hp StorageWorks Enterprise Virtual Array – прогнозирование реальной производительности без тестирования?...

Состав параметров, по которым происходит выбор дискового массива, достаточно широк, но одним из основных был и остается – его производительность. Попытка ответа на вопрос – "можно ли провести оценку пропускной способности модульного массива в условиях реальной нагрузки без сложного тестирования (ориентируясь на показатели производителя) и почему это возможно для hp EVA?" – данная публикация.

Введение

Вопрос оценки производительности дисковых массивов среди поставщиков, особенно корпоративного класса, остается "камнем преткновения" уже в течение многих лет. И хотя этот показатель при решении вопроса о закупке учитывается вкупе с другими (стоимость, надежность, совместимость, масштабируемость и др.) его детерминируемость и прогнозируемость имеют во многих случаях решающее значение. Основная проблема здесь состоит в оценке результатов на реальной нагрузке в сравнении с заявляемой производителем, которая в отличие от максимальной должна учитывать многочисленные факторы, оказывающие влияние на ее возможное снижение и прежде всего связанные с разбалансировкой нагрузки массива.

В реальных условиях эксплуатации, безусловно, можно "разгрузить" особенно "горячие" диски, но при постоянно меняющейся динамической многопрофильной нагрузке делать это крайне сложно. И даже, если в статике (в течение достаточно длительного периода времени) нагрузку удается выровнять, в динамике (в более короткие промежутки времени) она остается различной для разных групп дисков. И во многих случаях это обуславливается заданными условиями эксплуатации сетевого хранилища, менять которые по многим причинам еще сложнее. По сути оценка коэффициента недогрузки массива в реальных условиях нагрузки и есть основная задача предпроектного тестирования на производительность.

На сегодняшний день – hp StorageWorks Enterprise Virtual Array v.2 – один из немногих поставляемых модульных массивов корпоративного класса, для которого задача прогнозирования производительности решаема и с достаточно высокой степенью заданности вследствие его архитектурных особенностей.

Архитектурные особенности hp StorageWorks EVA

Дисковый массив hp StorageWorks Enterprise Virtual Array в отличие от аналогичных традиционных массивов в контексте данной публикации отличается прежде всего максимальным распараллеливанием и балансировкой запросов внутри массива. К примеру, если имеется 12 дисководов и на них организуют 3 группы RAID (включающих соответственно 6, 4 и 2 накопителя), то при физическом отражении их на структуру дискового массива все они будут "размазаны" на максимальное количество дисков, имеющихся в массиве (рис. 1). И это, помимо расширения функциональности (SN № 1 (10), 2002 г.), обеспечивает максимальное повышение его пропускной способности за счет динамической максимальной балансировки нагрузки.

Рис. 1. Физическое представление групп RAID в hp StorageWorks Enterprise Virtual Array.

Вследствие этого у потребителя появляется возможность прогнозирования уже реальной производительности в условиях равной доступности каждого накопителя в массиве и ограниченности накладных затрат, связанных с поддержанием подобного механизма распараллеливания запросов.

Результаты исследований

Основное преимущество, предоставляемое пользователю массивом hp EVA при оценке его производительности, это гораздо бóльшая детерминированность нагрузки в сравнении с традиционными массивами, что приводит к тому, что при множестве параллельно работающих задач с массивом каждая из задач практически не влияет на работу других до приближения к точке насыщения. Во всех тестах время реакции начинало превышать 30 мс (стандартная норма для систем midrange класса) при коэффициенте нагрузки свыше 92%. Т.е. в условиях генерирования множества параллельных нагрузок ко многим Vraid группам каждый из входных потоков практически "не ощущал" работу параллельных задач до коэффициента нагрузки дискового массива 0,92.

