Pavilion: 200 млн IOPS на стойку

В конце ноября 2020 г. компания Pavilion Data Systems объявила [1] о выпуке 3-й версии Pavilion HyperOS, в которой для NFS и S3 была добавлена поддержка множества массивов Pavilion HyperParallel Flash Array с единым глобальным пространством имен. На настоящий момент Pavilion HFA является одним из самых производительных массивов в своем классе с одновременной поддержкой блочного, файлового и объектного доступа.

Введение

Компания Pavilion Data Systems была образована в 2014 г., выход на рынок – 2019 г. Решния Pavilion были отмечены рядом наград на международных конференциях. Основные разработки: NVMe-массив Pavilion HyperParallel Flash Array и Pavilion HyperOS для его управления.

Гурприт Сингх (Gurpreet Singh), генеральный директор Pavilion Data, на видео-брифинге для прессы (ноябрь 2020 г.) сказал, что третий квартал 2020 г. Pavilion был блокбастером с почти 10-кратным ростом выручки по сравнению с прошлым годом, что эквивалентно всей выручке Pavilion за 2019 год (https://blocksandfiles.com/2020/11/18/pavilion-takes-on-wekaio-and-hpc-filers-with-hyperparallel-file-system/).

«Фабрика Cisco ACI с безопасностью нулевого доверия (zero trust security) и передовыми методами контроля перегрузки трафика идеально подходит для хранения данных с использованием СХД с поддержкой NVMe over Ethernet. В лаборатории Cisco по исследованиям и тестированию NVMe мы использовали массивы Pavilion HyperParallel Flash Arrays для тестирования сессий с десятками миллионов IOPS over NVMe-TCP с серверами Cisco UCS. Возможность изменять протоколы одним нажатием кнопки обеспечивает непревзойденную простоту использования. Pavilion HyperParallel Data Platform с поддержкой нескольких контроллеров, гибкостью, простотой использования и управления устанавливает отраслевой стандарт производительности для хранения данных NVMe» – Камал Бакши (Kamal Bakshi), директор, Cisco Systems.

«Уникальная гиперпараллельная архитектура Pavilion Data может одновременно поддерживать блочные, файловые и объектные протоколы с потрясающей производительностью, – сказал Эрик Бургенер (Eric Burgener), вице-президент по исследованиям группы инфраструктурных систем, платформ и технологий IDC. – Система способна вместить до 20 процессоров и 72 устройства хранения в форм-факторе 4U, эффективно обеспечивает чрезвычайно высокую производительность, низкую задержку, предсказуема при масштабировании, обеспечивая высокоэффективную сквозную (end-to-end) платформу на основе NVMe, подходящую как для облачных нагрузок высокой производительности, так и для традиционных рабочих нагрузкок, которая по производительности значительно опережают конкурентов, во многих случаях занимая меньше места на 50–75%».

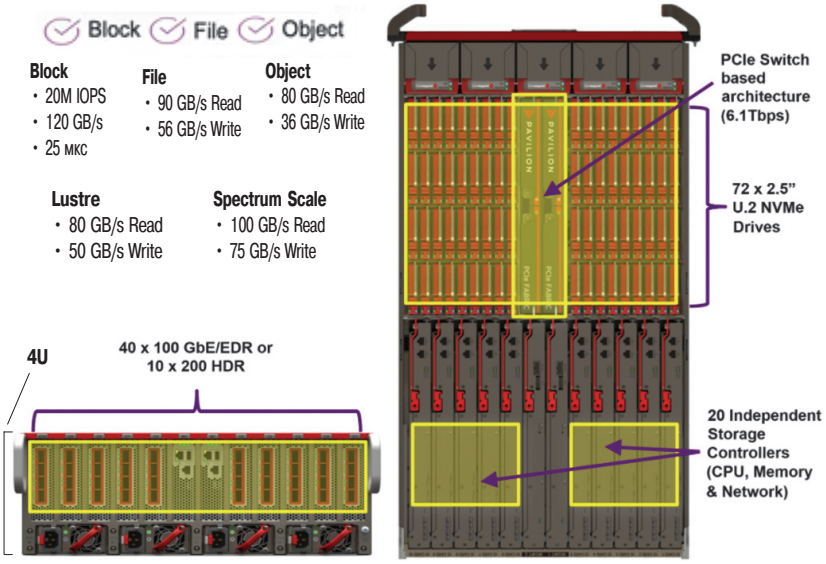

Рис. 2. Внешний вид и основные характеристики производительности (по информации разработчика) Pavilion HyperParallel Flash Array.

Pavilion HyperParallel Flash Array

Массив Pavilion HyperParallel Flash Array (Pavilion HFA) в полной комплектации представляет кластерное 4U хранилище, которое может масштабироваться от 2 до 20 контроллеров и от 4 до 4 100 GbE/GDR портов. Массив разработан на базе стандартных компонент, без использования специализированных ASIC или FPGA. Каждый контроллер запускает свой экземпляр HyperOS (включает в себя Pavilion HyperParallel File System™) для поддержания согласованности. Каждый Pavilion HFA в полной комплектации, по заявлениям разработчика, поддерживает до 20 млн IOPS и до 2,2 Пбайт сырой емкости (540 Тбайт/U; от 18 до 72 2,5" накопителей NVMe на основе технологий TLC и SCM). Можно отдельно масштабировать емкость или вычислительную обработку.

Управление контроллерами может поддерживаться через GUI, CLI, и APIs для интеграции с V-Center, Kubernetes CSI, RedFish™ и SwordFish™.

Архитектура массива и ПО HyperOS разработаны "с нуля" под особенности реализаций NVMe/NVMe-oF решений, благодаря чему удалось достичь высоких показателей производительности (рис. 1). Внешний вид Pavilion HFA представлен на рис. 2.

Pavilion HFA разработан на основе виртуализированного сетевого коммутатора NVMe-oF (рис.3). Компоновка без кеширования позволяет контроллерам записывать на любой диск, используя удаленный прямой доступ к памяти. Эффект – производительность, эквивалентная хранилищу с прямым подключением.

Внутренний коммутатор реализован на базе сдвоенных избыточных с "горячей" заменой 6,1 Тбит/с модулей.

