Выбираем NVMe over Fabrics

В течение 2017 г. ожидается окончательное утверждение спецификаций стандарта NVMe over Fabrics с использованием транспортов Fibre Channel, InfiniBand, RoCEv2 и iWARP, и, соответственно, конец 2017 г. – начало 2018 г. может стать началом массового выхода на рынок сетевых AFA с использованием NVMe over Fabrics.

Введение

AFA- и гибридные массивы, основанные на SCSI-протоколе, в настоящее время являются одними из основных компонентов мэйнстрима при создании и развитии датацентров. Между тем, в ближайшие несколько лет на смену SCSI-протоколу должен прийти протокол NVMe (Non-Volatile Memory Express, первая спецификация – март 2011 г., первоначально создавался с использованием протокола/шины PCI Express, PCIe), специально разработанный для твердотельных PCIe-модулей, в качестве нового высокопроизводительного интерфейса для серверной флеш-памяти. Это связано с тем, что NVMe поддерживает более низкие задержки и повышенную очередизацию запросов (до тысяч одновременно поддерживаемых операций в/в на тысячах устройств, прим. ред.), что, в свою очередь, обеспечивает гораздо более высокую производительность на случайных операциях в/в, более высокую потоковую производительность, а также более высокий параллелизм приложений, чем при использовании традиционного SCSI-протокола.

В конце 2016 г. прошло утверждение сетевой спецификации NVMe – NVMe over Fabrics, которая дает возможность использования различных транспортов для поддержки NVMe в сетевой архитектуре, включая: Fibre Channel, InfiniBand, RoCEv2 и iWARP. В данной публикации[1] будет проведено сравнение использования транспорта Fibre Channel для NVMe over Fabrics с другими вариантами.

Транспорты для NVMe over Fabrics

Является ли Fibre Channel фабрикой NVMe? – Да.

Fibre Channel является одной из фабрик, поддерживающих NVMe. По адресу *www.nvmexpress.org/wp-content/uploads/NVMe_Over_Fabrics.pdf* размещен технический документ (white paper) об NVMe over Fabrics. В этой статье явно перечислены два типа транспорта для сетевого NVMe – те, которые используют RDMA, и один, использующий Fibre Channel.

Является ли Ethernet столь же хорошо соответствующим для сетевого NVMe, как и FC? – Нет.

В том же самом документе (см. NVMExpress.org) отмечается, что "при использовании Fibre Channel для NVMe over Fabrics, одной из опций этого решения является наличие надежного механизма управления кредитами и надежного механизма доставки." Там же отмечается, что "управление потоками на основе кредитов является "родным" для транспортных сетей Fibre Channel, InfiniBand и PCI Express. Управление потоком на основе кредитов не является частью сетей Ethernet/IP". Это свидетельствует о том, что Fibre Channel является лучшей структурой для NVMe, чем любая из фабрик на основе Ethernet, iWARP[2] или RoCE.

Является ли RDMA ключом к NVMe фабрике? – Нет.

Последователи RDMA утверждают, что наличие технологии RDMA (Remote Direct Memory Access) – необходимое условие для построения рабочей фабрики NVMe. Но необходимо отметить, что указанный выше технический документ (white paper) не указывает на RDMA как на основной атрибут «идеального» транспорта для NVMe. В самой технологии RDMA нет ничего особенного, это просто один из способов обмена данными. Сообщество Infiniband приложило большие усилия к разработке технологий RDMA. Также это сообщество поддерживает тесные связи с техническими группами PCIe, где и появился протокол NVMe. Но ни протокол NVMe, ни протоколы NVMe over Fabrics не имеют зависимости от технологий RDMA.

Является ли SCSI единственным собственным FC-протоколом? – Нет.

Один из подходов сторонников RDMA – сравнить латентность NVMe over Ethernet/IP с латентностью «Fibre Channel». Это похоже на сравнение IP с Ethernet, потому что NVMe – это протокол верхнего уровня, а Fibre Channel – протокол канального уровня. Полное сравнение – это NVMe over Ethernet по сравнению с SCSI over Fibre Channel, что является правильным сравнением, если оно описано правильно. Теперь термин SCSI over Fibre Channel получил (несколько запутанное) имя «Fibre Channel Protocol» (FCP), после чего некоторые стали предлагать весь трафик Fibre Channel считать FCP. Но FCP – это не то же самое, что FC. Это всего лишь один протокол FC-4 (верхний уровень), аналогичный протоколу FICON (протокол, используемый в мейнфреймах для хранения данных), который может переноситься транспортом Fibre Channel. Сторонники RDMA продвигают идею, что FC – это только SCSI-транспорт с использованием "багажа" задержек на основе SCSI.

В результате возникло неправильное понимание: NVMe работает на Fibre Channel только в режиме эмуляции, после трансляции в команды SCSI (FCP). Скорее всего, это неправильное понимание было спровоцировано тем же техническим документом, который говорит нам, что идеальный транспорт NVMe должен позволять клиентам «отправлять и получать "родные" команды NVMe напрямую к/из фабрики, не используя уровень трансляции в SCSI». Это имеет смысл, поскольку NVMe оптимизируется с учетом латентности, а уровень трансляции (translation layer) увеличивает задержку.

