Флеш-СХД: все познается в сравнении

Сравнение по производительности, а также по ряду других параметров all-flash массивов от трех вендоров.

Владимир Колганов – руководитель направления систем хранения данных компании КРОК.

Введение

Флеш-память — это, пожалуй, наиболее оптимальный носитель с точки зрения производительности хранения данных. Уже сегодня аналитики рынка утверждают, что в ближайшие несколько лет флеш-накопители вытеснят с рынка высокопроизводительные жесткие диски, при этом основные корпоративные данные будут храниться на флеш, а жесткие диски SATA большой емкости останутся для хранения архивно-справочных данных и резервных копий.

Рост популярности и интереса к флеш-памяти обусловлены возрастающими требованиями корпоративных задач к производительности приложений и минимизации времени отклика от них, и в первую очередь это базы данных, аналитические системы, серверная виртуализация, VDI. Флеш-системы позволяют удовлетворить требования транзакционных баз по количеству операций ввода-вывода в секунду, в несколько раз сократить время построения аналитических отчетов. Носители на флеш-памяти также имеют значительно более низкое время отклика (0,5 мс против порядка 5 мс у дисковых систем) даже под нагрузкой, что положительно сказывается на качестве и эффективности работы вычислительных систем.

В сравнении с традиционными СХД флеш-хранилища кажутся довольно дорогим удовольствием, однако это справедливо, если рассматривать только стоимость хранения за терабайт (на сегодняшний день этот параметр — не «конек» флеш-накопителей), но если исходить из стоимости производительности (например, IOPS/$) становится понятно, что выгоднее флеш-СХД сегодня рынок не может предложить ничего другого. Кроме того, флэш-накопители позволяют достичь максимальной плотности производительности и энергоэффективности. Например, производительность одного SSD-диска сравнима с производительностью дисковой полки! При этом экономия на ресурсах ЦОДа может составлять до 100000$ на одной такой системе в год.

Технология флеш сегодня находится на достаточном уровне зрелости для использования в системах хранения корпоративного класса, и многие производители предлагают на базе нее свои решения. Необходимо также отметить, что рынок флеш-решений, конечно же, не ограничен тремя игроками, но проведенное тестирование на синтетических тестах может быть полезно с точки зрения получения дополнительной информации, хотя оно и не дает полного представления о характеристиках производительности флеш-массивов с учетом всех их особенностей в условиях реальной нагрузки.

Huawei Dorado — «крупная рыба»

Компания Huawei давно известна на рынке, особенно в секторе телекоммуникационного оборудования. Сегодня Huawei держит курс на развитие своего портфеля в области ИТ-оборудования и работы в корпоративном сегменте. Линейка систем хранения данных Huawei носит название OceanStor, а СХД на флеш-памяти называются Dorado. Линейка массивов Dorado состоит из двух моделей: 5100 и 2100 G2. Компания Huawei участвует в тестах SPC (Storage Performance Council), которые публикуются открыто: результат массива 2100 G2 составил 400 000 IOPs (SPC-1). СХД Dorado позиционируется как простое решение для высоконагруженных задач, оно может быть особенно интересно в свете текущих внешнеполитических рисков.

Рис. 1. Схема подключения Huawei OceanStor Dorado 2100 G2 (20 дисков по 200 Гбайт в одной дисковой полке) к нагрузочному серверу IBM Power 770 четырьмя путями по 8 Гбит/с.

Мы в своей лаборатории протестировали Huawei OceanStor Dorado 2100 G2 с двадцатью пятью дисками по 200 Гбайт в одной дисковой полке. Система хранения была подключена напрямую к нагрузочному серверу IBM Power 770 четырьмя путями по 8 Гбит/с (рис. 1).

