SAN Volume Controller на марше

В мае с.г. IBM анонсировала SAN Volume Controller ("in-the-data-path", или симметричное решение виртуализации с кодовым названием Lodestone) – первую часть своей программы виртуализации. IBM провела предварительное комплексное тестирование новых продуктов и представила их в региональных ДемоЦентрах. Результаты этого тестирования – тема публикации (в продолжение статьи прошлого номера: SN № 1 (15), 2003 г.).

Введение

Тестирование первого из решений, поставленных в Россию на базе SAN Volume Controller (SVC), проводилось в рамках программы Wave, в региональном ДемоЦентре IBM. Новое устройство SAN Volume Controller представляет собой комбинацию программных и аппаратных средств, предназначенных для создания промежуточного виртуализационного слоя, который производит отображение физических дисков в виртуальные.

Идея, заложенная в SVC, заключается в отделении серверов приложений от конкретных дисковых подсистем. Серверы приложений работают с виртуальными дисками. Со стороны приложения виртуальный диск воспринимается как физический диск, непосредственно подключенный к серверу. Реальные данные с виртуального диска располагаются на нескольких физических дисках, которые принадлежат одному или нескольким дисковым массивам. Виртуальные диски создаются из доступных, управляемых SVC, физических дисков (Managed Disk). Такими управляемыми дисками являются RAID группы, представляемые в SAN дисковыми массивами. Виртуальный диск, создаваемый SVC, физически может быть расположен на нескольких дисковых массивах, которые могут отличаться по емкости, производительности. В настоящий момент поддерживаются дисковые массивы IBM. Активно тестируются дисковые стойки HP, Hitachi и EMC.

Аппаратная реализация

Аппаратно SAN Volume Controller представляет собой кластер, в минимальной конфигурации состоящий из двух двухпроцессорных (node&node) Intel серверов с процессорами Pentium 4 2,4 GHz. Используется операционная система на базе ядра Linux, с частью кода, написанного специалистами IBM. Операционная система хранится на внутреннем жестком диске SCSI. Следует заметить, что этот диск не дублируется и в случае выхода из строя заменяется на новый, после чего узел грузится с Flash диска и по сети хранения копирует данные со второго узла кластера. Для передачи данных используется четыре 2 Gb/s порта на узел кластера, что в минимальной конфигурации дает пропускную способность 16 Gb/s. Все порты равноправны и подключаются к коммутатору SAN. Управление SVC может осуществляться двумя способами: через Ethernet порт на основе web-интерфейса; через serial порт на основе специально выделенной платформы и приложения. При этом управление SVC происходит как интегрированным ресурсом вне зависимости от числа узлов. Для этого один из узлов назначается главным и служит для управления всеми узлами. В случае выхода его из строя, все узлы получают об этом информацию и происходит автоматическое переназначение узла управления.

Механизм виртуализации, который при этом используется, называется симметричным (рис. 1). Его основное отличие от ассиметричного – в том, что все потоки данных проходят через устройство управления виртуализацией (в данном случае – Lodestone). А распределение дискового пула между хостами осуществляется зонированием на уровне коммутатора (рис. 2). Общая логическая топология Lodestone с основными особенностями в реализациях представлена на рис. 3.

Между всеми портами кластера осуществляется балансировка нагрузки. Балансировка осуществляется как на SVC, так и на серверах, и на дисковых массивах. Устройство имеет информацию о количестве FC портов у каждого сервера и дискового массива, и в случае изменения числа портов определяет, какие порты работают и перенаправляет ввод/вывод.

Балансировка нагрузки позволяет создавать масштабируемые интегрированные производительные решения уровня предприятия из небольших дисковых массивов. Большие и производительные массивы имеют тенденцию к увеличению времени реакции обрабатываемых запросов I/O при увеличении числа подключенных дисков. Это связано с архитектурным решением, использованным при создании дискового массива. Например, дисковый массив, построенный по