Проиллюстрируем это следующим примером. В условиях полностью детерминированной нагрузки (одного потока) для RAID традиционной организации, например со схемой 0, время реакции возрастает практически вертикально при достижении точки насыщения (рис. 2) и меняется только от количества накопителей в группе ("Проекти-

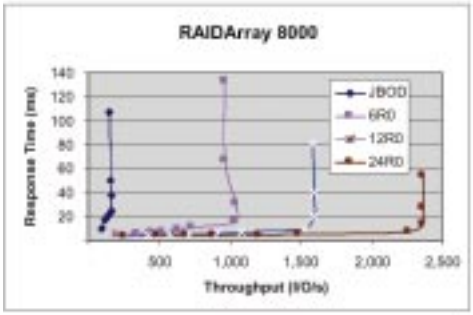


Рис. 2. Зависимость времени реакции дисковой системы от интенсивности входной нагрузки и числа накопителей (6, 12, 24) в RAID 0 (вх.нагрузка: 4 KB I/O, 50% Read, 2 GB Seek, Chunk=128KB, Writeback, Read-ahead).

рование дисковых массивов" – SN № 2 (7), 2001 г.). Поведение отдельных Vraid в массиве в отличие от данного примера более плавно, но до коэффициента нагрузки 0,92 его можно интерполировать с приведенным примером.

Заметим, что в условиях полностью нормального распределения входного потока запросов к устройству (в соответствии с теорией массового обслуживания), время реакции ($t_{реакц.}$) определяется по формуле:

$$t_{реакц.} = t_{обсл.}/(1 - k_{нагрузки}),$$

т.е., если время обслуживания запроса накопителем (устройством) сопоставимо с 10 мс, то при при коэффициенте 0,92 оно должно возрасти в 12,5 раз.

При оценке производительности массива с модульной структурой необходимо учитывать пропускную способность 4-х составляющих: хоста, контроллеров массива (front-end и back-end), интерфейса (между самими накопителями и контроллерами) и самих дисков. И в условиях предполагаемого максимального распараллеливания нагрузки вопрос оценки производительности

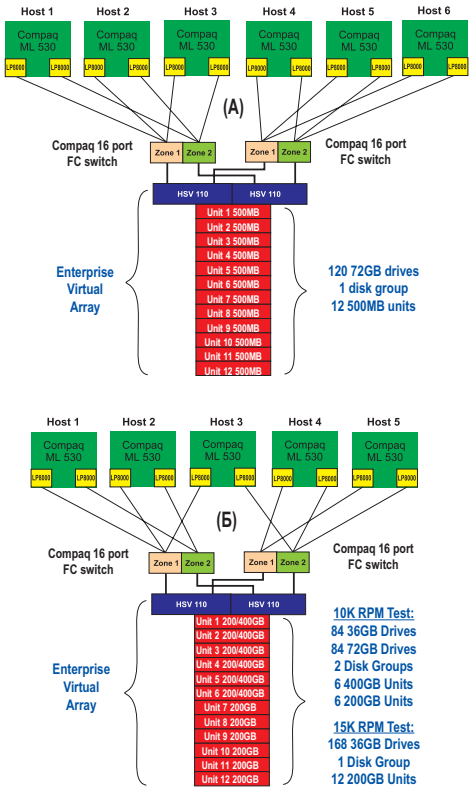


Рис. 3. Конфигурации тестирования hp EVA по первой (А) и второй (Б) серии тестов.

фактически сводится к определению "узкого" места среди названных выше составляющих в условиях заданной входной нагрузки.

Все тестовые серии проводились на двух базовых конфигурациях (рис. 3). Первая серия тестов имела целью предельные показатели пропускной способности массива в условиях его 100% загрузки, в которых время реакции или полное время обработки запроса игнорировалось. Вторая – ставила целью приблизить оценку показателей массива к работе с реальными приложениями и учетом условий, накладываемых при этом на работу массива.

Первая серия состояла из пяти последовательностей.

