

Гипервизоры и СХД

Обзор особенностей интеграции гипервизора VMware vSphere 5 с компонентами уровня хранения данных. Статья — первая из серии на эту тему.



Сергей Платонов — менеджер по продуктам, компания AVRORAID.

Введение

В 2011 г. производители средств серверной виртуализации анонсировали и выпустили новые версии гипервизоров. Покупателям уже доступны VMware vSphere 5, Citrix Xen Server 6. Следующая версия гипервизора от Microsoft появится одновременно с Windows 8. Однако уже сейчас благодаря наличию доступного Windows 8 Developer Preview можно оценить выполненные нововведения. По-прежнему в центре виртуализованных ЦОД находятся системы хранения данных. В данном цикле статей рассматриваются новые возможности гипервизоров в рамках их взаимодействия с СХД.

Часть 1. VMware vSphere

Лидером серверной виртуализации на данный момент является продукт компании VMware vSphere.

1.1 Файловая система VMFS

В VMFS-5 используется унифицированный блок размером 1 МБ, теперь большие файлы могут быть созданы при использовании блока 1 МБ. Ранее для этого приходилось использовать блок размером в 2-8 МБ. Данное нововведение упрощает администрирование виртуальной среды и

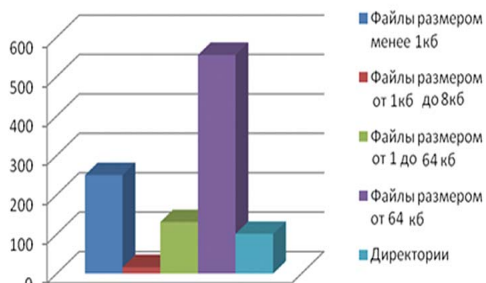


Рис. 1. Распределение размеров файлов файловой системы виртуальной инфраструктуры средней компании (400 человек)

предостерегает от ошибки неправильного выбора размера блока.

При выполнении обновления с VMFS-3 до VMFS-5 следует понимать, что размер блока остается первоначально установленным, то есть его изменение потребует переформатирования файловой системы.

Большинство нововведений для гипервизоров обусловлено необходимостью увеличения масштабируемости и снижения накладных расходов при использовании файлов небольшого размера. Одним из изменений является использование субблока меньшего размера, который снизился с 64 КБ до 8 КБ.

Субблок меньшего размера позволит использовать не более 8 КБ пространства для хранения небольших файлов размером от 1 до 8 КБ, в то время как с предыдущей версией файловой системы каждый файл требовал не менее 64 КБ дискового пространства. VMFS-5 способна предоставить 30000 восьмикилобайтных субблоков для хранения таких файлов, как журналы и метаданные виртуальных машин (.vmtx).

Для хранения файлов размером менее 1 КБ используется специальный небольшой блок, что позволяет еще больше снизить накладные расходы.

Нами было проведено исследование файловой системы виртуальной инфраструктуры средней компании (400 человек). Исследованная инфраструктура состоит из трех серверов гипервизоров, 20 хранилищ, развернутых на 2 СХД и 56 работающих виртуальных машин. Распределение размеров файлов показано на рис. 1.

Размер поддерживаемых логических томов был увеличен до 64 ТБ. Ранее для создания больших хранилищ требовалось выполнять расширение томов множеством “экстендов” размером в 2 ТБ. К сожалению, при реализации такого подхода многие администраторы столкнулись с рядом проблем и были вынуждены отказаться от использования расширенных томов. Проблема ограничения максимального размера устройства была связана

не только с VMFS-3, но и возможностями систем хранения данных. Производители СХД в новых прошивках и версиях своего ПО изменили ситуацию (табл. 1).

Табл. 1. Максимальный размер LUN наиболее популярных СХД.

Наименование СХД	Максимальный размер LUN
HP EVA 4400	64 TB
HP MSA P2000 G3	64 TB
Hitachi AMS 2100	60 TB
СХД на основе ПО AVRORAID	1 PB

В новой версии файловой системы VMFS-5 было учтено расширение возможностей популярных моделей СХД. Кроме того, добавился ряд полезных функций:

- поддержка большего количества файлов — более 100 000 файлов;
- расширена поддержка ATS.

