависимостей. Этот цикл — способ, которым SCM автоматически упорядочивает службы внутри группы в соответствии с их зависимостями DependOnService. SCM запускает службы, от которых зависят другие службы, в первых циклах, пропуская зависимые службы до последующих циклов. Обратите внимание на то, что SCM игнорирует значения Tag для служб Windows, которые вы можете встретить в подразделах раздела HKLM\SYSTEM\CurrentControlSet\Services. Диспетчер ввода-вывода учитывает значения Tag, чтобы упорядочить запуск драйверов устройств в группе для драйверов, запускаемых при загрузке и запуске системы. После того как SCM завершает этапы для всех групп, перечисленных в параметре ServiceGroupOrder\List, он выполняет этап для служб, принадлежащих группам, не указанным в этом параметре, а затем выполняет заключительный этап для служб без группы.

После обработки служб автозапуска SCM вызывает ScInitDelayStart, который ставит в очередь отложенный рабочий элемент, связанный с рабочим потоком, ответственным за обработку всех служб, которые ScAutoStartServices пропустил, поскольку они были помечены как отложенный автозапуск (с помощью параметра реестра DelayedAutostart). Этот рабочий поток будет выполнен после задержки. По умолчанию задержка составляет 120 с, но ее можно переопределить, создав параметр AutoStartDelay в HKLM\SYSTEM\CurrentControlSet\Control. SCM выполняет те же действия, что и при запуске служб без отложенного автозапуска.

Когда SCM завершит запуск всех служб и драйверов автозапуска, а также настройку отложенного рабочего элемента автозапуска, SCM сигнализирует о событии \BaseNamedObjects\SC_AutoStartComplete. Оно используется программой установки Windows для оценки хода запуска во время установки.

Вход службы в систему

Если во время процедуры запуска SCM не находит ни одной существующей записи образа, это означает, что необходимо создать хост-процесс. Действительно, новая служба не может применяться совместно, она выполняется первой, была перезапущена или является пользовательской службой. Перед запуском процесса SCM должен создать токен доступа для процесса узла службы. Цель функции LogonAndStartImage — создать токен и запустить хост-процесс службы. Процедура зависит от типа службы, которая будет запущена.

Пользовательские службы (точнее, их экземпляры) запускаются путем получения токена вошедшего в систему пользователя (с помощью функций, реализованных в библиотеке UserMgr.dll). В этом случае функция LogonAndStartImage дублирует токен пользователя и добавляет атрибут безопасности WIN://ScmUserService (его значение обычно установлено равным 0). Этот атрибут задействуется главным образом диспетчером управления службами при получении запросов на подключение от службы. Хотя SCM может распознать процесс, в котором размещается классическая служба, по SID службы (или SID системной учетной записи, если служба выполняется под локальной системной учетной записью), он использует атрибут безопасности SCM для идентификации процесса, в котором размещается пользовательская служба.

Для всех других типов служб SCM считывает учетную запись, под которой служба будет запущена, из реестра (из параметра ObjectName) и вызывает ScCreateServiceSids с целью создать SID службы для каждой службы, которая будет размещена в новом процессе. (SCM циклически переключается между каждой

службой в своей внутренней базе данных служб.) Обратите внимание: если служба запускается под учетной записью `LocalSystem` (без ограниченного или неограниченного SID), этот шаг не выполняется.

SCM регистрируется в службах, которые не запускаются под учетной записью System, вызывая функцию `LSASS LogonUserExEx`. Обычно она требует пароля, но чаще всего SCM указывает LSASS, что пароль хранится как секрет LSASS службы под разделом HKLM\SECURITY\Policy\Secrets в реестре. (Имейте в виду, что содержимое SECURITY обычно не отображается, поскольку его настройки безопасности по умолчанию разрешают доступ только из системной учетной записи.) Когда SCM вызывает `LogonUserExEx`, он указывает вход в службу в качестве типа входа, поэтому LSASS ищет пароль в подразделе `Secrets`, имя которого имеет вид `_SC_<имя_службы>`.

ПРИМЕЧАНИЕ Службам, работающим с учетной записью виртуальной службы, не требуется пароль для создания токена службы службой LSA. Для этих служб SCM не предоставляет пароль для API LogonUserExEx.

SCM предписывает LSASS хранить пароль для входа в систему в качестве секрета с помощью функции `LsaStorePrivateData`, когда SCP настраивает информацию для входа службы. При успешном входе в систему `LogonUserEx` возвращает дескриптор токена доступа вызывающей стороне. SCM добавляет необходимые идентификаторы безопасности службы в возвращенный токен и, если новая служба использует ограниченные SID, вызывает функцию `ScMakeServiceTokenWriteRestricted`, которая преобразует токен в токен с ограниченным доступом к записи (добавлением соответствующих ограниченных идентификаторов SID). Windows применяет токены доступа для представления контекста безопасности пользователя, а SCM позже связывает токен доступа с процессом, реализующим службу.

Затем SCM создает блок пользовательской среды и дескриптор безопасности для связи с новым процессом обслуживания. Если служба, которая будет запущена, является пакетной службой, SCM считывает из реестра всю информацию о пакете (его полное имя, источник и идентификатор модели пользователя приложения) и вызывает службу Appinfo, которая помечает токен необходимыми атрибутами безопасности `AppModel` и подготавливает процесс обслуживания для активации современного пакета. (Более подробную информацию об `AppModel` см. в разделе «Пакетные приложения» главы 8.)

После успешного входа в систему SCM загружает информацию профиля учетной записи, если она еще не загружена, вызывая функцию `LoadProfileBasic` библиотеки User Profile Basic Api DLL (`%SystemRoot%\System32\Profapi.dll`). Параметр HKLM\SOFTWARE\Microsoft\Windows NT\CurrentVersion\ProfileList\<раздел профиля пользователя>\ProfileImagePath содержит путь на диске к кусту реестра, который `LoadUserProfile` загружает в реестр, делая информацию в кусте разделом HKEY_CURRENT_USER для службы.

На следующем шаге `LogonAndStartImage` приступает к запуску процесса службы. SCM запускает процесс в приостановленном состоянии с помощью функции Windows `CreateProcessAsUser`. (За исключением процессов, размещающих службы под локальной системной учетной записью, которые создаются с помощью стандартного API `CreateProcess`. SCM уже работает с токеном SYSTEM, поэтому нет необходимости в каком-либо другом входе в систему.)

Прежде чем процесс возобновится, SCM создает структуру данных связи, которая позволяет приложению службы и SCM взаимодействовать через асинхронные

RPC. Структура данных содержит управляющую последовательность, указатель на буфер управления и ответа, данные хост-процесса и службы (например, PID, SID службы и т. д.), событие синхронизации и указатель на асинхронное состояние RPC.

SCM возобновляет процесс обслуживания через функцию `ResumeThread` и ожидает подключения службы к ее конвейеру SCM. Если он существует, параметр реестра `HKLM\SYSTEM\CurrentControlSet\Control\ServicesPipeTimeout` определяет продолжительность времени, в течение которого SCM ожидает, пока служба вызовет `StartServiceCtrlDispatcher` и подключится, прежде чем она сдастся, завершит процесс и придет к выводу, что служба не запустилась. (Обратите внимание на то, что в этом случае SCM завершает процесс, в отличие от случая, когда служба не отвечает на запрос запуска, описанный ранее в разделе «Запуск службы».) Если `ServicesPipeTimeout` не существует, SCM использует тайм-аут по умолчанию, равный 30 с. SCM использует одно и то же значение тайм-аута для всех коммуникаций служб.

Службы с отложенным автозапуском

Службы с отложенным автозапуском позволяют Windows справляться с растущим числом служб, которые запускаются при входе пользователя в систему, что замедляет процесс загрузки и увеличивает время до того, как пользователь сможет взаимодействовать с рабочим столом. Дизайн служб автозапуска в первую очередь предназначался для служб, необходимых на ранних этапах процесса загрузки, поскольку от них зависят другие службы. Хорошим примером является служба RPC, от которой зависят все остальные службы. Другое их назначение заключалось в том, чтобы разрешить автоматический запуск службы, такой как служба обновления Windows. Поскольку многие службы автозапуска попадают во вторую категорию, пометка их как служб с отложенным автозапуском позволяет критически важным службам быстрее запуститься и подготовить рабочий стол пользователя, когда он входит в систему сразу после загрузки. Кроме того, эти службы работают в фоновом режиме, что снижает приоритет их потоков, операций ввода-вывода и памяти. Для настройки службы с отложенным автозапуском требуется вызов API `ChangeServiceConfig2`. Вы можете проверить состояние флага службы, используя параметр `qc` файла sc.exe.

ПРИМЕЧАНИЕ Если служба с неотложенным автозапуском имеет в качестве одной из зависимостей службу с отложенным автозапуском, то флаг отложенного автозапуска игнорируется и служба запускается немедленно, чтобы удовлетворить зависимость.

Службы с запуском по триггеру

Некоторые службы необходимо запускать по требованию после возникновения определенных системных событий. По этой причине в Windows 7 появилась концепция службы запуска по триггеру. Программа управления службами может использовать API `ChangeServiceConfig2` (путем указания уровня информации `SERVICE_CONFIG_TRIGGER_INFO`) для настройки запуска или остановки службы по требованию после возникновения одного или нескольких системных событий. Примеры системных событий.

- К системе подключен определенный интерфейс устройства.
- Компьютер присоединяется к домену или покидает его.

- Порт TCP/IP открыт или закрыт в системном брандмауэре.
- Политика компьютера или пользователя была изменена.
- IP-адрес в сетевом стеке TCP/IP становится доступным или недоступным.
- Запрос RPC или пакет именованного конвейера поступает на определенный интерфейс.
- В системе создано событие ETW.

Вначале реализация служб триггерного запуска основывалась на Unified Background Process Manager (подробности в следующем разделе). В Windows 8.1 появилась инфраструктура брокеров, основная цель которой заключалась в управлении несколькими системными событиями, предназначенными для современных приложений. Таким образом, всеми ранее перечисленными событиями начали управлять в основном три брокера, которые являются частями инфраструктуры брокеров (за исключением агрегации событий): брокер активности рабочего стола, брокер системных событий и агрегация событий. Дополнительную информацию об инфраструктуре брокеров можно найти в разделе «Пакетные приложения» главы 8.

После завершения первой фазы `ScAutoStartServices` (которая обычно запускает критически важные службы, перечисленные в параметре реестра HKLM\SYSTEM\CurrentControlSet\Control\EarlyStartServices) SCM вызывает `ScRegisterServicesForTriggerAction` — функцию, отвечающую за регистрацию триггеров для каждой службы, запускаемой по триггеру. Процедура проходит по каждой из служб Win32, расположенных в базе данных SCM. Для каждой службы функция генерирует временное имя состояния WNF (с помощью нативного API `NtCreateWnfStateName`), защищенное соответствующим дескриптором безопасности, и публикует его со статусом службы, хранящимся в виде данных о состоянии. (Архитектура WNF описана в разделе «Средство уведомлений Windows» главы 8.) Это имя состояния WNF используется для публикации изменений состояния служб. Затем процедура запрашивает все триггеры службы из раздела реестра `TriggerInfo`, проверяя их достоверность и отключаясь, если они недоступны.

ПРИМЕЧАНИЕ Список поддерживаемых триггеров вместе с их параметрами задокументирован по адресу https://docs.microsoft.com/en-us/windows/win32/api/winsvc/ns-winsvc-service_trigger.

Если проверка прошла успешно, для каждого триггера SCM создает внутреннюю структуру данных, содержащую всю информацию о нем (например, имя целевой службы, SID, имя брокера и параметры триггера), и определяет правильный брокер на основе типа триггера: события внешних устройств управляются брокером системных событий, тогда как все остальные типы событий — брокером активности рабочего стола. На этом этапе SCM может вызвать соответствующую процедуру регистрации брокера. Процесс регистрации является частным и зависит от брокера: для каждого триггера и условия создается несколько частных имен состояний WNF, специфичных для брокера.

Брокер агрегации событий — это связующее звено между частными именами состояний WNF, опубликованными двумя брокерами, и диспетчером управления

службами. Он подписывается на все имена состояний WNF, соответствующие триггерам и условиям, с помощью API `RtlSubscribeWnfStateChangeNotification`. Когда передано достаточное количество имен состояний WNF, агрегация событий вызывает обратно SCM, который может запустить или остановить службу с запуском по триггеру.

В отличие от имен состояний WNF, используемых для каждого триггера, SCM всегда сама публикует имя состояния WNF для каждой службы Win32 независимо от того, зарегистрировала ли служба какие-то триггеры. Это связано с тем, что SCP может получать уведомление при изменении указанного статуса службы, вызывая API `NotifyServiceStatusChange`, который подписывается на имя состояния WNF статуса службы. Каждый раз, когда SCM вызывает событие, которое меняет статус службы, он публикует новые данные о состоянии в WNF «Изменение статуса службы», что активирует поток, выполняющий функцию обратного вызова изменения статуса в SCP.

Ошибки, возникающие при запуске

Если драйвер или служба сообщает об ошибке в ответ на команду запуска SCM, параметр `ErrorControl` раздела реестра службы определяет реакцию SCM. Если параметр `ErrorControl` равно `SERVICE_ERROR_IGNORE` (0) или не указано, SCM просто игнорирует ошибку и продолжает обработку запусков службы. Если параметр `ErrorControl` равен `SERVICE_ERROR_NORMAL` (1), SCM записывает в системный журнал событий событие, в котором говорится: «Службу *<имя службы>* не удалось запустить из-за следующей ошибки». SCM включает в запись журнала событий текстовое представление кода ошибки Windows, который служба вернула SCM, в качестве причины сбоя запуска. На рис. 10.18 показана запись журнала событий, сообщающая об ошибке запуска службы.

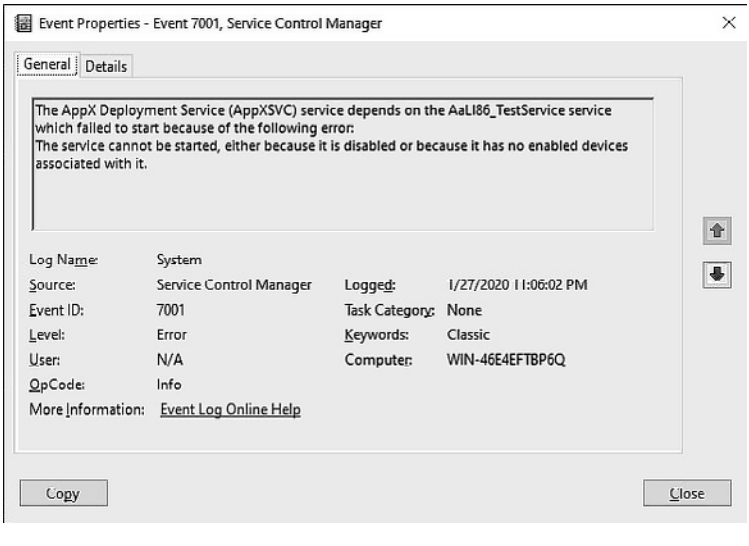


Рис. 10.18. Запись в журнале событий о сбое при запуске службы

Если служба с параметром ErrorControl, равным SERVICE_ERROR_SEVERE (2) или SERVICE_ERROR_CRITICAL (3), сообщает об ошибке запуска, SCM вносит запись в журнал событий, а затем вызывает внутреннюю функцию ScRevertToLastKnownGood. Она проверяет, включен ли функционал восстановления последнего известного исправного состояния, и, если да, переключает конфигурацию реестра системы на версию, указанную последней известной исправной, с которой система в последний раз успешно загружалась. Затем SCM перезапускает систему с помощью системной службы NtShutdownSystem, реализованной в исполнительной системе. Если система уже загружается с последней удачной конфигурацией или функционал восстановления последней удачной конфигурации не включен, SCM не делает ничего, кроме регистрации события в журнале.

Признание загрузки и последняя удачная конфигурация

Помимо запуска служб, система поручает SCM определить, когда конфигурация системного реестра HKLM\SYSTEM\CurrentControlSet должна быть сохранена в качестве последней удачной конфигурации. Раздел CurrentControlSet содержит подраздел Services, поэтому CurrentControlSet включает представление реестра базы данных SCM. Он также содержит раздел Control, в котором хранятся многие параметры конфигурации подсистемы режима ядра и пользовательского режима. По умолчанию успешная загрузка состоит из успешного запуска служб автозапуска и успешного входа пользователя в систему. Загрузка завершается сбоем, если система останавливается из-за того, что драйвер устройства приводит к сбою системы во время загрузки или если служба автозапуска с параметром ErrorControl, равным SERVICE_ERROR_SEVERE или SERVICE_ERROR_CRITICAL, сообщает об ошибке запуска.

Функционал восстановления последней удачной конфигурации обычно отключен в клиентской версии Windows. Его можно включить, установив для параметра реестра HKLM\SYSTEM\CurrentControlSet\Control\Session Manager\Configuration Manager\LastKnownGood\Enabled значение 1. В серверных версиях Windows этот параметр включен по умолчанию.

SCM знает, когда он завершил успешный запуск автозапускаемых служб, но Winlogon (%SystemRoot%\System32\Winlogon.exe) должен уведомить его об успешном входе в систему. Winlogon вызывает функцию NotifyBootConfigStatus, когда пользователь входит в систему, а NotifyBootConfigStatus отправляет сообщение в SCM. После успешного запуска автозапускаемых служб или получения сообщения от NotifyBootConfigStatus (в зависимости от того, что наступит позже), если включен функционал восстановления последнего известного исправного состояния, SCM вызывает системную функцию NtInitializeRegistry, чтобы сохранить текущую конфигурацию запуска реестра.

Сторонние разработчики программного обеспечения могут заменить определение успешного входа в систему Winlogon собственным определением. Например, система, на которой работает Microsoft SQL Server, может не считать загрузку успешной до тех пор, пока SQL Server не сможет принимать и обрабатывать транзакции. Разработчики навязывают свое определение успешной загрузки, создавая программу проверки загрузки, устанавливая ее и указывая ее расположение на диске с параметром, хранящимся в разделе реестра HKLM\SYSTEM\CurrentControlSet\Control\BootVerificationProgram. Кроме того, при установке программы проверки загрузки необходимо отключить вызов Winlogon к NotifyBootConfigStatus, задав для HKLM\

SOFTWARE\Microsoft\Windows NT\CurrentVersion\Winlogon\ReportBootOk значение 0. Когда программа проверки загрузки установлена, SCM запускает ее после завершения обработки автозапускаемых служб и ожидает вызова программы NotifyBootConfigStatus перед сохранением последнего известного исправного набора элементов управления.

Windows поддерживает несколько копий CurrentControlSet, и CurrentControlSet на самом деле представляет собой символическую ссылку реестра, указывающую на одну из копий. Наборы элементов управления имеют имена в формате HKLM\SYSTEM\ControlSet*nnn*, где *nnn* — это число, например, 001 или 002. Раздел HKLM\SYSTEM\Select содержит параметры, определяющие роль каждого набора элементов управления. Например, если CurrentControlSet указывает на ControlSet001, то текущий параметр в разделе Select имеет значение 1. Параметр LastKnownGood в разделе Select содержит номер последнего известного исправного набора элементов управления, который в последний раз использовался для успешной загрузки. Другой параметр, который может быть в вашей системе определен для раздела Select, — это Failed, который указывает на последний набор элементов управления, для которого загрузка была признана неудачной и прервана в пользу попытки загрузки с последним известным удачным набором элементов управления. На рис. 10.19 показаны наборы элементов управления системы Windows Server и параметры Select.

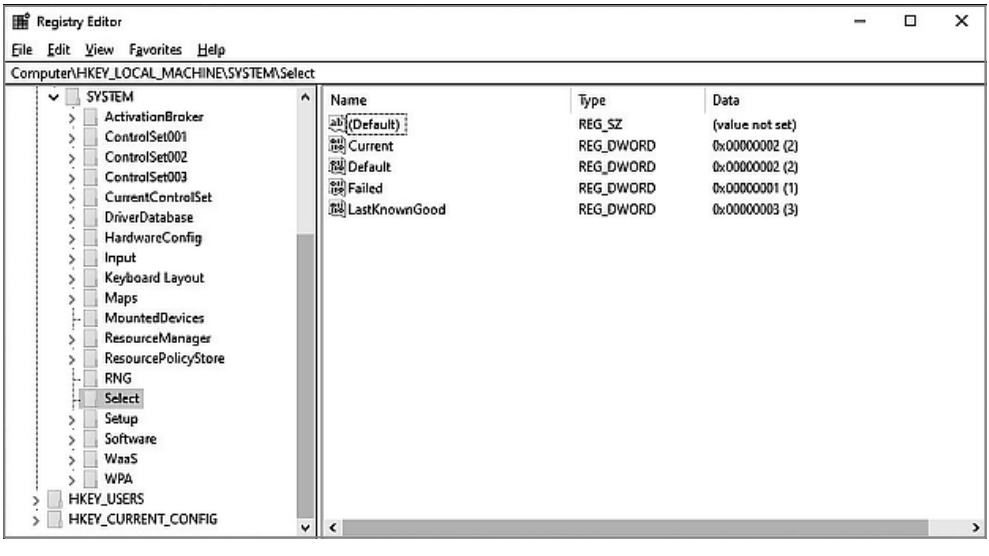


Рис. 10.19. Раздел выбора набора элементов управления в Windows Server 2019

NtInitializeRegistry берет содержимое последнего известного удачного набора элементов управления и синхронизирует его с содержимым дерева раздела CurrentControlSet. Если это была первая успешная загрузка системы, то последнего известного удачного набора не будет и система создаст для него новый набор элементов управления. Если последнее известное исправное дерево существует, система просто обновляет его, учитывая различия между ним и CurrentControlSet.

Последний известный удачный вариант полезен в ситуациях, когда изменение в CurrentControlSet, например изменение параметра настройки производительности

системы в разделе HKLM\SYSTEM\Control или добавление службы или драйвера устройства, приводит к сбою последующей загрузки. На рис. 10.20 показаны параметры запуска современного меню загрузки. Действительно, когда функция Последний удачный вариант (Last Known Good) включена и система находится в процессе загрузки, пользователи могут выбрать параметр Параметры запуска (Startup Settings) в разделе Устранение неполадок (Troubleshoot) современного меню загрузки (или в среде восстановления Windows), чтобы открыть другое меню, которое позволяет им ориентировать загрузку на использование последнего известного удачного набора элементов управления. Если в системе все еще используется устаревшее меню загрузки, пользователи должны нажать F8, чтобы включить Дополнительные параметры загрузки (Advanced Boot Options). Как показано на рисунке, когда выбран параметр Включить последнюю удачную конфигурацию (Enable Last Known Good Configuration), система загружается путем обновления реестра системы до конфигурации, существовавшей при последней успешной загрузке системы. В главе 12 более подробно описывается использование современного меню загрузки, среды восстановления Windows и других механизмов восстановления для устранения проблем при запуске системы.

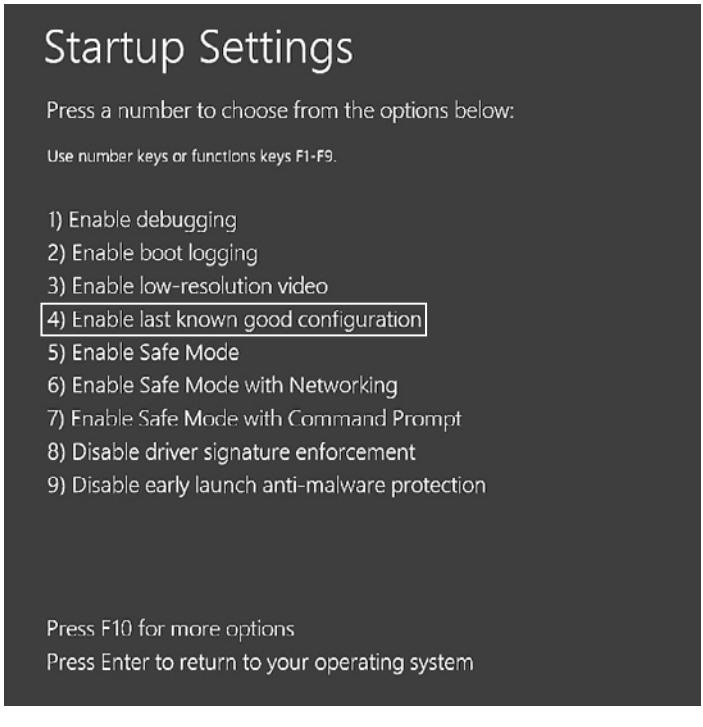


Рис. 10.20. Включение последней удачной конфигурации

Сбои служб

Служба может иметь дополнительные параметры `FailureActions` и `FailureCommand` в своем разделе реестра, который SCM записывает во время запуска службы. SCM

регистрируется в системе, чтобы система сигнализировала ему о завершении процесса службы. Когда процесс службы неожиданно завершается, SCM определяет, какие службы выполнялись в процессе, и предпринимает шаги по восстановлению, указанные в их параметрах реестра, связанных со сбоем. Кроме того, службы не ограничиваются только запросом действий при сбое или неожиданном завершении, поскольку привести к сбою службы могут и другие проблемы, такие как утечка памяти.

Если служба переходит в состояние `SERVICE_STOPPED` и код ошибки, возвращаемый в SCM, не равен `ERROR_SUCCESS`, SCM проверяет, установлен ли для службы флаг `FailactionActionsOnNonCrashFailures`, и выполняет такое же восстановление, как если бы служба вышла из строя. Чтобы использовать этот функционал, службу необходимо настроить с помощью API `ChangeServiceConfig2`, или системный администратор может применить утилиту Sc.exe с параметром `Failflag`, чтобы установить для `FailureActionsOnNonCrashFailures` значение `1`. При значении по умолчанию `0` SCM будет вести себя так же, как и в более ранних версиях Windows, для всех остальных служб.

Действия, которые служба может настроить для SCM, включают перезапуск службы, запуск некоей программы и перезагрузку компьютера. Кроме того, служба может указывать действия, которые будут происходить при первом сбое процесса службы, во втором и последующих случаях, а также может указывать период задержки, который SCM ожидает перед перезапуском службы, если та его запрашивает. Вы можете легко управлять действиями по восстановлению службы, используя вкладку Восстановление (Recovery) диалогового окна Свойства (Properties) службы в оснастке Службы (Services) MMC (рис. 10.21).

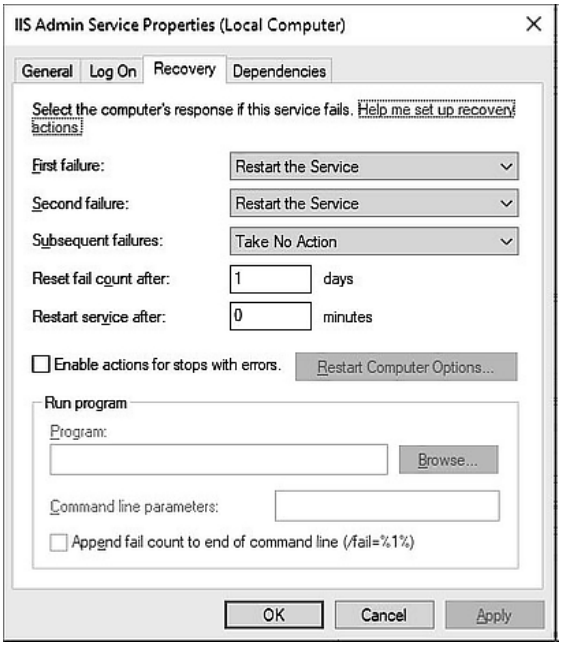


Рис. 10.21. Параметры вкладки Восстановление диалогового окна Службы

Обратите внимание: если действием при последующих сбоях является перезагрузка компьютера, SCM после запуска службы помечает хост-процесс как критический, вызывая нативный API `NtSetInformationProcess` с информационным классом `ProcessBreakOnTermination`. Неожиданно завершаясь, критический процесс приводит к сбою системы с кодом проверки ошибок `CRITICAL_PROCESS_DIED` (см. главу 2 тома 1).

Завершение работы службы

Когда Winlogon вызывает функцию Windows `ExitWindowsEx`, та отправляет сообщение Csrss — процессу подсистемы Windows — для вызова процедуры завершения его работы. Csrss просматривает активные процессы и уведомляет их о завершении работы системы. Csrss ожидает завершения каждого системного процесса, кроме SCM, в течение количества секунд, указанного в миллисекундах в HKCU\Control Panel\Desktop\WaitToKillTimeout (по умолчанию равно 5 с), прежде чем перейти к следующему процессу. Когда Csrss обнаруживает процесс SCM, он также уведомляет его о завершении работы системы, но использует тайм-аут, установленный отдельно для SCM. Csrss распознает SCM, задействуя идентификатор процесса Csrss, сохраненный при регистрации SCM в Csrss с помощью функции `RegisterServicesProcess` во время его инициализации. Тайм-аут SCM отличается от тайм-аута других процессов, поскольку Csrss знает, что SCM взаимодействует со службами, которым необходимо выполнить очистку при завершении, поэтому администратору может потребоваться настроить только тайм-аут SCM. Значение тайм-аута SCM в миллисекундах находится в параметре реестра HKLM\SYSTEM\CurrentControlSet\Control\WaitToKillServiceTimeout и по умолчанию составляет 20 с.

Обработчик завершения работы SCM отвечает за отправку уведомлений о завершении работы всем службам, которые запросили уведомление о завершении работы в ходе инициализации с помощью SCM. Функция SCM `ScShutdownAllServices` сначала запрашивает параметр HKLM\SYSTEM\CurrentControlSet\Control\ShutdownTimeout и устанавливает значение по умолчанию 20 с, если оно не установлено. Затем он обходит в цикле базу данных служб SCM. Для каждой службы он отменяет регистрацию возможных триггеров службы, определяет, желает ли служба получать уведомление о завершении работы, и отправляет команду завершения работы (`SERVICE_CONTROL_SHUTDOWN`), если это так. Обратите внимание на то, что все уведомления отправляются службам параллельно с использованием рабочих потоков пула потоков. Для каждой службы, которой SCM отправляет команду завершения работы, он записывает параметр подсказки ожидания службы, которое служба также указывает при регистрации в SCM. SCM отслеживает наибольшую подсказку ожидания из получаемых (если максимальное рассчитанное указание ожидания меньше тайм-аута завершения работы, заданного параметром реестра `ShutdownTimeout`, тайм-аут завершения работы считается максимальной подсказкой ожидания). После отправки сообщений о завершении работы SCM ожидает либо завершения работы всех служб, которые он уведомил об отключении, либо истечения времени, указанного в наибольшей подсказке ожидания.

Пока SCM сообщает службам о необходимости закрытия и ожидает их выхода, Csrss ожидает выхода SCM. Если срок ожидания истекает, а все службы не завершаются, SCM заканчивает работу, а Csrss продолжает процесс завершения работы. Если ожидание Csrss заканчивается без выхода SCM (время ожидания `WaitToKillServiceTimeout`

истекло), Csrss убивает SCM и продолжает процесс завершения работы. Таким образом, службы, которые не могут завершить работу своевременно, уничтожаются. Эта логика позволяет системе заканчивать работу при наличии служб, которые никогда не завершают работу из-за ошибочного проектирования, а также означает, что службы, которым требуется более 5 с, не завершат свои операции окончания работы.

Кроме того, поскольку порядок завершения работы недетерминированный, службы, которые могут зависеть от других служб для первоочередного завершения работы (так называемые зависимости завершения работы), не имеют возможности сообщить об этом SCM и, вероятно, никогда не смогут выполнить очистку.

Чтобы удовлетворить эти потребности, Windows реализует уведомления перед завершением работы и порядок завершения работы для борьбы с проблемами, вызванными этими двумя сценариями. Уведомление перед завершением работы отправляется службе, которая его запросила, через API `SetServiceStatus` (через принятый элемент управления `SERVICE_ACCEPT_PRESHUTDOWN`) с использованием того же механизма, который применяют и уведомления о завершении работы. Такие уведомления отправляются до выхода Wininit. SCM обычно ожидает их подтверждения.

Суть этих уведомлений состоит в том, чтобы пометить службы, очистка которых может занять много времени (например, службы сервера базы данных), и дать им больше времени для завершения работы. SCM отправляет запрос о ходе выполнения и ждет 10 с, пока служба ответит на это уведомление. Если служба не отвечает в течение этого времени, она уничтожается во время процедуры завершения работы, в противном случае она может работать столько, сколько необходимо, пока продолжает отвечать на запросы SCM.

Службы, участвующие в предварительном завершении работы, также могут указывать порядок завершения работы по отношению к другим службам предварительного завершения работы. Службы, которые зависят от первоочередного завершения работы других служб (например, службе групповой политики необходимо дождаться завершения работы Центра обновления Windows), могут указать свои зависимости завершения работы в параметре реестра HKLM\SYSTEM\CurrentControlSet\Control\PreshutdownOrder.

Процессы, общие для нескольких служб

Запуск каждой службы в отдельном процессе вместо совместного использования по возможности одного процесса приводит к затратам системных ресурсов впустую. Однако совместное использование процессов означает, что если в какой-либо из служб в процессе имеется ошибка, приводящая к его завершению, все службы в этом процессе завершают работу.

Некоторые из встроенных служб Windows работают в собственном процессе, а некоторые используют общий процесс совместно с другими службами. Например, процесс LSASS содержит службы, связанные с безопасностью, такие как служба диспетчера учетных записей безопасности (SamSs), служба сетевого входа в систему (Netlogon), служба шифрованной файловой системы (EFS) и служба изоляции ключей шифрования следующего поколения (CNG) (KeyIso).

Существует также неспецифический процесс с именем Service Host (SvcHost — %SystemRoot%\System32\Svchost.exe) для работы в нем нескольких служб. Несколько экземпляров SvcHost выполняются как разные процессы. Службы, которые

выполняются в процессах SvcHost, включают телефонию (TapiSrv), вызов удаленных процедур (RpcSs) и диспетчер подключений удаленного доступа (RasMan). Windows реализует службы, которые запускаются в SvcHost, как библиотеки DLL и включает определение ImagePath в форме %SystemRoot%\System32\svchost.exe–k netsvcs в разделе реестра службы. Раздел реестра службы также должен иметь параметр с именем `ServiceDll` в подразделе Parameters, который указывает на файл DLL службы.

Все службы, использующие общий процесс SvcHost, указывают один и тот же параметр (`-k netsvcs` в примере из предыдущего абзаца), чтобы у них была одна запись в базе данных образов SCM. Когда во время запуска службы SCM обнаруживает первую службу, имеющую SvcHost ImagePath с определенным параметром, он создает новую запись в базе данных образов и запускает процесс SvcHost с этим параметром. Параметр, указанный с помощью переключателя `-k`, является именем группы служб. Вся командная строка анализируется SCM при создании нового общего хост-процесса. Как обсуждалось в разделе «Вход службы в систему», если другая служба в базе данных использует тот же параметр `ImagePath`, ее SID службы будет добавлен в список SID группы нового хост-процесса.

Новый процесс SvcHost принимает группу служб, указанную в командной строке, и ищет параметр с тем же именем в разделе HKLM\SOFTWARE\Microsoft\Windows NT\CurrentVersion\Svchost. Он считывает содержимое параметра, интерпретируя его как список имен служб, и уведомляет SCM о том, что размещает эти службы, когда регистрируется в SCM.

Когда SCM обнаруживает другую общую службу (в ходе проверки параметра типа службы) во время запуска службы с `ImagePath`, совпадающим с записью, которая уже есть в базе данных образов, он не запускает второй процесс, а просто отправляет команду запуска службы в SvcHost, уже запущенный для этого параметра `ImagePath`. Существующий процесс SvcHost считывает параметр ServiceDll в разделе реестра службы, включает новый SID группы служб в своем токене и загружает DLL в свой процесс для запуска службы.

В табл. 10.12 перечислены все группы служб по умолчанию в Windows и некоторые службы, зарегистрированные для каждой из них.

Таблица 10.12. Основные группы служб

Группа служб	Службы	Примечания
LocalService	Интерфейс сетевого хранилища, узел диагностики Windows, время Windows, система событий COM+, служба автоматического прокси-сервера HTTP, уведомление пользовательского интерфейса платформы защиты программного обеспечения, служба порядка потоков, обнаружение LLDT, SSL, хост FDP, веб-клиент	Службы, которые запускаются под учетной записью локальной службы и могут использовать сеть на различных портах или не задействовать ее (следовательно, никаких ограничений)
LocalServiceAndNoImpersonation	UPnP и SSDP, смарт-карта, TPM, кэш шрифтов, обнаружение функций, AppID, qWAVE, Windows Connect Now, Media Center Extender, адаптивная яркость	Службы, которые запускаются под учетной записью локальной службы и используют сеть с фиксированным набором портов. Запускаются с применением токена с ограничением записи

Группа служб	Службы	Примечания
LocalServiceNetworkRestricted	DHCP, регистратор событий, Windows Audio, NetBIOS, Центр безопасности, родительский контроль, поставщик домашней группы	Службы, которые запускаются под учетной записью локальной службы и используют сеть с фиксированным набором портов
LocalServiceNoNetwork	Механизм диагностической политики, базовый механизм фильтрации, ведение журнала производительности и оповещения, брандмауэр Windows, автоконфигурация WWAN	Службы, которые запускаются под учетной записью локальной службы, но не используют сеть. Запускаются с применением токена с ограничением записи
LocalSystemNetworkRestricted	DWM, системный хост WDI, сетевые подключения, отслеживание распределенных каналов, конечная точка Windows Audio, автоконфигурация проводной/WLAN, Pnp-X, доступ к HID, служба платформы драйверов пользовательского режима, Superfetch, счетчик портативных устройств, прослушиватель домашней группы, ввод с планшета, совместимость программ, автономные файлы	Службы, которые запускаются под локальной системной учетной записью и используют сеть с фиксированным набором портов
NetworkService	Криптографические службы, DHCP-клиент, службы терминалов, рабочая станция, защита доступа к сети, NLA, DNS-клиент, телефония, сборщик событий Windows, WinRM	Службы, которые запускаются под учетной записью сетевой службы и используют сеть на различных портах (или не имеют принудительных сетевых ограничений)
NetworkServiceAndNoImpersonation	KTM для DTC	Службы, которые запускаются под учетной записью сетевой службы и используют сеть с фиксированным набором портов. Запускаются с применением токена с ограничением записи
NetworkServiceNetworkRestricted	Агент политики IPSec	Службы, которые запускаются под учетной записью сетевой службы и используют сеть с фиксированным набором портов

Разделение Svchost по службам

Как обсуждалось в предыдущем разделе, запуск службы в общем хост-процессе экономит системные ресурсы, но имеет большой недостаток: единственная необработанная ошибка в службе приводит к уничтожению всех других служб, совместно используемых в хост-процессе. Для решения этой проблемы в Windows 10 Creators Update (RS2) появилась функция разделения Svchost по службам.

При запуске SCM считывает из реестра три параметра, представляющие низкие, средние и жесткие глобальные ограничения фиксации служб. Эти параметры SCM использует для отправки сообщений о нехватке ресурсов в случае, если система работает в условиях нехватки памяти. Затем он считывает пороговое значение разделения службы Svchost из параметра реестра HKLM\SYSTEM\CurrentControlSet\Control\SvcHostSplitThresholdInKB. Параметр представляет собой минимальный объем физической памяти системы (в килобайтах), необходимый для включения разделения

службы Svchost (значение по умолчанию составляет 3,5 Гбайт в клиентских системах и около 3,7 Гбайт — в серверных). Затем SCM получает общий объем физической памяти системы с помощью API `GlobalMemoryStatusEx` и сравнивает его с пороговым значением, ранее считанным из реестра. Если общий объем физической памяти превышает пороговое значение, включается разделение службы Svchost путем установки внутренней глобальной переменной.

Разделение Svchost по службам, если оно активно, изменяет поведение, при котором SCM запускает хост-процесс Svchost общих служб. Как обсуждалось в разделе «Запуск службы» ранее в этой главе, SCM не ищет существующую запись образа в своей базе данных, если для службы разрешено разделение. Это означает, что, хотя служба помечена как общедоступная, она запускается с использованием своего частного хост-процесса и ее тип изменяется на `SERVICE_WIN32_OWN_PROCESS`. Разделение по службам допускается только при соблюдении следующих условий.

- Разделение Svchost по службам включено глобально.
- Служба не помечена как критическая. Она помечается как критическая, если действие по восстановлению при повторном сбое предполагает перезагрузку компьютера (это обсуждалось ранее в разделе «Сбои служб»).
- Имя хост-процесса службы — Svchost.exe.
- Разделение по службам не отключено явно для службы с помощью параметра реестра `SvcHostSplitDisable` в разделе управления службой.

Технологии диспетчера памяти, такие как сжатие памяти и объединение, помогают сохранить как можно большую часть рабочего набора системы. Это объясняет одну из причин включения разделения Svchost по службам. Несмотря на то что в системе создается множество новых процессов, диспетчер памяти обеспечивает то, что все физические страницы хост-процессов остаются общими и потребляют как можно меньше системных ресурсов. Объединение, сжатие и совместное использование памяти подробно описаны в главе 5 тома 1.

ЭКСПЕРИМЕНТ. Практическое разделение Svchost по службам

Если вы используете рабочую станцию с Windows 10, оснащенную 4 Гбайт или более памяти, при открытии диспетчера задач можете заметить, что в настоящее время выполняется множество экземпляров процесса Svchost.exe. Как поясняется в данном разделе, это не приводит к ненужным затратам памяти, но вас может заинтересовать отключение разделения Svchost. Сначала откройте диспетчер задач и подсчитайте, сколько экземпляров процесса Svchost в данный момент запущено в системе. В системе Windows 10 May 2019 Update (19H1) у вас должно быть около 80 экземпляров процесса Svchost. Легко подсчитать их, открыв административное окно PowerShell и введя следующую команду:

```
(get-process -Name "svchost" | measure).Count
```

В системе автора эта команда вернула значение 85.

Откройте редактор реестра, введя `regedit.exe` в поле поиска Cortana, и перейдите к разделу HKLM\SYSTEM\CurrentControlSet\Control. Обратите внимание на

текущие данные в DWORD-параметре SvcHostSplitThresholdInKB. Чтобы глобально отключить разделение службы Svchost, следует изменить параметр реестра, установив для его данных значение 0. (Вы можете изменить его, дважды щелкнув на параметре реестра и введя 0.) После изменения параметра реестра перезагрузите систему и повторите предыдущий шаг — подсчет количества экземпляров процесса Svchost. Теперь система работает с гораздо меньшим их количеством:

```
PS C:\> (get-process -Name "svchost" | measure).Count
26
```

Чтобы вернуться к прежнему поведению, необходимо восстановить прежнее содержимое параметра реестра SvcHostSplitThresholdInKB. Изменяя DWORD-параметр, вы также можете точно настроить объем физической памяти, необходимый для правильного включения разделения Svchost.

Теги служб

Одним из недостатков применения хост-процессов служб является то, что учет времени использования процессора, а также задействования ресурсов конкретной службой намного сложнее, поскольку каждая служба использует адресное пространство памяти, таблицу дескрипторов и отчеты процессоров по использованию их ресурсов процессами совместно с другими службами, входящими в ту же группу служб. Несмотря на то что внутри хост-процесса службы всегда есть поток, принадлежащий определенной службе, это соответствие не всегда легко установить. Например, служба может использовать рабочие потоки для выполнения своей операции или, возможно, начальный адрес и стек потока не раскрывают имя DLL службы, что затрудняет определение того, какую работу может выполнять поток и какой службе он может принадлежать.

Windows реализует атрибут службы, называемый тегом службы (не путать с тегом драйвера), который SCM генерирует, вызывая ScGenerateServiceTag при создании службы или формировании базы данных службы во время загрузки системы. Атрибут — это просто индекс, идентифицирующий службу. Тег службы хранится в поле SubProcessTag блока среды каждого потока (TEB) (дополнительную информацию о TEB см. в главе 3 тома 1) и распространяется по всем потокам, создаваемым основным потоком службы, за исключением созданных косвенно через API-интерфейсы пула потоков.

Хотя тег службы хранится внутри SCM, некоторые утилиты Windows, такие как Netstat.exe (утилита, которую можно использовать для отображения того, какие программы открыли какие порты в сети), задействуют недокументированные API для запроса тегов службы и сопоставления их со службой имен. Еще один инструмент, который можно применять для просмотра тегов служб, — это ScTagQuery от Winsider Seminars & Solutions Inc. (www.winsiders.com/tools/sctagquery/sctagquery.htm). Он может запрашивать у SCM сопоставление каждого тега службы и отображать их либо для всей системы, либо для каждого процесса. Он также может показать вам, каким службам принадлежат все потоки внутри хост-процесса. (Это возможно при условии, что эти потоки имеют связанный с ними тег службы.) Таким образом,

если у вас есть неконтролируемая служба, потребляющая много процессорного времени, ее можно определить, если начальный адрес потока или стек не имеют связанной с ними очевидной служебной DLL.

Пользовательские службы

Как обсуждалось в разделе «Запуск служб в альтернативных учетных записях», службу можно запустить с применением учетной записи локального пользователя системы. Служба, настроенная таким образом, всегда загружается с помощью указанной учетной записи пользователя независимо от того, вошел ли он в систему в данный момент. Это может стать ограничением в многопользовательских средах, где служба должна выполняться с токеном доступа текущего пользователя, вошедшего в систему. Кроме того, это может подвергнуть риску учетную запись пользователя, поскольку злоумышленники могут внедриться в процесс службы и задействовать его токен для доступа к ресурсам, к которым не должны его получать (имея возможность также пройти аутентификацию в сети).

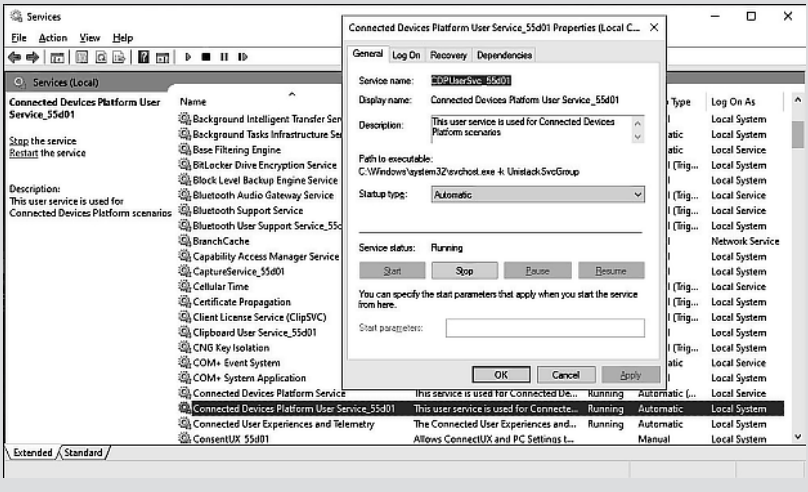
Пользовательские службы, доступные в Windows 10 Creators Update (RS2), позволяют службе запускаться с токеном текущего пользователя, вошедшего в систему. Пользовательские службы могут запускаться в отдельном процессе или разделять процесс с одной или несколькими другими службами, работающими под той же учетной записью пользователя, вошедшего в систему, как это происходит и со стандартными службами. Они запускаются, когда пользователь выполняет интерактивный вход в систему, и останавливаются, когда выходит из нее. SCM поддерживает два дополнительных флага типа — `SERVICE_USER_SERVICE` (64) и `SERVICE_USERSERVICE_INSTANCE` (128), которые идентифицируют шаблон пользовательской службы и экземпляр пользовательской службы соответственно. Одно из состояний конечного автомата Winlogon (подробную информацию о Winlogon и процессе загрузки см. в главе 12) выполняется, когда инициируется интерактивный вход в систему. Состояние создает сеанс входа в систему нового пользователя, оконную станцию, рабочий стол и среду, отображает куст реестра `HKEY_CURRENT_USER` и уведомляет подписчиков входа в систему (LogonUI и диспетчер пользователей). Служба диспетчера пользователей (Usermgr.dll) через RPC может обращаться к SCM для доставки события сеанса `WTS_SESSION_LOGON`.

SCM обрабатывает сообщение с помощью функции `ScCreateUserServicesForUser`, которая вызывает диспетчер пользователей для получения токена авторизованного в системе в настоящий момент пользователя. Затем он запрашивает список служб шаблонов пользователей из базы данных SCM и для каждого из них генерирует новое имя службы экземпляра пользователя.

ЭКСПЕРИМЕНТ. Наблюдение за пользовательскими службами

Отладчик ядра может легко показать атрибуты безопасности токена процесса. В этом эксперименте вам понадобится компьютер с Windows 10 с активированным и подключенным к хосту отладчиком ядра (локальный отладчик тоже подойдет). В этом эксперименте вы выбираете экземпляр пользовательской службы и анализируете токен его хост-процесса. Откройте инструмент Службы

(Services), введя его имя в поле поиска Cortana. Приложение отображает стандартные службы, а также экземпляры пользовательских служб (хотя ошибочно отображает локальную систему в качестве учетной записи пользователя), которые можно легко идентифицировать, поскольку они имеют локальный уникальный идентификатор (LUID, созданный диспетчером пользователей), прикрепленный к их отображаемым именам. В этом примере служба пользователя подключенного устройства отображается в приложении Службы (Services) как Connected Device User Service_55d01.



Если дважды щелкнуть на идентифицированной службе, инструмент покажет фактическое имя экземпляра пользовательской службы (в примере `CDPUserSvc_55d01`). Если служба размещена в общем процессе, например, выбранном в примере, вам следует задействовать редактор реестра для навигации по корневому разделу службы шаблона пользовательской службы, который имеет то же имя, что и экземпляр, но без LUID (в примере имя шаблона пользовательской службы — `CDPUserSvc`). Как пояснялось в эксперименте «Просмотр привилегий, запрашиваемых службами», в подразделе Parameters хранится имя DLL службы. Его следует использовать в Process Explorer для поиска правильного идентификатора хост-процесса (в последних версиях Windows 10 можно просто задействовать диспетчер задач).

После того как вы нашли PID хост-процесса, следует войти в отладчик ядра и ввести следующие команды (заменив `<ServicePid>` на PID хост-процесса службы):

```
!process <ServicePid> 1
```

Отладчик отображает несколько фрагментов информации, включая адрес связанного объекта токена безопасности:

```
Kd: 0> !process 0n5936 1
Searching for Process with Cid == 1730
PROCESS ffffe10646205080
```

```
    SessionId: 2 Cid: 1730     Peb: 81ebbd1000 ParentCid: 0344
    DirBase: 8fe39002 ObjectTable: ffffa387c2826340 HandleCount: 313.
    Image: svchost.exe
    VadRoot ffffe1064629c340 Vads 108 Clone 0 Private 962. Modified 214. Locked 0.
    DeviceMap ffffa387be1341a0
    Token                             ffffa387c2bdc060
    ElapsedTime                       00:35:29.441
    ...
    <Output omitted for space reasons>
```

Чтобы показать атрибуты безопасности токена, нужно просто использовать команду !token, за которой следует адрес объекта токена (он представлен структурой данных _TOKEN), возвращенный предыдущей командой. Вы можете легко убедиться, что в процессе размещена пользовательская служба, просмотрев атрибут безопасности WIN://ScmUserService, как показано в следующем выводе:

```
0: kd> !token ffffa387c2bdc060
_TOKEN 0xffffa387c2bdc060
TS Session ID: 0x2
User: S-1-5-21-725390342-1520761410-3673083892-1001
User Groups:
 00 S-1-5-21-725390342-1520761410-3673083892-513
    Attributes – Mandatory Default Enabled

... <Output omitted for space reason> ...

OriginatingLogonSession: 3e7
PackageSid: (null)
CapabilityCount: 0 Capabilities: 0x0000000000000000
LowboxNumberEntry: 0x0000000000000000
Security Attributes:
 00 Claim Name   : WIN://SCMUserService
    Claim Flags: 0x40 – UNKNOWN
    Value Type   : CLAIM_SECURITY_ATTRIBUTE_TYPE_UINT64
    Value Count: 1
    Value[0]   : 0
 01 Claim Name   : TSA://ProcUnique
    Claim Flags: 0x41 – UNKNOWN
    Value Type   : CLAIM_SECURITY_ATTRIBUTE_TYPE_UINT64
    Value Count: 2
    Value[0]   : 102
    Value[1]   : 352550
```

Process Hacker — системный инструмент, аналогичный Process Explorer и доступный по адресу https://processhacker.sourceforge.io/, способен извлекать ту же информацию.

Как обсуждалось ранее, имя экземпляра пользовательской службы генерируется объединением исходного имени службы и локального уникального идентификатора (LUID), сгенерированного диспетчером пользователей для идентификации интерактивного сеанса пользователя, называемого идентификатором контекста. Идентификатор контекста для интерактивного сеанса входа в систему хранится в изменяющемся параметре реестра HKLM\SOFTWARE\Microsoft\Windows NT\CurrentVersion\Winlogon\VolatileUserMgrKey\<Session ID>\<User SID>\contextLuid, где <Session ID> и <User SID> — идентификатор сеанса входа в систему и SID пользователя соответственно. Открыв редактор реестра и перейдя к этому разделу, вы найдете то же значение идентификатора контекста, которое использовалось для создания имени экземпляра пользовательской службы.

На рис. 10.22 показан пример экземпляра пользовательской службы Clipboard User Service, который запускается с помощью токена текущего авторизированного пользователя. Сгенерированный идентификатор контекста для сеанса 1 — 0x3a182, как показано в изменяющемся разделе реестра User Manager (подробности см. в предыдущем эксперименте). Затем SCM вызывает `ScCreateService`, который создает служебную запись в базе данных SCM. Новая запись службы представляет собой новый экземпляр пользовательской службы и сохраняется в реестре, как и обычные службы. Дескриптор безопасности службы, все зависимые службы и информация о триггерах копируются из шаблона пользовательской службы в новую службу экземпляра пользователя.

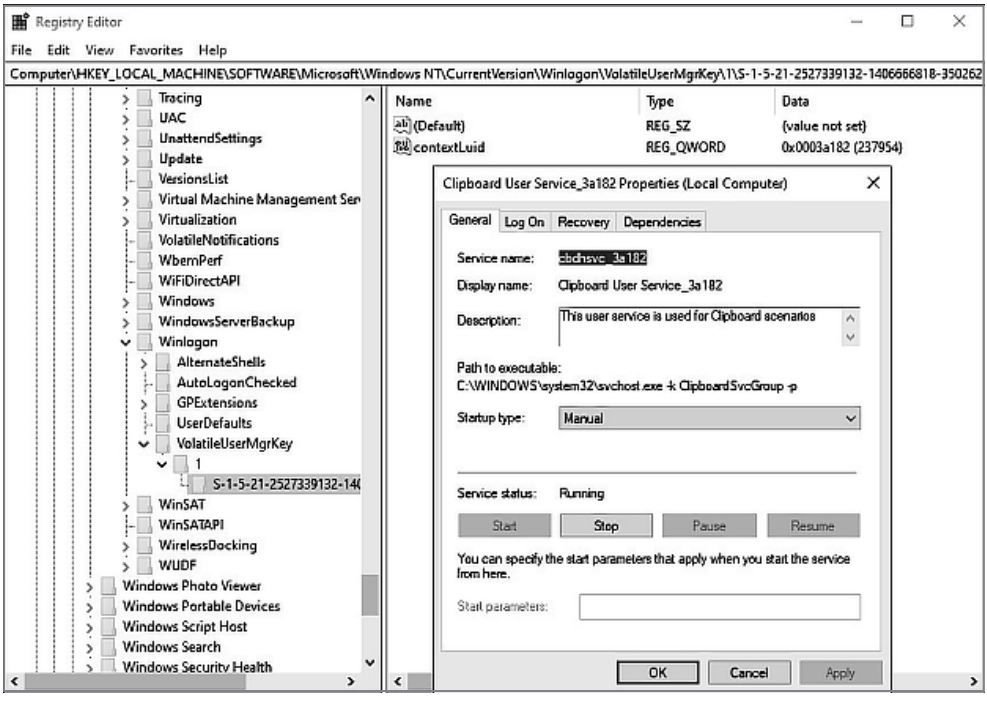


Рис. 10.22. Экземпляр пользовательской службы буфера обмена, работающий с идентификатором контекста 0x3a182

SCM регистрирует возможные триггеры службы (подробности см. в разделе «Службы с запуском по триггеру» ранее в этой главе), а затем запускает службу, если для ее типа запуска установлен параметр `SERVICE_AUTO_START`. Как обсуждалось в разделе «Вход службы в систему», когда SCM запускает процесс, размещающий пользовательскую службу, он назначает токен текущего вошедшего в систему пользователя и атрибут безопасности `WIN://ScmUserService`, применяемый SCM для распознавания того, что процесс действительно является хостом для службы. На рис. 10.23 показано, что после входа пользователя в систему подразделы экземпляра и шаблона сохраняются в корневом разделе Services, представляющем одну и ту же пользовательскую службу. Подраздел экземпляра удаляется при выходе пользователя из системы и игнорируется, если он все еще присутствует во время запуска системы.

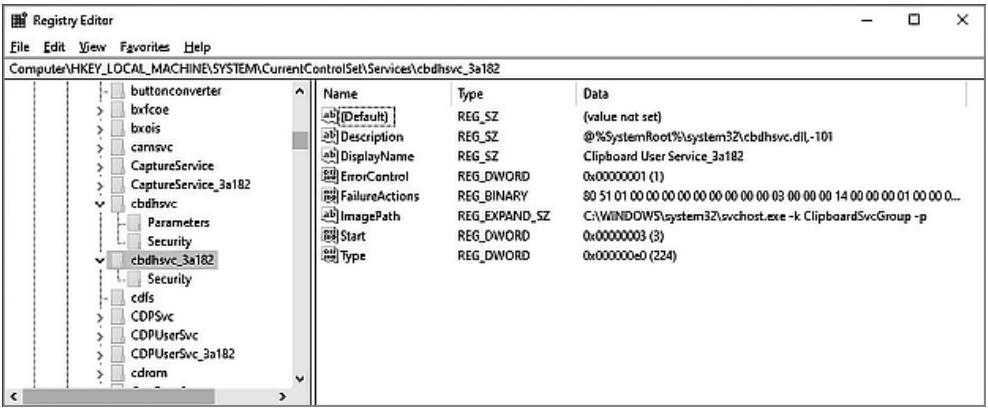


Рис. 10.23. Разделы реестра для экземпляра пользовательской службы и шаблона

Пакетные службы

Как упоминалось в разделе «Вход службы в систему», начиная с Windows 10 Anniversary Update (RS1), диспетчер управления службами поддерживает пакетные службы. Пакетная служба идентифицируется с помощью флага `SERVICE_PKG_SERVICE` (512), установленного в ее типе службы. Пакетные службы были разработаны в основном для поддержки стандартных настольных приложений Win32 (они могут работать со связанной службой), преобразованных в новую модель современных приложений. Desktop App Converter действительно способен преобразовать приложение Win32 в приложение Centennial, которое работает в легком контейнере, называющемся Helium. Более подробную информацию о модели современного приложения можно найти в разделе «Пакетные приложения» главы 8.

При запуске пакетной службы SCM считывает информацию о пакете из реестра и, как и для стандартных приложений Centennial, вызывает службу AppInfo. Последняя проверяет наличие информации о пакете в репозитории состояний и целостность всех файлов пакета приложения, а далее помечает токен хост-процесса новой службы правильными атрибутами безопасности. Затем процесс запускается из приостановленного состояния с использованием API `CreateProcessAsUser`, включая атрибут «Полное имя пакета», и создается контейнер Helium, который будет применять перенаправление реестра и виртуальную файловую систему (VFS), как для обычных приложений Centennial.

Защищенные службы

В главе 3 тома 1 подробно описана архитектура защищенных процессов и технология Protected Process Light (PPL). Диспетчер управления службами Windows 8.1 поддерживает защищенные службы. На момент написания этой книги служба может иметь четыре уровня защиты: Windows, Windows Light, Antimalware Light и App. Программа управления службами может указать защиту службы с помощью API `ChangeServiceConfig2` (с информационным уровнем `SERVICE_CONFIG_LAUNCH_PROTECTED`). Основной

исполняемый файл службы (или библиотека в случае общих служб) должен быть правильно подписан для работы в качестве защищенной службы по правилам для защищенных процессов (это означает, что система проверяет EKU цифровой подписи и корневой сертификат и генерирует максимальный уровень подписывающего лица, как описано в главе 3 тома 1).

Хост-процесс службы, запущенный как защищенный, гарантирует определенный вид защиты по отношению к другим незащищенным процессам. В зависимости от уровня защиты они не могут получить некоторые права доступа при попытке получить доступ к хост-процессу защищенной службы. (Механизм идентичен стандартным защищенным процессам. Классическим примером является незащищенный процесс, который не может внедрить какой-либо код в защищенную службу.)

Даже процессы, запущенные под учетной записью SYSTEM, не могут получить доступ к защищенному процессу. Однако SCM должен иметь полный доступ к хост-процессу защищенной службы. В результате Wininit.exe запускает SCM, указав максимальный уровень защиты пользовательского режима — WinTcb Light. На рис. 10.24 показана цифровая подпись основного исполняемого файла SCM, Services.exe, которая включает EKU компонента Windows TCB (1.3.6.1.4.1.311.10.3.23).

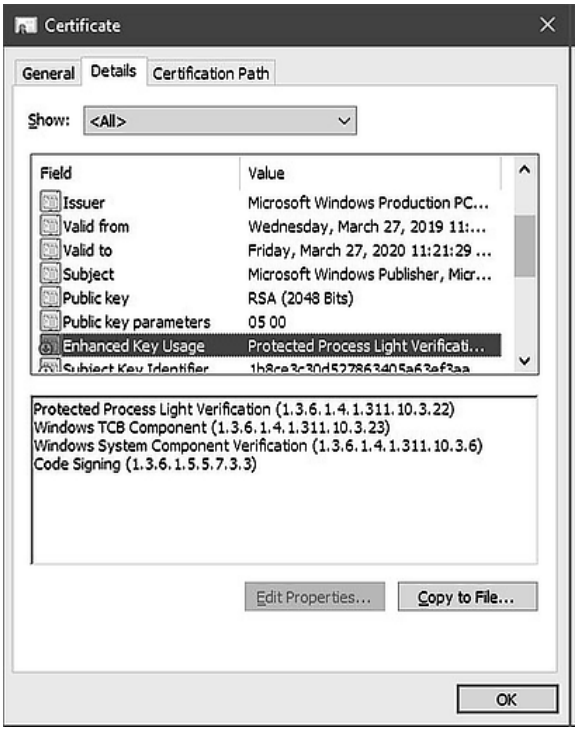


Рис. 10.24. Цифровой сертификат основного исполняемого файла Service Control Manager (service.exe)

Вторую часть защиты обеспечивает Service Control Manager. Когда клиент запрашивает действие, которое необходимо выполнить над защищенной службой,

SCM вызывает подпрограмму `ScCheckServiceProtectedProcess`, чтобы проверить, имеет ли вызывающая сторона достаточные права доступа для выполнения запрошенного действия над службой. В табл. 10.13 перечислены запрещенные операции, запрошенные незащищенным процессом защищенной службы.

Таблица 10.13. Список запрещенных операций при запросе от незащищенного клиента

Имя задействованного API	Операция	Описание
ChangeServiceConfig[2]	Изменить конфигурацию службы	Любое изменение конфигурации защищенной службы запрещено
SetServiceObjectSecurity	Установить новый дескриптор безопасности для службы	Применение нового дескриптора безопасности к защищенной службе запрещено. (Это может сузить способы атаки на службу.)
DeleteService	Удалить службу	Незащищенный процесс не может удалить защищенную службу
ControlService	Отправить управляющий код в службу	Для незащищенных вызывающих абонентов разрешены только код управления, определенный службой, и SERVICE_CONTROL_INTERROGATE. SERVICE_CONTROL_STOP разрешен для любого уровня защиты, кроме защиты от вредоносных программ

Функция `ScCheckServiceProtectedProcess` ищет запись службы по указателю службы, указанному вызывающей стороной, и если служба не защищена, всегда предоставляет доступ. В ином случае она олицетворяет токен процесса клиента, получает уровень защиты процесса и реализует следующие правила.

- Если запрос представляет собой запрос управления STOP и целевая служба не защищена на уровне защиты от вредоносного ПО, предоставьте доступ (службы, имеющие защиту от вредоносного ПО, невозможно остановить незащищенными процессами).
- Если SID службы `TrustedInstaller` присутствует в группах токенов клиента или установлен в качестве пользователя токена, SCM предоставляет доступ к защите процесса клиента.
- В противном случае он вызывает `RtlTestProtectedAccess`, который выполняет те же проверки, что и для защищенных процессов. Доступ предоставляется только в том случае, если уровень защиты клиентского процесса совместим с целевой службой. Например, процесс, защищенный Windows, всегда может работать на всех защищенных уровнях обслуживания, тогда как PPL для защиты от вредоносных программ может работать только со службами, защищенными на уровне защиты от вредоносных программ и приложений.

Примечательно, что последняя описанная проверка не выполняется ни для одного клиентского процесса, запущенного с учетной записью виртуальной службы `TrustedInstaller`. Это сделано специально. Когда Центр обновления Windows

устанавливает обновление, он должен иметь возможность запускать, останавливать и контролировать любые службы, не требуя применения надежной цифровой подписи, которая может стать для Центра обновления Windows нежелательной возможностью атаки.

ПЛАНИРОВАНИЕ ЗАДАЧ И UBPM

Различные компоненты Windows традиционно отвечают за управление размещенными или фоновыми задачами, поскольку операционная система усложняет свои функции: все, от диспетчера управления службами, описанного ранее, до средства запуска сервера DCOM и поставщика WMI, также несут ответственность за выполнение внепроцессного размещенного кода. Хотя современные версии Windows используют инфраструктуру фонового брокера для управления большинством фоновых задач современных приложений (более подробную информацию см. в главе 8), планировщик задач по-прежнему остается основным компонентом, управляющим задачами Win32. В Windows реализован унифицированный диспетчер фоновых процессов (UBPM), который обрабатывает задачи, управляемые планировщиком задач.

Служба планировщика задач (Schedule) реализована в библиотеке Schedsvc.dll и запускается в общем процессе Svchost. Она поддерживает базу данных задач и размещает UBPM, который запускает и останавливает задачи, а также управляет их действиями и триггерами. UBPM использует службы, предоставляемые брокером активности рабочего стола (DAB), брокером системных событий (SEB) и диспетчером ресурсов, для получения уведомлений при создании триггеров задач. (DAB и SEB размещаются в службе брокера системных событий, а диспетчер ресурсов — в службе инфраструктуры брокера.) И планировщик задач, и UBPM предоставляют общедоступные интерфейсы через RPC. Внешние приложения могут использовать COM-объекты для подключения к этим интерфейсам и взаимодействия с обычными задачами Win32.

Планировщик задач

Планировщик задач реализует хранилище для всех задач. Он также содержит службу простоя планировщика, способную определять, когда система входит в состояние ожидания или выходит из него, и поставщик ловушек событий, который помогает планировщику задач запускать задачу при изменении состояния компьютера и предоставляет внутреннюю систему триггеров журнала событий. Планировщик задач включает в себя и еще один компонент — прокси-сервер UBPM, который собирает все действия и триггеры задач, преобразует их дескрипторы в формат, понятный UBPM, и отправляет их в UBPM.

Обзор архитектуры планировщика задач показан на рис. 10.25. Как видно на рисунке, планировщик задач тесно взаимодействует с UBPM (оба компонента запускаются в службе планировщика задач, которая размещается в общем процессе Svchost.exe). UBPM управляет состояниями задач и получает уведомления от SEB, DAB и диспетчер ресурсов через состояния WNF.

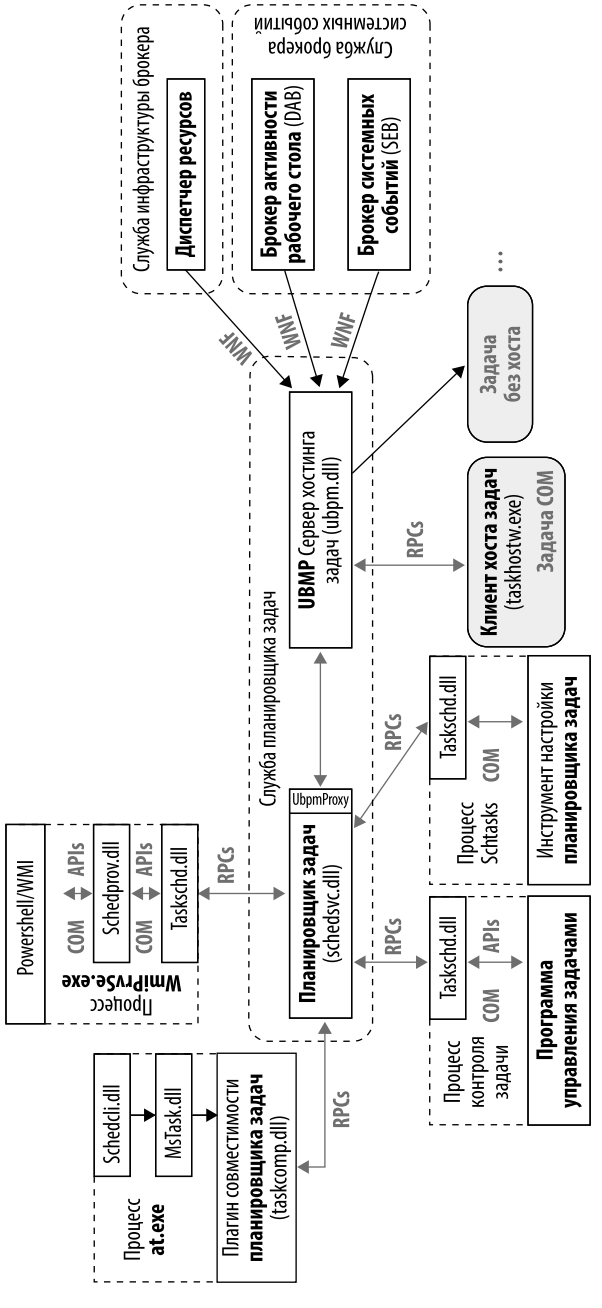


Рис. 10.25. Архитектура планировщика задач

Планировщик задач выполняет важную работу по предоставлению доступа к серверной части API планировщика задач COM. Когда программа управления задачами вызывает один из этих API, механизм COM загружает в адресное пространство приложения библиотеку COM API планировщика задач (Taskschd.dll). Библиотека запрашивает службы от имени программы управления задачами у планировщика задач через интерфейсы RPC.

Аналогичным образом поставщик WMI планировщика задач (Schedprov.dll) реализует классы и методы COM, способные взаимодействовать с библиотекой COM API планировщика задач. Его классы, свойства и события WMI можно вызывать из Windows PowerShell с помощью командлета ScheduledTasks (описан по адресу: https://docs.microsoft.com/en-us/powershell/module/scheduledtasks/). Обратите внимание на то, что планировщик задач включает в себя подключаемый модуль совместимости, который позволяет устаревшим приложениям, таким как AT-команда, работать с планировщиком задач. В обновлении Windows 10 (19H1) за май 2019 года инструмент AT объявлен устаревшим, вместо него следует использовать schtasks.exe.

Инициализация

При запуске диспетчера управления службами служба планировщика задач запускает процедуру инициализации. Она начинается с регистрации поставщика событий ETW на основе манифеста с глобальным уникальным идентификатором DE7B24EA-73C8-4A09-985D-5BDADCFA9017. Все события, генерируемые планировщиком задач, используются UBPM. Затем он инициализирует хранилище учетных данных, которое представляет собой компонент, применяемый для безопасного доступа к учетным данным пользователя, хранящимся в диспетчере учетных данных и хранилище задач. Этот компонент проверяет, что все дескрипторы задач XML, расположенные во вторичной теневой копии хранилища задач, сохраняемой по соображениям совместимости и обычно расположенной по пути `%SystemRoot%\System32\Tasks`, синхронизированы с дескрипторами задач, находящимися в кэше хранилища задач. Кэш хранилища задач представлен несколькими разделами реестра, корнем которых является HKLM\SOFTWARE\Microsoft\Windows NT\CurrentVersion\Schedule\TaskCache.

Следующим шагом при инициализации планировщика задач является инициализация UBPM. Служба планировщика задач использует API `UbpmInitialize`, экспортированный из UBPM.dll, для запуска основных компонентов UBPM. Функция регистрирует потребителя ETW поставщика событий планировщика задач и подключается к диспетчеру ресурсов. Диспетчер ресурсов — это компонент, загружаемый диспетчером состояния процесса (Psmsrv.dll в контексте службы инфраструктуры брокеров), который управляет политиками в отношении ресурсов на основе состояния компьютера и использования глобальных ресурсов. Диспетчер ресурсов помогает UBPM управлять задачами технического обслуживания. Задачи такого типа выполняются только при определенных состояниях системы, например, когда загрузка процессора рабочей станции низкая, игровой режим выключен, пользователь физически отсутствует и т. д. Затем код инициализации

UBPM получает от брокера системных событий имена состояний WNF, представляющие состояния задачи: питание от сети, неактивная рабочая станция, IP-адрес или доступность сети, переключение рабочей станции на питание от батареи. Эти условия отображаются на вкладке Условия (Conditions) диалогового окна Создать задачу (Create Task) подключаемого модуля MMC Планировщик заданий (Task Scheduler).

UBPM инициализирует рабочие потоки своего внутреннего пула потоков, получает варианты питания системы, считывает список действий по обслуживанию и критическим задачам (из раздела реестра HKLM\System\CurrentControlSet\Control\Ubpm и настроек групповой политики) и подписывается на уведомления о настройках питания системы (способ, которым UBPM узнает, что система меняет состояние питания).

Управление выполнением возвращается планировщику задач, который наконец регистрирует глобальные интерфейсы RPC — как свои, так и UBPM. Эти интерфейсы используются клиентской DLL API планировщика задач (Taskschd.dll), чтобы дать возможность клиентским процессам взаимодействовать через COM-интерфейсы планировщика задач, которые документированы по адресу https://docs.microsoft.com/en-us/windows/win32/api/taskschd/.

После завершения инициализации хранилище задач перечисляет все задачи, установленные в системе, и запускает их. Задачи хранятся в кэше разбитыми на четыре группы: загрузка, вход в систему, обычная задача и задача обслуживания. У каждой группы есть связанный подраздел, называемый разделом задач группы индекса и расположенный в корневом разделе реестра хранилища задач (HKLM\SOFTWARE\Microsoft\Windows NT\CurrentVersion\Schedule\TaskCache, как было указано ранее). Внутри каждого раздела группы задач индексирования имеется один подраздел для каждой задачи, идентифицируемый с помощью глобального уникального идентификатора (GUID). Планировщик задач перечисляет имена всех подразделов группы и для каждого из них открывает соответствующий главный раздел задачи, который находится в подразделе Tasks корневого раздела реестра хранилища задач. На рис. 10.26 показан пример задачи загрузки с GUID {0C7D8A27-9B28-49F1-979C-AD37C4D290B1}. GUID задачи указан на рисунке как одна из первых записей в групповом разделе индекса загрузки. Показан также главный раздел задачи, который хранит в реестре двоичные данные, полностью описывающие задачу.

Главный ключ задачи содержит всю информацию, описывающую задачу. Наиболее важны два свойства задачи: *триггеры*, которые описывают условия, запускающие задачу, и *действия*, которые описывают, что происходит при ее выполнении. Оба свойства хранятся в двоичных параметрах реестра, названных Triggers и Actions (см. рис. 10.26). Планировщик задач сначала считывает хеш всего дескриптора задачи (хранится в параметре реестра Hash), затем — все данные конфигурации задачи и двоичные данные для триггеров и действий. После анализа этих данных он добавляет каждый идентифицированный дескриптор триггера и действия во внутренний список.

Затем планировщик задач пересчитывает хеш SHA256 нового дескриптора задачи, который включает в себя все данные, считанные из реестра, и сравнивает его с ожидаемым значением. Если два хеша не совпадают, планировщик задач открывает XML-файл, связанный с задачей и содержащийся в теневой копии хранилища (папка `%SystemRoot%\System32\Tasks`), анализирует его данные, пересчитывает новый хеш и, наконец, заменяет дескриптор задачи в реестре. Действительно, задачи могут быть описаны с помощью двоичных данных, включенных в реестр, а также с помощью XML-файла, который соответствует четко определенной схеме, описанной по адресу https://docs.microsoft.com/en-us/windows/win32/taskschd/task-scheduler-schema.

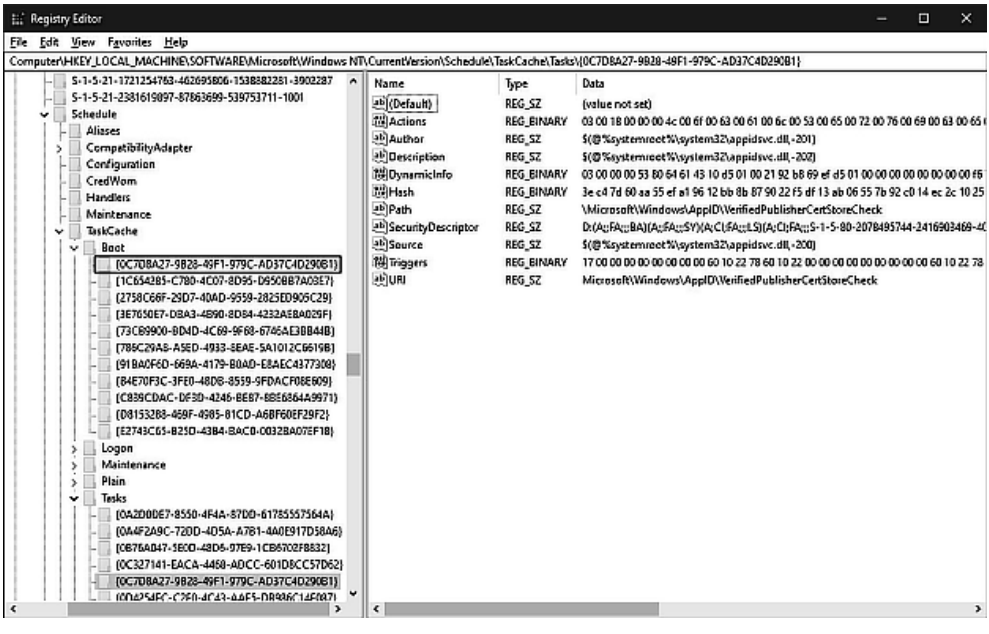
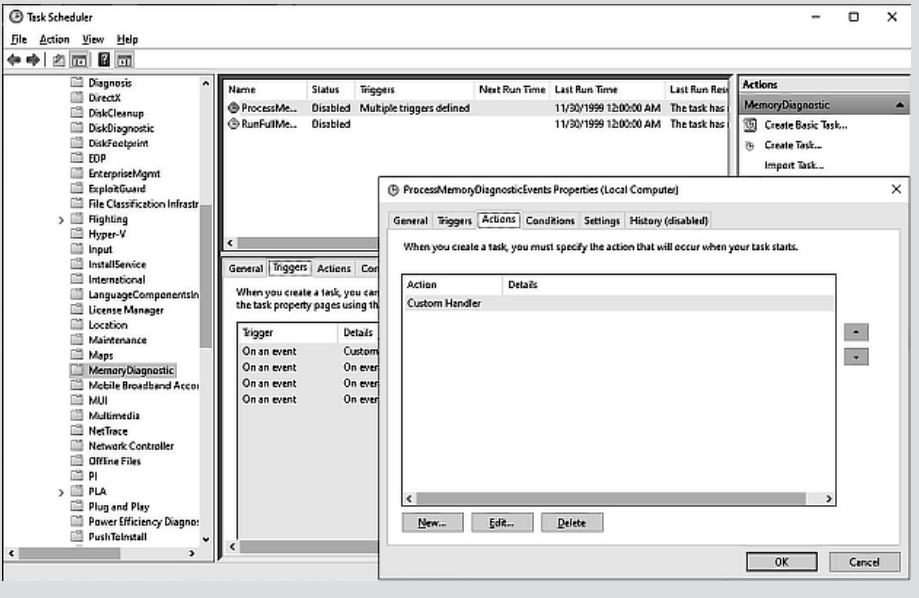


Рис. 10.26. Главный раздел задачи загрузки

ЭКСПЕРИМЕНТ. Просмотр XML-файла дескриптора задачи

Дескрипторы задач, как сказано в этом разделе, в хранилище задач хранятся в двух форматах: в файле XML и реестре. В этом эксперименте вы познакомитесь с обоими. Сначала откройте апплет Планировщик заданий (Task Scheduler), введя `Taskschd.msc` в поле поиска Cortana. Разверните узел Библиотека планировщика заданий (Task Scheduler Library) и все его подузлы, пока не дойдете до папки `Microsoft\Windows`. Исследуйте каждый подузел и найдите задачу, для которой

на вкладке Действия (Actions) установлено значение Пользовательский обработчик (Custom Handler). Тип действия используется для описания задач, размещенных в COM, которые не поддерживаются апплетом планировщика задач. В этом примере мы рассматриваем `ProcessMemoryDiagnosticEvents`, который можно найти в папке `MemoryDiagnostics`, но любая задача, для действия которой установлено значение Пользовательский обработчик (Custom Handler), работает хорошо.



Откройте окно администраторской командной строки, введя `CMD` в поле поиска Cortana и выбрав Запустить от имени администратора (Run As Administrator), затем введите следующую команду (заменив путь к задаче на выбранный вами):

```
schtasks /query /tn «Microsoft\Windows\MemoryDiagnostic\
ProcessMemoryDiagnosticEvents» /xml
```

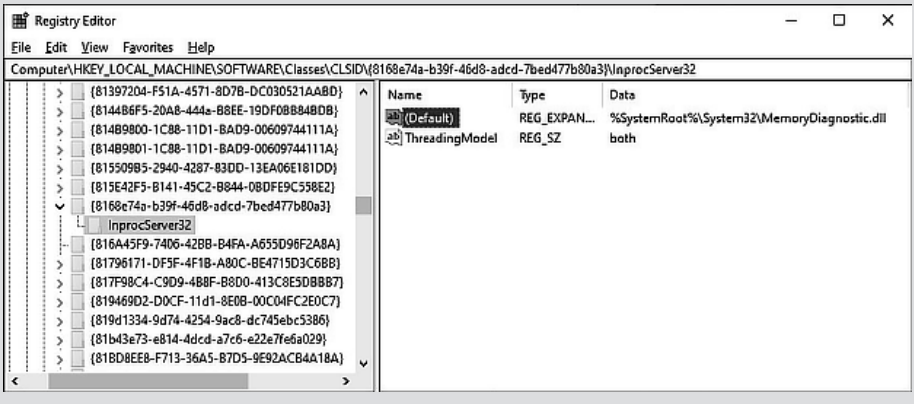
В выходных данных показан XML-дескриптор задачи, который включает в себя дескриптор безопасности задачи (применяемый для ее защиты от открытия неавторизованными пользователями), автора и описание задачи, субъекта безопасности, который должен ее запустить, настройки задачи, ее триггеры, а также действия:

```
<?xml version="1.0" encoding="UTF-16"?>
<Task xmlns="http://schemas.microsoft.com/windows/2004/02/mit/task">
  <RegistrationInfo>
    <Version>1.0</Version>
    <SecurityDescriptor>D:P(A;;FA;;;BA)(A;;FA;;;SY)(A;;FR;;;AU)</SecurityDescriptor>
    <Author>$(@%SystemRoot%\system32\MemoryDiagnostic.dll,-600)</Author>
```

```
    <Description>$(@%SystemRoot%\system32\MemoryDiagnostic.dll,-603)</Description>
    <URI>\Microsoft\Windows\MemoryDiagnostic\ProcessMemoryDiagnosticEvents</URI>
  </RegistrationInfo>
  <Principals>
    <Principal id="LocalAdmin">
      <GroupId>S-1-5-32-544</GroupId>
      <RunLevel>HighestAvailable</RunLevel>
    </Principal>
  </Principals>
  <Settings>
    <AllowHardTerminate>false</AllowHardTerminate>
    <DisallowStartIfOnBatteries>true</DisallowStartIfOnBatteries>
    <StopIfGoingOnBatteries>true</StopIfGoingOnBatteries>
    <Enabled>false</Enabled>
    <ExecutionTimeLimit>PT2H</ExecutionTimeLimit>
    <Hidden>true</Hidden>
    <MultipleInstancesPolicy>IgnoreNew</MultipleInstancesPolicy>
    <StartWhenAvailable>true</StartWhenAvailable>
    <RunOnlyIfIdle>true</RunOnlyIfIdle>
    <IdleSettings>
      <StopOnIdleEnd>true</StopOnIdleEnd>
      <RestartOnIdle>true</RestartOnIdle>
  </IdleSettings>
  <UseUnifiedSchedulingEngine>true</UseUnifiedSchedulingEngine>
</Settings>
<Triggers>
  <EventTrigger>
    <Subscription><QueryList><Query Id="0" Path="System"><Select Path=
     "System">*[System[Provider[@Name='Microsoft-Windows-WER-SystemErrorReporting']
     and (EventID=1000 or EventID=1001 or EventID=1006)]]</Select></Query>
     </QueryList></Subscription>
    </EventTrigger>
    . . . [часть кода удалена для краткости] . . .
  </Triggers>
  <Actions Context="LocalAdmin">
    <ComHandler>
      <ClassId>{8168E74A-B39F-46D8-ADCD-7BED477B80A3}</ClassId>
      <Data><![CDATA[Event]]></Data>
    </ComHandler>
  </Actions>
</Task>
```

В случае задачи `ProcessMemoryDiagnosticEvents` существуют несколько триггеров ETW, которые позволяют выполнять ее только при возникновении определенных диагностических событий. (Действительно, дескрипторы триггера включают запрос ETW, указанный в формате XPath.) Единственное зарегистрированное действие — это ComHandler, которое включает только CLSID (идентификатор класса) COM-объекта, представляющего задачу. Откройте редактор реестра и перейдите к разделу HKEY_LOCAL_MACHINE\SOFTWARE\Classes\CLSID. Выберите Найти (Find) в меню Правка (Edit), скопируйте и вставьте CLSID, расположенный после XML-тега ClassID дескриптора задачи (с фигурными скобками или без них). Вы сможете найти DLL, реализующую интерфейс ITaskHandler,

представляющий задачу, которая будет размещаться клиентским приложением Task Host (Taskhostw.exe), описанным далее в разделе «Клиент узла задач».



Если перейдете в раздел реестра HKLM\SOFTWARE\Microsoft\Windows NT\CurrentVersion\Schedule\TaskCache\Tasks, то сможете также увидеть GUID дескриптора задачи, хранящегося в кэше хранилища задач. Чтобы найти его, следует выполнить поиск, используя URI задачи. Действительно, GUID задачи не хранится в файле конфигурации XML. Данные, принадлежащие дескриптору задачи в реестре, идентичны тем, которые хранятся в файле конфигурации XML, расположенном в теневой копии хранилища (%systemroot%\System32\Tasks\Microsoft\Windows\MemoryDiagnostic\ProcessMemoryDiagnosticEvents). Меняется только двоичный формат, в котором они хранятся.

Включенные задачи должны быть зарегистрированы в UBPM. Планировщик задач вызывает функцию `RegisterTask` прокси-сервера Ubpm, которая сначала подключается к хранилищу учетных данных для получения таких данных, используемых для запуска задачи, а затем обрабатывает список всех действий и триггеров (хранится во внутреннем списке), преобразуя их в формат, понятный UBPM. Наконец, он вызывает API `UbpmTriggerConsumerRegister`, экспортированный из UBPM.dll. Задача готова к выполнению, когда проверена правильность условий.

Унифицированный диспетчер фоновых процессов

Традиционно UBPM отвечал в основном за управление жизненными циклами и состояниями задач (запуск, остановка, включение/выключение и т. д.), а также за обеспечение поддержки уведомлений и триггеров. В Windows 8.1 появилась инфраструктура брокеров, и все управление триггерами и уведомлениями было передано разным брокерам, которые могут использоваться как современными, так и стандартными приложениями Win32. Таким образом, в Windows 10 UBPM действует как прокси-сервер для стандартных триггеров задач Win32 и переадресует

запрос потребителей триггера правильному брокеру. UBPM по-прежнему отвечает за предоставление COM API, доступных приложениям для следующих действий:

- регистрации и отмены регистрации триггерного потребителя, а также открытия и закрытия его дескриптора;
- создания уведомления или триггера;
- отправки команды поставщику триггеров.

Подобно архитектуре планировщика задач, UBPM состоит из различных внутренних компонентов: сервера и клиента узла задач, библиотеки узла задач на базе COM и диспетчера событий.

Хост-сервер задач

Когда один из системных брокеров путем публикации изменения состояния WNF вызывает событие, зарегистрированное потребителем триггера UBPM, реализуется функция обратного вызова `UbpmTriggerArrived`. UBPM выполняет поиск во внутреннем списке зарегистрированных триггеров задачи на основе имени состояния WNF и, найдя правильный, обрабатывает действия задачи. На момент написания этой книги поддерживался только запуск исполняемого файла. Это актуально как для размещенных, так и для неразмещенных файлов. Неразмещенные исполняемые файлы — это обычные исполняемые файлы Win32, которые не взаимодействуют напрямую с UBPM, размещенные исполняемые файлы — это COM-классы, которые напрямую взаимодействуют с UBPM и должны располагаться в клиентском процессе узла задачи. После запуска размещенного исполняемого файла (taskhostw.exe) он может выполнять различные задачи, что зависит от связанного с ним токена. (Размещенные исполняемые файлы очень похожи на общие службы Svchost.)

Как и SCM, UBPM поддерживает различные типы токенов безопасности входа в систему для хост-процессов задачи. Функция `UbpmTokenGetTokenForTask` может создать новый токен на основе информации об учетной записи, хранящейся в дескрипторе задачи. Токен безопасности, созданный UBPM для задачи, может иметь одного из следующих владельцев: зарегистрированную учетную запись пользователя, учетную запись виртуальной службы, учетную запись сетевой службы или учетную запись локальной службы. В отличие от SCM, UBPM полностью поддерживает интерактивные токены. UBPM задействует службы, предоставляемые диспетчером пользователей (Usermgr.dll), для перечисления активных в данный момент интерактивных сеансов. Для каждого сеанса он сравнивает SID пользователя, указанный в дескрипторе задачи, с владельцем интерактивного сеанса. Если они совпадают, UBPM дублирует токен, прикрепленный к интерактивному сеансу, и задействует его для входа в новый исполняемый файл. В результате интерактивные задачи могут выполняться только с помощью стандартной учетной записи пользователя. (Неинтерактивные задачи могут выполняться со всеми типами учетных записей, перечисленными ранее.)

После создания токена UBPM запускает хост-процесс задачи. Если это размещенная задача COM, то функция `UbpmFindHost` выполняет поиск во внутреннем списке экземпляров процесса Taskhostw.exe (клиент узла задачи). Если он находит процесс, который выполняется с тем же контекстом безопасности, что и новая задача, то просто отправляет команду Запустить задачу (Start Task) (она включает имя

задачи COM и CLSID) через локальное соединение RPC узла задачи и ждет первого ответа. Клиентский процесс узла задачи и UBPM соединяются через статический канал RPC с именем `ubpmtaskhostchannel` и используют протокол соединения, аналогичный реализованному в SCM.

Если совместимый экземпляр клиентского процесса не найден или хост-процесс задачи является обычным исполняемым файлом, отличным от COM, UBPM создает новый блок среды, анализирует командную строку и создает новый процесс в приостановленном состоянии с помощью API `CreateProcessAsUser`. UBPM запускает хост-процесс каждой задачи в объекте задания, что позволяет быстро устанавливать состояние нескольких задач и точно настраивать ресурсы, выделенные для фоновых задач. UBPM выполняет поиск во внутреннем списке объектов заданий, содержащих хост-процессы, принадлежащие к одному и тому же идентификатору сеанса и задачам одного и того же типа (обычные, критические, на основе COM или неразмещенные). Если он находит совместимое задание, то просто назначает ему новый процесс с помощью API `AssignProcessToJobObject`. В противном случае он создает новый процесс и добавляет его в свой внутренний список.

После создания объекта задания задача наконец готова к запуску — поток исходного процесса возобновляется. Для задач, размещенных на COM, UBPM ожидает первоначального контакта от клиента узла задач (выполняется, когда клиент хочет открыть канал связи RPC с UBPM, аналогично тому, как приложения управления службами открывают канал к SCM) и отправляет команду Запустить задачу (Start Task). Наконец, UBPM регистрирует обратный вызов ожидания в хост-процессе задачи, что позволяет ему обнаружить неожиданное завершение последнего.

Клиент узла задач

Клиентский процесс узла задач получает команды от UBPM (сервера узла задач), находящегося в службе планировщика задач. Во время инициализации он открывает локальный интерфейс RPC, созданный UBPM во время его инициализации, и работает вечно, ожидая поступления команд по каналу. В настоящее время поддерживаются четыре команды, которые отправляются через API RPC `TaskHostSendResponseReceiveCommand`:

- остановка хоста;
- запуск задачи;
- остановка задачи;
- завершение задачи;

Все команды, основанные на задачах, реализуются внутри общей библиотеки задач COM и, по сути, приводят к созданию и уничтожению компонентов COM. В частности, размещенные задачи — это COM-объекты, наследующие от интерфейса ITaskHandler. Последний предоставляет только четыре обязательных метода, которые соответствуют переходам состояний различных задач: Start, Stop, Pause и Resume. Когда UBPM отправляет команду на запуск задачи своему клиентскому хост-процессу (Taskhostw.exe), последний создает новый поток для этой задачи. Новый рабочий поток задачи использует функцию `CoCreateInstance` для создания экземпляра COM-объекта ITaskHandler, представляющего задачу, и вызывает его метод Start. UBPM точно знает, какой CLSID (уникальный идентификатор класса)

идентифицирует конкретную задачу: CLSID задачи хранится в конфигурации задачи и указывается во время ее регистрации. Кроме того, размещенные задачи применяют функции, предоставляемые COM-интерфейсом `ITaskHandlerStatus`, для уведомления UBPM о своем состоянии выполнения. Интерфейс использует RPC для того, чтобы вызвать `UbpmReportTaskStatus` и сообщить UBPM о новом состоянии.

ЭКСПЕРИМЕНТ. Наблюдение за задачей на базе COM

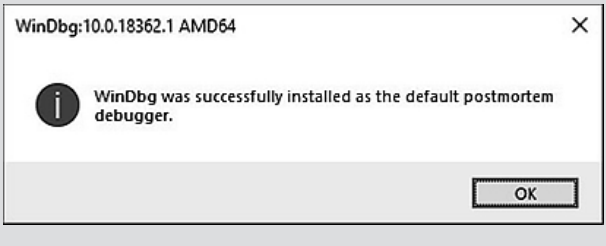
В этом эксперименте вы станете свидетелем того, как клиентский процесс узла задачи загружает DLL COM-сервера, реализующую задачу. Для него потребуются инструменты отладки, установленные в вашей системе. (Инструменты отладки можно найти в составе Windows SDK, который доступен по адресу https://developer.microsoft.com/en-us/windows/downloads/windows-10-sdk/.) Вы создадите точку останова отладчика в процедуре запуска задачи, для чего выполните следующее.

1. Настройте Windbg в качестве отладчика по умолчанию после остановки. (Можете пропустить этот шаг, если к целевой системе подключен отладчик ядра.) Для этого откройте административную командную строку и введите следующие команды:

   ```
   cd "C:\Program Files (x86)\Windows Kits\10\Debuggers\x64"
   Windbg.exe /I
   ```

   Обратите внимание: C:\Program Files (x86)\Windows Kits\10\Debuggers\x64 — это путь к инструментам отладки, который может меняться в зависимости от версии отладчика и программы установки.

2. Windbg должен запуститься и вывести следующее сообщение, подтверждающее успех операции.



3. После того как вы нажмете кнопку OK, WinDbg должен автоматически закрыться.
4. Откройте апплет Планировщик заданий (Task Scheduler), введя `Taskschd.msc` в командной строке.
5. Обратите внимание: если не подключен отладчик ядра, вы не сможете включить точку останова начальной задачи для неинтерактивных задач, иначе не получилось бы взаимодействовать с окном отладчика, которое будет открыто в другом неинтерактивном сеансе.

6. Изучая различные задачи (дополнительную информацию см. в предыдущем эксперименте «Обзор XML-дескриптора задачи»), вы должны найти интерактивную задачу COM с именем `CacheTask` по пути `\Microsoft\Windows\Wininet`. Помните, что на странице Действия (Actions) задачи должен отображаться Пользовательский обработчик (Custom Handler), в противном случае задача не является задачей COM.
7. Откройте редактор реестра, введя `regedit` в окне командной строки, и перейдите к разделу реестра HKLM\SOFTWARE\Microsoft\Windows NT\CurrentVersion\Schedule.
8. Щелкните правой кнопкой мыши на разделе Schedule и создайте новый параметр реестра, выбрав Мультистроковый параметр (Multi-String Value) в меню Создать (New).
9. Назовите новый параметр реестра `EnableDebuggerBreakForTaskStart`. Чтобы включить первоначальную точку останова задачи, вам следует вставить полный путь к задаче. В этом случае полный путь — `\Microsoft\Windows\Wininet\CacheTask`. В предыдущем эксперименте путь к задаче назывался URI задачи.
10. Закройте редактор реестра и вернитесь в планировщик задач.
11. Щелкните правой кнопкой мыши на задаче CacheTask и выберите Выполнить (Run).
12. Если вы все настроили правильно, должно появиться новое окно WinDbg.
13. Настройте символы, используемые отладчиком, выбрав пункт Путь к файлу символов (Symbol File Path) в меню Файл (File) и вставив действительный путь к серверу символов Windows (для получения более подробной информации см. https://docs.microsoft.com/en-us/windows-hardware/drivers/debugger/microsoft-public-symbols).
14. Вы сможете просмотреть стек вызовов процесса Taskhostw.exe непосредственно перед его прерыванием с помощью команды k:

```
0:000>k
# Child-SP          RetAddr           Call Site
00 000000a7`01a7f610 00007ff6`0b0337a8 taskhostw!ComTaskMgrBase::[ComTaskMgr]::
StartComTask+0x2c4
01 000000a7`01a7f960 00007ff6`0b033621 taskhostw!StartComTask+0x58
02 000000a7`01a7f9d0 00007ff6`0b033191 taskhostw!UbpmTaskHostWaitForCommands+0x2d1
3 000000a7`01a7fb00 00007ff6`0b035659 taskhostw!wWinMain+0xc1
04 000000a7`01a7fb60 00007ffa`39487bd4 taskhostw!__wmainCRTStartup+0x1c9
05 000000a7`01a7fc20 00007ffa`39aeced1 KERNEL32!BaseThreadInitThunk+0x14
06 000000a7`01a7fc50 00000000`00000000 ntdll!RtlUserThreadStart+0x21
```

15. Стек показывает, что клиент узла задачи только что был создан UBPM и получил команду Start, запрашивающую запуск задачи.
16. В консоли Windbg вставьте команду `~.` и нажмите Enter. Обратите внимание на текущий идентификатор выполняющегося потока.
17. Теперь вам следует поставить точку останова на COM API `CoCreateInstance` и возобновить выполнение, используя следующие команды:

```
bp combase!CoCreateInstance
g
```

18. После того как отладчик прервется, снова вставьте команду `~.` в консоли Windbg, нажмите Enter и обратите внимание на то, что идентификатор потока полностью изменился.
19. Это показывает, что клиент узла задачи создал новый поток для выполнения точки входа в задачу. Документированная функция `CoCreateInstance` используется для создания одного COM-объекта класса, связанного с определенным CLSID, указанным как параметр. Для этого эксперимента интересны два GUID: GUID класса COM, который представляет задачу, и идентификатор интерфейса, реализованного COM-объектом.
20. В 64-битных системах соглашение о вызовах определяет, что первые четыре параметра функции передаются через регистры, поэтому эти GUID легко получить:

```
0:004> dt combase!CLSID @rcx
 {0358b920-0ac7-461f-98f4-58e32cd89148}
   +0x000 Data1     : 0x358b920
   +0x004 Data2     : 0xac7
   +0x006 Data3     : 0x461f
   +0x008 Data4     : [8]  "???"
0:004> dt combase!IID @r9
 {839d7762-5121-4009-9234-4f0d19394f04}
   +0x000 Data1     : 0x839d7762
   +0x004 Data2     : 0x5121
   +0x006 Data3     : 0x4009
   +0x008 Data4     : [8]  "???"
```

Как видно из предыдущего вывода, CLSID COM-сервера — {0358b920-0ac7-461f-98f4-58e32cd89148}. Вы можете убедиться, что он соответствует GUID единственного действия COM, расположенного в XML-дескрипторе задачи CacheTask (подробности см. в предыдущем эксперименте). Запрошенный идентификатор интерфейса — {839d7762-5121-4009-9234-4f0d19394f04}, который соответствует GUID интерфейса действия обработчика задач COM (ITaskHandler).

COM-интерфейсы планировщика задач

Как мы обсуждали в предыдущем разделе, задача COM должна соответствовать четко определенному интерфейсу, который используется UBPM для управления переходом состояний задачи. В то время как UBPM решает, когда запускать задачу, и управляет всем ее состоянием, все остальные интерфейсы, применяемые для регистрации, удаления или просто запуска и остановки задачи вручную, реализуются планировщиком задач в его клиентской DLL (Taskschd.dll).

`ITaskService` — это центральный интерфейс, с помощью которого клиенты могут подключаться к планировщику задач и выполнять различные операции, например, перечислять зарегистрированные задачи, получать экземпляр хранилища задач, представленный COM-интерфейсом `ITaskFolder`, а также включать, отключать, удалять или регистрировать задачу и все связанные с ней триггеры и действия с помощью COM-интерфейса `ITaskDefinition`. Когда клиентское приложение впервые вызывает API планировщика задач через COM, система

загружает клиентскую DLL планировщика задач (Taskschd.dll) в адресное пространство клиентского процесса (COM-объекты планировщика задач находятся на внутрипроцессном COM-сервере, что соответствует контракту COM). API COM реализуются путем маршрутизации запросов через вызовы RPC в службу планировщика задач, которая обрабатывает каждый запрос и при необходимости пересылает его в UBPM. Архитектура COM планировщика задач позволяет пользователям взаимодействовать с ним через языки сценариев, такие как PowerShell (через командлет `ScheduledTasks`) или VBScript.

ИНСТРУМЕНТАРИЙ УПРАВЛЕНИЯ WINDOWS

Инструментарий управления Windows (Windows Management Instrumentation, WMI) — это реализация системы управления предприятием на основе использовании веб-технологии Web-Based Enterprise Management (WBEM) — стандарта, определенного промышленным консорциумом Distributed Management Task Force (DMTF). Стандарт WBEM включает в себя разработку расширяемого корпоративного средства сбора данных и управления ими, которое обладает гибкостью и расширяемостью, необходимыми для управления локальными и удаленными системами, состоящими из произвольных компонентов.

Архитектура WMI

WMI состоит из четырех основных компонентов: приложений управления, инфраструктуры WMI, поставщиков и управляемых объектов (рис. 10.27). Приложения управления — это приложения Windows, которые получают доступ к данным об управляемых объектах, отображают или обрабатывают их. Простым примером приложения управления является замена инструмента производительности, которая для получения информации о производительности использует WMI, а не Performance API. Более сложный пример — инструмент управления предприятием, позволяющий администраторам выполнять автоматическую инвентаризацию конфигурации программного и аппаратного обеспечения каждого компьютера на предприятии.

Разработчикам обычно приходится использовать приложения управления для сбора данных и управления конкретными объектами. Объект может представлять один компонент, например сетевой адаптер, или набор компонентов, например компьютер. (Объект-компьютер может содержать объект сетевого адаптера.) Поставщикам необходимо определить и экспортировать представление объектов, которые интересуют приложения управления. Например, поставщик сетевого адаптера может захотеть добавить специфичные для адаптера свойства в поддержку WMI сетевого адаптера, которую включает в себя Windows, для запроса и настройки состояния и поведения адаптера в соответствии с указаниями приложений управления. В некоторых случаях, например для драйверов устройств, Microsoft предоставляет поставщика, имеющего собственный API, чтобы помочь разработчикам использовать реализацию поставщика для собственных управляемых объектов, прилагая минимальные усилия для написания кода.

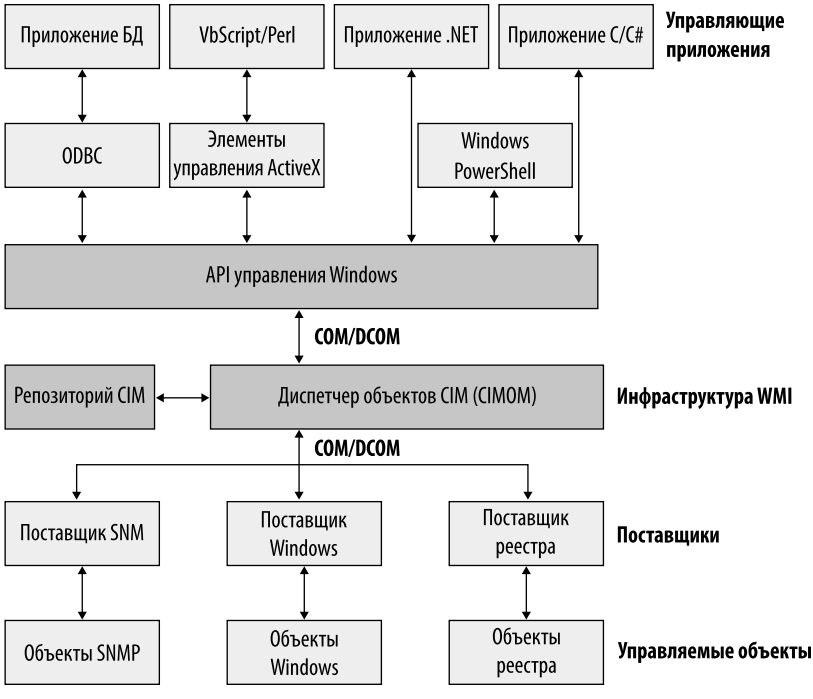


Рис. 10.27. Архитектура WMI

Инфраструктура WMI, центром которой является диспетчер объектов общей информационной модели (CIM) (CIMOM), — это связующее звено между приложениями управления и поставщиками. (CIM описана далее в этой главе.) Инфраструктура служит также хранилищем объектных классов и во многих случаях — диспетчером хранения постоянных свойств объектов. WMI реализует хранилище (репозиторий) в виде базы данных на диске, называемой репозиторием объектов CIMOM. В рамках своей инфраструктуры WMI поддерживает несколько API, посредством которых приложения управления получают доступ к данным объектов, а поставщики предоставляют данные и определения классов.

Программы и сценарии Windows, например Windows PowerShell, используют COM API WMI — основной API управления — для прямого взаимодействия с WMI. Другие API располагаются поверх COM API и включают адаптер Open Database Connectivity (ODBC) для приложения базы данных Microsoft Access. Разработчик базы данных задействует адаптер WMI ODBC для внедрения ссылок на данные объекта в базу данных разработчика. После этого разработчик сможет легко создавать отчеты с помощью запросов к базе данных, содержащих данные на основе WMI. Элементы управления ActiveX WMI поддерживают еще один многоуровневый API. Веб-разработчики используют элементы управления ActiveX для создания веб-интерфейсов для данных WMI. Еще один API управления — это API сценариев WMI, предназначенный для применения в приложениях на основе сценариев,

например Visual Basic Scripting Edition. Поддержка сценариев WMI существует для всех технологий языков программирования Microsoft.

Поскольку COM-интерфейсы WMI предназначены для приложений управления, они представляют собой основной API для поставщиков. Однако, в отличие от приложений управления, которые являются COM-клиентами, поставщики — это COM-серверы или серверы распределенного COM (DCOM), то есть они реализуют COM-объекты, с которыми взаимодействует WMI. Возможные варианты реализации поставщика WMI включают библиотеки DLL, которые загружаются в процесс диспетчера WMI, или автономные приложения Windows, или службы Windows. Microsoft включает в себя ряд встроенных поставщиков, которые предоставляют данные из известных источников, таких как Performance API, реестр, диспетчер событий, Active Directory, SNMP и современные драйверы устройств. WMI SDK позволяет разработчикам создавать сторонних поставщиков WMI.

Поставщики WMI

В основе WBEM лежит спецификация CIM, разработанная DMTF. CIM определяет, как системы управления представляют с точки зрения системного управления все, от компьютера до приложения или устройства на компьютере. Разработчики поставщиков используют CIM для представления компонентов — частей приложения, которым разработчики хотят разрешить управление. Разработчики используют язык Managed Object Format (MOF) для реализации представления CIM.

Помимо определения классов, представляющих объекты, поставщик должен связать WMI с объектами. WMI классифицирует поставщиков в соответствии с функциями интерфейса, предоставляемыми поставщиками. В табл. 10.14 приведена классификация поставщиков WMI. Обратите внимание: поставщик может реализовать одну или несколько функций, поэтому он может быть поставщиком, например, и классов, и событий. Чтобы определения функций в табл. 10.14 стали понятнее, давайте посмотрим на поставщика, который реализует некоторые из этих функций. Поставщик журнала событий поддерживает несколько объектов, включая компьютер журнала событий, запись журнала событий и файл журнала событий. Журнал событий является поставщиком экземпляров, поскольку он может определять несколько экземпляров для нескольких своих классов. Один из классов, для которого поставщик журнала событий определяет несколько экземпляров, — это класс файла журнала событий (`Win32_NTEventlogFile`). Поставщик журнала событий определяет экземпляр этого класса для каждого журнала событий системы: журнала системных событий, журнала событий приложений и журнала событий безопасности.

Поставщик журнала событий определяет данные экземпляра и позволяет приложениям управления перечислять записи. Чтобы приложения управления могли использовать WMI для резервного копирования и восстановления файлов журнала событий, поставщик последнего реализует методы резервного копирования и восстановления для объектов файлов журнала событий. Это действие классифицирует поставщика журнала событий как поставщика методов. Наконец, приложение управления может зарегистрироваться для получения уведомлений всякий раз, когда в один из журналов событий заносится новая запись. Таким образом, когда поставщик журнала событий использует уведомления о событиях, чтобы сообщить WMI о прибытии записей журнала событий, он выступает в качестве поставщика событий.

Таблица 10.14. Классификация поставщиков

Поставщик	Описание
Классов	Может предоставлять, изменять, удалять и перечислять классы, специфичные для поставщика. Также может поддерживать обработку запросов. Редкий пример службы, являющейся поставщиком классов, — Active Directory
Экземпляров	Может предоставлять, изменять, удалять и перечислять экземпляры системных классов и классов, специфичных для поставщика. Экземпляр представляет собой управляемый объект. Также может поддерживать обработку запросов
Свойств	Может предоставлять и изменять значения свойств отдельных объектов
Методов	Предоставляет методы для класса, специфичного для поставщика
Событий	Генерирует уведомления о событиях
Потребитель событий	Сопоставляет физического потребителя с логическим для поддержки уведомлений о событиях

Общая информационная модель и язык формата управляемых объектов

Общая информационная модель (Common Information Model, CIM) следует по стопам объектно-ориентированных языков наподобие C++ и C#, где разработчик модели проектирует представления в виде классов. Работа с классами позволяет использовать мощные методы моделирования наследования и композиции. Подклассы могут наследовать атрибуты родительского класса, а также добавлять собственные характеристики или переопределять унаследованные. Класс, который наследует свойства другого класса, является производным от последнего. Кроме того, возможно создание композиций, то есть разработчик может создать класс, включающий в себя другие классы. Классы CIM состоят из свойств и методов. Свойства описывают конфигурацию и состояние ресурса, управляемого WMI, а методы представляют исполняемые функции, которые позволяют выполнять действия с ним.

DMTF предоставляет несколько классов как часть стандарта WBEM. Эти классы составляют основу языка CIM и представляют собой объекты, применимые ко всем областям управления. Они являются частью базовой модели CIM. Пример базового класса — `CIM_ManagedSystemElement`. Он содержит несколько основных свойств, которые идентифицируют физические компоненты, например аппаратные устройства, и логические компоненты, такие как процессы и файлы. Каждое свойство включает в себя заголовок, описание, дату установки и статус. Например, классы `CIM_LogicalElement` и `CIM_PhysicalElement` наследуют атрибуты класса `CIM_ManagedSystemElement`. Они тоже входят в базовую модель CIM. Стандарт WBEM называет эти классы абстрактными, поскольку они существуют исключительно как родительские для других классов (иными словами, экземпляров объектов абстрактного класса не существует). Поэтому абстрактные классы можно воспринимать как шаблоны, определяющие свойства для использования в других классах.

Вторая категория классов представляет объекты, специфичные для областей управления, но независимые от конкретной реализации. Эти классы составляют общую модель и считаются расширением базовой модели. Пример класса общей модели — `CIM_FileSystem`, наследующий атрибуты `CIM_LogicalElement`. Поскольку

практически каждая операционная система, включая Windows, Linux и другие разновидности UNIX, использует структурированное хранилище на основе файловой системы, класс `CIM_FileSystem` является подходящим компонентом общей модели.

Последняя категория классов — расширенная модель — включает в себя технологические дополнения к общей модели. Windows определяет большой набор таких классов для представления объектов, специфичных для среды Windows. Поскольку все операционные системы хранят данные в файлах, в модель CIM входит класс `CIM_LogicalFile`. От последнего наследует класс `CIM_DataFile`, а Windows добавляет классы файлов `Win32_PageFile` и `Win32_ShortcutFile` для соответствующих типов файлов.

Windows включает в себя различные приложения управления WMI, которые позволяют администратору взаимодействовать с классами и пространствами имен WMI. Утилита командной строки WMI (WMIC.exe) и Windows PowerShell могут подключаться к WMI, выполнять запросы и вызывать методы объектов классов WMI. На рис. 10.28 показано окно PowerShell, где извлекается информация из класса `Win32_NTEventlogFile`, входящего в состав поставщика журнала событий. Этот класс является производным от `CIM_DataFile` и широко использует наследование от него. Файлы журнала событий — это файлы данных, имеющие дополнительные атрибуты, специфичные для журнала событий, такие как имя файла журнала (`LogfileName`) и количество записей, содержащихся в файле (`NumberOfRecords`). Класс `Win32_NTEventlogFile` основан на нескольких уровнях наследования, где `CIM_DataFile` является потомком `CIM_LogicalFile`, который, в свою очередь, происходит от `CIM_LogicalElement`, а он — от класса `CIM_ManagedSystemElement`.

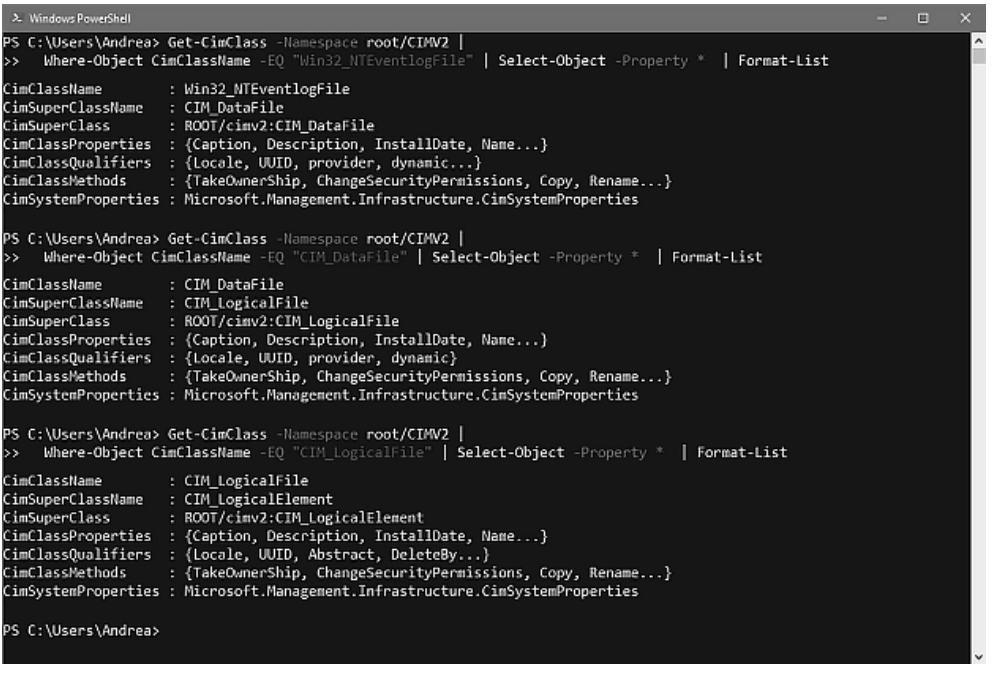


Рис. 10.28. Извлечение информации из класса Win32_NTEventlogFile посредством Windows PowerShell

Как говорилось ранее, разработчики поставщиков WMI создают свои классы на языке формата управляемых объектов (Managed Object Format, MOF). Следующий вывод текста показывает определение `Win32_NTEventlogFile` поставщика журнала событий, которое было запрошено на рис. 10.28:

```
[dynamic: ToInstance, provider("MS_NT_EVENTLOG_PROVIDER"): ToInstance,
SupportsUpdate,
Locale(1033): ToInstance, UUID("{8502C57B-5FBB-11D2-AAC1-006008C78BC7}"): ToInstance]
class Win32_NTEventlogFile : CIM_DataFile
{
    [Fixed: ToSubClass, read: ToSubClass] string LogfileName;
    [read: ToSubClass, write: ToSubClass] uint32 MaxFileSize;
    [read: ToSubClass] uint32 NumberOfRecords;
    [read: ToSubClass, volatile: ToSubClass, ValueMap{"0", "1..365", "4294967295"}:
     ToSubClass] string OverWritePolicy;
    [read: ToSubClass, write: ToSubClass, Range("0-365 | 4294967295"): ToSubClass]
     uint32 OverwriteOutDated;
    [read: ToSubClass] string Sources[];
    [ValueMap{"0", "8", "21", ".."}: ToSubClass, implemented, Privileges{
     "SeSecurityPrivilege", "SeBackupPrivilege"}: ToSubClass]
      uint32 ClearEventlog([in] string ArchiveFileName);
    [ValueMap{"0", "8", "21", "183", ".."}: ToSubClass, implemented, Privileges{
     "SeSecurityPrivilege", "SeBackupPrivilege"}: ToSubClass]
      uint32 BackupEventlog([in] string ArchiveFileName);
};
```

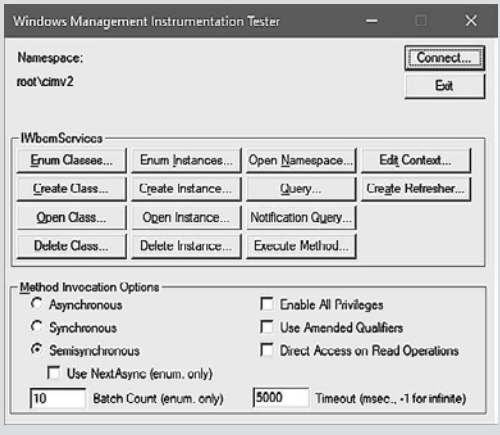
Стоит обратить внимание на еще один термин — «динамический», который является описательным обозначением класса `Win32_NTEventlogFile`, показанного в файле MOF в текстовом выводе ранее. Присутствие этого термина означает, что инфраструктура WMI запрашивает у поставщика WMI значения свойств, связанных с определением этого класса, всякий раз, когда приложение управления запрашивает свойства объекта. Статический же класс находится в репозитории WMI, и инфраструктура WMI обращается туда для получения значений вместо того, чтобы запрашивать их у поставщика. Поскольку обновление репозитория — довольно дорогостоящая операция, динамические поставщики более эффективны для объектов, свойства которых часто меняются.

ЭКСПЕРИМЕНТ. Просмотр определений на языке MOF у классов WMI

Вы можете просмотреть определение MOF для любого класса WMI с помощью утилиты тестирования инструментария управления Windows (WbemTest), поставляемой в составе Windows. В данном эксперименте рассмотрим определение MOF для класса `Win32_NTEventLogFile`.

1. Введите `Wbemtest` в поле поиска Cortana и нажмите Enter. Должен открыться тестер инструментария управления Windows.

2. Нажмите кнопку Подключиться (Connect), измените пространство имен на `root\cimv2` и подключитесь. Инструмент должен включить все командные кнопки, как показано на следующем снимке экрана.



3. Нажмите кнопку Перечислить классы (Enum Classes), выберите опцию Рекурсивный (Recursive) и нажмите OK.
4. Найдите `Win32_NTEventLogFile` в списке классов, а затем дважды щелкните на нем, чтобы просмотреть свойства его класса.
5. Нажмите кнопку Показать MOF (Show MOF), чтобы открыть окно с текстом MOF.

После создания классов в MOF разработчики WMI могут передать их определения в WMI несколькими способами. Разработчики драйверов WDM компилируют файл MOF в двоичный файл MOF (BMF) — более компактное двоичное представление — и могут динамически передавать файлы BMF в инфраструктуру WDM или статически включать их в свой двоичный файл. Другой способ — поставщик компилирует MOF и использует API-интерфейсы WMI COM для предоставления определений инфраструктуре WMI. Наконец, поставщик может задействовать утилиту MOF Compiler (Mofcomp.exe), чтобы напрямую предоставлять инфраструктуре WMI скомпилированное представление классов.

ПРИМЕЧАНИЕ Предыдущие выпуски Windows (до Windows 7) предоставляли графический инструмент под названием WMI CIM Studio, поставлявшийся вместе с инструментом администрирования WMI. Он мог графически отображать пространства имен, классы, свойства и методы WMI. В настоящее время этот инструмент не поддерживается и недоступен для загрузки, поскольку был заменен возможностями WMI в рамках Windows PowerShell. PowerShell предоставляет язык сценариев, не работающий с графическим интерфейсом. Некоторые сторонние инструменты имеют интерфейс, аналогичный CIM Studio. Один из таких — WMI Explorer, который можно загрузить с https://github.com/vinaypamnani/wmie2/releases.

Репозиторий общей информационной модели (CIM) хранится по пути `%SystemRoot%\System32\wbem\Repository` и включает в себя:

- **Index.btr** — индексный файл двоичного дерева (btree);
- **MappingX.map** — файлы управления транзакциями (X — число, начиная с 1);
- **Objects.data** — репозиторий CIM, в котором хранятся определения управляемых ресурсов.

Пространство имен WMI

Классы определяют объекты, предоставляемые поставщиком WMI. Объекты — это экземпляры классов в системе. WMI использует пространство имен, содержащее несколько подпространств имен, которые WMI иерархически упорядочивает для организации объектов. Приложение управления должно подключиться к пространству имен, прежде чем сможет получить доступ к объектам в нем.

WMI называет корневой каталог пространства имен ROOT. Все установки WMI имеют четыре предопределенных пространства имен, которые находятся в корневом каталоге: CIMV2, Default, Security и WMI. Внутри некоторых из этих пространств имен есть другие пространства имен. Например, CIMV2 включает пространства имен Applications и ms_409 в качестве подпространств. Иногда поставщики определяют собственные пространства имен, вы можете увидеть пространство имен WMI (которое определяет поставщик WMI драйвера устройства Windows) под ROOT в Windows.

В отличие от пространства имен файловой системы, имеющего иерархию каталогов и файлов, пространство имен WMI имеет только один уровень. Вместо использования имен, как это происходит в файловой системе, WMI применяет свойства объектов в качестве ключей для их идентификации. Приложения управления указывают имена классов с именами ключей для поиска конкретных объектов в пространстве имен. Таким образом, каждый экземпляр класса должен однозначно идентифицироваться по параметрам его ключей. Например, поставщик журнала событий применяет класс `Win32_NTLogEvent` для представления записей в журнале событий. У этого класса есть два ключа: `Logfile` — строка и `RecordNumber` — целое число без знака. Приложение управления, которое запрашивает у WMI экземпляры записей журнала событий, получает их от поставщика по паре ключей, идентифицирующих записи. Приложение обращается к записи, используя синтаксис, который вы можете увидеть в следующем примере пути к объекту:

```
\\ANDREA-LAPTOP\root\CIMV2:Win32_NTLogEvent.Logfile="Application", RecordNumber="1"
```

Первый компонент имени (`\\ANDREA-LAPTOP`) идентифицирует компьютер, на котором расположен объект, а второй компонент (`\root\CIMV2`) — это пространство имен, в котором находится объект. Имя класса стоит после двоеточия, а имена ключей и связанные с ними параметры — после точки. Запятая разделяет параметры ключей.

WMI предоставляет интерфейсы, которые позволяют приложениям перечислять все объекты определенного класса или выполнять запросы, возвращающие экземпляры класса, соответствующие указанному критерию.

Связи классов

Многие типы объектов тем или иным образом связаны друг с другом. Например, объект-компьютер имеет процессор, программное обеспечение, операционную систему, активные процессы и т. д. WMI позволяет поставщикам создавать классы-ассоциации для представления логической связи между двумя разными классами.

Классы-ассоциации связывают один класс с другим, поэтому имеют только два свойства: имя класса и модификатор Ref. Следующий текстовый вывод показывает связь, в рамках которой файл MOF поставщика журнала событий ассоциирует класс `Win32_NTLogEvent` с классом `Win32_ComputerSystem`. Получив объект, приложение управления может запрашивать связанные с ним объекты. Таким образом, поставщик может определять иерархию объектов:

```
[dynamic: ToInstance, provider("MS_NT_EVENTLOG_PROVIDER"): ToInstance,
EnumPrivileges{"SeSecurityPrivilege"}: ToSubClass,
Privileges{"SeSecurityPrivilege"}: ToSubClass, Locale(1033):
ToInstance, UUID("{8502C57F-5FBB-11D2-AAC1-006008C78BC7}"): ToInstance,
Association: DisableOverride ToInstance ToSubClass]
class Win32_NTLogEventComputer
{
    [key, read: ToSubClass] Win32_ComputerSystem ref Computer;
    [key, read: ToSubClass] Win32_NTLogEvent ref Record;
};
```

На рис. 10.29 показано окно PowerShell, отображающее первый экземпляр класса `Win32_NTLogEventComputer`, расположенный в пространстве имен CIMV2. Из экземпляра агрегированного класса пользователь может запросить связанный экземпляр объекта `Win32_ComputerSystem` WIN-46E4EFTBP6Q, который сгенерировал событие с номером записи 1031 в файле журнала приложения.

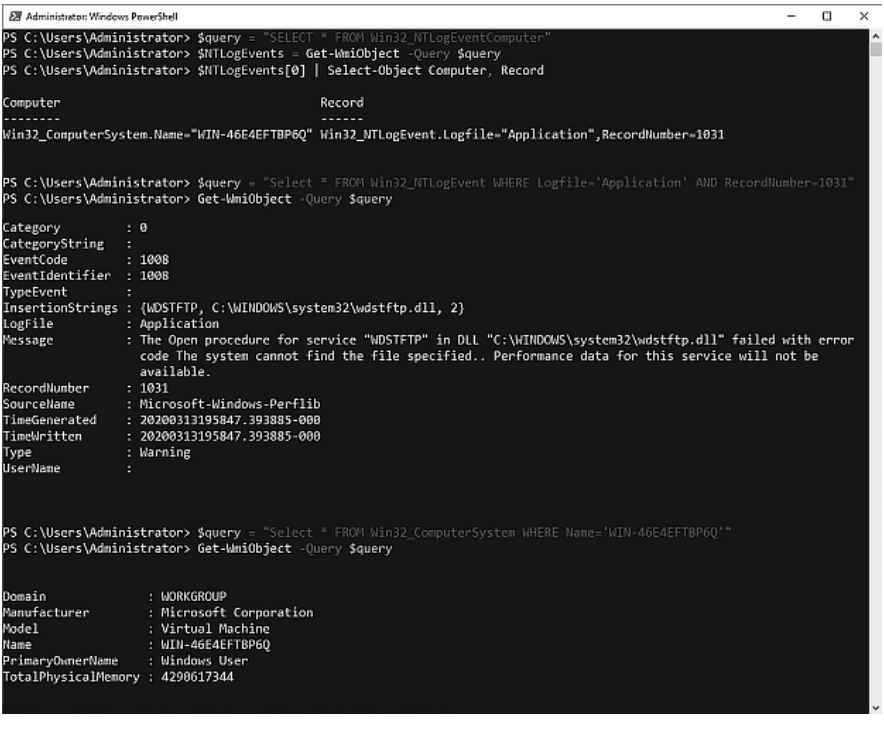


Рис. 10.29. Класс ассоциации Win32_NTLogEventComputer

ЭКСПЕРИМЕНТ. Использование сценариев WMI для управления системами

Мощным аспектом WMI является поддержка языков сценариев. Microsoft создала сотни сценариев, выполняющих общие административные задачи управления учетными записями пользователей, файлами, реестром, процессами и аппаратными устройствами. Веб-сайт Центра сценариев Microsoft TechNet служит центральным хранилищем сценариев Microsoft. Использовать сценарий из Центра сценариев так же просто, как скопировать его текст из интернет-браузера, сохранить его в файле с расширением .vbs и запустить с помощью команды `cscript script.vbs`, где `script` — это имя, которое вы ему дали. Cscript — это интерфейс командной строки для хоста сценариев Windows (WSH).

Вот пример сценария TechNet, который регистрируется для получения событий во время создания экземпляров объекта `Win32_Process`, что происходит при каждом запуске процесса, и выводит строку с именем процесса, представляющего объект:

```
strComputer = "."
Set objWMIService = GetObject("winmgmts:" _
    & "{impersonationLevel=impersonate}!\\" & strComputer & "\root\cimv2")
Set colMonitoredProcesses = objWMIService. _
    ExecNotificationQuery("SELECT * FROM __InstanceCreationEvent " _
        & " WITHIN 1 WHERE TargetInstance ISA 'Win32_Process'")
i = 0
Do While i = 0
    Set objLatestProcess = colMonitoredProcesses.NextEvent
    Wscript.Echo objLatestProcess.TargetInstance.Name
Loop
```

Строка, вызывающая `ExecNotificationQuery`, делает это с помощью параметра, включающего оператор `select`, который подчеркивает поддержку WMI, доступного только для чтения подмножества стандартного языка структурированных запросов (SQL) ANSI, известного как WQL, чтобы предоставить потребителям WMI гибкий способ указать информацию, которую они хотят получить от поставщиков WMI. Запуск примера сценария с помощью Cscript, а затем запуск Блокнота дает следующий вывод:

```
C:\>cscript monproc.vbs
Microsoft (R) Windows Script Host Version 5.812
Copyright (C) Microsoft Corporation. All rights reserved.

NOTEPAD.EXE
```

PowerShell поддерживает ту же функциональность с помощью команд `Register-WmiEvent` и `Get-Event`:

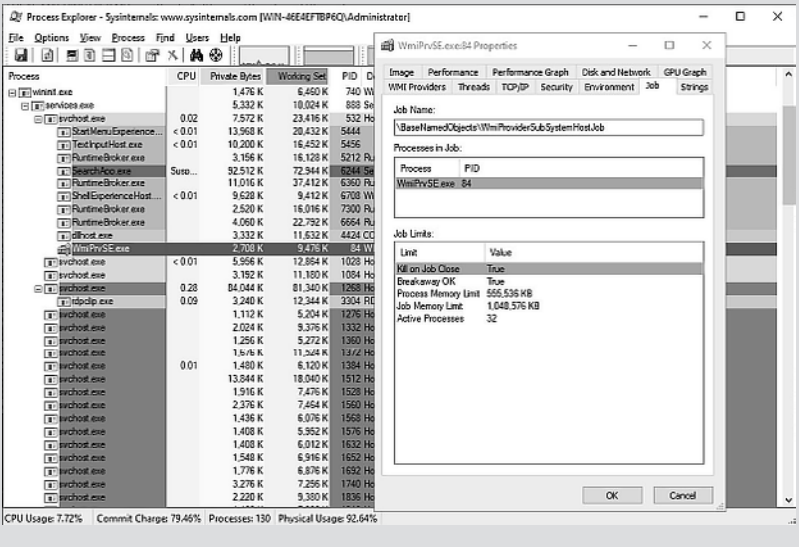
```
PS C:\> Register-WmiEvent -Query "SELECT * FROM __InstanceCreationEvent WITHIN 1
WHERE TargetInstance ISA 'Win32_Process'" -SourceIdentifier "TestWmiRegistration"

PS C:\> (Get-Event)[0].SourceEventArgs.NewEvent.TargetInstance | Select-Object
-Property ProcessId, ExecutablePath

ProcessId ExecutablePath
--------- --------------
    76016 C:\WINDOWS\system32\notepad.exe
PS C:\> Unregister-Event -SourceIdentifier "TestWmiRegistration"
```

Реализация WMI

Служба WMI действует в общем процессе Svchost, который выполняется под учетной записью локальной системы. Он загружает поставщиков в процесс их размещения WmiPrvSE.exe, который запускается как дочерний процесс средства запуска DCOM (служба RPC). WMI выполняет Wmiprvse в учетной записи локальной системы, локальной службы или сетевой службы в зависимости от значения свойства `HostingModel` экземпляра объекта WMI `Win32Provider`, который представляет реализацию поставщика. Процесс Wmiprvse завершается после удаления поставщика из кэша через одну минуту после получения последнего запроса от него.

ЭКСПЕРИМЕНТ. Наблюдение за созданием Wmiprvse

Вы можете увидеть, как создается WmiPrvSE, запустив Process Explorer и выполнив Wmic. Процесс WmiPrvSE появится под процессом Svchost, в котором размещается служба запуска DCOM. Если подсветка заданий Process Explorer включена, задание будет выделено цветом, поскольку, чтобы не дать неконтролируемому поставщику использовать все ресурсы виртуальной памяти в системе, Wmiprvse выполняется в объекте задания, который ограничивает количество дочерних процессов, которые он может создать, и объем виртуальной памяти, которую могут выделить все процессы задания. (Дополнительную информацию об объектах заданий см. в главе 5 тома 1.)



Большинство компонентов WMI по умолчанию находятся в %SystemRoot%\System32 и %SystemRoot%\System32\Wbem, включая файлы Windows MOF, встроенные библиотеки DLL поставщиков и библиотеки WMI приложений управления. Загляните в каталог %SystemRoot%\System32\Wbem, и вы найдете Ntevt.mof — MOF-файл

поставщика журнала событий. А также увидите Ntevt.dll — DLL поставщика журнала событий, которую использует служба WMI.

Поставщики обычно реализуются как динамически подключаемые библиотеки (DLL), дающие доступ к серверам COM, предоставляющим определенный набор интерфейсов (`IWbemServices` является центральным, обычно один поставщик реализуется как один COM-сервер). WMI включает множество встроенных поставщиков для семейства операционных систем Windows. Встроенные поставщики, также известные как стандартные, предоставляют данные и функции управления из известных источников операционной системы, таких как подсистема Win32, журналы событий, счетчики производительности и реестр. В табл. 10.15 перечислены некоторые стандартные поставщики WMI, входящие в состав Windows.

Таблица 10.15. Стандартные поставщики WMI, включенные в состав Windows

Поставщик	Бинарный образ	Пространство имен	Описание
Поставщик Active Directory	dsprov.dll	root\directory\ ldap	Сопоставляет объекты Active Directory с WMI
Поставщик журнала событий	ntevt.dll	root\cimv2	Управляет журналами событий Windows: чтение, резервное копирование, очистка, копирование, удаление, мониторинг, переименование, сжатие, распаковка и изменение настроек журнала событий
Поставщик счетчиков производительности	wbemperf.dll	root\cimv2	Предоставляет доступ к необработанным данным о производительности
Поставщик реестра	stdprov.dll	root\default	Считывает, записывает, перечисляет, отслеживает, создает и удаляет разделы и параметры в реестре
Поставщик виртуализации	vmmsprox.dll	root\ virtualization\v2	Предоставляет доступ к службам виртуализации, реализованным в vmms.exe, таким как управление виртуальными машинами в хост-системе и получение информации о периферийных устройствах хост-системы, из гостевой виртуальной машины
Поставщик WDM	wmiprov.dll	root\wmi	Предоставляет доступ к информации о драйверах устройств WDM
Поставщик Win32	cimwin32.dll	root\cimv2	Предоставляет информацию о компьютере, дисках, периферийных устройствах, файлах, папках, файловых системах, сетевых компонентах, операционной системе, принтерах, процессах, безопасности, службах, общих ресурсах, пользователях SAM, группах и т. д.
Поставщик установщика Windows	msiprov.dll	root\cimv2	Предоставляет доступ к информации об установленном программном обеспечении

Ntevt.dll, DLL поставщика журнала событий, представляет собой COM-сервер, зарегистрированный в разделе реестра HKLM\Software\Classes\CLSID с CLSID {F55C5B4C-517D-11d1-AB57-00C04FD9159E}. (Вы можете найти его в дескрипторе MOF.) В каталогах под %SystemRoot%\System32\Wbem хранятся репозиторий, файлы журналов и сторонние файлы MOF. WMI реализует репозиторий, называемый репозиторием объектов CIMOM, с использованием собственной версии СУБД Microsoft JET. Файл базы данных по умолчанию находится в папке SystemRoot%\System32\Wbem\Repository\.

WMI учитывает многочисленные параметры реестра, хранящиеся в разделе реестра службы HKLM\SOFTWARE\Microsoft\WBEM\CIMOM, такие как пороговые значения и максимальные значения для определенных параметров.

Драйверы устройств используют специальные интерфейсы для предоставления данных и приема команд, называемых командами управления (WMI System Control) системой WMI, из WMI. Эти интерфейсы — часть WDM, что объясняется в главе 6 тома 1. Поскольку интерфейсы являются кросс-платформенными, они относятся к пространству имен `\root\WMI`.

Безопасность WMI

WMI реализует безопасность на уровне пространства имен. Если приложение управления успешно подключается к пространству имен, то там ему разрешается просматривать все объекты и получать доступ к их свойствам. Администратор может задействовать приложение WMI Control, чтобы контролировать, какие пользователи могут получить доступ к пространству имен. На внутреннем уровне эта модель безопасности реализуется с помощью списков ACL и дескрипторов безопасности, которые являются частью стандартной модели безопасности Windows, реализующей проверки доступа. (Дополнительную информацию о проверках доступа см. в главе 7 тома 1.)

Чтобы запустить приложение WMI Control, откройте панель управления, введя Управление компьютером (Computer Management) в поле поиска Cortana. Затем откройте узел Службы и приложения (Services And Applications). Щелкните правой кнопкой мыши на элементе управления WMI и выберите Свойства (Properties), чтобы открыть диалоговое окно Свойства элемента управления WMI (WMI Control Properties) (рис. 10.30). Чтобы настроить безопасность пространств имен, перейдите на вкладку Безопасность (Security), выберите пространство имен и нажмите Безопасность (Security). Другие вкладки в диалоговом окне Свойства элемента управления WMI (WMI Control Properties) позволяют изменять параметры производительности и резервного копирования, хранящиеся в реестре.

ТРАССИРОВКА СОБЫТИЙ ДЛЯ WINDOWS

Трассировка событий для Windows (Event Tracing for Windows, ETW) — это основное средство, которое дает приложениям и драйверам режима ядра возможность предоставлять события журнала и трассировки, использовать их и управлять ими. События могут храниться в файле журнала или в циклическом буфере либо с ними можно работать в режиме реального времени. Их можно применять для отладки драйвера, платформы, такой как .NET CLR, или приложения, а также чтобы понять,

какие проблемы с производительностью возможны. Функция ETW в основном реализована в ядре NT, но приложение может задействовать и частные средства регистрации, которые вообще не переходят в режим ядра. Приложение, использующее ETW, может относиться к одной из следующих категорий.

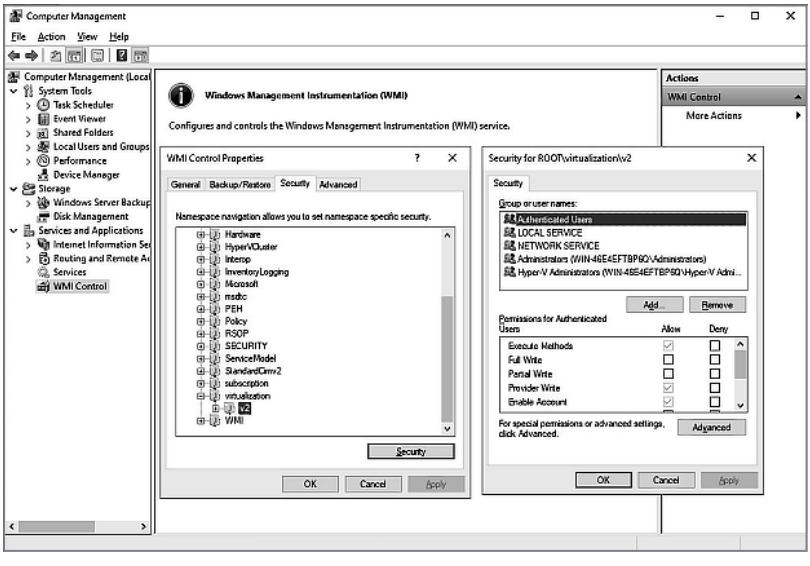


Рис. 10.30. Приложение Свойства элемента управления WMI и вкладка Безопасность пространства имен root\virtualization\v2

- **Контроллер.** Он запускает и останавливает сеансы отслеживания событий, управляет размером пулов буферов и позволяет поставщикам регистрировать события в сеансе. Примеры контроллеров — это монитор надежности и производительности и XPerf из набора средств Windows Performance Toolkit (теперь часть пакета Windows Assessment and Deployment Kit, доступного для загрузки по адресу https://docs.microsoft.com/en-us/windows-hardware/get-started/adk-install).
- **Поставщик.** Это приложение или драйвер, где представлены инструменты отслеживания событий. Поставщик регистрирует в ETW GUID поставщика (глобальный уникальный идентификатор), определяющий события, которые он может создавать. После регистрации поставщик способен генерировать события, которые могут быть включены или отключены приложением контроллера в соответствующем сеансе трассировки.
- **Потребитель.** Это приложение, которое выбирает один или несколько сеансов трассировки, для которых хочет прочитать данные трассировки. Потребители могут получать события, хранящиеся в файлах журналов, циклическом буфере или сеансах, которые доставляют события в реальном времени.

Важно отметить, что в ETW каждый поставщик, сеанс, признак и группа поставщика представлены GUID (дополнительная информация об этих концепциях — далее в этой главе). На базе ETW выстроены четыре технологии предоставления

событий. Они различаются в основном методом хранения и определения событий (хотя есть и другие различия).

- Поставщики MOF (или классические) устарели и используются, в частности, WMI. Поставщики MOF хранят дескрипторы событий в классах MOF, чтобы потребитель знал, как их применять.
- Поставщики WPP (программный обработчик трассировки Windows) используются для отслеживания операций приложения или драйвера (они являются расширением трассировки событий WMI) и задействуют файл формата сообщения трассировки TMF (trace message format), позволяющий потребителю декодировать события трассировки.
- Поставщики на основе манифеста используют файл манифеста XML для определения событий, которые могут быть декодированы потребителем.
- Поставщики TraceLogging, которые, как и поставщики WPP, используются для быстрого отслеживания работы приложения драйвера, задействуют события с самоописанием, которые содержат всю необходимую информацию для потребления контроллером.

Сразу после установки Windows уже включает в себя десятки поставщиков, которые используются каждым компонентом ОС для регистрации событий диагностики и трассировки производительности. Например, Hyper-V имеет несколько поставщиков, которые обеспечивают отслеживание событий для гипервизора, динамической памяти, драйвера Vid и стека виртуализации. Как показано на рис. 10.31, ETW реализован в ряде компонентов.

- Большая часть реализации ETW (создание глобального сеанса, регистрация и включение поставщика, основной поток журнала) находится в ядре NT.
- Библиотека API поиска узла для SCM/SDDL/LSA (sechost.dll) предоставляет приложениям основные API пользовательского режима, применяемые для создания сеанса ETW, включения поставщиков и потребления событий. Sechost использует службы, предоставляемые Ntdll, для вызова ETW в ядре NT. Некоторые API пользовательского режима ETW реализованы непосредственно в Ntdll, не раскрывая функциональность для Sechost. Регистрация поставщика и генерация событий являются примерами функций пользовательского режима, реализованных в Ntdll (но не в Sechost).
- Вспомогательная библиотека декодирования трассировки событий (TDH.dll) реализует службы, доступные потребителям для декодирования событий ETW.
- Библиотека потребления и настройки событий (WevtApi.dll) реализует API-интерфейсы журнала событий Windows, известные также как Evt API, которые доступны приложениям-потребителям для управления поставщиками и событиями на локальных и удаленных компьютерах. API журнала событий Windows поддерживают XPath 1.0 или структурированные XML-запросы, для анализа событий, создаваемых сеансом ETW.
- Ядро безопасности реализует базовые службы безопасности, способные взаимодействовать с ETW в ядре NT, которое находится в VTL 0. Это позволяет трастлетам и ядру безопасности использовать ETW для регистрации собственных безопасных событий.

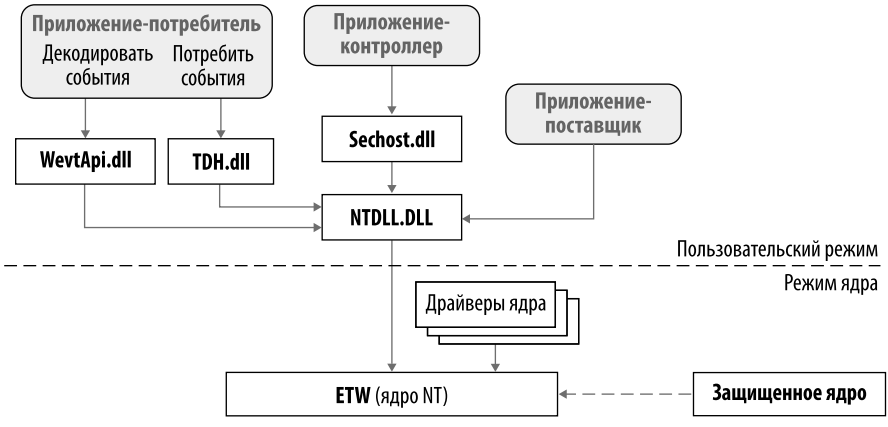


Рис. 10.31. Архитектура ETW

Инициализация ETW

Инициализация ETW начинается на раннем этапе запуска ядра NT (более подробную информацию об инициализации ядра NT см. в главе 12). Она управляется внутренней функцией `EtwInitialize` и проходит в три этапа. Фаза 0 инициализации ядра NT вызывает `EtwInitialize` для правильного размещения и инициализации специфичной для каждого хранилища ETW структуры данных, в которой хранится массив контекстов журнала, представляющих глобальные сеансы ETW (более подробную информацию см. в разделе «Сеансы ETW» далее в этой главе). Максимальное количество глобальных сеансов запрашивается из параметра реестра HKLM\System\CurrentControlSet\Control\WMI\EtwMaxLoggers, который должен находиться в диапазоне от 32 до 256 (если нужный параметр в реестре отсутствует, по умолчанию считается 64).

Позже, при запуске ядра NT, процедура `IoInitSystemPreDrivers` фазы 1 продолжает инициализацию ETW, в ходе чего выполняет следующие шаги.

1. Получает время запуска системы и основное время системы, рассчитывает частоту QPC.
2. Инициализирует раздел безопасности ETW и считывает сеанс по умолчанию и дескриптор безопасности поставщика.
3. Инициализирует глобальные структуры трассировки для каждого процессора, расположенные в PRCB.
4. Создает тип объекта-потребителя ETW в реальном времени, называемый `EtwConsumer`, который используется для того, чтобы позволить потребительскому процессу в режиме реального времени подключаться к основному потоку журнала ETW и типу объекта регистрации ETW (он называется `EtwRegistration`), которые позволяют поставщику регистрироваться из приложения пользовательского режима.
5. Регистрирует обратный вызов проверки ошибок ETW, применяемый для выгрузки данных сеансов журнала в дамп проверки ошибок.

6. Инициализирует и запускает сеансы Global logger и Autologgers на основе разделов реестра GlobalLogger и AutoLogger соответственно, расположенных в корневом разделе HKLM\System\CurrentControlSet\Control\WMI.
7. Использует API ядра `EtwRegister` для регистрации различных поставщиков событий ядра NT, таких как поставщик трассировки событий ядра, поставщик общих событий, процесс, сеть, диск, имя файла, ввод-вывод, память и т. д.
8. Публикует имя состояния WNF, инициализированное ETW, чтобы указать, что подсистема ETW инициализирована.
9. Записывает событие `SystemStart` как в поставщики ведения журнала глобальной трассировки, так и в поставщики общих событий. Событие, которое показано на рис. 10.32, записывает приблизительное время запуска ОС.
10. При необходимости загружает драйвер FileInfo, который предоставляет дополнительную информацию о вводе-выводе файлов в Superfetch (дополнительную информацию о превентивном управлении памятью можно найти в главе 5 тома 1).

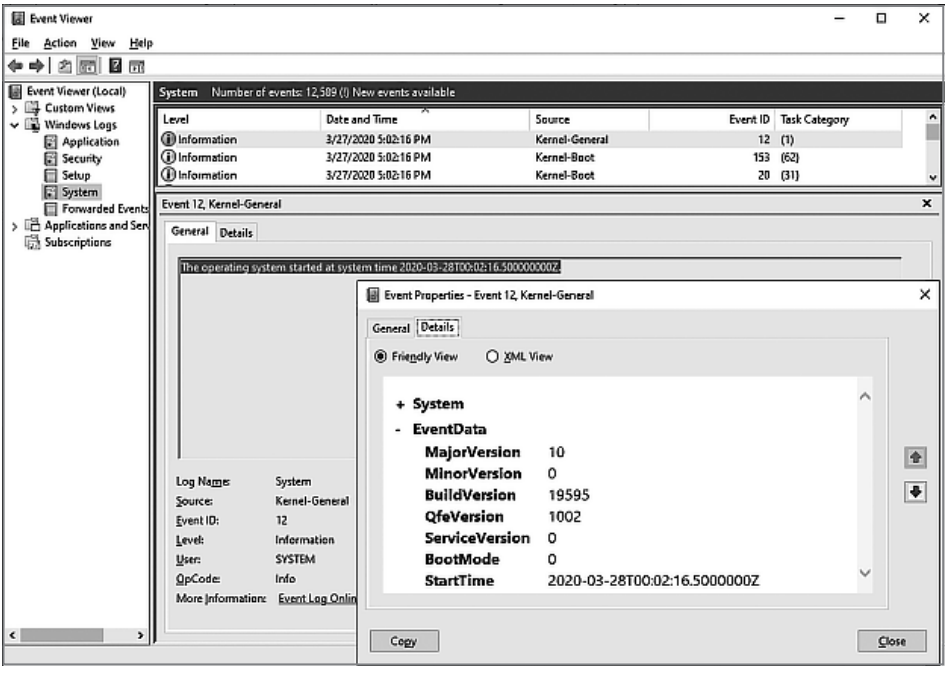


Рис. 10.32. Событие SystemStart системы ETW, отображаемое средством просмотра событий

На ранних этапах загрузки реестр Windows и подсистемы ввода-вывода еще не полностью инициализированы. Таким образом, ETW не может напрямую записывать файлы журналов. В конце процесса загрузки, после того как диспетчер сеансов (SMSS.exe) правильно инициализирует куст программного обеспечения, проходит последний этап инициализации ETW. Его цель — просто сообщить каждому уже зарегистрированному глобальному сеансу ETW о готовности файловой системы, чтобы они могли сбросить все события, записанные в буферах ETW, в файл журнала.

Сеансы ETW

Один из наиболее важных объектов ETW — сеанс, называемый экземпляром журнала, который является связующим звеном между поставщиками и потребителями. Сеанс трассировки событий записывает события от одного или нескольких поставщиков, включенных контроллером. Сеанс обычно содержит всю информацию, описывающую, какие события какими поставщиками должны записываться и как они должны обрабатываться. Например, сеанс может быть настроен на прием всех событий от поставщика Microsoft-Windows-Hyper-V-Hypervisor, который идентифицируется с помощью GUID {52fc89f8-995e-434c-a91e-199986449890}. Пользователь также может настроить фильтры. Каждое событие, генерируемое поставщиком или группой поставщиков, можно фильтровать по уровню (информация, предупреждение, ошибка или критическое), ключевому слову, идентификатору и иным характеристикам. Конфигурация сеанса может определять и другие детали, допустим, какой источник времени следует использовать для временных меток событий (например, QPC, TSC или системное время), для каких событий следует фиксировать трассировки стека и т. д. Для сеанса существует важное правило размещения потока журнала ETW, который является основным объектом, записывающим события в файл журнала или доставляющим их потребителю в реальном времени.

Сеансы создаются с помощью функции `StartTrace` и настраиваются с помощью функций `ControlTrace` и `EnableTraceEx2`. Инструменты командной строки, такие как xperf, logman, Tracelog и wevtutil, используют их для запуска сеансов трассировки или управления ими. Сеанс можно настроить и как частный для процесса, который его создает. В этом случае ETW применяется для потребления событий, созданных только тем приложением, которое выступает в роли поставщика. Таким образом, приложение избавляет от накладных расходов, связанных с переходом в режим ядра. Частные сеансы ETW могут только записывать события для потоков процесса, в котором они выполняются, и не могут использоваться с доставкой в реальном времени. Внутренняя архитектура частного ETW не входит в тематику данной книги.

При создании глобального сеанса функция `StartTrace` проверяет параметры и копирует их в структуру данных, которую функция `NtTraceControl` использует для вызова внутренней функции `EtwpStartLogger` в ядре. Сеанс ETW представлен внутри структуры данных `ETW_LOGGER_CONTEXT`, которая содержит важные указатели на буферы памяти сеанса, куда записываются события. Как обсуждалось в разделе «Инициализация ETW», система может поддерживать ограниченное количество сеансов ETW, которые хранятся в массиве, расположенном в глобальной структуре данных, для каждого хранилища. `EtwpStartLogger` проверяет глобальный массив сеансов, определяя, есть ли свободное место или уже существует сеанс с таким именем. В этом случае она завершает работу и сигнализирует об ошибке. В противном случае генерирует GUID сеанса (если он еще не указан вызывающей стороной), выделяет и инициализирует структуру данных `ETW_LOGGER_CONTEXT`, представляющую сеанс, назначает ей индекс и вставляет ее в массив для каждого хранилища.

ETW запрашивает дескриптор безопасности сеанса, расположенный в разделе реестра HKLM\System\CurrentControlSet\Control\Wmi\Security. Каждый параметр реестра в разделе называется GUID сеанса (раздел реестра содержит также GUID поставщика) и содержит двоичное представление собственного относительного дескриптора безопасности (рис. 10.33). Если дескриптора безопасности для сеанса не существует, для него возвращается дескриптор безопасности по умолчанию

(подробности см. в эксперименте «Наблюдение за дескриптором безопасности по умолчанию для сеансов ETW» далее в этой главе).

Функция `EtwpStartLogger` проверяет доступ к дескриптору безопасности сеанса, запрашивая право доступа `TRACELOG_GUID_ENABLE` (а также `TRACELOG_CREATE_REALTIME` или `TRACELOG_CREATE_ONDISK` в зависимости от режима файла журнала) с помощью токена доступа текущего процесса. Если проверка успешна, вычисляются размер по умолчанию и количество буферов событий, которые рассчитываются на основе размера физической памяти системы (размер буфера по умолчанию — 8, 16 или 64 Кбайт). Количество буферов зависит от количества процессоров системы и наличия флага режима логгера `EVENT_TRACE_NO_PER_PROCESSOR_BUFFERING`, который предотвращает запись в попроцессорный буфер событий, которые могут генерироваться разными процессорами.

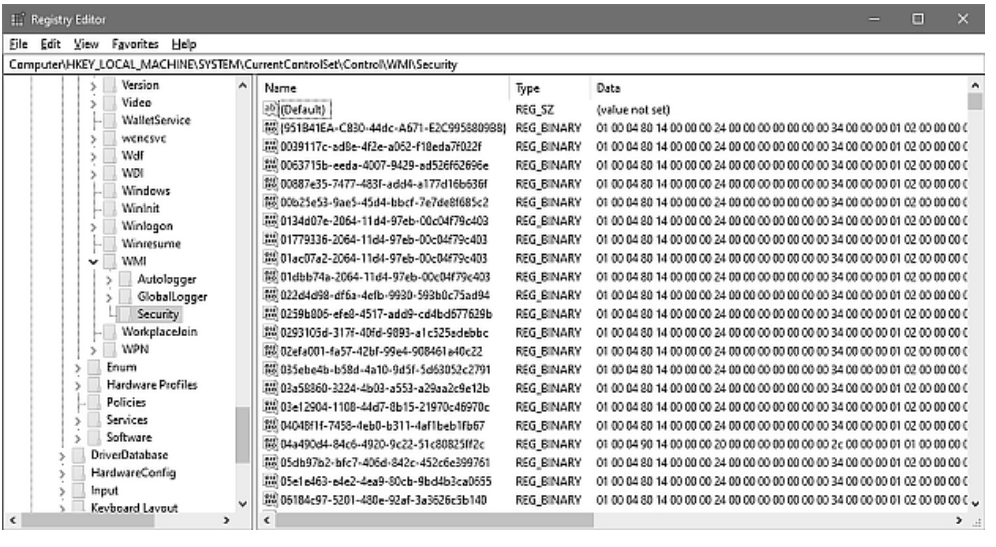


Рис. 10.33. Раздел реестра безопасности ETW

ETW получает начальную эталонную отметку времени сеанса. Сейчас поддерживаются три разрешения часов: счетчик производительности запросов (QPC, отметка времени с высоким разрешением, не зависящая от системных часов), системное время и счетчик циклов процессора. Функция `EtwpAllocateTraceBuffer` используется для выделения каждого буфера, связанного с сеансом журнала (количество буферов было рассчитано заранее или указано как входные данные пользователя). Буфер может быть выделен из выгружаемого пула, невыгружаемого пула или непосредственно из физических больших страниц в зависимости от режима ведения журнала. Каждый буфер хранится в нескольких внутренних списках для каждого сеанса, которые могут обеспечить быстрый поиск как для основного потока журнала ETW, так и для поставщиков ETW. Наконец, если для режима журнала не установлен циклический буфер, функция `EtwpStartLogger` запускает основной поток журнала ETW, цель которого — сбросить события, записанные поставщиками, связанными с сеансом, в файл журнала или потребителю в реальном времени. После запуска основного потока ETW отправляет уведомление о сеансе

зарегистрированному поставщику уведомлений о сеансе (GUID 2a6e185b-90de-4fc5-826c-9f44e608a427) — специальному поставщику, который позволяет своим потребителям получать информацию о возникновении определенных событий ETW (например, когда новый сеанс создается или уничтожается, создается новый файл журнала или возникает ошибка журнала).

ЭКСПЕРИМЕНТ. Перечисление сеансов ETW

В Windows 10 существует несколько способов перечисления активных сеансов ETW. В этом и всех последующих экспериментах, касающихся ETW, вы будете использовать утилиту XPERF — часть набора инструментов для повышения производительности Windows, распространяемого в составе пакета Windows Assessment and Deployment Kit (ADK), который можно бесплатно загрузить с https://docs.microsoft.com/en-us/windows-hardware/get-started/adk-install.