- ***64 KB sequential read (MB/s)*** – пять хостов, генерирующих последовательные команды чтения ввода/вывода, начинающиеся с LBN 0 (Logical Block Number) к каждому из модулей конфигурации. Счетчик байтов был установлен на 64 Кбайт. Тестирование выполнялось 8 раз, с числом команд IO к каждому модулю от каждого хоста, начиная с 1 и удвоением до 128 (1, 2, 4, 8, 16, и т.д.). Результат – самое высокое наблюдаемое значение.
- ***64 KB sequential write, mirrored cache (MB/s)*** – точно такая же последовательность, что и *64 KB sequential reads*, за исключением того, что использовались только операции записи.
- ***64 KB sequential write, non-mirrored cache (MB/s)*** – такая же, как и предыдущая серия, но без зеркалирования кэша.
- ***2 KB read cache latency (ms)*** – один хост, обращающийся за 0,5 Кбайт данных с диапазоном поиска 16 блоков на одном модуле. Число выданных команд варьировалось от 1 до 4. Результат – самое низкое наблюдаемое время ответа.
- ***0.5 KB read cache rate (Req/Sec)*** – 4 хоста, делающие случайные запросы чтения по 0,5 Кбайт в диапазоне 128 блоков, при условии, что каждый хост обращается к единственному модулю. Число выданных запросов ввода/вывода было различным – от 1 до 8 на модуль. Результат – самая высокая наблюдаемая скорость обработки запросов.

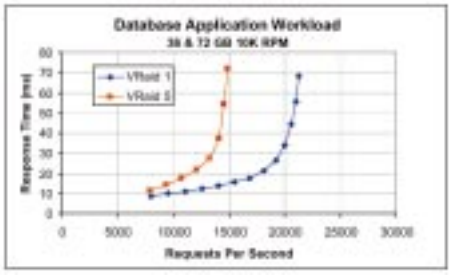
Результаты тестирования первой серии представлены в табл. 1.

Табл. 1. Результаты тестирования EVA на "насыщение".

Профиль рабочей нагрузки	Vraid 0	Vraid 1	Vraid 5
64 KB seq. read (MB/s)	527	528	529
64 KB seq. write, mirrored cache (MB/s)	173	163	155
64 KB seq. write, non-mirrored cache (MB/s)	559	505	355
2 KB read cache latency (ms)	0,115	0,116	0,115
0.5 KB read cache rate (Req/Sec)	162 000	162 000	162 000

Тест 1 (Database Workload).

Данная прикладная нагрузка была разработана для моделирования потока ввода/вывода, генерируемого высоконагруженными, многопользовательскими приложениями реляционных баз данных. Эти приложения имеют тенденцию к генерированию взрывообразного потока ввода/вывода. Дополнительные приложения в этой категории



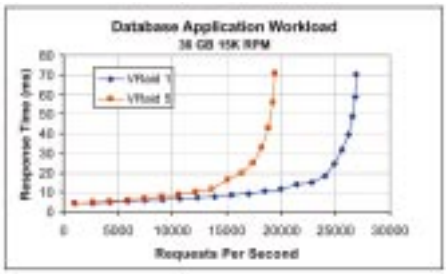


Рис. 4. Результаты тестировании EVA на "database" нагрузке с дисками 10K rpm и 15K rpm.

включают:

- транзакционные системы – финансовые, бронирования билетов и др.;
- ERP и MRP системы;
- системы электронной коммерции (e-commerce);
- системы обслуживания и поддержки пользователей.

Поток запросов случайных операций чтения и записи по 8 Кбайт, однородно распределенных по всем модулям (см.рис. 3), генерировался пятью хостами. *(Замечание: этот поток объемом свыше 1 Тбайт данных приводит к очень низкой вероятности удачного обращения в кэш чтения).* Из общего потока ввода/вывода: 67% – операции чтения, 33 % – записи. Тестирование выполнялось многократно с увеличивающимся потоком ввода/вывода, пока среднее время ответа не превышало установленный уровень. Результаты тестирования в динамике приведены на рис. 4, сводные итоги – в табл. 2.

Табл. 2. Результаты тестирования EVA на прикладных нагрузках.