На самом деле, Fibre Channel позволяет передавать команды NVMe напрямую. При этом не требуется какой-либо трансляции для транспортировки NVMe-команд. Реализация NVMe over Fibre Channel определяет новый тип трафика верхнего уровня FC-NVMe, который определяет специфические NVMe-фреймы.

Однако, некоторые разъяснения при этом все же требуются. Разработчики стандарта FC-NVMe признали, что огромное значение имеет одновременная поддержка как NVMe, так и SCSI-трафика в одной инфраструктуре. Они также признали, что это будет сделано наиболее эффективно (и просто) за счет использования существующих типов фреймов/кадров. Так обстоит дело с кадрами ввода/вывода. В стандарте FC-NVMe указано, что реализация NVMe поверх Fibre Channel будет использовать тот же тип фрейма ввода-вывода, который использует FCP. Поэтому, если вы захватываете и анализируйте соединение, работающее на NVMe over Fibre Channel, вы увидите смесь типов кадров FC-NVMe и FCP.

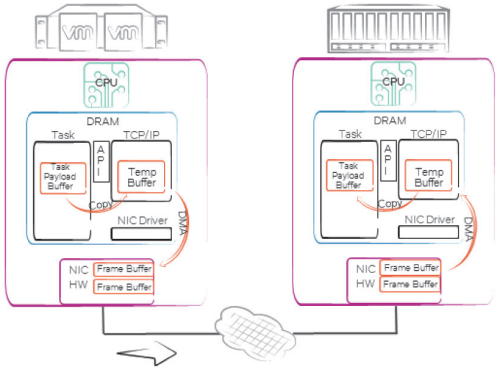
Длительное использование Fibre Channel в качестве многопротокольной структуры является хорошим показателем того, что Fibre Channel SAN будет одновременно поддерживать SCSI и NVMe очень надежно.

Является ли уровень трансляции плохим для NVMe? – Не всегда, это зависит от потребностей.

В техническом документе NVMe over Fabrics говорится, что идеальным атрибутом транспорта фабрики NVMe является то, что для него не требуется уровень трансляции. С точки зрения реализаций с нуля, ориентированных на наименьшую латентность, преобразование, скажем, от SCSI к NVMe с использованием слоя перевода, было бы неоптимальным. Лучше писать приложения, чтобы непосредственно использовать NVMe, и избегать шага перевода, что добавит нежелательные тактовые циклы задержки к каждому вводу-выводу. И, следовательно, идеальная структура не должна требовать перевода, и, как уже упоминалось, Fibre Channel поддерживает NVMe изначально, то есть без перевода. В то же время сообщество NVM Express осознало важность всех развернутых SCSI-приложений, когда они разработали документ SCSI Translation Reference. Этот документ предназначался для разработчиков приложений, чтобы они могли адаптировать свои про-

1) Публикация подготовлена на основе документа Why Fibre Channel Is the NVMe Fabric of Choice, Brocade, 53-1004983-01, 10 March 2017, *Brocade.com/nvme.*

2) iWARP (internet Wide Area RDMA Protocol) – сетевой протокол, который имплементирует remote direct memory access (RDMA) для повышения эффективности передачи данных по сетям на базе Internet протокола. Поскольку iWARP разбивается на IETF-стандартные протоколы с поддержкой перегрузки, такие как TCP и SCTP, он предъявляет мало требований к сети и может быть успешно развернут в широком диапазоне сред. *https://en.wikipedia.org/wiki/IWARP.*



When a task sends using traditional TCP/IP stack:

- API not given ownership of payload buffer, so copies payload to TCP/IP temp buffer.
- TCP/IP stack processing passes temp buffer address to NIC driver, which passes to NIC hardware.
- NIC hardware DMAs from temp buffer to NIC frame buffer for transmission.
- Frame transmitted over network.
- Receiving NIC captures in frame buffer, identifies as TCP/IP.
- NIC DMAs the frame buffer into TCP temp buffer.
- After processing, TCP/IP stack copies temp buffer into task payload buffer.
- Result: DRAM-to-DRAM copies required at each end.

Рис. 1. Традиционный TCP/IP делает одну промежуточную копию на каждой стороне.

дукты под использование NVMe. Но многие потенциальные пользователи NVMe не в состоянии перепроектировать приложения, которые они запускают. Они хотят вариант перехода на инфраструктуру NVMe на своих условиях, не зависящих от поставщиков приложений. Если взглянуть на это с такой точки зрения, то доступный уровень перевода – как опция – действительно хорош для внедрения NVMe. NVMe over Fibre Channel предлагает лучшее из двух "миров" – поставщики HBA предоставляют драйверы, которые предлагают перевод SCSI-to-NVMe, когда это необходимо, а также обеспечивается прямая поддержка NVMe для приложений, которые предназначены для такого использования.

Может ли Fibre Channel делать «Zero Copy»[3]? – Да.