Перечислить активные сеансы ETW можно несколькими способами. XPERF может сделать это во время выполнения с помощью следующей команды (обычно XPERF устанавливается в C:\Program Files (x86)\Windows Kits\10\Windows Performance Toolkit):

```
xperf -Loggers
```

Вывод команды может быть огромным, поэтому настоятельно рекомендуется перенаправить его в файл TXT:

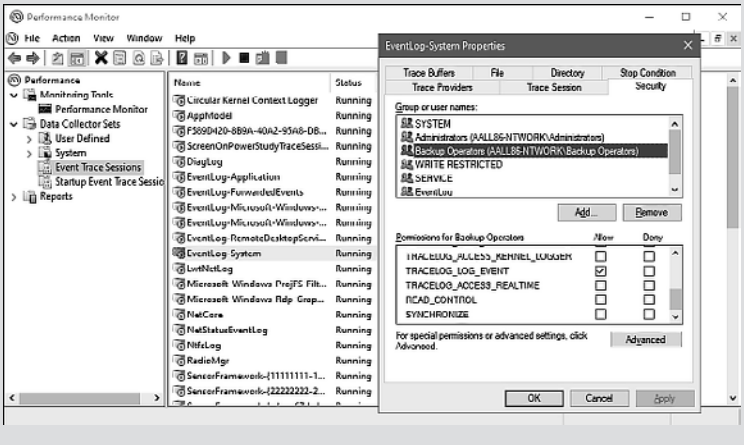
```
xperf -Loggers > ETW_Sessions.txt
```

Утилита может декодировать и отображать в удобочитаемой форме все данные конфигурации сеанса. В качестве примера приведен сеанс EventLog-Application, который используется службой регистрации событий (Wevtsvc.dll) для записи событий в файл `Application.evtx`, отображаемый средством просмотра событий:

```
Logger Name           : EventLog-Application
Logger Id             : 9
Logger Thread Id      : 000000000000008C
Buffer Size           : 64
Maximum Buffers       : 64
Minimum Buffers       : 2
Number of Buffers     : 2
Free Buffers          : 2
Buffers Written       : 252
Events Lost           : 0
Log Buffers Lost      : 0
Real Time Buffers Lost: 0
Flush Timer           : 1
Age Limit             : 0
Real Time Mode        : Enabled
Log File Mode         : Secure PersistOnHybridShutdown PagedMemory IndependentSession
                        NoPerProcessorBuffering
Maximum File Size     : 100
Log Filename          :
Trace Flags           : "Microsoft-Windows-CertificateServicesClient-Lifecycle-
                           User":0x8000000000000000:0xff+"Microsoft-Windows-
                           SenseIR":0x8000000000000000:0xff+
... (вывод сокращен для краткости)
```

Утилита позволяет также декодировать имя каждого поставщика, включенного в сеанс, и битовую маску категорий событий, которые поставщик должен записывать в сеансы. Интерпретация битовой маски (показана в строке `Trace Flags`) зависит от поставщика. Например, поставщик может определить, что категория 1 (установлен бит 0) указывает набор событий, генерируемых во время инициализации и очистки, категория 2 (установлен бит 1) — набор событий, генерируемых при выполнении ввода-вывода реестра, и т. д. Флаги трассировки интерпретируются по-разному для системных сеансов (более подробную информацию см. в разделе «Системные средства ведения журнала (логгеры)»). В этом случае флаги декодируются из включенных флагов ядра, которые определяют, какие события ядра должен регистрировать системный сеанс.

Монитор производительности Windows, помимо работы со счетчиками производительности системы, может легко подсчитывать сеансы ETW. Откройте Монитор производительности (Performance Monitor), введя `perfmon` в поле поиска Cortana, разверните Наборы сборщиков данных (Data Collector Sets) и щелкните на Сеансы трассировки событий (Event Trace Sessions). Приложение должно отображать те же сеансы, что и XPERF. Если вы щелкнете правой кнопкой мыши на имени сеанса и выберете Свойства (Properties), то сможете перемещаться между конфигурациями сеанса. В частности, страница свойств безопасности декодирует дескриптор безопасности сеанса ETW.



Наконец, вы можете использовать консольную утилиту Microsoft Logman (%SystemRoot%\System32\logman.exe) для перечисления активных сеансов ETW (с помощью аргумента командной строки `-ets`).

Поставщики ETW

Как говорилось в предыдущих разделах, поставщик — это компонент, создающий события, в то время как приложение, включающее поставщика, содержит инструменты отслеживания событий. ETW поддерживает различные типы поставщиков, которые используют схожие модели программирования. (Они различаются

в основном способом кодирования событий.) Поставщик должен быть изначально зарегистрирован в ETW, прежде чем он сможет генерировать какое-либо событие. Аналогично приложение-контроллер должно включить поставщика и связать его с сеансом ETW, чтобы иметь возможность получать от него события. Если ни один сеанс не активировал поставщика, он не будет генерировать никаких событий. Поставщик определяет свою интерпретацию включения или отключения. Как правило, включенный поставщик генерирует события, а отключенный — нет.

Регистрация поставщиков

Каждый тип поставщика имеет собственный API, который должен вызываться приложением поставщика или драйвером для регистрации поставщика. Например, поставщики на основе манифеста полагаются на функцию `EventRegister` для регистрации в пользовательском режиме и `EtwRegister` — для регистрации в режиме ядра. Все типы поставщиков в итоге вызывают внутреннюю функцию `EtwpRegisterProvider`, которая реально выполняет регистрацию и реализована как в ядре NT, так и в NTDLL. Процесс регистрации выделяет и инициализирует структуру данных `ETW_GUID_ENTRY`, которая представляет поставщика (одна и та же структура данных используется для уведомлений и признаков). Структура данных содержит важную информацию, такую как GUID поставщика, дескриптор безопасности, счетчик ссылок, информацию о включении (для каждого сеанса ETW, который включает поставщика) и список регистраций поставщика.

При регистрации поставщика в пользовательском режиме ядро NT проверяет доступ к токену вызывающего процесса, запрашивая право доступа `TRACELOG_REGISTER_GUIDS`. Если проверка успешна или запрос на регистрацию исходит из кода ядра, ETW вставляет новую структуру данных `ETW_GUID_ENTRY` в хеш-таблицу, расположенную в глобальной структуре данных каждого бункера ETW, применяя хеш GUID поставщика в качестве ключа таблицы (это позволяет быстро найти всех поставщиков, зарегистрированных в системе). Если запись с таким же GUID уже существует в хеш-таблице, ETW использует ее вместо новой. GUID уже мог существовать в хеш-таблице главным образом по двум причинам.

- Другой драйвер или приложение активировали поставщика до его фактической регистрации (дополнительную информацию см. в разделе «Активация поставщика» далее в этой главе).
- Поставщик был зарегистрирован ранее. Поддерживается множественная регистрация одного и того же GUID поставщика.

После успешного добавления поставщика в глобальный список ETW создает и инициализирует объект регистрации ETW, который представляет собой одну регистрацию. Объект инкапсулирует структуру данных `ETW_REG_ENTRY`, которая связывает поставщика с процессом и сеансом, запросившим его регистрацию. (ETW поддерживает также регистрацию из разных сеансов.) Объект вставляется в список, расположенный в `ETW_GUID_ENTRY` (тип объекта `EtwRegistration` был предварительно создан и зарегистрирован в диспетчере объектов NT во время инициализации ETW). На рис. 10.34 показаны две структуры данных и их отношения. Здесь процессы двух поставщиков: процесс А, находящийся в сеансе 4, и процесс Б, находящийся в сеансе 16, — зарегистрировались для поставщика 1. Таким образом, две структуры данных `ETW_REG_ENTRY` были созданы и связаны с `ETW_GUID_ENTRY`, представляющим поставщика 1.

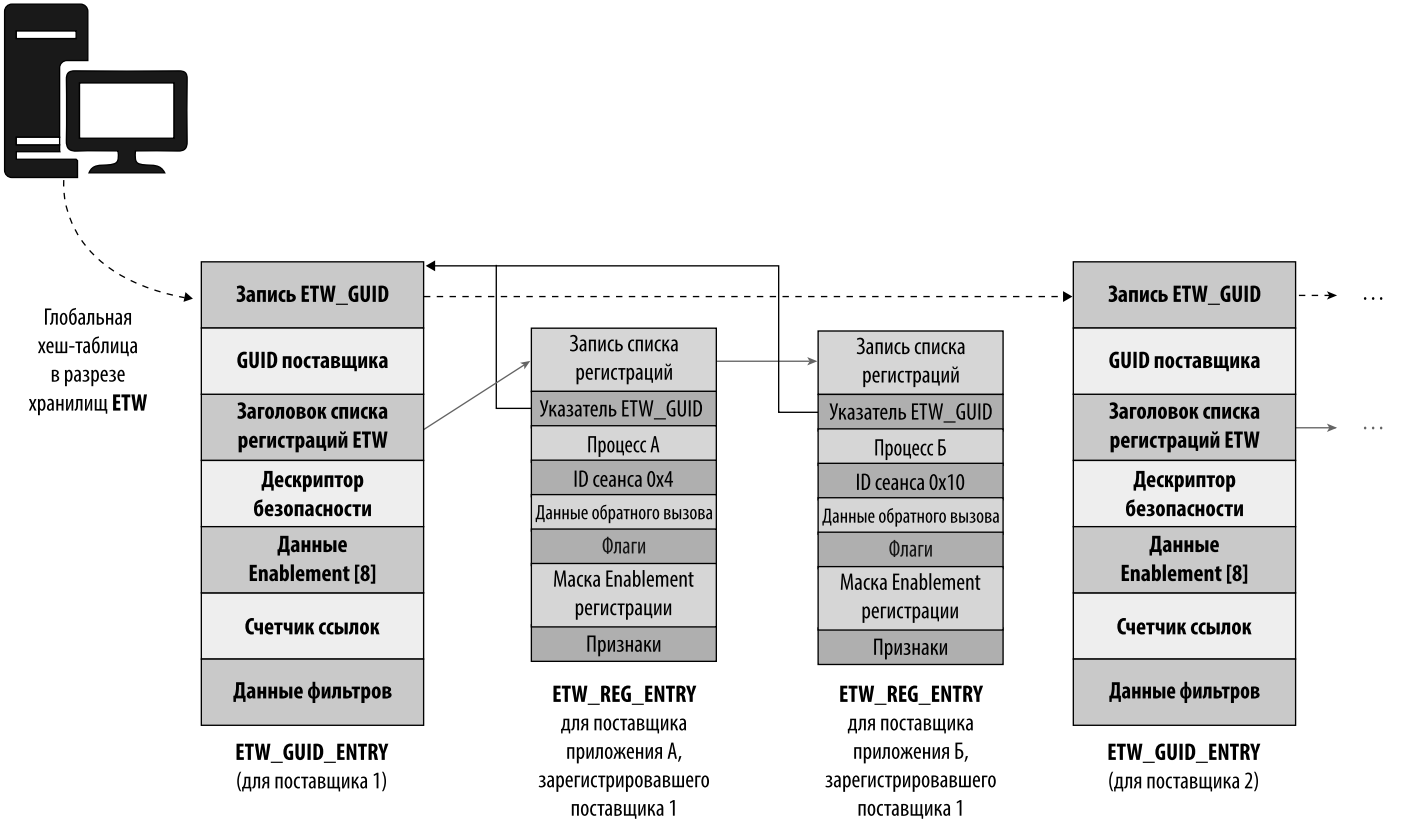


Рис. 10.34. Структуры данных ETW_GUID_ENTRY и ETW_REG_ENTRY

На этом этапе поставщик зарегистрирован и готов к включению в сеансе(ах), который его запросил (через функцию `EnableTrace`). Если поставщик уже был включен хотя бы в одном сеансе до его регистрации, ETW включает его (подробности см. в следующем разделе) и вызывает обратный вызов Enablement (активации), который может быть указан вызывающим API `EventRegister` (или `EtwRegister`), который начал процесс регистрации.

ЭКСПЕРИМЕНТ. Перечисление поставщиков ETW

Что касается сеансов ETW, то XPERF может перечислить список всех действующих зарегистрированных поставщиков (инструмент WEVTUTIL, устанавливаемый вместе с Windows, может делать то же самое). Откройте командную строку и перейдите по пути к Windows Performance Toolkit. Чтобы перечислить зарегистрированных поставщиков, используйте параметр команды `-providers`. Параметр поддерживает разные флаги. В этом эксперименте вас будут интересовать флаги `I` и `R`, которые сообщают XPERF о необходимости перечисления установленных или зарегистрированных поставщиков. Как мы обсудим в разделе «Расшифровка (декодирование) событий» далее в этой главе, разница состоит в том, что поставщик может быть зарегистрирован (путем указания GUID), но не установлен в разделе реестра HKLM\SOFTWARE\Microsoft\Windows\CurrentVersion\WINEVT\Publishers. Это не позволит любому потребителю декодировать событие с помощью процедур TDH. Следующие команды:

```
cd /d "C:\Program Files (x86)\Windows Kits\10\Windows Performance Toolkit"
xperf -providers R > registered_providers.txt
xperf -providers I > installed_providers.txt
```

создадут два текстовых файла с аналогичной информацией. Если вы откроете файл `registered_providers.txt`, то обнаружите смесь имен и GUID. Имена идентифицируют поставщиков, которые также установлены в разделе реестра Publisher, тогда как GUID представляет поставщиков, только что зарегистрированных с помощью функции `EventRegister`, упомянутой ранее в этом разделе. Все имена с соответствующими GUID присутствуют также в файле `installув_providers.txt`, но там вы не найдете ни одного GUID, указанного в первом текстовом файле, в списке установленных поставщиков.

XPERF также разрешает перечисление всех флагов ядра и групп, поддерживаемых системными средствами ведения журнала (обсуждаются в разделе «Системные средства ведения журнала (логгеры)» далее в этой главе), посредством флага `K`, который включает в себя функционал флагов `KF` и `KG`.

Активация поставщика

Как говорилось в предыдущем разделе, чтобы иметь возможность генерировать события, поставщик должен быть связан с сеансом ETW. Эта ассоциация называется *активацией поставщика* и может возникнуть как до, так и после регистрации поставщика. Приложение контроллера может включить поставщика в сеансе

с помощью функции EnableTraceEx. Она позволяет указать битовую маску ключевых слов, определяющих категорию событий, которые хочет получать сеанс. Точно так же функция поддерживает расширенные фильтры для других типов данных, таких как идентификаторы процессов, генерирующих события, идентификатор пакета, имя исполняемого файла и т. д. (Дополнительную информацию можно найти по адресу https://docs.microsoft.com/en-us/windows/win32/api/evntprov/ns-evntprov-event_filter_descriptor.)

Активацией поставщика управляет ETW в режиме ядра с помощью внутренней функции EtwpEnableGuid. Для запросов пользовательского режима функция проверяет доступ как к дескрипторам безопасности сеанса, так и к поставщику, запрашивая право доступа TRACELOG_GUID_ENABLE от имени токена вызывающего процесса. Если сеанс регистрации включает флаг SECURITY_TRACE, то EtwpEnableGuid требует, чтобы вызывающий процесс был PPL (более подробную информацию см. в разделе «Безопасность ETW» далее в этой главе). Если проверка успешна, функция выполняет задачу, аналогичную рассмотренной ранее при обсуждении регистрации поставщиков:

- выделяет и инициализирует структуру данных ETW_GUID_ENTRY для представления поставщика или использования структуры, уже связанной с глобальной структурой данных каждого хранилища ETW, если поставщик уже зарегистрирован;
- связывает поставщика с сеансом средства ведения журнала, добавляя информацию о включении сеанса в ETW_GUID_ENTRY.

Если поставщик не был ранее зарегистрирован, то объекта регистрации ETW, связанного со структурой данных ETW_GUID_ENTRY, не существует, поэтому процедура завершается. (Поставщик будет активирован после первой регистрации.) В противном случае поставщик будет включен.

В то время как устаревшие поставщики MOF и поставщики WPP могут быть активированы только для одного сеанса за раз, поставщики на основе манифеста и поставщики отслеживания трассировки могут быть активированы максимум в восьми сеансах. Как было показано на рис. 10.32, структура данных ETW_GUID_ENTRY содержит информацию об активации для каждого возможного сеанса ETW, который активировал поставщика (максимум восемь). На основе активированных сеансов функция EtwpEnableGuid вычисляет новую маску активации сеанса, сохраняя ее в структуре данных ETW_REG_ENTRY, представляющей регистрацию поставщика. Маска очень важна, поскольку является ключом для генерации событий. Когда приложение или драйвер записывает событие поставщику, выполняется проверка: если бит в маске активации равен 1, то событие должно быть записано в буфер, поддерживаемый конкретной сессией ETW, в противном случае сеанс пропускается и событие не записывается в его буфер.

Обратите внимание: для защищенных сеансов дополнительная проверка доступа выполняется перед обновлением маски активации сеанса при регистрации поставщика. Дескриптор безопасности сеанса ETW должен разрешать право доступа TRACELOG_LOG_EVENT к токену доступа вызывающего процесса. В противном случае относительный бит в маске включения не установлен в 1. (Целевой сеанс

ETW не получит никаких событий от регистрации поставщика.) Дополнительную информацию о безопасных сеансах можно найти в разделе «Безопасные логгеры» далее в этой главе.

Поставка событий

После регистрации одного или нескольких поставщиков ETW приложение поставщика может начать генерировать события. Обратите внимание на то, что события могут создаваться, даже если приложение контроллера не имело возможности включить поставщика в сеансе ETW. Способ, которым приложение или драйвер может генерировать события, зависит от типа поставщика. Например, приложения, которые записывают события поставщикам на основе манифеста, обычно напрямую создают дескриптор события (он соответствует XML-манифесту) и используют API `EventWrite` для записи события в сеансы ETW, в которых поставщик активирован. Приложения, которые управляют поставщиками MOF и WPP, вместо этого применяют API `TraceEvent`.

События, генерируемые поставщиками на основе манифеста, как обсуждалось ранее в разделе «Сеансы ETW», можно фильтровать несколькими способами. ETW находит структуру данных `ETW_GUID_ENTRY` из объекта регистрации поставщика, который предоставляется приложением через дескриптор. Внутренняя функция `EtwpEventWriteFull` использует маску активации сеанса регистрации поставщика для циклического переключения между всеми активированными сеансами ETW, связанными с поставщиком и представленными `ETW_LOGGER_CONTEXT`. Для каждого сеанса он проверяет, удовлетворяет ли событие всем фильтрам. Если да, то вычисляет полный размер полезной нагрузки события и проверяет, достаточно ли свободного места в текущем буфере сеанса.

Если свободного места нет, ETW проверяет, есть ли в сеансе еще один свободный буфер. Свободные буферы сохраняются в очереди FIFO («первым вошел — первым обслужен»). Если есть свободный буфер, ETW помечает старый буфер как «грязный» и переключается на новый, свободный. В этом подходе поток Logger может проснуться и сбросить весь буфер в файл журнала или доставить его потребителю в реальном времени. Если режим журнала сеанса представляет собой циклический журнал, поток журнала никогда не создается — ETW просто связывает старый полный буфер в конце очереди свободных буферов (в результате очередь никогда не будет пустой). Если же в очереди нет свободного буфера, ETW пытается выделить дополнительный буфер, прежде чем возвращать ошибку вызывающему объекту.

После того как в буфере найдено достаточно места, `EtwpEventWriteFull` атомарно записывает в него всю полезную нагрузку события и завершает работу. Обратите внимание: если маска активации сеанса равна 0, то с поставщиком не связано ни одного сеанса. В результате событие теряется и нигде не протоколируется.

События MOF и WPP проходят аналогичную процедуру, но поддерживают только один сеанс ETW и обычно меньшее количество фильтров. Для таких поставщиков выполняется дополнительная проверка связанного сеанса: если приложение контроллера пометило сеанс как безопасный, никто не сможет записывать какие-либо события. В этом случае вызывающей стороне возвращается ошибка (безопасные сеансы обсуждаются позже в разделе «Безопасные логгеры»).

ЭКСПЕРИМЕНТ. Просмотр активности процессов с помощью ETW

В этом эксперименте мы будем использовать ETW для мониторинга активности системных процессов. В Windows 10 есть два поставщика, которые могут отслеживать эту информацию: Microsoft-Windows-Kernel-Process и средство ведения журнала ядра NT с помощью флагов ядра `PROC_THREAD`. Используйте первый, который является классическим поставщиком и уже имеет всю информацию для расшифровки своих событий. Можете записать трассировку с помощью нескольких инструментов. Вам по-прежнему потребуется XPERF (также можно задействовать Windows Performance Monitor).

Откройте окно командной строки и введите следующие команды:

```
cd /d "C:\Program Files (x86)\Windows Kits\10\Windows Performance Toolkit"
xperf -start TestSession -on Microsoft-Windows-Kernel-Process -f c:\process_trace.etl
```

Команда запускает сеанс ETW под названием `TestSession` (можете заменить имя), который будет использовать события, созданные поставщиком ядра-процесса, и сохранять их в файле журнала `C:\process_trace.etl` (имя файла тоже можно заменить).

Чтобы убедиться, что сеанс действительно начался, повторите шаги, описанные ранее в эксперименте «Перечисление сеансов ETW». (Сеанс трассировки `TestSession` должен быть указан как в XPERF, так и в мониторе производительности Windows.) Теперь следует запустить несколько новых процессов или приложений, например Notepad или Paint.

Чтобы остановить сеанс ETW, примените команду

```
xperf -stop TestSession
```

Шаги, используемые для декодирования файла ETL, описаны далее в эксперименте «Декодирование файла ETL». Windows включает поставщиков почти для всех своих компонентов. Например, поставщик Microsoft-Windows-MSPaint генерирует события на основе функций Paint. Вы можете попробовать провести этот эксперимент, перехватывая события от поставщика MsPaint.

Поток ETW Logger

Поток Logger — один из наиболее важных объектов в ETW. Его основная цель — записывать события в файл журнала или доставлять их потребителю в реальном времени, отслеживая количество доставленных и потерянных. Поток ведения журнала запускается каждый раз при первоначальном создании сеанса ETW, но только в том случае, если сеанс не использует режим циклического журнала. Логика его выполнения проста. После запуска он связывается со структурой данных `ETW_LOGGER_CONTEXT`, представляющей связанный сеанс ETW, и ожидает двух основных объектов синхронизации. Событие Flush сигнализируется ETW каждый раз, когда буфер, принадлежащий сеансу, заполняется (это может произойти после того, как поставщик сгенерировал новое событие, как обсуждалось в предыдущем

разделе «Поставка событий»), когда новый потребитель реального времени запросил подключение или когда сеанс регистрации будет закончен. Таймер `TimeOut` инициализируется допустимым значением (обычно 1 с) только в том случае, если сеанс является сеансом реального времени или пользователь явно потребовал его при вызове API `StartTrace` для создания нового сеанса.

Когда один из двух объектов синхронизации получает сигнал, поток логгера перезагружает их и проверяет, готова ли файловая система. Иначе основной поток логгера снова переходит в спящий режим (на ранних этапах загрузки сеансы не должны сбрасываться). В противном случае он начинает сбрасывать каждый буфер, принадлежащий сеансу, в файл журнала или потребителю реального времени.

Для сеансов реального времени поток логгера сначала создает временный ETL-файл для каждого сеанса в папке %SystemRoot%\System32\LogFiles\WMI\RtBackup (рис. 10.35). Имя файла журнала генерируется добавлением префикса `EtwRT` к имени сеанса реального времени. Файл используется для сохранения временных событий перед их доставкой потребителю в реальном времени (в файле журнала могут храниться также потерянные события, которые не были доставлены потребителю в надлежащее время). При запуске автоматические логгеры в реальном времени восстанавливают потерянные события из файла журнала для доставки их потребителю.

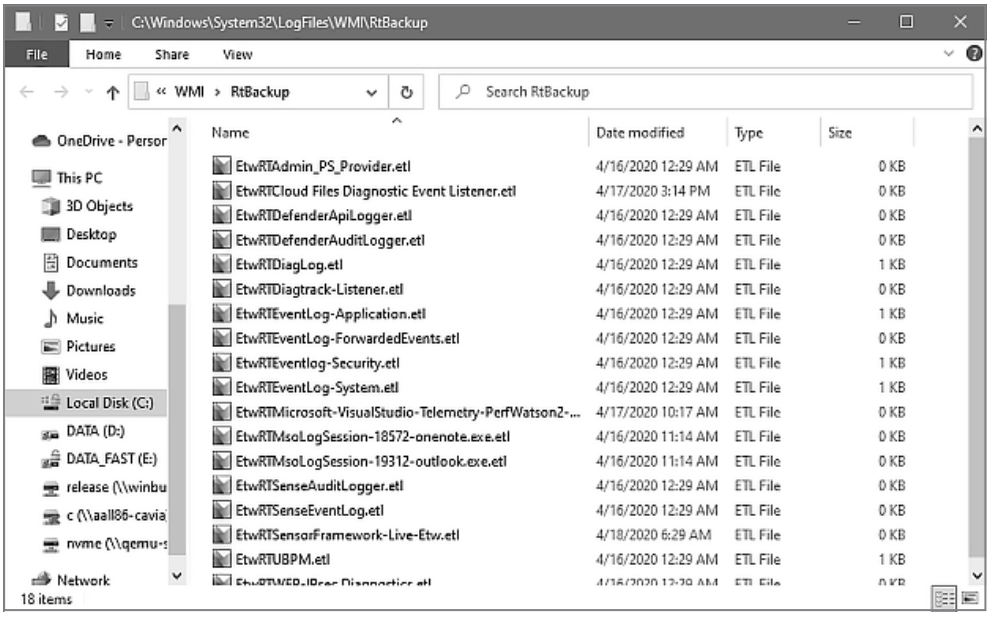


Рис. 10.35. Временные файлы журналов ETL реального времени

Поток логгера — единственный объект, способный установить соединение между потребителем реального времени и сеансом. Когда потребитель впервые вызывает API `ProcessTrace` для получения событий из сеанса реального времени, ETW устанавливает новый объект `RealTimeConsumer` и использует его для создания связи между потребителем и сеансом реального времени. Объект, который преобразуется

в структуру данных ETW_REALTIME_CONSUMER в ядре NT, позволяет внедрить события в адресное пространство процесса потребителя (еще один буфер пользовательского режима предоставляется приложением потребителя).

Для сеансов не в режиме реального времени поток журнала открывает исходный файл журнала ETL, указанный сущностью, создавшей сеанс (или создает файл, если его не существует). Поток журнала может создать и совершенно новый файл журнала, если в режиме журнала сеанса указан флаг EVENT_TRACE_FILE_MODE_NEWFILE и текущий файл журнала достигает максимального размера.

На этом этапе поток журнала ETW инициирует очистку всех буферов, связанных с сеансом, в текущий файл журнала, который, как обсуждалось ранее, может быть временным для сеансов реального времени. Сброс выполняется добавлением заголовка события к каждому событию в буфере и использованием API NtWriteFile для записи двоичного содержимого в файл журнала ETL. Для сеансов реального времени поток журнала, просыпаясь в следующий раз, может внедрить все события, хранящиеся во временном файле журнала, в целевое приложение-потребитель реального времени в пользовательском режиме. Таким образом, для сеансов в реальном времени события ETW никогда не доставляются синхронно.

Потребление событий

Потребление событий в ETW почти полностью выполняется приложением-потребителем в пользовательском режиме благодаря службам, предоставляемым Sechost.dll. Приложение-потребитель применяет функцию OpenTrace для открытия файла журнала ETL, созданного основным потоком средства ведения журнала, или для установления соединения со средством ведения журнала в реальном времени. Приложение определяет функцию обратного вызова событий, которая вызывается каждый раз, когда ETW обрабатывает одно событие. Кроме того, для сеансов в реальном времени приложение может предоставить дополнительную функцию обратного вызова буфера, которая получает статистику для каждого буфера, сбрасываемого ETW, и вызывается каждый раз, когда один буфер заполнен и доставлен потребителю.

Фактическое потребление событий запускается функцией ProcessTrace. Она работает как для стандартных сеансов, так и для сеансов реального времени в зависимости от флагов режима файла журнала, переданных ранее в OpenTrace.

Для сеансов реального времени API использует службы режима ядра, доступ к которым осуществляется через системный вызов NtTraceControl, чтобы убедиться, что сеанс ETW действительно является сеансом реального времени. Ядро NT проверяет, предоставляет ли дескриптор безопасности сеанса ETW право доступа TRACELOG_ACCESS_REALTIME к токену вызывающего процесса. Если доступа нет, API дает сбой и возвращает ошибку приложению контроллера. В противном случае он выделяет временный буфер пользовательского режима и битовую карту, применяемые для получения событий, и подключается к основному потоку журнала, который создает связанный объект EtwConsumer (подробности см. в разделе «Поток ETW Logger» ранее в этой главе). Как только соединение установлено, API ожидает поступления новых данных из потока журнала сеанса. Когда данные поступают, API перечисляет каждое событие и вызывает обратный вызов события.

Для обычных сеансов ETW не в реальном времени API `ProcessTrace` выполняет аналогичную обработку, но вместо подключения к потоку журнала он просто открывает и анализирует файл журнала ETL, считывая все буферы один за другим и вызывая обратный вызов события для каждого найденного события (они отсортированы в хронологическом порядке). В отличие от средств ведения журнала реального времени, которые можно использовать по одному, в этом случае API может работать даже с несколькими дескрипторами трассировки, созданными API `OpenTrace`, что означает: он может анализировать события из разных файлов журналов ETL.

События, принадлежащие сеансам ETW, использующим циклические буферы, не обрабатываются с помощью описанной методологии. (Действительно, не существует потока журнала, который записывает какое-либо событие.) Обычно приложение контроллера применяет `FlushTrace` API, когда хочет сохранить снимок текущих буферов, принадлежащих сеансу ETW, настроенному на использование циклического буфера, в файл журнала. API вызывает ядро NT с помощью системного вызова `NtTraceControl`, который находит сеанс ETW и проверяет, предоставляет ли его дескриптор безопасности право доступа `TRACELOG_CREATE_ONDISK` к токену доступа вызывающего процесса. Если да и если приложение контроллера указало допустимое имя файла журнала, ядро NT вызывает внутреннюю процедуру `EtwpBufferingModeFlush`, которая создает новый файл ETL, добавляет соответствующие заголовки и записывает все буферы, связанные с сеансом. Затем приложение-потребитель может анализировать события, записанные в новом файле журнала, с помощью API `OpenTrace` и `ProcessTrace`, как описано ранее.

Расшифровка (декодирование) событий

Когда функция `ProcessTrace` идентифицирует новое событие в буфере ETW, она вызывает обратный вызов события, который обычно находится в приложении-потребителе. Чтобы иметь возможность правильно обработать событие, приложение-потребитель должно декодировать передаваемые им данные. Вспомогательная библиотека декодирования трассировки событий (TDH.dll) предоставляет приложениям-потребителям услуги по декодированию событий. Как обсуждалось в предыдущих разделах, приложение-поставщик (или драйвер) должно включать информацию, описывающую, как декодировать события, генерируемые его зарегистрированными поставщиками.

Эта информация кодируется по-разному в зависимости от типа поставщика. Например, поставщики на основе манифеста компилируют XML-дескриптор своих событий в двоичный файл и сохраняют его в разделе ресурсов своего приложения-поставщика (или драйвера). В рамках регистрации поставщика приложение установки должно зарегистрировать двоичный файл поставщика в разделе реестра HKLM\SOFTWARE\Microsoft\Windows\CurrentVersion\WINEVT\Publishers. Последнее важно для декодирования событий, особенно по следующим причинам.

- Система обращается к разделу Publisher, когда хочет сопоставить имя поставщика с его GUID (с точки зрения ETW поставщики не имеют имени). Это позволяет таким инструментам, как Xperf, отображать читаемые имена поставщиков вместо их GUID.

- Вспомогательная библиотека декодирования трассировки обращается к разделу, чтобы получить двоичный файл поставщика, проанализировать его раздел ресурсов и прочитать двоичное содержимое дескриптора событий.

После получения дескриптора события вспомогательная библиотека декодирования трассировки получает всю необходимую информацию для декодирования события, анализируя двоичный дескриптор, и позволяет приложениям-потребителям использовать API `TdhGetEventInformation` для получения всех полей, составляющих полезную нагрузку события, и правильных значений, интерпретируя связанные с ними данные. TDH следует аналогичной процедуре для поставщиков MOF и WPP, в то время как TraceLogging включает все данные декодирования в полезные данные события, которые соответствуют стандартному двоичному формату.

Обратите внимание: все события изначально сохраняются ETW в файле журнала ETL, который имеет четко определенный несжатый двоичный формат и не содержит информации о декодировании событий. Это означает, что если файл ETL открыт другой системой, которая не получила трассировку, существует большая вероятность того, что она не сможет декодировать события. Чтобы решить эту проблему, средство просмотра событий использует другой двоичный формат — EVTX. Он включает в себя все события и информацию об их декодировании, и его легко анализирует любое приложение. Последнее может задействовать API журнала событий Windows `EvtExportLog` для сохранения событий, включенных в файл ETL, и информации об их декодировании в файле EVTX.

ЭКСПЕРИМЕНТ. Декодирование файла ETL

В Windows есть несколько инструментов, которые используют API `EvtExportLog` для автоматического преобразования файла журнала ETL и включения в него всей информации о декодировании. В этом эксперименте вы воспользуетесь netsh.exe, но TraceRpt.exe тоже подойдет.

1. Откройте командную строку, перейдите в папку, где находится файл ETL, созданный в результате предыдущего эксперимента («Просмотр активности процессов с помощью ETW»), и вставьте

   ```
   netsh trace convert input=process_trace.etl output=process_trace.txt dump=txt
   overwrite=yes
   ```

2. Здесь `process_trace.etl` — имя входного файла журнала, `aprocess_trace.txt` — имя выходного декодированного текстового файла.
3. Открыв текстовый файл, вы найдете все расшифрованные события (по одному на каждую строку) с описанием, например:

   ```
   [2]1B0C.1154::2020-05-01 12:00:42.075601200 [Microsoft-Windows-Kernel-Process]
   Process 1808 started at time 2020 – 05 – 01T19:00:42.075562700Z by parent 6924
   running in session 1 with name \Device\HarddiskVolume4\Windows\System32\
   notepad.exe.
   ```

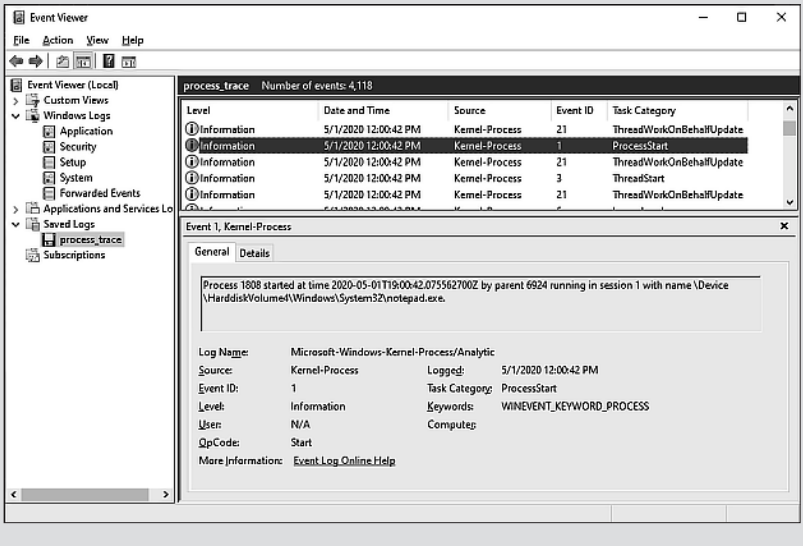
4. Анализируя журнал, вы обнаружите, что иногда события не расшифровываются полностью или не содержат никакого описания. Это связано с тем, что манифест поставщика не включает необходимую информацию (хороший

пример приведен в событии `ThreadWorkOnBehalfUpdate`). Вы можете избавиться от этих событий, получив трассировку, которая не включает их ключевое слово. Ключевое слово события хранится в файле CSV или EVTX.

5. Используйте netsh.exe для создания файла EVTX с помощью следующей команды:

```
netsh trace convert input=process_trace.etl output=process_trace.evtx dump=evtx
overwrite=yes
```

6. Откройте средство просмотра событий. В дереве консоли, расположенном в левой части окна, щелкните правой кнопкой мыши на корневом узле Просмотр событий (локальный) (Event Viewer (Local)) и выберите Открыть сохраненные журналы (Open Saved Logs). Выберите только что созданный файл `process_trace.evtx` и нажмите Открыть (Open).
7. В окне Открыть сохраненный журнал (Open Saved Log) необходимо дать журналу имя и выбрать папку для его отображения. (В примере принято имя по умолчанию process_trace и выбрана папка Сохраненные журналы (Saved Logs).)
8. Средство просмотра событий теперь должно отображать каждое событие, указанное в файле журнала. Щелкните на столбце Дата и время (Date and Time), чтобы упорядочить события по дате и времени в порядке возрастания (от старых к новым). Найдите ProcessStart нажатием Ctrl+F, чтобы найти событие, указывающее на создание процесса Notepad.exe.



9. Событие `ThreadWorkOnBehalfUpdate`, не имеющее понятного для человека описания, вызывает слишком много шума, и его следует убрать из трассировки. Если вы щелкнете на одном из этих событий и откроете вкладку Сведения (Details) в узле Система (System), то обнаружите, что событие принадлежит к категории `WINEVENT_KEYWORD_ WORK_ON_BEHALF`, битовая маска ключевого

слова которой — 0x8000000000002000. (Имейте в виду, что старшие 16 битов ключевых слов зарезервированы для категорий, определенных Microsoft.) Побитовая операция NOT для 64-битного значения 0x8000000000002000 — это 0x7FFFFFFFFFFFDFFF.

10. Закройте средство просмотра событий и запишите еще одну трассировку с помощью XPERF, используя следующую команду:

```
xperf -start TestSession -on Microsoft-Windows-Kernel-Process:0x7FFFFFFFFFFFDFFF
-f c:\process_trace.etl
```

11. Откройте Блокнот или другое приложение и остановите трассировку, как описано в эксперименте «Просмотр активности процессов с помощью ETW». Преобразуйте файл ETL в EVTX. На этот раз полученный декодированный журнал должен быть меньшего размера и не должен содержать событий `ThreadWorkOnBehalfUpdate`.

Системные средства ведения журнала (логгеры)

До сих пор мы описывали, как работают обычные сеансы и поставщики ETW. Начиная с Windows XP ETW поддерживает концепции системных средств ведения журнала, позволяющих ядру NT глобально генерировать события журнала, которые не привязаны к какому-либо поставщику и обычно используются для измерения производительности. На момент написания данной главы были доступны два основных системных логгера, представленные логгером ядра NT и циклическим логгером контекста ядра, в то время как глобальный логгер является подмножеством логгера ядра NT. Ядро NT поддерживает максимум восемь сеансов системного журнала. Каждый сеанс, который получает события от системного логгера, считается системным сеансом.

Чтобы запустить системный сеанс, приложение использует функцию `StartTrace`, но указывает флаг `EVENT_TRACE_SYSTEM_LOGGER_MODE` или GUID сеанса системного средства ведения журнала в качестве входных параметров. В табл. 10.16 перечислены системные средства ведения журнала и их GUID. Функция `EtwpStartLogger` в ядре NT распознает флаг или специальные идентификаторы GUID и выполняет дополнительную проверку по дескриптору безопасности логгера ядра NT, запрашивая право доступа `TRACELOG_GUID_ENABLE` от имени токена доступа к вызывающему процессу. Если проверка пройдена, ETW вычисляет индекс системного логгера и обновляет как маску группы логгеров, так и маску глобальной группы производительности системы.

Таблица 10.16. Системные логгеры

Индекс	Название	GUID	Обозначение
0	Логгер ядра NT	{9e814aad-3204-11d2-9a82-006008a86939}	SystemTraceControl-Guid
1	Глобальный логгер	{e8908abc-aa84-11d2-9a93-00805f85d7c6}	GlobalLoggerGuid
2	Циклический логгер контекста ядра	{54dea73a-ed1f-42a4-af71-3e63d056f174}	CKCLGuid

Последний шаг является ключевым, он управляет системными логгерами. Несколько низкоуровневых системных функций, которые могут выполняться с высоким IRQL (хороший пример — Context Swapper), анализируют маску группы производительности и решают, записывать ли событие в системный журнал. Приложение-контроллер может включать или отключать различные события, регистрируемые системным логгером, путем изменения битовой маски `EnableFlags`, используемой API `StartTrace` и API `ControlTrace`. События, регистрируемые системным логгером, сохраняются внутри в глобальной маске группы производительности в четко определенном порядке. Маска представляет собой массив из восьми 32-битных значений. Каждый индекс в массиве представляет набор событий. Наборы системных событий, также называемые группами, можно перечислить с помощью инструмента Xperf. В табл. 10.17 приведены события системного журнала и выполнена их классификация по группам. Большинство событий системного журнала задокументированы по адресу https://docs.microsoft.com/en-us/windows/win32/api/evntrace/ns-evntrace-event_trace_properties.

Таблица 10.17. События системного логгера (флаги ядра) и их группы

Имя	Описание	Группа
ALL_FAULTS	Все ошибки страниц, включая аппаратные, ошибки копирования при записи, ошибки нулевого требования и т. д.	Нет
ALPC	Расширенный вызов локальных процедур	Нет
CACHE_FLUSH	События очистки кэша	Нет
CC	События диспетчера кэша	Нет
CLOCKINT	События прерывания часов	Нет
COMPACT_CSWITCH	Компактное переключение контекста	Diag
CONTMEMGEN	Непрерывная генерация памяти	Нет
CPU_CONFIG	Топология NUMA, группа процессоров и индекс процессора	Нет
CSWITCH	Переключение контекста	IOTrace
DEBUG_EVENTS	Планирование событий отладчика	Нет
DISK_IO	Дисковый ввод-вывод	Все, кроме SysProf, ReferenceSet и Network
DISK_IO_INIT	Инициализация дискового ввода-вывода	Нет
DISPATCHER	Планировщик процессора	Нет
DPC	События DPC	Diag, DiagEasy и Latency
DPC_QUEUE	События очереди DPC	Нет
DRIVERS	События драйвера	Нет
FILE_IO	Время и результаты завершения операций файловой системы	FileIO

Продолжение ⇩

Таблица 10.17 *(продолжение)*

Имя	Описание	Группа
FILE_IO_INIT	Операция файловой системы (create/open/close/read/write)	FileIO
FILENAME	FileName (например, FileName create/delete/rundown)	Нет
FLT_FASTIO	Завершение обратного вызова мини-фильтра Fastio	Нет
FLT_IO	Завершение обратного вызова мини-фильтра	Нет
FLT_IO_FAILURE	Завершение обратного вызова мини-фильтра с ошибкой	Нет
FLT_IO_INIT	Инициирование обратного вызова мини-фильтра	Нет
FOOTPRINT	Поддержка анализа занимаемого пространства	ReferenceSet
HARD_FAULTS	Жесткие ошибки страниц	Все, кроме SysProf и Network
HIBERRUNDOWN	Сводка (-и) во время спящего режима	Нет
IDLE_STATES	Состояния простоя процессора	Нет
INTERRUPT	События прерывания	Diag, DiagEasy и Latency
INTERRUPT_STEER	События управления прерываниями	Diag, DiagEasy и Latency
IPI	События межпроцессорного прерывания	Нет
KE_CLOCK	События конфигурации часов	Нет
KQUEUE	Постановка/удаление из очереди ядра	Нет
LOADER	События загрузки/выгрузки образа ядра и пользовательского режима	Base
MEMINFO	Информация о списке памяти	Base, ResidentSet и ReferenceSet
MEMINFO_WS	Информация о рабочем наборе	Base и ReferenceSet
MEMORY	Трассировка памяти	ResidentSet и ReferenceSet
NETWORKTRACE	Сетевые события, например передача/получение TCP/UDP	Сеть
OPTICAL_IO	Оптический ввод/вывод	Нет
OPTICAL_IO_INIT	Инициализация оптического ввода/вывода	Нет
PERF_COUNTER	Счетчики производительности процесса	Diag и DiagEasy
PMC_PROFILE	События выборки PMC	Нет
POOL	Трассировка пула	Нет
POWER	События управления питанием	ResumeTrace
PRIORITY	События изменения приоритета	Нет

Имя	Описание	Группа
PROC_THREAD	Процесс создания/удаления потока	База
PROFILE	Пример профиля процессора	SysProf
REFSET	Поддержка анализа занимаемого пространства	ReferenceSet
REG_HIVE	Отслеживание куста реестра	Нет
REGISTRY	Отслеживание реестра	Нет
SESSION	События создания/удаления сеанса	ResidentSet и ReferenceSet
SHOULDYIELD	Отслеживание кооперативного механизма DPC	Нет
SPINLOCK	Коллизии спин-блокировки	Нет
SPLIT_IO	Разделение ввода/вывода	Нет
SYSCALL	Системные вызовы	Нет
TIMER	Настройки таймера и срок его действия	Нет
VAMAP	Информация о файле карты	ResidentSet и ReferenceSet
VIRT_ALLOC	Резерв и освобождение виртуального распределения	ResidentSet и ReferenceSet
WDF_DPC WDF	События WDF DPC	Нет
WDF_INTERRUPT	События прерываний WDF	Нет

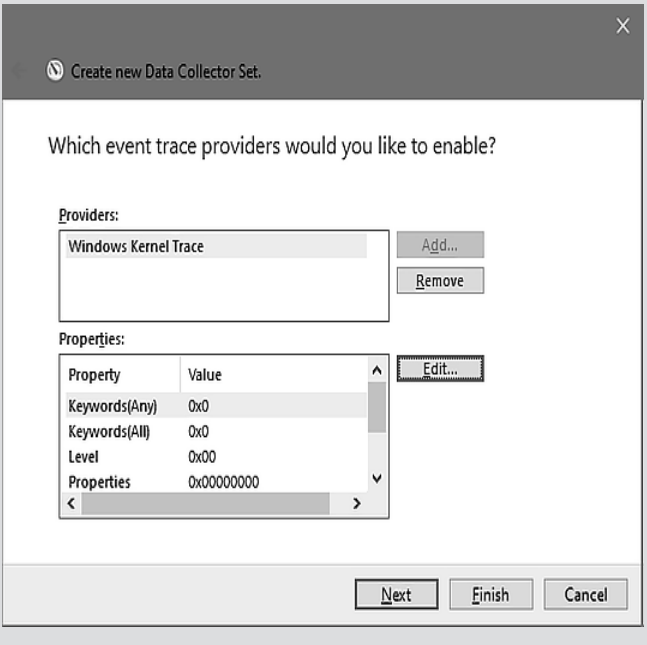
При запуске системного сеанса события немедленно регистрируются. Нет поставщика, которого нужно активировать. Это означает, что приложение-потребитель не имеет возможности декодировать событие в общем виде. События системного логгера используют точный формат кодирования событий, называемый NTPERF, который зависит от типа события. Однако большинство структур данных, представляющих различные события журнала ядра NT, обычно документируются в SDK платформы Windows.

ЭКСПЕРИМЕНТ. Отслеживание активности TCP/IP с помощью логгера ядра

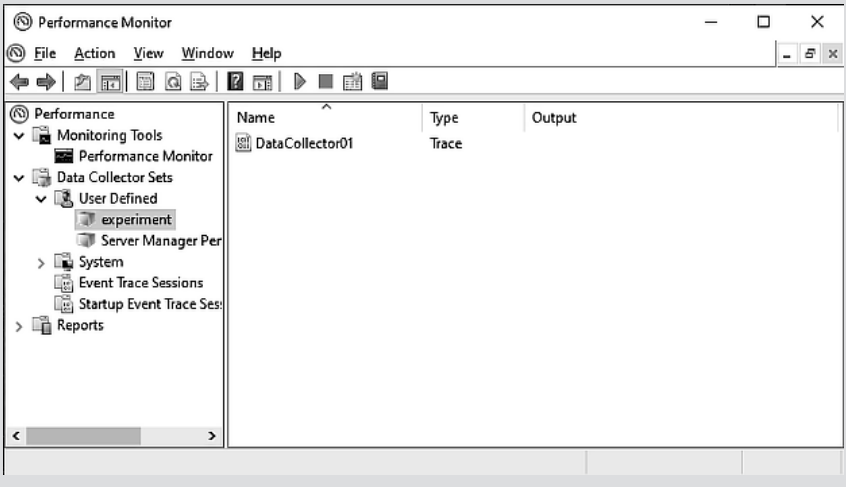
В этом эксперименте вы прослушаете события сетевой активности, генерируемые системным логгером, с помощью монитора производительности Windows. Как уже было показано в эксперименте «Перечисление сеансов ETW», этот графический инструмент способен не только получать данные от счетчиков производительности системы, но и запускать и останавливать сеансы ETW (включая системный сеанс), а также управлять ими. Чтобы включить логгер ядра и создать файл журнала активности TCP/IP, выполните следующие действия.

1. Запустите монитор производительности, введя `perfmon` в поле поиска Cortana, и выберите Наборы сборщиков данных (Data Collector Sets), а далее Определяемые пользователем (User Defined).

2. Щелкните правой кнопкой мыши на Определено пользователем (User Defined), выберите Создать (New), а затем Набор сборщиков данных (Data Collector Set).
3. При появлении запроса введите имя набора сборщиков данных, например `experiment`, и выберите Создать вручную (дополнительно) (Create Manually (Advanced)), прежде чем нажать Далее (Next).
4. В открывшемся диалоговом окне выберите Создать журналы данных (Create Data Log), установите флажок Данные трассировки событий (Event Trace Data) и нажмите на Далее (Next). В области Поставщики (Providers) нажмите Добавить (Add) и найдите Трассировка ядра Windows (Windows Kernel Trace). Нажмите OK. В списке Свойства (Properties) выберите Ключевые слова (любые) (Keywords (Any)) и нажмите на Изменить (Edit).



5. В списке, отображаемом в окне Свойства (Properties), выберите Автоматически (Automatic) и установите флажок Только сеть (Only net) для сети TCP/IP, а затем нажмите OK.
6. Нажмите на Далее (Next), чтобы выбрать место сохранения файлов. По умолчанию это расположение `%SystemDrive%\PerfLogs\Admin\experiment\`, если вы именно так назвали набор сборщиков данных. Нажмите на Далее (Next), в поле Запустить от имени (Run As) введите имя учетной записи администратора и установите соответствующий ему пароль. Нажмите на Готово (Finish). Вы должны увидеть окно, похожее на следующее.



7. Щелкните правой кнопкой мыши на имени, которое дали набору сборщиков данных (`experiment` в нашем примере), а затем нажмите кнопку Пуск (Start). Теперь создайте некоторую сетевую активность, открыв браузер и посетив веб-сайт.
8. Снова щелкните правой кнопкой мыши на узле набора сборщиков данных и выберите Стоп (Stop).

Выполнив шаги, перечисленные в эксперименте «Декодирование файла ETL», чтобы декодировать полученный файл трассировки ETL, вы обнаружите, что лучший способ прочитать результаты — использовать тип файла CSV. Это связано с тем, что сеанс системы не включает в себя никакой информации о декодировании событий, поэтому netsh.exe не имеет обычного способа кодирования настраиваемых структур данных, представляющих события в файле EVTX.

Наконец, вы можете повторить эксперимент, используя XPERF, с помощью следующей команды, при необходимости заменив файл C:\network.etl на предпочтительное имя:

```
xperf -on NETWORKTRACE -f c:\network.etl
```

После остановки сеанса трассировки системы и преобразования полученного файла трассировки вы получите события, аналогичные тем, которые были получены с помощью системного монитора.

Глобальный логгер и автологгеры

Некоторые сеансы журналирования запускаются автоматически при загрузке системы. Сеанс глобального логгера записывает события, происходящие на ранних этапах процесса загрузки операционной системы, включая события, генерируемые

логгером ядра NT. (Глобальный логгер на самом деле является системным логгером, как показано в табл. 10.16.) Приложения и драйверы устройств могут задействовать сеанс глобального логгера для записи трассировок до входа пользователя в систему (некоторые драйверы устройств, например драйверы дисковых устройств, не загружаются в момент начала сеанса глобального логгера). Хотя глобальный логгер в основном используется для захвата трассировок, создаваемых поставщиком ядра NT (см. табл. 10.17), автологгеры предназначены для захвата трассировок от классических поставщиков ETW, а не от логгера ядра NT.

Вы можете настроить глобальный логгер, задав правильные параметры реестра в разделе GlobalLogger, который находится в корневом разделе HKLM\SYSTEM\CurrentControlSet\Control\WMI. Аналогичным образом можно настроить автологгеры, создав подраздел реестра, названный как сеанс журналирования, в разделе Autologgers, расположенном в корневом разделе WMI. Процедура настройки и запуска автологгеров описана по адресу https://docs.microsoft.com/en-us/windows/win32/etw/configuring-and-starting-an-Autologger-session.

Как было сказано в разделе «Инициализация ETW» ранее в этой главе, ETW запускает глобальный логгер и автологгеры почти одновременно, на ранней стадии 1 инициализации ядра NT. Внутренняя функция `EtwStartAutoLogger` запрашивает все данные конфигурации логгера из реестра, проверяет их и создает сеанс логгера с помощью процедуры `EtwpStartLogger`, которая подробно обсуждалась в разделе «Сеансы ETW». Глобальный логгер — это системный логгер, поэтому после создания сеанса дополнительные поставщики не активируются. В отличие от глобального логгера, автологгеры требуют активации поставщиков. Они запускаются путем перечисления имени всех сеансов из раздела реестра Autologger. После создания сеанса ETW перечисляет поставщиков, которые должны быть включены в сеансе и перечислены как подразделы раздела Autologger (поставщик идентифицируется по GUID). На рис. 10.36 показаны несколько поставщиков, включенных в сеансе EventLog-System. Этот сеанс является одним из основных журналов Windows, отображаемых средством просмотра событий Windows (фиксируется службой журналирования событий).

После проверки данных конфигурации поставщика он активируется в сеансе с помощью внутренней функции `EtwpEnableTrace`, как в классических сеансах ETW.

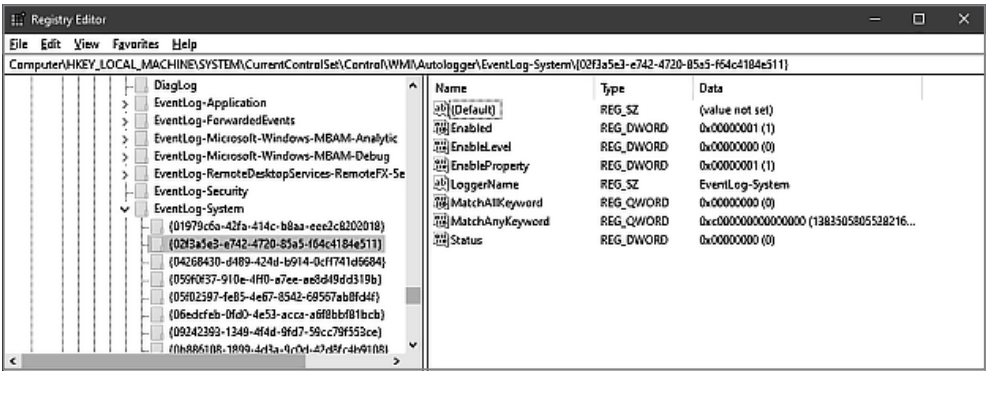


Рис. 10.36. Активированные поставщики автологгера EventLog-System

Безопасность ETW

Запуск и остановка сеанса ETW считаются операциями с высоким уровнем привилегий, поскольку события могут включать в себя системные данные, которые можно задействовать для атаки целостности системы (это особенно актуально для системных средств ведения журнала). Модель безопасности Windows была расширена для поддержки безопасности ETW. Как говорилось в предыдущих разделах, каждая операция, выполняемая ETW, требует четко определенного права доступа, которое должно быть предоставлено дескриптором безопасности, защищающим сеанс, поставщика или группу поставщиков (в зависимости от операции). В табл. 10.18 перечислены все новые права доступа, представленные для ETW, и их применение.

Таблица 10.18. Права доступа безопасности ETW и их использование

Значение	Описание	Сфера применения
WMIGUID_QUERY	Позволяет пользователю запрашивать информацию о сеансе трассировки	Сеанс
WMIGUID_NOTIFICATION	Позволяет пользователю отправлять уведомление поставщику уведомлений сеанса	Сеанс
TRACELOG_CREATE_REALTIME	Позволяет пользователю запускать или обновлять сеанс реального времени	Сеанс
TRACELOG_CREATE_ONDISK	Позволяет пользователю запускать или обновлять сеанс, который записывает события в файл журнала	Сеанс
TRACELOG_GUID_ENABLE	Разрешает пользователю активировать поставщика	Поставщик
TRACELOG_LOG_EVENT	Позволяет пользователю регистрировать события в сеансе трассировки, если тот работает в безопасном режиме	Сеанс
TRACELOG_ACCESS_REALTIME	Позволяет потребительскому приложению получать события в реальном времени	Сеанс
TRACELOG_REGISTER_GUIDS	Позволяет пользователю зарегистрировать поставщика, создавая объект EtwRegistration, поддерживаемый структурой данных ETW_REG_ENTRY	Поставщик
TRACELOG_JOIN_GROUP	Позволяет пользователю вставлять поставщика на основе манифеста или поставщика журнала трассировки в группу поставщиков (это часть свойств ETW, не описанных в этой книге)	Поставщик

Большинство прав доступа к ETW автоматически предоставляются системной учетной записи и членам групп «Администраторы», «Локальная служба» и «Сетевая служба». Это означает, что обычным пользователям нельзя взаимодействовать с ETW, если это не разрешено явно дескриптором безопасности сеанса и поставщика. Чтобы решить эту проблему, в Windows включена группа «Пользователи журнала

производительности», которая позволяет обычным пользователям взаимодействовать с ETW, особенно для управления сеансами трассировки. Хотя все права доступа ETW предоставляются дескриптором безопасности по умолчанию группе «Пользователи журнала производительности», Windows поддерживает другую группу, называемую «Пользователи системного монитора», которая предназначена только для получения уведомлений о сеансе или их отправки поставщику. Это связано с тем, что группа была разработана для доступа к счетчикам производительности системы, перечисляемым такими инструментами, как системный монитор и монитор ресурсов, а не для доступа ко всем событиям ETW. Эти два инструмента были описаны в разделе «Системный монитор и монитор ресурсов» главы 1 тома 1.

Как говорилось в разделе «Сеансы ETW» данной главы, все дескрипторы безопасности ETW хранятся в разделе реестра HKLM\System\CurrentControlSet\Control\Wmi\Security в двоичном формате. В ETW все, что представлено GUID, может быть защищено настраиваемым дескриптором безопасности. Для управления безопасностью ETW приложения обычно не взаимодействуют напрямую с дескрипторами безопасности, хранящимися в реестре, а используют API `EventAccessControl` и `EventAccessQuery`, реализованные в Sechost.dll.

ЭКСПЕРИМЕНТ. Наблюдение за дескриптором безопасности по умолчанию для сеансов ETW

Отладчик ядра может легко показать дескриптор безопасности по умолчанию, связанный с сеансами ETW, с которыми не связан конкретный дескриптор. В этом эксперименте вам понадобится компьютер с Windows 10 с уже подключенным отладчиком ядра, подключенный к хост-системе. В противном случае можете использовать локальный отладчик ядра или LiveKd (можно загрузить с https://docs.microsoft.com/en-us/sysinternals/downloads/livekd). После настройки правильных символов вы сможете создать дамп дескриптора безопасности по умолчанию с помощью команды:

```
!sd poi(nt!EtwpDefaultTraceSecurityDescriptor)
```

Текстовый вывод должен быть подобен следующему (сокращен для экономии места):

```
->Revision: 0x1
->Sbz1    : 0x0
->Control : 0x8004
            SE_DACL_PRESENT
            SE_SELF_RELATIVE
->Owner   : S-1-5-32-544
->Group   : S-1-5-32-544
->Dacl    :
->Dacl    : ->AclRevision: 0x2
->Dacl    : ->Sbz1       : 0x0
->Dacl    : ->AclSize    : 0xf0
->Dacl    : ->AceCount   : 0x9
->Dacl    : ->Sbz2       : 0x0
->Dacl    : ->Ace[0]: ->AceType: ACCESS_ALLOWED_ACE_TYPE
->Dacl    : ->Ace[0]: ->AceFlags: 0x0
->Dacl    : ->Ace[0]: ->AceSize: 0x14
->Dacl    : ->Ace[0]: ->Mask : 0x00001800
```

```
->Dacl    : ->Ace[0]: ->SID: S-1-1-0
->Dacl    : ->Ace[1]: ->AceType: ACCESS_ALLOWED_ACE_TYPE
->Dacl    : ->Ace[1]: ->AceFlags: 0x0
->Dacl    : ->Ace[1]: ->AceSize: 0x14
->Dacl    : ->Ace[1]: ->Mask : 0x00120fff
->Dacl    : ->Ace[1]: ->SID: S-1-5-18

->Dacl    : ->Ace[2]: ->AceType: ACCESS_ALLOWED_ACE_TYPE
->Dacl    : ->Ace[2]: ->AceFlags: 0x0
->Dacl    : ->Ace[2]: ->AceSize: 0x14
->Dacl    : ->Ace[2]: ->Mask : 0x00120fff
->Dacl    : ->Ace[2]: ->SID: S-1-5-19

->Dacl    : ->Ace[3]: ->AceType: ACCESS_ALLOWED_ACE_TYPE
->Dacl    : ->Ace[3]: ->AceFlags: 0x0
->Dacl    : ->Ace[3]: ->AceSize: 0x14
->Dacl    : ->Ace[3]: ->Mask : 0x00120fff
->Dacl    : ->Ace[3]: ->SID: S-1-5-20

->Dacl    : ->Ace[4]: ->AceType: ACCESS_ALLOWED_ACE_TYPE
->Dacl    : ->Ace[4]: ->AceFlags: 0x0
->Dacl    : ->Ace[4]: ->AceSize: 0x18
->Dacl    : ->Ace[4]: ->Mask : 0x00120fff
->Dacl    : ->Ace[4]: ->SID: S-1-5-32-544

->Dacl    : ->Ace[5]: ->AceType: ACCESS_ALLOWED_ACE_TYPE
->Dacl    : ->Ace[5]: ->AceFlags: 0x0
->Dacl    : ->Ace[5]: ->AceSize: 0x18
->Dacl    : ->Ace[5]: ->Mask : 0x00000ee5
->Dacl    : ->Ace[5]: ->SID: S-1-5-32-559

->Dacl    : ->Ace[6]: ->AceType: ACCESS_ALLOWED_ACE_TYPE
->Dacl    : ->Ace[6]: ->AceFlags: 0x0
->Dacl    : ->Ace[6]: ->AceSize: 0x18
->Dacl    : ->Ace[6]: ->Mask : 0x00000004
->Dacl    : ->Ace[6]: ->SID: S-1-5-32-558
```

Вы можете воспользоваться утилитой Psgetsid (доступна по адресу https://docs.microsoft.com/en-us/sysinternals/downloads/psgetsid), чтобы преобразовать SID в удобочитаемые имена. Из предыдущего вывода видно, что весь доступ к ETW предоставляется SYSTEM (S-1-5-18), LOCAL SERVICE (S-1-5-19), NETWORK SERVICE (S-1-5-18) и группе «Администраторы» (S-1-5-32-544). Как объяснялось в предыдущем разделе, группа «Пользователи журнала производительности» (S-1-5-32-559) имеет почти полный доступ к ETW, тогда как группа «Пользователи системного монитора» (S-1-5-32-558) — только право доступа `WMIGUID_NOTIFICATION`, предоставленное дескриптором безопасности сеанса по умолчанию:

```
C:\Users\andrea>psgetsid64 S-1-5-32-559

PsGetSid v1.45 – Translates SIDs to names and vice versa
Copyright (C) 1999-2016 Mark Russinovich

Sysinternals – www.sysinternals.com

Account for AALL86-LAPTOP\S-1-5-32-559:
Alias: BUILTIN\Performance Log Users
```

Логгер аудита безопасности

Логгер аудита безопасности — это сеанс ETW, используемый службой журнала событий Windows (wevtsvc.dll) для прослушивания событий, создаваемых поставщиком Security Lsass. Поставщик Security Lsass, который идентифицируется GUID {54849625-5478-4994-a5ba-3e3b0328c30d}, может быть зарегистрирован только ядром NT во время инициализации ETW и никогда не вставляется в хеш-таблицу глобального поставщика. Только логгер аудита безопасности и автологгеры, настроенные с параметром реестра `EnableSecurityProvider`, равным `1`, могут получать события от поставщика Security Lsass. Когда внутренняя функция `EtwStartAutoLogger` обнаруживает параметр, равный `1`, она включает флаг `SECURITY_TRACE` в связанном сеансе ETW, добавляя сеанс в список средств ведения журнала, которые могут получать события аудита безопасности.

Этот флаг имеет большое значение: приложения пользовательского режима больше не могут запрашивать, останавливать, сбрасывать или контролировать сеанс, если только они не выполняются как защищенный легкий процесс (на уровне защиты от вредоносных программ, Windows или WinTcb; дополнительные сведения о защищенном процессе есть в главе 3 тома 1).

Безопасные логгеры

Классические поставщики (MOF) и поставщики WPP не предназначены для поддержки всех функций безопасности, реализованных для поставщиков на основе манифестов и поставщиков журналов трассировки. Поэтому можно создать автологгер или общий сеанс ETW с флагом `EVENT_TRACE_SECURE_MODE`, который помечает сеанс как безопасный. Целью безопасного сеанса является обеспечение получения событий только от доверенных лиц. У флага имеются два основных эффекта.

- Он не позволяет классическим поставщикам (MOF) и поставщикам WPP записывать какие-либо события в защищенный сеанс. Если классический поставщик активирован в безопасном разделе, он не сможет генерировать какие-либо события.
- Требуется дополнительное право доступа `TRACELOG_LOG_EVENT`, которое должно быть предоставлено дескриптором безопасности сеанса токену доступа приложения-контроллера при одновременной активации поставщика для безопасного сеанса.

Право доступа `TRACE_LOG_EVENT` позволяет более детально указать безопасность в дескрипторе безопасности сеанса. Если дескриптор безопасности предоставляет недоверенному пользователю только `TRACELOG_GUID_ENABLE`, а сеанс ETW создается как защищенный другим объектом (более привилегированным приложением или драйвером ядра), ненадежный пользователь не может активировать ни одного поставщика в безопасном разделе. Если раздел создан как незащищенный, недоверенный пользователь может активировать в нем любых поставщиков.

ДИНАМИЧЕСКАЯ ТРАССИРОВКА

Как обсуждалось в предыдущем разделе, трассировка событий для Windows — это мощная технология отслеживания, интегрированная в ОС, но она *статична*, а это значит, что конечный пользователь может отслеживать и регистрировать только события, генерируемые четко определенными компонентами, принадлежащими операционной системе или сторонним платформам/приложениям, например .NET CLR. Чтобы преодолеть это ограничение, в майском обновлении Windows 10 (19H1) 2019 года было представлено DTrace — средство динамической трассировки, встроенное в Windows. DTrace могут применять администраторы работающих систем для проверки поведения как пользовательских программ, так и самой операционной системы. DTrace — это технология с открытым исходным кодом, которая была разработана для операционной системы Solaris (и ее потомка illumos, обе основаны на Unix) и портирована на ряд операционных систем, помимо Windows.

DTrace может динамически отслеживать части ОС и пользовательских приложений в определенных местах, называемых *зондами*. Зонд — это местоположение или действие двоичного кода, к которому DTrace может привязать запрос для выполнения набора действий, таких как регистрация сообщений, запись трассировки стека, метки времени и т. д. Когда срабатывает зонд, DTrace собирает с него данные и выполняет связанные с ним действия. И зонды, и действия указываются в файле сценария (или непосредственно в приложении DTrace через командную строку) с помощью языка программирования D. Поддержка зондов обеспечивается модулями ядра, называемыми *поставщиками*. Оригинальная версия illumos DTrace поддерживала около 20 поставщиков, которые были тесно связаны с этой ОС на базе Unix. На момент написания книги Windows поддерживает следующих поставщиков:

- **SYSCALL.** Позволяет отслеживать системные вызовы ОС (как при входе, так и при выходе), вызываемые из приложений пользовательского режима и драйверов режима ядра (через API Zw*);
- **FBT** (отслеживание границ функции). С помощью FBT системный администратор может отслеживать выполнение отдельных функций, реализованных во всех модулях, работающих в ядре NT;
- **PID** (отслеживание процессов в пользовательском режиме). Поставщик аналогичен FBT и позволяет отслеживать отдельные функции пользовательского процесса и приложения;
- **ETW** (отслеживание событий для Windows). DTrace может задействовать этого поставщика для присоединения к событиям на основе манифеста и событиям TraceLogging, запускаемым из механизма ETW. DTrace может определять новых поставщиков ETW и предоставлять связанные события ETW с помощью действия `etw_trace` (не является частью какого-либо поставщика);
- **PROFILE.** Предоставляет зонды, связанные с прерываниями по времени, срабатывающими каждый фиксированный заданный интервал времени;
- **DTRACE.** Встроенный поставщик, неявно включен в механизме DTrace.

Перечисленные поставщики позволяют системным администраторам динамически отслеживать практически все компоненты операционной системы Windows и приложений пользовательского режима.

ПРИМЕЧАНИЕ Существуют большие различия между первой версией DTrace для Windows, которая появилась в обновлении Windows 10 за май 2019 года, и нынешней стабильной версией, распространяемой на момент написания этой книги в выпуске Windows 10 за май 2021 года. Одним из наиболее заметных различий является то, что в первом выпуске требовалась настройка отладчика ядра для активации поставщика FBT. Более того, в том выпуске DTrace поставщик ETW не был полностью доступен.

ЭКСПЕРИМЕНТ. Включение DTrace и вывод списка установленных поставщиков

В этом эксперименте вы установите и включите DTrace и составляете список поставщиков, доступных для динамического отслеживания различных компонентов Windows. Вам нужна система с установленным обновлением Windows 10 за май 2020 года (20H1) или более поздней версии. Как объясняется в документации Microsoft (https://docs.microsoft.com/en-us/windows-hardware/drivers/devtest/dtrace), сначала следует включить DTrace, открыв административную командную строку и введя следующую команду (не забудьте отключить Bitlocker, если он включен):

```
bcdedit /set dtrace ON
```

После успешного выполнения команды можете загрузить пакет DTrace с сайта https://www.microsoft.com/download/details.aspx?id=100441 и установить его. Перезагрузите компьютер (или виртуальную машину) и откройте командную строку администратора, введя `CMD` в поле поиска Cortana и выбрав Запустить от имени администратора (Run As Administrator). Введите следующие команды, при необходимости заменив providers.txt другим именем файла:

```
cd /d "C:\Program Files\DTrace"
dtrace -l > providers.txt
```

Откройте сгенерированный файл (providers.txt в примере). Если DTrace успешно установлен и активирован, в выходном файле должен быть указан список зондов и поставщиков (DTrace, системный вызов и ETW). Зонды состоят из идентификатора и удобочитаемого имени. Имя состоит из четырех частей. Каждая из них может существовать, а может и не существовать в зависимости от поставщика. В целом поставщики стараются максимально точно следовать конвенции, но иногда смысл каждой части может быть перегружен чем-то другим.

- **Поставщик.** Имя поставщика DTrace, публикующего зонд.
- **Модуль.** Если зонд соответствует определенному местоположению программы, имя модуля, в котором он находится. Модуль используется только для PID (он не отображается в выводе команды `dtrace -l`) и поставщика ETW.

- **Функция.** Если зонд соответствует определенному местоположению программы, имя функции программы, в которой находится зонд.
- **Имя.** Последний компонент имени зонда — это имя, которое дает некоторое представление о его семантическом значении, например BEGIN или END.

При написании полного удобочитаемого имени зонда все его части разделяются двоеточиями, например:

```
syscall::NtQuerySystemInformation:entry
```

Это обозначает зонд для функции `NtQueryInformation`, предоставленной поставщиком системных вызовов. Обратите внимание на то, что в данном случае имя модуля пустое, поскольку поставщик системных вызовов не указывает никакого имени (все системные вызовы неявно предоставляются ядром NT).

Вместо этого поставщики PID и FBT динамически генерируют зонды на основе образа процесса или ядра, к которому они применяются (и на основе доступных в данный момент символов). Например, чтобы правильно составить список PID-зондов процесса, вы должны сначала получить идентификатор процесса, который хотите проанализировать (просто открыв диспетчер задач и выбрав список свойств Подробности (Details); в этом примере мы пользуемся Блокнотом, у которого в тестовой системе PID равен 8020). Затем выполните DTrace с помощью следующей команды:

```
dtrace -ln pid8020:::entry > pid_notepad.txt
```

Здесь перечислены все зонды записей функций, созданные поставщиком PID для процесса Блокнота. Текстовый вывод будет содержать много записей. Обратите внимание: если у вас не установлен путь к хранилищу символов, выходные данные не будут содержать никаких зондов, созданных частными функциями. Чтобы ограничить вывод, можете добавить имя модуля:

```
dtrace.exe -ln pid8020:kernelbase::entry >pid_kernelbase_notepad.txt
```

В результате все PID-зонды, сгенерированные для функциональных записей модуля kernelbase.dll, отображаются в Блокноте. Если вы повторите две предыдущие команды после установки пути к хранилищу символов с помощью следующей команды:

```
set _NT_SYMBOL_PATH=srv*C:\symbols*http://msdl.microsoft.com/download/symbols
```

то обнаружите, что выходные данные, а также зонды в приватных функциях сильно различаются.

Как поясняется в разделе «Поставщики трассировки границ функций (FBT) и процессов (PID)» далее в этой главе, поставщики PID и FBT могут применяться к любому смещению в коде функции. Следующая команда возвращает все смещения (всегда расположенные на границе инструкции), в которых поставщик PID может генерировать зонды для функции `SetComputerNameW` файла Kernelbase.dll:

```
dtrace.exe -ln pid8020:kernelbase:SetComputerNameW:
```

Внутренняя архитектура

Как объяснялось ранее в этой главе в эксперименте «Включение DTrace и вывод списка установленных поставщиков», в выпуске Windows 10 за май 2020 года (20H1) некоторые компоненты DTrace следует устанавливать через внешний пакет. Будущие версии Windows могут полностью интегрировать DTrace в образ ОС. Несмотря на то что DTrace глубоко интегрирован в операционную систему, для правильной работы требуются три внешних компонента. К ним относятся как реализация, специфичная для NT, так и исходный код DTrace, выпущенный по бесплатной общей лицензии на разработку и распространение (CDDL), которую можно загрузить по адресу https://github.com/microsoft/DTrace-on-Windows/tree/windows.

DTrace в Windows состоит из следующих компонентов (рис. 10.37):

- **DTrace.sys.** Драйвер расширения DTrace — это основной компонент, который выполняет действия, связанные с зондами, и сохраняет результаты в циклическом буфере, который приложение пользовательского режима получает через IOCTL;
- **DTrace.dll.** Модуль инкапсулирует LibDTrace — механизм пользовательского режима DTrace. Он реализует компилятор для сценариев D, отправляет IOCTL драйверу DTrace и является основным потребителем циклического буфера DTrace, где драйвер DTrace хранит выходные данные действий;
- **DTrace.exe.** Исполняемый файл точки входа, который отправляет в LibDTrace все возможные команды, указанные с помощью командной строки.

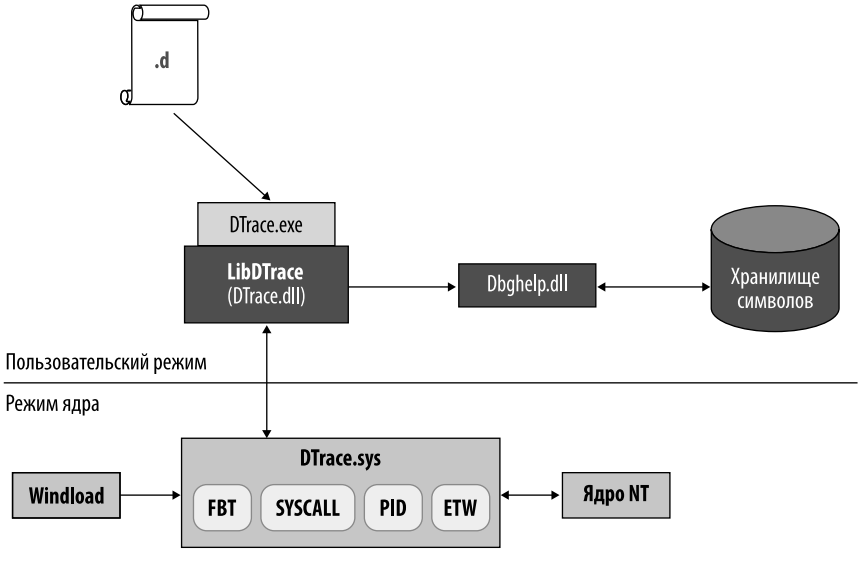


Рис. 10.37. Внутренняя архитектура DTrace

Чтобы запустить динамическую трассировку ядра Windows, драйвера или приложения пользовательского режима, пользователь просто вызывает основной исполняемый файл DTrace.exe, указав команду или внешний сценарий D. В обоих

случаях команда или файл содержат один или несколько зондов и дополнительных действий на языке программирования D. DTrace.exe анализирует входную командную строку и перенаправляет соответствующий запрос в LibDTrace, который реализован в DTrace.dll. Например, при запуске для активации одного или нескольких зондов исполняемый файл DTrace вызывает внутреннюю функцию `dtrace_program_fcompile`, реализованную в LibDTrace, которая компилирует сценарий D и создает байт-код промежуточного формата DTrace (DIF) в выходном буфере.

ПРИМЕЧАНИЕ Описание деталей байт-кода DIF и того, как компилируется сценарий D (или команды D), выходит за рамки книги. Заинтересованные читатели могут найти подробную документацию в книге спецификаций OpenDTrace, выпущенной Кембриджским университетом, которая доступна по адресу https://www.cl.cam.ac.uk/techreports/UCAM-CL-TR-924.pdf.

Хотя компилятор D полностью реализован в пользовательском режиме в LibDTrace, для выполнения скомпилированного байт-кода DIF модуль LibDtrace просто отправляет `IOCTL DTRACEIOC_ENABLE` драйверу DTrace, который реализует виртуальную машину DIF. Виртуальная машина DIF способна интерпретировать каждое предложение D, выраженное в байт-коде, и выполнять дополнительные действия, связанные с ними. Доступен ограниченный набор действий, которые выполняются с помощью собственного кода и не интерпретируются виртуальной машиной D.

Как показано на рис. 10.37, драйвер расширения DTrace реализует всех поставщиков. Прежде чем обсуждать, как работают основные поставщики, необходимо представить введение в инициализацию DTrace в ОС Windows.

Инициализация DTrace

Инициализация DTrace начинается на ранних этапах загрузки, когда загрузчик Windows загружает все модули, необходимые для правильного запуска ядра. Одной из важных частей, требующих загрузки и проверки, является файл набора API (apisetschema.dll) — ключевой компонент системы Windows. (Наборы API описаны в главе 3 тома 1.) Если в загрузочной записи установлен элемент BCD `DTRACE_ENABLED` (значение 0x26000145, которое можно установить через читаемое имя `dtrace`; дополнительные сведения об объектах BCD см. в главе 12), загрузчик Windows проверяет, присутствует ли драйвер Dtrace.sys в пути %SystemRoot%\System32\Drivers. Если это так, он создает новое расширение схемы набора API с именем `ext-ms-win-ntos-trace-l1-1-0`. Схема предназначена для драйвера Dtrace.sys и объединяется со схемой набора системных API (`OslApiSetSchema`).

Позже в процессе загрузки, когда ядро NT начинает фазу 1 инициализации, вызывается функция `TraceInitSystem` для инициализации подсистемы динамической трассировки. API импортируется в ядро NT через схему набора API `ext-ms-win-ntos-trace-l1-1-0.dll`. Это означает, что если DTrace не активирован загрузчиком Windows, разрешение имени завершится неудачно и функция, по сути, не будет работать.

У `TraceInitSystem` есть важная обязанность — вычисление содержимого массива вызовов трассировки, содержащего функции, которые будут вызываться ядром

NT при срабатывании зонда трассировки. Массив хранится в глобальном символе `KiDynamicTraceCallouts`, который в дальнейшем будет защищен Patchguard, чтобы предотвратить несанкционированное перенаправление вредоносными драйверами потока выполнения системных подпрограмм. Наконец, через функцию `TraceInitSystem` ядро NT отправляет драйверу DTrace еще один важный массив, который содержит частные системные интерфейсы, используемые драйвером DTrace для применения зондов. (Массив представлен в структуре данных контекста расширения трассировки.) Этот вид инициализации, при котором драйвер DTrace и ядро NT обмениваются частными интерфейсами, является основной причиной того, почему драйвер DTrace называется драйвером расширения.

Позже диспетчер Pnp запускает драйвер DTrace, который устанавливается в системе в качестве загрузочного драйвера, и вызывает его основную точку входа `DriverEntry`. Процедура регистрирует управляющее устройство \Device\DTrace и его символическую ссылку \GLOBAL??\DTrace. Затем он инициализирует внутреннее состояние DTrace, создавая первый встроенный поставщик DTrace. Наконец, он регистрирует всех доступных поставщиков, вызывая функцию инициализации каждого из них. Метод инициализации зависит от каждого поставщика и обычно заканчивается вызовом внутренней функции `dtrace_register`, которая регистрирует поставщика в платформе DTrace. Еще одно распространенное действие при инициализации поставщика — регистрация обработчика устройства управления. Приложения пользовательского режима могут взаимодействовать с DTrace и поставщиком через устройство управления DTrace, которое предоставляет виртуальные файлы (обработчики) поставщикам. Например, LibDTrace пользовательского режима напрямую взаимодействует с поставщиком PID, открывая дескриптор виртуального файла `\\.\DTrace\Fasttrap` (обработчика).

Поставщик системных вызовов

Когда поставщик системных вызовов активируется, DTrace вызывает подпрограмму `KeSetSystemServiceCallback`, чтобы активировать обратный вызов для системного вызова, указанного в зонде. Процедура доступна драйверу DTrace благодаря массиву системных интерфейсов NT. Последний компилируется ядром NT во время инициализации DTrace (более подробную информацию см. в предыдущем разделе) и инкапсулируется в структуру данных контекста расширения с именем `KiDynamicTraceContext`. При первом вызове `KeSetSystemServiceCallback` подпрограмма выполняет важную задачу — создает глобальную таблицу трассировки служб `KiSystemServiceTraceCallbackTable`, которая представляет собой красно-черное дерево, содержащее дескрипторы всех доступных системных вызовов. Каждый дескриптор включает в себя хеш имени системного вызова, его адрес, а также ряд параметров и флагов, указывающих, активирован ли обратный вызов при входе или выходе. Ядро NT включает статический список системных вызовов, предоставляемый через внутренний массив `KiServicesTab`.

После заполнения глобальной таблицы трассировки службы `KeSetSystemServiceCallback` вычисляет хеш имени системного вызова, указанного зондом, и ищет хеш в красно-черном дереве. Если совпадений нет, значит, зонд указал неправильное имя системного вызова, поэтому функция завершает работу, сигнализируя об

ошибке. В противном случае она изменяет флаги активации, расположенные в дескрипторе найденного системного вызова, и увеличивает количество активированных обратных вызовов трассировки, которое хранится во внутренней переменной.

Когда первый обратный вызов системного вызова DTrace включен, ядро NT устанавливает бит системного вызова в глобальной битовой маске `KiDynamicTraceMask`. Это очень важно, поскольку позволяет обработчику системных вызовов `KiSystemCall64` вызывать глобальные обработчики трассировки. (Системные вызовы и диспетчеризация системных служб подробно обсуждались в главе 8.)

Такая конструкция позволяет DTrace сосуществовать с механизмом обработки системных вызовов без снижения производительности. Если ни один зонд системного вызова DTrace неактивен, обработчики трассировки не вызываются. Обработчик трассировки может быть вызван при входе и выходе системного вызова. Его функциональность проста: он просматривает глобальную таблицу трассировки службы в поисках дескриптора системного вызова. Найдя дескриптор, он проверяет, установлен ли флаг активации, и, если да, выполняет правильный вызов, содержащийся в глобальном массиве вызовов динамической трассировки `KiDynamicTraceCallouts`, как сказано в предыдущем разделе. Вызов, реализованный в драйвере DTrace, использует общую внутреннюю функцию `dtrace_probe` для запуска зонда системного вызова и выполнения связанных с ним действий.

Поставщики трассировки границ функций (FBT) и процессов (PID)

Поставщики FBT и PID схожи, поскольку позволяют включать зонд в любой точке входа и выхода функции — не обязательно в системном вызове. Целевая функция может находиться в ядре NT или в драйвере (в этих случаях задействуется поставщик FBT) либо в модуле пользовательского режима, который должен выполняться процессом. (Поставщик PID может отслеживать приложения пользовательского режима.) Зонд FBT или PID активируется в системе с помощью кодов операций точки останова (INT 3 в x86, BRK в ARM64), которые записываются непосредственно в коде целевой функции. Из этого вытекают важные последствия.

- При срабатывании зонда PID или FBT DTrace нужна возможность повторно выполнить замененную инструкцию перед обратным вызовом целевой функции. Для этого DTrace использует эмулятор инструкций, который на момент написания статьи совместим с архитектурой AMD64 и ARM64. Эмулятор реализован в ядре NT и обычно вызывается обработчиком системных исключений при работе с исключением точки останова.
- DTrace нужен способ идентификации функций по имени. Имя функции (за исключением экспортируемых) никогда не компилируется в окончательный двоичный файл. Для этого DTrace использует несколько методов, которые будут обсуждаться в разделе «Библиотека типов DTrace» далее в этой главе.
- Одна функция может выйти несколькими способами из разных ветвей кода (закончить работу). Для определения точек выхода необходим анализатор графа функции, который дизассемблирует инструкции функции и находит каждую точку выхода. Несмотря на то что исходный анализатор графов функций был частью кода Solaris, реализация DTrace для Windows использует его

новую оптимизированную версию, которая до сих пор находится в библиотеке LibDTrace (DTrace.dll). В то время как функции пользовательского режима анализируются анализатором графа функций, DTrace задействует информацию о раскрутке PDATA v2 для надежного поиска точек выхода из функций режима ядра (более подробную информацию о раскрутке функций и диспетчеризации исключений можно найти в главе 8). Если модуль режима ядра не применяет информацию о раскрутке PDATA v2, поставщик FBT не будет создавать для него никаких зондов при выходе из функций.

DTrace устанавливает зонды FBT или PID, вызывая функцию `KeSetTracepoint` ядра NT, доступную через массив интерфейсов системы NT. Функция просматривает параметры (в частности, указатель обратного вызова) и для целей ядра проверяет, находится ли целевая функция в разделе исполняемого кода известного модуля режима ядра. Подобно поставщику системных вызовов структура данных `KI_TRACEPOINT_ENTRY` создается и используется для отслеживания активированных точек трассировки. Структура данных содержит процесс-владелец, режим доступа и адрес целевой функции. Он вставляется в глобальную хеш-таблицу `KiTpHashTable`, которая выделяется при первой активации зонда FBT или PID. Наконец, единственная инструкция, расположенная в целевом коде, анализируется (импортируется в эмулятор) и заменяется кодом операции точки останова. Устанавливается бит ловушки в глобальной битовой маске `KiDynamicTraceMask`.

Для целей режима ядра замена точки останова может произойти только в случае, если включена VBS (безопасность на основе виртуализации). Процедура `MmWriteSystemImageTracepoint` находит запись таблицы данных загрузчика, связанную с целевой функцией, и делает безопасный вызов `SECURESERVICE_SET_TRACEPOINT`. Secure Kernel (безопасное ядро) — единственный объект, способный взаимодействовать с HyperGuard и таким образом превращать вставку точки останова в законную модификацию кода. Как объяснялось в главе 7 тома 1, защита Kernel Patch (также известная как PatchGuard) предотвращает выполнение любых модификаций кода в ядре NT и некоторых важных драйверах ядра. Если VBS не включен в системе и отладчик не подключен, возвращается код ошибки и приложение зонда завершается сбоем. Если подключен отладчик ядра, код операции точки останова применяется ядром NT через функцию `MmDbgCopyMemory`. (Patchguard не действует в отлаживаемых системах.)

При вызове исключений отладчика, которые могут быть инициированы срабатыванием зонда FTB или PID DTrace, обработчик системных исключений `KiDispatchException` проверяет, установлен ли бит ловушки в глобальной битовой маске `KiDynamicTraceMask`. Если установлен, обработчик исключений вызывает функцию `KiTpHandleTrap`, которая выполняет поиск в `KiTpHashTable`, чтобы определить, произошло ли исключение из-за зарегистрированного срабатывания зонда FTB или PID. Для зондов пользовательского режима функция проверяет, является ли контекст процесса ожидаемым. Если да или если зонд работает в режиме ядра, функция напрямую производит обратный вызов DTrace `FbtpCallback`, который выполняет действия, связанные с зондом. После завершения обратного вызова обработчик вызывает эмулятор, который эмулирует исходную первую инструкцию целевой функции перед передачей ей контекста выполнения.

ЭКСПЕРИМЕНТ. Отслеживание динамической памяти

В этом эксперименте вы динамически отслеживаете динамическую память, примененную к виртуальной машине. Используя диспетчер Hyper-V, вам необходимо создать виртуальную машину поколения 2 и применить минимум 768 Мбайт и неограниченный максимальный объем динамической памяти (дополнительную информацию о динамической памяти и Hyper-V можно найти в главе 9). На виртуальной машине должно быть установлено обновление Windows 10 от мая 2019 года (19H1), мая 2020-го (20H1) или более поздней версии, а также пакет DTrace (который должен быть включен, как описано в эксперименте «Включение DTrace и вывод списка установленных поставщиков» ранее в этой главе).

Сценарий `dynamic_memory.d`, который можно найти в загружаемых ресурсах этой книги, необходимо скопировать в каталог DTrace и запустить, введя в окне административной командной строки следующие команды:

```
cd /d "c:\Program Files\DTrace"
dtrace.exe -s dynamic_memory.d
```

При использовании только предыдущих команд DTrace откажется компилировать скрипт из-за ошибки, подобной следующей:

```
dtrace: failed to compile script dynamic_memory.d: line 62: probe description
fbt:nt:MiRemovePhysicalMemory:entry does not match any probes
```

Это связано с тем, что в стандартных конфигурациях путь к хранилищу символов не задан. Сценарий присоединяет поставщика FBT к двум функциям ОС: `MmAddPhysicalMemory`, которая экспортируется из двоичного файла ядра NT, и `MiRemovePhysicalMemory`, которая не экспортируется и не публикуется в общедоступном WDK. В последнем случае у поставщика FBT нет возможности вычислить свой адрес в системе.

DTrace может получать информацию о типах и символах из разных источников, как описано в разделе «Библиотека типов DTrace» далее в этой главе. Чтобы поставщик FBT мог корректно работать с внутренними функциями ОС, следует установить путь к хранилищу символов, указывающий на общедоступный сервер символов Microsoft, используя следующую команду:

```
set _NT_SYMBOL_PATH=srv*C:\symbols*http://msdl.microsoft.com/download/symbols
```

Если после установки пути к хранилищу символов вы перезапустите DTrace, нацелив его на сценарий `Dynamic_memory.d`, он сможет правильно скомпилировать его и отобразить следующий вывод:

```
The Dynamic Memory script has begun.
```

Теперь следует смоделировать ситуацию с высокой нагрузкой на память. Вы можете сделать это несколькими способами, например, запустив предпочитаемый браузер и открыв множество вкладок, запустив 3D-игру или просто используя утилиту TestLimit с командным переключателем `-d`, который заставит систему последовательно выделять памяти ресурсы и писать в нее, пока все ресурсы

не будут исчерпаны. Рабочий процесс виртуальной машины в корневом разделе должен обнаружить сценарий и добавить память в дочернюю виртуальную машину. Это будет обнаружено DTrace:

```
Physical memory addition request intercepted. Start physical address 0x00112C00,
Number of pages: 0x00000400.
    Addition of 1024 memory pages starting at PFN 0x00112C00 succeeded!
```

Аналогично, если вы закроете все приложения на гостевой виртуальной машине и воссоздадите сценарий с высокой нехваткой памяти в своей хост-системе, сценарий сможет перехватывать запросы на удаление динамической памяти:

```
Physical memory removal request intercepted. Start physical address 0x00132000,
Number of pages: 0x00000200.
    Removal of 512 memory pages starting at PFN 0x00132000 succeeded!
```

После прерывания DTrace с помощью Ctrl+C скрипт распечатывает некоторую статистическую информацию:

```
Dynamic Memory script ended.
Numbers of Hot Additions: 217
Numbers of Hot Removals: 1602
Since starts the system has gained 0x00017A00 pages (378 MB).
```

Открыв сценарий `dynamic_memory.d` с помощью Блокнота, вы обнаружите, что он устанавливает в общей сложности шесть зондов (четыре FBT и два встроенных) и выполняет действия по журналированию и подсчету, например:

```
fbt:nt:MmAddPhysicalMemory:return
/ self->pStartingAddress != 0 /
```

Это устанавливает зонд в точках выхода функции `MmAddPhysicalMemory` только в том случае, если начальный физический адрес, полученный в точке входа функции, не равен 0. Дополнительную информацию о языке программирования D, применяемом к DTrace, можно найти в книге *The illumos Dynamic Tracing Guide*, которая находится в свободном доступе по адресу http://dtrace.org/guide/preface.html.

Поставщик ETW

DTrace поддерживает как поставщика ETW, который позволяет зондам срабатывать, когда определенные события ETW генерируются конкретными поставщиками, так и действие `etw_trace`, позволяющее сценариям DTrace генерировать новые настраиваемые события `TraceLogging` ETW. Действие `etw_trace` реализовано в LibDTrace, который использует API-интерфейсы `TraceLogging` для динамической регистрации нового поставщика ETW и создания связанных с ним событий. Дополнительная информация о ETW была представлена в разделе «Трассировка событий для Windows» ранее в этой главе.

Поставщик ETW реализован в драйвере DTrace. Когда механизм трассировки инициализируется диспетчером Pnp, он регистрирует всех поставщиков в механизме DTrace. Во время регистрации поставщик ETW настраивает сеанс ETW под названием `DTraceLoggingSession` на запись событий в циклический буфер. Когда DTrace запускается из командной строки, он отправляет IOCTL драйверу DTrace. Обработчик IOCTL вызывает функцию предоставления каждого поставщика, внутренняя функция `DtEtwpCreate` вызывает API `NtTraceControl` с кодом функции `EtwEnumTraceGuidList`. Это позволяет DTrace перечислить всех поставщиков ETW, зарегистрированных в системе, и создать зонд для каждого из них. (Также `dtrace -l` может отображать зонды ETW.)

Когда сценарий D, предназначенный для поставщика ETW, компилируется и выполняется, внутренняя подпрограмма `DtEtwEnable` вызывается с целью включения одного или нескольких зондов ETW. Сеанс журналирования, настроенный во время регистрации, запускается, если этого еще не было сделано. Через контекст расширения трассировки (который, как обсуждалось ранее, содержит частные системные интерфейсы) DTrace может регистрировать обратный вызов в режиме ядра, выполняемый каждый раз, когда новое событие регистрируется в сеансе журналирования DTrace. При первом запуске сеанса с ним не связаны никакие поставщики. Подобно поставщику системных вызовов и FBT для каждого зонда DTrace создает структуру данных отслеживания и вставляет ее в глобальное красно-черное дерево (`DtEtwpProbeTree`), представляющее все включенные зонды ETW. Структура данных отслеживания важна, поскольку она представляет собой связь между поставщиком ETW и его зондами. DTrace вычисляет правильный уровень активации и битовую маску ключевого слова для поставщика (более подробную информацию см. в разделе «Активация поставщика» ранее в этой главе) и активирует поставщика в сеансе, вызывая API `NtTraceControl`.

При создании события подсистема ETW вызывает процедуру обратного вызова, которая ищет в глобальном дереве зондов ETW правильную структуру данных контекста, представляющую зонд. При ее обнаружении DTrace может запустить зонд (все еще используя внутреннюю функцию `dtrace_probe`) и выполнить все связанные с ним действия.

Библиотека типов DTrace

DTrace работает с типами. Системные администраторы могут проверять внутренние структуры данных операционной системы и использовать их в предложениях D для описания действий, связанных с зондами. DTrace также поддерживает дополнительные типы данных вдобавок к поддерживаемым стандартным языком программирования D. Чтобы иметь возможность работать со сложными типами данных, зависящими от ОС, и позволить поставщикам FBT и PID устанавливать зонды для внутренних функций ОС и приложений, DTrace получает информацию из разных источников.

- Имена функций, сигнатуры и типы данных изначально извлекаются из информации, встроенной в двоичный исполняемый файл, который соответствует формату переносимого исполняемого файла (Portable Executable), например из таблицы экспорта и отладочной информации.

- В исходном проекте DTrace операционная система Solaris включала поддержку компактного формата типа C (Compact C Type Format, CTF) в своих исполняемых двоичных файлах, которые соответствуют формату исполняемых и связываемых файлов (Executable and Linkable Format, ELF). Это позволило ОС хранить отладочную информацию, необходимую DTrace для запуска, непосредственно в своих модулях (отладочная информация может храниться также с использованием формата сжатия deflate). Версия DTrace для Windows по-прежнему поддерживает частичный CTF, который был добавлен как раздел ресурсов библиотеки LibDTrace (Dtrace.dll). CTF в библиотеке LibDTrace хранит информацию о типах, содержащуюся в общедоступном WDK (наборе драйверов Windows) и SDK (наборе разработки программного обеспечения), и позволяет DTrace работать с базовыми типами данных ОС, не требуя какого-либо файла символов.
- Большинство частных типов и сигнатур внутренних функций ОС получаются из символов PDB. Общедоступные символы PDB для большинства модулей операционной системы можно загрузить с сервера символов Microsoft. (Это те же символы, которые использует отладчик Windows.) Эти символы широко применяет поставщик FBT для правильной идентификации внутренних функций ОС, а также DTrace — для получения правильного типа параметров для каждого системного вызова и функции.

Сервер символов DTrace

DTrace включает автономный сервер символов, который может загружать символы PDB из общедоступного хранилища символов Microsoft и предоставлять их подсистеме DTrace. Сервер символов реализован в основном в LibDTrace и может быть запрошен драйвером DTrace с использованием модели инвертированного вызова. В рамках регистрации поставщиков драйвер DTrace регистрирует псевдопоставщика `SymServer`. Последний — это не настоящий поставщик, а просто ярлык для регистрации обработчика symsrv для устройства управления DTrace.

Когда DTrace запускается из командной строки, библиотека LibDTrace запускает сервер символов, открывая дескриптор устройства управления `\\.\dtrace\symsrv` (с использованием стандартного API `CreateFile`). Запрос обрабатывается драйвером DTrace с помощью обработчика IRP сервера символов, который регистрирует процесс пользовательского режима, добавляя его во внутренний список процессов сервера символов. Затем LibDTrace запускает новый поток, который отправляет фиктивный IOCTL на устройство сервера символов DTrace и бесконечно ожидает ответа от драйвера. Драйвер помечает IRP как ожидающий и завершает его только тогда, когда поставщик или подсистема DTrace требует анализа новых символов.

Каждый раз, когда драйвер завершает ожидающий IRP, поток сервера символов DTrace просыпается и использует службы, предоставляемые вспомогательной библиотекой образов Windows (Dbghelp.dll), для правильной загрузки и анализа необходимого символа. Затем драйвер ожидает отправки нового фиктивного IOCTL из потока символов. На этот раз новый IOCTL будет содержать результаты процесса анализа символов. Поток пользовательского режима снова просыпается только тогда, когда этого требует драйвер DTrace.

ОТЧЕТЫ ОБ ОШИБКАХ WINDOWS

Отчеты об ошибках Windows (Windows Error Reporting, WER) — это сложный механизм, который автоматизирует отправку как сбоев процессов в пользовательском режиме, так и сбоев системы в режиме ядра. Несколько системных компонентов были разработаны для поддержки отчетов, генерируемых при сбое процесса пользовательского режима, защищенного процесса, трастлета или ядра.

Windows 10, в отличие от своих предшественников, не включает графическое диалоговое окно, где пользователь может настроить сведения, которые Windows Error Reporting получает при сбое приложения и отправляет в Microsoft или на внутренний сервер, настроенный системным администратором. В Windows 10 апплет «Безопасность и обслуживание» панели управления может показывать пользователю историю отчетов, созданных службой отчетов об ошибках Windows при сбое приложения или ядра (рис. 10.38). Апплет может отображать также некоторую основную информацию, содержащуюся в отчете.

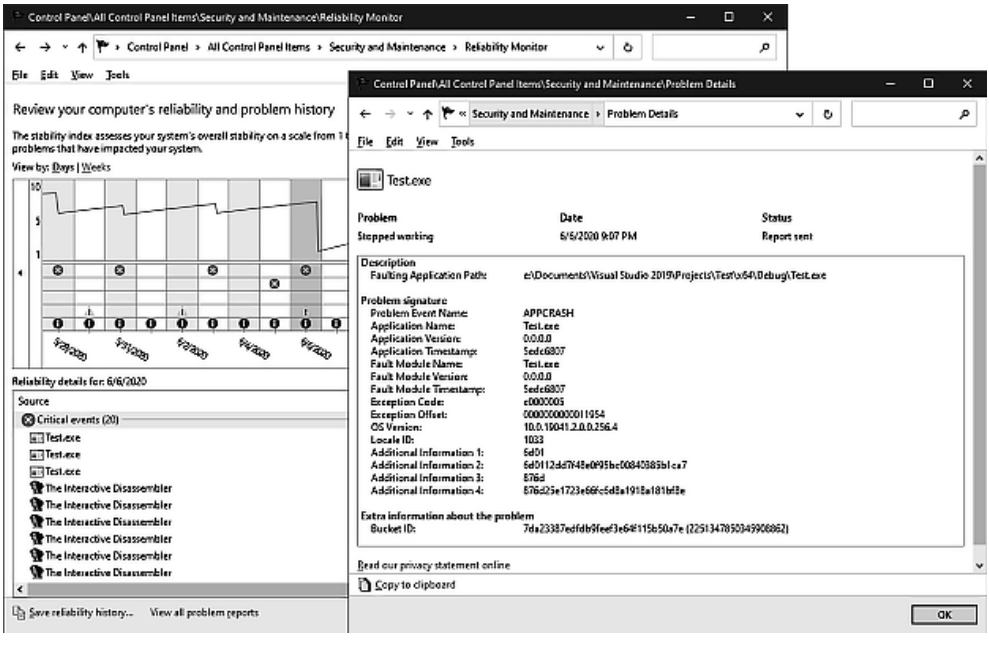


Рис. 10.38. Монитор надежности апплета «Безопасность и обслуживание» панели управления

Отчеты об ошибках Windows реализованы в нескольких компонентах ОС главным образом потому, что им необходимо справляться с различными видами сбоев.

- Служба отчетов об ошибках Windows (WerSvc.dll) — это основная служба, которая управляет созданием и отправкой отчетов при сбое процесса пользовательского режима, защищенного процесса или трастлета.
- Отчеты о сбоях Windows и безопасные отчеты о сбоях (WerFault.exe и WerFaultSecure.exe) в основном применяются для получения снимка сбойного

приложения и запуска создания отчета и его отправки на сайт Microsoft Online Crash Analysis или, если он настроен, на внутренний сервер отчетов об ошибках.

- Реальное создание и передача отчета выполняются с помощью библиотеки отчетов об ошибках Windows (Wer.dll). Она включает в себя все функции, используемые внутри механизма WER, а также некоторые экспортированные API, которые приложения могут применять для взаимодействия с отчетами об ошибках Windows (документировано по адресу https://docs.microsoft.com/en-us/windows/win32/api/_wer/). Обратите внимание на то, что некоторые API WER реализованы также в Kernelbase.dll и Faultrep.dll.
- Библиотека отчетов о сбоях пользовательского режима Windows (Faultrep.dll) содержит общий код-заглушку WER, который применяется системными модулями (Kernel32.dll, службой WER и т. д.) при сбое или зависании приложения пользовательского режима. Он включает в себя службы для создания сигнатуры сбоя и сообщает о зависании службе WER, управляя правильным контекстом безопасности для создания и передачи отчета (в том числе создание исполняемого файла WerFault под правильным токеном безопасности).
- Библиотека кодирования дампов отчетов об ошибках Windows (Werenc.dll) задействуется службой безопасных отчетов о сбоях для шифрования файлов дампов, создаваемых при сбое трастлета.
- Драйвер ядра отчетов об ошибках Windows (WerKernel.sys) — это библиотека ядра, которая экспортирует функции для захвата моментального дампа памяти ядра и отправки отчета на сайт Microsoft Online Crash Analysis. Кроме того, драйвер включает API для создания и отправки отчетов об ошибках пользовательского режима из драйвера режима ядра.

Описание всей архитектуры WER выходит за рамки книги. В этом разделе мы в основном рассматриваем отчеты об ошибках для приложений пользовательского режима и сбои ядра NT или драйвера ядра.

Сбой пользовательского приложения

Как обсуждалось в главе 3 тома 1, все потоки пользовательского режима в Windows начинаются с функции `RtlUserThreadStart`, расположенной в Ntdll. Функция не делает ничего, кроме вызова реальной процедуры запуска потока под управлением структурированного обработчика исключений. (Структурированная обработка исключений описана в главе 8.) Обработчик, защищающий процедуру реального запуска, называется обработчиком необработанных исключений, поскольку он последний, кто может управлять исключением, возникающим в потоке пользовательского режима (если сам поток этим не занимается). Обработчик, если он выполняется, обычно завершает процесс с помощью API `NtTerminateProcess`. Объектом, который решает, следует ли выполнять обработчик, является фильтр необработанных исключений `RtlpThreadExceptionFilter`. Примечательно, что фильтр необработанных исключений и обработчик выполняются только в аномальных условиях — обычно приложения должны управлять собственными исключениями с помощью внутренних обработчиков исключений.

При запуске процесса Win32 загрузчик Windows сопоставляет необходимые импортированные библиотеки. Процедура инициализации ядра устанавливает

для процесса собственный фильтр необработанных исключений — процедуру `UnhandledExceptionFilter`. Когда в потоке процесса происходит фатальное необработанное исключение, вызывается фильтр, чтобы определить, как обработать исключение. Фильтр необработанных исключений ядра собирает контекстную информацию (например, текущее значение регистров и стека машины, идентификатор сбойного процесса и идентификатор потока) и обрабатывает исключение.

- Если к процессу подключен отладчик, фильтр допускает возникновение исключения, возвращая `CONTINUE_SEARCH`. Таким образом отладчик может прервать выполнение и увидеть исключение.
- Если процесс является трастлетом, фильтр останавливает любую обработку и обращается в ядро для запуска системы безопасных отчетов о сбоях (WerFaultSecure.exe).
- Фильтр вызывает процедуру необработанного исключения CRT (если она существует), а если та не знает, как обработать исключение, вызывает внутреннюю функцию `WerpReportFault`, которая подключается к службе WER.

Прежде чем открыть соединение ALPC, `WerpReportFault` должна разбудить службу WER и подготовить наследуемый раздел общей памяти, где она хранит всю ранее полученную контекстную информацию. Служба WER — это служба прямого триггерного запуска, которую SCM запускает только в случае обновления состояния `WER_SERVICE_START` WNF или если событие записано в фиктивном поставщике ETW активации WER (с именем Microsoft-Windows-Feedback-Service-TriggerProvider). `WerpReportFault` обновляет относительное состояние WNF и ожидает события `\KernelObjects\SystemErrorPortReady`, которое сигнализируется службой WER, указывая, что она готова принимать новые подключения. После установления соединения Ntdll подключается к порту ALPC `\WindowsErrorReportingServicePort` службы WER, отправляет сообщение `WERSVC_REPORT_CRASH` и ожидает ответа в течение неопределенного времени.

Сообщение позволяет службе WER начать анализировать состояние сбойной программы и выполнить действия, необходимые для создания отчета о сбое. В большинстве случаев это означает запуск программы WerFault.exe. В случае сбоев в пользовательском режиме процесс отчета о сбоях Windows вызывается два раза с применением учетных данных процесса, ставшего причиной сбоя. В первый раз это делается для получения снимка процесса. Данная функция была введена в Windows 8.1, чтобы ускорить создание отчетов о сбоях приложений UWP, которые в то время были приложениями с одним экземпляром. Таким образом пользователь мог перезапустить сбойное приложение UWP, не дожидаясь создания отчета. (UWP и современный стек приложений обсуждаются в главе 8.)

Создание снимка

WerFault отображает раздел общей памяти, содержащий данные о сбое, и открывает сбойный процесс и поток. При вызове с аргументом командной строки `-pss`, используемым для запроса снимка процесса, он вызывает функцию `PssNtCaptureSnapshot`, экспортированную Ntdll. Последняя задействует нативные API для запроса обширной информации, касающейся сбойного процесса (например, базовой информации, сведений о задании, времени обработки, мерах безопасности, имени файла процесса

и разделе общих пользовательских данных). Кроме того, функция запрашивает информацию обо всех разделах памяти, затронутых файлом и отображаемых во всем адресном пространстве пользовательского режима процесса. Затем она сохраняет все полученные данные в структуре данных `PSS_SNAPSHOT`, представляющей снимок. И наконец, создает точную копию всего виртуального пространства сбойного процесса в другом фиктивном процессе (клонированном) с помощью API `NtCreateProcessEx`, предоставляя специальную комбинацию флагов. С этого момента исходный процесс можно завершить, а дальнейшие операции, необходимые для отчета, выполнить над клонированным процессом.

ПРИМЕЧАНИЕ WER не делает снимков для защищенных процессов и трастлетов. В этих случаях отчет создается путем получения данных из исходного процесса устранения сбоев, который приостанавливается и возобновляется только после завершения отчета.

Создание отчета о сбое

После создания моментального снимка управление выполнением возвращается к службе WER, которая инициализирует среду для создания отчета о сбое. Это делается в основном двумя способами.

- Если сбой произошел в обычном незащищенном процессе, служба WER напрямую вызывает подпрограмму `WerpInitiateCrashReporting`, экспортированную из DLL отчетов о сбоях пользовательского режима Windows (Faultrep.dll).
- В случае сбоев, принадлежащих защищенным процессам, требуется другой процесс-брокер, который создается под учетной записью системы, а не учетными данными сбойного процесса. Брокер выполняет некоторые проверки и вызывает ту же процедуру, которая применяется для сбоев, происходящих в обычных процессах.

Процедура `WerpInitiateCrashReporting`, вызываемая из службы WER, подготавливает среду для реализации правильного процесса сообщения о сбоях. Она использует API, экспортированные из библиотеки WER, для инициализации хранилища компьютера (в конфигурации по умолчанию оно находится в `C:\ProgramData\Microsoft\Windows\WER`) и загрузки всех настроек WER из реестра Windows. WER действительно содержит множество настраиваемых параметров, которые пользователь может настроить с помощью редактора групповой политики или путем внесения изменений в реестр вручную. На этом этапе WER выдает себя за пользователя, открывшего сбойное приложение, и запускает правильный процесс сообщения о сбое с помощью основного ключа командной строки `-u`, который указывает WerFault (или WerFaultSecure) обработать сбой пользователя и создать новый отчет.

ПРИМЕЧАНИЕ Если сбой произошел в современном приложении, работающем на низком уровне целостности или под токеном AppContainer, WER задействует службу диспетчера пользователей для создания нового токена среднего уровня целостности, представляющего пользователя, который запустил сбойное приложение.

В табл. 10.19 перечислены параметры конфигурации реестра WER, способы их применения и возможные значения. Последние расположены в подразделе HKLM\SOFTWARE\Microsoft\Windows\Windows Error Reporting для конфигурации компьютера и в эквивалентном пути в разделе HKEY_CURRENT_USER для индивидуальной конфигурации (некоторые значения могут присутствовать и в разделе \Software\Policies\Microsoft\Windows\Windows Error Reporting).

Таблица 10.19. Параметры конфигурации реестра WER

Параметр	Назначение	Значение
ConfigurationArchive	Содержимое архивных данных	1 для параметров, 2 для всех данных
Consent\DefaultConsent	Какие данные должны требовать согласия	1 для любых данных, 2 только для параметров, 3 для параметров и безопасных данных, 4 для всех данных
Consent\DefaultOverrideBehavior	Переопределяет ли DefaultConsent значения согласия подключаемого модуля WER	1, чтобы включить переопределение
Consent\PluginName	Значение согласия для конкретного подключаемого модуля WER	То же, что и DefaultConsent
CorporateWERDirectory	Каталог корпоративного хранилища WER	Строка, содержащая путь
CorporateWERPortNumber	Порт для использования в корпоративном хранилище WER	Номер порта
CorporateWERServer	Имя корпоративного хранилища WER	Строка, содержащая имя
CorporateWERUseAuthentication	Использование встроенной аутентификации Windows для корпоративного хранилища WER	1, чтобы включить встроенную аутентификацию
CorporateWERUseSSL	Использование Secure Sockets Layer (SSL) для корпоративного хранилища WER	1, чтобы включить SSL
DebugApplications	Список приложений, требующих от пользователя выбора между опциями Отладка и Продолжить	1, чтобы пользователь мог выбрать
DisableArchive	Включен ли архив	1, чтобы отключить архив
Отключено	Отключен ли WER	1, чтобы отключить WER
DisableQueue	Определяет, помещать ли отчеты в очередь	1, чтобы отключить очередь

Продолжение ⇩

Таблица 10.19 *(продолжение)*

Параметр	Назначение	Значение
DontShowUI	Отключает или включает пользовательский интерфейс WER	1 для отключения пользовательского интерфейса
DontSendAdditionalData	Предотвращает отправку дополнительных данных о сбое	1 — не отправлять
ExcludedApplications\ AppName	Список приложений, исключенных из WER	Строка, содержащая список приложений
ForceQueue	Следует ли отправлять отчеты в очередь пользователя	1 для отправки отчетов в очередь
LocalDumps\DumpFolder	Путь для хранения файлов дампа	Строка, содержащая путь
LocalDumps\DumpCount	Максимальное количество файлов дампа в пути	Число
LocalDumps\DumpType	Тип дампа, создаваемого во время сбоя	0 для пользовательского дампа, 1 для мини-дампа, 2 для полного дампа
LocalDumps\ CustomDumpFlags	Для пользовательских дампов указывает пользовательские параметры	Значения, определенные в MINIDUMP_TYPE (дополнительную информацию см. в главе 12)
LoggingDisabled	Включает или отключает ведение журнала	1, чтобы отключить ведение журнала
MaxArchiveCount	Максимальный размер архива (в файлах)	Значение от 1 до 5000
MaxQueueCount	Максимальный размер очереди	Значение от 1 до 500
QueuePesterInterval	Количество дней между запросами, чтобы пользователь проверил наличие решений	Количество дней

Процесс создания отчетов о сбоях Windows, запущенный с ключом `-u`, генерирует отчет: процесс снова сопоставляет раздел общей памяти, содержащий данные о сбое, идентифицирует запись и дескриптор исключения и получает ранее сделанный снимок. Если снимка не существует, процесс WerFault работает непосредственно с процессом, вызвавшим сбой, который приостанавливается. Сначала WerFault определяет природу процесса, вызывающего сбой: служебный, собственный, стандартный или процесс оболочки. Если сбойный процесс попросил систему не сообщать о каких-либо серьезных ошибках (с помощью функции `SetErrorMode`), процесс прерывается и отчет не создается. В противном случае WER проверяет, включен ли отладчик после сбоя по умолчанию, с помощью настроек, хранящихся в подразделе AeDebug (AeDebugProtected для защищенных процессов) корневого раздела реестра HKLM\SOFTWARE\Microsoft\Windows NT\CurrentVersion\. В табл. 10.20 описаны возможные параметры обоих разделов.

Таблица 10.20. Допустимые параметры реестра, используемые для корневых разделов AeDebug и AeDebugProtected

Имя параметра	Значение	Данные
Debugger	Укажите исполняемый файл отладчика, который будет запускаться при сбое приложения	Полный путь к исполняемому файлу отладчика с возможными аргументами командной строки. Переключатель –p, автоматически добавляемый WER, указывает на идентификатор процесса, вызывающего сбой
ProtectedDebugger	То же, что и Debugger, но только для защищенных процессов	Полный путь к исполняемому файлу отладчика. Недействительно для раздела AeDebug
Auto	Укажите режим автозапуска	1, чтобы разрешить запуск отладчика в любом случае, без согласия пользователя, 0 — в противном случае
LaunchNonProtected	Укажите, следует ли запускать отладчик как незащищенный. Применяется только к разделу AeDebugProtected	1 для запуска отладчика как стандартного процесса

Если для типа запуска отладчика установлен параметр `Auto`, WER запускает его и ожидает сигнала о событии отладчика, прежде чем продолжить создание отчета. Формирование отчета запускается с помощью внутренней процедуры `GenerateCrashReport`, реализованной в библиотеке DLL отчетов о сбоях пользовательского режима (Faultrep.dll). Последняя настраивает все плагины WER и инициализирует отчет с помощью функции `WerReportCreate`, экспортированной из WER.dll. (Обратите внимание на то, что на этом этапе отчет находится только в памяти.) Процедура `GenerateCrashReport` вычисляет идентификатор отчета и подпись и добавляет в отчет дополнительные диагностические данные: время обработки и параметры запуска, — или данные, определяемые приложением. После этого он проверяет конфигурацию WER, чтобы определить, какой тип дампа памяти создавать (по умолчанию создается мини-дамп). Затем вызывает экспортированный API `WerReportAddDump`, чтобы инициализировать получение дампа для процесса, вызвавшего сбой (он будет добавлен в окончательный отчет). Обратите внимание: если ранее был создан снимок, он используется для получения дампа.

API `WerReportSubmit`, экспортированный из WER.dll, — это центральная процедура, которая генерирует дамп сбойного процесса, создает все файлы, включенные в отчет, отображает пользовательский интерфейс (если это настроено параметром реестра `DontShowUI`) и отправляет отчет на сервер Online Crash. Отчет обычно включает в себя:

- файл мини-дампа сбойного процесса (обычно называется memory.hdmp);
- удобочитаемый текстовый отчет, включающий информацию об исключении, рассчитанную сигнатуру сбоя, информацию об ОС, список всех файлов, связанных с отчетом, и список всех модулей, загруженных во время сбоя (этот файл обычно называется report.wer);

- файл CSV (значения, разделенные запятыми), содержащий список всех активных процессов на момент сбоя и основную информацию, например количество потоков, размер частного рабочего набора, количество серьезных ошибок и т. д.;
- текстовый файл, содержащий информацию о состоянии глобальной памяти;
- текстовый файл, содержащий информацию о совместимости приложений.

Процесс сообщения о сбоях связывается через ALPC со службой WER и отправляет команды, позволяющие службе генерировать большую часть информации, представленной в отчете. После того как все файлы были сгенерированы, если они настроены соответствующим образом, процесс Windows Fault Reporting отображает пользователю диалоговое окно (рис. 10.39), уведомляющее о том, что в целевом процессе произошла критическая ошибка. (В Windows 10 эта функция по умолчанию отключена.)

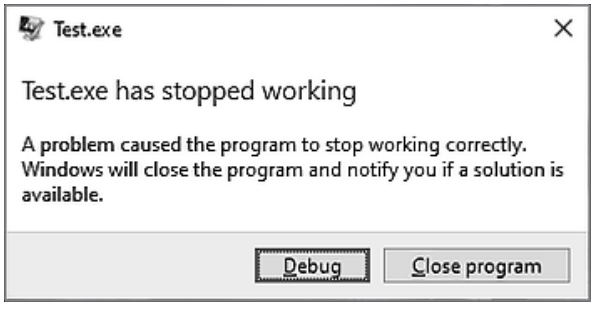


Рис. 10.39. Диалоговое окно Отчеты об ошибках Windows

В средах, где системы не подключены к Интернету или где администратор хочет контролировать, какие отчеты об ошибках отправляются в Microsoft, в качестве места назначения отчета об ошибках можно задать внутренний файловый сервер. Мониторинг ошибок рабочего стола System Center (часть пакета оптимизации рабочего стола Microsoft) понимает структуру каталогов, созданную отчетами об ошибках Windows, и предоставляет администратору возможность выборочно отправлять в Microsoft отчеты об ошибках.

Как обсуждалось ранее, служба WER использует порт ALPC для связи со сбойными процессами. Этот механизм задействует общесистемный порт ошибок, который служба WER регистрирует с помощью функции `NtSetInformationProcess`, использующей `DbgkRegisterErrorPort`. В результате все процессы Windows имеют порт ошибки, который на самом деле является объектом порта ALPC, зарегистрированным службой WER. Ядро и фильтр необработанных исключений в Ntdll применяют этот порт для отправки сообщения службе WER, которая затем анализирует сбойный процесс. Это означает, что даже в серьезных случаях повреждения состояния потока WER по-прежнему может получать уведомления и запускать WerFault.exe для регистрации подробной информации о критической ошибке в журнале событий Windows (или для отображения пользовательского интерфейса пользователю), поэтому нет необходимости выполнять эту работу внутри самого аварийного потока. Это решает все проблемы тихого завершения процесса: пользователей уведомляют, может выполняться отладка, а администраторы служб могут видеть событие сбоя.

ЭКСПЕРИМЕНТ. Включение пользовательского интерфейса WER

Начиная с первого выпуска Windows 10 пользовательский интерфейс, отображаемый WER при сбое приложения, по умолчанию отключен. В первую очередь это связано с появлением диспетчера перезапуска (часть технологии восстановления и перезапуска приложений). Он позволяет приложениям регистрировать обратный вызов перезапуска или восстановления, реализуемый при сбое приложения, зависании или просто при необходимости перезапуска для обслуживания обновления. В результате классические приложения, которые не регистрируют обратный вызов восстановления при возникновении необработанного исключения, просто завершают работу, не отображая никакого сообщения пользователю, но корректно регистрируя ошибку в системном журнале. Как обсуждалось в этом разделе, WER поддерживает пользовательский интерфейс, который можно включить, просто добавив параметр в один из разделов WER, используемых для хранения настроек. Для этого эксперимента вы повторно включите пользовательский интерфейс WER, применив глобальный системный раздел.

Из загружаемых ресурсов книги скопируйте исполняемый файл BuggedApp и запустите его. После нажатия клавиши приложение генерирует критическое необработанное исключение, которое WER перехватывает и о котором сообщает. В конфигурациях по умолчанию сообщение об ошибке не отображается. Процесс завершается, событие ошибки сохраняется в системном журнале, а отчет формируется и отправляется без какого-либо вмешательства пользователя. Откройте редактор реестра, введя `regedit` в поле поиска Cortana, и перейдите к разделу реестра HKLM\SOFTWARE\Microsoft\Windows\Windows Error Reporting. Если параметра `DontShowUI` не существует, создайте его, щелкнув правой кнопкой мыши на корневом разделе и выбрав Создать (New), параметр DWORD (32 бита) (DWORD (32 bit) Value) и присвоив ему значение 0.

Если вы перезапустите приложение с ошибкой и нажмете клавишу, WER отобразит пользовательский интерфейс, аналогичный показанному на рис. 10.39, прежде чем завершить работу приложения, вызывающего сбой. Вы можете повторить эксперимент, добавив отладчик в раздел `AeDebug`. Запуск Windbg с ключом `-I` автоматически выполняет регистрацию, как обсуждалось в эксперименте «Наблюдение за задачей на базе COM» ранее в этой главе.

Сбой в режиме ядра (системы)

Прежде чем обсуждать, как WER участвует в обработке сбоя ядра, нужно рассказать, как ядро записывает информацию о сбое. По умолчанию все системы Windows настроены на попытку записать информацию о состоянии системы до отображения «синего экрана смерти» (BSOD) и перезагрузки системы. Вы можете просмотреть эти настройки, открыв инструмент Свойства системы (System Properties) на панели управления (в разделе Система и безопасность (System and Security), Система (System), Дополнительные параметры системы (Advanced System Settings)), перейдя на вкладку Дополнительно (Advanced), а затем нажав кнопку Настройки (Settings) в разделе

Загрузка и восстановление (Startup and Recovery). Настройки по умолчанию для системы Windows показаны на рис. 10.40.

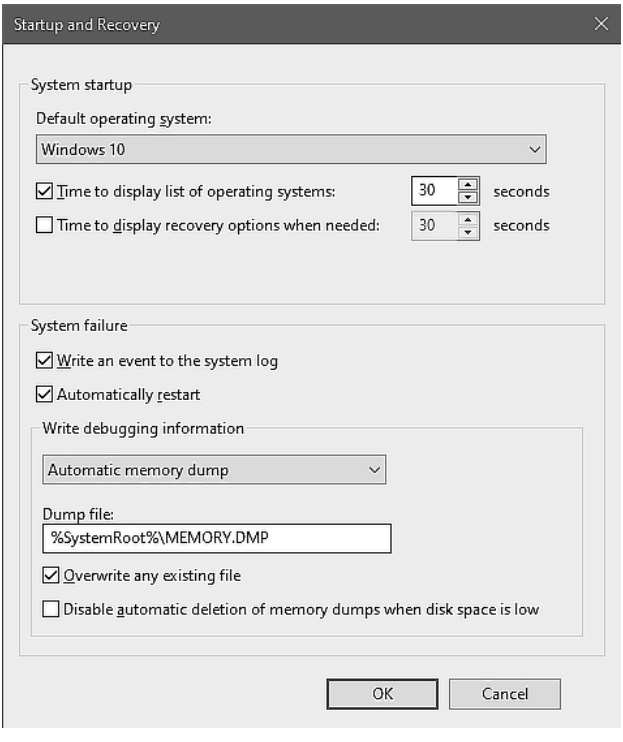


Рис. 10.40. Настройки аварийного дампа

Файлы аварийного дампа

При сбое системы может быть записана информация различных уровней.

- **Дамп активной памяти.** Дамп активной памяти содержит всю физическую память, доступную и используемую Windows на момент сбоя. Этот тип дампа является подмножеством полного дампа памяти, он просто отфильтровывает страницы, которые не имеют отношения к устранению неполадок на хост-компьютере. Этот тип дампа включает в себя память, выделенную для приложений пользовательского режима, и активные страницы, сопоставленные с ядром или пользовательским пространством, а также выбранные переходные, резервные и модифицированные страницы, такие как память, выделенная с помощью VirtualAlloc или разделы, поддерживаемые файлом подкачки. Активные дампы не включают страницы из списков свободных и обнуленных, файловый кэш, страницы гостевых ВМ и другие типы памяти, бесполезные во время отладки.
- **Полный дамп памяти.** Полный дамп памяти — это самый большой файл дампа режима ядра, содержащий все физические страницы, доступные Windows. Этот

тип дампа не поддерживается полностью на всех платформах (его заменяет дамп активной памяти). Windows требует, чтобы размер файла подкачки был не меньше объема физической памяти плюс 1 Мбайт для заголовка. Драйверы устройств могут добавить до 256 Мбайт для данных вторичного аварийного дампа, поэтому в целях безопасности рекомендуется увеличить размер файла подкачки еще на 256 Мбайт.

- **Дамп памяти ядра.** Дамп памяти ядра включает в себя только страницы режима ядра, выделенные операционной системой, HAL и драйверами устройств, которые присутствуют в физической памяти во время сбоя. Этот тип дампа не содержит страниц, принадлежащих пользовательским процессам. Но, поскольку лишь код режима ядра может напрямую вызвать сбой Windows, маловероятно, что страницы пользовательских процессов будут необходимы для отладки сбоя. Кроме того, все структуры данных, необходимые для анализа аварийного дампа, включая список запущенных процессов, стек режима ядра текущего потока и список загруженных драйверов, хранятся в невыгружаемой памяти, которая сохраняется в дампе памяти ядра. Невозможно предсказать размер дампа памяти ядра, поскольку его размер зависит от объема памяти режима ядра, выделенной операционной системой и драйверами, имеющимися на машине.
- **Автоматический дамп памяти.** Это настройка по умолчанию как для клиентских, так и для серверных систем Windows. Автоматический дамп памяти аналогичен дампу памяти ядра, но при нем сохраняются также некоторые метаданные активного процесса пользовательского режима (на момент сбоя). Кроме того, этот тип дампа позволяет лучше управлять размером системного файла подкачки. Windows может установить размер файла подкачки меньшим, чем размер оперативной памяти, но достаточно большим для того, чтобы обеспечить возможность записи дампа памяти ядра большую часть времени.
- **Небольшой дамп памяти.** Небольшой дамп памяти, размер которого обычно составляет от 128 Кбайт до 1 Мбайт, называется также мини-дампом или дампом сортировки, содержит код остановки и параметры, список загруженных драйверов устройств, структуры данных, описывающие текущий процесс и поток (они называются `EPROCESS` и `ETHREAD` и рассмотрены в главе 3 тома 1), стек ядра для потока, вызвавшего сбой, и дополнительную память, которая считается потенциально значимой из-за эвристики аварийного дампа, например страницы, на которые ссылаются регистры процессора, содержащие адреса памяти, и данные вторичного дампа, добавленные драйверами.

ПРИМЕЧАНИЕ Драйверы устройств могут зарегистрировать процедуру обратного вызова данных вторичного дампа, вызвав KeRegisterBugCheckReasonCallback. Ядро выполняет эти обратные вызовы после сбоя, а сами они для упрощения отладки могут добавлять в файл дампа сбоя дополнительные данные, такие как аппаратная память устройства или информация об устройстве. Все драйверы могут добавить до 256 Мбайт в масштабе всей системы в зависимости от места, необходимого для хранения дампа, и размера файла, в который записывается дамп. Каждый обратный вызов может добавить не более 1/8 доступного дополнительного пространства. Как только дополнительное пространство будет использовано, вызываемые впоследствии драйверы не смогут добавлять данные.

Отладчик сообщает, что при загрузке мини-дампу доступна ограниченная информация, а базовые команды, такие как `!process`, перечисляющие активные процессы, не имеют необходимых им данных. Дамп памяти ядра содержит дополнительную информацию, но переключение на сопоставления адресного пространства другого процесса не будет работать, поскольку в файле дампа отсутствуют необходимые данные. Хотя полный дамп памяти является расширенным набором других опций, его недостаток в том, что его размер соответствует объему физической памяти в системе, а потому он может оказаться довольно громоздким. Несмотря на то что код и данные пользовательского режима обычно не применяются при анализе большинства сбоев, дамп активной памяти преодолел это ограничение и сохраняет в дампе только ту память, которая фактически задействуется (исключая физические страницы, входящие в список свободных и обнуленных). Поэтому в дампе активной памяти можно переключать адресное пространство.

Преимуществом мини-дампа является его небольшой размер, что делает его удобным для обмена, например, по электронной почте. Кроме того, при каждом сбое в каталоге `%SystemRoot%\Minidump` создается файл с уникальным именем, состоящим из даты, количества миллисекунд, прошедших с момента запуска системы, и порядкового номера (например, 040712-24835-01.dmp). В случае конфликта система пытается создать дополнительные уникальные имена файлов, вызывая функцию Windows `GetTickCount`, чтобы вернуть обновленный системный счетчик тиков, а также увеличивает порядковый номер. По умолчанию Windows сохраняет последние 50 мини-дампов. Количество сохраняемых мини-дампов можно настроить, изменив параметр `MinidumpsCount` в разделе реестра HKLM\SYSTEM\CurrentControlSet\Control\CrashControl.

Существенным недостатком является то, что ограниченный объем данных, хранящихся в дампе, может помешать эффективному анализу. Вы также можете воспользоваться преимуществами мини-дампов, даже если настраиваете систему для создания дампов сбоя ядра, полных, активных или автоматических, открыв более крупный сбой с помощью WinDbg и используя команду `.dump/m` для извлечения мини-дампа. Обратите внимание на то, что мини-дамп создается автоматически, даже если в системе настроен полный дамп или дамп ядра.

ПРИМЕЧАНИЕ Вы можете использовать команду .dump из LiveKd для создания образа памяти работающей системы, который можно анализировать в автономном режиме, не останавливая систему. Этот подход полезен, когда в системе возникла проблема, но она все еще исполняет свои функции и вы хотите решить вопрос, не прерывая ее работы. Чтобы предотвратить создание образов сбоев, которые не обязательно целостны, поскольку содержимое разных областей памяти отражает разные моменты, LiveKd поддерживает флаг –m. Опция зеркального дампа создает согласованный снимок памяти в режиме ядра, используя API зеркалирования памяти диспетчера памяти, которые дают представление о системе на определенный момент.

Опция дампа памяти ядра предлагает практичную золотую середину. Поскольку он содержит всю физическую память, принадлежащую режиму ядра, то имеет тот же уровень данных, связанных с анализом, что и полный дамп памяти, но обычно он опускает нерелевантные данные и код пользовательского режима, следовательно, может быть значительно меньше. Например, в системе под управлением 64-разрядной версии Windows с 4 Гбайт ОЗУ размер дампа памяти ядра составлял 294 Мбайт.

Когда вы настраиваете дампы памяти ядра, система проверяет, достаточно ли велик файл подкачки, как описано ранее. Не существует надежного способа предсказать размер дампа памяти ядра. Поскольку размер дампа памяти ядра зависит от объема памяти режима ядра, используемой операционной системой и драйверами, присутствующими на компьютере во время сбоя, невозможно заранее предположить, каким он будет. Поэтому возможно, что в момент сбоя файл подкачки слишком мал, чтобы вместить дамп ядра, в этом случае система переключится на создание мини-дампа. Если вы хотите увидеть размер дампа ядра в свой системе, принудительно выполните сбой вручную, настроив параметр реестра, позволяющий инициировать сбой системы вручную с консоли (описано по адресу https://docs.microsoft.com/en-us/windows-hardware/drivers/debugger/forcing-a-system-crash-from-the-keyboard) или с помощью инструмента Notmyfault (https://docs.microsoft.com/en-us/sysinternals/downloads/notmyfault).

Автоматический дамп памяти решает данную проблему. Система действительно сможет создать файл подкачки достаточно большого размера, чтобы гарантировать возможность записи дампа памяти ядра большую часть времени. Если компьютер выходит из строя и размер файла подкачки недостаточно велик для сохранения дампа памяти ядра, Windows увеличивает размер файла подкачки как минимум до размера установленной физической оперативной памяти.

Чтобы ограничить объем дискового пространства, занимаемого аварийными дампами, Windows необходимо определить, следует ли ей сохранять последнюю копию ядра или полный дамп. После получения сообщения об ошибке ядра (описана далее) Windows использует следующий алгоритм, чтобы решить, нужно ли сохранять файл Memory.dmp. Если система является сервером, Windows всегда сохраняет файл дампа. В клиентской системе Windows только компьютеры, присоединенные к домену, по умолчанию всегда сохраняют аварийный дамп. Для машины, не присоединенной к домену, Windows сохраняет копию аварийного дампа только в том случае, если на целевом томе имеется более 25 Гбайт свободного дискового пространства (4 Гбайт на ARM64, настраивается через параметр реестра HKLM\SYSTEM\CurrentControlSet\Control\CrashControl\PersistDumpDiskSpaceLimit), то есть на томе, на который система настроена записывать файл Memory.dmp. Если система из-за ограничений дискового пространства не может сохранить копию файла аварийного дампа, в журнал системных событий записывается событие, говорящее о том, что файл дампа был удален (рис. 10.41). Это поведение можно переопределить, создав параметр `DWORD` реестра HKLM\SYSTEM\CurrentControlSet\Control\CrashControl\AlwaysKeepMemoryDump и установив для него значение `1`. В этом случае Windows всегда сохраняет аварийный дамп независимо от объема свободного дискового пространства.

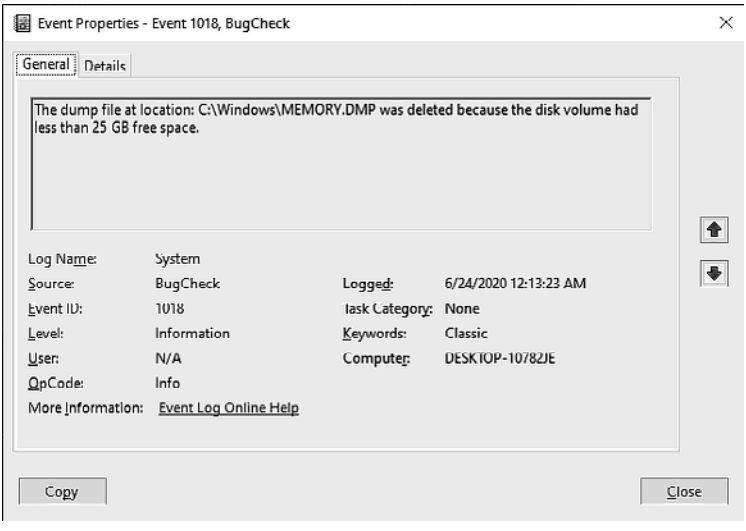


Рис. 10.41. Запись журнала событий удаления файла дампа

ЭКСПЕРИМЕНТ. Просмотр информации о файле дампа

Каждый файл дампа сбоя содержит заголовок дампа, который описывает код остановки и его параметры, тип системы, где произошел сбой (включая информацию о версии), а также список указателей на важные структуры режима ядра, необходимые во время анализа. Заголовок дампа содержит также тип записанного аварийного дампа и любую информацию, относящуюся к этому типу дампа. Команду отладчика `.dumpdebug` можно использовать для отображения заголовка файла аварийного дампа. Например, следующий вывод — результат сбоя системы, настроенной для автоматического дампа:

```
0: kd> .dumpdebug
----- 64 bit Kernel Bitmap Dump Analysis – Kernel address space is available,
      User address space may not be available.

DUMP_HEADER64:
MajorVersion        0000000f
MinorVersion        000047ba
KdSecondaryVersion  00000002
DirectoryTableBase  00000000`006d4000
PfnDataBase         ffffe980`00000000
PsLoadedModuleList  fffff800`5df00170
PsActiveProcessHead fffff800`5def0b60
MachineImageType    00008664
NumberProcessors    00000003
BugCheckCode        000000e2
BugCheckParameter1  00000000`00000000
BugCheckParameter2  00000000`00000000
BugCheckParameter3  00000000`00000000
BugCheckParameter4  00000000`00000000
KdDebuggerDataBlock fffff800`5dede5e0
```

```
SecondaryDataState  00000000
ProductType         00000001
SuiteMask           00000110
Attributes          00000000

BITMAP_DUMP:
DumpOptions         00000000
HeaderSize          16000
BitmapSize          9ba00
Pages               25dee

KiProcessorBlock at fffff800`5e02dac0
  3 KiProcessorBlock entries:
  fffff800`5c32f180 ffff8701`9f703180 ffff8701`9f3a0180
```

Команда `.enumtag` отображает все вторичные данные, хранящиеся в аварийном дампе, как показано далее. Для каждого обратного вызова вторичных данных отображаются тег, длина данных и сами данные (в байтах и формате ASCII). Разработчики могут задействовать API расширений отладчика для создания пользовательских расширений отладчика, которые также будут читать данные вторичного дампа (дополнительную информацию см. в файле справки «Инструменты отладки для Windows»):

```
{E83B40D2-B0A0-4842-ABEA71C9E3463DD1} – 0x100 bytes
  46 41 43 50 14 01 00 00 06 98 56 52 54 55 41 4C  FACP......VRTUAL
  4D 49 43 52 4F 53 46 54 01 00 00 00 4D 53 46 54  MICROSFT....MSFT
  53 52 41 54 A0 01 00 00 02 C6 56 52 54 55 41 4C  SRAT......VRTUAL
  4D 49 43 52 4F 53 46 54 01 00 00 00 4D 53 46 54  MICROSFT....MSFT
  57 41 45 54 28 00 00 00 01 22 56 52 54 55 41 4C  WAET(....."VRTUAL
  4D 49 43 52 4F 53 46 54 01 00 00 00 4D 53 46 54  MICROSFT....MSFT
  41 50 49 43 60 00 00 00 04 F7 56 52 54 55 41 4C  APIC`.....VRTUAL
...
```

Генерация аварийного дампа

Фаза 1 процесса загрузки системы позволяет диспетчеру ввода-вывода проверить настроенные параметры аварийного дампа, прочитав раздел реестра HKLM\SYSTEM\CurrentControlSet\Control\CrashControl. Если дамп настроен, диспетчер ввода-вывода загружает драйвер аварийного дампа (Crashdmp.sys) и вызывает его точку входа. Она передает обратно диспетчеру ввода-вывода таблицу функций управления, которые тот использует для взаимодействия с драйвером аварийного дампа. Диспетчер ввода-вывода также инициализирует безопасное шифрование, необходимое безопасному ядру (Secure Kernel) для хранения зашифрованных страниц в дампе. Одна из функций управления, помещенных в таблицу, инициализирует глобальную систему аварийного дампа. Он получает физические сектора (экстент файла), в которых хранятся файл подкачки и связанный с ним объект устройства тома.

Функция инициализации глобального аварийного дампа получает драйвер мини-порта, который управляет физическим диском, на котором хранится файл подкачки. Затем он использует процедуру `MmLoadSystemImageEx` для создания копии драйвера аварийного дампа и драйвера мини-порта диска, присваивая им исходные имена с префиксом строки `dump_`. Обратите внимание: это подразумевает также создание копий всех драйверов, импортированных драйвером мини-порта (рис. 10.42).

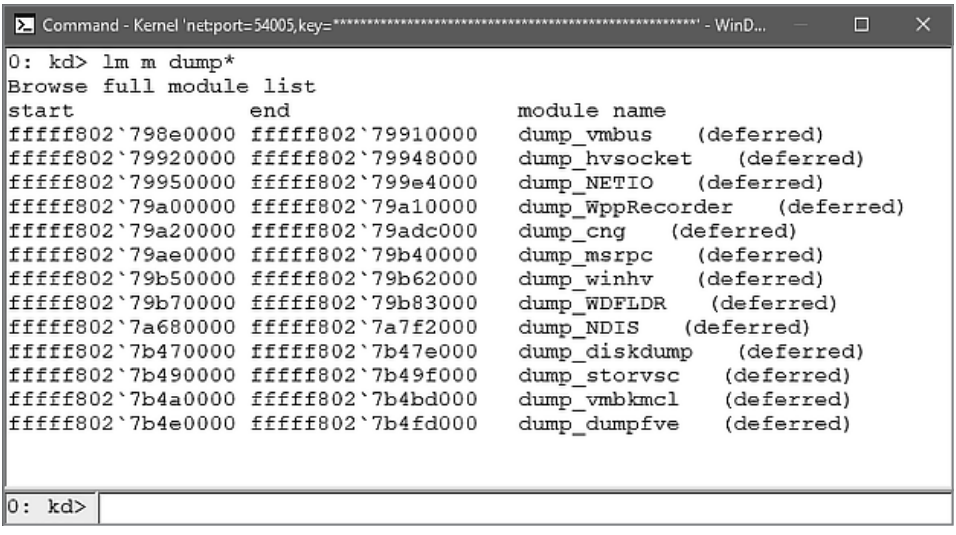


Рис. 10.42. Модули ядра, скопированные для создания и записи файла аварийного дампа

Система также запрашивает параметр `DumpFilters` для любых драйверов фильтров, необходимых для записи на том, например Dumpfve.sys — драйвера фильтра аварийного дампа шифрования диска BitLocker. Он собирает информацию, относящуюся к компонентам, участвующим в написании аварийного дампа, включая имя драйвера мини-порта диска, структуры диспетчера ввода-вывода, необходимые для записи дампа, и карту расположения файла подкачки на диске, и сохраняет две копии данных в структурах контекста дампа. Система готова сгенерировать и записать дамп, используя безопасный неповрежденный путь.

На деле при сбое системы драйвер аварийного дампа %SystemRoot%\System32\Drivers\Crashdmp.sys проверяет целостность двух структур контекста дампа, полученных при загрузке, выполняя сравнение памяти. Если совпадений нет, он не записывает аварийный дамп, поскольку это может привести к сбою или повреждению диска. В случае совпадения Crashdmp.sys при поддержке скопированного драйвера мини-порта диска и любых необходимых драйверов фильтров записывает информацию дампа непосредственно в сектора на диске, занятые файлом подкачки, минуя стек драйверов файловой системы и драйвера хранилища, которые могли быть повреждены или даже вызвать сбой.

ПРИМЕЧАНИЕ Поскольку файл подкачки открывается на ранней стадии запуска системы для использования аварийного дампа, большинство сбоев, вызванных ошибками в инициализации драйвера при запуске системы, приводят к созданию файла дампа. Сбои в ранних компонентах загрузки Windows, таких как HAL или инициализация загрузочных драйверов, происходят слишком рано, чтобы система могла иметь файл подкачки, поэтому использование другого компьютера для отладки процесса запуска — единственный способ выполнить анализ сбоев в таких случаях.

Во время процесса загрузки диспетчер сеансов (Smss.exe) проверяет параметр реестра HKLM\SYSTEM\CurrentControlSet\Control\Session Manager\Memory Management\ExistingPageFiles на наличие списка существующих файлов подкачки из предыдущей загрузки. (Дополнительную информацию о файлах подкачки см. в главе 5 тома 1.) Затем он циклически перемещается по списку, вызывая функцию SmpCheckForCrashDump для каждого существующего файла и проверяя, содержит ли он данные аварийного дампа. Он проверяет заголовок в верхней части каждого файла подкачки на предмет сигнатуры PAGEDUMP или PAGEDU64 в 32- или 64-битных системах соответственно. (Ее наличие указывает на то, что файл подкачки содержит информацию об аварийном дампе.) Если данные аварийного дампа присутствуют, диспетчер сеансов затем считывает набор параметров сбоя из раздела реестра HKLM\SYSTEM\CurrentControlSet\Control\CrashControl, включая параметр DumpFile, который содержит имя целевого файла дампа (обычно %SystemRoot%\Memory.dmp, если не указано иное).

Затем Smss.exe проверяет, находится ли целевой файл дампа на другом томе — не там, где файл подкачки. Если это так, он проверяет, достаточно ли свободного дискового пространства на целевом томе (размер, необходимый для аварийного дампа, хранится в заголовке дампа файла подкачки), прежде чем обрезать файл подкачки до размера данных сбоя и дать ему имя временного файла дампа. (Новый файл подкачки будет создан позже, когда диспетчер сеансов вызовет функцию NtCreatePagingFile.) Имя временного файла дампа имеет формат DUMP*xxxx*.tmp, где *xxxx* — текущее значение младшего слова счетчика тактов системы (она пытается найти неконфликтующее значение 100 раз). После переименования файла подкачки система удаляет из файла как скрытый, так и системный атрибуты и устанавливает соответствующие дескрипторы безопасности для защиты аварийного дампа.

Затем диспетчер сеансов создает изменчивый раздел реестра HKLM\SYSTEM\CurrentControlSet\Control\CrashControl\MachineCrash и сохраняет имя временного файла дампа в параметре DumpFile. Затем он записывает в параметр TempDestination значение типа DWORD, указывающее, является ли расположение файла дампа только временным местом назначения. Если файл подкачки находится на том же томе, что и целевой файл дампа, временный файл дампа не используется, поскольку файл подкачки усекается и напрямую переименовывается — ему дается имя целевого файла дампа. В этом случае параметр DumpFile будет соответствовать значению целевого файла дампа, а TempDestination будет равен 0.

Позднее в процедуре загрузки Wininit проверяет наличие раздела MachineCrash и, если он существует, запускает процесс отчета о сбоях Windows (Werfault.exe) с ключами командной строки -k -c (флаг k указывает на ошибку ядра, а флаг c — что полный дамп или дамп ядра должен быть преобразован в мини-дамп). WerFault считывает параметры TempDestination и DumpFile. Если для параметра TempDestination установлено значение 1, что указывает на использование временного файла, WerFault перемещает временный файл в целевое расположение и защищает целевой файл, разрешая доступ к нему только системной учетной записи и локальной группе администраторов. Затем WerFault записывает имя окончательного файла дампа в параметр FinalDumpFileLocation в разделе MachineCrash. Эти шаги показаны на рис. 10.43.

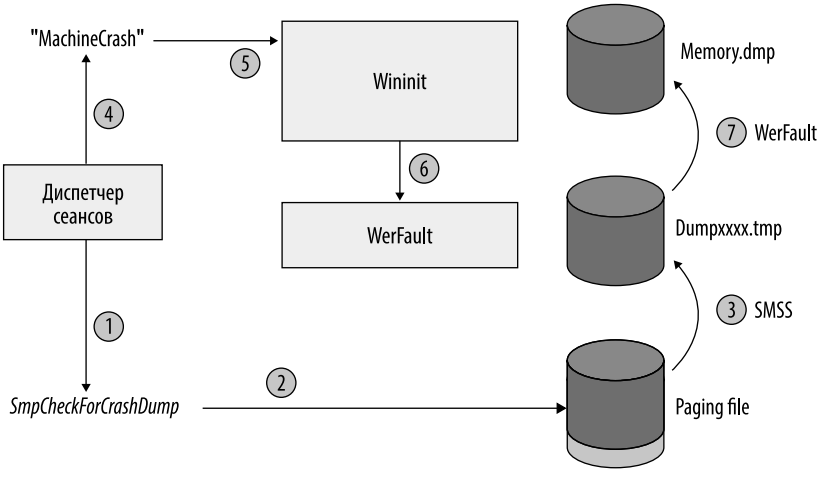


Рис. 10.43. Создание файла аварийного дампа

Чтобы обеспечить больший контроль над тем, куда записываются данные файла дампа (например, в системах, загружающихся из сети SAN, или в системах с недостаточным дисковым пространством на томе, где настроен файл подкачки), Windows также поддерживает использование выделенного файла дампа, который настроен в параметрах `DedicatedDumpFile` и `DumpFileSize` в разделе реестра HKLM\SYSTEM\CurrentControlSet\Control\CrashControl. Если указан выделенный файл дампа, драйвер аварийного дампа создает файл дампа указанного размера и записывает данные о сбое в него, а не в файл подкачки. Если параметр `DumpFileSize` не указан, Windows создает выделенный файл дампа, используя самый большой размер файла, который потребуется для хранения полного дампа. Windows рассчитывает требуемый размер как размер общего количества физических страниц памяти, присутствующих в системе, плюс размер, необходимый для заголовка дампа (одна страница в 32-разрядных системах и две страницы в 64-разрядных), плюс максимальное значение для данных вторичного аварийного дампа, составляющее 256 Мбайт. Если настроен полный дамп или дамп ядра, но на целевом томе недостаточно места для создания выделенного файла дампа необходимого размера, система возвращается к записи мини-дампа.

Отчеты ядра

После того как процесс WerFault запускается Wininit и правильно генерирует окончательный файл дампа, он создает отчет для отправки на сайт Microsoft Online Crash Analysis (или, если настроено, на внутренний сервер отчетов об ошибках). Создание отчета о сбое ядра — это процедура, которая включает в себя следующие действия.

1. Если созданный дамп не был мини-дампом, из файла дампа извлекается мини-дамп и сохраняется в расположение по умолчанию `%SystemRoot%\Minidump`, если иное не настроено с помощью параметра `MinidumpDir` в разделе HKLM\SYSTEM\CurrentControlSet\Control\CrashControl.
2. Имена файлов мини-дампа записываются в HKLM\SOFTWARE\Microsoft\Windows\Windows Error Reporting\KernelFaults\Queue.

3. Добавляется команда для выполнения WerFault.exe (%SystemRoot%\System32\WerFault.exe) с флагами `-k -rq` (флаг `rq` указывает на использование режима отчетов в очереди и необходимость перезапуска WerFault) в HKLM\SOFTWARE\Microsoft\Windows\CurrentVersion\RunOnce, чтобы WerFault выполнялся при первом входе пользователя в систему для реальной отправки отчета об ошибке.

Когда утилита WerFault запускается во время входа в систему, настроив себя на запуск, она вновь делает это с использованием флагов `-k -q` (флаг `q` указывает поочередный режим отчетов) и завершает предыдущий экземпляр. Это сделано для того, чтобы оболочка Windows не ждала WerFault и возвращала управление RunOnce как можно быстрее. Вновь запущенный экземпляр WerFault.exe проверяет раздел реестра HKLM\SOFTWARE\Microsoft\Windows\Windows Error Reporting\KernelFaults\Queue для поиска отчетов в очереди, которые могли быть добавлены на предыдущем этапе преобразования дампа. Он также проверяет, имеются ли ранее не отправленные отчеты о сбоях из предыдущих сеансов. Если они есть, WerFault.exe создает два файла в формате XML.

- Первый содержит базовое описание системы, включая версию операционной системы, список драйверов, установленных на машине, и список устройств, присутствующих в системе.
- Второй содержит метаданные, используемые службой OCA, включая тип события, вызвавшего запуск WER, и дополнительную информацию о конфигурации, например, производителя системы.

Затем `WerFault` отправляет копию двух XML-файлов и мини-дампа на сервер Microsoft OCA, который пересылает данные на ферму серверов для автоматического анализа. Автоматический анализ на ферме серверов использует тот же механизм анализа, который применяют отладчики ядра Microsoft, когда вы загружаете в них файл аварийного дампа. В результате анализа генерируется идентификатор слота, который является сигнатурой, идентифицирующей конкретный тип сбоя.

Обнаружение зависания процесса

Отчеты об ошибках Windows также используются, когда приложение зависает и прекращает работу из-за какого-либо дефекта или ошибки в его коде. Непосредственным следствием зависания приложения является то, что оно не реагирует ни на какие действия пользователя. Алгоритм обнаружения зависшего приложения, зависит от типа приложения: стек современных приложений обнаруживает, что приложение Centennial или UWP зависло, когда запрос, отправленный из HAM (диспетчера активности хоста), не обрабатывается по истечении точно определенного времени ожидания (обычно 30 с); диспетчер задач обнаруживает зависшее приложение, когда оно не отвечает на сообщение `WM_QUIT`; настольные приложения Win32 считаются не отвечающими и зависают, когда окно переднего плана перестает обрабатывать сообщения GDI более чем на 5 с.

Описание всех алгоритмов обнаружения зависаний выходит за рамки этой книги. Вместо этого мы рассмотрим наиболее вероятный случай — классическое настольное приложение Win32 перестало реагировать на любые действия пользователя. Обнаружение начинается с драйвера ядра Win32k, который после пятисекундного тайм-аута отправляет сообщение на порт ALPC DwmApiPort, созданный диспетчером

Windows рабочего стола (DWM.exe). DWM обрабатывает сообщение, используя сложный алгоритм, который в итоге создает призрачное окно поверх висящего окна. Призрак перерисовывает исходное содержимое окна, размывая его и добавляя в заголовок строку «Не отвечает». Призрачное окно обрабатывает сообщения GDI с помощью внутренней процедуры перекачки сообщений, которая перехватывает сообщения о закрытии, выходе и активации путем вызова функции `ReportHang`, экспортируемой DLL отчетов о сбоях пользовательского режима Windows (faultrep.dll). Та просто создает сообщение `WERSVC_REPORT_HANG` и отправляет его службе WER, чтобы та ожидала ответ.

Служба WER обрабатывает сообщение и инициализирует создание отчета о зависании, считывая значения параметров из корневого раздела реестра HKLM\Software\Microsoft\Windows\Windows Error Reporting\Hangs. В частности, параметр `MaxHangrepInstances` используется для указания того, сколько зависших отчетов может быть создано одновременно (если параметра не существует, то по умолчанию восемь), а параметр `TerminationTimeout` указывает время, которое должно пройти после попытки WER завершить процесс зависания, прежде чем будет считаться, что вся система зависла (по умолчанию 10 с). Это может произойти по разным причинам, например, если приложение имеет активный ожидающий IRP, который никогда не завершается драйвером ядра. Служба WER открывает зависший процесс и получает его токен и некоторую другую базовую информацию. Затем она создает объект раздела общей памяти для их хранения (аналогично сбоям пользовательского приложения, в этом случае общий раздел называется `Global\<Random GUID>`).

Процесс WerFault запускается в приостановленном состоянии с помощью токена сбойного процесса и ключа командной строки `-h`, который требует создания отчета для зависшего процесса. В отличие от случаев сбоев пользовательских приложений, снимок зависающего процесса создается из службы WER с помощью полного токена SYSTEM путем вызова функции `PssNtCaptureSnapshot`, экспортированной из Ntdll. Дескриптор снимка дублируется в приостановленном процессе WerFault, который возобновляется после успешного получения снимка. Когда WerFault запускается, он сигнализирует о событии, указывающем на начало формирования отчета. На этом этапе исходный процесс может быть прекращен. Информация для отчета получена из клонированного процесса.

Отчет для зависшего процесса аналогичен отчету, полученному для сбойного процесса: WerFault начинает с запроса значения параметра реестра отладчика, расположенного в корневом разделе глобального реестра HKLM\Software\Microsoft\Windows\Windows Error Reporting\Hangs. Если есть действующий отладчик, он запускается и присоединяется к исходному зависающему процессу. Если для параметра реестра `Disable` установлено значение `1`, процедура прерывается и процесс WerFault завершается без создания какого-либо отчета. В противном случае WerFault открывает раздел общей памяти, проверяет его и извлекает всю информацию, ранее сохраненную службой WER. Отчет инициализируется с помощью функции `WerReportCreate`, экспортированной в WER.dll и используемой также при сбое процессов. Диалоговое окно зависающего процесса (рис. 10.44) всегда отображается независимо от конфигурации WER. Наконец, функция `WerReportSubmit` (экспортированная в WER.dll) задействуется для создания всех файлов отчета, включая файл мини-дампа, аналогично случаям сбоев пользовательских приложений (см. раздел «Создание отчета о сбое» ранее в этой главе). Наконец, отчет отправляется на сервер онлайн-анализа сбоев.

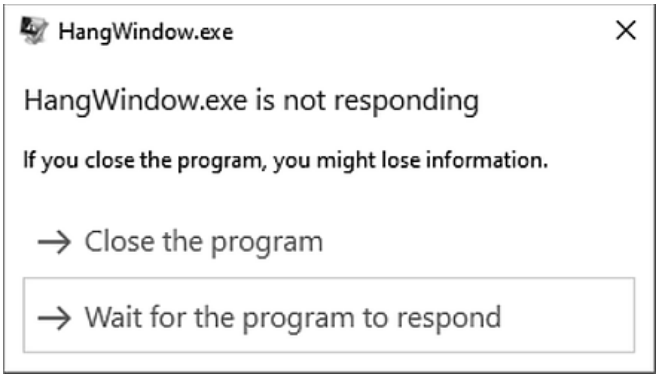


Рис. 10.44. Диалоговое окно Отчет об ошибках Windows для зависших приложений

После запуска создания отчета и возврата сообщения `WERSVC_HANG_REPORTING_STARTED` в DWM WER завершает зависший процесс с помощью API `TerminateProcess`. Если процесс не заканчивается за ожидаемое время (обычно 10 с, но его можно настроить с помощью параметра `TerminationTimeout`, как объяснялось ранее), служба WER перезапускает другой экземпляр WerFault, работающий под полным токеном SYSTEM, и ждет дольше (обычно тайм-аут 60 с, но его можно перенастроить с помощью параметра `LongTerminationTimeout`). Если процесс не завершается даже по истечении более длительного тайм-аута, у WER нет другого выбора, кроме как сделать в журнале событий приложения запись о событии ETW, сообщающую о невозможности завершить процесс. Событие ETW показано на рис. 10.45. Обратите внимание на то, что описание события вводит в заблуждение, поскольку WER не смог завершить зависшее приложение.

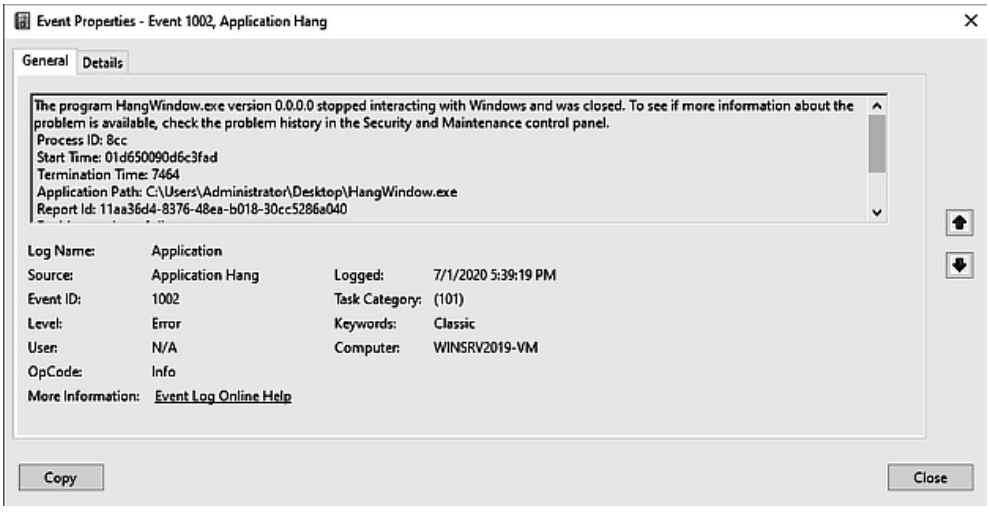


Рис. 10.45. Событие ошибки ETW, записанное в журнал приложений для незакрывающегося зависшего приложения

ГЛОБАЛЬНЫЕ ФЛАГИ

В Windows имеется набор флагов, хранящихся в двух общесистемных глобальных переменных `NtGlobalFlag` и `NtGlobalFlag2`, которые обеспечивают разнообразную внутреннюю поддержку отладки, трассировки и проверки в операционной системе. Две системные переменные инициализируются из раздела реестра HKLM\SYSTEM\CurrentControlSet\Control\Session Manager в параметрах `GlobalFlag` и `GlobalFlag2` во время загрузки системы (фаза 0 инициализации ядра NT). По умолчанию оба параметра реестра равны `0`, поэтому вполне вероятно, что в ваших системах не используются никакие глобальные флаги. Кроме того, каждый образ имеет набор глобальных флагов, которые включают также внутренний код трассировки и проверки, хотя разрядность этих флагов немного отличается от разрядности общесистемных глобальных флагов.

