Hewlett Packard
Enterprise

HPE J2000 Flash Enclosure

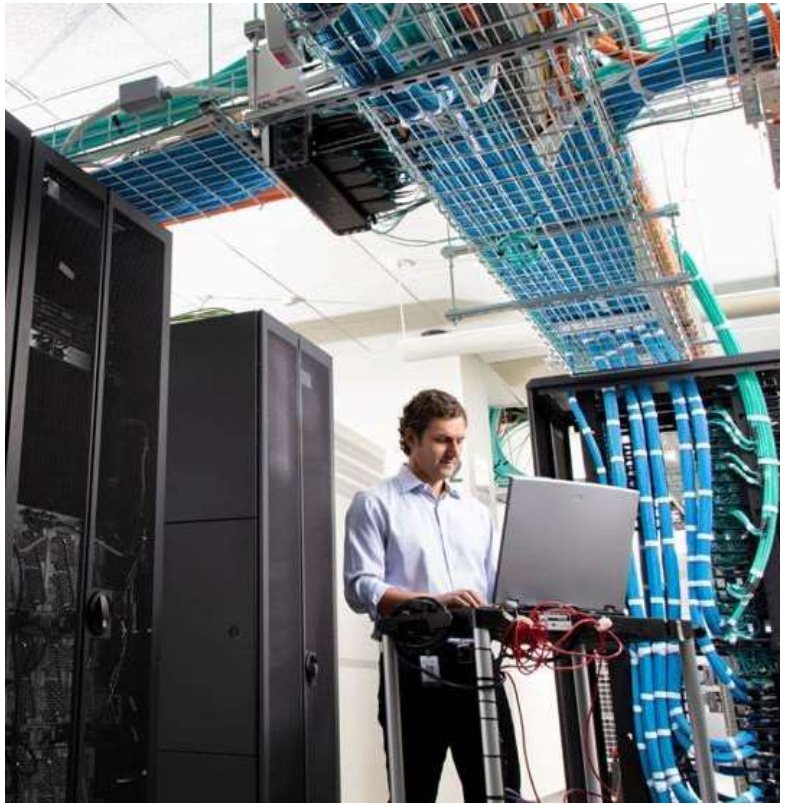
Илья Семухин, менеджер по продуктам, HPE в России

15 апреля 2021 г.

Новый этап развития СХД и технологии NVMe-oF™

Экстремальная производительность и низкие задержки передачи данных

Эффективность хранения благодаря возможности распределения ресурсов между разными серверами



Строительный блок для организации решения хранения и обработки данных

Представляем полку расширения HPE J2000 Flash Enclosure

Следующая ступень развития флэш-хранилищ

Особенности архитектуры

- Форм-фактор 2U, NVMe-over-Fabric JBOF
- Поддержка 24 SFF Mixed-Use NVMe SSDs
- Два модуля ввода-вывода (active-active) режим высокой доступности
- Использует протокол NVMe® / RoCEv2
- Прямое или коммутируемое соединение с серверами HPE через адаптеры 100GbE
- Графический интерфейс управления
- Поддержка Redfish и RESTful API

Преимущества

- Новая категория продукта в традиционном форм-факторе
- Решение для хранения данных только во флэш-памяти.
- Высокая доступность с возможностью распределения ресурсов
- Высокая производительность и низкие задержки
- Гибкое и масштабируемое подключение по сетевые 100GbE
- Соответствие отраслевым стандартам HPE Redfish / Swordfish
- Доступ к хранилищу с минимальными задержками

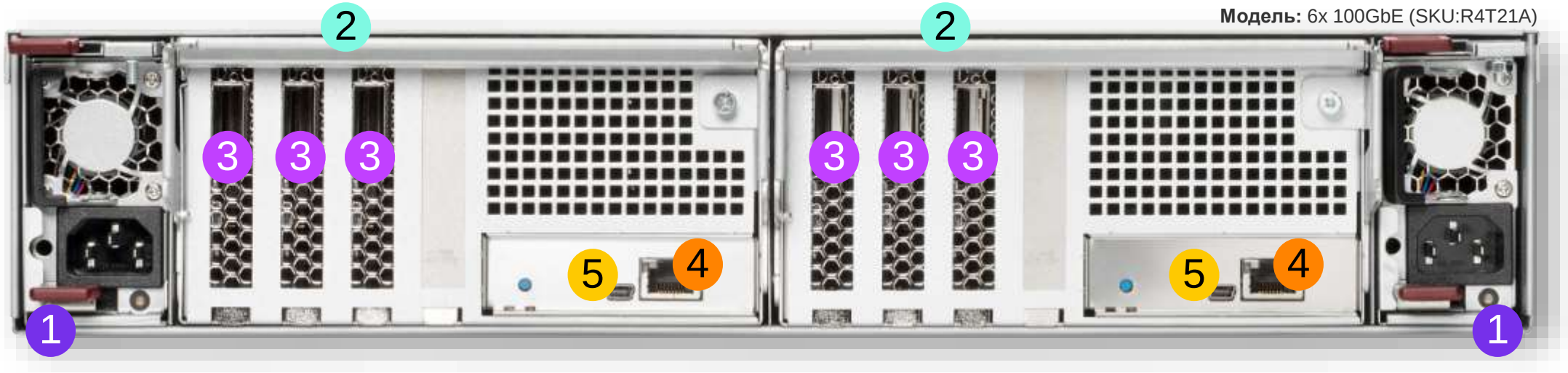
Архитектура HPE J2000 Flash Enclosures

Внешний вид и компоненты

1. Шасси 2U
2. SFF NVMe SSD
3. LED индикаторы состояния системы
4. LEDs индикаторы состояния накопителя