VMware обещает простоту обновления с VMFS 3 до VMFS 5. Однако в процессе обновления были обнаружены следующие “подводные камни”:

- размер блока и суперблока остается неизменным;
- копирование файлов между хранилищами с разными размерами блока происходит без использования VAAI, что может оказывать влияние на производительность процесса vMotion;
- ограничение на количество файлов остается прежним (чуть более 30 000 файлов);
- размеры начального сектора обновленной и вновь созданной файловой системы отличны;
- таблица разделов после обновления остается прежней (MBR) до тех пор,

ESX 3.5 Relative Performance with FC as the Baseline at 100% Using 75% Sequential, 60% Read and 4K Request Size

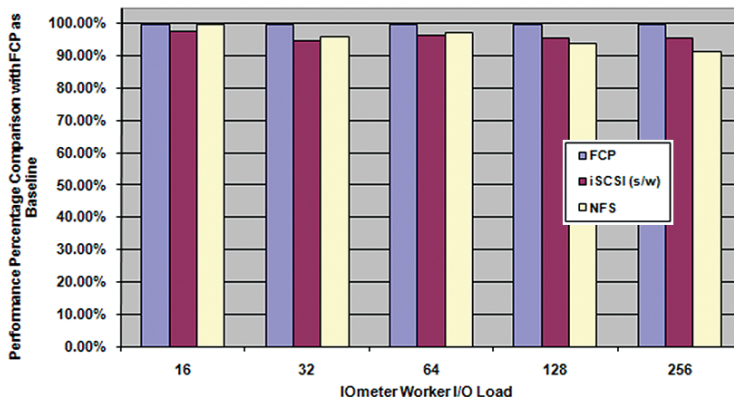


Рис. 2. Сравнение производительности протоколов (по данным компании NetApp).

пока размер файловой системы не превысит 2 ТБ.

В связи с этим, мы рекомендуем выполнять миграцию с VMFS-3 до VMFS-5 средствами Storage vMotion.

Теперь при использовании Passthru RDM могут быть использованы LUN размером 64 ТБ.

К сожалению, максимальный размер файла vmdk остался неизменным и равным 2 ТБ-512 Б, а использование RDM подразумевает ряд ограничений.

На наш взгляд, нововведения в VMFS-5 при глубоком рассмотрении не обрадуют пользователей, которым необходимы виртуальные машины с большими объемами данных и им по-прежнему придется использовать NFS системы хранения для размещения Datastores.

Эксперименты компании NetApp показали паритет производительности VMFS-3 с блочным хранилищем и хранилищем NAS (рис. 2).

Лаборатория AVRORAID проведет аналогичные эксперименты для сравнения производительности VMFS-5 и NAS на высокопроизводительных СХД с использованием программного обеспечения AVRORA.

Для поддержки устройств размером более 2 ТБ был изменен тип таблицы разделов с MBR на GPT.

1.2 Storage vMotion

Live Migration предназначен для перемещения виртуальных машин между различными аппаратными платформами без простоя в работе приложений. Для многих заказчиков данная функция является одной из значимых при выборе средств серверной виртуализации. Live Storage Migration может происходить между двумя СХД любого типа: Fibre Channel, iSCSI, NFS.

Live Storage Migration предназначена для следующих задач:

- выполнение обслуживания без простоя: пользователь может изменять параметры СХД и файловых систем, не прерывая работу приложений;
- автоматизированная и неавтоматизированная балансировка нагрузки на СХД;
- обеспечение мобильности виртуальных машин: пользователь больше “не привязан” к СХД.

В vSphere 5 был изменен механизм Storage vMotion, позволяющий перемещать виртуальные машины между различными LUN и системами хранения без простоя в работе.

Проследим, как происходили изменения в различных версиях vSphere. Механизмы выполнения Storage Vmotion можно оценивать по следующим параметрам.