К счастью, инструменты отладки содержат утилиту Gflags.exe, которую можно использовать для просмотра и изменения глобальных флагов системы (либо в реестре, либо в работающей системе), а также глобальных флагов образа. GFlags имеет как командную строку, так и графический интерфейс. Чтобы просмотреть флаги командной строки, введите `gflags /?`. Если вы запустите утилиту без каких-либо переключателей, появится диалоговое окно, показанное на рис. 10.46.

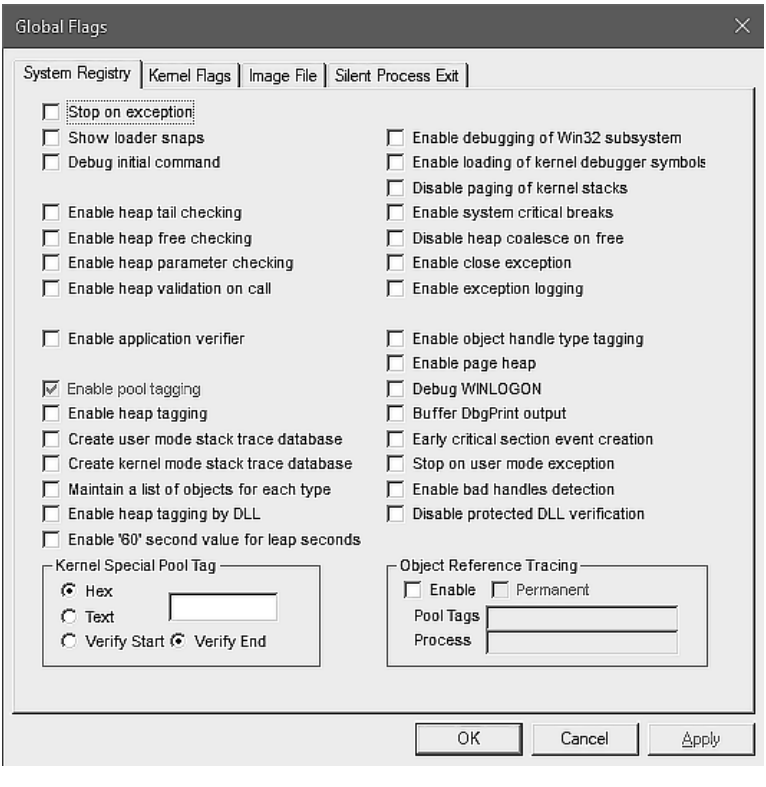


Рис. 10.46. Настройка параметров отладки системы с помощью GFlags

Флаги, относящиеся к глобальным флагам Windows, можно разделить на категории.

- Флаги ядра обрабатываются непосредственно различными компонентами ядра NT (диспетчером кучи, исключениями, обработчиками прерываний и т. д.).
- Флаги пользователя обрабатываются компонентами, работающими в приложениях пользовательского режима (обычно Ntdll).
- Флаги только загрузки обрабатываются лишь при запуске системы.
- Глобальные флаги файла образа (они имеют несколько иное значение, чем остальные) обрабатываются загрузчиком, WER и некоторыми другими компонентами пользовательского режима в зависимости от контекста процесса пользовательского режима, в котором они выполняются.

Названия групповых страниц, отображаемые инструментом GFlags, немного вводят в заблуждение. Флаги ядра, загрузочные и пользовательские флаги смешаны на каждой странице. Основное отличие состоит в том, что страница системного реестра позволяет пользователю устанавливать глобальные флаги для параметров реестра `GlobalFlag` и `GlobalFlag2`, анализируемых во время загрузки системы. Это означает, что возможные новые флаги будут включены только после перезагрузки системы. Страница Флаги ядра (Kernel Flags), несмотря на свое название, не позволяет на лету применять флаги ядра к работающей системе. Только определенные флаги пользовательского режима могут быть установлены или сняты (хорошим примером является флаг включения кучи страницы) без необходимости перезагрузки системы: инструмент GFlags устанавливает их с помощью собственного API `NtSetSystemInformation` (с информационным классом `SystemFlagsInformation`). Таким образом можно установить только флаги пользовательского режима.

ЭКСПЕРИМЕНТ. Просмотр и установка глобальных флагов

Вы можете применить команду отладчика ядра `!gflag` для просмотра и установки состояния переменной ядра `NtGlobalFlag`. Команда `!gflag` выводит список всех включенных флагов. Можно использовать `!gflag -?` чтобы получить весь список поддерживаемых глобальных флагов. На момент написания этой книги расширение `!gflag` не было обновлено для отображения содержимого переменной `NtGlobalFlag2`.

На странице Файл образа (Image File) вам необходимо ввести имя файла исполняемого образа. Используйте эту опцию, чтобы изменить набор глобальных флагов, которые применяются к отдельному образу, а не ко всей системе. Страница показана на рис. 10.47. Обратите внимание на то, что флаги отличаются от флагов операционной системы, показанных на рис. 10.46. Большинство флагов и настроек, доступных на страницах Файл образа (Image File) и Тихий выход из процесса (Silent Process Exit), применяются с сохранением новых параметров в подразделе с тем же именем, что и файл образа (то есть notepad.exe для случая, показанного на рис. 10.47) в разделе реестра HKLM\SOFTWARE\Microsoft\Windows NT\CurrentVersion\

Image File Execution Options, также известном как раздел IFEO. В частности, параметр `GlobalFlag` (и `GlobalFlag2`) представляет собой битовую маску всех доступных глобальных флагов для каждого образа.

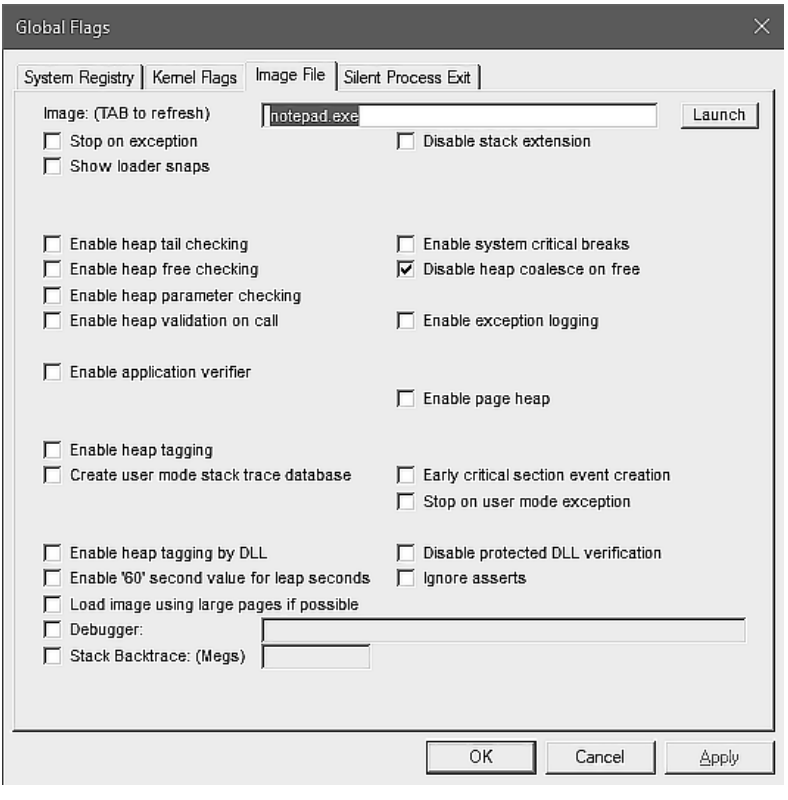


Рис. 10.47. Установка глобальных флагов для каждого образа с помощью GFlags

Когда загрузчик инициализирует новый, ранее созданный процесс и загружает все зависимые библиотеки основного базового исполняемого файла (более подробную информацию о запуске процесса см. в главе 3 тома 1), система обрабатывает глобальные флаги каждого образа. Внутренняя функция `LdrpInitializeExecutionOptions` открывает раздел IFEO на основе имени базового образа и анализирует все настройки и флаги для каждого образа. В частности, после того как глобальные флаги каждого образа извлекаются из реестра, они сохраняются в поле `NtGlobalFlag` (и `NtGlobalFlag2`) PEB процесса. Так к ним может легко получить доступ любой образ, отображаемый в процессе, включая Ntdll.

Большинство доступных глобальных флагов описаны по адресу https://docs.microsoft.com/en-us/windows-hardware/drivers/debugger/gflags-flag-table.

ЭКСПЕРИМЕНТ. Устранение проблем с загрузчиком Windows

В эксперименте «Наблюдение за загрузчиком образов» в главе 3 тома 1 мы использовали инструмент GFlags для отображения сведений времени выполнения для загрузчика Windows. Эта информация может быть полезна для понимания того, почему приложение вообще не запускается (без возврата какой-либо полезной информации об ошибке). Вы можете повторить этот эксперимент с mspaint.exe, переименовав файл Msftedit.dll (библиотека элемента управления редактированием форматированного текста), расположенный в `%SystemRoot%\system32`. Действительно, Paint косвенно зависит от этой DLL. Библиотека Msftedit загружается динамически с помощью MSCTF.dll. (Он не связан статически в исполняемом файле Paint.) Откройте окно командной строки администратора и введите следующие команды:

```
cd /d c:\windows\system32
takeown /f msftedit.dll
icacls msftedit.dll /grant Administrators:F
ren msftedit.dll msftedit.disabled
```

Затем включите снимки загрузчика с помощью инструмента GFlags, как указано в эксперименте «Наблюдение за загрузчиком образов». Если вы запустите mspaint.exe с помощью Windbg, снимки загрузчика смогут почти сразу выявить проблему, вернув следующий текст:

```
142c:1e18 @ 00056578 – LdrpInitializeNode – INFO: Calling init routine
00007FFC79258820 for
DLL "C:\Windows\System32\MSCTF.dll"142c:133c @ 00229625 – LdrpResolveDllName –
ENTER: DLL name: .\MSFTEDIT.DLL
142c:133c @ 00229625 – LdrpResolveDllName – RETURN: Status: 0xc0000135
142c:133c @ 00229625 – LdrpResolveDllName – ENTER: DLL name: C:\Program Files\
Debugging Tools for Windows (x64)\MSFTEDIT.DLL
142c:133c @ 00229625 – LdrpResolveDllName – RETURN: Status: 0xc0000135
142c:133c @ 00229625 – LdrpResolveDllName – ENTER: DLL name: C:\Windows\system32\
MSFTEDIT.DLL
142c:133c @ 00229625 – LdrpResolveDllName – RETURN: Status: 0xc0000135
. . .
C:\Users\test\AppData\Local\Microsoft\WindowsApps\MSFTEDIT.DLL
142c:133c @ 00229625 – LdrpResolveDllName – RETURN: Status: 0xc0000135
142c:133c @ 00229625 – LdrpSearchPath – RETURN: Status: 0xc0000135
142c:133c @ 00229625 – LdrpProcessWork – ERROR: Unable to load DLL: "MSFTEDIT.DLL",
Parent Module: "(null)", Status: 0xc0000135
142c:133c @ 00229625 – LdrpLoadDllInternal – RETURN: Status: 0xc0000135
142c:133c @ 00229625 – LdrLoadDll – RETURN: Status: 0xc0000135
```

ПРОСЛОЙКИ ЯДРА

Новые выпуски операционной системы Windows иногда способны создавать проблемы старым драйверам, которые могут испытывать трудности в ходе работы в новой среде, — вызывать зависания системы или синие экраны смерти. Чтобы решить эту проблему, в Windows 8.1 был представлен механизм Kernel Shim, способный динамически

изменять старые драйверы, которые могут продолжать работать в новой версии ОС. Механизм Kernel Shim реализован в основном в ядре NT. Прослойки драйвера регистрируются через реестр Windows и файл базы данных прослоек. Прослойки для драйверов предоставляются драйверами прослоек. Драйвер прослойки использует экспортированный API `KseRegisterShimEx` для регистрации прослойки, ее можно применить к целевым драйверам, которым она нужна. Механизм Kernel Shim поддерживает в основном два типа прослоек, применяемых к устройствам или драйверам.

Инициализация механизма прослоек

На ранних этапах загрузки ОС загрузчик Windows, загружая все нужные драйверы, считывает и сопоставляет файл базы данных их совместимости, расположенный в %SystemRoot%\apppatch\Drvmain.sdb (и, если он существует, также в файле Drvpatch.sdb). На первом этапе инициализации ядра NT диспетчер ввода-вывода запускает два этапа инициализации механизма Kernel Shim. Ядро NT копирует двоичное содержимое файлов базы данных в глобальный буфер, выделенный из выгружаемого пула, на который указывает внутренняя глобальная переменная KsepShimDb. Затем он проверяет, отключены ли глобальные прослойки ядра. Если система загрузилась в безопасном режиме или режиме WinPE либо включена проверка драйверов, механизм прослоек будет неактивен. Механизмом Kernel Shim можно управлять также с помощью системных политик или через параметр реестра HKLM\System\CurrentControlSet\Control\Compatibility\DisableFlags. Затем ядро NT собирает низкоуровневую системную информацию, необходимую в ходе применения прослоек устройства, как то информация BIOS и идентификатор OEM, путем проверки фиксированной таблицы дескрипторов системы ACPI (FADT). Механизм прослойки регистрирует первого встроенного поставщика прослойки с именем DriverScope с помощью API `KseRegisterShimEx`. Встроенные прослойки, предоставляемые Windows, перечислены в табл. 10.21. Некоторые из них реализованы непосредственно в ядре NT, а не в каком-либо внешнем драйвере. DriverScope — единственная прослойка, зарегистрированная на этапе 0.

Таблица 10.21. Встроенные прослойки ядра Windows

Shim Name	GUID	Purpose	Module
DriverScope	{BC04AB45-EA7E-4A11-A7BB-977615F4CAAE}	Прослойка области драйвера используется для сбора событий ETW, относящихся к работоспособности целевого драйвера. Его перехватчики не делают ничего, кроме записи события ETW до или после вызова исходных обратных вызовов без прослойки	Ядро NT
Version Lie	{3E28B2D1-E633-408C-8E9B-2AFA6F47FCC3} (7.1) {47712F55-BD93-43FC-9248-B9A83710066E} (8) {21C4FB58-D477-4839-A7EAAD6918FBC518} (8.1)	Прослойка подмены версии доступна для Windows 7.1, 8 и 8.1. Прослойка передает предыдущую версию ОС, если этого требует драйвер, в котором она применяется	Ядро NT

Shim Name	GUID	Purpose	Module
SkipDriverUnload	{3E8C2CA6-34E2-4DE6-8A1E-9692DD3E316B}	Прослойка заменяет процедуру выгрузки драйвера процедурой, которая ничего не делает, кроме регистрации событий ETW	Ядро NT
ZeroPool	{6B847429-C430-4682-B55FFD11A7B55465}	Заменяет API ExAllocatePool функцией, которая выделяет память пула и обнуляет ее	Ядро NT
ClearPCIDBits	{B4678DFF-BD3E-46C9-923B-B5733483B0B3}	Очищает биты PCID, когда некоторые антивирусные драйверы отображают физическую память, на которую ссылается ядро CR3	Ядро NT
Kaspersky	{B4678DFF-CC3E-46C9-923B-B5733483B0B3}	Прослойка исключительно для драйверов фильтров Касперского для маскировки реального параметра реестра UseVtHardware во избежание сбоев у старых версий антивируса	Ядро NT
Memcpy	{8A2517C1-35D6-4CA8-9EC8-98A12762891B}	Обеспечивает более безопасную (но более медленную) реализацию копирования памяти, которая всегда обнуляет целевой буфер и может использоваться с памятью устройства	Ядро NT
KernelPadSectionsOverride	{4F55C0DB-73D3-43F2-9723-8A9C7F79D39D}	Предотвращает освобождение отбрасываемых разделов любого модуля ядра диспетчером памяти и блокирует загрузку целевого драйвера, в котором применяется прослойка	Ядро NT
Прослойка NDIS	{49691313-1362-4e75-8c2a-2dd72928eba5}	Прослойка совместимости версий NDIS (возвращает 6.40, если применяется к драйверу)	Ndis.sys
SrbShim	{434ABAFD-08FA-4c3d-A88D-D09A88E2AB17}	Прослойка совместимости блока запросов SCSI, которая перехватывает IOCTL_STORAGE_QUERY_PROPERTY	Storport.sys
DeviceIdShim	{0332ec62-865a-4a39-b48fcda6e855f423}	Прослойка совместимости для RAID-устройств	Storport.sys
ATADeviceIdShim	{26665d57-2158-4e4b-a959-c917d03a0d7e}	Прослойка совместимости для устройств с последовательным интерфейсом ATA	Storport.sys
Прослойка питания фильтра Bluetooth	{6AD90DAD-C144-4E9DA0CF-AE9FCB901EBD}	Прослойка совместимости для драйверов фильтров Bluetooth	Bthport.sys
UsbShim	{fd8fd62e-4d94-4fc7-8a68-bff7865a706b}	Прослойка совместимости для старого USB-модема Conexant	Usbd.sys
Прослойка фильтра Nokia Usbser	{7DD60997-651F-4ECB-B893-BEC8050F3BD7}	Прослойка совместимости для драйверов фильтра Nokia Usbser (используется Nokia PC Suite)	Usbd.sys

Прослойка представлена посредством структуры данных KSE_SHIM, где KSE означает Kernel Shim Engine. Структура данных включает в себя GUID, удобочитаемое имя прослойки и массив коллекции хуков (структуры данных `KSE_HOOK_COLLECTION`). Прослойки драйверов поддерживают различные типы перехватчиков — перехваты функций, экспортируемых ядром NT, HAL и библиотеками драйверов, а также функции обратного вызова объекта драйвера. На этапе 1 инициализации механизм Shim Engine регистрирует поставщик Microsoft-Windows-Kernel-ShimEngine ETW (он имеет GUID {0bf2fb94-7b60-4b4d-9766-e82f658df540}), открывает базу данных прослоек драйверов и инициализирует оставшиеся встроенные прослойки, реализованные в ядре NT (см. табл. 10.21).

Чтобы зарегистрировать прослойку (через `KseRegisterShimEx`), ядро NT выполняет некоторые начальные проверки целостности как структуры данных KSE_SHIM, так и каждого перехватчика в коллекции (все перехватчики должны находиться в адресном пространстве вызывающего драйвера). Затем оно выделяет и заполняет структуру данных `KSE_REGISTERED_SHIM_ENTRY`, которая, как следует из названия, представляет зарегистрированную прослойку. Та содержит счетчик ссылок и указатель на объект драйвера (используется только в том случае, если прослойка не реализована в ядре NT). Выделенная структура данных связана с глобальным связанным списком, в котором отслеживаются все зарегистрированные прослойки в системе.

База данных прослоек

Формат файла базы данных прослоек (Shim Database, SDB) был представлен в Windows XP для обеспечения совместимости приложений. Первоначальной целью формата файла было хранение двоичной базы данных программ и драйверов в стиле XML, которым для правильной работы требовалась какая-то помощь со стороны операционной системы. Файл SDB был адаптирован для включения прослоек режима ядра. Формат файла описывает базу данных XML с использованием тегов. Тег — это двухбайтовая базовая структура данных, применяемая в качестве уникального идентификатора записей и атрибутов в базе данных. Он имеет четырехбитный тип, который определяет формат данных, связанных с тегом, и 12-битный индекс. Каждый тег указывает тип данных, размер и интерпретацию, соответствующую самому тегу. Файл SDB имеет 12-байтовый заголовок и набор тегов. Этот набор обычно определяет три основных блока в файле базы данных прослойки.

- Блок `INDEX` содержит индексные теги, которые служат для быстрого индексирования элементов в базе данных. Индексы в блоке `INDEX` хранятся в порядке возрастания. Поэтому поиск элемента в индексах (с помощью алгоритма двоичного поиска) — быстрая операция. В движке Kernel Shim элементы хранятся в блоке `INDEXES` с использованием восьмибайтового ключа, полученного из имени прослойки.
- Блок `DATABASE` содержит теги верхнего уровня, описывающие прослойки, драйверы, устройства и исполняемые файлы. Каждый тег верхнего уровня содержит дочерние теги, описывающие свойства или внутренние блоки, принадлежащие корневому объекту.

- Блок `STRING TABLE` содержит строки, на которые ссылаются теги нижнего уровня в блоке `DATABASE`. Теги в этом блоке обычно не описывают строку напрямую, а содержат ссылку на тег (он называется `STRINGREF`), описывающий строку, расположенную в таблице строк. Это позволяет базам данных, содержащим много общих строк, иметь небольшой размер.

Microsoft частично задокументировала формат файла SDB и API, используемые для его чтения и записи, по адресу https://docs.microsoft.com/en-us/windows/win32/devnotes/application-compatibility-database. Все API SDB реализованы в клиентской библиотеке совместимости приложений (apphelp.dll).

Прослойки драйверов

Диспетчер памяти NT решает, применять ли прослойку к драйверу уровня ядра во время его загрузки, используя функцию `KseDriverLoadImage` (загружаемые во время загрузки системы драйверы обрабатываются диспетчером ввода-вывода, как описано в главе 12). Процедура вызывается в нужный момент жизненного цикла модуля ядра, прежде чем к нему будут применены средства проверки драйверов, оптимизации импорта или защиты от исправлений ядра. (Это важно, иначе система аварийно завершится.) Список текущих модулей ядра с прослойками хранится в глобальной переменной. Процедура `KsepGetShimsForDriver` проверяет, присутствует ли в данный момент в списке модуль с тем же базовым адресом, что и у загружаемого. Если да, значит, целевой модуль уже получил прослойку, поэтому процедура прерывается. В противном случае, чтобы определить, следует ли обработать новый модуль, программа проверяет два разных источника.

- Запрашивает многострочный параметр Shims из раздела реестра, названного как загружаемый модуль и расположенного в корневом разделе HKLM\System\CurrentControlSet\Control\Compatibility\Driver. Параметр реестра содержит массив имен прослоек, которые будут применены к целевому модулю.
- Если параметра реестра для целевого модуля не существует, анализируется файл базы данных совместимости драйверов в поисках тега KDRIVER (индексируется блоком INDEX), который имеет то же имя, что и загружаемый модуль. Если драйвер найден в файле SDB, ядро NT сравнивает версии драйвера (TAG_SOURCE_OS, хранящегося в корневом теге KDRIVER), имени файла, пути (если соответствующие теги существуют в SDB) и низкоуровневой системной информации, собираемой во время инициализации ядра (чтобы определить, совместим ли драйвер с системой). В случае несоответствия какой-либо информации драйвер пропускается и прослойки не применяются. В противном случае список имен прослоек берется из тегов нижнего уровня KSHIM_REF, которые являются частью корневого KDRIVER. Теги относятся к KSHIM, расположенным в блоке базы данных SDB.

Если один из двух источников дает одно или несколько имен прослоек, которые нужно применить к целевому драйверу, файл SDB снова анализируется с целью проверки существования допустимого дескриптора KSHIM. Если нет тегов, связанных с указанным именем прослойки (это означает, что в базе данных

не существует дескриптора прослойки), процедура прерывается, что не позволяет администратору применять к драйверу случайные прослойки сторонних производителей. В противном случае массив структуры данных `KSE_SHIM_INFO` возвращается в `KsepGetShimsForDriver`.

На следующем шаге нужно определить, зарегистрированы ли в системе прослойки, описанные их дескрипторами. Для этого механизм Shim выполняет поиск в глобальном связанном списке зарегистрированных прослоек (заполняется каждый раз при регистрации новой прослойки, как объяснялось ранее в разделе «Инициализация механизма прослоек»). Если прослойка не зарегистрирована, механизм прослойки пытается загрузить драйвер, который ее предоставляет (его имя хранится в дочернем теге MODULE корневой записи KSHIM), и повторяет попытку. Когда прослойка применяется впервые, механизм Shim разрешает указатели всех перехватчиков, описанных массивом структур данных `KSE_HOOK_COLLECTION`, принадлежащим зарегистрированной прослойке (структура данных `KSE_SHIM`). Механизм прослойки выделяет и заполняет структуру данных `KSE_SHIMMED_MODULE`, представляющую целевой модуль, который нужно подставить (он включает базовый адрес), и добавляет его в глобальный список, проверенный вначале.

На этом этапе механизм прослойки применяет прослойку к целевому модулю с помощью внутренней процедуры `KsepApplyShimsToDriver`. Последний циклически переключается между каждым перехватчиком, описанным массивом `KSE_HOOK_COLLECTION`, и исправляет таблицу адресов импорта (IAT) целевого модуля, заменяя исходный адрес перехваченных функций новыми, описываемыми коллекцией перехватчиков. Обратите внимание на то, что функции обратного вызова объекта драйвера (обработчики IRP) на этом этапе не обрабатываются. Они изменяются позже диспетчером ввода-вывода перед вызовом процедуры `DriverInit` целевого драйвера. Подпрограммы обратного вызова IRP исходного драйвера сохраняются в расширении целевого драйвера. Таким образом, у подключенных функций есть простой способ при необходимости вернуться к исходным функциям.

ЭКСПЕРИМЕНТ. Наблюдение за прослойками на уровне ядра

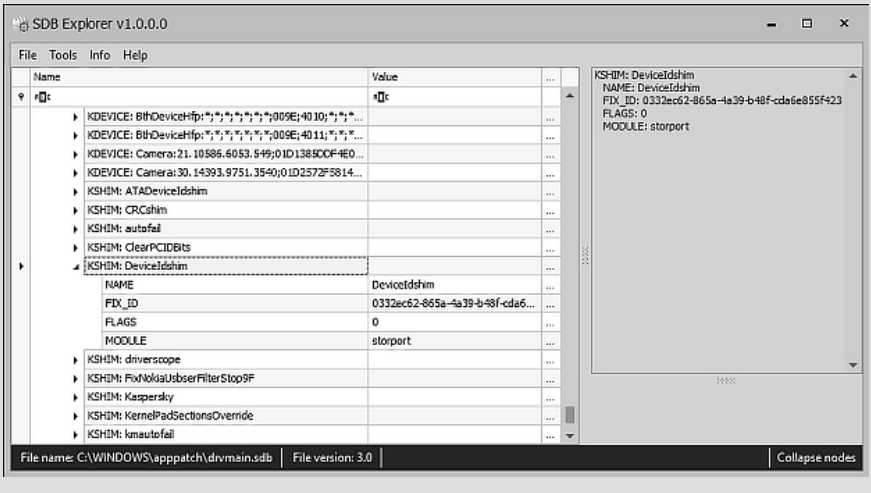
Хотя официальный набор средств обеспечения совместимости приложений Microsoft, поставляемый вместе с пакетом Windows Assessment and Deployment Kit, позволяет открывать, изменять и создавать файлы базы данных прослойки, он не работает с файлами системной базы данных, идентифицируемыми по их внутренним идентификаторам GUID, поэтому не сможет проанализировать все прослойки ядра, описанные в базе данных drvmain.sdb. Существует несколько сторонних анализаторов SDB. В частности, один из них, SDB explorer, можно бесплатно загрузить с сайта https://ericzimmerman.github.io/.

В этом эксперименте вы получите доступ к файлу системной базы данных drvmain и примените прослойку ядра к тестовому драйверу ShimDriver, который имеется в загружаемых ресурсах книги. Для этого эксперимента вам необходимо включить тестовую подпись (ShimDriver подписывается тестовым, самостоятельно заверенным сертификатом).

1. Откройте командную строку администратора и введите следующую команду:

   ```
   bcdedit /set testsigning on
   ```

2. Перезагрузите компьютер, загрузите SDB Explorer с веб-сайта, запустите его и откройте базу данных drvmain.sdb, расположенную в %SystemRoot%\apppatch.
3. В главном окне SDB Explorer вы можете просмотреть весь файл базы данных, организованный в виде трех основных блоков: индексы, базы данных и строковая таблица. Разверните корневой блок DATABASES и прокрутите вниз, пока не увидите список KSHIM (должен быть расположен после KDEVICE). Вы увидите окно, подобное следующему.



4. Примените одну из прослоек версии к нашему тест-драйверу. Сначала следует скопировать ShimDriver в папку %SystemRoot%\System32\Drivers. Затем нужно установить его, введя в командной строке администратора (предполагается, что ваша система 64-битная) следующую команду:

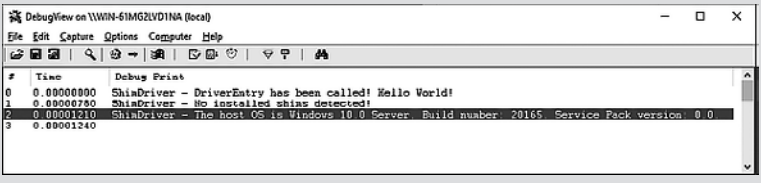
   ```
   sc create ShimDriver type= kernel start= demand error= normal binPath= c:\
   Windows\System32\ShimDriver64.sys
   ```

5. Перед запуском тестового драйвера необходимо загрузить и запустить утилиту DebugView, доступную на веб-сайте Sysinternals (https://docs.microsoft.com/en-us/sysinternals/downloads/debugview). Это необходимо, поскольку ShimDriver выдает отладочные сообщения.
6. Запустите ShimDriver с помощью команды

   ```
   sc start shimdriver
   ```

7. Проверьте выходные данные утилиты DebugView. Вы должны увидеть сообщения, подобные показанному на следующем рисунке. То, что вы увидите,

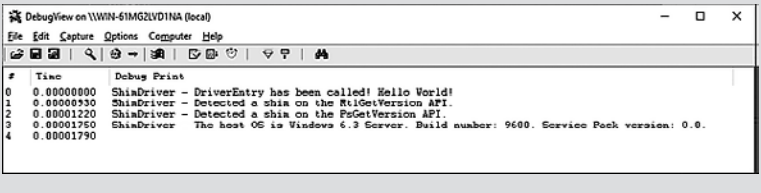
зависит от версии Windows, в которой запускаете драйвер. В этом примере мы запускаем драйвер на инсайдерской версии Windows Server 2022.



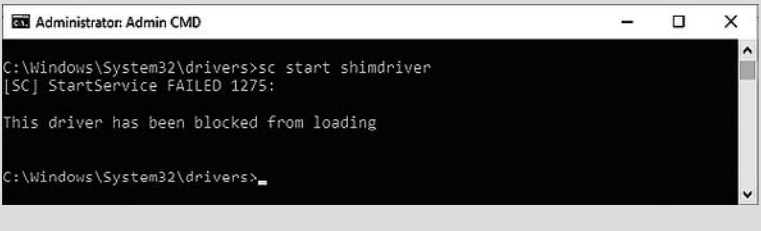
8. Теперь нужно остановить драйвер и включить одну из прослоек, имеющихся в базе данных SDB. В этом примере вы начнете с одной из прослоек подмены версии. Остановите целевой драйвер и установите прослойку, используя следующие команды (здесь `ShimDriver64.sys` — имя файла драйвера, установленного на предыдущем шаге):

```
sc stop shimdriver
reg add "HKLM\System\CurrentControlSet\Control\Compatibility\Driver\
    ShimDriver64.sys" /v Shims /t REG_MULTI_SZ /d
KmWin81VersionLie /f /reg:64
```

9. Последняя команда добавляет прослойку для маскировки под Windows 8.1, но можете свободно выбирать и другие версии.
10. Теперь, если вы перезапустите драйвер, то увидите сообщения, выводимые утилитой DebugView, как показано на следующем снимке экрана.



11. Это связано с тем, что механизм прослойки правильно применил перехваты к API-интерфейсам NT, используемым для получения информации о версии ОС (драйвер также может обнаружить прослойку). У вас будет возможность повторить эксперимент, применив другие прослойки, такие как `SkipDriverUnload` или `KernelPadSectionsOverride`, которые нейтрализуют процедуру выгрузки драйвера или предотвратят загрузку целевого драйвера, как показано на следующем снимке экрана.



Прослойки устройств

В отличие от прослоек драйвера, прослойки, применяемые к объектам устройств, загружаются и используются по требованию. Ядро NT экспортирует функцию `KseQueryDeviceData`, которая позволяет драйверам проверять, нужно ли применять прослойку к объекту устройства. (Обратите внимание на то, что заодно экспортируется функция `KseQueryDeviceFlags`. Однако она лишь аналог для более частных случаев.) Запрос прослоек устройства возможен и для приложений пользовательского режима через API `NtQuerySystemInformation`, применяемый с информационным классом `SystemDeviceDataInformation`. Прослойки устройства всегда хранятся в трех местах, которые рассматриваются в следующем порядке.

1. В корневом разделе реестра HKLM\System\CurrentControlSet\Control\Compatibility\Device с добавлением раздела, называемого идентификатором оборудования PNP устройства и заменой символа \ на !, чтобы не сбить с толку систему реестра. Параметры в разделе устройства определяют обработанные прослойкой запрашиваемые данные устройства (обычно это флаги для определенного класса устройства).
2. В кэше прослоек ядра. Механизм Kernel Shim реализует кэш-прослойку, доступный посредством структуры данных `KSE_CACHE`, с целью ускорения поиска флагов устройств и данных.
3. В файле базы данных прослоек с использованием корневого тега `KDEVICE`. Корневой тег, в отличие от многих других параметров (например, описания устройства, имени производителя, GUID и т. д.), включает в себя дочерний тег `NAME`, содержащий строку, составленную следующим образом: `<DataName:HardwareID>`. Дочерние теги `KFLAG` или `KDATA` включают параметры данных устройства, затрагиваемых прослойкой.

Если прослойка устройства отсутствует в кэше, а находится только в файле SDB, она всегда туда добавляется. Таким образом, следующие запросы будут проходить быстрее и не потребуют доступа к файлу базы данных прослоек.

ЗАКЛЮЧЕНИЕ

В этой главе мы описали наиболее важные функции операционной системы Windows, которые предоставляют средства управления, такие как реестр Windows, службы пользовательского режима, планирование задач, UBPM и инструментарий управления Windows (WMI). Кроме того, мы обсудили, как трассировка событий для Windows (ETW), DTrace, отчеты об ошибках Windows (WER) и глобальные флаги (GFlags) предоставляют службы, которые позволяют пользователям лучше отслеживать и диагностировать проблемы, возникающие из-за любых компонентов ОС или приложений пользовательского режима. Глава завершилась обзором механизма Kernel Shim, который помогает системе использовать стратегии совместимости, что позволяет корректно работать устаревшим компонентам, разработанным для более старых версий операционной системы.

В следующей главе рассматриваются различные файловые системы, доступные в Windows, совместно с глобальным кэшированием, предназначенным для ускорения доступа к файлам и данным.

ГЛАВА 11

Кэширование и файловые системы

Диспетчер кэша — это набор функций режима ядра и системных потоков, взаимодействующих с диспетчером памяти для кэширования данных всех драйверов, как локальных, так и сетевых, в файловой системе Windows. В этой главе рассматривается устройство диспетчера кэша, изучаются его основные внутренние структуры данных и функции, процедура определения его размера при инициализации системы, способы взаимодействия с другими компонентами операционной системы и объясняется, как можно наблюдать за его работой с помощью трекеров производительности. Рассматриваются также пять флагов Windows-функции `CreateFile`, влияющих на кэширование файлов, и тома DAX — диски прямого доступа, обходящие диспетчер кэша при определенных операциях ввода-вывода.

Службы диспетчера кэша используются всеми драйверами файловой системы Windows. Драйверы тесно взаимодействуют с диспетчером для достижения максимальной производительности операций ввода-вывода диска. Рассматриваются различные файловые системы, поддерживаемые Windows. В частности, глубоко анализируются NTFS и ReFS — две наиболее распространенные файловые системы. Изучаются их внутренняя архитектура и основные операции, включая взаимодействие с другими компонентами системы, такими как диспетчер памяти и диспетчер кэша.

В конце главы приводится обзор концепции дисковых пространств Storage Spaces — нового решения для хранения данных, призванного заменить динамические диски. Данная технология позволяет создавать многоуровневые и тонкие виртуальные диски, предоставляя возможности, которые могут быть использованы расположенной выше файловой системой.

ТЕРМИНОЛОГИЯ

Чтобы понять материал этой главы, необходимо знать базовую терминологию.

- *Диски* — физические устройства хранения данных: жесткие диски, оптические диски CD-ROM, DVD-ROM и Blu-ray, твердотельные диски SSD, диски с энергонезависимой памятью NVMe и флеш-накопители.
- *Секторы* — аппаратно-адресуемые блоки на носителе информации. Размеры секторов определяются аппаратным обеспечением. Секторы большинства жестких дисков имеют размер 4096 или 512 байт, а секторы дисков DVD-ROM

и Blu-ray — обычно 2048 байт. Таким образом, если размер сектора составляет 4096 байт, а операционная система хочет изменить 5120-й байт на диске, то она должна перезаписать 4096-байтный блок данных во втором секторе диска.

- *Разделы* — группы смежных секторов на диске. Таблица разделов или иная база данных управления дисками хранит информацию о начальном секторе раздела, его размере и прочих характеристиках и располагается на том же диске, что и раздел.
- *Тома* — объекты, представляющие секторы, которыми драйверы файловой системы всегда управляют как единым целым. Простые тома представляют секторы из одного раздела, а многораздельные тома — секторы из нескольких разделов. Многораздельные тома обеспечивают производительность, надежность и управление размерами разделов на уровне, которого нет у простых томов.
- *Форматы файловых систем* определяют способ хранения файловых данных на носителях и возможности файловой системы. Например, формат, не поддерживающий пользовательские разрешения доступа к файлам и каталогам, небезопасен. Применяемый формат файловой системы может определять также ограничения размеров файлов и устройств хранения, которые она поддерживает. Наконец, некоторые форматы файловых систем эффективно реализуют поддержку крупных или мелких файлов и дисков. NTFS, exFAT и ReFS — примеры форматов файловых систем, реализующих различные наборы функций и допускающих разные сценарии применения.
- *Кластеры* — адресуемые блоки, используемые во многих форматах файловых систем. Размер кластера всегда кратен размеру сектора, как показано на рис. 11.1, где каждый кластер состоит из восьми секторов. Форматы файловых систем применяют кластеры для более эффективного управления дисковым пространством. Размер кластера, превышающий размер сектора, делит диск на более гибкие в управлении блоки. Потенциальный недостаток увеличения размера кластера заключается в потере дискового пространства либо внутренней фрагментации, возникающей, когда размер файла не кратен размеру кластера.

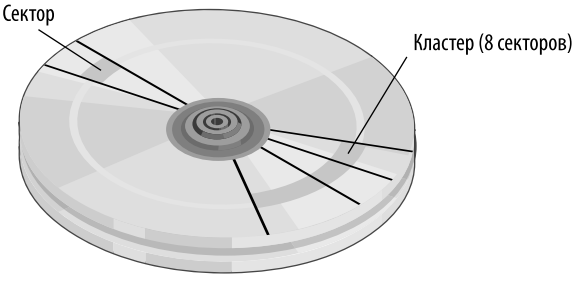


Рис. 11.1. Секторы и кластеры на классическом жестком диске

- *Метаданные* — информация, хранящаяся в томе для управления форматом файловой системы. Обычно они недоступны приложениям. К метаданным относятся, например, адреса файлов и каталогов в томе.

КЛЮЧЕВЫЕ ФУНКЦИИ ДИСПЕТЧЕРА КЭША

Диспетчер кэша имеет несколько ключевых особенностей.

- Поддерживает все типы файловых систем, как локальных, так и сетевых, устраняя необходимость в использовании разного кода для управления кэшем в разных системах.
- Задействует диспетчер памяти для управления тем, какие части каких файлов размещены в физической памяти, распределяя запросы к физической памяти между пользовательскими процессами и операционной системой.
- Кэширует данные на основе виртуальных блоков (смещения в файле), в отличие от многих систем, кэширующих данные на основе логических блоков (смещения в дисковом томе), обеспечивая заблаговременное и высокоскоростное чтение в кэш без привлечения драйверов файловой системы. Этот метод кэширования, называемый *быстрым вводом-выводом*, описан далее в этой главе.
- Поддерживает дополнительные сообщения от приложений, передаваемые ими при открытии файла, — например, произвольный или последовательный доступ, создание временного файла и т. д.
- Поддерживает восстанавливаемые файловые системы — например, с протоколированием транзакций — для восстановления данных после системного сбоя.
- Поддерживает твердотельные диски, NVMe и диски прямого доступа (direct access, DAX).

Далее в главе будет подробно рассмотрена работа перечисленных функций диспетчера кэша, а пока познакомимся с концепциями, лежащими в их основе.

Единый централизованный системный кэш

Некоторые операционные системы полагаются на способность каждой отдельной файловой системы кэшировать данные, что приводит либо к избыточному дублированию кода кэширования памяти и управления ею, либо к ограничениям типов данных, которые могут быть кэшированы. В отличие от этого подхода, в Windows реализована централизованная система кэширования, охватывающая все данные, хранящиеся извне: на локальных жестких дисках, съемных USB-накопителях, сетевых файловых серверах или приводах оптических дисков. Кэшироваться могут любые данные, как пользовательские потоки — содержимое файла и текущие операции чтения из этого файла или записи в него, так и *метаданные* файловой *системы*, например заголовки каталогов и файлов. Как будет видно в этой главе, способ, используемый Windows для доступа к кэшу, зависит от типа кэшируемых данных.

Диспетчер памяти

Один из необычных аспектов работы диспетчера кэша заключается в том, что он никогда не знает, сколько кэшированных данных реально находится в физической памяти. Это утверждение может показаться странным, поскольку цель кэша — хранить подмножество часто используемых данных в физической памяти для повышения производительности системы ввода-вывода. Причина, по которой диспетчеру кэша неизвестно, сколько данных находится в физической памяти, заключается

в том, что он получает доступ к данным путем сопоставления представлений файлов с виртуальными адресными пространствами системы, используя стандартные *объекты-секции*, или *объекты отображения файлов* в терминологии Windows API. Секции — базовые примитивы диспетчера памяти, они подробно рассматриваются в главе 5 книги. При доступе к адресам в этих отображенных представлениях диспетчер памяти подгружает блоки, которых нет в физической памяти. При нехватке памяти он удаляет страницы из кэша и, если данные изменились, помещает данные обратно в файлы.

Благодаря кэшированию на основе виртуальных адресных пространств и объектов сопоставления файлов отпадает необходимость в кэшировании диспетчером пакетов запросов ввода-вывода (I/O request packet, IRP) для доступа к данным в кэшируемых файлах. Вместо этого он просто копирует данные по виртуальным адресам, соответствующим отображенной части кэшируемого файла, и полагается на диспетчер памяти для ввода данных в физическую память или вывода из нее. Этот процесс позволяет диспетчеру памяти распределять оперативную память между системным кэшем и пользовательскими процессами. Диспетчер кэша также инициирует ввод-вывод, например позднюю запись, о которой речь пойдет далее в этой главе, но для записи страниц обращается к диспетчеру памяти. Кроме того, в следующем разделе будет рассмотрено, как такая конструкция позволяет процессам, взаимодействующим с кэшированными файлами, видеть те же данные, которые видят и другие процессы, отображающие те же файлы в свои пользовательские адресные пространства.

Когерентность кэша

Одна из важнейших функций диспетчера кэша — гарантировать, что процесс, обращающийся к кэшированным данным, получает их актуальную версию. Проблема может возникнуть, когда один процесс открывает файл и, следовательно, тот кэшируется, а другой процесс непосредственно отображает файл в свое адресное пространство с помощью функции Windows `MapViewOfFile`. Этой вероятной проблемы не возникает в Windows, так как и диспетчер кэша, и пользовательские приложения, отображающие файлы в свои адресные пространства, задействуют одни и те же службы отображения файлов и управления памятью. Поскольку диспетчер памяти гарантирует, что у него есть только одно представление каждого уникального отображаемого файла независимо от количества объектов-секций или отображаемых представлений, он отображает все представления файла, даже если они конфликтуют, в один набор страниц в физической памяти (рис. 11.2). Дополнительные сведения о том, как диспетчер памяти работает с отображаемыми файлами, см. в главе 5.

К примеру, если у процесса 1 есть представление файла, отображенное в его пользовательском адресном пространстве, а процесс 2 обращается к тому же представлению через системный кэш, то он видит любые изменения, которые процесс 1 вносит во время создания, а не сброса на диск. Диспетчер памяти не сбрасывает на диск *все* страницы, отображаемые пользователем, а только те, в которые была сделана запись согласно «грязному» биту. Поэтому в Windows все процессы, обращающиеся к файлу, видят его последнюю версию, даже если он открыт определенными процессами через систему ввода-вывода, а у других файл отображен в их адресное пространство с помощью функций отображения файлов Windows.

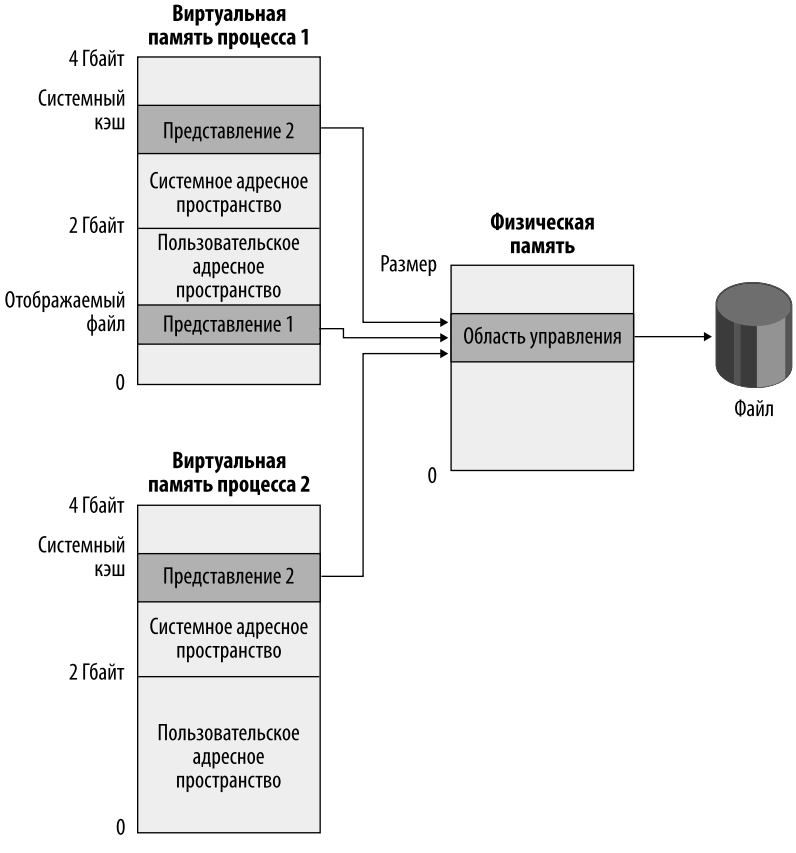


Рис. 11.2. Схема когерентного кэширования

ПРИМЕЧАНИЕ Когерентность кэша в данном случае означает когерентность между пользовательскими данными и кэшированным вводом-выводом, а не между некэшированным и кэшированным аппаратным доступом и вводом-выводом, которые почти гарантированно будут некогерентными. Кроме того, когерентность кэша несколько сложнее для сетевых перенаправителей, чем для локальных файловых систем, поскольку сетевые перенаправители должны реализовать дополнительные операции очистки для обеспечения когерентности кэша при доступе к сетевым данным.

Виртуальное блочное кэширование

Диспетчер кэша Windows использует метод, известный как *виртуальное блочное кэширование*, при котором диспетчер кэша отслеживает, какие части каких *файлов* находятся в кэше. Диспетчер кэша может отслеживать эти части файлов путем отображения 256 Кбайт представлений файлов в системные виртуальные адресные пространства с помощью специальных системных подпрограмм кэша, расположенных в диспетчере памяти. Этот подход имеет следующие ключевые преимущества.

- Открывает возможность упреждающего чтения. Поскольку кэш отслеживает, какие части каких файлов находятся в кэше, он может предсказать, что именно будет затребовано в следующий раз.
- Позволяет системе ввода-вывода не обращаться к файловой системе при запросах данных, уже находящихся в кэше, то есть реализован быстрый ввод-вывод. Поскольку диспетчер кэша знает, какие части каких файлов находятся в кэше, он может вернуть адрес кэшированных данных, чтобы удовлетворить запрос ввода-вывода без обращения к файловой системе.

Подробнее о том, как работают опережающее чтение и быстрый ввод-вывод, рассказывается далее в этой главе в разделах «Чтение с упреждением и отложенная запись» и «Быстрый ввод-вывод» соответственно.

Потоковое кэширование

Диспетчер кэша предназначен также для *кэширования потоков*, а не файлов. *Поток* — это последовательность байтов в файле. В некоторых файловых системах, таких как NTFS, файл может содержать более одного потока. Диспетчер кэша позволяет таким файловым системам кэшировать каждый поток независимо. NTFS может использовать эту возможность, организуя свою главную таблицу файлов, описанную далее в этой главе в разделе «Главная файловая таблица», в потоки и кэшируя и эти потоки. Хотя можно сказать, что диспетчер кэша кэширует файлы, но на самом деле он кэширует потоки — все файлы имеют по крайней мере один поток данных, — идентифицирующиеся как именем файла, так и именем потока, если в файле существует более одного потока.

ПРИМЕЧАНИЕ Диспетчер кэша реализован так, что он не знает имен файлов или потоков, но использует указатели на эти структуры.

Поддержка восстанавливаемых файловых систем

Восстанавливаемые файловые системы, такие как NTFS, предназначены для восстановления структуры тома диска после сбоя системы. Эта возможность означает, что операции ввода-вывода, выполнявшиеся в момент сбоя системы, должны быть либо полностью завершены, либо полностью удалены с диска при перезапуске системы. Наполовину завершенные операции ввода-вывода могут повредить дисковый том и даже сделать его недоступным.

Чтобы избежать этой проблемы, восстанавливаемая файловая система ведет журнал, в котором записывает каждое планируемое обновление своей структуры — метаданных файловой системы — перед записью изменений в том. Если система выходит из строя, прерывая выполняемые изменения тома, то восстанавливаемая файловая система использует информацию, хранящуюся в журнале, для повторного выполнения обновлений тома.

Чтобы гарантировать успешное восстановление тома, каждая запись журнала, документирующая его обновление, должна быть целиком записана на диск до применения самого обновления к тому. Поскольку записи на диск кэшируются,

диспетчер кэша и файловая система должны координировать обновления метаданных, обеспечивая очистку файла журнала перед обновлениями метаданных. В целом перечисленные действия выполняются последовательно.

1. Файловая система заносит в журнал запись, документирующую обновление метаданных, которое она собирается произвести.
2. Файловая система вызывает диспетчер кэша, чтобы сохранить запись, сделанную в журнале, из кэша на диск.
3. Файловая система записывает обновление тома в кэш, то есть изменяет свои кэшированные метаданные.
4. Диспетчер кэша сбрасывает измененные метаданные из кэша на диск, обновляя структуру тома. На самом деле записи журнала, как и изменения томов, компонуются в группы перед отправкой на диск.

ПРИМЕЧАНИЕ Термин «метаданные» применяется только к изменениям в структуре файловой системы: созданию, переименованию и удалению файлов и каталогов.

Когда файловая система записывает данные в кэш, она может предоставить *логический номер последовательности* (logical sequence number, LSN), идентифицирующий запись в ее журнале, соответствующую обновлению кэша. Диспетчер кэша отслеживает эти номера, записывая связанные с каждой страницей в кэше самый низкий и самый высокий LSN, представляющие самые старые и самые новые записи журнала. Кроме того, потоки данных, защищенные записями журнала транзакций, помечаются NTFS как запрещенные для записи, чтобы программа записи отображенных страниц не могла нечаянно записать эти страницы до того, как будут сделаны соответствующие записи журнала. Когда программа записи отображенных страниц видит страницу, помеченную таким образом, она перемещает ее в специальный список, который диспетчер кэша затем очищает в нужное время, например, когда происходит отложенная запись.

Когда диспетчер кэша готовится выгрузить на диск группу «грязных» страниц, он определяет наивысший LSN, связанный со страницами, которые нужно выгрузить, и сообщает об этом файловой системе. После этого файловая система может вызвать диспетчер кэша и дать ему указание сбросить данные журнала до точки, представленной сообщенным LSN. После того как диспетчер кэша сбросит журнал на диск до этого LSN, он сбрасывает соответствующие обновления структуры тома на диск, обеспечивая тем самым протоколирование того, что он собирается сделать, прежде чем выполнить это на самом деле. Такое взаимодействие между файловой системой и диспетчером кэша гарантирует возможность восстановления дискового тома после системного сбоя.

Усовершенствования рабочего набора NTFS MFT

Как описывалось в предыдущих абзацах, механизм, используемый диспетчером кэша для кэширования файлов, аналогичен общим интерфейсам ввода-вывода отображенной памяти, предоставляемым диспетчером памяти операционной системе. Для доступа к файлу или его кэширования диспетчер кэша отображает представление файла в виртуальном адресном пространстве системы. Затем доступ к содержимому

осуществляется простым считыванием из сопоставленного диапазона виртуальных адресов. Когда кэшированное содержимое файла больше не нужно — по разным причинам, подробности изложены в следующих абзацах, — диспетчер кэша прекращает отображение файла. Эта стратегия хорошо работает для любых типов файлов данных, но возникают некоторые проблемы с метаданными, поддерживаемыми файловой системой для правильного хранения файлов в томе.

Когда файловый дескриптор закрывается или контролирующий его процесс умирает, диспетчер кэша убеждается, что кэшированные данные больше не находятся в рабочем наборе. Файловая система NTFS обращается к главной файловой таблице (Master File Table, MFT) как к большому файлу, кэшируемому диспетчером кэша так же, как и любой другой пользовательский файл. Проблема с MFT заключается в следующем: поскольку это системный файл, отображаемый и обрабатываемый в контексте процесса System, никто никогда не закроет его дескриптор, если только том не будет размонтирован, поэтому система никогда не прекратит отображения кэшированного представления MFT. Процесс, первоначально вызвавший отображение определенного представления MFT, мог закрыть дескриптор или закончить работу, оставив потенциально нежелательные представления MFT, все еще отображенные в памяти и потребляющие дорогостоящий системный кэш. Эти представления прекратят отображаться, только если система столкнется с нехваткой памяти.

В Windows 8.1 эта проблема решена за счет хранения счетчика ссылок на каждую запись MFT в динамически выделяемом многоуровневом массиве, хранящемся в структуре Volume Control Block (VCB) файловой системы NTFS. Каждый раз, когда создается структура данных File Control Block (FCB) (подробнее FCB и VCB описаны далее в этой главе), файловая система увеличивает счетчик соответствующей индексной записи MFT. Аналогично, когда FCB уничтожается, все дескрипторы файла или каталога, на которые ссылается запись MFT, закрываются, NTFS убирает ссылку на относительный счетчик и вызывает процедуру диспетчера кэша `CcUnmapFileOffsetFromSystemCache`, прекращающую отображение более ненужной части MFT.

Поддержка разделов памяти

В Windows 10 с целью обеспечения поддержки контейнеров Hyper-V и игрового режима появилось понятие разделов. Разделы памяти описывались в главе 5. Как там было показано, разделы памяти представлены большой структурой данных `MI_PARTITION`, поддерживающей структуры управления памятью, связанные с разделом, такие как списки страниц (резервных, измененных, нулевых, свободных и т. д.), резервирование физической памяти под виртуальную, рабочий набор, обрезка страниц, запись измененных страниц и поток нулевых страниц. Для поддержки разделов диспетчер кэша должен работать совместно с диспетчером памяти. Во время первой фазы инициализации ядра NT система создает и инициализирует раздел диспетчера кэша, который будет частью раздела `MemoryPartition0` System Executive. Подробнее об инициализации ядра Windows см. главу 12. Код диспетчера кэша подвергся объемному рефакторингу для поддержки разделов, все глобальные структуры данных и переменные диспетчера кэша были перенесены в структуру данных раздела диспетчера кэша `CC_PARTITION`.

Раздел диспетчера кэша содержит данные, связанные с кэшем, такие как список глобальных карт общего кэша, список рабочих потоков (упреждающего чтения, отложенной записи, дополнительной отложенной записи, отложенного чтения, сканирования отложенного чтения, асинхронного чтения), событий сканирования отложенного чтения, массив, хранящий всю историю отложенной записи, верхний и нижний пределы порога «грязных» страниц, количество «грязных» страниц и т. д. Когда инициализируется системный раздел диспетчера кэша, все необходимые системные потоки запускаются в контексте процесса System, принадлежащего этому разделу. Каждый раздел всегда имеет ассоциированный минимальный процесс System, создаваемый во время создания раздела с помощью API `NtCreatePartition`.

Когда система формирует новый раздел с помощью API `NtCreatePartition`, она всегда создает и инициализирует пустой объект `MI_PARTITION`. Память перемещается из родительского раздела в дочерний или добавляется на лету позже с помощью функции `NtManagePartition`. Объект раздела диспетчера кэша создается только по требованию. Если в контексте нового раздела не создается никаких файлов, то создавать объект раздела диспетчера кэша не нужно. Когда файловая система создает или открывает файл для доступа к кэшу, функция `CcInitializeCacheMap(Ex)` проверяет, к какому разделу принадлежит файл и есть ли у этого раздела корректная ссылка на раздел диспетчера кэша. Если раздел диспетчера кэша отсутствует, то система создает и инициализирует новый раздел с помощью процедуры `CcCreatePartition`. В новом разделе запускаются связанные с диспетчером кэша отдельные потоки — упреждающее чтение, отложенные записи и т. д. — и вычисляются новые значения порога «грязных» страниц, основанные на количестве страниц, принадлежащих конкретному разделу.

Файловый объект содержит ссылку на раздел, к которому он принадлежит, задействуя свою область управления, первоначально выделяемую драйвером файловой системы при создании и отображении блока управления потоком (Stream Control Block, SCB). Раздел целевого файла хранится в расширении файлового объекта типа `MemoryPartitionInformation` и проверяется диспетчером памяти при создании объекта раздела для SCB. В общем случае файлы применяются совместно, поэтому драйверы файловой системы не могут автоматически привязать файл к разделу, отличному от системного. Приложение может установить другой раздел для файла с помощью API `NtSetInformationFileKernel`, используя новый класс `FileMemoryPartitionInformation`.

УПРАВЛЕНИЕ ВИРТУАЛЬНОЙ ПАМЯТЬЮ КЭША

Поскольку диспетчер кэша системы Windows кэширует данные на виртуальной основе, он задействует области виртуального адресного пространства системы вместо физической памяти и управляет ими в структурах, называемых *блоками управления виртуальными адресами* (virtual address control block, VACB). VACB определяют эти области адресного пространства в 256-килобайтные слоты, называемые *представлениями*. Когда диспетчер кэша инициализируется в процессе загрузки, он выделяет начальный массив VACB для описания кэшированной памяти. С ростом требований к кэшированию и увеличением необходимого объема памяти диспетчер кэша выделяет дополнительные массивы VACB по мере надобности. Он также может сокращать виртуальное адресное пространство по мере того, как другие запросы оказывают давление на систему.

При первой операции ввода-вывода — чтения или записи, — относящейся к файлу, диспетчер кэша отображает в свободный слот адресного пространства системного кэша представление области файла с выравниванием 256 Кбайт, содержащей запрашиваемые данные. Например, если в файл было считано 10 байт, начинающихся со смещения 300 000 байт, то отображаемое представление будет начинаться со смещения 262 144 — вторая область файла с выравниванием 256 Кбайт — и простираться на 256 Кбайт.

Диспетчер кэша отображает представления файлов в слоты адресного пространства кэша по кругу, отображая первое запрошенное представление в первый слот размером 256 Кбайт, второе представление — во второй слот размером 256 Кбайт и т. д. (рис. 11.3). В этом примере файл Б был отображен первым, файл А — вторым, а файл В — третьим, поэтому отображенный кусок файла Б занимает первый слот в кэше. Обратите внимание: была отображена только первая часть файла Б размером 256 Кбайт, это связано с тем, что доступ был получен только к части файла. Поскольку размер файла В всего 100 Кбайт и, следовательно, меньше, чем одно из представлений в системном кэше, ему требуется собственный слот в кэше размером 256 Кбайт.

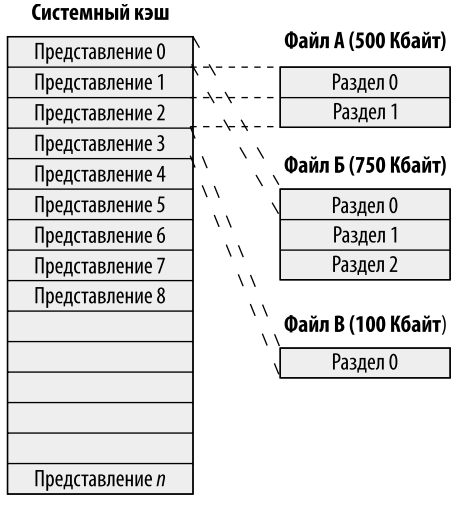


Рис. 11.3. Отображение файлов разного размера в системном кэше

Диспетчер кэша гарантирует, что представление отображается до тех пор, пока оно активно, хотя представления могут оставаться отображенными и после того, как становятся неактивными. Однако представление помечается активным только во время операции чтения из файла или записи в него. Если процесс не открывает файл, указав параметр `FILE_FLAG_RANDOM_ACCESS` в вызове `CreateFile`, диспетчер кэша прекращает отображение неактивных представлений файла по мере создания новых представлений для файла, если обнаруживает, что к файлу обращаются последовательно. Страницы для снятых с отображения представлений отправляются в списки резервных или измененных в зависимости от того, были ли они изменены, а поскольку диспетчер памяти предоставляет специальный интерфейс диспетчеру кэша, тот может направлять страницы для размещения в конце или начале этих списков. Страницы, соответствующие представлениям файлов, открытых с флагом `FILE_FLAG_SEQUENTIAL_SCAN`, перемещаются в начало списков, а все остальные — в конец. Эта схема поощряет повторное использование страниц, принадлежащих последовательно прочитанным файлам, и позволяет затронуть лишь небольшую часть физической памяти операцией копирования большого файла. Флаг тоже влияет на снятие отображения. Диспетчер кэша будет агрессивно снимать отображения, если он установлен.

Если диспетчеру кэша нужно отобразить представление файла, а в кэше больше нет свободных слотов, то он снимет наиболее давнее отображение неактивного

представления и использует этот слот. Если доступных представлений нет, то возвращается ошибка ввода-вывода, показывающая недостаток системных ресурсов для выполнения операции. Однако, учитывая, что представления помечаются как активные только во время операции чтения или записи, такой сценарий крайне маловероятен, поскольку для возникновения подобной ситуации потребовался бы одновременный доступ к тысячам файлов.

РАЗМЕР КЭША

В следующих разделах объясним, как Windows вычисляет размер системного кэша виртуально и физически. Как и большинство вычислений, связанных с управлением памятью, размер системного кэша зависит от ряда факторов.

Виртуальный размер кэша

В 32-разрядной системе Windows виртуальный размер системного кэша ограничен исключительно объемом виртуального адресного пространства режима ядра и разделом реестра `SystemCacheLimit`, который может быть настроен опционально. Дополнительные сведения об ограничении размера виртуального адресного пространства ядра см. в главе 5. Это означает, что размер кэша ограничен системным адресным пространством в 2 Гбайт, но обычно он значительно меньше, поскольку системное адресное пространство используется совместно с другими ресурсами, включая системные записи выгружаемых таблиц (paged table entry, PTE), невыгружаемый и выгружаемый пул, а также таблицы страниц. Максимальный размер виртуального кэша составляет 64 Тбайт в 64-битной Windows, и даже в этом случае ограничение привязано к размеру системного адресного пространства: в будущих системах, поддерживающих 56-битный режим адресации, ограничение составит 32 Пбайт.

Размер рабочего набора кэша

Как уже говорилось, одним из ключевых отличий в конструкции диспетчера кэша в Windows от других операционных систем является делегирование управления физической памятью глобальному диспетчеру памяти. В связи с этим существующий код, обрабатывающий расширение и обрезание рабочего набора и управление измененными и резервными списками, применяется также для управления размером системного кэша, динамически распределяя требования к физической памяти между процессами и операционной системой.

Системный кэш не имеет собственного рабочего набора, а использует единый системный набор, включающий данные кэша, выгружаемые пул, код ядра и код драйвера. Как объясняется в разделе «Системные рабочие наборы» в главе 5, этот единый рабочий набор в обиходе называется *рабочим набором системного кэша*, хотя системный кэш — лишь один из компонентов, вносящих в него свой вклад. В рамках данной книги этот рабочий набор будет называться просто *системным рабочим набором*. Также в главе 5 объясняется, что если параметр реестра `LargeSystemCache` равен `1`, то диспетчер памяти отдает предпочтение системному рабочему набору, а не процессам, запущенным в системе.

Физический размер кэша

Хотя рабочий набор системы включает в себя объем физической памяти, отображаемый на представления в виртуальном адресном пространстве кэша, он не обязательно отражает общий объем файловых данных, кэшируемых в физической памяти. Между этими двумя значениями может быть расхождение, поскольку дополнительные файловые данные могут находиться в резервных или измененных списках страниц диспетчера памяти.

Вспомните из главы 5, что во время обрезки рабочего набора или замены страниц диспетчер памяти может перемещать «грязные» страницы из рабочего набора в список либо резервных, либо измененных страниц в зависимости от того, содержит ли страница данные, которые должны быть записаны в файл подкачки или другой файл, прежде чем она может быть использована повторно. Если бы диспетчер памяти не реализовывал эти списки, то при обращении какого-либо процесса к данным, ранее удаленным из его рабочего набора, диспетчеру памяти приходилось бы жестко подставлять их с диска. Вместо этого если данные, к которым осуществляется доступ, присутствуют в любом из этих списков, то диспетчер памяти просто мягко возвращает страницу в рабочий набор процесса. Таким образом, списки служат кэшами данных в памяти, хранящихся в файле подкачки, исполняемых образах или файлах данных. То есть общий объем файловых данных, кэшируемых в системе, включает в себя не только размер рабочего набора системы, но и суммарные размеры списков резервных и измененных страниц.

Следующий пример иллюстрирует, как диспетчер кэша может заставить кэшировать в физической памяти гораздо больше файловых данных, чем содержится в рабочем наборе системы. Рассмотрим систему, работающую как выделенный файловый сервер. Клиентское приложение получает доступ к файловым данным по сети, а сервер, например драйвер файлового сервера %SystemRoot%\System32\Drivers\Srv2.sys, описанный далее в главе, использует интерфейсы диспетчера кэша для чтения и записи файловых данных от имени клиента. Если клиент прочитывает несколько тысяч файлов по 1 Мбайт каждый, то диспетчеру кэша придется начать повторно задействовать представления, когда закончится место для отображения, и он не сможет увеличить область отображения VACB. Для каждого последующего прочитанного файла диспетчер кэша снимает отображение представлений и заново отображает их для новых файлов. Когда диспетчер кэша снимает отображение, диспетчер памяти не удаляет данные файла в рабочем наборе кэша, соответствующие представлению, а перемещает их в список резервных страниц. При отсутствии каких-либо других потребностей в физической памяти этот список может занять почти всю физическую память, остающуюся за пределами рабочего набора системы. Другими словами, практически вся физическая память сервера будет использоваться для кэширования файловых данных (рис. 11.4).

Поскольку общий объем кэшируемых файловых данных включает размеры системного рабочего набора и списков измененных и резервных страниц, которые контролируются диспетчером памяти, он в некотором смысле является настоящим диспетчером кэша. Подсистема диспетчера кэша просто предоставляет удобные интерфейсы для доступа к файловым данным через диспетчер памяти. Он с его политиками упреждающего чтения и отложенной записи играет важную роль, влияя на то, какие данные диспетчер памяти сохраняет в физической памяти, а также управляя системными виртуальными адресными представлениями пространства.

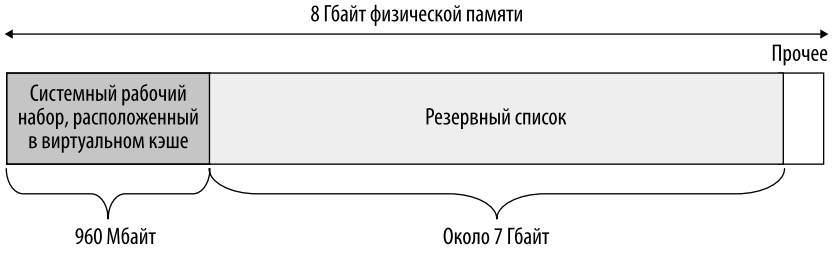


Рис. 11.4. Пример, в котором почти вся физическая память используется файловым кэшем

Чтобы попытаться точно отразить общий объем файловых данных, кэшируемых в системе, диспетчер задач показывает на вкладке быстродействия значение «Кэшировано», отражающее совокупный размер рабочего набора системы и списков резервных и измененных страниц. Process Explorer разбивает эти значения на Cache WS (рабочий набор системного кэша), Standby и Modified. На рис. 11.5 показаны представление информации о системе в Process Explorer и значение Cache WS в области Physical Memory (в левом нижнем углу), а также размер резервного и измененного списков в области Paging Lists (в середине). Обратите внимание на то, что значение «Кэшировано» в диспетчере задач включает также Paged WS, Kernel WS и Driver WS, показанные в Process Explorer. Когда эти значения были выбраны, подавляющее большинство System WS поступало из Cache WS. Сегодня это уже не так, но анахронизм остался в диспетчере задач.

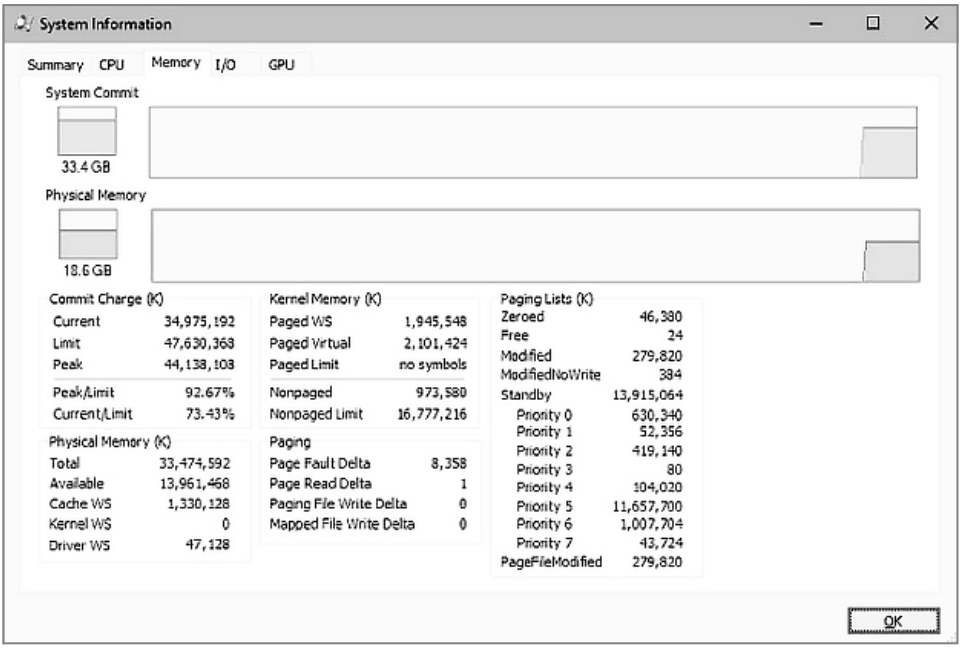


Рис. 11.5. Диалоговое окно с информацией о системе программы Process Explorer

СТРУКТУРЫ ДАННЫХ КЭША

Диспетчер кэша использует следующие структуры данных для отслеживания кэшированных файлов.

- Каждый слот системного кэша размером 256 Кбайт описывается VACB.
- Каждый отдельно открытый кэшированный файл имеет личную карту кэша, содержащую информацию, используемую для управления упреждающим чтением. Это будет обсуждаться далее в разделе «Интеллектуальное упреждающее чтение».
- Каждый кэшируемый файл имеет одну общую структуру карты кэша, указывающую на слоты в системном кэше, содержащие отображенные представления файла.

Эти структуры и их взаимосвязи описаны в следующих разделах.

Общесистемные структуры данных кэша

Как говорилось ранее, диспетчер кэша отслеживает состояние представлений в системном кэше с помощью массива структур данных, называемых *массивами блоков управления виртуальными адресами* (virtual address control block, VACB) и хранящихся в невыгружаемом пуле. В 32-разрядной системе размер каждого VACB составляет 32 байта, а массива VACB — 128 Кбайт, что дает 4096 VACB на массив. В 64-битной системе размер VACB составляет 40 байт, что дает 3276 VACB на массив. Диспетчер кэша выделяет начальный массив VACB во время инициализации системы и связывает его с общесистемным списком массивов VACB, который называется `CcVacbArrays`. Каждый VACB соответствует одному представлению размером 256 Кбайт в системном кэше (рис. 11.6). Структура VACB показана на рис. 11.7.

Кроме того, каждый массив VACB состоит из двух видов VACB: с *низким* и *высоким приоритетом отображения*. Система выделяет 64 начальных VACB с высоким приоритетом для каждого массива VACB. Высокоприоритетные VACB отличаются тем, что их представления предварительно распределены из системного адресного пространства. Когда у диспетчера памяти нет представлений, которые можно было бы передать диспетчеру кэша во время отображения каких-то данных, и если запрос на отображение помечен как высокоприоритетный, диспетчер кэша использует одно из предварительно выделенных представлений, присутствующих в высокоприоритетном VACB. Он применяет эти высокоприоритетные VACB, например, для критических метаданных файловой системы, а также для удаления данных из кэша. Однако после того как VACB с высоким приоритетом будут исчерпаны, любая операция, требующая представления VACB, завершится неудачей из-за недостатка ресурсов. Обычно приоритет отображения устанавливается по умолчанию на низкий, но, применяя флаг `PIN_HIGH_PRIORITY` при закреплении кэшированных данных (описано далее), файловые системы могут запросить использование вместо него VACB с высоким приоритетом, если он необходим.

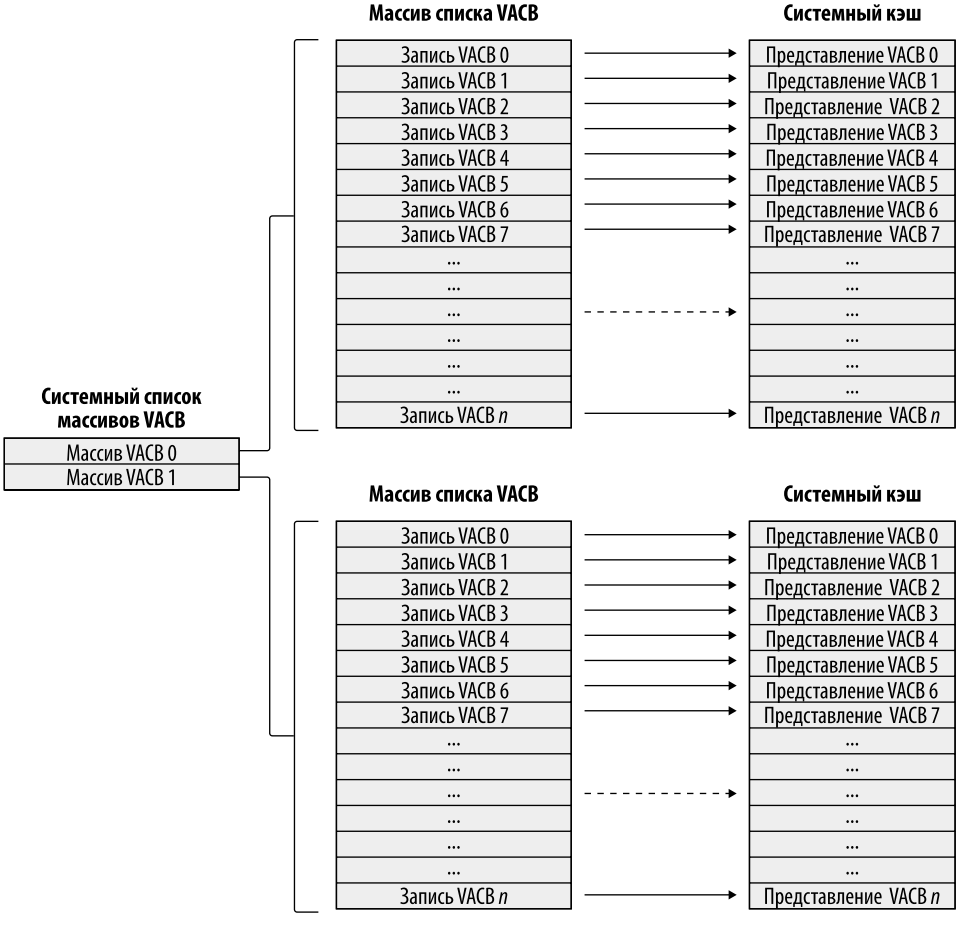


Рис. 11.6. Системный массив VACB

Как видно на рис. 11.7, первое поле в VACB — это виртуальный адрес данных в системном кэше. Второе поле — указатель на структуру карты общего кэша, определяющей, какой файл кэшируется. Третье поле определяет смещение в файле, с которого начинается представление, всегда на основе гранулярности в 256 Кбайт. При такой гранулярности младшие 16 бит смещения файла всегда будут нулевыми, поэтому они применяются для хранения количества ссылок на представление — сколько случаев активных чтения или записи обращаются к представлению. Четвертое поле связывает VACB со списком наименее часто используемых (least recently used, LRU) VACB. Освобождая VACB, диспетчер

Виртуальный адрес данных в системном кэше	
Указатель на карту кэша с общим доступом	
Смещение файла	Активный счетчик
Запись ссылки на начало списка LRU	
Указатель на владеющий массив VACB	

Рис. 11.7. Структура данных VACB

кэша сначала проверяет этот список при выделении нового VACB. Наконец, пятое поле связывает этот VACB с заголовком массива VACB, представляющим массив, в котором хранится VACB.

Во время операции ввода-вывода из файла счетчик ссылок VACB увеличивается, а затем уменьшается, когда эта операция завершается. Когда счетчик ссылок ненулевой, VACB *активен*. Для получения доступа к метаданным файловой системы активный счетчик представляет, сколько драйверов файловой системы заблокировали страницы этого представления в памяти.

ЭКСПЕРИМЕНТ. Просмотр VACB и статистики VACB

Диспетчер кэша скрыто отслеживает различные значения, полезные разработчикам и инженерам поддержки при отладке дампов аварий. Все эти отладочные переменные начинаются с префикса `CcDbg`, что позволяет легко увидеть весь список благодаря команде x:

```
1: kd> x nt!*ccdbg*
fffff800`d052741c nt!CcDbgNumberOfFailedWorkQueueEntryAllocations = <no type
information>
fffff800`d05276ec nt!CcDbgNumberOfNoopedReadAheads = <no type information>
fffff800`d05276e8 nt!CcDbgLsnLargerThanHint = <no type information>
fffff800`d05276e4 nt!CcDbgAdditionalPagesQueuedCount = <no type information>
fffff800`d0543370 nt!CcDbgFoundAsyncReadThreadListEmpty = <no type information>
fffff800`d054336c nt!CcDbgNumberOfCcUnmapInactiveViews = <no type information>
fffff800`d05276e0 nt!CcDbgSkippedReductions = <no type information>
fffff800`d0542e04 nt!CcDbgDisableDAX = <no type information>
...
```

В некоторых системах имена переменных могут различаться в зависимости от 32- или 64-битной реализации. В данном эксперименте точные имена переменных не имеют значения — вместо этого сосредоточьтесь на методологии, которая будет объяснена. Используя эти переменные и свои знания о структурах данных заголовков массива VACB, можно задействовать отладчик ядра, чтобы перечислить все заголовки массива VACB.

Переменная `CcVacbArrays` представляет собой массив указателей на заголовки массивов VACB, которые будут разыменованы, чтобы выгрузить содержимое `_VACB_ARRAY_HEADER`. Сначала получите наивысший индекс массива:

```
1: kd> dd nt!CcVacbArraysHighestUsedIndex  l1
fffff800`d0529c1c  00000000
```

Теперь можно разыменовывать каждый индекс вплоть до максимального. В данной системе, и это нормально, максимальный индекс равен 0, иначе говоря, есть только один заголовок для разыменовывания:

```
1: kd> ?? (*((nt!_VACB_ARRAY_HEADER***)@@(nt!CcVacbArrays)))[0]
struct _VACB_ARRAY_HEADER * 0xffffc40d`221cb000
   +0x000 VacbArrayIndex   : 0
   +0x004 MappingCount     : 0x302
   +0x008 HighestMappedIndex : 0x301
   +0x00c Reserved         : 0
```

Если бы их было больше, то можно было бы изменить индекс массива в конце команды на бóльшее число, пока не будет достигнут самый высокий используемый индекс. Вывод показывает, что в системе имеется только один массив VACB с 770 (0x302) активными VACB.

Наконец, переменная `CcNumberOfFreeVacbs` хранит количество VACB в списке свободных VACB. При дампе этой переменной в системе, используемой для эксперимента, их получается 2506 (0x9ca):

```
1: kd> dd nt!CcNumberOfFreeVacbs  l1
fffff800`d0527318  000009ca
```

Как и ожидалось, сумма свободных (0x9ca — 2506 десятичных) и активных VACB (0x302 — 770 десятичных) в 64-битной системе с одним массивом VACB равна 3276, то есть количеству VACB в одном массиве VACB. Если бы в системе закончились свободные VACB, то диспетчер кэша попытался бы выделить новый массив VACB. Из-за изменчивой ситуации в этом эксперименте система может создать или освободить дополнительные VACB между двумя этапами — дампы сначала активных, а затем свободных VACB. Это может привести к тому, что общее количество свободных и активных VACB не будет точно соответствовать 3276. Попробуйте быстро повторить эксперимент несколько раз, если это произойдет, хотя можно никогда не получить стабильные значения, особенно если в системе велика активность файловой системы.

Структуры данных кэша для отдельного файла

Каждый открытый дескриптор файла имеет соответствующий файловый объект. Файловые объекты подробно описаны в главе 6. Если файл кэшируется, то файловый объект указывает на структуру *закрытой карты кэша*, содержащую местоположение двух последних чтений, чтобы диспетчер кэша мог выполнять интеллектуальное упреждающее чтение (описано далее в одноименном разделе). Кроме того, все закрытые карты кэша для открытых экземпляров файла связаны.

Каждый кэшированный файл, в отличие от файлового объекта, имеет структуру *общей карты кэша*, описывающей состояние кэшированного файла, включая раздел, к которому он принадлежит, его размер и действительную длину данных. Функция поля действительной длины данных объясняется в разделе «Кэширование с отложенной записью и поздняя запись». Общая карта кэша указывает также на *секционный объект* (поддерживается диспетчером памяти и описывает отображение файла в виртуальную память) — список закрытых карт кэша, связанных с этим файлом, а также все VACB, описывающие текущие отображенные представления файла в системном кэше. (Подробнее об указателях секционных объектов см. в главе 5.) Все открытые общие карты кэша для разных файлов собраны в глобальном связанном списке, хранящемся в структуре данных раздела диспетчера кэша. Взаимоотношения между этими структурами данных кэша для каждого файла показаны на рис. 11.8.

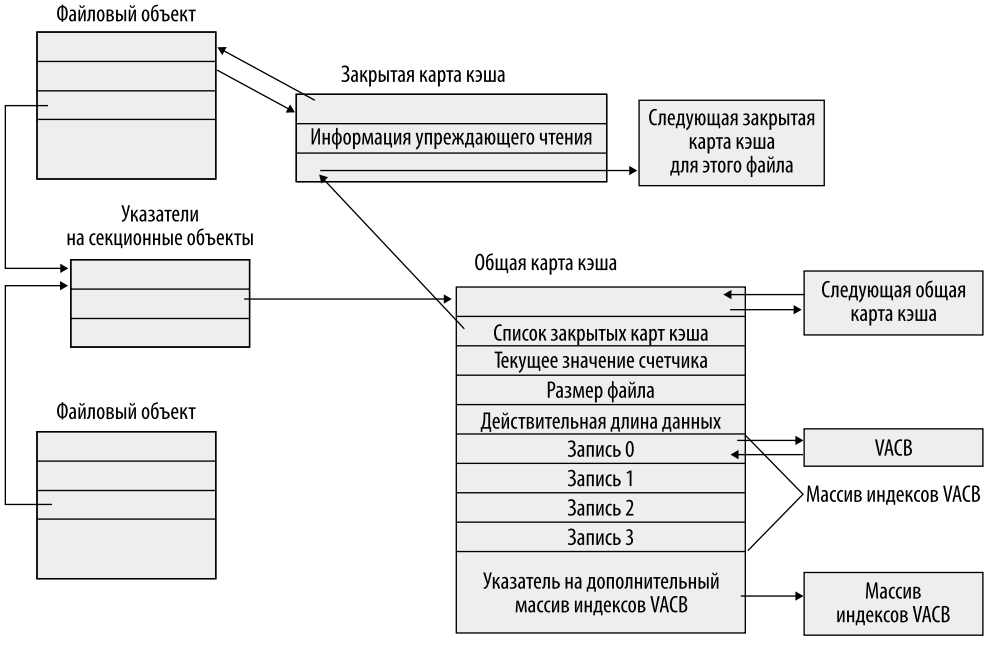


Рис. 11.8. Структуры данных кэша для отдельного файла

При запросе на чтение из определенного файла диспетчер кэша должен получить ответы на следующие вопросы.

1. Находится ли файл в кэше?
2. Если да, то какой VACB, если он есть, относится к запрашиваемому месту?

Другими словами, диспетчер кэша должен выяснить, отображено ли представление файла по нужному адресу в системном кэше. Если ни один VACB не содержит нужного смещения файла, то запрашиваемые данные в текущий момент не отображены в системный кэш.

Чтобы отслеживать, какие представления для данного файла отображаются в системный кэш, диспетчер кэша поддерживает массив указателей на VACB, называющийся *индексным массивом VACB*. Первая запись в индексном массиве VACB относится к первым 256 Кбайт файла, вторая запись — ко вторым 256 Кбайт и т. д. На рис. 11.9 показаны четыре различных раздела из трех разных файлов, в настоящее время отображающихся в системном кэше.

Когда процесс обращается к определенному файлу в определенном месте, диспетчер кэша просматривает соответствующую запись в массиве индексов VACB файла, чтобы узнать, были ли запрошенные данные отображены в кэш. Если запись в массиве ненулевая и, следовательно, содержит указатель на VACB, то область файла, на которую ссылаются, находится в кэше. VACB, в свою очередь, указывает на место в системном кэше, где отображается представление файла. Если запись равна нулю, то диспетчер кэша должен найти свободный слот в системном кэше и, следовательно, свободный VACB, чтобы отобразить требуемое представление.

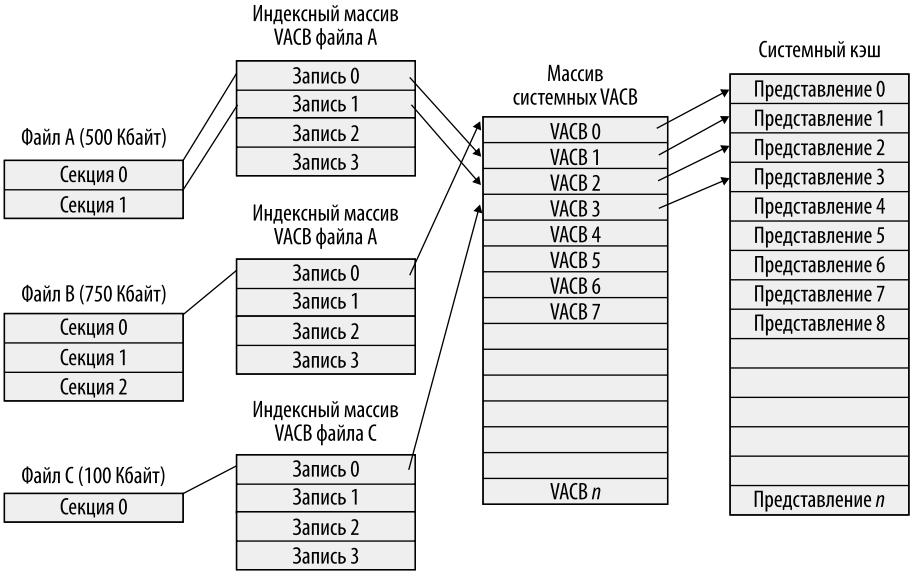


Рис. 11.9. Индексные массивы VACB

В качестве оптимизации размера общая карта кэша содержит индексный массив VACB размером в четыре записи. Поскольку каждый VACB описывает 256 Кбайт, записи в этом небольшом индексном массиве фиксированного размера могут указывать на записи массива VACB, вместе описывающие файл размером до 1 Мбайт. Если размер файла превышает 1 Мбайт, то из невыгружаемого пула выделяется отдельный индексный массив VACB, основанный на размере файла, деленном на 256 Кбайт и округленном в большую сторону, если получается остаток. Общая карта кэша тогда указывает на эту отдельную структуру.

Дальнейшая оптимизация выражается в том, что индексный массив VACB, выделенный из невыгружаемого пула, становится разреженным многоуровневым индексным массивом, если размер файла превышает 32 Мбайт, причем каждый такой массив состоит из 128 записей. Количество уровней, необходимых для файла, можно рассчитать по формуле:

$$(x - 18) / 7,$$

где x — количество битов, необходимых для представления размера файла.

Округлите результат вычисления до ближайшего целого числа. Значение 18 в уравнении обусловлено тем, что один VACB представляет 256 Кбайт, а это 2^{18}. Значение 7 определяется тем, что каждый уровень в массиве имеет 128 записей, а это 2^7. Таким образом, для файла максимального размера, который можно описать 63 битами (это наибольший размер, поддерживаемый диспетчером кэша), потребуется всего семь уровней. Массив разрежен, потому что диспетчер кэша выделяет только те ветви, для которых есть активные представления на самом нижнем уровне индексного массива. На рис. 11.10 показан пример многоуровневого массива VACB для разреженного файла, достаточно большого для того, чтобы потребовать трех уровней.

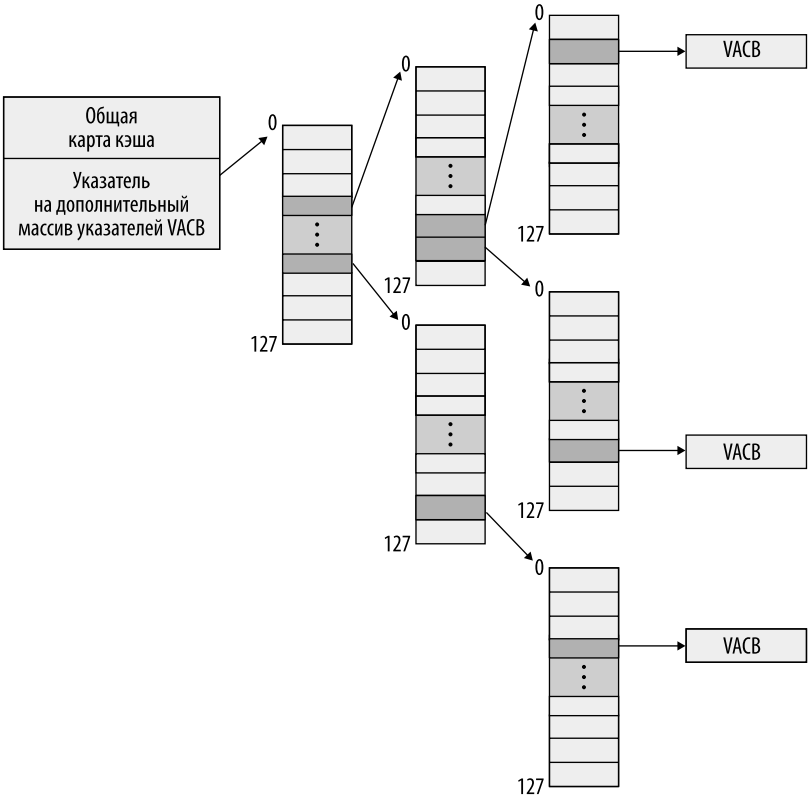


Рис. 11.10. Многоуровневые массивы VACB

Эта схема необходима для эффективной работы с разреженными файлами, которые могут быть очень объемными и содержать лишь небольшую часть достоверных данных, поскольку массив выделяется только для работы с текущими отображенными представлениями файла. Например, для 32-гигабайтного разреженного файла, для которого в виртуальном адресном пространстве кэша отображено только 256 Кбайт, потребуется массив VACB с тремя выделенными индексными массивами, поскольку лишь одна ветвь массива имеет отображение, а для 32-гигабайтного файла требуется трехуровневый массив. Если бы диспетчер кэша не использовал оптимизацию многоуровневого индексного массива VACB для этого файла, ему пришлось бы выделить индексный массив VACB со 128 000 записей, что эквивалентно 1000 индексных массивов VACB.

ИНТЕРФЕЙСЫ ФАЙЛОВОЙ СИСТЕМЫ

При первом обращении к данным файла для чтения или записи в кэш драйвер файловой системы должен определить, отображена ли часть файла в системном кэше. Если нет, то он должен вызвать функцию `CcInitializeCacheMap`, чтобы установить структуры данных для каждого файла, описанные в предыдущем разделе.

Когда файл настроен на кэшированный доступ, драйвер файловой системы вызывает одну из нескольких функций для доступа к данным в файле. Существует три основных метода доступа к кэшированным данным, каждый из которых предназначен для определенной ситуации.

- Метод копирования копирует пользовательские данные между буферами кэша в системном пространстве и буфером процесса в пространстве пользователя.
- Метод сопоставления и закрепления использует виртуальные адреса для чтения данных непосредственно из буферов кэша и записи в них.
- Метод доступа к физической памяти применяет физические адреса для чтения данных напрямую из буферов кэша и записи в них.

Драйверы файловых систем должны предоставлять две версии операции чтения файла — с кэшем и без кэша — для предотвращения бесконечного цикла, когда диспетчер памяти обрабатывает отказ страницы. Когда диспетчер памяти разрешает отказ страницы, вызывая файловую систему для получения данных из файла, разумеется, через драйвер устройства, он должен указать это как операцию чтения с подкачкой, установив в IRP флаги `no cache` и `paging IO`.

На рис. 11.11 показано типичное взаимодействие между диспетчером кэша, диспетчером памяти и драйверами файловой системы в ответ на чтение файла или запись в него пользователем. Диспетчер кэша вызывается файловой системой с помощью интерфейсов копирования — путей `CcCopyRead` и `CcCopyWrite`. Например, для обработки чтения `CcFastCopyRead` или `CcCopyRead` диспетчер кэша создает в кэше представление для отображения части читаемого файла и считывает данные из него в пользовательский буфер путем копирования из этого представления. Операция копирования вызывает отказы страницы при обращении к каждой ранее недопустимой странице в представлении, в ответ диспетчер памяти инициирует некэшируемый ввод-вывод в драйвер файловой системы для получения данных, соответствующих части файла, отображенной на страницу, породившую отказ.

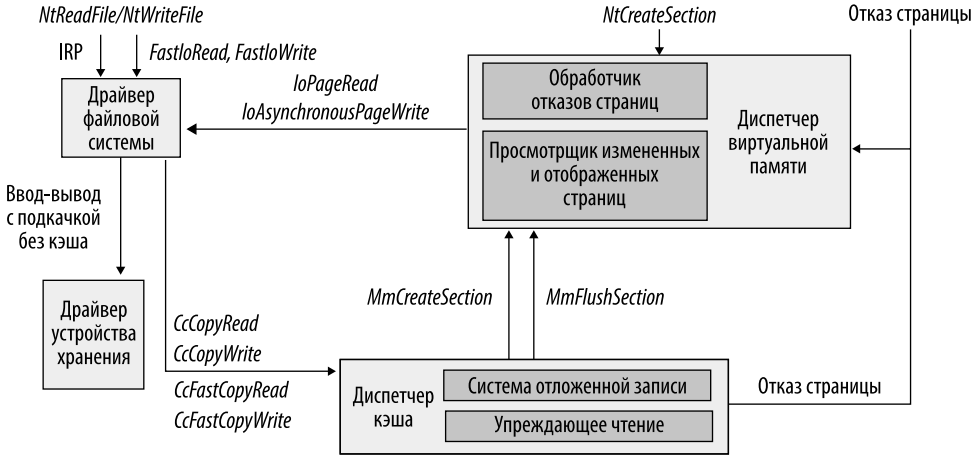


Рис. 11.11. Взаимодействие файловой системы с диспетчерами кэша и памяти

В следующих трех разделах рассказывается об этих механизмах доступа к кэшу, их назначении и использовании.

Копирование в кэш и из него

Поскольку системный кэш находится в системном пространстве, он отображается в адресное пространство каждого процесса. Однако, как и все страницы системного пространства, страницы кэша недоступны из пользовательского режима, поскольку это было бы потенциальной брешью в системе безопасности. Например, процесс может не иметь прав на чтение файла, данные которого в этот момент содержатся в какой-то части системного кэша. Таким образом, чтение файлов пользовательских приложений из кэша и запись в него должны обслуживаться процедурами режима ядра, копирующими данные между буферами кэша в системном пространстве и буферами приложения в адресном пространстве процесса.

Кэширование с помощью интерфейсов отображения и закрепления

Точно так же, как пользовательские приложения читают данные из файлов на диске и записывают в них, драйверам файловой системы необходимо читать и записывать данные, описывающие сами файлы, — метаданные, или данные структуры тома. Однако, поскольку драйверы файловой системы работают в режиме ядра, они могут (если диспетчер кэша будет должным образом проинформирован) изменять данные непосредственно в системном кэше. Для такой оптимизации диспетчер кэша предоставляет функции, позволяющие драйверам файловой системы находить в виртуальной памяти метаданные файловой системы, что дает возможность выполнять прямые изменения без использования промежуточных буферов.

Если драйверу файловой системы нужно прочитать метаданные файловой системы в кэше, то он обращается к интерфейсу отображения диспетчера кэша, чтобы получить виртуальный адрес нужных данных. Диспетчер кэша обращается ко всем запрошенным страницам, чтобы доставить их в память, а затем возвращает управление драйверу файловой системы. После этого драйвер файловой системы может получить прямой доступ к данным.

Если драйверу файловой системы нужно изменить страницы кэша, он обращается к службам закрепления диспетчера кэша, которые сохраняют страницы активными в виртуальной памяти, чтобы их нельзя было вернуть. На самом деле страницы не фиксируются в памяти, например, когда драйвер устройства фиксирует страницы для передачи данных при прямом доступе к памяти. Чаще всего драйвер файловой системы помечает поток метаданных как запрещенный для записи, что дает указание устройству записи отображенных страниц диспетчера памяти, описанному в главе 5, не записывать страницы на диск до тех пор, пока это не будет сделано по явному указанию. Когда драйвер файловой системы снимает с них закрепление, диспетчер кэша освобождает свои ресурсы, чтобы отложенно сбросить все изменения на диск и освободить представление кэша, которое занимали метаданные.

Интерфейсы отображения и закрепления решают одну сложную проблему реализации файловой системы — управление буферами. Без непосредственного

манипулирования кэшированными метаданными файловая система должна предсказывать максимальное количество буферов, которые ей понадобятся при обновлении структуры тома. Позволяя файловой системе обращаться к метаданным и обновлять их непосредственно в кэше, диспетчер кэша устраняет необходимость в буферах, просто обновляя в виртуальной памяти структуру тома, предоставляемую диспетчером памяти. Единственное ограничение, с которым сталкивается файловая система, — это объем доступной памяти.

Кэширование с помощью интерфейсов прямого доступа к памяти

В дополнение к интерфейсам отображения и привязки, используемым для доступа к метаданным непосредственно в кэше, диспетчер кэша предоставляет третий интерфейс для доступа к кэшированным данным — *прямой доступ к памяти* (direct memory access, DMA). Функции DMA применяются для чтения из страниц кэша или записи в них без промежуточных буферов — например, когда сетевая файловая система выполняет передачу по сети.

Интерфейс DMA возвращает файловой системе физические адреса кэшированных пользовательских данных (а не виртуальные, возвращаемые интерфейсами отображения и закрепления), которые затем могут быть задействованы для передачи данных непосредственно из физической памяти в сетевое устройство. Для передачи небольших объемов данных, от 1 до 2 Кбайт, можно применять обычные интерфейсы копирования на основе буфера, а при передаче больших объемов интерфейс DMA может привести к значительному повышению производительности сетевого сервера, обрабатывающего файловые запросы от удаленных систем. Для описания этих ссылок на физическую память используется *список дескрипторов памяти* (memory descriptor list, MDL). О MDL рассказывается в главе 5.

БЫСТРЫЙ ВВОД-ВЫВОД

По возможности чтение из кэшированных файлов и запись в них обрабатываются с помощью высокоскоростного механизма, называемого *быстрым вводом-выводом*. Так называется средство чтения из кэшированного файла или записи в него без необходимости создавать IRP. При быстром вводе-выводе диспетчер ввода-вывода вызывает подпрограмму быстрого ввода-вывода драйвера файловой системы, чтобы проверить, можно ли выполнить эту процедуру непосредственно из диспетчера кэша без создания IRP.

Поскольку диспетчер кэша спроектирован поверх подсистемы виртуальной памяти, драйверы файловых систем могут использовать его для доступа к файловым данным, просто копируя их на страницы, сопоставленные с реальным файлом, на который ссылаются, или из этих страниц, не тратя времени на создание IRP.

Быстрый ввод-вывод происходит не всегда. Например, первое чтение из файла или запись в него требует подготовки файла для кэширования — его отображения в кэш и настройки структур данных кэша, как объяснялось ранее в разделе «Структуры данных кэша». Кроме того, если вызывающая сторона указала асинхронное чтение или запись, то быстрый ввод-вывод не используется, поскольку вызывающая сторона может быть остановлена во время операций ввода-вывода с подкачкой, необходимых для копирования буфера в системный кэш или из него, и, таким образом, на самом

деле не обеспечивает запрошенную асинхронную операцию ввода-вывода. Но даже при синхронной операции ввода-вывода драйвер файловой системы может решить, что он не способен обработать операцию ввода-вывода с помощью механизма быстрого ввода-вывода, например, если в данном файле заблокирован диапазон байтов в результате вызовов функций Windows `LockFile` и `UnlockFile`. Поскольку диспетчер кэша не знает, какие части файлов заблокированы, драйвер файловой системы должен проверить достоверность чтения или записи, что требует создания IRP. Дерево решений для быстрого ввода-вывода показано на рис. 11.12.

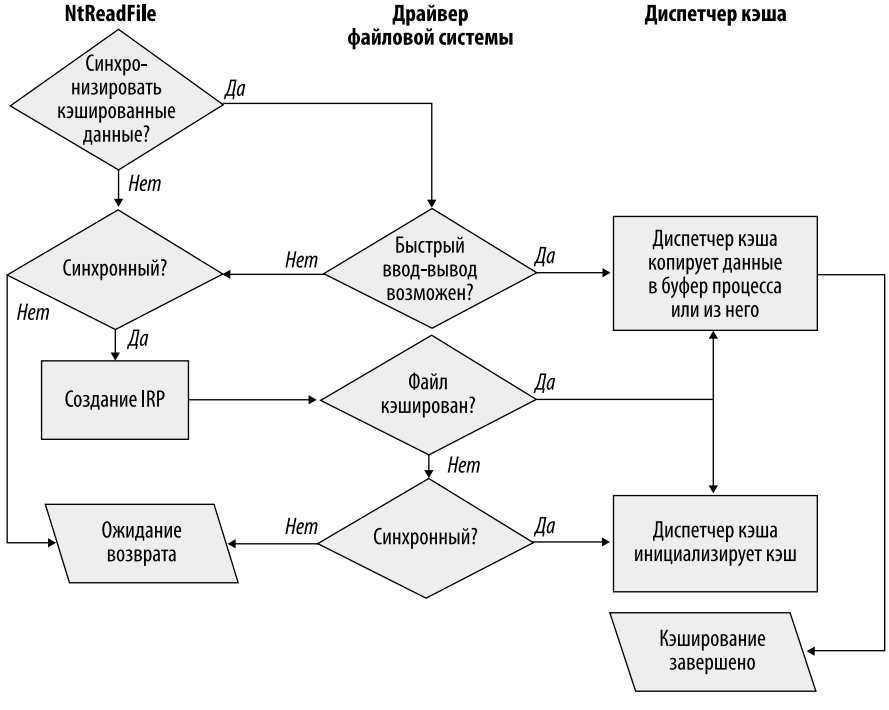


Рис. 11.12. Дерево решений по быстрому вводу-выводу

При обслуживании чтения или записи с помощью быстрого ввода-вывода выполняются следующие шаги.

1. Поток выполняет операцию чтения или записи.
2. Если файл кэширован и ввод-вывод синхронный, запрос переходит к точке входа быстрого ввода-вывода в стеке драйвера файловой системы. Если файл не кэширован, то драйвер файловой системы подготавливает его для кэширования, чтобы в следующий раз для удовлетворения запроса на чтение или запись можно было использовать быстрый ввод-вывод.
3. Если процедура быстрого ввода-вывода драйвера файловой системы определяет, что быстрый ввод-вывод возможен, она вызывает процедуру чтения или записи диспетчера кэша, чтобы получить доступ к данным файла непосредственно в кэше. Если быстрый ввод-вывод невозможен, драйвер файловой системы

возвращается к системе ввода-вывода, которая создает IRP для ввода-вывода и в конечном счете вызывает обычную процедуру чтения файловой системы.

4. Диспетчер кэша преобразует указанное смещение файла в виртуальный адрес в кэше.
5. При чтении диспетчер кэша копирует данные из кэша в буфер процесса, запрашивающего их, при записи копирует данные из буфера в кэш.
6. Происходит одно из следующих действий.
   - Для чтения, при котором `FILE_FLAG_RANDOM_ACCESS` не был указан при открытии файла, обновляется информация об упреждающем чтении в закрытой карте кэша вызывающей стороны. Упреждающее чтение может вызываться также для файлов, для которых не указан флаг `FO_RANDOM_ACCESS`.
   - Для записи устанавливается «грязный» бит любой измененной страницы в кэше, чтобы система отложенной записи знала, что нужно сбросить ее на диск.
   - Для файлов со сквозной записью все изменения сбрасываются на диск.

ЧТЕНИЕ С УПРЕЖДЕНИЕМ И ОТЛОЖЕННАЯ ЗАПИСЬ

В этом разделе будет показано, как диспетчер кэша реализует чтение и запись файловых данных от имени драйверов файловой системы. Помните, что диспетчер кэша участвует в файловом вводе-выводе только тогда, когда файл открывается без флага `FILE_FLAG_NO_BUFFERING`, а затем считывается или записывается с помощью ввода-вывода Windows, например с помощью функций Windows `ReadFile` и `WriteFile`. Отображенные файлы не проходят через диспетчер кэша, как и файлы, открытые с установленным флагом `FILE_FLAG_NO_BUFFERING`.

ПРИМЕЧАНИЕ Когда приложение использует флаг FILE_FLAG_NO_BUFFERING для открытия файла, его файловый ввод-вывод должен начинаться с выровненных по устройству смещений и иметь размеры, кратные размеру выравнивания. Его входные и выходные буферы также должны быть выровненными по устройству виртуальными адресами. Для файловых систем это обычно соответствует размеру сектора, типично 4096 байт в NTFS и 2048 байт в CDFS. Одним из преимуществ диспетчера кэша, помимо собственно производительности кэширования, является то, что он выполняет промежуточную буферизацию, позволяя осуществлять ввод-вывод с произвольным выравниванием и размером.

Интеллектуальное упреждающее чтение

Диспетчер кэша использует принцип пространственной локальности для *интеллектуального упреждающего чтения*, предсказывая, какие данные вызывающий процесс может прочитать следующими, основываясь на данных, которые он читает в данный момент. Поскольку системный кэш основан на виртуальных адресах, смежных с конкретным файлом, то не имеет значения, расположены ли они рядом в физической памяти. Предугадывание чтения файлов для кэширования логических блоков более сложно и требует тесного сотрудничества между драйверами файловой системы и блочным кэшем, поскольку система кэша основана на относительном расположении данных на диске, а файлы, разумеется, не всегда хранятся на диске

рядом друг с другом. Можно посмотреть на активность упреждающего чтения через счетчик производительности Cache: Read Aheads/sec или системную переменную `CcReadAheadIos`.

Чтение следующего блока файла, к которому осуществляется последовательный доступ, обеспечивает очевидное повышение производительности, но при этом вызывает лишние перемещения считывающих головок жесткого диска. Чтобы распространить преимущества упреждающего чтения на случаи доступа к длинной цепочке последовательных данных как в прямом, так и в обратном направлении, диспетчер кэша отслеживает историю двух последних запросов на чтение в закрытой карте кэша для файла, к которому осуществляется доступ, — этот метод известен как *асинхронное упреждающее чтение с историей*. Если из произвольных на первый взгляд чтений вызывающего пользователя можно выявить какую-то закономерность, то диспетчер кэша экстраполирует ее. Например, если вызывающая сторона читает страницу 4000, а затем страницу 3000, то диспетчер кэша предполагает, что следующей страницей, которая потребуется вызывающей стороне, будет страница 2000, и предварительно читает ее.

ПРИМЕЧАНИЕ Хотя вызывающая сторона должна выполнить как минимум три операции чтения, чтобы установить предсказуемую последовательность, только две из них сохраняются в закрытой карте кэша.

Чтобы сделать чтение с упреждением еще более эффективным, функция Win32 `CreateFile` предоставляет флаг, указывающий на прямой последовательный доступ к файлу, — `FILE_FLAG_SEQUENTIAL_SCAN`. Если он установлен, то диспетчер кэша не хранит историю чтения для вызывающей стороны для предсказания, а вместо этого выполняет последовательное упреждающее чтение. Однако по мере чтения файла в рабочий набор кэша диспетчер кэша снимает с отображения представления файла, которые больше неактивны, и, если они не изменены, направляет диспетчер памяти на размещение страниц, принадлежащих этим представлениям, в начале списка ожидания, чтобы они могли быть быстро использованы повторно. Он также считывает вперед в два раза больше данных, например 2 Мбайт вместо 1 Мбайт. По мере того как вызывающая сторона продолжает чтение, диспетчер кэша предварительно считывает дополнительные блоки данных, всегда оставаясь впереди вызывающей стороны примерно на одно чтение размером с текущее.

Чтение диспетчера кэша является асинхронным, поскольку выполняется в потоке, отдельном от потока вызывающей стороны, и происходит одновременно с его выполнением. При вызове для извлечения кэшированных данных диспетчер кэша сначала обращается к запрошенной виртуальной странице, чтобы удовлетворить запрос, а затем ставит в очередь дополнительный запрос ввода-вывода для извлечения дополнительных данных в рабочий поток системы. Затем рабочий поток действует в фоновом режиме, считывая дополнительные данные в ожидании следующего запроса на чтение. Предварительно прочитанные страницы загружаются в память во время выполнения программы, так что, когда вызывающая сторона запрашивает данные, они уже находятся в памяти.

Для приложений, у которых нет предсказуемой схемы чтения, при вызове функции `CreateFile` можно установить флаг `FILE_FLAG_RANDOM_ACCESS`. Он указывает диспетчеру кэша не пытаться предсказать, где приложение будет читать дальше,

и таким образом отключает упреждающее чтение. Флаг также не позволяет диспетчеру кэша агрессивно снимать с отображения представления файла по мере обращения к нему, чтобы минимизировать активность отображения или снятия с него файла при повторном обращении приложения к его частям.

Усовершенствования, связанные с упреждающим чтением

В Windows 8.1 были внесены некоторые улучшения в функциональность диспетчера кэша для чтения. Драйверы файловых систем и сетевые перенаправляющие устройства могут определять размер и рост для интеллектуального упреждающего чтения с помощью API-функции `CcSetReadAheadGranularityEx`. Клиент диспетчера кэша может принимать следующие решения.

- **Гранулярность упреждающего чтения.** Устанавливает минимальный размер единицы упреждающего чтения и смещение конечного файла следующего упреждающего чтения. Диспетчер кэша устанавливает гранулярность по умолчанию в 4 Кбайт — размер страницы памяти, но каждая файловая система определяет это значение по-своему. NTFS, например, устанавливает гранулярность кэша в 64 Кбайт.

  На рис. 11.13 показан пример упреждающего чтения для файла размером 200 Кбайт, где гранулярность кэша установлена в 64 Кбайт. Если пользователь запрашивает чтение без выравнивания на 1 Кбайт по смещению 0x10800 и последовательное чтение уже было обнаружено, то интеллектуальное упреждающее чтение выдаст ввод-вывод, охватывающий 64 Кбайт данных со смещения 0x10000 по 0x20000. Если уже было более двух последовательных чтений, то диспетчер кэша выполняет еще одно дополнительное чтение со смещения 0x20000 по смещение 0x30000, то есть 192 Кбайт.

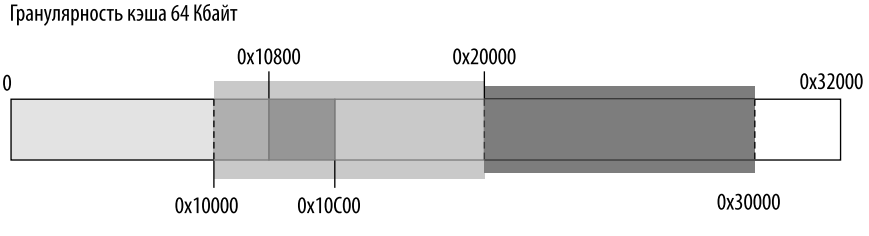


Рис. 11.13. Упреждающее чтение для файла размером 200 Кбайт при гранулярности 64 Кбайт

- **Размер конвейера.** Для некоторых драйверов удаленных файловых систем может быть разумно разделить большие операции ввода-вывода с упреждением чтения на более мелкие фрагменты, которые будут выдаваться параллельно рабочими потоками диспетчера кэша. Сетевая файловая система может достичь значительно большей пропускной способности с помощью этой техники.
- **Агрессивность упреждающего чтения.** Драйверы файловых систем могут указать процент, используемый диспетчером кэша для принятия решения об увеличении размера упреждающего чтения после обнаружения третьего последовательного чтения. Предположим, что приложение читает большой файл,

используя размер ввода-вывода 1 Мбайт. После десятого чтения приложение уже прочитало 10 Мбайт — диспетчер кэша, возможно, уже предзагрузил некоторые из них. Интеллектуальное управление чтением теперь решает, как увеличить размер ввода-вывода при чтении. Если файловая система задала 60 % роста, то применяется следующая формула:

(количество последовательных чтений × размер последнего чтения) × (процент роста / 100).

Это означает, что следующий размер упреждающего чтения составит 6 Мбайт вместо 2 Мбайт, если предположить, что гранулярность составляет 64 Кбайт, а размер ввода-вывода — 1 Мбайт. Процент роста по умолчанию составляет 50 %, если он не изменен ни одним клиентом диспетчера кэша.

Кэширование с отложенной записью и поздняя запись

Диспетчер кэша реализует кэш с отложенной записью и поздней записью. Это означает, что данные, записанные в файлы, сначала хранятся в памяти на страницах кэша, а затем записываются на диск. Таким образом, операции записи накапливаются в течение короткого времени, а затем сразу сбрасываются на диск, что сокращает общее количество операций ввода-вывода на диск.

Диспетчер кэша должен явно вызывать диспетчер памяти для сброса страниц кэша, потому что в противном случае диспетчер памяти записывает содержимое памяти на диск только тогда, когда спрос на физическую память превышает предложение, как и положено для изменяющихся данных. Однако кэшированные данные файлов представляют собой неизменяющиеся данные на диске. Если процесс изменяет кэшированные данные, то пользователь ожидает, что их содержимое будет своевременно отражено на диске.

Кроме того, диспетчер кэша имеет возможность наложить вето на поток записи с отображением диспетчера памяти. Поскольку измененный список (более подробную информацию см. в главе 5) не сортируется в порядке логических адресов блоков (logical block address, LBA), попытки диспетчера кэша сгруппировать страницы для более крупных последовательных операций ввода-вывода на диск не всегда успешны и фактически приводят к повторным передвижениям головок. Для борьбы с этим эффектом диспетчер кэша имеет возможность наложить агрессивное вето на поток записи с отображением и осуществлять запись в порядке виртуального смещения байта (virtual byte offset, VBO), который гораздо ближе к порядку LBA на диске. Поскольку диспетчер кэша теперь владеет этими записями, он также может применять собственные алгоритмы планирования и дросселирования, чтобы предпочесть упреждающее чтение, а не отложенную запись и меньше влиять на систему.

Решение о том, как часто сбрасывать кэш на диск, очень важно. Если он сбрасывается слишком часто, то производительность системы будет снижена из-за ненужных операций ввода-вывода. Если сбрасывается слишком редко, есть риск потерять измененные данные файла в случае системного сбоя (эта потеря особенно раздражает пользователей, которые знают, что они просили приложение сохранить изменения) и исчерпать физическую память, потому что она используется избытком измененных страниц.

Чтобы сгладить эти проблемы, функция *сканирования системы поздней записи* диспетчера кэша выполняется в системном рабочем потоке 1 раз в секунду. У сканирования системы поздней записи есть другие обязанности.

- Проверяет среднее количество доступных и «грязных» страниц, принадлежащих текущему разделу, и обновляет нижний и верхний пределы порога «грязных» страниц. Сам порог тоже обновляется, в основном на основе общего количества «грязных» страниц, записанных в предыдущем цикле (подробнее — в следующих абзацах). Если нет «грязных» страниц для записи, то поток переходит в спящий режим.
- Вычисляет количество «грязных» страниц для записи на диск с помощью внутренней процедуры `CcCalculatePagesToWrite`. Если количество «грязных» страниц превышает 256 (1 Мбайт данных), то диспетчер кэша ставит в очередь для записи на диск 1/8 часть от общего количества «грязных» страниц. Если скорость создания таких страниц превышает количество, которое система поздней записи определила для себя, то она записывает дополнительное количество «грязных» страниц, которое, по ее расчетам, необходимо, чтобы соответствовать этой скорости.
- Циклически переходит от одной общей карты кэша к другой (они хранятся в связанном списке, принадлежащем текущему разделу) и, используя внутреннюю процедуру `CcShouldLazyWriteCacheMap`, определяет, нужно ли сбрасывать на диск текущий файл, описанный общей картой кэша. Существуют различные причины, по которым файл не следует сбрасывать на диск, например: ввод-вывод мог быть уже инициализирован другим потоком, файл может быть временным или, что еще проще, в карте кэша может не быть «грязных» страниц. Если процедура определила, что файл должен быть удален, то сканер системы поздней записи проверяет, достаточно ли еще свободных страниц для записи, и если да, то посылает рабочий элемент рабочим потокам системы диспетчера кэша.

ПРИМЕЧАНИЕ Сканирование системы поздней записи использует некоторые исключения при определении количества отображаемых конкретной картой общего кэша «грязных» страниц для записи (не всегда записываются все «грязные» страницы файла): если целевой файл является потоком метаданных с более чем 256 Кбайт «грязных» страниц, то диспетчер кэша записывает только 1/8 часть всех страниц. Еще одно исключение используется для файлов, содержащих больше «грязных» страниц, чем общее количество страниц, которые может сбросить на диск сканирование системы поздней записи.

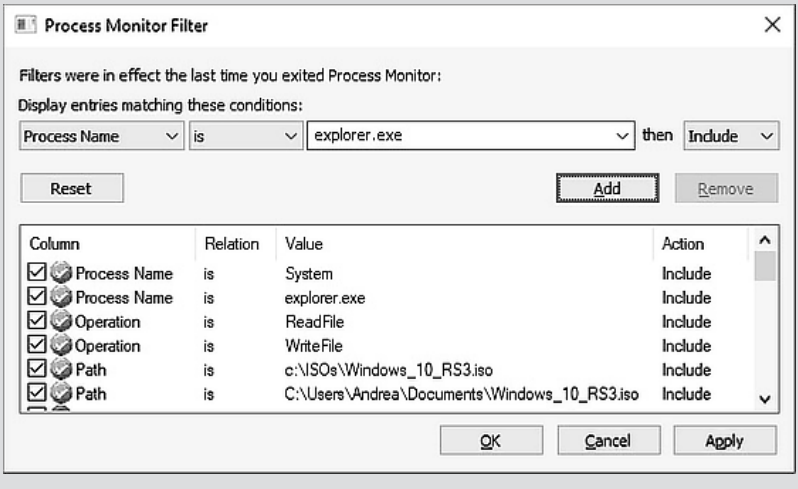
Рабочие потоки системы поздней записи из общесистемного пула критических рабочих потоков фактически выполняют операции ввода-вывода. Система поздней записи также знает о том, когда записывающее устройство отображаемых страниц диспетчера памяти уже выполняет сброс на диск. В таких случаях она задерживает отложенную запись в тот же поток, чтобы избежать ситуации, когда два сбрасывателя на диск пишут в один и тот же файл.

ПРИМЕЧАНИЕ Диспетчер кэша предоставляет драйверам файловой системы возможность отслеживать, когда и сколько данных было записано в файл. После того как система поздней записи сбрасывает «грязные» страницы на диск, диспетчер кэша уведомляет файловую систему, инструктируя ее обновить свое представление о действительной длине данных файла. Диспетчер кэша и файловые системы отдельно отслеживают в памяти действительную длину данных для файла.

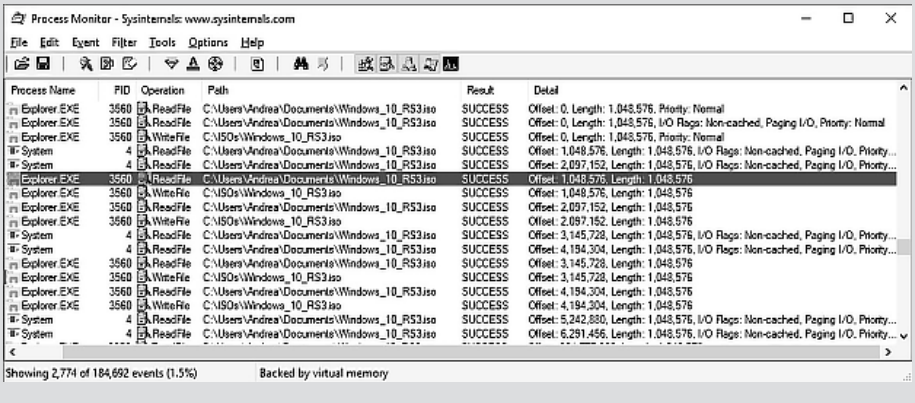
ЭКСПЕРИМЕНТ. Наблюдение за работой диспетчера кэша в действии

В этом эксперименте используется Process Monitor для просмотра активности файловой системы, включая чтение и запись диспетчера кэша, когда в Проводнике Windows из одного локального каталога в другой копируется большой файл — образ DVD.

Сначала настройте фильтр Process Monitor на включение путей к исходному и целевому файлам, процессов Explorer.exe и System, а также операций ReadFile и WriteFile. В этом примере файл C:\Users\Andrea\Documents\Windows_10_RS3.iso был скопирован в C:\ISOs\Windows_10_RS3.iso, поэтому фильтр настроен следующим образом:

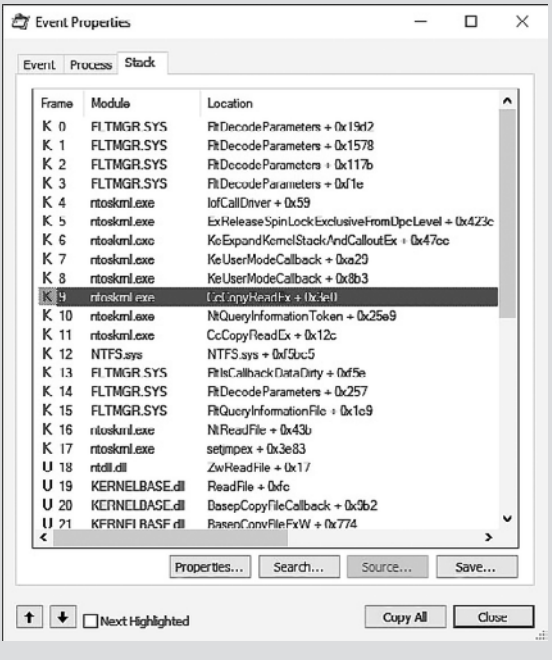


После копирования файла должна появиться трассировка Process Monitor, подобная показанной далее:



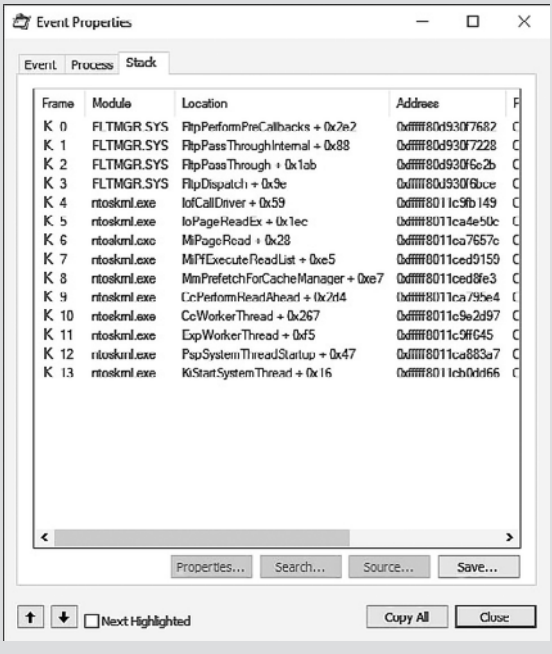
Первые несколько записей показывают начальную обработку ввода-вывода, выполняемую механизмом копирования, и первые операции диспетчера кэша. Вот часть из них, которые можно увидеть.

- Первоначальное кэшированное чтение 1 Мбайт из Проводника при первой записи. Размер этого чтения зависит от внутреннего расчета матрицы, основанной на размере файла, и может варьироваться от 128 Кбайт до 1 Мбайт. Поскольку этот файл был большим, механизм копирования выбрал 1 Мбайт.
- За чтением 1 Мбайт следует еще одно чтение 1 Мбайт без кэша. Некэшированные чтения обычно указывают на активность, связанную с ошибками страниц или доступом к диспетчеру кэша. Более внимательное изучение трассировки стека для этих событий, которую можно увидеть, дважды щелкнув запись и выбрав вкладку Стек (Stack), показывает, что, действительно, процедура диспетчера кэша `CcCopyRead`, вызываемая процедурой чтения драйвера NTFS, заставляет диспетчер памяти перевести исходные данные в физическую память через отказ страницы.

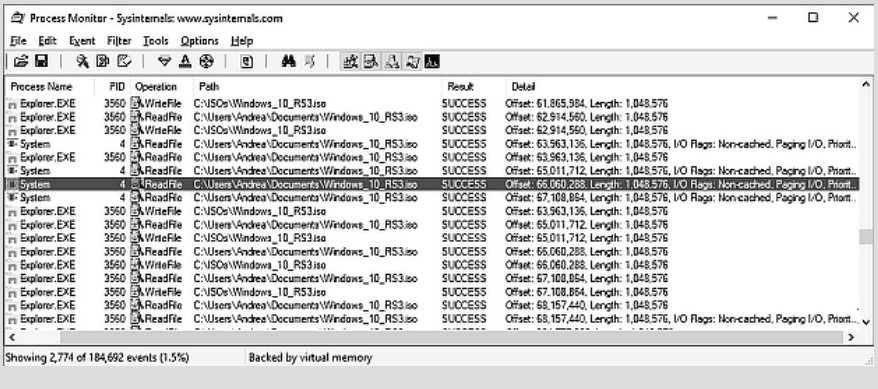


- После этого 1-мегабайтного ввода-вывода по отказу страницы механизм упреждающего чтения диспетчера кэша начинает чтение файла, включающее последующее некэшируемое чтение процесса System на 1 Мбайт со смещением в 1 Мбайт. Учитывая размеры файла и ввода-вывода при чтении в Проводнике, диспетчер кэша выбрал 1 Мбайт в качестве оптимального размера упреждающего чтения. Трассировка стека для одной из операций

чтения с упреждением, показанная далее, подтверждает, что один из рабочих потоков диспетчера кэша выполняет чтение с упреждением.

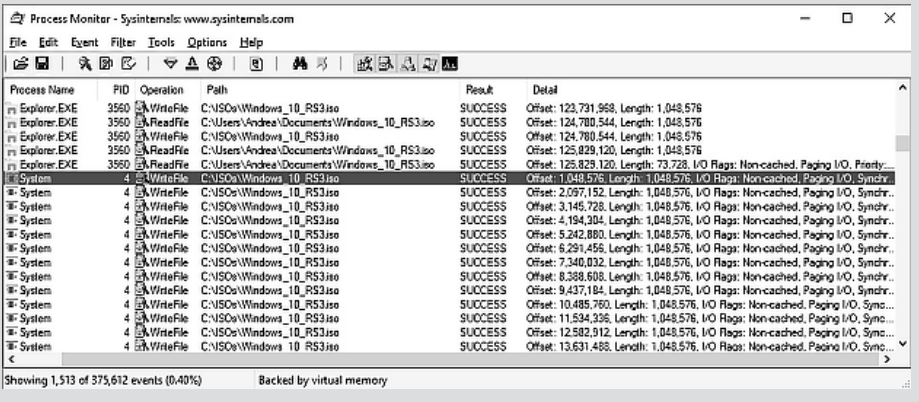


- После этого 1-мегабайтные чтения в Проводнике не сопровождаются отказами страниц, поскольку поток чтения с упреждением опережает Проводник, предварительно получая данные файла с помощью своих 1-мегабайтных некэшированных чтений. Однако время от времени поток упреждающего чтения не успевает собрать достаточно данных, и возникают кластерные страничные ошибки, проявляющиеся как ввод-вывод синхронной подкачки.



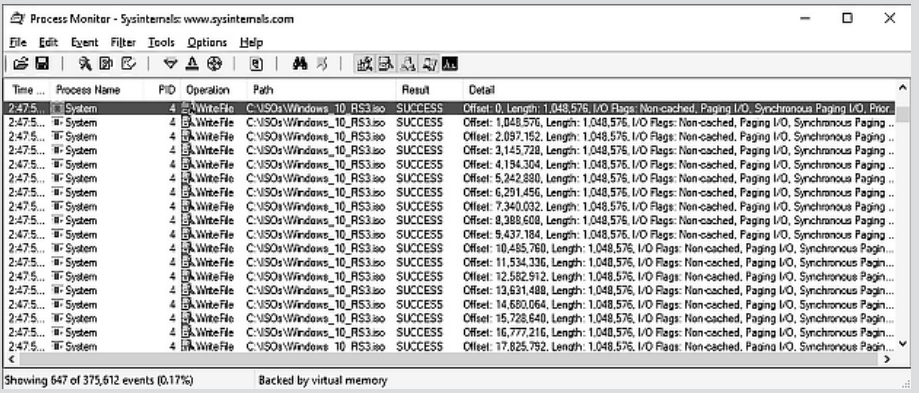
Если посмотреть на стек этих записей, то можно увидеть, что вместо `MmPrefetchForCacheManager` вызываются процедуры `MmAccessFault/MiIssueHardFault`.

Как только начинается чтение, Проводник выполняет запись в целевой файл. Это последовательные записи с кэшированием по 1 Мбайт. После примерно 124 Мбайт чтения реализуется первая операция `WriteFile` из процесса System, как показано на снимке экрана:

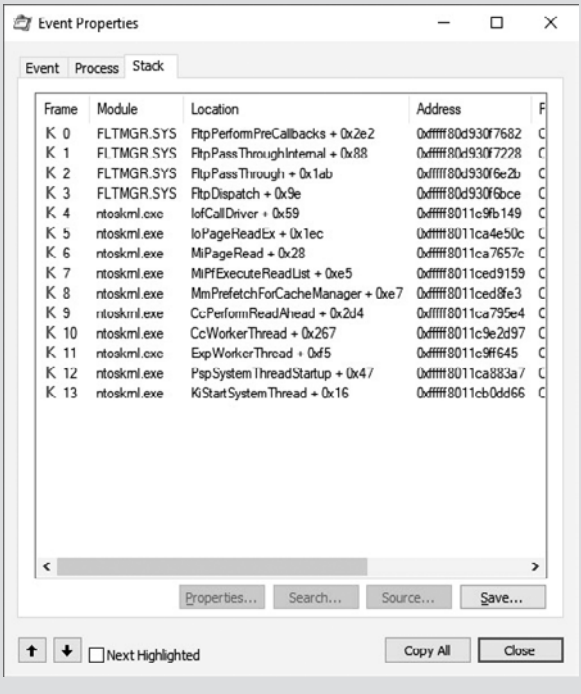


Трассировка стека операции записи, показанная здесь, демонстрирует, что на самом деле за запись отвечал поток *системы записи отображенной страницы* диспетчера памяти. Это происходит потому, что для первых нескольких мегабайт данных диспетчер кэша еще не начал выполнять отложенную запись, так что поток записи отображенных страниц диспетчера памяти начал сбрасывать на диск измененные данные целевого файла. Дополнительные сведения о программе записи отображенных страниц см. в главе 10.

Чтобы получить более четкое представление об операциях диспетчера кэша, удалите Проводник из фильтра Process Monitor, чтобы были видны только операции процесса System, как показано на снимке экрана:



В таком виде гораздо проще увидеть операции диспетчера кэша по отложенной записи на 1 Мбайт. Максимальный размер записи составляет 1 Мбайт в клиентских версиях Windows и 32 Мбайт — в серверных (данный эксперимент проводился на клиентской системе). Трассировка стека для одной из операций отложенной записи, показанной на снимке экрана, проверяет, что рабочий поток диспетчера кэша выполняет отложенную запись:



В качестве дополнительного эксперимента попробуйте повторить этот процесс с удаленным копированием с одной системы Windows на другую и копированием файлов разного размера. Будут заметны некоторые различия в поведении механизма копирования и диспетчера кэша на стороне как получателя, так и отправителя.

Отключение поздней записи для файла

Если создать временный файл, указав флаг `FILE_ATTRIBUTE_TEMPORARY` в вызове функции Windows `CreateFile`, то система поздней записи не будет записывать «грязные» страницы на диск, пока не возникнет острая нехватка физической памяти или файл не будет явно сброшен на диск. Эта особенность систем поздней записи повышает производительность системы — система поздней записи не записывает на диск данные, которые в конечном счете могут быть выброшены. Приложения обычно удаляют временные файлы вскоре после их закрытия.

Принудительная запись кэша на диск

Поскольку некоторые приложения не переносят даже кратковременных задержек между записью файла и появлением обновлений на диске, диспетчер кэша поддерживает сквозное кэширование для отдельного файлового объекта — изменения записываются на диск сразу же после их внесения. Чтобы включить сквозное кэширование, установите флаг `FILE_FLAG_WRITE_THROUGH` в вызове функции `CreateFile`. В качестве альтернативы поток может явно сбросить на диск открытый файл с помощью функции Windows `FlushFileBuffers`, когда достигает точки, в которой данные должны быть записаны на диск.

Сброс отображенных файлов на диск

Если система поздней записи должна записать данные на диск из представления, также отображенного в адресное пространство другого процесса, ситуация немного усложняется, потому что диспетчер кэша будет знать только о страницах, которые изменил сам. Страницы, измененные другим процессом, известны только этому процессу, поскольку «грязный» бит в записях таблицы страниц для измененных страниц хранится в личных таблицах страниц процессов. Чтобы решить эту проблему, диспетчер памяти сообщает диспетчеру кэша, когда пользователь отображает файл. Когда такой файл сбрасывается в кэш, например, в результате вызова функции Windows `FlushFileBuffers`, диспетчер кэша записывает «грязные» страницы в кэш, а затем проверяет, не сопоставлен ли этот файл другому процессу. Если диспетчер кэша видит, что файл отображен также другим процессом, он сбрасывает на диск все представления секции, чтобы записать страницы, которые второй процесс мог изменить. Если пользователь отображает представление файла, также открытого в кэше, то при снятии отображения измененные страницы помечаются как «грязные». Поэтому, когда поток системы поздней записи позже будет сбрасывать на диск представление, эти «грязные» страницы будут записаны на диск. Процедура работает до тех пор, пока действия выполняются в следующем порядке.

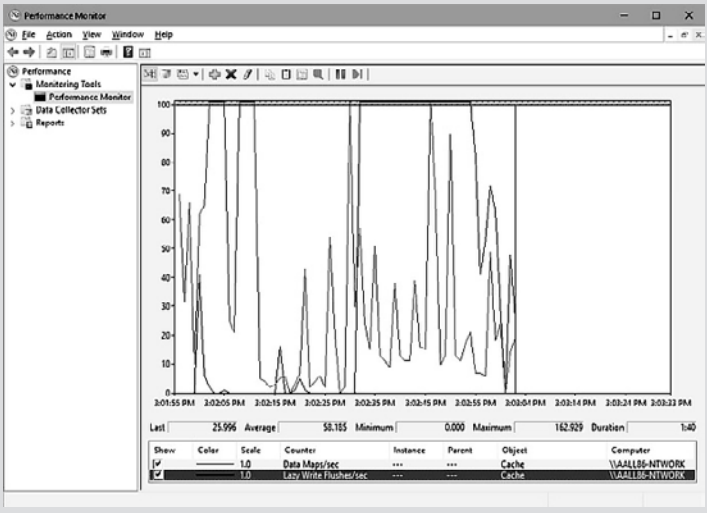
1. Пользователь снимает с отображения представление.
2. Процесс сбрасывает на диск файловые буферы.

Если эта последовательность не соблюдена, то нельзя предсказать, какие страницы будут записаны на диск.

ЭКСПЕРИМЕНТ. Наблюдение за сбросом кэша на диск

Запустив Performance Monitor и добавив счетчики Data Maps/sec и Lazy Write Flushes/sec, можно увидеть, как диспетчер кэша отображает представления в системный кэш и сбрасывает страницы на диск. Эти счетчики можно найти в группе Кэш (Cache). Затем скопируйте большой файл из одного места в другое. Более высокая строка на следующем снимке экрана показывает Data Maps/sec,

а вторая — Lazy Write Flushes/sec. Во время копирования файла параметр Lazy Write Flushes/sec значительно увеличился.



Дросселирование записи

Файловая система и диспетчер кэша должны определить, повлияет ли запрос на запись в кэш на производительность системы, а затем запланировать все отложенные записи. Сначала файловая система спрашивает у диспетчера кэша, можно ли записать определенное количество байтов прямо сейчас без ущерба для производительности, используя параметр `CcCanIWrite`, и при необходимости блокирует эту запись. Для асинхронного ввода-вывода файловая система устанавливает обратный вызов с диспетчером кэша для автоматической записи байтов, когда запись снова разрешена вызовом `CcDeferWrite`. В противном случае она просто блокируется и ожидает продолжения записи по вызову `CcCanIWrite`. Получив уведомление о предстоящей операции записи, диспетчер кэша определяет, сколько «грязных» страниц находится в кэше и сколько физической памяти доступно. Если свободных физических страниц мало, диспетчер кэша на мгновение блокирует поток файловой системы, запрашивающий запись данных в кэш. Поздняя запись диспетчера кэша сбрасывает часть «грязных» страниц на диск, а затем позволяет заблокированному потоку файловой системы продолжить работу. Такое *дросселирование записи* предотвращает снижение производительности системы из-за нехватки памяти, когда файловая система или сетевой сервер выполняют большие операции записи.

ПРИМЕЧАНИЕ Результаты дросселирования записи зависят от объема, поэтому если пользователь копирует большой файл, скажем, на твердотельный накопитель RAID-0 и одновременно переносит документ на портативный USB-накопитель, то запись на USB-диск не приведет к дросселированию записи при передаче на SSD.

Порог «грязных» страниц — это количество страниц, которое системный кэш позволяет «загрязнить» перед тем, как отключить запись в кэш. Это значение вычисляется при инициализации раздела диспетчера кэша (системный раздел создается и инициализируется на фазе 1 запуска ядра NT) и зависит от типа продукта — клиент или сервер. Как видно из предыдущих абзацев, вычисляются и два других значения — *верхний* и *нижний* порог «грязных» страниц. В зависимости от потребления памяти и скорости обработки «грязных» страниц сканер системы поздней записи вызывает внутреннюю функцию `CcAdjustThrottle`, на серверных системах выполняющую динамическую настройку текущего порога на основе вычисленных верхнего и нижнего значений. Эта регулировка выполняется для сохранения кэша чтения при высокой нагрузке записи, неизбежно приводящей к переполнению кэша и дросселированию. В табл. 11.1 перечислены алгоритмы, используемые для расчета пороговых значений «грязных» страниц.

Таблица 11.1. Алгоритмы расчета пороговых значений «грязных» страниц

Тип продукта	Порог «грязной» страницы		
	Текущий	Верхний	Нижний
Клиент	Физические страницы/8	Физические страницы/8	Физические страницы/8
Сервер	Физические страницы/2	Физические страницы/2	Физические страницы/8

Дросселирование записи также полезно для сетевых перенаправляющих устройств, передающих данные по медленным линиям связи. Предположим, что локальный процесс записывает большой объем данных в удаленную файловую систему по медленной линии 640 Кбит/с. Данные не записываются на удаленный диск до тех пор, пока поздняя запись диспетчера кэша не сбросит кэш на диск. Если в перенаправляющей системе накопилось много «грязных» страниц, которые сразу сбрасываются на диск, то получатель может получить сетевой тайм-аут до завершения передачи данных. С помощью функции `CcSetDirtyPageThreshold` диспетчер кэша позволяет сетевым перенаправителям устанавливать ограничение на количество «грязных» страниц кэша, которое они могут допустить для каждого потока, тем самым предотвращая этот сценарий. Ограничивая количество «грязных» страниц, перенаправитель гарантирует, что операция сброса кэша на диск не приведет к сетевому тайм-ауту.

Системные потоки

Как уже говорилось, диспетчер кэша выполняет операции ввода-вывода с поздней записью и упреждающим чтением, направляя запросы в общий пул потоков критических исполнителей системы. Однако он ограничивает использование этих потоков до числа на единицу меньшего, чем общее количество потоков критических исполнителей системы. В клиентских системах всего пять потоков критических исполнителей системы, в то время как в серверных системах их десять.

Внутри системы диспетчер кэша организует свои рабочие запросы в четыре списка, хотя их обслуживает один и тот же набор исполнительных рабочих потоков.

- Экспресс-очередь используется для операций упреждающего чтения.
- Обычная очередь применяется для сканирования поздней записи (для сброса на диск «грязных» данных), отложенной записи и поздних закрытий.

- Очередь быстрого уничтожения используется, когда диспетчер памяти ожидает освобождения секции данных, принадлежащей диспетчеру кэша, чтобы вместо нее открыть файл с секцией образа, что заставляет `CcWriteBehind` сбросить на диск весь файл и уничтожить общую карту кэша.
- Очередь после тиков задействуется диспетчером кэша для внутренней регистрации уведомлений после каждого тика потока системы поздней записи — другими словами, в конце каждого прохода.

Чтобы отслеживать рабочие элементы, которые должны выполнить рабочие потоки, диспетчер кэша создает собственный внутренний *ассоциативный список для каждого процессора* — список фиксированной длины структур элементов очереди рабочих потоков, по одному для каждого процессора. Ассоциативные списки рассматриваются в главе 5. Количество элементов рабочей очереди зависит от типа системы: 128 для клиентских систем и 256 для серверных. Для обеспечения межпроцессорной производительности диспетчер кэша также выделяет *глобальный ассоциативный список* тех же размеров, что и описанный ранее.

Агрессивная отложенная запись и низкоприоритетная поздняя запись

С целью повышения производительности диспетчера кэша и обеспечения совместимости с низкоскоростными дисковыми устройствами, например дисками eMMC, поздняя запись диспетчера кэша подверглась существенным улучшениям в Windows 8.1 и более поздних версиях.

Как видно из предыдущих абзацев, сканирование системы поздней записи корректирует порог «грязных» страниц, а также его верхний и нижний пределы. Многочисленные корректировки пределов выполняются путем анализа истории общего количества доступных страниц. Другие корректировки выполняются для самого порога «грязных» страниц проверкой того, смогла ли система поздней записи записать ожидаемое общее количество страниц за последний цикл выполнения — 1 раз в секунду. Если общее количество записанных страниц за последний цикл меньше ожидаемого, вычисленного процедурой `CcCalculatePagesToWrite`, значит, базовое дисковое устройство не смогло поддержать порожденную скорость ввода-вывода, поэтому порог «грязных» страниц снижается. Это означает, что выполняется большее дросселирование ввода-вывода и некоторые клиенты диспетчера кэша будут ждать при вызове API `CcCanIWrite`. В противоположном случае, когда от последнего цикла не осталось страниц, сканирование системы поздней записи может легко повысить порог. В обоих случаях он должен оставаться в диапазоне, заданном нижней и верхней границами.

Наибольшее улучшение было достигнуто благодаря рабочим потокам *Extra Write Behind*. В серверных SKU максимальное количество этих потоков — девять, что соответствует общему количеству критических системных рабочих потоков минус один, а в клиентских редакциях — всего один. Когда диспетчер кэша запрашивает системное сканирование поздней записи, система проверяет, способствуют ли «грязные» страницы увеличению нагрузки на память, по простой формуле, проверяющей, что количество «грязных» страниц меньше 1/4 порога «грязных» страниц и меньше 1/2 доступных страниц. Если это так, то процедура системного пула потоков диспетчера кэша `CcWorkerThread` использует сложный алгоритм,

определяющий, может ли она добавить еще один поток поздней записи, который будет записывать «грязные» страницы на диск параллельно с остальными.

Чтобы правильно понять, можно ли добавить еще один поток, который будет выдавать дополнительные операции ввода-вывода, не ухудшая производительность системы, диспетчер кэша рассчитывает пропускную способность диска для старых циклов поздней записи и следит за их производительностью. Если пропускная способность текущих циклов та же или лучше, чем у предыдущих, значит, диск может поддерживать общий уровень ввода-вывода, поэтому имеет смысл добавить еще один поток системы поздней записи, который в этом случае называется *экстрапоздней записью*. Если же текущая пропускная способность ниже, чем в предыдущем цикле, значит, базовый диск не в состоянии поддерживать дополнительные параллельные записи, поэтому поток экстрапоздней записи удаляется. Эта функция называется *агрессивной поздней записью*.

В клиентских редакциях Windows диспетчер кэша включает оптимизацию, предназначенную для работы с низкоскоростными дисками. Когда запрашивается сканирование поздней записи и драйверы файловой системы записывают данные в кэш, диспетчер кэша применяет алгоритм, чтобы решить, должны ли потоки поздней записи выполняться с низким приоритетом. Дополнительные сведения о приоритетах потоков см. в главе 4. Диспетчер кэша по умолчанию применяет низкий приоритет для систем поздней записи, если выполняются следующие условия (в противном случае диспетчер кэша по-прежнему использует обычный приоритет):

- вызывающий абонент не ожидает завершения текущего позднего сканирования:
- общий размер «грязных» страниц раздела менее 32 Мбайт.

Если эти два условия выполняются, то диспетчер кэша ставит рабочие элементы для систем поздней записи в очередь с низким приоритетом. Системы поздней записи запускаются рабочим потоком системы, выполняющейся с приоритетом 6 — наименьшим. Кроме того, система поздней записи устанавливает свой приоритет ввода-вывода как наименьший непосредственно перед передачей фактического ввода-вывода соответствующему драйверу файловой системы.

Динамическая память

Как видно из предыдущего абзаца, порог «грязной» страницы рассчитывается динамически на основе доступного объема физической памяти. Диспетчер кэша использует порог, чтобы решить, когда дросселировать входящие запросы на запись, а когда быть более агрессивным с поздней записью.

До появления разделов расчет производился проверкой глобального значения `MmNumberOfPhysicalPages` в процедуре `CcInitializeCacheManager`, выполнявшейся во время фазы 1 инициализации ядра. Теперь функция инициализации раздела диспетчера памяти проводит вычисления на основе доступных физических страниц памяти, принадлежащих связанному с ним разделу памяти. Более подробную информацию о разделах диспетчера кэша см. в разделе «Поддержка разделов памяти» в начале этой главы. Однако этого недостаточно, поскольку Windows поддерживает также горячее добавление физической памяти — функцию, широко использующуюся HyperV для поддержки динамической памяти для дочерних виртуальных машин.

Во время инициализации фазы 0 диспетчера памяти `MiCreatePfnDatabase` вычисляет максимально возможный размер базы данных PFN. В 64-разрядных системах диспетчер памяти предполагает, что максимально возможный объем установленной физической памяти равен всему адресуемому диапазону виртуальной памяти, например 256 Тбайт в не-LA57-системах. Система просит диспетчер памяти зарезервировать объем виртуального адресного пространства, необходимый для хранения PFN для каждой виртуальной страницы во всем адресном пространстве. Размер этой гипотетической базы данных PFN около 64 Гбайт. Затем `MiCreateSparsePfnDatabase` циклически переходит от одного допустимого диапазона физической памяти к другому, обнаруженному Winload, и заносит допустимые PFN в базу данных. База данных PFN использует разреженную память. Когда процедура `MiAddPhysicalMemory` обнаруживает новую физическую память, она создает PFN, просто выделяя новые области в базах данных PFN. Динамическая память более подробно описана в главе 9.

Диспетчер кэша должен обнаружить новую добавленную или удаленную наживую память и адаптироваться к новой конфигурации системы, иначе могут возникнуть многочисленные проблемы.

- В случаях, когда новая память была добавлена без отключения ПК, диспетчер кэша может решить, что в системе меньше памяти, поэтому его порог «грязных» страниц ниже, чем должен быть. В результате диспетчер кэша не кэширует столько «грязных» страниц, сколько должен, и поэтому дросселирует запись гораздо раньше.
- Если значительная часть доступной памяти заблокирована или больше не доступна, то выполнение кэшированного ввода-вывода в системе может уменьшить скорость ответа других приложений, у которых после удаления наживую практически не останется памяти.

Чтобы справиться с этой ситуацией, диспетчер кэша не регистрирует обратный вызов в диспетчере памяти, а реализует адаптивную коррекцию в потоке сканирования системы поздней записи (lazy writer scan, LWS). Кроме сканирования списка общей карты кэша и принятия решения о том, какую «грязную» страницу записать, поток LWS имеет возможность изменять порог «грязных» страниц в зависимости от скорости переднего плана, скорости записи и доступной памяти. LWS хранит историю среднего количества доступных физических страниц и «грязных» страниц, принадлежащих разделу. Каждую секунду поток LWS обновляет эти списки и вычисляет суммарные значения. Используя последние, LWS может реагировать на изменения объема памяти, поглощая скачки и постепенно изменяя верхнее и нижнее пороговые значения.

Учет дисковых операций ввода-вывода в диспетчере кэша

До Windows 8.1 невозможно было точно определить общий объем операций ввода-вывода, выполняемых одним процессом. Причин было несколько.

- Поздние запись и чтение не происходят в контексте процесса или потока, вызвавшего ввод-вывод. Диспетчер кэша выполняет позднюю запись данных, завершая запись в другом контексте потока, первоначально записавшего файл, обычно в системном контексте. Фактический ввод-вывод может происходить

даже после завершения процесса. Аналогичным образом диспетчер кэша может выбрать опережающее чтение, получая из файла больше данных, чем запросил процесс.

- Асинхронным вводом-выводом по-прежнему управляет диспетчер кэша, но есть случаи, когда он не задействован, например при некэшируемом вводе-выводе.
- Некоторые специализированные приложения могут выполнять низкоуровневый дисковый ввод-вывод с помощью драйвера нижнего уровня в дисковом стеке.

В хвосте IRP Windows хранит указатель на поток, выдавший запрос ввода-вывода. Этот поток не всегда тот, который запустил запрос ввода-вывода. В результате учет ввода-вывода часто ошибочно связывался с процессом System. В Windows 8.1 эта проблема решена за счет появления API `PsUpdateDiskCounters`, используемого как диспетчером кэша, так и драйверами файловой системы, которые должны тесно взаимодействовать между собой. Эта функция хранит общее количество прочитанных и записанных байтов и количество операций ввода-вывода в основной структуре данных `EPROCESS`, с помощью которой ядро NT описывает процесс. Более подробно об этом — в главе 3.

Диспетчер кэша обновляет счетчики дисков процесса, вызывая функцию `PsUpdateDiskCounters`, при выполнении кэшированных операций чтения и записи с помощью всех своих открытых интерфейсов файловой системы и во время выполнения операций ввода-вывода с упреждающим чтением через экспортируемый API `CcScheduleReadAheadEx`. Драйверы файловых систем NTFS и ReFS вызывают функцию `PsUpdateDiskCounters` при выполнении операций ввода-вывода без кэширования и с подкачкой.

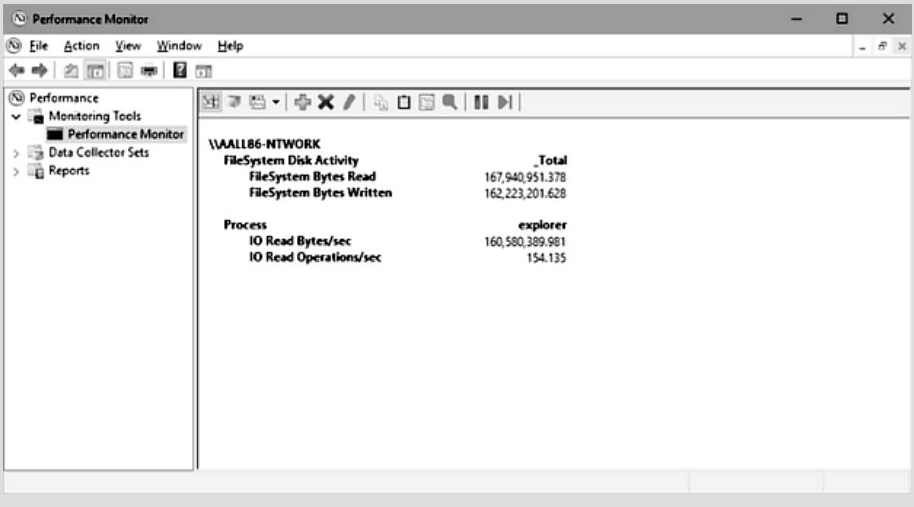
Как и `CcScheduleReadAheadEx`, многие API диспетчеров кэша были расширены, чтобы принимать указатель на поток, который запросил ввод-вывод и должен заплатить за это (`CcCopyReadEx` и `CcCopyWriteEx` — хорошие примеры). Таким образом, обновленные драйверы файловых систем могут даже контролировать, с какого потока взимать плату в случае асинхронного ввода-вывода.

Помимо счетчиков для каждого процесса, диспетчер кэша ведет также счетчик глобального дискового ввода-вывода, отслеживающий все операции ввода-вывода, выполненные файловыми системами в стеке хранения. Счетчик обновляется каждый раз, когда через драйверы файловых систем выдается некэшируемый и подкачиваемый ввод-вывод. Таким образом, этот глобальный счетчик, вычитаемый из общего объема ввода-вывода, выданного на конкретное дисковое устройство (значение, которое приложение может получить с помощью управляющего кода `IOCTL_DISK_PERFORMANCE`), представляет собой ввод-вывод, который не может быть отнесен к какому-либо конкретному процессу, — например, пейджинговый ввод-вывод, выданный измененной системой записи страниц, или ввод-вывод, выполненный внутри драйверов мини-фильтров.

Новые дисковые счетчики для каждого процесса открываются через API `NtQuerySystemInformation` с помощью информационного класса `SystemProcessInformation`. Именно этот метод используют такие средства диагностики, как Диспетчер задач или Process Explorer, для запроса точного количества операций ввода-вывода, относящихся к процессам, запущенным в системе в данный момент.

ЭКСПЕРИМЕНТ. Подсчет дисковых операций ввода-вывода

Можно увидеть точный подсчет общего количества системных операций ввода-вывода с помощью различных счетчиков, представленных в Performance Monitor. Откройте Performance Monitor и добавьте счетчики FileSystem Bytes Read и FileSystem Bytes Written, находящиеся в группе FileSystem Disk Activity. Кроме того, для этого эксперимента нужно добавить счетчики дискового ввода-вывода для каждого процесса, относящиеся к группе Process и называющиеся IO Read Bytes/sec и IO Write Bytes/sec. Когда два последних счетчика добавлены, убедитесь, что выбран процесс Проводник (Explorer) в поле Экземпляры выбранного объекта (Instances Of Selected Object).



Когда начинается копирование большого файла, видно, что значения счетчиков, принадлежащих процессам Проводник (Explorer), увеличиваются, пока не достигнут значений, показанных в глобальной файловой активности System Disk.

ФАЙЛОВЫЕ СИСТЕМЫ

В этом разделе дается обзор форматов файловых систем, поддерживаемых Windows. Затем описываются типы драйверов файловых систем и основные принципы их работы, включая взаимодействие с другими компонентами системы, такими как диспетчер памяти и диспетчер кэша. После этого подробно рассматриваются функциональность и структура данных двух наиболее важных файловых систем — NTFS и ReFS. Обсуждение начинается с анализа внутренней архитектуры этих систем, а затем сосредоточивается на их дисковой компоновке и расширенных возможностях, таких как сжатие, способность восстанавливаться, шифрование, поддержка многоуровневой структуры, создание снимков файлов и т. д.

Форматы файловой системы Windows

Windows поддерживает следующие форматы файловых систем:

- CDFS;
- UDF;
- FAT12, FAT16 и FAT32;
- exFAT;
- NTFS;
- ReFS.

Каждый из них лучше всего подходит для определенных ситуаций, как будет видно в следующих разделах.

CDFS

CDFS (%SystemRoot%\System32\Drivers\Cdfs.sys), или файловая система CD-ROM, — это драйвер файловой системы только для чтения, поддерживающий расширение формата ISO-9660 и расширение дискового формата Joliet. Формат ISO-9660 довольно прост и имеет такие ограничения, как необходимость использовать имена только в кодировке ASCII в верхнем регистре и максимальной длины 32 символа, а Joliet более гибок и поддерживает имена произвольной длины в Юникоде. Если на диске присутствуют структуры для обоих форматов для обеспечения максимальной совместимости, то CDFS использует формат Joliet. У CDFS есть несколько ограничений:

- максимальный размер файла — 4 Гбайт;
- не более 65 535 каталогов.

CDFS считается устаревшим форматом, поскольку в отрасли уже принят универсальный формат дисков UDF в качестве стандарта для оптических носителей.

UDF

Реализация файловой системы Universal Disk Format (UDF) в Windows соответствует стандарту OSTA (Optical Storage Technology Association) для UDF. UDF — это подмножество формата ISO-13346 с расширениями для таких форматов, как CD-R и DVD-R/RW. OSTA выбрала UDF в 1995 году как формат, который должен заменить ISO-9660 для магнитооптических носителей, в основном для DVD-ROM. UDF включен в спецификацию DVD и более гибок, чем CDFS. Формат файловой системы UDF обладает следующими характеристиками.

- Имена каталогов и файлов могут содержать до 254 символов ASCII или 127 символов Юникода.
- Файлы могут быть разреженными. Определение разреженных файлов приводится далее в этой главе в разделе «Сжатие и разреженные файлы».
- На размер файла выделено 64 бита.

- Есть поддержка списков управления доступом (access control list, ACL).
- Есть поддержка альтернативных потоков данных.

Драйвер UDF поддерживает версии UDF вплоть до 2.60. Формат UDF был разработан с ориентацией на перезаписываемые носители. Драйвер UDF в Windows %SystemRoot%\System32\Drivers\Udfs.sys обеспечивает поддержку чтения-записи для приводов Blu-ray, DVD-RAM, CD-R/RW и DVD±R/RW при использовании UDF 2.50 и только чтения для UDF 2.60. Однако в Windows не реализована поддержка некоторых функций UDF, таких как именованные потоки и списки управления доступом.

FAT12, FAT16 и FAT32

Windows поддерживает файловую систему FAT в основном для совместимости с другими ОС в мультизагрузочных конфигурациях, а также в качестве формата для флеш-накопителей или карт памяти. Драйвер FAT в Windows реализован в файле %SystemRoot%\System32\Drivers\Fastfat.sys.