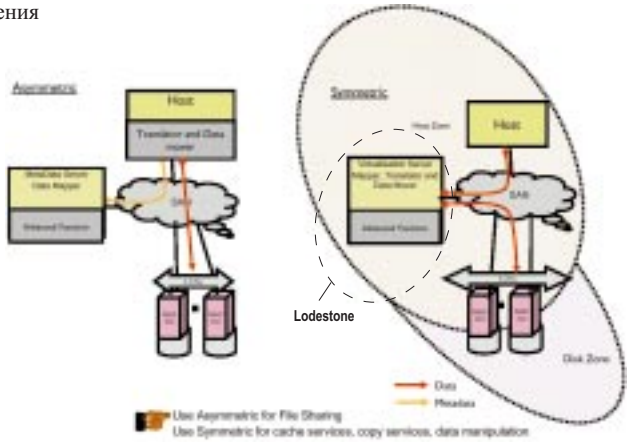

Рис. 1. Концепция виртуализации SVC является симметричной.

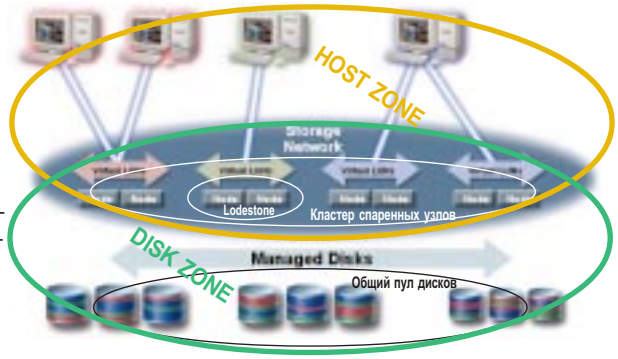

Рис. 2. Зонирование серверов и дискового пула на основе Lodestone.

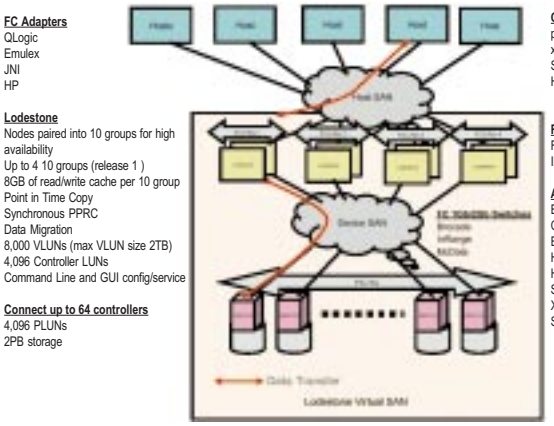

Рис. 3. Логическая топология Lodestone.

технологии Fibre Channel-Arbitrated Loop (FC-AL), в один момент времени позволяет работать только с одним диском в FC кольце. При увеличении числа дисков неизбежно начинают возникать простои при обращении к нескольким дискам, подключенным к одному кольцу. Устройство SAN Volume Controller позволяет распределить данные по разным дисковым массивам и обращаться к ним одновременно. В сочетании с общим для двух узлов кэшем на 8 GB это позволяет обеспечить очень высокую производительность. Использование общего кэша позволяет повысить производительность для приложений, активно работающих с кэшем (cash-friendly). При большой нагрузке в случае, когда кэша не хватает, производительность слегка снижается. На реальных тестах, проведенных в ДемоЦентре, при большой нагрузке производительность SVC была ниже на 1% по сравнению с непосредственным подключением дискового массива (см. тест 2), однако при средних нагрузках снижения производительности не было, а на FAStT 200 был даже некоторый прирост за счет большого кэша.

При создании устройства большое внимание уделялось надежности, именно поэтому минимальная конфигурация – это кластер из двух узлов. Помимо этого каждый узел кластера имеет собственный источник бесперебойного питания (UPS). Каждый UPS имеет внутренний код, разработанный в IBM, который постоянно тестирует различные параметры и сообщает о них узлу. При пропадании питания или различных поломках SVC узнает о происходящем и сообщает оператору. Кроме защиты от потери питания в каждый узел кластера установлен специальный механизм тестирования состояния системы (WatchDog), который отслеживает сбои и автоматически перегружает систему. Для конфигурирования и управления каждый узел имеет встроенный сервисный процессор.

Основное преимущество, которое дает SVC – это отделение серверов от конкретных дисковых массивов. Данные можно переносить с устройства на устройство, динамически менять размеры виртуального диска, копировать виртуальные диски независимо от приложений. Преимущества использования виртуальных дисков заключаются в возможности динамического переноса данных с одного массива на другой без останова приложения (это позволяет проводить плановое обслуживание дисковых подсистем без отключения пользователей) и динамического изменения размеров виртуального диска.

