Преимущества использования SCM-кэша в составе внешних СХД HPE

Обзор особенностей и преимуществ использования PCIe-карт Intel Optane в качестве SCM-кэша в составе HPE 3PAR и Nimble.

Алексей Казьмин – менеджер по продуктам, отдел гибридных ИТ, HPE в России.

HPE давно занимается развитием новых типов хранилищ данных и оптимизацией доступа к хранилищам с целью ускорить работу приложений своих заказчиков. Пока совершенно новая архитектура вычислительных систем The Machine (т.н. памяте-центричная архитектура) еще находится в процессе создания, но мы понимаем, что ускоряться нужно уже сейчас. Давайте поговорим о том, что можно сделать сегодня и что появится у HPE завтра.

В первую очередь, речь идет о сильном ускорении наших СХД 3PAR и Nimble с помощью умного и относительно бюджетного кэширования на Storage Class Memory (SCM) в форме Intel Optane.

С появлением на рынке продуктов на базе Intel Optane применение SCM для корпоративных задач, несомненно, начнется в промышленных масштабах повсеместно. Существуют два варианта подобных применений на производствах:

– точечные, требующие размещения новых SCM в серверах в непосредственной близости к процессорам (с подключением по шине DDR или PCIe), с побайтной адресацией этого нового слоя хранения данных;

– широкого назначения, когда приложения могут получать выгоду от SCM в блочном режиме работы SCM как от кэширующего или основного слоя хранения – внутри серверов или во внешних СХД (с подключением по шине PCIe).

Первый вариант зачастую требует значительной доработки кода приложений для получения желаемого прироста производительности, в то время как второй вариант позволяет получать прирост проще – с обновлением текущих СХД 3PAR и Nimble (в случае HPE) или приобретением в будущем новых (в случае ряда других вендоров) СХД с поддержкой SCM.

Говоря о первом варианте, имеющем характерную для него потребность во внутрисерверном хранилище с низким временем отклика и большим ресурсом, то еще до появления Intel Optane «в железе» HPE совместно с Intel, Microsoft, Red Hat и другими партнерами проводила не менее важную работу, чем создание нового типа твердотельных накопителей. Так, в частности, специалисты поработали над программной экосистемой. На базе собственных продуктов HPE Persistent Memory и Scalable Persistent Memory компания тестировала и проверяла под реальными нагрузками у заказчиков способность понимать новый тип энергонезависимой памяти у доработанных ОС Microsoft и Linux, а также у ключевых корпоративных приложений. Сегодня HPE Persistent Memory нативно понимается (*https://access.redhat.com/documentation/en-us/red_hat_enterprise_linux/7/html/7.3_release_notes/new_features_storage*) ядром Linux и Windows Server 2012R2 и 2016. Для MS SQL Server 2016 еще в 2016 году был получен (*https://h20195.www2.hpe.com/V2/getpdf.aspx/A00006919ENW.pdf*) прирост производительности с помощью «эмуляции» SCM. Также кратный прирост был получен (*https://habr.com/company/hpe/blog/282395*) для PostgreSQL.

Эти наработки – ключ к ускоренному включению Intel Optane в портфель серверов HPE и инфраструктуры наших заказчиков.

Вернемся ко второму, простому и быстрому, варианту использования SCM. Есть несколько способов повысить производительность виртуализованного приложения, работающего с данными на внешней СХД:

– посмотреть, что «сдерживает» приложение сейчас. Возможно, дело совсем не в СХД, а в ожидании процессора, во внутренней логике работы с данными или в не оптимально написанных запросах;

– если большие задержки со стороны ожидания данных (IO), то сначала стоит проверить соблюдены ли все рекомендации по настройке связки «приложение–ОС–драйверы» (SCSI, HBA и т.п.);

– решить возможную проблему в сети SAN (Ethernet, FC);

– также можно поискать неполадки в СХД: в железе контроллера (состояние кэша, степень загрузки процессора) или в ОС контроллера и драйверах, в шине данных или в дисках.

Психологически комфортнее задуматься, в первую очередь, о производительности дисков в СХД. Все остальные стадии «жизненного пути» данных в процессе чтения-записи довольно сложны, и разобраться в них стоит труда. Часто специалисты, использующие гибридное хранилище, думают, что решением проблем производительности может стать all-flash. Но что делать, если all-flash уже внедрен? Тогда в поле внимания попадает реклама некоторых известных брендов СХД, рассказывающая о преимуществах хранилища с «NVMe-дисками».

Как правило, стоят такие решения дорого, и к тому же требуется ещё и купить новую СХД вместо апгрейда текущей. Но необходимость вынуждает.