Название каждого формата FAT включает число, показывающее количество битов, используемых данным форматом для идентификации кластеров на диске. 12-битный идентификатор кластера в FAT12 ограничивает хранение в разделе максимум 212 (4096) кластеров. Windows допускает размеры кластеров от 512 байт до 8 Кбайт, что ограничивает размер тома FAT12 значением 32 Мбайт.

ПРИМЕЧАНИЕ Все виды FAT резервируют первые два и последние 16 кластеров тома, поэтому количество доступных кластеров для тома FAT12, например, немного меньше 4096.

FAT16 с 16-битным идентификатором кластера может адресовать 216, то есть 65 536 кластеров. В Windows размер кластеров FAT16 варьируется от 512 байт (размер сектора) до 64 Кбайт на дисках с размером сектора 512 байт, что ограничивает размер тома FAT16 значением 4 Гбайт. Диски с размером сектора 4096 байт позволяют создавать кластеры размером 256 Кбайт. Используемый Windows размер кластера зависит от размера тома. Различные размеры перечислены в табл. 11.2. Если отформатировать том размером менее 16 Мбайт в формате FAT с помощью команды `format` или приложения Disk Management, то Windows применит формат FAT12 вместо FAT16.

Таблица 11.2. Размеры кластеров FAT16 по умолчанию в Windows

Размер тома	Размер кластера
< 8 Мбайт	Не поддерживается
8–32 Мбайт	512 байт
32–64 Мбайт	1 Кбайт
64–128 Мбайт	2 Кбайт
128–256 Мбайт	4 Кбайт

Продолжение ⇩

Таблица 11.2 *(продолжение)*

Размер тома	Размер кластера
256–512 Мбайт	8 Кбайт
512 –1024 Мбайт	16 Кбайт
1–2 Гбайт	32 Кбайт
2–4 Гбайт	64 Кбайт
> 16 Гбайт	Не поддерживается

Том в формате FAT разделен на несколько областей (рис. 11.14). Таблица размещения файлов, благодаря которой FAT (file allocation table) получил свое название, содержит по одной записи для каждого кластера в томе. Поскольку таблица размещения файлов необходима для успешной интерпретации содержимого тома, в FAT хранятся две копии таблицы, чтобы можно было прочитать ее из второй, если драйвер файловой системы или программа проверки целостности, например Chkdsk, не может получить доступ к одной из них, допустим, из-за поврежденного сектора диска.

Рис. 11.14. Структура формата FAT

Записи в таблице размещения файлов определяют цепочки размещения файлов (рис. 11.15) для файлов и каталогов, где звенья цепочки являются индексами для перехода к следующему кластеру данных файла. В записи для файла хранится его начальный кластер. Последняя запись цепочки размещения файла — это зарезервированное значение 0xFFFF для FAT16 и 0xFFF для FAT12. Записи FAT для неиспользуемых кластеров имеют значение 0. На рис. 11.15 видно, что файлу FILE1 назначены кластеры 2, 3 и 4, FILE2 фрагментирован и использует кластеры 5, 6 и 8, а FILE3 — только кластер 7. Чтение файла с тома FAT может потребовать чтения больших частей таблицы размещения файлов для прохода цепочек распределения файла.

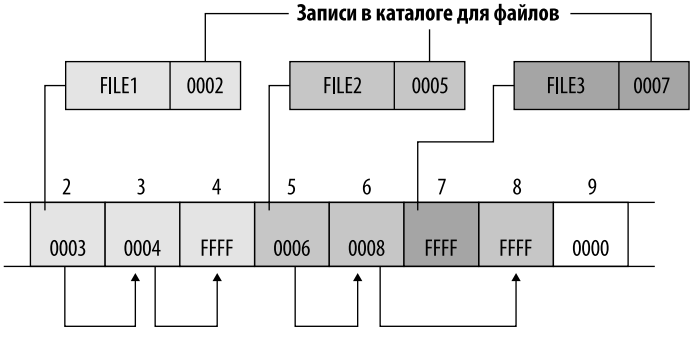


Рис. 11.15. Образец цепочки размещения файлов в FAT

В корневом каталоге томов FAT12 и FAT16 в начале тома выделяется место, достаточное для хранения 256 записей каталога, что определяет верхний предел количества файлов и каталогов, которые можно хранить в корневом каталоге. Для корневых каталогов FAT32 не существует никаких ограничений по размеру и объему. Запись в FAT занимает 32 байта и хранит имя файла, его размер, начальный кластер и информацию о временной метке — последнее обращение, создание и т. д. Если у файла имя в кодировке Юникод или не соответствует соглашению об именах, принятому в MS-DOS 8.3, то для хранения длинного имени файла выделяются дополнительные записи. Они предшествуют основной записи файла. На рис. 11.16 показан пример записи каталога для файла с именем `The quick brown fox`. Система создала представление имени `THEQUI~1.FOX` в стиле 8.3 (в записи нет точки, потому что предполагается, что она идет после восьмого символа) и использовала еще две записи каталога для хранения длинного имени файла в кодировке Юникод. Каждая строка на рисунке состоит из 16 байтов.

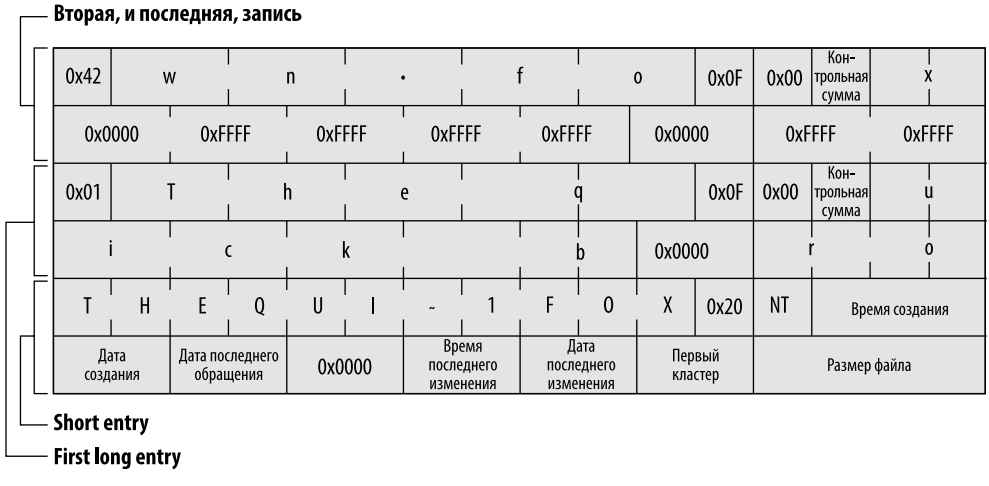


Рис. 11.16. Запись каталога FAT

FAT32 использует 32-битные идентификаторы кластеров, но резервирует старшие четыре бита, поэтому фактически имеет 28-битные идентификаторы кластеров. Поскольку размер кластера FAT32 может достигать 64 Кбайт, теоретически FAT32 может работать с 16-терабайтными томами. Хотя Windows работает с уже существующими томами FAT32 большего размера, созданными в других операционных системах, она ограничивает создаваемые в ней новые тома размером 32 Гбайт в FAT32. Потенциально большее количество кластеров в FAT32 позволяет ей управлять дисками более эффективно, чем с FAT16, — она может использовать тома объемом до 128 Гбайт с 512-байтными кластерами. В табл. 11.3 приведены размеры кластеров по умолчанию для томов FAT32.

Помимо более высокого ограничения на количество кластеров к преимуществам FAT32 по сравнению с FAT12 и FAT16 относится то, что корневой каталог FAT32 не хранится в заранее определенном месте тома и не имеет верхнего ограничения на размер, а для надежности FAT32 хранит вторую копию загрузочного сектора.

Ограничение FAT32, такое же, как и для FAT16, заключается в том, что максимальный размер файла составляет 4 Гбайт, поскольку размеры файлов хранятся как 32-битные значения.

Таблица 11.3. Размеры кластера по умолчанию для томов FAT32

Размер раздела	Размер кластера
< 32 Мбайт	Не поддерживается
32–64 Мбайт	512 байт
64–128 Мбайт	1 Кбайт
128–256 Мбайт	2 Кбайт
256 Мбайт — 8 Гбайт	4 Кбайт
8–16 Гбайт	8 Кбайт
16–32 Гбайт	16 Кбайт
> 32 Гбайт	Не поддерживается

exFAT

Разработанная компанией Microsoft файловая система Extended File Allocation Table (exFAT, также называемая FAT64) является усовершенствованием традиционной системы FAT и предназначена специально для флеш-накопителей. Основная цель exFAT — предоставить некоторые расширенные функциональные возможности, предлагаемые NTFS, без накладных расходов на структуру метаданных и ведение журнала метаданных, обеспечивающие ведение записей в режимах, не подходящих для многих устройств с флеш-носителями. В табл. 11.4 перечислены размеры кластеров по умолчанию для exFAT.

Таблица 11.4. Размеры кластеров по умолчанию для томов exFAT с секторами по 512 байт

Размер тома	Размер кластера
< 256 Мбайт	4 Кбайт
256 Мбайт — 32 Гбайт	32 Кбайт
32–512 Гбайт	128 Кбайт
512 Гбайт — 1 Тбайт	256 Кбайт
1–2 Тбайт	512 Кбайт
2–4 Тбайт	1 Мбайт
4–8 Тбайт	2 Мбайт
8–16 Тбайт	4 Мбайт
16–32 Тбайт	8 Мбайт
32–64 Тбайт	16 Мбайт
≥ 64 Тбайт	32 Мбайт

Как следует из названия FAT64, ограничение на размер файла увеличено до 264, что позволяет хранить файлы размером до 16 Эбайт (эксабайт). Соответственно

увеличен максимальный размер кластера, который в настоящее время составляет 32 Мбайт, но может достигать 2255 секторов. Также exFAT добавляет битовую карту, отслеживающую свободные кластеры, что повышает производительность операций выделения и удаления. Наконец, exFAT позволяет хранить более 1000 файлов в одном каталоге. Эти характеристики обеспечивают повышенную масштабируемость и поддержку дисков большого размера.

Кроме того, в exFAT реализованы некоторые функции, ранее доступные только в NTFS, например поддержка списков контроля доступа (access control list, ACL) и транзакций — так называемый Transaction-Safe FAT (TFAT). Хотя реализация exFAT в Windows Embedded CE включает эти функции, но в версии exFAT в Windows их нет.

ПРИМЕЧАНИЕ ReadyBoost (описывается в главе 5) может работать с флеш-накопителями, отформатированными в exFAT, чтобы поддерживать файлы кэша размером гораздо больше 4 Гбайт.

NTFS

Как отмечалось в начале главы, файловая система NTFS — это один из собственных форматов файловой системы Windows. NTFS применяет 64-разрядные номера кластеров. Это дает ей возможность адресовать тома размером до 16 эксакластеров, однако Windows ограничивает размер тома NTFS размером, адресуемым 32-битными кластерами, что составляет чуть меньше 8 Пбайт при использовании кластеров размером 2 Мбайт. В табл. 11.5 приведены размеры кластеров по умолчанию для томов NTFS. При форматировании тома NTFS значения по умолчанию можно изменить. NTFS также поддерживает $2^{32} - 1$ файлов на том. Формат NTFS позволяет создавать файлы размером 16 Эбайт, но практическая реализация ограничивает максимальный размер файла до 16 Тбайт.

Таблица 11.5. Размеры кластеров по умолчанию для томов NTFS

Размер тома	Размер кластера
< 7 Мбайт	Не поддерживается
7 Мбайт — 16 Тбайт	4 Кбайт
16–32 Тбайт	8 Кбайт
32–64 Тбайт	16 Кбайт
64–128 Тбайт	32 Кбайт
128–256 Тбайт	64 Кбайт
256–512 Тбайт	128 Кбайт
512–1024 Тбайт	256 Кбайт
1–2 Пбайт	512 Кбайт
2–4 Пбайт	1 Мбайт
4–8 Пбайт	2 Мбайт

NTFS содержит ряд дополнительных функций, таких как защита файлов и каталогов, альтернативные потоки данных, дисковые квоты, разреженные файлы,

сжатие файлов, символические (мягкие) и жесткие ссылки, поддержка транзакционной семантики, точки пересечения и шифрование. Одной из наиболее важных особенностей является *возможность восстановления*. Если система неожиданно остановлена, то метаданные тома FAT могут остаться несогласованными, что приведет к повреждению большого количества данных о файлах и каталогах. NTFS регистрирует изменения метаданных транзакционным способом, поэтому структуры файловой системы могут быть восстановлены до согласованного состояния без потери информации о структуре файлов и каталогов. Хотя файловые данные могут быть потеряны, если пользователь не задействует TxF, о котором речь пойдет далее в этой главе. Кроме того, драйвер NTFS в Windows реализует *самовосстановление* — механизм, с помощью которого выполняет большинство мелких исправлений повреждений структур файловой системы на диске во время работы Windows, не требуя перезагрузки.

ПРИМЕЧАНИЕ На момент написания книги обычный размер физического сектора дисковых устройств составляет 4 Кбайт. Даже для таких дисковых устройств стек хранения данных по соображениям совместимости предоставляет драйверам файловых систем логический размер сектора 512 байт. Вычисления, выполняемые драйвером NTFS для определения правильного размера кластера, используют логический размер сектора, а не реальный физический размер.

Начиная с Windows 10 NTFS самостоятельно поддерживает тома DAX. Они рассматриваются далее в этой главе в разделе «Тома DAX». Драйвер файловой системы NTFS поддерживает также ввод-вывод в этот тип тома с использованием больших страниц. Сопоставление файла, находящегося в томе DAX, с применением больших страниц возможно двумя способами: NTFS может автоматически выровнять файл по границе кластера размером 2 Мбайт или том может быть отформатирован с использованием кластера размером 2 Мбайт.

ReFS

Устойчивая файловая система (Resilient File System, ReFS) — еще одна система, которую Windows поддерживает самостоятельно. Она была разработана в первую очередь для больших серверов хранения данных с целью преодоления некоторых ограничений NTFS, таких как отсутствие возможности самовосстановления или ремонта тома в режиме онлайн, а также отсутствие поддержки снимков файлов. ReFS — это файловая система типа «запись в новое место», что означает: метаданные тома всегда обновляются путем записи новых данных на базовый носитель и пометки старых метаданных как удаленных. Нижний уровень файловой системы ReFS, понимающий структуру данных на диске, использует библиотеку хранения объектов под названием Minstore, предоставляющую вызывающим ее пользователям интерфейс таблицы типа «ключ — значение». Minstore похожа на современный движок базы данных, является переносимой и применяет другие структуры данных и алгоритмы по сравнению с NTFS — B^+-деревья.

Одной из важных целей разработки ReFS была поддержка огромных томов, которые могли быть созданы с помощью Storage Spaces. Как и NTFS, ReFS задействует 64-битные номера кластеров и может адресовать тома до 16 эксакластеров. В ReFS нет ограничений на размер адресуемых значений, поэтому теоретически

ReFS может управлять томами размером до 1 Йбайт при использовании кластеров размером 64 Кбайт.

В отличие от NTFS, Minstore не нуждается в центральном месте для хранения собственных метаданных в томе, хотя таблицу объектов можно считать в какой-то степени централизованной, и не имеет ограничений на адресуемые значения, поэтому нет необходимости поддерживать множество кластеров разного размера. ReFS работает только с кластерами размером 4 Кбайт и 64 Кбайт. На момент написания этой книги ReFS не поддерживает тома DAX.

Структуры данных NTFS и ReFS и их дополнительные возможности подробно описаны далее в этой главе.

Архитектура драйвера файловой системы

Драйверы файловой системы (file system drivers, FSD) управляют форматами файловой системы. Хотя FSD работают в режиме ядра, они отличаются от стандартных драйверов режима ядра по ряду признаков. Пожалуй, самое важное — они должны регистрироваться как FSD в диспетчере ввода-вывода и более активно взаимодействовать с диспетчером памяти. Для повышения производительности драйверы файловой системы обычно полагаются на услуги диспетчера кэша. Таким образом, они применяют расширенный набор экспортируемых функций Ntoskrnl.exe, которыми пользуются стандартные драйверы. Как и для стандартных драйверов режима ядра, для создания драйверов файловой системы необходимо иметь Windows Driver Kit (WDK). Дополнительные сведения о WDK см. в главе 1 и на сайте http://www.microsoft.com/whdc/devtools/wdk.

В Windows есть два типа FSD:

- *локальные FSD* управляют томами, непосредственно подключенными к компьютеру;
- *сетевые FSD* позволяют пользователям получать доступ к томам данных, подключенным к удаленным компьютерам.

Локальные FSD

К локальным FSD относятся Ntfs.sys, Refs.sys, Refsv1.sys, Fastfat.sys, Exfat.sys, Udfs.sys, Cdfs.sys и RAW FSD, встроенный в Ntoskrnl.exe. На рис. 11.17 показано упрощенное представление того, как локальные FSD взаимодействуют с диспетчером ввода-вывода и драйверами устройств хранения. Локальный FSD отвечает за регистрацию в диспетчере ввода-вывода. После регистрации FSD диспетчер может обратиться к нему для выполнения распознавания томов при первоначальном обращении приложений или системы к томам. Распознавание тома включает в себя проверку загрузочного

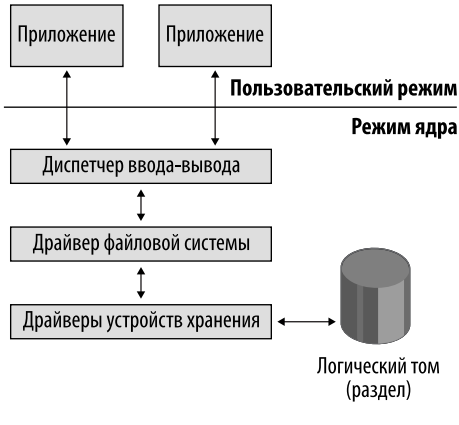


Рис. 11.17. Локальный FSD

сектора тома и нередко метаданных файловой системы в качестве проверки целостности. Если ни одна из зарегистрированных файловых систем не распознает том, система назначает ему драйвер файловой системы RAW, а затем выводит пользователю диалоговое окно с вопросом о необходимости форматирования тома. Если пользователь не форматирует том, драйвер файловой системы RAW предоставляет доступ к нему, но только на уровне секторов, то есть пользователь может читать и записывать только полные сектора.

Цель распознавания файловой системы заключается в том, чтобы дать ОС дополнительный вариант действительной, но не распознанной файловой системы, отличной от RAW. Для этого система определяет фиксированный тип структуры данных `FILE_SYSTEM_RECOGNITION_STRUCTURE`, которая записывается в первый сектор тома. Если эта структура данных присутствует, она будет распознана операционной системой, которая затем сообщит пользователю, что том содержит действительную, но не распознанную файловую систему. Система по-прежнему будет загружать файловую систему RAW для тома, но не станет предлагать пользователю отформатировать его. Пользовательское приложение или драйвер режима ядра может запросить копию `FILE_SYSTEM_RECOGNITION_STRUCTURE`, задействуя новый код управления вводом-выводом файловой системы `FSCTL_QUERY_FILE_SYSTEM_RECOGNITION`.

Первый сектор каждого формата файловой системы, поддерживаемого Windows, зарезервирован как загрузочный сектор тома. Загрузочный сектор содержит достаточно информации, чтобы локальный FSD мог идентифицировать том, в котором находится этот сектор, как содержащий формат, управляемый FSD, и найти любые другие метаданные, необходимые для определения места хранения метаданных в томе.

Когда локальный FSD (см. рис. 11.17) распознает том, он создает объект устройства, представляющий формат смонтированной файловой системы. Диспетчер ввода-вывода через *блок параметров тома* (volume parameter block, VPB) устанавливает связь между объектом устройства тома, созданным драйвером устройства хранения, и объектом устройства, созданным FSD. В результате соединения VPB диспетчер ввода-вывода перенаправляет адресованные объекту устройства тома запросы ввода-вывода объекту FSD.

Для повышения производительности локальные FSD обычно используют диспетчер кэша для кэширования данных файловой системы, включая метаданные. FSD также интегрируются с диспетчером памяти для корректной реализации отображаемых файлов. Например, FSD должны запрашивать диспетчер памяти каждый раз, когда приложение пытается обрезать файл, чтобы убедиться, что ни один процесс не отобразил часть файла за точкой обрезки. Дополнительные сведения о диспетчере памяти см. в главе 5. Windows не позволяет удалять файловые данные, отображенные приложением, ни путем обрезки, ни путем удаления файла.

Локальные FSD поддерживают также операции демонтажа файловой системы, позволяющие системе отсоединить FSD от объекта тома. Демонтаж происходит всякий раз, когда приложению требуется прямой доступ к содержимому тома на диске или когда связанный с томом носитель изменяется. При первом обращении приложения к носителю после демонтажа диспетчер ввода-вывода повторно инициирует операцию монтирования тома для носителя.

Сетевые FSD

Каждый сетевой FSD состоит из двух компонентов — клиента и сервера. Сетевой FSD на стороне клиента позволяет приложениям получать доступ к сетевым файлам и каталогам. Компонент клиентского FSD принимает запросы ввода-вывода от приложений и преобразует их в команды протокола сетевой файловой системы, например SMB, которые FSD отправляет по сети компоненту сервера, являющемуся удаленным FSD. FSD на стороне сервера прослушивает команды, поступающие по сетевому соединению, и выполняет их, выдавая запросы ввода-вывода локальному FSD, управляющему томом, в котором находится файл или каталог, для которого предназначена команда.

В состав Windows входят клиентский удаленный FSD под названием LANMan Redirector (обычно его называют просто *редиректором*) и серверный удаленный FSD под названием LANMan Server (%SystemRoot%\System32\Drivers\Srv2.sys). На рис. 11.18 показана взаимосвязь между клиентом, получающим удаленный доступ к файлам сервера через редиректор, и серверными FSD.

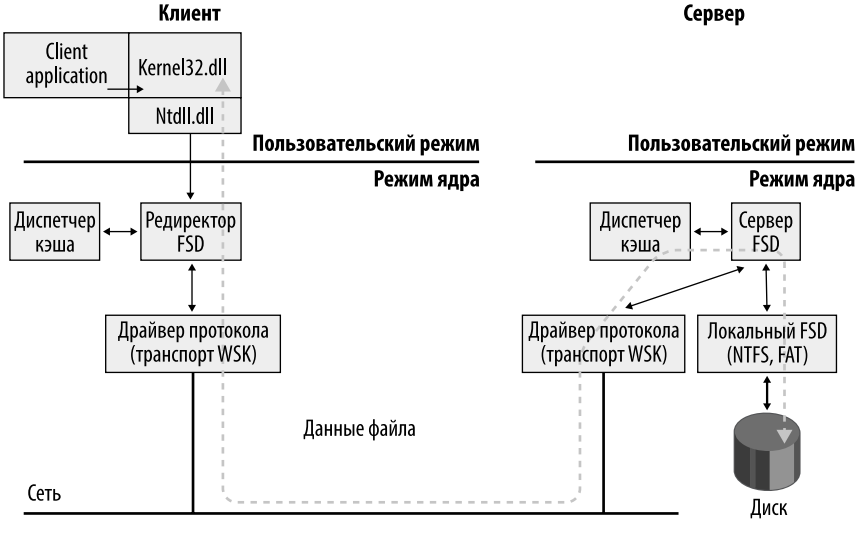


Рис. 11.18. Доступ к файлам через общую файловую интернет-систему

Windows использует протокол общей файловой интернет-системы (Common Internet File System, CIFS) для форматирования сообщений, передаваемых между редиректором и сервером. CIFS — это версия протокола Server Message Block (SMB) от Microsoft. Для получения дополнительной информации о SMB перейдите по ссылке https://docs.microsoft.com/en-us/windows/win32/fileio/microsoft-smb-protocol-and-cifs-protocol-overview.

Как и локальные FSD, клиентские сетевые FSD обычно используют службы диспетчера кэша для локального кэширования файловых данных, принадлежащих удаленным файлам и каталогам, и в таких случаях оба типа должны реализовать механизм распределенной блокировки как на клиенте, так и на сервере. Сетевые

FSD со стороны SMB-клиента реализуют протокол когерентности распределенного кэша, называемый *oplock* (оппортунистическая блокировка), чтобы данные, которые видит приложение при обращении к удаленному файлу, совпадали с данными, которые видят приложения, работающие на других компьютерах, обращающихся к тому же файлу. Сторонние файловые системы могут использовать протокол oplock или реализовать собственный протокол. Хотя сетевые FSD на стороне сервера участвуют в поддержании когерентности кэша своих клиентов, они не кэшируют данные из локальных FSD, поскольку локальные FSD кэшируют собственные данные.

В любом случае, когда ресурс может одновременно задействоваться несколькими пользователями, необходимо предусмотреть механизм сериализации для контроля записи в него, чтобы гарантировать, что только один пользователь записывает в ресурс в любой момент времени. Без этого механизма ресурс может быть поврежден. Механизмы блокировки, применяемые всеми файловыми серверами, реализующими протокол SMB, — это oplock и lease. Какой из них применяется, зависит от возможностей сервера и клиента, при этом lease — предпочтительный механизм.

Oplock. Функциональность oplock реализована в библиотеке исполнительной среды файловой системы — функции `FsRtlXxx` и может использоваться любым драйвером файловой системы. Клиент удаленного файлового сервера применяет oplock для динамического определения стратегии кэширования на стороне клиента для минимизации сетевого трафика. Драйвер файловой системы или редиректор запрашивает oplock для файла, использующегося совместно, от имени приложения, когда оно пытается открыть файл. Предоставление oplock позволяет клиенту кэшировать файл, а не отправлять каждое чтение или запись на файловый сервер по сети. Например, клиент может открыть файл для исключительного доступа, что позволит ему кэшировать все операции чтения и записи в файл, а затем копировать обновления на файловый сервер, когда файл будет закрыт. В отличие от этого, если сервер не предоставляет клиенту oplock, все операции чтения и записи должны быть отправлены на сервер.

После получения oplock клиент может начать кэшировать файл, при этом тип oplock определяет, какой тип кэширования разрешен. Oplock не обязательно остается активным до тех пор, пока клиент не закончит работу с файлом, и может быть сброшен в любой момент, если сервер получит операцию, несовместимую с предоставленными блокировками. Это означает, что клиент должен уметь быстро реагировать на сброс oplock и динамически менять свою стратегию кэширования.

До версии SMB 2.1 существовали четыре типа oplock.

- **Уровень 1, исключительный доступ.** Эта блокировка позволяет клиенту открыть файл для исключительного доступа. Клиент может выполнять буферизацию с опережающим чтением и кэширование при чтении или записи.
- **Уровень 2, общий доступ.** Эта блокировка разрешает нескольким клиентам одновременно читать файл и запрещает кому-либо писать в него. Клиент может выполнять буферизацию с опережающим чтением и кэширование данных и атрибутов файла. Запись в файл приведет к тому, что владельцы блокировки получат уведомление о том, что блокировка была сброшена.
- **Пакетный, исключительный доступ.** Эта блокировка получила свое название от блокировки, используемой при обработке пакетных файлов .bat, открывающихся и закрывающихся для обработки каждой строки в файле. Клиент может держать

файл открытым на сервере, даже если приложение закрыло его, пусть и временно. Эта блокировка поддерживает чтение и запись, а также кэширование заголовков.

- **Фильтрованный, исключительный доступ.** Эта блокировка предоставляет приложениям и фильтрам файловой системы механизм для отказа от блокировки, когда другие клиенты пытаются получить доступ к тому же файлу, но, в отличие от блокировки уровня 2, файл не может быть открыт для удаления и другой клиент не получит сообщения о нарушении совместного доступа. Эта блокировка поддерживает кэширование при чтении и записи.

Проще говоря, если несколько клиентских систем кэшируют один и тот же файл с совместным доступом на сервере, то до тех пор, пока каждое приложение, обращающееся к файлу с любого клиента или сервера, пытается только читать файл, это может выполняться из локального кэша каждой системы. Это значительно снижает сетевой трафик, поскольку содержимое файла не пересылается каждой системе с сервера. Информация о блокировке все равно должна передаваться между клиентскими системами и сервером, но для этого требуется очень низкая пропускная способность сети. Однако если хотя бы один из клиентов откроет файл для чтения и записи или исключительно записи, то ни один из клиентов не сможет использовать свои локальные кэши и все операции ввода-вывода в файл должны будут немедленно отправляться на сервер, *даже если в файл так и не будет записи*. Режимы блокировки зависят от того, как открыт файл, а не от отдельных запросов ввода-вывода.

Пример, показанный на рис. 11.19, поможет проиллюстрировать работу oplock. Сервер автоматически предоставляет oplock первого уровня первому клиенту, открывшему файл на сервере для доступа. Редиректор на клиенте кэширует данные файла как для чтения, так и для записи в файловом кэше клиентской машины. Если второй клиент открывает файл, то он тоже запрашивает oplock первого уровня. Но поскольку теперь к одному и тому же файлу обращаются два клиента, сервер должен принять меры, чтобы представить обоим согласованное представление данных файла. Если первый клиент записал данные в файл, как показано на рисунке, то сервер отзывает свой oplock и не предоставляет oplock ни одному из клиентов. Когда oplock первого клиента отзывается или *разрушается*, клиент передает все данные, которые он кэшировал для файла, обратно на сервер.

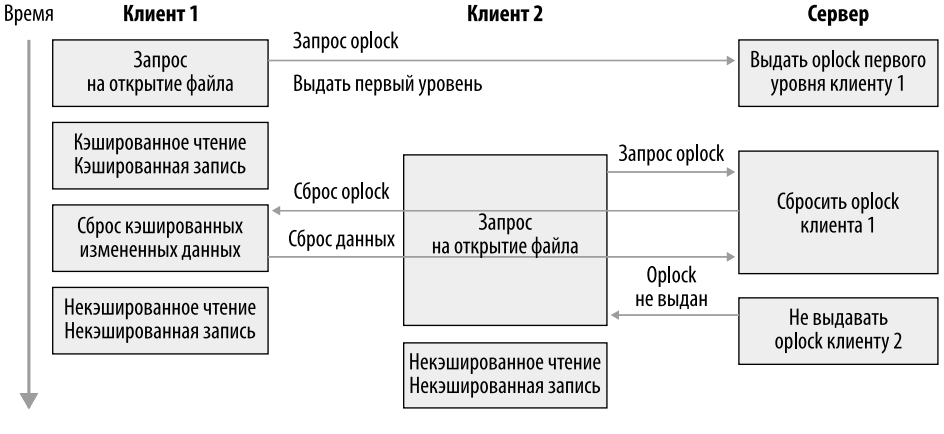
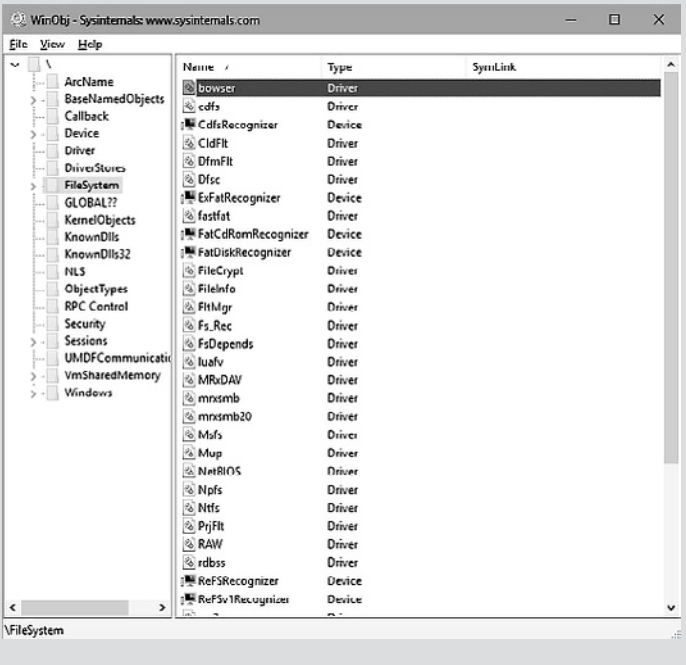


Рис. 11.19. Пример с oplock

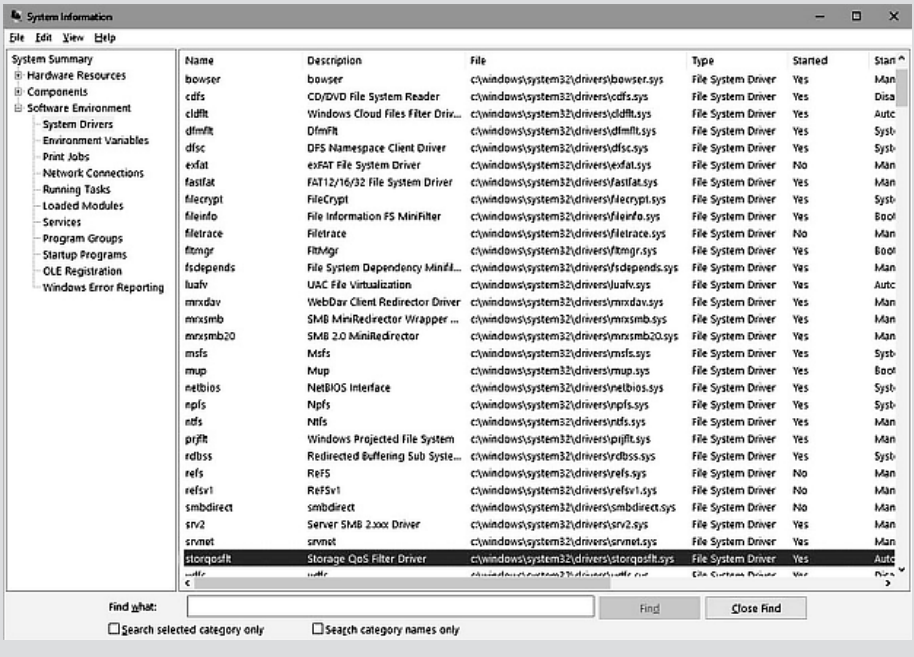
Если бы первый клиент не записывал в файл, то его oplock был бы сброшен до oplock второго уровня, являющегося тем же типом oplock, который сервер предоставил бы второму клиенту. Теперь оба клиента могут кэшировать чтение, но если кто-то из них запишет в файл, сервер отменит их oplock и начнутся операции без кэширования. После того как oplock сброшены, они больше не предоставляются для одного и того же открытого экземпляра файла. Однако если клиент закрывает файл, а затем открывает его снова, то сервер заново оценивает, какой уровень oplock ему предоставить, основываясь на том, у кого из других клиентов открыт файл и записывал ли хотя бы один из них в файл.

ЭКСПЕРИМЕНТ. Просмотр списка зарегистрированных файловых систем

Когда диспетчер ввода-вывода загружает драйвер устройства в память, он обычно называет создаваемый для представления драйвера объект так, чтобы он был помещен в каталог `\Driver` диспетчера объектов. Объекты для любого драйвера, загружаемого диспетчером ввода-вывода, у которых значение атрибута Type равно `SERVICE_FILE_SYSTEM_DRIVER` (2), помещаются диспетчером ввода-вывода в каталог `\FileSystem`. Таким образом, используя инструмент WinObj от Sysinternals, можно увидеть файловые системы, зарегистрированные в ОС, как показано на следующем снимке экрана. Обратите внимание на то, что *драйверы фильтров* файловых систем тоже будут отображаться в этом списке. Драйверы фильтров описаны далее в этом разделе.



Другой способ посмотреть зарегистрированные файловые системы — средство просмотра информации о системе. Запустите Msinfo32 из диалогового окна Выполнить (Run) меню Пуск (Start) и выберите Системные драйверы (System Drivers) в разделе Программная среда (Software Environment). Отсортируйте список драйверов, щелкнув на столбце Тип (Type), и драйверы с атрибутом Type `SERVICE_FILE_SYSTEM_DRIVER` сгруппируются.



Обратите внимание: даже если драйвер регистрируется как драйвер файловой системы, это не означает, что он является локальным или сетевым FSD. Например, NPFS (Named Pipe File System) — это драйвер, реализующий именованные каналы через частное пространство имен, подобное файловой системе. Как уже упоминалось, в этот список войдут также драйверы фильтров файловой системы.

Lease. До версии SMB 2.1 протокол SMB предполагал свободное от ошибок сетевое соединение между клиентом и сервером и не допускал разрывов сети, вызванных временными сбоями в сети, перезагрузкой сервера или отказом кластера. Когда клиент получал событие обрыва сетевого соединения, он отключал от контекста все дескрипторы, открытые на затронутом сервере, и все последующие операции ввода-вывода с этими дескрипторами были неудачными. Аналогичным образом сервер освобождал все открытые дескрипторы и ресурсы, связанные с отключенной пользовательской сессией. Такое поведение приводило к потере состояния приложений и возникновению ненужного сетевого трафика.

В SMB 2.1 появилась концепция *lease* как нового типа клиентского механизма кэширования, аналогичного oplock. Назначение lease и oplock одинаково, но lease обеспечивает бóльшую гибкость и гораздо лучшую производительность.

- **Чтение (Read, R), общий доступ.** Разрешает *нескольким* приложениям одновременно читать файл, но никому не позволяет записывать в него. Этот lease дает возможность клиенту выполнять буферизацию с упреждающим чтением и кэширование чтения.
- **Чтение — дескриптор (Read — Handle, RH), общий доступ.** Похоже на oplock второго уровня с дополнительным преимуществом: клиент может держать файл открытым на сервере, даже если требующее доступа приложение на клиенте закрыло его. Диспетчер кэша будет отложенно стирать незаписанные данные и очищать неизмененные страницы кэша в зависимости от доступности памяти. Это лучше, чем oplock второго уровня, поскольку не нужно разрывать lease между открытиями и закрытиями файлового дескриптора. В этом отношении он обеспечивает семантику, аналогичную пакетному oplock. Этот тип lease особенно полезен для файлов, которые многократно открываются и закрываются, поскольку кэш не аннулируется при закрытии файла и заполняется заново при его повторном открытии, что значительно повышает производительность сложных приложений с интенсивным вводом-выводом.
- **Чтение — запись (Read — Write, RW), исключительный доступ.** Этот lease позволяет клиенту открыть файл для исключительного доступа. Данная блокировка дает возможность клиенту выполнять буферизацию с опережающим чтением и кэширование чтения или записи.
- **Чтение — запись — дескриптор (Read — Write — Handle, RWH), исключительный доступ.** Блокировка позволяет клиенту открыть файл для исключительного доступа. Этот lease поддерживает чтение и запись, а также кэширование дескриптора аналогично lease типа «чтение — дескриптор».

Еще одно преимущество lease перед oplock заключается в том, что файл может быть кэширован, даже если на клиенте открыто несколько дескрипторов этого файла, что не редкость во многих приложениях. Это выполняется с помощью *ключа* для lease, реализованного посредством GUID, который создается клиентом и ассоциируется с блоком управления файлом (File Control Block, FCB) для кэшируемого файла, позволяя всем дескрипторам одного файла использовать одно и то же состояние lease, что обеспечивает кэширование по файлу, а не по дескрипторам. До введения lease oplock разрывался всякий раз, когда к файлу открывался новый дескриптор, даже от одного и того же клиента. На рис. 11.20 показано поведение oplock, а на рис. 11.21 — поведение нового lease.

До версии SMB 2.1 oplock можно было только предоставлять или разрывать, но lease можно также *преобразовывать*. Например, lease на чтение может быть преобразован в lease на чтение и запись, что значительно снижает сетевой трафик, поскольку кэш для конкретного файла не нужно аннулировать и пополнять, как было бы в случае разрыва oplock второго уровня, за которым следовали бы запрос и предоставление oplock первого уровня.

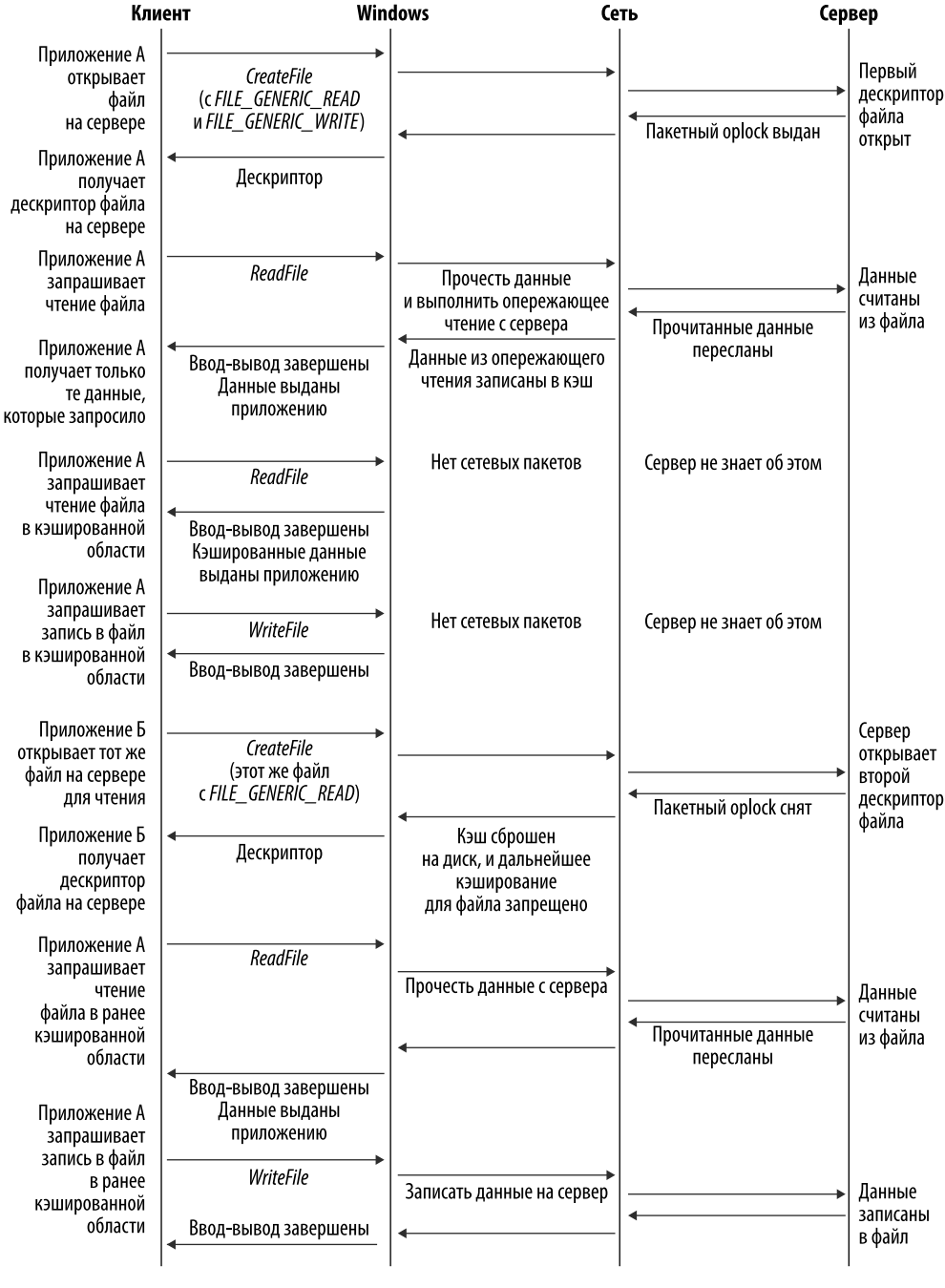


Рис. 11.20. Oplock с несколькими дескрипторами от одного клиента

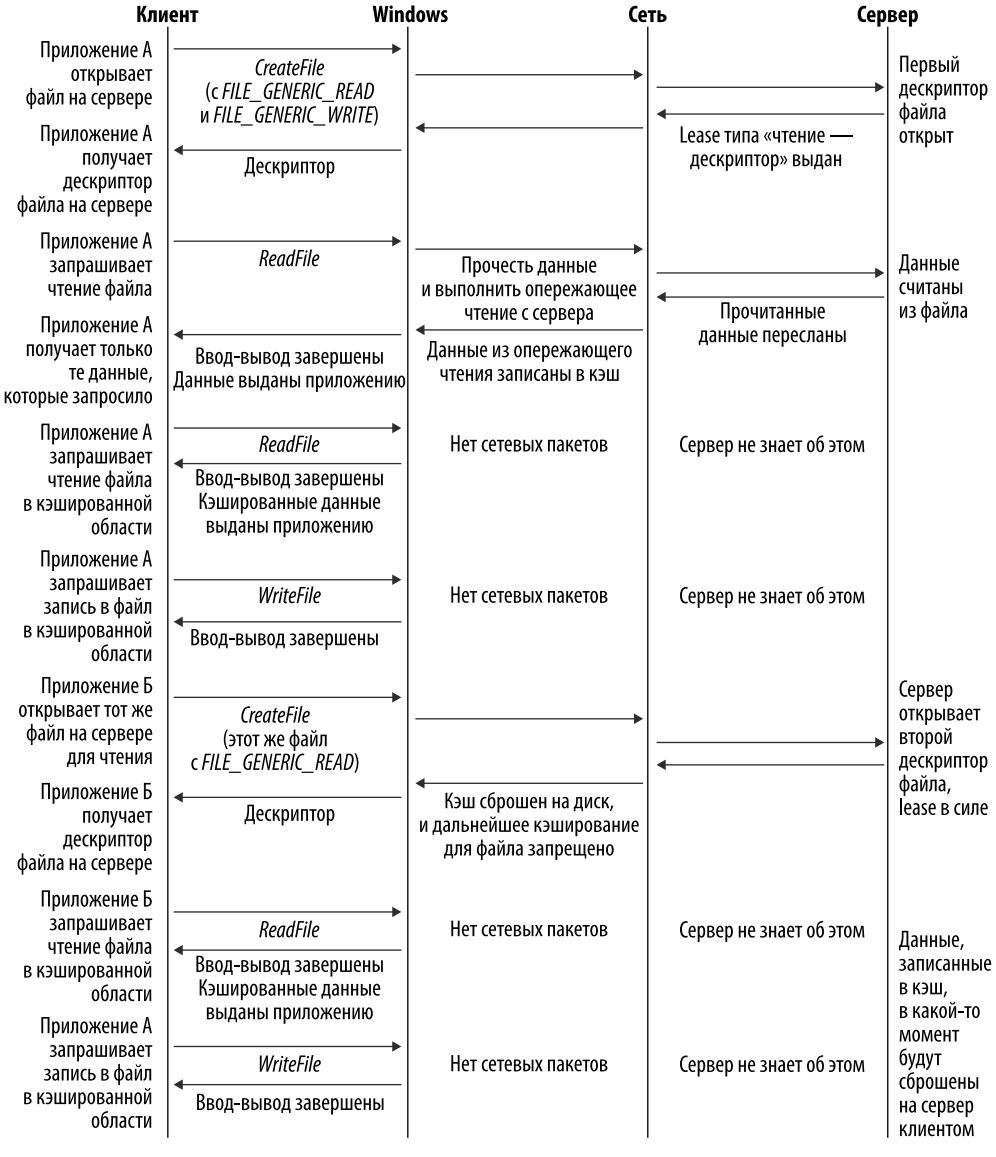


Рис. 11.21. Lease с несколькими дескрипторами от одного клиента

Операции с файловой системой

Приложения и система обращаются к файлам двумя способами: напрямую, с помощью функций файлового ввода-вывода, таких как `ReadFile` и `WriteFile`, и косвенно, путем чтения или записи части своего адресного пространства, представляющей собой отображенную часть файла. Дополнительные сведения об отображенных файлах см. в главе 5. На рис. 11.22 показана упрощенная диаграмма, на которой обозначены

компоненты, участвующие в этих операциях файловой системы, и способы их взаимодействия. Как видно на диаграмме, FSD может быть вызван несколькими способами:

- из пользовательского или системного потока, выполняющего явный файловый ввод-вывод;
- из системы записей измененных и отображенных страниц диспетчера памяти;
- косвенно из системы поздних записей диспетчера кэша;
- косвенно из потока упреждающего чтения диспетчера кэша;
- из обработчика отказов страниц диспетчера памяти.

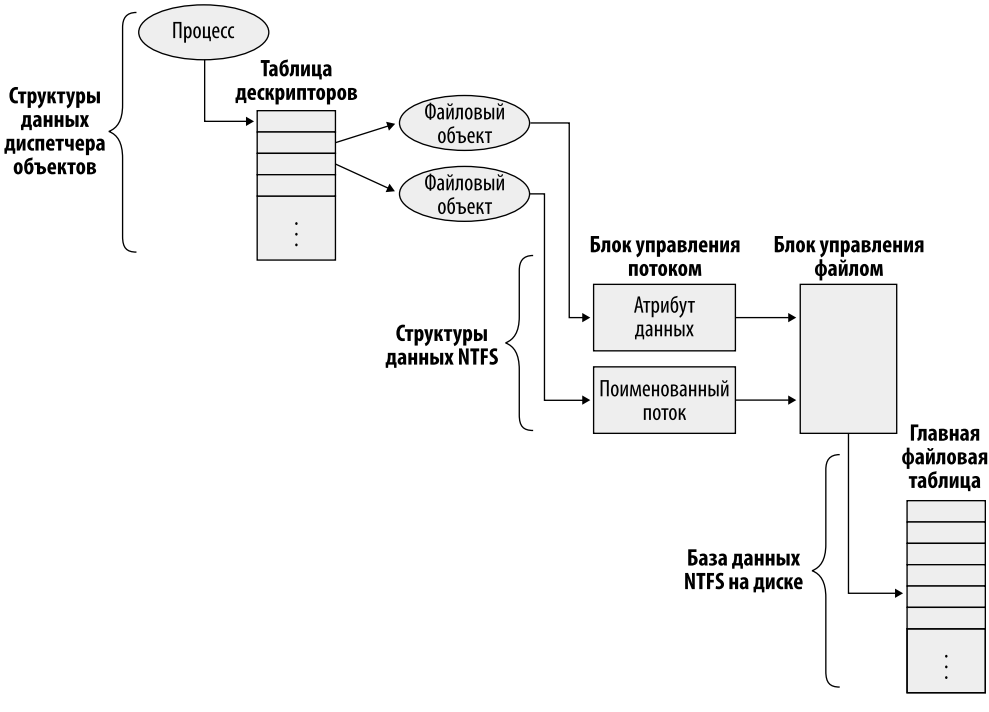


Рис. 11.22. Компоненты, участвующие во вводе-выводе файловой системы

В следующих разделах описаны обстоятельства, связанные со всеми этими сценариями, и шаги, обычно предпринимаемые FSD в ответ на каждый из них. Станет понятно, насколько сильно FSD полагаются на диспетчер памяти и диспетчер кэша.

Явный файловый ввод-вывод

Самый очевидный способ доступа приложения к файлам — это вызов функций ввода-вывода Windows, таких как `CreateFile`, `ReadFile` и `WriteFile`. Приложение открывает файл с помощью `CreateFile`, а затем читает, записывает или удаляет его, передавая дескриптор, полученный от `CreateFile`, другим функциям Windows. Функция `CreateFile`, реализованная в клиентской DLL-библиотеке Windows

Kernel32.dll, вызывает собственную функцию `NtCreateFile`, формирующую полное имя относительно корня для переданного ей приложением пути, обрабатывая символы «.» и «..» в имени, и добавляет слева `\??`, например `\??\C:\Daryl\Todo.txt`.

Системная служба `NtCreateFile` применяет для открытия файла команду `ObOpenObjectByName`, которая разбирает имя, начиная с корневого каталога диспетчера объектов и первого компонента имени пути, `??`. Глава 8 содержит подробное описание разрешения имен диспетчером объектов и использования им таблиц устройств процесса, а здесь будет рассмотрена только последовательность его действий с упором на определение букв дисков томов.

Первым шагом диспетчера объектов является преобразование `\??` в сессионный каталог пространства имен процесса, на который ссылается поле `DosDevicesDirectory` структуры таблицы устройств в объекте процесса, которое было передано от первого процесса в сеансе входа в систему с помощью поля `logon session references` в токене сеанса входа в систему. В каталоге сеанса обычно хранятся только имена томов сетевых ресурсов и буквы дисков, отображенные утилитой Subst.exe, поэтому на тех системах, где имя (в данном примере `C:`) отсутствует в сессионном каталоге, диспетчер объектов возобновляет поиск в каталоге, на который ссылается поле `GlobalDosDevicesDirectory` таблицы устройств, связанной с сессионным каталогом. Поле `GlobalDosDevicesDirectory` всегда указывает на каталог `\GLOBAL??`, в котором Windows хранит буквы дисков для локальных томов. (Дополнительные сведения см. в разделе «Пространство имен сессии» главы 8.) Процессы могут иметь собственную таблицу устройств, что является важной характеристикой при реализации по таким протоколам, как RPC.

Символическая ссылка на букву диска тома указывает на объект устройства тома в разделе `\Device`, поэтому, когда диспетчер объектов встречает объект тома, он передает остальную часть имени пути функции разбора `IopParseDevice`, которую диспетчер ввода-вывода зарегистрировал для объектов устройств. В томах на динамических дисках символическая ссылка ведет на промежуточную символическую ссылку, указывающую на объект устройства тома. На рис. 11.23 показано, как осуществляется доступ к объектам тома с помощью пространства имен диспетчера объектов. Здесь видно, как символическая ссылка `\GLOBAL??\C:` указывает на объект устройства тома `\Device\HarddiskVolume6`.

После фиксирования контекста безопасности вызывающей стороны и получения информации о безопасности из токена вызывающей стороны `IopParseDevice` создает пакет запроса ввода-вывода (I/O request packet, IRP) типа `IRP_MJ_CREATE`, затем создает файловый объект, хранящий имя открываемого файла, следует за VPB объекта устройства тома, чтобы найти смонтированный объект устройства файловой системы тома, и использует `IoCallDriver` для передачи IRP драйверу файловой системы, которому принадлежит объект устройства файловой системы.

Когда FSD получает `IRP_MJ_CREATE` IRP, он ищет указанный файл, выполняет проверку безопасности и, если файл существует и у пользователя есть разрешение на доступ к нему запрашиваемым способом, возвращает код успешного выполнения. В таблице дескрипторов процесса диспетчер объектов создает дескриптор для файлового объекта, и он распространяется обратно по цепочке вызовов, в конце концов достигая приложения в виде возвращаемого параметра из `CreateFile`. Если файловая система не справляется с операцией создания, то диспетчер ввода-вывода удаляет файловый объект, созданный для файла.

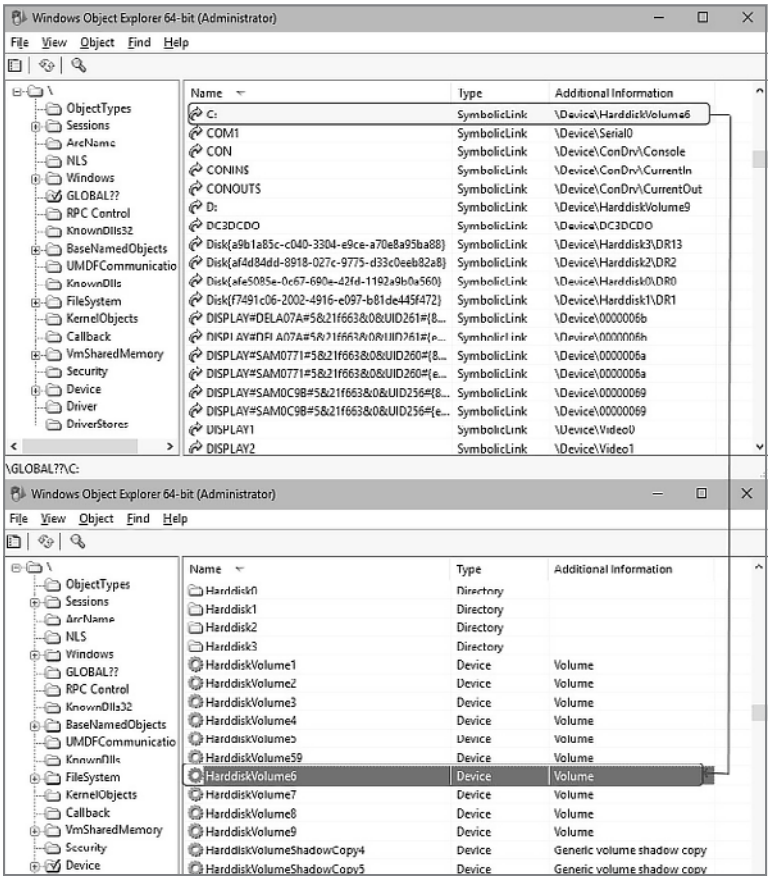


Рис. 11.23. Определение буквенных имен дисков

Здесь опущены подробности того, как FSD определяет местоположение открываемого файла в томе, но операция вызова функции `ReadFile` повторяет многие взаимодействия FSD с диспетчером кэша и драйвером хранилища. И `ReadFile`, и `CreateFile` — это системные вызовы, отображающиеся на функции диспетчера ввода-вывода, но системной службе `NtReadFile` не нужно выполнять поиск имени. Она обращается к диспетчеру объектов, чтобы преобразовать дескриптор, переданный из `ReadFile`, в указатель объекта файла. Если дескриптор указывает, что вызывающая сторона получила разрешение на чтение файла при его открытии, то `NtReadFile` создает IRP типа `IRP_MJ_READ` и отправляет его в FSD для тома, в котором находится файл. `NtReadFile` получает объект устройства FSD, хранящийся в файловом объекте, и вызывает `IoCallDriver`, а диспетчер ввода-вывода находит FSD по объекту устройства и отдает IRP в FSD.

Если читаемый файл может быть кэширован, то есть флаг `FILE_FLAG_NO_BUFFERING` не был передан `CreateFile` при открытии файла, то FSD проверяет, не было ли уже инициализировано кэширование для этого файлового объекта. Поле `PrivateCacheMap` в файловом объекте указывает на структуру данных частной карты кэша (о ней

рассказывалось в предыдущем разделе), если для файлового объекта инициализировано кэширование. Если FSD не инициализировал кэширование для файлового объекта, что происходит при первом чтении или записи файлового объекта, то поле `PrivateCacheMap` будет равно нулю. FSD вызывает функцию `CcInitializeCacheMap` диспетчера кэша для инициализации кэширования, в ходе чего диспетчер кэша создает частную карту кэша, а также общую карту кэша и объект секции, если другой файловый объект, ссылающийся на тот же файл, не инициализировал кэширование.

Убедившись, что кэширование для файла включено, FSD копирует запрошенные данные файла из виртуальной памяти диспетчера кэша в буфер, переданный потоком функции `ReadFile`. Файловая система выполняет копирование в блоке `try-except`, чтобы перехватить все ошибки, возникшие из-за недействительного буфера приложения. Файловая система использует для копирования функцию `CcCopyRead` диспетчера кэша. `CcCopyRead` принимает в качестве параметров файловый объект, смещение файла и длину.

Когда диспетчер кэша выполняет `CcCopyRead`, он получает указатель на общую карту кэша, хранящуюся в файловом объекте. Напомним, что в общей карте кэша хранятся указатели на блоки управления виртуальными адресами (virtual address control blocks, VACB) с одной записью VACB на каждый блок файла размером 256 Кбайт. Если указатель VACB для части считываемого файла равен нулю, то `CcCopyRead` выделяет VACB, резервируя представление размером 256 Кбайт в виртуальном адресном пространстве диспетчера кэша, и отображает с помощью `MmMapViewInSystemCache` указанную часть файла в это представление. Затем `CcCopyRead` просто копирует данные файла из отображенного представления в первоначально переданный ему от `ReadFile` буфер. Если данные файла не находятся в физической памяти, то операция копирования вызывает отказы страниц, обслуживающиеся функцией `MmAccessFault`.

Когда происходит отказ страницы, `MmAccessFault` проверяет виртуальный адрес, вызвавший отказ, и находит дескриптор виртуального адреса (virtual address descriptor, VAD) в дереве VAD процесса, вызвавшего отказ. (Дополнительные сведения о деревьях VAD см. в главе 5.) В данном сценарии VAD описывает отображенное представление диспетчера кэша о считываемом файле, поэтому `MmAccessFault` вызывает `MiDispatchFault` для обработки отказа страницы на действительном адресе виртуальной памяти. `MiDispatchFault` находит область управления, на которую указывает VAD, а через нее — файловый объект, представляющий открытый файл. Если файл был открыт несколько раз, то может существовать список файловых объектов, связанных указателями в их частных картах кэша.

Имея файловый объект, `MiDispatchFault` вызывает функцию диспетчера ввода-вывода `IoPageRead` для создания IRP типа `IRP_MJ_READ` и отправляет IRP в FSD, которому принадлежит объект устройства, на который указывает файловый объект. Таким образом, происходит повторный вход в файловую систему для чтения данных, которые она запросила через `CcCopyRead`, но на этот раз IRP помечен как некэшируемый и страничный ввод-вывод. Эти флаги сигнализируют FSD, что он должен получить данные файла непосредственно с диска, и он делает это, определяя, какие кластеры на диске содержат запрошенные данные (точный механизм зависит от файловой системы), и посылая IRP диспетчеру томов, владеющему объектом устройства тома, в котором находится файл. Поле блока параметров тома (volume parameter block, VPB) в объекте устройства FSD указывает на объект устройства тома.

Диспетчер памяти ждет, пока FSD завершит чтение IRP, а затем возвращает управление диспетчеру кэша, который продолжает операцию копирования, прерванную из-за ошибки страницы. Когда `CcCopyRead` завершается, FSD возвращает управление потоку, вызвавшему `NtReadFile`, скопировав запрашиваемые данные файла с помощью диспетчера кэша и диспетчера памяти в буфер потока.

Путь для `WriteFile` почти аналогичен, за исключением того, что системная служба `NtWriteFile` генерирует IRP типа `IRP_MJ_WRITE`, а FSD вызывает `CcCopyWrite` вместо `CcCopyRead`. `CcCopyWrite`, как и `CcCopyRead`, гарантирует, что записываемые части файла будут отображены в кэш, а затем копирует в кэш буфер, переданный `WriteFile`.

Если данные файла уже находятся в кэше в рабочем наборе системы, то существует несколько вариантов только что описанного сценария. Если данные файла уже хранятся в кэше, то `CcCopyRead` не приводит к ошибкам страницы. Кроме того, при определенных условиях `NtReadFile` и `NtWriteFile` вызывают точку входа быстрого ввода-вывода FSD вместо того, чтобы сразу создавать и отправлять IRP в FSD. Вот некоторые из этих условий: считываемая часть файла должна находиться в первых 4 Гбайт файла, файл не должен иметь блокировок, читаемая или записываемая часть файла не должна быть больше его текущего размера.

Точки входа быстрого ввода-вывода при чтении и записи для большинства FSD вызывают функции `CcFastCopyRead` и `CcFastCopyWrite` диспетчера кэша. Эти варианты стандартных процедур копирования гарантируют, что данные файла отображены в кэш файловой системы перед выполнением операции копирования. Если это условие не выполняется, то `CcFastCopyRead` и `CcFastCopyWrite` указывают, что быстрый ввод-вывод невозможен. Когда быстрый ввод-вывод невозможен, `NtReadFile` и `NtWriteFile` возвращаются к созданию IRP. Более полное описание быстрого ввода-вывода см. в разделе «Быстрый ввод-вывод», приведенном ранее.

Запись измененной и отображенной страницы диспетчера памяти

Потоки системы записи измененных и отображенных страниц диспетчера памяти просыпаются периодически (или когда доступная память заканчивается) для сброса измененных страниц в их основное хранилище на диске. Потоки вызывают `IoAsynchronousPageWrite`, чтобы создать IRP типа `IRP_MJ_WRITE` и записать страницы либо в файл подкачки, либо в измененный после отображения файл. Как и IRP, создаваемые `MiDispatchFault`, эти IRP помечаются как некэшируемые и выгружаемые операции ввода-вывода. Таким образом FSD обходит кэш файловой системы и выдает IRP непосредственно драйверу хранилища для записи памяти на диск.

Система поздней записи диспетчера кэша

Поток системы поздней записи диспетчера кэша также играет роль в записи измененных страниц, поскольку он периодически сбрасывает представления секций файлов, отображенных в кэше, об изменении которых ему известно. Выполняемая диспетчером кэша операция сброса, вызывая `MmFlushSection`, заставляет диспетчер памяти записывать на диск все измененные страницы в сбрасываемой части секции. Как и для записи измененных и отображенных страниц, `MmFlushSection` использует `IoSynchronousPageWrite` для отправки данных в FSD.

Поток упреждающего чтения диспетчера кэша

Кэш применяет два аспекта того, как программы ссылаются на код и данные: временную и пространственную локальность. Концепция временно́й локальности заключается в том, что если на ячейку памяти ссылаются, то, вероятно, скоро к ней снова обратятся. Суть пространственной локальности заключается в том, что если на участок памяти ссылаются, то, вероятно, скоро будут ссылаться и на близлежащие места. Таким образом, кэш обычно очень хорошо ускоряет доступ к участкам памяти, к которым обращались в недалеком прошлом, но плохо ускоряет доступ к участкам памяти, к которым еще не обращались, — у него нулевая способность предсказывать будущее. В попытке заполнить кэш данными, которые, вероятно, будут использованы в ближайшее время, диспетчер кэша реализует два механизма: поток упреждающего чтения и Superfetch.

Как говорилось в предыдущем разделе, диспетчер кэша включает в себя поток, отвечающий за попытку чтения данных из файлов до того, как приложение, драйвер или системный поток явно запросят их. Поток упреждающего чтения использует хранящуюся в частной карте кэша файлового объекта историю операций чтения, выполненных над файлом, чтобы определить, сколько данных нужно прочитать. Когда поток выполняет упреждающее чтение, он просто отображает в кэш часть файла, которую хочет прочитать, при необходимости выделяя VACB, и обращается к отображенным данным. Отказы страниц, вызванные доступом к памяти, вызывают обработчик отказа страниц, который считывает страницы в рабочий набор системы.

Ограничение потока чтения с упреждением заключается в том, что он работает только с открытыми файлами. Superfetch был введен в Windows для активного добавления файлов в кэш еще до их открытия. В частности, диспетчер памяти отправляет информацию об использовании страниц в службу Superfetch, %SystemRoot%\System32\Sysmain.dll, а мини-фильтр файловой системы предоставляет данные о разрешении имен файлов. Служба Superfetch пытается обнаружить закономерности в применении файлов, например: расчет заработной платы выполняется каждую пятницу в 12.00 или Outlook запускается каждое утро в 8.00. Когда эти закономерности выявлены, информация сохраняется в базе данных и запрашиваются таймеры. Незадолго до того момента, когда файл, вероятно, будет использован, срабатывает таймер и указывает диспетчеру памяти считать файл в низкоприоритетную память, задействуя низкоприоритетный дисковый ввод-вывод. Если файл затем открывается, данные уже находятся в памяти и нет необходимости ждать, пока они будут считаны с диска. Если файл не открыт, низкоприоритетная память будет освобождена системой. Внутреннее устройство и полное описание службы Superfetch рассматриваются в главе 5.

Обработчик отказов страниц диспетчера памяти

Ранее говорилось о том, как обработчик страничных ошибок используется в контексте явного файлового ввода-вывода и упреждающего чтения диспетчера кэша. Также он вызывается всякий раз, когда любое приложение обращается к виртуальной памяти, которая является представлением отображенного файла, и сталкивается со страницами, представляющими части файла, еще не находящиеся в памяти. Обработчик `MmAccessFault` диспетчера памяти выполняет те же

действия, что и диспетчер кэша при возникновении отказа страницы в результате `CcCopyRead` или `CcCopyWrite`, отправляя IRP через `IoPageRead` в файловую систему, в которой хранится файл.

Драйверы фильтров файловой системы и мини-фильтры

Драйвер фильтра, который накладывается на драйвер файловой системы, называется *драйвером фильтра файловой системы*. Модель ввода-вывода Windows поддерживает два следующих типа драйверов фильтров файловой системы.

- Устаревшие драйверы фильтров файловой системы обычно создают один или несколько объектов устройств и прикрепляют их к устройству файловой системы с помощью API `IoAttachDeviceToDeviceStack`. Устаревшие драйверы фильтров перехватывают все запросы, поступающие от диспетчера кэша или диспетчера ввода-вывода, и должны реализовывать как стандартные функции диспетчеризации IRP, так и путь быстрого ввода-вывода. Из-за сложности разработки такого драйвера (проблемы с синхронизацией, недокументированные интерфейсы, зависимость от исходной файловой системы и т. д.) Microsoft разработала унифицированную модель фильтрации, использующую специальные драйверы, называемые мини-фильтрами, и прекратила работать с устаревшими драйверами файловой системы. Функция API `IoAttachDeviceToDeviceStack` не работает, когда ее вызывают для томов DAX.
- Драйверы мини-фильтров являются клиентами диспетчера фильтров файловой системы Fltmgr.sys. Диспетчер фильтров файловой системы — это устаревший драйвер фильтра файловой системы, который предоставляет обширный документированный интерфейс для создания фильтров файловой системы, скрывая всю сложность взаимодействия между драйверами файловой системы и диспетчером кэша. Мини-фильтры регистрируются в диспетчере фильтров с помощью API `FltRegisterFilter`. Вызывающая сторона обычно указывает процедуру настройки экземпляра объекта и различные обратные вызовы операций. Настройка экземпляра объекта вызывается диспетчером фильтров для каждого действительного устройства тома, которым управляет файловая система. Мини-фильтр имеет возможность решить, присоединяться ли к тому. Мини-фильтры могут задавать обратные вызовы операций типов «до» и «после» для всего основного кода функции IRP, а также некоторых псевдоопераций, описывающих семантику внутреннего диспетчера памяти или диспетчера кэша и имеющих отношение к шаблонам доступа к файловой системе. Обратный вызов типа «до» выполняется перед обработкой ввода-вывода драйвером файловой системы, а обратный вызов типа «после» — по завершении операции ввода-вывода. Диспетчер фильтров также предоставляет собственные средства связи, которые могут применяться для общения драйверов мини-фильтров и связанных с ними приложений пользовательского режима.

Возможность видеть все запросы к файловой системе и по желанию изменять или завершать их позволяет задействовать целый ряд приложений, включая службы удаленной репликации файлов, шифрование файлов, эффективное резервное копирование и лицензирование. Любой продукт для защиты от вредоносного ПО обычно

включает в себя как минимум драйвер мини-фильтра, перехватывающий попытки приложений открыть или изменить файлы. Например, прежде чем передать IRP драйверу файловой системы, которому адресована команда, сканер вредоносных программ проверяет открываемый файл на предмет его чистоты. Если файл чист, то он передает IRP дальше, а если тот заражен — помещает в карантин или очищает его. Если файл не может быть очищен, драйвер отклоняет IRP, обычно с ошибкой запрета доступа, чтобы вредоносная программа не могла стать активной.

Подробное описание всей архитектуры мини-фильтров и устаревшего драйвера фильтра выходит за рамки тематики этой главы. Более подробную информацию об архитектуре устаревшего драйвера фильтров можно найти в главе 6. О мини-фильтрах можно почитать в MSDN (https://docs.microsoft.com/en-us/windows-hardware/drivers/ifs/file-system-minifilter-drivers).

Секции сканирования данных

Начиная с версии Windows 8.1 диспетчер фильтров взаимодействует с драйверами файловой системы для предоставления объектов секции сканирования данных, которые могут использоваться продуктами защиты от вредоносного программного обеспечения. Эти объекты похожи на стандартные объекты-секции (дополнительные сведения о них см. в главе 5), за исключением следующего.

- Объекты-секции сканирования данных могут быть созданы из функций обратного вызова мини-фильтра, конкретно из обратных вызовов, управляющих кодом функции `IRP_MJ_CREATE`. Эти обратные вызовы вызываются диспетчером фильтров, когда приложение открывает или создает файл. Антивирусный сканер может создать секцию сканирования данных и начать сканирование до завершения обратного вызова.
- `FltCreateSectionForDataScan`, API, используемый для создания секций сканирования данных, принимает указатель `FILE_OBJECT`. Это означает, что вызывающей стороне не нужно предоставлять файловый дескриптор. Обычно файлового дескриптора еще не существует, и его приходится повторно создавать с помощью API `FltCreateFile`, который затем создал бы другие IRP создания файла, еще раз рекурсивно взаимодействуя с фильтрами файловой системы более низкого уровня. С новым API этот процесс станет происходить гораздо быстрее, поскольку дополнительные рекурсивные вызовы не будут генерироваться.

Секция сканирования данных может быть отображена как обычная секция с помощью традиционного API. Это позволяет антивирусным приложениям реализовать свой механизм сканирования либо в пользовательском режиме, либо в драйвере режима ядра. При отображении секции сканирования данных в драйвере мини-фильтра по-прежнему генерируются события `IRP_MJ_READ`, но это не проблема, поскольку мини-фильтру вовсе не обязательно включать обратный вызов чтения.

Фильтрация именованных каналов и мейлслоты

Когда процессу, принадлежащему пользовательскому приложению, нужно связаться с другим объектом — процессом, драйвером ядра или удаленным приложением, он может воспользоваться средствами, предоставляемыми операционной системой. Чаще всего задействуются именованные каналы и мейлслоты, поскольку

они портируются и в другие операционные системы. Именованный канал — это именованный односторонний канал связи между канальным сервером и одним или несколькими канальными клиентами. Все экземпляры именованного канала используют одно и то же имя канала, но каждый экземпляр имеет собственные буферы и дескрипторы и предоставляет отдельный канал для связи между клиентом и сервером. Именованные каналы реализуются с помощью драйвера файловой системы, конкретно — драйвера NPFS Npfs.sys.

Мейлслот — это многосторонний канал связи между сервером мейлслотов и одним или несколькими клиентами. Сервер мейлслотов — это процесс, который создает мейлслот с помощью API `CreateMailslot` Win32 и может читать только небольшие сообщения (не более 424 байт) при пересылке между удаленными компьютерами, созданные одним или несколькими клиентами. Клиенты — это процессы, пишущие сообщения в мейлслот. Клиенты подключаются к мейлслоту через стандартный API `CreateFile` и отправляют сообщения посредством функции `WriteFile`. Мейлслоты обычно используются для широковещания внутри домена. Если несколько серверных процессов в домене создают мейлслот с одним и тем же именем, то каждое сообщение, адресованное ему и отправленное в домен, будет получено всеми участвующими процессами. Мейлслоты реализуются с помощью драйвера файловой системы мейлслотов Msfs.sys.

И драйвер мейлслотов, и драйвер NPFS реализуют простые файловые системы. Они управляют пространствами имен, состоящими из файлов и каталогов, которые поддерживают безопасность, могут быть открыты, закрыты, прочитаны, записаны и т. д. Описание реализации этих двух драйверов выходит за рамки тематики главы.

Начиная с Windows 8 мейлслоты и именованные каналы поддерживаются диспетчером фильтров. Мини-фильтры могут подключаться к томам мейлслотов и именованных каналов `\Device\NamedPipe` и `\Device\Mailslot`, не являющимся реальными томами, с помощью флага `FLTFL_REGISTRATION_SUPPORT_NPFS_MSFS`, указанного во время регистрации. После этого мини-фильтр может перехватывать и изменять все операции ввода-вывода именованных каналов и мейлслотов, протекающие между локальным и удаленным процессами, а также между пользовательским приложением и его драйвером ядра. Кроме того, мини-фильтры могут открывать или создавать именованный канал или мейлслот без генерации рекурсивных событий с помощью API `FltCreateNamedPipeFile` или `FltCreateMailslotFile`.

ПРИМЕЧАНИЕ Одна из причин, объясняющих, почему драйверы файловых систем именованных каналов и мейлслотов проще по сравнению с NTFS и ReFs, заключается в том, что они не слишком активно взаимодействуют с диспетчером кэша. Драйвер именованных каналов реализует путь быстрого ввода-вывода, но без поддержки кэшированного чтения или записи. Драйвер мейлслотов вообще не взаимодействует с диспетчером кэша.

Управление поведением точки повторной обработки

Файловая система NTFS поддерживает концепцию *точек повторной обработки* — блоков по 16 Кбайт прикладных и системных данных повторной обработки, которые могут быть связаны с отдельными файлами. Точки повторной обработки подробнее рассматриваются в нескольких разделах далее в этой главе. Некоторые типы точек повторной обработки, например точки монтирования тома или символические

ссылки, содержат связь между исходным файлом или пустым каталогом, используемым в качестве заглушки, и другим файлом, который может быть даже расположен в другом томе. Когда драйвер файловой системы NTFS встречает на своем пути точку повторной обработки, он возвращает код ошибки верхнему драйверу в стеке устройств. Тот — а это может быть другой драйвер фильтра — анализирует содержимое точки повторной обработки и в случае символической ссылки повторно выполняет ввод-вывод в нужное устройство тома.

Этот процесс сложен и обременителен для любого драйвера фильтра. Драйверы мини-фильтров могут перехватывать код ошибки `STATUS_REPARSE` и заново открывать точку повторной обработки с помощью нового API `FltCreateFileEx2`, принимающего список дополнительных параметров создания (Extra Create Parameters, ECP), используемых для тонкой настройки поведения процесса открытия или создания целевого файла в контексте мини-фильтра. В общем случае диспетчер фильтров поддерживает различные ECP, и каждый из них однозначно идентифицируется через GUID. Диспетчер фильтров предоставляет несколько документированных API, работающих с ECP и списками ECP. Обычно мини-фильтры выделяют ECP с помощью функции `FltAllocateExtraCreateParameter`, заполняют его и вставляют в список через `FltInsertExtraCreateParameter` перед вызовом API ввода-вывода диспетчера фильтров.

Дополнительный параметр создания `FLT_CREATEFILE_TARGET` позволяет диспетчеру фильтров автоматически управлять созданием файла, расположенного на нескольких томах, для этого вызывающая сторона должна установить флаг. Мини-фильтрам не нужно выполнять никаких других сложных операций.

С целью поддержки изоляции контейнеров можно также установить точку повторной обработки для непустых каталогов и, чтобы поддержать изоляцию контейнеров, создавать новые файлы, имеющие точки повторной обработки каталогов. Поведение файловой системы по умолчанию при встрече с точкой повторной обработки непустого каталога зависит от того, применяется ли та в последнем компоненте полного пути к файлу. Если да, то файловая система возвращает код ошибки `STATUS_REPARSE`, как в случае пустого каталога, в противном случае продолжает идти по пути.

Диспетчер фильтров может корректно работать с этим новым типом точек повторной обработки с помощью другого ECP, `TYPE_OPEN_REPARSE`. Этот ECP включает список дескрипторов, структур данных `OPEN_REPARSE_LIST_ENTRY`, каждый из которых описывает тип точки повторной обработки через тег повторной обработки и поведение, которое система должна применять, когда встречает точку повторной обработки этого типа во время разбора пути. Мини-фильтры, после того как они правильно инициализировали список дескрипторов, могут применять новое поведение следующими способами.

- С помощью функции `FltCreateFileEx2` выдать новую операцию открытия или создания файла, находящегося на пути, включающем точку повторной обработки в любом из своих компонентов. Эта процедура аналогична используемой в ECP `FLT_CREATEFILE_TARGET`.
- Глобально применить новое поведение точки повторной обработки к любому файлу, перехватываемому обратным вызовом предварительного создания. API `FltAddOpenReparseEntry` и `FltRemoveOpenReparseEntry` можно применять для

указания поведения точки повторной обработки для целевого файла до его фактического создания — обратный вызов предварительного создания перехватывает запрос на создание файла до самого создания. Драйвер мини-фильтра изоляции контейнера Wcifs.sys использует эту стратегию.

Process Monitor

Process Monitor (Procmon), утилита мониторинга активности системы из Sysinternals, которая используется на протяжении всей книги, — это пример пассивного драйвера мини-фильтра, не изменяющего поток IRP между приложениями и драйверами файловой системы.

Process Monitor работает, извлекая драйвер устройства мини-фильтра файловой системы из исполняемого образа, хранящегося как ресурс внутри Procmon.exe, при первом запуске после загрузки, устанавливая драйвер в память и затем удаляя образ драйвера с диска, если не настроен постоянный мониторинг во время загрузки. С помощью графического интерфейса Process Monitor можно указать драйверу отслеживать активность файловой системы в локальных томах, которым назначены буквы дисков, в сетевых ресурсах, именованных каналах и мейлслотах. Когда драйвер получает команду начать мониторинг тома, он регистрирует обратные вызовы фильтрации в диспетчере фильтров, прикрепляемом к объекту устройства, представляющему смонтированную файловую систему в томе. После операции присоединения диспетчер ввода-вывода перенаправляет IRP, адресованный соответствующему объекту устройства, драйверу, владеющему присоединенным устройством, в данном случае диспетчеру фильтров, который отправляет событие зарегистрированным драйверам мини-фильтров, здесь Process Monitor.

Когда драйвер Process Monitor перехватывает IRP, он записывает в буфер ядра без подкачки информацию о команде IRP, включая имя целевого файла и другие параметры, характерные для команды, например, длину и смещение чтения и записи. Каждые 500 мс графический интерфейс Process Monitor посылает IRP объекту интерфейсного устройства Process Monitor, который запрашивает копию буфера, содержащего последнюю активность, а затем показывает активность в своем окне вывода. Process Monitor показывает всю файловую активность по мере ее возникновения, что делает его идеальным инструментом для устранения неполадок в файловой системе, связанных с системными сбоями и сбоями в работе приложений. Чтобы запустить Process Monitor в первый раз в системе, учетная запись должна обладать привилегиями Load Driver и Debug. После загрузки драйвер остается резидентным, поэтому для последующих запусков требуется только привилегия Debug.

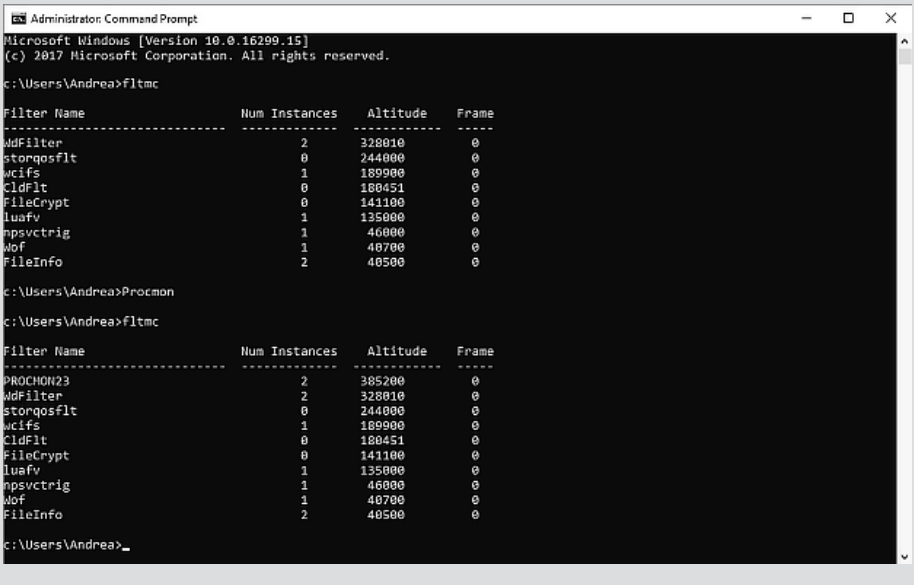
Process Monitor запускается в базовом режиме, показывающем активность файловой системы, наиболее часто полезную для устранения неполадок. В базовом режиме Process Monitor не показывает некоторые операции с файловой системой, включая:

- ввод-вывод в файлы метаданных NTFS;
- ввод-вывод в файл подкачки;
- ввод-вывод, порождаемый процессом System;
- ввод-вывод, порождаемый процессом Process Monitor.

В базовом режиме Process Monitor также сообщает об операциях файлового ввода-вывода, показывая понятные имена, а не представляющие их типы IRP. Например, обе операции, `IRP_MJ_WRITE` и `FASTIO_WRITE`, показываются как `WriteFile`, а операции `IRP_MJ_CREATE` — как Open, если они представляют собой операцию открытия, и как Create для создания новых файлов.

ЭКСПЕРИМЕНТ. Просмотр драйвера мини-фильтра Process Monitor

Чтобы узнать, какие драйверы мини-фильтров файловой системы загружены, запустите командную строку администрирования и программу управления Filter Manager (`%SystemRoot%\System32\Fltmc.exe`). Запустите Process Monitor (ProcMon.exe) и снова запустите Fltmc. Можно видеть, что драйвер фильтра Process Monitor (PROCMON20) загружен и имеет ненулевое значение в столбце Instances. Теперь выйдите из Process Monitor и снова запустите Fltmc. На этот раз можно видеть, что драйвер фильтра Process Monitor все еще загружен, но теперь количество его экземпляров равно нулю.



ФАЙЛОВАЯ СИСТЕМА NT

В этом разделе анализируется внутренняя архитектура файловой системы NT (NTFS), начиная с требований, лежавших в основе ее разработки. Будут рассмотрены структуры данных на диске, а затем расширенные возможности файловой системы, такие как поддержка восстановления, многоуровневые тома и система шифрования файлов (EFS).

Требования к файловой системе высокого класса

NTFS была разработана с учетом функций, необходимых для файловой системы корпоративного уровня. Чтобы свести к минимуму потерю данных в случае неожиданного сбоя или отказа системы, файловая система должна обеспечивать постоянную целостность метаданных, а для защиты конфиденциальных данных от несанкционированного доступа — иметь встроенную модель безопасности. Наконец, она должна обеспечивать программное резервирование данных в качестве недорогой альтернативы аппаратным решениям для защиты пользовательских данных. В этом разделе будет объяснено, как NTFS реализует каждую из этих возможностей.

Восстанавливаемость

Чтобы удовлетворить требования к надежности хранения данных и доступа к ним, NTFS обеспечивает восстановление файловой системы на основе концепции *атомарных транзакций*. Атомарные транзакции — это техника обработки изменений в базе данных, благодаря которой системные сбои не влияют на корректность или целостность базы данных. Основной принцип атомарных транзакций заключается в том, что некоторые операции с базой данных, называемые *транзакциями*, являются предложениями типа «все или ничего». *Транзакция* определяется как операция ввода-вывода, изменяющая данные файловой системы или структуру каталогов тома. Отдельные обновления диска, составляющие транзакцию, должны выполняться атомарно — если транзакция начинает выполняться, то все ее обновления диска должны быть завершены. Если системный сбой прерывает транзакцию, то завершенная часть должна быть отменена. Операция отмены возвращает базу данных в ранее известное и согласованное состояние, как если бы транзакция никогда не выполнялась.

NTFS использует атомарные транзакции для реализации функции восстановления файловой системы. Если программа начинает операцию ввода-вывода, которая изменяет структуру тома NTFS, то есть структуру каталогов, расширяет файл, выделяет место для нового файла и т. д., то NTFS рассматривает эту операцию как атомарную транзакцию. Она гарантирует, что транзакция будет либо завершена, либо отменена в случае сбоя системы при ее выполнении. Подробности того, как NTFS это делает, изложены в разделе «Поддержка восстановления NTFS» далее в этой главе. Кроме того, NTFS использует резервированное хранение для жизненно важной информации файловой системы, так что, если сектор на диске испортится, NTFS все равно сможет получить доступ к важным данным файловой системы тома.

Безопасность

Безопасность в NTFS заимствована непосредственно из объектной модели Windows. Файлы и каталоги защищены от доступа неавторизованных пользователей. (Дополнительные сведения о безопасности Windows см. в главе 7.) Открытый файл реализуется как файловый объект с дескриптором безопасности, хранящимся на диске в скрытом метафайле `$Secure`, в потоке с именем `$SDS` (Security Descriptor Stream). Перед тем как процесс сможет открыть дескриптор любого объекта, включая файловый, система безопасности Windows проверяет, что он имеет соответствующие полномочия для этого. Дескриптор безопасности в сочетании с требованием, чтобы

пользователь вошел в систему и представил идентифицирующий его пароль, гарантирует, что ни один процесс не сможет получить доступ к файлу, если ему не будет дано специальное разрешение на это системным администратором или владельцем файла. (Дополнительные сведения о дескрипторах безопасности см. в разделе «Дескрипторы безопасности и управление доступом» главы 7.)

Резервирование данных и отказоустойчивость

Помимо возможности восстановления данных файловой системы, некоторые клиенты требуют, чтобы их данные не подвергались опасности при отключении питания или катастрофическом отказе диска. Возможности восстановления NTFS гарантируют, что файловая система в томе останется доступной, но не дают гарантий полного восстановления пользовательских файлов. Защита приложений, которые не могут рисковать потерей файловых данных, обеспечивается за счет резервирования данных.

Резервирование данных для пользовательских файлов реализуется с помощью многоуровневого драйвера Windows, обеспечивающего поддержку отказоустойчивых дисков. NTFS взаимодействует с диспетчером томов, который, в свою очередь, взаимодействует с драйвером диска для записи данных на диск. Диспетчер томов может *зеркалировать* данные — дублировать их с одного диска на другой, чтобы всегда можно было получить резервную копию. Эта поддержка обычно называется *RAID уровня 1*. Диспетчеры томов также позволяют записывать данные в виде *полос* на три диска и более, используя эквивалент одного диска для хранения информации о четности. Если данные на одном диске потеряны или стали недоступными, то драйвер может восстановить содержимое диска с помощью операций исключающего «или». Такая поддержка называется *RAID уровня 5*.

В Windows 7 резервирование данных для NTFS, реализованное с помощью многоуровневого драйвера Windows, обеспечивалось Dynamic Disks. У Dynamic Disks было множество ограничений, которые были устранены в Windows 8.1 благодаря внедрению новой технологии виртуализации оборудования для хранения данных под названием Storage Spaces. Она может создавать виртуальные диски, которые обеспечивают резервирование данных и отказоустойчивость. Диспетчер томов не делает различий между виртуальным и реальным дисками, поэтому компоненты пользовательского режима не видят разницы между ними. Драйвер файловой системы NTFS сотрудничает с Storage Spaces для поддержки многоуровневых дисков и виртуальных конфигураций RAID. Storage Spaces и Spaces Direct будут рассмотрены далее в этой главе.

Дополнительные возможности NTFS

Помимо того что NTFS является восстанавливаемой, безопасной, надежной и эффективной для критически важных систем, она включает в себя следующие расширенные возможности, позволяющие ей поддерживать широкий спектр приложений. Некоторые из этих возможностей предоставляются в виде API для использования приложениями, а другие являются внутренними функциями:

- несколько потоков данных;
- имена, основанные на Юникоде;
- общее средство индексирования;
- динамическое перераспределение плохих кластеров;
- жесткие ссылки;
- символические (мягкие) ссылки и соединения;
- сжатие и разреженные файлы;
- регистрация изменений;
- пользовательские квоты для томов;
- отслеживание ссылок;
- шифрование;
- поддержка POSIX;
- дефрагментация;
- поддержка режима «только чтение» и динамическое разбиение на разделы;
- многоуровневая поддержка томов.

В следующих разделах представлен обзор этих функций.

Несколько потоков данных

В NTFS каждая единица информации, связанная с файлом: его имя, владелец, временные метки, содержимое и т. д., — реализуется как атрибут файла, атрибут объекта NTFS. Каждый атрибут состоит из одного *потока*, то есть простой последовательности байтов. Такая общая реализация позволяет легко добавлять в файл дополнительные атрибуты и, соответственно, потоки. Поскольку данные файла — это просто еще один его атрибут и можно добавлять новые атрибуты, файлы и каталоги NTFS могут содержать несколько потоков данных.

Файл NTFS имеет один поток данных по умолчанию, у которого нет имени. Приложение может создавать дополнительные именованные потоки данных и обращаться к ним, ссылаясь на их имена. Чтобы не изменять API ввода-вывода Windows, принимающие строку в качестве аргумента имени файла, имя потока данных задается добавлением двоеточия «:» к имени файла. Поскольку двоеточие — зарезервированный символ, оно может служить разделителем между именем файла и именем потока данных, как показано в этом примере:

```
myfile.dat:stream2
```

Каждый поток имеет отдельный выделенный размер (определяющий, сколько дискового пространства было зарезервировано для него), фактический размер (сколько байтов использовано вызывающей стороной) и действительную длину данных (какая часть потока была инициализирована). Кроме того, каждому потоку присваивается отдельная файловая блокировка, применяемая для блокировки диапазонов байтов и разрешения одновременного доступа.

Одним из компонентов Windows, использующих несколько потоков данных, является служба Attachment Execution Service, вызываемая всякий раз, когда стандартный Windows API для сохранения вложений из Интернета задействуют такие приложения, как Edge или Outlook. В зависимости от того, из какой *зоны* был загружен файл — например, из зоны «Мой компьютер», «Интранет» или «Недоверенной зоны», — Проводник Windows может предупредить пользователя о том, что файл был получен из недоверенного места, или даже полностью заблокировать доступ к файлу.

На рис. 11.24 показано диалоговое окно, появляющееся при запуске Process Explorer после его загрузки с сайта Sysinternals. Этот тип потока данных называется $Zone.Identifier и в просторечии именуется меткой Сети.

ПРИМЕЧАНИЕ Если снять флажок Always ask before opening this file, то поток данных идентификатора зоны будет удален из файла.

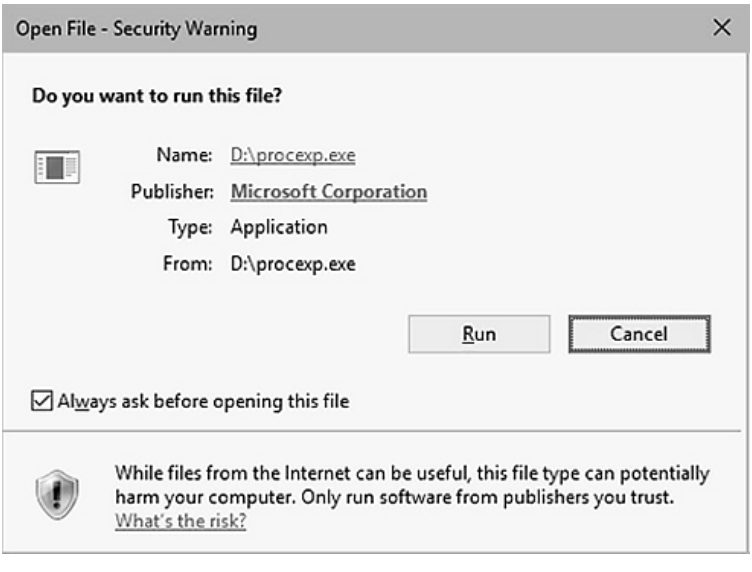


Рис. 11.24. Предупреждение о безопасности для файлов, загруженных из Интернета

Функция нескольких потоков данных может использоваться и другими приложениями. Например, утилита резервного копирования может задействовать дополнительный поток данных для хранения временных меток файлов, специфичных для резервного копирования. Или утилита архивации может реализовать иерархическое хранение, при котором файлы старше определенной даты или без доступа в течение определенного времени перемещаются в автономное хранилище. Утилита может скопировать файл в автономное хранилище, установить поток данных файла по умолчанию на 0 и добавить поток данных, определяющий место хранения файла.

ЭКСПЕРИМЕНТ. Смотрим на потоки

Большинство приложений Windows не рассчитаны на работу с альтернативными именованными потоками, но команды `echo` и `more` работают с ними. Таким образом, простой способ увидеть потоки в действии — это создать именованный поток с помощью `echo`, а затем отобразить его с помощью `more`. Такая последовательность команд создает файл с именем `test` и потоком с именем `stream`:

```
c:\Test>echo Hello from a named stream! > test:stream
c:\Test>more < test:stream
Hello from a named stream!

c:\Test>
```

Если вызывается список содержимого каталога, то размер файла `test` не отражает данные, хранящиеся в альтернативном потоке, поскольку NTFS возвращает размер только безымянного потока данных для операций запроса файла, включая перечисление содержимого каталога:

```
c:\Test>dir test
 Volume in drive C is OS.
 Volume Serial Number is F080-620F

 Directory of c:\Test

12/07/2018  05:33 PM                 0 test
               1 File(s)              0 bytes
               0 Dir(s)  18,083,577,856 bytes free
c:\Test>
```

Можно определить, какие файлы и каталоги в системе имеют альтернативные потоки данных, с помощью утилиты Streams от Sysinternals (см. следующий вывод) или ключа `/r` в команде `dir`:

```
c:\Test>streams test

streams v1.60 – Reveal NTFS alternate streams.
Copyright (C) 2005-2016 Mark Russinovich
Sysinternals – www.sysinternals.com

c:\Test\test:
          :stream:$DATA 29
```

Имена на основе Юникода

Как и Windows в целом, NTFS поддерживает 16-разрядные символы формата Юникод 1.0/UTF-16 для хранения имен файлов, каталогов и томов. Юникод позволяет уникально представлять каждый символ в любом из основных языков мира — даже эмодзи, маленькие рисунки, что помогает легко переносить данные из одной страны в другую. Юникод — это шаг вперед по сравнению с традиционным представлением международных символов, использующим двухбайтовую схему кодирования, в которой одни символы хранятся в 8 битах, а другие — в 16 битах,

что требует загрузки различных кодовых страниц для определения доступных символов. Поскольку Юникод имеет уникальное представление для каждого символа, нет зависимости от того, какая кодовая страница загружена. Каждое имя каталога и файла в пути может содержать до 255 символов, а также символы Юникода, пробелы и больше одной точки.

Общее средство индексирования

Архитектура NTFS построена таким образом, что позволяет индексировать любой атрибут файла на дисковом томе с помощью структуры B-дерева. Создание индексов по произвольным атрибутам не передается пользователям. Эта структура позволяет файловой системе эффективно находить файлы, соответствующие определенным критериям, — например, все файлы в определенном каталоге. В отличие от этого файловая система FAT индексирует имена файлов, но не сортирует их, что делает поиск в больших каталогах медленным.

Несколько функций NTFS используют преимущества общего индексирования, в том числе консолидированные дескрипторы безопасности, при которых дескрипторы безопасности файлов и каталогов тома хранятся в едином внутреннем потоке с удаленными дубликатами и индексируются с помощью внутреннего идентификатора безопасности, определяемого NTFS. (Использование индексации с помощью этих функций описано в разделе «Структура NTFS на диске» далее в этой главе.)

Динамическое перераспределение плохих кластеров

Обычно, если программа пытается считать данные из поврежденного сектора диска, операция чтения завершается неудачей и данные в выделенном кластере становятся недоступными. Но если диск отформатирован как отказоустойчивый том NTFS, то диспетчер томов Windows или Storage Spaces в зависимости от компонента, обеспечивающего резервирование данных, динамически извлекает хорошую копию данных, хранившихся в плохом секторе, и отправляет NTFS предупреждение о том, что сектор поврежден. После этого NTFS выделяет новый кластер, заменяя тот, в котором находился плохой сектор, и копирует туда данные. Она добавляет плохой кластер в список плохих кластеров на этом томе, хранящийся в скрытом файле метаданных `$BadClus`, и больше не использует его. Такое восстановление данных и динамическое перераспределение плохих кластеров особенно полезны для файловых серверов и отказоустойчивых систем или любых приложений, которые не могут позволить себе потерять данные. Если диспетчер томов или Storage Spaces не используются, когда сектор поврежден, например, на ранней стадии загрузки, то NTFS все равно заменяет кластер и не задействует его повторно, но не может восстановить данные, которые были в поврежденном секторе.

Жесткие ссылки

Жесткая ссылка позволяет нескольким путям ссылаться на один и тот же файл. Такие ссылки не поддерживаются для каталогов. Если создать жесткую ссылку с именем C:\Documents\Spec.doc на существующий файл C:\Users\Administrator\

Documents\Spec.doc, то оба пути ссылаются на один и тот же файл на диске и можно вносить изменения в файл, используя любой из них. Процессы могут создавать жесткие ссылки с помощью функции Windows `CreateHardLink`.

NTFS реализует жесткие ссылки, ведя подсчет ссылок на фактические данные, где каждый раз, когда для файла создается жесткая ссылка, на него ссылается дополнительное имя файла. Это означает, что, если есть несколько жестких ссылок на файл, можно удалить его исходное имя, ссылавшееся на данные (C:\Users\Administrator\Documents\Spec.doc в нашем примере), а остальные жесткие ссылки, такие как C:\Documents\Spec.doc, останутся и будут указывать на данные. Но поскольку жесткие ссылки — это локальные ссылки на данные на диске, представленные *номером файловой записи*, они способны существовать только в пределах одного тома и не могут работать между томами или компьютерами.

ЭКСПЕРИМЕНТ. Создание жесткой ссылки

Существуют два способа создания жесткой ссылки: команда `fsutil hardlink create` или утилита `mklink` с опцией `/H`. В этом эксперименте используется `mklink`, потому что позже она будет применяться и для создания символической ссылки. Сначала создайте файл `test.txt` и добавьте в него текст, как показано далее:

```
C:\>echo Hello from a Hard Link > test.txt
```

Теперь создайте жесткую ссылку под названием `hard.txt` следующим образом:

```
C:\>mklink hard.txt test.txt /H
Hardlink created for hard.txt <<===>> test.txt
```

Если запросить список содержимого каталога, то можно заметить, что эти два файла будут идентичны во всем — с одинаковыми датой создания, разрешениями и размером, только с разными именами:

```
c:\>dir *.txt
 Volume in drive C is OS
 Volume Serial Number is F080-620F

 Directory of c:\

12/07/2018  05:46 PM                26 hard.txt
12/07/2018  05:46 PM                26 test.txt
               2 File(s)             52 bytes
               0 Dir(s)  15,150,333,952 bytes free
```

Символические (мягкие) ссылки и соединения

Помимо жестких ссылок, NTFS поддерживает и другой тип псевдонимов имен файлов, называемый *символическими* или *мягкими ссылками*. В отличие от жестких ссылок, символические представляют собой строки, которые интерпретируются динамически и могут быть относительными или абсолютными путями, ссылающимися

на местоположение на любом устройстве хранения, включая другое локальное устройство тома или даже общий ресурс в другой системе. Это означает, что символические ссылки на самом деле не увеличивают количество ссылок на исходный файл, поэтому удаление исходного файла приведет к потере данных, а символическая ссылка, указывающая на несуществующий файл, будет оставлена. Наконец, в отличие от жестких ссылок, символические могут указывать на каталоги, а не только на файлы, что дает им дополнительное преимущество.

Например, если путь C:\Drivers является символической ссылкой на каталог, перенаправляющей на %SystemRoot%\System32\Drivers, то приложение, читающее C:\Drivers\Ntfs.sys, на самом деле читает %SystemRoot%\System\Drivers\Ntfs.sys. Символические ссылки на каталоги — это полезный способ поднять каталоги, находящиеся в глубине дерева, на более удобную глубину, не нарушая структуру и содержимое исходного дерева. В только что приведенном примере Drivers поднимается до корневого каталога тома, уменьшая глубину Ntfs.sys с трех уровней до одного, когда доступ к Ntfs.sys осуществляется через символическую ссылку. Символические ссылки на файлы работают точно так же — можно считать их ярлыками, только они реализованы в файловой системе, а не являются файлами .lnk, управляемыми Проводником Windows. Как и жесткие ссылки, символические можно создавать с помощью утилиты mklink без опции `/H` или через API `CreateSymbolicLink`.

Поскольку некоторые устаревшие приложения могут вести себя небезопасно в присутствии символических ссылок, особенно на другие машины, для создания таких ссылок требуется привилегия `SeCreateSymbolicLink`, обычно предоставляемая только администраторам. Начиная с Windows 10, и только если включен режим разработчика, пользователи API `CreateSymbolicLink` могут дополнительно установить флаг `SYMBOLIC_LINK_FLAG_ALLOW_UNPRIVILEGED_CREATE`, чтобы преодолеть данное ограничение. Это позволяет обычному пользователю по-прежнему создавать символические ссылки из окна командной строки. Файловая система также имеет опцию поведения под названием `SymLinkEvaluation`, которая может быть настроена следующей командой:

```
fsutil behavior set SymLinkEvaluation
```

По умолчанию политика разбора символических ссылок Windows разрешает только символические ссылки «локальный — локальный» и «локальный — удаленный», но не наоборот, как показано далее:

```
D:\>fsutil behavior query SymLinkEvaluation
Local to local symbolic links are enabled
Local to remote symbolic links are enabled.
Remote to local symbolic links are disabled.
Remote to Remote symbolic links are disabled.
```

Символические ссылки реализуются с помощью механизма NTFS, называемого *точками повторной обработки*. Они рассматриваются в разделе «Точки повторной обработки» далее в этой главе. Точка повторной обработки — это файл или каталог, с которым связан блок данных, называемый *данными повторной обработки*. Данные повторной обработки — это пользовательские данные о файле

или каталоге, такие как его состояние или местоположение, которые могут быть считаны из точки повторной обработки приложением, создавшим эти данные, драйвером фильтра файловой системы или диспетчером ввода-вывода. Когда NTFS встречает точку повторной обработки во время поиска файла или каталога, она возвращает код состояния `STATUS_REPARSE`, дающий сигнал подключенным к тому драйверам фильтров файловой системы и диспетчеру ввода-вывода изучить данные повторной обработки. Каждый тип точки повторной обработки имеет уникальную *метку повторной обработки*. Она позволяет компоненту, ответственному за интерпретацию данных повторной обработки, распознать точку повторной обработки, не проверяя данные. Владелец метки повторной обработки, то есть драйвер фильтра файловой системы или диспетчер ввода-вывода, может выбрать один из следующих вариантов распознавания данных повторной обработки.

- Владелец тега повторной обработки может изменить имя пути, указанное в операции ввода-вывода файла, пересекающей точку повторной обработки, и позволить операции ввода-вывода пройти с измененным именем пути. Этот подход используется в соединениях, о которых будет рассказано в дальнейшем, например для перенаправления поиска каталога.
- Владелец метки повторной обработки может удалить точку повторной обработки из файла, изменить его каким-либо образом, а затем снова выполнить операцию ввода-вывода файла.

В Windows нет функций для создания точек повторной обработки. Вместо этого процессы должны использовать управляющий код файловой системы `FSCTL_SET_REPARSE_POINT` с функцией Windows `DeviceIoControl`. Процесс может запросить содержимое точки повторной обработки с помощью управляющего кода файловой системы `FSCTL_GET_REPARSE_POINT`. Флаг `FILE_ATTRIBUTE_REPARSE_POINT` устанавливается в атрибутах файла точки повторной обработки, поэтому приложения могут проверять наличие таких точек с помощью функции Windows `GetFileAttributes`.

Другой тип точки повторной обработки, поддерживаемый NTFS, — это *соединение* (junction), или *точка монтирования тома*. Соединение — это унаследованная концепция NTFS, работающая почти так же, как символические ссылки каталогов, за исключением того, что оно может быть только локальным для тома. Нет никаких преимуществ в использовании соединения вместо символической ссылки каталога, за исключением того, что соединения совместимы со старыми версиями Windows, а символические ссылки каталогов — нет.

Как было показано в предыдущем разделе, современные версии Windows позволяют создавать точки повторной обработки, указывающие на непустые каталоги. Поведение системы, которым можно управлять с помощью драйверов мини-фильтров, зависит от положения точки повторной обработки в полном пути целевого файла. Драйверы диспетчера фильтров и файловых систем NTFS и ReFS используют доступную функцию API `FsRtlIsNonEmptyDirectoryReparsePointAllowed` для определения того, разрешен ли тип точки повторной обработки для непустых каталогов.

ЭКСПЕРИМЕНТ. Создание символической ссылки

Этот эксперимент показывает основное отличие символической ссылки от жесткой, даже если речь идет о файлах на одном томе. Создайте символическую ссылку `soft.txt`, указывающую на файл `test.txt`, созданный в предыдущем эксперименте:

```
C:\>mklink soft.txt test.txt
symbolic link created for soft.txt <<===>> test.txt
```

Если теперь посмотреть на содержимое каталога, то можно заметить, что символическая ссылка не имеет размера файла и ее тип определяется как `<SYMLINK>`. Кроме того, видно, что время создания относится к символической ссылке, а не к целевому файлу. Также символьная ссылка может иметь разрешения безопасности, отличные от разрешений целевого файла:

```
C:\>dir *.txt
 Volume in drive C is OS
 Volume Serial Number is 38D4-EA71

 Directory of C:\

05/12/2012  11:55 PM                 8 hard.txt
05/13/2012  12:28 AM    <SYMLINK>      soft.txt [test.txt]
05/12/2012  11:55 PM                 8 test.txt
               3 File(s)             16 bytes
               0 Dir(s)  10,636,480,512 bytes free
```

Наконец, если удалить оригинальный файл `test.txt`, то можно убедиться, что жесткая и символическая ссылки по-прежнему существуют, но символическая ссылка больше не указывает на действительный файл, в то время как жесткая ссылка ссылается на данные файла.

Сжатие и разреженные файлы

NTFS поддерживает сжатие файловых данных. Поскольку она выполняет процедуры сжатия и распаковки прозрачно, приложения не нужно изменять, чтобы воспользоваться этой возможностью. Каталоги также могут быть сжаты, что означает: все файлы, созданные в них впоследствии, будут сжаты.

Приложения сжимают и распаковывают файлы, передавая в `DeviceIoControl` код управления файловой системой `FSCTL_SET_COMPRESSION`. Они запрашивают состояние сжатия файла или каталога с помощью кода управления файловой системой `FSCTL_GET_COMPRESSION`. У сжатого файла или каталога в атрибутах установлен флаг `FILE_ATTRIBUTE_COMPRESSED`, поэтому приложения могут определить состояние сжатия файла или каталога с помощью `GetFileAttributes`.

Второй тип сжатия известен как *разреженные файлы*. Если файл помечен как разреженный, то NTFS не выделяет место в томе для тех его частей, которые приложение обозначает как пустые. NTFS возвращает заполненные нулями буферы, когда приложение читает из пустых областей разреженного файла. Этот тип сжатия

может быть полезен для клиент-серверных приложений, в которых реализовано протоколирование с круговым буфером, при котором сервер записывает информацию в файл, а клиенты асинхронно считывают ее. Поскольку информация, записываемая сервером, не нужна после того, как клиент ее прочитал, нет необходимости хранить эту информацию в файле. Сделав такой файл разреженным, клиент может указывать части файла, которые он читает, как пустые, освободив место в томе. Сервер может продолжать добавлять в файл новую информацию, не опасаясь, что тот разрастется и займет все свободное место в томе.

Как и в случае со сжатыми файлами, NTFS управляет разреженными файлами прозрачно. Приложения задают состояние разреженности файла, передавая код управления файловой системой `FSCTL_SET_SPARSE` в `DeviceIoControl`. Чтобы пометить некий диапазон файла как пустой, приложения используют код `FSCTL_SET_ZERO_DATA`. Они могут запросить у NTFS описание того, какие части файла являются разреженными, с помощью управляющего кода `FSCTL_QUERY_ALLOCATED_RANGES`. Одним из применений разреженных файлов является *журнал изменений* NTFS, описанный далее.

Регистрация изменений

Многие типы приложений нуждаются в мониторинге томов на предмет изменений файлов и каталогов. Например, программа автоматического резервного копирования может выполнить первоначальное полное резервное копирование, а затем делать только инкрементные резервные копии на основе изменений файлов. Очевидным способом мониторинга тома на предмет изменений будет такой: сканировать том и записать состояния файлов и каталогов, а при следующем сканировании обнаружить различия. Однако этот процесс может отрицательно сказаться на производительности системы, особенно на компьютерах с тысячами или десятками тысяч файлов.

Альтернативный подход заключается в том, что приложение регистрирует уведомление о каталоге с помощью функций Windows `FindFirstChangeNotification` или `ReadDirectoryChangesW`. В качестве входного параметра приложение указывает имя каталога, который хочет отслеживать, и функция возвращается всякий раз, когда его содержимое изменяется. Хотя этот подход более эффективен, чем сканирование томов, он требует, чтобы приложение было запущено постоянно. Использование этих функций может потребовать от приложения также сканирования каталогов, поскольку функция `FindFirstChangeNotification` не указывает, что именно изменилось, — только отмечает тот факт, что в каталоге что-то изменилось. Приложение может передать в `ReadDirectoryChangesW` буфер, который FSD заполняет записями об изменениях. Однако если он переполнится, то приложение должно быть готово вернуться к сканированию каталога.

NTFS предлагает третий подход, нивелирующий недостатки первых двух: приложение может настроить средство журнала изменений NTFS, используя код управления файловой системой `FSCTL_CREATE_USN_JOURNAL` функции `DeviceIoControl` (update sequence number (USN) — номер последовательности обновления), чтобы NTFS записывала информацию об изменениях файлов и каталогов во внутренний

файл, называемый *журналом изменений*. Журнал изменений обычно достаточно велик для того, чтобы практически гарантировать, что приложения получат шанс обработать изменения, не пропустив ни одного. Приложения используют управляющий код файловой системы `FSCTL_QUERY_USN_JOURNAL` для чтения записей из журнала изменений и могут указать, что функция `DeviceIoControl` не будет завершена до тех пор, пока не появятся новые записи.

Пользовательские квоты в томе

Системным администраторам часто требуется отслеживать или ограничивать задействование пользователем дискового пространства на общих томах хранения, поэтому в NTFS включена поддержка управления квотами. Она позволяет определять квоты для каждого пользователя, что полезно для контроля применения пороговых значений для предупреждений и ограничений и отслеживания их достижения пользователем. NTFS можно настроить на регистрацию событий в системном журнале событий, если пользователь превысил порог для предупреждения. Аналогично, если он пытается задействовать объем памяти, превышающий допустимую квоту, то NTFS может зарегистрировать событие в системном журнале событий и отменить ввод-вывод файла приложения, который привел бы к нарушению квоты, с кодом ошибки «диск заполнен».

NTFS отслеживает применение тома пользователем, полагаясь на то, что она помечает файлы и каталоги идентификатором безопасности (security ID, SID) создавшего их пользователя. (Определение SID см. в главе 7.) Логические размеры файлов и каталогов, которыми владеет пользователь, учитываются при определении квоты, установленной администратором. Таким образом, пользователь не может обойти ограничение, создав пустой разреженный файл размером больше, чем позволяет квота, а затем заполнив его ненулевыми данными. Аналогично, хотя файл размером 50 Кбайт может быть сжат до 10 Кбайт, для учета квоты используются все 50 Кбайт.

По умолчанию в томах не включено отслеживание квот. Чтобы включить квоты, указать пороги предупреждения и ограничения по умолчанию и настроить поведение NTFS для случая, когда пользователь достигает порога предупреждения или ограничения, нужно задействовать вкладку Квота (Quota) диалогового окна Свойства (Properties) тома (рис. 11.25). Инструмент Quota Entries, который можно запустить из этого диалогового окна, позволяет администратору задать различные ограничения и поведение для каждого пользователя. Приложения, желающие взаимодействовать с управлением квотами NTFS, применяют COM-интерфейсы квот, включая `IDiskQuotaControl`, `IDiskQuotaUser` и `IDiskQuotaEvents`.

Отслеживание ссылок

Ярлыки оболочки позволяют пользователям размещать в пространствах имен оболочки, например на рабочем столе, файлы, ссылающиеся на файлы, расположенные в пространстве имен файловой системы. В меню Пуск Windows широко применяются ярлыки оболочки. Аналогично ссылки для связывания и внедрения объектов (object linking and embedding, OLE) позволяют прозрачно встраивать документы одного приложения в документы других приложений. Продукты пакета Microsoft Office, включая PowerPoint, Excel и Word, используют OLE-ссылки.

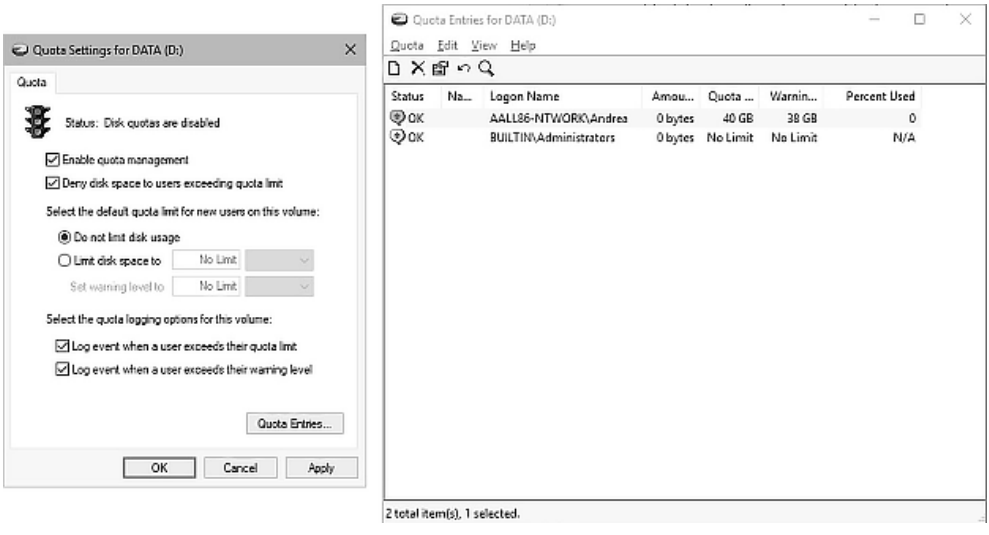


Рис. 11.25. Диалог настроек квоты, доступный из окна свойств тома

Хотя ссылки оболочки и OLE обеспечивают простой способ соединения файлов друг с другом и с пространством имен оболочки, ими может быть сложно управлять, если пользователь перемещает источник ссылки оболочки или OLE. Источник ссылки — это файл или каталог, на который указывает ссылка. NTFS в Windows включает поддержку служебного приложения, называемого *отслеживанием распределенных ссылок*, которое поддерживает целостность ссылок оболочки и OLE при перемещении целей ссылок. С помощью поддержки отслеживания ссылок NTFS служба отслеживания ссылок может прозрачно проследить за перемещением и обновить ссылку, чтобы отразить изменения, если цель ссылки, расположенная в томе NTFS, перемещается на любой другой том NTFS в домене исходного тома.

Поддержка отслеживания ссылок в NTFS основана на необязательном атрибуте файла, известном как *идентификатор объекта*. Приложение может назначить идентификатор объекта файлу с помощью управляющих кодов файловой системы `FSCTL_CREATE_OR_GET_OBJECT_ID` (назначает идентификатор, если он еще не назначен) и `FSCTL_SET_OBJECT_ID`. Идентификаторы объектов запрашиваются с помощью управляющих кодов файловой системы `FSCTL_CREATE_OR_GET_OBJECT_ID` и `FSCTL_GET_OBJECT_ID`. Управляющий код файловой системы `FSCTL_DELETE_OBJECT_ID` позволяет приложениям удалять идентификаторы объектов из файлов.

Шифрование

Корпоративные пользователи часто хранят на своих компьютерах конфиденциальную информацию. Хотя данные, хранящиеся на серверах компании, обычно надежно защищены с помощью соответствующих настроек сетевой безопасности и контроля физического доступа, данные, хранящиеся на ноутбуке, могут быть раскрыты в случае его потери или кражи. Разрешения на файлы NTFS не обеспечивают защиты, поскольку доступ к томам NTFS можно получить, не обращая внимания на ограничения системы безопасности, с помощью программ для чтения файлов

NTFS, не требующих запуска Windows. Более того, разрешения на файлы NTFS становятся бесполезными, когда альтернативная установка Windows используется для доступа к файлам из учетной записи администратора. Как объяснялось в главе 6, учетная запись администратора имеет право присвоить себе владение файлом и провести резервное копирование — и то и другое позволяет ей получить доступ к любому защищенному объекту, переопределив параметры безопасности объекта.

В NTFS есть средство под названием «система шифрования файлов» (Encrypting File System, EFS), которое пользователи могут применять для шифрования конфиденциальных данных. Работа EFS, как и сжатие файлов, полностью прозрачна для приложений, что означает: данные файла автоматически расшифровываются, когда их читает приложение, запущенное под учетной записью пользователя, имеющего право на их просмотр, и автоматически шифруются, когда авторизованное приложение их изменяет.

ПРИМЕЧАНИЕ NTFS не разрешает шифровать файлы, расположенные в корневом каталоге системного тома или каталоге \Windows, поскольку многие файлы в этих местах требуются в процессе загрузки, а в это время EFS неактивна. BitLocker — технология, гораздо лучше подходящая для сред, в которых это является обязательным условием, поскольку она поддерживает шифрование всего тома. Как будет показано далее, BitLocker работает совместно с NTFS для поддержки шифрования файлов.

EFS полагается на криптографические службы, предоставляемые Windows в пользовательском режиме, поэтому она состоит из компонента режима ядра, тесно интегрированного с NTFS, и библиотек DLL пользовательского режима, взаимодействующих с сервисом проверки подлинности локальной системы безопасности (Local Security Authority Subsystem Service, LSASS) и криптографическими библиотеками DLL.

Получить доступ к зашифрованным файлам можно только с помощью закрытого ключа из пары закрытых и открытых ключей EFS учетной записи, а эти ключи блокируются с помощью пароля учетной записи. Таким образом, доступ к зашифрованным с помощью EFS файлам на потерянных или украденных ноутбуках невозможно получить никакими средствами (кроме криптографической атаки методом перебора) без пароля учетной записи, имеющей право на просмотр данных.

Приложения могут использовать функции `EncryptFile` и `DecryptFile` Windows API для шифрования и расшифровки файлов, а также `FileEncryptionStatus` для получения атрибутов файла или каталога, связанных с EFS, — например, зашифрован ли файл или каталог. У зашифрованного файла или каталога в атрибутах установлен флаг `FILE_ATTRIBUTE_ENCRYPTED`, поэтому приложения могут также определить состояние шифрования файла или каталога с помощью `GetFileAttributes`.

Семантика удаления в стиле POSIX

Подсистема POSIX устарела и уже недоступна в операционной системе Windows. Подсистема Windows для Linux (Windows Subsystem for Linux, WSL) заменила оригинальную подсистему POSIX. Драйвер файловой системы NTFS был обновлен

для унификации различий между операциями ввода-вывода, поддерживаемыми в Windows и Linux. Одно из этих различий обеспечивается командами Linux `unlink` или `rm` для удаления файла или папки. В Windows приложение не может удалить файл, используемый другим приложением, владеющим открытым дескриптором для него. В Linux, наоборот, это обычно поддерживается — другие процессы продолжают нормально работать с исходным удаленным файлом. Для поддержки WSL драйвер файловой системы NTFS в Windows 10 поддерживает новую операцию POSIX Delete — удаление в стиле POSIX.

Функция API Win32 `DeleteFile` реализует стандартное удаление файлов. Целевой файл открывается, то есть создается новый дескриптор, а затем к файлу прикрепляется метка удаления через собственный API `NtSetInformationFile`. Метка просто сообщает драйверу файловой системы NTFS, что файл будет удален. Драйвер файловой системы проверяет, равно ли единице количество ссылок на FCB (File Control Block), что означает: для файла нет других открытых дескрипторов. Если оно равно, то драйвер файловой системы помечает файл как удаленный при закрытии и возвращается. Только когда дескриптор для файла закрыт, процедура удаления `IRP_MJ_CLEANUP` физически удаляет файл с базового носителя.

Подобная архитектура несовместима с командой Linux `unlink`. Подсистема WSL, когда ей нужно стереть файл, использует удаление в стиле POSIX. Она вызывает собственный API `NtSetInformationFile` с новым информационным классом `FileDispositionInformationEx`, указывая флаг `FILE_DISPOSITION_POSIX_SEMANTICS`. Драйвер файловой системы NTFS помечает файл как POSIX-удаленный, вставляя флаг в свой блок управления контекстом CCB (Context Control Block) — структуры данных, представляющей контекст открытого экземпляра объекта на диске. Затем он снова открывает файл с помощью специальной внутренней процедуры и прикрепляет новый дескриптор, который здесь будет называться PosixDeleted-дескриптором, к блоку управления потоком (Stream Control Block, SCB). Когда оригинальный дескриптор закрыт, драйвер файловой системы NTFS обнаруживает наличие PosixDeleted-дескриптора и ставит в очередь рабочий элемент для его закрытия. Когда рабочий элемент завершается, процедура очистки обнаруживает, что дескриптор помечен как POSIX-удаленный, и физически перемещает файл в скрытый каталог `\$Extend\$Deleted`. Другие приложения все еще могут работать с исходным файлом, который больше не находится в исходном пространстве имен и будет удален только после закрытия последнего дескриптора файла (первый запрос на удаление пометил FCB как удаляемый при закрытии).

Если система не может удалить целевой файл по какой-то необычной причине — из-за зависшей ссылки в неисправном драйвере ядра или внезапного прерывания питания, то в следующий раз, когда файловая система NTFS получит возможность смонтировать том, она проверит каталог `\$Extend\$Deleted` и удалит все входящие в него файлы с помощью стандартных процедур удаления файлов.

ПРИМЕЧАНИЕ Начиная с обновления от мая 2019 года (19H1) Windows 10 использует POSIX-удаление в качестве метода удаления файлов по умолчанию. Это означает, что API DeleteFile применяет новое поведение.

ЭКСПЕРИМЕНТ. Наблюдение за POSIX-удалением

В этом эксперименте можно будет наблюдать POSIX-удаление с помощью приложения FsTool, доступного в загружаемых ресурсах книги. Убедитесь, что используется копия Windows Server 2019 (RS5) — в новых клиентских версиях Windows реализуют POSIX-удаление по умолчанию. Начните с открытия окна командной строки. Задействуйте аргумент командной строки `/touch` при запуске FsTool, чтобы создать txt-файл, который будет применяться исключительно приложением:

```
D:\>FsTool.exe /touch d:\Test.txt
NTFS / ReFS Tool v0.1
Copyright (C) 2018 Andrea Allievi (AaL186)

Touching "d:\Test.txt" file... Success.
   The File handle is valid... Press Enter to write to the file.
```

В ответ на запрос вместо нажатия клавиши Enter откройте другое окно командной строки и попробуйте открыть и удалить файл:

```
D:\>type Test.txt
The process cannot access the file because it is being used by another process.

D:\>del Test.txt

D:\>dir Test.txt
 Volume in drive D is DATA
 Volume Serial Number is 62C1-9EB3

 Directory of D:\

12/13/2018  12:34 AM                49 Test.txt
               1 File(s)             49 bytes
               0 Dir(s)  1,486,254,481,408 bytes free
```

Как и ожидалось, нельзя открыть файл, пока FsTool имеет к нему эксклюзивный доступ. Когда делается попытка удалить файл, система помечает его на удаление, но не может удалить из пространства имен файловой системы. Если попробовать удалить файл еще раз с помощью File Explorer, то можно наблюдать такое же поведение. Когда вы нажимаете Enter в первом окне командной строки и выходите из приложения FsTool, файл действительно удаляется драйвером файловой системы NTFS.

Следующим шагом будет использование POSIX-удаления для избавления от файла. Это можно сделать, указав аргумент командной строки `/pdel` приложению FsTool. В первом окне командной строки перезапустите FsTool с аргументом командной строки `/touch` — исходный файл уже был помечен на удаление и повторно удалить его нельзя. Прежде чем нажать Enter, переключитесь на второе окно и выполните следующую команду:

```
D:\>FsTool /pdel Test.txt
NTFS / ReFS Tool v0.1
Copyright (C) 2018 Andrea Allievi (AaL186)

Deleting "Test.txt" file (Posix semantics)... Success.
Press any key to exit...
```

```
D:\>dir Test.txt
 Volume in drive D is DATA
 Volume Serial Number is 62C1-9EB3

 Directory of D:\

File Not Found
```

В данном случае файл `Test.txt` полностью удален из пространства имен файловой системы, но все еще действителен. Если нажать Enter в первом окне командной строки, то FsTool все еще сможет записывать в него данные. Это происходит потому, что файл был перемещен в скрытый системный каталог `\$Extend\$Deleted`.

Дефрагментация

Несмотря на то что NTFS старается сохранить целостность файлов при выделении блоков для их увеличения, файлы тома все равно могут со временем стать фрагментированными, особенно когда файл расширяется несколько раз или мало свободного места. Файл фрагментирован, если его данные занимают несмежные кластеры. Например, на рис. 11.26 показан файл, состоящий из пяти фрагментов. Однако, как и большинство файловых систем, включая версии FAT в Windows, NTFS не прилагает особых усилий для сохранения целостности файлов (этим занимается встроенный дефрагментатор), за исключением резервирования области дискового пространства, известной как зона *главной файловой таблицы* (Master File Table, MFT). NTFS позволяет другим файлам выделять место из зоны MFT, когда свободное пространство в томе заканчивается. Если сохранять эту область свободной для MFT, она может оставаться единой, но может и стать фрагментированной. Дополнительные сведения о MFT см. в разделе «Главная файловая таблица» далее в этой главе.

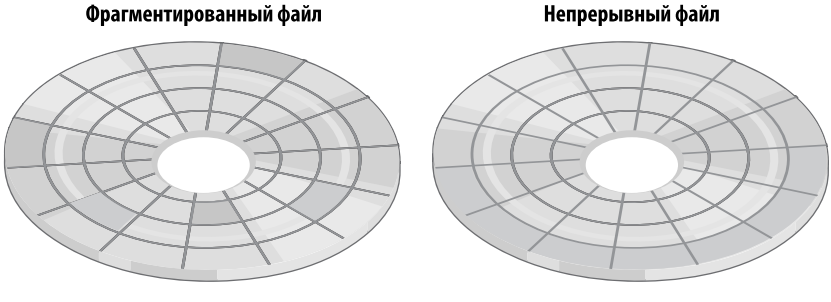


Рис. 11.26. Фрагментированный и непрерывный файлы

Чтобы облегчить разработку сторонних средств дефрагментации диска, Windows включает API дефрагментации, который эти средства может задействовать для перемещения файловых данных таким образом, чтобы файлы занимали смежные кластеры. API состоит из элементов управления файловой системой,

позволяющих приложениям получать карту свободных и используемых кластеров тома (`FSCTL_GET_VOLUME_BITMAP`), карту применения кластеров файлом (`FSCTL_GET_RETRIEVAL_POINTERS`) и перемещать файл (`FSCTL_MOVE_FILE`).

В Windows есть встроенный инструмент дефрагментации, доступный с помощью утилиты Optimize Drives, %SystemRoot%\System32\Dfrgui.exe (рис. 11.27), а также интерфейс командной строки %SystemRoot%\System32\Defrag.exe, который можно запускать в интерактивном режиме или по расписанию, но который не выдает подробные отчеты и не позволяет контролировать процесс дефрагментации, например исключать файлы или каталоги.

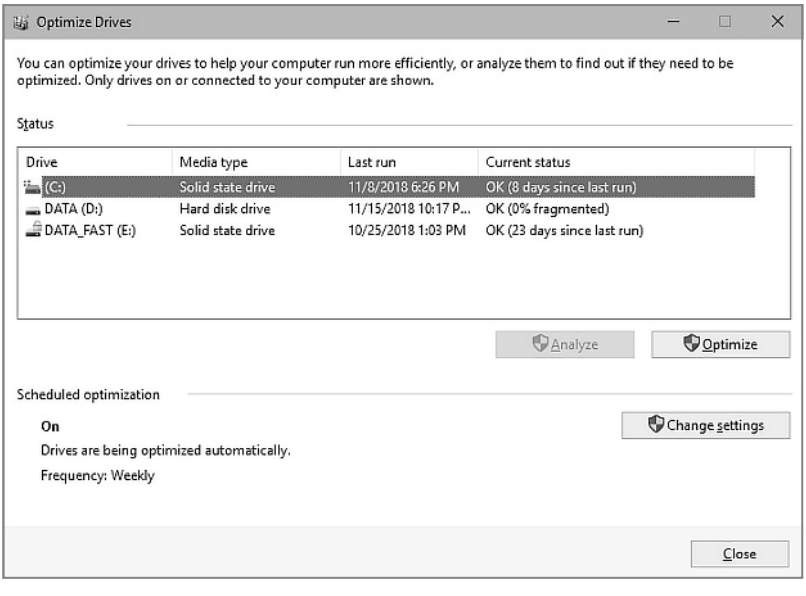


Рис. 11.27. Инструмент Optimize Drives

Единственное ограничение, накладываемое реализацией дефрагментации в NTFS, заключается в том, что файлы подкачки и файлы журналов NTFS не могут быть дефрагментированы. Инструмент Optimize Drives является развитием Disk Defragmenter, который был доступен в Windows 7. Инструмент был обновлен для поддержки многоуровневых томов, SMR- и SSD-дисков. Механизм оптимизации реализован в службе Optimize Drive, Defragsvc.dll, открывающей COM-интерфейс `IDefragEngine`, используемый как графическим инструментом, так и интерфейсом командной строки.

Для SSD-дисков инструмент реализует также операцию `retrim`. Чтобы понять, как она выполняется, необходимо вкратце рассказать об архитектуре твердотельного диска. SSD-диски хранят данные в ячейках флеш-памяти, сгруппированных в страницы размером от 4 до 16 Кбайт, объединенные в блоки, обычно состоящие из 128–512 страниц. Записывать данные в ячейки флеш-памяти напрямую можно, только когда они пусты. Если же содержат данные, то перед операцией записи их содержимое необходимо стереть. Операция записи на твердотельный накопитель может быть выполнена на одной странице, но из-за аппаратных ограничений

команды стирания всегда затрагивают целые блоки, поэтому запись данных на пустые страницы на SSD происходит очень быстро, но значительно замедляется, когда требуется перезаписать ранее записанные страницы. В этом случае сначала содержимое всего блока сохраняется в кэш, а затем весь блок стирается с твердотельного накопителя. Переписанная страница записывается в кэш-блок, и, наконец, весь обновленный блок записывается на флеш-носитель. Чтобы решить проблему скорости, драйвер файловой системы NTFS пытается отправить контроллеру SSD команду `TRIM` при каждом удалении кластеров диска, которые могут частично или полностью принадлежать файлу. В ответ на команду `TRIM` твердотельный накопитель, если это возможно, начинает асинхронно стирать целые блоки. Примечательно, что контроллер SSD ничего не может сделать, если удаляемая область соответствует лишь некоторым страницам блока.

Операция `retrim` анализирует SSD-диск и начинает отправлять команду `TRIM` каждому кластеру в свободное пространство фрагментами размером 1 Мбайт. Это объясняется следующими мотивами.

- Команды `TRIM` не всегда выполняются. Файловая система не очень строго относится к обрезке.
- Файловая система NTFS выдает команды `TRIM` на страницы, но не на блоки SSD. Дисковый оптимизатор с помощью операции `retrim` ищет фрагментированные блоки. Для них он сначала перемещает действительные данные обратно в некоторые временные блоки, дефрагментируя исходные и даже вставляя страницы, принадлежащие другим фрагментированным блокам. Наконец, он выдает команды `TRIM` на исходные очищенные блоки.

ПРИМЕЧАНИЕ Способ, которым дисковый оптимизатор выдает команды TRIM на свободное пространство, довольно сложен: он выделяет пустой разреженный файл и ищет кусок свободного пространства размером от 128 Кбайт до 1 Гбайт. Затем вызывает файловую систему через управляющий код FSCTL_MOVE_FILE и перемещает в пустое пространство данные из разреженного файла, имеющего размер 1 Гбайт, но на самом деле не содержащего никаких корректных данных. Базовая файловая система фактически стирает содержимое одного или нескольких блоков SSD — разреженные файлы, не содержащие осмысленных данных, при чтении возвращают фрагменты из нулей. Это реализация команды TRIM, выполняемой прошивкой SSD.

Для многоуровневых и SMR-дисков инструмент Optimize Drives поддерживает две дополнительные операции: Slabify (Slab Consolidation) и Tier Optimization. Большие файлы, хранящиеся на многоуровневых томах, могут состоять из различных экстентов, расположенных на разных уровнях. Операция Slab Consolidation не только дефрагментирует таблицу экстентов файла (этап называется консолидацией), но и перемещает содержимое файла в конгруэнтные слэбы. Слэб — это единица распределения диска с тонким резервированием (дополнительную информацию см. в разделе «Storage Spaces» далее в этой главе). Конечная цель Slab Consolidation — позволить файлам использовать меньшее количество слэбов. Tier Optimization перемещает файлы с частым доступом, включая явно зафиксированные, с уровня хранения на уровень производительности и, наоборот, перемещает файлы с меньшим доступом с уровня производительности на уровень хранения. Для этого оптимизационный механизм обращается к механизму распределения по

уровням. Тот предоставляет экстенты файлов, которые должны быть перемещены на уровень емкости, и файлы, которые должны быть перемещены на уровень производительности, на основе карты применения для каждого файла, к которому обращался пользователь.

ПРИМЕЧАНИЕ Многоуровневые диски и механизм многоуровневой обработки подробно рассматриваются в следующих разделах данной главы.

ЭКСПЕРИМЕНТ. Retrim на томе SSD

Выполнить Retrim на быстром SSD или NVMe-томе можно с помощью команды `defrag.exe /L`, как в следующем примере.

```
D:\>defrag /L c:
Microsoft Drive Optimizer
Copyright (c) Microsoft Corp.

Invoking retrim on (C:)...

The operation completed successfully.

Post Defragmentation Report:

        Volume Information:
               Volume size                = 475.87 GB
                Free space                = 343.80 GB
        Retrim:
                Total space trimmed        = 341.05 GB
```

В этом примере размер тома составлял 475,87 Гбайт, а свободного места было 343,80 Гбайт. Только 341 Гбайт было стерто и обрезано. Очевидно, что при выполнении этой команды на томах на классическом жестком диске будет выдана ошибка. Запрашиваемая операция не поддерживается аппаратной основой тома.

Динамическая разбивка на разделы

Драйвер NTFS позволяет динамически изменять размер любого раздела, включая системный, либо уменьшая, либо увеличивая его при наличии достаточного пространства. Расширить раздел легко, если на диске достаточно места, и выполняется это с помощью кода управления файловой системой `FSCTL_EXPAND_VOLUME`. Сокращение раздела — более сложный процесс, поскольку он требует перемещения всех данных файловой системы, находящихся в тот момент в области, подлежащей удалению, в область, которая останется после процесса сокращения. Этот механизм аналогичен дефрагментации. Процесс уменьшения раздела реализуется двумя компонентами: *механизмом уменьшения* и драйвером файловой системы.

Механизм уменьшения размера реализован в пользовательском режиме. Он взаимодействует с NTFS, чтобы определить максимальное количество потенциально восстанавливаемых байтов, то есть сколько данных можно переместить из области, размер которой будет изменен, в область, которая останется. Механизм уменьшения

размера применяет описанный ранее стандартный механизм дефрагментации, который не поддерживает перемещение используемых фрагментов файла подкачки или любых других файлов, помеченных как неперемещаемые с помощью управляющего кода файловой системы FSCTL_MARK_HANDLE, например файла спящего режима. Резервная копия главной файловой таблицы ($MftMirr), журнал транзакций метаданных NTFS ($LogFile) и файл меток тома ($Volume) не могут быть перемещены, что ограничивает минимальный размер уменьшаемого тома и вызывает появление неиспользуемого пространства.

Код уменьшения драйвера файловой системы отвечает за то, чтобы том оставался в связном состоянии на протяжении всего процесса уменьшения. Для этого он предоставляет интерфейс, использующий следующие три запроса, описывающие текущую операцию, которые отправляются через управляющий код FSCTL_SHRINK_VOLUME.

- Запрос ShrinkPrepare, который должен быть выдан перед любой другой операцией. Он принимает желаемый размер нового тома в секторах и используется для того, чтобы файловая система могла блокировать дальнейшее выделение за границей нового тома. Запрос ShrinkPrepare не проверяет, действительно ли том может быть уменьшен на указанную величину, но гарантирует, что та численно допустима и не выполняются другие операции уменьшения. Заметьте, что после операции подготовки файловый дескриптор тома оказывается связанным с запросом на сжатие. Если файловый дескриптор закрыт, то предполагается, что операция прервана.
- Запрос ShrinkCommit, который механизм сжатия выдает после запроса ShrinkPrepare. В этом состоянии файловая система пытается удалить запрошенное в последнем ShrinkPrepare количество кластеров. Если было отправлено несколько запросов с разными размерами, то определяющим является последний. Запрос ShrinkCommit предполагает, что механизм уменьшения размера завершил работу, и потерпит неудачу, если в области, подлежащей уменьшению, остались выделенные блоки.
- Запрос ShrinkAbort, который может быть выдан механизмом сокращения или вызван такими событиями, как закрытие файлового дескриптора тома. Этот запрос отменяет операцию ShrinkCommit, возвращая разделу исходный размер, и позволяет и далее выделять новые места за пределами уменьшенной области. Однако изменения в ходе дефрагментации, сделанные механизмом уменьшения, сохраняются.

Если система перезагружается во время операции уменьшения, то NTFS восстанавливает файловую систему в связное состояние с помощью механизма восстановления метаданных, о котором рассказывается далее в этой главе. Поскольку операция уменьшения выполняется только после завершения всех остальных операций, исходный размер тома не изменяется и сохраняются только операции дефрагментации, уже сброшенные на диск.

Наконец, уменьшение тома влияет и на механизм теневого копирования тома. Напомним, что механизм копирования при записи позволяет VSS просто сохранять части файла, которые были изменены, не затрагивая при этом ссылки на исходные данные файла. Для удаленных файлов эти данные не будут связаны с видимыми файлами, а станут отображаться как свободное пространство, которое, скорее всего, окажется в области, подлежащей сокращению. Поэтому механизм сокращения

связывается с VSS, чтобы вовлечь его в процесс сокращения. В общих чертах задача механизма VSS заключается в копировании данных удаленных файлов в область разграничения и увеличении последней по мере необходимости для размещения дополнительных данных. Эта деталь важна, поскольку она накладывает еще одно ограничение на размер, до которого могут уменьшиться даже тома с большим количеством свободного пространства.

Поддержка NTFS для многоуровневых томов

Многоуровневые тома состоят из различных типов устройств хранения и базовых носителей и обычно создаются поверх одного физического или виртуального диска. Storage Spaces предоставляет виртуальные диски, состоящие из нескольких физических дисков, которые могут быть разных типов и иметь разную производительность: быстрые NVMe-диски, SSD и обычные жесткие диски. Виртуальный диск такого типа называется *многоуровневым*. В Storage Spaces используется название *Storage Tiers*. В то же время многоуровневые тома могут быть созданы поверх физических SMR-дисков, которые имеют обычную быструю зону произвольного доступа и зону только последовательной емкости. Все многоуровневые тома характеризуются тем, что состоят из уровня производительности, поддерживающего быстрый произвольный ввод-вывод, и уровня хранения, который не обязательно поддерживает быстрый произвольный ввод-вывод, медленнее и имеет большую емкость.

ПРИМЕЧАНИЕ Диски SMR, многоуровневые тома и Storage Spaces более подробно будут рассмотрены далее в этой главе.

Драйвер файловой системы NTFS поддерживает многоуровневые тома следующими способами.

- Том разделен на две зоны, соответствующие многоуровневым дисковым областям хранения и производительности.
- Новый атрибут `$DSC` типа `$LOGGED_UTILITY_STREAM` указывает, на каком уровне должен храниться файл. NTFS предоставляет новый интерфейс закрепления, который позволяет заблокировать файл на определенном уровне и предотвратить его перемещение системой уровней.
- Служба Storage Tiers Management играет центральную роль в поддержке многоуровневых томов. Драйвер файловой системы NTFS записывает события ETW типа «горячо» при каждом чтении или записи файлового потока. Механизм управления уровнями потребляет эти события, накапливает их во фрагментах по 1 Мбайт и записывает в базу данных один раз в час. Каждые 4 часа система уровней обрабатывает эту базу данных и с помощью сложного алгоритма решает, какой файл считается недавним (горячим), а какой — старым (холодным). Механизм распределения файлов по уровням перемещает файлы между уровнями производительности и емкости на основе рассчитанных данных.

Кроме того, распределитель NTFS был изменен, чтобы выделять файловые кластеры на основе уровня, указанного в атрибуте `$DSC`. Распределитель NTFS

использует специальный алгоритм, чтобы решить, из какого уровня выделять кластеры тома. Алгоритм работает путем выполнения проверок в следующем порядке.

1. Если файл является журналом USN для тома, он всегда выделяется на уровне хранения.
2. Записи MFT и файлы системных метаданных всегда выделяются на уровне производительности.
3. Если файл был ранее явно закреплен, то есть имеет атрибут `$DSC`, то он выделяется на указанном уровне хранения.
4. Если система работает под управлением клиентской версии Windows, то всегда отдается предпочтение уровню производительности, в противном случае — уровню хранения.
5. Если на уровне производительности нет места, то используется уровень хранения.

Приложение может указать желаемый уровень для файла с помощью API `NtSetInformationFile` с информационным классом `FileDesiredStorageClass-Information`. Эта операция называется *закреплением файла*, и если она выполняется для дескриптора вновь созданного файла, то центральный распределитель выделит содержимое нового файла на указанном уровне. В противном случае, то есть если файл уже существует и расположен на неправильном уровне, механизм распределения перемещает его на нужный уровень при следующем запуске. Эта операция называется *оптимизацией уровня* и может быть инициирована запланированной задачей механизма управления уровнями или задачей `SchedulerDefrag`.

ПРИМЕЧАНИЕ Важно отметить, что описанная здесь поддержка многоуровневых томов в NTFS полностью отличается от поддержки, предоставляемой драйвером файловой системы ReFS.

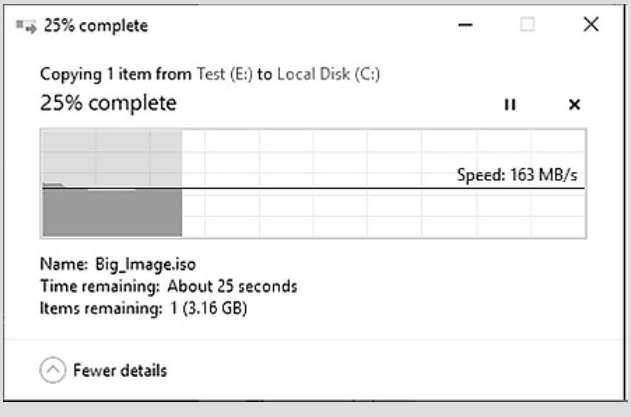
ЭКСПЕРИМЕНТ. Наблюдение за закреплением файлов в многоуровневых томах

Как говорилось в предыдущем разделе, распределитель NTFS использует специальный алгоритм для принятия решения о том, на каком уровне выделять ресурсы. В этом эксперименте большой файл копируется в многоуровневый том и рассматриваются последствия операции закрепления файла. После завершения копирования откройте административное окно PowerShell, щелкнув правой кнопкой мыши на значке меню Пуск и выбрав Windows PowerShell (Admin), и используйте команду `Get-FileStorageTier`, чтобы получить информацию об уровнях для файла:

```
PS E:\> Get-FileStorageTier -FilePath 'E:\Big_Image.iso' | FL FileSize,
DesiredStorageTierClass, FileSizeOnPerformanceTierClass,
FileSizeOnCapacityTierClass,
PlacementStatus, State
```

```
FileSize                        : 4556566528
DesiredStorageTierClass         : Unknown
FileSizeOnPerformanceTierClass : 0
FileSizeOnCapacityTierClass     : 4556566528
PlacementStatus                 : Unknown
State                           : Unknown
```

Пример показывает, что файл `Big_Image.iso` был выделен на уровне хранения. Пример выполнен в системе Windows Server. Чтобы убедиться в этом, просто скопируйте файл с многоуровневого диска на быстрый SSD-том. Скорость передачи данных будет низкой, обычно от 160 до 250 Мбайт/с в зависимости от скорости вращения диска.



Теперь можно выполнить запрос закрепления с помощью команды `Set-FileStorageTier`, как в следующем примере:

```
PS E:\> Get-StorageTier -MediaType SSD | FL FriendlyName, Size, FootprintOnPool,
UniqueId

FriendlyName    : SSD
Size            : 128849018880
FootprintOnPool : 128849018880
UniqueId        : {448abab8-f00b-42d6-b345-c8da68869020}

PS E:\> Set-FileStorageTier -FilePath 'E:\Big_Image.iso'
-DesiredStorageTierFriendlyName
'SSD'
PS E:\> Get-FileStorageTier -FilePath 'E:\Big_Image.iso' | FL FileSize,
DesiredStorageTierClass, FileSizeOnPerformanceTierClass,
FileSizeOnCapacityTierClass,
PlacementStatus, State

FileSize                        : 4556566528
DesiredStorageTierClass         : Performance
FileSizeOnPerformanceTierClass : 0
FileSizeOnCapacityTierClass     : 4556566528
PlacementStatus                 : Not on tier
State                           : Pending
```

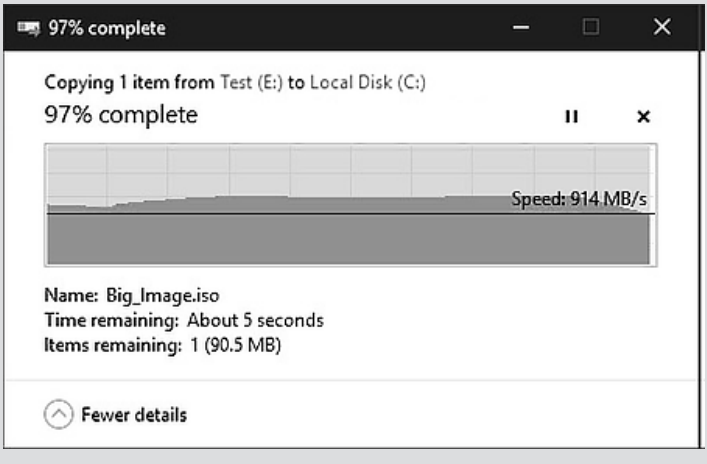
В приведенном примере показано, что файл был правильно помещен на уровень производительности, но его содержимое все еще на уровне хранения. Когда выполняется запланированное задание механизма управления уровнями, оно перемещает экстенты файлов с уровня хранения на уровень производительности. Можно принудительно оптимизировать уровень, запустив оптимизатор дисков через встроенный инструмент defrag.exe /g:

```
PS E:> defrag /g /h e:
Microsoft Drive Optimizer
Copyright (c) Microsoft Corp.

Invoking tier optimization on Test (E:)...

Pre-Optimization Report:

        Volume Information:
                Volume size                = 2.22 TB
                Free space                 = 1.64 TB
                Total fragmented space     = 36%
                Largest free space size    = 1.56 TB

        Note: File fragments larger than 64MB are not included in the
fragmentation
        statistics.

The operation completed successfully.

Post Defragmentation Report:

        Volume Information:
                Volume size                = 2.22 TB
                Free space                 = 1.64 TB

        Storage Tier Optimization Report:
                % I/Os Serviced from Perf Tier Perf Tier Size Required
                100%                           28.51 GB *
                95%                            22.86 GB
...
                20%                            2.44 GB
                15%                            1.58 GB
                10%                            873.80 MB
                5%                             361.28 MB

        * Current size of the Performance tier: 474.98 GB
        Percent of total I/Os serviced from the Performance tier: 99%

        Size of files pinned to the Performance tier: 4.21 GB
        Percent of total I/Os: 1%

        Size of files pinned to the Capacity tier: 0 bytes
        Percent of total I/Os: 0%
```

Программа Drive Optimizer подтвердила закрепление файла. Можно еще раз проверить статус закрепления, выполнив команду `Get-FileStorageTier` и снова скопировав файл на SSD-том. На этот раз скорость передачи данных должна

быть гораздо выше, поскольку все содержимое файла находится на уровне производительности:

```
PS E:\> Get-FileStorageTier -FilePath 'E:\Big_Image.iso' | FL FileSize,
DesiredStorageTierClass, FileSizeOnPerformanceTierClass,
FileSizeOnCapacityTierClass,
PlacementStatus, State

FileSize                        : 4556566528
DesiredStorageTierClass         : Performance
FileSizeOnPerformanceTierClass : 0
FileSizeOnCapacityTierClass     : 4556566528
PlacementStatus                 : Completely on tier
State                           : OK
```



Можно повторить эксперимент в клиентской версии Windows 10, поместив файл на уровень хранения, — клиентские версии Windows 10 по умолчанию выделяют кластеры файлов на уровне производительности. Аналогичная функция закрепления реализована в приложении FsTool, доступном в загружаемых ресурсах этой книги, с помощью которого можно скопировать файл непосредственно на нужный уровень.

ДРАЙВЕР ФАЙЛОВОЙ СИСТЕМЫ NTFS

Как говорилось в главе 6, в рамках системы ввода-вывода Windows файловая система NTFS и другие файловые системы являются загружаемыми драйверами устройств, которые работают в режиме ядра. Они вызываются косвенно приложениями, использующими Windows или другие API ввода-вывода. Как показано на рис. 11.28, подсистемы среды Windows вызывают системные службы Windows, которые, в свою очередь, находят соответствующие загруженные драйверы и вызывают их. Описание диспетчеризации системных служб см. в разделе «Диспетчеризация системных служб» главы 8.

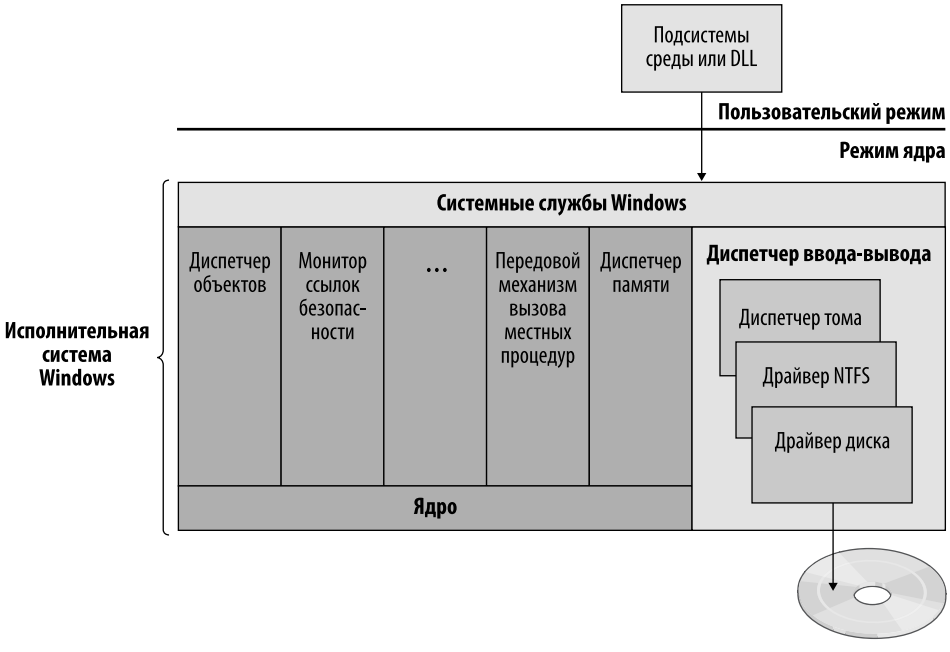


Рис. 11.28. Компоненты системы ввода-вывода Windows

Многоуровневые драйверы передают друг другу запросы ввода-вывода, обращаясь к диспетчеру ввода-вывода исполнительного блока Windows. Использование диспетчера ввода-вывода в качестве посредника позволяет каждому драйверу сохранять независимость, чтобы его можно было загружать или выгружать, не влияя на другие драйверы. Кроме того, драйвер NTFS взаимодействует с тремя другими компонентами исполнительной системы Windows, показанными в левой части рис. 11.29, которые тесно связаны с файловыми системами.

Служба файла журнала (log file service, LFS) — это часть NTFS, предоставляющая услуги по ведению журнала записей на диск. Файл журнала, записываемый LFS, используется для восстановления тома, отформатированного в NTFS, в случае системного сбоя (см. раздел «Служба файла журнала» далее в этой главе).

Как уже говорилось, диспетчер кэша — это компонент исполнительной системы Windows, который предоставляет общесистемные услуги кэширования для NTFS и других драйверов файловых систем, включая драйверы сетевых файловых систем — серверы и редиректоры. Все файловые системы, реализованные в Windows, обращаются к кэшированным файлам, отображая их в системное адресное пространство и затем обращаясь к виртуальной памяти. Для этого диспетчер кэша предоставляет специализированный интерфейс файловой системы к диспетчеру памяти Windows. Когда программа пытается получить доступ к части файла, не загруженной в кэш, — *пропуску кэша*, диспетчер памяти вызывает NTFS для обращения к драйверу диска и получения содержимого файла с диска. Диспетчер кэша оптимизирует дисковый ввод-вывод, используя свои потоки отложенной записи для вызова диспетчера памяти, чтобы сбросить содержимое кэша на диск в качестве фоновой активности (асинхронная запись на диск).

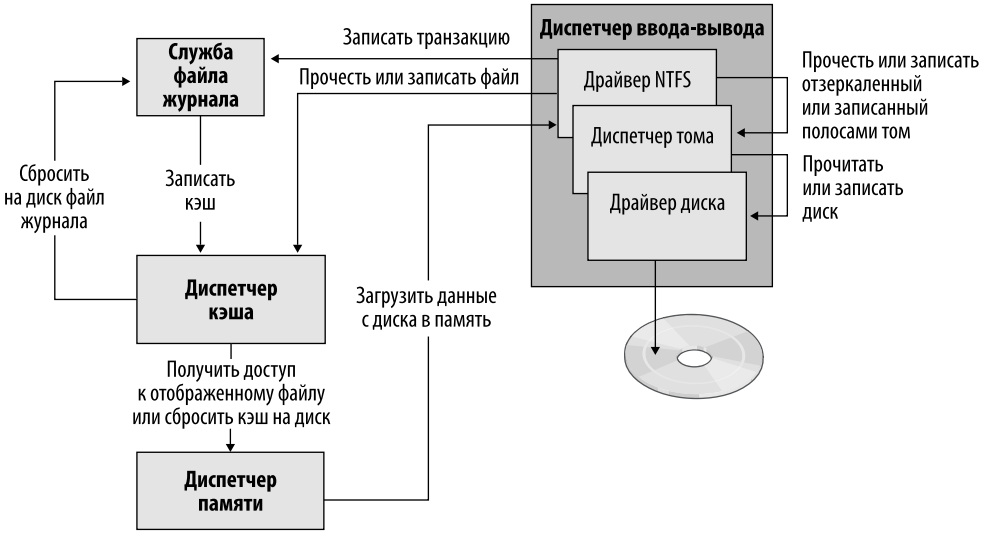


Рис. 11.29. NTFS и связанные с ней компоненты

NTFS, как и другие файловые системы, участвует в объектной модели Windows, реализуя файлы как объекты. Это позволяет совместно организовать общий доступ к файлам и защищать их диспетчером объектов — компонентом Windows, который управляет всеми объектами исполнительного уровня. (Диспетчер объектов описан в разделе «Диспетчер объектов» в главе 8.)

Приложение создает файлы и получает к ним доступ так же, как и в случае с другими объектами Windows, — с помощью объектных дескрипторов. К тому моменту, когда запрос ввода-вывода достигает NTFS, диспетчер объектов Windows и система безопасности уже проверили, что вызывающий процесс имеет право доступа к файловому объекту тем способом, которым пытается это сделать. Система безопасности сравнила маркер доступа вызывающего процесса с записями в списке контроля доступа к файловому объекту. (Дополнительные сведения о списках контроля доступа см. в главе 7.) Диспетчер ввода-вывода преобразовал файловый дескриптор в указатель на файловый объект. NTFS использует информацию в файловом объекте для доступа к файлу на диске.

На рис. 11.30 показаны структуры данных, связывающие файловый дескриптор со структурой файловой системы на диске.

NTFS использует несколько указателей, чтобы перейти от объекта файла к его местоположению на диске. Как показано на рис. 11.30, файловый объект, представляющий собой один вызов системной службы open-file, указывает на *блок управления потоком* (stream control block, SCB) для атрибута файла, который вызывающая сторона пытается прочитать или записать. Здесь процесс открыл для файла как неименованный атрибут данных, так и именованный поток — альтернативный атрибут данных. SCB представляют отдельные атрибуты файла и содержат информацию о том, как найти определенные атрибуты в файле. Все SCB для файла указывают на общую структуру данных, называемую *блоком управления файлом* (file control

block, FCB). FCB содержит указатель (фактически индекс в MFT, как объясняется в разделе «Номера файловых записей» далее в этой главе) на запись файла в дисковой таблице главных файлов (master file table, MFT), которая подробно описана в следующем разделе.