Архитектура HPE J2000 Flash Enclosures

Внешний вид и компоненты



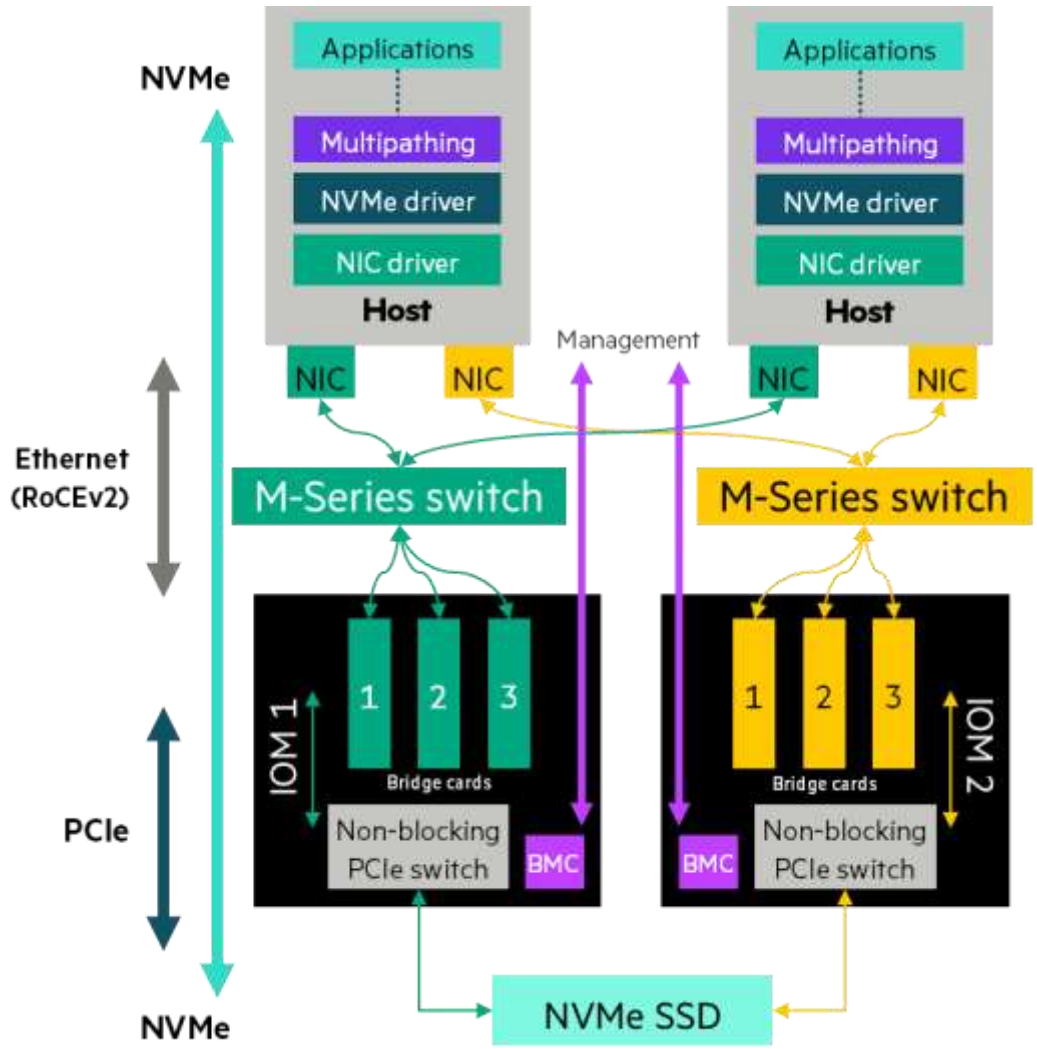
Модель: 6x 100GbE (SKU:R4T21A)

1. Блоки питания и модули охлаждения
2. Модули ввода-вывода (IOMs)
3. Сетевой адаптер
4. 1GbE порт управления
5. Mini-USB серийный порт[1] (CLI)

[1] Ограниченные возможности, например, просмотр/изменение IP-адреса

Архитектура HPE J2000 Flash Enclosures

- HPE J2000 может быть подключена к серверам HPE ProLiant и HPE Apollo напрямую или через коммутаторы
- Хосты подключаются к накопителям по протоколу RoCEv2
- Каждому накопителю в J2000 присваивается NVMe Qualified Name (NQN), в отличие от SAN, где уникальный идентификатор присваивается хост-порту контроллера или СХД в целом