В дополнение к поддержке разнородных дисковых подсистем SAN Volume Controller поддерживает расширенные функции копирования. Если раньше возможности по созданию мгновенных копий (FlashCopy) и удаленного зеркалирования данных (Remote Mirroring) ограничивались рамками одного массива, то с появлением SVC стало возможным использовать эти функции для любых дисковых подсистем, подключенных к SAN. Например, при тестировании данные зеркалировались между дисковым массивом ESS и FAStT. Преимущества этого очевидны. Раньше для создания отказоустойчивого решения необходимо было приобретать дисковые подсистемы парами. При потребности в высокой производительности это влекло за собой большие затраты, так как фактически приобретались два высокопроизводительных и, соответственно, дорогих устройства. Теперь нет необходимости приобретать вторую высокопроизводительную дисковую стойку для удаленной площадки. Достаточно иметь ее на основной площадке и зеркалировать с недорогой системой на удаленной. Поддержка мгновенной копии данных (FlashCopy) тоже распространяется на всю сеть хранения. С SVC можно создавать мгновенную копию любого виртуального раздела, который может быть расположен на любом устройстве в SAN, причем сама мгновенная копия может хранится также на любом массиве. Расширение этих функций на всю сеть хранения позволяет не заказывать функции PPRC и FlashCopy вместе с дисковыми массивами, что существенно снижает затраты. Так, например, стоимость функции удаленного зеркалирования с двумя дисковыми стойками ESS приблизительно равна стоимости устройства SAN Volume Controller.

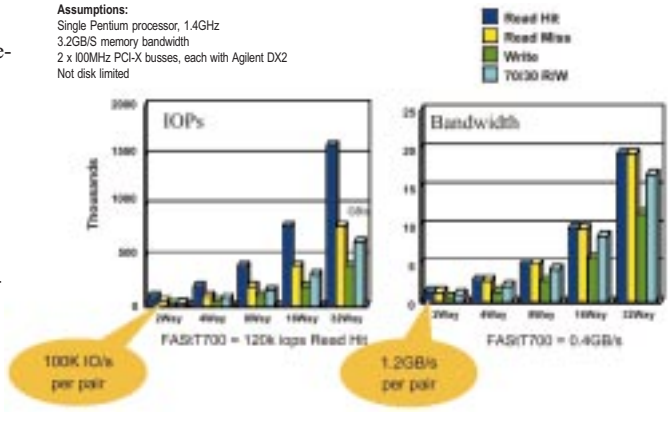

Рис. 4. Производительность SVC потоковая и на случайных операциях.

Результаты тестирования

Проводимые тесты ставили целью определить предельные показатели пропускной способности SVC и характер его поведения при изменении нагрузки. Всего было выполнено два эксперимента по тестированию SVC.

Тест 1. Определение максимальной пропускной способности SVC.

Данный тест был основным и ставил задачу определить предельные показатели производительности SVC на случайных операциях и потоковой нагрузке с изменяющимися количеством узлов в конфигурации и условиями нагрузки. К SVC был подключен дисковый массив FAStT 700. Измерения проводились для 2, 4, 8, 16 и 32-узловой конфигурации. В настоящий момент официально объявлено о поддержке 8 узлов, но в дальнейшем будет обеспечиваться поддержка до 32 узлов в кластере, и эта конфигурация тестируется.

Общие результаты тестирования приведены на рис. 4 и в табл. 1 отдельно для 2- и 8-узловой конфигураций.

Табл. 1. Значения максимальных показателей производительности SVC для 2 и 8 узлов

Нагрузка	2 Nodes (op/sec)	8 Nodes (op/sec)	2 Nodes (M/sec)	8 Nodes (MB/sec)
Read Cache Hit	100 000	400 000	1 500	6 000
Read Cache Miss	50 000	200 000	1 000	4 000
Writes into cache *(no destage)*	40 000	160 000	600	2 400
Writes into cache *(incl. destage)*	25 000	100 000	500	2 000
70%/30% Read/Write *(incl. stage/destage)*	40 000	160 000	750	3 000

Обозначение Read Hit и Read Miss (см. рис. 4) относится к наличию и соответственно отсутствию требуемых данных в кэше SAN Volume Controller. На том же рисунке видно, что при полном “непопадании” в кэш SVC устройство выдает 100 000 IO/s, в конфигурации из 2-х узлов. Отметим, что производительность FAStT 700 при непосредственном подключении и 100% “попадании” в кэш составляет 120 000 IO/s.