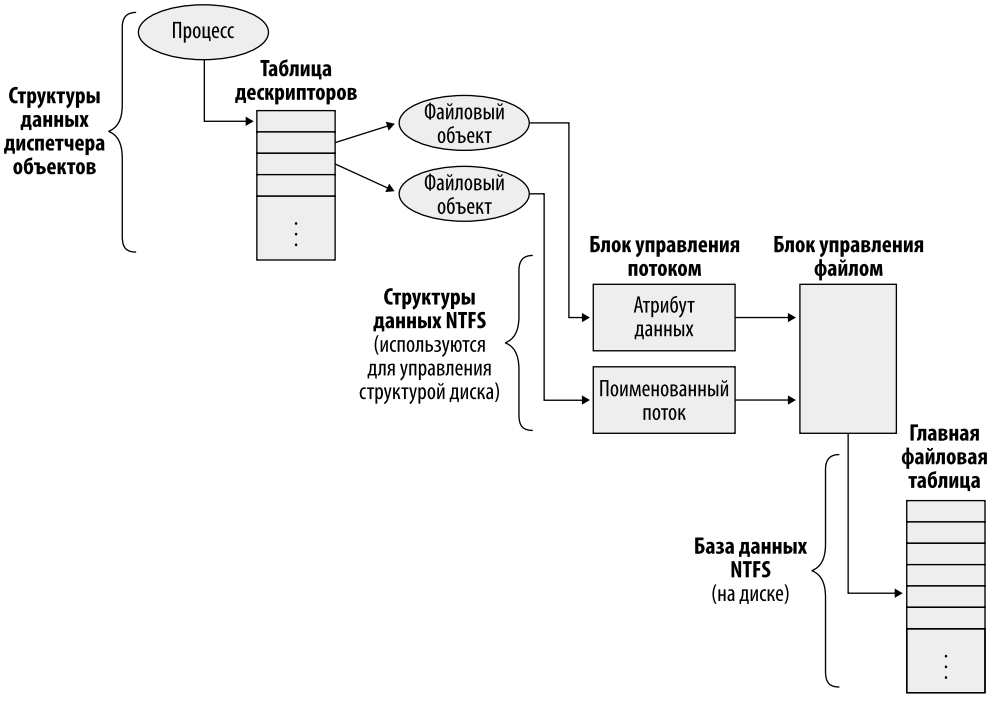


Рис. 11.30. Структуры данных NTFS

СТРУКТУРА NTFS НА ДИСКЕ

В этом разделе описывается структура тома NTFS на диске, в том числе то, как дисковое пространство организуется и делится на кластеры, как файлы организуются в каталоги, как фактические данные файлов и информация об атрибутах хранятся на диске и, наконец, как работает сжатие данных NTFS.

Тома

Структура NTFS начинается с *тома*, который соответствует логическому разделу на диске и создается при форматировании диска или его части под NTFS. Также можно создать виртуальный диск RAID, охватывающий несколько физических дисков, с помощью Storage Spaces (доступ к нему осуществляется через оснастку панели управления Manage Storage Spaces) или команд Storage Spaces, доступных в Windows PowerShell, например команды `New-StoragePool`, используемой для

создания нового пула хранения. Полный список команд PowerShell для Storage Spaces доступен по следующей ссылке: https://docs.microsoft.com/en-us/powershell/module/storagespaces/.

На диске может быть один том или несколько. NTFS обрабатывает каждый том независимо от других. Три примера конфигурации диска для жесткого диска емкостью 2 Тбайт показаны на рис. 11.31.

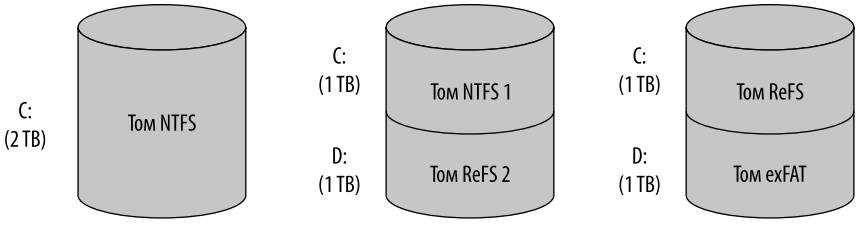


Рис. 11.31. Образцы конфигураций дисков

Том состоит из серии файлов и дополнительного нераспределенного пространства, оставшегося на разделе диска. Во всех файловых системах FAT том содержит также области, специально отформатированные для использования файловой системой. Однако в томах NTFS или ReFS все данные файловой системы, такие как битовые карты и каталоги, и даже загрузочный файл системы хранятся как обычные файлы.

ПРИМЕЧАНИЕ Дисковый формат томов NTFS в Windows 10 и Windows Server 2019 имеет версию 3.1, как и в Windows XP и Windows Server 2003. Номер версии тома хранится в файле метаданных $Volume.

Кластеры

Размер кластера в томе NTFS, или *кластерный фактор*, определяется, когда пользователь форматирует том с помощью команды `format` или оснастки MMC Disk Management. Кластерный фактор по умолчанию зависит от размера тома, но представляет собой целое число физических секторов, всегда равное степени двойки, — один сектор, два сектора, четыре сектора, восемь секторов и т. д. Кластерный фактор записывается как число байтов в кластере, например 512 байт, 1 Кбайт, 2 Кбайт и т. д.

NTFS оперирует только кластерами. Однако NTFS формирует низкоуровневые операции ввода-вывода томов таким образом, что кластеры выровнены по секторам и имеют длину, кратную размеру сектора. NTFS использует кластер в качестве единицы выделения, чтобы сохранить независимость от размеров физических секторов. Эта независимость позволяет NTFS эффективно поддерживать очень большие диски за счет применения большего кластерного фактора или поддерживать новые диски с размером сектора, отличным от 512 байт. В больших томах использование большего кластерного фактора может уменьшить фрагментацию и ускорить выделение, но за счет неиспользованного дискового пространства.

Если размер кластера 64 Кбайт, а размер файла всего 16 Кбайт, то 48 Кбайт будут потрачены впустую. Как команда *форматирования*, доступная в командной строке, так и пункт меню Форматировать (Format) в опции Все задачи (All Tasks) меню Действие (Action) оснастки Disk Management MMC выбирают кластерный фактор по умолчанию, основанный на размере тома, но этот размер можно переопределить.

NTFS обозначает физические места на диске с помощью *номеров логических кластеров* (logical cluster number, LCN). LCN — это просто нумерация всех кластеров от начала тома до конца. Для преобразования LCN в физический адрес диска NTFS умножает LCN на кластерный фактор, чтобы получить физическое смещение байта в томе, как того требует интерфейс драйвера диска. NTFS обозначает данные в файле с помощью *номеров виртуальных кластеров* (virtual cluster number, VCN). VCN нумеруют кластеры, принадлежащие определенному файлу, от 0 до *m*. Однако VCN не обязательно являются физически смежными, они могут быть отображены на любое количество LCN в томе.

Главная файловая таблица

В NTFS в файлах содержатся все данные, хранящиеся в томе, включая структуры данных, используемые для поиска и извлечения файлов, загрузочные данные и битовую карту, записывающую состояние выделения всего тома, — метаданные NTFS. Хранение всего этого в файлах позволяет файловой системе легко находить и обслуживать данные, а каждый отдельный файл может быть защищен дескриптором безопасности. Кроме того, если какая-то часть диска выходит из строя, NTFS может переместить файлы метаданных, чтобы диск не стал недоступным.

Главная файловая таблица (master file table, MFT) — это сердце структуры тома NTFS. Она реализована в виде массива файловых записей. Размер каждой файловой записи может быть 1 Кбайт или 4 Кбайт, как определено при форматировании тома, и зависит от типа базового физического носителя: новые физические диски с размером собственных секторов 4 Кбайт и многоуровневые диски обычно используют файловые записи размером 4 Кбайт, а старые диски с размером секторов 512 байт — размером 1 Кбайт. Размер каждой записи MFT не зависит от размера кластеров и может быть изменен во время форматирования тома с помощью команды `Format /l`. Структура файловой записи описана в разделе «Файловые записи» далее в этой главе. Логически MFT содержит по одной записи для каждого файла в томе, включая запись для самой MFT. Помимо MFT, каждый том NTFS включает в себя набор файлов метаданных, содержащих информацию, использующуюся для реализации структуры файловой системы. Каждый из файлов метаданных NTFS имеет имя, которое начинается со знака доллара `$` и является скрытым. Например, имя файла MFT — `$MFT`. Остальные файлы в томе NTFS — это обычные пользовательские файлы и каталоги (рис. 11.32).

Обычно каждая запись MFT соответствует отдельному файлу. Однако если файл имеет большое количество атрибутов или сильно фрагментирован, то для одного файла может потребоваться более одной записи. В таких случаях первая запись MFT, в которой хранится местоположение остальных, называется *записью базового файла*.

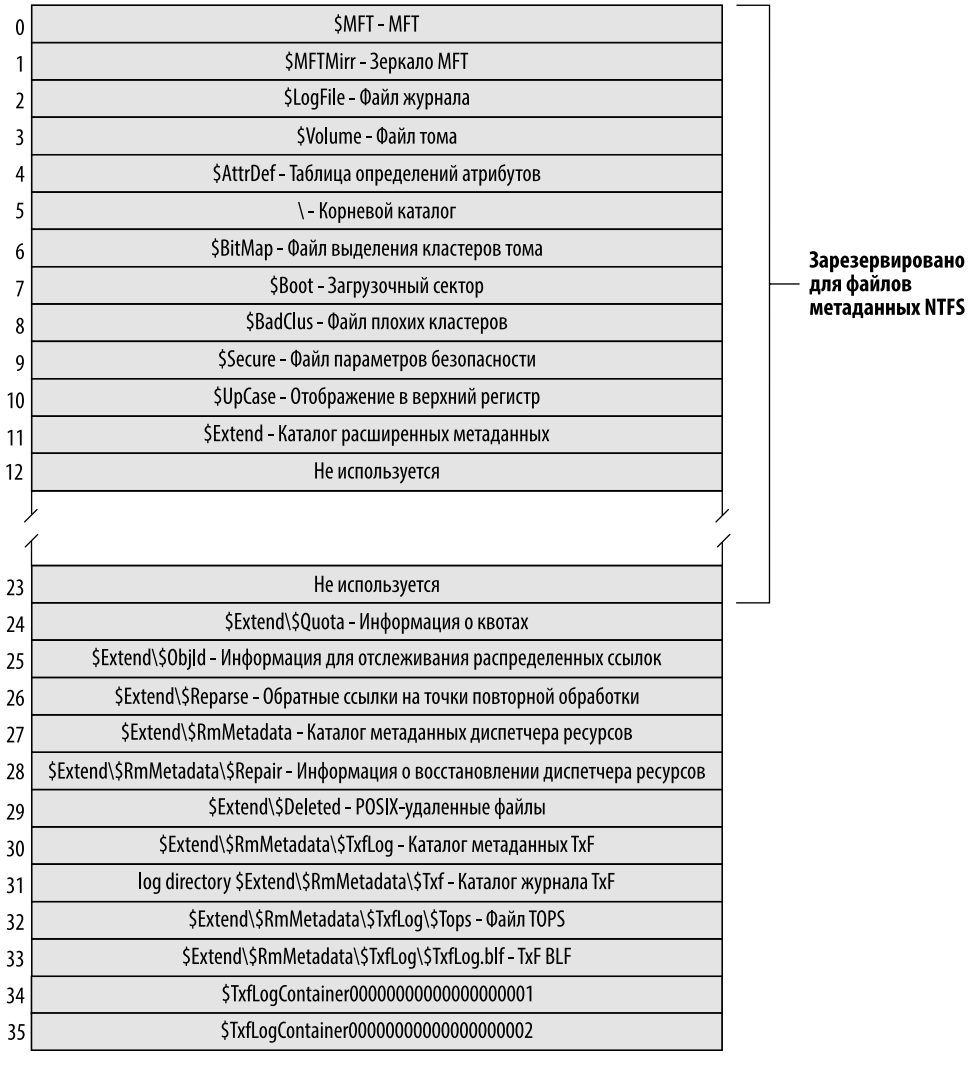


Рис. 11.32. Файловые записи для файлов метаданных NTFS в MFT

При первом обращении к тому NTFS должна *смонтировать* его, то есть прочитать метаданные с диска и создать внутренние структуры данных, чтобы можно было обрабатывать обращения к файловой системе приложений. Чтобы смонтировать том, NTFS просматривает загрузочную запись тома (volume boot record, VBR), расположенную по адресу LCN 0, которая содержит структуру данных, называемую блоком параметров загрузки (boot parameter block, BPB), чтобы найти физический адрес диска MFT. Файловая запись MFT — это первая запись в таблице. Вторая файловая запись указывает на файл `$MFTMirr`, находящийся в середине диска и называемый *зеркалом MFT*, который содержит копию первых четырех строк MFT. Эта

частичная копия MFT используется для поиска файлов метаданных, если часть файла MFT по какой-то причине не может быть прочитана.

Когда NTFS находит файловую запись для MFT, она получает информацию о сопоставлении VCN и LCN в атрибуте данных файловой записи и сохраняет ее в памяти. У каждого прогона (что это такое, объясняется далее в этой главе, в разделе «Резидентные и нерезидентные атрибуты») есть сопоставление VCN и LCN и длина, поскольку это вся информация, необходимая для нахождения LCN для любого VCN. Информация о сопоставлении сообщает NTFS, где на диске находятся прогоны, содержащие MFT. Затем NTFS обрабатывает записи MFT для еще нескольких файлов метаданных и открывает эти файлы. После этого NTFS выполняет операцию восстановления файловой системы, описанную в разделе «Восстановление» далее в этой главе, и, наконец, открывает оставшиеся файлы метаданных. Теперь том доступен для пользователей.

ПРИМЕЧАНИЕ Для большей ясности в тексте и на диаграммах в этой главе прогон изображается включающим VCN, LCN и длину. На самом деле NTFS сжимает эту информацию на диске в пару из LCN и следующего VCN. Учитывая начальный VCN, NTFS может определить длину прогона, вычитая начальный VCN из следующего VCN.

На протяжении работы системы NTFS записывает данные в другой важный файл метаданных — *файл журнала* `$LogFile`. NTFS использует файл журнала для записи всех операций, которые влияют на структуру тома NTFS, включая создание файлов или любые команды, такие как *копирование*, которые изменяют структуру каталогов. Файл журнала применяется для восстановления тома NTFS после системного сбоя (он также описывается в разделе «Восстановление»).

Еще одна запись в MFT отведена под *корневой каталог*, также известный как `\`, например `C:\`. Его файловая запись содержит индекс файлов и каталогов, хранящихся в корневой части структуры каталогов NTFS. Когда у NTFS впервые запрашивается открытие файла, она начинает его поиск в файловой записи корневого каталога. После открытия файла NTFS сохраняет номер записи MFT файла, чтобы в дальнейшем при чтении и записи файла иметь прямой доступ к записи MFT.

NTFS записывает состояние распределения тома в *файл битовой карты* `$BitMap`. Атрибут данных для файла битовой карты содержит битовую карту, каждый из битов которой представляет кластер в томе и определяет, свободен ли этот кластер или был выделен под файл.

В *файле безопасности* `$Secure` хранится база дескрипторов безопасности всего тома. Файлы и каталоги NTFS имеют индивидуально настраиваемые дескрипторы безопасности, но для экономии места NTFS хранит настройки в общем файле, что позволяет файлам и каталогам, имеющим одинаковые настройки безопасности, ссылаться на один и тот же дескриптор безопасности. В большинстве сред целые деревья каталогов имеют одинаковые параметры безопасности, поэтому такая оптимизация обеспечивает значительную экономию дискового пространства.

В другом системном файле, *загрузочном*, `$Boot`, хранится код начальной загрузки Windows, если том является системным. В несистемных томах есть код, который выводит на экран сообщение об ошибке при попытке загрузки с этого тома. Чтобы система загрузилась, код начальной загрузки должен быть расположен по

определенному адресу на диске, чтобы диспетчер загрузки мог его найти. Во время форматирования команда `format` определяет эту область как файл, создавая для нее файловую запись. Все файлы находятся в MFT, а все кластеры либо свободны, либо выделены под файл — в NTFS нет скрытых файлов или кластеров, хотя некоторые файлы (метаданные) не видны пользователям. Загрузочный файл, а также файлы метаданных NTFS могут быть индивидуально защищены с помощью дескрипторов безопасности, которые применяются ко всем объектам Windows. Использование модели «все на диске — это файл» также означает, что при обычном файловом вводе-выводе загрузочный файл может быть изменен, хотя он и защищен от редактирования.

NTFS также ведет *файл плохих кластеров* `$BadClus` для записи всех плохих мест в дисковом томе и файл `$Volume`, известный как *файл тома*, содержащий имя тома, версию NTFS, под которую отформатирован том, и ряд битов-флагов, указывающих на его состояние и работоспособность, например бит, указывающий на то, что том поврежден и должен быть восстановлен утилитой Chkdsk. (Эта утилита более подробно рассматривается далее в главе.) *Файл верхнего регистра* `$UpCase` содержит таблицу перевода между строчными и прописными символами. NTFS поддерживает файл `$AttrDef`, содержащий *таблицу определений атрибутов*, которая определяет типы атрибутов, поддерживаемых в томе, и указывает, могут ли они быть проиндексированы, восстановлены во время операции восстановления системы и т. д.

ПРИМЕЧАНИЕ На рис. 11.32 приведена главная файловая таблица тома NTFS и указаны конкретные записи, в которых находятся файлы метаданных. Стоит отметить, что записи файлов в позиции ниже 16 гарантированно фиксированы. Файлы метаданных, расположенные в записях выше 16, зависят от порядка, в котором NTFS их создает. Действительно, инструмент форматирования не создает файлы метаданных выше позиции 16, это обязанность драйвера файловой системы NTFS при первом монтировании тома после завершения форматирования. Порядок следования файлов метаданных, создаваемых драйвером файловой системы, не гарантируется.

NTFS хранит в каталоге *расширений* метаданных `$Extend` несколько файлов метаданных, включая *файл идентификатора объекта* `$ObjId`, *файл квот* `$Quota`, *файл журнала изменений* `$UsnJrnl`, *файл точек повторной обработки* `$Reparse`, каталог поддержки *POSIX-удаления* `$Deleted`, а также *каталог диспетчера ресурсов по умолчанию* `$RmMetadata`. В этих файлах хранится информация, связанная с расширенными возможностями NTFS. В файле идентификаторов объектов хранятся идентификаторы файловых объектов, в файле квот — информация о квотах и поведении томов с включенными квотами, в файле журнала изменений — изменения файлов и каталогов, а в файле точек повторной обработки — информация о том, какие файлы и каталоги в томе содержат данные точек повторной обработки.

Каталог POSIX-удаления `$Deleted` содержит невидимые для пользователя файлы, которые были удалены с помощью новой семантики POSIX. Эти файлы будут перемещены в данный каталог, когда приложение, первоначально запросившее удаление файла, закроет его дескриптор. Другие приложения, которые могут попрежнему иметь действительную ссылку на файл, продолжают работать, пока имя файла удаляется из пространства имен. Подробная информация о POSIX-удалении была представлена в предыдущем разделе.

Каталог диспетчера ресурсов по умолчанию содержит каталоги, связанные с поддержкой транзакционной NTFS (TxF), включая *каталог журнала транзакций* `$TxfLog`, *каталог изоляции транзакций* `$Txf` и *каталог восстановления транзакций* `$Repair`. Каталог журнала транзакций содержит *базовый файл журнала TxF* `$TxfLog.blf` и любое количество файлов — контейнеров журнала в зависимости от размера журнала транзакций, но всегда как минимум два: один для потока журнала Kernel Transaction Manager (KTM) (файл `$TxfLogContainer0000000000000001`) и один для потока журнала TxF (файл `$TxfLogContainer0000000000000002`). Каталог журнала транзакций также содержит *поток старых страниц TxF* `$Tops`, который будет описан позже.

ЭКСПЕРИМЕНТ. Просмотр информации NTFS

Можно использовать встроенную программу командной строки Fsutil.exe для просмотра информации о томе NTFS, включая размещение и размер MFT и зоны MFT:

```
d:\>fsutil fsinfo ntfsinfo d:
NTFS Volume Serial Number :        0x48323940323933f2
NTFS Version   :                   3.1
LFS Version    :                   2.0
Number Sectors :                   0x000000011c5f6fff
Total Clusters :                   0x00000000238bedff
Free Clusters  :                   0x000000001a6e5925
Total Reserved :                   0x00000000000011cd
Bytes Per Sector  :                512
Bytes Per Physical Sector :        4096
Bytes Per Cluster :                4096
Bytes Per FileRecord Segment    :  4096
Clusters Per FileRecord Segment :  1
Mft Valid Data Length :            0x0000000646500000
Mft Start Lcn  :                   0x00000000000c0000
Mft2 Start Lcn :                   0x0000000000000002
Mft Zone Start :                   0x00000000069f76e0
Mft Zone End   :                   0x00000000069f7700
Max Device Trim Extent Count :     4294967295
Max Device Trim Byte Count :       0x10000000
Max Volume Trim Extent Count :     62
Max Volume Trim Byte Count :       0x10000000
Resource Manager Identifier :      81E83020-E6FB-11E8-B862-D89EF33A38A7
```

В этом примере том `D:` использует файловые записи в 4 Кбайт — MFT-записи — на диске с секторами в стандартные 4 Кбайт, эмулирующем старые сектора в 512 байтов, и применяет кластеры в 4 Кбайт.

Номера файловых записей

Файл в томе NTFS идентифицируется 64-битным значением, называемым *номером файловой записи*, который состоит из номера файла и номера последовательности. Номер файла соответствует позиции файловой записи файла в MFT минус 1 или позиции базовой файловой записи минус 1, если файл имеет более одной файловой

записи. Порядковый номер, увеличивающийся каждый раз, когда позиция файловой записи MFT используется повторно, позволяет NTFS выполнять внутренние проверки согласованности. Номер файловой записи показан на рис. 11.33.

Рис. 11.33. Номер записи файла

Файловые записи

Вместо того чтобы рассматривать файл как хранилище текстовых или двоичных данных, NTFS хранит файл как набор пар атрибутов и значений, и одной из этих пар будут содержащиеся в файле данные — так называемый *неименованный атрибут данных*. Другие атрибуты, составляющие файл, включают его имя, информацию о временной метке и, возможно, дополнительные именованные атрибуты данных. На рис. 11.34 показана запись MFT для небольшого файла.

Каждый атрибут файла хранится в файле в виде отдельного потока байтов. Строго говоря, NTFS читает и записывает не файлы, а потоки атрибутов. NTFS предоставляет следующие операции с атрибутами: создание, удаление, чтение диапазона байтов и запись диапазона байтов. Службы чтения и записи обычно работают с неименованным атрибутом данных файла. Однако вызывающая сторона может указать другой атрибут данных, используя синтаксис именованного потока данных.

В табл. 11.6 перечислены атрибуты файлов в томе NTFS. Не все они применяются в отдельно взятом файле. Каждый атрибут в файловой системе NTFS может быть безымянным или иметь имя. Пример именованного атрибута — `$LOGGED_UTILITY_STREAM`, его используют для различных целей разные компоненты NTFS. В табл. 11.7 перечислены возможные имена атрибута `$LOGGED_UTILITY_STREAM` и соответствующие им назначения.

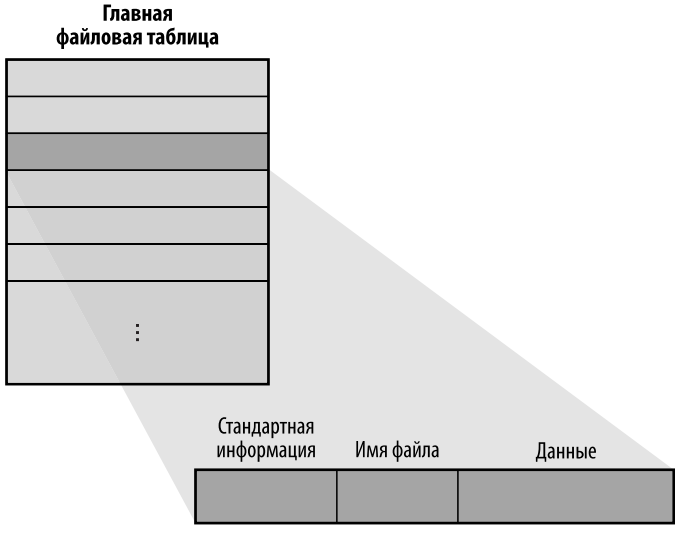


Рис. 11.34. Запись MFT для небольшого файла

Таблица 11.6. Атрибуты файлов NTFS

Атрибут	Тип атрибута	Резидентный?	Описание
Информация о томе	$VOLUME_INFORMATION, $VOLUME_NAME	Всегда, всегда	Эти атрибуты присутствуют только в файле метаданных $Volume. Они хранят информацию о версии и метке тома
Стандартная информация	$STANDARD_INFORMATION	Всегда	Атрибуты файлов, такие как «только для чтения», «архив» и т. д., и временные метки, включая время создания или последнего изменения файла
Имя файла	$FILE_NAME	Возможно	Имя файла в символах Юникода 1.0. Файл может иметь несколько атрибутов имени файла, как происходит, когда существует жесткая ссылка на файл или файл с длинным именем имеет автоматически созданное короткое имя для обеспечения доступа приложений MS-DOS и 16-битной Windows
Дескриптор безопасности	$SECURITY_DESCRIPTOR	Возможно	Этот атрибут присутствует для обратной совместимости с предыдущими версиями NTFS и редко используется в текущей версии NTFS (3.1). NTFS хранит почти все дескрипторы безопасности в файле метаданных $Secure, разделяя их между файлами и каталогами, имеющими одинаковые настройки. Предыдущие версии NTFS хранили частную информацию о дескрипторах безопасности в каждом файле и каталоге. Некоторые файлы по-прежнему содержат атрибут $SECURITY_DESCRIPTOR, например $Boot
Данные	$DATA	Возможно	Содержимое файла. В NTFS файл по умолчанию имеет один неименованный атрибут данных и, возможно, дополнительные именованные атрибуты данных, то есть у него может быть несколько потоков данных. Каталог не имеет атрибута данных по умолчанию, но может иметь дополнительные именованные атрибуты данных. Именованные потоки данных могут использоваться даже для определенных системных целей. Например, поток Storage Reserve Area Table (SRAT) $SRAT задействуется службой хранения для резервирования пространства в томе. Этот атрибут применяется только к файлу метаданных $Bitmap. Резервы хранения описываются далее в этой главе

Продолжение ⇩

Таблица 11.6 *(продолжение)*

Атрибут	Тип атрибута	Резидентный?	Описание
Корень индекса, выделение индекса	$INDEX_ROOT, $INDEX_ ALLOCATION	Всегда, никогда	Три атрибута для реализации структур данных B-дерева, применяемых в каталогах, файлах безопасности, квотах и других метаданных
Список атрибутов	$ATTRIBUTE_LIST	Возможно	Список атрибутов, составляющих файл, и номер записи файла в MFT, где находится каждый атрибут. Этот атрибут присутствует, если для файла требуется более одной записи файла MFT
Битовая карта индекса	$BITMAP	Возможно	Этот атрибут служит для разных целей: для нерезидентных каталогов, где $INDEX_ALLOCATION всегда существует, битовая карта по мере роста B-дерева записывает, какие индексные блоки размером 4 Кбайт уже им используются, а какие свободны для применения в будущем. В MFT есть безымянный атрибут $Bitmap, который отслеживает, какие сегменты MFT используются, а какие свободны для применения в дальнейшем новыми файлами или существующими, которым требуется больше места
Идентификатор объекта	$OBJECT_ID	Всегда	16-байтовый идентификатор (GUID) для файла или каталога. Служба отслеживания ссылок присваивает идентификаторы объектов файлам ярлыков оболочки и исходным файлам ссылок OLE. NTFS предоставляет API, чтобы файлы и каталоги можно было открывать с помощью идентификатора объекта, а не имени файла
Информация о повторной обработке	$REPARSE_POINT	Возможно	В этом атрибуте хранятся данные о точке повторной обработки файла. Стыки NTFS и точки монтирования включают этот атрибут
Расширенные атрибуты	$EA, $EA_ INFORMATION	Возможно, всегда	Расширенные атрибуты представляют собой пары «имя — значение» и обычно не используются, но предоставляются для обратной совместимости с приложениями OS/2
Вспомогательный поток журналирования	$LOGGED_ UTILITY_STREAM	Возможно	Этот тип атрибутов может применяться для различных целей разными компонентами NTFS. Более подробную информацию см. в табл. 11.7

Таблица 11.7. Атрибут $LOGGED_UTILITY_STREAM

Атрибут	Тип атрибута	Резидентный?	Описание
Поток зашифрованных файлов	$EFS	Возможно	EFS хранит в этом атрибуте данные, которые используются для управления шифрованием файла, например зашифрованную версию ключа, необходимого для расшифровки файла, и список пользователей, имеющих право доступа к файлу
Резервная копия при шифровании онлайн	$EfsBackup	Возможно	Атрибут применяется при шифровании EFS онлайн для хранения фрагментов исходного потока зашифрованных данных
Транзакционные данные NTFS	$TXF_DATA	Возможно	Когда файл или каталог становится частью транзакции, TxF сохраняет также данные транзакции в атрибуте $TXF_DATA, например уникальный идентификатор транзакции файла
Желаемый класс хранения	$DSC	Да	Нужный класс хранения используется для закрепления файла на желательном уровне хранения. Дополнительные сведения см. в разделе «Поддержка NTFS для многоуровневых томов»

В табл. 11.6 приведены имена атрибутов, однако на самом деле атрибутам соответствуют численные коды типов, которые NTFS использует для их упорядочивания в файловой записи. Атрибуты файла в записи MFT упорядочены по этим кодам типов в порядке возрастания, причем некоторые типы атрибутов встречаются не один раз, например, если файл имеет несколько атрибутов данных или несколько имен. Все возможные типы атрибутов и их имена перечислены в файле метаданных `$AttrDef`.

Каждый атрибут в файловой записи идентифицируется кодом типа атрибута, имеет значение и необязательное имя. Значение атрибута — это поток байтов, составляющих атрибут. Например, значение атрибута `$FILE_NAME` — это имя файла, значение атрибута `$DATA` — байты, которые пользователь сохранил в файле.

Большинство атрибутов не имеют имен, хотя у тех, которые связаны с индексами, и у атрибута `$DATA` часто они есть. Имена различают несколько атрибутов одного типа, которые могут быть включены в файл. Например, файл с именованным потоком данных имеет два атрибута `$DATA`: неименованный атрибут `$DATA`, хранящий неименованный поток данных по умолчанию, и именованный атрибут `$DATA`, имеющий имя альтернативного потока и хранящий данные именованного потока.

Имена файлов

И NTFS, и FAT допускают, чтобы каждое имя файла в пути состояло из 255 символов. Имена файлов могут содержать символы Юникода, а также несколько точек и встроенных пробелов. Однако файловая система FAT, поставляемая с MS-DOS, ограничивает имя восемью неюникодными символами, за которыми следуют точка

и трехсимвольное расширение. На рис. 11.35 наглядно представлены различные файлы пространства имен, которые поддерживает Windows, и показано, как они пересекаются.

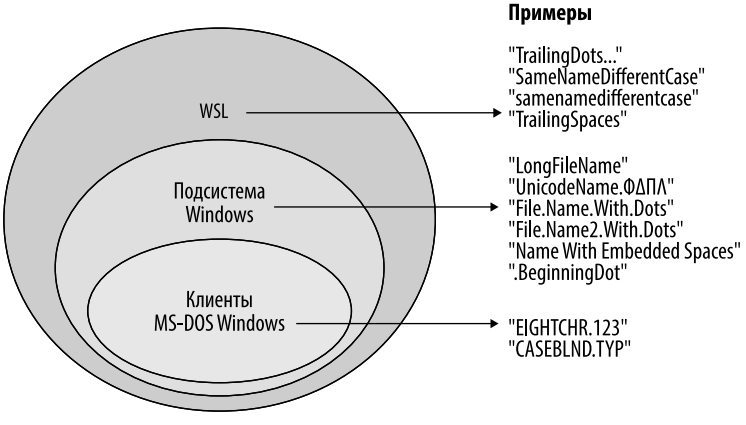


Рис. 11.35. Пространства имен файлов Windows

Windows Subsystem for Linux (WSL) требует самого большого пространства имен из всех сред выполнения приложений, поддерживаемых Windows, поэтому пространство имен NTFS эквивалентно пространству имен WSL. WSL может создавать имена, которые не видны приложениям Windows и MS-DOS, включая имена с точками или пробелами в конце. Как правило, создание файла с применением большого пространства имен POSIX не проблема, потому что это делается, только когда предполагается, что приложения WSL будут использовать этот файл.

Однако отношения между 32-разрядными приложениями Windows и приложениями MS-DOS и 16-разрядной Windows гораздо теснее. Область Windows на рис. 11.35 представляет имена файлов, которые подсистема Windows может создавать в томе NTFS, но их не видят приложения MS-DOS и 16-разрядной Windows. В эту группу входят имена файлов, длина которых превышает формат 8.3 для имен MS-DOS, имена с символами Юникода, имена с несколькими символами точки или начальной точкой, а также имена с пробелами внутри. Когда создается файл с таким именем, по соображениям совместимости NTFS автоматически дает ему альтернативное имя в стиле MS-DOS. Windows показывает короткие имена, когда используется параметр `/x` в команде `dir`.

Имена файлов MS-DOS являются полнофункциональными псевдонимами файлов NTFS и хранятся в том же каталоге, что и длинные имена файлов. Запись MFT для файла с автоматически созданным именем файла MS-DOS показана на рис. 11.36.

Имя NTFS и созданное имя MS-DOS хранятся в одной файловой записи и, следовательно, относятся к одному и тому же файлу. Имя MS-DOS можно использовать для открытия, чтения, записи или копирования файла. Если пользователь переименовывает файл, задействуя либо длинное, либо короткое имя, то новое имя заменяет оба существующих. Если новое имя не является действительным

именем MS-DOS, то NTFS создает другое имя MS-DOS для файла. Обратите внимание на то, что NTFS создает имена файлов в стиле MS-DOS только для первого имени файла.

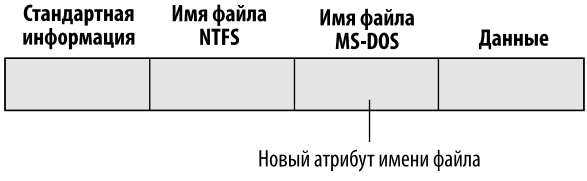


Рис. 11.36. Запись файла MFT с атрибутом имени файла MS-DOS

ПРИМЕЧАНИЕ Жесткие ссылки работают аналогичным образом. Когда создается жесткая ссылка на файл, NTFS добавляет еще один атрибут имени файла в файловую запись MFT файла и запись в атрибут Index Allocation каталога, в котором находится новая ссылка. Однако эти ситуации различаются в одном отношении. Когда пользователь удаляет файл, имеющий несколько имен (жестких ссылок), запись файла и он сам остаются на месте. Файл и его запись удаляются только при удалении последнего имени файла (жесткой ссылки). Но если файл имеет как имя NTFS, так и автоматически созданное имя MS-DOS, то пользователь может удалить его с помощью любого из них.

Вот алгоритм, который NTFS использует для создания имени MS-DOS из длинного имени файла. Он реализован в функции ядра `RtlGenerate8dot3Name` и может измениться в будущих версиях Windows. Эту функцию применяют и другие драйверы, такие как CDFS, FAT и файловые системы сторонних производителей.

1. Удалите из длинного имени все символы, недопустимые для имен MS-DOS, включая пробелы и символы Юникода. Удалите точки в начале и конце имени. Удалите все точки внутри, кроме последней.
2. Обрежьте часть строки перед точкой, если таковая имеется, до шести символов — их может быть уже шесть или меньше, поскольку этот алгоритм применяется и тогда, когда в имени присутствует любой символ, недопустимый в MS-DOS. Если символов два или меньше, сгенерируйте и прибавьте четырехсимвольную шестнадцатеричную строку контрольной суммы. Добавьте справа строку `~n`, где `n` — число, начиная с 1, применяемое для различения разных файлов, усеченных до одного и того же имени. Обрежьте часть строки после точки, если она есть, до трех символов.
3. Запишите результат прописными буквами. MS-DOS не различает регистров, и этот шаг гарантирует, что NTFS не создаст новое имя, отличающееся от старого только регистром.
4. Если получившееся имя дублирует существующее в каталоге, то увеличьте строку `~n`. Если `n` больше 4, а контрольная сумма еще не была прибавлена, то обрежьте часть строки перед точкой до двух символов, сгенерируйте и прибавьте четырехсимвольную шестнадцатеричную контрольную сумму.

В табл. 11.8 приведены длинные имена файлов Windows с рис. 11.35 и их NTFS-версии в стиле MS-DOS. Алгоритм и примеры, приведенные на рисунке, должны дать представление о том, как выглядят имена файлов, созданные NTFS в стиле MS-DOS.

ПРИМЕЧАНИЕ Начиная с Windows 8.1 по умолчанию во всех незагружаемых томах NTFS отключено создание коротких имен. Его можно отключить даже в старых версиях Windows, установив для параметра реестра HKLM\SYSTEM\CurrentControlSet\Control\FileSystem\NtfsDisable8dot3NameCreation DWORD-значение 1 и перезагрузив машину. Однако это может нарушить совместимость со старыми приложениями.

Таблица 11.8. Создаваемые NTFS имена файлов

Длинное имя Windows	Созданное NTFS короткое имя
LongFileName	LONGFI~1
UnicodeName.FDPL	UNICOD~1
File.Name.With.Dots	FILENA~1.DOT
File.Name2.With.Dots	FILENA~2.DOT
File.Name3.With.Dots	FILENA~3.DOT
File.Name4.With.Dots	FILENA~4.DOT
File.Name5.With.Dots	FIF596~1.DOT
Name With Embedded Spaces	NAMEWI~1
.BeginningDot	BEGINN~1
25¢.two characters	255440~1.TWO
©	6E2D~1

Туннелирование

NTFS использует концепцию *туннелирования* для обеспечения совместимости со старыми программами, которые зависят от файловой системы, кэширующей определенные метаданные файла в течение определенного времени даже после того, как файл исчезнет, например, когда он удален или переименован. При туннелировании любой новый файл, созданный с тем же именем, что и исходный, и в течение определенного времени, будет сохранять часть метаданных. Суть этого заключается в том, чтобы повторить поведение, ожидаемое программами MS-DOS при использовании метода *безопасного сохранения*, при котором измененные данные копируются во временный файл, оригинальный файл удаляется, а затем временному файлу дается имя оригинального. В этом случае ожидается, что переименованный временный файл будет выглядеть так же, как исходный, в противном случае время создания будет постоянно обновляться при каждом изменении — именно так сохраняется время изменения.

NTFS использует туннелирование таким образом, что когда имя файла удаляется из каталога, его длинное и короткое имена, а также время создания сохраняются в кэше. Когда в каталог добавляется новый файл, кэш просматривается на предмет наличия туннелированных данных для восстановления. Поскольку эти операции применяются к каталогам, каждый экземпляр каталога имеет собственный кэш, удаляемый при удалении каталога.

NTFS будет применять туннелирование для следующей серии операций, если используемые имена приводят к удалению и повторному созданию одного и того же имени файла:

- удалить + создать;
- удалить + переименовать;
- переименовать + создать;
- переименовать + переименовать.

По умолчанию NTFS сохраняет кэш туннелирования в течение 15 с, но можно изменить это время, задав новый параметр `MaximumTunnelEntryAgeInSeconds` в разделе реестра HKLM\SYSTEM\CurrentControlSet\Control\FileSystem. Туннелирование можно полностью отключить, создав новый параметр `MaximumTunnelEntries` и установив его равным `0`. Однако в этом случае старые приложения могут перестать работать, если они полагаются на совместимое поведение. В томах NTFS, у которых отключено создание коротких имен (см. предыдущий раздел), туннелирование отключено по умолчанию.

Чтобы увидеть туннелирование в действии, проведите следующий простой эксперимент в командной строке.

1. Создайте файл с именем `file1`.
2. Ждите более 15 с — тайм-аут туннельного кэша по умолчанию.
3. Создайте файл с именем `file2`.
4. Выполните команду `dir /TC`. Обратите внимание на время создания.
5. Переименуйте `file1` в `file`.
6. Переименуйте `file2` в `file1`.
7. Выполните команду `dir /TC`. Обратите внимание на то, что время создания одно и то же.

Резидентные и нерезидентные атрибуты

Если файл небольшой, то все его атрибуты и их значения, например данные, помещаются в файловую запись, описывающую файл. Если значение атрибута хранится в MFT — либо в основной файловой записи файла, либо в расположенной в другом месте MFT записи расширения, — то атрибут называется *резидентным*. На рис. 11.37, например, все атрибуты резидентные. Несколько атрибутов определены как всегда резидентные, чтобы NTFS могла находить нерезидентные атрибуты. Например, атрибуты стандартной информации и корня индекса всегда резидентны.

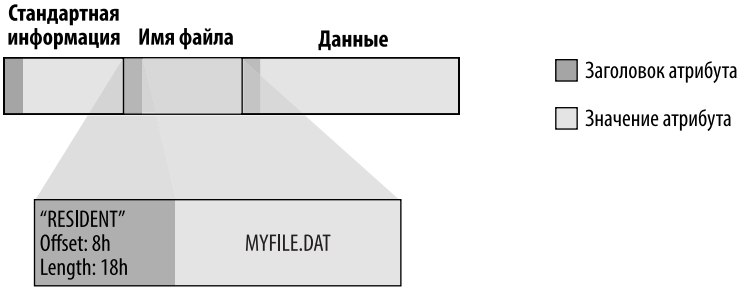


Рис. 11.37. Заголовок и значение резидентного атрибута

Каждый атрибут начинается со стандартного заголовка, содержащего информацию о нем, которую NTFS использует для управления атрибутами в общем виде. В заголовке, который всегда резидентен, записывается, резидентно или нерезидентно значение атрибута. Для резидентных атрибутов заголовок содержит также смещение от заголовка к значению атрибута и длину значения атрибута, как показано на рис. 11.37 для атрибута имени файла.

Когда значение атрибута хранится непосредственно в MFT, время, необходимое NTFS для доступа к нему, значительно сокращается. Вместо того чтобы искать файл в таблице, а затем читать последовательность данных о размещении, чтобы найти данные файла, как это делает, например, файловая система FAT, NTFS обращается к диску один раз и немедленно извлекает данные.

Атрибуты небольшого каталога или файла могут быть резидентными в MFT (рис. 11.38). Для небольшого каталога атрибут корня индекса содержит индекс номеров записей файлов и подкаталогов в каталоге, организованный в виде B-дерева.

Рис. 11.38. Запись файла в MFT для небольшого каталога

Конечно, многие файлы и каталоги невозможно втиснуть в запись MFT размером 1 Кбайт или 4 Кбайт. Если значение определенного атрибута, например атрибута данных файла, слишком велико, чтобы уместить его в запись файла MFT, то NTFS выделяет кластеры для значения атрибута за пределами MFT. Непрерывная группа кластеров называется *прогоном* или *экстентом*. Если значение атрибута впоследствии увеличивается, например, из-за того что пользователь добавляет данные в файл, то NTFS выделяет еще один прогон для дополнительных данных. Атрибуты, значения которых хранятся в прогонах, а не в MFT, называются *нерезидентными*. Файловая система решает, резидентным или нерезидентным является конкретный атрибут. Местоположение этих данных прозрачно для обращающегося к ним процесса.

Если атрибут нерезидентный, как, например, атрибут данных для большого файла, то его заголовок содержит информацию, которая нужна NTFS для поиска значения атрибута на диске. На рис. 11.39 показан нерезидентный атрибут данных, хранящийся в двух прогонах.

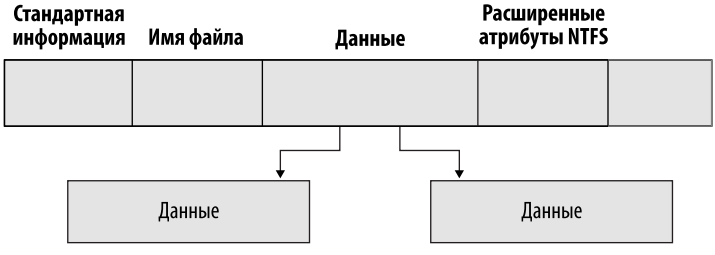


Рис. 11.39. Запись файла в MFT для большого файла с двумя прогонами данных

Среди стандартных атрибутов нерезидентными могут быть только те, которые способны расти. Для файлов такими являются данные и список атрибутов (не показаны на рис. 11.39). Стандартные атрибуты информации и имени файла всегда резидентны.

В большом каталоге также могут быть нерезидентные атрибуты или части атрибутов (рис. 11.40). В этом примере в файловой записи MFT недостаточно места для хранения B-дерева, содержащего индекс файлов, находящихся в этом большом каталоге. Часть индекса хранится в корневом атрибуте индекса, а остальная часть — в нерезидентных прогонах, называемых *индексными распределениями*. Атрибуты index root, index allocation и bitmap показаны здесь в упрощенном виде. Более подробно они описаны в следующем разделе. Атрибуты стандартной информации и имени файла всегда резидентны. Заголовок и по крайней мере часть значения атрибута корня индекса также резидентны для каталогов.

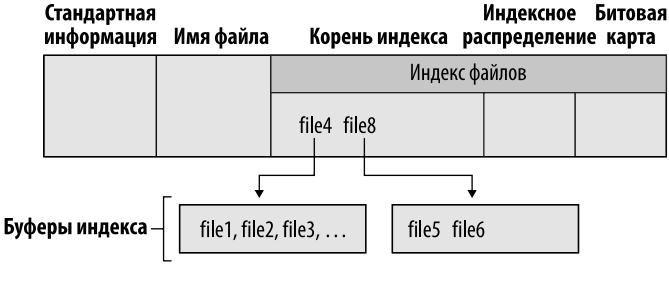


Рис. 11.40. Запись файла в MFT для большого каталога с нерезидентным индексом имени файла

Когда значение атрибута не помещается в файловую запись MFT и требуется его выделить, NTFS отслеживает прогоны с помощью пар сопоставлений VCN и LCN. LCN представляют последовательность кластеров во всем томе от 0 до *n*. VCN нумеруют кластеры, принадлежащие определенному файлу, от 0 до *m*. Например, кластеры в прогонах нерезидентного атрибута данных нумеруются, как показано на рис. 11.41.

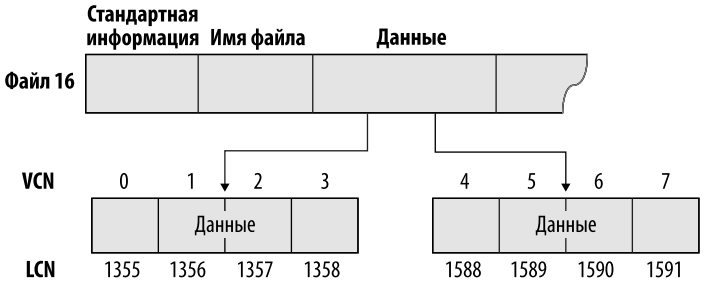


Рис. 11.41. VCN для нерезидентного атрибута данных

Если бы в этом файле было больше двух прогонов, то нумерация третьего прогона начиналась бы с VCN 8. Как показано на рис. 11.42, заголовок атрибута данных содержит сопоставления VCN и LCN для двух прогонов, что позволяет NTFS легко найти распределение на диске.

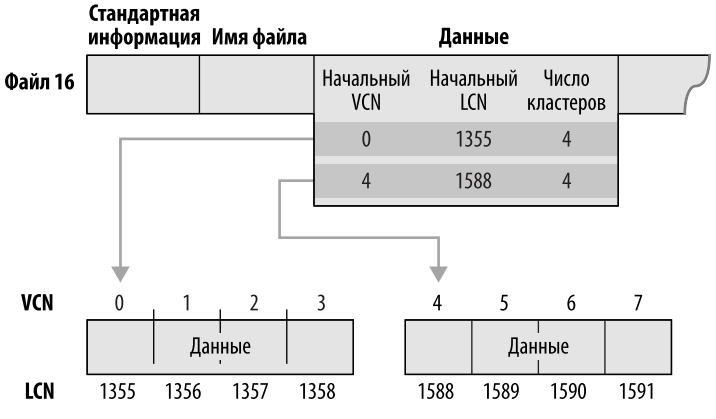


Рис. 11.42. Сопоставление VCN и LCN для нерезидентного атрибута данных

Хотя на рис. 11.41 показаны только прогоны данных, в них могут храниться и другие атрибуты, если в файловой записи MFT недостаточно места для их размещения. Если же в конкретном файле слишком много атрибутов, чтобы уместить их в записи MFT, то для размещения дополнительных атрибутов или заголовков нерезидентных атрибутов используется вторая запись MFT. В этом случае добавляется атрибут, называемый *списком атрибутов*. Он содержит имя и код типа каждого из атрибутов файла, а также номер записи для файла в MFT, в которой находится атрибут. Этот атрибут предназначен для тех случаев, когда все атрибуты файла не помещаются в файловую запись файла или когда он становится настолько большим или фрагментированным, что одна запись MFT не может содержать множества сопоставлений VCN и LCN, необходимых для поиска всех его прогонов. Файлы с более чем 200 прогонами обычно требуют использования списка атрибутов. Подведем итог: заголовки атрибутов всегда содержатся внутри записей файла в MFT, но значение атрибута может быть расположено за пределами MFT в одном или нескольких экстентах.

Сжатие данных и разреженные файлы

NTFS поддерживает сжатие файлов, каталогов или томов с помощью разновидности алгоритма LZ77, известной как LZNT1. Сжатие NTFS выполняется только для пользовательских данных, но не для метаданных файловой системы. В Windows 8.1 и более поздних версиях файлы можно сжимать также с помощью нового набора алгоритмов, включающего XPRESS и наиболее компактный LZX, в том числе с применением блоков размером 4, 8 или 16 Кбайт, в порядке возрастания скорости. Этот тип сжатия, который доступен с помощью команды `compact` оболочки и API File Provider, использует драйвер фильтра файловой системы (Windows Overlay Filter, WOF) Wof.sys, задействующий альтернативный поток данных NTFS и разреженные файлы, и не является частью драйвера NTFS как такового. WOF выходит за рамки тематики этой книги, подробнее о нем можно прочитать по адресу https://devblogs.microsoft.com/oldnewthing/20190618-00/?p=102597.

Узнать, сжат ли том, можно с помощью функции Windows `GetVolumeInformation`. Чтобы получить фактический размер сжатого файла, используйте функцию Windows `GetCompressedFileSize`. Наконец, чтобы проверить или изменить настройки сжатия для файла или каталога, задействуйте функцию Windows `DeviceIoControl`. См. коды управления файловой системой `FSCTL_GET_COMPRESSION` и `FSCTL_SET_COMPRESSION`. Помните: хотя установка состояния сжатия файла приводит к его немедленному сжатию или распаковке, установка состояния сжатия каталога или тома не вызывает их немедленного сжатия или распаковки. Вместо этого она задает состояние сжатия по умолчанию, которое будет присвоено всем вновь созданным файлам и подкаталогам в этом каталоге или томе. Хотя если задать сжатие каталога с помощью страницы свойств каталога в Проводнике, то содержимое всего дерева каталогов будет сжато немедленно.

В следующем разделе рассмотрим сжатие в NTFS на примере простого случая сжатия разреженных данных. В последующих разделах поговорим о сжатии обычных файлов и файлов с разреженными данными.

ПРИМЕЧАНИЕ Сжатие NTFS не поддерживается для томов DAX или зашифрованных файлов.

Сжатие разреженных данных

Разреженные данные часто велики, но содержат небольшое количество ненулевых данных относительно своего размера. Один из примеров разреженных данных — разреженная матрица. Как говорилось ранее, NTFS использует VCN от 0 до *m* для перечисления кластеров файла. Каждый VCN сопоставляется с соответствующим LCN, который определяет расположение кластера на диске. На рис. 11.43 показаны прогоны (дисковые распределения) по умолчанию обычного несжатого файла, включая VCN и LCN, которым они соответствуют.

Рис. 11.43. Прогоны несжатого файла

Этот файл хранится в трех прогонах, каждый длиной четыре кластера, то есть всего 12 кластеров. На рис. 11.44 показана запись MFT для этого файла. Как говорилось ранее, для экономии места атрибут данных записи MFT, содержащий сопоставления VCN и LCN, записывает только одно сопоставление для каждого прогона, а не одно для каждого кластера. Однако обратите внимание на то, что с каждым VCN от 0 до 11 связан соответствующий LCN. Первая запись начинается с VCN 0 и охватывает четыре кластера, вторая запись начинается с VCN 4 и охватывает четыре кластера и т. д. Такой формат записи типичен для несжатого файла.

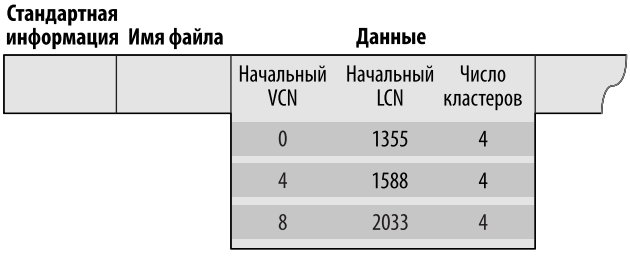


Рис. 11.44. Запись MFT для несжатого файла

Выбрав файл в томе NTFS для сжатия, пользователь применяет один из методов сжатия NTFS, который заключается в удалении из него длинных последовательностей нулей. Если данные файла разрежены, он обычно сжимается так, что занимает лишь часть дискового пространства, которое заполнял бы без сжатия. В дальнейшем при записи в файл NTFS выделяет пространство только для прогонов, содержащих ненулевые данные.

На рис. 11.45 показаны прогоны сжатого файла, содержащего разреженные данные. Обратите внимание на то, что для некоторых диапазонов VCN файла (16–31 и 64–127) на диске не выделено место.

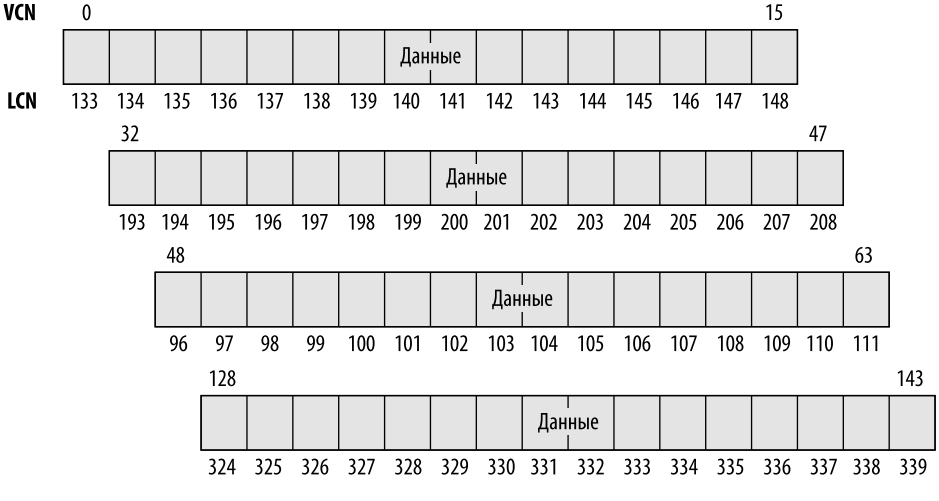


Рис. 11.45. Прогоны сжатого файла с разреженными данными

В записи MFT для этого сжатого файла отсутствуют блоки VCN, которые содержат нули и поэтому не имеют выделенного для них физического хранилища. Первая запись данных на рис. 11.46, например, начинается с VCN 0 и охватывает 16 кластеров. Вторая запись переходит к VCN 32 и охватывает 16 кластеров.

Когда программа считывает данные из сжатого файла, NTFS проверяет запись MFT, чтобы определить, покрывает ли сопоставление VCN и LCN считываемое место. Если программа читает из нераспределенной «дыры» в файле, это означает, что данные в этой части файла состоят из нулей, поэтому NTFS возвращает нули, не обращаясь в дальнейшем к диску. Если программа записывает ненулевые данные в «дыру», то NTFS прозрачно выделяет место на диске, а затем записывает туда данные. Этот прием очень эффективен для разреженных файлов, содержащих много нулевых данных.

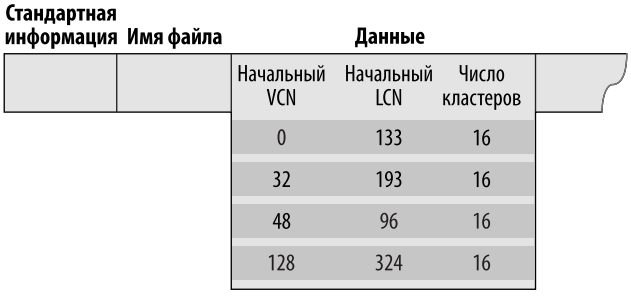


Рис. 11.46. Запись MFT для сжатого файла с разреженными данными

Сжатие неразреженных данных

Приведенный пример сжатия разреженного файла несколько искусственный. Он описывает сжатие, которое действует на целые части файла, заполненные нулями, но не затрагивает остальные данные файла. Однако в большинстве файлов данные неразреженные, но их все равно можно сжать, применив алгоритм сжатия.

В NTFS пользователь может задать сжатие для отдельных файлов или для всех файлов в каталоге. Новые файлы, созданные в каталоге, помеченном для сжатия, сжимаются автоматически — существующие файлы должны быть сжаты отдельно при программном включении сжатия с помощью команды `FSCTL_SET_COMPRESSION`. При сжатии файла NTFS делит необработанные данные файла на *единицы сжатия* длиной 16 кластеров — например, 128 Кбайт для кластеров по 8 Кбайт. Определенные последовательности данных в файле могут сжиматься не очень сильно или вообще не сжиматься, поэтому для каждой единицы сжатия в файле NTFS определяет, сэкономит ли ее сжатие хотя бы один кластер на диске. Если сжатие блока не освободит хотя бы один кластер, то NTFS выделяет прогон в 16 кластеров и записывает данные из этого блока на диск без сжатия. Если данные в блоке из 16 кластеров будут сжаты до 15 или менее кластеров, то NTFS выделяет только такое количество кластеров, которое необходимо для размещения сжатых данных, а затем записывает их на диск. На рис. 11.47 показано сжатие файла с четырьмя прогонами. Незаштрихованные области представляют собой реальные места хранения, которые файл занимает после сжатия. Первый, второй и четвертый прогоны были сжаты,

а третий — нет. Даже с одним несжатым прогоном сжатие этого файла позволило сэкономить 26 кластеров дискового пространства, то есть 41 %.

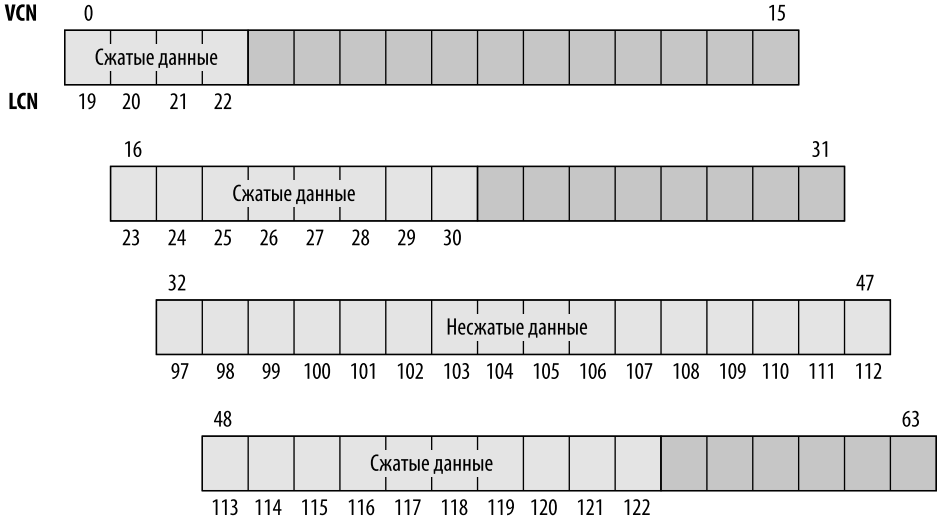


Рис. 11.47. Прогоны данных в сжатом файле

ПРИМЕЧАНИЕ Хотя на диаграммах в этой главе показаны смежные LCN, единицы сжатия не обязательно должны храниться в физически смежных кластерах. Прогоны, занимающие несмежные кластеры, создают несколько более сложные записи MFT, чем показанные на рис. 11.47.

При записи данных в сжатый файл NTFS гарантирует, что каждый прогон начинается на виртуальной границе 16 кластеров. Таким образом, начальный VCN каждого прогона кратен 16, а прогоны не длиннее 16 кластеров. При обращении к сжатым файлам NTFS считывает и записывает по крайней мере одну единицу сжатия за раз. Однако при записи сжатых данных NTFS старается хранить единицы сжатия в физически непрерывных местах, чтобы можно было прочитать их все за одну операцию ввода-вывода. Размер единицы сжатия NTFS в 16 кластеров был выбран для уменьшения внутренней фрагментации: чем больше единица сжатия, тем меньше суммарного дискового пространства требуется для хранения данных. Размер единицы сжатия в 16 кластеров представляет собой компромисс между созданием более компактных сжатых файлов и замедлением операций чтения для программ, которые обращаются к файлам случайно. Для каждого промаха кэша необходимо распаковать эквивалент 16 кластеров — пропуск кэша более вероятен при случайном доступе к файлу. На рис. 11.48 показана запись MFT для сжатого файла из рис. 11.47.

Отличие этого сжатого файла от сжатого файла с разреженными данными из предыдущего примера заключается в том, что длина трех сжатых прогонов в нем меньше 16 кластеров. Чтение этой информации из файловой записи MFT позволяет NTFS узнать, сжаты ли данные в файле. Любой прогон длиной менее 16 кластеров

содержит сжатые данные, которые NTFS должна распаковать при первом чтении данных в кэш. Прогон длиной ровно 16 кластеров не содержит сжатых данных и поэтому не требует распаковки.

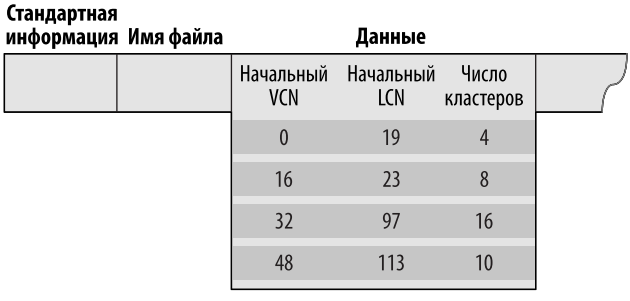


Рис. 11.48. Запись MFT для сжатого файла

Если данные в прогоне были сжаты, NTFS распаковывает их во временный буфер, а затем копирует в буфер вызывающей стороны. NTFS также загружает распакованные данные в кэш, что делает последующее чтение из того же прогона таким же быстрым, как и любое другое чтение из кэша. NTFS записывает в кэш все обновления файла, предоставляя системе отложенной записи асинхронно сжимать и записывать измененные данные на диск. Эта стратегия гарантирует, что запись в сжатый файл не вызовет более значительной задержки, чем запись в несжатый файл.

NTFS выделяет диск для сжатого файла по соседству, когда это возможно. Как показывают LCN, первые два прогона сжатого файла (см. рис. 11.47) физически непрерывны, как и два последних. Когда два или более прогона являются непрерывными, NTFS выполняет упреждающее чтение с диска, как это делается с данными в других файлах. Поскольку чтение и распаковка данных смежных файлов происходят асинхронно, до того как программа запросит данные, последующие операции чтения получают данные непосредственно из кэша, что значительно повышает производительность чтения.

Разреженные файлы

Разреженные файлы (отдельный тип файла в NTFS, отличный от описанных ранее файлов с разреженными данными) — это, по сути, сжатые файлы, для которых NTFS не применяет сжатие к неразреженным данным. Однако NTFS управляет данными прогонов разреженного файла в записи MFT так же, как и в случае со сжатыми файлами, состоящими из разреженных и неразреженных данных.

Файл журнала изменений

Файл журнала изменений `\$Extend\$UsnJrnl` — это разреженный файл, в котором NTFS хранит записи об изменениях файлов и каталогов. Такие приложения, как служба репликации файлов Windows (File Replication Service, FRS) и служба поиска Windows, используют журнал для реагирования на изменения файлов и каталогов по мере их возникновения.

Журнал хранит записи об изменениях в потоке данных `$J` и максимальный размер журнала в потоке данных `$Max`. Записи версионированы и содержат следующую информацию об изменении файла или каталога:

- время изменения;
- причина изменения (табл. 11.9);
- атрибуты файла или каталога;
- имя файла или каталога;
- номер записи файла или каталога в MFT;
- номер записи файла родительского каталога файла;
- идентификатор безопасности;
- порядковый номер обновления записи (update sequence number, USN);
- дополнительная информация об источнике изменений — пользователь, FRS и т. д.

Таблица 11.9. Причины изменения журнала изменений

Идентификатор	Причина
USN_REASON_DATA_OVERWRITE	Данные в файле или каталоге перезаписаны
USN_REASON_DATA_EXTEND	Данные добавлены в файл или каталог
USN_REASON_DATA_TRUNCATION	Данные в файле или каталоге урезаны
USN_REASON_NAMED_DATA_OVERWRITE	Данные в потоке данных файла перезаписаны
USN_REASON_NAMED_DATA_EXTEND	Данные в потоке данных файла расширены
USN_REASON_NAMED_DATA_ TRUNCATION	Данные в потоке данных файла урезаны
USN_REASON_FILE_CREATE	Создан новый файл или каталог
USN_REASON_FILE_DELETE	Файл или каталог удален
USN_REASON_EA_CHANGE	Расширенные атрибуты файла или каталога изменились
USN_REASON_SECURITY_CHANGE	Дескриптор безопасности для файла или каталога изменен
USN_REASON_RENAME_OLD_NAME	Файл или каталог переименован, это старое имя
USN_REASON_RENAME_NEW_NAME	Файл или каталог переименован, это новое имя
USN_REASON_INDEXABLE_CHANGE	Состояние индексирования для файла или каталога изменено независимо от того, будет ли служба индексирования обрабатывать этот файл или каталог
USN_REASON_BASIC_INFO_CHANGE	Атрибуты файла или каталога или метки времени изменены
USN_REASON_HARD_LINK_CHANGE	Для файла или каталога добавлена или удалена жесткая ссылка

Идентификатор	Причина
USN_REASON_COMPRESSION_CHANGE	Состояние сжатия для файла или каталога изменено
USN_REASON_ENCRYPTION_CHANGE	Состояние шифрования (EFS) включено или выключено для этого файла или каталога
USN_REASON_OBJECT_ID_CHANGE	Идентификатор объекта для этого файла или каталога изменен
USN_REASON_REPARSE_POINT_CHANGE	Точка повторной обработки для файла или каталога изменена либо новая точка повторной обработки (например, символическая ссылка) добавлена или удалена для файла или каталога
USN_REASON_STREAM_CHANGE	Новый поток данных добавлен в файл, удален из него или переименован
USN_REASON_TRANSACTED_CHANGE	Это значение добавляется через OR к причине изменения, чтобы указать, что изменение было результатом недавней фиксации транзакции TxF
USN_REASON_CLOSE	Дескриптор для файла или каталога закрыт, то есть это последнее изменение файла в данной серии операций
USN_REASON_INTEGRITY_CHANGE	Содержимое экстента (прогона) файла изменилось, поэтому связанный с ним поток целостности обновлен новой контрольной суммой. Этот идентификатор создается файловой системой ReFS
USN_REASON_DESIRED_STORAGE_CLASS_CHANGE	Событие создается драйвером файловой системы NTFS, когда поток перемещается с уровня хранения на уровень производительности или в обратном направлении

ЭКСПЕРИМЕНТ. Чтение журнала изменений

Встроенный инструмент %SystemRoot%\System32\Fsutil.exe можно использовать для создания, удаления или запроса информации о журнале с помощью встроенной утилиты Fsutil.exe, как показано далее:

```
d:\>fsutil usn queryjournal d:
Usn Journal ID   : 0x01d48f4c3853cc72
First Usn        : 0x0000000000000000
Next Usn         : 0x0000000000000a60
Lowest Valid Usn : 0x0000000000000000
Max Usn          : 0x7fffffffffff0000
Maximum Size     : 0x0000000000a00000
Allocation Delta : 0x0000000000200000
Minimum record version supported : 2
Maximum record version supported : 4
Write range tracking: Disabled
```

Вывод показывает максимальный размер журнала изменений для тома, 10 Мбайт, и его текущее состояние. В качестве простого эксперимента, чтобы увидеть, как NTFS записывает изменения в журнал, создайте файл `Usn.txt` в текущем каталоге, переименуйте его в `UsnNew.txt`, а затем сделайте дамп журнала с помощью Fsutil:

```
d:\>echo Hello USN Journal! > Usn.txt
d:\>ren Usn.txt UsnNew.txt
d:\>fsutil usn readjournal d:
...

Usn               : 2656
File name         : Usn.txt
File name length  : 14
Reason            : 0x00000100: File create
Time stamp        : 12/8/2018 15:22:05
File attributes   : 0x00000020: Archive
File ID           : 0000000000000000000c000000617912
Parent file ID    : 00000000000000000018000000617ab6
Source info       : 0x00000000: *NONE*
Security ID       : 0
Major version     : 3
Minor version     : 0
Record length     : 96

Usn               : 2736
File name         : Usn.txt
File name length  : 14
Reason            : 0x00000102: Data extend | File create
Time stamp        : 12/8/2018 15:22:05
File attributes   : 0x00000020: Archive
File ID           : 0000000000000000000c000000617912
Parent file ID    : 00000000000000000018000000617ab6
Source info       : 0x00000000: *NONE*
Security ID       : 0
Major version     : 3
Minor version     : 0
Record length     : 96

Usn               : 2816
File name         : Usn.txt
File name length  : 14
Reason            : 0x80000102: Data extend | File create | Close
Time stamp        : 12/8/2018 15:22:05
File attributes   : 0x00000020: Archive
File ID           : 0000000000000000000c000000617912
Parent file ID    : 00000000000000000018000000617ab6
Source info       : 0x00000000: *NONE*
Security ID       : 0
Major version     : 3
Minor version     : 0
Record length     : 96

Usn               : 2896
File name         : Usn.txt
File name length  : 14
Reason            : 0x00001000: Rename: old name
Time stamp        : 12/8/2018 15:22:15
```

```
File attributes   : 0x00000020: Archive
File ID           : 0000000000000000000c000000617912
Parent file ID    : 00000000000000000018000000617ab6
Source info       : 0x00000000: *NONE*
Security ID       : 0
Major version     : 3
Minor version     : 0
Record length     : 96

Usn               : 2976
File name         : UsnNew.txt
File name length  : 20
Reason            : 0x00002000: Rename: new name
Time stamp        : 12/8/2018 15:22:15
File attributes   : 0x00000020: Archive
File ID           : 0000000000000000000c000000617912
Parent file ID    : 00000000000000000018000000617ab6
Source info       : 0x00000000: *NONE*
Security ID       : 0
Major version     : 3
Minor version     : 0
Record length     : 96

Usn               : 3056
File name         : UsnNew.txt
File name length  : 20
Reason            : 0x80002000: Rename: new name | Close
Time stamp        : 12/8/2018 15:22:15
File attributes   : 0x00000020: Archive
File ID           : 0000000000000000000c000000617912
Parent file ID    : 00000000000000000018000000617ab6
Source info       : 0x00000000: *NONE*
Security ID       : 0
Major version     : 3
Minor version     : 0
Record length     : 96
```

Эти записи отражают отдельные операции по изменению, задействованные в операциях, лежащих в основе операций командной строки. Если журнал изменений не включен для тома — это происходит, в частности, с несистемными томами, где никакие приложения не запрашивали уведомления об изменении файлов или создании журнала USN, — то его можно легко создать с помощью следующей команды (в данном примере запрошен журнал объемом 10 Мбайт):

```
d:\ >fsutil usn createJournal d: m=10485760 a=2097152
```

Журнал является разреженным, поэтому никогда не переполняется. Когда размер журнала на диске превышает максимальный размер, определенный для файла, NTFS просто начинает обнулять данные файла, предшествующие окну информации об изменениях, размер которого равен максимальному размеру журнала (рис. 11.49). Чтобы предотвратить непрерывное изменение размера, когда приложение постоянно превышает размер журнала, NTFS уменьшает журнал только тогда, когда он в два раза превышает максимальный сконфигурированный размер, определенный приложением.

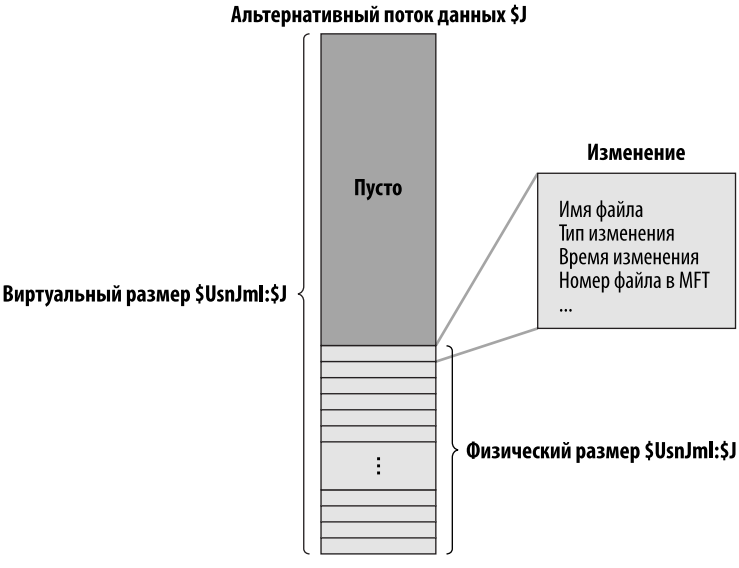


Рис. 11.49. Выделение пространства для журнала изменений $UsnJrnl

Индексирование

В NTFS каталог файлов — это просто указатель их имен, то есть коллекция имен файлов вместе с их номерами записей, организованная в виде B-дерева. Чтобы создать каталог, NTFS индексирует атрибуты имен файлов в каталоге. Запись MFT для корневого каталога тома показана на рис. 11.50.

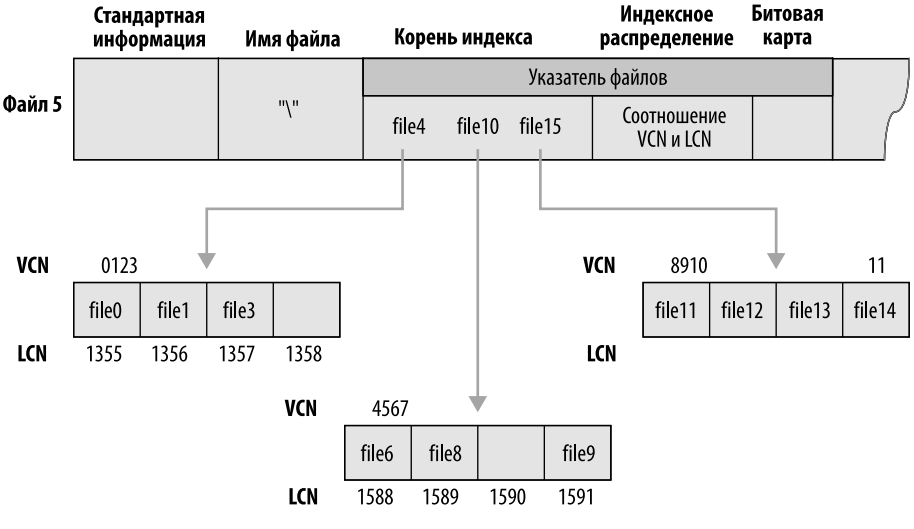


Рис. 11.50. Указатель имен файлов для корневого каталога тома

Концептуально запись MFT для каталога содержит в его атрибуте корня индекса отсортированный список файлов в каталоге. Однако для больших каталогов имена файлов фактически хранятся в индексных буферах фиксированного размера объемом 4 Кбайт, являющихся нерезидентными значениями атрибута *index allocation*, которые содержат и упорядочивают имена файлов. Индексные буферы реализуют структуру данных в виде B-дерева, которая минимизирует количество обращений к диску, необходимых для поиска определенного файла, особенно для больших каталогов. Атрибут *index root* определяет первый уровень B-дерева и корневые подкаталоги и указывает на индексные буферы, содержащие следующий уровень, например дополнительные подкаталоги или файлы.

На рис. 11.50 показаны только имена файлов в атрибуте index root и буферах индекса (например, *file6*), но каждая запись в индексе содержит также номер записи в MFT, в которой описан файл, и информацию о временной метке и размере файла. NTFS дублирует временные метки и информацию о размере файла из записи MFT. Эта техника, используемая в FAT и NTFS, требует записи обновленной информации в двух местах. Тем не менее она значительно повышает скорость просмотра каталогов, так как позволяет файловой системе показывать временные метки и размер каждого файла, не открывая каждый файл в каталоге.

Атрибут index allocation сопоставляет VCN прогонов индексных буферов с LCN, указывающими местоположение индексных буферов на диске, а атрибут bitmap отслеживает, какие VCN в индексных буферах используются, а какие свободны. На рис. 11.50 показана одна запись файла на VCN, то есть на кластер, но на самом деле записи имен файлов упакованы в каждый кластер. Каждый индексный буфер размером 4 Кбайт обычно содержит от 20 до 30 записей имен файлов в зависимости от длины этих имен в каталоге.

Структура данных типа B-дерево — это тип сбалансированного дерева, которое идеально подходит для организации отсортированных данных, хранящихся на диске, поскольку минимизирует количество обращений к диску, необходимых для поиска записи. В MFT атрибут index root каталога содержит несколько имен файлов, выступающих в качестве индексов второго уровня B-дерева. В атрибуте index root каждое имя файла имеет связанный необязательный указатель, который ведет на индексный буфер. Индексный буфер содержит имена файлов с лексикографическими значениями, которые меньше его собственного. На рис. 11.50, например, `файл4` — это запись первого уровня в B-дереве. Она указывает на индексный буфер, содержащий имена файлов, которые лексикографически меньше нее самой, — `файл0`, `файл1` и `файл3`. Обратите внимание на то, что имена `файл1`, `файл3` и т. д., используемые в этом примере, не являются буквальными именами файлов — это имена для обозначения взаимного расположения файлов, которые лексикографически упорядочены в соответствии с показанной последовательностью.

Хранение имен файлов в B-деревьях дает несколько преимуществ. Поиск в каталоге происходит быстро, поскольку имена хранятся в отсортированном порядке. И когда программное обеспечение более высокого уровня перечисляет файлы в каталоге, NTFS возвращает уже отсортированные имена. Наконец, поскольку B-деревья склонны расти вширь, а не вглубь, скорость поиска в NTFS не уменьшается по мере роста каталогов.

NTFS обеспечивает также общую поддержку индексирования данных, помимо имен файлов, а некоторые функции NTFS, в том числе идентификаторы объектов, отслеживание квот и консолидированная безопасность, используют индексирование для управления внутренними данными.

Указатели в виде B-деревьев — это обобщенный механизм NTFS, они применяются для организации дескрипторов безопасности, идентификаторов безопасности, идентификаторов объектов, записей дисковых квот и точек повторной обработки. Каталоги называют *указателями имен файлов*, в то время как другие типы указателей известны как *указатели представлений*.

Идентификаторы объектов

Помимо присвоенного файлу или каталогу идентификатора объекта в атрибуте `$OBJECT_ID` записи MFT, NTFS хранит также соответствие между идентификаторами объектов и номерами записей файлов в указателе `$O` файла метаданных `\$Extend\$ObjId`. В нем записи сопоставляются по идентификатору объекта, который является GUID, из-за чего NTFS может быстро находить файл по его идентификатору. Эта возможность позволяет приложениям открывать файл или каталог по идентификатору объекта, используя встроенный API `NtCreateFile` с флагом `FILE_OPEN_BY_FILE_ID`. На рис. 11.51 показано соответствие файла метаданных `$ObjId` и атрибутов `$OBJECT_ID` в записях MFT.

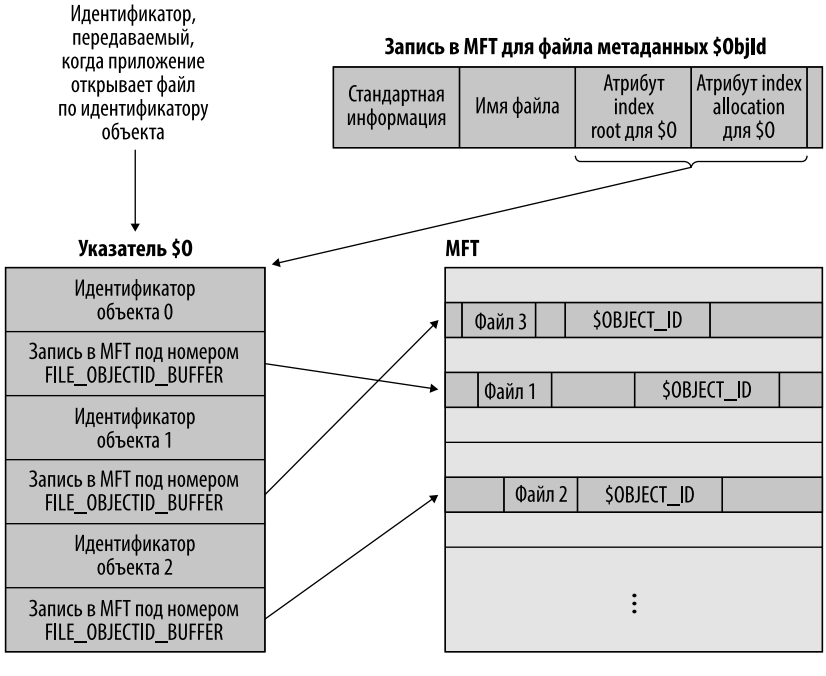


Рис. 11.51. Соотношение между $ObjId и $OBJECT_ID

Контроль над квотами

NTFS хранит информацию о квотах в файле метаданных `\$Extend\$Quota`, который состоит из именованных атрибутов index root указателей `$O` и `$Q`. На рис. 11.52 показана организация этих указателей. Подобно тому как NTFS присваивает каждому дескриптору безопасности уникальный внутренний идентификатор безопасности, она присваивает каждому пользователю уникальный идентификатор пользователя. Когда администратор определяет информацию о квотах для пользователя, NTFS назначает идентификатор пользователя, который соответствует его SID. В указателе `$O` NTFS создает запись, сопоставляющую SID с идентификатором пользователя, и сортирует указатель по SID, а в указателе `$Q` `NTFS` создает запись управления квотами. Эта запись содержит значение квоты пользователя, а также объем дискового пространства, занимаемого им в томе.

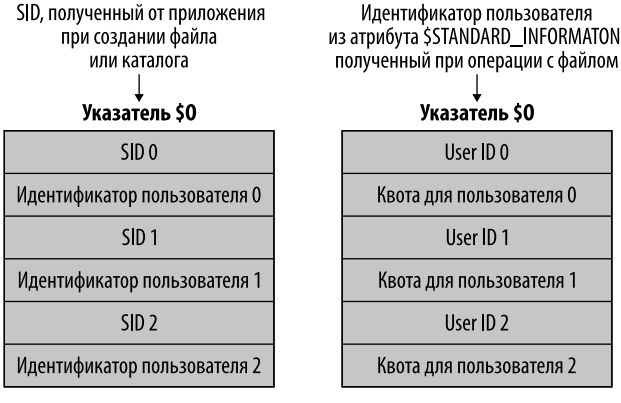


Рис. 11.52. Индексирование $Quota

Когда приложение создает файл или каталог, NTFS получает SID пользователя приложения и ищет связанный с ним идентификатор пользователя в указателе `$O`. Далее она записывает идентификатор пользователя в атрибут `$STANDARD_INFORMATION` нового файла или каталога, который засчитывает все дисковое пространство, выделенное файлу или каталогу, в квоту этого пользователя. Затем NTFS просматривает запись о квоте в указателе `$Q` и определяет, приводит ли новое выделение к превышению пользователем порога предупреждения или лимита. Если новое выделение приводит к тому, что пользователь превышает пороговое значение, то NTFS предпринимает соответствующие действия, например, регистрирует событие в системном журнале событий или не разрешает пользователю создавать файл или каталог. При изменении размера файла или каталога NTFS обновляет запись контроля над квотами, связанную с идентификатором пользователя и хранящуюся в атрибуте `$STANDARD_INFORMATION`. NTFS применяет индексацию обобщенного B-дерева для эффективного сопоставления идентификаторов пользователей с SID учетных записей, а при известном идентификаторе пользователя — для эффективного поиска информации о контроле над квотами пользователя.

Консолидированная безопасность

NTFS всегда поддерживает систему безопасности, что позволяет администратору определять, какие пользователи могут, а какие не могут получить доступ к отдельным файлам и каталогам. NTFS оптимизирует использование диска для дескрипторов безопасности, задействуя центральный файл метаданных `$Secure` для хранения только одного экземпляра каждого дескриптора безопасности в томе.

Файл `$Secure` содержит два индексных атрибута: `$SDH` (Security Descriptor Hash) и `$SII` (Security ID Index) — и атрибут потока данных `$SDS` (Security Descriptor Stream) (рис. 11.53). NTFS присваивает каждому уникальному дескриптору безопасности в томе внутренний идентификатор безопасности NTFS (не путайте с идентификатором SID в Windows, однозначно идентифицирующим компьютеры и учетные записи пользователей) и хеширует дескриптор безопасности в соответствии с простым алгоритмом хеширования. Хеш — это потенциально неуникальное сокращенное представление дескриптора. Записи в `$SDH` сопоставляют хеши дескрипторов безопасности с местом хранения дескриптора безопасности в атрибуте данных `$SDS`, а записи в `$SII` сопоставляют идентификаторы безопасности NTFS с местом хранения дескриптора безопасности в атрибуте данных `$SDS`.

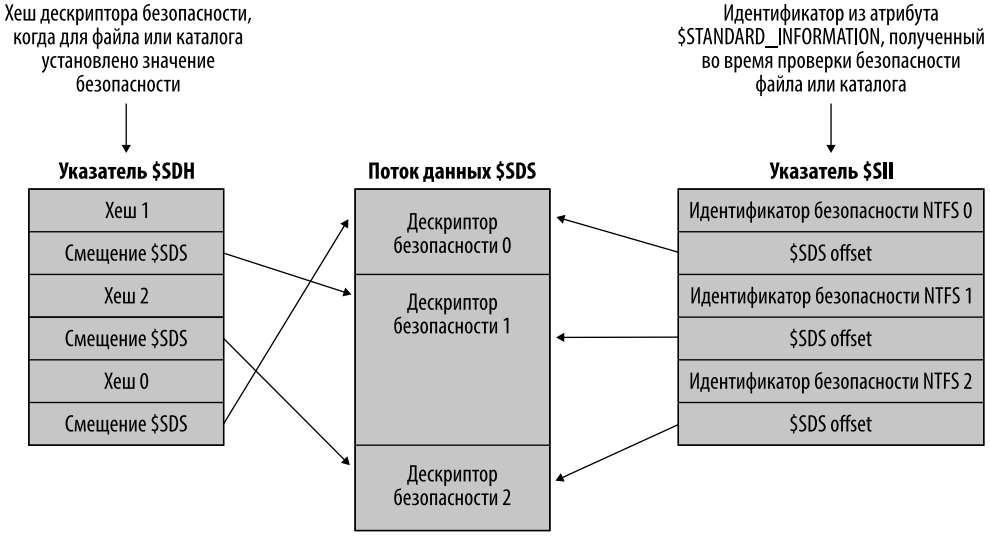


Рис. 11.53. Индексирование $Secure

Когда дескриптор безопасности применяется к файлу или каталогу, NTFS получает хеш дескриптора и просматривает `$SDH` в поисках совпадений. Она сортирует записи индекса `$SDH` в соответствии с хешем дескриптора безопасности и сохраняет записи в B-дереве. Обнаружив совпадение для дескриптора в `$SDH`, NTFS находит смещение дескриптора безопасности записи из значения смещения записи и считывает дескриптор безопасности из атрибута `$SDS`. Если хеши совпадают, а дескрипторы безопасности — нет, NTFS ищет другую подходящую запись в `$SDH`. Когда точное совпадение найдено, файл или каталог, к которому применяется

дескриптор безопасности, может ссылаться на существующий дескриптор безопасности в атрибуте $SDS. NTFS создает эту ссылку, считывая идентификатор безопасности NTFS из записи $SDH, и сохраняет ее в атрибуте $STANDARD_INFORMATION файла или каталога. Атрибут NTFS $STANDARD_INFORMATION, который есть у всех файлов и каталогов, хранит основную информацию о файле, включая его атрибуты, информацию о метке времени и идентификатор безопасности.

Если NTFS не находит в $SDH записи с дескриптором безопасности, соответствующим применяемому дескриптору, то последний является уникальным для тома и NTFS присваивает дескриптору новый внутренний идентификатор безопасности. Внутренние идентификаторы безопасности NTFS — это 32-битные значения, тогда как SID обычно в несколько раз больше, поэтому представление SID идентификаторами безопасности NTFS экономит место в атрибуте $STANDARD_INFORMATION. Затем NTFS добавляет дескриптор безопасности в конец атрибута данных $SDS и добавляет в $SDH и $SII записи, ссылающиеся на смещение дескриптора в данных $SDS.

Когда приложение пытается открыть файл или каталог, NTFS использует $SII для поиска дескриптора безопасности файла или каталога. Она считывает внутренний идентификатор безопасности файла или каталога из атрибута $STANDARD_INFORMATION записи MFT, а затем задействует $SII файла $Secure, чтобы найти запись идентификатора в атрибуте данных $SDS. Смещение для атрибута $SDS позволяет NTFS прочитать дескриптор безопасности и завершить проверку безопасности. NTFS хранит 32 последних запрошенных дескриптора безопасности и записи об их $SII в кэше, чтобы обращаться к файлу $Secure только тогда, когда $SII не кэширован.

NTFS не удаляет записи из файла $Secure, даже если ни один файл или каталог в томе не ссылается на них. То, что эти записи не удаляются, не приводит к значительному сокращению дискового пространства, поскольку большинство томов, даже те, которые используются в течение длительного времени, имеют довольно мало уникальных дескрипторов безопасности.

Применение в NTFS обобщенной индексации с помощью B-деревьев позволяет файлам и каталогам с одинаковыми настройками безопасности эффективно обмениваться дескрипторами безопасности. $SII позволяет NTFS быстро найти дескриптор безопасности в файле $Secure при выполнении проверок безопасности, а $SDH — быстро определить, хранится ли дескриптор безопасности, применяемый к файлу или каталогу, уже в файле $Secure и может ли к нему существовать общий доступ.

Точки повторной обработки

Как было сказано ранее в этой главе, *точка повторной обработки* — это блок данных повторной обработки размером до 16 Кбайт, определяемых приложением, и 32-битный тег повторной обработки, которые хранятся в атрибуте $REPARSE_POINT файла или каталога. Каждый раз, когда приложение создает или удаляет точку повторной обработки, NTFS обновляет файл метаданных \$Extend\$Reparse, в котором она хранит записи, идентифицирующие номера файловых записей файлов и каталогов, содержащих точки повторной обработки. Хранение этих записей в центре управления позволяет NTFS предоставлять приложениям интерфейсы для перечисления всех точек повторной обработки тома или только определенных типов точек повторной обработки, например точек монтирования. Файл \$Extend\$Reparse

использует обобщенный механизм индексации с помощью B-деревьев NTFS, сопоставляя записи файла в указателе с именем $R по меткам точек повторной обработки и номерам записей файла.

ЭКСПЕРИМЕНТ. Исследование различных точек повторной обработки

Точка повторной обработки файла или каталога может содержать любые произвольные данные. В этом эксперименте используется встроенный инструмент fsutil.exe для анализа содержимого точки повторной обработки символической ссылки и AppExecutionAlias приложения Modern, как в эксперименте в главе 8. Сначала нужно создать символическую ссылку:

```
C:\>mklink test_link.txt d:\Test.txt
symbolic link created for test_link.txt <<===>> d:\Test.txt
```

Затем можно применить команду fsutil reparsePoint query для изучения содержимого точки повторной обработки:

```
C:\>fsutil reparsePoint query test_link.txt
Reparse Tag Value : 0xa000000c
Tag value: Microsoft
Tag value: Name Surrogate
Tag value: Symbolic Link

Reparse Data Length: 0x00000040
Reparse Data:
0000:  16 00 1e 00 00 00 16 00  00 00 00 00 64 00 3a 00  ............d.:.
0010:  5c 00 54 00 65 00 73 00  74 00 2e 00 74 00 78 00  \.T.e.s.t...t.x.
0020:  74 00 5c 00 3f 00 3f 00  5c 00 64 00 3a 00 5c 00  t.\.?.?.\.d.:.\.
0030:  54 00 65 00 73 00 74 00  2e 00 74 00 78 00 74 00  T.e.s.t...t.x.t.
```

Как и ожидалось, содержимое представляет собой простую структуру данных REPARSE_DATA_BUFFER, документированную в Microsoft Docs, которая содержит цель символической ссылки и выведенное имя файла. Можно даже удалить точку повторной обработки с помощью команды fsutil reparsePoint delete:

```
C:\>more test_link.txt
This is a test file!

C:\>fsutil reparsePoint delete test_link.txt

C:\>more test_link.txt
```

Если удалить точку повторной обработки, размер файла станет 0 байт. Это сделано специально, потому что безымянный поток данных $DATA в файле ссылок пуст. Можно повторить эксперимент с AppExecutionAlias установленного приложения Modern (в следующем примере использован Spotify):

```
C:\>cd C:\Users\Andrea\AppData\Local\Microsoft\WindowsApps
C:\Users\andrea\AppData\Local\Microsoft\WindowsApps>fsutil reparsePoint query
Spotify.exe
Reparse Tag Value : 0x8000001b
Tag value: Microsoft
```

```
Reparse Data Length: 0x00000178
Reparse Data:
0000:  03 00 00 00 53 00 70 00  6f 00 74 00 69 00 66 00  ....S.p.o.t.i.f.
0010:  79 00 41 00 42 00 2e 00  53 00 70 00 6f 00 74 00  y.A.B...S.p.o.t.
0020:  69 00 66 00 79 00 4d 00  75 00 73 00 69 00 63 00  i.f.y.M.u.s.i.c.
0030:  5f 00 7a 00 70 00 64 00  6e 00 65 00 6b 00 64 00  _.z.p.d.n.e.k.d.
0040:  72 00 7a 00 72 00 65 00  61 00 30 00 00 00 53 00  r.z.r.e.a.0...S
0050:  70 00 6f 00 74 00 69 00  66 00 79 00 41 00 42 00  p.o.t.i.f.y.A.B.
0060:  2e 00 53 00 70 00 6f 00  74 00 69 00 66 00 79 00  ..S.p.o.t.i.f.y.
0070:  4d 00 75 00 73 00 69 00  63 00 5f 00 7a 00 70 00  M.u.s.i.c._.z.p.
0080:  64 00 6e 00 65 00 6b 00  64 00 72 00 7a 00 72 00  d.n.e.k.d.r.z.r.
0090:  65 00 61 00 30 00 21 00  53 00 70 00 6f 00 74 00  e.a.0.!.S.p.o.t.
00a0:  69 00 66 00 79 00 00 00  43 00 3a 00 5c 00 50 00  i.f.y...C.:.\.P.
00b0:  72 00 6f 00 67 00 72 00  61 00 6d 00 20 00 46 00  r.o.g.r.a.m. .F.
00c0:  69 00 6c 00 65 00 73 00  5c 00 57 00 69 00 6e 00  i.l.e.s.\.W.i.n.
00d0:  64 00 6f 00 77 00 73 00  41 00 70 00 70 00 73 00  d.o.w.s.A.p.p.s.
00e0:  5c 00 53 00 70 00 6f 00  74 00 69 00 66 00 79 00  \.S.p.o.t.i.f.y.
00f0:  41 00 42 00 2e 00 53 00  70 00 6f 00 74 00 69 00  A.B...S.p.o.t.i.
0100:  66 00 79 00 4d 00 75 00  73 00 69 00 63 00 5f 00  f.y.M.u.s.i.c._.
0110:  31 00 2e 00 39 00 34 00  2e 00 32 00 36 00 32 00  1...9.4...2.6.2.
0120:  2e 00 30 00 5f 00 78 00  38 00 36 00 5f 00 5f 00  ..0._.x.8.6._._.
0130:  7a 00 70 00 64 00 6e 00  65 00 6b 00 64 00 72 00  z.p.d.n.e.k.d.r.
0140:  7a 00 72 00 65 00 61 00  30 00 5c 00 53 00 70 00  z.r.e.a.0.\.S.p.
0150:  6f 00 74 00 69 00 66 00  79 00 4d 00 69 00 67 00  o.t.i.f.y.M.i.g.
0160:  72 00 61 00 74 00 6f 00  72 00 2e 00 65 00 78 00  r.a.t.o.r...e.x.
0170:  65 00 00 00 30 00 00 00                           e...0...
```

В этой выдаче наблюдается другой тип точки повторной обработки, `AppExecutionAlias`, используемый приложениями Modern. Дополнительная информация есть в главе 8.

Storage Reserves и NTFS Reservations

Windows Update и приложение Windows Setup должны иметь возможность корректно применять важные обновления безопасности, даже когда системный том почти заполнен, — им нужно убедиться, что на диске достаточно места. В Windows 10 для достижения этой цели были введены Storage Reserves. Прежде чем будет рассмотрена эта система, необходимо понять, как работает резервирование NTFS и зачем оно нужно.

Когда файловая система NTFS монтирует том, она подсчитывает объем его используемого и свободного пространства. Для отслеживания этих двух счетчиков на диске не существует атрибутов, NTFS поддерживает и хранит на диске битовую карту тома, которая представляет состояние всех кластеров в нем. Код монтирования NTFS сканирует битовую карту и подсчитывает количество применяемых кластеров, у которых бит в битовой карте установлен в 1, и с помощью простого уравнения — общее количество кластеров тома минус количество использованных — вычисляет количество свободных кластеров. Два вычисленных счетчика хранятся в структуре данных, называемой *блоком управления томом* (volume control block, VCB), которая представляет смонтированный том и существует только в памяти до тех пор, пока том не будет демонтирован.

Во время нормальной активности ввода-вывода томов NTFS должна контролировать общее количество зарезервированных кластеров. Этот счетчик должен существовать по следующим причинам.

- При записи в сжатые и разреженные файлы система должна убедиться, что весь файл доступен для записи, поскольку работающее с ним приложение может сохранить корректные несжатые данные во всем файле.
- При создании открытой для записи секции, для которой сохранен образ, файловая система должна зарезервировать пространство для всего ее размера, даже если в томе еще не выделено физическое пространство.
- Журнал USN и TxF используют этот счетчик, чтобы убедиться в наличии свободного места для транзакций журнала USN и NTFS.

Во время обычных операций ввода-вывода NTFS поддерживает еще один счетчик, *Total Free Available Space* (общее свободное пространство), который представляет собой последнее пространство, которое пользователь может видеть и задействовать для хранения новых файлов или данных. Эти три понятия являются частью NTFS Reservations. Важная особенность NTFS Reservations заключается в том, что счетчики — это только изменяемые представления в памяти, которые будут уничтожены при демонтаже тома.

Storage Reserve — это функция, основанная на NTFS Reservations, которая позволяет файлам иметь назначенную область в Storage Reserve. Она определяет 15 областей резервирования (две из них резервируются операционной системой), которые задаются и хранятся как в памяти, так и в структурах данных NTFS на диске.

Чтобы использовать новое резервирование на диске, приложение определяет область Storage Reserve тома с помощью управляющего кода файловой системы `FSCTL_QUERY_STORAGE_RESERVE`, который указывает в структуре данных общий объем зарезервированного пространства и идентификатор области. Это приведет к обновлению нескольких счетчиков в VCB (области Storage Reserve хранятся в памяти) и вставке новых данных в именованный поток данных `$SRAT` файла метаданных `$Bitmap`. Поток данных `$SRAT` содержит структуру данных, которая отслеживает каждую резервную область, включая количество зарезервированных и использованных кластеров. Приложение может запрашивать информацию об областях Storage Reserve с помощью управляющего кода файловой системы `FSCTL_QUERY_STORAGE_RESERVE` и удалять Storage Reserve с помощью кода `FSCTL_DELETE_STORAGE_RESERVE`.

После определения области Storage Reserve приложению гарантируется, что это пространство больше не будет использоваться другими компонентами. Затем приложения могут направлять файлы и каталоги в область Storage Reserve с помощью встроенного API `NtSetInformationFile` с информационным классом `FileStorageReserveIdInformationEx`. Драйвер файловой системы NTFS обрабатывает запрос, обновляя счетчики зарезервированных и использованных кластеров в зарезервированной области, а также общее количество зарезервированных кластеров тома, принадлежащих резервированию NTFS. Он также сохраняет и обновляет на диске атрибут `$STANDARD_INFO` целевого файла. Последний содержит четыре бита для хранения идентификатора области Storage Reserve. Таким способом система может быстро перечислить каждый файл, принадлежащий резервной области, просто

разобрав записи MFT. NTFS реализует перечисление в функции диспетчеризации кода `FSCTL_QUERY_FILE_LAYOUT`. Пользователь может перечислить файлы, принадлежащие Storage Reserve, с помощью команды `fsutil storageReserve findByID`, указав имя пути к тому и идентификатор Storage Reserve, который его интересует.

Некоторые базовые файловые операции имеют новые побочные эффекты, связанные со Storage Reserve, — например, создание и переименование файлов. Вновь созданные файлы или каталоги автоматически наследуют идентификатор Storage Reserve своего родителя, то же самое относится к файлам или каталогам, которые переименовываются или перемещаются к новому родителю. Поскольку операция переименования может изменить идентификатор Storage Reserve файла или каталога, это означает, что операция может завершиться неудачей из-за нехватки дискового пространства. Перемещение непустого каталога к новому родителю подразумевает, что новый идентификатор Storage Reserve будет рекурсивно применен ко всем файлам и подкаталогам. Когда зарезервированное пространство Storage Reserve заканчивается, система начинает использовать свободное пространство тома, поэтому нет гарантии, что операция всегда будет успешной.

ЭКСПЕРИМЕНТ. Наблюдение за Storage Reserve

Начиная с обновления Windows 10 от мая 2019 года (19H1) можно просматривать существующие резервы NTFS с помощью встроенного инструмента fsutil.exe:

```
C:\>fsutil storagereserve query c:
Reserve ID:       1
Flags:            0x00000000
Space Guarantee:  0x0               (0 MB)
Space Used:       0x0               (0 MB)

Reserve ID:       2
Flags:            0x00000000
Space Guarantee:  0x0               (0 MB)
Space Used:       0x199ed000        (409 MB)
```

Windows Setup определяет два резерва NTFS: жесткий резерв ID 1, используемый приложением Setup для хранения своих файлов, которые не могут быть удалены или заменены другими приложениями, и мягкий резерв ID 2, применяемый для хранения временных файлов, например системных журналов и загруженных файлов Windows Update. В предыдущем примере приложение Setup уже смогло установить все свои файлы и обновление Windows не выполняется, поэтому жесткий резерв пуст, а в мягком выделено все зарезервированное пространство. Можно перечислить все файлы, принадлежащие резерву, с помощью команды `fsutil storagereserve findById`. Имейте в виду, что результат будет очень большим, поэтому есть смысл перенаправить его в файл с помощью оператора >:

```
C:\>fsutil storagereserve findbyid c: 2
...

********* File 0x0002000000018762 *********
File reference number   : 0x0002000000018762
File attributes         : 0x00000020: Archive
```

```
File entry flags        : 0x00000000
Link (ParentID: Name)   : 0x0001000000001165: NTFS Name     :
Windows\System32\winevt\Logs\OAlerts.evtx
Link (ParentID: Name)   : 0x0001000000001165: DOS Name      : OALERT~1.EVT
Creation Time           : 12/9/2018 3:26:55
Last Access Time        : 12/10/2018 0:21:57
Last Write Time         : 12/10/2018 0:21:57
Change Time             : 12/10/2018 0:21:57
LastUsn                 : 44,846,752
OwnerId                 : 0
SecurityId              : 551
StorageReserveId        : 2
Stream                  : 0x010 ::$STANDARD_INFORMATION
    Attributes          : 0x00000000: *NONE*
    Flags               : 0x0000000c: Resident | No clusters allocated
    Size                : 72
    Allocated Size      : 72
Stream                  : 0x030 ::$FILE_NAME
    Attributes          : 0x00000000: *NONE*
    Flags               : 0x0000000c: Resident | No clusters allocated
    Size                : 90
    Allocated Size      : 96
Stream                  : 0x030 ::$FILE_NAME
    Attributes          : 0x00000000: *NONE*
    Flags               : 0x0000000c: Resident | No clusters allocated
    Size                : 90
    Allocated Size      : 96
Stream                  : 0x080 ::$DATA
    Attributes          : 0x00000000: *NONE*
    Flags               : 0x00000000: *NONE*
    Size                : 69,632
    Allocated Size      : 69,632
    Extents             : 1 Extents
                        : 1: VCN: 0 Clusters: 17 LCN: 3,820,235
```

Поддержка транзакций

Используя поддержку диспетчера транзакций ядра (Kernel Transaction Manager, KTM), а также средства, предоставляемые общей системой журнальных файлов (Common Log File System), NTFS реализует транзакционную модель, называемую *транзакционной NTFS*, или *TxF*. TxF предоставляет набор API пользовательского режима, который приложения могут применять для транзакционных операций над файлами и каталогами, а также интерфейс управления файловой системой (file system control, FSCTL) для управления его диспетчерами ресурсов.

ПРИМЕЧАНИЕ В Windows Vista поддержка TxF была добавлена как средство внедрения атомарных транзакций в Windows. Драйвер NTFS был изменен без фактического изменения формата структур данных NTFS, поэтому номер версии формата NTFS, 3.1, остался таким же, какой был в Windows XP и Windows Server 2003. TxF добивается обратной совместимости за счет повторного применения типа атрибута $LOGGED_UTILITY_STREAM, который ранее использовался только для поддержки EFS, вместо добавления нового.

TxF — это мощный API, но из-за сложности и различных проблем, которые необходимо учитывать разработчикам, его используют лишь небольшое количество приложений. На момент написания этой книги Microsoft рассматривает возможность отказа от TxF API в одной из будущих версий Windows. Для полноты картины здесь приводится лишь общий обзор архитектуры TxF.

Общая архитектура TxF (рис. 11.54) состоит из следующих компонентов:

- транзакционных API, реализованных в библиотеке Kernel32.dll;
- библиотеки для чтения журналов TxF `%SystemRoot%\System32\Txfw32.dll`;
- COM-компонента для ведения журнала TxF `%SystemRoot\System32\Txflog.dll`;
- транзакционной библиотеки NTFS внутри драйвера NTFS;
- инфраструктуры CLFS для чтения и записи журнала.

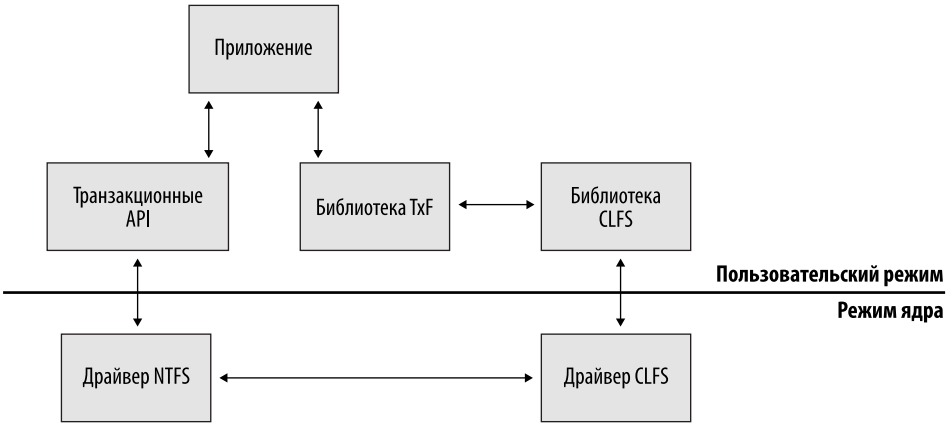


Рис. 11.54. Архитектура TxF

Изоляция

Хотя транзакционные файловые операции, как и транзакционные операции с реестром (TxR), описанные в главе 10, являются опциональными, TxF влияет на обычные приложения, которые не знают о транзакциях, поскольку обеспечивает *изоляцию* транзакционных операций. Например, если антивирусная программа сканирует файл, который в данный момент изменяется другим приложением с помощью транзакционной операции, то TxF должен гарантировать, что сканер прочитает данные до транзакции, а приложения, которые обращаются к файлу в рамках транзакции, будут работать с измененными данными. Эта модель называется *изоляцией с фиксацией чтения*.

Изоляция с фиксацией чтения включает в себя концепцию *транзакционных писателей* и *транзакционных читателей*. Первые всегда просматривают самую последнюю версию файла, включая все изменения, сделанные транзакцией, которая в данный момент связана с файлом. Транзакционные же читатели имеют доступ только к зафиксированной версии файла в момент его открытия. Таким

образом, они изолированы от изменений, внесенных транзакционными писателями. Это позволяет читателям иметь согласованное представление о файле, даже когда транзакционный писатель фиксирует свои изменения. Чтобы увидеть обновленные данные, транзакционный читатель должен открыть новый дескриптор к измененному файлу.

В то же время нетранзакционным писателям запрещают открывать файл и транзакционные писатели, и транзакционные читатели, поэтому они не могут вносить изменения в файл, не участвуя в транзакции. Нетранзакционные читатели действуют аналогично транзакционным читателям, поскольку видят только то содержимое файла, которое было зафиксировано в последний раз, когда был открыт дескриптор файла. Однако, в отличие от транзакционных читателей, они не получают изоляцию с фиксацией чтения, поэтому всегда имеют обновленное представление последней зафиксированной версии транзакционного файла без необходимости открывать новый файловый дескриптор. Это позволяет нетранзакционным приложениям вести себя так, как ожидается.

Подводя итог, можно сказать, что модель изоляции TxF с фиксацией чтения имеет следующие характеристики.

- Изменения изолированы от транзакционных читателей.
- Изменения откатываются (отменяются), если связанная с ними транзакция откатывается, или машина катастрофически завершает работу, или том принудительно демонтируется.
- Изменения сбрасываются на диск, если соответствующая транзакция зафиксирована.

Транзакционные API

TxF реализует транзакционные версии API файлового ввода-вывода Windows, использующие суффикс *Transacted*:

- **API создания** — `CreateDirectoryTransacted`, `CreateFileTransacted`, `CreateHardLinkTransacted`, `CreateSymbolicLinkTransacted`;
- **API поиска** — `FindFirstFileNameTransacted`, `FindFirstFileTransacted`, `FindFirstStreamTransacted`;
- **API запросов** — `GetCompressedFileSizeTransacted`, `GetFileAttributesTransacted`, `GetFullPathNameTransacted`, `GetLongPathNameTransacted`;
- **API удаления** — `DeleteFileTransacted`, `RemoveDirectoryTransacted`;
- **API копирования и перемещения или переименования** — `CopyFileTransacted`, `MoveFileTransacted`;
- **API присвоения** — `SetFileAttributesTransacted`.

Кроме того, некоторые API автоматически участвуют в транзакционных операциях, когда передаваемый им файловый дескриптор является частью транзакции, например, созданной API `CreateFileTransacted`. В табл. 11.10 перечислены API-интерфейсы Windows, которые ведут себя иначе, работая с транзакционным файловым дескриптором.

Таблица 11.10. Поведение API, измененное TxF

Имя API	Изменение
CloseHandle	Транзакции не фиксируются до тех пор, пока все приложения не закроют транзакционные дескрипторы файла
CreateFileMapping, MapViewOfFile	Изменения отображенных представлений файловой части транзакции ассоциируются с самой транзакцией
FindNextFile, ReadDirectoryChanges, GetInformationByHandle, GetFileSize	Если файловый дескриптор является частью транзакции, то для этих операций применяются правила изоляции чтения
GetVolumeInformation	Функция возвращает FILE_SUPPORTS_TRANSACTIONS, если том поддерживает TxF
ReadFile, WriteFile	Операции чтения и записи в транзакционный файловый дескриптор являются частью транзакции
SetFileInformationByHandle	Изменения классов FileBasicInfo, FileRenameInfo, FileAllocationInfo, FileEndOfFileInfo и FileDispositionInfo транзакционны, если файловый дескриптор является частью транзакции
SetEndOfFile, SetFileShortName, SetFileTime	Изменения транзакционны, если файловый дескриптор является частью транзакции

Реализация на диске

Как было показано в табл. 11.7, TxF использует тип атрибута `$LOGGED_UTILITY_STREAM` для хранения дополнительных данных для файлов и каталогов, которые являются или являлись частью транзакции. Этот атрибут называется `$TXF_DATA` и содержит важную информацию, позволяющую TxF сохранять активные автономные данные для файловой части транзакции. Атрибут постоянно хранится в MFT, то есть даже после того, как файл перестает быть частью транзакции, поток сохраняется по причинам, которые будут объяснены далее. Основные компоненты атрибута показаны на рис. 11.55.

Рис. 11.55. Атрибут $TXF_DATA

Первое показанное поле — это номер файловой записи корня диспетчера ресурсов, ответственного за транзакцию, связанную с этим файлом. Для диспетчера ресурсов по умолчанию номер файловой записи равен 5 — это номер файловой записи для корневого каталога \ в MFT (см. рис. 11.31). TxF нуждается в этой информации, когда создает FCB для файла, чтобы связать его с правильным диспетчером ресурсов, который, в свою очередь, должен создать фиксацию для транзакции, когда NTFS получает запрос на транзакционный файл.

Другой важной частью данных, хранящихся в атрибуте `$TXF_DATA`, является *идентификатор файла TxF*, или TxID, и это объясняет, почему атрибуты `$TXF_DATA` никогда не удаляются. Поскольку NTFS заносит имена файлов в свои записи при внесении в журнал транзакций, ей необходим способ уникальной идентификации файлов в одном каталоге, которые могут иметь одинаковое имя. Например, если файл sample.txt удаляется из каталога в ходе транзакции, а позже в том же каталоге и в рамках той же транзакции создается новый файл с тем же именем, то TxF необходим способ однозначной идентификации двух экземпляров sample.txt. Идентификация обеспечивается 64-битным уникальным номером TxID, который TxF увеличивает, когда новый файл или экземпляр файла становится частью транзакции. Поскольку TxID не может быть использован повторно, он постоянен, поэтому атрибут `$TXF_DATA` никогда не будет удален из файла.

И последнее, но не менее важное: для каждой файловой части транзакции хранятся три LSN CLFS (Common Logging File System). Всякий раз, когда транзакция активна, например во время операций создания, переименования или записи, TxF вносит запись в журнал CLFS. Каждой записи присваивается LSN, он заносится в соответствующее поле в атрибуте `$TXF_DATA`. Первый LSN используется для хранения записи журнала, которая идентифицирует изменения метаданных NTFS по отношению к данному файлу. Например, если стандартные атрибуты файла изменяются в рамках транзакционной операции, то TxF должен обновить соответствующую запись файла MFT, и LSN для описывающей это изменение записи журнала сохраняется. TxF задействует второй LSN, когда изменяются данные файла. Наконец, TxF использует третий LSN, когда указатель имен файлов для каталога требует изменения, связанного с транзакцией, в которой участвовал файл, или когда каталог был частью транзакции и получил TxID.

Атрибут `$TXF_DATA` хранит также внутренние флаги, описывающие информацию о состоянии для TxF, и указатель записи USN, которая была применена к файлу при фиксации. Транзакция TxF может охватывать несколько записей USN, которые могут быть частично обновлены механизмом восстановления NTFS, описанным далее, поэтому указатель сообщает TxF, сколько еще записей USN должно быть применено после восстановления.

TxF использует диспетчер ресурсов *по умолчанию*, один для каждого тома, для отслеживания состояния транзакций, а также поддерживает дополнительные диспетчеры ресурсов, называемые вторичными. Они могут быть определены авторами приложений, и их метаданные располагаются в любом каталоге по выбору приложения, определяя собственные транзакционные рабочие единицы для операций отмены, резервного копирования, восстановления и повторного выполнения. Как диспетчер ресурсов по умолчанию, так и вторичные диспетчеры ресурсов содержат ряд файлов и каталогов метаданных, описывающих их текущее состояние.

- Каталог `$Txf`, расположенный в каталоге `$Extend\$RmMetadata`, с которым связываются файлы, когда они удаляются или перезаписываются в результате транзакционных операций.
- Файл `$Tops` (TxF Old Page Stream, TOPS), содержащий поток данных по умолчанию и альтернативный поток данных `$T`. Поток по умолчанию в файле TOPS содержит метаданные о диспетчере ресурсов, такие как его GUID, политика журнала CLFS и LSN, с которого должно начаться восстановление. Поток `$T`

содержит данные файла, частично перезаписывающиеся транзакционной записью, в отличие от полной перезаписи, при которой файл перемещается в каталог `$Txf`.

- Файлы журнала TxF, представляющие собой файлы журнала CLFS, хранящие записи транзакций. Для диспетчера ресурсов по умолчанию эти файлы находятся в каталоге `$TxfLog`, но вторичные диспетчеры могут хранить их в любом месте. TxF использует мультиплексированный базовый файл журнала под названием `$TxfLog.blf`. Файл `\$Extend\$RmMetadata\$TxfLog\$TxfLog` содержит два потока: KtmLog для записей метаданных диспетчера транзакций ядра и TxfLog, содержащий записи в журнале для TxF.

ЭКСПЕРИМЕНТ. Запрос информации о диспетчере ресурсов

Можно использовать встроенную программу командной строки Fsutil.exe для запроса информации о диспетчере ресурсов по умолчанию, а также для создания, запуска и остановки вторичных диспетчеров ресурсов и настройки их политик и поведения при регистрации. Информацию о диспетчере ресурсов по умолчанию, который идентифицируется корневым каталогом `\`, запрашивает следующая команда:

```
d:\>fsutil resource info \
Resource Manager Identifier :      81E83020-E6FB-11E8-B862-D89EF33A38A7
KTM Log Path for RM:  \Device\HarddiskVolume8\$Extend\$RmMetadata\$TxfLog\
$TxfLog::KtmLog
Space used by TOPS:   1 Mb
TOPS free space:      100%
RM State:             Active
Running transactions: 0
One phase commits:    0
Two phase commits:    0
System initiated rollbacks: 0
Age of oldest transaction:  00:00:00
Logging Mode:         Simple
Number of containers: 2
Container size:       10 Mb
Total log capacity:   20 Mb
Total free log space: 19 Mb
Minimum containers:   2
Maximum containers:   20
Log growth increment: 2 container(s)
Auto shrink:          Not enabled

RM prefers availability over consistency.
```

Как уже упоминалось, у команды `fsutil resource` есть множество опций для настройки диспетчеров ресурсов TxF, включая возможность создания вторичного диспетчера ресурсов в любом каталоге по выбору. Например, можно использовать команду `fsutil resource create c:\rmtest` для создания вторичного диспетчера ресурсов в каталоге Rmtest, а затем команду `fsutil resource start c:\rmtest` для его запуска. Обратите внимание на наличие в этой папке файлов `$Tops` и `$TxfLogContainer*`, а также каталогов TxfLog и `$Txf`.

Реализация протоколирования

Как уже говорилось, каждый раз, когда в результате текущей транзакции на диске происходит изменение, TxF вносит запись о нем в свой журнал. TxF использует различные типы записей журнала для отслеживания транзакционных изменений, но независимо от типа все записи журнала имеют обобщенный заголовок, содержащий информацию, идентифицирующую тип записи, действие, связанное с записью, TxID, к которому относится запись, и GUID транзакции KTM, с которой связана запись.

Запись повтора определяет, как повторно применить часть изменения транзакции, которая уже была зафиксирована в томе, если на самом деле транзакция никогда не сбрасывалась из кэша на диск. *Запись отмены* определяет, как отменить часть изменения транзакции, не зафиксированную на момент отката. Некоторые записи предназначены только для повтора, то есть они не содержат эквивалентных данных для отмены, в то время как другие записи включают в себя информацию как для повтора, так и для отмены.

В файле TOPS TxF хранит две важные части данных: *базовый LSN* и *LSN перезапуска*. Базовый LSN определяет LSN первой действительной записи в журнале, а LSN перезапуска указывает, с какого LSN должно начинаться восстановление при запуске диспетчера ресурсов. Когда TxF вносит *запись перезапуска*, она обновляет эти два значения, указывая на то, что изменения были внесены в том и сброшены на диск — это означает, что файловая система полностью согласована до нового LSN перезапуска.

TxF также делает *компенсирующие записи журнала* (compensating log records, CLR). В них хранятся действия, выполняемые во время отката транзакции. В основном они используются для хранения *LSN следующей отмены*, который позволяет процессу восстановления избежать повторения операций отмены, обходя уже обработанные записи отмены, что может произойти, если система дает сбой на этапе восстановления и уже выполнила часть прохода отмены. Наконец, TxF имеет дело с *записями подготовки*, *отмены* и *фиксации*, описывающими состояние транзакций KTM, связанных с TxF.

ПОДДЕРЖКА ВОССТАНОВЛЕНИЯ NTFS

Поддержка восстановления NTFS гарантирует, что в случае сбоя питания или отказа системы ни одна операция (транзакция) файловой системы не будет оставлена незавершенной, а структура дискового тома останется неповрежденной, причем не потребуется запускать утилиту восстановления диска. Утилита NTFS Chkdsk используется для устранения катастрофических повреждений диска, вызванных ошибками ввода-вывода (например, плохими секторами диска, сбоями в его работе или перебоями с питанием) или программными ошибками. Но при наличии возможностей восстановления NTFS утилита Chkdsk требуется редко.

Как упоминалось ранее в разделе «Восстанавливаемость», NTFS использует схему обработки транзакций для реализации восстанавливаемости. Эта стратегия обеспечивает восстановление всего диска, причем чрезвычайно быстрое, порядка нескольких секунд даже для самых больших дисков. NTFS ограничивает свои процедуры восстановления данными файловой системы, чтобы гарантировать, что пользователь по крайней мере никогда не потеряет том из-за повреждения файловой

системы. Однако если приложение не предпримет специальных действий, например не сбросит кэшированные файлы на диск, то поддержка восстановления NTFS не гарантирует полного обновления пользовательских данных в случае сбоя. За это отвечает транзакционная NTFS (TxF).

В следующих разделах подробно описывается схема регистрации транзакций, которую NTFS использует для записи изменений в структурах данных файловой системы, и объясняется, как NTFS восстанавливает том в случае сбоя системы.

Конструкция

В NTFS реализована конструкция *восстанавливаемой файловой системы*. Такие файловые системы обеспечивают согласованность томов с помощью методов ведения журнала, иногда называемых *журналированием*, изначально разработанных для обработки транзакций. При сбое операционной системы восстанавливаемая файловая система отстраивает согласованность заново, выполняя процедуру восстановления, которая обращается к информации, хранящейся в файле журнала. Поскольку файловая система регистрирует свои записи на диск, процедура восстановления занимает всего несколько секунд независимо от размера тома, в отличие от файловой системы FAT, где время восстановления зависит от размера тома. Процедура восстановления для файловой системы является точной и гарантирует, что том будет приведен в согласованное состояние.

Восстанавливаемая файловая система несет определенные затраты на обеспечение безопасности. Каждая транзакция, изменяющая структуру тома, требует внесения в файл журнала одной записи для каждой подоперации транзакции. Эти накладные расходы на ведение журнала компенсируются тем, что файловая система *пакетно обрабатывает* записи журнала — вносит много записей в файл журнала за одну операцию ввода-вывода. Кроме того, восстанавливаемая файловая система может использовать методы оптимизации файловой системы с отложенной записью. Она даже может увеличить длительность интервалов между сбросами кэша на диск, поскольку метаданные файловой системы могут быть восстановлены, если система аварийно завершит работу до того, как изменения в кэше будут выгружены на диск. Этот выигрыш по сравнению с производительностью кэширования в файловых системах с отложенной записью компенсирует, а зачастую и превосходит накладные расходы на ведение журнала в восстанавливаемой файловой системе.

Ни осторожная запись, ни отложенная запись файловых систем не гарантируют защиту данных пользовательских файлов. Если во время записи файла приложением произойдет сбой системы, файл может быть потерян или поврежден. Хуже того, сбой способен повредить файловую систему с отложенной записью, уничтожив существующие файлы или даже сделав недоступным весь том.

Восстанавливаемая файловая система NTFS реализует несколько стратегий, которые повышают ее надежность по сравнению с традиционными файловыми системами. Во-первых, восстанавливаемость NTFS гарантирует, что структура тома не будет повреждена, поэтому все файлы останутся доступными после сбоя системы. Во-вторых, хотя NTFS не гарантирует защиты пользовательских данных в случае сбоя системы — некоторые изменения в кэше могут быть потеряны, — приложения могут использовать возможности NTFS по прямой записи и сбросу кэша, чтобы обеспечить запись изменений файлов на диск через соответствующие промежутки времени.

Эффективными операциями являются как *прямая запись через кэш*, то есть принудительная немедленная запись на диск, так и *сброс кэша*, то есть принудительная запись его содержимого на диск. NTFS не нужно выполнять дополнительные операции ввода-вывода на диск для сброса изменений нескольких различных структур данных файловой системы, поскольку они заносятся в файл журнала за одну операцию записи. Если произойдет сбой и содержимое кэша будет утеряно, изменения файловой системы можно будет восстановить из журнала. Кроме того, в отличие от файловой системы FAT, NTFS гарантирует, что пользовательские данные будут согласованы и доступны сразу после операции прямой записи или очистки кэша, даже если система впоследствии выйдет из строя.

Регистрация метаданных

NTFS обеспечивает возможность восстановления файловой системы за счет использования той же техники протоколирования, которую применяет и TxF, заключающейся в записи всех операций, изменяющих метаданные файловой системы, в файл журнала. Однако, в отличие от TxF, встроенная поддержка восстановления файловой системы NTFS использует не CLFS, а внутреннюю реализацию протоколирования, называемую *службой файла журнала*, не являющуюся фоновым процессом службы, как говорилось в главе 10. Еще одно отличие заключается в том, что если TxF задействуется только при выборе транзакционных операций, то NTFS записывает все изменения метаданных, чтобы файловая система могла быть согласована в случае системного сбоя.

Служба файла журнала

Служба файла журнала (log file service, LFS) — это серия процедур режима ядра внутри драйвера NTFS, которые NTFS использует для доступа к файлу журнала. NTFS передает LFS указатель на объект открытого файла, определяющий файл журнала, к которому нужно получить доступ. LFS либо инициализирует новый файл журнала, либо вызывает диспетчер кэша Windows для доступа к существующему через кэш (рис. 11.56). Обратите внимание на то, что, хотя LFS и CLFS имеют похожие названия и работают сходным образом, это разные реализации ведения журнала, используемые для разных целей.

LFS делит файл журнала на две области: *область перезапуска* и «бесконечную» *область регистрации* (рис. 11.57).

NTFS вызывает LFS для чтения и записи области перезапуска. Она использует область перезапуска для хранения контекстной информации, например местоположения в области регистрации, с которого начинает считывать данные при восстановлении после сбоя системы. LFS хранит вторую копию данных перезапуска на случай, если первая будет повреждена или недоступна по иной причине. Оставшаяся часть файла журнала — это область регистрации, содержащая записи транзакций, вносимые NTFS для восстановления тома в случае системного сбоя. LFS делает файл журнала бесконечным, используя его по кругу и гарантируя при этом, что нужная информация не будет перезаписана. Как и CLFS, LFS применяет LSN для идентификации записей, внесенных в файл журнала. Циклически просматривая файл, LFS увеличивает значения LSN. NTFS использует 64 бита для представления LSN, поэтому количество возможных LSN практически бесконечно.

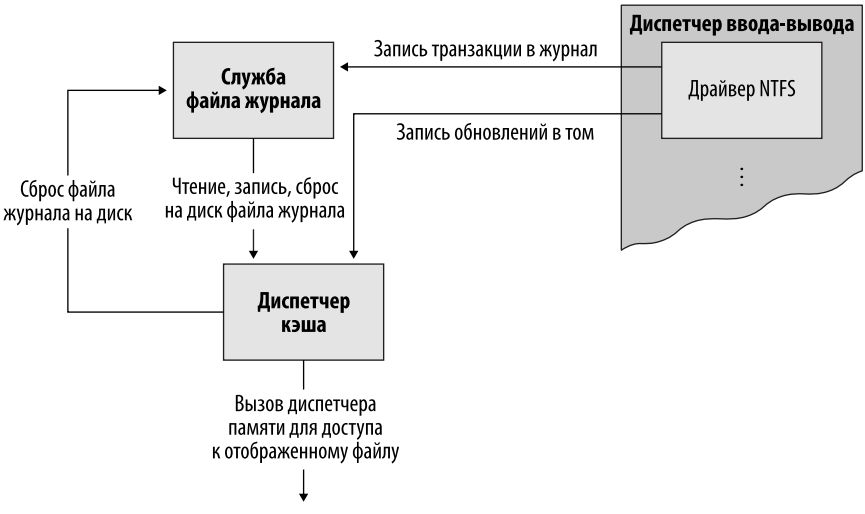


Рис. 11.56. Служба файлов журнала (LFS)

Рис. 11.57. Области файла журнала

NTFS никогда не считывает транзакции из файла журнала и не записывает их в него напрямую. LFS предоставляет услуги, которые NTFS вызывает для открытия файла журнала, внесения в него записей, чтения записей в прямом или обратном порядке, сброса на диск записей до заданного LSN или установки начала файла журнала на более высокий LSN. Во время восстановления NTFS вызывает LFS для выполнения тех же действий, которые описаны в разделе о восстановлении TxF: прохода повторов для не сброшенных на диск зафиксированных изменений, а затем прохода отмены для незафиксированных изменений.

Система гарантирует, что том может быть восстановлен, следующим образом.

1. Сначала NTFS вызывает LFS для записи в кэшированный файл журнала всех операций, которые будут изменять структуру тома.
2. NTFS изменяет том в кэше.
3. Диспетчер кэша просит LFS сбросить файл журнала на диск. LFS выполняет сброс, вновь обращаясь к диспетчеру кэша и сообщая ему, какие страницы памяти необходимо сбросить. (См. последовательность вызовов на рис. 11.56.)
4. Сбросив файл журнала на диск, диспетчер кэша скидывает туда же изменения тома, то есть сами операции с метаданными.

Эти шаги гарантируют, что если в итоге изменения файловой системы окажутся неудачными, то соответствующие транзакции можно будет извлечь из файла журнала и либо повторить, либо отменить в рамках процедуры восстановления файловой системы.

Восстановление файловой системы начинается автоматически при первом использовании тома после перезагрузки системы. NTFS проверяет, были ли транзакции, записанные в файл журнала до сбоя, применены к тому, и если нет, то выполняет их заново. NTFS также гарантирует, что транзакции, не полностью записанные в журнал до сбоя, будут отменены, чтобы их не было в томе.

Типы записей в журнале

Механизм восстановления NTFS задействует те же типы записей журнала, что и механизм восстановления TxF: *записи обновления*, соответствующие используемым TxF записям повтора и отмены, и *записи контрольной точки*, похожие на применяемые TxF записи перезапуска. На рис. 11.58 показаны три записи обновления в файле журнала. Каждая запись представляет собой одну подоперацию транзакции, создающую новый файл. Запись повтора в каждой записи обновления указывает NTFS, как повторно применить подоперацию к тому, а запись отмены — как откатить (отменить) подоперацию.

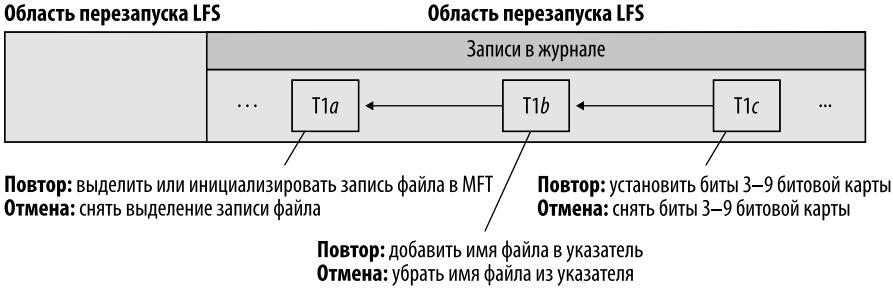


Рис. 11.58. Записи обновления в файле журнала

После регистрации транзакции в данном примере путем вызова LFS для внесения трех записей обновления в файл журнала NTFS выполняет подоперации на самом томе в кэше. По завершении обновления кэша она вносит еще одну запись в файл журнала, считая всю транзакцию завершенной, — эта подоперация называется *фиксацией* транзакции. После фиксации транзакции NTFS гарантирует, что вся транзакция появится в томе, даже если операционная система впоследствии даст сбой.

При восстановлении после системного сбоя NTFS просматривает файл журнала и повторяет каждую зафиксированную транзакцию. Хотя NTFS завершила зафиксированные транзакции до сбоя системы, она не знает, своевременно ли диспетчер кэша сбрасывал изменения тома на диск. Обновления могли исчезнуть из кэша при сбое системы, поэтому NTFS выполняет зафиксированные транзакции снова, чтобы убедиться, что диск обновлен.

После повторного выполнения зафиксированных транзакций во время восстановления файловой системы NTFS находит в файле журнала все транзакции,

не зафиксированные во время сбоя, и откатывает каждую подоперацию, записанную в журнал. На рис. 11.58 NTFS сначала отменит подоперацию T1*c*, затем проследует по обратному указателю к T1*b* и отменит ее. Она будет двигаться по обратным указателям, отменяя подоперации, пока не достигнет первой в транзакции. Следуя указателям, NTFS знает, какие записи обновления нужно отменить, чтобы откатить транзакцию, и сколько.

Информация о повторах и отменах может быть выражена как физически, так и логически. Как самый нижний уровень программного обеспечения, поддерживающий структуру файловой системы, NTFS вносит записи обновления с *физическими описаниями*, определяющими обновления тома в терминах конкретных диапазонов байтов на диске, которые должны быть изменены, перемещены и т. д., в отличие от TxF, которая использует *логические описания*, определяющие обновления в терминах операций, таких как «удалить файл A.dat». NTFS вносит записи обновлений, обычно сразу несколько, для каждой из следующих операций:

- создание файла;
- удаление файла;
- расширение файла;
- усечение файла;
- внесение информации о файле;
- переименование файла;
- изменение установок безопасности, применяемых к файлу.

Информация об отмене и повторе в записи обновления должна быть тщательно продумана, поскольку, хотя NTFS отменяет транзакцию, восстанавливает систему после сбоя или даже работает нормально, она может попытаться повторить транзакцию, которая уже выполнена, или, наоборот, отменить транзакцию, которая никогда не выполнялась или уже была отменена. Аналогично NTFS может попытаться повторить или отменить транзакцию, состоящую из нескольких записей обновления, только часть которых завершена на диске. Формат записей обновления должен гарантировать, что выполнение избыточных операций повторения или отмены будет *идемпотентным*, то есть иметь нейтральное воздействие. Например, установка бита, который уже установлен, не имеет никакого эффекта, а переключение бита, который уже был переключен, имеет. Также файловая система должна правильно обрабатывать промежуточные состояния томов.

В дополнение к записям обновления NTFS периодически вносит в файл журнала запись контрольной точки (рис. 11.59).

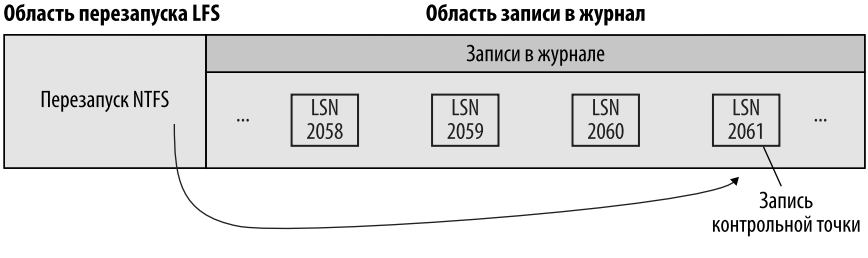


Рис. 11.59. Запись контрольной точки в файле журнала

Запись контрольной точки помогает NTFS определить, какие действия потребуются для восстановления тома, если авария произойдет прямо сейчас. Используя информацию, хранящуюся в записи контрольной точки, NTFS знает, например, как далеко назад в файле журнала ей нужно вернуться, чтобы начать восстановление. После записи контрольной точки NTFS сохраняет LSN этой записи в области перезапуска, чтобы быстро найти последнюю запись контрольной точки, когда начнется восстановление файловой системы после сбоя. Это похоже на LSN перезапуска, который используется TxF по той же причине.

Хотя LFS представляет файл журнала NTFS, как будто он бесконечно велик, это не так. Значительный размер файла журнала и частое внесение записей контрольных точек — обычно освобождающая место в файле журнала операция — делают возможность переполнения файла журнала исключительно теоретической. Тем не менее LFS, как и CLFS, учитывает эту возможность, отслеживая следующие рабочие параметры:

- доступное пространство журнала;
- объем пространства, необходимый для внесения входящей записи в журнал и отмены записи, если это потребуется;
- объем пространства, необходимый для отката всех активных (незафиксированных) транзакций, если это потребуется.

Если в файле журнала недостаточно свободного места для размещения двух последних элементов, то LFS возвращает ошибку «файл журнала переполнен», а NTFS выдает исключение. Обработчик исключений NTFS откатывает текущую транзакцию и помещает ее в очередь для последующего перезапуска.

Чтобы освободить место в файле журнала, NTFS должна на короткое время запретить дальнейшие операции с файлами. Для этого она блокирует создание и удаление файлов, а затем запрашивает исключительный доступ ко всем системным файлам и общий доступ ко всем файлам пользователя. Постепенно активные транзакции либо успешно завершаются, либо получают исключение «файл журнала переполнен». NTFS откатывает и ставит в очередь транзакции, получившие исключение.

После того как NTFS заблокировала транзакционную активность в файлах, как было описано ранее, она вызывает диспетчер кэша, чтобы сбросить на диск незаписанные данные, включая незаписанные данные файла журнала. После того как все безопасно сброшено на диск, NTFS больше не нужны данные из файла журнала. Она переустанавливает начало этого файла на текущую позицию, делая его пустым. Затем NTFS перезапускает транзакции, поставленные в очередь. Помимо короткой паузы в обработке операций ввода-вывода ошибка переполнения файла журнала не оказывает никакого влияния на работающие программы.

Этот сценарий — один из примеров того, как NTFS использует файл журнала не только для восстановления файловой системы, но и для восстановления после ошибок во время нормальной работы. Подробнее об этом говорится в следующем разделе.

Восстановление

NTFS автоматически выполняет восстановление диска при первом обращении программы к тому NTFS после загрузки системы. Если восстановление не требуется, процесс тривиален. Восстановление зависит от двух таблиц, которые NTFS ведет в памяти: таблицы транзакций, действующей так же, как и таблица TxF, и *таблицы «грязных» страниц*, отмечающей, какие страницы в кэше содержат еще не записанные на диск изменения в структуре файловой системы. Эти данные должны быть выгружены на диск во время восстановления.

NTFS вносит запись контрольной точки в файл журнала раз в 5 с. Непосредственно перед этим она вызывает LFS для сохранения текущей копии таблицы транзакций и таблицы «грязных» страниц в файле журнала. Затем NTFS вносит в запись контрольной точки LSN записей журнала, содержащих скопированные таблицы. При восстановлении после сбоя системы NTFS обращается к LFS, чтобы найти записи журнала, содержащие последнюю запись контрольной точки и последние копии таблиц транзакций и «грязных» страниц. Затем она копирует эти таблицы в память.

В файле журнала обычно содержится больше записей обновлений, следующих за последней записью контрольной точки. Эти записи представляют изменения тома, произошедшие после того, как была сделана последняя запись контрольной точки. NTFS должна обновить таблицы транзакций и «грязных» страниц, чтобы включить эти операции. После обновления NTFS использует эти таблицы и содержимое файла журнала для обновления самого тома.

Для восстановления тома NTFS сканирует файл журнала три раза, загружая его в память во время первого прохода, чтобы минимизировать дисковые операции ввода-вывода. Каждый проход имеет определенную цель.

1. Анализ.
2. Повтор транзакций.
3. Отмена транзакций.

Проход анализа

Во время *прохода анализа* (рис. 11.60) NTFS сканирует файл журнала вперед от начала последней операции контрольной точки, чтобы найти записи обновления и использовать их для обновления таблиц транзакций и «грязных» страниц, которые она скопировала в память. Обратите внимание на то, что на рисунке операция контрольной точки сохраняет три записи в файле журнала, а записи обновления могут находиться между ними. Поэтому NTFS должна начинать сканирование в начале операции контрольной точки.

Большинство записей обновления, появляющихся в файле журнала после начала операции контрольной точки, представляют собой изменение либо таблицы транзакций, либо таблицы «грязных» страниц. Если запись обновления является, например, записью типа «транзакция зафиксирована», то транзакция, которую она представляет, должна быть удалена из таблицы транзакций. Аналогично если запись обновления является записью обновления страницы, изменяющей структуру

данных файловой системы, то таблица «грязных» страниц должна быть обновлена, чтобы отразить это изменение.

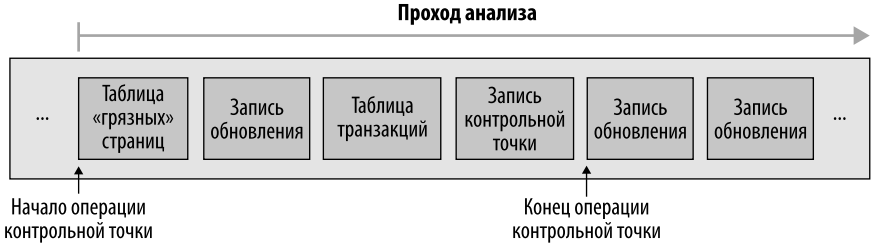


Рис. 11.60. Проход анализа

Когда таблицы обновлены в памяти, NTFS сканирует их, чтобы определить LSN самой старой записи обновления, регистрирующей еще не выполненную на диске операцию. Таблица транзакций содержит LSN незафиксированных (незавершенных) транзакций, а таблица «грязных» страниц включает в себя LSN записей в кэше, которые не были сброшены на диск. LSN самой старой записи обновления, найденной NTFS в этих двух таблицах, определяет, с чего начнется проход повтора. Но если последняя запись контрольной точки старше, то NTFS начнет проход повтора с этого места.

ПРИМЕЧАНИЕ В модели восстановления TxF нет отдельного прохода анализа. Вместо этого TxF выполняет эквивалентную работу в проходе повтора (см. раздел о восстановлении TxF).

Проход повтора

Во время *прохода повтора* (рис. 11.61) NTFS сканирует журнал вперед, начиная с LSN самой старой записи обновления, найденной во время прохода анализа. Она ищет записи обновления страниц, содержащие изменения тома, записанные до сбоя системы, но, возможно, не выведенные на диск. NTFS повторяет эти обновления в кэше.

Когда NTFS достигает конца файла журнала, она обновляет кэш необходимыми изменениями тома, и система отложенной записи диспетчера кэша может начать запись содержимого кэша на диск в фоновом режиме.

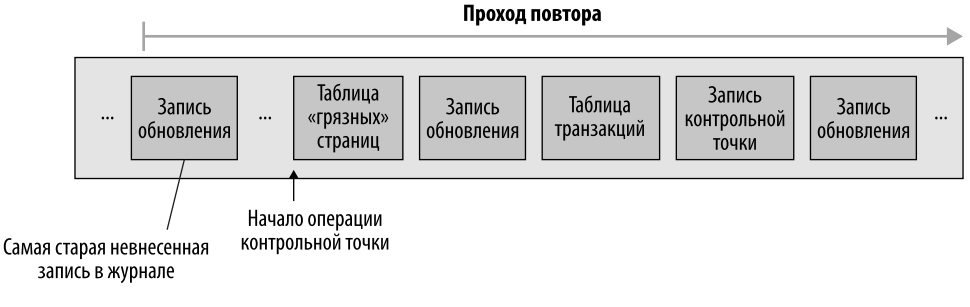


Рис. 11.61. Проход повтора

Проход отмены

После того как NTFS завершит проход повтора, она начинает *проход отмены*, в ходе которого откатывает все транзакции, которые не были зафиксированы во время сбоя системы. На рис. 11.62 показаны две транзакции, находящиеся в файле журнала. Транзакция 1 была выполнена до сбоя питания, а транзакция 2 — нет. NTFS должна отменить транзакцию 2.

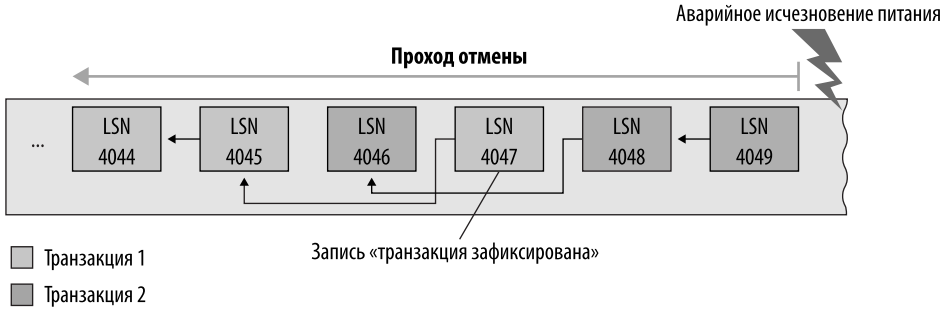


Рис. 11.62. Проход отмены

Предположим, что транзакция 2 создала файл — это операция, состоящая из трех подопераций, каждая из которых имеет свою запись обновления. Записи обновления транзакции связаны с обратными указателями в файле журнала, поскольку обычно они не смежные.

В таблице транзакций NTFS указан LSN последней зарегистрированной записи обновления для каждой незафиксированной транзакции. В этом примере таблица транзакций определяет LSN 4049 как последнюю запись обновления, зарегистрированную для транзакции 2. Как показано на рис. 11.63 (справа налево), NTFS откатывает транзакцию 2.

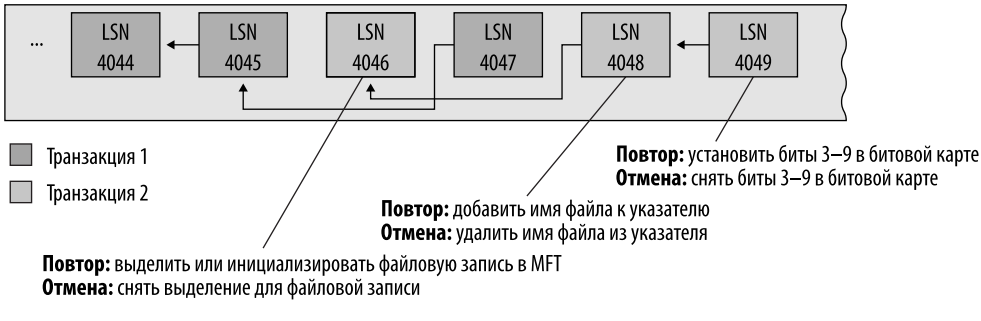


Рис. 11.63. Отмена транзакции

Отыскав LSN 4049, NTFS находит информацию об отмене и выполняет ее, очищая биты с 3-го по 9-й в своей битовой карте распределения. Затем NTFS следует по обратному указателю к LSN 4048, предписывающему удалить новое имя файла из соответствующего указателя имен файлов. Наконец, она следует по

последнему обратному указателю и убирает выделение для файловой записи в MFT, зарезервированной для этого файла, как уточняет запись обновления с LSN 4046. Откат транзакции 2 завершен. Если есть другие незафиксированные транзакции, которые нужно отменить, NTFS выполняет ту же процедуру для их отката. Поскольку отмена транзакций влияет на структуру файловой системы тома, NTFS должна регистрировать операции отмены в файле журнала. В конце концов, во время восстановления питание может снова отключиться, и NTFS придется заново выполнять операции отмены.

После завершения прохода отмены в процедуре восстановления том будет восстановлен до согласованного состояния. В этот момент NTFS готова сбросить изменения кэша на диск, чтобы обеспечить актуальность тома. Однако перед этим она выполняет обратный вызов, зарегистрированный TxF для уведомлений о сбросах LFS. Поскольку TxF и NTFS используют журналирование с упреждающей записью, TxF должен скинуть на диск свой журнал через CLFS до того, как будет скинут журнал NTFS, чтобы обеспечить согласованность собственных метаданных. Аналогично файл TOPS должен быть скинут до файлов журналов, управляемых CLFS. Затем NTFS записывает пустую область перезапуска LFS, чтобы указать, что том согласован и нет необходимости восстанавливать его, если система снова выйдет из строя. Восстановление завершено.

NTFS гарантирует, что восстановление вернет том в некоторое предварительно существовавшее согласованное состояние, но не обязательно в состояние непосредственно перед сбоем системы. Гарантию последнего NTFS дать не может, поскольку для повышения производительности в ней используется алгоритм отложенной фиксации, то есть файл журнала не скидывается на диск сразу после внесения в него записи с фиксацией транзакции. Вместо этого множество записей с фиксацией транзакций собираются в пакет и записываются вместе, либо когда диспетчер кэша вызывает LFS для сброса файла журнала на диск, либо когда LFS вносит в файл журнала запись контрольной точки — раз в 5 с. Еще одна причина, по которой восстановленный том может быть не совсем актуальным, такова: несколько параллельных транзакций могут быть активными в момент сбоя системы и одни их записи с фиксацией транзакций могут попасть на диск, а другие — нет. Связный том, создаваемый восстановлением, включает все обновления тома, чьи записи с фиксацией транзакций попали на диск, и ни одного обновления, чьи записи с фиксацией транзакций туда не попали.

NTFS задействует файл журнала для восстановления тома после сбоя системы, используя при этом важную бесплатную возможность, которую получает от регистрации транзакций. Файловые системы обязательно содержат много кода, посвященного восстановлению после ошибок файловой системы, возникающих в процессе обычного ввода-вывода файлов. Поскольку NTFS регистрирует каждую транзакцию, изменяющую структуру тома, она может использовать файл журнала для восстановления при возникновении ошибки файловой системы и тем самым значительно упростить код обработки ошибок. Описанная ранее ошибка переполнения файла журнала — один из примеров применения файла журнала для восстановления после ошибок.

Большинство ошибок ввода-вывода, которые получает программа, не являются ошибками файловой системы и поэтому не могут быть целиком решены NTFS. Например, при вызове на создание файла NTFS может начать с создания файловой записи в MFT, а затем ввести имя нового файла в индекс каталога. Но когда NTFS

попытается выделить место для файла в своей битовой карте, она может обнаружить, что диск переполнен и запрос на создание выполнить не получится. В этом случае NTFS использует информацию из файла журнала, чтобы отменить часть операции, которую она уже выполнила, и убрать выделение структур данных, зарезервированных для файла. Затем она возвращает ошибку «диск переполнен» вызывающей стороне, которая должна соответствующим образом отреагировать на нее.

Восстановление плохих кластеров NTFS

Диспетчер томов, входящий в состав Windows (VolMgr), может восстановить данные из поврежденного сектора на диске отказоустойчивого тома, но если жесткий диск не выполняет переадресацию плохих секторов или на нем закончились свободные сектора, то диспетчер тома не может заменить плохие сектора. Вместо этого, когда файловая система читает из такого сектора, диспетчер томов восстанавливает данные и возвращает файловой системе предупреждение о том, что существует только одна копия данных.

Файловая система FAT не реагирует на это предупреждение. Более того, ни FAT, ни диспетчер томов не отслеживают поврежденные сектора, поэтому пользователь должен запустить утилиту Chkdsk или Format, чтобы диспетчер томов не восстанавливал многократно данные для файловой системы. И Chkdsk, и Format не слишком подходят для удаления поврежденных секторов. У Chkdsk поиск и удаление плохих секторов могут занять много времени, а Format стирает все данные из раздела, который форматирует.

Эквивалентом замены плохого сектора диспетчером томов в файловой системе операции является то, что NTFS динамически заменяет кластер, содержащий плохой сектор, и отслеживает последний, чтобы он не использовался повторно. Напомним, что NTFS сохраняет переносимость, обращаясь к логическим кластерам, а не к физическим секторам. NTFS выполняет эти функции, когда диспетчер томов не может заменить плохой сектор. Когда диспетчер томов выдает предупреждение о плохом секторе или драйвер жесткого диска выдает ошибку плохого сектора, NTFS выделяет новый кластер взамен содержащего плохой сектор. Она копирует данные, восстановленные диспетчером томов, в новый кластер, чтобы восстановить резервирование данных.

На рис. 11.64 показана запись MFT для пользовательского файла с плохим кластером в одном из его прогонов данных в том виде, в котором она существовала до того, как кластер испортился. При получении ошибки плохого сектора NTFS переназначает кластер, содержащий сектор, в свой файл плохого кластера `$BadClus`. Это предотвращает выделение плохого кластера для другого файла. Затем NTFS выделяет новый кластер для файла и изменяет сопоставления VCN и LCN файла, чтобы они указывали на новый кластер. Этот процесс переадресации плохих кластеров, представленный ранее в данной главе, показан на рис. 11.64. Кластер 1357, содержащий плохой сектор, должен быть заменен хорошим кластером.

Ошибки плохих секторов нежелательны, но когда они возникают, сочетание NTFS и диспетчера томов обеспечивает наилучшее возможное решение. Если поврежденный сектор находится на резервированном томе, диспетчер томов восстанавливает данные и заменяет сектор, если может. Если он не может этого сделать, то возвращает NTFS предупреждение и NTFS заменяет кластер, содержащий плохой сектор.

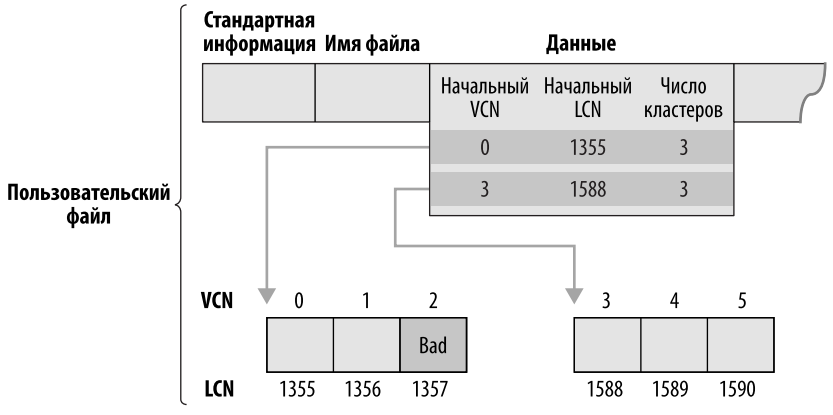


Рис. 11.64. Запись MFT для пользовательского файла с плохим кластером

Если том не настроен как резервированный, данные в плохом секторе восстановить невозможно. Когда том отформатирован в системе FAT и диспетчер томов не может восстановить данные, чтение из плохого сектора дает неопределенные результаты. Если некоторые управляющие структуры файловой системы находятся в поврежденном секторе, то может быть потерян весь файл или группа файлов, а возможно, и весь диск. В лучшем случае будет потеряна часть данных в поврежденном файле, чаще — все данные из файла за пределами поврежденного сектора. Более того, файловая система FAT, скорее всего, перераспределит поврежденный сектор на тот же или другой файл в томе, что приведет к повторному возникновению проблемы.

Как и другие файловые системы, NTFS не может восстановить данные из поврежденного сектора без помощи диспетчера томов. Однако NTFS значительно ограничивает ущерб, который может нанести поврежденный сектор. Если NTFS обнаруживает плохой сектор во время операции чтения, то она перераспределит кластер, в котором находится сектор (рис. 11.65). Если том не настроен как резервированный, то NTFS возвращает вызывающей программе ошибку чтения данных. Хотя данные, находившиеся в этом кластере, будут потеряны, остальная часть файла и файловая система останутся нетронутыми. Вызывающая программа сможет соответствующим образом отреагировать на потерю данных, а плохой кластер не будет повторно использоваться в будущих выделениях диска. Если NTFS обнаруживает плохой кластер при записи, а не при чтении, то она перераспределяет кластер перед записью и, таким образом, не теряет данные и не порождает сообщение об ошибке.

Те же процедуры восстановления применяются, если данные файловой системы хранятся в секторе, который испортился. Если поврежденный сектор находится на резервированном томе, NTFS заменяет кластер динамически, используя данные, восстановленные диспетчером томов. Если том нерезервированный, данные не могут быть восстановлены, поэтому NTFS устанавливает бит в файле метаданных `$Volume`, указывающий на повреждение тома. Утилита NTFS Chkdsk проверяет этот бит при следующей перезагрузке системы, и если он установлен, то она запускается, восстанавливая повреждение файловой системы путем реконструкции метаданных NTFS.

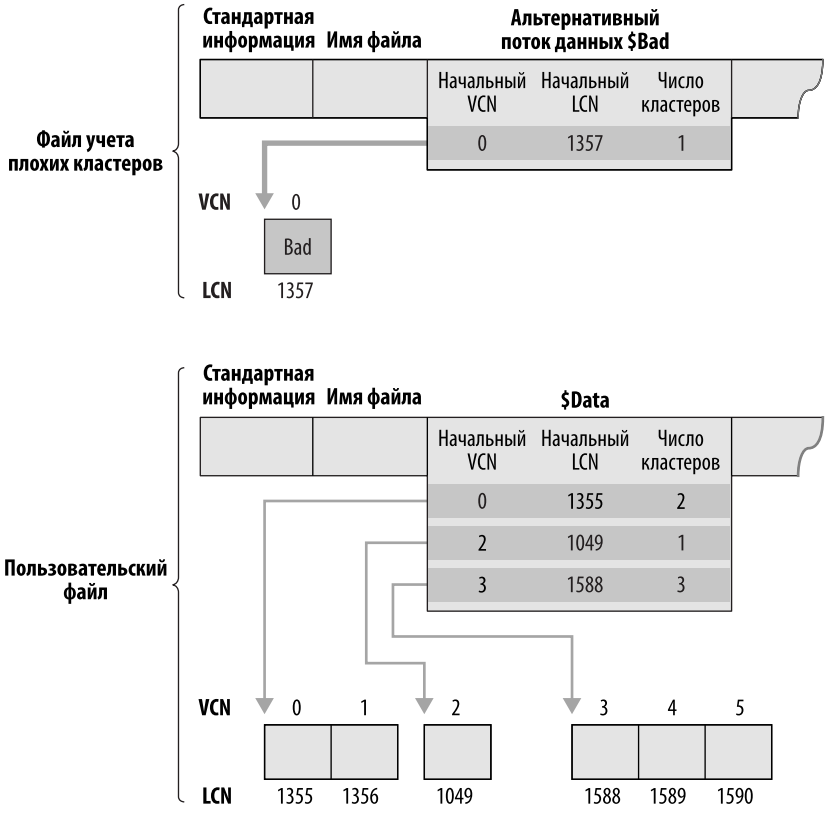


Рис. 11.65. Перераспределение плохих кластеров

В редких случаях повреждение файловой системы может произойти даже на дисках с отказоустойчивой конфигурацией. Двойная ошибка может уничтожить как данные файловой системы, так и способы их восстановления. Если система аварийно завершает работу, когда NTFS делает зеркальную копию записи файла MFT, например указателя имен файлов или файла журнала, то зеркальная копия данных файловой системы может оказаться не полностью обновленной. Если система перезагрузится и на первичном диске произойдет ошибка плохого сектора в том же месте, где находится неполная запись на зеркале диска, NTFS не сможет восстановить корректные данные с зеркала. В NTFS реализована специальная схема обнаружения таких повреждений в данных файловой системы. Если она обнаруживает несоответствие, то устанавливает бит повреждения в файле тома, что заставляет Chkdsk восстанавливать метаданные NTFS при следующей перезагрузке системы. Поскольку повреждения файловой системы на дисках с отказоустойчивой конфигурацией случаются редко, Chkdsk нужна нечасто. Она поставляется в качестве меры предосторожности, а не как первая стратегия восстановления данных.

Использование Chkdsk в NTFS значительно отличается от ее применения в файловой системе FAT. Прежде чем записать что-либо на диск, FAT устанавливает «грязный» бит тома, а после завершения изменений сбрасывает его. Если в момент

сбоя системы выполняется какая-либо операция ввода-вывода, «грязный» бит остается установленным и Chkdsk запускается при перезагрузке системы. В NTFS Chkdsk запускается только при обнаружении неожиданных или нечитаемых данных файловой системы, и NTFS не может восстановить данные с резервированного тома или из резервированных структур файловой системы на одном томе. Загрузочный сектор системы дублируется в последнем секторе тома, как и части MFT (`$MftMirr`), необходимые для загрузки системы и выполнения процедуры восстановления NTFS. Такое резервирование гарантирует, что NTFS всегда сможет загрузиться и восстановиться.