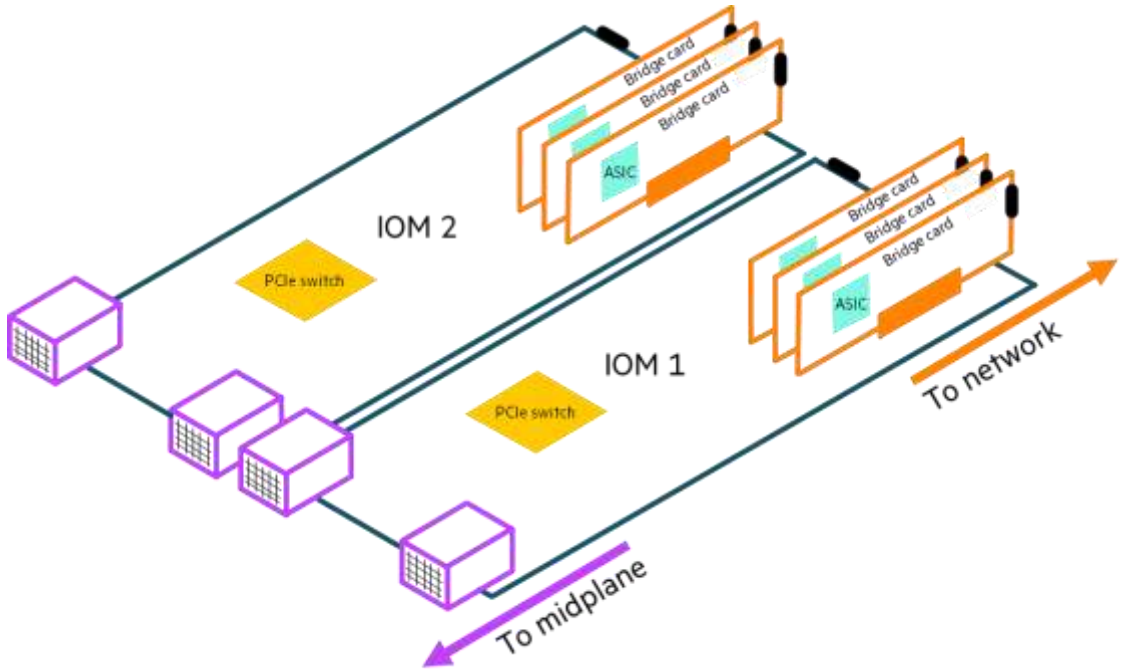
Примечание:

- На момент анонса кластеризованные файловые системы не поддерживаются
- Совместное использование - это возможность подключения накопителей по двум портам к одному хосту, а также возможность подключения нескольких хостов к одному шасси или накопителю

Архитектура HPE J2000 Flash Enclosures

Модули ввода-вывода

- Модули ввода-вывода (IOM) транслируют данные между сетевыми адаптерами (bridge cards) и NVMe накопителями
- HPE J2000 поставляется с двумя модулями ввода-вывода
  - В зависимости от модели каждый модуль поставляется с одним или тремя сетевыми адаптерами, которые устанавливаются на заводе
- Каждый модуль ввода-вывода содержит:
  - Контроллер управления
  - Неблокируемый PCIe коммутатор, который передает данные от накопителей к сетевым адаптерам

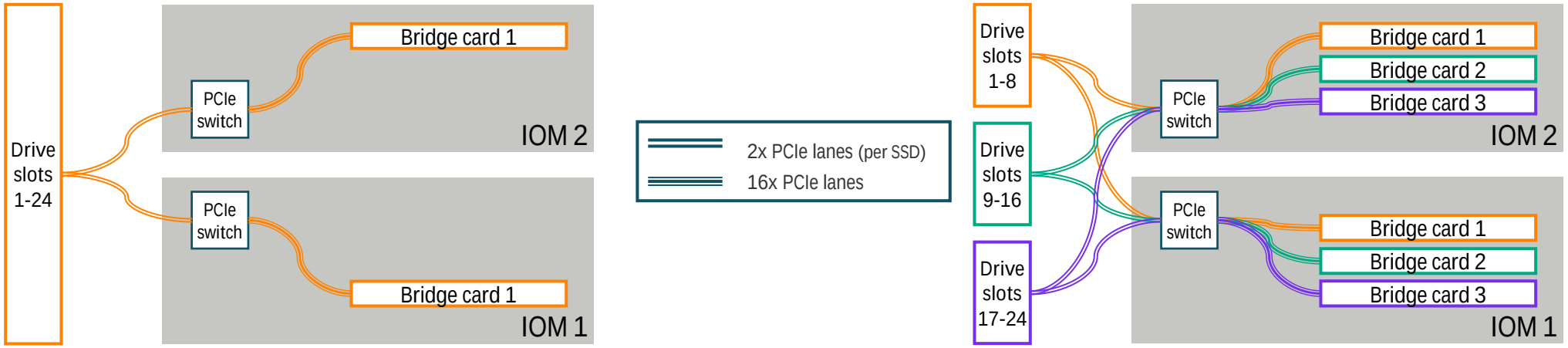


Примечание:

- Модули ввода-вывода могут быть заменены целиком, замена отдельных сетевых карт ввода-вывода не поддерживается

Архитектура HPE J2000 Flash Enclosures

Подключение NVMe накопителей



HPE J2000 2x100GbE (SKU: R4T20A)

- Два сетевых адаптера обеспечивают высокую производительность по доступной цене
- Доступ ко всем накопителям осуществляется через обе сетевые карты, позволяя сократить расходы на коммутацию
- 48x PCIe линий (downlinks) к накопителям и 16x PCIe линий (uplinks) до сетевых адаптеров

HPE J2000 6x100GbE (SKU: R4T21A)

- Шесть сетевых адаптеров позволяют обеспечить максимальную производительность
- Каждая сетевая карта подключена к восьми накопителям, что требует дополненных расходов на коммутацию
- 48x PCIe линий (downlinks) к накопителям 48x PCIe линий (uplinks) до сетевых адаптеров