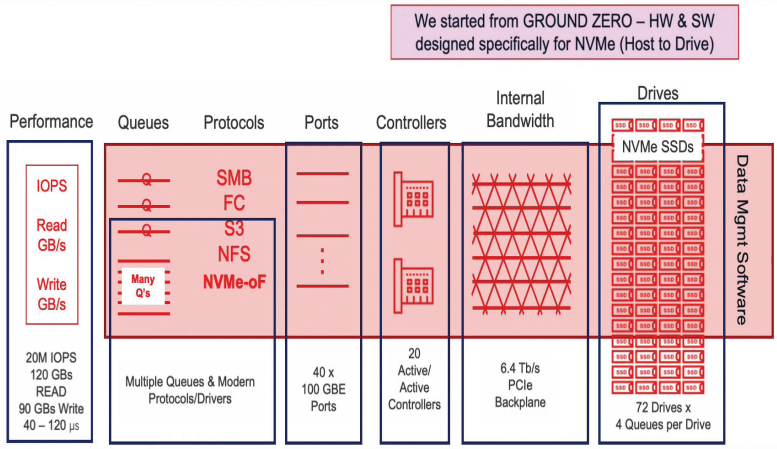

Рис. 1. Аппаратное и программное обеспечение Pavilion HFA разработаны "с нуля" с учетом всех преимуществ решений NVMe.

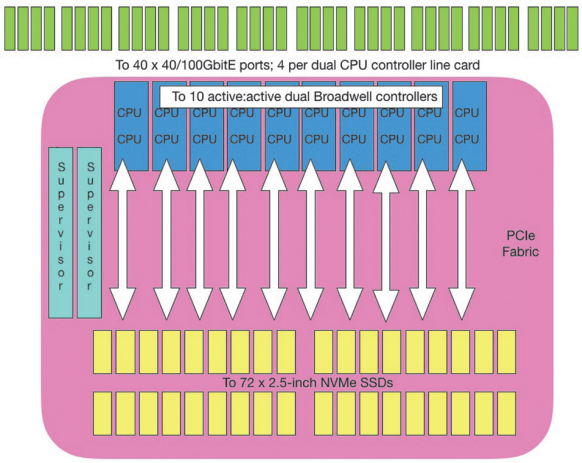

Рис. 3. Структурная схема Pavilion HFA.

Один HFA обеспечивает пропускную способность 120 ГБ/с, задержку 25 мкс и 2,2 ПБ хранилища в форм-факторе 4U. На уровне стойки он поддерживает емкость 20,2 ПБ, скорость чтения 1,2 ТБ/с, пропускную способность записи 900 ГБ/с и 200 млн IOPS при той же задержке 25 мкс.

Pavilion заявляет, что производительность Pavilion HFA в несколько раз больше производительности по чтению и записи для данных NFS и S3 по сравнению с существующими разработчиками СХД и превышает производительность СХД ряда новых игроков (Qumulo, StorOne, WekaIO и VAST Data) в этой области.

Массивы Pavilion поддерживают iSCSI, NVMe-oF/TCP, NVMe-oF/RoCE, NFS и S3. Они также могут связываться с внешней файловой системой, такой как IBM Spectrum Scale, Lustre или BeeGFS. Один HFA поддерживает до 20 контроллеров и 40 портов Ethernet или InfiniBand 100 Гбит/с.

Компания Pavilion сообщила, что ее протестированная производительность позволяет линейно масштабировать до восьми стоек флэш-памяти и до 160 контроллеров [2]. Одна из отличительных особенностей Pavilion HFA: высокая плотность производительности – до 5 млн IOPS/U.

Pavilion HyperOS 3.0

Pavilion HyperOS 3.0 распараллеливает вычислительные ресурсы, флэш-память NVMe и сеть хранения. Pavilion HFA использует собственный протокол TCP для кластеризации нескольких контроллеров в глобальном пространстве имен для блочного, файлового и объектного доступа. Пространство имен можно кластеризовать по нескольким шасси и стойкам [2].

Pavilion HyperOS на 100% соответствует стандартам. Она использует стандартные сетевые интерфейсы Ethernet/InfiniBand, стандартные твердотельные накопители NVMe, стандартные стеки программного обеспечения хоста и многое другое. Pavilion не требует установки специального программного обеспечения на хостах приложений или сетевых коммутаторах. Pavilion HyperOS поддерживает стандартные 40Gb, 100Gb или 200Gb адаптеры Ethernet/InfiniBand для NVMe/Ethernet (RoCE v2), NVMe/TCP, NVMe/InfiniBand (RDMA), любой совместимый с v1.x драйвер NVMe-Over-Fabrics. iSCSI, NFS v3/v4, pNFS и S3 любые или все эти протоколы и фабрики могут работать в одной системе одновременно [3].

Pavilion HyperOS обеспечивает гибкое подключение клиентов к ряду клиентов для максимальной простоты использования и подключения. Клиентские плагины включают NFS v3/v4, pNFS, NFS RDMA, собственный клиент Gluster, Hadoop и Apache Spark.

Сервисы данных корпоративного уровня

Все сервисы хранения, API-интерфейсы и расширенные сервисы данных встроены и включены в каждый массив. Услуги корпоративных данных, такие как многоуровневое хранение, репликация, моментальные снимки, клонирование, безопасность, шифрование, тонкое выделение ресурсов и проактивная поддержка, включены в каждый экземпляр Pavilion HyperOS 3.0.

Сервисы данных S3 включают в себя единое или глобальное пространство имен, средства управления доступом, многоуровневое хранение, репликацию, моментальные снимки, безопасность, шифрование, плагины приложений и многое другое.

Интеграции для S3 включают наиболее популярные решения для поставщиков удостоверений, мониторинга и оповещения, целей notificaiton, федерации, оркестровки, балансировки нагрузки и резервного копирования. Интеграции машинного обучения и больших данных включают Apache Spark, Hadoop, TensorFlow, kafka, splunk, Greenplum, Teradata и другие.

Масштабирование

Pavilion HyperOS обеспечивает линейное, независимое масштабирование производительности между контроллерами и массивами. Масштабирование до 10 систем на стойку для достижения производительности 1,2 Тбайт/с на стойку или 200 млн IOPS на стойку.

Для NFS-нагрузок на одну систему поддерживается скорость чтения до 70 Гбайт/с и запись 50 Гбайт/с, для рабочих нагрузок S3 – до 60 Гбайт/с при чтении и 40 Гбайт/с при записи. При этом количестве систем может масштабироваться неограниченно, а емкость или производительность меняться линейно и независимо (рис. 4).