В табл. 11.11 кратко изложено, что происходит при повреждении сектора на дисковом томе, отформатированном под одну из файловых систем, поддерживаемых Windows, в соответствии с различными условиями, описанными в этом разделе.

Если том, на котором появился плохой сектор, отказоустойчивый — зеркальный RAID-1 или RAID-5/RAID-6, а жесткий диск поддерживает замену плохого сектора и на нем не закончились свободные сектора, то не имеет значения, какая файловая система применяется, FAT или NTFS. Диспетчер томов заменяет поврежденный сектор без вмешательства пользователя или файловой системы.

Если поврежденный сектор находится на жестком диске, не поддерживающем замену поврежденного сектора, то файловая система отвечает за замену (перераспределение) поврежденного сектора, а в случае NTFS — кластера, в котором находится поврежденный сектор. Файловая система FAT не обеспечивает перераспределения секторов или кластеров. Преимущества перераспределения кластеров в NTFS заключаются в том, что плохие участки в файле можно исправить без вреда для файла или файловой системы в зависимости от ситуации и плохой кластер больше никогда не будет использоваться.

Таблица 11.11. Сводка сценариев восстановления данных NTFS

Сценарий	С диском, поддерживающим перераспределение плохих секторов и имеющим запасные сектора	С диском, на котором не выполняется перераспределение плохих секторов или нет запасных секторов
Отказоустойчивый том[1]	1. Диспетчер томов восстанавливает данные. 2. Диспетчер томов заменяет плохой сектор. 3. Файловая система остается в неведении об ошибке	1. Диспетчер томов восстанавливает данные. 2. Диспетчер томов отправляет данные и ошибку плохого сектора в файловую систему. 3. NTFS выполняет перераспределение кластеров
Неотказоустойчивый том	1. Диспетчер томов не может восстановить данные. 2. Диспетчер томов отправляет ошибку плохого сектора в файловую систему. 3. NTFS выполняет перераспределение кластеров. Данные теряются[2]	1. Диспетчер томов не может восстановить данные. 2. Диспетчер томов отправляет ошибку плохого сектора в файловую систему. 3. NTFS выполняет перераспределение кластеров. Данные теряются

[1] Отказоустойчивый том — это либо зеркальный набор RAID-1, либо набор RAID-5.

[2] При операции записи данные не теряются — NTFS перераспределяет кластер перед записью.

Самовосстановление

На современных многотерабайтных устройствах хранения данных вывод тома в автономный режим для проверки согласованности может привести к многочасовому перерыву в обслуживании. Понимая, что многие повреждения диска локализуются в одном файле или части метаданных, NTFS реализует функцию самовосстановления для устранения повреждений, пока том остается подключенным. Когда NTFS обнаруживает повреждение, она предотвращает доступ к поврежденному файлу или файлам и создает рабочий поток системы, который исправляет поврежденные структуры данных, подобно Chkdsk, разрешая доступ к восстановленным файлам после завершения работы. Во время этой операции доступ к другим файлам происходит в обычном режиме, что сводит к минимуму перебои в работе службы.

С помощью команды `fsutil repair set` можно просмотреть и задать параметры восстановления тома, приведенные в табл. 11.12. Утилита Fsutil использует код управления файловой системой `FSCTL_SET_REPAIR` для установки этих параметров, сохраняющихся в VCB для тома.

Во всех случаях, в том числе при отключении визуального предупреждения, присутствующем по умолчанию, NTFS будет регистрировать все операции самовосстановления в журнале событий системы.

Таблица 11.12. Поведение самовосстановления NTFS

Флаг	Поведение
SET_REPAIR_ENABLED	Включает самовосстановление для тома
SET_REPAIR_WARN_ ABOUT_DATA_LOSS	Если процесс самовосстановления не может полностью восстановить файл, то указывает, следует ли визуально предупреждать пользователя
SET_REPAIR_ DISABLED_AND_ BUGCHECK_ON_ CORRUPTION	Если параметр реестра NTFS NtfsBugCheckOnCorrupt задан с помощью поведенческого набора fsutil NtfsBugCheckOnCorrupt 1 и этот флаг установлен, то система завершится со STOP-ошибкой 0x24, указывающей на повреждение файловой системы. Этот параметр автоматически снимается во время загрузки, чтобы избежать повторных циклов перезагрузки

Помимо периодического автоматического самовосстановления, NTFS поддерживает циклы самовосстановления, инициируемые вручную (этот тип самовосстановления называется *проактивным*) с помощью управляющих кодов `FSCTL_INITIATE_REPAIR` и `FSCTL_WAIT_FOR_REPAIR`, которые могут быть инициированы командами `fsutil repair initiate` и `fsutil repair wait` соответственно. Это позволяет пользователю принудительно восстановить определенный файл и дождаться завершения этого процесса.

Чтобы проверить состояние механизма самовосстановления, используйте управляющий код `FSCTL_QUERY_REPAIR` или команду `fsutil repair query`, как показано далее:

```
C:\>fsutil repair query c:
Self healing state on c: is: 0x9

 Values: 0x1 — Enable general repair.
         0x9 — Enable repair and warn about potential data loss.
        0x10 — Disable repair and bugcheck once on first corruption.
```

Проверка диска в режиме онлайн и быстрое восстановление

В редких случаях, когда повреждения диска не устраняются драйвером файловой системы NTFS с помощью самовосстановления, службы файлов журналов и т. д., необходимо запустить инструмент Windows Check Disk и перевести том в автономный режим. Возможны различные причины повреждения диска: будь то ошибки носителя с жесткого диска или временные ошибки памяти, повреждения в метаданных файловой системы случаются. На больших файловых серверах, занимающих несколько терабайт дискового пространства, полная проверка диска может занять несколько дней. Столь длительное отключение тома в подобных ситуациях обычно недопустимо.

До Windows 8 в NTFS была реализована более простая модель, в которой том файловой системы обозначался либо как нормально работающий, либо нет — через «грязный» бит в атрибуте `$VOLUME_INFORMATION`. В этой модели том выводился из сети на время, необходимое для устранения повреждений файловой системы и приведения тома в нормальное состояние. Время простоя было прямо пропорционально количеству файлов в томе. В Windows 8 с целью сокращения или исключения времени простоя, вызванного повреждением файловой системы, были переработаны модель здоровья NTFS и проверка диска.

В новой модели появились компоненты, которые совместно обеспечивают работу инструмента проверки диска в режиме онлайн и значительно сокращают время простоя в случае обнаружения серьезных повреждений файловой системы. Драйвер файловой системы NTFS способен выявить несколько типов повреждений во время обычного ввода-вывода системы. Если повреждение обнаружено, NTFS пытается самостоятельно восстановить его (см. ранее). Если это не удается, драйвер файловой системы NTFS вносит новую *запись о повреждении* в поток `$Verify` файла `\$Extend\$RmMetadata\$Repair`.

Запись о повреждении — это общая структура данных, которую NTFS задействует для описания повреждений метаданных. Она используется как в памяти, так и на диске. Запись о повреждении представлена заголовком фиксированного размера, содержащим информацию о версии и флаги и уникально представляющим тип записи через GUID, описание переменного размера типа произошедшего повреждения и необязательный контекст.

После корректного добавления записи NTFS издает событие ETW через собственный поставщик событий Microsoft-Windows-Ntfs-UBPM. Это событие потребляется диспетчером управления службами, который запускает службу Spot Verifier (подробнее о запускаемых службах можно прочитать в главе 10).

Служба Spot Verifier, реализованная в библиотеке Svsvc.dll, проверяет, не является ли сигнал о повреждении ложным срабатыванием. Некоторые повреждения оказываются временными из-за проблем с памятью и могут не быть результатом реального повреждения диска. Записи в потоке `$Verify` удаляются во время проверки Spot Verifier. Если повреждение, о котором есть запись, — это не ложное срабатывание, то Spot Verifier устанавливает бит проактивного сканирования (P-бит) в атрибуте `$VOLUME_INFORMATION` тома, который запускает онлайн-сканирование файловой системы. Онлайн-сканирование выполняется Proactive Scanner, запускаемым в подходящее время в качестве задачи обслуживания планировщиком задач Windows (задача находится в `Microsoft\Windows\Chkdsk`, как показано на рис. 11.66).

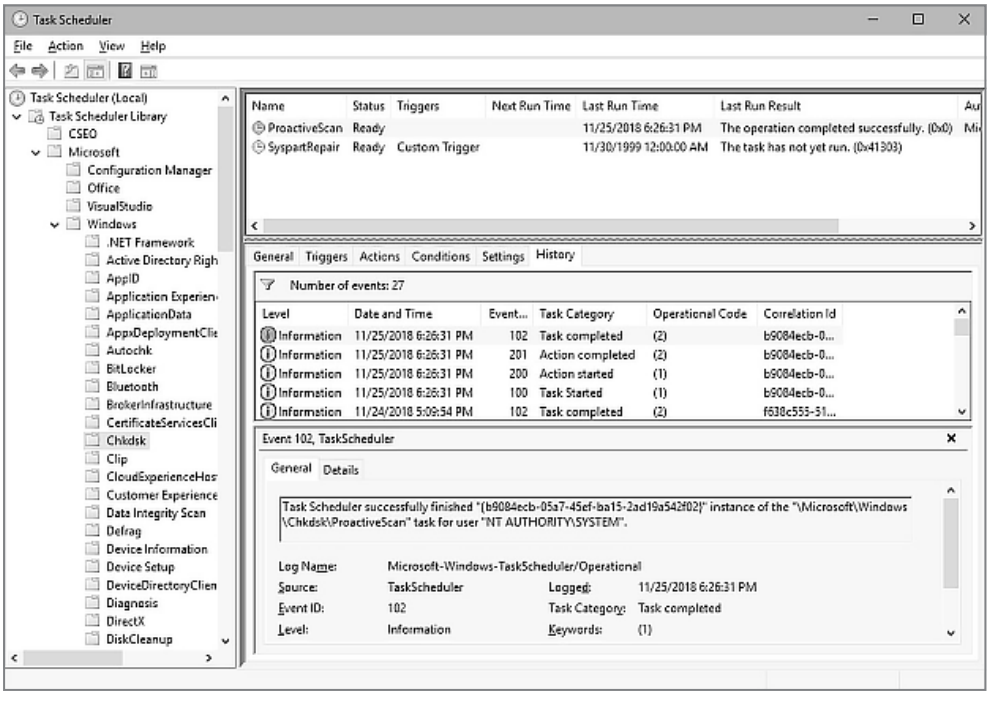


Рис. 11.66. Задача обслуживания Proactive Scan

Proactive Scanner реализован в библиотеке Untfs.dll, импортируемой инструментом Windows Check Disk, Chkdsk.exe. При запуске Proactive Scanner делает снимок целевого тома с помощью службы Volume Shadow Copy и запускает полную проверку Check Disk на теневом томе. Теневой том доступен только для чтения. Код проверки диска обнаруживает это и вместо того, чтобы напрямую исправлять ошибки, использует функцию самовосстановления NTFS, чтобы попытаться автоматически исправить повреждения. Если это не удается, он отправляет код `FSCTL_CORRUPTION_HANDLING` драйверу файловой системы, который, в свою очередь, создает запись в потоке `$Corrupt` файла метаданных `\$Extend\$RmMetadata\$Repair` и устанавливает «грязный» бит тома.

Этот «грязный» бит имеет несколько иное значение по сравнению с предыдущими версиями Windows. Атрибут `$VOLUME_INFORMATION` корневого пространства имен NTFS по-прежнему содержит «грязный» бит, а также P-бит, используемый для затребования проактивного сканирования, и F-бит, применяемый для затребования полной проверки диска из-за серьезности конкретного повреждения. «Грязный» бит устанавливается в 1 драйвером файловой системы, если включен P-бит или F-бит либо поток `$Corrupt` содержит одну или несколько записей о повреждениях.

Если повреждение все еще не устранено, то на данном этапе нет других возможностей исправить его, когда том находится в автономном режиме, — это не обязательно требует немедленного размонтирования тома. Spot Fixer — новый компонент, общий для инструментов Check Disk и Autocheck. Он потребляет записи, вставленные Proactive Scanner в поток `$Corrupt`. Во время загрузки встроенное

приложение Autocheck обнаруживает, что том «загрязнен», но вместо полной проверки диска исправляет только поврежденные записи, расположенные в потоке `$Corrupt`, и эта операция занимает всего несколько секунд. На рис. 11.67 показана сводка различных методик исправления, реализованных в описанных ранее компонентах файловой системы NTFS.

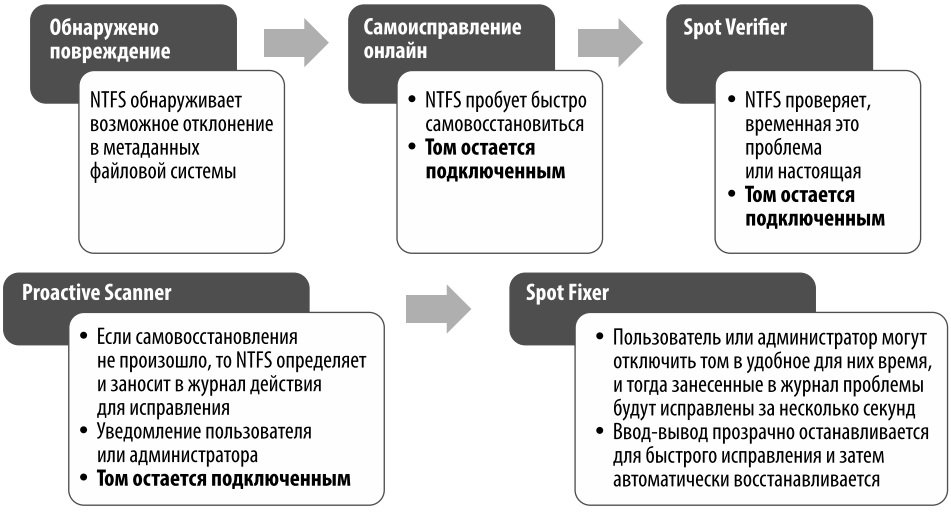


Рис. 11.67. Схема, описывающая компоненты, взаимодействующие для обеспечения проверки диска онлайн и быстрого восстановления повреждений томов NTFS

Проактивное сканирование можно запустить вручную для тома с помощью команды `chkdsk/scan`. Таким же образом Spot Fixer может быть запущен инструментом Check Disk с помощью аргумента командной строки `/spotfix`.

ЭКСПЕРИМЕНТ. Опробование онлайн-проверки диска

Можно опробовать онлайн-проверку диска, проведя простой эксперимент. Предполагая, что проверяется том `D:`, начните с воспроизведения большого видеопотока с диска `D`. Тем временем откройте административное окно командной строки и запустите проверку с помощью следующей команды:

```
C:\>chkdsk d: /scan
The type of the file system is NTFS.
Volume label is DATA.

Stage 1: Examining basic file system structure ...
  4041984 file records processed.
File verification completed.
  3778 large file records processed.
  0 bad file records processed.

Stage 2: Examining file name linkage ...
Progress: 3454102 of 4056090 done; Stage: 85%; Total: 51%; ETA:    0:00:43 ..
```

Будет видно, что видеопоток не останавливается и продолжает плавно воспроизводиться. Если интерактивная проверка диска обнаружит ошибку, которую не сможет исправить при смонтированном томе, та будет помещена в поток `$Corrupt` системного файла `$Repair`. Для исправления ошибок потребуется демонтаж тома, но исправление будет очень быстрым. В этом случае можно просто перезагрузить машину или вручную запустить Spot Fixer посредством командной строки:

```
C:\>chkdsk d: /spotfix
```

Если применить Spot Fixer, то видеопоток будет прерван, поскольку том необходимо размонтировать.

ЗАШИФРОВАННАЯ ФАЙЛОВАЯ СИСТЕМА

В Windows есть функция шифрования всего тома под названием Windows BitLocker Drive Encryption. BitLocker шифрует и защищает тома от офлайн-атак, но после загрузки системы ее работа завершается. Система шифрования файлов (Encrypting File System, EFS) защищает отдельные файлы и каталоги от других аутентифицированных пользователей системы. При подборе способа защиты данных нельзя выбирать между BitLocker и EFS — каждый из них обеспечивает защиту от определенных непересекающихся угроз. Вместе BitLocker и EFS обеспечивают эшелонированную оборону данных в системе.

Парадигма, используемая EFS, заключается в шифровании файлов и каталогов с помощью симметричного шифрования, когда один и тот же ключ применяется для шифрования и расшифровки файла. Ключ симметричного шифрования затем шифруется с помощью асимметричного шифрования для каждого пользователя, которому предоставлен доступ к файлу. В асимметричном шифровании применяется один ключ (открытый) для шифрования и другой (закрытый) для расшифровки. Подробное теоретическое описание этих методов выходит за рамки тематики данной книги, хороший учебник можно найти по адресу https://docs.microsoft.com/en-us/windows/desktop/SecCrypto/cryptography-essentials.

EFS работает с API Windows Cryptography Next Generation (CNG) и поэтому может быть настроена на использование любого алгоритма, поддерживаемого CNG или добавленного в него. По умолчанию EFS применяет Advanced Encryption Standard (AES) для симметричного шифрования с 256-битным ключом и алгоритм открытого ключа Ривеста — Шамира — Адлемана (RSA) для несимметричного шифрования с 2048-битным ключом.

Пользователи могут зашифровать файлы с помощью Проводника Windows, открыв диалоговое окно свойств файла и нажав кнопку Дополнительные атрибуты (Advanced), а затем выбрав опцию Шифровать содержимое для защиты данных (Encrypt Contents To Secure Data) (рис. 11.68). Файл может быть или зашифрован, или сжат, но не то и другое одновременно. Пользователи могут шифровать файлы также с помощью утилиты командной строки под названием Cipher, %SystemRoot%\System32\Cipher.exe, или программно, используя API Windows, такие как `EncryptFile` и `AddUsersToEncryptedFile`.

Рис. 11.68. Шифрование файлов с помощью диалогового окна дополнительных атрибутов

Windows автоматически шифрует файлы, находящиеся в каталогах, обозначенных как зашифрованные. Когда файл зашифрован, EFS генерирует для него случайное число, которое называет *ключом шифрования файла* (File Encrypton Key, FEK). EFS использует FEK для шифрования содержимого файла с помощью симметричного шифрования. Затем EFS шифрует FEK с помощью асимметричного открытого ключа пользователя и сохраняет зашифрованный FEK в альтернативном потоке данных `$EFS` для файла. Источник открытого ключа может быть административно задан как исходящий из присвоенного сертификата X.509, или это может быть смарт-карта, или он может быть сгенерирован случайным образом. В последнем случае источник затем будет добавлен в хранилище сертификатов пользователя, которое можно просмотреть с помощью диспетчера сертификатов `%SystemRoot%\System32\Certmgr.msc`. После того как EFS выполнит эти действия, файл будет защищен и другие пользователи не смогут расшифровать данные без расшифрованного FEK файла, а расшифровать FEK они не смогут без закрытого ключа пользователя.

Алгоритмы симметричного шифрования обычно очень быстрые, что позволяет применять их для шифрования больших объемов данных, например файловых данных. Однако у симметричных алгоритмов шифрования есть слабое место: можно обойти их защиту, если получить ключ. Если несколько пользователей хотят совместно работать с одним зашифрованным файлом, защищенным лишь с помощью симметричного шифрования, то каждому из них потребуется доступ к FEK этого файла. Если оставить FEK незашифрованным, то это, очевидно, будет проблемой с безопасностью, но если зашифровать FEK один раз, то всем пользователям потребуется один и тот же ключ расшифровки FEK — еще одна потенциальная проблема с безопасностью.

Обеспечение безопасности FEK — сложная проблема, которую EFS решает с помощью архитектуры шифрования, основанной на открытом ключе. Шифрование FEK для отдельных пользователей, имеющих доступ к файлу, позволяет нескольким пользователям совместно применять зашифрованный файл. EFS может шифровать FEK открытым ключом каждого пользователя и хранить зашифрованный FEK каждого пользователя в потоке данных `$EFS` файла. Любой может получить доступ к открытому ключу пользователя, но никто не может применять этот ключ для расшифровки данных, зашифрованных с помощью открытого ключа. Расшифровать файл можно только с помощью закрытого ключа, к которому операционная система должна иметь доступ. Закрытый ключ пользователя расшифровывает его зашифрованную копию FEK. Алгоритмы на основе открытых ключей обычно медленные, но EFS применяет их только для шифрования FEK. Разделение

управления ключами на общедоступный открытый и закрытый ключи делает его немного проще, чем в алгоритмах симметричного шифрования, и решает дилемму обеспечения безопасности FEK.

Несколько компонентов действуют вместе, чтобы обеспечить работу EFS, как показано на диаграмме архитектуры EFS на рис. 11.69. Поддержка EFS включена в драйвер NTFS. Всякий раз, когда NTFS сталкивается с зашифрованным файлом, выполняются содержащиеся в ней функции EFS. Они шифруют и расшифровывают данные файла, когда приложения обращаются к зашифрованным файлам. Хотя EFS хранит FEK вместе с данными файла, открытые ключи пользователей шифруют FEK. Чтобы зашифровать или расшифровать данные файла, EFS должна расшифровать FEK файла с помощью служб управления ключами CNG, которые находятся в пользовательском режиме.

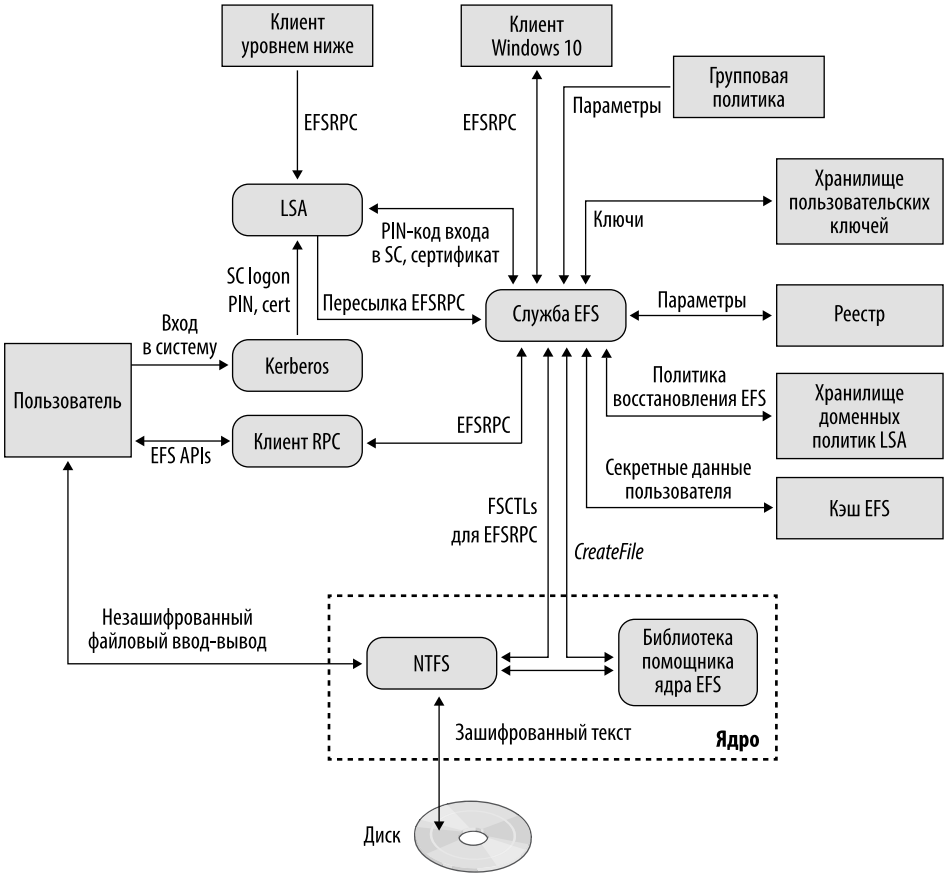


Рис. 11.69. Архитектура EFS

Подсистема Local Security Authority Subsystem (LSASS, %SystemRoot%\System32\Lsass.exe) управляет сеансами входа в систему, а также содержит службу EFS,

Efssvc.dll. Например, когда EFS нужно расшифровать FEK для расшифровки данных файла, к которому хочет получить доступ пользователь, NTFS отправляет запрос службе EFS внутри LSASS.

Первичное шифрование файла

Драйвер NTFS вызывает свои вспомогательные функции EFS, когда встречает зашифрованный файл. Атрибуты файла фиксируют, что файл зашифрован, точно так же как файл фиксирует, что он сжат (об этом говорилось ранее в этой главе). NTFS имеет специальные интерфейсы для преобразования файла из незашифрованного состояния в зашифрованное, но в основном этим процессом управляют компоненты пользовательского режима. Как было сказано ранее, Windows позволяет зашифровать файл двумя способами: с помощью утилиты командной строки `cipher` или установив флажок Шифровать содержимое для защиты данных (Encrypt Contents To Secure Data) в диалоговом окне Дополнительные атрибуты (Advanced Attributes) для файла в Проводнике Windows. И Проводник Windows, и команда `cipher` применяют API Windows `EncryptFile`.

EFS хранит в зашифрованном файле только один блок информации, он содержит записи для всех пользователей, совместно применяющих файл. Эти записи называются *ключевыми*, и EFS хранит их в поле расшифровки данных (data decryption field, DDF) в данных EFS файла. Коллекция из нескольких ключевых записей называется *кольцом ключей*, поскольку, как уже говорилось, EFS позволяет нескольким пользователям совместно задействовать зашифрованные файлы.

На рис. 11.70 показаны формат информации EFS и формат записи ключа для файла. EFS хранит в первой части записи ключа достаточно информации, чтобы точно описать открытый ключ пользователя. Эти данные включают идентификатор безопасности пользователя SID (обратите внимание на то, что присутствие SID не гарантируется), имя контейнера, в котором хранится ключ, имя криптопровайдера и хеш сертификата асимметричной пары ключей. В процессе расшифровки применяется только хеш сертификата асимметричной пары ключей. Вторая часть записи ключа содержит зашифрованную версию FEK. EFS использует CNG для шифрования FEK с помощью выбранного алгоритма асимметричного шифрования и открытого ключа пользователя.

EFS хранит информацию о записях ключей восстановления в поле восстановления данных (data recovery field, DRF) файла. Формат записей DRF идентичен формату записей DDF. Назначение DRF — позволить назначенным учетным записям, так называемым агентам восстановления, расшифровать файл пользователя, когда администратор должен получить доступ к его данным. Предположим, сотрудник компании забыл свой пароль для входа в систему. Администратор может сбросить пароль пользователя, но без агентов восстановления никто не сможет восстановить его зашифрованные данные.

Агенты восстановления определяются с помощью политики безопасности Encrypted Data Recovery Agents на локальном компьютере или в домене. Эта политика доступна из приложения Local Security Policy MMC (рис. 11.71). Когда используется мастер добавления агента восстановления (для этого щелкните правой кнопкой мыши на пункте Шифрование файловой системы (Encrypting File System), затем выберите Добавить агент восстановления данных (Add Data Recovery Agent)),

можно добавить агентов восстановления и указать, какие пары закрытых и открытых ключей, назначенные их сертификатами, они используют для восстановления EFS. Lsasrv — служба Local Security Authority, о которой рассказывается в главе 7, интерпретирует политику восстановления при инициализации и получении уведомления об изменении политики восстановления. EFS создает ключевую запись DRF для каждого агента восстановления с помощью криптопровайдера, зарегистрированного для восстановления EFS.

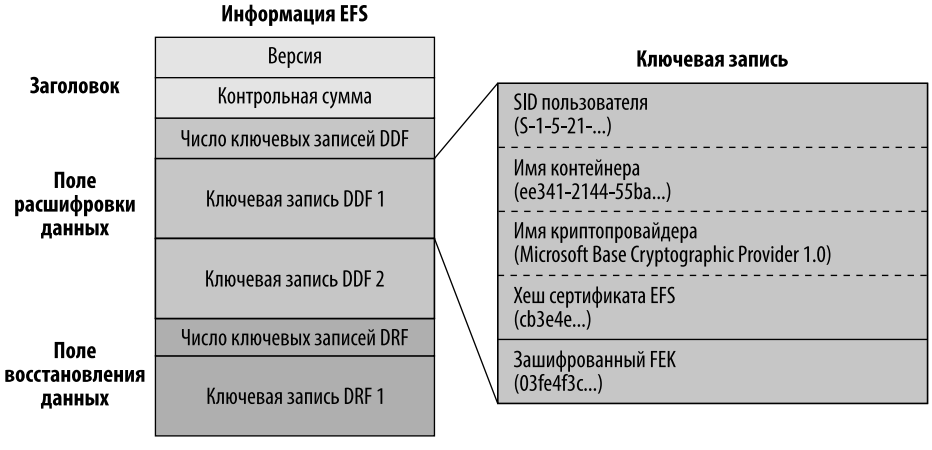


Рис. 11.70. Формат информации EFS и ключевых записей

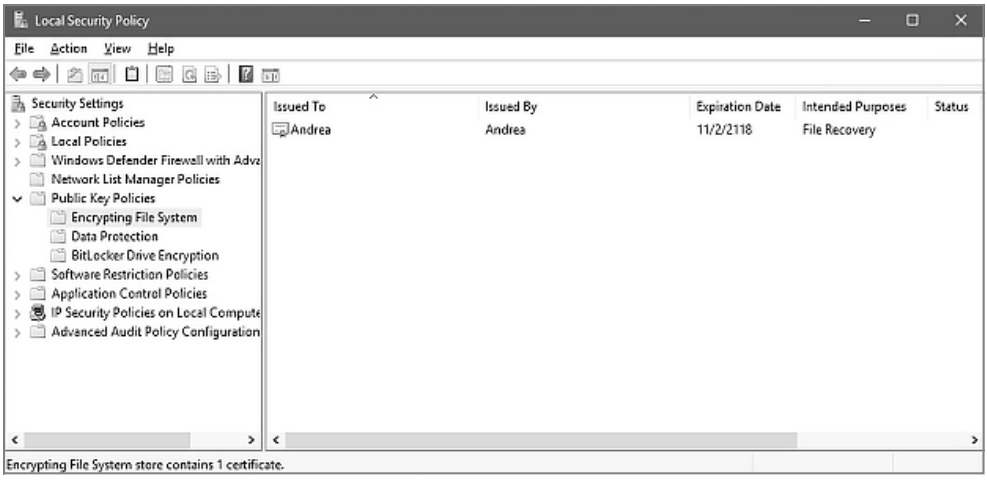


Рис. 11.71. Групповая политика Encrypted Data Recovery Agents

Пользователь может создать собственный сертификат Data Recovery Agent (DRA) с помощью команды `cipher/r`. Созданный файл частного сертификата может быть импортирован через Recovery Agent Wizard и приложение Certificates

контроллера домена или машины, на которой администратор должен иметь возможность расшифровывать зашифрованные файлы.

На последнем этапе создания EFS-информации для файла Lsasrv вычисляет контрольную сумму для DDF и DRF, используя хеш-функцию MD5 из Base Cryptographic Provider 1.0. Lsasrv сохраняет результат вычисления контрольной суммы в заголовке информации EFS. EFS проверяет эту контрольную сумму при расшифровке, чтобы убедиться, что содержимое информации EFS файла не было повреждено или подделано.

Шифрование данных файла

Когда пользователь шифрует существующий файл, происходит следующее.

1. Служба EFS открывает файл для исключительного доступа.
2. Все потоки данных в файле копируются во временный файл с открытым текстом во временном каталоге системы.
3. FEK генерируется случайным образом и применяется для шифрования файла с помощью AES-256.
4. Создается DDF, содержащий FEK, зашифрованный с помощью открытого ключа пользователя. EFS автоматически получает открытый ключ пользователя из его сертификата шифрования файлов X.509 версии 3.
5. Если агент восстановления был назначен с помощью групповой политики, то создается DRF, содержащий FEK, зашифрованный с помощью RSA и открытого ключа агента восстановления.
6. EFS автоматически получает открытый ключ агента восстановления для восстановления файлов из его сертификата X.509 версии 3, который хранится в политике восстановления EFS. Если есть несколько агентов восстановления, то копия FEK шифруется с помощью открытого ключа каждого агента и создается DRF для хранения каждого зашифрованного FEK.

ПРИМЕЧАНИЕ Свойство восстановления файлов в сертификате является примером поля расширенного использования ключей (EKU). Расширение EKU и расширенное свойство определяют и ограничивают допустимые варианты применения сертификата. Восстановление файлов — одно из полей EKU, определенных компанией Microsoft в рамках инфраструктуры открытых ключей Microsoft (public key infrastructure, PKI).

---

7. EFS записывает зашифрованные данные вместе с DDF и DRF обратно в файл. Поскольку симметричное шифрование не добавляет данных, увеличение размера файла после шифрования минимально. Метаданные, состоящие в основном из зашифрованных FEK, обычно занимают менее 1 Кбайт. Размер файла в байтах до и после шифрования обычно указывается одинаковый.
8. Временный файл с открытым текстом удаляется.

Когда пользователь сохраняет файл в папке, настроенной на шифрование, процесс происходит аналогично, за исключением того, что временный файл не создается.

Процесс расшифровки

Когда приложение получает доступ к зашифрованному файлу, расшифровка происходит следующим образом.

1. NTFS распознает, что файл зашифрован, и отправляет запрос драйверу EFS.
2. Драйвер EFS извлекает DDF и передает его службе EFS.
3. Служба EFS извлекает закрытый ключ пользователя из его профиля и применяет его для расшифровки DDF и получения FEK.
4. Служба EFS передает FEK обратно драйверу EFS.

ПРИМЕЧАНИЕ Когда приложение открывает файл, расшифровываются только те его разделы, которые использует приложение, поскольку EFS задействует режим сцепления блоков шифротекста. Если пользователь удаляет атрибут шифрования из файла, то поведение будет другим. В этом случае весь файл расшифровывается и переписывается в виде обычного текста.

---

5. Драйвер EFS применяет FEK для расшифровки разделов файла, необходимых для работы приложения.
6. Драйвер EFS возвращает расшифрованные данные в NTFS, которая затем отправляет их запрашивающему устройству.

Резервное копирование зашифрованных файлов

Важным аспектом конструкции любого средства шифрования файлов является то, что данные файла должны быть доступны в незашифрованном виде только приложениям, которые обращаются к нему с помощью средства шифрования. Это ограничение особенно сильно влияет на утилиты резервного копирования, используемые архивными системами для хранения файлов. EFS решает эту проблему, предоставляя утилитам резервного копирования возможность создавать резервные копии и восстанавливать файлы в зашифрованном виде. Таким образом, утилитам не нужно уметь расшифровывать данные файлов и нет необходимости шифровать их в своих процедурах резервного копирования.

Утилиты резервного копирования применяют функции EFS API `OpenEncryptedFileRaw`, `ReadEncryptedFileRaw`, `WriteEncrypted FileRaw` и `CloseEncryptedFileRaw` в Windows для доступа к зашифрованному содержимому файла. После того как утилита открывает файл для прямого доступа во время операции резервного копирования, она вызывает `ReadEncryptedFileRaw` для получения его данных. Все API-интерфейсы утилит резервного копирования EFS работают путем выдачи FSCTL файловой системе NTFS. Например, API `ReadEncryptedFileRaw` сначала считывает поток `$EFS`, выдавая `FSCTL_ENCRYPTION_FSCTL_IO` драйверу NTFS, а затем считывает все потоки файла, включая поток `$DATA` и возможные альтернативные потоки данных. Если поток зашифрован, API `ReadEncryptedFileRaw` использует управляющий код `FSCTL_READ_RAW_ENCRYPTED` для отправки запроса зашифрованных данных потока драйверу файловой системы.

ЭКСПЕРИМЕНТ. Просмотр информации EFS

EFS имеет несколько API-функций, которые приложения могут применять для работы с зашифрованными файлами. Например, приложения используют API-функцию `AddUsersToEncryptedFile`, чтобы предоставить дополнительным пользователям доступ к зашифрованному файлу, и `RemoveUsersFromEncryptedFile`, чтобы отозвать его. Приложения задействуют функцию `QueryUsersOnEncryptedFile` для получения информации о связанных с файлом ключевых полях DDF и DRF. `QueryUsersOnEncryptedFile` возвращает SID, хеш-значение сертификата и информацию, которую содержит каждое ключевое поле DDF и DRF. Далее приведен вывод утилиты EFSDump от Sysinternals, когда в качестве аргумента командной строки указан зашифрованный файл:

```
C:\Andrea>efsdump Test.txt
EFS Information Dumper v1.02
Copyright (C) 1999 Mark Russinovich
Systems Internals – http://www.sysinternals.com

C:\Andrea\Test.txt:
DDF Entries:
    WIN-46E4EFTBP6Q\Andrea:
        Andrea(Andrea@WIN-46E4EFTBP6Q)
    Unknown user:
        Tony(Tony@WIN-46E4EFTBP6Q)
DRF Entry:
    Unknown user:
        EFS Data Recovery
```

Видно, что в файле Test.txt есть две записи DDF для пользователей Andrea и Tony и одна запись DRF для агента восстановления данных EFS, который является единственным агентом восстановления, зарегистрированным в системе в настоящее время. Можно применять инструмент шифрования для добавления или удаления пользователей в DDF-записях файла. Например, команда

```
cipher /adduser /user:Tony Test.txt
```

позволяет пользователю Tony получить доступ к зашифрованному файлу Test.txt, добавляя запись в DDF этого файла.

Копирование зашифрованных файлов

При копировании зашифрованного файла система не расшифровывает его и не зашифровывает заново в месте назначения — она просто копирует зашифрованные данные и альтернативный поток данных EFS в указанное место назначения. Однако если место назначения не поддерживает альтернативные потоки данных — если это не том NTFS (например, том FAT) или сетевой ресурс (даже если сетевой ресурс является томом NTFS), — то копирование не может проходить нормально, поскольку альтернативные потоки данных будут потеряны. Если копирование выполняется с помощью Проводника, то диалоговое окно сообщает пользователю, что конечный том не поддерживает шифрование, и спрашивает, следует ли скопировать файл

в пункт назначения в незашифрованном виде. Если пользователь соглашается, то файл будет расшифрован и скопирован в указанное место. Если копирование выполняется из командной строки, то команда копирования завершится неудачно и вернет сообщение об ошибке «Указанный файл не удалось зашифровать».

Передача шифрования в BitLocker

Драйвер файловой системы NTFS задействует службы, предоставляемые шифрующей файловой системой (Encrypting File System, EFS), для шифрования и расшифровки файлов. Эти службы режима ядра, которые взаимодействуют с пользовательской службой шифрования файлов Efssvc.dll, предоставляются NTFS через обратные вызовы. Когда пользователь или приложение шифрует файл в первый раз, служба EFS отправляет управляющий код `FSCTL_SET_ENCRYPTION` драйверу NTFS. Этот драйвер применяет обратный вызов EFS write для выполнения шифрования в памяти данных, находящихся в исходном файле. Сам процесс шифрования выполняется путем разбиения содержимого файла, которое обычно обрабатывается блоками по 2 Мбайт, на небольшие 512-байтовые фрагменты. Библиотека EFS использует API `BCryptEncrypt` для фактического шифрования фрагмента. Как уже говорилось, механизм шифрования обеспечивается драйвером CNG ядра Cng.sys, который поддерживает алгоритмы AES или 3DES, используемые EFS, а также многие другие. Поскольку EFS шифрует каждый 512-байтовый фрагмент — это наименьший физический размер стандартных секторов жесткого диска, — при каждом раунде она обновляет вектор инициализации, применяя байтовое смещение текущего блока. Вектор инициализации (initialization vector, IV), также известный как *соль*, — это 128-битное число, используемое для рандомизации схемы шифрования.

В Windows 10 производительность шифрования увеличилась благодаря *передаче шифрования* в BitLocker. Когда BitLocker включен, стек хранения уже содержит устройство, созданное драйвером шифрования всего тома, Fvevol.sys, который, если том зашифрован, выполняет шифрование и расшифровку в реальном времени на физических секторах диска, а в противном случае просто пропускает через себя запросы ввода-вывода.

Драйвер NTFS может отложить шифрование файла с помощью расширений IRP. Эти расширения предоставляются диспетчером ввода-вывода (подробнее о нем можно прочитать в главе 6) и представляют собой способ хранения различных типов дополнительной информации в IRP. Во время создания файла драйвер EFS проверяет стек устройств на наличие объекта управляющего устройства BitLocker (control device object, CDO) с помощью управляющего кода `IOCTL_FVE_GET_CDOPATH` и, если он имеется, устанавливает флаг в SCB, указывающий, что поток может поддерживать передачу шифрования.

Каждый раз при чтении или записи зашифрованного файла или при первом его шифровании драйвер NTFS, основываясь на ранее установленном флаге, определяет, нужно ли ему шифровать или расшифровывать каждый блок файла. Если включена передача шифрования, то NTFS пропускает обращение к EFS, а вместо этого добавляет расширение IRP к IRP, который будет отправлен на соответствующее устройство тома для выполнения физического ввода-вывода. В расширении IRP драйвер файловой системы NTFS хранит начальное виртуальное байтовое смещение блока файла, который драйвер хранилища собирается прочитать или записать, его

размер и некоторые флаги. Наконец, драйвер NTFS передает данные ввода-вывода на соответствующее устройство тома с помощью API `IoCallDriver`.

Диспетчер томов разбирает IRP и отправляет его нужному драйверу систем хранения. Драйвер BitLocker распознает расширение IRP и шифрует данные, которые NTFS отправила в стек устройств, используя собственные процедуры, работающие с физическими секторами. BitLocker как драйвер фильтра томов не реализует концепцию файлов и каталогов. Некоторые драйверы систем хранения, например драйвер диспетчера логических дисков VolmgrX.sys, обеспечивающий поддержку динамических дисков, являются драйверами фильтров, которые подключаются к объектам устройств тома. Эти драйверы находятся ниже диспетчера томов, но выше драйвера BitLocker и могут обеспечивать резервирование данных, чередование или виртуализацию хранилища, которые обычно реализуются путем разделения исходного IRP на несколько вторичных, которые будут передаваться на разные физические дисковые устройства. В этом случае вторичные операции ввода-вывода, перехваченные драйвером BitLocker, приведут к тому, что данные будут зашифрованы с помощью другого значения соли, которое может повредить данные файла.

Расширения IRP поддерживают концепцию распространения IRP, которая автоматически изменяет виртуальное байтовое смещение файла, хранящегося в расширении IRP, при каждом разбиении исходного IRP. Обычно драйвер EFS шифрует блоки файлов по 512-байтовым границам, и IRP не может быть разделен на выравнивание меньшее, чем размер сектора. В результате BitLocker может правильно зашифровать и расшифровать данные, гарантируя отсутствие повреждений.

Многие процедуры драйвера BitLocker не переносят ошибок памяти. Но поскольку расширение IRP динамически выделяется из невыгружаемого пула при разделении IRP, выделение может быть неудачным. Диспетчер ввода-вывода решает эту проблему с помощью процедуры `IoAllocateIrpEx`. Драйверы ядра могут использовать ее для выделения IRP, как и устаревшую `IoAllocateIrp`. Но новая процедура выделяет дополнительное место в стеке и сохраняет в нем все расширения IRP. Драйверам, которые запрашивают расширение IRP для IRP, выделенных новым API, больше не нужно выделять новую память из невыгружаемого пула.

ПРИМЕЧАНИЕ Драйвер хранилища может решить разделить IRP по разным причинам в зависимости от того, нужно ли ему отправить несколько операций ввода-вывода на несколько физических устройств. Драйвер теневого копирования тома Volsnap.sys, например, разделяет ввод-вывод, когда ему нужно прочитать файл из теневой копии тома, использующей копирование при записи, если файл находится в разных разделах: на живом томе и в файле различий теневой копии, который располагается в скрытом каталоге System Volume Information.

Поддержка шифрования в режиме онлайн

Когда файловый поток зашифровывается или расшифровывается, он монопольно блокируется драйвером файловой системы NTFS. Это означает, что никакие приложения не могут получить доступ к файлу в течение всего процесса шифрования или расшифровки. Для больших файлов такое ограничение может нарушить доступность файла на многие секунды или даже минуты. Очевидно, что это неприемлемо для серверов с большими файлами.

Чтобы решить эту проблему, в последних версиях Windows 10 появилась поддержка онлайн-шифрования. Благодаря правильной синхронизации драйвер NTFS может выполнять шифрование и расшифровку файлов без сохранения монопольного доступа к ним. EFS включает онлайн-шифрование, только если целевой поток шифрования является потоком данных, именованным или неименованным, и если он нерезидентен. В противном случае начинается стандартный процесс шифрования. Если оба условия выполнены, то служба EFS посылает управляющий код `FSCTL_SET_ENCRYPTION` драйверу NTFS для установки флага, разрешающего шифрование в режиме онлайн.

Онлайн-шифрование реализуется благодаря атрибуту `$EfsBackup` типа `$LOGGED_UTILITY_STREAM` и введению *блокировок диапазона* — новой возможности, позволяющей драйверу файловой системы блокировать в монопольном или совместном режиме только часть файла. Когда включено онлайн-шифрование, внутренняя функция `NtfsEncryptDecryptOnline` запускает процесс шифрования и расшифровки, создавая атрибут `$EfsBackup` и его SCB и получая совместную блокировку на первый двухмегабайтный диапазон файла. Совместная блокировка означает, что другие пользователи могут читать из этого диапазона файла, но чтобы записать новые данные, должны дождаться окончания операции шифрования или расшифровки.

Драйвер NTFS выделяет двухмегабайтный буфер из невыгружаемого пула и резервирует несколько кластеров из тома, которые необходимы для представления 2 Мбайт свободного пространства. Общее количество кластеров зависит от размера кластера тома. Функция онлайн-шифрования считывает исходные данные с физического диска и сохраняет их в выделенном буфере. Если передача шифрования BitLocker не включена (см. предыдущий раздел), то буфер шифруется с помощью служб EFS, в противном случае драйвер BitLocker шифрует данные при записи буфера в ранее зарезервированные кластеры.

На этом этапе NTFS блокирует весь файл на короткое время — лишь на то, которое необходимо для удаления кластеров, содержащих незашифрованные данные, из таблицы экстентов исходного потока, назначения их нерезидентному атрибуту `$EfsBackup` и замены удаленного диапазона таблицы экстентов исходного потока новыми кластерами, содержащими зашифрованные данные. Перед тем как снять монопольную блокировку, драйвер NTFS вычисляет новое значение *последней метки* (high watermark) и сохраняет его как в SCB оригинального файла в памяти, так и в полезной нагрузке EFS альтернативного потока данных `$EFS`. Затем NTFS снимает монопольную блокировку. Кластеры, содержащие исходные данные, сначала обнуляются, а затем, если больше нет блоков для обработки, освобождаются. В противном случае цикл онлайн-шифрования возобновляется со следующего фрагмента в 2 Мбайт.

Последняя метка хранит смещение файла, которое представляет собой границу между зашифрованными и незашифрованными данными. Любая одновременная запись за пределами метки может происходить в исходной форме, но другие одновременные записи до метки должны быть зашифрованы, прежде чем смогут завершиться. Запись в заблокированный диапазон не допускается. На рис. 11.72 показан пример текущей операции онлайн-шифрования для файла в 16 Мбайт. Первые два блока размером 2 Мбайт уже зашифрованы, а последняя метка установлена на 4 Мбайт, разделяя файл на зашифрованные и незашифрованные данные. На блок размером 2 Мбайт, следующий за последней меткой, устанавливается блокировка

диапазона. Приложения по-прежнему могут читать из него, но не могут записывать новые данные, для этого им нужно подождать. Данные блока шифруются и хранятся в зарезервированных кластерах. Когда устанавливается монопольное владение файлом, кластеры исходного блока переставляются в поток `$EfsBackup` путем удаления или разделения их записи в таблице экстентов исходного файла и вставки новой записи в атрибут `$EfsBackup`, а новые кластеры вставляются на место предыдущих. Значение последней метки увеличивается, блокировка файла снимается, а процесс онлайн-шифрования переходит на следующий этап, начиная со смещения 6 Мбайт. Предыдущие кластеры, расположенные в потоке `$EfsBackup`, одновременно обнуляются и могут повторно использоваться для новых этапов.

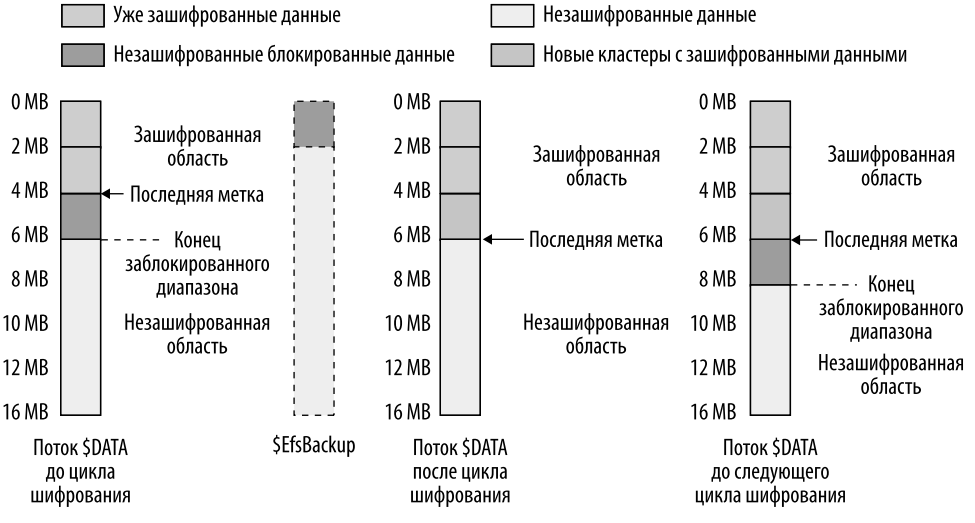


Рис. 11.72. Пример онлайн-шифрования для файла размером 16 Мбайт

Эта новая реализация позволяет NTFS шифровать или расшифровывать на месте, избавляясь от временных файлов (подробности см. ранее в разделе «Шифрование данных файла»). Что еще более важно, она позволяет NTFS выполнять шифрование и расшифровку файлов, в то время как другие приложения все еще могут применять и изменять целевой файловый поток. Время, затрачиваемое на удержание монопольной блокировки, невелико и не ощущается приложением, которое пытается использовать файл.

ДИСКИ С ПРЯМЫМ ДОСТУПОМ

Постоянная память — это развитие технологии твердотельных дисков — нового вида энергонезависимых носителей информации, которые имеют характеристики, схожие с характеристиками оперативной памяти (низкая задержка и высокая пропускная способность), размещаются на шине памяти DDR и могут использоваться как обычное дисковое устройство.

Диски прямого доступа (Direct Access Disk, DAX) — это термин, применяемый операционной системой Windows для обозначения такой технологии постоянной

памяти. Другой распространенный термин — *память класса хранения* (storage class memory, SCM). Примером нового типа памяти является *энергонезависимый модуль двойной линейной памяти* (Non-Volatile Dual In-Line Memory Module, NVDIMM) (рис. 11.73). NVDIMM — это тип памяти, который сохраняет свое содержимое даже при отключении питания. Слова Dual In-Line указывают на то, что память использует корпус DIMM. На момент написания книги существуют три типа NVDIMM: NVIDIMM-F содержит только флеш-память, наиболее распространенный NVDIMM-N является объединением флеш-памяти и традиционных чипов DRAM в одном модуле, а NVDIMM-P имеет постоянные чипы DRAM, которые не теряют данные в случае отключения питания.

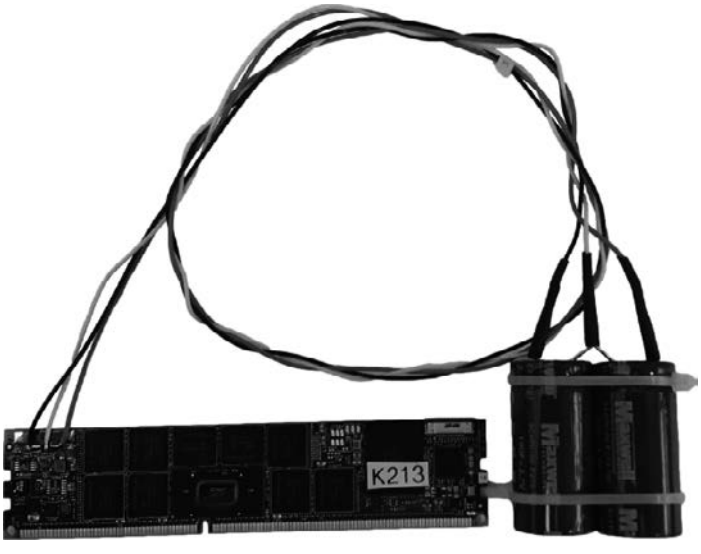


Рис. 11.73. NVDIMM, содержащий флеш-память и микросхемы DRAM. Для сохранения данных в микросхемах DRAM требуются внешняя батарея или встроенные суперконденсаторы

Одной из основных характеристик DAX, залогом его высокой производительности является поддержка *нулевого доступа* к постоянной памяти. Это означает, что многие компоненты, такие как драйвер файловой системы и диспетчер памяти, должны быть обновлены для поддержки прорывной технологии DAX.

Windows Server 2016 стала первой операционной системой Windows, поддерживающей DAX: новая модель хранения данных обеспечивает совместимость с большинством существующих приложений, которые могут работать на дисках DAX без каких-либо изменений. Для достижения максимальной производительности файлы и каталоги в томе DAX должны быть отображены в памяти с помощью API, а сам том должен быть отформатирован в специальном режиме DAX. На момент написания этой книги только NTFS поддерживает тома DAX.

В следующих разделах рассматривается принцип работы дисков прямого доступа и подробно описываются архитектура новой модели драйверов и изменения основных компонентов, отвечающих за поддержку томов DAX: драйвера NTFS, диспетчера памяти, диспетчера кэша и диспетчера ввода-вывода. Кроме того,

драйверы фильтров файловой системы, родные и сторонних производителей, включая мини-фильтры, также следует обновить, чтобы в полной мере использовать преимущества DAX.

Модель драйвера DAX

Для поддержки томов DAX Windows потребовалось внедрить совершенно новую модель драйвера хранилища. SCM Scmbus.sys — это новый драйвер шины, который перечисляет имеющиеся в системе физические и логические устройства постоянной памяти (persistent memory, PM), подключенные к ее шине памяти. Перечисление осуществляется благодаря таблице *NFIT* ACPI. Драйвер шины, который не считается частью пути ввода-вывода, является *первичным* драйвером шины, управляемым перечислителем ACPI, который предоставляется HAL (уровнем абстракции оборудования) с помощью раздела реестра базы данных оборудования HKLM\SYSTEM\CurrentControlSet\Enum\ACPI. Подробнее о перечислении устройств Plug and Play можно прочитать в главе 6.

На рис. 11.74 показана архитектура модели драйвера устройства хранения SCM. Драйвер шины SCM создает два типа объектов устройств.

- Объекты физических устройств (physical device objects, PDO) представляют физические устройства PM. Устройство NVDIMM обычно состоит из одного или нескольких чередующихся модулей NVDIMM-N. В первом случае драйвер шины SCM создает только один объект физического устройства, представляющий блок NVDIMM, во втором случае — два отдельных устройства, представляющих каждый модуль NVDIMM-N. Драйвер мини-порта Nvdimm.sys управляет всеми физическими устройствами, в том числе NVDIMM, и отвечает за контроль его состояния.
- Функциональные объекты устройств (functional device objects, FDO) представляют собой отдельные диски DAX, которыми управляет драйвер постоянной памяти Pmem.sys. Он управляет любыми побайтово адресуемыми наборами чередования и отвечает за все операции ввода-вывода, направленные в том DAX. Драйвер постоянной памяти является драйвером класса для каждого диска DAX. Он заменяет Disk.sys в классическом стеке хранения данных.

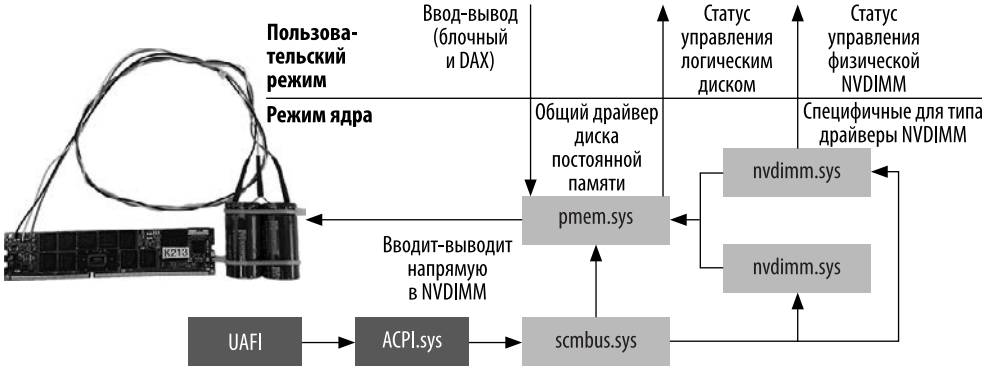


Рис. 11.74. Модель драйвера устройства хранения SCM

Драйвер шины SCM и драйвер мини-порта NVDIMM предоставляют некоторые интерфейсы для взаимодействия с драйвером класса PM. Эти интерфейсы доступны через основную функцию `IRP_MJ_PNP` с помощью запроса `IRP_MN_QUERY_INTERFACE`. Когда запрос получен, драйвер шины SCM знает, что он должен показать свой коммуникационный интерфейс, поскольку вызывающие стороны указывают GUID интерфейса {8de064ff-b63042e4-ea88-6f24c8641175}. Аналогично драйвер постоянной памяти требует интерфейс связи с устройствами NVDIMM через GUID {0079c21b-917e-405e-cea9-0732b5bbcebd}.

В новой модели драйверов устройств хранения данных реализовано четкое разделение ответственности: драйвер класса постоянной памяти управляет функциональностью логического диска (открытие, закрытие, чтение, запись, отображение памяти и т. д.), а драйверы NVDIMM — физическим устройством и его состоянием. В будущем будет легко добавить поддержку новых типов NVDIMM, просто обновив драйвер Nvdimm.sys. Pmem.sys менять не нужно.

Тома DAX

Модель драйвера устройства хранения DAX вводит новый тип томов — тома DAX. Когда пользователь впервые форматирует раздел с помощью инструмента Format, он может указать аргумент `/DAX` в командной строке. Если базовым носителем является диск DAX и он разбит на разделы по схеме GPT, то перед созданием базовой структуры данных диска, необходимой для файловой системы NTFS, инструмент заносит флаг `GPT_BASIC_DATA_ATTRIBUTE_DAX` в записи раздела GPT целевого тома, что соответствует биту 58. Хорошая справка по таблице разделов GUID имеется по адресу https://ru.wikipedia.org/wiki/Таблица_разделов_GUID.

Когда драйвер NTFS впоследствии монтирует том, он распознает флаг и отправляет управляющий код `STORAGE_QUERY_PROPERTY` базовому драйверу системы хранения. IOCTL распознается драйвером шины SCM, который отвечает драйверу файловой системы другим флагом, указывающим, что базовый диск является диском DAX. Только драйвер шины SCM может установить данный флаг. Если проверка этих двух условий оказалась успешной и поддержка DAX не отключена с помощью параметра реестра HKLM\System\CurrentControlSet\Control\FileSystem\NtfsEnableDirectAccess, NTFS включает поддержку томов DAX.

Тома DAX отличаются от стандартных томов главным образом тем, что поддерживают доступ к постоянной памяти с нулевым копированием. Файлы с отображением в памяти предоставляют приложениям прямой доступ к базовым секторам аппаратного диска через отображение, что означает: никакие промежуточные компоненты не будут перехватывать ввод-вывод. Эта характеристика обеспечивает высочайшую производительность, но, как уже говорилось, может повлиять на драйверы фильтров файловой системы, включая мини-фильтры.

Когда приложение создает раздел с отображением в памяти, основанный на файле, находящемся в томе DAX, диспетчер памяти спрашивает файловую систему, следует ли создавать раздел в режиме DAX. Так должно быть только в случае, если том тоже был отформатирован в режиме DAX. При последующем отображении файла с помощью API `MapViewOfFile` диспетчер памяти запрашивает у файловой системы диапазон физической памяти для заданного диапазона файла. Драйвер файловой системы преобразует запрошенный диапазон файла в один или несколько относительных экстентов тома (смещение и длина сектора) и просит

драйвер класса PM диска преобразовать экстенты тома в диапазоны физической памяти. Диспетчер памяти, получив диапазоны физической памяти, обновляет таблицы страниц целевого процесса, чтобы сопоставить раздел непосредственно с постоянным хранилищем. Это действительно доступ к хранилищу с нулевым копированием: приложение имеет прямой доступ к постоянной памяти. Никакого чтения или записи с подкачкой не происходит. Важно то, что диспетчер кэша в этом не участвует. Последствия будут рассмотрены далее в главе.

Приложения могут распознать тома DAX с помощью API `GetVolumeInformation`. Если возвращаемые флаги включают `FILE_DAX_VOLUME`, значит, том отформатирован в DAX-совместимой файловой системе — на момент написания этой книги только NTFS. Таким же образом приложение может определить, находится ли файл на DAX-диске, используя API `GetVolumeInformationByHandle`.

Кэшированный и некэшированный ввод-вывод в томах DAX

Несмотря на то что ввод-вывод с отображением в памяти для томов DAX обеспечивает доступ к базовому хранилищу с нулевым копированием, тома DAX по-прежнему поддерживают ввод-вывод стандартными средствами через классические API `ReadFile` и `WriteFile`. Как говорилось в начале главы, Windows предоставляет два вида обычного ввода-вывода — кэшированный и некэшированный. Они имеют существенные различия в работе с томами DAX.

Кэшированный ввод-вывод по-прежнему требует взаимодействия с диспетчером кэша, который, создавая общую карту кэша для файла, запрашивает у диспетчера памяти создание объекта секции, непосредственно сопоставляемого с аппаратным обеспечением постоянной памяти. NTFS может сообщить диспетчеру кэша, что целевой файл находится в DAX-режиме, с помощью новой процедуры `CcInitializeCacheMapEx`. После этого диспетчер кэша скопирует данные из пользовательского буфера в постоянную память. Таким образом, кэшированный ввод-вывод имеет однократный доступ к постоянному хранилищу. Обратите внимание на то, что кэшированный ввод-вывод по-прежнему согласован с другими видами ввода-вывода, отображаемыми в памяти, — диспетчер кэша использует тот же раздел. Как и в случае ввода-вывода с отображением в памяти, чтения и записи с подкачкой по-прежнему не происходит, поэтому поток системы отложенной записи и интеллектуальное упреждающее чтение не включаются.

Одним из следствий прямого отображения является то, что диспетчер кэша напрямую записывает данные на диск DAX, как только завершается выполнение функции `NtWriteFile`. Это означает, что кэшированный ввод-вывод по сути является некэшированным. По этой причине запросы ввода-вывода без кэширования напрямую преобразуются файловой системой в кэшированный ввод-вывод, при этом диспетчер кэша все так же копирует данные непосредственно между пользовательским буфером и постоянной памятью. Этот вид ввода-вывода по-прежнему согласован с кэшированным и отображаемым в памяти вводом-выводом.

NTFS продолжает задействовать стандартный ввод-вывод при обработке обновлений файлов метаданных. Режим ввода-вывода DAX для каждого файла определяется во время создания потока установкой флага в блоке управления потоком. Если файл является системным файлом метаданных, то этот атрибут никогда не устанавливается, поэтому диспетчер кэша при отображении такого

файла создает стандартную секцию с поддержкой файлов без DAX, которая будет использовать стандартный стек хранения для выполнения операций чтения и записи с подкачкой. В итоге каждый ввод-вывод обрабатывается драйвером Pmem так же, как и в случае блочных томов, с помощью алгоритма атомарности секторов. (Подробнее см. в разделе «Блоковые тома».) Такое поведение необходимо для поддержания совместимости с журналированием с упреждающей записью. Метаданные не должны сохраняться на диске до того, как соответствующий журнал будет сброшен. Таким образом, если бы файл метаданных был отображен на DAX, то это требование ведения журнала с опережением записи было бы нарушено.

Влияние на функциональность файловой системы

Отсутствие регулярного ввода-вывода с подкачкой и возможность прямого доступа приложений к постоянной памяти устраняют традиционные точки привязки, используемые файловыми системами и связанными с ними фильтрами для реализации различных функций. На томах с поддержкой DAX невозможно поддерживать множество функций, таких как шифрование файлов, сжатые и разреженные файлы, моментальные снимки и поддержка журнала USN.

В режиме DAX файловая система больше не знает, когда файл, отображенный в памяти для записи, был изменен. Когда раздел памяти создается впервые, драйвер файловой системы NTFS обновляет время изменения файла и время доступа к нему и помечает файл как измененный в журнале изменений USN. В то же время он сообщает об изменении каталога. Тома DAX больше не совместимы ни с какими устаревшими драйверами фильтров и оказывают большое влияние на мини-фильтры — клиенты диспетчеров фильтров. Такие компоненты, как BitLocker и драйвер теневого копирования тома Volsnap.sys, не работают с томами DAX и удалены из стека устройств. Поскольку мини-фильтр больше не знает, был ли файл изменен, то антивирусный сканер доступа к файлам наподобие описанного ранее не может знать, следует ли ему проверять файл на вирусы. Ему необходимо предполагать, что при любом закрытии дескриптора могло произойти изменение. Это, в свою очередь, значительно снижает производительность, поэтому мини-фильтры должны сознательно выбирать поддержку томов DAX.

Отображение исполняемых образов

Когда загрузчик Windows размещает исполняемый образ в памяти, он использует службы отображения памяти, предоставляемые диспетчером памяти. Загрузчик создает отображенную в памяти секцию образа, передавая API `NtCreateSection` флаг `SEC_IMAGE`. Этот флаг указывает загрузчику отобразить секцию как образ, применив все необходимые исправления. В режиме DAX этого делать нельзя, иначе все перемещения и исправления будут применены к исходному файлу образа на диске постоянной памяти. Для корректного решения этой проблемы диспетчер памяти применяет следующие стратегии при отображении исполняемого образа, хранящегося в томе в режиме DAX.

- Если область управления, представляющая собой секцию данных двоичного файла, уже существует (это означает, что приложение открыло образ для чтения двоичных данных), то диспетчер памяти создает пустую секцию образа

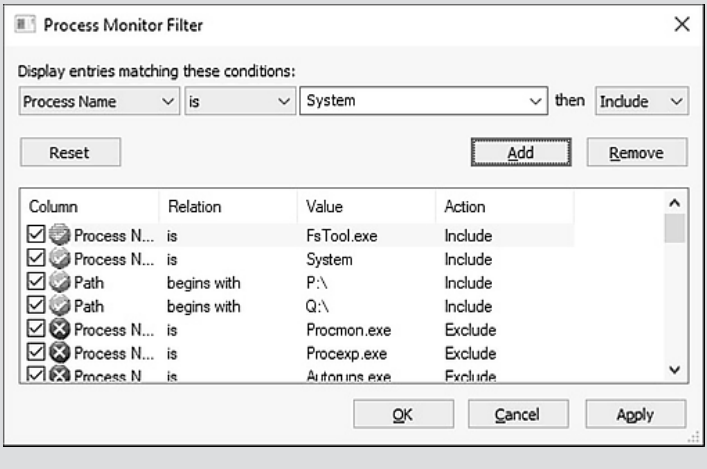
с поддержкой в памяти и копирует данные из существующей секции данных во вновь созданную секцию образа, а затем применяет необходимые исправления.

- Если для файла нет секций данных, то диспетчер памяти создает обычную секцию образа, не относящуюся к DAX, которая затем создает стандартные недопустимые прототипы PTE (подробнее см. главу 5). В этом случае диспетчер памяти использует стандартные процедуры чтения и записи драйвера Pmem для возврата данных в память при возникновении отказа страницы при некорректном доступе по адресу, принадлежащему секции с образом.

На момент написания этой книги Windows 10 не поддерживает выполнение на месте, что означает: загрузчик не может напрямую выполнить образ из хранилища DAX. Однако это не проблема, поскольку тома в режиме DAX изначально были разработаны для хранения данных с высокой производительностью. Выполнение на месте для томов DAX будет поддерживаться в будущих версиях Windows.

ЭКСПЕРИМЕНТ. Наблюдение за вводом-выводом DAX с помощью Process Monitor

Можно наблюдать за вводами-выводами DAX с помощью Process Monitor от SysInternals и приложения FsTool.exe, доступного в загружаемых ресурсах этой книги. Когда приложение читает из файла, отображенного в памяти и находящегося в томе в режиме DAX, или пишет в него, система не порождает никаких подкачек ввода-вывода, поэтому ничего не видно ни драйверу NTFS, ни минифильтрам, подключенным выше или ниже него. Чтобы убедиться в таком поведении, просто откройте Process Monitor и, предполагая, что есть два разных тома, установленных как диски P: и Q:, настройте фильтры так, как показано на следующем рисунке (диск Q: — это том в режиме DAX).

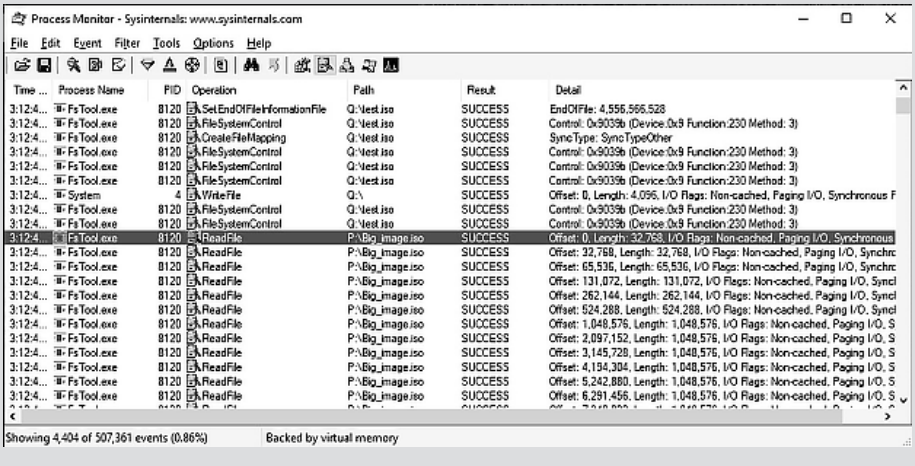


Для порождения ввода-вывода в томах в режиме DAX необходимо смоделировать копирование DAX с помощью приложения FsTool. В следующем примере

образ ISO, расположенный в томе P: блочного режима DAX — для эксперимента подойдет даже стандартный том, созданный поверх обычного диска, — копируется на диск Q: режима DAX:

```
P:\>fstool.exe /daxcopy p:\Big_image.iso q:\test.iso
NTFS / ReFS Tool v0.1
Copyright (C) 2018 Andrea Allievi (AaL186)

Starting DAX copy...
   Source file path: p:\Big_image.iso.
   Target file path: q:\test.iso.
   Source Volume: p:\ – File system: NTFS – Is DAX Volume: False.
   Target Volume: q:\ – File system: NTFS – Is DAX Volume: True.

   Source file size: 4.34 GB

Performing file copy... Success!
   Total execution time: 8 Sec.
   Copy Speed: 489.67 MB/Sec

Press any key to exit...
```

Process Monitor зафиксировал трассировку операции копирования DAX, которая подтверждает ожидаемые результаты.



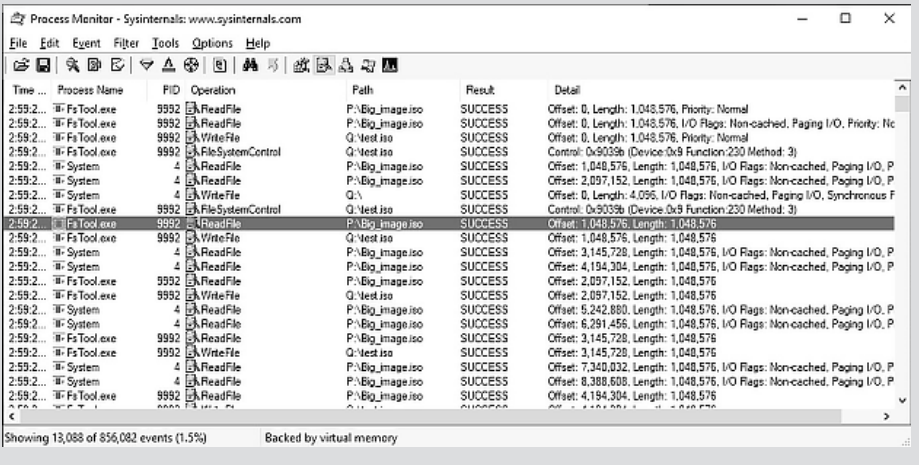
Из приведенной трассировки видно, что на целевом файле `Q:\test.iso` была перехвачена только операция `CreateFileMapping` — событий `WriteFile` не видно. Во время копирования Process Monitor обнаружил только ввод-вывод с подкачкой на исходном файле. Эти вводы-выводы с подкачкой были сгенерированы диспетчером памяти, которому нужно было считать данные обратно с исходного тома, поскольку приложение порождало страничные ошибки при обращении к файлу, отображенному в памяти.

Чтобы увидеть разницу между вводом-выводом с отображением в памяти и стандартным вводом-выводом с кэшированием, нужно снова скопировать

файл с помощью стандартной операции копирования. (Чтобы увидеть ввод-вывод с подкачкой для данных исходного файла, обязательно перезагрузите систему, в противном случае исходные данные останутся в кэше.)

```
P:\>fstool.exe /copy p:\Big_image.iso q:\test.iso
NTFS / ReFS Tool v0.1
Copyright (C) 2018 Andrea Allievi (AaL186)

Copying "Big_image.iso" to "test.iso" file... Success.
   Total File-Copy execution time: 13 Sec – Transfer Rate: 313.71 MB/s.
Press any key to exit...
```

Если сравнить трассировку, полученную Process Monitor, с предыдущей, то можно убедиться, что кэшированный ввод-вывод — это операция с одним копированием. Диспетчер кэша по-прежнему копирует фрагменты памяти между буфером, предоставляемым приложением, и системным кэшем, отображаемым непосредственно на диск DAX. Это подтверждается тем, что для целевого файла снова не выделяется ввод-вывод с подкачкой.

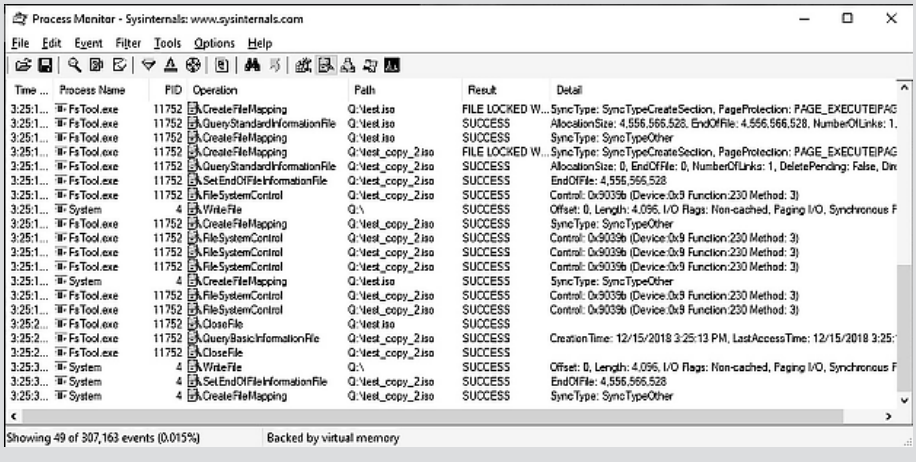


В качестве последнего эксперимента можно попробовать запустить DAX-копирование между двумя файлами, находящимися в одном томе с режимом DAX или в двух разных томах с режимом DAX:

```
P:\>fstool /daxcopy q:\test.iso q:\test_copy_2.iso
TFS / ReFS Tool v0.1
Copyright (C) 2018 Andrea Allievi (AaL186)

Starting DAX copy...
   Source file path: q:\test.iso.
   Target file path: q:\test_copy_2.iso.
   Source Volume: q:\ – File system: NTFS – Is DAX Volume: True.
   Target Volume: q:\ – File system: NTFS – Is DAX Volume: True.
Great! Both the source and the destination reside on a DAX volume.
Performing a full System Speed Copy!
   Source file size: 4.34 GB
```

```
Performing file copy... Success!
   Total execution time: 8 Sec.
   Copy Speed: 501.60 MB/Sec

Press any key to exit...
```

Трассировка, собранная в последнем эксперименте, показывает, что ввод-вывод с отображением в памяти в томах DAX не генерирует никакого ввода-вывода с подкачкой. Ни для исходного, ни для целевого файла не видно событий `WriteFile` или `ReadFile`.



Блоковые тома

Не все ограничения, накладываемые томами DAX, приемлемы в определенных сценариях. Windows обеспечивает обратную совместимость для оборудования постоянной памяти с помощью томов в блочном режиме, которыми управляет устаревший стек ввода-вывода как обычными томами, используемыми обычными и SSD-дисками. Блоковые тома сохраняют существующую семантику хранения: все операции ввода-вывода проходят через стек хранения на пути к драйверу класса диска постоянной памяти. (Драйверы мини-портов отсутствуют, поскольку они не нужны.) Они полностью совместимы со всеми существующими приложениями, устаревшими фильтрами и драйверами мини-фильтров.

Хранилища постоянной памяти способны выполнять ввод-вывод с гранулярностью в 1 байт. Точнее, ввод-вывод выполняется с гранулярностью кэш-линии, которая зависит от архитектуры, но обычно составляет 64 байта. Однако тома блочного режима извне показываются как стандартные тома, выполняющие ввод-вывод с гранулярностью сектора, 512 байт или 4 Кбайт. Если на томе DAX идет запись и на диске неожиданно происходит сбой питания, то блок данных (сектор) будет содержать смесь старых и новых данных. Приложения не готовы к такому сценарию. В блочном режиме атомарность секторов гарантируется драйвером класса диска постоянной памяти, который реализует алгоритм таблицы перевода блоков (Block Translation Table, BTT).

Алгоритм BTT, разработанный Intel, разделяет доступное дисковое пространство на фрагменты размером до 512 Гбайт, называемые *аренами*. Для каждой арены он ведет таблицу перевода блоков — простую карту перенаправления или поиска, которая сопоставляет LBA с внутренним блоком, принадлежащим арене. Для каждой 32-битной записи в карте алгоритм использует два старших бита (most significant bits, MSB) для хранения статуса блока: действительный, обнуленный или ошибка. Хотя в таблице хранится статус каждого LBA, алгоритм BTT обеспечивает атомарность сектора, предоставляя область `flog` (`"free list"` + `"log"`), содержащую массив из блоков `nfree` (`"number"` + `"free"`).

Блок `nfree` содержит все данные, необходимые алгоритму для обеспечения атомарности сектора. В массиве 256 записей `nfree`, размер одной записи составляет 32 байта, поэтому область `flog` занимает 8 Кбайт. Каждый `nfree` используется одним процессором, поэтому количество блоков `nfree` описывает число одновременных атомарных операций ввода-вывода, которые арена может обрабатывать одновременно. На рис. 11.75 показана схема диска DAX, отформатированного в блочном режиме. Структуры данных, применяемые в алгоритме BTT, не видны драйверу файловой системы. Алгоритм BTT устраняет вероятность выполнения оборванных записей в подсекторы и, как сказано ранее, необходим даже в томах, отформатированных в DAX, для поддержки записи метаданных файловой системы.

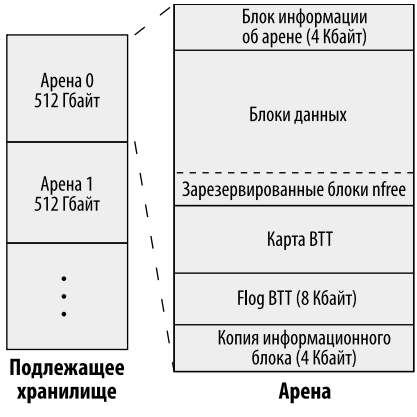


Рис. 11.75. Карта диска DAX, поддерживающего атомарность секторов (алгоритм BTT)

Тома блочного режима не имеют флага `GPT_BASIC_DATA_ATTRIBUTE_DAX` в записи раздела. NTFS ведет себя так же, как и с обычными томами, полагаясь на диспетчер кэша для выполнения кэшированного ввода-вывода и обрабатывая некэшированный ввод-вывод с помощью драйвера класса диска постоянной памяти. Драйвер Pmem предоставляет функции чтения и записи, выполняющие передачу данных с прямым доступом к памяти (direct memory access, DMA) путем построения списка дескрипторов памяти (memory descriptor list, MDL) для пользовательского буфера и адреса физического блока устройства. (MDL более подробно описаны в главе 5.) Алгоритм BTT обеспечивает атомарность секторов. На рис. 11.76 показаны стеки ввода-вывода обычного тома, тома DAX и блокового тома.

Драйверы фильтров файловой системы и DAX

Устаревшие драйверы фильтров и мини-фильтры не работают с томами DAX. Эти типы драйверов обычно дополняют функциональность файловой системы, часто взаимодействуя со всеми операциями, которыми управляет драйвер файловой системы. Существуют различные классы фильтров, предоставляющих новые возможности или изменяющих существующую функциональность драйвера файловой системы: антивирусы, шифрование, репликация, сжатие, иерархическое управление

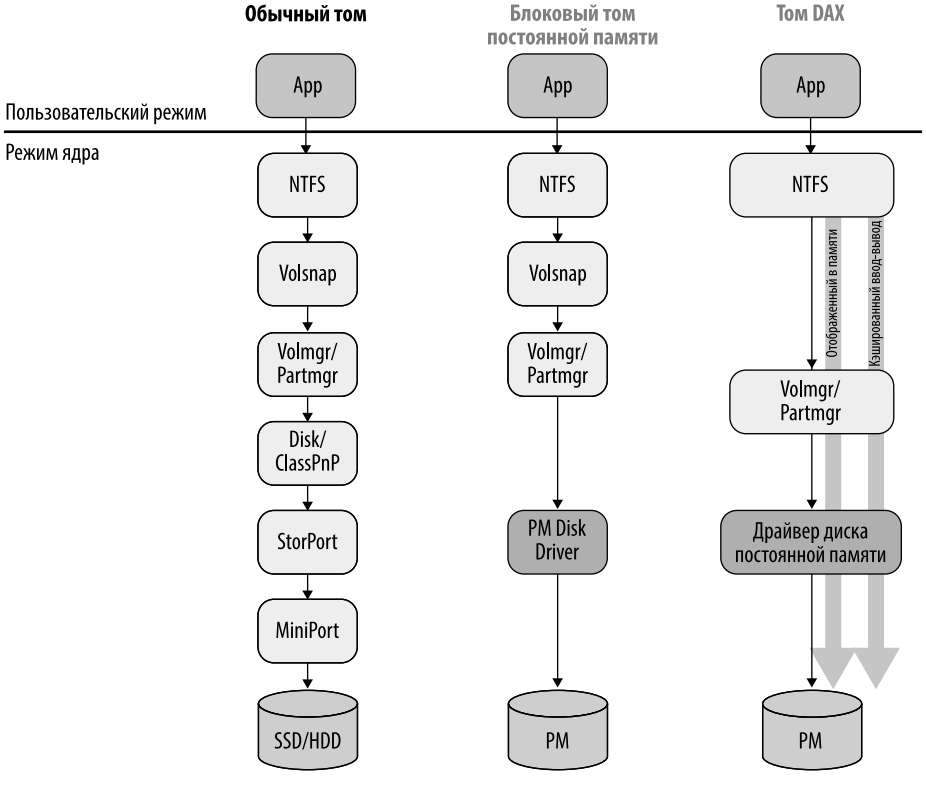


Рис. 11.76. Сравнение стеков ввода-вывода обычного тома, блокового тома и тома DAX

хранением (Hierarchical Storage Management, HSM) и т. д. Модель драйвера DAX существенно изменяет взаимодействие томов DAX с такими компонентами.

Как уже говорилось в этой главе, когда файл отображается в памяти, файловая система в режиме DAX не получает никаких запросов ввода-вывода на чтение или запись, равно как и все драйверы фильтров, расположенные выше или ниже драйвера файловой системы. Это означает, что драйверы фильтров, полагающиеся на перехват данных, работать не будут. Чтобы свести к минимуму возможные проблемы совместимости, существующие мини-фильтры не будут получать уведомление через обратный вызов `InstanceSetup`, когда монтируется том DAX. Новые и обновленные драйверы мини-фильтров, которые по-прежнему хотят работать с томами DAX, должны указать флаг `FLTFL_REGISTRATION_SUPPORT_DAX_VOLUME` при регистрации в диспетчере фильтров через `FltRegisterFilter` API ядра.

Мини-фильтры, решившие поддерживать тома DAX, имеют ограничение: они не могут перехватывать любые формы подкачки ввода-вывода. Фильтры преобразования данных, обеспечивающие шифрование или сжатие, не имеют шансов корректно работать с файлами, отображаемыми в памяти. Фильтры защиты от вредоносного программного обеспечения подвержены влиянию, как описано ранее, поскольку теперь они должны выполнять сканирование при каждом открытии

и закрытии, теряя возможность определить, действительно ли произошла запись. Влияние в основном связано с определением времени последнего обновления файла. Устаревшие фильтры больше не совместимы: если драйвер вызывает API стека `IoAttachDeviceToDevice` или аналогичные функции, то диспетчер ввода-вывода просто отклоняет запрос и регистрирует событие ETW.

Сброс ввода-вывода в режиме DAX

Обычные диски (HDD, SSD, NVme) всегда содержат кэш, который повышает их общую производительность. При выполнении операций ввода-вывода драйвером устройства хранения фактические данные сначала передаются в кэш, который впоследствии будет записан на постоянный носитель. Операционная система обеспечивает правильный сброс на диск (это гарантирует, что данные будут записаны в конечное хранилище), и упорядочивание по времени, которое обеспечивает запись данных в правильном порядке. Для обычного кэшированного ввода-вывода приложение может вызвать API `FlushFileBuffers`, чтобы гарантировать, что данные будут сохранены на диске. При этом будет создан IRP с кодом основной функции `IRP_MJ_FLUSH_BUFFERS`, который будет реализован драйвером NTFS. Некэшируемый ввод-вывод записывается NTFS непосредственно на диск, поэтому упорядочивание и сброс не проблема.

С томами в режиме DAX это уже невозможно. После отображения файла в памяти драйвер NTFS не имеет представления о данных, которые будут записаны на диск. Если приложение записывает некоторые критически важные структуры данных в том DAX, а питание отключается, то у приложения нет никаких гарантий, что все структуры данных правильно записаны на базовый носитель. Более того, у него нет гарантий, что порядок записи данных был правильным. Это происходит потому, что с точки зрения процессора хранилище постоянной памяти реализовано как классическая физическая память. Процессор использует механизм кэширования процессора, который при чтении или записи в тома DAX задействует собственные механизмы кэширования.

В результате в новых версиях Windows 10 пришлось внедрить новые API для сброса отображенных в томах DAX областей, выполняющие необходимую работу по оптимальному сбросу содержимого постоянной памяти из кэша процессора. Эти API доступны как для приложений пользовательского режима, так и для драйверов режима ядра и отличаются высокой степенью оптимизации с применением процессорной архитектуры — в стандартных системах x64 используются, например, коды операций CLFLUSH и CLWB. Приложение, которому требуются упорядочение ввода-вывода и сброс в томах DAX, может вызвать `RtlGetNonVolatileToken` для области, отображенной в PM. Эта функция возвращает энергонезависимый токен, который впоследствии может быть задействован с помощью API `RtlFlushNonVolatileMemory` или `RtlFlushNonVolatileMemoryRanges`. Обе функции выполняют реальный сброс данных из кэша процессора на базовое устройство постоянной памяти.

Операции копирования памяти, выполняемые с помощью стандартных функций операционной системы, по умолчанию являются *темпоральными* операциями копирования, то есть данные всегда проходят через кэш процессора, сохраняя порядок выполнения. *Нетемпоральные* же операции копирования используют

специализированные процессорные коды операций, чтобы обойти кэш процессора, — конкретика тоже зависит от архитектуры процессора, в x64 CPU применяется код операции MOVNTI. В этом случае упорядочивание не сохраняется, но выполнение происходит быстрее. `RtlWriteNonVolatileMemory` предоставляет операции копирования памяти в энергонезависимую память и из нее. По умолчанию API выполняет классические темпоральные операции копирования, но приложение с помощью флага `WRITE_NV_MEMORY_FLAG_NON_TEMPORAL` может запросить нетемпоральное копирование и таким образом выполнить более быструю операцию копирования.

Поддержка больших и огромных страниц

Чтение или запись файла в томе в режиме DAX через разделы, отображенные в памяти, обрабатывается диспетчером памяти аналогично разделам без DAX: если во время отображения указан флаг `MEM_LARGE_PAGES`, то диспетчер памяти обнаруживает, что один или несколько экстентов файла указывают на достаточно выровненное непрерывное физическое пространство (NTFS выделяет экстенты файлов), и использует большие (2 Мбайт) или огромные (1 Гбайт) страницы для отображения физического пространства DAX. Более подробно о диспетчере памяти и больших страницах можно прочитать в главе 5. Большие и огромные страницы имеют различные преимущества по сравнению с обычными страницами размером 4 Кбайт. В частности, они повышают производительность в процессе работы с файлами DAX, поскольку требуют меньшего количества обращений к структурам таблиц страниц процессора и записей в буфере стороннего преобразования (translation lookaside buffer, TLB). В приложениях с большим объемом памяти, которые обращаются к памяти случайным образом, процессор может тратить много времени на поиск записей в TLB, а также на чтение и запись иерархии таблицы страниц в случае пропусков TLB. Кроме того, использование больших или огромных страниц может привести к значительной экономии на фиксировании, так как обрабатываются только родительский каталог подкачки и сам каталог подкачки только для больших файлов, не для огромных. Нет особых расходов на пространство таблицы страниц — 4 Кбайт на 2 Мбайт пространства листа VA. В результате при отображении файла размером 2 Тбайт система может сэкономить 4 Гбайт фиксированной памяти за счет использования больших и огромных страниц.

Драйвер NTFS взаимодействует с диспетчером памяти, чтобы обеспечить поддержку огромных и больших страниц при отображении файлов, расположенных в томах DAX, следующим образом.

- По умолчанию каждый раздел DAX выравнивается по границам 2 Мбайт.
- NTFS поддерживает кластеры размером 2 Мбайт. Том DAX, отформатированный с применением кластеров размером 2 Мбайт, гарантированно использует только большие страницы для каждого файла, хранящегося в томе.
- Кластеры размером 1 Гбайт не поддерживаются NTFS. Если размер файла, хранящегося в томе DAX, превышает 1 Гбайт и один или несколько экстентов файла хранятся в достаточном непрерывном физическом пространстве, то диспетчер памяти отобразит файл с помощью огромных страниц. Такие страницы задействуют только два уровня отображения страниц, в то время как большие страницы — три.

Как говорилось в главе 5, для обычных секций на основе памяти диспетчер памяти использует большие и огромные страницы только в том случае, если экстент, описывающий страницы постоянной памяти, правильно выровнен в томе DAX. Выравнивание производится относительно LCN тома, а не VCN файла. Для больших страниц это означает, что экстент должен начинаться с границы 2 Мбайт, а для огромных — 1 Гбайт. Если файл в томе DAX выровнен не полностью, то диспетчер памяти использует большие или огромные страницы только для выровненных блоков, а для всех остальных — стандартные страницы размером 4 Кбайт.

Чтобы упростить использование больших страниц и увеличить его частоту, файловая система NTFS предоставляет управляющий код `FSCTL_SET_DAX_ALLOC_ALIGNMENT_HINT`, который приложение может применять для установки предпочтительного выравнивания для новых экстентов файлов. Код управления вводом-выводом принимает значение, определяющее предпочтительное выравнивание — начальное смещение, которое позволяет указать, где начинаются требования к выравниванию, и некоторые флаги. Обычно приложение отправляет IOCTL драйверу файловой системы после создания нового файла, но до его отображения. Таким образом, выделяя место для файла, NTFS захватывает свободные кластеры, которые попадают в границы предпочтительного выравнивания.

Если запрашиваемое выравнивание недоступно, например, из-за значительной фрагментации тома, то IOCTL может указать иное поведение файловой системы: отменить запрос или вернуться к запасному варианту выравнивания, который может быть указан в качестве входного параметра. IOCTL можно применить даже к уже существующему файлу для задания выравнивания новых экстентов. Приложение может запросить выравнивание всех экстентов, принадлежащих файлу, с помощью управляющего кода `FSCTL_QUERY_FILE_REGIONS` или инструмента командной строки `fsutil dax queryfilealignment`.

ЭКСПЕРИМЕНТ. Выравнивание файлов DAX

Можно наблюдать различные виды выравнивания файлов DAX с помощью приложения FsTool, доступного в загружаемых ресурсах книги. Для этого эксперимента на машине должен быть том DAX. Откройте окно командной строки и выполните копирование большого файла (рекомендуется не менее 4 Гбайт) в том DAX с помощью этого инструмента. В следующем примере два диска DAX монтируются как тома P: и Q:. Файл `Big_Image.iso` копируется в том Q: DAX с помощью стандартной операции копирования, запущенной приложением FsTool:

```
D:\>fstool.exe /copy p:\Big_DVD_Image.iso q:\test.iso
NTFS / ReFS Tool v0.1
Copyright (C) 2018 Andrea Allievi (AaLl86)

Copying "Big_DVD_Image.iso" to "test.iso" file... Success.
   Total File-Copy execution time: 10 Sec – Transfer Rate: 495.52 MB/s.
Press any key to exit...
```

Можно проверить выравнивание нового файла `test.iso` с помощью аргумента командной строки `/queryalign` приложения FsTool.exe или аргумента `queryFileAlignment` встроенного инструмента fsutil.exe, доступного в Windows:

```
D:\>fsutil dax queryFileAlignment q:\test.iso

  File Region Alignment:

    Region        Alignment      StartOffset      LengthInBytes
    0             Other          0                0x1fd000
    1             Large          0x1fd000         0x3b800000
    2             Huge           0x3b9fd000       0xc0000000
    3             Large          0xfb9fd000       0x13e00000
    4             Other          0x10f7fd000      0x17e000
```

Как видно из результатов работы инструмента, первый фрагмент файла хранится в выровненных по 4 Кбайт кластерах. Смещение, показанное инструментом, — это смещение не относительно тома, то есть LCN, а относительно файла, то есть VCN. Это важное различие, поскольку выравнивание, необходимое для отображения больших и огромных страниц, осуществляется относительно смещения страницы тома. По мере роста файла некоторые его кластеры будут выделяться со смещением тома, выровненным по 2 Мбайт или 1 Гбайт. При таком подходе эти части файла могут быть отображены диспетчером памяти с использованием больших и огромных страниц. Теперь, как и в предыдущем эксперименте, попробуем выполнить копирование DAX, указав целевое выравнивание:

```
P:\>fstool.exe /daxcopy p:\Big_DVD_Image.iso q:\test.iso /align:1GB
NTFS / ReFS Tool v0.1
Copyright (C) 2018 Andrea Allievi (AaLl86)

Starting DAX copy...
   Source file path: p:\Big_DVD_Image.iso.
   Target file path: q:\test.iso.
   Source Volume: p:\ – File system: NTFS – Is DAX Volume: True.
   Target Volume: q:\ – File system: NTFS – Is DAX Volume: False.

   Source file size: 4.34 GB
   Target file alignment (1GB) correctly set.

Performing file copy... Success!
   Total execution time: 6 Sec.
   Copy Speed: 618.81 MB/Sec

Press any key to exit...

P:\>fsutil dax queryFileAlignment q:\test.iso

  File Region Alignment:

    Region        Alignment      StartOffset        LengthInBytes
    0             Huge           0                  0x100000000
    1             Large          0x100000000        0xf800000
    2             Other          0x10f800000        0x17b000
```

В последнем случае файл сразу же выделяется на следующем кластере с выравниванием по 1 Гбайт. Первые 4 Гбайт (0x100000000 байт) содержимого файла хранятся в непрерывном пространстве. Когда диспетчер памяти отображает эту часть файла, ему нужно использовать только четыре записи таблицы указателей директора страниц (page director pointer table, PDPT), а не 2048 таблиц страниц. Это позволит сэкономить место в физической памяти и значительно повысить производительность при обращении процессора к данным, расположенным в секции DAX.

Чтобы убедиться, что копия действительно была выполнена с применением больших страниц, можно подключить к машине отладчик ядра — достаточно даже локального — и воспользоваться ключом `/debug` приложения FsTool:

```
P:\>fstool.exe /daxcopy p:\Big_DVD_Image.iso q:\test.iso /align:1GB /debug
NTFS / ReFS Tool v0.1
Copyright (C) 2018 Andrea Allievi (AaL186)

Starting DAX copy...
   Source file path: p:\Big_DVD_Image.iso.
   Target file path: q:\test.iso.
   Source Volume: p:\ – File system: NTFS – Is DAX Volume: False.
   Target Volume: q:\ – File system: NTFS – Is DAX Volume: True.

   Source file size: 4.34 GB
   Target file alignment (1GB) correctly set.

Performing file copy...
 [Debug] (PID: 10412) Source and Target file correctly mapped.
         Source file mapping address: 0x000001F1C0000000 (DAX mode: 1).
         Target file mapping address: 0x000001F2C0000000 (DAX mode: 1).
         File offset : 0x0 – Alignment: 1GB.

Press enter to start the copy...

 [Debug] (PID: 10412) File chunk's copy successfully executed.
Press enter go to the next chunk / flush the file...
```

Эффективное отображение памяти можно увидеть с помощью расширения `!pte` отладчика. Сначала необходимо перейти в контекст нужного процесса с помощью команды `.process`, а затем проанализировать виртуальный адрес, показанный FsTool:

```
8: kd> !process 0n10412 0
Searching for Process with Cid == 28ac
PROCESS ffffd28124121080
    SessionId: 2  Cid: 28ac    Peb: a29717c000  ParentCid: 31bc
    DirBase: 4cc491000  ObjectTable: ffff950f94060000  HandleCount:  49.
    Image: FsTool.exe

8: kd> .process /i ffffd28124121080
You need to continue execution (press 'g' <enter>) for the context
to be switched. When the debugger breaks in again, you will be in
the new process context.

8: kd> g
```

```
Break instruction exception - code 80000003 (first chance)
nt!DbgBreakPointWithStatus:
fffff804`3d7e8e50 cc              int     3

8: kd> !pte 0x000001F2C0000000
                                           VA 000001f2c0000000
PXE at FFFFB8DC6E371018    PPE at FFFFB8DC6E203E58    PDE at FFFFB8DC407CB000
contains 0A0000D57CEA8867  contains 8A000152400008E7  contains 0000000000000000
pfn d57cea8   ---DA--UWEV  pfn 15240000  --LDA--UW-V  LARGE PAGE pfn 15240000

PTE at FFFFB880F9600000
contains 0000000000000000
LARGE PAGE pfn 15240000
```

Команда отладчика `!pte` подтвердила, что первый 1 Гбайт пространства файла DAX отображается с помощью огромных страниц. Действительно, и каталог страниц, и таблица страниц отсутствуют. Приложение FsTool можно использовать также для установки выравнивания уже существующих файлов. Управляющий код `FSCTL_SET_DAX_ALLOC_ALIGNMENT_HINT` фактически не перемещает никаких данных, он просто предоставляет подсказку для новых выделенных экстентов файла, поскольку тот продолжит расти в будущем:

```
D:\>fstool e:\test.iso /align:2MB /offset:0
NTFS / ReFS Tool v0.1
Copyright (C) 2018 Andrea Allievi (AaLl86)

Applying file alignment to "test.iso" (Offset 0x0)... Success.
Press any key to exit...

D:\>fsutil dax queryfileAlignment e:\test.iso

  File Region Alignment:

    Region        Alignment      StartOffset         LengthInBytes
    0             Huge           0                   0x100000000
    1             Large          0x100000000         0xf800000
    2             Other          0x10f800000         0x17b000
```

Поддержка виртуальных дисков постоянной памяти и Storage Spaces

Постоянная память была специально разработана для серверных систем и критически важных приложений, таких как огромные базы данных SQL, которым требуются быстрое время отклика и обработка тысяч запросов в секунду. Часто на таких серверах приложения запускаются в виртуальных машинах, предоставляемых HyperV. Windows Server 2019 поддерживает новый тип виртуальных жестких дисков — виртуальные диски постоянной памяти. Они поддерживаются файлом VHDPMEM, который на момент написания книги можно создать или преобразовать из обычного VHD-файла только с помощью Windows PowerShell. Виртуальные диски постоянной памяти напрямую отображают фрагменты пространства, расположенные на реальном диске DAX, установленном на хосте, через файл VHDPMEM, который должен находиться в этом томе DAX.

При подключении к виртуальной машине HyperV открывает для гостя виртуальное устройство постоянной памяти (VPMEM). Оно описывается таблицей интерфейса прошивки NVDIMM (NVDIMM Firmware interface table, NFIT), расположенной в виртуальном UEFI BIOS. (Более подробную информацию о таблице NFIT можно найти в спецификации ACPI 6.2.) Драйвер шины SCM считывает таблицу и создает обычные объекты устройств, представляющие виртуальное устройство NVDIMM и диск постоянной памяти. Драйвер класса диска Pmem управляет виртуальными дисками постоянной памяти так же, как и обычными дисками постоянной памяти, и создает на них виртуальные тома. (Подробности о гипервизоре Windows и его компонентах можно найти в главе 9.) На рис. 11.77 показан стек постоянной памяти для виртуальной машины, использующей виртуальное устройство постоянной памяти. Темно-серые компоненты являются частями виртуализированного стека, а светло-серые компоненты одинаковы как в гостевом, так и в хост-разделе.

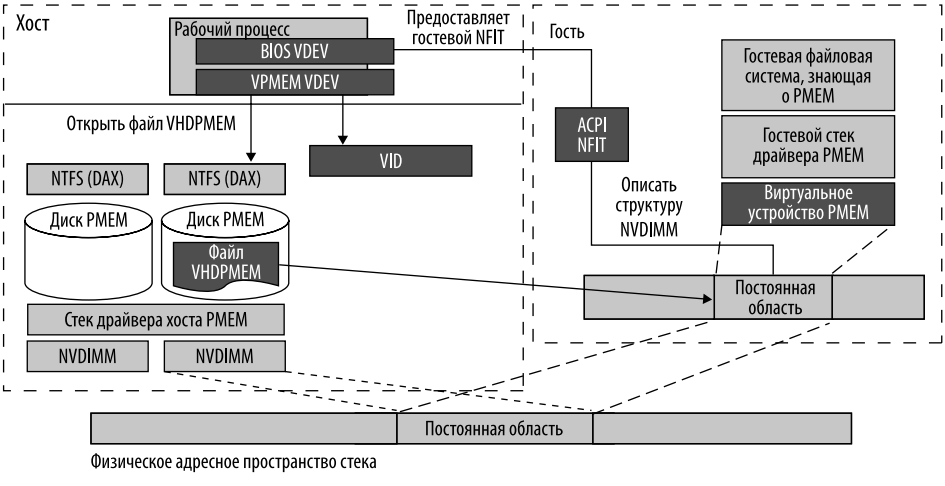


Рис. 11.77. Архитектура виртуальной постоянной памяти

Виртуальное устройство постоянной памяти предоставляет непрерывное адресное пространство, виртуализированное от хоста, то есть файлы VHDPMEM хоста не обязательно должны быть непрерывными. Оно поддерживает как DAX, так и блочный режим, между которыми, как и в случае с хостом, нужно выбрать во время форматирования тома, а также разрешает работу с большими и огромными страницами, которые используются так же, как и в хостовой системе. Только виртуальные машины второго поколения поддерживают виртуальные устройства постоянной памяти и отображение файлов VHDPMEM.

Storage Spaces Direct в Windows Server 2019 поддерживает также диски DAX в своих виртуальных пулах хранения. Один или несколько дисков DAX могут быть частью агрегированного массива дисков смешанного типа. Диски постоянной памяти в массиве могут быть настроены на обеспечение уровня хранения либо уровня производительности более крупного многоуровневого виртуального диска или на работу в качестве высокопроизводительного кэша. Более подробные сведения о Storage Spaces приводятся далее в этой главе.

ЭКСПЕРИМЕНТ. Создание и монтирование образа VHDPMEM

Как говорилось ранее, виртуальные диски постоянной памяти можно создавать, преобразовывать и назначать виртуальной машине HyperV с помощью PowerShell. Для этого эксперимента понадобятся диск DAX и виртуальная машина второго поколения с установленной Windows 10 October Update — RS5 или более поздние версии. Описание создания виртуальной машины выходит за рамки данного эксперимента. Откройте административную строку Windows PowerShell, перейдите на диск с режимом DAX и создайте виртуальный диск постоянной памяти (в этом примере диск DAX расположен в Q:):

```
PS Q:\> New-VHD VmPmemDis.vhdpmem -Fixed -SizeBytes 256GB -PhysicalSectorSizeBytes
4096

ComputerName             : 37-4611k2635
Path                     : Q:\VmPmemDis.vhdpmem
VhdFormat                : VHDX
VhdType                  : Fixed
FileSize                 : 274882101248
Size                     : 274877906944
MinimumSize              :
LogicalSectorSize        : 4096
PhysicalSectorSize       : 4096
BlockSize                : 0
ParentPath               :
DiskIdentifier           : 3AA0017F-03AF-4948-80BE-B40B4AA6BE24
FragmentationPercentage  : 0
Alignment                : 1
Attached                 : False
DiskNumber               :
IsPMEMCompatible         : True
AddressAbstractionType   : None
Number                   :
```

Виртуальные диски постоянной памяти могут быть только фиксированного размера, то есть все пространство выделено под виртуальный диск — так задумано. На втором этапе необходимо создать виртуальный контроллер постоянной памяти и подключить его к виртуальной машине. Убедитесь, что виртуальная машина выключена, и введите следующую команду (замените `TestPmVm` на конкретное имя виртуальной машины):

```
PS Q:\> Add-VMPmemController -VMName "TestPmVm"
```

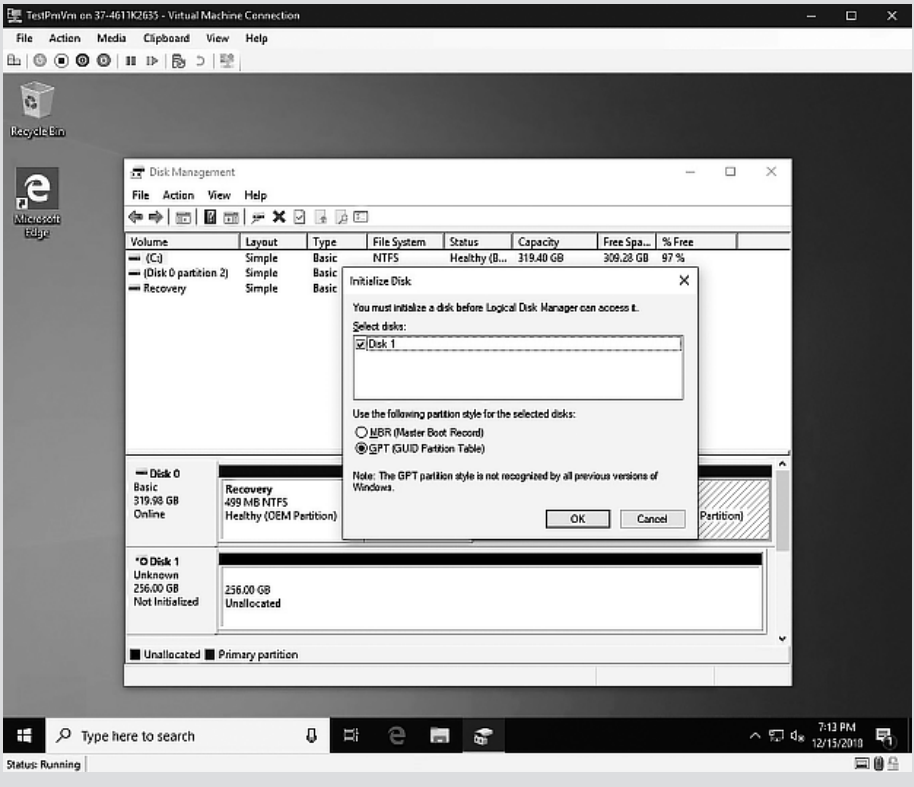
Наконец, нужно прикрепить созданный виртуальный диск постоянной памяти к контроллеру постоянной памяти виртуальной машины:

```
PS Q:\> Add-VMHardDiskDrive "TestVm" PMEM -ControllerLocation 1 -Path 'Q:\
VmPmemDis.vhdpmem'
```

Проверить результат операции можно с помощью команды `Get-VMPmemController`:

```
PS Q:\> Get-VMPmemController -VMName "TestPmVm"

VMName     ControllerNumber Drives
------     ---------------- ------
TestPmVm   0                {Persistent Memory Device on PMEM controller number 0
                             at location 1}
```

Если включить виртуальную машину, то можно увидеть, что Windows обнаружила новый виртуальный диск. На виртуальной машине откройте оснастку MMC Disk Management (diskmgmt.msc) и инициализируйте диск, используя разбиение на разделы GPT. Затем создайте простой том, назначьте ему букву диска, но не форматируйте его.



Нужно отформатировать виртуальный диск PM в режиме DAX. Откройте административное окно командной строки на виртуальной машине. Предполагая, что буква диска виртуального диска PM — E:, выполните следующую команду:

```
C:\>format e: /DAX /fs:NTFS /q
The type of the file system is RAW.
The new file system is NTFS.

WARNING, ALL DATA ON NON-REMOVABLE DISK
DRIVE E: WILL BE LOST!
Proceed with Format (Y/N)? y
QuickFormatting 256.0 GB
Volume label (32 characters, ENTER for none)? DAX-In-Vm
Creating file system structures.
Format complete.
     256.0 GB total disk space.
     255.9 GB are available.
```

Убедиться в том, что виртуальный диск был отформатирован в режиме DAX, можно с помощью встроенного инструмента fsutil.exe, указав аргументы командной строки `fsinfo volumeinfo`:

```
C:\>fsutil fsinfo volumeinfo C:
Volume Name : DAX-In-Vm
Volume Serial Number : 0x1a1bdc32
Max Component Length : 255
File System Name : NTFS
Is ReadWrite
Not Thinly-Provisioned
Supports Case-sensitive filenames
Preserves Case of filenames
Supports Unicode in filenames
Preserves & Enforces ACL's
Supports Disk Quotas
Supports Reparse Points
Returns Handle Close Result Information
Supports POSIX-style Unlink and Rename
Supports Object Identifiers
Supports Named Streams
Supports Hard Links
Supports Extended Attributes
Supports Open By FileID
Supports USN Journal
Is DAX Volume
```

УСТОЙЧИВАЯ ФАЙЛОВАЯ СИСТЕМА

В Windows Server 2012 R2 появилась новая усовершенствованная файловая система — Resilient File System (ReFS). Она является частью новой архитектуры хранения под названием Storage Spaces, которая, помимо прочего, позволяет создавать многоуровневые виртуальные тома, состоящие из твердотельного накопителя и обычного диска. (Введение в Storage Spaces и многоуровневое хранение представлено далее в этой главе.) ReFS — это файловая система типа «запись в новое место», что означает: метаданные файловой системы *никогда не обновляются* на своем месте, обновленные метаданные записываются в новое место, а старые помечаются как удаленные. Это свойство очень важно и является одной из функций, обеспечивающих целостность данных. Первоначальные цели ReFS были следующими.

1. Самовосстановление, проверка и восстановление томов в режиме онлайн, что обеспечивает практически нулевую недоступность из-за повреждения файловой системы, а также поддержка сквозной записи (о ней речь пойдет далее в этом разделе).
2. Целостность всех пользовательских данных, аппаратных и программных.
3. Эффективные и быстрые снимки файлов (клонирование блоков).
4. Поддержка очень больших томов (эксабайтных размеров) и файлов.
5. Автоматическая многоуровневая система хранения данных и метаданных, поддержка магнитной записи внахлест (shingled magnetic recording, SMR) и будущие твердотельные диски.

Существуют различные версии ReFS. Описанная в этой книге называется ReFS v2 и впервые была реализована в Windows Server 2016. На рис. 11.78 показан обзор различий высокоуровневых реализаций NTFS и ReFS. Вместо того чтобы полностью переписывать файловую систему NTFS, ReFS использует другой подход, разделяя реализацию NTFS на две части: одна часть понимает формат на диске, а другая — нет.

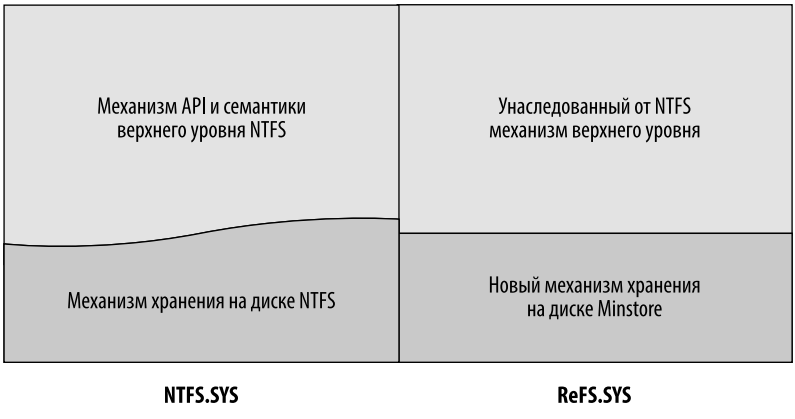


Рис. 11.78. Высокоуровневая реализация ReFS в сравнении с NTFS

ReFS заменяет механизм хранения данных на диске на *Minstore*. Minstore — это библиотека восстанавливаемого хранилища объектов, которая предоставляет вызывающим ее пользователям интерфейс таблиц пар ключей и значений, реализует семантику выделения при записи для изменения этих таблиц и интегрируется с диспетчером кэша Windows. По сути, это библиотека, реализующая ядро современной масштабируемой файловой системы с функцией копирования и записи. Minstore применяется в ReFS для реализации файлов, каталогов и т. д. Понимание основ Minstore необходимо для рассмотрения ReFS, поэтому начнем с описания этой библиотеки.

Архитектура Minstore

Все в Minstore является таблицей. Она состоит из множества строк — пар «ключ — значение». Таблицы Minstore при хранении на диске представляются с помощью B⁺-деревьев. При хранении в энергозависимой памяти RAM они представляются с помощью хеш-таблиц. B⁺-деревья, также известные как сбалансированные деревья, обладают следующими важными свойствами.

1. Обычно у них большое количество детей на один узел.
2. Они хранят указатели данных — указатели на блок дискового файла, содержащие значение ключа, — только на листьях, но не на внутренних узлах.
3. Все пути от корневых узлов к листовым одинаковой длины.

Другие файловые системы, например NTFS, обычно используют B-деревья (другая структура данных, обобщающая двоичное дерево поиска, не путать с двоичным деревом) для хранения указателя данных вместе с ключом в каждом узле дерева. Эта техника значительно уменьшает количество записей, которые могут быть упакованы в узел B-дерева, тем самым способствуя увеличению количества уровней в B-дереве, а значит, увеличивая время поиска записи.

На рис. 11.79 приведен пример B^+-дерева. В нем корень и внутренний узел содержат только ключи, которые используются для правильного доступа к данным, расположенным в узлах листьев. Узлы листьев находятся на одном уровне и, как правило, связаны между собой. Поэтому нет необходимости выполнять множество операций ввода-вывода для поиска элемента в дереве.

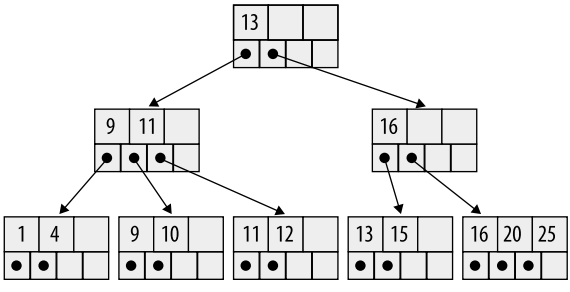


Рис. 11.79. Пример B^+-дерева. Только узлы листьев содержат указатели данных. Узлы маршрутизации содержат только ссылки на дочерние узлы

Предположим, что Minstore нужно получить доступ к узлу с ключом 20. Корневой узел содержит один ключ, используемый в качестве индекса. Ключи со значением, бóльшим или равным 13, хранятся в одном из дочерних узлов, индексируемых правым указателем, а ключи со значением меньше 13 — в одном из левых дочерних узлов. Когда Minstore достигает листа, содержащего фактические данные, он может легко получить доступ к данным и для узла с ключами 16 и 25, не выполняя полного сканирования дерева.

Кроме того, узлы листьев обычно связаны между собой с помощью связных списков. Это означает, что в огромных деревьях Minstore может, например, исследовать все файлы в папке, обратившись к корню и промежуточным узлам только один раз, при условии, что на рисунке все файлы представлены значениями, хранящимися в листьях. Как упоминалось ранее, Minstore обычно использует B^+-дерево для представления объектов, отличных от файлов или каталогов.

В этой книге применяются термины «B^+-дерево» и «B^+-таблица» для обозначения одного и того же понятия. В Minstore определены различные типы таблиц. Таблица может быть создана, строки могут быть добавлены в нее, удалены из нее или обновлены внутри нее. Внешний объект может перечислить таблицу или найти одну строку. Ядро Minstore представлено таблицей объектов. Таблица объектов — это указатель расположения всех корневых, не вложенных B^+-деревьев в томе.

B^+-деревья могут быть встроены в другие деревья, при этом корень дочернего дерева хранится в строке родительского дерева.

Каждая таблица в Minstore определяется композитом и схемой. Композит — это просто набор правил, которые описывают поведение корневого узла (иногда и дочерних) и то, как находить каждый узел B^+-таблицы и управлять им. Minstore поддерживает следующие два вида корневых узлов, управляемых соответствующими композитами.

- **Копируемый при записи (Copy on Write, CoW).** Корневой узел этого типа перемещается при изменении дерева. Это означает, что в случае изменения записывается совершенно новое B^+-дерево, а старое помечается на удаление. Чтобы работать с такими узлами, соответствующий композит должен хранить идентификатор объекта, который будет использоваться при записи таблицы.
- **Встроенный.** Этот тип корневого узла хранится в данных — в значении узла листа — индексной записи другого B^+-дерева. Встроенный композит поддерживает ссылку на индексную запись, в которой хранится встроенный корневой узел.

Указание *схемы* при создании таблицы сообщает Minstore, какой тип ключа используется, какого размера должны быть корневой и листовой узлы таблицы и как располагаются в ней строки. В ReFS применяются разные схемы для файлов и каталогов. Каталоги — это объекты B^+-таблицы, на которые ссылается таблица объектов, которая может содержать три вида строк: файлы, ссылки и идентификаторы файлов. В ReFS ключ каждой строки представляет собой имя файла, ссылки или идентификатора файла. Файлы — это таблицы, содержащие в своих строках атрибуты, то есть пары кодов и значений атрибутов.

Каждая операция, которая может быть выполнена над таблицей: закрытие, изменение, запись на диск или удаление, — представлена транзакцией Minstore. Она похожа на транзакцию базы данных — это единица работы, иногда состоящая из нескольких операций, которые могут завершиться или не завершиться только атомарным способом. Таблицы записываются на диск в ходе процесса, известного как *обновление дерева*. Когда запрашивается обновление дерева, транзакции из него удаляются и ни одна транзакция не может быть запущена до тех пор, пока обновление не завершится.

Одна из важных концепций, используемых в ReFS, — это встроенная таблица, B^+-дерево, корневой узел которого расположен в строке другого B^+-дерева. В ReFS встроенные таблицы применяются широко. Например, каждый файл — это B^+-дерево, корни которого находятся в строке каталогов. Встроенные таблицы поддерживают операцию перемещения, которая изменяет родительскую таблицу. Размер корневого узла фиксирован и берется из схемы таблицы.

Физическая схема B^+-дерева

В Minstore B^+-дерево состоит из *корзин*. Так называется эквивалент обобщенных узлов B^+-дерева в Minstore. Листовые корзины содержат данные, которые хранит дерево, промежуточные корзины называются узлами маршрутизации и используются только для прямого поиска на следующем уровне дерева. На рис. 11.79 каждый узел — это корзина. Поскольку узлы маршрутизации применяются только

для направления трафика к дочерним корзинам, им не обязательно иметь точные копии ключа в дочерней корзине — они могут выбрать значение между двумя корзинами и работать с ним. В ReFS ключом обычно является сжатое имя файла. Данные промежуточной корзины вместо этого содержат логический номер кластера (LCN) и контрольную сумму корзины, на которую указывают. Контрольная сумма позволяет ReFS реализовать функции самовосстановления. Промежуточные узлы таблицы Minstore можно рассматривать как дерево Меркла, в котором каждый листовой узел помечен хешем блока данных, а каждый нелистовой узел — криптографическим хешем меток его дочерних узлов.

Каждая корзина состоит из описывающего ее *индексного заголовка* и подвала, представляющего собой массив смещений, указывающих на индексные записи в правильном порядке. Индексные записи представляют собой строку в B⁺-таблице и располагаются между заголовком и подвалом. Строка — это простая структура данных, в которой указаны местоположение и размер ключа и данных, находящихся в одной и той же корзине. На рис. 11.80 приведен пример листовой корзины, содержащей три строки, проиндексированные по смещениям, расположенным в подвале. В листовых страницах каждая строка содержит ключ и фактические данные или корневой узел другого встроенного дерева.

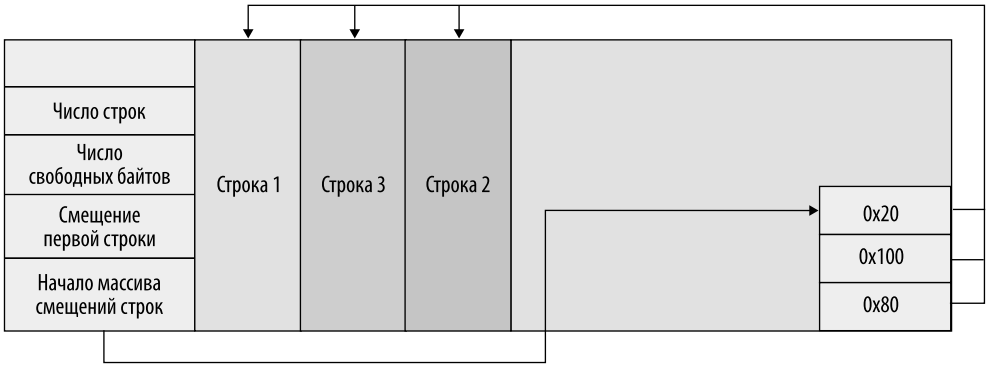


Рис. 11.80. Листовая корзина с тремя индексными записями, упорядоченными по массиву смещений в подвале

Распределители

Когда файловая система просит Minstore выделить корзину (B⁺-таблица запрашивает ее с помощью процесса, называемого *прикреплением корзины*), ему нужен способ отслеживать свободное пространство на устройстве хранения. В первой версии Minstore применялся иерархический распределитель, что означало наличие нескольких объектов-распределителей, каждый из которых выделял пространство из своего родительского распределителя. Когда корневой распределитель отображал все пространство тома, каждый распределитель становился B⁺-деревом, которое использовало схему таблицы *lcn-count*. Эта схема описывает ключ строки как диапазон LCN, который распределитель взял из своего родительского узла, а значение

строки — как *область распределителя*. В первоначальной реализации область распределителя описывала состояние каждого своего фрагмента по отношению к его дочерним узлам: свободен он или выделен, — а также идентификатор владельца объекта, которому он принадлежит.

На рис. 11.81 показана упрощенная версия оригинальной реализации иерархического распределителя. Здесь у большого распределителя установлена только одна единица распределения — место, представленное битом, было выделено для среднего распределителя, который в данный момент пуст. В этом случае средний распределитель является дочерним для большого.

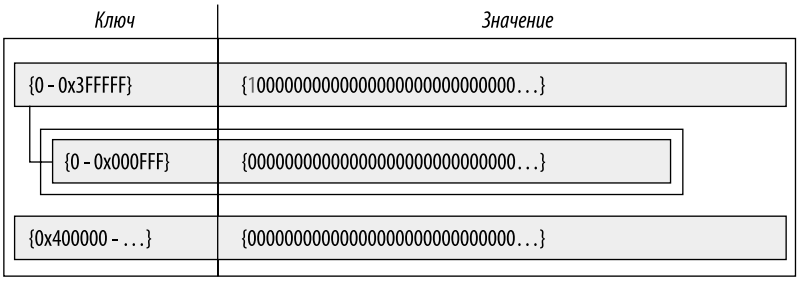


Рис. 11.81. Старый иерархический распределитель

B⁺-таблицы сильно зависят от распределителей при получении новых корзин и поиске места для копий существующих корзин посредством копирования во время записи. В последней версии Minstore иерархический распределитель был заменен на распределитель, управляемый политиками, с целью поддержки центрального расположения в файловой системе, которое могло бы поддерживать многоуровневое размещение. Уровень — это разновидность устройства хранения, например, SSD, NVMe или обычный жесткий диск. (Многоуровневая структура рассматривается далее в этой главе.) По сути, это возможность поддерживать диск, состоящий из быстрой зоны произвольного доступа, которая обычно меньше медленной зоны только последовательного доступа.

Новый распределитель, управляемый политикой, представляет собой оптимизированную версию, которая поддерживает очень большое количество распределений в секунду и определяет различные области распределения в зависимости от запрашиваемого уровня, то есть типа базового устройства хранения. Когда файловая система запрашивает место для новых данных, центральный распределитель решает с помощью управляемого политикой механизма, из какой области его выделить. Этот механизм политики учитывает уровни, что означает: метаданные всегда записываются на уровни производительности и никогда — на уровни хранения SMR, поскольку метаданные записываются в случайном порядке. Он также поддерживает диапазоны ReFS и реализует логику отложенного распределения (deferred allocation logic, DAL). Логика отложенного распределения основывается на том, что когда файловая система создает файл, она обычно также выделяет необходимое пространство для его содержимого. Minstore вместо того, чтобы возвращать

базовой файловой системе диапазон LCN, возвращает маркер, содержащий резервирование пространства, которое дает гарантию от заполнения диска. Когда файл в конечном счете записывается, распределитель назначает LCN для содержимого файла и обновляет метаданные. Это решает проблемы с дисками SMR, о которых речь пойдет далее в этой главе, и позволяет ReFS создавать даже огромные файлы, 64 Тбайт и более, менее чем за 1 с.

Распределитель, управляемый политикой, состоит из трех центральных распределителей, реализованных на диске в виде глобальных B^+-таблиц. Однако при загрузке в память распределители представляются в виде AVL-деревьев. AVL-дерево — это еще один вид самобалансирующегося двоичного дерева, который не рассматривается в этой книге. Хотя каждая строка в B^+-таблице по-прежнему индексируется по диапазону, часть данных в строке может содержать битовую карту или в качестве оптимизации — только количество выделенных кластеров в случае, если выделенное пространство является непрерывным. Три распределителя используются для следующих целей.

- Средний распределитель (MAA) является распределителем для каждого файла в пространстве имен, за исключением некоторых B^+-таблиц, выделенных из других распределителей. Средний распределитель сам является B^+-таблицей, поэтому ему необходимо найти место для обновления своих метаданных, которые все еще придерживаются стратегии записи в новое место. Это роль малого распределителя.
- Малый распределитель (SAA) выделяет место для себя, для среднего распределителя и для двух таблиц: таблицы состояния целостности, позволяющей ReFS поддерживать потоки целостности, и таблицы счетчика ссылок на блоки, позволяющей ReFS поддерживать клонирование блоков файла.
- Распределитель контейнеров (CAA) применяется для выделения места для таблицы контейнеров — фундаментальной таблицы, обеспечивающей кластерную виртуализацию в ReFS, а также для уплотнения контейнеров. Кроме того, распределитель контейнеров содержит одну или несколько записей для описания пространства, используемого им самим.

Когда Format формирует базовые структуры данных для ReFS, он создает три распределителя. Средний изначально описывает все кластеры тома. Место для метаданных SAA и CAA, которые являются B^+-таблицами, выделяется из MAA — это единственный случай за все время существования тома. В SAA вставляется запись для описания пространства, используемого средним распределителем. После создания распределителей дополнительные записи для SAA и CAA больше не выделяются из среднего распределителя, за исключением случаев, когда ReFS обнаруживает повреждения в самих распределителях.

Чтобы выполнить операцию записи в новое место для файла, ReFS сначала должна обратиться к распределителю MAA, чтобы найти место для записи. В многоуровневой конфигурации она делает это с учетом уровней. После успешного завершения операции она обновляет таблицу экстентов потока файла, чтобы отразить новое местоположение этого экстента, и обновляет метаданные файла. Затем новое

B^+-дерево записывается на диск в блок свободного пространства, а старая таблица преобразуется в свободное пространство. Если запись помечена как сквозная, то есть должна быть обнаруживаемой после сбоя, то ReFS делает запись в журнал для регистрации операции записи в новое место. (Более подробную информацию см. в разделе «Сквозная запись в ReFS» далее в этой главе.)

Таблица страниц

Когда Minstore обновляет корзину в B^+-дереве, например, чтобы переместить дочерний узел или даже добавить строку в таблицу, ему обычно нужно обновить родительские или маршрутизирующие узлы. Точнее, Minstore использует различные ссылки, которые указывают на новую и старую дочернюю корзину для каждого узла. Это происходит потому, что, как говорилось ранее, каждый узел маршрутизации содержит контрольную сумму своих листьев. Кроме того, листовой узел мог быть перемещен или даже удален. Это вызывает проблемы с синхронизацией. Например, представьте поток, который читает B^+-дерево, в то время как строка удаляется. Блокировка дерева и запись каждого изменения на физический носитель будут стоить непомерно дорого. Minstore нужен удобный и быстрый способ отслеживать информацию о дереве. *Таблица страниц Minstore*, не имеющая отношения к таблице страниц процессора, — это хеш-таблица в памяти, персональная для корневой таблицы каждого Minstore (обычно это таблица каталогов и файлов), которая отслеживает, какая корзина «загрязнена», освобождена или удалена. Эта таблица никогда не хранится на диске. В Minstore термины *«корзина»* и *«страница»* взаимозаменяемые: страница обычно находится в памяти, а корзина хранится на диске, но они выражают одну и ту же концепцию высокого уровня. Понятия *«деревья»* и *«таблицы»* также используются как взаимозаменяемые, что объясняет, почему таблица страниц называется именно так. Строки таблицы страниц состоят из LCN целевой корзины в качестве ключа и структуры данных, которая отслеживает состояние страниц и помогает синхронизации B^+-дерева в качестве значения.

При первом чтении или создании страницы в хеш-таблицу, представляющую собой таблицу страниц, вставляется новая запись. Она может быть удалена только при выполнении всех следующих условий.

- Нет активных транзакций, обращающихся к странице.
- Страница чистая и не имеет изменений.
- Страница не является копией, сделанной во время записи предыдущей страницы.

Благодаря этим правилам чистые страницы обычно постоянно попадают в таблицу страниц и удаляются из нее, в то время как «грязная» страница остается там до тех пор, пока B^+-дерево не будет обновлено и записано на диск. Процесс записи дерева на стабильный носитель в значительной степени зависит от состояния таблицы страниц в каждый момент времени. Как видно из рис. 11.82, таблица страниц используется Minstore в качестве кэша в памяти, создавая неявную машину состояний, которая описывает каждое состояние страницы.

Рис. 11.82. На диаграмме показаны состояния «грязной» страницы (корзины) в таблице страниц. Новая страница создается в результате копирования во время записи старой страницы или когда B⁺-дерево растет и требует больше места для хранения корзины

Ввод-вывод Minstore

В Minstore чтение из B⁺-дерева и запись в него на конечном физическом носителе выполняются иначе: чтение дерева обычно реализуется частями, то есть операция чтения может включать, например, лишь некоторые листья и происходит как часть транзакционного доступа или как действие упреждающей предварительной выборки. После того как корзина прочитана в кэш, Minstore все еще не может интерпретировать ее данные, поскольку необходимо проверить контрольную сумму корзины. Ожидаемая контрольная сумма хранится в родительском узле: когда драйвер ReFS, который находится над Minstore, перехватывает прочитанные данные, он знает, что узел еще нуждается в проверке. Родительский узел уже расположен в кэше (дерево уже пройдено для достижения дочернего) и содержит контрольную сумму дочернего. У Minstore есть вся необходимая информация для проверки того, что корзина содержит достоверные данные. Обратите внимание: в таблице страниц могут быть страницы, к которым никогда не обращались. Это происходит потому, что их контрольная сумма все еще нуждается в проверке.

Minstore обновляет дерево, записывая все B⁺-дерево в виде одной транзакции. Процесс обновления дерева записывает «загрязненные» страницы B⁺-дерева на физический диск. Обновление дерева может происходить по разным причинам: приложение явно сбрасывает свои изменения, система работает в условиях нехватки памяти или других подобных, диспетчер кэша сбрасывает кэшированные

данные на диск и т. д. Стоит отметить, что Minstore обычно записывает новые обновленные деревья отложенно с помощью потока системы отложенной записи. Как было показано в предыдущем разделе, существует несколько триггеров для запуска систем отложенной записи, например, когда количество «грязных» страниц достигает определенного порога.

Minstore не знает о фактической причине запроса на обновление дерева. Первое, что он делает, — убеждается, что никакие другие транзакции не изменяют дерево, используя сложные примитивы синхронизации. После начальной синхронизации Minstore начинает записывать «грязные» страницы и старые удаленные страницы. В реализации записи в новое место новая страница представляет собой корзину, которая была изменена и содержимое которой заменено, а освобожденная страница — это старая страница, которую нужно отвязать от родительской. Если транзакция хочет изменить листовой узел, она копирует в память корневую корзину и листовую страницу. Затем Minstore создает соответствующие записи в таблице страниц, не изменяя ни одной ссылки.

Алгоритм обновления дерева перечисляет каждую страницу в таблице страниц. Однако эта таблица не имеет представления о том, на каком уровне B^+-дерева находится страница, поэтому алгоритм проверяет все B^+-дерево, начиная с внешнего узла, обычно листа, и заканчивая корневыми узлами. Для каждой страницы алгоритм выполняет следующие шаги.

1. Проверяет состояние страницы. Если это свободная страница, он пропускает ее. Если это «грязная» страница, обновляет указатель на родителя и контрольную сумму и помещает страницу во внутренний список страниц для записи.
2. Удаляет старую страницу.

Когда алгоритм достигает корневого узла, Minstore обновляет его родительский указатель и контрольную сумму непосредственно в таблице объектов и, наконец, помещает корневой узел в список страниц для записи. Теперь Minstore может записать новое дерево в свободное пространство базового тома, сохранив старое дерево на прежнем месте. Старое дерево только помечается как освобожденное, но все еще присутствует на физическом носителе. Это важная характеристика, которая обобщает стратегию записи в новое пространство и позволяет файловой системе ReFS, находящейся выше Minstore, поддерживать расширенные функции восстановления в режиме онлайн. На рис. 11.83 приведен пример процесса обновления дерева для B^+-таблицы, содержащей две новые листовые страницы, A’ и B’. Здесь страницы, расположенные в таблице страниц, более светлые, а старые страницы — более темные.

Сохранение монопольного доступа к дереву при его обновлении может быть проблемой для производительности, потому что никто другой не может читать из монопольно заблокированного B^+-дерева или записывать в него. В последних версиях Windows 10 B^+-деревья в Minstore стали *поколенческими* — к каждому привязан номер поколения. Это означает, что страница в дереве может быть «грязной» по отношению к определенному поколению. Если страница изначально «грязная» только для определенного поколения дерева, то ее можно обновить напрямую, не копируя при записи, поскольку окончательное дерево еще не записано на диск.

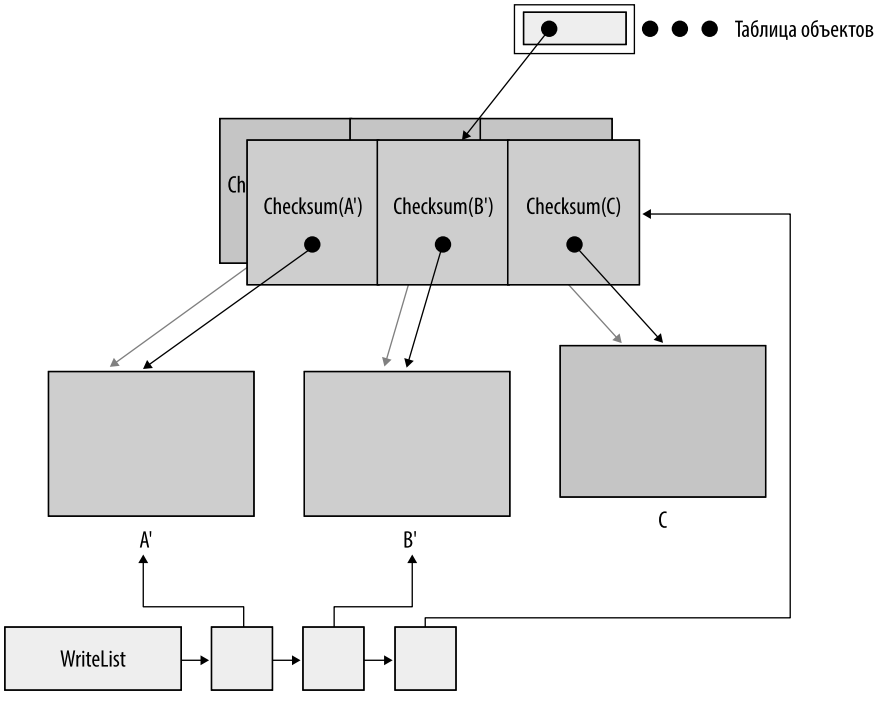


Рис. 11.83. Процесс обновления дерева Minstore

В новой модели процесс обновления дерева обычно делится на следующие этапы.

- **Этап с допустимым сбоем.** Minstore получает монопольную блокировку дерева, увеличивает номер его поколения, вычисляет и выделяет память, необходимую для обновления дерева, и, наконец, меняет блокировку на совместную.
- **Этап без сбоев.** Эта фаза выполняется с совместной блокировкой, то есть другие операции ввода-вывода могут читать из дерева. Minstore обновляет связи узлов маршрутизации и все контрольные суммы дерева и записывает окончательное дерево на базовый диск. Если другая транзакция захочет изменить дерево во время его записи на диск, то обнаружит, что номер поколения дерева выше, и снова скопирует дерево во время записи.

Согласно новой схеме Minstore удерживает монопольную блокировку только на этапе возможного отказа. Это означает, что обновление дерева может выполняться параллельно с другими транзакциями Minstore, что значительно повышает общую производительность.

Архитектура ReFS

Как говорилось в предыдущих разделах, ReFS (Resilient file system) — это гибрид реализации NTFS и Minstore, где каждый файл и каталог представляет собой B^+-дерево, сконфигурированное по определенной схеме. Том файловой системы

представляет собой плоское пространство имен каталогов. Как уже известно, NTFS состоит из следующих компонентов.

- **Поддержка основной файловой системы.** Описывает интерфейс между файловой системой и другими компонентами системы, такими как диспетчер кэша и подсистема ввода-вывода, и раскрывает понятия создания, открытия, чтения, записи, закрытия файлов и т. д.
- **Поддержка высокоуровневых функций файловой системы.** Описывает высокоуровневые функции современной файловой системы, например сжатие файлов, ссылки на файлы, отслеживание квот, точки повторной обработки, шифрование файлов, поддержка восстановления и т. д.
- **Зависимые от диска компоненты и структуры данных.** MFT и файловые записи, кластеры, пакет индексов, резидентные и нерезидентные атрибуты и т. д. (подробнее см. раздел «Файловая система NT» в начале этой главы).

ReFS сохраняет первые две части практически без изменений и заменяет остальные компоненты, зависящие от диска, на Minstore (рис. 11.84).

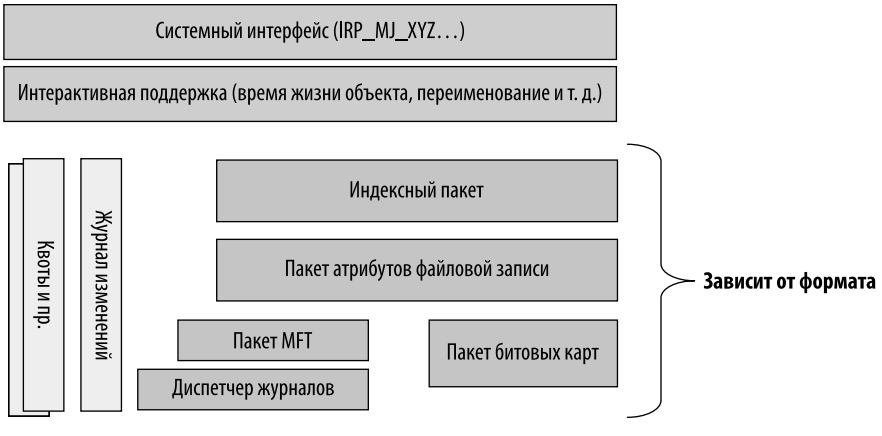


Рис. 11.84. Схема архитектуры ReFS

В разделе «Драйвер файловой системы NTFS» этой главы были представлены структуры, связывающие файловый дескриптор со структурой файловой системы на диске. В драйвере файловой системы ReFS эти структуры данных (блок управления потоком, представляющий атрибут NTFS, который пытается прочитать вызывающая сторона, и блок управления файлом, содержащий указатель на запись файла в MFT диска) по-прежнему актуальны, но имеют несколько иное значение в смысле их хранения на носителе. Изменения, вносимые в эти объекты, проходят через Minstore вместо того, чтобы напрямую транслироваться в изменения MFT на диске. В ReFS происходит следующее (рис. 11.85).

- Блок управления файлами (file control block, FCB) представляет один файл или каталог и содержит указатель на B^+-дерево Minstore, ссылку на блок управления потоком родительского каталога и ключ, имя каталога. На FCB указывает файловый объект посредством поля `FsContext2`.

- Блок управления потоком (stream control block, SCB) представляет собой открытый поток файлового объекта. Структура данных, используемая в ReFS, — это упрощенная версия структуры NTFS. Но когда SCB представляет каталоги, SCB содержит ссылку на индекс каталога, который находится в B⁺-дереве, соответствующем каталогу. На SCB указывает файловый объект с помощью поля `FsContext`.
- Блок управления томом (volume control block, VCB) представляет собой смонтированный в данный момент том, отформатированный ReFS. Когда драйвер ReFS опознает правильно отформатированный том, структура данных VCB создается, встраивается в расширение объекта устройства тома и связывается со списком, расположенным в глобальной структуре данных, которую драйвер файловой системы ReFS выделяет во время инициализации. VCB содержит таблицу всех FCB каталогов, в данный момент открытых в томе, с индексацией по идентификатору ссылки.

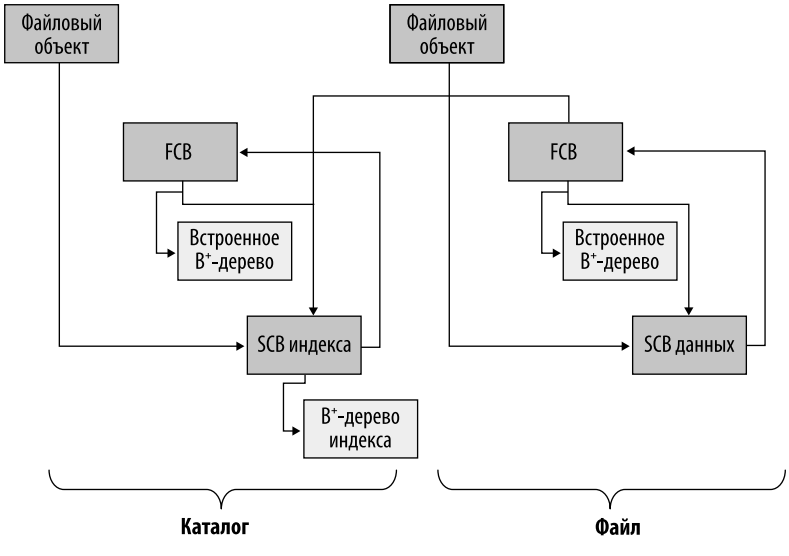


Рис. 11.85. Структуры данных ReFS файлов и каталогов в памяти

В ReFS каждый открытый файл имеет в памяти один FCB, на который могут указывать разные SCB в зависимости от количества открытых потоков. В отличие от NTFS, где FCB нужно знать только MFT-запись файла, чтобы правильно изменить атрибут, в ReFS FCB должен указывать на B⁺-дерево, которое представляет запись файла. Каждая строка в B⁺-дереве файла представляет собой атрибут файла, например ID, полное имя, таблицу экстентов и т. д. Ключом каждой строки является код атрибута (это целочисленное значение).

Файловые записи — это записи в каталоге, в котором находятся файлы. Корневой узел B⁺-дерева, представляющий файл, встроен в данные значения записи каталога и никогда не появляется в таблице объектов. Потоки данных файла, представленные таблицей экстентов, являются встроенными B⁺-деревьями в записи файла. Таблица экстентов индексируется по диапазону. Это означает, что каждая строка

в таблице экстентов имеет диапазон VCN в качестве ее ключа и LCN экстента файла в качестве значения. В ReFS таблица экстентов может быть очень большой — она действительно представляет собой обычное B⁺-дерево. Это позволяет ReFS поддерживать огромные файлы, обходя ограничения NTFS.

На рис. 11.86 показаны таблица объектов, файлы, каталоги и таблица экстентов файлов, которые в ReFS представлены в виде B⁺-деревьев и обеспечивают пространство имен файловой системы.

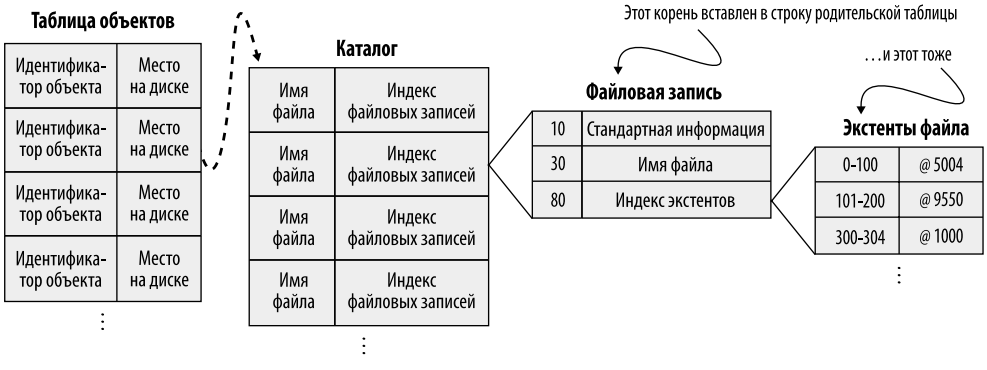


Рис. 11.86. Файлы и каталоги в ReFS

Каталоги — это B⁺-деревья Minstore, отвечающие за единое плоское пространство имен. Каталог ReFS может содержать:

- файлы;
- ссылки на каталоги;
- ссылки на другие файлы (идентификаторы файлов).

Строки в B⁺-дереве каталога состоят из пары <ключ, <тип, значение>>, где ключ — это имя записи, а значение зависит от типа записи каталога. С целью поддержки запросов и другой высокоуровневой семантики Minstore хранит также некоторые внутренние данные в невидимых строках каталога. Ключ таких строк начинается с нулевого символа Юникода. Еще одна строка, о которой стоит упомянуть, — это файловая строка каталога. В каждом каталоге есть запись, в ReFS она хранится как файловая строка в том же каталоге с использованием хорошо известного нулевого ключа. Это оказывает некоторое влияние на структуры данных в памяти, которые ReFS поддерживает для каталогов. В NTFS каталог — это в действительности свойство файловой записи, обозначенное посредством атрибутов Index Root и Index Allocation. В ReFS каталог — это файловая запись, хранящаяся в самом каталоге и называемая индексной записью каталога. Поэтому всякий раз, когда ReFS что-то делает с файлами в каталоге, она должна убедиться, что индекс каталога открыт и находится в памяти. Чтобы иметь возможность обновлять каталог, ReFS хранит указатель на индексную запись каталога в открытом блоке управления потоком.

Описанная конфигурация B⁺-деревьев ReFS не решает важную проблему. Каждый раз, когда система хочет перечислить файлы в каталоге, ей необходимо открыть и проанализировать B⁺-дерево каждого файла. Это означает, что нужно выполнить

множество запросов ввода-вывода к различным местам на базовом носителе. Если носитель — обычный жесткий диск, то производительность будет довольно низкой.

Чтобы решить эту проблему, ReFS хранит структуру данных `STANDARD_INFORMATION` в корневом узле встроенной таблицы файла, вместо того чтобы содержать ее в строке B⁺-таблицы дочернего файла. Данные `STANDARD_INFORMATION` содержат всю информацию, необходимую для перечисления файла, например время доступа к файлу, размер, атрибуты, идентификатор дескриптора безопасности, номер последовательности обновления и т. д. Встроенный корневой узел файла хранится в листовой корзине B⁺-дерева родительского каталога. Благодаря тому что структура данных находится во встроенном корневом узле файла, когда система перечисляет файлы в каталоге, ей нужно проанализировать только записи в B⁺-дереве каталога, не обращаясь к B⁺-таблицам, описывающим отдельные файлы. B⁺-дерево, представляющее каталог, уже находится в таблице страниц, поэтому перечисление происходит довольно быстро.

Дисковая структура ReFS

В этом разделе структура тома ReFS на диске описывается аналогично NTFS в предыдущем разделе. Здесь рассматриваются различия между NTFS и ReFS и не затрагиваются концепции, о которых говорилось ранее.

Загрузочный сектор тома ReFS состоит из небольшой структуры данных, которая, как и в NTFS, содержит основную информацию о томе (серийный номер, размер кластера и т. д.), идентификатор файловой системы (строку и версию ReFS OEM) и размер контейнера ReFS (подробнее об этом говорится в разделе «Тома магнитной записи внахлест» далее в этой главе). Наиболее важная структура данных в томе — суперблок тома. Он содержит смещение последних записей контрольной точки тома и реплицируется в трех разных кластерах. Чтобы смонтировать том, ReFS считывает одну из его контрольных точек, проверяет и разбирает ее (запись контрольной точки содержит контрольную сумму) и, наконец, получает смещение каждой глобальной таблицы.

Процесс монтирования тома открывает таблицу объектов и получает информацию, необходимую для чтения корневого каталога, в котором содержатся все деревья каталогов, составляющие пространство имен тома. Таблица объектов вместе с таблицей контейнеров — это одни из самых важных структур данных, которые являются отправной точкой для всех метаданных тома. Таблица контейнеров раскрывает пространство имен виртуализации, поэтому без нее ReFS не смогла бы правильно определить конечное расположение любого кластера. Minstore опционально позволяет клиентам хранить информацию в строках таблицы объектов. Значения строк таблицы объектов (рис. 11.87) состоят из двух частей: части,

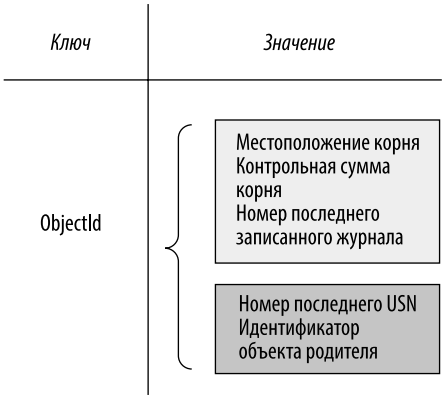


Рис. 11.87. Запись таблицы объектов состоит из части ReFS *(нижний прямоугольник)* и части Minstore *(верхний прямоугольник)*

принадлежащей Minstore, и части, принадлежащей ReFS. ReFS хранит информацию о родителях, а также последнюю метку для номеров USN в каталоге. (Подробнее см. в разделе «Безопасность и журнал изменений» далее в главе.)

Идентификаторы объектов

Еще одна проблема, которую необходимо решить в ReFS, связана с идентификаторами файлов. По разным причинам, в первую очередь для эффективного отслеживания и хранения метаданных о файлах без привязки информации к пространству имен, ReFS должна поддерживать приложения, открывающие файл по его идентификатору, например, с помощью API `OpenFileById`. В NTFS для этого используется файл `$Extend\$ObjId` с помощью корневого атрибута индекса `$0` (подробнее об этом рассказывалось ранее). В ReFS присвоение идентификатора каждому каталогу — тривиальная задача. Действительно, Minstore хранит объектный идентификатор каталога в таблице объектов. Проблема возникает, когда системе нужно присвоить идентификатор файлу, потому что в ReFS нет центрального хранилища идентификаторов файлов, как в NTFS. Чтобы найти правильный идентификатор файла, расположенного в дереве каталогов, ReFS разделяет пространство идентификаторов файлов на две части: каталог и файл. Идентификатор каталога занимает часть каталога и индексируется в ключе строки таблицы объектов. Файловая часть назначается из внутреннего пространства идентификаторов файлов каталога. Идентификатор, представляющий каталог, обычно имеет ноль в файловой части, а все файлы внутри каталога пользуются одной и той же его частью. ReFS поддерживает концепцию идентификаторов файлов, добавляя в B^+-дерево каталога отдельную строку, состоящую из пары `<FileId, FileName>`, которая сопоставляет идентификатор файла с именем файла в каталоге.

Когда системе требуется открыть расположенный в томе ReFS файл, используя его идентификатор, ReFS удовлетворяет запрос следующим образом.

1. Открывает каталог, указанный частью каталога.
2. Запрашивает строку `FileId` в B^+-дереве каталога, содержащую ключ, соответствующий части файла.
3. Ищет в B^+-дереве каталога имя файла, найденного при последнем просмотре.

Внимательные читатели могли заметить, что алгоритм не объясняет, что происходит, когда файл переименовывается или перемещается. Идентификатор переименованного файла должен совпадать с его предыдущим местоположением, даже если идентификатор нового каталога отличается в части каталога идентификатора файла. ReFS решает эту проблему, заменяя оригинальную запись ID файла, расположенную в старом B^+-дереве каталога, новой записью в роли «надгробного камня», которая вместо указания имени целевого файла в своем значении содержит вновь назначенный ID переименованного файла с измененными частями каталога и файла. В новом B^+-дереве каталога выделяется и еще одна новая запись File ID, которая позволяет присвоить переименованному файлу новый локальный ID. Если файл затем перемещается в другой каталог, то во втором каталоге его ID-запись удаляется, поскольку она больше не нужна. Для любого файла существует максимум один «надгробный камень».

Безопасность и журнал изменений

Механика поддержки безопасности объектов Windows в файловой системе заключается в основном в компонентах более высоких уровней, которые реализуются частями файловой системы, оставшимися неизменными со времен NTFS. Базовая реализация на диске была изменена для поддержки того же набора семантик. В ReFS дескрипторы безопасности объектов хранятся в B^+-таблице глобального каталога безопасности тома. Для каждого дескриптора безопасности в таблице вычисляется хеш по непубличному алгоритму, который работает только с автономными дескрипторами безопасности, и каждому присваивается идентификатор.

Когда система прикрепляет к файлу новый дескриптор безопасности, драйвер ReFS вычисляет его хеш и проверяет, присутствует ли он уже в глобальной таблице безопасности. Если хеш имеется в таблице, то ReFS получает его идентификатор и сохраняет его в структуре данных `STANDARD_INFORMATION`, расположенной во встроенном корневом узле B^+-дерева файла. Если хеша еще не существует в глобальной таблице безопасности, ReFS выполняет аналогичную процедуру, но сначала добавляет новый дескриптор безопасности в глобальное B^+-дерево и создает его новый идентификатор.

Строки глобальной таблицы безопасности имеют формат `<<hash, ID>, <security descriptor, ref. count>>`, где `hash` и `ID` описаны ранее, `security descriptor` — это необработанная байтовая полезная нагрузка самого дескриптора безопасности, а `ref. count` — приблизительная оценка того, сколько объектов в томе используют этот дескриптор безопасности.

Как описано в предыдущем разделе, в NTFS реализована функция журнала изменений, которая предоставляет приложениям и службам возможность просматривать внесенные ранее изменения файлов в томе. В ReFS предусмотрен совместимый с NTFS журнал изменений, реализованный несколько иначе. Журнал ReFS хранит записи об изменениях в файле журнала изменений, расположенном в другом глобальном B^+-дереве Minstore тома, в таблице каталогов метаданных. ReFS открывает и анализирует файл журнала изменений тома только после того, как том смонтирован. Максимальный размер журнала хранится в атрибуте `$USN_MAX` файла журнала. В ReFS каждый файл и каталог содержат свой последний номер последовательности обновления (update sequence number, USN) в структуре данных `STANDARD_INFORMATION`, хранящейся во встроенном корневом узле родительского каталога. С помощью файла журнала и USN каждого файла и каталога ReFS может предоставить следующие три FSCTL для чтения и перечисления файла журнала тома.

- **FSCTL_READ_USN_JOURNAL** считывает журнал USN напрямую. Вызывающие стороны указывают идентификатор журнала, который читают, и номер записи USN, которую ожидают прочитать.
- **FSCTL_READ_FILE_USN_DATA** получает информацию журнала изменений USN для указанного файла или каталога.
- **FSCTL_ENUM_USN_DATA** сканирует все файловые записи и перечисляет только те, которые в последний раз обновляли журнал USN с помощью записи, чей USN находится в диапазоне, указанном вызывающей стороной. ReFS может

удовлетворить запрос, просканировав таблицу объектов, затем просканировав каждый каталог, на который ссылается таблица объектов, и вернув файлы, находящиеся в этих каталогах, которые попадают в указанный диапазон времени. Это медленно, потому что каждый каталог нужно открывать, исследовать и т. д. B^+-деревья каталогов могут быть разбросаны по всему диску. Способ оптимизации ReFS заключается в том, что он хранит самый высокий USN всех файлов в каталоге в записи таблицы объектов этого каталога. При таком подходе ReFS может удовлетворить запрос, посещая только те каталоги, которые, как ей известно, находятся в указанном диапазоне.

РАСШИРЕННЫЕ ВОЗМОЖНОСТИ REFS

В этом разделе описываются расширенные возможности ReFS, которые объясняют, почему эта файловая система лучше подходит для больших серверных систем, подобных тем, что используются в инфраструктуре облака Azure.

Клонирование блоков файлов (поддержка моментальных снимков) и разреженный VDL

Традиционно системы хранения реализуют функции моментальных снимков и клонирования на уровне томов (см., например, динамические тома). В современных центрах обработки данных, где сотни виртуальных машин работают и хранятся на одном томе, такие методы уже не могут масштабироваться. Одними из первоначальных целей разработки ReFS были поддержка моментальных снимков на уровне файлов и масштабируемая поддержка клонирования (виртуальная машина обычно сопоставляется с одним или несколькими файлами в базовом хранилище хоста), что означало: ReFS должна была обеспечить быстрый метод клонирования всего файла или только его фрагментов. Клонирование диапазона блоков из одного файла в диапазон другого файла позволяет не только создавать снимки на уровне файлов, но и выполнять более тонкое клонирование для приложений, которым необходимо перемешивать блоки внутри одного или нескольких файлов. Одним из примеров является слияние VHD-дисков.

ReFS предоставляет новую функцию `FSCTL_DUPLICATE_EXTENTS_TO_FILE` для дублирования диапазона блоков из файла в другой диапазон того же файла или в другой файл. После операции клонирования запись в клонированные диапазоны любого из файлов будет выполняться по принципу записи в новое место с сохранением клонированного блока. Когда остается только одна ссылка, блок может быть записан на месте. В качестве параметров указываются дескрипторы исходного и целевого файлов, а также все детали того, откуда должен быть клонирован блок, какие блоки клонировать из источника, и целевой диапазон.