Однако есть и другие варианты, которые нужно рассматривать, и вот почему. Большинство NVMe SSD на рынке в настоящее время — это тот же самый тип носителя, NAND-flash, только подключенный к контроллеру не по протоколу Serial Attached SCSI (SAS), а по новому протоколу NVMe, который в действительности только начинает “взрослую жизнь”. Вот некоторые факты о нём:

– доступно 64 000 очередей с 64 000 потоков каждая — IOPS предостаточно;

– контроллер находится прямо в CPU — таким образом, нагрузка на процессор становится ниже;

– каждое ядро процессора «видит» каждый SSD напрямую, поэтому задержки на доступ низкие.

При полной замене SCSI-протокола на всем пути от приложения до дисков возможно значительно снизить задержку доступа. Но, как правило, производители на рынке предлагают нам «NVMe-диски», т.е. вся цепочка до самого контроллера СХД остается та же — SCSI. Затем контроллер просто перепаковывает SCSI в NVMe и так взаимодействует с подключенными NAND SSD.

Результат от использования такого решения виден на графике (рис. 1). Как можно заметить, выигрыш в задержке минимальный, хотя по пиковым IOPS он, действительно, может быть очень заметен. Здесь можно привести традиционную аналогию: вам нужна машина, которая может быстро разгоняться для обгона за 5 секунд, или машина, которая в идеальных условиях может за 10 минут разогнаться до 300 км/ч? Оба варианта хороши, но чаще выбирают первый.

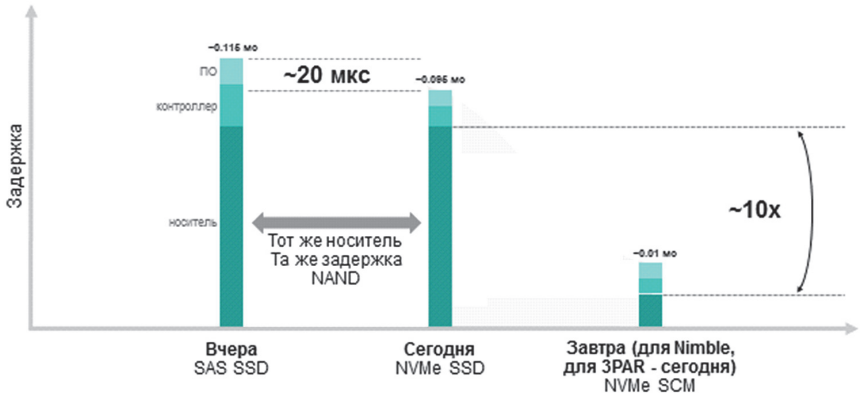

Рис. 1. Сравнение задержек для различных типов флэш-накопителей с подключением по SAS- и NVMe-интерфейсу.

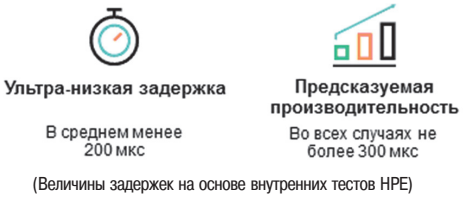

Рис. 2. Преимущества использования накопителей Intel Optane в качестве кэша на чтение в контроллерах СХД 3PAR и Nimble.

Реальное положение вещей сегодня таково, что прирост от NVMe NAND сегодня мало заметен для реальных приложений и, на наш взгляд, совсем не стоит той разницы в цене и проигрыше в доступной емкости по сравнению с SAS SSD.

Вместо простой замены «последней мили» с SAS на NVMe HPE предлагает использовать подключенные по NVMe совершенно новые накопители Intel Optane в качестве кэша на чтение в контроллерах СХД 3PAR и Nimble.

Вот несколько причин, почему HPE решила пойти по этому пути:

– возможность предложить заказчикам обновление уже закупленных СХД (конкретно 3PAR 9450, 20450, 20850 и Nimble AF60 и AF80 — все топовые all-flash);

– возможное снижение максимальной задержки примерно в несколько раз, и средней – на 30–40% (IOPS также растут) очень простым способом (добавлением карты расширения с Optane на борту в каждый контроллер). К тому же, задержка не будет скакать от декларируемых некоторыми производителями «NVMe-СХД» «от 0,2 мс» до бесконечности, а будет становится гораздо стабильнее (рис. 2);

– от такого снижения задержки на массиве, например, для Oracle можно ожидать сокращения IO wait в среднем на 37% и ускорения выполнения SQL SELECT-ов на 27% (согласно внутренним тестам HPE);

– кэш используется на чтение, а не на запись, потому что и в 3PAR, и в Nimble уже многие годы в качестве кэша на запись используется оперативная память DRAM (в случае Nimble — энергонезависимая NVRAM). Она, в свою очередь, в разы быстрее NVMe-устройств, и до появления Gen-Z или аналогичных новых протоколов будет оставаться таковой. Таким образом, запись ускорять через NVMe не нужно.