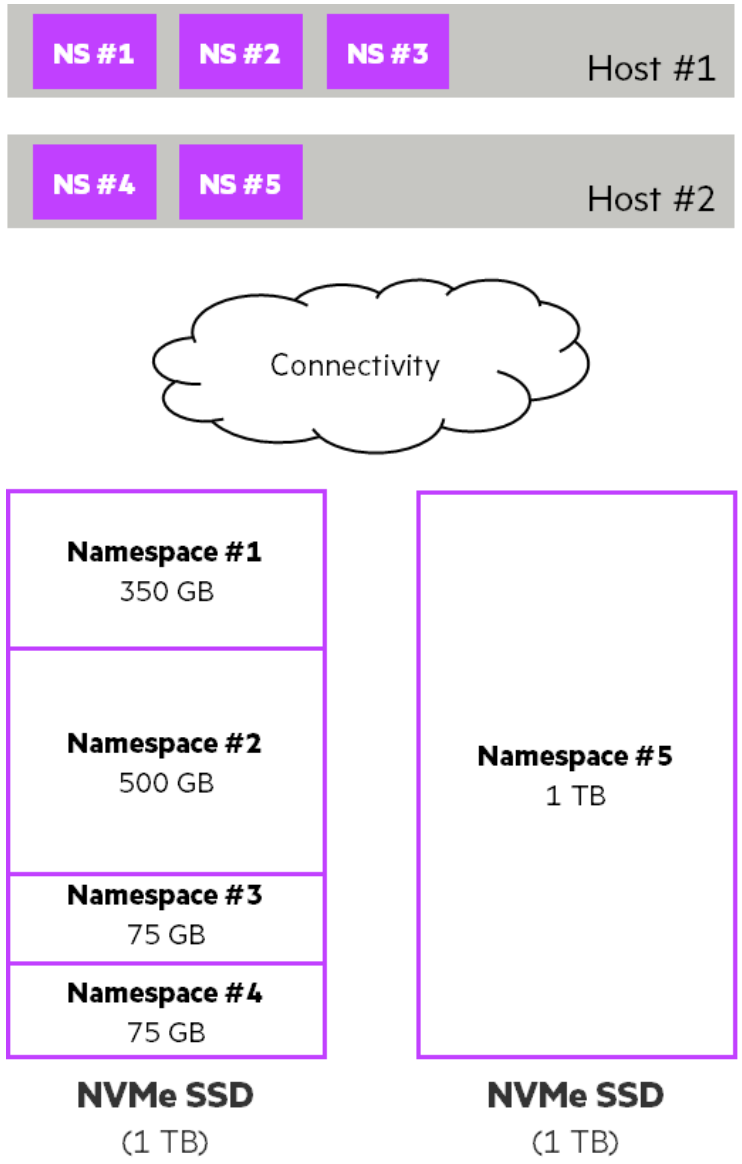
Архитектура HPE J2000 Flash Enclosures

Namespaces

- *Namespace* это логический сегмент NVMe накопителя
  - Можно сравнить с созданием раздела на жестком диске
- На данный момент поддерживается презентация *namespace* только одному хосту, но несколько хостов могут совместно использовать один SSD накопитель
  - Для оптимизации производительности HPE рекомендует презентовать один накопитель одному хосту
  - Планируется добавить поддержку файловых систем с функцией “*fused Compare and Write”*, таких как VMware VFMS

Примечание:

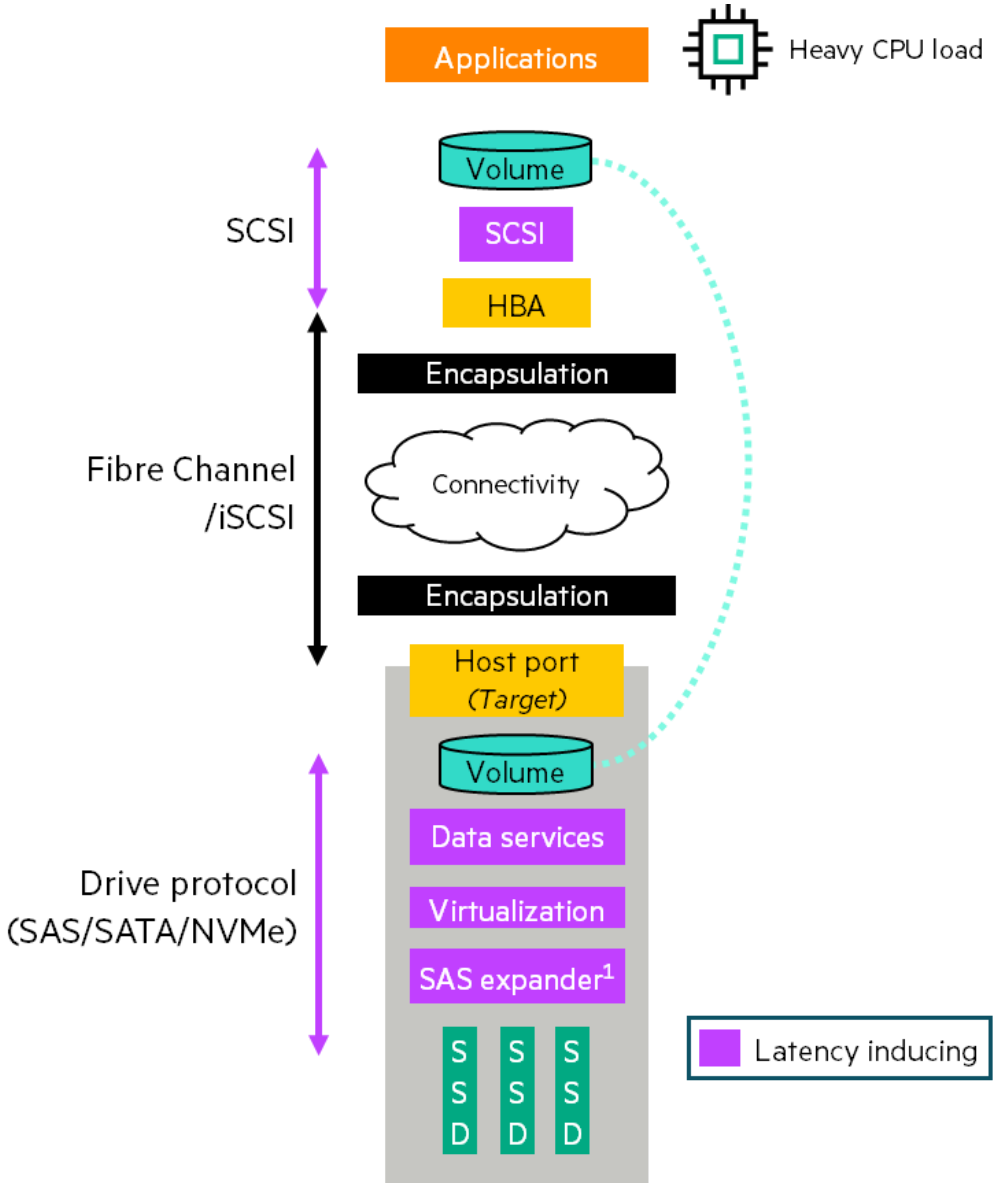
- Поддерживается создание до 4 *Namespaces* на одном накопителе
- *Namespaces* могут быть только созданы или удалены
- *Namespaces* не поддерживают объединение
- *Namespaces* не имеют собственного NQN, они используют NQN диска, на котором они размещены.
- *Namespace* не имеет собственной избыточности или отказоустойчивости