Когда IP-стек разрабатывался в 1980-х годах, он был предназначен для работы с большим числом протоколов верхнего уровня и множеством сетей уровня 2 – от Token Ring до телефонных линий. Чистое разделение сетевых уровней имело идеальный смысл для взаимодействия, и одним из способов достижения этой цели было использование промежуточной буферизации, что делало буфер для копий общим местом. Однако по мере увеличения скорости передачи большинство буферных копирований было удалено в целях оптимизации, за исключением случаев, когда это нарушало бы обратную совместимость.

В начале 1990-х годов хороший сетевой стек мог обеспечить эффективность при однократном промежуточном копировании данных. NIC (network interface card – сетевая интерфейсная карта) получал кадры и записывал их (используя DMA – direct memory access[4]), прямой доступ к памяти) в буферы DRAM, связанные с сетевым стеком. Затем стек обрабатывает кадр, определяет, какое приложение должно получить полезную нагрузку (payload), и копирует его в буфер DRAM приложения (поскольку этап NIC DMA не является копией DRAM-to-DRAM, он не учитывается.) В то время такая архитектура с одной промежуточной копией была оптимальной (рис. 1).

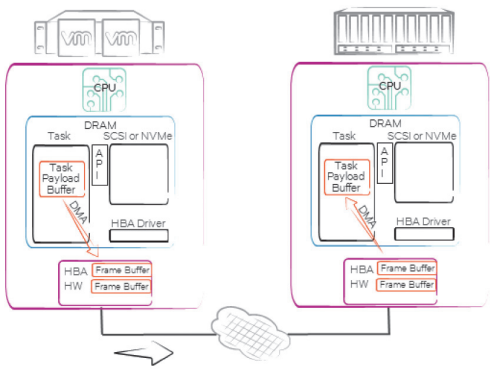
Но в середине 1990-х годов, когда Fibre Channel стал воплощаться в продуктах, ситуация изменилась. Основной претензией FC была скорость, поэтому давление для оптимизации было высоким. Технология на основе чипа поддерживала бОльшую сложность, а стек FC/SCSI имел меньшее количество уровней и не был ограничен теми же проблемами обратной совместимости, с которыми сталкивались IP-стеки. Поэтому Fibre Channel хорошо позиционировался для реализации архитектуры адаптера / драйвера / стека, которая устраняла промежуточную копию. Интерфейс прикладного программирования (API) требует, чтобы приложение определяло желаемые ячейки памяти в терминах «Scatter Gather List» (SGL), которые адаптер FC (FC HBA) использует для записи полезных нагрузок. Fibre Channel в течение последних двух десятилетий "спокойно" поддерживал функцию «zero copy» (рис. 2).

Некоторые сведения о Remote DMA (удаленном DMA)

Спустя несколько лет после того, как Fibre Channel стал мэйнстримом, два направления развития в отрасли "Future I/O" и "Next Generation I/O" объединились в виде стандарта InfiniBand, ориентированного на объединении серверов в кластер. По мере "созревания" InfiniBand, основное его внимание фокусировалось на Remote DMA (RDMA) в качестве протокола для повышения производительности обмена данными с кластерами серверов, и в дальнейшем RDMA получил широкое распространение в средах высокопроизводительных вычислений (HPC). RDMA уменьшает задержку для передачи данных между серверами по сравнению с более ранними протоколами, в частности, когда данные очень динамичны (например, результат вычислений). RDMA работает, передавая «список сборок рассеяния» (SGL) адресов блоков памяти с локального сервера на удаленный сервер, эффективно разделяя права владения локальной памятью с удаленным сервером, позволяя удаленному серверу напрямую читать или записывать из/в память локального сервера. InfiniBand, так же, как и Fibre Channel перед ним, находился в ситуации (доступность чип-технологии, отсутствие обратной совместимости), где эффективность с нулевым копированием (zero-copy) была наибольшей.

Требуется ли RDMA для Zero Copy на IP? – Нет.

После того как RDMA завоевал популярность при построении кластеров серверов, были предприняты усилия по его распространению на сети, использующие стандарт iWARP (Internet Wide Area RDMA Protocol, опубликован в 2007 г., официально описывается пятью стандартами группы IETF – Internet Engineering Task Force: RFCs 5040-5044). iWARP был разработан поверх TCP (transmission control



When a task sends using SCSI or NVMe stack and an FC HBA:

- API gives ownership of payload buffer to storage stack; no temp buffer is needed.
- SCSI or NVMe stack processing passes payload buffer address to HBA driver, which passes to HBA hardware.
- HBA hardware DMAs the payload from task buffer to HBA frame buffer for transmission.
- Frame transmitted over network.
- Receiving HBA captures in frame buffer. HBA is storage-centric, so behavior is not protocol dependent.
- HBA DMAs the payload from the frame buffer into task buffer.
- Result: No copies needed on either end.

Рис. 2. Fiber Channel выполняет нулевое копирование.