Как говорилось в предыдущем разделе, ReFS индексирует LCN, составляющие поток данных файла, в таблицу индексов экстентов — встроенное B^+-дерево, расположенное в одной из строк записи файла. Для поддержки клонирования блоков в Minstore используется новое глобальное индексное B^+-дерево, называемое

таблицей ссылок на счетчики блоков и отслеживающее счетчики ссылок на каждый экстент блоков, которые в данный момент клонируются. Поначалу индекс пуст. Первая успешная операция клонирования добавляет в таблицу одну или несколько строк, указывая, что блоки теперь имеют значение счетчика ссылок, равное двум. Если одно из представлений этих блоков будет удалено, то и строки будут удалены. Этот индекс используется в операциях записи, чтобы определить, требуется ли запись в новое место или можно выполнить ее на месте. К нему обращаются также перед тем, как помечать свободные блоки в распределителе. При освобождении кластеров, принадлежащих файлу, счетчик ссылок кластера-диапазона уменьшается. Если количество ссылок в таблице достигает нуля, то пространство действительно помечается как освобожденное.

На рис. 11.88 приведен пример клонирования файлов. После клонирования файла (файл 1 и файл 2 на рисунке) исходный и новый файлы имеют идентичные таблицы экстентов, а таблица ссылок на счетчики блока Minstore показывает две ссылки на оба экстента тома.

Minstore автоматически объединяет строки в таблице подсчета ссылок на блоки, когда это возможно, для уменьшения размера таблицы. В Windows Server 2016 HyperV использует новую функцию клонирования FSCTL. В результате дублирование виртуальной машины и объединение нескольких ее снимков происходят очень быстро.

ReFS поддерживает концепцию действительной длины данных в файле (Valid Data Length, VDL) аналогично NTFS. С помощью потока данных файла `$$ZeroRangeInStream` ReFS отслеживает состояние каждого выделенного блока данных файла — действительное или недействительное. Все новые выделения, запрашиваемые для файла, недействительны, а первая запись в файл делает их действительными. На запросы чтения из недействительных диапазонов файлов ReFS возвращает обнуленное содержимое. Эта техника похожа на DAL, о чем рассказывалось ранее в этой главе. Приложения могут логически обнулить часть файла без фактической записи данных с помощью управляющего кода файловой системы `FSCTL_SET_ZERO_DATA`. Эта функция используется HyperV для быстрого создания VHD-дисков фиксированного размера.

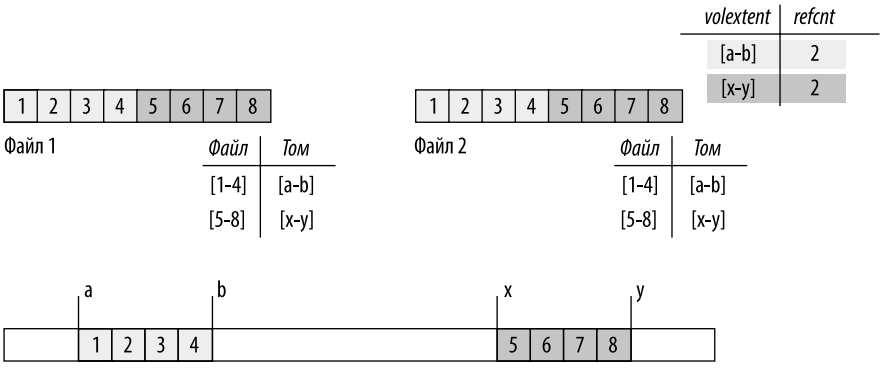
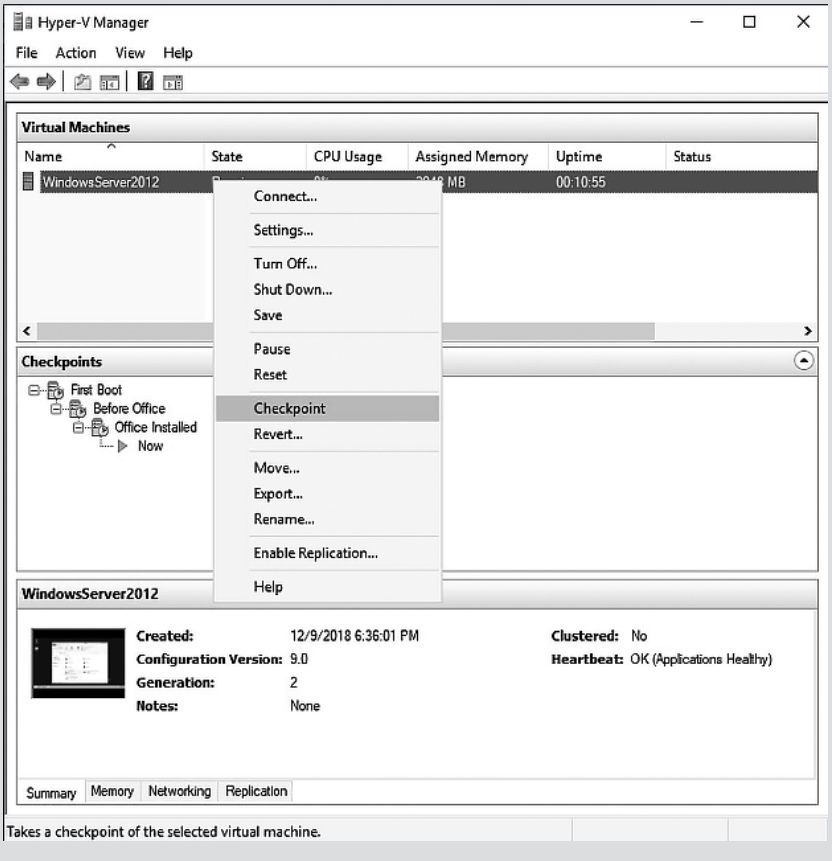


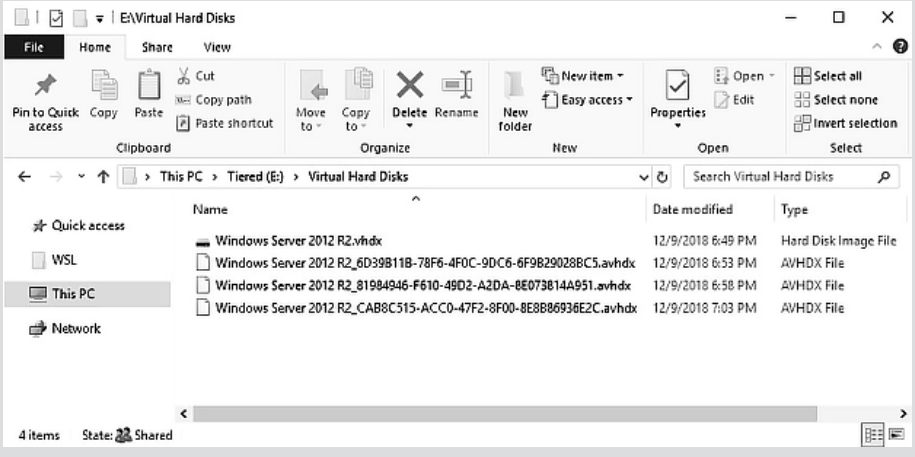
Рис. 11.88. Клонирование файла ReFS

ЭКСПЕРИМЕНТ. Проверка поддержки моментальных снимков ReFS с помощью HyperV

В этом эксперименте HyperV используется для проверки поддержки моментальных снимков томов в ReFS. С помощью диспетчера HyperV нужно создать виртуальную машину и установить на нее любую операционную систему. При первой загрузке создайте контрольную точку на виртуальной машине, щелкнув правой кнопкой мыши на ее имени и выбрав пункт меню Контрольная точка (Checkpoint). Затем установите на виртуальную машину несколько приложений — в примере далее показана машина с Windows Server 2012 и Office — и создайте еще одну контрольную точку.



Если выключить виртуальную машину и, используя File Explorer, найти место, где находится файл виртуального жесткого диска, то обнаружатся виртуальный жесткий диск и несколько других файлов, представляющих собой разницу между текущей и предыдущей контрольными точками.



Если снова открыть диспетчер HyperV и удалить все дерево контрольных точек, щелкнув правой кнопкой мыши на первой корневой контрольной точке и выбрав пункт меню Удалить поддерево контрольной точки (Delete Checkpoint Subtree), то обнаружится, что весь процесс слияния длится всего несколько секунд. Это объясняется тем, что HyperV использует поддержку блочного клонирования в ReFS с помощью кода управления вводом-выводом `FSCTL_DUPLICATE_EXTENTS_TO_FILE` для правильного объединения содержимого контрольных точек в базовый файл виртуального жесткого диска. Как объяснялось в предыдущих абзацах, клонирование блоков на самом деле не перемещает никаких данных. Если повторить этот эксперимент с томом, отформатированным с использованием файловой системы exFAT или NTFS, то обнаружится, что время, необходимое для объединения контрольных точек, значительно больше.

Сквозная запись в ReFS

Одной из целей создания ReFS было обеспечение практически нулевой недоступности из-за повреждения файловой системы. В следующем разделе будут описаны все доступные методы онлайн-ремонта, которые ReFS использует для восстановления после повреждения диска. Прежде чем описывать их, необходимо понять, как ReFS реализует сквозную запись при записи транзакций на базовый носитель.

Термин *«сквозная запись»* относится к любой вносящей изменения примитивной операции, например, созданию файла, расширению файла или записи блока, которая не должна завершаться до тех пор, пока система не даст разумную гарантию того, что результаты операции будут видны после восстановления после сбоя. Производительность сквозной записи критична для различных сценариев ввода-вывода, которые можно разделить на два вида операций файловой системы: с данными и метаданными.

Когда ReFS выполняет обновление файла на месте, не требуя мутации метаданных, например, когда система изменяет содержимое уже выделенного файла, не увеличивая его длину, дополнительные ресурсы для сквозной записи почти не нужны. Поскольку в ReFS для метаданных используется выделение при записи, давать гарантии сквозной записи для других сценариев, когда метаданные меняются, очень трудно. Например, гарантия того, что файл был переименован, подразумевает, что блоки метаданных от корня файловой системы до блока, описывающего имя файла, должны быть записаны в новое место. Свойство ReFS выделения при записи заключается в том, что данные не изменяются на месте. В отличие от NTFS, при восстановлении системы никогда не придется отменять какие-либо операции.

Чтобы добиться сквозной записи, Minstore использует журналирование с записью с упреждением (write-ahead-logging, WAL). В этой схеме (рис. 11.89) система добавляет записи в логически бесконечно длинный журнал, а при восстановлении журнал считывается и воспроизводится. В Minstore ведется журнал логических записей транзакций повтора для всех таблиц, кроме таблицы распределителей. Каждая запись в журнале описывает целую транзакцию, которая должна быть воспроизведена во время восстановления. Каждая запись транзакции имеет одну или несколько записей повторных операций, описывающих фактическую высокоуровневую операцию для выполнения, например, «вставить пару [ключ K/значение V] в таблицу X». Запись транзакции позволяет восстанавливать отдельные транзакции и является единицей атомарности — ни одна транзакция не будет частично повторена. С точки зрения логики запись в журнал принадлежит каждой транзакции ReFS, а небольшой буфер журнала содержит запись журнала. Если транзакция зафиксирована, то буфер журнала добавляется в журнал тома в памяти, который позже будет записан на диск. В противном случае, если транзакция прерывается, внутренний буфер журнала выбрасывается. Транзакции со сквозной записью ожидают от системы журналирования подтверждения того, что журнал был зафиксирован до этого момента, в то время как транзакции без сквозной записи могут продолжаться без подтверждения.

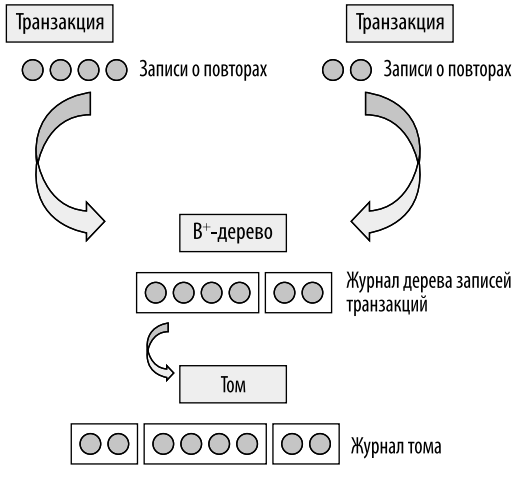


Рис. 11.89. Схема протоколирования с записью с упреждением в Minstore

Кроме того, ReFS использует контрольные точки для фиксации некоторых представлений системы на базовом диске, что делает ненужными некоторые из ранее внесенных записей журнала. Записи журнала повторного выполнения для транзакции больше не нужно повторять, когда контрольная точка фиксирует представление затронутых деревьев на диске. Это означает, что контрольная точка будет отвечать за определение диапазона записей журнала, которые могут быть выброшены движком журнала.

Поддержка восстановления ReFS

Чтобы обеспечить постоянную доступность тома файловой системы, ReFS использует различные стратегии восстановления. Хотя NTFS имеет аналогичную поддержку восстановления, цель ReFS — избавиться от автономных утилит проверки диска, таких как инструмент Chkdsk в NTFS, которые могут выполняться на огромных дисках в течение многих часов и требуют перезагрузки операционной системы. Существуют четыре стратегии восстановления ReFS.

- Повреждения метаданных обнаруживаются с помощью контрольных сумм и корректирующих кодов. Потоки целостности проверяют и поддерживают целостность данных файла с помощью контрольной суммы фактического содержимого файла (она хранится в строке таблицы B^+-дерева файла), что поддерживает целостность самого файла, а не только его метаданных в файловой системе.
- ReFS обдуманно восстанавливает все данные, которые признаны поврежденными, если доступна другая действительная копия. Другие копии могут быть предоставлены самой ReFS, которая хранит дополнительные копии собственных метаданных для критических структур, таких как таблица объектов, или с помощью резервирования томов, обеспечиваемого Storage Spaces (см. раздел «Storage Spaces» далее в этой главе).
- В ReFS реализована операция *спасения*, которая удаляет поврежденные данные из пространства имен файловой системы, пока та находится в режиме онлайн.
- ReFS восстанавливает утраченные метаданные наилучшим из возможных способов.

Первая и вторая стратегии — это свойства библиотеки Minstore, от которой зависит ReFS. (Подробнее о потоках целостности рассказано далее в этом разделе.) Таблица объектов и все глобальные таблицы B^+-деревьев Minstore содержат контрольную сумму для каждой ссылки, указывающей на дочерние или маршрутизирующие узлы, хранящиеся в разных блоках диска. Когда Minstore обнаруживает, что блок не соответствует его ожиданиям, он автоматически пытается восстановить его из одной из своих дублированных копий, если она доступна. Если копия недоступна, Minstore возвращает ошибку на верхний уровень ReFS. ReFS реагирует на ошибку, инициализируя *онлайн-спасение*.

Термин *«спасение»* относится к любым исправлениям, необходимым для восстановления максимального количества данных, когда ReFS обнаруживает повреждение метаданных в B^+-дереве каталогов. Спасение — это развитие техники *выжигания*. Целью выжигания было вернуть том в рабочее состояние, даже если это могло привести к потере поврежденных данных. При этом все поврежденные метаданные удалялись из пространства имен файлов, которое после восстановления становилось доступным.

Предположим, что маршрутизирующий узел B^+-дерева каталога поврежден. В этом случае операция выжигания исправит родительский узел, переписав все ссылки на дочерний и восстановив баланс дерева, но данные, на которые изначально указывал поврежденный узел, будут полностью потеряны. Minstore не знает, как восстановить записи, на которые указывал поврежденный маршрутизирующий узел.

Чтобы решить эту проблему и правильно восстановить дерево каталогов в процессе спасения, ReFS необходимо знать идентификаторы подкаталогов, даже если

сама таблица каталогов недоступна, например, из-за поврежденного узла маршрутизации. Восстановить части утраченного дерева каталогов возможно благодаря введению глобальной таблицы тома, называемой таблицей *родителей и детей*, обеспечивающей информационную избыточность каталога.

Ключ в таблице родителей и детей представляет собой идентификатор родительской таблицы, а данные содержат список идентификаторов дочерних таблиц. Процедура спасения сканирует эту таблицу, считывает список дочерних таблиц и создает новое, неповрежденное B^+-дерево, содержащее все подкаталоги поврежденного узла. Чтобы полностью восстановить поврежденный родительский каталог, ReFS, помимо идентификаторов дочерних таблиц, по-прежнему нужны имена дочерних таблиц, которые изначально хранились в ключах родительского B^+-дерева. В дочерней таблице есть автоматическая запись с этой информацией типа *«ссылка на каталог»* (подробнее см. предыдущий раздел). Процесс спасения открывает восстановленную дочернюю таблицу, считывает автоматическую запись и заново вставляет ссылку на каталог в родительскую таблицу. Эта стратегия позволяет ReFS восстановить все подкаталоги поврежденного маршрутизирующего или корневого узла, но не файлы. На рис. 11.90 приведен пример операций выжигания и спасения для поврежденного корневого узла, представляющего каталог Bar. С помощью операции спасения ReFS удается быстро вернуть файловую систему в рабочее состояние, потеряв только два файла в каталоге.

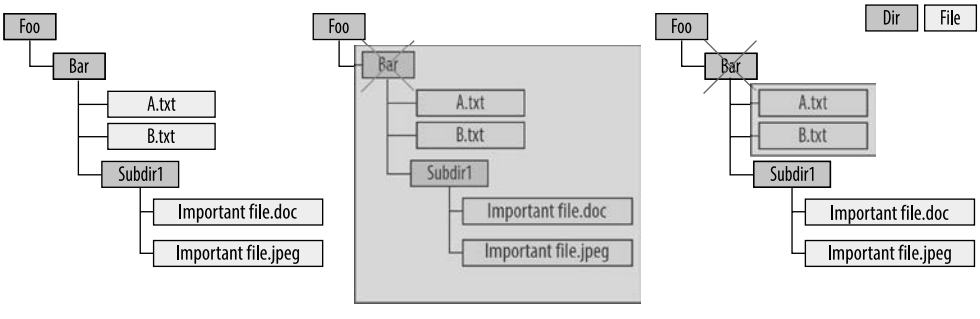


Рис. 11.90. Сравнение между операциями выжигания и спасения

Файловая система ReFS после завершения спасения пытается восстановить недостающую информацию, используя наилучшие из доступных методов. Например, она может восстановить недостающие идентификаторы файлов, считывая информацию из других корзин, благодаря правилу объединения, разделяющему идентификаторы файлов и таблиц. Кроме того, ReFS вносит в объектную таблицу Minstore небольшое количество дополнительной информации, чтобы ускорить восстановление. Несмотря на то что в ReFS есть эти эвристики, важно понимать, что ReFS в первую очередь полагается на резервирование, обеспечиваемое метаданными и стеком хранения, чтобы исправить повреждение без потери данных.

В очень редких случаях, когда повреждены критические метаданные, ReFS может смонтировать том в режиме «только для чтения», но не для поврежденных таблиц. Например, при повреждении таблицы контейнеров и всех ее дубликатов том нельзя будет смонтировать в режиме «только для чтения». Пропуская эти таблицы, файловая система может просто игнорировать применение таких глобальных таблиц,

например распределителя, сохраняя при этом возможность для пользователя восстановить свои данные.

Наконец, ReFS поддерживает также потоки целостности файлов, в которых контрольная сумма используется, чтобы гарантировать целостность данных файла, а не только метаданных файловой системы. Для потоков целостности ReFS хранит контрольную сумму каждого прогона, составляющего таблицу экстентов файла, — она содержится в секции данных строки таблицы экстентов. Контрольная сумма позволяет ReFS проверять целостность данных перед доступом к ним. Перед возвратом любых данных, для которых включены потоки целостности, ReFS вычисляет их контрольную сумму и сравнивает ее с контрольной суммой, содержащейся в метаданных файла. Если контрольные суммы не совпадают, значит, данные повреждены.

Файловая система ReFS предоставляет управляющий код `FSCTL_SCRUB_DATA`, который используется *скруббером*, также известным как сканер целостности данных. Он реализован в библиотеке Discan.dll и представлен в виде задачи планировщика заданий, которая выполняется при запуске системы и дополнительно раз в неделю. Когда скруббер отправляет FSCTL драйверу ReFS, тот начинает проверку целостности всего тома: загрузочного раздела, всех глобальных B^+-деревьев и метаданных файловой системы.

ПРИМЕЧАНИЕ Операция спасения в режиме онлайн, описанная в этом разделе, отличается от ее аналога в режиме офлайн. Эту операцию поддерживает инструмент refsutil.exe, входящий в состав Windows. Он используется, когда том настолько поврежден, что его невозможно подключить даже в режиме «только для чтения», — редкое явление. Операция автономного спасения проходит по всем кластерам томов в поисках страниц метаданных и применяет наилучшие из доступных методов, чтобы восстановить их.

Обнаружение утечек

Утечка кластера — это ситуация, когда кластер помечен как выделенный, но ссылок на него нет. В ReFS утечки кластеров могут происходить по разным причинам. Когда в каталоге обнаруживается повреждение, онлайновое спасение способно изолировать его и перестроить дерево, потеряв в итоге лишь некоторые файлы, находившиеся в самом корневом каталоге. Сбой системы до того, как алгоритм обновления дерева запишет транзакцию Minstore на диск, может привести к потере имени файла. В этом случае данные файла правильно записаны на диск, но в ReFS нет метаданных, указывающих на него. Таблица B^+-дерева, представляющая сам файл, все еще может существовать где-то на диске, но ее встроенная таблица больше не связана ни с одним B^+-деревом каталога.

Встроенный инструмент refsutil.exe, доступный в Windows, поддерживает операцию *обнаружения утечек*, которая может сканировать весь том и с помощью Minstore перемещаться по всему пространству имен тома. Затем она строит список всех найденных в пространстве имен B^+-деревьев (каждое из них идентифицируется известной структурой данных, содержащей идентификационный заголовок) и, запрашивая распределители Minstore, сравнивает список каждого идентифицированного дерева со списком деревьев, отмеченных распределителем как валидные. Если обнаруживается несоответствие, то средство обнаружения утечек уведомляет драйвер файловой системы ReFS, который пометит кластеры, выделенные для обнаруженного потерянного дерева, как освобожденные.

Другой вид утечки, которая может произойти в томе, затрагивает таблицу счетчиков ссылок на блоки, например, когда диапазон кластера, расположенный в одной из его строк, имеет больший номер счетчика ссылок, чем реальные файлы, которые на него ссылаются. Утилита нижнего регистра способна подсчитать правильное количество ссылок и устранить проблему.

Для корректного выявления и устранения утечек инструмент обнаружения утечек должен работать с автономным томом, но, используя технику, аналогичную онлайн-сканированию NTFS, он может работать со снимком целевого тома, доступным только для чтения, который предоставляет служба Volume Shadow Copy.

ЭКСПЕРИМЕНТ. Использование Refsutil для поиска и устранения утечек в томе ReFS

В этом эксперименте в томе ReFS применяется встроенный инструмент refsutil.exe для поиска и устранения утечек в кластере, которые могут произойти в этом томе. По умолчанию он не требует размонтирования тома, поскольку работает с его снимком, доступным только для чтения. Чтобы позволить инструменту исправлять найденные утечки, можно отменить эту настройку с помощью аргумента командной строки `/x`. Переключатель `/v` включает подробный вывод данных. В этом примере в качестве диска `E:` был смонтирован том ReFS объемом 1 Тбайт. Откройте административную командную строку и введите следующую команду:

```
C:\>refsutil leak /v e:
Creating volume snapshot on drive \\?\Volume{92aa4440-51de-4566-8c00-
bc73e0671b92}...
Creating the scratch file...
Beginning volume scan... This may take a while...
Begin leak verification pass 1 (Cluster leaks)...
End leak verification pass 1. Found 0 leaked clusters on the volume.

Begin leak verification pass 2 (Reference count leaks)...
End leak verification pass 2. Found 0 leaked references on the volume.

Begin leak verification pass 3 (Compacted cluster leaks)...
End leak verification pass 3.

Begin leak verification pass 4 (Remaining cluster leaks)...
End leak verification pass 4. Fixed 0 leaks during this pass.

Finished.
Found leaked clusters: 0
Found reference leaks: 0
Total cluster fixed  : 0
```

Тома магнитной записи внахлест

На момент написания этой книги одна из самых больших проблем для обычных жестких дисков связана с физическими ограничениями, с которыми сталкивается процесс записи. Для увеличения емкости диска необходимо постоянно увеличивать плотность записи дисковых пластин, а для чтения и записи крошечных единиц

информации физический размер головок дисков продолжает уменьшаться. Это приводит к снижению энергетического барьера для изменения битов, а значит, тепловая энергия окружающей среды с большей вероятностью может случайно изменить бит, что снижает целостность данных. Твердотельные накопители (solid state drives, SSD) получили распространение во многих потребительских системах, а для больших серверов хранения данных требуется больше места и меньшая стоимость, которую обычные жесткие диски все еще обеспечивают. Для преодоления проблемы жестких дисков было разработано множество решений. Наиболее эффективное из них называется *магнитной записью внахлест* (shingled magnetic recording, SMR) (рис. 11.91). В отличие от перпендикулярной магнитной записи (perpendicular magnetic recording, PMR), в которой дорожки расположены параллельно, головка для чтения данных на дисках SMR меньше, чем головка для записи. Более крупная пишущая головка позволяет эффективнее намагничивать носитель, то есть записывать на него без ущерба для читаемости и стабильности.

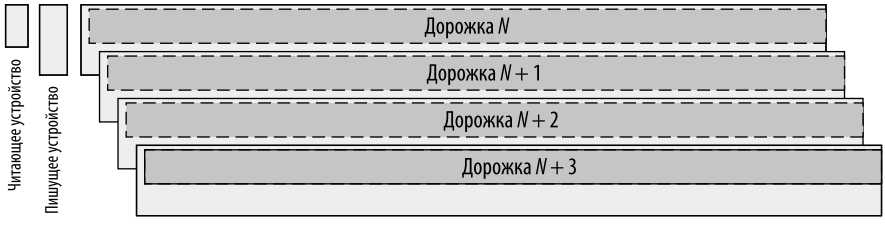


Рис. 11.91. На дисках SMR дорожка записи больше дорожки считывания

Новая конфигурация создает некоторые логические проблемы. Практически невозможно сделать запись на дорожку диска, не заменив частично данные на соседней дорожке. Чтобы решить эту проблему, SMR-диски разделяют на зоны, которые технически называются полосами. Существуют два основных вида зон:

- обычные, или быстрые, зоны работают как традиционные PMR-диски, на которых разрешена запись в произвольное место;
- зоны с указателем записи — полосы, которые имеют собственный указатель записи и требуют строго последовательной записи. Конечно, это не совсем так, поскольку SMR-диски с поддержкой хоста поддерживают также концепцию зон *предпочтительной записи*, в которых запись в произвольное место все еще можно сделать. Однако этот тип зон не используется в ReFS.

Каждая полоса на SMR-диске обычно составляет 256 Мбайт и работает как базовая единица ввода-вывода. Это означает, что система может записывать в одну полосу, не вмешиваясь в работу следующей полосы. Существуют три типа SMR-дисков.

- **Управляемый диском.** Диск кажется хосту идентичным обычному диску без записи внахлест. Хосту не нужно следовать какому-либо специальному протоколу, поскольку вся обработка данных, наличие дисковых зон и ограничений последовательной записи регулируются прошивкой устройства. Этот тип SMR-дисков отлично совместим, но имеет ряд ограничений: дисковый кэш, используемый для преобразования записей в произвольное место в последовательные, ограничен,

очистка зон сложна, а обнаружить последовательные записи непросто. Эти ограничения снижают производительность.

- **Управляемый хостом.** Устройство требует от хоста строгого соблюдения специальных правил ввода-вывода. Хост должен писать последовательно, чтобы не уничтожить существующие данные. Диск отказывается выполнять команды, нарушающие это требование. Управляемые хостом диски поддерживают только зоны последовательной записи и обычные зоны, где последними могут быть любые носители, включая не-SMR, управляемые диском SMR, и флеш-память.
- **Знающий о хосте.** Комбинация управляемого диском и управляемого хостом. Диск может управлять записью внахлест и будет выполнять любую команду, которую ему даст хост, независимо от того, является ли она последовательной. Однако хост знает о том, что на диске ведется запись внахлест, и может запросить у него информацию о зоне SMR. Это позволяет хосту оптимизировать запись с учетом ее характера, а также обеспечивает гибкость и обратную совместимость диска. Такие диски поддерживают концепцию зон предпочтительной последовательной записи.

На момент написания этой книги ReFS — единственная файловая система, которая может поддерживать управляемые хостом SMR-диски. Стратегия, используемая ReFS для поддержки таких дисков, которые могут быть очень емкими — 20 Тбайт и более, аналогична той, что применяется для многоуровневых томов, обычно создаваемых Storage Spaces. (Более подробную информацию о Storage Spaces см. в последнем разделе.)

Поддержка ReFS для многоуровневых томов и SMR

Многоуровневые тома похожи на SMR-диски, знающие о хосте. Они состоят из быстрой области произвольного доступа, обычно предоставляемой SSD, и более медленной области последовательной записи. Однако это необязательное условие, многоуровневые диски могут состоять из разных дисков с произвольным доступом, даже с одинаковой скоростью. ReFS способна правильно управлять многоуровневыми томами и SMR-дисками, предоставляя новый логический косвенный слой между файлами и пространством имен каталогов поверх пространства имен тома. Этот новый слой делит том на непересекающиеся логические контейнеры, поэтому отдельный кластер одновременно присутствует только в одном контейнере. Контейнер представляет собой область в томе, все контейнеры в томе всегда одинакового размера, который определяется в зависимости от типа базового диска: 64 Мбайт для стандартных многоуровневых дисков и 256 Мбайт для дисков SMR. Контейнеры называются *полосами ReFS*, поскольку при их использовании с дисками SMR размер контейнеров становится точно таким же, как размер полос SMR, и каждый контейнер сопоставляется один к одному с каждой полосой SMR.

Косвенный слой настраивается и обеспечивается глобальной таблицей контейнеров (рис. 11.92). Строки этой таблицы состоят из ключей, которые хранят идентификатор и тип контейнера. В зависимости от типа контейнера, который может быть уплотненным или сжатым, данные в строке различаются. Для неуплотненных контейнеров (подробности об уплотнении ReFS см. в следующем разделе) данные строки представляют собой структуру данных, содержащую отображение диапазона кластеров, к которым обращается контейнер. Это обеспечивает ReFS отображение виртуальных LCN на реальные LCN в пространстве имен.

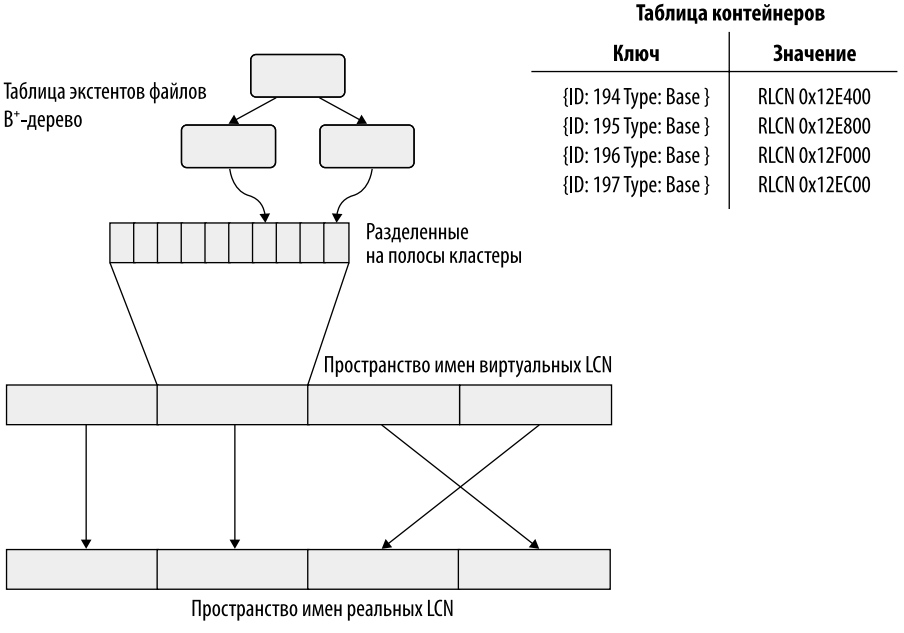


Рис. 11.92. Таблица контейнеров создает косвенный слой отображения виртуальных LCN на реальные LCN

Таблица контейнеров очень важна: все данные, управляемые ReFS и Minstore, за небольшими исключениями, должны проходить через таблицу контейнеров, поэтому ReFS поддерживает несколько ее копий. Для выполнения ввода-вывода блока ReFS сначала должна найти местоположение контейнера экстента, чтобы обнаружить реальное местоположение данных. Это достигается с помощью таблицы экстентов, которая содержит целевой виртуальный LCN диапазона кластера в разделе данных своих строк. Идентификатор контейнера выводится из LCN с помощью математического соотношения. Новый уровень непрямой связи позволяет ReFS перемещать местоположение контейнеров без обращения к таблицам экстентов файлов или их изменения.

ReFS использует уровни, созданные Storage Spaces, аппаратные многоуровневые тома и SMR-диски. ReFS перенаправляет небольшие случайные операции ввода-вывода на часть более быстрых уровней и отправляет эти записи партиями на более медленные уровни с помощью последовательной записи — вывод происходит с гранулярностью контейнера. Действительно, в ReFS термин *«быстрый уровень»* (*fast tier* или *flash tier*) относится к зоне произвольного доступа, которая может быть обеспечена обычными полосами SMR-диска или всей совокупностью устройств SSD или NVMe. Термин *«медленный уровень»* (*slow tier* или *HDD tier*) относится к полосам последовательной записи или обычному жесткому диску. ReFS использует различные модели поведения в зависимости от класса базового носителя. Для дисков без SMR не предъявляются требования к последовательности, поэтому кластеры можно выделять из любого места тома. Для дисков SMR, как уже говорилось, предъявляются строгие требования к последовательности, поэтому ReFS никогда не записывает случайные данные на медленный ярус.

По умолчанию все метаданные, которые использует ReFS, должны оставаться на быстром уровне. ReFS пытается применять его даже при обработке общих запросов на запись. На дисках без SMR по мере заполнения флеш-контейнеров ReFS перемещает контейнеры из флеш-памяти на HDD. Это означает, что при обработке задачи непрерывной записи ReFS постоянно перемещает контейнеры из флеш-памяти на жесткий диск. Она также способна при необходимости делать обратное — выбирать контейнеры с HDD и перемещать их во флеш-память для последующей записи. Эта функция называется *ротацией контейнеров* и реализуется в два этапа. После того как драйвер хранилища скопировал фактические данные, ReFS изменяет отображение LCN контейнеров, показанное ранее. Никаких изменений в таблице экстентов файлов не требуется.

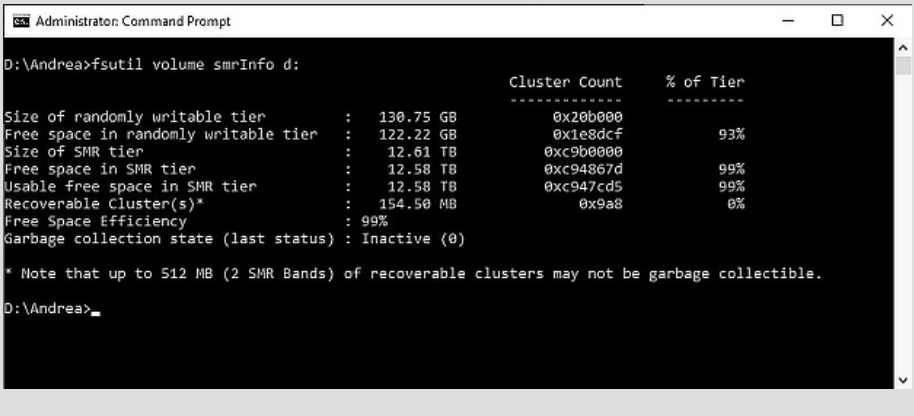
Ротация контейнеров реализована только для дисков без SMR. Это важно, поскольку на дисках SMR драйвер файловой системы ReFS никогда автоматически не перемещает данные между уровнями. Приложения, поддерживающие диски SMR и желающие записывать данные на уровень хранения SMR, могут использовать контрольный код `FSCTL_SET_REFS_FILE_STRICTLY_SEQUENTIAL`. Если приложение отправляет управляющий код на файловый дескриптор, то драйвер ReFS записывает все новые данные на уровень хранения тома.

ЭКСПЕРИМЕНТ. Изучение уровней SMR-дисков

Можно использовать инструмент FsUtil, предоставляемый Windows, чтобы запросить информацию о SMR-диске — например, размер каждого уровня, доступное и свободное пространство и т. д. Для этого просто запустите инструмент в административной командной строке. Можно запустить командную строку от имени администратора, выполнив поиск `cmd` в окне поиска Cortana и выбрав Запустить от имени администратора (Run As Administrator) после щелчка правой кнопкой мыши на ярлыке командной строки. Введите следующие параметры:

```
fsutil volume smrInfo <VolumeDrive>
```

заменяя часть `<VolumeDrive>` буквой имеющегося диска SMR.



Кроме того, можно запустить сборку мусора с помощью следующей команды:

```
fsutil volume smrGc <VolumeDrive> Action=startfullspeed
```

Сборку мусора можно остановить или приостановить с помощью относительного параметра `Action`. Можно запустить более точную сборку мусора, указав параметр `IoGranularity`, который задает гранулярность ввода-вывода сборки мусора, и используя значение `start` вместо `startfullspeed`.

Уплотнение контейнеров

Ротация контейнеров сопряжена с проблемами с производительностью, особенно при хранении небольших файлов, обычно не занимающих целую полосу. Более того, на дисках SMR ротация контейнеров никогда не выполняется, как объяснялось ранее. Напомним, что каждая полоса SMR имеет связанный с ней указатель записи, реализованный аппаратно, который определяет место для последовательной записи. Если система будет записывать до или после указателя записи непоследовательно, это приведет к повреждению данных, расположенных в других кластерах, поэтому прошивка SMR должна запретить такую запись.

ReFS поддерживает два типа контейнеров: базовые контейнеры, отображающие диапазон виртуального кластера непосредственно на физическое пространство, и уплотненные контейнеры, отображающие виртуальный контейнер на множество различных базовых контейнеров. Для корректного отображения соответствия между пространством, отображаемым уплотненным контейнером, и составляющими его базовыми контейнерами в ReFS реализована битовая карта распределения, хранящаяся в строках глобальной индексной таблицы контейнеров (это еще одна таблица, в которой каждая строка описывает один уплотненный контейнер). Битовая карта имеет бит, установленный в 1, если относительный кластер выделен, в противном случае он устанавливается в 0.

На рис. 11.93 показан пример базового контейнера (C32), сопоставляющего диапазон виртуальных LCN (от 0x8000 до 0x8400) с LCN реального тома (от 0xB800 до 0xBC00, идентифицируется R46). Как говорилось ранее, идентификатор контейнера данного диапазона виртуальных LCN определяется по начальному номеру виртуального кластера, все контейнеры виртуально смежные. Таким образом, ReFS никогда не нужно искать идентификатор контейнера для заданного диапазона контейнеров. В контейнере C32 из 1024 кластеров непрерывны лишь 560 (0x230). Только свободное пространство в конце базового контейнера может быть задействовано ReFS. И для дисков без SMR большой фрагмент пространства, расположенный в середине базового контейнера, в случае освобождения тоже может быть использован повторно. Даже для дисков без SMR важно, чтобы пространство было непрерывным.

Если контейнер становится фрагментированным из-за того, что некоторые небольшие экстенты файлов в конечном счете освобождаются, то ReFS может преобразовать базовый контейнер в уплотненный. Эта операция позволяет ReFS повторно использовать свободное пространство контейнера, не перераспределяя ни одной строки в таблице экстентов файлов, которые задействуют кластеры, описанные самим контейнером.

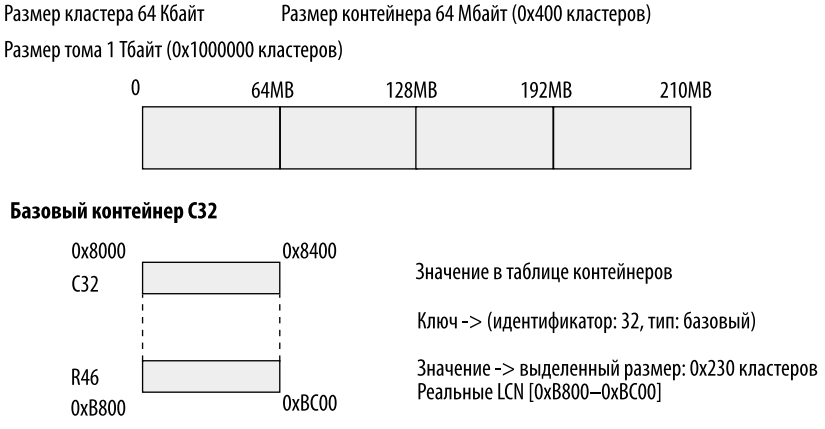


Таблица экстентов

Диапазон VCN	LCN	Идентификатор контейнера
[0 - 0x400]	0x18400	97
[0x400 - 0x800]	0x32000	200
[0x800 - 0xA00]	0x61E00	391
[0xA00 - 0xC00]	0x11200	68
[0xC00 - 0xD20]	0x8110	32

Рис. 11.93. Пример базового контейнера, адресуемого файлом размером 210 Мбайт. Контейнер C32 использует только 35 Мбайт из своих 64 Мбайт

ReFS предоставляет возможность дефрагментировать фрагментированные контейнеры. В ходе обычной активности ввода-вывода в системе возникает множество мелких файлов или фрагментов данных, которые необходимо обновить или создать. В результате контейнеры, расположенные на медленном уровне, могут содержать небольшие участки освобожденных кластеров и быстро фрагментироваться. Уплотнение контейнеров — это название функции, которая создает новые пустые полосы на медленном уровне, позволяя контейнерам правильно дефрагментироваться. Уплотнение контейнеров выполняется только на уровне хранения многоуровневого тома и было разработано со следующими двумя целями.

- **Уплотнение — это сборщик мусора для SMR-дисков.** В SMR ReFS может записывать данные на уровень хранения только последовательно. Небольшие данные не могут быть отдельно обновлены в контейнере, расположенном на медленном уровне. Эти данные не находятся в том месте, на которое ведет указатель записи SMR, поэтому любой ввод-вывод такого рода может повредить другие данные, принадлежащие этому контейнеру. В этом случае данные копируются в новую полосу. У дисков без SMR нет такой проблемы, ReFS обновляет данные, находящиеся на медленном уровне, напрямую.
- **В многоуровневых томах без SMR уплотнение является генератором для ротации контейнеров.** Образовавшиеся свободные контейнеры можно использовать в качестве целей для прямой ротации при перемещении данных с быстрого уровня на медленный.

Во время форматирования тома ReFS выделяет несколько базовых контейнеров с уровня хранения только для уплотнения, они называются *уплотненными зарезервированными контейнерами*. Уплотнение работает так: вначале на медленном уровне выполняется поиск фрагментированных контейнеров. ReFS считывает такой контейнер в системной памяти и дефрагментирует его. Затем дефрагментированные данные сохраняются в уплотненном зарезервированном контейнере, расположенном на уровне хранения, как описано ранее. Исходный контейнер, адресуемый таблицей размеров файлов, становится уплотненным. Диапазон, описывающий его, становится виртуальным (уплотнение добавляет еще один уровень перенаправления), указывая на виртуальные LCN, описываемые другим базовым контейнером — зарезервированным. По окончании уплотнения исходный физический контейнер помечается как освобожденный и может быть использован повторно для других целей. Также он может стать новым уплотненным зарезервированным контейнером. Поскольку контейнеры, расположенные на медленном уровне, обычно сильно фрагментируются за относительно небольшое время, уплотнение может привести к образованию большого количества пустых полос на этом уровне.

Кластеры, выделенные уплотненным контейнером, могут храниться в разных базовых контейнерах. Чтобы правильно управлять такими кластерами, ReFS использует еще один дополнительный непрямой уровень, обеспечиваемый глобальной таблицей индексов контейнеров и иной компоновкой уплотненного контейнера. На рис. 11.94 показан тот же фрагментированный контейнер, что и на рис. 11.93, который был уплотнен — освобождены 272 из 560 кластеров. В таблице контейнеров строка, описывающая уплотненный контейнер, хранит сопоставление между диапазоном кластеров, описываемым уплотненным контейнером, и виртуальными кластерами, описываемыми базовыми контейнерами. Компактные контейнеры поддерживают максимум четыре различных диапазона, называемых *ножками*. Эти четыре ножки создают второй непрямой уровень и позволяют ReFS эффективно выполнять дефрагментацию контейнера. Битовая карта распределения уплотненного контейнера обеспечивает также второй непрямой уровень. Проверяя положение выделенных кластеров, которым соответствует единица в битовой карте, ReFS может правильно отобразить каждый фрагментированный кластер уплотненного контейнера.

В примере на рис. 11.94 первый бит, установленный в 1, находится в позиции 17, что в шестнадцатеричном исчислении равно 0x11. В примере один бит соответствует 16 кластерам, но в реальной реализации один бит соответствует только одному кластеру. Это означает, что первый кластер, выделенный по смещению 0x110 в уплотненном контейнере C32, хранится в виртуальном кластере 0x1F2E0 в базовом контейнере C124. Свободное место, оставшееся после кластера по смещению 0x230 в уплотненном контейнере C32, отображается в базовый контейнер C56. Физический контейнер R46 был переотображен ReFS и превратился в пустой уплотненный зарезервированный контейнер, отображенный базовым контейнером C180.

На дисках SMR процесс, запускающий уплотнение, называется сборкой мусора. Для них приложение может в любой момент вручную запустить, остановить или приостановить сборку мусора с помощью управляющего кода файловой системы `FSCTL_SET_REFS_SMR_VOLUME_GC_PARAMETERS`.

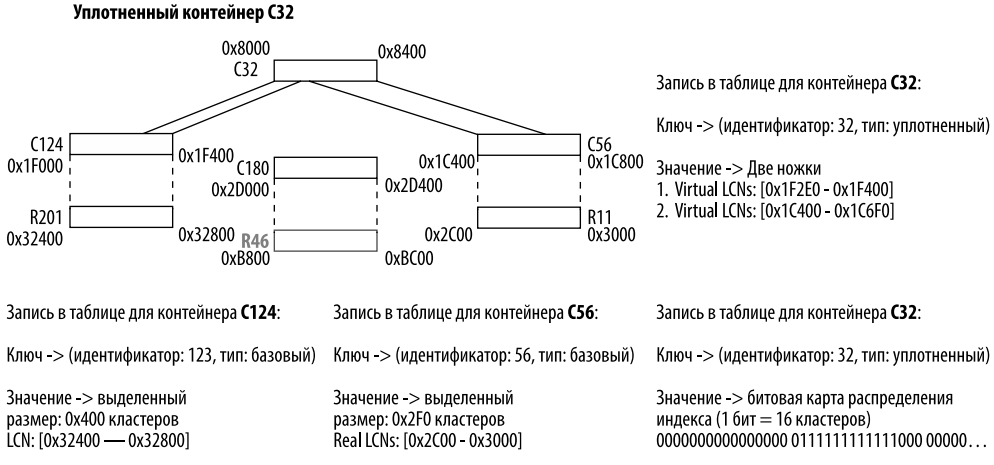


Рис. 11.94. Контейнер C32 был уплотнен в базовых контейнерах C124 и C56

В отличие от NTFS, на дисках без SMR механизм анализа томов ReFS может автоматически запускать процесс уплотнения контейнеров. ReFS отслеживает свободное пространство медленного и быстрого уровней, а также доступное свободное пространство для записи на медленном уровне. Если разница между свободным и доступным пространством превышает пороговое значение, то механизм анализа томов срабатывает и запускает процесс уплотнения. Кроме того, если базовое хранилище предоставляется Storage Spaces, то уплотнение контейнеров происходит периодически и выполняется выделенным потоком.

Сжатие и фантомные файлы

ReFS не поддерживает собственное сжатие файловой системы, но на многоуровневых томах файловая система может сохранить больше свободных контейнеров на медленном уровне благодаря сжатию контейнеров. Каждый раз, когда ReFS выполняет уплотнение контейнеров, она считывает в память исходные данные, находящиеся во фрагментированном базовом контейнере. На этом этапе, если включено сжатие, ReFS сжимает данные и записывает их в сжатый контейнер. Она поддерживает четыре алгоритма сжатия: LZNT1, LZX, XPRESS и XPRESS_HUFF.

Многие программные решения для иерархического управления хранением (hierarchical storage management, HMR) поддерживают концепцию *фантомного* файла. Это состояние может быть вызвано разными причинами. Например, когда HSM переносит пользовательский файл или некоторые его фрагменты в облачный сервис, а пользователь позже изменяет копию, находящуюся в облаке, с помощью другого устройства, драйверу фильтра HSM необходимо отслеживать, какая часть файла изменилась, и устанавливать состояние фантомности для каждого измененного диапазона файла. Обычно HSM-фильтры отслеживают фантомное состояние с помощью своих драйверов фильтров. В ReFS этого не требуется, поскольку файловая система ReFS предоставляет новый код управления вводом-выводом, `FSCTL_GHOST_FILE_EXTENTS`. Драйверы фильтров могут отправить IOCTL драйверу

ReFS, чтобы определить часть файла как фантомную. Кроме того, они могут запрашивать диапазоны файла, находящегося в фантомном состоянии, с помощью другого кода управления вводом-выводом — `FSCTL_QUERY_GHOSTED_FILE_EXTENTS`.

ReFS реализует фантомные файлы, сохраняя информацию о новом состоянии непосредственно в таблице экстентов файла, реализуемой через встроенную таблицу в файловой записи, как объяснялось в предыдущем разделе. Драйвер фильтра может установить фантомное состояние для каждого диапазона файла, который должен быть выровнен по кластерам. Когда драйвер ReFS перехватывает запрос на чтение экстента, являющегося фантомным, он возвращает вызывающей стороне код ошибки `STATUS_GHOSTED`, а драйвер фильтра, перехватив его, может перенаправить чтение в нужное место — в облако, как в предыдущем примере.

STORAGE SPACES

Storage Spaces — это технология, которая заменяет динамические диски и обеспечивает виртуализацию физического оборудования для хранения данных. Изначально она была разработана для крупных серверов хранения данных, но теперь доступна даже в клиентских редакциях Windows 10. Storage Spaces позволяет также создавать виртуальные диски, состоящие из различных базовых физических носителей. Эти носители могут иметь различные характеристики производительности.

На момент написания книги Storage Spaces может работать с четырьмя типами устройств хранения: NVMe (Nonvolatile memory express), постоянной памятью (PM) в виде флеш-дисков, твердотельными накопителями SATA и SAS (SSD) и обычными жесткими дисками (HDD). NVMe считается самым быстрым, а HDD — самым медленным. Система Storage Spaces была разработана, чтобы обеспечивать следующие четыре характеристики.

- **Производительность.** В StorageSpaces реализована поддержка встроенного кэша на стороне сервера для повышения производительности хранилища, а также поддержка многоуровневых дисков и конфигурации RAID 0.
- **Надежность.** Помимо разделенных томов (RAID 0), Storage Spaces поддерживает конфигурации Mirror (RAID 1 и 10) и Parity (RAID 5, 6, 50, 60), когда данные распределяются по разным физическим дискам или узлам кластера.
- **Гибкость.** Storage Spaces позволяет системе создавать виртуальные диски, которые можно автоматически перемещать между узлами кластера и автоматически сжимать или расширять в зависимости от реального потребления пространства.
- **Доступность.** Тома Storage Spaces обладают встроенной отказоустойчивостью. Это означает, что если диск или даже весь сервер, входящий в кластер, выходит из строя, то Storage Spaces может перенаправить трафик ввода-вывода на другие работающие узлы без вмешательства пользователя и определенным образом. У Storage Spaces нет единой точки отказа.

Storage Spaces Direct — это развитие технологии Storage Spaces. Storage Spaces Direct предназначена для крупных центров обработки данных, где несколько серверов, содержащих разные медленные и быстрые диски, используются вместе для создания пула. Прежняя технология не поддерживала кластеры серверов,

не подключенных к дисковым массивам JBOD, поэтому к названию был добавлен термин *direct*. Все серверы подключаются через быстрое Ethernet-соединение — например, 10 или 40 Гбайт/с. Представление удаленных дисков как локальных для системы обеспечивается двумя драйверами: драйвером кластерного мини-порта Clusport.sys и драйвером кластерного блочного фильтра Clusbflt.sys, рассмотрение которых выходит за рамки данной главы. Все физические единицы хранения, локальные и удаленные диски добавляются в *пул хранения*, являющийся основной единицей управления, агрегации и изоляции, на основе которого могут быть созданы виртуальные диски.

Весь кластер хранения отображается внутри Storage Spaces с помощью XML-файла *BluePrint*. Он автоматически создается графическим интерфейсом Storage Spaces и описывает весь кластер с помощью дерева различных объектов хранения данных: *стоек*, *шасси*, *машин*, *JBOD* (Just a Bunch of Disks) и *дисков*. Эти сущности составляют каждый уровень кластера. Сервер (машина) может быть подключен к различным JBOD или иметь различные диски, подключенные непосредственно к нему. В этом случае JBOD абстрагируется и представляется только одной сущностью. Точно так же несколько машин могут быть расположены на одном шасси, которое может быть частью серверной стойки. Наконец, кластер может состоять из нескольких серверных стоек. Используя представление BluePrint, Storage Spaces может работать со всеми дисками кластера и перенаправлять трафик ввода-вывода на нужную замену в случае сбоя на диске, JBOD или машине. Storage Spaces Direct может выдерживать максимум два одновременных сбоя.

Внутренняя архитектура Storage Spaces

Одно из самых больших различий между Storage Spaces и динамическими дисками заключается в том, что Storage Spaces создает объекты виртуальных дисков, которые представляются системе как реальные объекты дисковых устройств драйвером хранилища Storage Spaces, Spaceport.sys. Динамические диски работают на более высоком уровне — объекты виртуальных томов открыты для системы. Это означает, что приложения пользовательского режима все еще могут обращаться к оригинальным дискам. Диспетчер томов — это компонент, отвечающий за создание единого тома, состоящего из нескольких динамических томов. Драйвер Storage Spaces — это драйвер фильтра (полный драйвер фильтра, а не мини-фильтр), находящийся между диспетчером разделов Partmgr.sys и драйвером класса диска.

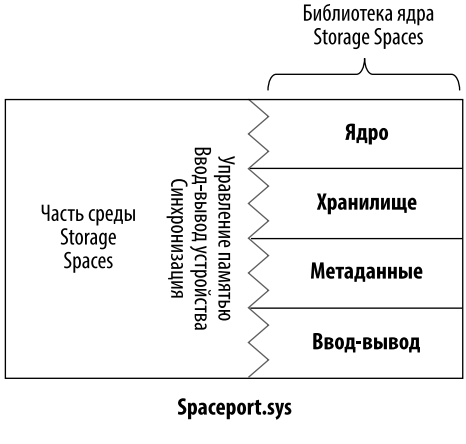


Рис. 11.95. Архитектура Storage Spaces

Архитектура Storage Spaces (рис. 11.95) состоит в основном из двух частей: платформонезависимой библиотеки, реализующей ядро Storage Spaces, и части среды, зависящей от платформы и связывающей ядро Storage Spaces с текущей средой. Уровень среды обеспечивает Storage Spaces базовую функциональность,

реализующуюся по-разному в зависимости от платформы, на которой она работает. Поскольку Storage Spaces может использоваться в качестве загружаемых объектов, загрузчик и диспетчер загрузки Windows должны знать, как анализировать пространства хранения, так что существует необходимость в реализации как в UEFI, так и в Windows. Основная базовая функциональность включает в себя процедуры управления памятью (alloc, free, lock, unlock и т. д.), процедуры ввода-вывода устройств (Control, Pnp, Read и Write), а также методы синхронизации. Как правило, эти функции являются обертками для конкретных системных процедур. Например, сервис чтения на платформах Windows реализуется путем создания IRP типа `IRP_MJ_READ` и отправки его нужному драйверу диска, а в средах UEFI — с помощью `BLOCK_IO_PROTOCOL`.

Помимо загрузки и реализации в ядре Windows, Storage Spaces должен быть доступен во время аварийных дампов, что обеспечивается драйвером фильтра аварийных дампов Spacedump.sys. Storage Spaces доступен даже в виде библиотеки пользовательского режима Backspace.dll, совместимой с устаревшими операционными системами Windows, которым необходимо работать с виртуальными дисками, созданными Storage Spaces, особенно с файлами VHD, и даже в виде драйвера UEFI DXE HyperSpace.efi, который может быть выполнен UEFI BIOS в случаях, когда даже системный раздел EFI присутствует в объекте пространства хранения. Некоторые новые устройства Surface продаются с большим твердотельным диском, который состоит из двух или более быстрых NVMe-дисков.

Ядро Storage Spaces реализовано в виде статической библиотеки, не зависящей от платформы, и импортируется всеми слоями окружения. Оно состоит из четырех слоев: ядро, хранилище, метаданные и ввод-вывод. Ядро — это самый верхний уровень, реализующий все сервисы, которые предоставляет Storage Spaces. Хранилище — это компонент, считывающий и записывающий записи, принадлежащие базе данных кластера, созданной из файла BluePrint. Метаданные интерпретируют двоичные записи, считанные хранилищем, и раскрывают всю базу данных кластера через различные объекты: *Pool*, *Drive*, *Space*, *Extent*, *Column*, *Tier* и *Metadata*. Компонент ввода-вывода, являющийся самым нижним уровнем, может выдавать запросы ввода-вывода на нужное устройство в кластере в правильном последовательном порядке благодаря данным, разобранным на более высоких уровнях.

Услуги, предоставляемые Storage Spaces

Storage Spaces поддерживает различные конфигурации типов дисков. С его помощью пользователь может создавать виртуальные диски, полностью состоящие из быстрых дисков (SSD, NVMe и PM), медленных дисков или даже из всех четырех поддерживаемых типов дисков (гибридная конфигурация). В случае гибридных развертываний, где используется смесь различных классов устройств, Storage Spaces поддерживает следующие функции, позволяющие кластеру быть быстрым и эффективным.

- **Серверный кэш.** Storage Spaces может скрыть быстрый диск от кластера и задействовать его в качестве кэша для более медленных дисков. Storage Spaces поддерживает PM-диски для использования в качестве кэша для дисков NVMe или SSD, NVMe-диски — для применения в качестве кэша SSD-дисков и SSD-диски — для использования в качестве кэша обычных HDD-дисков. В отличие

от многоуровневых дисков, кэш невидим для файловой системы, расположенной над виртуальным томом. Это означает, что кэш не имеет представления о том, что к данному файлу обращались чаще, чем к другому. Storage Spaces реализует быстрый кэш для виртуального диска с помощью журнала, отслеживающего горячие и холодные блоки. Горячие блоки представляют собой части файлов — экстенты, к которым система обращается часто, в то время как холодные блоки — это части файлов, к которым практически не обращаются. Журнал реализует кэш как очередь, в которой горячие блоки всегда находятся в начале, а холодные — в конце. Таким образом, холодные блоки могут быть удалены из кэша, если он переполнен, или храниться в более медленном хранилище, а горячие блоки обычно остаются в кэше на более длительное время.

- **Многоуровневость.** Storage Spaces может создавать многоуровневые диски, которые управляются ReFS и NTFS. Если ReFS поддерживает SMR-диски, то NTFS — только многоуровневые диски, предоставляемые Storage Spaces. Файловая система отслеживает горячие и холодные блоки и чередует полосы в зависимости от использования файла (см. раздел «Поддержка ReFS для многоуровневых томов и SMR» ранее в этой главе). Storage Spaces предоставляет драйверу файловой системы поддержку закрепления — функции, позволяющей *закрепить* файл на быстром уровне и заблокировать его там до тех пор, пока он не будет отсоединен. В этом случае ротация полос никогда не выполняется. Windows задействует функцию закрепления для хранения новых файлов на быстром уровне при выполнении обновления ОС.

Как говорилось ранее, одна из главных целей Storage Spaces — гибкость. Storage Spaces поддерживает создание виртуальных дисков, которые могут расширяться и занимают только выделенное пространство на устройствах базового кластера. Такой тип виртуального диска называется *тонким диском*. В отличие от дисков с фиксированным резервированием, где все пространство выделяется базовому кластеру хранения, тонкие диски выделяют только то пространство, которое реально используется. Таким образом, можно создавать виртуальные диски, размер которых значительно превышает размер базового кластера хранения. Когда свободного места становится мало, системный администратор может динамически добавлять диски в кластер. Storage Spaces автоматически включает новые физические диски в пул и перераспределяет выделенные блоки между новыми дисками.

Storage Spaces поддерживает тонкие диски с помощью *слэбов*. Слэб — это единица выделения, похожая на концепцию контейнера ReFS, но применяемая к стеку более низкого уровня. Слэб — единица выделения виртуального диска, а не концепция файловой системы. По умолчанию размер каждого слэба составляет 256 Мбайт, но он может быть и больше, если это позволяет базовый кластер хранения, то есть когда в кластере много свободного места. Ядро Storage Spaces отслеживает каждый слэб в виртуальном диске и может динамически выделять или освобождать их с помощью собственного распределителя. Стоит отметить, что каждый слэб является точкой надежности: в зеркальных и паритетных конфигурациях данные, хранящиеся в слэбе, автоматически реплицируются по всему кластеру.

Когда создается тонкий диск, необходимо указать его размер. Это значение будет использоваться файловой системой для правильного форматирования нового тома и создания необходимых метаданных. Когда том готов, Storage Spaces выделяет слэбы

только при реальной записи новых данных на диск — этот метод называется выделением при записи. Обратите внимание на то, что тип диска не виден файловой системе, которая находится поверх тома, и поэтому не знает, является ли базовый диск тонким.

Storage Spaces позволяет избавиться от единой точки отказа за счет применения зеркалирования и сопряжения. В больших кластерах хранения данных, состоящих из нескольких дисков, в качестве решения четности обычно применяется RAID 6. RAID 6 допускает отказ максимум двух базовых устройств и поддерживает беспрепятственное восстановление данных без вмешательства пользователя. К сожалению, когда кластер сталкивается с одной или двумя точками отказа, время, необходимое для восстановления массива — среднее время восстановления (mean time to repair, MTTR), — очень велико, что часто приводит к серьезному снижению производительности.

Storage Spaces решает эту проблему с помощью алгоритма локального кода реконструкции (local reconstructon code, LRC), уменьшающего количество чтений, необходимых для восстановления большого дискового массива, ценой одной дополнительной единицы четности. Как показано на рис. 11.96, алгоритм LRC делает это, разделяя дисковый массив на строки и добавляя единицу четности для каждой строки. Если диск выходит из строя, то необходимо прочитать только остальные диски этого ряда. В результате восстановление вышедшего из строя массива проходит гораздо быстрее и эффективнее.

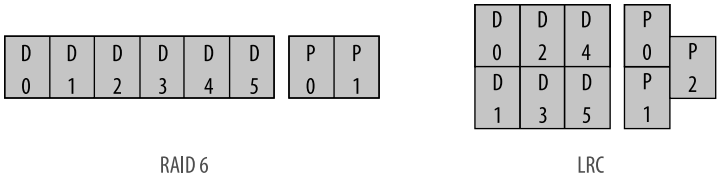


Рис. 11.96. RAID 6 и четность LRC

На рис. 11.96 показано сравнение типичной реализации RAID 6 с контролем четности и реализации LRC на кластере из восьми дисков. В конфигурации RAID 6 при отказе одного или двух дисков для правильного восстановления недостающей информации необходимо прочитать остальные шесть дисков. В LRC же нужно прочитать только диски, которые принадлежат к тому же ряду, что и отказавший диск.

ЭКСПЕРИМЕНТ. Создание многоуровневых томов

Storage Spaces поддерживается как серверными, так и клиентскими версиями Windows 10. Многоуровневые диски можно создавать с помощью графического интерфейса пользователя, а также Windows PowerShell. В этом эксперименте для создания виртуального многоуровневого диска понадобится рабочая станция, на которой, кроме загрузочного диска Windows, есть также пустой SSD и пустой обычный диск HDD. Для проверки можно эмулировать подобную конфигурацию с помощью HyperV. В этом случае один файл виртуального диска должен располагаться на SSD, а другой — на обычном диске.

Сначала нужно открыть как администратор Windows PowerShell, щелкнув правой кнопкой мыши на значке меню Пуск и выбрав Windows PowerShell (администратор) (Windows PowerShell (Admin)). Убедитесь, что система уже определила тип установленных дисков:

```
PS C:\> Get-PhysicalDisk | FT DeviceId, FriendlyName, UniqueID, Size, MediaType, CanPool

DeviceId FriendlyName            UniqueID                       Size MediaType CanPool
-------- ------------            --------                       ---- --------- -------
2        Samsung SSD 960 EVO 1TB eui.0025385C61B074F7 1000204886016 SSD         False
0        Micron 1100 SATA 512GB  500A071516EBA521      512110190592 SSD         True
1        TOSHIBA DT01ACA200      500003F9E5D69494     2000398934016 HDD         True
```

В предыдущем примере система уже определила два SSD и один обычный жесткий диск. Необходимо убедиться в том, что для пустых дисков значение `CanPool` установлено равным `True`. В противном случае диск содержит допустимые разделы, которые необходимо удалить. Если используется виртуализированная среда, то часто система не может правильно определить тип носителя базового диска:

```
PS C:\> Get-PhysicalDisk | FT DeviceId, FriendlyName, UniqueID, Size, MediaType, CanPool

DeviceId FriendlyName      UniqueID                                 Size MediaType    CanPool
-------- ------------      --------                                 ---- ---------    -------
2        Msft Virtual Disk 600224802F4EE1E6B94595687DDE774B 137438953472 Unspecified     True
1        Msft Virtual Disk 60022480170766A9A808A30797285D77 1099511627776 Unspecified    True
0        Msft Virtual Disk 6002248048976A586FE149B00A43FC73 274877906944 Unspecified    False
```

В этом случае необходимо вручную указать тип диска с помощью команды `Set-PhysicalDisk -UniqueId (Get-PhysicalDisk)[<IDX>].UniqueID -MediaType <Type>`, где `IDX` — номер строки в предыдущем выводе, а `MediaType` — SSD или HDD в зависимости от типа диска, например:

```
PS C:\> Set-PhysicalDisk -UniqueId (Get-PhysicalDisk)[0].UniqueID -MediaType SSD
PS C:\> Set-PhysicalDisk -UniqueId (Get-PhysicalDisk)[1].UniqueID -MediaType HDD
PS C:\> Get-PhysicalDisk | FT DeviceId, FriendlyName, UniqueID, Size, MediaType, CanPool

DeviceId FriendlyName      UniqueID                                 Size MediaType     CanPool
-------- ------------      --------                                 ---- ---------     -------
2        Msft Virtual Disk 600224802F4EE1E6B94595687DDE774B 137438953472 SSD             True
1        Msft Virtual Disk 60022480170766A9A808A30797285D77 1099511627776 HDD             True
0        Msft Virtual Disk 6002248048976A586FE149B00A43FC73 274877906944 Unspecified    False
```

На этом этапе нужно создать пул хранения, который будет содержать все физические диски, из которых будет состоять новый виртуальный диск. Затем следует создать уровни хранения. В этом примере пул хранения назван DefaultPool:

```
PS C:\> New-StoragePool -StorageSubSystemId (Get-StorageSubSystem).UniqueId -FriendlyName
DeafultPool -PhysicalDisks (Get-PhysicalDisk -CanPool $true)

FriendlyName OperationalStatus HealthStatus IsPrimordial IsReadOnly    Size AllocatedSize
------------ ----------------- ------------ ------------ ----------    ---- -------------
Pool         OK                Healthy      False             1.12 TB        512 MB

PS C:\> Get-StoragePool DefaultPool | New-StorageTier -FriendlyName SSD -MediaType SSD
...
PS C:\> Get-StoragePool DefaultPool | New-StorageTier -FriendlyName HDD -MediaType HDD
...
```

Наконец, можно создать виртуальный многоуровневый том, присвоив ему имя и указав нужный размер каждого уровня. В этом примере создается многоуровневый том TieredVirtualDisk, состоящий из уровня производительности 120 Гбайт и уровня хранения 1000 Гбайт:

```
PS C:\> $SSD = Get-StorageTier -FriendlyName SSD
PS C:\> $HDD = Get-StorageTier -FriendlyName HDD
PS C:\> Get-StoragePool Pool | New-VirtualDisk -FriendlyName "TieredVirtualDisk"
-ResiliencySettingName "Simple" -StorageTiers $SSD, $HDD -StorageTierSizes 128GB, 1000GB
...
PS C:\> Get-VirtualDisk | FT FriendlyName, OperationalStatus, HealthStatus, Size,
FootprintOnPool

FriendlyName      OperationalStatus HealthStatus          Size FootprintOnPool
------------      ----------------- ------------          ---- ---------------
TieredVirtualDisk OK                Healthy      1202590842880   1203664584704
```

После создания виртуального диска необходимо создать разделы и отформатировать новый том стандартными средствами, например, с помощью оснастки «Управление дисками» или инструмента Format. После завершения форматирования тома можно проверить, действительно ли полученный том многоуровневый, с помощью инструмента fsutil.exe:

```
PS E:\> fsutil tiering regionList e:
Total Number of Regions for this volume: 2
Total Number of Regions returned by this operation: 2

   Region # 0:
        Tier ID: {448ABAB8-F00B-42D6-B345-C8DA68869020}
        Name: TieredVirtualDisk-SSD
        Offset: 0x0000000000000000
        Length: 0x0000001dff000000

   Region # 1:
        Tier ID: {16A7BB83-CE3E-4996-8FF3-BEE98B68EBE4}
        Name: TieredVirtualDisk-HDD
        Offset: 0x0000001dff000000
        Length: 0x000000f9ffe00000
```

ЗАКЛЮЧЕНИЕ

Windows поддерживает широкий спектр форматов файловых систем, доступных как для локальной системы, так и для удаленных клиентов. Архитектура драйвера фильтра файловой системы обеспечивает чистый способ расширения и упрощения доступа к файловой системе, а NTFS и ReFS предоставляют надежный, безопасный и масштабируемый формат файловой системы для хранения локальных файлов. ReFS является относительно новой файловой системой и реализует некоторые расширенные функции, предназначенные для больших серверных сред. NTFS также была обновлена с поддержкой новых типов устройств и новых функций, таких как POSIX-удаление, онлайн-проверка диска и шифрование.

Диспетчер кэша обеспечивает высокоскоростной интеллектуальный механизм для сокращения дискового ввода-вывода и повышения общей пропускной

способности системы. Благодаря кэшированию на основе виртуальных блоков диспетчер кэша может выполнять интеллектуальное опережающее чтение, в том числе на удаленных сетевых файловых системах. Опираясь на примитив отображенного файла диспетчера глобальной памяти для доступа к файловым данным, диспетчер кэша может обеспечить специальный механизм быстрого ввода-вывода для сокращения процессорного времени, необходимого для операций чтения и записи. При этом все вопросы, связанные с управлением физической памятью, остаются за диспетчером памяти Windows, что позволяет сократить дублирование кода и повысить эффективность.

Благодаря поддержке дисков DAX и постоянной памяти, Storage Spaces и Storage Spaces Direct, многоуровневых томов и совместимости с дисками SMR Windows продолжает оставаться в авангарде архитектур хранения следующего поколения, предназначенных для обеспечения высокой доступности, надежности, производительности и масштабируемости на уровне облака.

В следующей главе рассматриваются запуск и завершение работы в Windows.

ГЛАВА 12

Запуск и завершение работы системы

В этой главе описываются шаги, необходимые для загрузки Windows, и параметры, которые могут повлиять на запуск системы. Понимание деталей процесса загрузки поможет диагностировать проблемы, возникающие во время нее. Обсуждаются детали новой прошивки UEFI и улучшения, которые она дает по сравнению с историческим BIOS. Рассказывается о роли диспетчера загрузки, загрузчика Windows, ядра NT и всех компонентов, участвующих в стандартной загрузке и в новом процессе Secure Launch, который обнаруживает любые виды атак на загрузку. Затем объясняется, что может пойти не так в ходе загрузки и как это исправить. Наконец, поясняется, что происходит во время нормального выключения системы.

ПРОЦЕСС ЗАГРУЗКИ

Описание процесса загрузки Windows начинается с установки Windows и переходит к выполнению файлов поддержки загрузки. Драйверы устройств являются важной частью процесса загрузки, поэтому объясняется, как они контролируют момент, когда загружаются и инициализируются. Затем описывается, как инициализируются исполнительные подсистемы и как ядро запускает пользовательскую часть Windows, начиная с процесса Session Manager Smss.exe, который запускает первые две сессии — 0 и 1. Попутно отмечаются моменты, когда появляются различные сообщения на экране, чтобы помочь читателю соотнести внутренний процесс загрузки с тем, что он видит.

Ранние фазы процесса загрузки существенно различаются в системах с интерфейсом расширяемой прошивки (Extensible Firmware Interface, EFI) и старых системах с базовой системой ввода-вывода (Basic Input/Output System, BIOS). EFI — это новый стандарт, позволяющий отказаться от большей части устаревшего 16-битного кода, используемого в системах BIOS, и дающий возможность загружать предзагрузочные программы и драйверы для поддержки фазы загрузки операционной системы. Стандарт EFI 2.0, известный как Unified EFI или UEFI, применяют подавляющее большинство производителей компьютеров. В следующих разделах описывается часть процесса загрузки, характерная для машин на базе UEFI.

Для поддержки различных реализаций прошивок Windows предоставляет архитектуру загрузки, которая скрывает многие различия от пользователей и разработчиков, чтобы обеспечить согласованную среду и работу независимо от типа прошивки в установленной системе.

Загрузка UEFI

Процесс загрузки начинается не тогда, когда включается компьютер или нажимается кнопка перезагрузки. Это происходит, когда на компьютер устанавливается Windows. В определенный момент во время выполнения программы установки основной жесткий диск системы подготавливается таким образом, чтобы он был понятен диспетчеру загрузки и прошивке UEFI. Прежде чем перейти к рассмотрению того, что делает код диспетчера загрузки, вкратце рассмотрим интерфейс платформы UEFI.

UEFI — это набор программного обеспечения, предусматривающего первый базовый программный интерфейс к платформе. Термином *«платформа»* обозначаются материнская плата, чипсет, процессор и другие компоненты, составляющие «двигатель» машины. Как показано на рис. 12.1, спецификации UEFI предоставляют четыре базовых сервиса, которые имеются в большинстве существующих архитектур процессоров — x86, ARM и т. д. Для этого краткого введения используется архитектура x86-64.

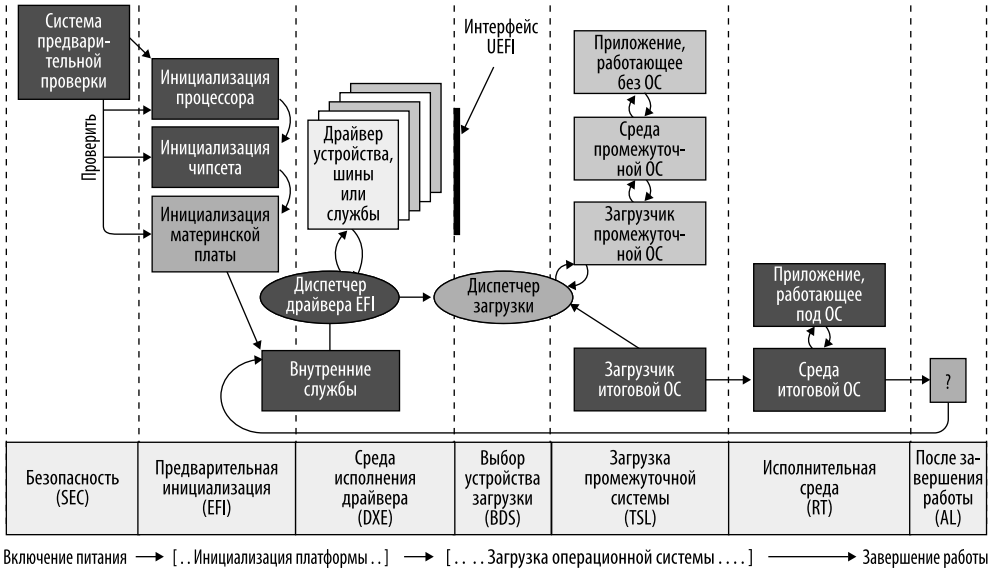


Рис. 12.1. Структура UEFI

- **Включение питания.** При включении питания платформы фаза безопасности UEFI обрабатывает событие перезапуска платформы, проверяет код модулей предварительной инициализации EFI и переключает процессор из 16-битного реального режима в 32-битный плоский режим, по-прежнему без поддержки подкачки.
- **Инициализация платформы.** Фаза предварительной инициализации (Pre EFI Initialization, PEI) инициализирует процессор, код ядра UEFI, чипсет и в конце передает управление фазе среды выполнения драйверов (Driver Execution Environment, DXE). Фаза DXE — это первый код, который полностью

выполняется в 64-битном режиме. Последний модуль фазы PEI, называемый DXE IPL, переключает режим выполнения в 64-битный длинный режим. Эта фаза выполняет поиск в томе прошивки, хранящемся на микросхеме перезаписываемой SPI-памяти, и выполняет DXE-драйверы запуска каждого периферийного устройства. Безопасная загрузка, Secure Boot — важная функция безопасности, о которой речь пойдет позже в этой главе, в разделе «Безопасная загрузка», — реализована в виде DXE-драйвера UEFI.

- **Загрузка ОС.** После завершения фазы UEFI DXE управление передается фазе выбора загрузочного устройства (Boot Device Selection, BDS). Она отвечает за реализацию загрузчика UEFI Boot Loader. Фаза BDS находит и запускает диспетчер загрузки UEFI Windows, установленный программой Setup.
- **Выключение.** В прошивке UEFI реализованы некоторые службы времени выполнения, доступные даже для ОС, которые помогают выключить платформу. Windows обычно не использует их, а полагается на интерфейсы ACPI (Advanced Configuration and Power Interface).

Описание всей структуры UEFI выходит за рамки тематики этой книги. После завершения фазы UEFI BDS прошивка по-прежнему контролирует платформу, предоставляя загрузчику ОС следующие услуги.

- **Службы загрузки.** Обеспечивают базовую функциональность загрузчика и других приложений EFI, такую как базовое управление памятью, синхронизация, текстовый и графический консольный ввод-вывод, дисковый и файловый ввод-вывод. Службы загрузки реализуют некоторые процедуры, способные перечислять и запрашивать установленные протоколы, то есть интерфейсы EFI. Эти службы доступны, только пока прошивка контролирует платформу, и удаляются из памяти после того, как загрузчик вызовет `ExitBootServices` из API исполнительной среды EFI.
- **Службы исполнительной среды.** Предоставляют услуги даты и времени, капсульного обновления (обновления прошивки) и методы, способные получить доступ к данным NVRAM, например переменные UEFI. Эти службы остаются доступными, пока операционная система запущена.
- **Данные конфигурации платформы.** Системные таблицы ACPI и SMBIOS всегда доступны через структуру UEFI.

Диспетчер загрузки UEFI может читать данные с жестких дисков компьютера и записывать на них и понимает основные файловые системы, такие как FAT, FAT32 и El Torito, используемую для загрузки с CD-ROM. Согласно спецификациям, загрузочный жесткий диск должен быть разбит по схеме GPT (GUID partition table), которая задействует GUID для идентификации различных разделов и их роли в системе. Схема GPT лишена всех ограничений старой схемы MBR и допускает максимум 128 разделов, используя 64-битный режим адресации LBA, в результате чего поддерживаются разделы огромного размера. Каждый раздел идентифицируется с помощью уникального 128-битного GUID-значения. Еще один GUID служит для определения типа раздела. Хотя UEFI определяет только три типа разделов, каждый разработчик операционной системы устанавливает собственные типы GUID разделов. Стандарт UEFI требует наличия как минимум одного системного раздела EFI, отформатированного в файловой системе FAT32.

Приложение Setup инициализирует диск и обычно создает не менее четырех разделов.

- Системный раздел EFI, куда копируются диспетчер загрузки Bootmgrfw.efi, приложение для тестирования памяти Memtest.efi, политики блокировки системы для систем с поддержкой Device Guard Winsipolicy.p7b и файл ресурсов загрузки Bootres.dll.
- Раздел восстановления, где хранятся файлы boot.sdi и Winre.wim, необходимые для загрузки среды восстановления в случае проблем с запуском. Он отформатирован с помощью NTFS.
- Зарезервированный раздел Windows, который Setup применяет в качестве быстрой восстанавливаемой области для хранения временных данных. Кроме того, некоторые системные инструменты используют зарезервированный раздел для переадресации поврежденных секторов загрузочного тома. Зарезервированный раздел не содержит файловой системы.
- Загрузочный раздел с загрузочными файлами, на который устанавливается Windows. Обычно он не совпадает с системным разделом. Этот раздел форматируется в NTFS — единственной поддерживаемой файловой системе, с которой Windows может загружаться при установке на фиксированный диск.

Программа установки, разместив файлы Windows в загрузочном разделе, копирует диспетчер загрузки в системный раздел EFI и скрывает содержимое загрузочного раздела от остальной части системы. Спецификация UEFI определяет некоторые глобальные переменные, которые могут находиться в энергонезависимой оперативной памяти системы NVRAM и быть доступны даже на этапе выполнения, когда ОС получила полный контроль над платформой. Некоторые другие переменные UEFI могут находиться в оперативной памяти системы. Программа установки конфигурирует платформу UEFI для загрузки диспетчера загрузки с помощью настроек переменных UEFI Boot000*X*, где *X* — уникальное число, зависящее от номера опции загрузки, и BootOrder. Когда система перезагружается после завершения установки, диспетчер загрузки UEFI автоматически может выполнять код диспетчера загрузки.

В табл. 12.1 приведены файлы, участвующие в процессе загрузки UEFI. На рис. 12.2 показан пример разметки жесткого диска по схеме раздела GPT. Файлы, расположенные в загрузочном разделе Windows, хранятся в каталоге `\Windows\System32`.

Таблица 12.1. Компоненты процесса загрузки UEFI

Компонент	Обязанности	Расположение
bootmgfw.efi	Считывает базу данных конфигурации загрузки BCD, если требуется, представляет меню загрузки и позволяет выполнять предзагрузочные программы, такие как приложение Memory Test (Memtest.efi)	Системный раздел EFI
Winload.efi	Загружает Ntoskrnl.exe, его зависимости (SiPolicy.p7b, hvloader.dll, hvix64.exe, Hal.dll, Kdcom.dll, Ci.dll, Clfs.sys, Pshed.dll) и драйверы устройств начальной загрузки	Загрузочный раздел Windows

Компонент	Обязанности	Расположение
Winresume.efi	При возобновлении работы после спящего режима вместо обычной загрузки возобновляется из файла Hiberfil.sys, созданного при уходе в спящий режим	Загрузочный раздел Windows
Memtest.efi	При выборе из меню Boot Immersive или из диспетчера загрузки запускается и предоставляет графический интерфейс для сканирования памяти и обнаружения поврежденной оперативной памяти	Системный раздел EFI
Hvloader.dll	Если этот модуль обнаружен диспетчером загрузки и правильно включен, то он является программой запуска гипервизора (hvloader.efi в предыдущей версии Windows)	Загрузочный раздел Windows
Hvix64.exe или hvax64.exe	Гипервизор Windows (Hyper-V). В зависимости от архитектуры процессора этот файл может называться по-разному. Это основной компонент системы безопасности на основе виртуализации (VBS)	Загрузочный раздел Windows
Ntoskrnl.exe	Инициализирует исполнительные подсистемы и драйверы устройств загрузки и запуска системы, подготавливает ее к запуску собственных приложений и запускает Smss.exe	Загрузочный раздел Windows
Securekernel.exe	Защищенное ядро Windows. Предоставляет службы режима ядра для безопасной среды VTL 1 и некоторые базовые средства связи с обычной средой (см. главу 9)	Загрузочный раздел Windows
Hal.dll	DLL режима ядра, обеспечивающая взаимодействие Ntoskrnl и драйверов с аппаратным обеспечением. Также выступает в роли драйвера для материнской платы, поддерживая компоненты, которые не управляются другим драйвером	Загрузочный раздел Windows
Smss.exe	Начальный экземпляр запускает свою копию для инициализации каждого сеанса. Экземпляр сеанса 0 загружает драйвер подсистемы Windows Win32k.sys и запускает процесс подсистемы Windows Csrss.exe и процесс инициализации Windows Wininit.exe. Все остальные экземпляры для каждой сессии запускают процессы Csrss и Winlogon	Загрузочный раздел Windows
Wininit.exe	Запускает диспетчер управления службами SCM, процесс Local Security Authority (LSASS) и диспетчер локальных сеансов LSM. Инициализирует остальную часть реестра и выполняет задачи инициализации пользовательского режима	Загрузочный раздел Windows
Winlogon.exe	Координирует вход в систему и безопасность пользователей, запускает Bootim и LogonUI	Загрузочный раздел Windows
Logonui.exe	Представляет интерактивный диалог входа в систему	Загрузочный раздел Windows
Bootim.exe	Представляет графическое интерактивное меню загрузки	Загрузочный раздел Windows

Продолжение ⇩

Таблица 12.1 *(продолжение)*

Компонент	Обязанности	Расположение
Services.exe	Загружает и инициализирует драйверы устройств с автоматическим запуском и службы Windows	Загрузочный раздел Windows
TcbLaunch.exe	Организует безопасный запуск операционной системы в системе, поддерживающей новую технологию Intel TXT	Загрузочный раздел Windows
TcbLoader.dll	Содержит код загрузчика Windows, который запускается в контексте безопасного запуска	Загрузочный раздел Windows

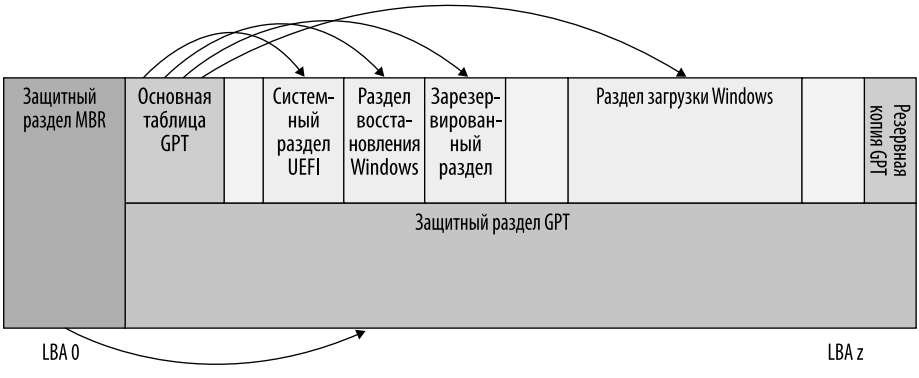


Рис. 12.2. Примерная схема жесткого диска UEFI

Еще одна роль Setup заключается в подготовке BCD, который в системах UEFI хранится в файле `\EFI\Microsoft\Boot\BCD` в корневом каталоге системного тома. Этот файл содержит параметры для запуска версии Windows, которую устанавливает Setup, и всех предыдущих установок Windows. Если файл BCD уже существует, Setup просто добавляет записи, относящиеся к новой установке. Дополнительные сведения о BCD см. в главе 10.

Все спецификации UEFI, включающие фазы PEI и BDS, безопасную загрузку и многие другие концепции, доступны на сайте https://uefi.org/specifications.

Процесс загрузки BIOS

Из-за недостатка места в этом издании не рассматривается старый процесс загрузки BIOS. Полное описание процесса предварительной и основной загрузки BIOS приводится в томе 2 предыдущего издания книги.

Безопасная загрузка

Как говорилось в главе 7, Windows была разработана для защиты от вредоносного программного обеспечения. Все старые системы BIOS были уязвимы для развитых постоянных угроз (Advanced Persistent Threats, APT), которые использовали буткит для скрытного выполнения кода. Буткит — это особый тип вредоносного программного обеспечения, которое запускается до диспетчера загрузки Windows

и позволяет основному модулю инфекции работать, не будучи обнаруженным антивирусными программами. Начальные части буткита BIOS обычно располагаются в секторе Master Boot Record (MBR) или Volume Boot Record (VBR) системного жесткого диска. Поэтому старые системы BIOS при включении выполняют код буткита вместо основного кода ОС. Первоначальный загрузочный код ОС шифруется и хранится в других областях жесткого диска и обычно выполняется вредоносным кодом на более поздней стадии. Этот тип буткита даже способен изменить код ОС в памяти на любом этапе загрузки Windows.

Как обнаружили специалисты по информационной безопасности, первые выпуски спецификации UEFI все еще были уязвимы для этой проблемы, поскольку прошивка, загрузчик и другие компоненты не проверялись. Поэтому злоумышленник, получивший доступ к компьютеру, мог подделать эти компоненты и заменить загрузчик на вредоносный. Действительно, для загрузки системы могло использоваться любое корректно зарегистрированное в относительной переменной загрузки EFI-приложение — портативный или TE (terse executable) исполняемый файл. Более того, даже DXE-драйверы не проверялись правильно, что позволяло внедрить вредоносный EFI-драйвер в SPI-флеш-память. Windows не могла правильно идентифицировать изменение процесса загрузки.

Эта проблема заставила консорциум UEFI разработать и внедрить технологию безопасной загрузки. Secure Boot — это функция UEFI, которая гарантирует, что каждый загружаемый компонент имеет цифровую подпись и подтверждение. Secure Boot гарантирует, что при загрузке компьютера применяется только программное обеспечение, которому доверяет производитель компьютера или пользователь. В Secure Boot прошивка отвечает за проверку всех компонентов — DXE-драйверов, диспетчеров загрузки UEFI, загрузчиков и т. д. — перед их загрузкой. Если какой-либо компонент не проходит проверку, пользователю выдается сообщение об ошибке, а процесс загрузки прерывается.

Проверка с помощью алгоритмов с открытым ключом, например RSA, для цифровой подписи осуществляется по базе данных принятых и непринятых сертификатов или хешей, содержащейся в прошивке UEFI. В таких алгоритмах используются два разных ключа.

- Открытый ключ применяется для *расшифровки* зашифрованного дайджеста — так называется хеш двоичных данных исполняемого файла. Этот ключ хранится в цифровой подписи файла.
- Закрытый ключ задействуется для *шифрования* хеша двоичного исполняемого файла и хранится в надежном секретном месте. Цифровая подпись исполняемого файла состоит из следующих этапов.
  - Вычисление дайджеста содержимого файла с помощью алгоритма сильного хеширования, например SHA256. Сильное хеширование должно создавать дайджест сообщения, который является уникальным и относительно небольшим представлением полных исходных данных — чем-то вроде сложной контрольной суммы. Алгоритмы хеширования — это одностороннее шифрование, то есть из дайджеста невозможно получить весь файл.
  - Шифрование вычисленного дайджеста с помощью закрытой части ключа.
  - Хранение зашифрованного дайджеста, открытой части ключа и названия хеширующего алгоритма в цифровой подписи файла.

При этом подходе система, когда хочет проверить целостность файла, пересчитывает хеш файла и сравнивает его с дайджестом, который был расшифрован из цифровой подписи. Никто, кроме владельца закрытого ключа, не может изменить зашифрованный дайджест, хранящийся в цифровой подписи.

Эта упрощенная модель может быть расширена до создания цепочки сертификатов, каждому из которых прошивка доверяет. Действительно, если открытый ключ, находящийся в определенном сертификате, неизвестен прошивке, но доверенное лицо вторично подписало сертификат промежуточным или корневым сертификатом, то прошивка может предположить, что даже внутренний открытый ключ должен считаться доверенным. Этот механизм (рис. 12.3) называется *цепочкой доверия*. Он основан на том, что цифровой сертификат, используемый для подписи кода, может быть подписан с помощью открытого ключа другого доверенного сертификата более высокого уровня — корневого или промежуточного сертификата. Модель здесь упрощена, поскольку полное описание всех деталей выходит за рамки тематики этой книги.

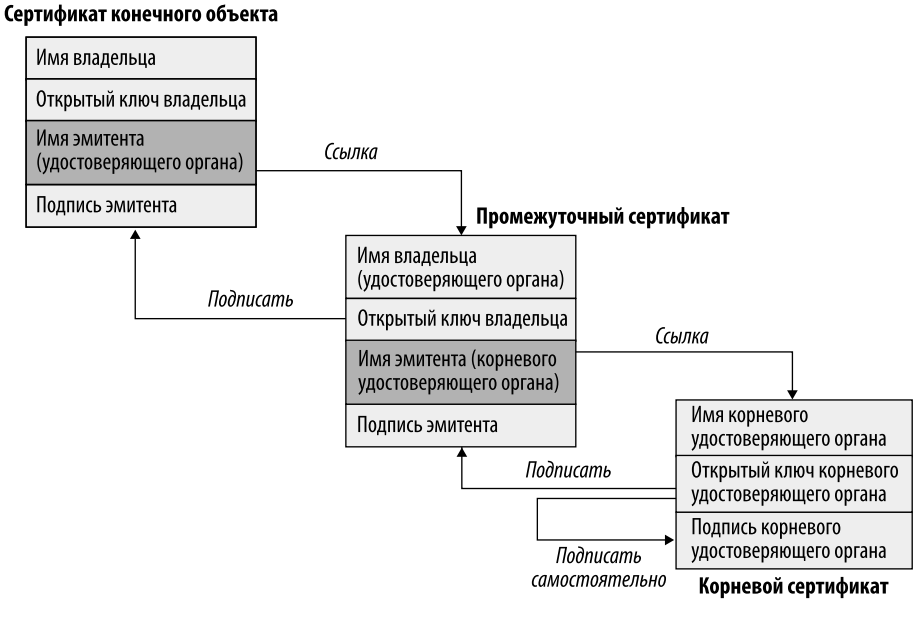


Рис. 12.3. Упрощенное представление цепочки доверия

Разрешенные и отозванные сертификаты и хеши UEFI должны установить некоторую иерархию доверия с помощью объектов (рис. 12.4), которые хранятся в переменных UEFI.

- **Ключ платформы (platform key, PK).** Представляет собой корень доверия и используется для защиты базы данных ключей обмена ключами. Поставщик платформы помещает открытую часть PK в прошивку UEFI во время производства. Его закрытая часть остается у производителя.

- **Ключ обмена ключами (key exchange key, KEK).** База данных обмена ключами содержит доверенные сертификаты, которым разрешено изменять базу данных разрешенных подписей (DB), базу данных запрещенных подписей (DBX) и базу данных подписей временных меток (DBT). База данных KEK обычно содержит сертификаты поставщика операционной системы (operating system vendor, OSV) и защищена PK.

Хеши и подписи, используемые для проверки загрузчиков и других предзагрузочных компонентов, хранятся в трех различных базах данных. База данных разрешенных подписей содержит хеши конкретных двоичных файлов или сертификатов либо их хешей, применявшихся для создания сертификатов подписи кода, которыми были подписаны загрузчик и другие предзагрузочные компоненты в соответствии с моделью цепочки доверия. База данных запрещенных подписей содержит хеши конкретных двоичных файлов или сертификатов либо их хешей, которые были скомпрометированы или отозваны. База данных подписей временных меток содержит сертификаты временных меток, используемые при подписании образов загрузчика. Все три базы данных заблокированы через KEK от редактирования.

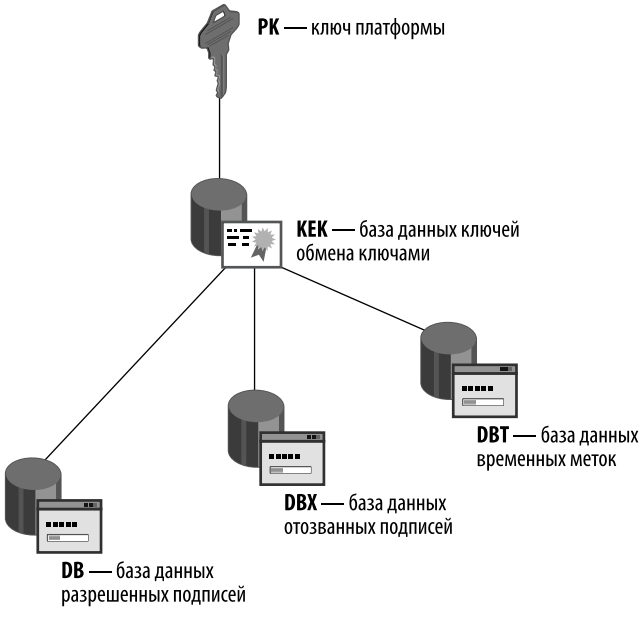


Рис. 12.4. Цепочка доверия, используемая в UEFI Secure Boot

Чтобы правильно зафиксировать ключи Secure Boot, прошивка не должна разрешать их обновление, пока пытающийся их обновить объект не докажет с помощью цифровой подписи на определенном пакете, называемом *дескриптором аутентификации*, что он обладает закрытой частью ключа, использованного для создания переменной. Этот механизм реализован в UEFI с помощью аутентифицированных переменных. На момент написания этой книги спецификации UEFI

допускают только два типа ключей подписи: X509 и RSA2048. Аутентифицированная переменная может быть очищена путем записи пустого обновления, которое все еще должно содержать действительный дескриптор аутентификации. Когда аутентифицированная переменная создается, она хранит как открытую часть ключа, который ее создал, так и начальное значение времени или монотонный отсчет и будет принимать только последующие обновления, подписанные этим ключом и имеющие тот же тип обновления. Например, переменная KEK создается с помощью PK и может быть обновлена только дескриптором аутентификации, подписанным этим PK.

ПРИМЕЧАНИЕ Способ, которым прошивка UEFI использует аутентифицированные переменные в средах Secure Boot, может привести к некоторой путанице. Действительно, только базы данных PK, KEK и подписей хранятся с помощью аутентифицированных переменных. Остальные переменные загрузки UEFI, в которых хранятся данные конфигурации загрузки, остаются обычными переменными времени исполнения. Это означает, что в среде Secure Boot пользователь все еще легко может обновлять или изменять конфигурацию загрузки и даже порядок загрузки. Это не проблема, поскольку безопасная проверка всегда производится на любом типе загрузочного приложения независимо от его источника или порядка. Secure Boot не предназначен для предотвращения изменения конфигурации загрузки системы.

Диспетчер загрузки Windows

Как уже говорилось, прошивка UEFI считывает и выполняет диспетчер загрузки Windows Bootmgfw.efi. Прошивка EFI передает управление Bootmgr в длинном режиме с включенной подкачкой, а пространство памяти, определенное картой памяти UEFI, отображается взаимно однозначно. Таким образом, в отличие от систем wBIOS, здесь нет необходимости переключать контекст выполнения. Диспетчер загрузки действительно является первым приложением, которое вызывается при запуске или возобновлении работы Windows из полностью выключенного состояния питания или спящего режима (состояние питания S4). Диспетчер загрузки был полностью переработан начиная с Windows Vista, чтобы:

- поддержать загрузку различных операционных систем, использующих сложные и разнообразные технологии загрузки;
- разделить коды запуска, специфичного для ОС, на отдельное загрузочное приложение Windows Loader и приложение Resume (Winresume);
- изолировать общие загрузочные сервисы и предоставить их загрузочным приложениям. Эта роль возложена на загрузочные библиотеки.

Несмотря на то что конечная цель диспетчера загрузки кажется очевидной, его архитектура сложна. С этого момента используется термин *«загрузочное приложение»* для обозначения любого загрузчика операционной системы, например загрузчика Windows и др. У Bootmgr есть несколько ролей.

- Он инициализирует журналирование загрузки и основные системные службы, необходимые для работы загрузочного приложения, которые будут рассмотрены далее в этом разделе.

- Инициализирует функции безопасности, такие как Secure Boot и Measured Boot, загружает их системные политики и проверяет собственную целостность.
- Находит, открывает и считывает хранилище данных конфигурации загрузки.
- Создает список загрузки и показывает базовое меню загрузки, если политика меню загрузки установлена как Legacy.
- Управляет TPM и разблокировкой зашифрованных BitLocker дисков, показывая экран разблокировки BitLocker и предоставляя метод восстановления в случае проблем с получением ключа расшифровки.
- Запускает специальное загрузочное приложение и управляет последовательностью восстановления в случае неудачной загрузки (среда восстановления Windows).

Одни из первых выполняемых действий — настройка средства журналирования загрузки и инициализация загрузочных библиотек. Приложения загрузки включают в себя стандартный набор библиотек, которые инициализируются при запуске диспетчера загрузки. После инициализации стандартных библиотек их основные службы становятся доступными для всех загрузочных приложений. Эти службы включают базовый диспетчер памяти (поддерживает трансляцию адресов и выделение страниц и кучи), параметры прошивки (например, записи о загрузочном устройстве и диспетчере загрузки в BCD), систему уведомления о событиях для измеренной загрузки, время, журналирование загрузки, криптографические модули, модуль доверенной платформы TPM (Trusted Module Platform), сеть, драйвер дисплея, систему ввода-вывода и базовый загрузчик PE. Читатель может представить себе загрузочные библиотеки как особый вид базового уровня аппаратной абстракции HAL (Hardware Abstraction Layer) для диспетчера загрузки и загрузочных приложений. На ранних этапах инициализации библиотеки инициализируется компонент загрузочной библиотеки System Integrity. Цель службы System Integrity — предоставить платформу для сообщения и регистрации системных событий, имеющих отношение к безопасности, таких как загрузка нового кода, присоединение отладчика и т. д. Это достигается с помощью функциональности, предоставляемой TPM, и используется, в частности, для измеренной загрузки. Эта функция будет описана далее в разделе «Измеренная загрузка».

Для правильного выполнения функции инициализации диспетчера загрузки *BmMain* необходима структура данных Application Parameters, которая, как следует из названия, описывает параметры запуска, например устройство загрузки, GUID объекта BCD и т. д. Для составления этой структуры данных диспетчер загрузки применяет службы прошивки EFI с целью получения полного относительного пути к собственному исполняемому файлу и параметров начальной загрузки, хранящихся в активной переменной загрузки EFI BOOT000X. Согласно спецификациям EFI, переменная загрузки EFI должна содержать краткое описание загрузочной записи, полный путь к устройству и файлу диспетчера загрузки, а также некоторые необязательные данные. Windows использует необязательные данные для хранения GUID объекта BCD, который описывает сам себя.

ПРИМЕЧАНИЕ Дополнительные данные могут включать любые другие параметры загрузки, которые диспетчер загрузки будет разбирать на более поздних этапах. Это позволяет настраивать диспетчер загрузки по переменным UEFI без использования реестра Windows.

ЭКСПЕРИМЕНТ. Исследуем переменные загрузки UEFI

Можно использовать утилиту UefiTool, которая включена в интернет-ресурсы этой книги, для создания дампа всех загрузочных переменных UEFI системы. Для этого просто запустите утилиту в административной командной строке и укажите параметр командной строки `/enum`. Можно запустить командную строку от имени администратора, найдя `cmd` в поисковой строке Cortana и после щелчка правой кнопкой мыши на Командная строка (Command Prompt) выбрав Запустить от имени администратора (Run As Administrator). В обычной системе используется множество переменных UEFI. Инструмент поддерживает фильтрацию всех переменных по имени и GUID. Можно даже экспортировать все имена переменных и данные в текстовый файл с помощью параметра `/out`.

Начните с дампа всех переменных UEFI в текстовый файл:

```
C:\Tools>UefiTool.exe /enum /out Uefi_Variables.txt
UEFI Dump Tool v0.1
Copyright 2018 by Andrea Allievi (AaLl86)

Firmware type: UEFI
Bitlocker enabled for System Volume: NO

Successfully written "Uefi_Variables.txt" file.
```

Список переменных загрузки UEFI можно получить с помощью следующего фильтра:

```
C:\Tools>UefiTool.exe /enum Boot
UEFI Dump Tool v0.1
Copyright 2018 by Andrea Allievi (AaLl86)

Firmware type: UEFI
Bitlocker enabled for System Volume: NO

EFI Variable  "BootCurrent"
   Guid      : {8BE4DF61-93CA-11D2-AA0D-00E098032B8C}
   Attributes: 0x06 ( BS RT )
   Data size : 2 bytes
   Data:
   00 00                                            |

EFI Variable  "Boot0002"
   Guid      : {8BE4DF61-93CA-11D2-AA0D-00E098032B8C}
   Attributes: 0x07 ( NV BS RT )
   Data size : 78 bytes
   Data:
   01 00 00 00 2C 00 55 00 53 00 42 00 20 00 53 00 |    , U S B   S
   74 00 6F 00 72 00 61 00 67 00 65 00 00 00 04 07 | t o r a g e
   14 00 67 D5 81 A8 B0 6C EE 4E 84 35 2E 72 D3 3E | g ü¿ l Nä5.r >
   45 B5 04 06 14 00 71 00 67 50 8F 47 E7 4B AD 13 | E q gPÅG K¡
   87 54 F3 79 C6 2F 7F FF 04 00 55 53 42 00       | çT≤y /   USB

EFI Variable "Boot0000"
   Guid      : {8BE4DF61-93CA-11D2-AA0D-00E098032B8C}
   Attributes: 0x07 ( NV BS RT )
   Data size : 300 bytes
```

```
   Data:
   01 00 00 00 74 00 57 00 69 00 6E 00 64 00 6F 00 |    t W I n d o
   77 00 73 00 20 00 42 00 6F 00 6F 00 74 00 20 00 | w s   B o o t
   4D 00 61 00 6E 00 61 00 67 00 65 00 72 00 00 00 | M a n a g e r
   04 01 2A 00 02 00 00 00 00 A0 0F 00 00 00 00 00 | * á
   00 98 0F 00 00 00 00 00 84 C4 AF 4D 52 3B 80 44 |  ÿ     ä "MR;ÇD
   98 DF 2C A4 93 AB 30 B0 02 02 04 04 46 00 5C 00 | ÿ ,ñô¼0 F \
   45 00 46 00 49 00 5C 00 4D 00 69 00 63 00 72 00 | E F I \ M i c r
   6F 00 73 00 6F 00 66 00 74 00 5C 00 42 00 6F 00 | o s o f t \ B o
   6F 00 74 00 5C 00 62 00 6F 00 6F 00 74 00 6D 00 | o t \ b o o t m
   67 00 66 00 77 00 2E 00 65 00 66 00 69 00 00 00 | g f w . e f i
   7F FF 04 00 57 49 4E 44 4F 57 53 00 01 00 00 00 |    WINDOWS
   88 00 00 00 78 00 00 00 42 00 43 00 44 00 4F 00 | ê   x   B C D O
   42 00 4A 00 45 00 43 00 54 00 3D 00 7B 00 39 00 | B J E C T = { 9
   64 00 65 00 61 00 38 00 36 00 32 00 63 00 2D 00 | d e a 8 6 2 c -
   35 00 63 00 64 00 64 00 2D 00 34 00 65 00 37 00 | 5 c d d – 4 e 7
   30 00 2D 00 61 00 63 00 63 00 31 00 2D 00 66 00 | 0 – a c c 1 – f
   33 00 32 00 62 00 33 00 34 00 34 00 64 00 34 00 | 3 2 b 3 4 4 d 4
   37 00 39 00 35 00 7D 00 00 00 6F 00 01 00 00 00 | 7 9 5 }   o
   10 00 00 00 04 00 00 00 7F FF 04 00             |

EFI Variable "BootOrder"
   Guid      : {8BE4DF61-93CA-11D2-AA0D-00E098032B8C}
   Attributes: 0x07 ( NV BS RT )
   Data size : 8 bytes
   Data:
   02 00 00 00 01 00 03 00 |
<Полный вывод сокращен по соображениям экономии места>
```

Утилита может даже интерпретировать содержимое каждой переменной загрузки. Для этого используется параметр `/enumboot`:

```
C:\Tools>UefiTool.exe /enumboot
UEFI Dump Tool v0.1
Copyright 2018 by Andrea Allievi (AaLl86)

Firmware type: UEFI
Bitlocker enabled for System Volume: NO

System Boot Configuration
   Number of the Boot entries: 4
   Current active entry: 0
   Order: 2, 0, 1, 3

Boot Entry #2
   Type: Active
   Description: USB Storage

Boot Entry #0
   Type: Active
   Description: Windows Boot Manager
   Path: Harddisk0\Partition2 [LBA: 0xFA000]\\EFI\Microsoft\Boot\bootmgfw.efi
   OS Boot Options: BCDOBJECT={9dea862c-5cdd-4e70-acc1-f32b344d4795}

Boot Entry #1
   Type: Active
   Description: Internal Storage

Boot Entry #3
   Type: Active
   Description: PXE Network
```

Если инструмент способен разобрать путь загрузки, то он выводит относительную строку пути. То же самое относится к опциям загрузки ОС Winload. Спецификации UEFI определяют различные интерпретации поля пути в загрузочной записи, которые зависят от аппаратного интерфейса. Можно изменить порядок загрузки системы, просто установив значение переменной BootOrder или воспользовавшись параметром командной строки `/setbootorder`. Имейте в виду, что это может сделать недействительным главный ключ тома BitLocker (эта концепция объясняется далее в разделе «Измеренная загрузка»):

```
C:\Tools>UefiTool.exe /setvar bootorder {8BE4DF61-93CA-11D2-AA0D-00E098032B8C}
0300020000000100
UEFI Dump Tool v0.1
Copyright 2018 by Andrea Allievi (AaLl86)

Firmware type: UEFI
Bitlocker enabled for System Volume: YES

Warning, The "bootorder" firmware variable already exist.
Overwriting it could potentially invalidate the system Bitlocker Volume Master
Key.
Make sure that you have made a copy of the System volume Recovery Key.
Are you really sure that you would like to continue and overwrite its content?
[Y/N] y
The "bootorder" firmware variable has been successfully written.
```

После создания структуры данных Application Parameters и получения всех путей загрузки (\EFI\Microsoft\Boot — основной рабочий каталог) диспетчер загрузки открывает и анализирует файл конфигурационных данных загрузки BCD (Boot Configuration Data). Этот файл является кустом реестра, который содержит все дескрипторы загрузочных приложений и обычно отображается в виртуальном разделе HKLM\BCD00000000 после полного запуска системы. Диспетчер загрузки использует библиотеку загрузки для открытия и чтения файла BCD. Библиотека задействует службы EFI для чтения физических секторов с жесткого диска и их записи и на момент создания книги реализует облегченную версию различных файловых систем, таких как NTFS, FAT, ExFAT, UDFS, El Torito, а также виртуальные файловые системы, поддерживающие сетевой загрузочный ввод-вывод, ввод-вывод VMBus для виртуальных машин Hyper-V и ввод-вывод образов WIM. Разбирается куст конфигурационных данных загрузки, находится описывающий диспетчер загрузки объект BCD с помощью его GUID, и все записи, представляющие аргументы загрузки, добавляются в раздел запуска структуры данных Application Parameters. Записи в BCD могут включать необязательные аргументы, которые интерпретируются Bootmgr, Winload и другими компонентами, участвующими в загрузке. В табл. 12.2 содержится список этих опций и описано их влияние на Bootmgr, в табл. 12.3 показан список опций BCD, доступных для всех загрузочных приложений, а в табл. 12.4 — опции BCD для загрузчика Windows. В табл. 12.5 показаны опции BCD, управляющие выполнением гипервизора.

Таблица 12.2. Параметры BCD для диспетчера загрузки Bootmgr

Читаемое имя	Значение	Код элемента BCD[1]	Назначение
bcdfilepath	Путь	BCD_FILEPATH	Указывает на файл BCD (обычно \Boot\BCD) на диске
displaybootmenu	Бинарное	DISPLAY_BOOT_MENU	Определяет, будет ли диспетчер загрузки показывать меню загрузки или автоматически выбирать пункт по умолчанию
noerrordisplay	Бинарное	NO_ERROR_DISPLAY	Заглушает вывод сообщений об ошибках, с которыми столкнулся диспетчер загрузки
resume	Бинарное	ATTEMPT_RESUME	Указывает, следует ли пытаться возобновить работу из спящего режима. Этот параметр автоматически устанавливается при переходе Windows в спящий режим
timeout	Секунды	TIMEOUT	Количество секунд, в течение которых диспетчер загрузки должен подождать, прежде чем выбрать запись по умолчанию
resumeobject	GUID	RESUME_OBJECT	Идентификатор того, какое загрузочное приложение должно использоваться для возобновления работы системы после спящего режима
displayorder	Список	DISPLAY_ORDER	Определение порядка показа элементов списка диспетчера загрузки
toolsdisplay order	Список	TOOL_DISPLAY_ ORDER	Определение порядка показа элементов списка инструментов диспетчера загрузки
bootsequence	Список	BOOT_SEQUENCE	Определение последовательности однократной загрузки
default	GUID	DEFAULT_OBJECT	Загрузочная запись по умолчанию для запуска
customactions	Список	CUSTOM_ACTIONS_ LIST	Определение настраиваемых действий, которые будут выполняться при вводе определенной последовательности клавиш
processcustomac-tionsfirst	Бинарное	PROCESS_CUSTOM_ ACTIONS_FIRST	Указывает, должен ли диспетчер загрузки запускать настраиваемые действия до начала загрузки
bcddevice	GUID	BCD_DEVICE	Идентификатор устройства, на котором расположено хранилище BCD

Продолжение ⇩

[1] Все коды элементов BCD диспетчера загрузки Windows начинаются с BCDE_BOOTMGR_TYPE, но эта часть опущена для экономии места.

Таблица 12.2 *(продолжение)*

Читаемое имя	Значение	Код элемента BCD[1]	Назначение
hiberboot	Бинарное	HIBERBOOT	Указывает, была ли эта загрузка гибридной
fverecoveryurl	Текстовая строка	FVE_RECOVERY_URL	URL восстановления BitLocker
fverecoverymessage	Текстовая строка	FVE_RECOVERY_MESSAGE	Сообщение о восстановлении BitLocker
flightedbootmgr	Бинарное	BOOT_FLIGHT_BOOTMGR	Указывает, должно ли выполнение происходить через предварительную версию Bootmgr

Таблица 12.3. Параметры библиотеки BCD для загрузочных приложений (действительны для всех типов объектов)

Читаемое имя	Значение	Код элемента[1]	Назначение
advancedoptions	Бинарное	DISPLAY_ADVANCED_OPTIONS	Если равно false, то по умолчанию при неудачной загрузке запускается загрузочный элемент с командой автовосстановления, в противном случае отображается ошибка загрузки и пользователю предлагается меню дополнительных параметров загрузки, связанных с этим загрузочным элементом. Эквивалентно нажатию клавиши F8
avoidlowmemory	Целое число	AVOID_LOW_PHYSICAL_MEMORY	Заставляет загрузчик по возможности избегать физических адресов ниже указанного значения. Иногда требуется на устаревших устройствах, например ISA, где доступны или видны только первые 16 Мбайт памяти
badmemoryaccess	Бинарное	ALLOW_BAD_MEMORY_ACCESS	Принудительное использование страниц памяти из списка плохих страниц. (Более подробную информацию о списках страниц см. в главе 5)
badmemorylist	Массив номеров кадров страниц (PFN)	BAD_MEMORY_LIST	Указывает список физических страниц в системе, про которые известно, что они повреждены из-за неисправной оперативной памяти
baodrate	Скорость передачи данных, бит/с	DEBUGGER_BAUDRATE	Задает переопределение скорости передачи данных по умолчанию (19200), с которой удаленный хост отладчика ядра будет подключаться через последовательный порт

[1] Все коды элементов BCD для Boot Applications начинаются с BCDE_LIBRARY_TYPE, но эта часть опущена для экономии места.

Читаемое имя	Значение	Код элемента[1]	Назначение
bootdebug	Бинарное	DEBUGGER_ ENABLED	Включает удаленную отладку загрузки для загрузчика. Если эта опция активна, можно использовать Kd.exe или Windbg.exe для подключения к загрузчику
bootems	Бинарное	EMS_ENABLED	Заставляет Windows включить Emergency Management Services (EMS) для загрузочных приложений, чтобы передавать информацию о загрузке и принимать команды управления системой через последовательный порт
busparams	Текстовая строка	DEBUGGER_BUS_ PARAMETERS	Если для отладки ядра используется физическое отладочное устройство PCI, то указывает шину PCI, функцию и номер или индекс таблицы ACPI DBG для устройства
channel	Номер канала от 0 до 62	DEBUGGER_1394_ CHANNEL	Применяется в сочетании с <debugtype> 1394 для указания канала IEEE 1394, по которому будет идти общение при отладке ядра
configaccesspolicy	Default, DisallowMmConfig	CONFIG_ACCESS_ POLICY	Определяет, будет ли система использовать отображаемый в памяти ввод-вывод для доступа к конфигурационному пространству PCI-разработчика или вернется к применению процедур HAL доступа к портам ввода-вывода. Иногда может быть полезен для решения проблем с платформенными устройствами
debugaddress	Адрес оборудования	DEBUGGER_ PORT_ADDRESS	Указывает аппаратный адрес последовательного (COM) порта, используемого для отладки
debugport	Номер COM-порта	DEBUGGER_ PORT_NUMBER	Определяет переопределение последовательного порта по умолчанию (обычно COM2 в системах как минимум с двумя последовательными портами), к которому подключается удаленный хост отладчика ядра
debugstart	Active, AutoEnable, Disable	DEBUGGER_ START_POLICY	Определяет настройки отладчика при включенной отладке ядра. AutoEnable активирует отладчик при возникновении точки останова или исключения ядра, включая сбои ядра

Продолжение ⇩

Таблица 12.3 *(продолжение)*

Читаемое имя	Значение	Код элемента[1]	Назначение
debugtype	Serial, 1394, USB, Net	DEBUGGER_TYPE	Указывает, через какой порт будет осуществляться отладка ядра: последовательный, FireWire (IEEE 1394), USB или Ethernet. По умолчанию используется последовательный порт
hostip	IP-адрес	DEBUGGER_NET_HOST_IP	Указывает целевой IP-адрес, к которому нужно подключиться, когда отладчик ядра включен через Ethernet
порт	Целое число	DEBUGGER_NET_PORT	Указывает номер целевого порта для подключения к нему при включении отладчика ядра через Ethernet
key	Текстовая строка	DEBBUGGER_NET_KEY	Указывает ключ шифрования, применяемый для шифрования пакетов отладчика при использовании отладчика ядра через Ethernet
emsbaudrate	Скорость передачи данных, бит/с	EMS_BAUDRATE	Указывает скорость передачи данных, используемую для EMS
emsport	Номер COM-порта	EMS_PORT_NUMBER	Указывает последовательный (COM) порт, который будет применяться для EMS
extendedinput	Бинарное	CONSOLE_EXTENDED_INPUT	Позволяет загрузочным приложениям использовать возможности BIOS для расширенного консольного ввода
keyringaddress	Физический адрес	FVE_KEYRING_ADDRESS	Указывает физический адрес, по которому находится набор ключей BitLocker
firstmegabyte-policy	UseNone, UseAll, UsePrivate	FIRST_MEGABYTE_POLICY	Определяет, как младший 1 Мбайт физической памяти используется HAL, чтобы уменьшить повреждения от BIOS при управлении питанием
fontpath	Текстовая строка	FONT_PATH	Указывает путь к OEM-шрифту, который должен применяться загрузочным приложением
graphicsmodedisabled	Бинарное	GRAPHICS_MODE_DISABLED	Отключает графический режим для загрузочных приложений
graphicsresulton	Разрешение	GRAPHICS_RESOLUTION	Устанавливает графическое разрешение для загрузочных приложений
initialconsoleinput	Бинарное	INITIAL_CONSOLE_INPUT	Задает начальный символ, который система вставляет в буфер ввода клавиатуры PC/AT
integrityservices	Default, Disable, Enable	SI_POLICY	Включает или отключает службы целостности кода, которые используются для подписи кода в режиме ядра. По умолчанию установлено значение Enable

Читаемое имя	Значение	Код элемента[1]	Назначение
locale	Строка локализации	PREFERRED_ LOCALE	Устанавливает локаль для загрузочного приложения, например EN-US
noumex	Бинарное	DEBUGGER_ IGNORE_ USERMODE_ EXCEPTIONS	Отключает исключения пользовательского режима при включенной отладке ядра. Если при загрузке в режиме отладки наблюдаются зависания системы, попробуйте включить эту опцию
recoveryenabled	Бинарное	AUTO_ RECOVERY_ ENABLED	Включает последовательность восстановления, если таковая имеется. Используется при новой установке Windows для показа PE-интерфейса запуска и восстановления
recoverysequence	Список	RECOVERY_ SEQUENCE	Определяет последовательность восстановления (описана ранее)
relocatephysical	Физический адрес	RELOCATE_ PHYSICAL_ MEMORY	Перемещает физическую память автоматически выбранного узла NUMA на указанный физический адрес
targetname	Текстовая строка	DEBUGGER_USB_ TARGETNAME	Определяет целевое имя для USB-отладчика при использовании отладки по USB2 или USB3, когда для параметра debugtype установлено значение USB
testsigning	Бинарное	ALLOW_ PRERELEASE_ SIGNATURES	Включает режим тестовой подписи, позволяющий разработчикам драйверов загружать локально подписанные 64-битные драйверы. Этот вариант приводит к показу предупреждения на Рабочем столе
truncatememory	Адрес, байт	TRUNCATE_ PHYSICAL_ MEMORY	Игнорирует физическую память выше указанного физического адреса

Таблица 12.4. Параметры BCD для загрузчика Winload

Элемент BCD	Значение	Код элемента[1]	Назначение
bootlog	Бинарное	LOG_ INITIALIZATION	Заставляет Windows записывать журнал загрузки в файл %SystemRoot%\ Ntbtlog.txt
bootstatuspolicy	DisplayAll-Failures, ignoreAllFailures, IgnoreShutdownFailures, IgnoreBootFailures	BOOT_STATUS_ POLICY	Переопределяет поведение системы по умолчанию, в котором пользователю предлагается меню загрузки для устранения неполадок, если система не завершила предыдущую загрузку или выключение

Продолжение ⇩

[1] Все коды элементов BCD для загрузчика ОС Windows начинаются с BCDE_OSLOADER_TYPE, но эта часть опущена для экономии места.

Таблица 12.4 *(продолжение)*

Элемент BCD	Значение	Код элемента[1]	Назначение
bootux	Disabled, Basic, Standard	BOOTUX_POLICY	Определяет графику загрузки, которую будет видеть пользователь. Disabled означает, что во время загрузки не будет видно никакой графики, только черный экран, Basic отображает только индикатор прогресса. Standard отображает обычную анимацию логотипа Windows во время загрузки
bootmenupolicy	Legacy, Standard	BOOT_MENU_POLICY	Тип загрузочного меню, которое будет показано при нескольких загрузочных записях (см. раздел «Меню загрузки» далее в этой главе)
clustermodeaddressing	Количество процессоров	CLUSTERMODE_ADDRESSING	Определяет максимальное количество процессоров, которые можно включить в один кластер Advanced Programmable Interrupt Controller (APIC)
configflags	Флаги	PROCESSOR_CONFIGURATION_FLAGS	Указывает флаги конфигурации, специфичные для процессора
dbgtransport	Имя образа транспорта	DBG_TRANSPORT_PATH	Отменяет использование одного из стандартных отладочных транспортов ядра (Kdcom.dll, Kd1394, Kdusb.dll) и вместо этого задействует указанный файл, позволяя применять специализированные отладочные транспорты, которые обычно не поддерживаются Windows
debug	Бинарное	KERNEL_DEBUGGER_ENABLED	Включает отладку в режиме ядра
detecthal	Бинарное	DETECT_KERNEL_AND_HAL	Включает динамическое обнаружение HAL
driverloadfailurepolicy	Fatal, UseErrorControl	DRIVER_LOAD_FAILURE_POLICY	Описывает поведение загрузчика при неудачной загрузке драйвера загрузки. Fatal предотвращает загрузку, а UseErrorControl заставляет систему применять стандартное поведение драйвера при ошибках, указанное в его служебном ключе
ems	Бинарное	KERNEL_EMS_ENABLED	Указывает ядру также использовать EMS. Если применяется только bootems, то лишь загрузчик будет задействовать EMS

Элемент BCD	Значение	Код элемента[1]	Назначение
evstore	Текстовая строка	EVSTORE	Сохраняет местоположение предварительно загруженной загрузочной ветки
groupaware	Бинарное	FORCE_GROUP_AWARENESS	Заставляет систему задействовать группы, отличные от нуля, при привязке сида группы к новым процессам. Используется только в 64-разрядных Windows
groupsize	Целое число	GROUP_SIZE	Устанавливает максимальное количество логических процессоров, которые могут входить в группу, — максимум 64. Может использоваться для принудительного создания групп в системе, которая обычно не требует их существования. Должно быть степенью двойки, применяется только в 64-разрядных Windows
hal	Имя образа HAL	HAL_PATH	Переопределяет имя файла по умолчанию для образа HAL (Hal.dll). Этот параметр может быть полезен при загрузке комбинации проверенного HAL и проверенного ядра. Требуется также указать элемент ядра
halbreakpoint	Бинарное	DEBUGGER_HAL_BREAKPOINT	Приводит к остановке HAL в точке останова в начале инициализации HAL. Первое, что делает ядро Windows при инициализации, — инициализация HAL, поэтому данная точка останова является самой ранней из возможных, если не используется отладка загрузки. Если этот переключатель применяется без ключа /DEBUG, то система выдаст синий экран со STOP-кодом 0x00000078 (PHASE0_EXCEPTION)
novesa	Бинарное	BCDE_OSLOADER_TYPE_DISABLE_VESA_BIOS	Отключает использование режимов дисплеев VESA
optionsedit	Бинарное	OPTIONS_EDIT_ONE_TIME	Включает редактор опций в диспетчере загрузки. С помощью этой опции Boot Manager позволяет пользователю интерактивно задавать желаемые параметры и переключатели командной строки для текущей загрузки. Эквивалентно нажатию клавиши F10

Продолжение ⇩

Таблица 12.4 *(продолжение)*

Элемент BCD	Значение	Код элемента[1]	Назначение
osdevice	GUID	OS_DEVICE	Указывает устройство, на котором установлена операционная система
pae	Default, ForceEnable, ForceDisable	PAE_POLICY	Значение Default позволяет загрузчику определить, поддерживает ли система PAE, и загрузить ядро PAE. ForceEnable делает это принудительно, а ForceDisable заставляет загрузчик загружать иную, не PAE, версию ядра Windows, даже если система определяется как поддерживающая x86 PAE и имеет более 4 Гбайт физической памяти. Однако в Windows 10 ядра x86 без PAE больше не поддерживаются
pciexpress	Default, ForceDisable	PCI_EXPRESS_ POLICY	Может использоваться для отключения поддержки шины и устройств PCI Express
perfmem	Размер, Мбайт	PERFORMANCE_ DATA_MEMORY	Размер буфера, выделяемого для журналирования данных о производительности. Эта опция действует аналогично элементу removememory, поскольку не позволяет Windows воспринимать размер, указанный как доступная память
quietboot	Бинарное	DISABLE_BOOT_ DISPLAY	Указывает Windows не инициализировать видеодрайвер VGA, отвечающий за отображение растровой графики в процессе загрузки. Этот драйвер применяется для отображения информации о ходе загрузки, поэтому его отключение лишает Windows возможности показывать эту информацию
ramdiskimage-length	Длина, байт	RAMDISK_IMAGE_ LENGTH	Указанный размер ramdisk
ramdiskima-geoffset	Смещение, байт	RAMDISK_IMAGE_ OFFSET	Если ramdisk содержит другие данные перед виртуальной файловой системой, например заголовок, то указывает загрузчику, откуда начинать чтение файла ramdisk
ramdiskdipath	Имя файла образа	RAMDISK_SDI_ PATH	Указывает имя загружаемого SDI ramdisk
ramdisktftp-blocksize	Размер блока	RAMDISK_TFTP_ BLOCK_SIZE	При загрузке WIM ramdisk с сетевого сервера Trivial FTP (TFTP) указывает размер используемого блока

Элемент BCD	Значение	Код элемента[1]	Назначение
ramdisktftpclientport	Номер порта	RAMDISK_TFTP_CLIENT_PORT	При загрузке WIM ramdisk с сетевого сервера TFTP указывает порт
ramdisktftpwindowsize	Размер окна	RAMDISK_TFTP_WINDOW_SIZE	При загрузке WIM ramdisk с сетевого сервера TFTP указывает размер используемого окна
removememory	Размер, байт	REMOVE_MEMORY	Указывает объем памяти, который не будет задействовать Windows
restrictapiccluster	Номер кластера	RESTRICT_APIC_CLUSTER	Определяет наибольший номер кластера APIC, используемого системой
resumeobject	GUID объекта	ASSOCIATED_RESUME_OBJECT	Описывает, какое приложение следует применять для возобновления работы из спящего режима (обычно это Winresume.exe)
safeboot	Minimal, Network, DsRepair	SAFEBOOT	Указывает параметры для загрузки в безопасном режиме. Minimal соответствует безопасному режиму без работы в сети, Network — безопасному режиму с работой в сети, а DsRepair — безопасному режиму с режимом восстановления служб каталогов. (См. раздел «Безопасный режим» далее в этой главе)
safebootalternateshell	Бинарное	SAFEBOOT_ALTERNATE_SHELL	Приказывает Windows использовать в качестве графической оболочки программу, указанную в параметре HKLM\SYSTEM\CurrentControlSet\Control\SafeBoot\AlternateShell, а не программу по умолчанию — Проводник Windows. Эта опция называется безопасным режимом с командной строкой в меню альтернативной загрузки
sos	Бинарное	SOS	Заставляет Windows показать список драйверов устройств, отмеченных для загрузки, а затем номер версии системы, включая номер сборки, объем физической памяти и количество процессоров
systemroot	Текстовая строка	SYSTEM_ROOT	Указывает путь относительно osdevice, по которому установлена операционная система
targetname	Имя	KERNEL_DEBUGGER_USB_TARGETNAME	Для отладки по USB присваивает имя отлаживаемой машине

Продолжение ⇩

Таблица 12.4 *(продолжение)*

Элемент BCD	Значение	Код элемента[1]	Назначение
tpmbootentropy	Defaut, ForceDisable, ForceEnable	TPM_BOOT_ ENTROPY_POLICY	Заставляет загрузчик выбирать определенную политику TPM Boot Entropy и передавать ее ядру. При использовании TPM Boot Entropy генератор случайных чисел ядра инициализируется данными, полученными от TPM, если тот присутствует
usefirmwarepci-settings	Бинарное	USE_FIRMWARE_ PCI_SETTINGS	Прекращает динамическое назначение Windows ресурсов IO/IRQ для PCI-устройств и оставляет для них настройку BIOS. (Дополнительные сведения см. в статье 148501 базы знаний Microsoft)
uselegacyapic-mode	Бинарное	USE_LEGACY_ APIC_MODE	Заставляет использовать базовую функциональность APIC, даже если чипсет сообщает о наличии расширенной функциональности APIC. Применяется в случаях аппаратных ошибок или несовместимости
usephysicaldes-tination	Бинарное	USE_PHYSICAL_ DESTINATION	Принуждает использовать APIC в режиме физического назначения
useplatform-clock	Бинарное	USE_PLATFORM_ CLOCK	Принуждает задействовать источник тактовых импульсов платформы в качестве счетчика производительности системы
vga	Бинарное	USE_VGA_DRIVER	Заставляет Windows использовать драйвер дисплея VGA вместо высокопроизводительного драйвера стороннего производителя
winpe	Бинарное	WINPE	Эта опция применяется в Windows PE и заставляет диспетчер конфигурации загружать ветвь SYSTEM реестра как ветвь, для которой изменения в памяти не сохраняются в образе ветви
x2apicpolicy	Disabled, Enabled, Default	X2APIC_POLICY	Указывает, следует ли использовать расширенную функциональность APIC, если чипсет ее поддерживает. Disabled эквивалентно установке uselegacyapicmode, a Enabled принудительно включает функциональность ACPI, даже если обнаружены ошибки. По умолчанию используются заявленные возможности чипсета, если нет ошибок

Элемент BCD	Значение	Код элемента[1]	Назначение
xsavepolicy	Целое число	XSAVEPOLICY	Принудительная загрузка заданной политики XSAVE из драйвера ресурсов политики XSAVE Hwpolicy.sys
xsaveaddfeature0-7	Целое число	XSAVEADDFEATURE0-7	Используется при тестировании поддержки XSAVE на современных процессорах Intel, позволяет симулировать наличие определенных возможностей процессора, когда на самом деле их нет. Это помогает увеличить размер структуры CONTEXT и подтверждает, что приложения корректно работают с расширенными функциями, которые могут появиться в дальнейшем. Однако никаких дополнительных функций не будет
xsaveremovefeature	Целое число	XSAVEREMOVEFEATURE	Заставляет не сообщать ядру о введенной функциональности XSAVE, даже если процессор ее поддерживает
xsaveprocessorsmask	Целое число	XSAVEPROCESSORSMASK	Битовая маска, указывающая, для каких процессоров применяется политика XSAVE
xsavedisable	Бинарное	XSAVEDISABLE	Отключает поддержку функциональности XSAVE, даже если процессор ее поддерживает

Таблица 12.5. Параметры BCD для загрузчика гипервизора hvloader

Элемент BCD	Значение	Код элемента[1]	Назначение
hypervisorlaunchtype	Off, Auto	HYPERVISOR_LAUNCH_TYPE	Разрешает загрузку гипервизора в системе Hyper-V или принудительно отключает его
hypervisordebug	Бинарное	HYPERVISOR_DEBUGGER_ENABLED	Включает или выключает отладчик гипервизора
hypervisordebugtype	Serial, 1394, None, Net	HYPERVISOR_DEBUGGER_TYPE	Указывает тип отладчика гипервизора — через последовательный порт, IEEE1394 или сетевой интерфейс
hypervisoriommupolicy	Default, Enable, Disable	HYPERVISOR_IOMMU_POLICY	Включает или отключает DMA Guard гипервизора — функции, которая блокирует прямой доступ к памяти (DMA) для всех портов PCI с горячим подключением до тех пор, пока пользователь не войдет в Windows

Продолжение ⇩

[1] Все коды элементов BCD для загрузчика гипервизора Windows начинаются с BCDE_OSLOADER_TYPE, но эта часть опущена для экономии места.

Таблица 12.5 *(продолжение)*

Элемент BCD	Значение	Код элемента[1]	Назначение
hypervisormsr-filterpolicy	Disable, Enable	HYPERVISOR_MSR_FILTER_POLICY	Контролирует, разрешен ли корневому разделу доступ к ограниченным MSR — регистрам, специфичным для конкретной модели
hypervisormmionxpolicy	Disable, Enable	HYPERVISOR_MMIO_NX_POLICY	Включает или отключает No Execute (NX) — защиту для областей памяти служебного кода и данных UEFI
hypervisorenforcedcodeintegrity	Disable, Enable, Strict	HYPERVISOR_ENFORCED_CODE_INTEGRITY	Включает или отключает Hypervisor Enforced Code Integrity (HVCI) — функцию, которая не позволяет ядру корневого раздела выделять неподписанные исполняемые страницы памяти
hypervisorschedulertype	Корень классического ядра	HYPERVISOR_SCHEDULER_TYPE	Указывает тип планировщика разделов гипервизора
hypervisordisableslat	Бинарное	HYPERVISOR_SLAT_DISABLED	Заставляет гипервизор игнорировать наличие функции трансляции адресов второго уровня (second layer address translation, SLAT), если она поддерживается процессором
hypervisornumproc	Целое число	HYPERVISOR_NUM_PROC	Указывает максимальное количество логических процессоров, доступных гипервизору
hypervisorrootprocpernode	Целое число	HYPERVISOR_ROOT_PROC_PER_NODE	Указывает общее количество корневых виртуальных процессоров на узел
hypervisorrootproc	Целое число	HYPERVISOR_ROOT_PROC	Указывает максимальное количество виртуальных процессоров в корневом разделе
hypervisorbaudrate	Скорость передачи данных в бит/с	HYPERVISOR_DEBUGGER_BAUDRATE	Если используется последовательная отладка гипервизора, то указывает применяемую скорость передачи данных
hypervisorchannel	Номер канала от 0 до 62	HYPERVISOR_DEBUGGER_1394_CHANNEL	Если используется отладка гипервизора через FireWire (IEEE 1394), то указывает номер канала
hypervisordebugport	Номер COM-порта	HYPERVISOR_DEBUGGER_PORT	Если задействуется последовательная отладка гипервизора, то указывает используемый COM-порт
hypervisoruselargevtlb	Бинарное	HYPERVISOR_USE_LARGE_VTLB	Позволяет гипервизору задействовать большее количество записей виртуального TLB

Элемент BCD	Значение	Код элемента[1]	Назначение
hypervisorhostip	IP-адрес (двоичный формат)	HYPERVISOR_DEBUGGER_NET_HOST_IP	Указывает IP-адрес целевой машины (отладчика), применяемой при отладке гипервизора по сети
hypervisorhostport	Целое число	HYPERVISOR_DEBUGGER_NET_HOST_PORT	Указывает сетевой порт, используемый при отладке гипервизора по сети
hypervisorusekey	Текстовая строка	HYPERVISOR_DEBUGGER_NET_KEY	Указывает ключ шифрования, применяемый для шифрования передаваемых по кабелю отладочных пакетов
hypervisorbusparams	Текстовая строка	HYPERVISOR_DEBUGGER_BUSPARAMS	Указывает номер шины, устройства и функции сетевого адаптера, используемых для отладки гипервизора
hypervisordhcp	Бинарное	HYPERVISOR_DEBUGGER_NET_DHCP	Указывает, должен ли отладчик гипервизора применять DHCP для получения IP-адреса сетевого интерфейса

Все записи, находящиеся в хранилище BCD, играют ключевую роль в последовательности запуска. Внутри каждой загрузочной записи (загрузочная запись — это объект BCD) перечислены все параметры загрузки, которые хранятся в кусте как подразделы реестра (рис. 12.5). Эти параметры называются элементами BCD. Диспетчер загрузки может добавить или удалить любой параметр загрузки как в физическом кусте, так и только в памяти. Это важно потому, что, как будет сказано далее в разделе «Меню загрузки», не все опции BCD должны находиться в физическом кусте.

Если куст Boot Configuration Data поврежден или при разборе его загрузочных записей произошла какая-то ошибка, то диспетчер загрузки повторно выполняет операцию с помощью куста Recovery BCD. Он обычно хранится в папке \EFI\Microsoft\Recovery\BCD. Система может быть настроена на прямое использование этого хранилища, минуя обычное, с помощью параметра `recoverybcd`, хранящегося в загрузочной переменной UEFI, или файла Bootstat.log.

Система готова к загрузке политик Secure Boot, показу меню загрузки, если это необходимо, и запуску загрузочного приложения. Список сертификатов загрузки, которым прошивка может доверять или не доверять, находится в переменных `db` и `dbx`, аутентифицированных UEFI. Загрузочная библиотека целостности кода считывает и анализирует переменные UEFI, но они управляют только тем, может ли быть загружен определенный модуль диспетчера загрузки. После запуска диспетчер загрузки позволяет дополнительно настроить или расширить конфигурацию Secure Boot, обеспечиваемую UEFI, с помощью списка сертификатов, предоставляемого Microsoft. Файл политики Secure Boot (хранится в \EFI\Microsoft\Boot\SecureBootPolicy.p7b), файлы политик манифеста платформы (файлы .pm) и дополнительные политики (файлы `.pol`) анализируются и объединяются с политиками, хранящимися в переменных UEFI. Поскольку в итоге за дело берется

механизм обеспечения целостности кода ядра, дополнительные политики содержат информацию и сертификаты, специфичные для конкретной ОС. В этом подходе защищенная версия Windows, например S-версия, может проверить ряд сертификатов, не расходуя драгоценные ресурсы UEFI. Это создает основу для доверия, поскольку файлы, в которых указываются новые настраиваемые списки сертификатов, подписываются цифровым сертификатом, содержащимся в базе данных разрешенных подписей UEFI.

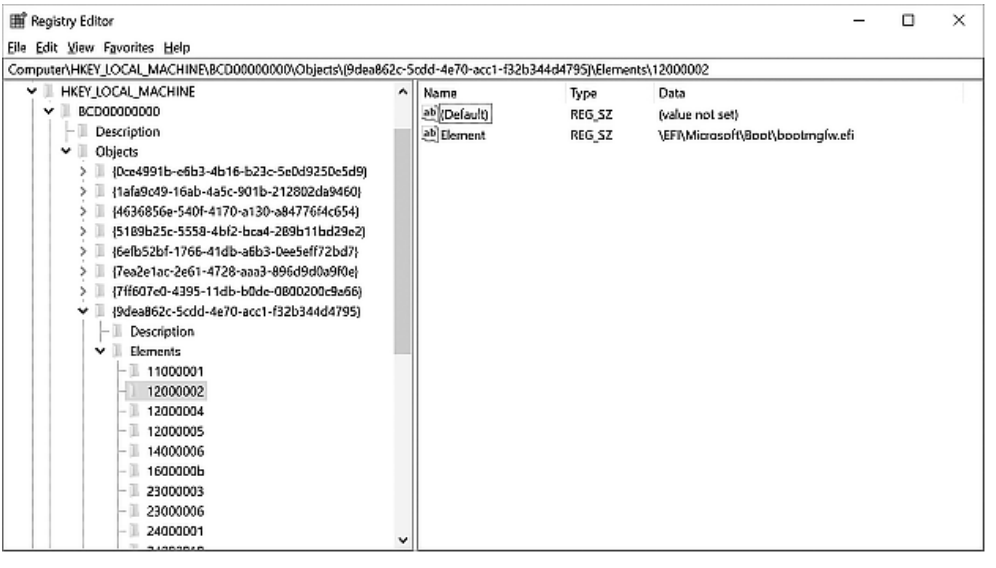


Рис. 12.5. Объекты BCD в окне диспетчера загрузки и связанные с ними опции загрузки (элементы BCD)

Если это не запрещено опциями загрузки `nointegritycheck` или `testsigning` или политикой Secure Boot, диспетчер загрузки выполняет самопроверку целостности: открывает собственный файл с жесткого диска и проверяет его цифровую подпись. При включенной функции Secure Boot цепочка подписи проверяется на соответствие политикам подписи Secure Boot.

Диспетчер загрузки инициализирует отладчик загрузки и проверяет через системную ACPI-таблицу BGRT, нужно ли ему показывать растровое изображение для OEM. Если нужно, то очищает экран и показывает логотип. Если Windows включила параметр BCD для информирования Bootmgr о возобновлении из спящего режима или гибридной загрузки, это сокращает процесс загрузки, запуская приложение Windows Resume Application, Winresume.efi, считывающее содержимое файла спящего режима в память и передающее управление коду в ядре, который возобновляет работу системы, находящейся в спящем режиме. Этот код отвечает за перезапуск драйверов, которые были активны в момент выключения системы. Hiberfil.sys действителен только в том случае, если последнее выключение компьютера было уходом в спящий режим или гибридной загрузкой. Это связано с тем, что файл спящего режима аннулируется после возобновления работы, чтобы избежать появления нескольких возобновлений работы из одной и той же точки. BCD-объект Windows Resume Application связан с дескриптором диспетчера загрузки через

специальный BCD-элемент `resumeobject`, описанный в разделе «Спящий режим и быстрый запуск» далее в этой главе.

Bootmgr определяет, зарегистрированы ли специфичные для OEM загрузочные действия через относительный элемент BCD, и если да, то обрабатывает их. На момент написания книги единственным поддерживаемым действием специфичной загрузки является запуск OEM-последовательности загрузки. При таком подходе OEM-производители могут зарегистрировать специфичную последовательность восстановления, вызываемую определенной клавишей, нажатой пользователем при запуске.

Меню загрузки

В Windows 8 и более поздних версиях в стандартных конфигурациях загрузки классическое, устаревшее меню загрузки никогда не показывается, поскольку внедрена новая технология — современная загрузка. Она дает Windows богатый графический интерфейс, сохраняя при этом возможность более глубокого погружения в настройки загрузки. В такой конфигурации конечный пользователь может выбрать операционную систему, которую хочет запустить, даже на устройствах с сенсорным управлением, не имеющих нормальной клавиатуры и мыши. Новое загрузочное меню отрисовывается поверх подсистемы Win32. Ее архитектура описана позже в этой главе в разделе «Smss, Csrss и Wininit».

Опция загрузки `bootmenupolicy` управляет тем, должна ли программа загрузки использовать старую или новую технологию для показа меню. Если нет OEM-последовательностей загрузки, то Bootmgr перечисляет GUID системных загрузочных записей, связанных с опцией `displayorder` в диспетчере. Если это значение пустое, то Bootmgr применяет запись по умолчанию. Для каждого найденного GUID Bootmgr открывает соответствующий объект BCD и запрашивает тип загрузочного приложения, его устройство запуска и читаемое описание. Все три атрибута должны существовать, в противном случае загрузочная запись считается недействительной и будет пропущена. Если Bootmgr не находит ни одного корректного загрузочного приложения, он выдает пользователю сообщение об ошибке и весь процесс загрузки прерывается. Алгоритм показа загрузочного меню начинается здесь. Одна из ключевых функций, `BmpProcessBootEntry`, используется следующим образом для принятия решения о том, показывать ли устаревшее меню загрузки.

- Если политика загрузочного меню приложения по умолчанию, а не записи Bootmgr, явно установлена на тип Modern, то алгоритм немедленно завершает работу и запускает вариант по умолчанию посредством функции `BmpLaunchBootEntry`. Примечательно, что в этом случае не проверяется нажатие пользователем клавиш, поэтому невозможно принудительно остановить процесс загрузки. Если в системе имеется несколько загрузочных записей, то в список вариантов загрузки в памяти приложения, загружаемого по умолчанию, добавляется специальная опция[1]. Этим способом на более поздних этапах запуска системы Winlogon может распознать данную опцию и показать современное меню.

[1] Эта мультизагрузочная специальная опция не имеет названия. Ее код элемента BCDE_LIBRARY_TYPE_MULTI_BOOT_SYSTEM (0x16000071).

- В противном случае, если политика загрузки для загрузочного приложения по умолчанию старая или вообще не установлена и существует только одна запись, `BmpProcessBootEntry` проверяет, нажал ли пользователь клавишу F8 или F10. Они описаны в файле ресурсов `bootmgr.xsl` как клавиши Advanced Options и Boot Options. Если Bootmgr обнаруживает, что одна из клавиш нажата во время запуска, то добавляет соответствующий элемент BCD в список опций загрузки в памяти приложения загрузки по умолчанию. Элемент BCD не записывается на диск. Эти два варианта загрузки обрабатываются позже в загрузчике Windows. Наконец, `BmpProcessBootEntry` проверяет с помощью относительной BCD-опции `displaybootmenu`, должна ли система показать загрузочное меню даже при наличии только одной записи.
- Если возможных загрузок несколько, то проверяется значение тайм-аута, которое хранится в виде BCD-опции, и если оно равно `0`, то сразу запускается приложение по умолчанию. В противном случае с помощью функции `BmDisplayBootMenu` показывается старое меню загрузки.

При показе старого меню Bootmgr перечисляет установленные инструменты загрузки, указанные в опции `toolsdisplayorder` диспетчера.

Запуск загрузочного приложения

Последняя задача диспетчера загрузки — правильно запустить загрузочное приложение, даже если оно находится на диске с шифрованием через BitLocker, и управлять последовательностью восстановления, если что-то пойдет не так. `BmpLaunchBootEntry` получает GUID и список вариантов загрузки для приложения, которое необходимо запустить. Одно из первых действий функции — проверка того, является ли указанная запись записью восстановления Windows (WinRE) с помощью элемента BCD. Эти типы загрузочных приложений используются в ходе работы с последовательностью восстановления. Если запись относится к типу WinRE, то системе необходимо определить загрузочное приложение, которое пытается восстановить WinRE. В этом случае начальное устройство такого загрузочного приложения определяется, а затем разблокируется, если зашифровано.

Процедура `BmTransferExecution` использует службы, предоставляемые библиотекой загрузки, чтобы открыть устройство загрузочного приложения, определить, зашифровано ли устройство, и если да, расшифровать его и прочитать файл загрузчика целевой ОС. Если целевое устройство зашифровано, то диспетчер загрузки сначала пытается получить мастер-ключ от TPM. В этом случае TPM открывает мастер-ключ только при соблюдении определенных условий. При этом подходе если изменилась какая-то конфигурация запуска, например включение Secure Boot, то TPM не сможет выдать ключ. Если извлечение ключа из TPM не удалось, диспетчер загрузки Windows показывает экран, подобный изображенному на рис. 12.6, и просит пользователя ввести ключ разблокировки, даже если в политике меню загрузки установлено значение Modern, поскольку на данном этапе система не имеет возможности запустить пользовательский интерфейс современной загрузки. На момент написания книги Bootmgr поддерживает четыре метода разблокировки: с помощью PIN-кода, парольной фразы, внешнего носителя и ключа восстановления. Если пользователь не может предоставить ключ, то процесс запуска прерывается и запускается последовательность восстановления Windows.

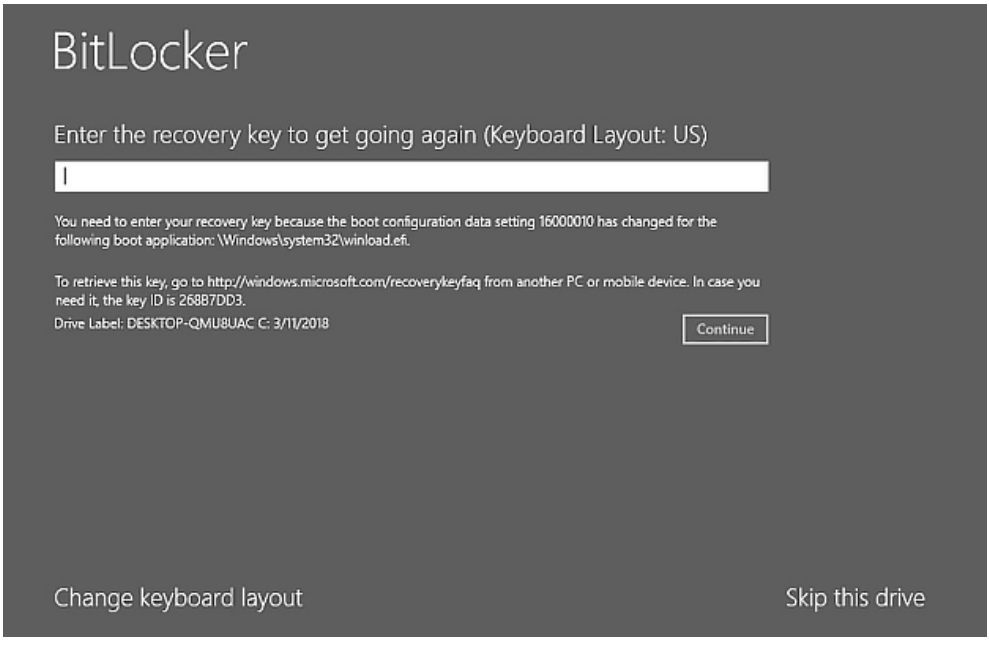


Рис. 12.6. Процедура восстановления BitLocker, которая была запущена из-за того, что в конфигурации загрузки что-то изменилось

Прошивка используется для чтения и проверки загрузчика целевой ОС. Проверка осуществляется с помощью библиотеки Code Integrity, которая применяет к цифровой подписи файла политики безопасной загрузки, как системные, так и все настраиваемые. Перед тем как передать выполнение целевому загрузочному приложению, диспетчер загрузки должен уведомить зарегистрированные компоненты, в частности ETW и измеренной загрузки, о том, что загрузочное приложение запускается. Кроме того, он должен убедиться в том, что TPM не может быть применен для разблокировки чего-либо еще.

Наконец, выполнение кода передается загрузчику Windows с помощью `BlImgStartBootApplication`. Эта процедура возвращает исполнение только в случае определенных ошибок. Как и раньше, диспетчер загрузки справляется с такой ситуацией, запуская последовательность восстановления Windows.

Измеренная загрузка

В конце 2006 года компания Intel представила технологию Trusted Execution Technology (TXT), которая гарантирует, что подлинная операционная система запускается в доверенной среде и не будет изменена внешним агентом, например вредоносным программным обеспечением. TXT использует TPM (Trusted Platform Module) и криптографические методы для оценки программного обеспечения и компонентов платформы (UEFI). Windows 8.1 и более поздние версии поддерживают новую функцию под названием «измеренная загрузка» (Measured Boot), которая измеряет каждый компонент, начиная с прошивки и заканчивая драйверами запуска,

сохраняет эти измерения в TPM машины, а затем предоставляет журнал, который можно исследовать удаленно, чтобы проверить состояние загрузки клиента. Этой технологии не существовало бы без TPM. Термин «измерение» означает процесс вычисления криптографического хеша чего-либо, например кода, структуры данных, конфигурации и вообще всего, что может быть загружено в память. Измерения применяются для различных целей. Измеренная загрузка предоставляет антивирусным программам надежный, устойчивый к подделке и фальсификации журнал всех компонентов загрузки, которые запускались перед Windows. Антивирусное программное обеспечение использует этот журнал, чтобы определить, являются ли компоненты, запущенные до него, надежными или зараженными вредоносным ПО. Программное обеспечение на локальной машине отправляет журнал на удаленный сервер для исследования. Работая с TPM и программным обеспечением, не относящимся к Microsoft, измеренная загрузка позволяет доверенному серверу в сети проверять целостность процесса запуска Windows.

Перечислим основные правила TPM.

- Необходимо обеспечить безопасное энергонезависимое хранилище для защиты секретов.
- Следует предоставить регистры конфигурации платформы (Platform Configuration Registers, PCR) для хранения измерений.
- Необходимо предоставить аппаратные криптографические механизмы и надежный генератор случайных чисел.

TPM хранит измерения загрузки в регистрах конфигурации платформы. Каждый регистр предоставляет область хранения, которая позволяет хранить неограниченное количество измерений в фиксированном объеме. Эта возможность обеспечивается свойством криптографических хешей. Диспетчер загрузки или загрузчик на более поздних этапах никогда не записывает данные непосредственно в регистр — он расширяет его содержимое. Операция расширения берет текущее значение регистра, добавляет новое измеренное значение и вычисляет криптографический хеш объединенного значения, обычно SHA-1 или SHA-256. Результат хеширования является новым значением регистра. Метод расширения обеспечивает зависимость измерений от порядка. Одним из свойств криптографических хешей является их зависимость от порядка. Это означает, что хеширование двух значений, A и B, приводит не к тому результату, который дает хеширование B и A. Поскольку регистры расширяются, а не перезаписываются, то даже если вредоносная программа сможет расширить регистр, единственным результатом станет то, что он будет содержать недействительное измерение. Еще одно свойство криптографических хешей заключается в том, что невозможно создать блок данных, который выдавал бы заданный хеш. Таким образом, невозможно расширить регистр, чтобы получить заданный результат, кроме как измеряя те же объекты в точно таком же порядке.

На ранних этапах процесса загрузки модуль System Integrity загрузочной библиотеки регистрирует различные функции обратного вызова. Каждая из них будет вызываться позже в различных точках последовательности запуска для управления событиями загрузки, такими как включение проверочного подписывания, включение отладчика загрузки, загрузка PE-образа, запуск загрузочного приложения, хеширование, запуск, выход, разблокировка BitLocker. Каждая функция решает,

какие данные хешировать и расширять в регистры TPM PCR. Например, каждый раз, когда диспетчер загрузки или загрузчик Windows запускает внешний исполняемый образ, он создает три измеряемых события, соответствующих различным фазам загрузки образа: `LoadStarting`, `ApplicationHashed` и `ApplicationLaunched`. В этом случае измеряемые сущности, которые отправляются в регистры 11 и 12 TPM, следующие: хеш образа, хеш цифровой подписи образа, база образа и размер.

Все измерения будут использованы в Windows позже, после полного запуска системы, для процедуры, называемой *аттестацией*. Благодаря свойству уникальности криптографических хешей можно задействовать значения регистров и их журналы для точной идентификации выполняемой версии программного обеспечения и его окружения. На этом этапе Windows использует TPM для создания цитаты TPM, где TPM подписывает значения регистров PCR, чтобы подтвердить, что они не были злонамеренно или случайно изменены при передаче. Это гарантирует подлинность измерений. Цитируемые измерения отправляются в центр сертификации — доверенную стороннюю организацию, способную подтвердить подлинность значений PCR и перевести эти значения, сравнив их с базой данных известных хороших значений. Описание всех моделей, применяемых для аттестации, выходит за рамки тематики этой книги. Конечная цель состоит в том, чтобы удаленный сервер подтвердил, является ли клиент доверенным лицом или мог быть изменен каким-либо вредоносным компонентом.

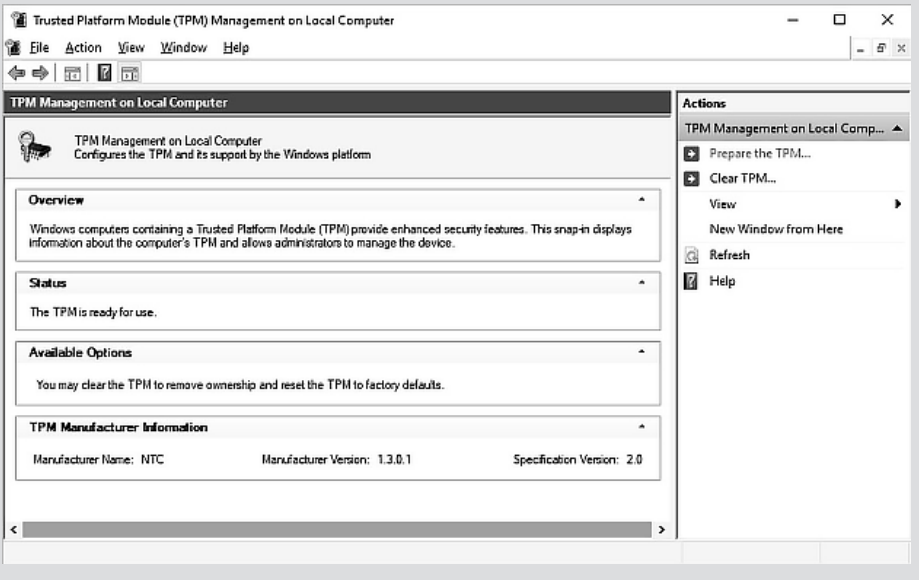
Ранее рассказывалось, как диспетчер загрузки может автоматически разблокировать зашифрованный с помощью BitLocker том запуска. В этом случае система использует еще одну важную услугу, предоставляемую TPM, — безопасное энергонезависимое хранилище. Энергонезависимая память с произвольным доступом (Nonvolatile Random Access Memory, NVRAM) TPM сохраняет данные при отключении питания и имеет больше функций безопасности, чем системная память. При выделении NVRAM TPM система должна указать следующее.

- **Права доступа на чтение.** Надо указать, какой уровень привилегий TPM, называемый *локальностью*, может читать данные. Что важнее, надо указать, должны ли какие-либо регистры PCR содержать определенные значения для считывания данных
- **Права доступа на запись.** То же самое, но для доступа на запись.
- **Атрибуты и разрешения.** Надо указать необязательные значения авторизации для чтения или записи (например, пароль) и временной или постоянной блокировки, то есть память может быть заблокирована для доступа на запись.

При первом шифровании пользователем загрузочного тома BitLocker шифрует главный ключ тома (Volume Master Key, VMK) другим случайным симметричным ключом, а затем запечатывает этот ключ, задействуя в качестве условия запечатывания расширенные значения регистров TPM, в частности регистров 7 и 11, которые измеряют BIOS и последовательность загрузки Windows. *Запечатывание* — это действие, при котором TPM шифрует блок данных таким образом, что он может быть расшифрован лишь тем же TPM, который его зашифровал, только если указанные регистры PCR имеют правильные значения. При последующих загрузках, если распечатывание запрашивается скомпрометированной последовательностью загрузки или другой конфигурацией BIOS, TPM отклоняет запрос на распечатывание и раскрытие ключа шифрования VMK.