Время миграции

Суммарное время миграции должно быть снижено, при этом по-прежнему необходимо обеспечивать идентичность копий виртуальных дисков и соответствие данных. Предсказуемость времени миграции также важна для планирования обслуживания.

Влияние на гостевые ОС

Влияние на гостевые ОС выражается во времени простоя приложений и потерях производительности дисковой подсистемы (в I/Ops).

Атомарность

Алгоритм Live Migration должен гарантировать атомарность переключения между источником и конечным томом СХД, что значительно повышает надежность, делает миграцию на больших расстояниях более безопасной. Алгоритм необходим для критических приложений.

Все варианты реализации Live Storage Migration одинаковы: виртуальный диск копируется с источника на приемник, однако, если во время копирования работающие виртуальные машины изменяют данные, эти изменения необходимо отражать на приемнике. После того как копии станут идентичными необходимо заново выполнять подключение работающих виртуальных машин к приемнику.

Для выполнения перемещения в vSphere 3.5 использовалась технология снапшотов. Изначально эта функция появилась как способ миграции с VMFS-2 до VMFS-3. При этом тома VMFS-2 работали в режиме только на чтение.

Первоначально создается снапшот для диска виртуальной машины, и происходит копирование основного диска на приемник. После завершения копирования образа диска создается второй снапшот, и происходит повтор транзакций, выполненных во время перемещения. После завершения повтора транзакций выполняется очередной снапшот и повтор транзакций. Данная процедура выполняется до тех пор, пока размер снапшота не станет меньше определенного порога. Как только порог достигнут, виртуальная машина останавливается, выполняется консолидацию последнего снапшота. Далее виртуальная машина стартует, используя ресурсы конечного тома.

Механизм снапшотов является простым и использует хорошо отработанную технологию, благодаря чему миграция виртуальных машин более надежна и устойчива к сбоям, чем механизм, используемый в vSphere 4.x

Процесс миграции с использованием снапшотов не атомарен. Выход из строя любого из томов приводит к завершению работы виртуальной машины. Снапшоты не подходят для выполнения миграции на значительные расстояния.

Для виртуальных машин с несколькими снапшотами возникают дополнительные накладные расходы на производительность и дисковое пространство (во время миграции виртуальная машина использует в три раза больше пространства).

При большой нагрузке на дисковую подсистему виртуальных машин дельта-файл снапшота увеличивается до огромных размеров, и дальнейший повтор выполняется слишком долго.

В версиях 4.x механизм Storage vMotion был изменен, и использовалась функциональность Changed Block Tracking (CBT) или Dirty Block Tracking. Это позволило отказаться от механизма снапшотов и по-

сле копирования основного файла в несколько приемов синхронизировать измененные за время копирования блоки. Для хранения информации об измененных блоках, называемых также “грязными” блоками, используется битовая карта.

Сначала происходит копирование диска на конечный том. В это время все измененные блоки регистрируются фильтром, находящимся на уровне между VMFS и VSCSI. По завершению копирования битовая карта используется для копирования измененных блоков, а затем очищается. Процесс повторяется до тех пор, пока количество “грязных” блоков в каждом цикле не станет постоянным и не будет видна положительная тенденция либо пока не будет достигнуто пороговое значение, учитывающее максимально возможное время простоя.

Переключение на новый датастор происходит при помощи Suspend / Resume.

Фильтр не отслеживает “грязные” блоки, которые еще не были скопированы, что значительно повышает быстродействие миграции, особенно при выполнении первой итерации.

Использование CBT / DBT предоставляет новые возможности по оптимизации алгоритмов благодаря переходу на более низкий уровень объектов (блоки вместо снапшотов). Также гарантируется атомарность переключения между начальным и конечным томами, что позволяет машине оставаться работающей в случае отключения конечного тома и выполнять миграцию на больших расстояниях через WAN.

Однако при большой нагрузке на СХД для выполнения Storage vMotion может потребоваться длительное время. Также при нагрузке на начальный том большей, чем пропускная способность миграции, изменения никогда не смогут “слиться”.