Для симуляции нагрузки использовалась программа FIO, которая хорошо подходит для тестирования систем хранения под Linux ОС. Генерировалась нагрузка в 100 процессов случайного чтения, затем один процесс начинал выполнять операции 100% случайной записи блоками по 4 Кбайт, после чего к нему подключались по одному дополнительные процессы (рис. 2).

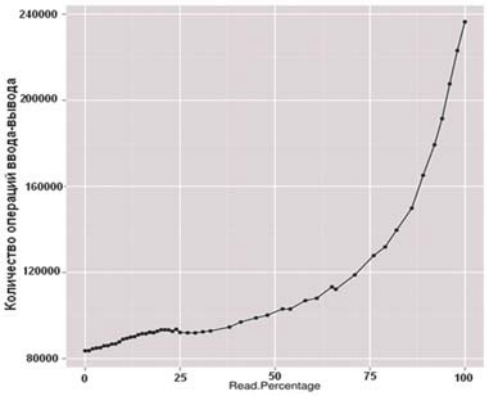

Рис. 2. Производительность Huawei OceanStor Dorado 2100 G2 при изменении процентного соотношения операций чтения и записи.

Зная максимальные значения и нужные нам соотношения читающих и пишущих процессов, мы стали варьировать нагрузку на систему хранения. В итоге получили весь диапазон производительности системы хранения данных на разной нагрузке (рис. 3).

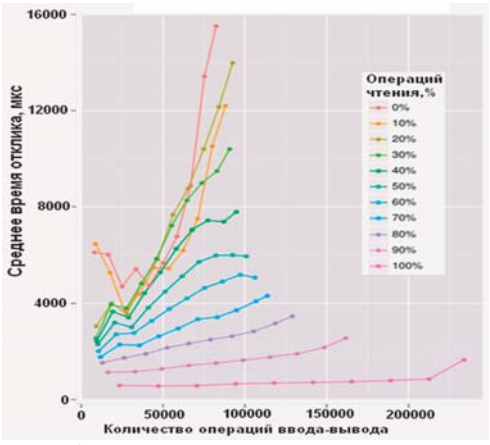

Рис. 3. Изменение времени отклика Huawei OceanStor Dorado 2100 G2 в зависимости от производительности и соотношения операций чтения и записи.

«Экстремальные» флеш-системы EMC

В отличие от рассмотренной выше системы Huawei, система EMC ExtremIO использует в своей архитектуре стандартные компоненты и не имеет никаких проприетарных контроллеров или флеш-карт. Все «железные» компоненты стандартны: процессоры x86, накопители eMLC от ведущих производителей привычного форм-фактора 2,5”объемом 400 или 800 Гбайт. Такой выбор позволяет обеспечить преемственность поколений и использовать в будущем новые процессоры и флеш-диски. Разработчики компании при этом сосредотачиваются на разработке ПО и функционала.

Помимо того, что системы EMC ExtremIO работают без единой точки отказа, имеют active-active контроллеры и собственные ИБП, одной из ее ключевых особенностей является Scale-out архитектура, то есть при масштабировании наращивается и емкость, и производительность.

Система может поставляться в конфигурации от 1 до 6 узлов, т.н. Xbrick, которые между собой связаны по шине Infiniband. В каждом из кластеров хранятся только уникальные данные системы: дедупликация работает на глобальном уровне. Набор стандартных компонентов Xbrick включает в себя два контроллера x86

с двумя восьмиядерными процессорами и 250 Гбайт памяти для обработки I/O и сервисов, соединенных с 24-дисковой полкой. Бесперебойная работа обеспечивается резервным питанием от двух одноюнитовых ИБП. Один Xbrick обеспечивает производительность от 100 000 до 250 000 уникальных операций ввода вывода в секунду в зависимости от профиля нагрузки.

В своем центре решений на базе технологий EMC мы тестировали однобриковый массив XtremIO с помощью сервера IBMx5 3850 с 64-мя ядрами, большим количеством памяти и четырьмя однопортовыми HBA-адаптерами. Сервер подключался напрямую к массиву, чтобы предотвратить возможное влияние Fibre Channel коммутаторов на результаты теста.