Табл. 1. Сервисы данных, поддерживаемые Pavilion HyperOS™ 3.0.

Поддержка блочного, файлового и объектного доступа	Многопротокольность • Приложениям необходим доступ к большому количеству данных, но большая их часть находится в разрозненных хранилищах. Унифицируйте разрозненные хранилища для совместной работы с блочными, файловыми и объектными данными. Одновременно используйте NVMe / RDMA, NVMe / TCP, iSCSI, NFS и S3 для всех ваших приложений. • Получайте все преимущества от производительности благодаря встроенной поддержке любого поддерживаемого протокола на любой комбинации контроллеров в неограниченном количестве массивов для полного выбора и управления.
Обеспечение безопасности и надежности данных	Распределенный RAID с двойной четностью • Полностью заполненный массив объединяет твердотельные накопители NVMe в 4 группы по 18 дисков. Pavilion реализует распределенный RAID 6 внутри группы, что обеспечивает защиту 16 + 2 RAID 6. Накладные расходы для RAID 6 менее 12%. При желании можно определить «горячий» резерв, что приведет к защите 15 + 2 + 1 RAID 6. Автоматическая перестройка «SWARM» • В случае отказа диска несколько контроллеров устанавливают заменяемый диск параллельно, чтобы обеспечить быстрое восстановление. Диск емкостью 2 ТБ восстанавливается менее чем за 5 минут. С помощью Pavilion HyperParallel Flash Array твердотельные накопители приложений полностью перестраиваются с защитой RAID 6 в 12 раз быстрее, чем при использовании DAS или массива all-flash. Снимок / Клоны • Можно сделать несколько логических устойчивых к сбоям копий любого тома и по отдельности передать их различным приложениям, например процессу резервного копирования. Эти копии являются согласованными, компактными, мгновенными и доступными для записи. Создается мгновенно, без необходимости физического копирования данных; блоки затем записываются как копии, или оригинал изменяется с течением времени. Непрерывные операции с многопутевостью • Достижение необходимого уровня доступности и надежности для современных массово-параллельных приложений требует больших затрат. Недостаточно загруженная и неэффективная емкость серверов, а также операционные издержки на управление этими изолированными серверами являются постоянной проблемой для ИТ-отдела. Нет единой точки отказа, и ввод-вывод может выполняться несколькими контроллерами хранилища, что увеличивает доступность данных. Шифрование данных в состоянии покоя • Безопасность является обязательной, поэтому Pavilion реализует шифрование хранимых данных в соответствии с требованиями FIPS и гарантирует, что постоянное шифрование не влияет на производительность. Обновление программного обеспечения без прерывания работы • Соблюдение требований соответствия - фундаментальная часть проектирования системы. Все обновления Pavilion HyperOS можно применять без прерывания текущих операций. CRC и управление версиями • Pavilion HyperOS включает в себя дифференциальную проверку избыточности T10 для защиты от битового гниения. Основываясь на этом, Pavilion также использует запатентованную технологию для включения номера версии для каждой записи для защиты от неудачной записи на носитель. Это защищает организации от редких, но случайных и значительных проблем, которые становятся все более распространенными по мере увеличения размера дисков.
Контроль роста объемов данных и контроль затрат	Тонкое выделение ресурсов (Thin Provisioning) • Тома представляют собой логические диски NVMe и используются как обычное блочное устройство. Том предоставляется с тонкой подготовкой из группы носителей. Пользователь получает предоставленный размер тома, но пространство выделяется только по мере необходимости для максимального использования. Объемы могут варьироваться от менее одного ТБ до 100 ТБ и могут быть переназначены или разделены между системами по мере необходимости. • Управляйте своими данными из единой системы для полного контроля с упреждающим мониторингом со стороны Pavilion Data
Экономичность и гибкость	Линейная масштабируемость • Гиперпараллельный флеш-массив Pavilion поддерживает требования нескольких приложений благодаря своей облачной масштабируемости. Когда емкость приложения необходимо увеличить, DevOps может добавить твердотельный накопитель NVMe без снижения производительности. И наоборот, когда приложению требуется более высокая производительность ввода-вывода, можно добавить дополнительные контроллеры хранения, не влияя на емкость. Multi-Fabric • Гиперпараллельный флеш-массив Pavilion поддерживает различные порты физической структуры NVMe-oF, устраняя ненужные трансляции протоколов и обеспечивая семантику NVMe для малой задержки и высокого IOPS в различных топологиях, включая NVMe / Ethernet, NVMe / TCP, NFS, iSCSI, S3 и NVMe / InfiniBand. Интеграция • Гиперпараллельный флэш-массив Pavilion позволяет вам управлять массивом по своему усмотрению. Любите ли вы использовать полнофункциональный веб-интерфейс и интерфейс командной строки, которые обеспечивают глубокое понимание показателей производительности на уровне тома / приложения, контроллера, порта или системы. Вы также можете использовать его API на основе REST для интеграции с DTMF / Redfish и Swordfish или использовать плагин Pavilion HyperParallel Flash Array OpenStack и vCenter, или интегрироваться с Kubernetes с его плагином FlexVolume и поддержкой CSI для постоянных томов (PV). Проактивная поддержка • Pavilion HyperParallel Flash Array сообщает о проблемах на облачном портале поддержки Pavilion для анализа компанией Pavilion. Портал поддержки позволяет Pavilion проактивно реагировать на любые проблемы, выполняя операции по их устранению удаленно.