3) **Zero-copy** (досл. ноль копирований) – описывает операции, в ходе которых процессор не выполняет задачу копирования данных из одной области памяти в другую. Термин применяется для описания технологий, которые помогли уменьшить количество копирований между промежуточными буферами.
Версии zero copy для элементов операционной системы, таких как драйверы устройств, файловые системы и стеки сетевых протоколов, значительно увеличивают производительность определенных прикладных программ и более эффективно используют системные ресурсы. Производительность улучшается за счет предоставления возможности процессору переходить к другим задачам во время копирования данных, выполняемого параллельно в другой части машины. Кроме того, операции zero copy снижают число затратных по времени переключений между режимами ядра и пользователя. Системные ресурсы используются более эффективно, так как использование такого сложного устройства как процессор для выполнения операций копирования, что само по себе является довольно простой задачей, весьма расточительно, если прочие более простые компоненты системы самостоятельно могут выполнить копирование.
Способы создания ПО с поддержкой zero copy включают в себя использование копирование на основе технологии DMA и отображение в памяти (memory mapping) через блок управления памятью (MMU). Эти особенности требуют специфической аппаратной поддержки и обычно включают в себя определенные требования к выделению памяти.
Протоколы с zero copy очень важны для высокоскоростных сетей, в которых ёмкость сетевого соединения приближается к или превосходит возможности обработки процессором. В этом случае процессор проводит почти все время копируя передаваемые данные, и таким образом становится узким местом («бутылочным горлышком»), устанавливающим ограничение скорости соединения ниже его возможностей. Приближенный подсчет, используемый в индустрии, говорит, что примерно один тактовый цикл процессора требуется для обработки одного бита входящих данных. Например, процессор с тактовой частотой в 1 ГГц может обрабатывать сетевое соединение с пропускной способностью в 1 Гбит/с при обычном копировании данных, но этот же самый процессор «захлебнется», работая с 10-гигабитным соединением. Именно поэтому ПО с поддержкой zero copy становится крайне необходимым. Сетевые соединения свыше 1 Гбит/с и, следовательно, и сетевое ПО с поддержкой zero copy, на данный момент ограничены использованием лишь в суперкомпьютерных кластерах, крупных индустриальных (особенно государственных, научных и коммерческих) центрах данных (ЦОД) и так далее. Однако, по мере развития информационных технологий и по мере того, как сети с пропускной способностью в 1 Гбит/с, 10 Гбит/сек и даже 100 Гбит/сек становятся все более распространенными, решения с zero copy также начинают пользоваться все большим спросом, так как пропускная способность сетей растет быстрее производительности процессоров.
Протоколы zero copy обладают некоторыми первоначальными накладными расходами, связанными с подготовкой регионов памяти для DMA-операций, так что отказ от программного ввода-вывода (PIO) приемлем только для больших пакетов данных, либо для больших потоков и адаптированного ПО.
В основе протоколов RDMA (*Remote Direct Memory Access*) лежат методики zero copy.
Некоторые операционные системы (включая Linux) поддерживают технологии zero copy для передачи файлов в сеть за счет специфических API-функций, как например, sendfile и sendfile64, splice, vmsplice. *https://ru.wikipedia.org/wiki/Zero-copy.*

4) DMA – прямой доступ к памяти (*direct memory access, DMA*) – режим обмена данными между устройствами компьютера или же между устройством и основной памятью, в котором центральный процессор (ЦП) не участвует. Так как данные не пересылаются в ЦП и обратно, скорость передачи увеличивается.

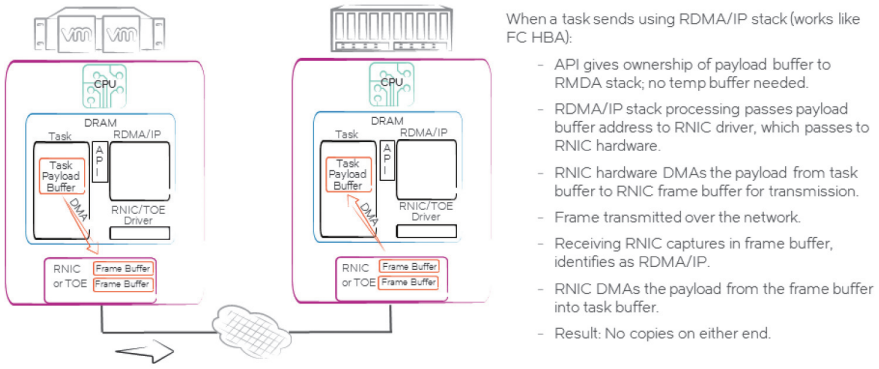


Рис. 3. Достижение Zero Copy при использовании iWARP и RoCE.

protocol – протокол управления передачей) – потокового транспортного протокола, который использует подтверждения и ретрансмиссии, когда это необходимо для обеспечения надежной доставки. TCP также включает «оконный» алгоритм для дозирования передачи, чтобы избежать превышения пропускной способности сети между отправителем и получателем («потоковый» аспект TCP позволяет собирать последовательные порции полезной нагрузки и отправлять их вместе, что означает, что приемники должны обрабатывать все ранние фрагменты, чтобы понять, где начинаются более поздние фрагменты). Первый стандарт – RFC 5040 – описывает, как RDMA использует «прямое размещение данных», (DDP) для достижения эффективности с нулевым копированием Fibre Channel и InfiniBand. В последнем стандарте RFC 5044 описывается «выровненное фреймирование PDU маркера для TCP», которое эффективно отключает «коалесцирующее» (coalescing) поведение TCP, так что принимающая NIC может легче обрабатывать фрагменты, что обеспечивает практическую аппаратную поддержку DDP.