Сравнение SAN и HPE J2000

Традиционная SAN архитектура

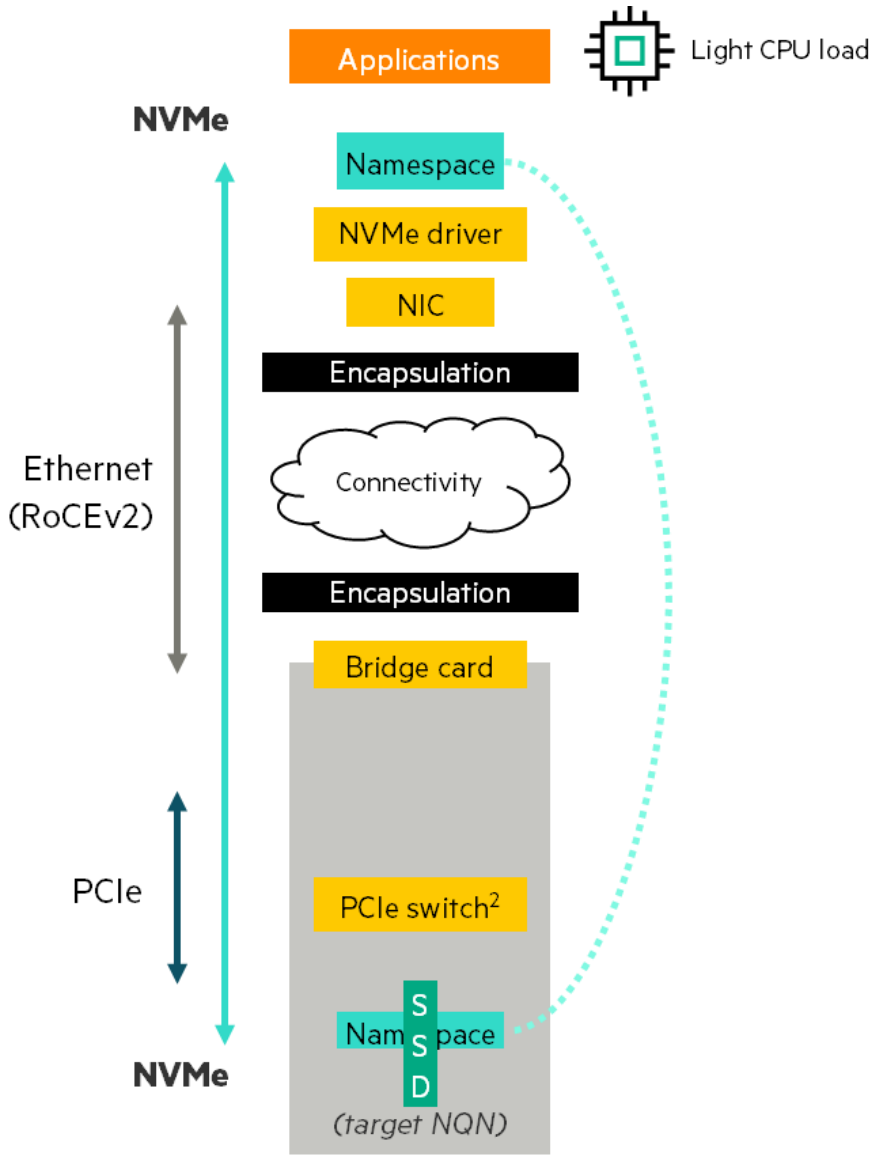
- Традиционные SAN массивы используют набор дополнительных функций, которые привносят задержки при передаче данных, по сравнению с твердотельными NVMe накопителями, установленными в сервер и/или подключенными по протоколу RoCEv2
  - Виртуализация данных, например RAID
  - Сервисы работы с данными такие как дедупликация и компрессия
  - Инкапсуляция протоколов, например SCSI в NVMe
  - Накладные расходы SCSI
  - Последовательный стек команд
- Некоторые SAN массивы поддерживают установку NVMe накопителей и технологию «End-to-End NVMe-oF», однако задержки все равно будут присутствовать:
  - Использование виртуализации и сервисов по работе с данными
  - Инкапсуляция протокола NVMe в SCSI