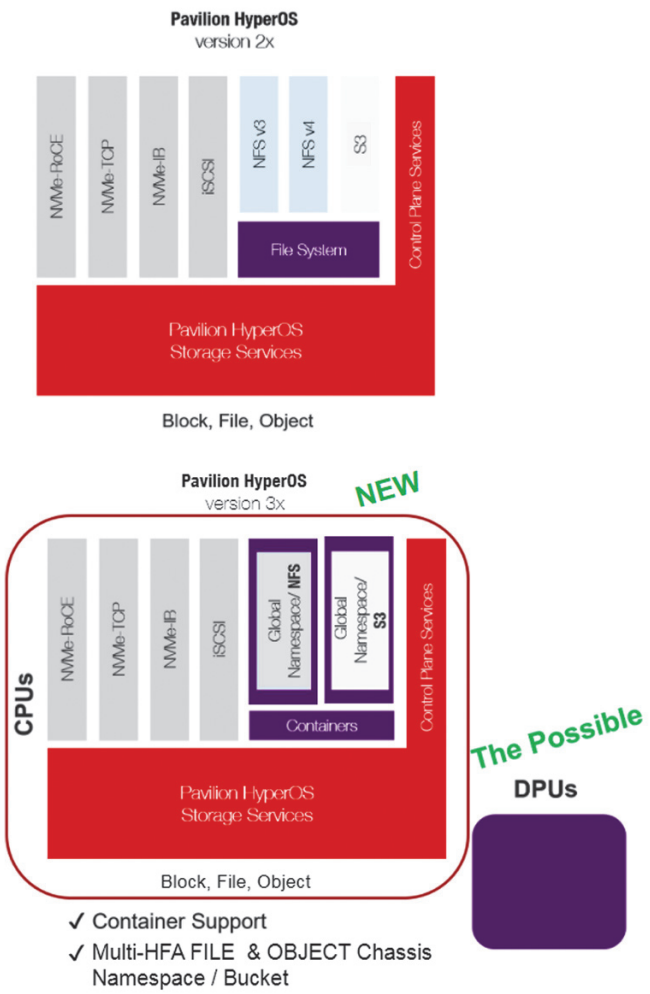

Рис. 4. Отличие версии 3.x от 2.x Pavilion HyperOS.

Повышенное качество обслуживания

Четыре уровня поддержки QoS позволяют организациям управлять доступом к данным так, что наиболее важные приложения всегда могли получать нужные данные. Элементы управления на основе ограничения гарантируют, что данные всегда будут доступны тем, кто в них нуждается.

Полный список сервисов, поддерживаемых Pavilion HyperOS 3.0™, представлен в табл. 1.

Тестирование Pavilion HFA с NVIDIA DGX-A100 с GPUDirect

В конце января 2021 г. Pavilion Data представила (*https://blocksandfiles.com/2021/01/26/pavilion-data-gpudirect-performance/*) результаты тестирования массива Pavilion HFA с NVIDIA DGX-A100 при использовании GPUDirect (прямая передача данных от GPU к внешнему хранилищу без CPU), подтвержденные NVIDIA.

Как показали результаты измерений (рис. 5), производительность Pavilion HFA <-> NVIDIA DGX-A100 (*https://pavilion.io/company/partner-resources/, страница партнерских ресурсов Pavilion – «Nvidia Pavilion Summary of GPU Direct Testing»*) при блочном чтении (с Pavilion HFA) составила 182 ГБ/с и 149 ГБ/с при записи. Скорость передачи файловых данных составляла 191 ГБ/с при чтении и 118 ГБ/с при записи. Это наилучший результат среди других протестированных систем Nvidia DGX-A100 при использовании GPUDirect (на 10% быстрее, чем DDN и VAST Data, и на 26% быстрее, чем WekaIO):

- Pavilion – 191 ГБ/с;
- DDN – Lustre – 173,9 ГБ/с;
- VAST Data – NFS – 173,9 ГБ/с;
- WekaIO – Matrix – 152 ГБ/с.

Pavilion HFA и GPFS

Spectrum Scale Storage

Spectrum Scale (также известных как GPFS, General Parallel File System) – очень популярная высокопроизводительная кластерная файловая система на рынке высокопроизводительных кластерных (HPC) вычислений и суперкомпьютеров. Она разработана, лицензирована и поддерживается IBM.

Spectrum Scale обеспечивает одновременный высокоскоростной доступ к файлам для приложений, работающих на нескольких узлах кластера. Он работает с кластерами AIX 5L, кластерами Linux, Microsoft Windows Server или гетерогенными кластерами AIX, Linux и Windows.

Spectrum Scale дополнительно предоставляет инструменты для управления и администрирования кластеров Spectrum Scale, обеспечивая при этом общий доступ к файловым системам с удаленных кластеров Spectrum Scale. Ее можно развернуть в трех вариантах (режимах):

- в качестве распределенной параллельной системы без совместного использования ресурсов (SNDPM, shared-nothing distributed parallel modes);
- как хранилище с общим диском (SD, shared-disk);
- в качестве сетевых дисков с общим доступом (NSD, network-shared-disk).