Для технологии CBT было сделано несколько оптимизаций, одной из которых стало исключение “горячих” блоков.

Определяются часто перезаписываемые блоки и их копирование откладывается. Данная функция не вошла в конечную поставку vSphere по ряду причин, но VMware планирует использовать указанную оптимизацию при применении технологии Mirror Driver.

При работе реальных приложений запросы ввода-вывода имеют некоторую локальность, временную и адресную.

VMware выполнила анализ трассировок VSCSI при эмуляции работы 100 пользователей Microsoft Exchange Server, под-

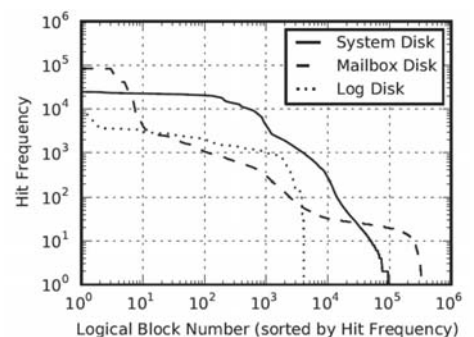


Рис. 3. Частота обращения к определенным блокам при эмуляции работы Microsoft Exchange Server.

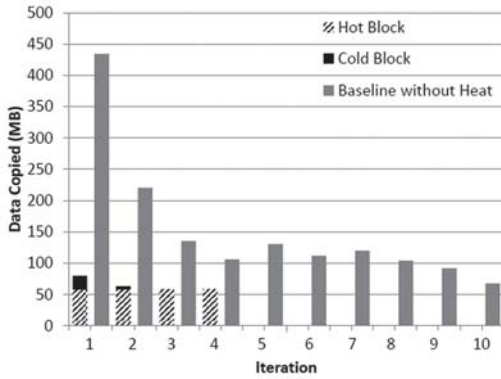


Рис. 4. Результаты оптимизации механизма СВТ. Количество скопированных блоков для каждой итерации. На левых колонках изображен результат оптимизации. На правых — работа без оптимизации.

твердив необходимость внедрения подобных оптимизаций (рис. 3). Применение алгоритма дало существенные преимущества (рис. 4).

В Storage vMotion 5.0 изменен механизм работы на однопроходный вместо множества итеративных копий. Storage vMotion 5.0 использует механизм Mirror Driver, который выполняет зеркалирование запросов записи на конечный том во время копирования основного диска (рис. 5).

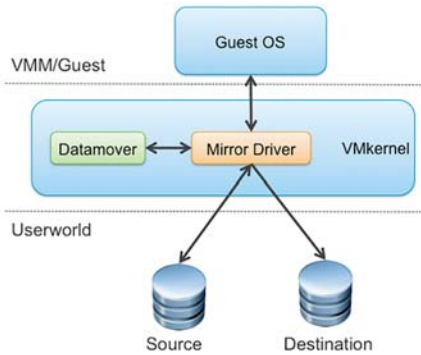


Рис. 5. Диаграмма работы Mirror Driver.

Перед включением Mirror Driver машина должна быть остановлена и запущена вновь после включения.

Основной процесс копирования выполняется с использованием VMkernel datamover. Во избежание проблем с некорректностью данных драйвер фильтра Ю Mirroring filter организует синхронизацию таким образом, чтобы предотвратить одновременный доступ data mover и виртуальной машины к одному и тому же региону, используя механизм блокировок.

Фильтр классифицирует запросы следующим образом:

- записи на регион, который уже был скопирован;
- записи на регион, который копируется;
- записи на регион, который будет скопирован.

Записи, относящиеся к уже скопированному региону, будут немедленно зеркалированы на конечный том.

Записи, обращенные на копируемый регион, поставлены в очередь. Сразу после завершения копирования региона DM с него снимается блокировка, и регион обновляется.

Еще не скопированные регионы не зеркалируются, запись выполняется сразу на конечный том, и чтение данных регионов будет выполняться только с конечного тома.

При использовании механизма Mirror Driver выполняется требование атомарности переключения между томами, гарантируется окончание синхронизации запросов на конечный том. Кроме того, значительно улучшается предсказуемость времени миграции машины.