– Intel Optane — это новейший тип носителя, отстающий от NAND по плотности, но на порядок быстрее по отклику. Также Optane обладает практически неисчерпаемым ресурсом на перезапись. Для нагруженных систем стоимость транзакции на Optane гораздо ниже, чем на NAND NVMe. А кэш — это слой очень нагруженный со всех сторон. В него копируются с более медленного слоя «горячие» данные (поэтому для этого нужен ресурс), с него идет чтение, если данные не найдены в RAM-кэше контроллера (поэтому нужен быстрый отклик, чтобы выход за пределы RAM-кэша не становился долгим и утомительным);

– SCM-кэш уже сейчас дает заметный прирост производительности при относительно небольшом бюджете на апгрейд текущих СХД. Его можно и нужно использовать — пока NVMe NAND еще есть, куда дешеветь, сам протокол NVMe еще развивается (multi-pathing появился в стандарте только в марте 2018 года, и еще сильно отстает от стабильности SCSI), и в целом экосистема NVMe от приложения до дисков еще не развита (NVMe over fabric только только что получил первую версию стандарта, производители все еще спорят, как это должно выглядеть и развиваться, драйверы обладают минимальным функционалом).

Небольшая иллюстрация причин выбора роли кэша для SCM в HPE Nimble – на рис. 3.

А с помощью инструмента HPE InfoSight (*https://habr.com/company/hpe/blog/329202*) вы всегда будете знать, где искать задержку (рис. 4).

Подводя итог, можно сказать: являясь обладателем 3PAR 9000 или 20000, то можете заказать 3PAR 3D Cache на базе Intel Optane прямо сейчас. Если же вы собираетесь приобретаеть массив Nimble All-flash, то это надежная база для защиты инвестиций в будущем. Начать можно с SAS NAND SSD сейчас, подключить All Flash Turbo-кэш на базе SCM позже, а затем поменять диски на NVMe.

Алексей Казьмин,
HPE в России

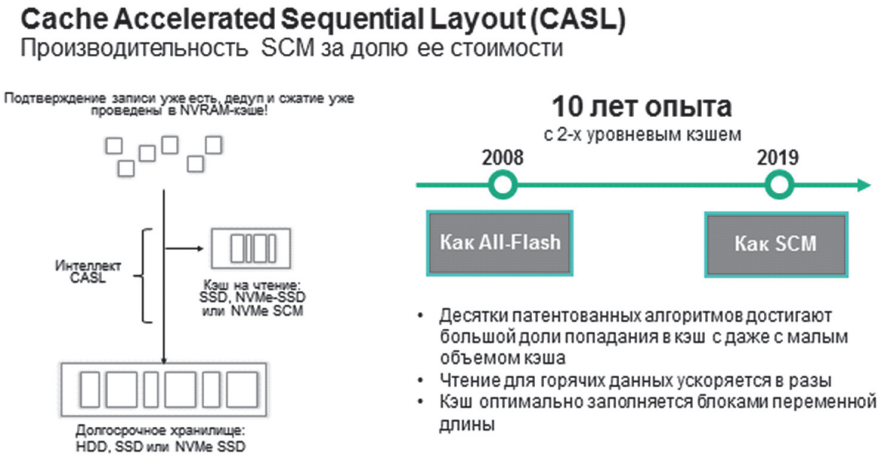

Рис. 3. Причины выбора SCM в качестве кэша для HPE Nimble.

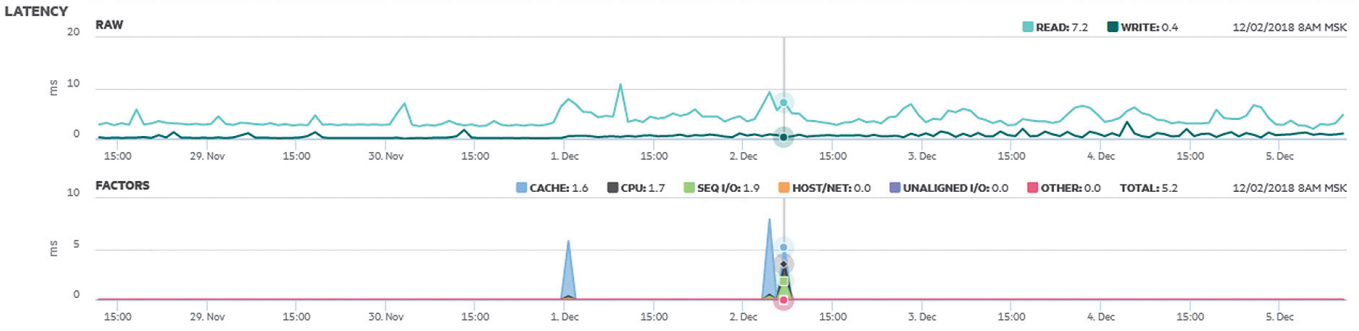

Рис. 4. Визуализация задержек с помощью HPE InfoSight.