Для тестирования код массива был обновлен до последней версии XtremIO 3.0 (3.0.0-44). В отличие от 2-й версии этот код не только дедуплицирует, но и сжимает данные. В результате все сводилось к тому, что для определения «честной» производительности СХД нам нужно было писать на массив недедуплицируемые и несжимаемые данные. Для этого был выбран хорошо известный IOmeter (его последняя версия). На массиве мы создали 4 луна по одному терабайту. Нагрузка на луны увеличивалась постепенно, а каждый следующий тест требовал больше производительности от массива.

Пришло время установить и настроить операционную систему. Большинство наших заказчиков используют различные ОС семейства Windows, это и определило наш выбор. Для тестирования на хост была установлена Windows Server 2012 R2 с последним набором обновлений. Настройки ОС по умолчанию не могут в полной мере раскрыть производительность флеш-СХД, поэтому потребовалась оптимизация. Была увеличена глубина очереди на HBA-адаптерах, настроен нативный мультипасинг, на тестируемых дисках отключен кэш ОС, а на лунах СХД созданы разделы с блоком 4 Кбайт.

Было проведено 3 теста (случайные операции в/в, блоками 4 Кбайт): 1) 100% записи; 2) 100% чтения; 3) 70% чтение, 30% запись.

Снимались следующие показатели: количество операций ввода/вывода и время отклика массива. В этом тесте мы не измеряли пропускную способность массива, так как для достижения ее максимального значения нужно использовать блоки большего размера. Все показатели были сняты средством мониторинга массива (табл. 1).

Таким образом, на текущий момент XTremIO является единственным массивом, который умеет глобально на лету дедуплицировать данные. Коэффициент дедупликации всегда разный, но наибольших значений он достигает на задачах по развертыванию VDI.

Табл. 1. Результаты тестирования EMC XtremIO.

100% чтение		100% запись		70% чтение, 30% запись	
Кол-во операций в/в	Время отклика EMC XtremIO, мкс	Кол-во операций в/в	Время отклика EMC XtremIO, мкс	Кол-во операций в/в	Время отклика EMC XtremIO, мкс
60000	550	30000	900	40000	650
105000	600	45000	1100	70000	800
140000	700	55000	1600	85000	950
180000	950	60000	1900	100000	1100
195000	1100	65000	2100	110000	1300
210000	1350	70000	2500	115000	1400
215000	1500	72000	2900	120000	1600
220000	1750	74000	3300	125000	1800
225000	2150	75000	3700	130000	2000
230000	2300	76000	4000	135000	2200

Гибридное решение Hitachi

Часто разработанные с нуля флеш-массивы обладают ограниченным функционалом, а купленные сторонние решения на флеш не всегда гармонично работают с системами мониторинга и производительности в сложившейся линейке. Компания Hitachi, развивая свои флеш-СХД, пошла другим путем — она изменила собственную прошивку массивов старшего уровня, сделав из нее специализированную прошивку под флеш, оптимизированную под работу с задержками не более 1 миллисекунды. При этом в данной прошивке унаследован весь функционал традиционных систем: снэпшоты, репликация, тонкое выделение ресурсов, полная интеграция с виртуальными массивами. Для флеш Hitachi предлагает использовать массив HUS VM. Если его производительности не хватает, можно использовать старшие системы VSPG1000/VSP. Компания Hitachi также участвует в тестах SPC*).

Заявленная производителем производительность одного флеш-накопителя от Hitachi составляет 100 000 IOPs при задержке 0,2 мс, в продуктивных системах такие решения показывают в среднем 20 000 — 25 000 IOPs.

У нас в этом году был открыт центр решений Hitachi, и тестирование HUS VM мы проводили именно в нем. Стенд состоял из двух блэйд-серверов ComputeBlade 520XB1, которые четырьмя путями были подключены к двум FC-свитчам. В них же был подключен тестируемый нами HUS VM восемью путями, по 4 на каждый свитч.