В общем случае процедура storage vMotion выглядит следующим образом:

1. Рабочая папка виртуальной машины (VM) копируется посредством VРХА на целевой том.
2. “Теневая” VM, использующая скопированные файлы, запускается на целевом хранилище. “Теневая” VM после этого простаивает и ожидает завершения копирования vmdk файлов.
3. Storage vMotion подключает Storage vMotion Mirror драйвер для зеркалирования операции записи уже скопированных, но измененных на исходном хранилище, блоков на целевое хранилище.
4. Копирование файлов VM производится в один проход, в процессе зеркалирования операций I/O.
5. Storage vMotion вызывает Fast Suspend и далее Resume VM (как при vMotion), для того чтобы переместить работающую VM на “теневую” VM.
6. После завершения Fast Suspend и Resume старая директория и файлы VM удаляются с исходного хранилища.

В лаборатории AVRORAID было проведено тестирование быстродействия Storage vMotion с использованием СХД AVRORA.

В качестве системы хранения данных использовались две СХД, построенные на основе программного обеспечения AVRORA. Данные системы задействуют 16 дисков SAS 15 K RPM, организованные в высокопроизводительный RAID 6. СХД были подключены к серверу с гипервизором напрямую без использования SAN-коммутатора по двум портам FC 8Gb каждая, что ограничивало пропускную способность каждой СХД 1600 МБ/с (при реальной производительности СХД 2400 МБ/с).

В результате тестирования было создано 2 хранилища размером 1999GB. Работающая виртуальная машина содержала диск размером 500GB под нагрузкой приложения iometer, который имитировал работу СУБД.

Измерялось время миграции на различных версиях vSphere относительно нагрузки на виртуальный диск при постоянном размере тестового файла iometer и относительно размера тестового файла при постоянной нагрузке.

На рис. 6, 7 представлены графики, отображающие время миграции на предыдущих версиях гипервизора относительно версии ESX 5.1.

Далее будет подробно рассмотрен процесс миграции с использованием модуля Deep View, а также проанализированы другие показатели, критичные для миграции виртуальных машин, такие, как

Продолжительность выполнения миграции виртуальной машины гипервизорами ESXi 3.5 и ESXi 4.1 относительно ESXi 5

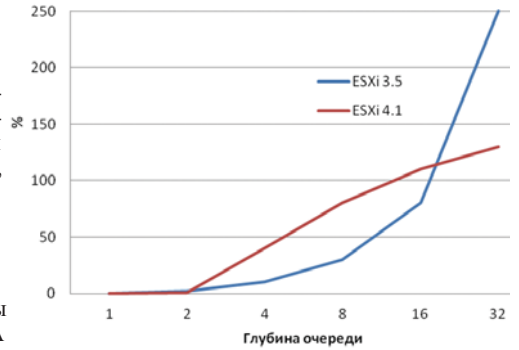


Рис. 6. Зависимость времени миграции от глубины очереди. Зависимость показана в виде отставания предыдущих версий гипервизора от ESXi 5 в процентном соотношении.

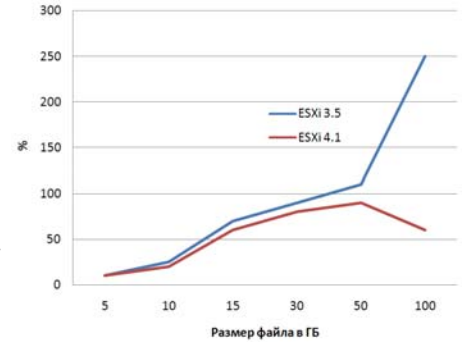


Рис. 7. Зависимость времени миграции от размера файла. Зависимость показывает отставание предыдущих версий гипервизора от ESXi 5 в процентном соотношении.

время простоя, влияние на производительность, и выполним сравнение не только между версиями VMware, но и между технологиями конкурентов.

Также Storage vMotion теперь поддерживает машины со снапшотами / linked clones.