Сравнение SAN и HPE J2000

NVMe over RoCEv2

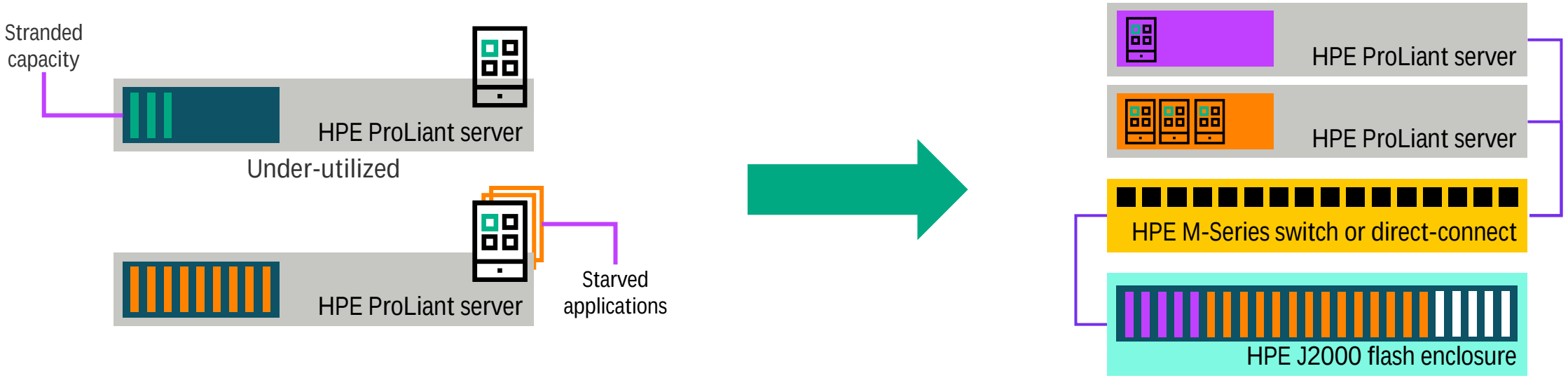
- Протокол RoCEv2 позволяет построить «end-to-end NVMe» инфраструктуру, передавая данные на значительные расстояния с низкой задержкой
- Технологии и архитектуры, которые обеспечивают производительность схожую с дисковыми полками прямого подключения (DAS)[1].
  - Протокол Remote Direct Memory Access (RDMA) снижает задержку и улучшает пропускную способность, не увеличивая нагрузку на операционную систему и центральный процессор
  - Устранение уровня SCSI, вызывающего высокую задержку
  - Исключение сервисов передачи данных или виртуализации хранилища любого типа
  - Неблокируемая PCIe архитектура = отсутствие конкуренции за внутреннюю полосу пропускания
  - Поддержки до 64K параллельных очередей ввода-вывода



[1] NVMe SSDs напрямую подключение к хост-шине PCIe

Примеры использования HPE J2000 Flash Enclosures

Новый подход к построению DAS решений



Недостатки внутренних накопителей

- Ограниченное количество
- Накопители необходимо физически перемещать между серверами в случае необходимости
- Может привести к приобретению незадействованных накопителей для выполнения краткосрочных задач
- Постоянный поиск компромисса между емкостью и производительностью

Преимущества дезагрегированного подхода

- Эффективное использование доступной емкости
- Гибкое перераспределение ресурсов
- Снижение совокупной стоимости владения
- Производительность сопоставимая с локально установленными дисками

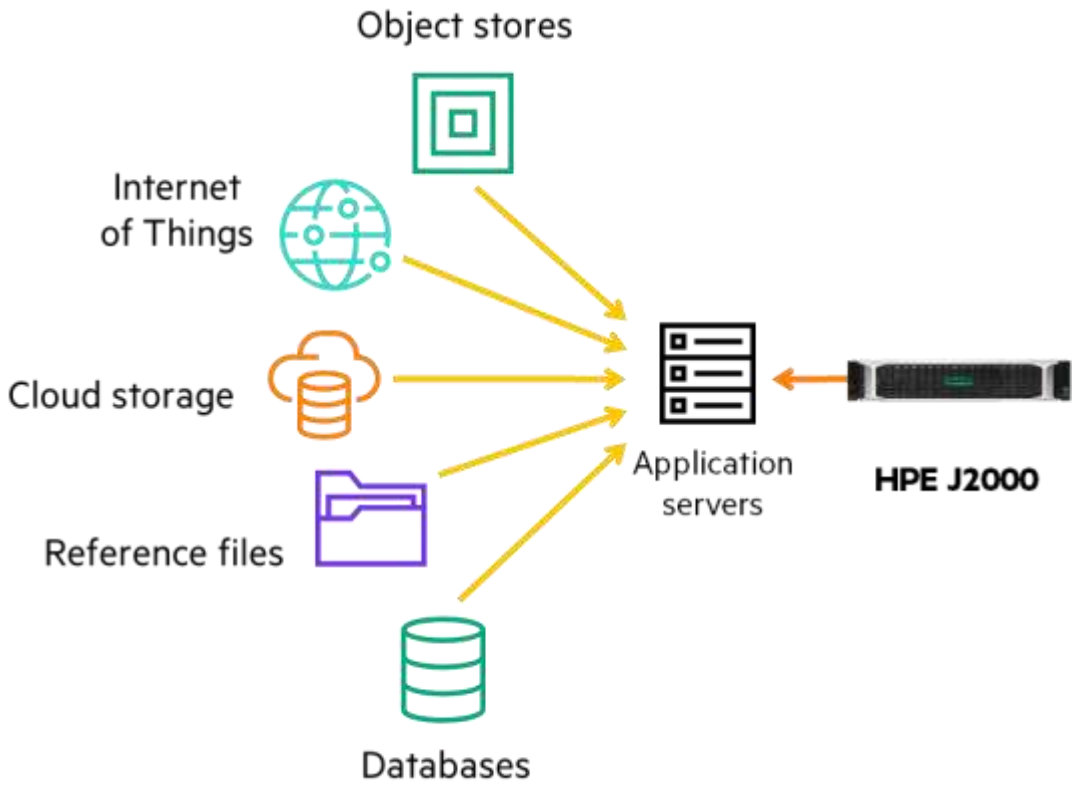
Примеры использования HPE J2000 Flash Enclosures