Вышеупомянутые стандарты RFC обеспечили основу для эффективной реализации zero-copy, но традиционные сетевые адаптеры (NIC) не обладали возможностями для обработки TCP. Программная реализация этого функционала обеспечивала взаимодействие, но не давала обещанной производительности RDMA. Для этого требовались новые сетевые адаптеры, называемые TCP Offload Engines (TOEs). Ранние TOE не соответствовали iWARP, поэтому были разработаны TOE с RDMA, которые могли реализовывать DDP в аппаратных средствах. Такие TOE способны уже были обеспечить такую же эффективную реализацию zero-copy, как и Fibre Channel (см. рис. 3).

Скачок в сетевизации RDMA

Примерно в 2009 году, когда NVMe стал "набирать обороты", группы IETF "TRansparent Interconnection of Lots of Links" (TRILL) и IEEE Data Center Bridging (DCB) начали разрабатывать стандарт, чтобы сделать lossless-Ethernet (Ethernet без потерь данных) (TRILL был создан для того, чтобы сделать доступной любую топологию Ethernet, не поддерживаемую протоколом IEEE's Spanning Tree Protocol. DCB строилась на "трех китах": Priority-based Flow Control – управление потоком на основе приоритетов, Enhanced Transmission Selection – расширенный выбор передачи и Data Center Bridging eXchange). Торговая ассоциация InfiniBand (IBTA, InfiniBand Trade Association) увидела в этом новые возможности для развития бизнеса и разработала спецификацию RDMA over Converged Ethernet (RoCE, Converged Ethernet – «конвергентный Ethernet» был более ранним термином для DCB). Так же, как и iWARP, который нуждался в поддержке специализированных TOE для обеспечения эффективности zero copy, RoCE зависит от NIC (RNICS) с поддержкой RDMA для достижения этой производительности. IBTA продвигала протокол RoCE как более высоко производительный по сравнению с iWARP. При этом отмечалось, что протокол TCP, являющийся основой iWARP, не был идеальным протоколом для коммуникаций с малой задержкой, отчасти из-за поведения «медленного старта» TCP, которое соединение инициировано или когда соединение было свободно в течение длительного периода времени. Поскольку Ethernet не обеспечивает надежную транспортную способность на базе TCP, стандарт RoCE реализовал ее на более высоком уровне в стеке. Когда RoCE был объявлен, существовало ограничение IPv4-адресов, и TRILL обещал радикально расширить масштаб Layer 2 Ethernet-сетей и, следовательно, IP-подсетей. По-видимому, IBTA считало, что у RoCE есть все, что необходимо для создания крупномасштабного высокопроизводительного RDMA.

В противовес этому, основные игроки на рынке гиперскалярных сетевых архитектур и программно-определяемых сетей стали продвигать концепцию "Layer 3 to Top of Rack" («уровень 3 на уровне стойки»). В результате, широкое распространение получили подсети IP на уровне стойки, что потребовало от IBTA создать RoCEv2 (иногда называемый «Routable RoCE»). В отличие от iWARP на основе TCP, RoCEv2 работает поверх UDP, который не имеет режима дозирования с "медленным стартом" (slow-start throttling). Переход к UDP означает, что фреймы RoCEv2 несовместимы с фреймами RoCEv1 (хотя сетевые платы с поддержкой RDMA, поддерживающие RoCEv2, обычно могут быть настроены на использование формата RoCEv1). Поскольку в UDP отсутствует поддержка TCP для IETF Explicit Congestion Notification (ECN, RFC 3168, 4301, 6040), IBTA указала, что RoCEv2 также поддерживает IETF ECN, реализуя управление потоком в транспортном слое IB по UDP.

Текущее положение дел заключается в том, что iWARP и RoCEv2 соперничают за владение сетевым рынком NVMe на базе Ethernet, каждый из которых имеет обоснованную критику в отношении другого транспорта.

Табл. 1. Сравнение семантик для памяти, флэш, СХД и NVMe.