1.3 VAAI

В API интеграции с системами хранения данных vSphere Storage APIs for Array Integration (VAAI) добавлены новые примитивы, позволяющие значительно улучшить технологию выделения пространства по требованию (Thin Provisioning) и решить следующие проблемы:

- при удалении или перемещении виртуальной машины из датастора массив не определяет освободившиеся блоки, при этом средства управления массивом показывают гораздо меньшее свободное пространство, чем есть на самом деле;
- перерасход пространства влияет на все машины, находящиеся в датаSTORE.

Что именно изменилось в vSphere 5-й версии:

- при перерасходе пространства тома с Thin Provisioning приостанавливаются только те виртуальные машины, которые требуют дополнительного места хранения, все остальные машины продолжают работать в стандартном режиме;
- новый примитив VAAI (использующий команду SCSI UNMAP) позволяет гипервизору сообщать системе хранения данных о том, что место, занимаемое виртуальной машиной, было освобождено. Однако использование данного примитива может оказать

влияние на производительность хранилища. В базе знаний VMware уже появилась информация о способах отключения данной функции;

- благодаря VAAI реализована система предупреждений vCenter, посредством которой пользователь общается о том, что том заполнен более чем на 75%, что позволяет администратору вовремя увеличить размер тома или выполнить миграцию виртуальных машин с тома и избежать перерасхода;
- усовершенствован примитив Atomic Test&Set (ATS).

Данный примитив используется для реализации подсистемы блокировок VMFS-хранилища, значительно превосходя по функциональности стандартный механизм SCSI Reservation. Одним из улучшений стало внедрение механизма ATS Retty, который позволяет отказаться от гипервизора SCSI Reservation в случае возникновения коллизий при использовании ATS.

Важная новость для администраторов, использующих файловые хранилища (NAS): в vSphere 5 реализован набор примитивов аппаратной конфигурации для сетевых СХД. Ранее API был доступен только для блочных устройств.

Были добавлены следующие примитивы:

- *Full File Clone* – аналог примитива аппаратной поддержки для блочных СХД "Full Copy". Позволяет выполнять копирование файлов виртуальных дисков средствами NAS;
- *Native Snapshot Support* – позволяет выполнять снапшоты виртуальных машин средствами СХД;
- *Extended Statistics* – обеспечивает прозрачность использования дискового пространства на NAS-хранилищах данных, особенно для Thin Provisioning;
- *Reserve Space* – позволяет создавать "толстые" виртуальные диски. Ранее при использовании сетевых хранилищ было возможно создавать только "тонкие" диски, что в свою очередь сказывалось на быстродействии дисковой подсистемы. При создании "толстого" диска пользователю предлагается на выбор два типа зачистки пространства: lazy-zero и eager-zero

Кроме того, теперь примитивы Full Copy, Block Zeroing и Hardware Assisted Locking совместимы со стандартом SCSI, который разрабатывается T10.

1.4 Storage DRS (SDRS)

SDRS является совершенно новой функцией в vSphere 5. SDRS позволяет выполнять интеллектуальное размещение и балансировку виртуальных машин, учитывая требования по дисковому пространству и производительности в рамках кластера датасторов.

Основным элементом SDRS является кластер хранилищ. Под кластером подразумевается папка датасторов. Кластер используется для агрегации ресурсов хранилищ и повышения эффективности их использования. Кластер может содержать хранилища VMFS3 и VMFS5, однако мы не рекомендуем использовать в одном

кластере хранилища с различными параметрами. Также теперь возможно включать в один кластер хранилища блочного и файлового доступа (VMFS и NFS).

Когда в кластере хранилища включена функция SDRS (включается по умолчанию при создании кластера), он превращается в домен балансировки нагрузки.

При создании миграции виртуальной машины есть возможность выбора кластера. SDRS при этом выбирает наиболее подходящее хранилище, учитывая имеющееся свободное пространство и задержку ввода-вывода.

Хранилище выбирается автоматически (если опция включена), однако Вам по-прежнему доступна возможность ручного выбора хранилища при создании или миграции машины.