Быстрое хранилище для аналитических рабочих нагрузок

- В HPC - задачах, связанных с моделированием сложных физических процессов, относительно небольшие объемы данных могут быстро разрастаться до сотен терабайт или даже нескольких петабайт. Для таких типов нагрузок HPE J2000 будет отличным решением - эта система позволит быстро сохранять и получать доступ к очень горячим данным

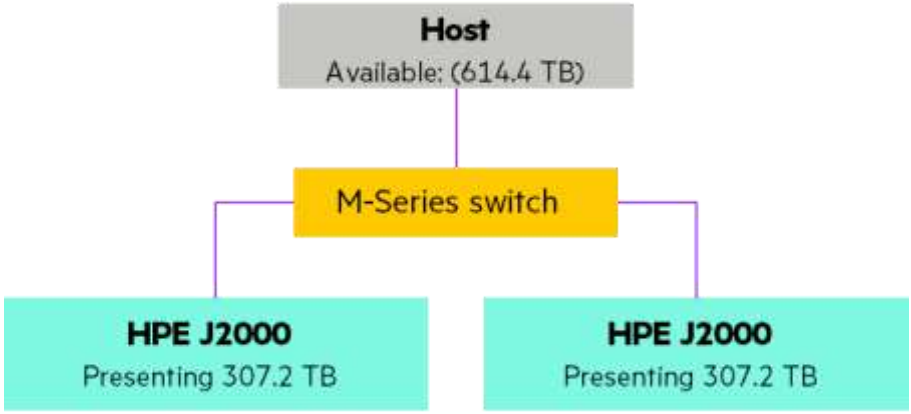
- Рендеринг, анализ, добавление информации и кодирование видеоконтента на лету требует быстрого перемещения больших объемов данных

- Строительный блок, для очень быстрых систем хранения, в которых требуется массивный и быстрый ввод и вывод



Масштабируемость HPE J2000 Flash Enclosures

- Поддерживает установку до 24 NVMe SFF накопителей
- Максимально возможная емкость ограничена емкостью накопителей
- На данный момент поддерживаются следующие накопители:
  - 1.6 ТБ
  - 3.2 ТБ
  - 6.4 ТБ
  - 12.8 ТБ
- Максимальная емкость одной полки 307.2 ТБ
- Дальнейшее расширение емкости осуществляется посредством подключения нескольких дисковых полок J2000 к одному хосту



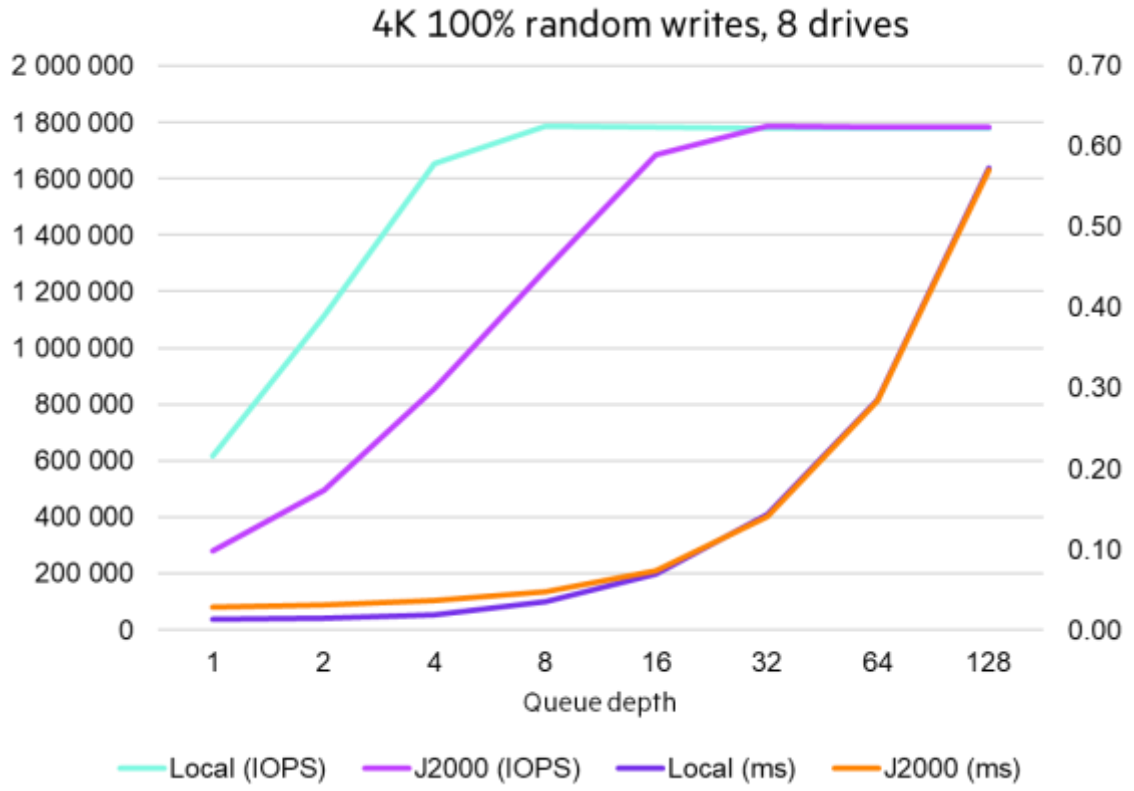
Примечание:

