НПО “БАУМ” начинает применять NVMe и адаптирует СХД под NVMeOF

Тестируем локальную NVMe СХД

Компания АО НПО «БАУМ» начала работу по адаптации ПО унифицированных и all-flash СХД для работы с применением технологий NVMe и RDMA. На текущий момент в лаборатории проходят нагрузочные тесты по уже реализованному протоколу iSER в сетях Mellanox. Как отмечают в «БАУМ»: “Современные быстрые флэш-массивы ограничены пропускной способностью интерфейса SAS, который в большинстве случаев является узким местом и причиной задержек. Мы попросили протестировать одного из наших заказчиков, у которого были жалобы на задержки и скорость передачи данных с СХД, «альфа»-версию нашего ПО, поддерживающего iSER. Стенд состоял из следующего оборудования:

- коммутатор Mellanox SwitchX-2;
- дисковый массив BAUM U42;
- карты ConnectX-4, установленные в дисковый массив;
- установленные в качестве кэш-памяти контроллеров твердотельные накопители NVMe PCIe SSD HGST – SN100 объемом 1600 Гбайт.

Результаты тестирования полностью удовлетворили заказчика. По сравнению с ранее тестируемыми all-flash массивами использование BAUM с поддержкой iSCSI RDMA (iSER) позволило снизить в десятки раз при многопоточной нагрузке.

В ходе тестирований было определено, что при использовании PCIe NVMe нагрузка на ядра процессора не такая большая как при использовании твердотельных накопителей с интерфейсом SAS, что позволяет параллельно обрабатывать тысячи запросов с минимальной задержкой.

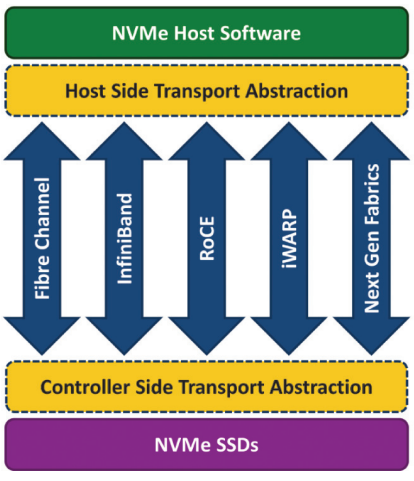

Рис. 1. Компоненты развиваемой архитектуры NVMe over Fabrics.

NVMe over Fabrics

Стандарт NVM Express over Fabrics (NVMeOF, *http://www.nvmexpress.org/wp-content/uploads/NVMe_Over_Fabrics.pdf*) определяет общую архитектуру, которая поддерживает ряд сетевых сред хранения данных для блочного протокола NVMe (рис. 1). Стандарт определяет интерфейс доступа к СХД, масштабирование большого количества устройств NVMe и увеличение расстояния внутри центра обработки данных, в котором можно получить доступ к устройствам NVMe и подсистемам NVMe.

Работа над спецификацией NVMeOF началась в 2014 году с целью распространения NVMe на такие фабрики, как: Ethernet, Fibre Channel и InfiniBand®. NVMeOF предназначена для работы с любыми подходящими технологиями хранения данных. Эта спецификация была опубликована в июне 2016 г.

В настоящее время разрабатываются два вида транспортов для NVMe:

- NVMeOF с использованием RDMA;
- NVMeOF с использованием Fibre Channel (FC-NVMe).

Использование RDMA с NVMeOF включает в себя любую из технологий RDMA: InfiniBand, RoCE и iWARP. Разработка NVMeOF с RDMA определяется технической подгруппой организации NVM Express.

FC-NVMe развивается комитетом INCITS T11, который разрабатывает все стандарты интерфейса Fibre Channel. В рамках этого комитета по развитию FC-NVMe также ожидается, что будут разрабатываться стандарты с использованием Fibre Channel over Ethernet (FCoE).

Целью разработки NVMeOF является обеспечение дистанционной связи с устройствами NVMe с дополнительной задержкой не более 10 мкс (мкс) к задержке при доступе к собственным устройствам NVMe внутри сервера.

Есть несколько вариантов использования NVMeOF. *Первый* представляет собой систему хранения, состоящую из многих устройств NVMe, использующих NVMeOF с интерфейсом RDMA или Fibre Channel, что обеспечивает полное законченное NVMe-решение для хранения. Эта система обеспечивала бы чрезвычайно высокую производительность, сохраняя при этом очень низкую задержку, доступную через NVMe.

Вторая реализация будет использовать NVMeOF для достижения низкой латентности при подключении к подсистеме хранения, которая использует более традиционные протоколы для обработки ввода/вывода на каждом из SSD в этой системе. Это позволило бы использовать преимущества упрощенного стека программного обеспечения хоста и более низкую задержку по сравнению с прямым подключением, используя преимущества существующих технологий подсистем хранения.

В настоящее время решения могут масштабироваться до сотен устройств NVMe, в так называемых «рэк-масштабируемых» (“rack scale”) решениях для совместного хранения данных. В будущем решения NVMeOF смогут масштабироваться до тысяч устройств NVMe в большом совместном хранилище, обеспечивая данными сотни и/или тысячи приложений (рис. 2).