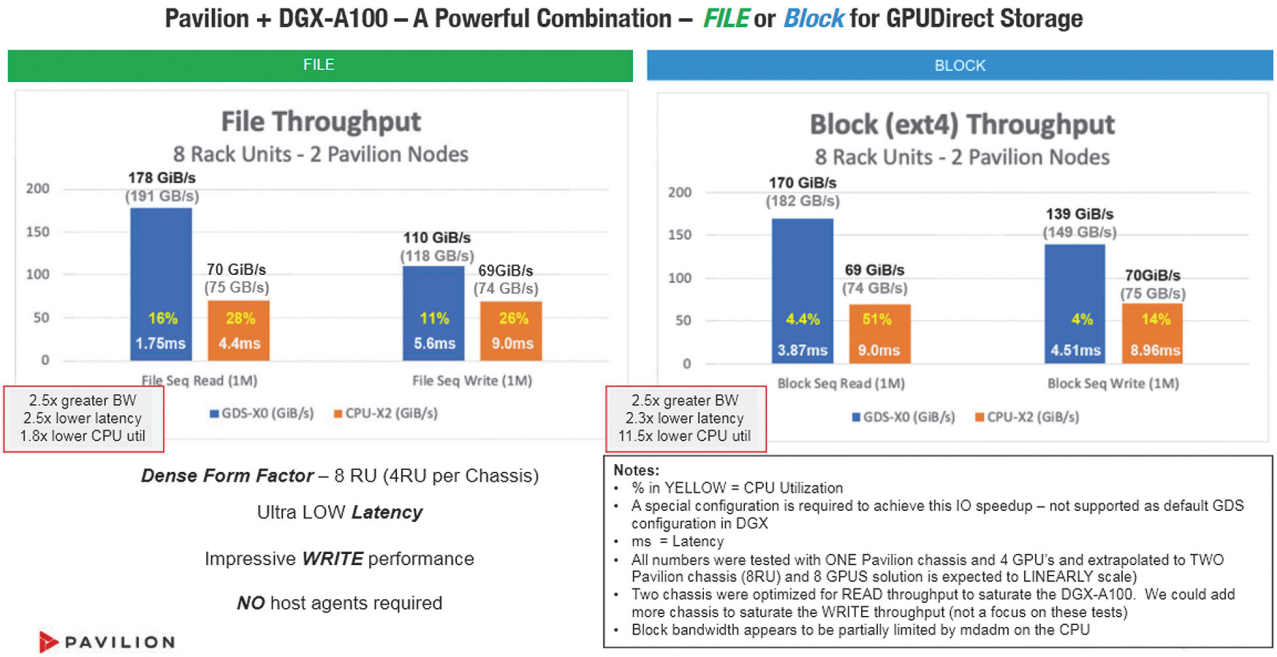

Рис. 5. Результаты тестирования Pavilion HFA с NVIDIA DGX-A100 при использовании GPUDirect.

Распределенный параллельный режим без общего доступа (SNDPM)

Распределенный параллельный режим без общего доступа – это термин Spectrum Scale для DAS или хранилища с прямым подключением. DAS использует внутреннее хранилище в каждом узле кластера Spectrum Scale. Этим носителем могут быть диски класса памяти (SCM, storage class memory), флэш-накопители NVMe, твердотельные накопители (SSD) SATA/SAS или жесткие диски (HDD). В кластере Spectrum Scale каждый узел является частью кластерной параллельной файловой системы. Каждый узел сервера Spectrum Scale в кластере имеет доступ к своим данным и данным каждого другого узла.

Реализация DAS Spectrum Scale на первый взгляд привлекательна тем, что не требует внешних систем хранения. Spectrum Scale хранит данные в виде стандартных блоков и предоставляет несколько встроенных функций защиты данных, включая Spectrum Scale RAID (аналогично RAID системы хранения), который виртуализирует тома для избыточности и параллельного доступа. Последние выпуски Spectrum Scale включают Erasure Coding со сквозной контрольной суммой по сети. Хотя он улучшает устойчивость данных, он оказывает заметное пагубное влияние на производительность как чтения, так и записи. Он также обеспечивает асинхронную репликацию файлов между томами для обеспечения доступа к данным даже в случае сбоя узла и множественные репликации для защиты от сбоев нескольких узлов, которые потребляют слишком много ресурсов.

Распределенные параллельные режимы без общего доступа исторически были самой популярной реализацией Spectrum Scale. Это относительно легко реализовать. Нет необходимости покупать или внедрять внешние системы хранения. Диски хранения физически ближе к вычислению, обеспечивая меньшую скорость небольшой задержки в узле; хотя гораздо более высокие задержки для доступа к данным на других узлах. Задержка влияет на чтение, запись и транзакции первого байта. Стоимость диска для серверов явно ниже: средняя цена составляет 1/3 от стоимости такого же диска в общей системе хранения после всех скидок.

Проблемы с распределенными параллельными режимами без общего доступа (SNDPM)

Непостоянная производительность – главная проблема для SNDPM. Пока все данные локализованы на узле, осуществляющем доступ к данным, производительность остается хорошей. Но как только требуемые данные находятся в узле, отличном от узла запрашивающего, производительность падает. Непостоянная производительность убивает продуктивность и приносит неудобства пользователям.

Кластеры Spectrum Scale по определению распределяют данные по узлам файловой системы. Каждый серверный узел Spectrum Scale имеет ограниченное количество доступных слотов хранения. Тип носителя может варьироваться в зависимости от этого ограничения. Накопители памяти класса хранения NVMe (SCM) или твердотельные флэш-накопители (SSD) имеют меньше слотов, чем флэш-накопители SATA и жесткие диски (HDD). Ограничения слотов ограничивают емкость на кластер и на узел. Это, в свою очередь, затрудняет локализацию всех запрошенных данных на запрашивающем узле. Spectrum Scale защищает данные на нескольких узлах от сбоев узлов и/или дисков с помощью Spectrum Scale RAID. Поскольку он виртуализирует тома на нескольких узлах, производительность ограничивается диском с самой низкой производительностью в этом виртуальном томе, что еще больше снижает производительность. Еще один редко признаваемый аспект SNDPM – это количество ресурсов, потребляемых (циклы, память и ввод-вывод) сервисами хранения Spectrum Scale, такими как множественный RAID. Есть причина, по которой в общей системе хранения есть выделенные контроллеры. Восстановление RAID невероятно ресурсоемко, занимает много времени, заметно снижает производительность, делая ее крайне непоследовательной.

Последняя версия Spectrum Scale также включает Erasure Coding, которое IBM называет Erasure Coding Edition (ECE). ECE – это декластеризованный RAID. Он помещает большие номера дисков на несколько серверных узлов Spectrum Scale в одну группу. Плюс ECE в том, что он делает данные очень устойчивыми. А поскольку четность распределяется между множеством дисков, восстановление выполняется относительно быстро по сравнению с традиционным RAID. Плохая новость в том, что ECE снижает общую производительность. Производительность любого отдельного тома зависит от скорости самого медленного диска в этом томе. Кроме того, ECE распределяется по большему количеству дисков в большем количестве узлов для защиты от сбоев одного или нескольких серверных узлов.

Возможные способы снижения влияния этих проблем следующие:

- максимально возможная локализация данных вручную. Локализация – нетривиальная трудоемкая задача, выполняемая вручную, и серьезно усложняет защиту данных RAID или ECE;
- использование межсоединений с очень высокой пропускной способностью, таких как 100 Гбит/с Ethernet или Infiniband, и удаленный прямой доступ к памяти (RDMA) для минимизации задержек и увеличения пропускной способности между узлами.

Все эти сложности обычно увеличивают количество узлов кластера Spectrum Scale, значительно превышающее то, что требуется для обработки приложений. Это, в свою очередь, увеличивает пространство в стойке, напольные плитки, сетевые карты, кабели, приемопередатчики, коммутаторы, питание, охлаждение, операции, управление, устранение неполадок, оперативные исправления, исправления, обновления, обслуживание, обучение персонала, время администратора и многое другое. Все это увеличивает совокупную стоимость владения (TCO).