Feature	"Ideal Memory"	"Ideal Storage"	Flash is like...	NVMe Semantic
Read Bandwidth	Very high	Medium	Memory	Memory
Write Bandwidth	Very high	Medium	Storage	Memory
Read Latency	Very high	Medium	Memory	50/50
Write Latency	Very high	Medium	Storage	50/50
Read Granularity	High	Low	Memory	Storage
Write Granularity	High	Low	Storage	Storage
Scale	GB to TB	TB to EB	GB to PB	Storage
Random Access	Very High	Medium	Memory	Memory
Persistence	Low	Very high	Storage	Storage
Rewritability	High	Medium	Storage	N/A
Reliability	High	Very high	Memory	N/A
Metadata Linkage	Low	Medium	Memory	Storage

Начало NVMe

Сообщество NVMe (Nonvolatile Memory Express, «энергонезависимая память», подключаемая по шине PCI Express.) возникло после того, как в 2007 году Форум разработчиков Intel предложил стандартизировать интерфейс для флеш-модулей на шине PCI Express. Некоторые аспекты полученной в результате спецификации NVMe, возможно, более ориентированы на семантику хранения, чем на семантику памяти, в основном из-за блочно-ориентированной природы flash-технологий (табл. 1).

Преимущество NVMe над SCSI и ATA

До разработки NVMe напрямую подключенные твердотельные диски (SSD), в том числе флэш-диски, обычно подключались через Serial Attached SCSI (SAS) или Serial ATA (оба этих интерфейса были последовательными версиями параллельных интерфейсов дисков, изначально определенных в 1980-е годы для индустрии ПК в эпоху DOS). По мере "созревания" операционных систем и приложений мало внимания уделялось сложности и задержкам служебных данных в этих унаследованных протоколах, так как время ожидания вращения и время поиска дисковых накопителей преобладали над общей задержкой ввода-вывода любого диска. Зрелость экосистемы флеша изменила эту ситуацию. С самого начала флэш-технология приносила огромные преимущества в задержках при чтении, особенно при чтении с произвольным доступом, где архитектура дисков наиболее уязвима. Кэши с записью позволяли нивелировать медленную скорость записи на флеш, но проблемы выносливости при записи флеша и ограниченная плотность сдерживали его пригодность к специализированным нишам в первых реализациях. По мере того как возрастала плотность флеш-технологий и появления умных алгоритмов для "смягчения" проблем выносливости при записи флеш становилась полностью жизнеспособной альтернативой вращающемуся диску. В результате, зависимость индустрии от интерфейса SCSI, ориентированного на диск, слабла и NVMe был ответом на эту тенденцию.

NVMe расширяется

Параллельно с усилиями по расширению RDMA, сторонники NVMe также стремились расширить и свое присутствие. По-

скольку флэш-накопители вытесняли жесткие диски на сервере, сообщество NVMe пришло к выводу, что флеш-память вскоре вытеснит дисковые накопители и в сетевом хранилище. Предваряя необходимость, сообщество NVMe приступило к разработке спецификации NVMe over Fabrics.

Как упоминалось ранее, спецификация семантики протокола NVMe ориентирована, скорее, на хранение, чем на память. Все три из транспортов, идентифицированные в спецификации NVMe over Fabrics – InfiniBand, iWARP и RoCEv2 – используют уровень RDMA в своих реализациях, который ориентирован, в первую очередь, на работу с памятью. В этой ситуации все эти три RDMA-транспорта "сталкиваются" с Fibre Channel, бесспорным действующим протоколом для высокопроизводительного сетевого хранилища.

RDMA требует исправления iWARP

Хотя RDMA является мощным протоколом для высокопроизводительных кластерных серверных приложений с разделяемой памятью, это не очень эффективный протокол хранения данных, особенно для небольших операций записи. Представим, что серверу необходимо записать 1024 байта в хранилище. В этом случае модель RDMA требует, чтобы сервер отправил сообщение на запоминающее устройство и сообщил: «Эй, я хочу записать 1024 байта на свой том. Байты расположены в моей памяти по адресам в подключенном SGL (scatter-gather list)". Устройство хранения получает сообщение и связанный SGL, затем разворачивается и выдает запрос на считывание RDMA в память сервера, после чего сервер отправляет 1024 байта. Для этой транзакции требуются три сообщения, в отличие от небольшого сообщения на базе SCSI, доступного в Fibre Channel. Сообщество NVMe признало эту слабость и, "закрыв глаза" на удаленные принципы DMA, определило механизм «капсулы» для отправки данных полезной нагрузки в одно сообщение с помощью команд.

Эффективная схема капсулирования NVMe с поддержкой «write-this-data» помогла ей повысить производительность, но не-RDMA-аспект сделал ее несовместимой с основной реализацией RDMA: iWARP. Сообщество iWARP "занялось" этим, и в 2014 году IETF опубликовал RFC 7306, который описывает «RDMA запись с немедленными данными».

Ethernet и IP добавляют риск для корпоративного хранилища

Сложность стека

Одна из причин того, что протокол NVMe более эффективен, чем протокол SCSI, – значительно более простой стек протокола. Поскольку простота стека важна, давайте рассмотрим стеки протоколов для различных NVMe фабрик. Графически стеки для Fibre Channel, RoCEv2 и iWARP показаны на рис. 4.