В системе хранения использовалось 33-FMD модуля емкостью 1,6 Тбайт каждый, 32 из которых были собраны в RAID-группы по схеме RAID-5 (3D+1P). Один модуль был оставлен нами в качестве модуля горячей замены. Емкость мы разбили на устройства емкостью 2,5 Тбайт, из которых выделялись устройства (LDEV) по 1 Тбайт, которые были отданы серверам. Суммарно на сервер выделялось по 4 устройства (рис. 4).

Тестирование СХД проводилось программой IOmeter, часто использующейся для проведения синтетических тестов на системах хранения. Была применена версия программы, позволяющая генерировать полностью случайные данные, чтобы поставить систему хранения в наиболее сложные условия. На каждый жесткий диск, который виделся в операционной системе Windows 2008 R2, было назначено по 2 инициатора нагрузки (Worker).

Тестирование состояло из трех групп тестов на случайных операциях ввода/вывода (блоками по 4 Кбайт): 1) 100% чтение; 2) 70% чтение, 30% запись; 3) 100% запись.

Инициаторы нагрузки увеличивали ее ступенчато: вдвое на каждом интервале тестирования от 1 операции ввода-вывода до 128 операций на каждом инициаторе.

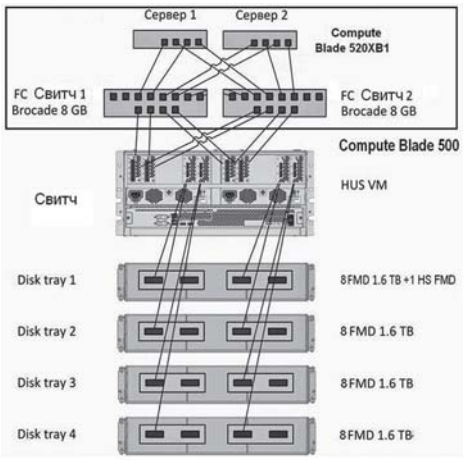

Рис. 4. Схема соединения блэйд-серверов ComputeBlade 520XB1 с массивом HUS VM.

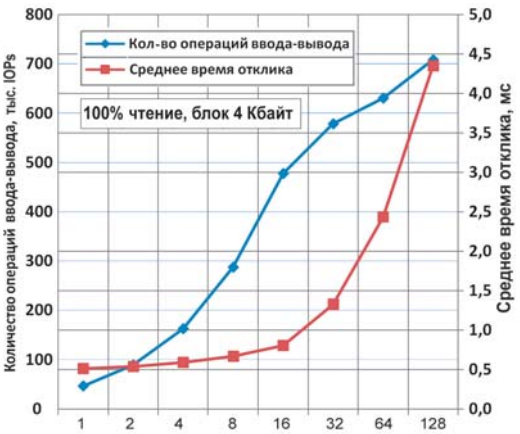

Рис. 5. Производительность и среднее отклика HUS VM в зависимости от числа операций в/в на каждом инициаторе (x2) – 100% чтение, блок 4 Кбайт.

Рис. 6. Производительность и среднее отклика HUS VM в зависимости от числа операций в/в на каждом инициаторе (x2) – 70% чтение, 30% запись, блок 4 Кбайт.

Рис. 6. Производительность и среднее отклика HUS VM в зависимости от числа операций в/в на каждом инициаторе (x2) – 100% запись, блок 4 Кбайт.