Профиль рабочей нагрузки	Vraid 1 10K rpm	Vraid 5 10K rpm	Vraid 1 15K rpm	Vraid 5 15K rpm
Database (8 KB, 67% reads, random)	19 600	13 500	25 400	17 900
Web server (8 KB, 100% reads, random)	21 200	21 100	29 900	29 700

Проанализируем полученные результаты. Воспользуемся данными уже упоминавшейся публикации (SN № 2 (7), 2001 г.). Соотношение отдельных составляющих времени обслуживания для накопителей с разным числом оборотов в первом приближении представлено на рис. 5, на основании которого можно оценить, что общее время обслуживания запроса блока размером

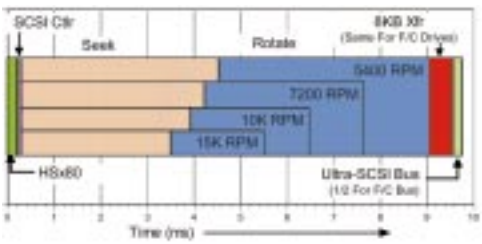


Рис. 5. Соотношение времен для дисков с различными скоростью вращения и типами интерфейсов.

8 Кбайт накопителем с интерфейсом FC составляет примерно 7 мс и 6 мс соответственно со скоростью 10K rpm и 15K rpm, или, что соответствует примерно обработке 142 запросов/с и 166 запросов/с. Максимальная потоковая пропускная способность EVA составляет около 520 Мбайт/с (см. табл. 1), что значительно выше потока 319 Мбайт/с (319 = 30x8x1,33 при случайном потоке запросов 30K в секунду, блоке 8 Кбайт и Vraid 1), следовательно, общая пропускная способность (при снятых прочих ограничениях) будет определяться только производительностью накопителей.

Необходимо учитывать, что реальный поток запросов на физическом уровне для схем RAID 1 и 5 будет больше из-за дополнительных операций записи. Поэтому, чтобы привести реальную нагрузку в соответствие с измерениями, представленными на рис. 4, ее нужно скорректировать соответственно на коэффициент 1,33 для VRAID 1 и на 1,99 – для VRAID 5

В результате получаем максимальный прогноз нагрузки (при максимальном распараллеливании) для VRAID 1/5 (10K rpm):

$$\text{Макс.нагр.}_{\text{Vraid 1/10K}} = (142\text{x}168 \text{ дисков})/1{,}33 = 17\,937 \text{ запросов/с}$$

$$\text{Макс.нагр.}_{\text{Vraid 5/10K}} = (142\text{x}168 \text{ дисков})/1{,}99 = 11\,988 \text{ запросов/с}$$

$$\text{Макс.нагр.}_{\text{Vraid 1/15K}} = (167\text{x}168 \text{ дисков})/1{,}33 = 21\,094 \text{ запросов/с}$$

$$\text{Макс.нагр.}_{\text{Vraid 5/15K}} = (167\text{x}168 \text{ дисков})/1{,}99 = 14\,014 \text{ запросов/с}$$

В среднем имеем отклонение в 20% (при отсутствии точных сведений о накопителях), но зато отклонение всего в 3% (!) соотношения полученных измерений для VRAID 1 и VRAID 5 (коэффициент отличия равен 1,99/1,33 = 1,496).

Во всех случаях нагрузка (при времени реакции 30 мс) отличалась от максимальной на коэффициент 0,92.

Имея результаты тестирования насыщения массива по передаче данных (см. табл. 1), можно достаточно просто определить предельную пропускную способность, в частности, для Vraid 1 при ограничении по передаче данных и размере блока, например, 24 Кбайт. Положим, что работа происходит при незеркалированном кэше и т.к. максимальные скорости передачи по чтению и записи близки (528 Мбайт/с и 505 Мбайт/с), усредним их на значение 516. Тогда максимальный входной поток запросов ($x_{вх}$) определится из выражения:

$$x_{вх} = 516/(24 \times 1{,}33) = 16165 \text{ запросов/с}$$

Тест 2 (Web Server-1).