Инженеры VMware рекомендуют использовать SDRS следующим образом: запустить кластер в ручном режиме и просмотреть рекомендации vSphere по размещению виртуальных машин. Если данные рекомендации совпадают с реальным положением дел на СХД, выполняйте переключение на автоматизированное управление.

При балансировке размещения машин на основе свободного пространства порог установлен в 80% (но при этом может быть изменен пользователем). При достижении данного порога SDRS попытается выполнить миграцию машин на другие хранилища в кластере с использованием Storage Vmotion. В ручном режиме администратору будут предложены рекомендации по перемещению, которые он сможет применить.

По-умолчанию SDRS будет выполнять миграцию машин на другое хранилище только в том случае, если разница в утилизации пространства хранилищ достигает минимум 5%. Конечно, и этот параметр можно изменить.

При балансировке виртуальных машин на основе метрик производительности SDRS использует информацию о времени отклика-латентности.

SDRS использует Storage I/O Control (SIOC) для определения параметров хранилища и собирает информацию о латентности операции ввода-вывода для всего кластера, создавая различную нагрузку.

SIOC во время простоя выполняет запросы на случайное чтение: сначала 1 запрос и получает время задержки, затем 3 параллельных запроса и так далее. На основе полученных данных создается профиль производительности хранилища.

SDRS постоянно использует SIOC для сбора информации о задержках. Если время отклика превышает определенный порог (15 мс по умолчанию) для значительного числа запросов за определенный промежуток времени, SDRS попытается сбалансировать машины внутри кластера либо предоставит администратору рекомендации по перемещению виртуальных машин.

В vSphere 5 был усовершенствован SIOC, который используется для контроля нагрузки на дисковую подсистему и предотвращения ситуации, когда машина с неприоритетным приложением исполь-

зует все ресурсы СХД. Кроме того, SIOC теперь поддерживает файловые СХД.

Для создания первых рекомендаций SDRS потребуется не менее 16 часов сбора информации. Проверка дисбаланса нагрузки выполняется каждые 8 часов.

Так как Storage I/O Control в vSphere поддерживается и для файловых хранилищ, SDRS может применяться и для NAS-СХД.

Многие администраторы ошибочно сравнивают SDRS с решениями автоматизированного многоуровневого хранения Auto-Taering (EMC FAST, AVORAIID+LSI cache cade, которые были представлены в прошлом номере), обнаруживая в SDRS такие недостатки как хранение всей машины на более быстром СХД, а не только горячих данных. Но данный подход является в корне неверным, так как SDRS разработан для балансировки нагрузки внутри кластера, а не для перемещения виртуальных машин по СХД различной производительности. VMware рекомендует использовать в кластере только тома с одинаковыми характеристиками.

При использовании SDRS с СХД, поддерживающими автоматизированное многоуровневое хранение, необходимо быть осторожным и следовать рекомендациям производителя системы хранения. При миграции процесс определения горячих и теплых данных будет нарушен и весь vmdk окажется на одном уровне хранения, что может привести к временному снижению производительности.

При тестировании Auto-Taering хранилища SIOC, скорее всего, будет неправильно составлен профиль производительности, так как невозможно предугадать, с какой из уровней будут сняты показатели задержек.

1.5 vSphere Storage Appliance

Одним из самых спорных нововведений компании VMware в этом году стало создание собственного storage Appliance VSA. Данный продукт ориентирован на рынок SMB, который не может приобрести SAN или NAS СХД.

Использование данного Appliance позволит клиенту использовать такие функции как vSphere HA и vMotion. Для создания кластера VSA требуется от 2-х серверов ESXi, не имеющих развернутых виртуальных машин. VSA экспортирует реплицированные тома через VPN всем хостам ESXi и обладает отказоустойчивостью. VSA разделяет локальное пространство на 2 части. При этом одна из частей используется как источник реплики, а другая – как зеркало для другого источника.

Таким образом, все ESX имеют общее хранилище.

Для управления VSA используется VSA Manager расширение vCenter Server.

Обзор функций VSA и сравнение его с конкурентами и продуктами-заменителями является темой одной из дальнейших статей.