Различия между локальным протоколом NVMe и NVMeOF

Примерно 90% протокола NVMe over Fabrics совпадает с локальным протоколом NVMe. Это включает пространства имен NVMe, ввод/вывод и административные команды, регистры и свойства, состояния питания, асинхронные события, резервирование и другие. Основные различия заключаются в четырех областях, перечисленных в табл. 1.

Эти различия в первую очередь интересны разработчикам продуктов NVMe, поскольку их драйверы устройств должны правиль-

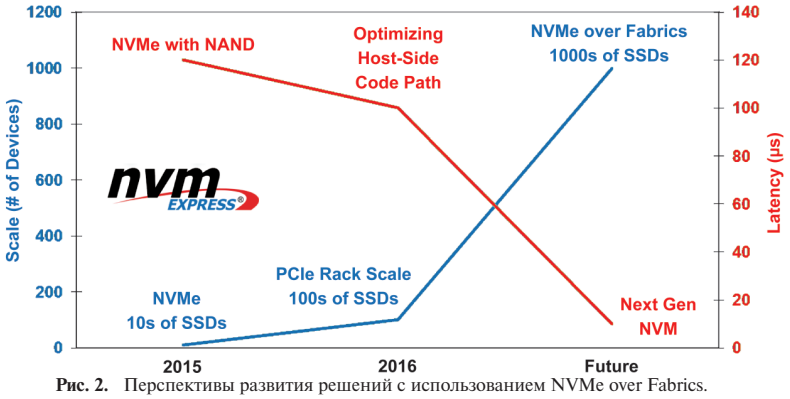

Рис. 2. Перспективы развития решений с использованием NVMe over Fabrics.

Табл.1. Отличительные особенности локалального подключения NVMe-устройств – по PCIe и сетевого – по NVMe over Fabrics.

Differences	PCI Express (PCIe)	NVMe over Fabrics
Identifier	Bus/Device/Function	NVMe Qualified Name (NQN)
Discovery	Bus Enumeration	Discovery and Connect commands
Queueing	Memory-based	Message-based
Data Transfers	PRPs or SGLs	SGLs only, added Key

PRP: Physical Region Page (physical memory page address, PCIe transport only)
SGL: Scatter-Gather List (list of locations and lengths for read or write requests)

но обрабатывать как локальные устройства NVMe, так и удаленные устройства NVMe. Некоторые из этих элементов, например, идентификатор, могут быть доступны конечным пользователям, чтобы помочь идентифицировать специфичные устройства NVMe для конкретных приложений. Механизм обнаружения предназначен для работы с несколькими типами транспорта для NVMe over Fabrics.

Транспортное мапирование NVMe

В локальной реализации NVMe команды и ответы NVMe отображаются в общую память на хосте через интерфейс PCIe. Фабрики основаны на концепции отправки и получения сообщений без общей памяти между конечными точками. Поэтому транспортные сообщения NVMe-фабрик инкапсулируются в «капсулу», которые включают одну или несколько команд или ответов NVMe. Капсулы или комбинация капсул и данных не зависят от конкретной технологии фабрики и отправляются и принимаются по выбранной технологии фабрики (рис. 3).

Для NVMe over Fabrics поддерживается вся многоочередная модель NVMe, использующая обычные очереди и очереди завершения NVMe, но инкапсулированные в транспортное сообщение (message-based transport). Очередь ввода/вывода NVMe (представление и завершение) предназначена для многоядерных процессоров, и этот эффективный дизайн с низкой латентностью поддерживается в NVMe over Fabrics.

При отправке сложных сообщений на устройство NVMe с использованием NVMe over Fabric, капсулы позволяют отправлять несколько небольших сообщений в виде одного сообщения, что повышает эффективность передачи и уменьшает задержку. Капсула представляет собой либо запись очереди отправки, либо запись очереди завершения в сочетании с некоторым количеством данных, метаданных или списков сборок Scatter-Gather (SGL). Содержимое элементов совпадает с локальным протоколом NVMe, но капсула – это способ объединить их вместе для повышения эффективности.

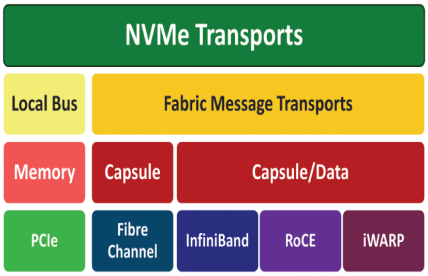

Рис. 3. При локальном подключении NVMe-устройств для NVMe-команд используется общая память, при подключении по NVMe over Fabrics – капсулы.

RDMA и iSER

Прямой доступ к памяти (Direct Memory Access, DMA) – это способность устройства напрямую обращаться к памяти хоста без вмешательства центрального процессора *(https://community.mellanox.com/docs/DOC-1963)*. Удаленный прямой доступ к памяти (Remote DMA, RDMA) – это способность доступа (чтение, запись) к памяти на удаленном компьютере без прерывания обработки процессора(ов) в системе, от которой поступает запрос.