Из полученных данных видно, что производительность SVC растет линейно с увеличением числа узлов кластера. Это говорит о высокой масштабируемости и эффективности использования ресурсов за счет распараллеливания потоков данных. Заплатив в четыре раза больше, компания приобретает в четыре раза большую производительность, что весьма существенно.

Тест 2. Определение процента накладных затрат для поддержания механизма распараллеливания при распределении группы RAID нескольким массивам.

Механизм балансировки нагрузки виртуального тома в SVC – концептуальное положение подхода IBM в реализации ее программы виртуализации. И определение “дани”, которую приходится платить за дополнительный сервис – чрезвычайно важный показатель, фактически определяющий возможность использования SVC совместно с другими продуктами хранения.

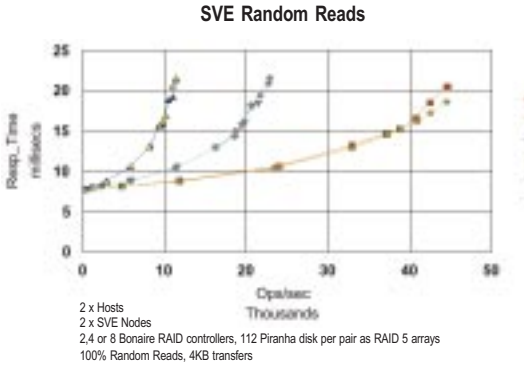

Рис. 5. Изменение времени ответа на операциях чтения при прямом подключении массива и через SVE.

Как показали исследования IBM еще на этапе разработки, подключение к SVC непосредственно RAID не оправдано из-за резкого увеличения нагрузки на ЦП SVC. Поэтому было принято решение о том, что RAID должен реализовываться средствами дисковых массивов.

Тестирование проводилось на 2, 4 и 8 RAID контроллерах Bonaire со 112 дисками Piranha на пару контроллеров. Диски были сконфигурированы в RAID 5 группы. Читались случайные блоки данных размером 4 Кбайт. Результаты тестирования представлены на рис. 5. Во всех случаях накладные затраты по обслуживанию балансировки не превысили 1%.

Возможные решения на основе SVC с учетом результатов тестирования

SAN Volume Controller позволяет строить высокопроизводительные решения среднего и высокого уровней на небольших и немощных дисковых массивах, которые по объему и производительности будут сравнимы с дисковыми стойками High-End класса. Производительность будет следствием высокого распараллеливания потоков данных и своего рода кэша “второго уровня”, так как помимо кэширования на дисковом массиве, SVC тоже кэширует данные. Такое решение обеспечивает большие возможности по динамическому наращиванию доступного дискового пространства без снижения производительности, которая неизбежно возникает с ростом числа дисков на одном устройстве. Что еще важно – использование SVC позволяет получить интегрированное решение из совершенно разных дисковых подсистем. Это решение дает возможность администратору избавить конечных пользователей от проблем, возникающих в сети хранения данных.

Другим возможным решением является построение резервного центра на основе дискового массива средней производительности. До появления SVC для построения такого решения необходимо было приобретать дисковые стойки парами, так как функция зеркалирования томов поддерживалась только в рамках одного устройства. SVC осуществляет зеркалирование виртуальных дисков, физически расположенных на любых доступных дисковых массивах.

Заключение

В целом новое устройство позволяет расширить возможности по использованию функций FlashCopy и PPRC и существенно снизить затраты на их приобретение и использование. Реализация концепции виртуальных дисков значительно упрощает администрирование, уменьшает простои, связанные с техническим обслуживанием дисковых массивов и намного увеличивает диапазон характеристик дисковых систем.