Сложное управление – еще одна существенная проблема SNDPM. Настройка, инициализация, защита и управление томами не предназначены для простого использования в распределенных параллельных режимах без общего доступа, и это не так. Это не для новичков в области хранения данных, которым требуются знания, навыки, обучение и, желательно, опыт работы с физическими серверами, сетями, хранилищами и Spectrum Scale. Даже замена неисправного диска может быть обременительной. Часто требуется отключение серверного узла Spectrum Scale и открытие корпуса. Серверные узлы не предназначены для горячей замены их накопителей в оперативном режиме без нарушения работы сервера. Это означает, что необходимо запланировать замену привода в нерабочее время, что усложняет простейшее обслуживание.

Масштабирование емкости или вычислительной мощности – это проблема SNDPM из-за большой зернистости (coarse grain). Крупнозернистость означает сложность добавления вычислительных ресурсов или емкости хранения без добавления других компонент. Да, можно добавить диски к конкретному узлу Spectrum Scale. Однако каждый серверный узел имеет серьезные ограничения на количество дисков, которые фактически могут быть добавлены. У них как низкая плотность емкости хранилища, так и низкая плотность производительности хранилища. Обычно это означает добавление дополнительных серверных узлов Spectrum Scale просто для увеличения емкости, даже когда дополнительная обработка не требуется. Низкая плотность приводит к увеличенному пространству в стойке, межкомпонентным соединениям, коммутаторам, портам коммутатора, кабелям, приемопередатчикам, кабелепроводу, питанию, охлаждению, обслуживанию и т.д. Все это приводит к непредвиденным расходам. Или это означает переключение хранилища на режимы сетевого диска (NSD) или общего диска (SD).

Когда требуется дополнительная обработка, добавляются дополнительные узлы Spectrum Scale с дополнительной емкостью, даже если эта емкость не требуется. Это увеличивает потребление ресурсов и усложняет кластер, а также увеличивает совокупную стоимость владения.

TCO представляет собой неожиданную проблему для SNDPM, потому что она, как правило, значительно выше, чем это общепринято. Согласно общепринятому мнению, это должна быть самая низкая стоимость, потому что она основана на стандартном (COTS, commodity-off-the-shelf) серверном оборудовании. Это оборудование является массово производимым. Накопители в серверах дешевле систем хранения на 67%. Но следует иметь в виду, что общепринятое мнение основано на ценообразовании, а не на совокупной стоимости владения. TCO – это гораздо больше, чем просто предварительная цена. Причина, по которой общие затраты, как правило, намного больше, чем ожидалось, заключается в чрезвычайно низкой плотности производительности хранилища и плотности емкости хранилища.

Обычно упускаемые из виду затраты включают:

- *увеличение SKU (Stock Keeping Unit, складских единиц)*. Производители серверных узлов COTS меняют конфигурации или артикулы серверов не реже одного раза в год, а часто и несколько раз в год. Конфигурация серверного узла для Spectrum Scale, купленная в прошлом году, с такой же вероятностью не будет существовать, когда потребуется другой. Новые серверные узлы

Spectrum Scale всегда будут иметь разные конфигурации с новейшими процессорами, памятью, SCM, твердотельными накопителями NVMe, твердотельными накопителями, жесткими дисками, микрокодом и т.д. Многие SKU со временем могут создать кошмар. Количество комбинаций, запасных частей, сложность управления, неудобства при поиске и устранении неисправностей часто являются непосильными и дорогостоящими;

– *неожиданное потребление ресурсов сервером защиты данных*. Spectrum Scale SNDPM "крадет" циклы ЦП, память и ввод-вывод с каждого серверного узла для RAID или ECE. Моментальные снимки также сильно нагружают ЦП и память, забирая ресурсы сервера у приложений, использующих Spectrum Scale. Это приводит к тому, что для кластера требуется больше серверных узлов. Эту "боль" усугубляет то, что каждый узел добавляет меньшую производительность, чем предыдущий узел, так как он усложняет кластер.

SNDPM, это путь наименьшего сопротивления для большинства пользователей Spectrum Scale, особенно для начинающих пользователей. Похоже, что это самая простая и легкая реализация системы хранения Spectrum Scale. Реальность часто доказывает, что это сложнее и дороже.

Режим общего диска

В режиме общего диска (SD) Spectrum Scale имеет файловую систему хранилища SAN, подключенное к каждому узлу сервера Spectrum Scale через сетевой коммутатор хранения. Преимущество SD заключается в централизации данных, совместно используемых всеми Spectrum Scale серверными узлами Enterprise Content Managers (ECM) в кластере. Каждый ECM запускает собственное чтение и запись ввода-вывода непосредственно в хранилище SAN без посредников. Управляющая информация передается по сети одновременно, поддерживая согласованность данных.

Использование сетевых карт или адаптеров NVMe-oF в каждом серверном узле Spectrum Scale и хранилище SAN обеспечивает чрезвычайно низкие задержки для хранилища SAN с NVMe-oF. Фактически, эта низкая задержка NVMeoF позволяет настроить серверные узлы Spectrum Scale с ECM, чтобы они были полностью бездисковыми и загружались с хранилища SAN. Это позволяет создавать очень маленькие недорогие серверные узлы Spectrum Scale без внутренней памяти.

Еще одно преимущество SD – возможность мелкозернистой масштабируемости – вычислительные ресурсы и хранилище SAN можно масштабировать отдельно друг от друга.

Проблемы режима с общим диском (SD)

SD для более крупных кластеров Spectrum Scale может становиться все более сложным с уменьшением предельной отдачи.

Недостаточная производительность для удовлетворения потребностей менеджеров корпоративного контента (ECM, Enterprise Content Managers) в кластере HPC Spectrum Scale. Большинство высокопроизводительных систем хранения SAN предназначены для различных одновременных рабочих нагрузок, которые обычно встречаются на уровне предприятия или среднего уровня. Приложения Spectrum Scale HPC обычно намного более требовательны к производительности.

Кроме того, существует проблема узкого места контроллера хранилища SAN. Каждый контроллер хранилища SAN ограничен своими процессорами Intel x86 Xeon или IBM POWER. Основным препятствием для производительности сегодня является ЦП. Следует помнить, что процессоры в этих системах не являются результатом удвоения производительности каждые два года, как предлагал закон Мура на протяжении последних четырех десятилетий. Закон Мура на самом деле не закон. Это было наблюдение, которое показало, что количество транзисторов удваивается каждые 18 месяцев, а позже – каждые два года. Теперь, когда квантовая физика продемонстрировала ограничения на это наблюдение, это удвоение значительно уменьшилось.