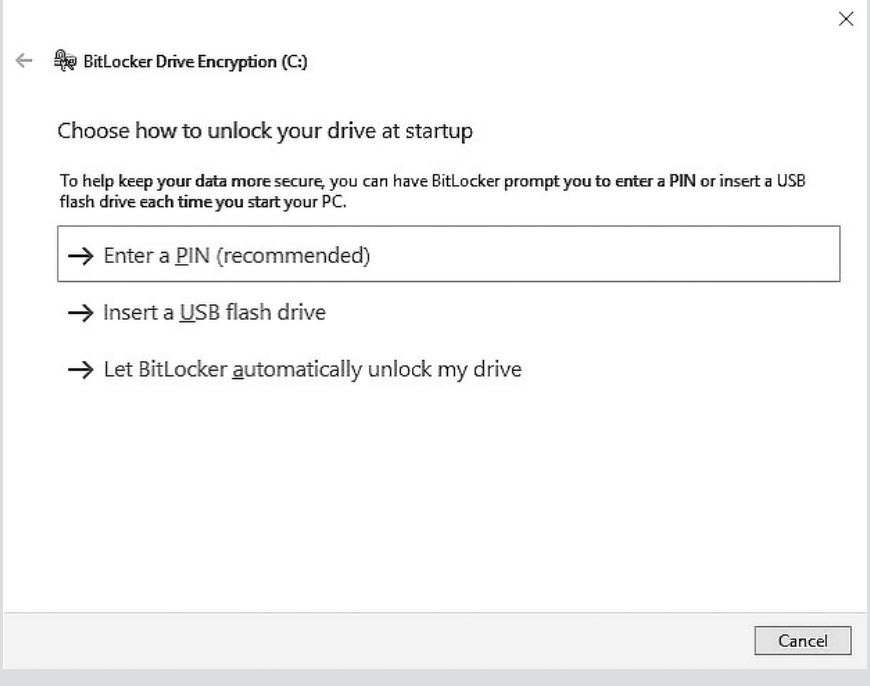
ЭКСПЕРИМЕНТ. Сделайте измерения TPM недействительными

В этом эксперименте исследуется быстрый способ аннулировать измерения TPM путем аннулирования конфигурации BIOS. Перед измерением последовательности запуска, драйверов и данных измеренная загрузка начинается со статического измерения конфигурации BIOS (хранится в регистре PCR1). Измеряемые данные конфигурации BIOS жестко зависят от производителя оборудования и иногда даже включают список порядка загрузки UEFI. Перед началом эксперимента убедитесь, что в системе установлен корректный TPM. Наберите `tpm.msc` в поисковой строке меню Пуск и запустите программу. Должна появиться консоль управления модулем TPM. Убедитесь, что TPM присутствует и включен в системе, проверив, что в поле Статус (Status) установлено значение TPM готов для использования (The TPM Is Ready For Use).



Запустите шифрование системного тома с помощью BitLocker. Если системный том уже зашифрован, этот шаг можно пропустить. Однако нужно обязательно сохранить ключ восстановления. Его можно проверить, выбрав пункт Создать резервную копию ключа восстановления (Back Up Your Recovery Key), находящийся в апплете шифрования диска, через BitLocker в панели управления. Откройте диспетчер файлов, щелкнув на его значке на панели задач, и перейдите в раздел Мой компьютер. Щелкните правой кнопкой мыши на системном томе (томе, содержащем все файлы Windows, обычно C) и выберите Включить BitLocker (Turn On BitLocker). После выполнения первоначальных проверок выберите Позволить BitLocker автоматически разблокировать мой диск (Let BitLocker Automatically Unlock My Drive), когда появится запрос на странице Выбрать способ

разблокировки диска при запуске (Choose How to Unlock Your Drive at Startup). Таким образом, VMK будет запечатан TPM с использованием загрузочных измерений в качестве ключа распечатывания. Не забудьте сохранить ключ восстановления — он понадобится на следующем этапе. Иначе вы больше не сможете получить доступ к своим файлам. Для всех остальных параметров оставьте значения по умолчанию.



После завершения шифрования выключите компьютер, запустите его заново и войдите в конфигурацию UEFI BIOS. Эта процедура различается для ПК разных производителей, поэтому смотрите инструкции по входу в настройки UEFI BIOS в руководстве пользователя оборудования. На страницах конфигурации BIOS измените порядок загрузки, а затем перезагрузите компьютер. Порядок загрузки при старте можно изменить с помощью утилиты UefiTool, которая находится в загружаемых файлах для этой книги. Если производитель оборудования включил порядок загрузки в параметры TPM, то должно появиться сообщение BitLocker о восстановлении перед загрузкой Windows. В противном случае, чтобы аннулировать измерения TPM, просто вставьте установочный DVD-диск Windows или флеш-накопитель перед включением рабочей станции. Если порядок загрузки настроен правильно, то запускается код начальной загрузки Windows Setup, который выводит сообщение Нажмите любую клавишу для загрузки с CD или DVD (Press Any Key For Boot From CD Or DVD). Если не нажать ни одной клавиши,

то система переходит к загрузке следующего элемента в списке. В этом случае последовательность запуска изменилась, и TPM-измерения иные. В результате TPM не сможет распечатать VMK.

Можно аннулировать результаты измерений TPM и получить те же результаты, если безопасная загрузка включена и попытаться ее отключить. Этот эксперимент демонстрирует, что измеренная загрузка привязана к конфигурации BIOS.

Доверенное исполнение

Хотя измеренная загрузка дает возможность удаленному субъекту подтвердить целостность процесса загрузки, она не решает важную проблему: диспетчер загрузки по-прежнему доверяет прошивке машины и использует ее службы для эффективного взаимодействия с TPM и запуска всей платформы. На момент написания этой книги атаки на прошивку ядра UEFI были продемонстрированы множество раз. Технология Trusted Execution Technology (TXT) была усовершенствована для поддержки еще одной важной функции, называемой безопасным запуском (Secure Launch). Безопасный запуск, также известный как безопасная загрузка (Trusted Boot или TBOOT в номенклатуре Intel), обеспечивает безопасные модули

аутентифицированного кода (Authenticated Code Modules, ACM), которые подписываются производителем процессора и выполняются чипсетом, а не прошивкой. Безопасный запуск обеспечивает поддержку динамических измерений в регистры PCR, которые могут быть сброшены без перезагрузки платформы. В этом случае операционная система предоставляет специальный модуль безопасной загрузки, который используется для инициализации платформы для работы в защищенном режиме и инициирования процесса безопасного запуска.

Модуль аутентифицированного кода (Authenticated Code Module, ACM) — это часть кода, предоставляемого производителем чипсета. ACM подписан производителем, и его код выполняется на одном из самых высоких уровней привилегий в специальной защищенной памяти, встроенной в процессор. Модули ACM вызываются с помощью специальной инструкции `GETSEC`. Существует два типа ACM: BIOS и SINIT. BIOS ACM измеряет BIOS и выполняет некоторые функции безопасности BIOS, а SINIT ACM используется для измерения и запуска модуля TCB операционной системы (TBOOT). Как BIOS, так и SINIT ACM обычно содержатся в образе системного BIOS (это требование не строгое), но при необходимости могут быть обновлены и заменены ОС. (Подробности см. в разделе «Безопасный запуск» далее в этой главе.)

ACM — это основа доверенных измерений. Поэтому он работает на самом высоком уровне безопасности и должен быть защищен от всех типов атак. Микрокод процессора копирует модуль ACM в защищенную память и выполняет различные проверки, прежде чем разрешить его выполнение. Процессор проверяет, что ACM разработан для работы с целевым чипсетом. Кроме того, он проверяет целостность, версию и цифровую подпись ACM, которая сверяется с открытым ключом, записанным в предохранители чипсета. Инструкция `GETSEC` не выполняет ACM, если одна из предыдущих проверок не прошла.

Еще одной ключевой особенностью безопасного запуска является поддержка в TPM динамического измерения корня доверия (Dynamic Root of Trust Measurement, DRTM). Как говорилось в предыдущем разделе, 16 различных регистров TPM PCR, с 0-го по 15-й, обеспечивают хранение результатов измерений загрузки. Диспетчер загрузки может расширить эти регистры, но очистить их содержимое до следующего сброса платформы или включения питания невозможно. Это объясняет, почему такие измерения называются статическими. Динамические измерения — это измерения, выполненные в PCR, которые могут быть сброшены без перезагрузки платформы. Существует шесть динамических PCR, используемых безопасным запуском и доверенной операционной системой. (На самом деле их восемь, но два зарезервированы и не применяются операционной системой.)

В типичной последовательности загрузки TXT процессор загрузки после проверки целостности ACM выполняет код запуска ACM, который измеряет критические компоненты BIOS, выходит из безопасного режима ACM и переходит к коду запуска UEFI BIOS. Затем BIOS измеряет весь оставшийся код, конфигурирует платформу и проверяет результаты измерений, выполняя инструкцию `GETSEC`. Эта TXT-инструкция загружает модуль BIOS ACM, который

проверяет безопасность и замораживает конфигурацию BIOS. На этом этапе UEFI BIOS может измерить код каждой опции в постоянной памяти (для каждого устройства) и начальную загрузку программы (Initial Program Load, IPL). Платформа теперь приведена в состояние готовности к загрузке операционной системы — конкретно через код IPL.

Последовательность загрузки TXT — это часть статического измерения корня доверия (Static Root of Trust Measurement, SRTM), поскольку доверенный код BIOS и диспетчер загрузки уже проверены и находятся в известном хорошем состоянии, которое не изменится до следующего сброса платформы. Как правило, в ОС с поддержкой TXT вместо первого загружаемого модуля ядра используется специальный модуль TCB (TBOOT). Его цель — инициализировать платформу для работы в безопасном режиме и инициировать безопасный запуск. Модуль TBOOT называется TcbLaunch.exe. Перед началом безопасного запуска он должен быть проверен модулем SINIT ACM. Таким образом, нужны компоненты, которые выполняют инструкции `GETSEC` и запускают DRTM. В модели безопасного запуска Windows таким компонентом является загрузочная библиотека.

Прежде чем система перейдет в безопасный режим, она должна перевести платформу в известное состояние. В нем все процессоры, кроме загрузочного, находятся в специальном состоянии простоя, поэтому никакой другой код не может быть выполнен. Загрузочная библиотека выполняет инструкцию `GETSEC`, указывая операцию `SENTER`. Это заставляет процессор сделать следующее.

1. Проверить модуль SINIT ACM и загрузить его в защищенную память процессора.
2. Запустить DRTM, очистив все относительные динамические регистры PCR, а затем измерить SINIT ACM.
3. Выполнить код SINIT ACM, измеряющий код доверенной операционной системы и запускающий политику управления запуском. Политика определяет, позволяют ли текущие измерения, которые находятся в некоторых динамических регистрах PCR, считать ОС доверенной.

Если одна из этих проверок не проходит, считается, что машина атакована, и ACM выдает TXT-сброс, который не позволит выполнять любую программу до тех пор, пока платформа не будет перезагружена. В противном случае ACM включает безопасный запуск, выходя из режима ACM и переходя к точке входа в доверенную операционную систему, — в Windows это функция `TcbMain` модуля TcbLaunch.exe. После этого доверенная операционная система берет управление на себя. Она может расширять и сбрасывать динамические PCR для каждого измерения, которое ей необходимо, или применять другой механизм, обеспечивающий цепочку доверия.

Описание всей архитектуры безопасного запуска выходит за рамки тематики этой книги. Смотрите спецификации TXT в руководствах Intel. Описание того, как реализовано доверенное выполнение в Windows, см. в разделе «Безопасный запуск» далее в этой главе. На рис. 12.7 показаны все компоненты, задействованные в технологии Intel TXT.

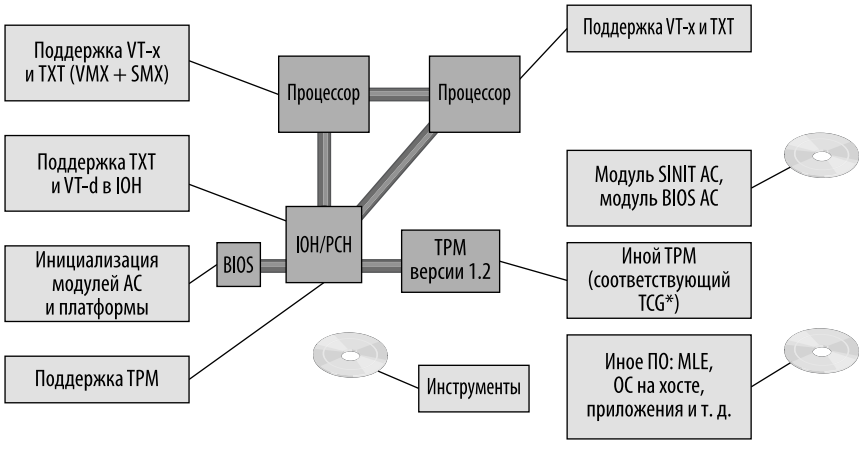


Рис. 12.7. Компоненты технологии Intel TXT

Загрузчик операционной системы Windows

Загрузчик операционной системы Windows (Winload) — это приложение, запускаемое диспетчером загрузки с целью загрузки и правильного выполнения ядра Windows. Этот процесс включает в себя несколько основных задач.

- Создание среды выполнения ядра. Включает инициализацию и использование таблиц страниц ядра, а также создание карты памяти. Загрузчик операционной системы EFI создает и инициализирует стеки ядра, страницу совместного доступа, GDT, IDT, TSS и селекторы сегментов.
- Загрузка в память всех модулей, которые необходимо выполнить или к которым нужно получить доступ до инициализации дискового стека. К ним относятся ядро и HAL, поскольку они выполняют раннюю инициализацию основных служб после передачи управления от загрузчика ОС. Критичные для загрузки драйверы и системный куст реестра также загружаются в память.
- Определение того, следует ли выполнять Hyper-V и безопасное ядро (VSM), и если да, то их правильная загрузка и запуск.
- Отрисовка первой фоновой анимации с использованием новой библиотеки загрузочной графики высокого разрешения BGFX, которая заменяет старый драйвер bootvid.dll.
- Организация последовательности загрузки Secure Launch в системах с поддержкой Intel TXT. (Полное описание измеренной загрузки, безопасного запуска и Intel TXT см. в соответствующих разделах ранее в этой главе.) Изначально эта задача была реализована в загрузчике гипервизора, но начиная с Windows 10 October Update (RS5) переместилась в другое место.

Загрузчик неоднократно улучшался и изменялся в каждом выпуске Windows. `OslMain` — это главная функция загрузчика, вызываемая диспетчером загрузки,

которая повторно инициализирует загрузочную библиотеку и вызывает внутреннюю `OslpMain`. На момент написания книги загрузочная библиотека поддерживает два контекста выполнения.

- Контекст прошивки означает, что подкачка отключена. На самом деле она не отключена, но обеспечивается прошивкой, которая выполняет взаимно однозначное отображение физических адресов, и для управления памятью используются только службы прошивки. Windows задействует этот контекст выполнения в диспетчере загрузки.
- Контекст приложения означает, что подкачка включена и обеспечивается операционной системой. Это контекст, используемый загрузчиком Windows.

Диспетчер загрузки непосредственно перед передачей исполнения загрузчику операционной системы создает и инициализирует четырехуровневую иерархию таблиц страниц x64, которую будет использовать ядро Windows, создавая только записи для самоотображения и однозначного отображения. Перед самым запуском `OslMain` переключается на контекст выполнения Application. Процедура `OslPrepareTarget` фиксирует состояние загрузки или выключения при последней загрузке, считывая данные из файла bootstat.dat в корневом каталоге системы.

Если последняя загрузка не удалась более двух раз, процедура возвращается к диспетчеру загрузки для запуска среды восстановления. В противном случае она считывает куст реестра SYSTEM, \Windows\System32\Config\System, чтобы определить, какие драйверы устройств необходимо загрузить для выполнения загрузки. Куст — это файл, содержащий поддеревья реестра. Подробнее о реестре говорилось в главе 10. Затем инициализируется библиотека дисплея BGFX, отрисовывая первое фоновое изображение, и при необходимости выводится меню выбора продвинутых опций (см. раздел «Меню загрузки» ранее в этой главе). Блок загрузчика, одна из важнейших структур данных, необходимых для загрузки ядра NT, выделяется и заполняется такой базовой информацией, как базовый адрес и размер системного куста, случайное число, по возможности запрашиваемое у TPM, и т. д.

Блок `OslInitializeLoaderBlock` содержит код, который запрашивает ACPI BIOS системы для получения основной информации об устройствах и конфигурации, включая информацию о времени и дате событий, хранящуюся в CMOS системы. Эта информация собирается во внутренние структуры данных, которые будут храниться под разделом реестра HKLM\HARDWARE\DESCRIPTION позже при загрузке. Это устаревший раздел, который существует только для совместимости. В настоящее время информация об аппаратном обеспечении хранится в базе данных диспетчера Plug and Play.

Далее Winload начинает загрузку с загрузочного тома файлов, необходимых для начала инициализации ядра. Загрузочный том — это том, соответствующий разделу, на котором находится системный каталог загружаемой установки, обычно \Windows. Winload выполняет следующие шаги.

1. Определяет с помощью BCD-опции `hypervisorlaunchtype` и политики VSM, нужно ли загружать гипервизор или безопасное ядро, и если да, то начинается фаза 0 установки гипервизора. На этой фазе предварительно загружается модуль загрузчика HV Hvloader.dll в память RAM и выполняется процедура

его инициализации `HvlLoadHypervisor`. Она загружает и отображает в памяти образ гипервизора — Hvix64.exe, Hvax64.exe или Hvaa64.exe в зависимости от архитектуры — и все его зависимости.

2. Перечисляет все перечисляемые через прошивку диски и вставляет список в блок параметров загрузчика. Кроме того, загружает куст синтетической начальной конфигурации Imc.hiv, если он указан в конфигурационных данных, и прикрепляет его к блоку загрузчика.
3. Инициализирует модуль целостности кода ядра CI.dll и создает блок CI Loader. Затем модуль целостности будут совместно использовать ядро NT и безопасное ядро.
4. Обрабатывает все запланированные обновления прошивки. Windows 10 поддерживает обновления, распространяемые через Windows Update.
5. Загружает соответствующие образы ядра и HAL, по умолчанию Ntoskrnl.exe и Hal.dll. Если Winload не удается загрузить ни один из этих файлов, то он выводит сообщение об ошибке. Перед правильной загрузкой зависимостей этих модулей Winload проверяет их содержимое по цифровым сертификатам и загружает системный файл API Set Schema. Это дает возможность обрабатывать импорты API Set.
6. Инициализирует отладчик, загружая нужный транспорт отладчика.
7. Загружает модуль обновления прошивки процессора Mcupdate.dll, если необходимо.
8. `OslpLoadAllModules` загружает модули, от которых зависят ядро NT и HAL, драйверы ELAM, расширения ядра, драйверы TPM и все остальные загрузочные драйверы, соблюдая порядок загрузки — драйверы файловой системы загружаются первыми. Драйверы загрузочных устройств — это драйверы, необходимые для загрузки системы. Их конфигурация хранится в кусте реестра SYSTEM. Каждый драйвер устройства имеет подраздел реестра в разделе HKLM\SYSTEM\CurrentControlSet\Services. Например, в Services есть подраздел rdyboost для драйвера ReadyBoost, который показан на рис. 12.8. (Подробное описание записей реестра Services см. в разделе «Службы Windows» главы 10.) Все драйверы загрузки имеют начальное значение `SERVICE_BOOT_START` (0).

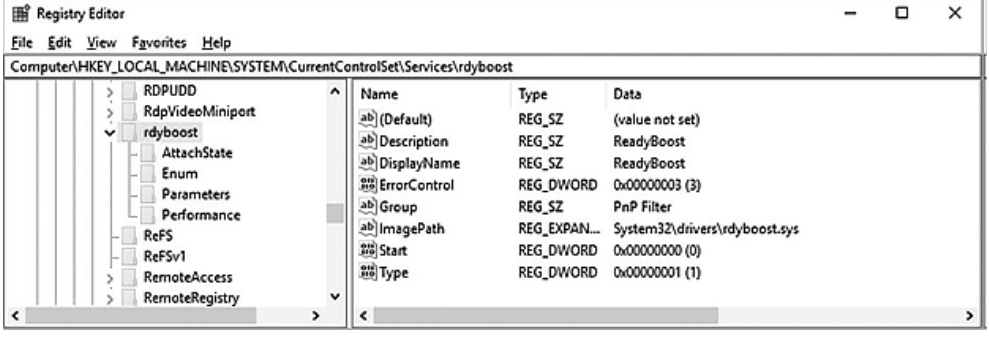


Рис. 12.8. Настройки службы драйвера ReadyBoost

9. На этом этапе для правильного выделения физической памяти Winload все еще использует службы, предоставляемые прошивкой EFI, — процедуру службы загрузки `AllocatePages`. Трансляцией же виртуальных адресов управляет загрузочная библиотека, запущенная в контексте выполнения приложения.
10. Считывает файлы NLS (National Language System), используемые для интернационализации. По умолчанию это `l_intl.nls`, `C_1252.nls` и `C_437.nls`.
11. Если оцениваемые политики требуют запуска VSM, то выполняется фаза 0 установки безопасного ядра, в ходе которой определяются местоположения процедур поддержки загрузчика VSM, экспортируемых модулем Hvloader.dll, а также загружаются модуль безопасного ядра Securekernel.exe и все его зависимости.
12. Для версии S Windows определяет минимальный настраиваемый уровень подписи целостности кода в пользовательском режиме для приложений.
13. Вызывает подпрограмму `OslArchpKernelSetupPhase0`, которая выполняет действия с памятью, необходимые для перехода ядра, такие как выделение GDT, IDT и TSS, отображение виртуального адресного пространства HAL, выделение стеков ядра, общей пользовательской страницы и передачи старой системы USB. Winload применяет функцию UEFI `GetMemoryMap` для получения полной карты физической памяти системы и отображает каждую физическую страницу, принадлежащую коду и данным исполнительной среды EFI, в виртуальное пространство памяти. Полная физическая карта будет передана ядру ОС.
14. Выполняет фазу 1 настройки VSM, копируя все необходимые таблицы ACPI из памяти VTL0 в память VTL1. На этом этапе создаются также таблицы страниц VTL1.
15. Модуль трансляции виртуальной памяти теперь работает весь, поэтому Winload вызывает функцию `ExitBootServices` UEFI, чтобы избавиться от служб загрузки прошивки, и перераспределяет все оставшиеся службы исполнительной среды UEFI в созданное виртуальное адресное пространство, используя функцию `SetVirtualAddressMap` исполнительной среды UEFI.
16. При необходимости запускает гипервизор и безопасное ядро именно в таком порядке. В случае успеха управление выполнением возвращается к Winload в контексте корневого раздела Hyper-V. (Подробные сведения о Hyper-V см. в главе 9.)
17. Передает выполнение в ядро через процедуру `OslArchTransferToKernel`.

Загрузка из iSCSI

Устройства интернет-SCSI (iSCSI) представляют собой разновидность сетевых хранилищ, в которых удаленные физические диски подключаются к адаптеру шины iSCSI Host Bus Adapter (HBA) или через сеть. Эти устройства отличаются от традиционных сетевых хранилищ NAS тем, что предоставляют доступ к дискам на уровне блоков, а не логический доступ через сетевую файловую систему, как в NAS. Поэтому подключенный к iSCSI диск неотличим от любого другого диска как для загрузчика, так и для ОС, если для обеспечения доступа через Ethernet-соединение применяется Microsoft iSCSI Initiator. Используя диски с поддержкой iSCSI вместо локальных хранилищ, можно сэкономить на расходах на помещение, энергопотребление и охлаждение.

Традиционно Windows поддерживала загрузку только с локально подключенных дисков или сетевую загрузку через PXE, но современные версии Windows способны загружаться и с устройств iSCSI с помощью механизма под названием iSCSI Boot. Как показано на рис. 12.9, загрузчик Winload.efi определяет, поддерживает ли система загрузочные устройства iSCSI, читая таблицу iSCSI Boot Firmware Table (iBFT), которая должна присутствовать в физической памяти и обычно открывается через ACPI. Благодаря таблице iBFT Winload знает местоположение удаленного диска, путь к нему и информацию об аутентификации. Если таблица присутствует, то Winload открывает и загружает предоставленный производителем драйвер сетевого интерфейса, помеченный флагом загрузки `CM_SERVICE_NETWORK_BOOT_LOAD` (0x1).

Кроме того, Windows Setup может читать эту таблицу, чтобы определить загрузочные устройства iSCSI и позволить прямую установку на такое устройство, что не потребует создания образа. В сочетании с инициатором Microsoft iSCSI Initiator это все, что требуется для загрузки Windows с iSCSI.

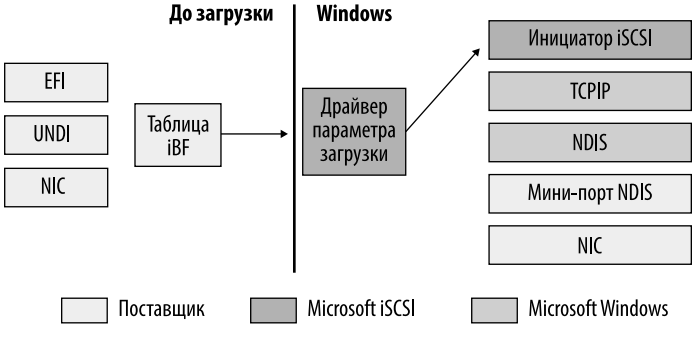


Рис. 12.9. Архитектура загрузки iSCSI

Загрузчик гипервизора

Загрузчик гипервизора — это загрузочный модуль Hvloader.dll, используемый для правильной загрузки и запуска гипервизора Hyper-V и безопасного ядра. (Полное описание Hyper-V и безопасного ядра см. в главе 9.) Модуль загрузчика гипервизора глубоко интегрирован в загрузчик Windows и имеет две основные задачи.

- Определение аппаратной платформы, загрузка и запуск нужной версии гипервизора Windows — Hvix64.exe для систем Intel, Hvax64.exe для систем AMD и Hvaa64.exe для систем ARM64.
- Разбор политики виртуального безопасного режима (Virtual Secure Mode, VSM), загрузка и запуск безопасного ядра.

В Windows 8 этот модуль представлял собой внешний исполняемый файл, загружаемый Winload по требованию. В то время единственной задачей загрузчика гипервизора были загрузка и запуск Hyper-V. С появлением VSM и доверенной загрузки архитектура была переработана, чтобы добиться лучшей интеграции всех компонентов.

Как уже говорилось, настройка гипервизора состоит из двух фаз. Первая фаза начинается в Winload сразу после инициализации блока NT Loader. HvLoader

определяет целевую платформу с помощью некоторых инструкций CPUID, копирует карту физической памяти UEFI и обнаруживает IOAPIC и IOMMU. Затем HvLoader загружает в память нужный образ гипервизора и все зависимости, например транспорт отладчика, и проверяет, соответствует ли информация о версии гипервизора ожидаемой. Это объясняет, почему HvLoader не может запустить другую версию Hyper-V. На данном этапе HvLoader выделяет блок загрузчика гипервизора — важную структуру данных, используемую для передачи системных параметров между HvLoader и самим гипервизором аналогично блоку загрузчика Windows. Важнейший этап фазы 1 — построение иерархии таблиц страниц гипервизора. Таблицы только что созданных страниц включают в себя лишь отображение образа гипервизора и его зависимостей и системных физических страниц в первом мегабайте. Страницы сопоставлены с идентификаторами и используются переходным кодом запуска (эта концепция объясняется далее в текущем разделе).

Вторая фаза начинается на последних этапах Winload: службы загрузки прошивки UEFI больше не используются, поэтому код HvLoader копирует диапазоны физических адресов служб исполнительной среды UEFI в блок загрузчика гипервизора, фиксирует состояние процессора, отключает прерывания, отладчик и подкачку и вызывает `HvlpTransferToHypervisorViaTransitionSpace` для переноса выполнения кода на физическую страницу в первом мегабайте. Находящийся здесь переходный код может переключить таблицы страниц, снова включить подкачку и перейти к коду гипервизора, что создает два разных адресных пространства. После запуска гипервизора он использует сохраненный контекст процессора, чтобы правильно передать выполнение кода Winload в контексте новой виртуальной машины, называемой корневым разделом. (Более подробная информация об этом доступна в главе 9.)

Запуск виртуального безопасного режима разделен на три этапа, поскольку некоторые действия необходимо выполнить после запуска гипервизора.

1. Первый этап очень похож на первую фазу настройки гипервизора. Данные копируются из блока загрузчика Windows в только что выделенный блок загрузчика VSM, генерируются главный ключ, ключ IDK и ключ Crashdump, в память загружается модуль SecureKernel.exe.
2. Второй этап инициируется Winload на поздних стадиях работы `OslPrepareTarget`, когда гипервизор уже инициализирован, но не запущен. Аналогично второй фазе настройки гипервизора, в блок загрузчика VSM копируются диапазоны физических адресов служб исполнительной среды UEFI, таблицы ACPI, данные о целостности кода, полная карта физической памяти системы и страница кода гипервызова. Наконец, здесь с помощью функции `OslpVsmBuildPageTables` строится иерархия таблиц защищенных страниц, используемых для защищенного пространства памяти VTL1, и создаются необходимые GDT.
3. Третий этап — последний в процессе запуска. Гипервизор уже запущен. Здесь выполняются последние проверки. Проверяется, например, наличие IOMMU и имеет ли корневой раздел привилегии VSM. IOMMU очень важен для VSM. На этом этапе также задается зашифрованная область для аварийного дампа гипервизора, копируются ключи шифрования VSM, а выполнение передается в точку входа безопасного ядра `SkiSystemStartup`. Код точки входа безопасного ядра исполняется в VTL 0. VTL 1 запускается кодом безопасного ядра на более поздних этапах с помощью гипервызова `HvCallEnablePartitionVtl`. (Подробности см. в главе 9.)

Политика запуска VSM

Во время запуска загрузчик Windows должен определить, нужно ли ему запускать виртуальный безопасный режим (Virtual Secure Mode, VSM). Чтобы предотвратить все попытки вредоносных программ отключить этот новый уровень защиты, система использует специальную политику для блокировки параметров запуска VSM. В конфигурации по умолчанию при первой загрузке после завершения копирования файлов Windows приложением Windows Setup загрузчик применяет процедуру `OslSetVsmPolicy` для чтения и закрытия конфигурации VSM, которая хранится в корневом разделе реестра VSM HKLM\SYSTEM\CurrentControlSet\Control\DeviceGuard.

VSM может быть включен с помощью:

- **сценариев защиты устройств.** Каждый сценарий хранится как подраздел в корневом разделе VSM. DWORD-параметр `Enabled` определяет, включен ли сценарий. Если один или несколько сценариев активны, то VSM включен;
- **глобальных параметров.** Хранятся в параметре реестра `EnableVirtualizationBasedSecurity`;
- **политик целостности кода HVCI.** Хранятся в файле политики целостности кода Policy.p7b.

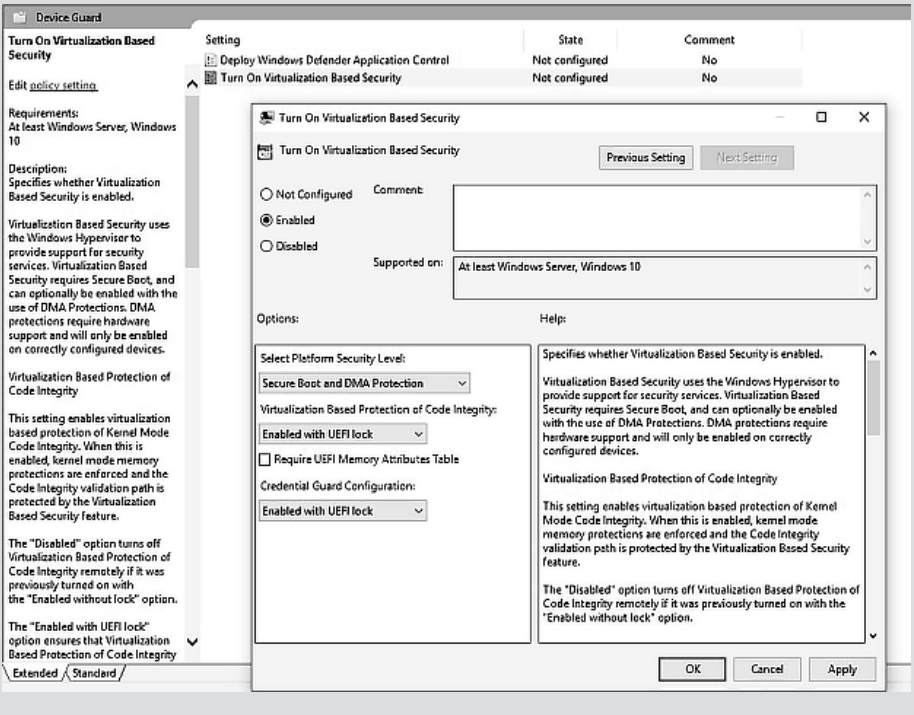
Кроме того, по умолчанию VSM автоматически включается при включении гипервизора, за исключением случаев, когда существует параметр реестра `HyperVVirtualizationBasedSecurityOptOut`.

Каждый источник активации VSM задает политику блокировки. Если режим блокировки включен, загрузчик создает переменную безопасной загрузки `VbsPolicy` и сохраняет в ней режим активации VSM и конфигурацию платформы. Часть конфигурации платформы VSM создается динамически на основе обнаруженного аппаратного обеспечения системы, а другая часть считывается из параметра реестра `RequirePlatformSecurityFeatures`, хранящегося в корневом разделе VSM. Переменная безопасной загрузки считывается при каждой последующей загрузке. Конфигурация, хранящаяся в переменной, всегда заменяет конфигурацию, находящуюся в реестре Windows.

При таком подходе, даже если вредоносное программное обеспечение сможет изменить реестр Windows, чтобы отключить VSM, то Windows просто проигнорирует это изменение и сохранит безопасность пользовательской среды. Вредоносное программное обеспечение не сможет изменить переменную безопасной загрузки VSM, поскольку согласно спецификации безопасной загрузки только новая переменная, подписанная проверенной цифровой подписью, способна изменить или удалить исходную. Microsoft предоставляет специальный подписанный инструмент, который может отключить защиту VSM. Это средство представляет собой специальное загрузочное приложение EFI, которое устанавливает другую подписанную переменную безопасной загрузки под названием `VbsPolicyDisabled`. Эта переменная распознается при запуске загрузчиком Windows. Если она существует, то Winload удаляет защищенную переменную `VbsPolicy` и изменяет реестр для отключения VSM — изменяются как глобальные настройки, так и каждая активация сценария.

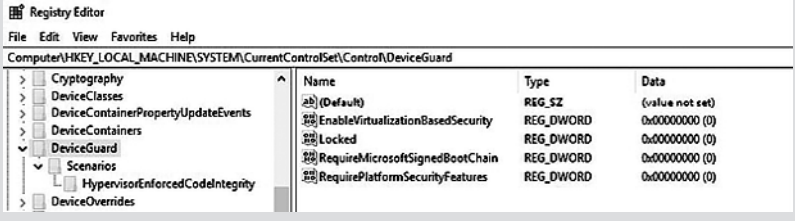
ЭКСПЕРИМЕНТ. Знакомство с политикой VSM

В этом эксперименте исследуется, как обеспечивается устойчивость безопасного запуска ядра к внешнему вмешательству. Сначала включите безопасность на основе виртуализации (Virtualization Based Security, VBS) в совместимой версии Windows — обычно хорошо работают версии Pro и Business. В них можно быстро проверить, включена ли VBS, с помощью диспетчера задач, и если да, то на вкладке Подробности (Details) должен быть виден процесс Secure System. Даже если VBS уже работает, проверьте, включена ли блокировка UEFI. Введите `Редактировать групповую политику` или `gpedit.msc` в строке поиска меню Пуск и запустите редактор группы локальных политик. Выберите Конфигурация компьютера (Computer Configuration) ▸ Административные шаблоны (Administrative Templates) ▸ Система (System) ▸ Защита устройств (Device Guard) и дважды щелкните на Включить безопасность на основе виртуализации (Turn On Virtualization Based Security). Убедитесь, что для политики установлено значение `Enabled`, а параметры настроены, как показано на рисунке.

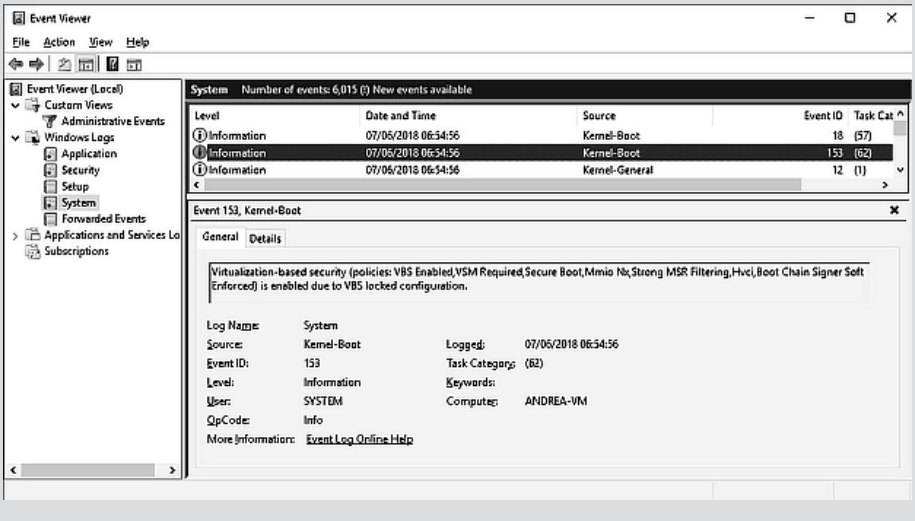


Убедитесь, что безопасная загрузка включена (для подтверждения можно использовать утилиту System Information или средство настройки BIOS системы), и перезагрузите систему. Опция `Enabled With UEFI Lock` обеспечивает защиту от несанкционированного доступа даже в контексте администратора. После перезагрузки системы отключите VBS с помощью того же редактора

групповой политики (убедитесь, что все параметры отключены) и удалите все параметры, расположенные в разделе реестра HKEY_LOCAL_MACHINE\SYSTEM\CurrentControlSet\Control\DeviceGuard (установка их в 0 дает тот же эффект). Используйте редактор реестра, чтобы правильно удалить все параметры.



Отключите гипервизор, выполнив команду `bcdedit /set {current} hypervisorlaunchtype off` из командной строки. Затем снова перезагрузите компьютер. После перезагрузки системы, даже если VBS и гипервизор обязаны быть отключены, должно быть видно, что процессы Secure System и Lsalso все еще присутствуют в диспетчере задач. Это связано с тем, что защищенная переменная UEFI `VbsPolicy` по-прежнему содержит оригинальную политику, поэтому вредоносная программа или пользователь не смогут легко отключить этот дополнительный уровень защиты. Чтобы достоверно подтвердить это, откройте средство просмотра системных событий, набрав `eventvwr`, и перейдите в раздел Журналы Windows (Windows Logs) ▸ Система (System). Если прокрутить список событий, то должно быть видно событие, описывающее тип активации VBS (ее источником будет Kernel-Boot).

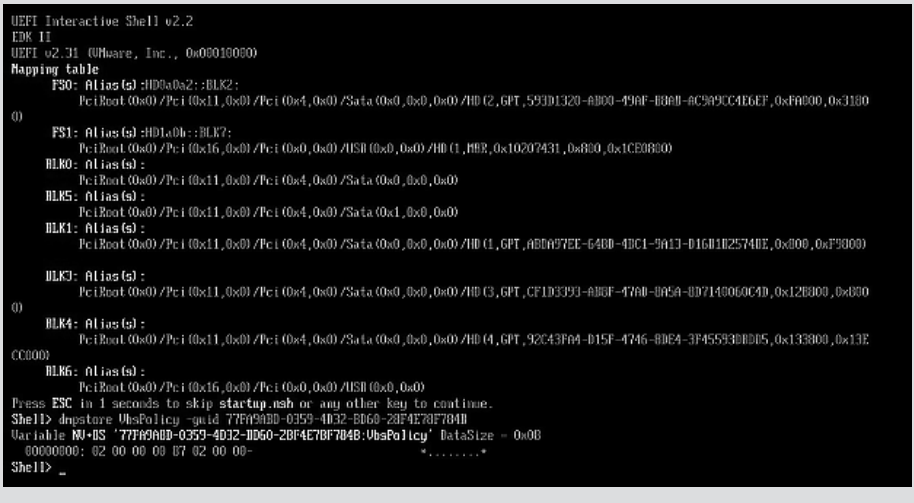


`VbsPolicy` — это переменная UEFI, аутентифицированная службами загрузки, поэтому она не видна после переключения ОС в режим исполнительной среды. Утилита UefiTool, использовавшаяся в предыдущем эксперименте, не способна

показывать подобные переменные. Чтобы изучить содержимое переменной VbsPolicy, перезагрузите компьютер, отключите безопасную загрузку и воспользуйтесь оболочкой Efi Shell. Эта оболочка (есть в ресурсах для скачивания к книге на сайте https://github.com/tianocore/edk2/tree/UDK2018/ShellBinPkg/UefiShell/X64) должна быть скопирована на USB-накопитель с FAT32 в файл с именем `bootx64.efi` и помещена по адресу `efi\boot`. Теперь можно загрузиться с USB-накопителя, что приведет к запуску оболочки Efi Shell. Выполните следующую команду:

```
dmpstore VbsPolicy -guid 77FA9ABD-0359-4D32-BD60-28F4E78F784B
```

77FA9ABD-0359-4D32-BD60-28F4E78F784B — это GUID частного пространства имен безопасной загрузки.



Безопасный запуск

Если доверительное исполнение (Trusted Execution) включено через определенное значение в политике VSM и система совместима, то Winload активирует новый путь загрузки, который немного отличается от обычного. Этот путь называется безопасным запуском. Он реализует технологию доверительной загрузки (TXT) от Intel — SKINIT на машинах AMD64. Доверительная загрузка реализована в двух компонентах: загрузочной библиотеке и файле TcbLaunch.exe. Загрузочная библиотека при инициализации определяет, что доверительная загрузка включена, и регистрирует обратный вызов загрузки, который перехватывает различные события: запуск приложения загрузки, вычисление хеша и завершение работы приложения загрузки. Загрузчик Windows на ранних этапах выполняет три этапа настройки безопасного запуска вместо загрузки гипервизора. В дальнейшем настройка безопасного запуска будет называться настройкой TCB.

Как уже говорилось, конечной целью безопасного запуска является запуск безопасной последовательности загрузки, в которой процессор является единственным источником информации, которой можно доверять. Для этого система должна

избавиться от всех зависимостей от прошивки. Чтобы добиться этого, Windows создает RAM-диск, отформатированный в файловой системе FAT и содержащий Winload, гипервизор, модуль VSM и все компоненты загрузочной ОС, необходимые для запуска системы. Загрузчик Winload считывает файл TcbLaunch.exe с загрузочного диска системы в память, используя процедуру `BlImgLoadBootApplication`. Та запускает три события, управляемые обратным вызовом загрузки TCB. Обратный вызов сначала подготавливает измеренную среду запуска (Measured Launch Environment, MLE), проверяя модули ACM, таблицу ACPI и отображая необходимые области TXT, затем заменяет точку входа загрузочного приложения на специальную процедуру TXT MLE.

Загрузчик Windows на последних этапах выполнения процедуры `OslExecuteTransition` не начинает последовательность запуска гипервизора. Вместо этого он передает выполнение последовательности запуска TCB, которая довольно проста. Загрузочное приложение TCB запускается с помощью той же процедуры `BlImgStartBootApplication`, которая описана в предыдущем абзаце. Измененная точка входа загрузочного приложения вызывает процедуру запуска TXT MLE, которая выполняет инструкцию `GETSEC(SENTER) TXT`. Эта инструкция измеряет исполняемый файл TcbLaunch.exe в памяти (модуль TBOOT), и если измерение прошло успешно, то процедура запуска MLE передает выполнение кода реальной точке входа загрузочного приложения `TcbMain`.

Функция `TcbMain` — это первый код, выполняемый в среде безопасного запуска. Реализация проста: заново инициализируется библиотека загрузки, регистрируется событие для получения уведомления о запуске или возобновлении виртуализации и вызывается `TcbLoadEntry` из модуля Tcbloader.dll на защищенном RAM-диске. Модуль Tcbloader.dll представляет собой мини-версию доверенного загрузчика Windows. Его задача — загрузить, проверить и запустить гипервизор, настроить страницу гипервызова и запустить безопасное ядро. Безопасный запуск на этом этапе заканчивается, поскольку гипервизор и безопасное ядро берут на себя проверку ядра NT и других модулей, обеспечивая цепочку доверия. Затем выполнение возвращается к загрузчику Windows, который переходит к ядру с помощью стандартной процедуры `OslArchTransferToKernel`.

На рис. 12.10 показаны схема безопасного запуска и все задействованные в ней компоненты. Пользователь может включить безопасный запуск с помощью редактора политики локальной группы, настроив параметр Включить безопасность на основе виртуализации (Turn On Virtualization Based), который находится в разделе Конфигурация компьютера (Computer Configuration) ▸ Административные шаблоны (Administrative Templates) ▸ Система (System) ▸ Защита устройств (Device Guard).

ПРИМЕЧАНИЕ Модули ACM доверенной загрузки предоставляются компанией Intel и зависят от чипсета. Большая часть интерфейса TXT отображается в физической памяти. Это означает, что HvLoader может получить доступ даже к области SINIT, проверить версию SINIT ACM и при необходимости обновить ее. Для этого Windows использует специальный заархивированный WIM-файл Tcbres.wim, который содержит все известные модули SINIT ACM для каждого чипсета. При необходимости этап подготовки MLE открывает заархивированный файл, извлекает нужный модуль в виде двоичного кода и заменяет содержимое исходной прошивки SINIT в области TXT. Когда вызывается процедура безопасного запуска, процессор загружает SINIT ACM в защищенную память, проверяет целостность цифровой подписи и сравнивает хеш открытого ключа с записанным в чипсете.

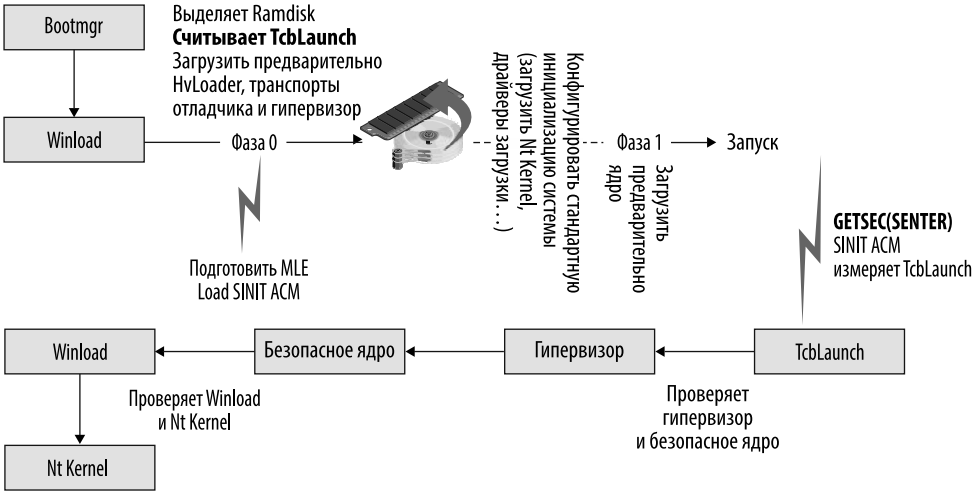


Рис. 12.10. Схема безопасного запуска. Обратите внимание: гипервизор и безопасное ядро запускаются с RAM-диска

Безопасный запуск на платформах AMD

На машинах Intel безопасный запуск возможен благодаря TXT, а обновление Windows 10 Spring 2020 поддерживает SKINIT — аналогичную технологию, разработанную AMD для проверяемого запуска доверенного программного обеспечения, начиная с изначально недоверенного режима работы.

SKINIT имеет ту же цель, что и Intel TXT, и используется для загрузки с помощью безопасного запуска. Но есть и отличия: основой SKINIT является небольшой тип программного обеспечения — безопасный загрузчик (Secure Loader, SL), который в Windows реализован в бинарном файле amdsl.bin, включенном в раздел ресурсов библиотеки Amddrtm.dll, предоставляемой компанией AMD. Инструкция SKINIT переинициализирует процессор для создания безопасной среды выполнения и запускает выполнение SL таким образом, что его невозможно подделать. Защищенный загрузчик находится в блоке безопасного загрузчика — 64-килобайтной структуре, которая передается в TPM инструкцией SKINIT. TPM измеряет целостность SL и передает выполнение в его точку входа.

SL проверяет состояние системы, расширяет измерения в PCR и передает выполнение процедуре запуска AMD MLE, которая находится в отдельном двоичном файле, включенном в модуль TcbLaunch.exe. Подпрограмма MLE инициализирует IDT и GDT и строит таблицу страниц для переключения процессора в длинный режим. MLE на машинах AMD выполняется в 32-битном защищенном режиме, чтобы сделать код в TCB как можно меньше. Наконец, она возвращается в TcbLaunch, который, как и в системах Intel, повторно инициализирует библиотеку загрузки, регистрирует событие для получения сигнала запуска или возобновления виртуализации и вызывает `TcbLoadEntry` из модуля tcbloader.dll. С этого момента ход загрузки идентичен реализации безопасного запуска для систем Intel.

Инициализация ядра и исполнительных подсистем

Вызвав Ntoskrnl, Winload передает ему структуру данных, называемую блоком параметров загрузчика (Loader Parameter Block). Этот блок содержит пути к системному и загрузочному разделам, указатель на таблицы памяти, созданные Winload для описания физической памяти системы, дерево физического оборудования, которое впоследствии будет использовано для создания изменяемого куста реестра HARDWARE, копию куста реестра SYSTEM в памяти и указатель на список загрузочных драйверов, загруженных Winload. Содержит он и другую информацию, связанную с обработкой загрузки, выполненной до этого момента.

ЭКСПЕРИМЕНТ. Блок параметров загрузчика

Во время загрузки ядро хранит указатель на блок параметров загрузчика в переменной KeLoaderBlock. Ядро отбрасывает блок параметров после первой фазы загрузки, поэтому единственный способ увидеть содержимое этой структуры — подключить отладчик ядра перед загрузкой и прерваться в начальной точке его останова. Если удастся это сделать, можно будет использовать команду dt для дампа блока, как показано далее:

```
kd> dt poi(nt!KeLoaderBlock) nt!LOADER_PARAMETER_BLOCK
   +0x000 OsMajorVersion   : 0xa
   +0x004 OsMinorVersion   : 0
   +0x008 Size             : 0x160
   +0x00c OsLoaderSecurityVersion : 1
   +0x010 LoadOrderListHead : _LIST_ENTRY [ 0xfffff800`2278a230 –
0xfffff800`2288c150 ]
   +0x020 MemoryDescriptorListHead : _LIST_ENTRY [ 0xfffff800`22949000 –
0xfffff800`22949de8 ]
   +0x030 BootDriverListHead : _LIST_ENTRY [ 0xfffff800`22840f50 –
0xfffff800`2283f3e0 ]
   +0x040 EarlyLaunchListHead : _LIST_ENTRY [ 0xfffff800`228427f0 –
0xfffff800`228427f0 ]
   +0x050 CoreDriverListHead : _LIST_ENTRY [ 0xfffff800`228429a0 –
0xfffff800`228405a0 ]
   +0x060 CoreExtensionsDriverListHead : _LIST_ENTRY [ 0xfffff800`2283ff20 –
0xfffff800`22843090 ]
   +0x070 TpmCoreDriverListHead : _LIST_ENTRY [ 0xfffff800`22831ad0 –
0xfffff800`22831ad0 ]
   +0x080 KernelStack      : 0xfffff800`25f5e000
   +0x088 Prcb             : 0xfffff800`22acf180
   +0x090 Process          : 0xfffff800`23c819c0
   +0x098 Thread           : 0xfffff800`23c843c0
   +0x0a0 KernelStackSize  : 0x6000
   +0x0a4 RegistryLength   : 0xb80000
   +0x0a8 RegistryBase     : 0xfffff800`22b49000 Void
   +0x0b0 ConfigurationRoot : 0xfffff800`22783090 _CONFIGURATION_COMPONENT_DATA
   +0x0b8 ArcBootDeviceName : 0xfffff800`22785290 "multi(0)disk(0)rdisk(0)
partition(4)"
   +0x0c0 ArcHalDeviceName : 0xfffff800`22785190 "multi(0)disk(0)rdisk(0)
partition(2)"
   +0x0c8 NtBootPathName   : 0xfffff800`22785250 "\WINDOWS\"
   +0x0d0 NtHalPathName    : 0xfffff800`22782bd0 "\"
```

```
    +0x0d8 LoadOptions      : 0xfffff800`22772c80 "KERNEL=NTKRNLMP.EXE NOEXECUTE=OPTIN
                                  HYPERVISORLAUNCHTYPE=AUTO DEBUG ENCRYPTION_KEY=
                                                      **** DEBUGPORT=NET
                                  HOST_IP=192.168.18.48 HOST_PORT=50000 NOVGA"
    +0x0e0 NlsData          : 0xfffff800`2277a450 _NLS_DATA_BLOCK
    +0x0e8 ArcDiskInformation : 0xfffff800`22785e30 _ARC_DISK_INFORMATION
    +0x0f0 Extension        : 0xfffff800`2275cf90 _LOADER_PARAMETER_EXTENSION
    +0x0f8 u                : <unnamed-tag>
    +0x108 FirmwareInformation : _FIRMWARE_INFORMATION_LOADER_BLOCK
    +0x148 OsBootstatPathName : (null)
    +0x150 ArcOSDataDeviceName : (null)
    +0x158 ArcWindowsSysPartName : (null)
```

Также можно использовать команду `!loadermememorylist` для поля `MemoryDescriptorListHead` для сброса диапазонов физической памяти:

```
kd> !loadermemorylist 0xfffff800`22949000
Base         Length      Type
0000000001  0000000005  (26) HALCachedMemory     ( 20 Kb )
0000000006  000000009a  ( 5) FirmwareTemporary   ( 616 Kb )
...
0000001304  0000000001  ( 7) OsloaderHeap        ( 4 Kb )
0000001305  0000000081  ( 5) FirmwareTemporary   ( 516 Kb )
0000001386  000000001c  (20) MemoryData          ( 112 Kb )
...
0000001800  0000000b80  (19) RegistryData        ( 11 Mb 512 Kb )
0000002380  00000009fe  ( 9) SystemCode          ( 9 Mb 1016 Kb )
0000002d7e  0000000282  ( 2) Free                ( 2 Mb 520 Kb )
0000003000  0000000391  ( 9) SystemCode          ( 3 Mb 580 Kb )
0000003391  0000000068  (11) BootDriver          ( 416 Kb )
00000033f9  0000000257  ( 2) Free                ( 2 Mb 348 Kb )
0000003650  00000008d2  ( 5) FirmwareTemporary   ( 8 Mb 840 Kb )
000007ffc9  0000000026  (31) FirmwareData        ( 152 Kb )
000007ffef  0000000004  (32) FirmwareReserved    ( 16 Kb )
000007fff3  000000000c  ( 6) FirmwarePermanent   ( 48 Kb )
000007ffff  0000000001  ( 5) FirmwareTemporary   ( 4 Kb )
NumberOfDescriptors: 90

Summary
Memory Type         Pages
Free                000007a89c   (    501916)   ( 1 Gb 936 Mb 624 Kb )
LoadedProgram       0000000370   (       880)   ( 3 Mb 448 Kb )
FirmwareTemporary   0000001fd4   (      8148)   ( 31 Mb 848 Kb )
FirmwarePermanent   000000030e   (       782)   ( 3 Mb 56 Kb )
OsloaderHeap        0000000275   (       629)   ( 2 Mb 468 Kb )
SystemCode          0000001019   (      4121)   ( 16 Mb 100 Kb )
BootDriver          000000115a   (      4442)   ( 17 Mb 360 Kb )
RegistryData        0000000b88   (      2952)   ( 11 Mb 544 Kb )
MemoryData          0000000098   (       152)   ( 608 Kb )
NlsData             0000000023   (        35)   ( 140 Kb )
HALCachedMemory     0000000005   (         5)   ( 20 Kb )
FirmwareCode        0000000008   (         8)   ( 32 Kb )
FirmwareData        0000000075   (       117)   ( 468 Kb )
FirmwareReserved    0000000044   (        68)   ( 272 Kb )
                    ==========    ==========
Total               000007FFDF   (    524255) = ( ~2047 Mb )
```

Расширение Loader Parameter может показывать полезную информацию об аппаратном обеспечении системы, характеристики процессора и тип загрузки:

```
kd> dt poi(nt!KeLoaderBlock) nt!LOADER_PARAMETER_BLOCK Extension
   +0x0f0 Extension : 0xfffff800`2275cf90 _LOADER_PARAMETER_EXTENSION
kd> dt 0xfffff800`2275cf90 _LOADER_PARAMETER_EXTENSION
nt!_LOADER_PARAMETER_EXTENSION
   +0x000 Size             : 0xc48
   +0x004 Profile          : _PROFILE_PARAMETER_BLOCK
   +0x018 EmInfFileImage   : 0xfffff800`25f2d000 Void
   ...
   +0x068 AcpiTable        : (null)
   +0x070 AcpiTableSize    : 0
   +0x074 LastBootSucceeded : 0y1
   +0x074 LastBootShutdown : 0y1
   +0x074 IoPortAccessSupported : 0y1
   +0x074 BootDebuggerActive : 0y0
   +0x074 StrongCodeGuarantees : 0y0
   +0x074 HardStrongCodeGuarantees : 0y0
   +0x074 SidSharingDisabled : 0y0
   +0x074 TpmInitialized   : 0y0
   +0x074 VsmConfigured    : 0y0
   +0x074 IumEnabled       : 0y0
   +0x074 IsSmbboot        : 0y0
   +0x074 BootLogEnabled   : 0y0
   +0x074 FeatureSettings  : 0y0000000 (0)
   +0x074 FeatureSimulations : 0y000000 (0)
   +0x074 MicrocodeSelfHosting : 0y0
   ...
   +0x900 BootFlags        : 0
   +0x900 DbgMenuOsSelection : 0y0
   +0x900 DbgHiberBoot     : 0y1
   +0x900 DbgSoftRestart   : 0y0
   +0x908 InternalBootFlags : 2
   +0x908 DbgUtcBootTime   : 0y0
   +0x908 DbgRtcBootTime   : 0y1
   +0x908 DbgNoLegacyServices : 0y0
```

Затем Ntoskrnl начинает фазу 0, первую в двухфазном процессе инициализации (фаза 1 — вторая). Большинство исполнительных подсистем имеют функцию инициализации, которая принимает параметр, определяющий, какая фаза выполняется.

Во время фазы 0 прерывания отключены. Ее цель — создать рудиментарные структуры, нужные для вызова сервисов, необходимых в фазе 1. Функция запуска Ntoskrnl, `KiSystemStartup`, вызывается в каждом контексте системного процессора (подробнее об этом позже в этой главе, в разделе «Фаза 1 инициализации ядра»). Она инициализирует загрузочные структуры процессора и создает таблицу глобальных дескрипторов (Global Descriptor Table, GDT) и таблицу дескрипторов прерываний (Interrupt Desctriptor Table, IDT). При вызове из загрузочного процессора процедура запуска инициализирует функции проверки Control Flow Guard (CFG) и совместно с диспетчером памяти инициализирует KASLR. Инициализация KASLR должна выполняться на ранних стадиях запуска системы, тогда ядро может назначить произвольные диапазоны VA для различных областей виртуальной памяти, таких как база данных PFN и системные области PTE. (Более подробная информация о KASLR

есть в разделе «Рандомизация образов» главы 5.) `KiSystemStartup` инициализирует также отладчик ядра, область процессора XSAVE и при необходимости KVA Shadow. Затем вызывается `KiInitializeKernel`. Если `KiInitializeKernel` запущен на загрузочном процессоре, то он выполняет общесистемную инициализацию ядра, например инициализацию внутренних списков и других структур данных, которые совместно используются всеми процессорами. Он создает и уплотняет таблицу дескрипторов системных служб (System Service Descriptor Table, SSDT) и вычисляет случайные значения для внутренних переменных `KiWaitAlways` и `KiWaitNever`, которые применяются для кодирования указателей ядра. Также проверяется, была ли запущена виртуализация, и если была, то создается карта страницы гипервызова и запускаются просветления процессора. (Подробнее о просветлениях гипервизора можно прочитать в главе 9.)

`KiInitializeKernel`, если она выполняется совместимыми процессорами, играет важную роль в инициализации и запуске технологии принудительного контроля (Control Enforcement Technology, CET). Эта аппаратная функция относительно новая, по сути, она реализует аппаратный теневой стек, используемый для обнаружения и предотвращения ROP-атак. Технология задействуется для защиты как приложений пользовательского режима, так и драйверов режима ядра — только при наличии VSM. `KiInitializeKernel` инициализирует Idle-процесс и поток и вызывает `ExpInitializeExecutive`. `KiInitializeKernel` и `ExpInitializeExecutive` обычно выполняются на каждом системном процессоре. Когда он выполняется загрузочным процессором, `ExpInitializeExecutive` полагается на функцию `InitBootProcessor`, отвечающую за организацию фазы 0, в то время как последующие процессоры вызывают только `InitOtherProcessors`.

ПРИМЕЧАНИЕ Ориентированное на возврат программирование (Return-Oriented Programming, ROP) — это техника взлома, при которой злоумышленник получает контроль над стеком вызовов программы с целью перехвата ее потока управления и выполняет тщательно подобранные последовательности машинных инструкций, называемые гаджетами, которые уже присутствуют в памяти машины. Соединяясь вместе, несколько гаджетов позволяют злоумышленнику выполнять произвольные действия на машине.

`InitBootProcessor` начинает работу с проверки загрузчика. Если версия загрузчика, используемая для запуска Windows, не соответствует нужному ядру, функция аварийно завершает работу системы с кодом проверки ошибки `LOADER_BLOCK_MISMATCH` (0x100). В противном случае она инициализирует сторонние указатели пула для начального процессора, проверяет наличие опции загрузки BCD `burnmemory` и отбрасывает тот объем физической памяти, который указан в ее значении. Затем выполняется инициализация файлов NLS, загруженных Winload, достаточная, чтобы обеспечить трансляцию Юникода в ANSI и OEM для работы. Далее происходят инициализация архитектуры аппаратных ошибок Windows (Windows Hardware Error Architecture, WHEA) и вызов функции HAL `HalInitSystem`, которая дает HAL возможность получить контроль над системой, прежде чем Windows выполнит дальнейшую инициализацию. `HalInitSystem` отвечает за инициализацию и запуск различных компонентов HAL, таких как таблицы ACPI, дескрипторы отладчика, DMA, прошивка, MMU ввода-вывода, системные таймеры, топология процессора, счетчики производительности и шина PCI. Одна из важных обязанностей

`HalInitSystem` — подготовка каждого контроллера прерываний процессора к приему прерываний и настройка прерывания интервального таймера, который используется для учета времени работы процессора. (Подробнее об учете времени процессора говорится в разделе «Квантование» главы 4.)

Когда `HalInitSystem` завершает работу, `InitBootProcessor` приступает к вычислению обратного значения для истечения таймера часов. Обратные значения используются для оптимизации деления в большинстве современных процессоров. Они позволяют быстрее выполнять умножение, а поскольку Windows должна разделить текущее 64-битное значение времени, чтобы узнать, какие таймеры должны истечь, этот статический расчет уменьшает задержку прерывания при срабатывании тактового интервала. `InitBootProcessor` использует вспомогательную процедуру `CmInitSystem0` для получения параметров реестра из управляющего вектора куста SYSTEM. Эта структура данных содержит более 150 параметров настройки ядра, являющихся частью раздела реестра HKLM\SYSTEM\CurrentControlSet\Control, включая такую информацию, как данные о лицензировании и версии для установки. Все настройки предварительно загружаются и сохраняются в глобальных переменных. Затем `InitBootProcessor` продолжает установку корневого пути системы и поиск в образе ядра строк сообщений о сбоях, которые он показывает на синих экранах, кэшируя их расположение, чтобы не искать их во время сбоя, что было бы опасно и ненадежно. Далее `InitBootProcessor` инициализирует подсистему таймера и страницу общих пользовательских данных.

Теперь `InitBootProcessor` готов к вызову процедур инициализации фазы 0 для исполнителя, верификатора драйверов и диспетчера памяти. Эти компоненты выполняют следующие задачи инициализации.

1. Исполнитель инициализирует различные внутренние блокировки, ресурсы, списки и переменные и проверяет правильность типа набора программных продуктов в реестре, предотвращая произвольное изменение реестра для обновления до SKU Windows, которая на самом деле не была приобретена. Это лишь одна из многих подобных проверок в ядре.
2. Верификатор драйверов, если он включен, инициализирует различные настройки и поведение в зависимости от параметров проверки и текущего состояния системы, например, от того, включен ли безопасный режим. Он также выбирает, на какие драйверы ориентироваться в тестах со случайным выбором драйверов.
3. Диспетчер памяти создает таблицы страниц, базу данных PFN и внутренние структуры данных, необходимые для предоставления основных услуг памяти. Он также обеспечивает ограничение максимального поддерживаемого объема физической памяти, создает и резервирует область для системного файлового кэша. Затем создает области памяти для страничных и нестраничных пулов, описанных в главе 5. Другие исполнительные подсистемы, ядро и драйверы устройств используют эти два пула памяти для выделения своих структур данных. Наконец, создается UltraSpace — область размером 16 Тбайт, которая обеспечивает поддержку быстрого и недорогого отображения страниц, не требующего сброса на диск TLB.

Далее `InitBootProcessor` включает динамическое разделение процессора для гипервизора, если оно имеется и правильно лицензировано, и вызывает

`HalInitializeBios` для настройки старой части HAL, содержащей код эмуляции BIOS. Этот код применяется для предоставления или эмуляции доступа к 16-битным прерываниям реального режима и памяти, которые в основном используются bootvid — этот драйвер был заменен на BGFX, но все еще существует по соображениям совместимости.

В этот момент `InitBootProcessor` перечисляет загрузочные драйверы, которые были загружены Winload, и вызывает `DbgLoadImageSymbols`, чтобы сообщить отладчику ядра, если он подключен, о необходимости загрузить символы для каждого из этих драйверов. Если хост-отладчик настроил опцию прерывания при загрузке символов, то это будет самая ранняя точка, когда отладчик ядра сможет получить контроль над системой. `InitBootProcessor` теперь вызывает `HvlPhase1Initialize`, выполняющий окончание инициализации HVL, которую невозможно было завершить на предыдущих этапах. Когда функция возвращается, она вызывает `HeadlessInit` для инициализации последовательной консоли, если машина была сконфигурирована для службы управления чрезвычайными ситуациями (Emergency Management Services, EMS).

Далее `InitBootProcessor` создает информацию о версиях, которая будет использоваться позже в процессе загрузки, такую как номер сборки, версия пакета обновления и статус бета-версии. Затем он копирует таблицы NLS, ранее загруженные Winload в страничный пул, переинициализирует их и создает базу данных трассировки стека ядра, если глобальные флаги требуют этого. (Подробнее о глобальных флагах говорится в главе 6.)

Наконец, `InitBootProcessor` вызывает диспетчер объектов, монитор ссылок безопасности, диспетчер процессов, платформу отладки в пользовательском режиме и диспетчер Plug and Play. Эти компоненты выполняют следующие шаги инициализации.

1. Во время инициализации диспетчера объектов определяются объекты, необходимые для построения пространства имен диспетчера, чтобы другие подсистемы могли вставлять в него объекты. Создаются таблицы системных процессов и глобальных дескрипторов ядра, чтобы можно было начать отслеживание ресурсов. Вычисляется значение, используемое для шифрования заголовка объекта, и создаются типы объектов Directory и SymbolicLink.
2. Монитор ссылок безопасности инициализирует глобальные переменные безопасности, например системные SID и Privilege LUID, и базу данных в памяти, а также создает объект типа токена. Затем он создает первый токен локальной системной учетной записи и подготавливает его для назначения начальному процессу. (Описание локальной системной учетной записи см. в главе 7.)
3. Диспетчер процессов выполняет большую часть своей инициализации в фазе 0, определяя типы объектов процессов, потоков, заданий и разделов и устанавливая списки для отслеживания активных процессов и потоков. Общесистемные параметры устранения рисков процессов инициализируются и объединяются с параметрами, указанными в параметре реестра HKLM\SYSTEM\CurrentControlSet\Control\Session Manager\Kernel\MitigationOptions. Затем диспетчер процессов создает объект раздела исполнительной системы, который называется `MemoryPartition0`. Название вводит в заблуждение, поскольку на самом деле этот объект является объектом исполнительного раздела — новым типом объекта Windows, который

включает в себя раздел памяти и раздел диспетчера кэша для поддержки новых контейнеров приложений.

4. Диспетчер процессов создает также объект процесса для начального процесса и называет его `Idle`. В качестве последнего шага диспетчер процессов создает защищенный процесс `System` и системный поток для выполнения процедуры `Phase1Initialization`. Этот поток запускается не сразу, поскольку прерывания все еще отключены. Процесс `System` создается как защищенный, чтобы получить защиту от атак в пользовательском режиме, поскольку его виртуальное адресное пространство применяется для отображения конфиденциальных данных, задействуемых системой и драйвером целостности кода. Кроме того, в таблице дескрипторов системного процесса хранятся дескрипторы ядра.
5. Структура отладки в пользовательском режиме создает определение типа объекта отладки, который применяется для прикрепления отладчика к процессу и получения событий отладки. (Дополнительные сведения об отладке в пользовательском режиме см. в главе 8.)
6. Затем происходит инициализация диспетчера Plug and Play в фазе 0, которая включает в себя инициализацию исполнительного ресурса, применяемого для синхронизации доступа к ресурсам шины.

Когда управление возвращается в `KiInitializeKernel`, последним шагом будет выделение стека DPC для текущего процессора, повышение IRQL до уровня диспетчеризации и включение прерываний. Затем управление переходит в цикл `Idle`, который заставляет системный поток, созданный на четвертом шаге, начать выполнение фазы 1. Вторичные процессоры ожидают начала своей инициализации до шага 11 фазы 1 (см. следующий раздел).

Фаза 1 инициализации ядра

Как только поток `Idle` получает возможность выполнения, начинается фаза 1 инициализации ядра. Она состоит из следующих этапов.

1. `Phase1InitializationDiscard`, как следует из названия, удаляет код, являющийся частью секции INIT образа ядра, чтобы сохранить память.
2. Поток инициализации устанавливает свой приоритет на 31, самый высокий из возможных, чтобы предотвратить перехват.
3. Оценивается параметр BCD, задающий максимальное количество виртуальных процессоров `hypervisorrootproc`.
4. Создаются топологические связи «NUMA — группа», с помощью которых система пытается найти наиболее оптимизированное отображение между логическими процессорами и процессорными группами, принимая во внимание локальность и расстояния NUMA, если это не отменяется соответствующими настройками BCD.
5. `HalInitSystem` выполняет фазу 1 инициализации системы и подготавливает последнюю к приему прерываний от внешних периферийных устройств.
6. Инициализируется прерывание системных часов и включается генерация их тактовых импульсов.

7. Инициализируется старый загрузочный видеодрайвер bootvid. Он используется только для выдачи отладочных сообщений и сообщений собственных приложений, которые запускаются SMSS, такими как NT chkdsk.
8. Ядро строит различные строки и информацию о версии, их показывают на загрузочном экране через bootvid, если была включена опция загрузки `sos`. Сюда входят полная информация о версии, количество поддерживаемых процессоров и объем поддерживаемой памяти.
9. Вызывается инициализация диспетчера питания.
10. Системное время инициализируется вызовом `HalQueryRealTimeClock`, а затем сохраняется как время загрузки системы.
11. В многопроцессорной системе оставшиеся процессоры инициализируются с помощью `KeStartAllProcessors` и `HalAllProcessorsStarted`. Количество инициализированных и поддерживаемых процессоров зависит от сочетания фактического физического количества, лицензионной информации для установленной SKU Windows, параметров загрузки, таких как `numproc` и `bootproc`, а в серверных системах также от того, включено ли динамическое разбиение на разделы. После инициализации всех доступных процессоров привязка системного процесса обновляется, чтобы включить все процессоры.
12. Диспетчер объектов инициализирует глобальный системный бункер, списки и сторонние дескрипторы для каждого процессора, а также базовый аудит, если он разрешен вектором управления системой. Затем он создает корневой каталог пространства имен \, каталог \KernelObjects, каталог \ObjectTypes и каталог сопоставления имен устройств DOS \Global??, в котором создаются ссылки Global и GLOBALROOT. Затем диспетчер объектов создает карту устройства системного бункера, которая будет управлять сопоставлением имен устройств DOS и прикреплять его к системному процессу. Он создает старую символическую ссылку \DosDevices, сохраняемую по соображениям совместимости, которая указывает на каталог сопоставления имен устройств подсистемы Windows. Наконец, диспетчер объектов вставляет каждый зарегистрированный тип объекта в объект \ObjectTypes каталога.
13. Исполнитель вызывается для создания типов исполнительных объектов, включая семафор, мьютекс, событие, таймер, событие с ключом, блокировку push и рабочий пул потоков.
14. Диспетчер ввода-вывода вызывается для создания типов объектов диспетчера ввода-вывода, включая объекты устройств, драйверов, контроллеров, адаптеров, завершения ввода-вывода, завершения ожидания и файлов.
15. Ядро инициализирует системные контроллеры. Существует два основных типа контроллеров: контроллер DPC, который проверяет, чтобы процедура DPC не выполнялась дольше определенного времени, и контроллер CPU Keep Alive, который проверяет, чтобы каждый процессор всегда отвечал на запросы. Контроллеры не инициализируются, если система выполняется гипервизором.
16. Ядро инициализирует структуру данных каждого блока управления процессором (KPRCB), вычисляет массив затрат NUMA и, наконец, рассчитывает длительность системного такта и кванта времени.

17. Библиотека отладчика ядра завершает инициализацию отладочных настроек и параметров независимо от того, был ли отладчик запущен до этого момента.
18. Диспетчер транзакций создает свои типы объектов, такие как фиксация, управление ресурсами, а также типы диспетчеров транзакций.
19. Структуры данных отладочной библиотеки пользовательского режима Dbgk инициализируются для глобального системного бункера.
20. Если верификатор драйверов включен и в зависимости от опций активна верификация пула, то для системного процесса запускается трассировка дескрипторов объектов.
21. Монитор ссылок безопасности создает каталог \Security в пространстве имен диспетчера объектов, защищая его дескриптором безопасности, в котором только учетная запись SYSTEM имеет полный доступ, и инициализирует структуры данных аудита, если тот включен. Кроме того, монитор ссылок безопасности инициализирует библиотеку SDDL режима ядра и создает событие, которое будет сигнализировано после инициализации LSA, \Security\LSA_AUTHENTICATION_INITIALIZED. Наконец, Security Reference Monitor впервые инициализирует компонент Kernel Code Integrity, Ci.dll, вызывая внутреннюю процедуру `CiInitialize`, которая инициализирует все обратные вызовы Code Integrity Callbacks и сохраняет список загрузочных драйверов для дальнейшего аудита и проверки.
22. Диспетчер процессов создает системный дескриптор для исполнительного системного раздела. Он никогда не будет разыменован, поэтому системный раздел не может быть уничтожен. Затем диспетчер процессов инициализирует поддержку дополнительного расширения ядра (подробнее об этом говорится на шаге 26). Он регистрирует вызовы хоста для различных служб ОС, таких как Background Activity Moderator (BAM), Desktop Activity Moderator (DAM), Multimedia Class Scheduler Service (MMCSS), Kernel Hardware Tracing и Windows Defender System Guard. Наконец, если VSM включен, он создает первый минимальный процесс — системный процесс IUM и присваивает ему название Secure System.
23. Создается символическая ссылка \SystemRoot.
24. Диспетчер памяти вызывается для выполнения фазы 1 его инициализации. Эта фаза создает тип объекта Section, инициализирует все связанные с ним структуры данных, например область управления, и создает объект секции \Device\PhysicalMemory. Затем инициализируется поддержка Control Flow Guard (CFG) ядра и создаются секции с поддержкой страничного файла, которые будут применяться для описания битовой карты CFG пользовательского режима. Подробнее о Control Flow Guard читайте в главе 7. Диспетчер памяти инициализирует поддержку анклава памяти для SGX-совместимых систем, поддержку горячих обновлений, структуры данных, объединяющие страницы, и системные события памяти. Наконец, он порождает три рабочих системных потока диспетчера памяти — Balance Set Manager, Process Swapper и Zero Page Thread, о которых рассказывается в главе 5, и создает секционный объект, используемый для отображения буфера памяти схемы API Set в системном пространстве, которое было предварительно выделено загрузчиком Windows. Только что созданные системные потоки имеют возможность выполниться позже, в конце фазы 1.

25. Таблицы NLS отображаются в системное пространство, чтобы их легко могли отобразить процессы пользовательского режима.
26. Диспетчер кэша инициализирует структуры данных кэша файловой системы и создает свои рабочие потоки.
27. Диспетчер конфигурации создает ключевой объект \Registry в пространстве имен диспетчера объектов и открывает куст SYSTEM в памяти как обычный файл куста. Затем он копирует исходные данные дерева оборудования, переданные Winload, в изменяемый куст HARDWARE.
28. Система инициализирует дополнительные расширения ядра. Эта функция была введена в Windows 8.1 с целью экспорта частных системных компонентов и данных загрузчика Windows (например, требований к кэшированию памяти, указателей служб исполнительной среды UEFI, карты памяти UEFI, данных SMBIOS, политик безопасной загрузки и данных о целостности кода) в различные компоненты ядра (например, безопасное ядро) без использования стандартного экспорта портативных исполняемых программ.
29. Диспетчер ошибок инициализирует и сканирует реестр на наличие информации об ошибках, а также базу данных INF (файл установки драйвера, описан в главе 6), содержащую ошибки для различных драйверов.
30. Выполняется обработка настроек, связанных с производителем. Режим производителя — это специальный режим операционной системы, который можно использовать для выполнения задач, связанных с производством, например для тестирования компонентов и поддержки. Эта функция задействуется, в частности, в мобильных системах и обеспечивается подсистемой UEFI. Если прошивка указывает операционной системе через специальный протокол UEFI, что этот специальный режим включен, то Windows считывает и записывает всю необходимую информацию из раздела реестра HKLM\System\CurrentControlSet\Control\ManufacturingMode.
31. Инициализируются `Superfetch` и `prefetcher`.
32. Инициализируется диспетчер виртуальных хранилищ ядра. Этот компонент является частью системы сжатия памяти.
33. Инициализируется компонент VM. Он представляет собой опциональное расширение ядра, используемое для связи с гипервизором.
34. Инициализируется и устанавливается информация о текущем часовом поясе.
35. Инициализируются глобальные структуры данных драйвера файловой системы.
36. Инициализируется механизм сжатия NT Rtl.
37. Поддержка отладчика гипервизора, если она необходима, настраивается таким образом, чтобы остальная часть системы не использовала собственное устройство.
38. Фаза 1 информации, специфичной для транспорта отладчика, выполняется вызовом процедуры `KdDebuggerInitialize1` в зарегистрированном транспорте, таком как Kdcom.dll.
39. Подсистема расширенного локального вызова процедур (Advanced Local Procedure Call, ALPC) инициализирует объекты типа порта ALPC и типа порта с ожиданием ALPC. Старые объекты LPC устанавливаются в качестве псевдонимов.

40. Если система была загружена с протоколированием загрузки с опцией BCD `bootlog`, то инициализируется файл журнала загрузки. Если система была загружена в безопасном режиме, то выясняется, нужно ли запускать альтернативную оболочку, как в случае загрузки в безопасном режиме с командной строкой.
41. Исполнитель вызывается для выполнения второй фазы инициализации, в которой он настраивает часть функциональности лицензирования Windows в ядре, например, проверяет параметры реестра, в которых хранятся данные лицензии. Кроме того, если присутствуют постоянные данные от загрузочных приложений, например результаты диагностики памяти или информация о возобновлении работы из спящего режима, то соответствующие файлы журналов и информация записываются на диск или в реестр.
42. Создаются разделы реестра MiniNT/WinPE, если это загрузка такого рода, и в пространстве имен создается каталог объектов NLS, который в дальнейшем будет использоваться для размещения объектов секций для различных файлов NLS, отображаемых в памяти.
43. Политики целостности кода ядра Windows, например список доверенных подписчиков и хеши сертификатов, и параметры отладки инициализируются, а все соответствующие настройки копируются из блока загрузчика в модуль CI ядра Ci.dll.
44. Диспетчер питания снова вызывается для инициализации. На этот раз он устанавливает поддержку запросов на питание, контроллеров питания, канала ALPC для уведомлений о яркости, а также обратных вызовов профиля.
45. Теперь происходит инициализация диспетчера ввода-вывода. Этот этап — сложная фаза запуска системы, на которую приходится большая часть времени загрузки. Сначала диспетчер ввода-вывода инициализирует различные внутренние структуры и создает типы объектов драйвера и устройства, а также свои корневые каталоги: \Driver, \FileSystem, \FileSystem\Filters и \UMDFCommunicationPorts для системы драйвера UMDF. Затем инициализируется Kernel Shim Engine, вызываются диспетчер Plug and Play, диспетчер питания и HAL, чтобы начать различные этапы динамического перечисления и инициализации устройств. (Все подробности этого сложного процесса рассмотрены в главе 6.) Затем инициализируется подсистема инструментария управления Windows (Windows Management Instrumentation, WMI), которая обеспечивает поддержку WMI для драйверов устройств. (Дополнительные сведения см. в разделе «Инструментарий управления Windows» главы 10.) Также инициализируется отслеживание событий Windows (Event Tracing for Windows, ETW) и записываются все события ETW с постоянными данными загрузки, если таковые имеются. Диспетчер ввода-вывода запускает специфичный для платформы драйвер ошибок и инициализирует глобальную таблицу источников аппаратных ошибок. Эти два компонента жизненно важны для инфраструктуры аппаратных ошибок Windows. Затем он выполняет первый вызов безопасного ядра и просит его выполнить последний этап инициализации в VTL 1. Также инициализируется драйвер зашифрованного безопасного дампа путем считывания части его конфигурации из реестра Windows HKLM\System\CurrentControlSet\Control\CrashControl. Все загрузочные драйверы перечисляются и упорядочиваются с учетом их

зависимостей и очередности загрузки. (Подробности обработки информации о контроле загрузки драйверов в реестре также рассматриваются в главе 6.) Все связанные библиотеки DLL режима ядра инициализируются встроенным драйвером файловой системы RAW. На этом этапе диспетчер ввода-вывода отображает Ntdll.dll, Vertdll.dll и WOW64-версию Ntdll в системное адресное пространство. Наконец, вызываются все драйверы начальной загрузки для выполнения их инициализации, а затем запускаются драйверы устройств начальной загрузки системы. Имена устройств подсистемы Windows создаются как символические ссылки в пространстве имен диспетчера объектов.

46. Диспетчер конфигурации регистрирует и запускает собственный провайдер журналирования трассировки ETW для реестра Windows. Это делает возможной трассировку всего диспетчера конфигурации.
47. Диспетчер транзакций устанавливает препроцессор трассировки программного обеспечения Windows (WPP) и регистрирует своего провайдера ETW.
48. Теперь, когда загрузочные и системные драйверы загружены, диспетчер ошибок загружает базу данных INF с ошибками драйверов и начинает их разбор, куда входит применение обходных путей конфигурации PCI в реестре.
49. Если компьютер загружается в безопасном режиме, то этот факт фиксируется в реестре.
50. Если это явно не запрещено в реестре, то включается подкачка кода в режиме ядра в Ntoskrnl и драйверах.
51. Диспетчер питания вызывается для завершения его инициализации.
52. Инициализируется поддержка тактового таймера ядра.
53. Перед тем как INIT-секция Ntoskrnl будет удалена, в частный системный раздел копируется остальная лицензионная информация системы, включая текущие настройки политик, которые хранятся в реестре. Затем устанавливается время истечения срока действия системы.
54. Диспетчер процессов вызывается для настройки ограничения скорости выполнения заданий и времени создания системного процесса. Он инициализирует статическое окружение для защищенных процессов и ищет различные системные точки входа в системных библиотеках пользовательского режима, ранее отображенных диспетчером ввода-вывода. Обычно это Ntdll.dll, Ntdll32.dll и Vertdll.dll.
55. Монитор контроля безопасности вызывается для создания потока командного сервера, который взаимодействует с LSASS. На этом этапе создается командный порт монитора, используемый LSA для отправки команд на SRM. (Дополнительные сведения о том, как обеспечивается безопасность в Windows, см. в разделе «Компоненты системы безопасности» главы 7.)
56. Если VSM включен, то зашифрованные ключи VSM сохраняются на диске. Системные библиотеки пользовательского режима отображаются в процесс Secure System Process. В этом подходе защищенное ядро получает всю необходимую информацию о системных библиотеках DLL VTL 0.
57. Запускается процесс Session Manager Smss, представленный в главе 2. Smss отвечает за создание среды пользовательского режима, которая обеспечивает

видимый интерфейс Windows — его инициализация рассматривается в следующем разделе.

58. Драйвер bootvid включается для того, чтобы инструмент проверки диска NT мог показывать текстовый вывод.
59. Запрашиваются значения энтропии загрузки TPM. Они могут быть запрошены только один раз за загрузку, и обычно к этому моменту системный драйвер TPM должен был запросить их, но если он не был запущен по какой-то причине — возможно, пользователь отключил его, то незапрошенные значения все еще будут доступны. Поэтому ядро вручную запрашивает их, чтобы избежать этой ситуации. В нормальных сценариях собственный запрос ядра не дал бы результатов.
60. Теперь освобождена вся память, используемая блоком параметров загрузчика, и все его ссылки, например код инициализации Ntoskrnl и всех загрузочных драйверов, которые находятся в секциях INIT.

В качестве последнего шага перед тем, как считать инициализацию исполнителя и ядра завершенной, поток инициализации фазы 1 устанавливает флаг критического прерывания при завершении для нового процесса Smss. В результате, если процесс Smss по какой-то причине завершается, ядро перехватывает это, входит в подключенный отладчик, если таковой имеется, и аварийно завершает систему с кодом останова `CRITICAL_PROCESS_DIED`.

Если пятисекундное ожидание закончилось, то есть прошло 5 с, считается, что диспетчер сеансов успешно запущен и поток инициализации фазы 1 завершается. Таким образом, процессор загрузки выполняет один из системных потоков диспетчера памяти, созданных на шаге 22, или возвращается в цикл `Idle`.

Smss, Csrss и Wininit

Smss похож на любой другой процесс пользовательского режима, но у него есть два отличия. Во-первых, Windows считает Smss доверенной частью операционной системы. Во-вторых, это *собственное* приложение системы. Поскольку Smss является доверенным компонентом операционной системы, он работает как защищенный процесс (protected process light, PPL; рассматриваются в главе 3) и может выполнять действия, недоступные другим процессам, например создавать токены безопасности. Поскольку Smss — собственное приложение, он использует не Windows API, а только основные исполнительные API, известные как собственные API Windows, которые обычно открываются Ntdll. Smss не применяет Win32 API, потому что подсистема Windows не выполняется при запуске Smss. На самом деле одной из первых задач Smss и является запуск подсистемы Windows.

Инициализация Smss рассматривалась в разделе «Диспетчер сеансов» главы 2. Обращайтесь к ней за всеми подробностями инициализации. Когда главный Smss создает дочерние процессы Smss, он передает им в качестве параметров заголовки двух объектов секций. Эти два объекта представляют собой общие буферы, используемые для обмена данными между несколькими экземплярами Smss и Csrss. Один применяется для связи между родительским и дочерними процессами Smss, а другой — для связи с процессом клиентской подсистемы. Главный Smss порождает

дочерний процесс с помощью процедуры `RtlCreateUserProcess`, устанавливая флаг, чтобы приказать диспетчеру процессов создать новую сессию. В этом случае функция ядра `PspAllocateProcess` вызывает диспетчер памяти для создания нового адресного пространства сессии.

Имя исполняемого файла, который запускает дочерний Smss в конце своей инициализации, хранится в общей секции, и, как говорилось в главе 2, обычно это Wininit.exe для сессии 0 и Winlogon.exe для любых интерактивных сессий. Важно помнить, что перед тем, как новый Smss сеанса 0 запустит Wininit, он подключается к главному Smss через порт `SmApiPort` ALPC и загружает и инициализирует все подсистемы.

Диспетчер сессий получает привилегию загрузочного драйвера и просит ядро загрузить и отобразить драйвер Win32k в новое адресное пространство сессии, используя собственный API `NtSetSystemInformation`. Затем он запускает процесс клиент-серверной подсистемы Csrss.exe, указывая в командной строке имя корневого каталога объектов `\Windows`, заголовки объектов разделяемых разделов, имя подсистемы (Windows) и следующие DLL-библиотеки подсистемы:

- **Basesrv.dll** — серверная часть процесса подсистемы;
- **Sxssrv.dll** — модуль расширения поддержки подсистемы;
- **Winsrv.dll** — модуль поддержки многопользовательской подсистемы.

Процесс клиент-серверной подсистемы выполняет частичную инициализацию: он включает некоторые опции упорядочивания процессов, удаляет ненужные привилегии из своего маркера, запускает собственный провайдер ETW и инициализирует связанный список структур данных `CSR_PROCESS` для отслеживания всех процессов Win32, которые будут запущены в системе. Затем разбирает свою командную строку, захватывает заголовки общих секций и создает два порта ALPC.

- **Командный порт CSR API** (\Sessions\<ID>\Windows\ApiPort). Будет задействоваться каждым процессом Win32 для связи с подсистемой Csrss. Kernelbase.dll подключается к нему в ходе своей процедуры инициализации.
- **Порт API диспетчера сеансов подсистемы** (\Sessions\<ID>\Windows\SbApiPort). Используется диспетчером сеансов для отправки команд в Csrss.

Csrss создает два потока, применяемых для диспетчеризации команд, полученных через порты ALPC. В итоге он соединяется с диспетчером сессий через другой порт ALPC, \SmApiPort, который ранее был создан в ходе инициализации Smss — на шаге 6 процедуры инициализации, описанной в главе 2. Во время соединения процесс Csrss отправляет Session Manager имя только что созданного API-порта. С этого момента можно начинать новые интерактивные сеансы. В конце главный поток Csrss завершается.

После порождения процесса подсистемы дочерний Smss запускает начальный процесс — Wininit или Winlogon, а затем выходит из системы. Активным остается только главный экземпляр Smss. Главный поток Smss постоянно ожидает заголовок процесса Csrss, в то время как другие потоки ALPC ожидают сообщений о создании новых сеансов или подсистем. Если Wininit или Csrss неожиданно завершаются, то

ядро аварийно завершает систему, поскольку эти процессы помечены как *критические*. Если Winlogon неожиданно завершается, связанная с ним сессия выходит из системы.

ЗАПЛАНИРОВАННЫЕ ОПЕРАЦИИ ПЕРЕИМЕНОВАНИЯ ФАЙЛОВ Тот факт, что исполняемые образы и библиотеки DLL при использовании размещаются в памяти, делает невозможным обновление основных системных файлов после завершения загрузки Windows, если только не применяется технология горячего обновления, но это касается только патчей Microsoft для операционной системы. В API Windows MoveFileEx есть возможность указать, что перемещение файла должно быть отложено до следующей загрузки. Пакеты обновлений и исправления, которые должны обновлять отображенные в памяти используемые файлы, устанавливают файлы-заменители в систему во временные позиции и задействуют API MoveFileEx, чтобы они заменяли применяемые файлы. При использовании этой опции MoveFileEx просто записывает команды в параметры PendingFileRenameOperations и PendingFileRenameOperations2 в разделе HKLM\SYSTEM\CurrentControlSet\Control\Session Manager. Эти параметры имеют тип MULTI_SZ, где каждая операция задается парами имен файлов: первое — это исходное расположение, второе — целевое расположение. Операции удаления используют пустую строку в качестве целевого пути. Для просмотра зарегистрированных отложенных команд переименования и удаления можно взять утилиту Pendmoves из пакета Windows Sysinternals (https://docs.microsoft.com/en-us/sysinternals/).

Wininit выполняет шаги по запуску, как описано в разделе «Процесс инициализации Windows» главы 2, например, создает начальную оконную станцию и объекты рабочего стола. Он также устанавливает пользовательское окружение, запускает RPC-сервер выключения и интерфейс WSI (подробнее см. раздел «Выключение» далее) и создает процесс диспетчера управления службами (Service Control Manager, SCM) Services.exe, который загружает все службы и драйверы устройств, помеченные для автоматического запуска. В это время запускается служба локального диспетчера сеансов Lsm.dll, которая работает в общем процессе Svchost. Далее Wininit проверяет, не было ли предшествующего сбоя системы, и если был, то вырезает дамп сбоя и запускает процесс доклада об ошибках Windows werfault.exe для дальнейшей обработки. Наконец, он запускает службу Local Security Authentication Subsystem Service %SystemRoot%\System32\Lsass.exe, а если включен Credential Guard, то запускает Isolated LSA Trustlet, Lsaiso.exe и постоянно ждет запроса на выключение системы.

Для сеанса 1 и прочих запускается Winlogon. В то время как Wininit создает неинтерактивную оконную станцию для сеанса 0, Winlogon создает станцию для интерактивного сеанса по умолчанию, называемую WinSta0, и два рабочих стола: защищенный рабочий стол Winlogon и рабочий стол пользователя по умолчанию. Затем Winlogon запрашивает информацию о загрузке системы с помощью API `NtQuerySystemInformation` — только при первом интерактивном сеансе входа в систему. Если конфигурация загрузки включает изменяемый флаг меню выбора ОС, то запускается система GDI, порождающая хост-процесс UMDF fontdrvhost.exe, и запускается современное приложение загрузочного меню Bootim.exe. Изменяемый флаг меню выбора ОС устанавливается на ранних этапах загрузки программой Bootmgr только в том случае, если ранее была обнаружена мультизагрузочная среда (подробнее см. раздел «Меню загрузки» в начале этой главы).

Bootim — это приложение с графическим интерфейсом, которое отрисовывает современное меню загрузки. Новая современная загрузка задействует подсистему Win32 (графический драйвер и вызовы GDI+) с целью поддержки высоких разрешений для показа вариантов загрузки и дополнительных опций. Поддерживаются даже сенсорные экраны, так что пользователь может выбрать, какую операционную систему запустить, простым прикосновением. Winlogon запускает новый процесс Bootim и ожидает его завершения. Когда пользователь делает выбор, Bootim завершает работу. Winlogon проверяет код выхода. Таким образом он может определить, выбрал ли пользователь операционную систему или средство загрузки или просто запросил завершение работы системы. Если пользователь выбрал операционную систему, отличную от текущей, то Bootim добавляет однократную опцию BCD `bootsequence` в основное хранилище загрузки системы (подробнее о хранилище BCD см. в разделе «Диспетчер загрузки Windows» ранее в этой главе). Новая последовательность загрузки распознается (а опция BCD удаляется) диспетчером загрузки Windows после того, как Winlogon перезагрузит машину с помощью API `NtShutdownSystem`. Winlogon помечает предыдущую загрузочную запись как хорошую перед перезагрузкой системы.

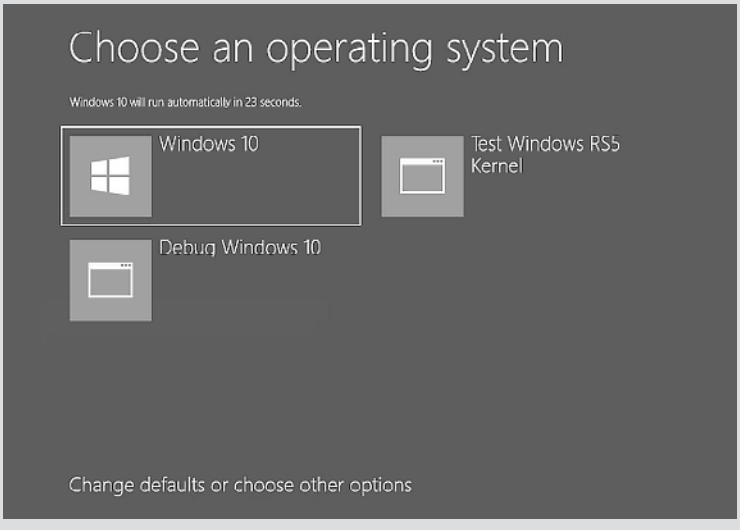
ЭКСПЕРИМЕНТ. Исследование современного загрузочного меню

Современное приложение загрузочного меню, вызываемое Winlogon после запуска Csrss, — на самом деле классическое приложение Win32 GUI, что этот эксперимент и продемонстрирует. Здесь лучше начать с правильно настроенной мультизагрузочной системы, в противном случае не получится увидеть несколько пунктов современного меню загрузки.

Откройте консольное окно, набрав `cmd` в строке поиска меню Пуск, и перейдите по пути \Windows\System32 к загрузочному тому, набрав `cd /d C:\Windows\System32`, где `C` — буква загрузочного тома. Затем введите `Bootim.exe` и нажмите Enter. Должно появиться окно, похожее на современное загрузочное меню, в котором будет только опция Выключить компьютер (Turn Off Your Computer). Это связано с тем, что процесс Bootim был запущен под стандартным неадминистративным токеном, который создается для системы контроля над пользовательскими учетными записями. Процесс попросту не может получить доступ к данным конфигурации загрузки системы. Нажмите Ctrl+Alt+Delete, чтобы запустить диспетчер задач и завершить процесс BootIm, или просто выберите вариант Выключить компьютер (Turn Off Your Computer). Реальный процесс выключения запускается не BootIm, а вызывающим процессом, которым в исходной последовательности загрузки является Winlogon.

Теперь необходимо запустить консольное окно с правами администратора, щелкнув правой кнопкой мыши на его значке на панели задач или на элементе Командная строка (Command Prompt) в строке поиска Windows и выбрав пункт Запустить от имени администратора (Run As Administrator). В новой административной подсказке запустите исполняемый файл BootIm. На этот раз появится

настоящее современное загрузочное меню со всеми опциями и инструментами загрузки, похожее на показанное на следующем рисунке.



Во всех остальных случаях Winlogon ждет инициализации процесса LSASS и службы LSM. Затем он порождает новый экземпляр процесса DWM (Desktop Windows Manager — компонент, используемый для отрисовки современного графического интерфейса) и загружает зарегистрированных поставщиков учетных данных для системы (по умолчанию поставщик учетных данных Microsoft поддерживает вход в систему на основе пароля, PIN-кода и биометрии) в дочерний процесс LogonUI %SystemRoot%\System32\Logonui.exe, отвечающий за показ интерфейса входа в систему. (Подробнее о последовательности запуска Wininit, Winlogon и LSASS рассказано в разделе «Инициализация Winlogon» главы 7.)

После запуска процесса LogonUI Winlogon запускает свой внутренний автомат конечных состояний. Он используется для управления всеми возможными состояниями, возникающими при различных типах входа в систему, таких как стандартный интерактивный вход, терминальный сервер, быстрое переключение пользователей и загрузка из спящего режима. При стандартном интерактивном входе Winlogon показывает экран приветствия и ожидает уведомления об интерактивном входе от поставщика учетных данных, при необходимости настраивая последовательность SAS. Когда пользователь вводит свои учетные данные — это может быть пароль, PIN-код или биометрическая информация, — Winlogon создает LUID сеанса входа в систему и проверяет вход с помощью пакетов аутентификации, зарегистрированных в Lsass — процессе, о котором подробнее говорилось в разделе «Шаги входа пользователя в систему» главы 7. Даже если аутентификация не удалась, Winlogon на этом этапе отмечает текущую загрузку как удачную. Если аутентификация прошла успешно, то Winlogon проверяет сценарий последовательного входа в случае клиентских SKU, при котором каждый раз может быть создана только одна сессия,

а если это не так и активна другая сессия, то спрашивает пользователя, как действовать дальше. Затем он загружает куст реестра из профиля пользователя, входящего в систему, отображая его в HKCU. Он добавляет необходимые ACL к Windows Station и Desktop нового сеанса и создает переменные окружения пользователя, которые хранятся в HKCU\Environment.

Далее Winlogon дожидается процесса Sihost и запускает оболочку, запуская исполняемый файл или файлы, указанные в разделе HKLM\SOFTWARE\Microsoft\Windows NT\CurrentVersion\WinLogon\Userinit (несколько исполняемых файлов разделяются запятыми), который по умолчанию указывает на \Windows\System32\Userinit.exe. Новый процесс Userinit будет жить на рабочем столе Winsta0\Default. Userinit.exe выполняет следующие действия.

1. Создает раздел сеанса Проводника HKCU\Software\Microsoft\Windows\CurrentVersion\Explorer\SessionInfo\.
2. Обрабатывает пользовательские сценарии, указанные в HKCU\Software\Policies\Microsoft\Windows\System\Scripts, и сценарии входа в систему в HKLM\SOFTWARE\Policies\Microsoft\Windows\System\Scripts. Поскольку машинные сценарии запускаются после пользовательских, они могут отменять настройки пользователя.
3. Запускает оболочку или оболочки, разделенные запятыми, указанные в HKCU\Software\Microsoft\Windows NT\CurrentVersion\Winlogon\Shell. Если этого параметра не существует, то Userinit.exe запускает оболочку или оболочки, указанные в HKLM\SOFTWARE\Microsoft\Windows NT\CurrentVersion\Winlogon\Shell, — по умолчанию Explorer.exe.
4. Если в групповой политике указана квота профиля пользователя, то запускается файл %SystemRoot%\System32\Proquota.exe, чтобы установить квоту для текущего пользователя.

Затем Winlogon уведомляет зарегистрированных сетевых провайдеров о том, что пользователь вошел в систему, запуская процесс mpnotify.exe. Сетевой провайдер Microsoft Multiple Provider Router, %SystemRoot%\System32\Mpr.dll, восстанавливает постоянное отображение букв дисков и принтеров пользователя, хранящихся в HKCU\Network и HKCU\Printers соответственно. На рис. 12.11 показано дерево процессов, видимое на мониторе процессов после входа в систему, с использованием возможности ведения журнала загрузки. Обратите внимание на затемненные процессы Smss — это означает, что они уже завершились. Они относятся к порожденным копиям, которые инициализируют каждый сеанс.

ReadyBoot

Если в системе менее 400 Мбайт свободной памяти, то Windows применяет стандартный логический префетчер загрузки (описан в главе 5), но если в системе 400 Мбайт и более свободной оперативной памяти, то она использует кэш в памяти для оптимизации процесса загрузки. Размер кэша зависит от общего объема доступной оперативной памяти, но он довольно велик для того, чтобы создать кэш достаточного размера и при этом предоставить системе память, необходимую для плавной загрузки. ReadyBoot реализован в двух отдельных бинарных файлах: драйвере ReadyBoost (Rdyboost.sys) и службе Sysmain (Sysmain.dll, которая также реализует SuperFetch).

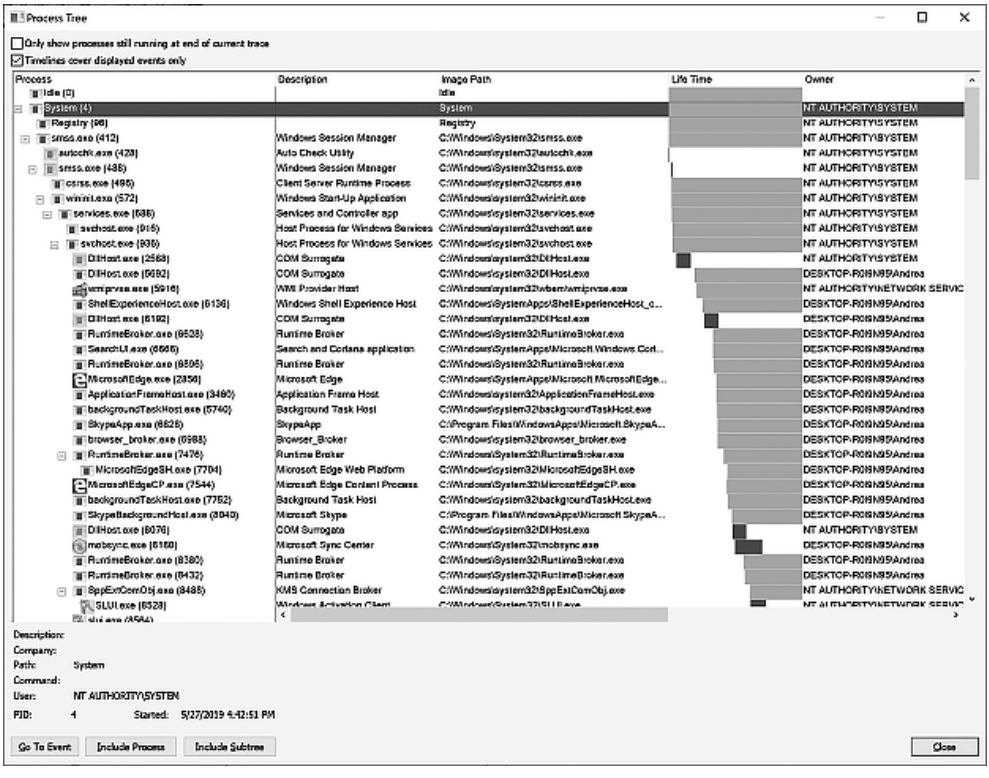


Рис. 12.11. Дерево процессов при входе в систему

Кэш реализуется диспетчером хранилища в том же драйвере устройства, который выполняет кэширование ReadyBoost, Rdyboost.sys, но наполнение кэша определяется планом загрузки, предварительно сохраненным в реестре. Хотя загрузочный кэш может быть сжат, как и кэш ReadyBoost, еще одно различие между управлением кэшем ReadyBoost и ReadyBoot заключается в том, что в режиме ReadyBoot кэш не шифруется. Служба ReadyBoost удаляет кэш через 50 с после запуска службы или если это обусловлено другими требованиями к памяти.

Когда система загружается, в фазе 1 инициализации ядра NT драйвер ReadyBoost (драйвер фильтра томов) перехватывает создание загрузочного тома и решает, включать ли кэш. Кэш включается только в том случае, если целевой том зарегистрирован в параметре реестра HKLM\System\CurrentControlSet\Services\rdyboost\Parameters\ReadyBootVolumeUniqueId. Этот параметр содержит идентификатор загрузочного тома. Если функция `ReadyBoot` включена, то драйвер ReadyBoost начинает регистрировать все загрузочные входы и выходы тома через ETW, а если предыдущий план загрузки зарегистрирован в двоичном параметре реестра `BootPlan`, то он порождает системный поток, который заполняет весь кэш, используя асинхронное чтение тома. Когда устанавливается новая ОС Windows, при первой загрузке системы этих двух параметров реестра не существует, поэтому ни кэш, ни трассировка журнала не включаются.

В этой ситуации служба Sysmain, запускаемая SCM позже в ходе загрузки, определяет, нужно ли включать кэш, проверяя конфигурацию системы и запущенную

Windows SKU. Бывают ситуации, когда ReadyBoot полностью отключен — например, когда загрузочным диском является твердотельный накопитель. Если проверка дает положительный результат, то Sysmain активирует ReadyBoot, записывая идентификатор загрузочного тома в относительный параметр реестра `ReadyBootVolumeUniqueId` и включая WMI ReadyBoot Autologger в разделе реестра HKLM\SYSTEM\CurrentControlSet\Control\WMI\AutoLogger\Readyboot. При следующей загрузке системы драйвер ReadyBoost регистрирует все операции ввода-вывода тома, но без заполнения кэша. План загрузки все еще отсутствует.

После каждой последующей загрузки служба Sysmain использует незанятое процессорное время для расчета плана кэширования при следующей загрузке. Она анализирует записанные события ввода-вывода ETW и определяет, к каким файлам был осуществлен доступ и где они расположены на диске. Затем обработанная трассировка сохраняется в %SystemRoot%\Prefetch\Readyboot как файлы `.fx` и рассчитывается новый план кэширующей загрузки с использованием файлов трассировки пяти предыдущих загрузок. Служба Sysmain сохраняет новый сгенерированный план в параметре реестра (рис. 12.12). Драйвер загрузки ReadyBoost считывает план загрузки и заполняет кэш, минимизируя общее время запуска загрузки.

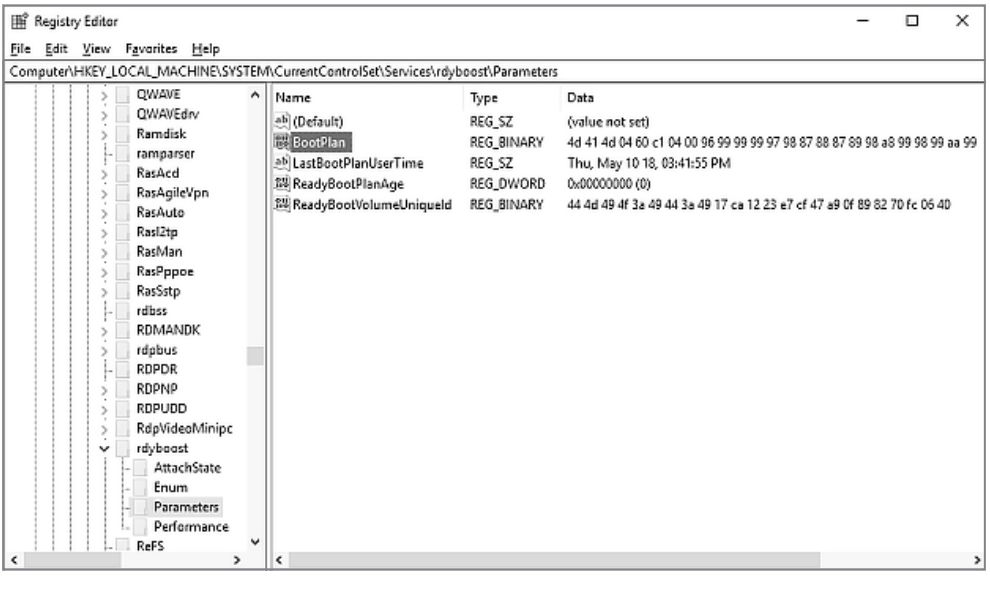


Рис. 12.12. Конфигурация и статистика ReadyBoot

Автоматически запускаемые образы

Помимо параметров реестра Userinit и Shell, в разделе Winlogon существует множество других мест и каталогов реестра, которые компоненты системы по умолчанию проверяют и обрабатывают для автоматического запуска процессов во время загрузки и входа в систему. Утилита Msconfig (%SystemRoot%\System32\Msconfig.exe) показывает образы, настроенные в некоторых таких местах. Утилита Autoruns (рис. 12.13), которую можно загрузить с сайта Sysinternals, проверяет большее количество мест

расположения, чем Msconfig, и показывает больше информации об образах, настроенных на автоматический запуск. По умолчанию Autoruns показывает только те места, которые настроены на автоматическое выполнение хотя бы одного образа, но если выбрать пункт Включить пустые местоположения (Include Empty Locations) в меню Параметры (Options), то Autoruns покажет все места, которые он проверяет.

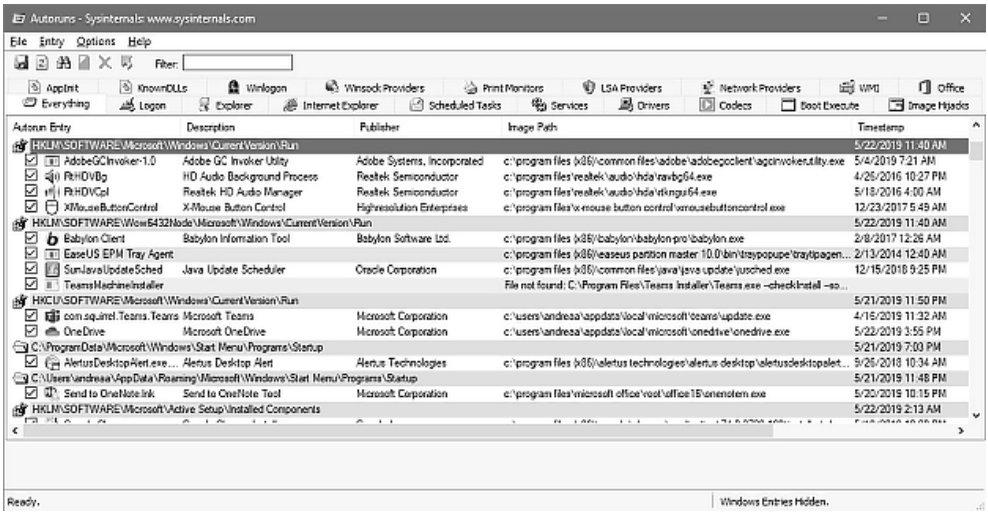


Рис. 12.13. Инструмент Autoruns, доступный в Sysinternals

В меню Параметры (Options) также есть опции, позволяющие направить Autoruns на сокрытие записей Microsoft, но всегда следует сочетать эту опцию с Проверка подписей образов (Verify Image Signatures), в противном случае есть риск скрыть вредоносные программы, содержащие ложную информацию о названии создавших их компаний.

Завершение работы

При завершении работы системы задействуются различные компоненты. Wininit, выполнив все свои инициализации, ожидает выключения системы.

Если кто-то вошел в систему и процесс инициирует завершение работы, вызвав функцию Windows `ExitWindowsEx`, то в Csrss сеанса отправляется сообщение с указанием завершить работу. Csrss, в свою очередь, выдает себя за вызывающего пользователя и посылает RPC-сообщение в Winlogon, указывая ему выполнить завершение работы системы. Winlogon проверяет, находится ли система в процессе перехода к гибридной загрузке (подробнее о гибридной загрузке см. в разделе «Спящий режим и быстрый запуск» далее в этой главе), затем выдает себя за текущего вошедшего в систему пользователя (он может иметь и тот же контекст безопасности, что у пользователя, инициировавшего выключение системы, и иной), просит LogonUI погасить экран (настраивается с помощью параметра реестра HKLM\Software\Microsoft\Windows NCurrentVersion\Winlogon\FadePeriodConfiguration) и вызывает `ExitWindowsEx` со специальными внутренними флагами. Этот вызов еще раз инициирует отправку сообщения процессу Csrss внутри этой сессии с запросом на завершение работы системы.

На этот раз Csrss видит, что запрос поступил от Winlogon, и перебирает все процессы в сеансе входа интерактивного пользователя (опять же не того, кто запросил завершение работы) в обратном порядке по *уровню их завершения*. Процесс может указать уровень завершения, сообщающий системе, когда именно он хочет выйти по отношению к другим процессам, вызвав `SetProcessShutdownParameters`. Допустимые уровни отключения находятся в диапазоне от 0 до 1023, а уровень по умолчанию равен 640. Проводник, например, устанавливает уровень выключения 2, а Диспетчер задач — 1. Для каждого активного процесса, владеющего окном верхнего уровня, Csrss отправляет сообщение `WM_QUERYENDSESSION` каждому потоку в процессе, имеющему цикл сообщений Windows. Если поток возвращает значение `TRUE`, то выключение системы может быть продолжено. Затем Csrss отправляет потоку Windows сообщение `WM_ENDSESSION` с запросом на выход. Csrss ожидает выхода потока в течение нескольких секунд, определенных в параметре HKCU\Control Panel\Desktop\HungAppTimeout. По умолчанию это 5000 мс.

Если поток не завершается до истечения этого срока, Csrss гасит экран и выводит на него окно с зависшей программой (рис. 12.14). Можно отключить этот экран, создав в реестре параметр HKCU\Control Panel\Desktop\AutoEndTasks и установив его равным `1`. Он показывает, какие программы запущены в данный момент и каково их текущее состояние, если оно известно. Windows указывает, какая программа не закрыта своевременно, и предоставляет пользователю выбор: либо «убить» процесс, либо прервать процесс выключения. На экране нет тайм-аута, поэтому запрос на завершение работы может застрять тут навсегда. Кроме того, сторонние приложения могут добавлять собственную информацию о состоянии — например, программа виртуализации может показывать количество активно работающих виртуальных машин с помощью API `ShutdownBlockReasonCreate`.

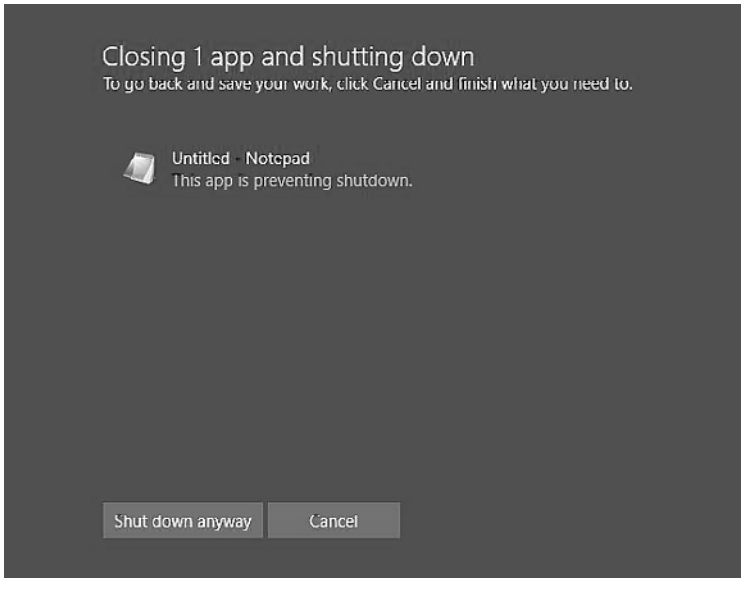
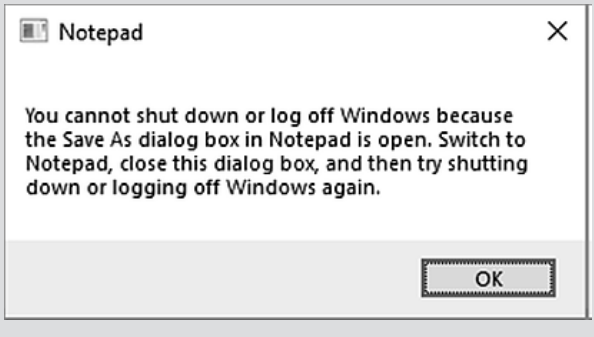


Рис. 12.14. Экран с сообщением о зависшем приложении

ЭКСПЕРИМЕНТ. Наблюдение за HungAppTimeout

Использование параметра реестра `HungAppTimeout` можно проследить, запустив Блокнот, введя в нем текст и выйдя из системы. По истечении времени, указанного параметром реестра HungAppTimeout, Csrss.exe выводит приглашение, в котором спрашивает, нужно ли завершить процесс Блокнота, который еще не завершился, потому что ждал, когда ему прикажут, сохранять ли введенный текст в файл. Если выбрать Отмена (Cancel), Csrss.exe прервет завершение работы.

В качестве второго эксперимента можно попробовать выключить программу снова при все еще открытом диалоговом окне запроса Блокнота. После этого Блокнот выведет собственное окно с сообщением о том, что завершение работы не может быть выполнено без ошибок. Однако это диалоговое окно — всего лишь информационное сообщение, призванное помочь пользователям: Csrss.exe все равно будет считать, что Блокнот завис, и отобразит пользовательский интерфейс для завершения не реагирующих на запросы процессов.



Если поток завершается до истечения тайм-аута, то Csrss продолжает отправлять пары сообщений `WM_QUERYENDSESSION` и `WM_ENDSESSION` другим потокам процесса, владеющим окнами. Как только все потоки, владеющие окнами в процессе, закрылись, Csrss завершает процесс и переходит к следующему процессу в интерактивной сессии.

Если Csrss находит консольное приложение, то он вызывает обработчик управления консолью, посылая событие `CTRL_LOGOFF_EVENT`. Только служебные процессы получают событие `CTRL_SHUTDOWN_EVENT` при завершении работы. Если обработчик возвращает `FALSE`, то Csrss завершает процесс. Если обработчик возвращает `TRUE` или не отвечает в течение нескольких секунд, определенных HKCU\Control Panel\Desktop\WaitToKillTimeout (по умолчанию 5000 мс), то Csrss показывает экран зависшей программы, как на рис. 12.14.

Далее автомат Winlogon вызывает `ExitWindowsEx`, чтобы Csrss завершил все COM-процессы, которые являются частью интерактивной сессии пользователя. К этому моменту все процессы в интерактивной сессии пользователя уже завершены. Далее Wininit вызывает `ExitWindowsEx`, который теперь выполняется

в контексте системного процесса. Это заставляет Wininit отправить сообщение в часть Csrss сеанса 0, где находятся службы. Затем Csrss просматривает все процессы, принадлежащие системному контексту, выполняет их и отправляет сообщения `WM_QUERYENDSESSION` и `WM_ENDSESSION` потокам GUI, как и раньше. Однако вместо отправки `CTRL_LOGOFF_EVENT` он отправляет `CTRL_SHUTDOWN_EVENT` консольным приложениям, которые зарегистрировали обработчики управления. Обратите внимание на то, что SCM — это консольная программа, которая регистрирует обработчик управления. Когда она получает запрос на выключение, то, в свою очередь, посылает сообщение управления выключением службы всем службам, которые зарегистрировались, чтобы получать уведомления об отключении. (Подробнее об отключении служб, например о тайм-ауте отключения, который Csrss использует для SCM, читайте в разделе «Службы Windows» в главе 10.)

Хотя Csrss выполняет те же тайм-ауты, что и при завершении пользовательских процессов, он не выводит никаких диалоговых окон и не останавливает процессы. Параметры реестра для тайм-аутов системных процессов взяты из пользовательского профиля по умолчанию. Эти тайм-ауты просто дают системным процессам шанс очистить и завершить работу до того, как система выключится. Поэтому в действительности при выключении системы многие системные процессы все еще работают — например, Smss, Wininit, Services и LSASS.

После того как Csrss завершает свою работу, уведомляя системные процессы о завершении работы системы, Wininit просыпается, ждет 60 с, пока все сессии не будут закрыты, а затем при необходимости вызывает восстановление системы. На этом этапе ни один пользовательский процесс в системе неактивен, поэтому приложение восстановления может обработать все необходимые файлы, которые могли применяться ранее. Wininit завершает процесс выключения, закрывая LogonUi и вызывая функцию исполнительной подсистемы `NtShutdownSystem`. Эта функция вызывает функцию `PoSetSystemPowerState`, чтобы организовать выключение драйверов и остальных исполнительных подсистем — диспетчера Plug and Play, диспетчера питания, исполнительной системы, диспетчера ввода-вывода, диспетчера конфигурации и диспетчера памяти.

Например, `PoSetSystemPowerState` вызывает диспетчер ввода-вывода для отправки пакетов ввода-вывода о выключении всем драйверам устройств, которые запросили уведомление о выключении. Это действие дает драйверам устройств возможность выполнить любую специальную обработку, которая может потребоваться их устройству перед выходом из Windows. Стеки рабочих потоков подкачиваются в память, диспетчер конфигурации сбрасывает все измененные данные реестра на диск, а диспетчер памяти записывает все измененные страницы, содержащие данные файлов, обратно в соответствующие файлы. Если включена опция очистки файла подкачки при завершении работы, диспетчер памяти очищает его в этот момент. Диспетчер ввода-вывода вызывается во второй раз, чтобы сообщить драйверам файловой системы о завершении работы системы. Завершение работы системы происходит в диспетчере питания. Действия диспетчера питания зависят от того, затребовал ли пользователь выключение, перезагрузку или отключение питания.

Все современные приложения полагаются на интерфейс выключения Windows (Windows Shutdown Interface, WSI), чтобы правильно завершить работу системы. API WSI по-прежнему задействует RPC для взаимодействия между процессами и поддерживает льготный период. Льготный период — это механизм, с помощью которого пользователя информируют о предстоящем выключении до того, как оно начнется. Этот механизм применятся даже в том случае, если системе необходимо установить обновления. Advapi32 использует WSI для связи с Wininit. Wininit ставит в очередь таймер, который срабатывает по окончании льготного периода и вызывает Winlogon для инициализации запроса на выключение. Winlogon вызывает `ExitWindowsEx`, а остальная часть процедуры идентична предыдущей. Все UWP-приложения и даже новое меню Пуск задействуют модуль ShutdownUX для выключения системы. ShutdownUX управляет переходом в режим питания для UWP-приложений и связан с Advapi32.dll.

Спящий режим и быстрый запуск

Чтобы улучшить время запуска системы, в Windows 8 ввели новую функцию под названием «быстрый запуск» (Fast Startup), известную также как гибридная загрузка. В предыдущих редакциях Windows, если оборудование поддерживало состояние питания системы S4 (подробнее о диспетчере питания читайте в главе 6), то Windows позволяла пользователю переводить систему в спящий режим. Чтобы правильно понять, что такое быстрый запуск, необходимо полно описать процесс перехода в спящий режим.

Когда пользователь или приложение вызывает API `SetSuspendState`, диспетчеру питания отправляется рабочий элемент. Он содержит всю информацию, необходимую ядру для инициализации перехода в состояние питания. Диспетчер питания сообщает префетчеру о невыполненном запросе на переход в спящий режим и ждет завершения всех ожидающих операций ввода-вывода. Затем он вызывает API ядра `NtSetSystemPowerState`.

`NtSetSystemPowerState` — это ключевая функция, которая организует весь процесс перехода в спящий режим. Эта процедура проверяет, что вызывающий токен содержит привилегию выключения, синхронизируется с диспетчером Plug and Play, реестром и диспетчером питания — таким образом исключается риск вмешательства других транзакций — и выполняет циклический обход всех загруженных драйверов, отправляя каждому из них IRP-запрос `IRP_MN_QUERY_POWER`. Таким способом диспетчер питания сообщает каждому драйверу, что операция питания запущена, поэтому устройства драйвера не должны больше запускать никаких операций ввода-вывода или предпринимать какие-либо другие действия, которые могли бы помешать успешному завершению процесса перехода в спящий режим. Если один из запросов не выполняется — возможно, некий драйвер находится в процессе важного ввода-вывода, то процедура прерывается.

Диспетчер питания использует внутреннюю процедуру, которая изменяет данные конфигурации загрузки системы (boot configuration data, BCD), чтобы включить приложение загрузки Windows Resume, которое, как следует из названия, пытается

возобновить работу системы после спящего режима. (Более подробную информацию см. в разделе «Диспетчер загрузки Windows» в начале этой главы.) Диспетчер питания выполняет следующие действия.

- Открывает объект BCD, используемый для загрузки системы, и считывает связанный с ним GUID приложения Windows Resume, хранящийся в специальном безымянном элементе BCD, имеющем значение 0x23000003.
- Ищет объект Resume в хранилище BCD, открывает его и проверяет описание. Записывает BCD-элементы `device` и `path`, связывая их с файлом `\Windows\System32\winresume.efi`, расположенным на загрузочном диске, и распространяет параметры загрузки из основного системного BCD-объекта, например параметры отладчика загрузки. Наконец, добавляет путь к файлу спящего режима и дескриптор устройства в элементы BCD `filepath` и `filedevice`.
- Обновляет корневой BCD-объект Boot Manager: записывает BCD-элемент `resumeobject` с GUID обнаруженного приложения загрузки Windows Resume, устанавливает элемент `resume` в `1`, а в случае использования спящего режима для быстрого запуска устанавливает элемент `hiberboot` в `1`.

Далее диспетчер питания сбрасывает данные BCD на диск, вычисляет все диапазоны физической памяти, которые необходимо записать в файл спящего режима (сложная операция, которая здесь не описывается), и посылает новый IRP на питание каждому драйверу с помощью функции `IRP_MN_SET_POWER`. На этот раз драйверы должны перевести свое устройство в спящий режим и не имеют возможности отменить запрос и остановить процесс перехода в спящий режим. Теперь система готова к спящему режиму, поэтому диспетчер питания запускает спящий поток, единственной целью которого является выключение питания машины. Затем он ожидает событие, сигнал о котором будет подан только после завершения возобновления работы и перезапуска системы пользователем.

Спящий поток с помощью процедур DPC останавливает все процессоры, кроме собственного, фиксирует системное время, отключает прерывания и сохраняет состояние процессора. Наконец, он вызывает реализованную в HAL процедуру обработчика состояния питания, выполняющую машинный код ACPI, необходимый для перевода всей системы в спящий режим, и вызывает процедуру, которая записывает все страницы физической памяти на диск. Спящий поток использует драйвер хранения аварийного дампа для выполнения низкоуровневых дисковых операций ввода-вывода, необходимых для записи данных в файл гибернации.

Диспетчер загрузки Windows на ранних этапах загрузки распознает BCD-элемент возобновления, хранящийся в BCD-дескрипторе диспетчера загрузки, открывает BCD-объект приложения возобновления загрузки Windows и считывает сохраненные данные спящего режима. Наконец, он передает выполнение загрузочному приложению Windows Resume, Winresume.efi. `HbMain` (процедура точки входа Winresume) повторно инициализирует загрузочную библиотеку и выполняет следующие проверки файла спящего режима.

- Проверяет, что файл был записан той же архитектурой исполняющего процессора.

- Проверяет, существует ли действительный файл подкачки и правильного ли он размера.
- Проверяет, сообщила ли прошивка о каких-либо изменениях конфигурации оборудования с помощью таблиц FADT и FACS ACPI.
- Проверяет целостность файла спящего режима.

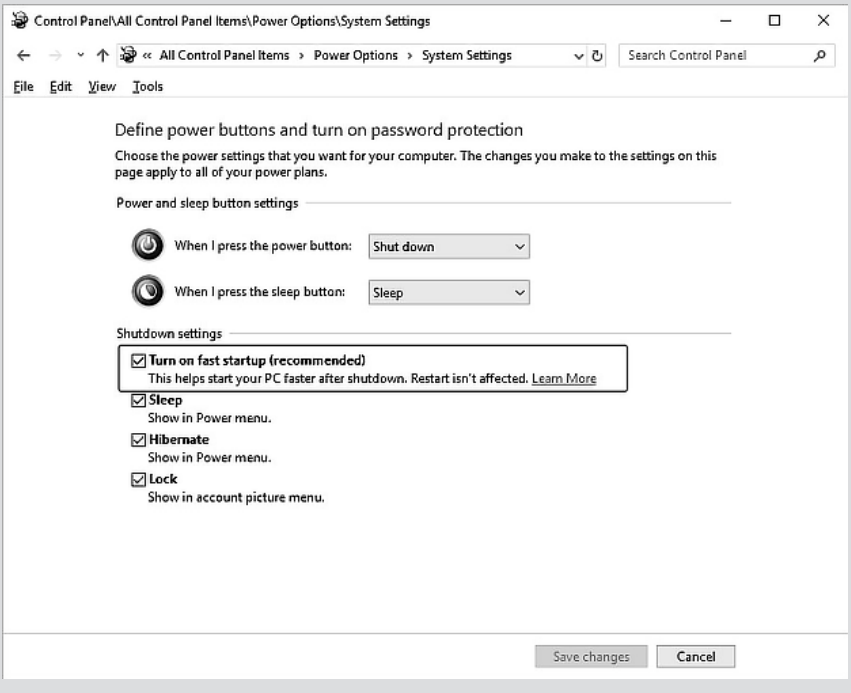
Если одна из этих проверок не проходит, Winresume завершает выполнение и возвращает управление диспетчеру загрузки, который выбрасывает файл спящего режима и перезапускает стандартную холодную загрузку. Если же все предыдущие проверки пройдены, Winresume считывает файл спящего режима, используя загрузочную библиотеку UEFI, и восстанавливает все сохраненное содержимое физических страниц. Далее он перестраивает необходимые таблицы страниц и структуры данных памяти, копирует нужную информацию в контекст операционной системы и, наконец, передает выполнение ядру Windows, восстанавливая исходный контекст процессора. Код ядра Windows перезапускается из того же спящего потока диспетчера питания, который первоначально ввел систему в спящий режим. Диспетчер питания снова включает прерывания и размораживает все остальные процессоры системы. Затем обновляет системное время, считывая его из CMOS, переставляет все системные таймеры и контрольные процессы и посылает еще один `IRP_MN_SET_POWER` каждому системному драйверу, прося их перезапустить свои устройства. Наконец, он перезапускает префетчер и отправляет ему журнал загрузчика для дальнейшей обработки. Теперь система полностью функциональна, а состояние питания системы — S0, она включена.

Быстрый запуск — это технология, реализованная с использованием спящего режима. Когда приложение передает флаг `EWX_HYBRID_SHUTDOWN` в API `ExitWindowsEx` или когда пользователь нажимает кнопку стартового меню Shutdown, если система поддерживает состояние питания S4 (спящий режим) и в ней включен файл спящего режима, то запускается гибридное завершение работы. После того как Csrss выключит все интерактивные сеансовые процессы, службы сеанса 0 и COM-серверы (подробности о самом процессе выключения см. в разделе «Выключение»), Winlogon обнаруживает, что в запросе на выключение установлен флаг `Hybrid`, и вместо того, чтобы пробудить код выключения Winint, идет другим путем. Новое состояние Winlogon использует системный API `NtPowerInformation` для отключения монитора, затем оно информирует LogonUI о незавершенном гибридном выключении и, наконец, вызывает API `NtInitializePowerAction`, запрашивая уход системы в спящий режим. В дальнейшем процедура не отличается от вызова спящего режима системы.

ЭКСПЕРИМЕНТ. Изучение гибридного отключения

Можно наблюдать эффект гибридного выключения, вручную смонтировав хранилище BCD после выключения системы с помощью внешней операционной системы. Сначала убедитесь, что в системе включена функция быстрого запуска. Для этого в строке поиска меню Пуск введите `Панель управления`, выберите Система и безопасность (System and Security), а затем Электропитание (Power

Options). После нажатия кнопки Действия кнопок питания (Choose What The Power Button does), расположенной в верхней левой части окна Электропитание (Power Options), должно появиться следующее окно.



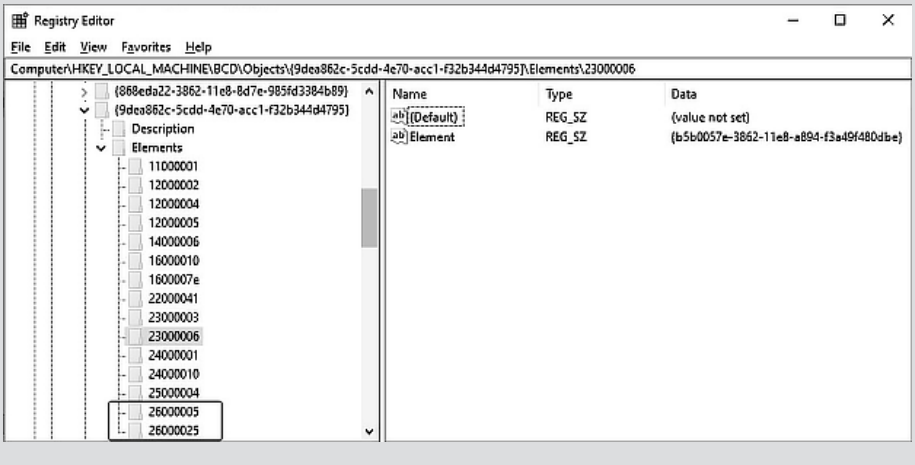
Как показано на рисунке, убедитесь, что опция Включить быстрый запуск (Turn On Fast Startup) выбрана. В противном случае система выполнит стандартное выключение. Выключить рабочую станцию можно с помощью кнопки питания, расположенной в левой части меню Пуск. Перед тем как компьютер выключится, необходимо вставить DVD или USB-накопитель с внешней операционной системой — хорошо подойдет копия Live Linux. Для этого эксперимента нельзя использовать программу установки Windows или любые среды на базе WinRE, поскольку процедура установки очищает все данные спящего режима перед монтированием системного тома.

При включении рабочей станции выполните загрузку с внешнего DVD или USB-накопителя. Эта процедура различается у разных производителей оборудования и обычно требует доступа к интерфейсу BIOS. (Инструкции по доступу к BIOS и выполнению загрузки с внешнего накопителя см. в руководстве пользователя рабочей станции.) Например, в ноутбуках Surface Pro и Surface Book для входа в конфигурацию BIOS обычно достаточно нажать и удерживать кнопку увеличения громкости, а затем нажать и отпустить кнопку питания. Когда новая операционная

система будет готова, установите основной системный раздел UEFI с помощью инструмента разметки в зависимости от типа операционной системы. Эта процедура здесь не описывается. После того как системный раздел будет правильно смонтирован, скопируйте системный файл данных конфигурации загрузки, находящийся в папке \EFI\Microsoft\Boot\BCD, на внешний накопитель или на тот же USB-накопитель, который использовался для загрузки. После этого можно перезагрузить компьютер и дождаться выхода Windows из спящего режима.

После перезагрузки компьютера запустите редактор реестра и откройте корневой раздел реестра HKEY_LOCAL_MACHINE. Затем в меню Файл (File) выберите пункт Загрузить куст (Load Hive). Найдите сохраненный файл BCD, выберите Открыть (Open) и присвойте имя ключа BCD новому загруженному кусту. Теперь нужно определить основной объект BCD диспетчера загрузки. Во всех системах Windows этот корневой объект BCD имеет GUID {9DEA862C5CDD-4E70-ACC1-F32B344D4795}. Откройте относительный раздел и его подраздел `Elements`. Если система была правильно выключена с помощью гибридного выключения, то должны присутствовать BCD-элементы `resume` и `hiberboot` (имена соответствующих ключей — 26000005 и 26000025, подробнее см. табл. 12.2) с параметром `Element` реестра, равным 1.

Чтобы правильно найти элемент BCD, соответствующий конкретной установке Windows, задействуйте элемент `displayorder` (ключ 24000001), в котором перечислены все загрузочные записи установленной операционной системы. В параметре реестра `Element` содержится список всех GUID BCD-объектов, которые описывают установленные загрузчики операционных систем. Проверьте BCD-объект, описывающий приложение Windows Resume, прочитав значение GUID элемента BCD `resumeobject`, который соответствует ключу 23000006. BCD-объект с этим GUID включает путь к файлу спящего режима в элемент `filepath`, соответствующий ключу 22000002.



Среда восстановления Windows (WinRE)

Среда восстановления Windows предоставляет набор инструментов и технологий автоматического восстановления для устранения наиболее распространенных проблем при запуске. Она включает следующие шесть основных инструментов.

- **Восстановление системы.** Позволяет восстановить предыдущую точку восстановления в случаях, когда не получается загрузить установку Windows, чтобы сделать это, даже в безопасном режиме.
- **Восстановление образа системы.** В предыдущих версиях Windows это восстановление называлось полным восстановлением ПК или автоматизированным восстановлением системы (Automated System Recovery, ASR). Оно восстанавливает установку Windows из полной резервной копии, а не только из точки восстановления системы, которая может не содержать всех поврежденных файлов и потерянных данных.
- **Startup Repair.** Автоматический инструмент, который обнаруживает наиболее распространенные проблемы с запуском Windows и пытается их устранить.
- **PC Reset.** Утилита, которая удаляет все приложения и драйверы, не входящие в стандартную установку Windows, восстанавливает все настройки по умолчанию и возвращает Windows в исходное состояние после установки. Пользователь может выбрать, сохранить все файлы личных данных или все удалить. В последнем случае Windows будет автоматически переустановлена с нуля.
- **Командная строка.** В случаях, когда устранение неполадок или ремонт требуют ручного вмешательства, например копирования файлов с другого диска или манипуляций с BCD, можно использовать командную строку, чтобы получить полную оболочку Windows, которая способна запустить практически любую программу Windows при условии соблюдения необходимых зависимостей, в отличие от консоли восстановления из ранних версий Windows, которая поддерживала ограниченный набор специализированных команд.
- **Средство диагностики памяти Windows.** Выполняет диагностику памяти, проверяя признаки неисправности оперативной памяти. Неисправная оперативная память может быть причиной произвольных сбоев ядра и приложений, а также нестабильного поведения системы.

Когда система загружается с DVD или загрузочных дисков, установщик предоставляет выбор: установить Windows или восстановить существующую установку. Если выбрано восстановление установки, то система показывает экран, похожий на современное загрузочное меню (рис. 12.15), в котором представлены различные варианты выбора.

Пользователь может выбрать загрузку с другого устройства, задействовать другую операционную систему, если она правильно зарегистрирована в системном хранилище BCD, или выбрать средство восстановления. Все описанные средства

восстановления, кроме средства диагностики памяти, находятся в разделе устранения проблем.

Приложение установки Windows также устанавливает WinRE на раздел восстановления при чистой установке системы. Доступ к WinRE можно получить, удерживая нажатой клавишу Shift при перезагрузке компьютера с помощью кнопки относительного выключения, расположенной в меню Пуск. Если система использует старое меню загрузки, то WinRE можно запустить с помощью клавиши F8, чтобы получить доступ к расширенным параметрам загрузки во время выполнения Bootmgr. Если показывается опция Repair Your Computer, то на компьютере есть локальная копия жесткого диска. Кроме того, если система не смогла загрузиться в результате повреждения файлов или по любой другой причине, которую Winload может понять, она даст указание Bootmgr автоматически запустить WinRE при следующем цикле перезагрузки. Вместо диалогового окна, показанного на рис. 12.15, среда восстановления автоматически запускает инструмент ремонта запуска (рис. 12.16).

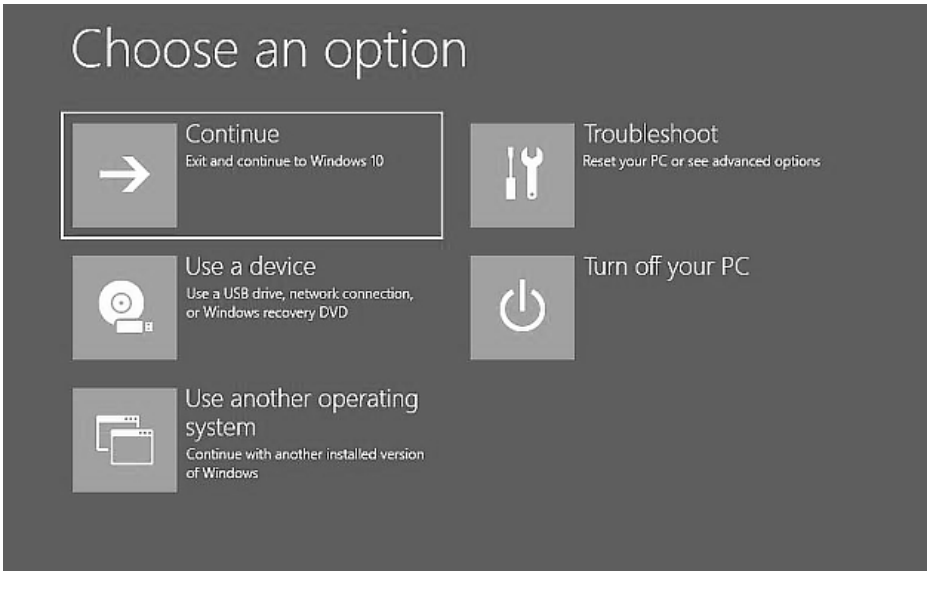


Рис. 12.15. Начальный экран среды восстановления Windows

По окончании цикла сканирования и восстановления инструмент автоматически пытается устранить все обнаруженные повреждения, в том числе заменить системные файлы с установочного носителя. Если инструмент восстановления запуска не может автоматически устранить повреждения, есть возможность попробовать другие методы, после чего снова показывается диалоговое окно с параметрами восстановления системы.

Инструмент диагностики памяти можно запустить из рабочей системы или из командной строки, открытой в WinRE, с помощью исполняемого файла

mdsched.exe. Инструмент спрашивает пользователя, нужно ли перезагрузить компьютер для выполнения проверки. Если система применяет старое загрузочное меню, то инструмент можно запустить с помощью клавиши Tab для перехода к разделу Инструменты (Tools).

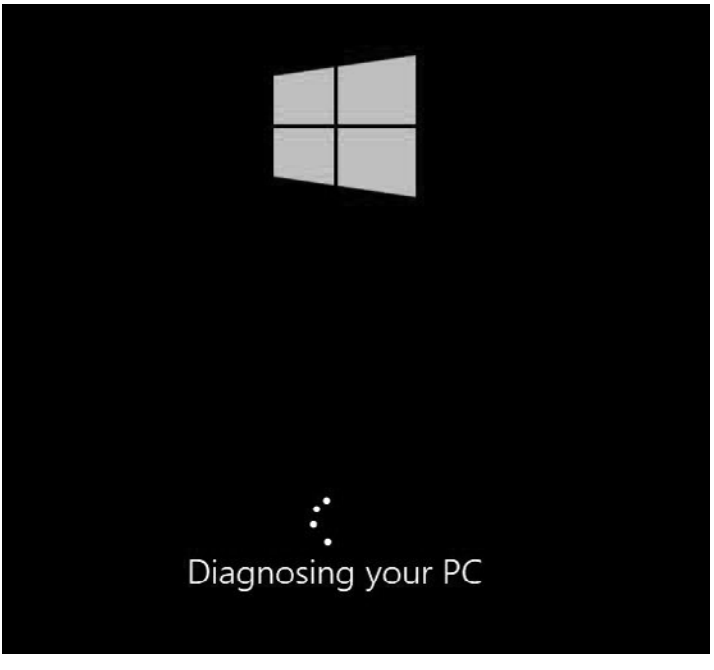


Рис. 12.16. Инструмент восстановления при запуске

Безопасный режим

Пожалуй, самая распространенная причина, по которой системы Windows становятся незагружаемыми, — это сбой драйвера устройства во время загрузки. Поскольку конфигурация программного или аппаратного обеспечения способна меняться со временем, скрытые ошибки могут всплыть в драйверах в любой момент. Windows предлагает администратору способ решения проблемы — загрузку в *безопасном режиме*. Безопасный режим — это конфигурация загрузки, состоящая из минимального набора драйверов и служб устройств. Полагаясь только на драйверы и службы, необходимые для загрузки, Windows избегает загрузки сторонних и других ненужных драйверов, которые могут привести к сбою.

Войти в безопасный режим можно следующими способами.

- Загрузите систему в WinRE и выберите Параметры запуска (Startup Settings) в опциях Advanced (рис. 12.17).
- В мультизагрузочных средах выберите Изменить настройки по умолчанию или выбрать другие параметры (Change Defaults Or Choose Other Options)

в современном меню загрузки и перейдите в раздел Устранение неполадок (Troubleshoot), чтобы выбрать кнопку Параметры запуска (Startup Settings), как в предыдущем случае.

- Если в системе используется старое меню загрузки, нажмите клавишу F8, чтобы войти в меню Расширенные параметры загрузки (Advanced Boot Options).

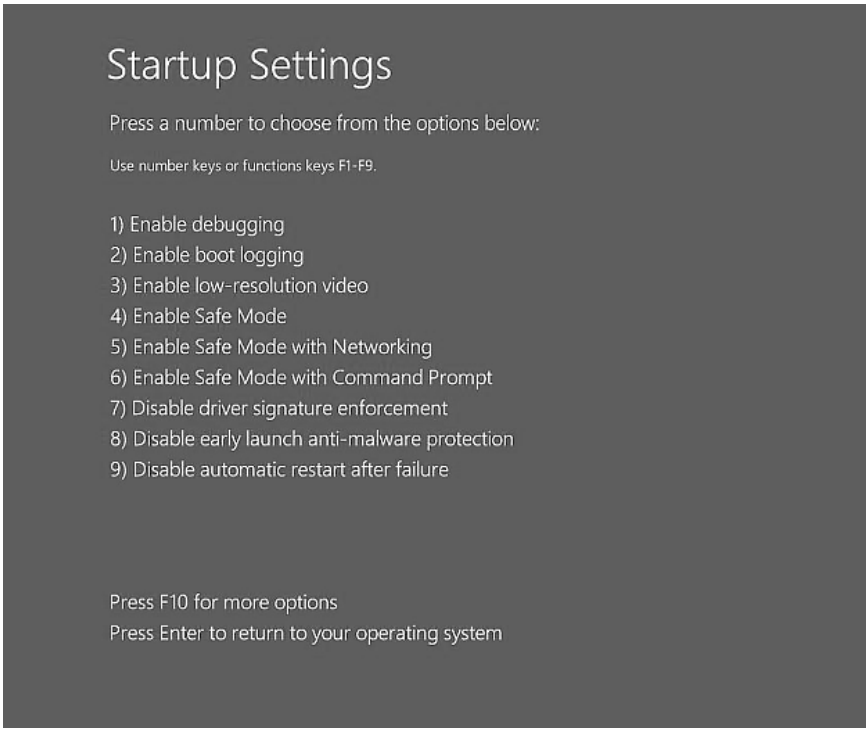


Рис. 12.17. Экран параметров запуска, на котором пользователь может выбрать три вида безопасного режима

Обычно выбирается один из трех вариантов безопасного режима: безопасный режим, безопасный режим с сетью и безопасный режим с командной строкой. Стандартный безопасный режим включает минимальное количество драйверов устройств и служб, необходимых для успешной загрузки. Безопасный режим с поддержкой сети добавляет сетевые драйверы и службы к драйверам и службам, которые входят в стандартный безопасный режим. Наконец, безопасный режим с командной строкой идентичен стандартному безопасному режиму, за исключением того, что Windows запускает командную консоль Cmd.exe вместо Проводника Windows в качестве оболочки, когда система включает режим графического интерфейса.

В Windows предусмотрен и четвертый безопасный режим — режим восстановления служб каталогов (Directory Services Restore), который отличается от стандартного и сетевого безопасных режимов. Он используется для загрузки системы

в режим, в котором служба Active Directory контроллера домена находится в автономном состоянии и не открыта. Это позволяет выполнять операции по ремонту базы данных или восстанавливать ее с резервных носителей. При загрузке в этом режиме загружаются все драйверы и службы, за исключением службы Active Directory. Когда не получается войти в систему из-за повреждения базы данных Active Directory, этот режим позволяет устранить повреждение.

Загрузка драйвера в безопасном режиме

Как Windows узнает, какие драйверы устройств и службы входят в стандартный и сетевой безопасный режимы? Ответ кроется в разделе реестра HKLM\SYSTEM\CurrentControlSet\Control\SafeBoot. Он содержит подразделы Minimal и Network. Каждый из них содержит дополнительные подразделы, в которых указаны имена драйверов устройств, служб или групп драйверов. Например, подразделом BasicDisplay.sys определяется драйвер устройства базового дисплея, который включен в конфигурацию запуска. Драйвер базового дисплея обеспечивает базовые графические службы для любого PC-совместимого адаптера дисплея. Система задействует этот драйвер в качестве драйвера дисплея для безопасного режима вместо того, который может использовать преимущества расширенных аппаратных возможностей адаптера, но при этом препятствовать загрузке системы. Каждый подраздел раздела SafeBoot имеет параметр по умолчанию, который описывает, что определяет этот подраздел. Параметр по умолчанию для подраздела BasicDisplay.sys — `Driver`.

Подраздел Boot файловой системы имеет параметр по умолчанию `Driver Group`. Когда разработчики создают сценарий установки драйвера устройства, файл `.inf`, они могут указать, что драйвер устройства принадлежит к группе драйверов. Группы драйверов, которые определяет система, перечислены в параметре `List` раздела HKLM\SYSTEM\CurrentControlSet\Control\ServiceGroupOrder. Разработчик определяет драйвер как члена группы, чтобы указать Windows, в какой момент процесса загрузки он должен запускаться. Основное назначение раздела ServiceGroupOrder — определить порядок загрузки групп драйверов. Некоторые типы драйверов должны загружаться либо до, либо после других типов драйверов. Параметр `Group` в разделе реестра конфигурации драйвера связывает драйвер с группой.

Ключи конфигурации драйверов и служб находятся в разделе HKLM\SYSTEM\CurrentControlSet\Services. Если заглянуть в этот раздел, то обнаружится подраздел BasicDisplay для драйвера базового дисплея, который, как видно в реестре, входит в группу Video. Любые драйверы файловой системы, которые Windows требует для доступа к системному диску Windows, автоматически загружаются как часть файловой системной группы Boot. Другие драйверы файловой системы входят в группу File System, которая также включена в стандартную конфигурацию и конфигурацию безопасного режима с поддержкой сетевых технологий.

Когда идет загрузка в конфигурации безопасного режима, загрузчик Winload передает ядру Ntoskrnl.exe связанный с ним переключатель в качестве параметра командной строки, а также все переключатели, которые были указаны в файле BCD для загружаемой установки. Если идет загрузка в любом безопасном режиме, то Winload устанавливает для опции `safeboot` BCD значение, описывающее тип

выбранного безопасного режима. Для стандартного безопасного режима Winload устанавливает `minimal`, а для безопасного режима с поддержкой сети добавляет `network`. Winload добавляет `minimal` и устанавливает `alternateshell` для безопасного режима с командной строкой и `dsrepair` для режима восстановления служб каталогов.

ПРИМЕЧАНИЕ Существует исключение, касающееся драйверов, которые безопасный режим исключает из загрузки. Winload, а не ядро загружает любые драйверы с параметром Start, равным 0, в соответствующем разделе реестра, что указывает на загрузку драйверов во время загрузки. Winload не проверяет раздел SafeBoot, поскольку предполагает, что любой драйвер с параметром Start, равным 0, необходим для успешной загрузки системы. Поскольку Winload не проверяет раздел реестра SafeBoot, чтобы определить, какие драйверы загружать, Winload загружает все драйверы, запускаемые при загрузке, а позже Ntoskrnl запускает их.

Ядро Windows сканирует параметры загрузки в поисках переключателей безопасного режима в конце фазы 1 процесса загрузки (`Phase1InitializationDiscard`, см. раздел «Фаза 1 инициализации ядра» ранее в этой главе) и присваивает внутренней переменной `InitSafeBootMode` значение, отражающее найденные переключатели. Во время выполнения функции `InitSafeBoot` ядро записывает значение `InitSafeBootMode` в параметр реестра HKLM\SYSTEM\CurrentControlSet\Control\SafeBoot\Option\OptionValue, чтобы компоненты пользовательского режима, такие как SCM, могли определить, в каком режиме загрузки находится система. Кроме того, если система загружается в безопасном режиме с командной строкой, то ядро устанавливает параметр HKLM\SYSTEM\CurrentControlSet\Control\SafeBoot\Option\UseAlternateShell равным `1`. Ядро записывает параметры, которые Winload передает ему, в параметр HKLM\SYSTEM\CurrentControlSet\Control\SystemStartOptions.

Когда подсистема ядра диспетчера ввода-вывода загружает драйверы устройств, указанные в HKLM\SYSTEM\CurrentControlSet\Services, диспетчер ввода-вывода выполняет функцию `IopLoadDriver`. Когда диспетчер Plug and Play обнаруживает новое устройство и хочет динамически загрузить драйвер устройства для обнаруженного устройства, диспетчер Plug and Play выполняет функцию `PipCallDriverAddDevice`. Обе эти функции вызывают функцию `IopSafebootDriverLoad` перед загрузкой драйвера. Она проверяет значение `InitSafeBootMode` и определяет, нужно ли загружать драйвер. Например, если система загружается в стандартном безопасном режиме, то `IopSafebootDriverLoad` ищет группу драйвера, если она есть, в подразделе Minimal. Если `IopSafebootDriverLoad` находит группу драйвера в списке, то `IopSafebootDriverLoad` указывает вызывающей стороне, что драйвер может быть загружен. В противном случае `IopSafebootDriverLoad` ищет имя драйвера в подразделе Minimal. Если имя драйвера указано в качестве подраздела, то драйвер может загружаться. Если `IopSafebootDriverLoad` не может найти группу драйверов или подраздел с именем драйвера, то драйвер не будет загружен. Если система загружается в безопасном режиме с поддержкой сетевых технологий, то `IopSafebootDriverLoad` выполняет поиск по подразделу Network. Если система не загружается в безопасном режиме, то `IopSafebootDriverLoad` позволяет загрузить все драйверы.

Пользовательские программы с поддержкой безопасного режима

Когда компонент пользовательского режима SCM, который реализует Services.exe, инициализируется во время процесса загрузки, SCM проверяет параметр HKLM\SYSTEM\CurrentControlSet\Control\SafeBoot\Option\OptionValue, чтобы определить, выполняет ли система загрузку в безопасном режиме. Если да, то SCM зеркально повторяет действия `IopSafebootDriverLoad`. Хотя SCM обрабатывает службы, перечисленные в HKLM\SYSTEM\CurrentControlSet\Services, он загружает только те из них, которые соответствующий подраздел безопасного режима указывает по имени. (Более подробную информацию о процессе инициализации SCM можно найти в разделе «Службы» главы 10.)

Userinit — компонент, инициализирующий среду пользователя при входе в систему (%SystemRoot%\System32\Userinit.exe), является еще одним компонентом пользовательского режима, которому необходимо знать, загружается ли система в безопасном режиме. Он проверяет значение параметра HKLM\SYSTEM\CurrentControlSet\Control\SafeBoot\Option\UseAlternateShell. Если оно установлено, то Userinit запускает программу, указанную в качестве оболочки пользователя в параметре HKLM\SYSTEM\CurrentControlSet\Control\SafeBoot\AlternateShell, а не выполняет Explorer.exe. Во время установки Windows записывает имя программы Cmd.exe в параметр `AlternateShell`, делая командную строку Windows оболочкой по умолчанию для безопасного режима с командной строкой. Несмотря на то что командная строка — это оболочка, можно набрать Explorer.exe в командной строке, чтобы открыть Проводник, и запустить любую другую программу с графическим интерфейсом из командной строки.

Как приложение определяет, загружается ли система в безопасном режиме? С помощью вызова функции Windows `GetSystemMetrics(SM_CLEANBOOT)`. Пакетные сценарии, которым необходимо выполнить определенные операции при загрузке системы в безопасном режиме, ищут переменную среды `SAFEBOOT_OPTION`, поскольку система определяет ее только при загрузке в безопасном режиме.

Файл состояния загрузки

Windows использует *файл состояния загрузки* %SystemRoot%\Bootstat.dat для записи факта прохождения различных этапов жизненного цикла системы, включая загрузку и выключение. Это позволяет диспетчеру загрузки, загрузчику Windows и инструменту восстановления запуска обнаружить аномальное завершение работы или невозможность завершить ее без ошибок и предложить пользователю варианты восстановления и диагностической загрузки, как в среде восстановления Windows. Этот двоичный файл содержит информацию, с помощью которой система сообщает об успешном прохождении следующих фаз жизненного цикла системы:

- загрузка;
- выключение и гибридное выключение;
- возобновление работы из спящего или приостановленного режима.

В файле состояния загрузки также указывается, была ли обнаружена проблема при последней попытке загрузки операционной системы и какие варианты восстановления были показаны. Это свидетельствует о том, что пользователь был поставлен в известность о проблеме и принял меры. API-интерфейсы библиотеки среды выполнения Rtl в Ntdll.dll содержат частные интерфейсы, которые Windows применяет для чтения из файла и записи в него. Как и BCD, он не может быть отредактирован пользователем.

ЗАКЛЮЧЕНИЕ

В этой главе подробно рассмотрены шаги, связанные с запуском Windows и завершением ее работы как в обычном режиме, так и в случае ошибок. Было разработано и внедрено множество новых технологий безопасности, цель которых — обеспечить безопасность системы даже на ранних стадиях запуска и сделать ее неуязвимой для различных внешних атак. Рассмотрены общая структура Windows и основные системные механизмы, которые запускают систему, поддерживают ее работу и в конечном счете завершают ее, включая быстрые способы.

Марк Руссинович, Дэвид Соломон, Алекс Ионеску, Андреа Аллиеви

Внутреннее устройство Windows. Ключевые компоненты и возможности

7-е издание

Перевели с английского Д. Лебедев, И. Двойных

Руководитель дивизиона	*Ю. Сергиенко*
Ведущий редактор	*Н. Гринчик*
Литературный редактор	*Н. Рощина*
Художественный редактор	*В. Мостипан*
Корректоры	*Т. Никифорова, Е. Павлович*
Верстка	*Г. Блинов*

Изготовлено в России. Изготовитель: ООО «Прогресс книга».
Место нахождения и фактический адрес: 194044, Россия, г. Санкт-Петербург,
Б. Сампсониевский пр., д. 29А, пом. 52. Тел.: +78127037373.

Дата изготовления: 12.2024. Наименование: книжная продукция. Срок годности: не ограничен.

Налоговая льгота — общероссийский классификатор продукции ОК 034-2014, 58.11.12 — Книги печатные профессиональные, технические и научные.

Импортер в Беларусь: ООО «ПИТЕР М», 220020, РБ, г. Минск, ул. Тимирязева, д. 121/3, к. 214, тел./факс: 208 80 01.

Подписано в печать 11.10.24. Формат 70×100/16. Бумага офсетная. Усл. п. л. 79,980. Тираж 800. Заказ 0000.