Данная нагрузка была разработана для моделирования динамичного ввода/вывода с доминированием операций чтения, характерного для web-серверов. Нагрузка приближается к использованной в тесте 1, за исключением того, что отсутствуют операции записи.

Результаты тестирования приведены на рис. 6 и в табл. 2. Все прогнозные оценки рассчитываются как и в предыдущем тесте, за исключением понижающих коэффициентов, вследствие отсутствия операций записи, благодаря чему, графики для VRAID 1 и 5 совпадают. Отклонения от прогнозных оценок находятся уже в диапазоне 3-13% (!). Соотношение между допустимой и максимальной нагрузкой осталось на том же уровне – 0,92-0,93.

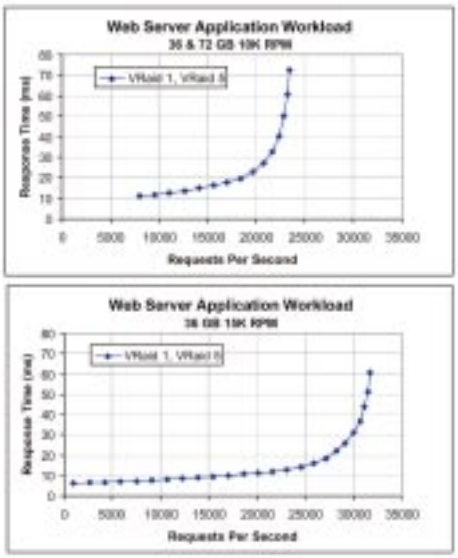


Рис. 6. Результаты тестировании EVA на "web-server" нагрузке с дисками 10K rpm и 15K rpm.

Имея результаты по чтению из кэша (см. табл. 1) при известной вероятности выборки из него можно прогнозировать производительность с поправкой на кэш.

Здесь уместно сделать одно важное замечание. Так, официально заявляемое максимальное значение обработки запросов (IOPS), обеспечиваемое только накопителями у EVA (55K), практически совпадает со значением xp1024 (57K). Поэтому приобретая дисковый массив, нужно правильно понимать его позиционирование. И если при заданном объеме кэша вероятность выборки из него невелика, то при превалирующих случайных обращениях будут неэффективно использоваться дорогостоящие high-end массивы.

Тест 3 (Web Server-2).

Данный тест являлся контрольным для подтверждения правильности возможности предложенного подхода прогнозирования пропускной способности hp EVA.

В качестве нагрузки использовалась смесь *web server*, как в предыдущем тесте. Условия проведения эсперимента отличались тем, что нагрузка генерировалась только одним сервером, а в качестве атрибута при определении максимальной производительности (и условии, что время реакции не превышает 30 мс) использовалось количество дисков в RAID группе. Результаты проведенных измерений (рис. 7) полностью подтвердили теоретическое изменение производительности как самого массива, так и метода его прогнозирования. При этом величина отклонения находилась в пределах 15%.

Рис. 7. Результаты тестировании EVA на "web-server" нагрузке с дисками 10K rpm при изменении их количества в RAID группе.

Заключение

Проведенные результаты измерений в условиях заданных ограничений позволяют для модульных массивов hp EVA сделать следующие выводы:

- *массивы hp StorageWorks EVA поддерживают производительность, близкую к максимальной без использования специальных программных средств или дополнительных усилий со стороны администратора;*
- *имеется возможность достаточно точно прогнозировать их производительность для заданного профиля нагрузки;*
- *при заданной входной нагрузке можно достаточно точно прогнозировать масштабируемость массивов;*
- *с достаточно точной степенью прогнозируемости можно развивать IT инфраструктуру на их основе с учетом заданных нормативных коэффициентов на используемость ресурсов.*

Большáя детерминированность поведения массива и предсказуемость результатов hp StorageWorks EVA в показателях производительности – фактор, который в условиях неспокойной экономики может дать уверенность в завтрашнем дне и оказаться фактором, не самым последним при развитии IT инфраструктуры организации.