Взаимодействие между узлами сервера Spectrum Scale и хранилищем SD SAN часто бывает неадекватным. Многие системы хранения SAN используют межсоединения с меньшей пропускной способностью, такие как оптоволоконный канал (FC) или FICON, которые не очень распространены в средах Spectrum Scale. Или они предоставляют слишком мало портов со слишком малой пропускной способностью, например, Ethernet 10 Гбит/с или меньше, превращая межсоединение в серьезную узкую точку. Некоторые из них не могут использовать технологию RDMA с низкой задержкой NVMe over Fabric (NVMe-oF), замедляющую транзакции и производительность первого байта до неприемлемых уровней.

Низкая плотность производительности увеличивает затраты на SD. Большинство систем хранения SAN не предназначены для обеспечения высокой плотности производительности. Как правило, они предназначены для обеспечения отличной производительности, не обязательно высокой плотности.

Network Shared Disk (NSD)

Spectrum Scale NSD – это блочного уровня интерфейсный протокол, предназначенный для работы в сети. NSD – это, по сути, архитектура клиент-сервер. Сервер NSD предоставляет доступ к хранилищу данных, к которому не подключены клиенты NSD. Хороший способ концептуально представить себе NSD – это то же самое, что и приложения или пользователи, обращающиеся к файловому хранилищу. NSD – это не файловое хранилище. Это блочное хранилище по сети TCP/IP, которое "чувствует" как файловое хранилище настроено. Сервер NSD предоставляет клиентам NSD свое внутреннее или внешнее хранилище. Информация управления Spectrum Scale сначала проходит по сети TCP/IP, а затем фактическое чтение или запись данных клиента NSD в подключенное хранилище сервера NSD выполняется серверами NSD.

Конфигурации NSD обычно имеют только несколько серверов NSD, расположенных перед подключенным хранилищем SAN.

Конфигурация Spectrum Scale NSD имеет несколько преимуществ. Одно из преимуществ – более простое управление хранилищем. Хранилище, подключенное к серверу NSD, может быть внутренним или внешним. В большинстве случаев это внешнее хранилище SAN. Хранилище SAN предназначено для упрощения оперативного обслуживания, установки исправлений и обновления без прерывистой перезагрузки системы. Диски поддерживают горячую замену.

Еще одним преимуществом является размер клиент-серверных узлов Spectrum Scale NSD. Они намного меньше с номинальными требованиями к внутренней памяти в первую очередь для загрузки. Это сокращает пространство в стойке и эксплуатационные расходы (обслуживание, питание, охлаждение, накладные расходы и т.д.).

Масштабирование вычислений и хранилища NSD похоже на SD в том, что оно является мелкозернистым. Оба могут масштабироваться отдельно друг от друга.

Проблемы с общим сетевым диском (NSD)

Недостаточная производительность. Сервер NSD находится перед внутренним или внешним блочным хранилищем. Этот сервер увеличивает задержку для внутреннего блочного хранилища. Задержки складывается. Существует задержка от клиента NSD по сети TCP/IP до сервера NSD. Затем существует задержка от сервера NSD к контроллеру хранилища SAN, которая может иметь, а может и не иметь задержку сетевого коммутатора хранения SAN. Кроме того, существует задержка между контроллером хранилища SAN и носителем в хранилище SAN. Эта задержка существует в обоих направлениях. RDMA между сервером NSD и контроллером хранилища SAN для NVMe-oF снижает задержку.

Также существует задержка при использовании внутренней программно-определяемой памяти (SDS) сервера NSD Spectrum Scale. Следует помнить, что сервисы хранения Spectrum Scale требуют ЦП, памяти и ввода-вывода точно так же, как программное обеспечение для обслуживания хранилищ в системе хранения SAN. Серверы NSD имеют двойную производительность по сравнению с SD. Производительность ограничена как на сервере NSD, так и в подключенном хранилище SAN из-за тех же проблем с процессором.

Одна из неизбежных причин снижения производительности сервера NSD заключается в том, что он является узким местом для подключенного хранилища SAN или внутреннего хранилища DAS для сервера NSD. Все запросы на чтение и запись проходят через сервер NSD от нескольких клиентов NSD. Высокая активность чтения и/или записи снизит производительность приложения, поскольку сервер NSD обрабатывает эти запросы. Самый распространенный обходной путь – добавить больше серверов NSD и увеличить расходы. Однако этот путь лишь несколько снижает проблему производительности добавляя сложность балансировки нагрузки.

Проблема с пропускной способностью сети: связь между клиентами NSD Spectrum Scale и сервером NSD. Большинство из них имеют пропускную способность 10 Гбит/с Ethernet или меньше. Даже те, у которых соединение с более высокой пропускной способностью

ограничено количеством сетевых портов, которые сервер NSD может обрабатывать одновременно – обычно 4 или меньше.

Низкая плотность производительность увеличивает затраты на NSD. Как правило, хранилища SAN имеют низкую плотность производительности, что значительно увеличивает затраты на NSD.

Решение Pavilion HFA

Как Pavilion HFA решает проблемы Spectrum Scale

Хранилище Pavilion HFA спроектировано так, чтобы исключить описанные выше проблемы со Spectrum Scale, за счет:

– интеграции до 20 контроллеров с двойной резервной структурой PCIe с несколькими ТБ/с в одной системе, что устраняет проблему узких мест, связанных с производительностью ЦП;

– предоставления до 40 высокоскоростных портов Ethernet или Infiniband 40/100 Гбит/с на систему, что устраняет узкие места в сети и использует наиболее распространенные соединения серверных узлов Spectrum Scale. Таким образом, устраняется необходимость в новом незнакомом оборудовании, навыках, опыте или обучении при обеспечении требуемой пропускной способности с меньшим количеством портов на узел Spectrum Scale. 40/100 Гбит/с Ethernet и Infiniband могут использоваться в системе одновременно (не на одних и тех же портах);

– использования NVMe-oF по этому межсоединению с высокой пропускной способностью он обеспечивает наименьшую возможную задержку в архитектуре SD или NSD. Использование отраслевых стандартов NVMe / TCP и NVMe / RoCE v2 через Ethernet или RDMA через InfiniBand означает, что новые навыки не требуются;

– объединения всего вышеперечисленного в корпусе 4 RU плюс (до 72 флэш-накопителей NVMe с возможностью горячей замены), что обеспечивает как очень высокую плотность производительности хранения, так и требуемую плотность емкости хранения.