Сложность IP/Ethernet, относительно Fibre Channel, не является случайной или безвозмездной. В протоколах есть несколько ключевых отличий, которые по ходу развития протокола привели к такой сложности:

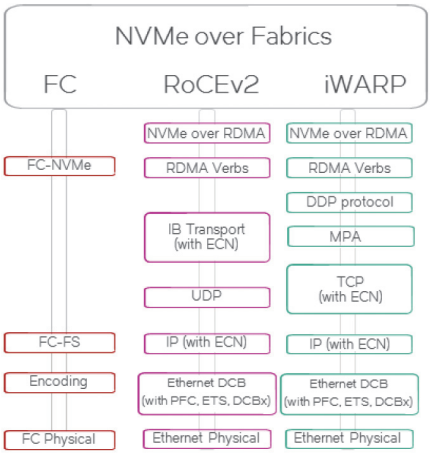


Рис. 4. Сравнение стека для различных NVMe фабрик.

– Ethernet и IP (и TCP / UDP) реализуют транспорт на уровнях, которые были разработаны гораздо более независимо друг от друга по сравнению с Fibre Channel. Задачи назначения адресов и маршрутизации в глобальном масштабе с буквально миллиардами узлов, которые должны поддерживаться сетями Ethernet /IP, требуют много сложных уровней и алгоритмов. Fibre Channel был разработан для масштабирования центров обработки данных, проблема, которая имеет свои сложности, но намного проще, чем глобальное масштабирование сетей Ethernet/IP;
– Ethernet разрабатывался в самом начале появления сетей для работы в средах с совместным использованием и разделением ресурсов. Протокол развивал различные разрозненные механизмы для loop avoidance, flooding, address learning и т.д. Управление потоком интегрировалось (сшивалось) постепенно по годам. Напротив, разработчикам Fibre Channel удалось извлечь пользу из этих ранних уроков и таким образом создать более целостно согласованный протокол;
– Ethernet и IP "выросли" для работы в самых разнообразных средах – от LAN к MAN, к кампусу, к небольшому офису и до дома. Такая функциональность как plug-and-play, обратная совместимость имеют важное значение для внедрения в большинстве этих сред. Эта реальность предъявляет огромные требования к стекам протоколов. Fibre Channel по-прежнему сосредоточился на использовании в центрах обработки данных премиум-класса и, следовательно, не был вынужден резко эволюционировать, что значительно упрощает проблемы совместимости.

Здесь уместно признать, что сложности стеков iWARP и RoCEv2 необязательно добавляют значительные задержки. Большая часть сложности стека обрабатывается «аппаратными средствами» (хотя часто используются процессорные ядра на основе ASIC) – специализированными сетевыми картами с поддержкой RDMA или механизмами TCP снижения нагрузки. Но сложные стеки вызывают проблемы с конфигурацией, управлением, взаимодействием, устранением неполадок и анализом.

Многоуровневое управление потоком является проблематичным

Fibre Channel всегда был сетью без потерь. По каждому линку отправитель и получатель используют буферные кредиты, точно зная, сколько пакетов можно отправить безопасно, гарантируя, что пакеты не будут отбрасываться из-за нехватки места. Фабрики Fibre Channel также реализованы с несколькими каналами передачи, работающими в параллель в одном слое, что позволяет передавать управление потоком через каждое устройство. Оба протокола – iWARP и RoCEv2 – рекомендуют использовать Ethernet без потерь для подключения Layer 2. DCB Ethernet добилась прогресса в уменьшении потерь пакетов традиционных Ethernet, но DCB по-прежнему страдает от проблем совместимости, и механизмы управления потоками не распространяются через маршрутизаторы. IETF определял Explicit Congestion Control (явный контроль перегрузки) для сквозного контроля перегрузками. Тем не менее, ECN создает конфликт интересов в том смысле, что "неправильные" конечные узлы могут несправедливо использовать больше, чем их доля в пропускной способности сети, что подчеркивает преимущества делегирования управления потоком для чистой одноуровневой фабрики, которую обеспечивает Fibre Channel.

Сложность стеков означает сложную конфигурацию

Ethernet и IP получили развитие за счет низкой стоимости и простого одноразмерного (one-size-fits-all) подхода, основанного на оптимальной доставке и механизмах восстановления TCP. Выгоды от разделения на слои были такими, как способность внедрять сетевую виртуализацию, например VXLAN, для поддержки мобильности рабочей нагрузки. Но схемы инкапсуляции, необходимые для виртуализации сети, оказывают негативное влияние. IP-сети должны рассматривать «максимальный блок данных протокола» (MAXPDU) для каждой линии связи. Когда маршрутизаторы добавляют дополнительные заголовки к фреймам, посылая их от одного линка к другому, MAXPDU между этими двумя линками может быть разным, что может привести (для IPv4) к необходимости того, чтобы маршрутизаторы фрагментировали фрейм (IPv6 обрабатывает фрагментацию на конечных узлах). Кроме того, MAXPDU может управляться через все линки для уменьшения фрагментации.