*) Результаты тестирования Hitachi VSP с FMD: *http://www.storageperformance.org/benchmark_results_files/SPC-1/HDS/A00136_Hitachi-VSP-Flash/a00136_Hitachi_VSP-HAF_SPC-1_executive-summary_revision-1.pdf* и *http://www.storageperformance.org/benchmark_results_files/SPC-1/HDS/A00136_Hitachi-VSP-Flash/a00136_Hitachi_VSP-HAF_SPC-1_full-disclosure-report_revision-1.pdf*

Результаты тестирования Hitachi VSP G1000 с FMD / HNAS: *https://www.spec.org/sfs2008/results/res2014q2/sfs2008-20140331-00245.html.*

IBM: новые системы на базе технологий OpenPOWER для больших данных

Октябрь 2014 г. – Компания IBM представила новую линейку вычислительных систем, способных справиться с огромными массивами данных. Новые системы превышают показатели «стоимость/технические характеристики» серверов общего назначения на 20% *(показатели производительности основаны на предварительной оценке опубликованных результатов SPECcpu2006 – SPECfp_rate2006 – от 26 сент. 2014 г., http://www.specbench.org/).*

Новые серверы IBM Power S824L созданы на базе процессора IBM POWER8 – первого в мире чипа, оптимизированного для наиболее требовательных нагрузок в сфере обработки больших данных. В новые системы надежно интегрированы технологии IBM и других членов консорциума OpenPOWER, включая GPU-ускоритель NVIDIA, с целью предоставить очень высокие вычислительные мощности на базе параллельной обработки для того, чтобы, например: 1) банки могли лучше оценивать риски, 2) энергетические компании более точно определять нефтяные месторождения, 3) ученые более оперативно подбирать правильные методы лечения пациентов.

Процессор POWER8 – первый в своем роде процессор с дифференцированными техническими характеристиками, созданный для работы как со структурированными, так и неструктурированными данными. Список данных характеристик включает архитектуру CAPI (интерфейс коллективного ускорителя) с такими возможностями, как: 1) CAPI Flash Access Efficiency, 2) сокращение объема хранилища с помощью CAPI Attached Compression Accelerator, 3) ускорение обработки информации и сокращение задержек с помощью CAPI Attached Mellanox RDMA Fabric.

Системы Power S824L предоставляют возможность запускать задачи с интенсивным обменом данными на процессоре POWER8, при этом снижая нагрузки на другие вычислительные задачи, связанные с обработкой больших данных с использованием GPU-ускорителей NVIDIA.

В рамках дальнейшей адаптации GPU-ускорителей под Power Systems компания IBM планирует оптимизировать все портфолио корпоративных приложений для работы с большими данными, включая базы данных IBM DB2 с поддержкой BLU Acceleration. Более того, IBM также работает над оптимизацией архитектуры Power под GPU-ускорение для приложений, использующихся в биоинформатике, финансовом и оборонном секторах, молекулярной динамике, моделировании погоды – в том числе SOAP3, NAMD, GROMACS, библиотеках FFTW и Quantum Espresso.

Следующие поколения IBM Power Systems будут поддерживать технологию NVIDIA NVLink, которая обеспечивает скоростной обмен данными между центральным и графическим процессорами с помощью интерфейса PCI Express. Это обеспечит GPU-ускорителям Nvidia доступ к памяти процессоров IBM POWER, а также позволит увеличить производительность многочисленных корпоративных приложений. Подобные системы будут доступны для заказа в начале 2016 г.

SAP Fraud Management for Banking на российском рынке

Ноябрь 2014 г. – Компания SAP СНГ объявила о выводе решения SAP Fraud Management for Banking на российский рынок. Решение помогает банковским и финансовым организациям предотвратить потери от злонамеренных действий, сократить расходы на расследование подобных случаев и ложных сигналов за счет мгновенного анализа больших данных. Инструмент содержит преднастроенные сценарии выявления злоупотреблений, связанных с противодействием легализации доходов, полученных преступным путем, и финансированию терроризма (ANTI-MONEY LAUNDERING). Скорость работы инструмента была протестирована одним из ведущих банков Европы и показала увеличение скорости работы в 1200 раз среди 20 млн счетов клиентов и 1 млрд транзакций.