- Дисковые полки не могут быть подключены последовательно (daisy-chained), как традиционные SAS JBOD

Производительность HPE J2000 Flash Enclosures

По сравнению с внутренними дисками

- NVMe накопители подключаются напрямую к PCIe шине (исключая SAS или SATA), позволят добиться максимальной производительности операций случайного доступа к флэш-памяти, также обладают упрощенным набор команд с очередью в 64 КБ
- NVMe-over-Fabric расширяет возможности использования NVMe накопителей
- NVMe/RoCEv2 обеспечивает передачу данных с минимальными задержками на большие расстояния
- Результаты тестирования производительности J2000 показаны на графике справа[1]
- Более 10M IOPS на операциях случайного чтения и 5M на операциях случайной записи
- Более 70 ГБ/с пропускная способность на чтение и 55 ГБ/с на запись
- Стоит принимать во внимание: общая производительность решения будет зависеть не только от JBOF, включая сервер, сетевые интерфейсы и топологию, а также возможности приложений



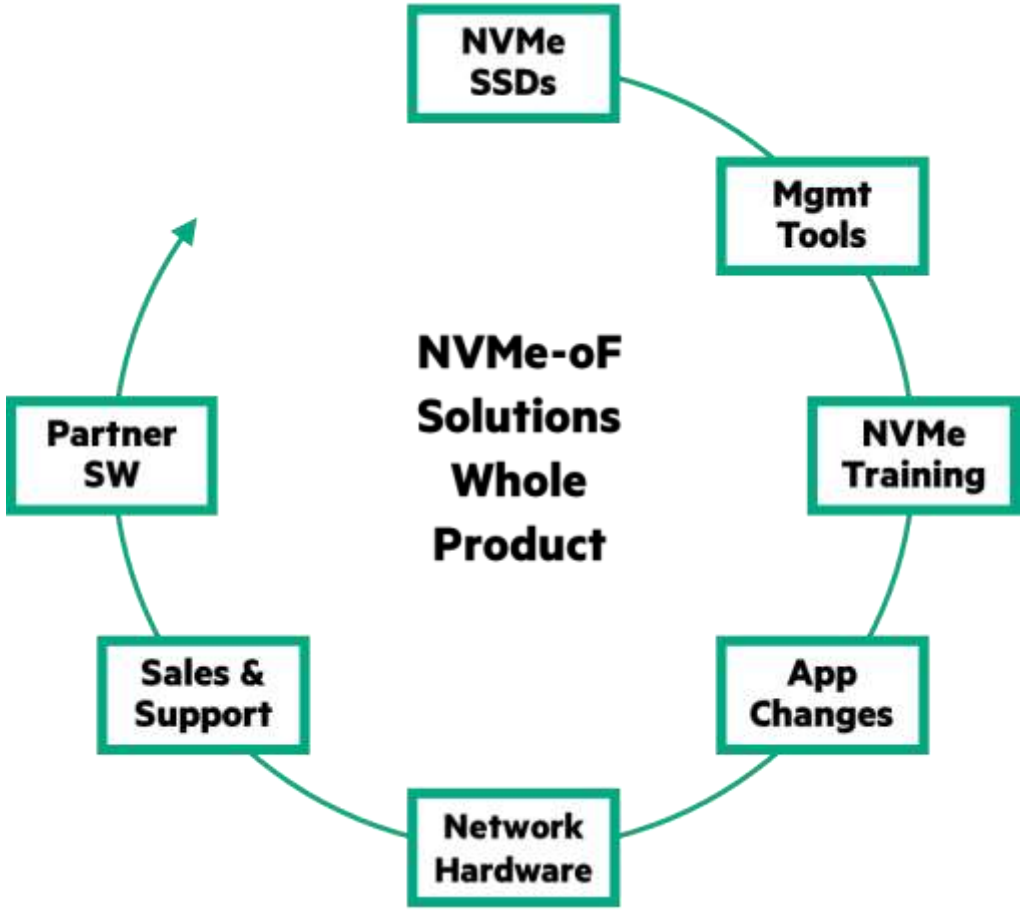
[1] - Тестирование проводилось в HPE Lab использовались серверы HPE DL380 Gen10 с RHEL 8, два 100GbE NICs подключенных через SN2100M 16-портовый коммутатор подключенный к J2000 с 6 сетевыми адаптерами и установленными 3.2TB MU SSDs

Экосистема NVMe и NVMe-oF только начинает свое развитие

Мы в начале трансформации, которая изменит системы хранения данных

- Потребуются годы, чтобы повторить путь SAS / SATA
- Спецификации NVMe-oF являются новыми и идет активная их разработка
- Не все приложения могут используют возможности, которые предоставляют NVMe накопители
- Для NVMe-oF существует несколько инструментов управления сетью
- Растет опыт работы с NVMe технологиями
- Возможности HPE J2000 будут развиваться вместе с развитием экосистемой NVMe

- HPE предлагает поддержку комплексных NVMe решений, включая серверы, сетевые коммутаторы, решения для хранения данных и управления, в сочетании с экспертизой и поддержкой продуктов



HPE занимает лидирующие позиции в области решений для хранения данных NVMe-oF!

Спасибо за внимание!

Илья Семухин, менеджер по продуктам, HPE в России
Ilya.semoukhin@hpe.com