Каждая система хранения Pavilion HFA подключается к узлам сервера Spectrum Scale напрямую или через коммутатор сети хранения. Каждый серверный узел Spectrum Scale видит все совместно используемое хранилище, как если бы это был внутренний накопитель NVMe с прямым подключением, что обеспечивает гораздо большую плотность кластера. Серверные узлы Spectrum Scale могут использовать очень маленький загрузочный диск внутри, помещая все свои данные ECM в Pavilion HFA. Это позволяет узлам сервера Spectrum Scale использовать 1 RU, микросерверы или блейд-серверы для минимального использования места в стойке, питания, охлаждения, накладных расходов, обслуживания и т.д., что дополнительно снижает затраты.

Pavilion HFA отличается от любой другой системы хранения SD тем, что позволяет клиентам Spectrum Scale использовать стандартные серверные флэш-накопители NVMe в HFA. Другими словами, клиентам HFA не нужно покупать диски в Pavilion (хотя они могут это сделать, если захотят). Они могут купить их у своего поставщика SSD и использовать в HFA. Поскольку серверные флэш-накопители NVMe обычно стоят на 67% меньше, чем их эквивалент, в системе хранения данных, что позволяет сэкономить много денег. Это однозначно обеспечивает заказчику Spectrum Scale более низкую стоимость SNDPM в управляемой системе SD.

Кластеры Spectrum Scale являются либо аналитическими и ориентированными на пропускную способность, либо транзакционными и ориентированными на ввод-вывод. Именно здесь Pavilion HFA не имеет себе равных. Он не требует компромиссов между пропускной способностью и IOPS с устойчивой пропускной способностью до 120 Гбит/с и до 20 млн IOPS. Когда требуется больше, просто добавьте еще 4 RU HFA. Эта непревзойденная производительность хранилища и плотность производительности хранилища позволяют ECM обрабатывать и анализировать больше данных за меньшее время, ускоряя получение результатов и увеличивая прибыль.

Pavilion HFA обеспечивает высоку степень защиты данных с помощью RAID 6 (плюс горячий резерв при желании), поддержку T.10 DIF (Data Integrity Field – подход к защите целостности данных в компьютерном хранилище данных от повреждения, *https://en.wikipedia.org/wiki/Data_Integrity_Field*), обнаружение повреждения данных, шифрование с аппаратным ускорением и практически мгновенную экономию места в моментальных снимках и клонах. И это оказывает минимальное влияние на производительность. Благодаря тому, что количество контроллеров хранилища, в основном, в несколько раз больше, чем у любой другой системы хранения, эта дополнительная мощность означает, что ресурсов ЦП более чем достаточно для моментальных снимков, восстановления поврежденных данных и восстановления RAID без заметного влияния на производительность. При восстановлении RAID-массива Pavilion HFA используется сверхбыстрое адаптивное восстановление «роя» с помощью мультиконтроллера, которое может восстанавливать твердотельные накопители со скоростью до 5 минут на каждый ТБ. Сокращение времени восстановления еще больше снижает влияние на производительность и риск потери данных.

Использование защиты данных Pavilion HFA избавляет Spectrum Scale от необходимости делать это. Это экономит циклы Spectrum Scale для приложений, что приводит к уменьшению количества серверных узлов Spectrum Scale, необходимых для удовлетворения требований обработки, что еще больше снижает затраты. Это также может уменьшить или устранить необходимость в хранении дополнительных реплик в целях обеспечения доступности.

Одна из самых больших затрат на систему хранения SD – это диски. Компания Pavilion позволяет своим клиентам приобретать собственные флэш-накопители NVMe у поставщиков серверов или использовать ранее приобретенные серверные накопители HPC. Это не только позволяет избежать привязки к поставщику, но и снижает стоимость накопителей на две трети и более.

Дополнительная экономия средств достигается за счет увеличения производительности, емкости и плотности вычислений. Большая плотность означает меньшее количество систем, стоек, мощности, охлаждения и выделенных накладных расходов. Это сокращение может быть таким же значительным, как и экономия от своих собственных серверных дисков. 100%-ые NVMe флеш-накопители сами по себе обычно сокращают операционные расходы. Это связано с тем, что у накопителей нет движущихся частей, они служат дольше и не так хрупки, как жесткие диски, а также потребляют гораздо меньше энергии и охлаждения. Тонкое выделение ресурсов также снижает требования к емкости, что еще больше снижает затраты.

Вся эта экономия средств приводит к очень разумной совокупной стоимости владения, конкурирующей с Spectrum Scale SNDPM, но с преимуществом централизованного управления, большей простотой и более высокой производительностью.

Прямое сравнение с ведущми поставщиками, предлагающими СХД для Spectrum Scale, показывает, что Pavilion HFA имеет лучшую производительность хранилища, плотность производительности хранилища, плотность межсоединений и плотность вычислений на рынке [4].

Источники, доп. ресурсы

1. Pavilion HyperOS 3.0 Launches, Powering the Most Performant, Dense, Scalable, and Flexible Storage Platform in the Universe – *https://pavilion.io/press/pavilion-hyperos-3-0-launches-powering-the-most-performant-dense-scalable-and-flexible-storage-platform-in-the-universe/*
KIOXIA_Introducing_SEF_Technology_White_Paper.pdf.
2. Dense processing fuels Pavilion Data storage performance, Garry Kranz, Senior News Writer – *https://searchstorage.techtarget.com/news/252492592/Dense-processing-fuels-Pavilion-Data-storage-performance*.
3. Pavilion HyperOS™ 3.0, Data Sheet – *https://pavilion.io/app/uploads/2019/11/Pavilion-HyperOS-3-v4.pdf*.
4. Unprecedented Spectrum Scale Performance. How Pavilion Optimizes GPFS Deployments, Marc Staimer, President & CDS Dragon Slayer Consulting – *https://www.hpcinnovation.com/downloads/pavilion-data-hyperparallel-spectrum-scale-whitepaper.pdf*.