В России проблема с безопасным использованием данных стоит особенно остро, в частности, в вопросах, касающихся внутрикорпоративного использования данных. По данным исследования юридической компании Vegas Lex, свыше трети российских компаний в 2013 г. выявили у себя признаки корпоративного мошенничества (38%), при этом в отдельных случаях ущерб составил до миллиарда долларов.

Точкой отсчета, по которой измеряется производительность флеш-массивов, считается производительность системы при среднем времени отклика в 1 мс. В нашем тестировании на 100% случайном чтении достигнут результат чуть больше 500 тыс. IOPs (рис. 5).

Во втором тестировании (70% чтение, 30% запись) получен результат в почти 150 тыс. IOPs (рис. 6).

В последнем тестировании производилась нагрузка системы хранения операциями 100% случайной записи (рис. 7). Система хранения почти сразу же вышла на пик своей мощности, что очень неплохо. При этом она показала производительность в 60 тыс. IOPs.

Заключение

Сегодня флеш-технология апробирована, остается лишь выбрать конкретную систему, наиболее подходящую по своим техническим характеристикам для текущих задач. А что это будет – доступные и импортозамещающие системы Huawei, устраняющие дубли СХД EMC или особое по архитектуре решение Hitachi – остается на усмотрение самого заказчика.

Однако стоит заметить, что за скобками проведенного исследования остались многие особенности флеш-массивов, которые могут оказать существенное влияние как на производительность системы, так и на окончательный ее выбор. В частности, следует учитывать следующее:

- *производительность может падать по мере заполнения массива;*
- *производительность массивов на случайных операциях ввода/вывода зависит не только от соотношения чтение/запись, но и от длительности записи. И в тех случаях, где это время достаточно велико, на ряде массивов производительность может падать;*
- *коэффициенты дедупликации и сжатия могут существенно варьироваться от 1 до 30 в зависимости от нагрузки и, соответственно, для тех решений, где есть встроенная дедупликация/сжатие, производительность может существенно возрастать при его увеличении;*
- *производительность может зависеть от размера блока. Например, на флеш-модулях Hitachi блок размером 128 Кбайт может «на лету» сжиматься до 8 Кбайт за счет встроенной компрессии, что выгодно будет отличать это решение по производительности от других, где эта технология отсутствует;*
- *когда на флеш записываются шифрованные данные, использование технологий сжатия и дедупликации должно быть очень взвешенным, чтобы не нарушить целостность данных;*
- *производительность на единицу занимаемой площади/объема/единицу хранимой информации может в разы отличаться у разных решений;*
- *объем хранимых данных на единицу занимаемой площади/объема может в разы отличаться у разных решений;*
- *большое значение на производительность в лучшую сторону может оказать наличие встроенной функциональности на уровне флеш;*
- *флеш-технологии (на уровне хранения) существенно отличаются по уровню надежности. Например, специализированные флеш-модули Hitachi, по заявлениям производителя, поддерживают уровень надежности 99,9999%. Это достигается за счет: 1) алгоритма коррекции ошибок (ECC), способного исправить до 48 бит на 1,4 Кбит; 2) перепроверки всех записанных данных каждые 2 дня; 3) автоматического обновления всех записанных данных один раз в 30 дней; 4) механизма автоматического контроля износа ячеек памяти.*

  При отсутствии встроенных механизмов поддержания надежности, это приходиться решать (в случае необходимости) за счет, например, избыточности компонент;
- *если предполагается использовать флеш в более высокоуровневых системах, например, для тайринга или/и поддержания катастрофоустойчивости, то оценку производительности необходимо производить с учетом этих факторов.*

Число особенностей флеш, влияющих на выбор, можно расширять, поэтому в полной мере повлиять на выбор системы может только ее тестирование на реальных нагрузках с учетом всех других требований.

Владимир Колганов,
компания КРОК.