Аналогично, многие продукты Ethernet поддерживают "jumbo frames", которые позволяют передавать полезную нагрузку до 8 Кбайт в одном пакете, сокращая накладные расходы заголовков пакетов. Поскольку "jumbo frames" не поддерживаются повсеместно, преимущества обычно ограничиваются специализированными средами. Когда поддержка jumbo-фрейма несовместима, маршрутизаторы вынуждены запускать свои алгоритмы обработки MAXPDU. Сторонники фабрик IP/Ethernet иногда подчеркивают возможность использования "jumbo frames" в качестве преимущества, но эксперты (такие как Demartek) не рекомендуют включать "jumbo frames" для RoCEv2.

Это наследие сложности IP / Ethernet представляет собой проблему в премиум-среде без потерь данных: поведение оборудования по умолчанию, а также опыт и обучение вспомогательного персонала в основном ориентированное на рынок

мэйнстрима. Хотя должна быть возможность сконфигурировать Ethernet-оборудование поставщика и IP-оборудование для премиум-операции, такая операция не является нормальной по умолчанию и обычно не является необходимой конфигурацией для того же поставщика в другой роли в сети. Напротив, Fibre Channel всегда разрабатывался как сеть премиум-класса, и это будет так же верно в контексте NVMe, как и для сред SCSI на протяжении десятилетий.

Новые стеки создают новые угрозы

Одним из преимуществ хранения данных в Fibre Channel SAN было то, что к таким фабрикам трудно получить доступ через Интернет. Просто нет общепринятого пути к инфраструктуре, чтобы добраться из протокола Интернета до стабильного стека протоколов Fibre Channel. Злоумышленники не могут отправлять кадры Fibre Channel через Интернет, чтобы исследовать SAN. Таким образом, повторяющиеся мелкие ошибки безопасности, которые регулярно возникают, не приводят к катострафическим последствиям для хранимых данных (zero-day exposures). Сложные и относительно непродуманные стеки RoCEv2 и iWARP открывают новые возможности для угроз, которые доступны через Интернет, в результате чего добавляется сложность в управление межсетевыми экранами и другими механизмами безопасности во всей IP-сети организации.

NVMe over Fabrics и доступность

Многие функции хранилищ премиум-класса, такие как: active-active multipath I/O, автоматическое переключение на альтернативные пути доступа (failover) и обновления без прерывания работы зависят от глубокого тестирования производителем. Это тестирование также важно для производителя, чтобы обеспечить поддержку корпоративного класса продаваемого оборудования. Полнота тестирования оборудования, функционирующего с использованием Fibre Channel, была частично обеспечена простой структурой стека FC, а также отношениями между поставщиками хранилищ и поставщиками оборудования FC. По мере того, как клиенты корпоративных хранилищ будут внедрять NVMe, они должны будут работать со своими производителями. Построение архитектуры NVMe на основе Fibre Channel использует традиционную модель тестирования и поддержки корпоративных хранилищ. Использование же Ethernet/IP для NVMe может привести к снижению качества тестирования и, соответственно, к уменьшению доступности.

Параллельная инфраструктура Ethernet означает риски

Создание параллельной сети Ethernet SAN является рискованным подходом при внедрении сетевых систем хранения на базе NVMe. Хотя это может показаться легким способом начать работу в новой архитектуре, вопросы могут возникнуть по мере развертывания продуктивного использования: “Что такое SLA для новой SAN? Как мы можем перенести активы между нашим Ethernet SAN и нашим FC SAN? Какие варианты отката возможны, если мы выполняем миграцию между SAN? Как мы можем прогнозировать спрос, когда у нас есть два набора инфраструктуры?” Признавая, что переход от SCSI к NVMe займет годы, ясно, что этот переход – не просто событие, а многолетний процесс, на который будут сильно влиять краткосрочные решения по инфраструктуре.

Заключение

NVMe over Fibre Channel предлагает производительность и надежность транспорта Fibre Channel, а также возможность одновременного запуска протоколов FCP и FC-NVMe в одной инфраструктуре. Такой подход с использованием двух протоколов позволяет ИТ-организациям плавно переводить свои объемы хранения со SCSI на NVMe, либо на разных массивах, либо, когда массивы с двумя протоколами становятся доступными, в одном массиве.

При использовании NVMe over Fibre Channel нет необходимости в замене имеющегося оборудования SAN (rip-and-replace), а также в создании дорогой параллельной инфраструктуры, когда начинается внедрение NVMe. Двухпортовые HBA и универсальные стеки драйверов означают, что каждое приложение хранения может быть перемещено по мере необходимости. Активы SCSI можно переносить из SCSI в NVMe по томам. В первую очередь рекомендуется переносить тома с наименее критичными данными, при этом требующие наибольшей скорости доступа. Тома с критически важными данными могут быть перенесены позже. Кроме того, можно поддержать интересные сценарии, в которых основные копии ключевых активов создаются и поддерживаются на системах хранения корпоративного класса, а операционные копии могут публиковаться на недорогих массивах в одной и той же сети хранения данных для использования другими приложениями.

Дополнительная информация по адресу: Brocade.com/nvme.

Перевод материала компании Brocade подготовлен сотрудниками Российского филиала Brocade Николаем Умновым и Сергеем Целиковым.