Преимущества RDMA:

– *zero-copy* – приложения могут выполнять передачу данных без участия сетевого программного стека. Данные отправляются и принимаются непосредственно в буферы без копирования между сетевыми уровнями;

– *обход ядра (kernel bypass)* – приложения могут выполнять передачу данных непосредственно из пользовательского пространства без участия ядра;

– *отсутствие участия ЦП* – приложения могут обращаться к удаленной памяти, не затрачивая процессорного времени на удаленном сервере. Удаленный сервер памяти будет считан без участия удаленного процесса (или процессора). Более того, кэши удаленного CPU не будут заполняться доступным содержимым памяти.

Для использования RDMA нужны: 1) сетевой адаптер с возможностью RDMA (например, семейство адаптеров Mellanox Connect-X); 2) протокол канального уровня сети – может быть либо Ethernet, либо InfiniBand – оба могут передавать приложения на основе RDMA. RDMA поддерживается в операционных системах Linux, Windows и VMware. В других операционных системах (или для расширенных функций) вам может понадобиться загрузка и установка соответствующего пакета драйверов и настройка его соответствующим образом.

Примеры имплементации RDMA: Storage RDMA Protocol (SRP); iSCSI Extensions for RDMA (iSER); Windows SMB Direct; Lustre (File System); IBM GPFS (Platform Computing); IPoIB; Apache Hadoop (UDA & R4H); Gluster; Ceph; NFSoRDMA; NVMeOF.

iSER расшифровывается как «расширение iSCSI для RDMA» *(https://community.mellanox.com/docs/DOC-1466)*. Это расширение модели передачи данных iSCSI, стандарта сети хранения данных для TCP/IP. Он использует компоненты iSCSI, пользуясь преимуществами набора протоколов RDMA. iSER – это транспорт RDMA для iSCSI, протокол связи может быть Ethernet или InfiniBand на любой поддерживаемой скорости (10,40,56,100 Gb/s).

iSER использует набор протоколов RDMA для обеспечения более высокой пропускной способности для передачи блоков хранения (режим копирования с нулевым временем). К тому же, это устраняет накладные расходы при обработке TCP/IP, сохраняя при этом совместимость с протоколом iSCSI.

Кроме того, iSER имеет: самую низкую задержку и минимальное использование ЦП; обладает стабильностью и преимуществами протокола iSCSI, такими как безопасность, высокая доступность и др.; быстрее, чем iSCSI, FC, FCoE и проще в управлении, чем SRP.

Среди поддерживаемых iSER целевых СХД: 1) на Linux – Linux IO (LIO), Linux iSCSI target framework (TGT), Generic SCSI target subsystem for Linux (SCST); 2) Oracle ZFS, Violin Memory, Zadara Saratoga Speed, HP SL4540 Moonshot server и др.

Планы на будущее

Унифицированные all-flash СХД «БАУМ» с поддержкой iSER будут доступны в продаже к 3-му кварталу 2017 года.

Говоря о перспективах, «БАУМ» сообщил редакции, что переход исключительно на NVMe накопители пока не представляется возможным из-за неготовности аппаратной платформы. Использование PCIe NVMe в системах хранения «БАУМ» пока планируется использовать только в качестве кэш-памяти второго уровня для чтения, что позволит увеличить производительность и надежность.

«БАУМ» планирует также в ближайшие полтора года осуществить поддержку NVMeOF на своих массивах с использованием RDMA и Fibre Channel: FC-NVMe, iWARP, NFS-over-RDMA.

С большим интересом изучаем возможности технологии BlueField™ Multicore System on Chip. Семейство продуктов BlueField – это высокоинтегрированная система на кристалле (SoC), оптимизированная для систем хранения NVMe, виртуализации сетевых функций (NFV), систем безопасности и встроенных устройств. BlueField интегрирует межсоединения Mellanox ConnectX® и архитектуры процессоров ARM в одно устройство. Решения BlueField получат применение на следующем этапе развития наших унифицированных систем хранения, когда станут доступны модули хранения JBOF, построенных исключительно на PCIe NVMe накопителях, таких как HGST SN100 и SN200. All-flash СХД «БАУМ» с использованием BlueField SoC в будущем позволят сократить задержки до 150 наносекунд и повысить пропускную способность до 128Гбайт/с.

Семейство устройств SoC BlueField интегрирует: массив 64-разрядных ядер ARMv8 A72, соединенных сетью; контроллеры памяти DDR4; несколько портов Ethernet/InfiniBand, поддерживающих 10/25/40/50/100 Gb/s; интегрированный PCIe-коммутатор с портами PCIe Gen 3.0/4.0 и поддержкой функциональности EP SR-IOV и RC. SoC предназначен для высоконагруженных приложений и интенсивно использующих ввод/вывод, которые объединяют сервисы данных и сетевые сервисы.

На текущий момент платформа позволяет подключать до 16 SSD и имеет 2 PCIe x16. Внутренние слоты PCIe также могут использоваться для задач компрессии и дедупликации.