Михаил Воробьев
demo.ibm@storagenews.ru

Новые предложения IBM в области Grid-систем

28 апреля 2003 г. **-** IBM анонсировала новые предложения, которые еще больше расширят возможности применения Grid-вычислений*) в коммерческих организациях, включая четыре отраслевые решения для нефтяной и электронной промышленности, высшего образования и агрохимии. Кроме того, IBM объявила, что более 35 компаний, в том числе лидер рынка сетевого оборудования Cisco Systems, будут совместно с IBM работать над построением фундамента экосистемы Grid для более широкого использования Grid-вычислений в бизнесе.

Построение экосистемы Grid

Ни одна компания самостоятельно не может обеспечить внедрение Grid-вычислений в коммерческом секторе, поэтому IBM представила инициативу по построению экосистемы Grid, включающей производителей ПО и бизнес-партнеров, которые помогут разрабатывать коммерческие решения Grid.

Cisco присоединилась к более 35 новым и существующим бизнес-партнерам и другим производителям, которые строят фундамент для разработки экосистемы IBM Grid. IBM и Cisco совместно разрабатывают расширенные Grid-сервисы для сетей хранения данных Storage Area Networks (SAN). Интеллектуальная многоуровневая сетевая архитектура поможет заложить основу для создания глобального масштабируемого доступа к данным Grid. Интеграция интеллектуальных сервисов в сеть поможет упростить доступ к данным, совместное управление ресурсами и управление ими в масштабах Grid. Ранее в этом году IBM заключила соглашение с Cisco о поставках коммутаторов для SAN MDS 9000. Cisco планирует усовершенствовать эти коммутаторы для реализации глобального масштабируемого доступа к данным через Grid.

“Масштабируемый и защищенный доступ к хранящимся данным является критическим элементом Grid-вычислений и позволяет компаниям и их партнерам управлять огромными объемами данных и сотрудничать по всему миру, – сказала Сони Джандани (Soni Jiandani), вице-президент по маркетингу Storage Technology Group в Cisco. – Решение IBM и Cisco – это важный шаг в реализации преимуществ Grid и вычислений по требованию для компаний любого масштаба”.

Сегодняшний анонс расширяет существующий глобальный альянс между Cisco и IBM. Две компании объединят свои разработки в области инфраструктуры Internet, систем и сервисов электронного бизнеса для реализации полного решения Internet для предприятий и сервис-провайдеров.

К инициативе Cisco и IBM по построению экосистемы Grid присоединились следующие разработчики прикладного программного обеспечения:

- **Accelrys**, планирующая обеспечить решения НИОКР для наук о жизни.
- **Cadence**, планирующая обеспечить решения по автоматизации проектирования для электронной промышленности.
- **Calypso Technology** использует Grid-вычисления для ускорения выполнения финансового анализа для банковского сектора.
- **Force10 Networks Inc.** планирует совместно с IBM реализовать для всех отраслей оптимизацию предприятий с поддержкой значительной нагрузки на сеть, которые создают Grid-вычисления в Grid-среде.
- **Landmark Graphics Corporation**, которая интегрирует все аспекты добычи нефти и газа с помощью ведущего программного обеспечения и сервисов, планирует совместно с IBM разработать решения Research and Development Grid для своих заказчиков из нефтяной и газовой промышленности.
- **Mercury Interactive** планирует совместно с IBM обеспечить высокое качество программного обеспечения для оптимизации предприятий всех секторов.
- **MSC.Software** планирует совместно с IBM разрабатывать решения проектирования для промышленности.

Более двадцати бизнес-партнеров IBM также помогают формировать экосистему IBM Grid.

Решения Grid для промышленности

IBM, анонсировавшая в январе стратегию вывода на рынок вертикальных решений Grid для аэрокосмической и автомобильной промышленности, финансового и государственного сектора и наук о жизни, объявила о новых решениях еще для четырех отраслей.

- ***Агрохимия*** – Analytics Acceleration Grid и Information Access Grid (ориентированы на исследования, разработку и бизнес-аналитику). Analytics Acceleration Grid поможет ускорить научные открытия, включая создание высокоурожайных сортов, за счет улучшения вычислительных ресурсов и систем хранения. Information Access Grid поможет добиться максимальной эффективности использо-

*) GRID-вычисления активно используются в науке. Одной из наиболее масштабных реализаций был запуск в эксплуатацию в 2002 г. TeraGRID системы *(www.teragrid.org)* для научных исследований – *прим. ред.*