1.6 VASA

Storage API for Storage Awareness (VASA) – это совершенно новая функциональность vSphere 5.

Функции перемещения данных между LUN и системами хранения данных Storage vMotion для обеспечения мобильности данных уже упоминались ранее. Благодаря использованию Distributed Resource Scheduler (DRS) vSphere позволила автоматизировать процедуру миграции данных, предоставляя vSphere самостоятельно решить необходимость переноса данных.

В vSphere 5 VASA очень прост. VASA – протокол, благодаря которому vSphere опрашивает LUN или NFS-папку через так называемые "провайдеры" о "характеристиках" (capabilities) СХД: дубликации, репликации, уровне массива, типах дисков, характеристиках производительности (MBps/IOps). Состав и формат ответов никак не регламентируется и остается на усмотрение вендоров. "Провайдеры" получают информацию о характеристиках СХД и передают ее vCenter, после чего она отображается в пользовательском интерфейсе.

В дальнейшем администраторы vSphere смогут использовать наборы характеристик для создания профилей СХД.

VASA позволяет администраторам vCenter полностью контролировать все элементы виртуальной инфраструктуры, а не обращаться за помощью к администраторам СХД, как это требовалось ранее.

Провайдеры VASA могут иметь несколько форм: плагин для vCenter или отдельно стоящее приложение на физическом или виртуальном сервере.

Объекты VASA Storage Capability состоят из имени характеристики и ассоциированного с ним описания. Например, с именем Performance (производительность) могут быть ассоциированы такие характеристики, как количество и тип дисков, IOps, MBps.

Определение характеристик СХД является не единственной функцией VASA. Также она используется для мониторинга состояния СХД и определения степени утилизации пространства. Кроме того, SDRS при перемещении машин использует подсказки VASA, например, предупреждение о том, что перемещение на том с Thin Provisioning может вызвать перерасход пространства.

1.7 Profile Driven Storage

При описании VASA мы упоминали, что на основе характеристик СХД администраторы смогут создавать профили СХД.

Profile Driven Storage – это функция, позволяющая корректно и просто выбирать хранилище для размещения виртуальной машины исходя из имеющихся характеристик. Данная функция используется для проверки соответствия требований к хранению виртуальной машины с реальными возможностями хранилища, на котором машина размещена.

На протяжении жизненного цикла виртуальной машины она может быть перемещена на СХД, которая наиболее полно удовлетворяет требованиям данной машины.

Перечень характеристик СХД можно объединять в профили.

Перед созданием профилей следует получить набор характеристик, автоматически связанных с хранилищами с помощью VASA, либо создать их вручную. Затем создаются профили, включающие в себя одну или несколько характеристик. Следующим шагом будет присвоение хранилищам определенных характеристик. Затем при создании машин или при изменении требований к ним задаются актуальные требования к профилю хранения данных машин.

Во время всего жизненного цикла виртуальной машины PDS проверяет соответствие требований виртуальной машины к текущему хранилищу.

Для корректной совместной работы DRS и Profile Driven Storage необходимо убедиться в том, что все хранилища в кластере имеют одинаковые характеристики.

Виртуальные машины, имеющие несколько vmdk, могут быть связаны с несколькими профилями.

1.8 All Path Down

В ESX 5 изменилось поведение системы при возникновении All Path Down (APD). Данная ситуация может произойти в том случае, если устройство было некорректно удалено или вышло из строя. При этом гипервизор не знает, когда последний раз устройство было доступно. В итоге остаются доступными пути к несуществующему устройству.

Состояние APD может быть как временным, так и постоянным. Ранее очередь IO к потерянному устройству сохранялась на неопределенный срок. В итоге при рескане SAN hostd, ожидающий ответов от устройства ответов, зависает и не доступен для таких сервисов как vCenter, который отвечает за коммуникацию с vCenter. В итоге мы получаем ситуацию, когда хост с гипервизором не доступен.

Hostd также может остановиться без пересканирования SAN из-за ограничений на количество рабочих потоков.

В предыдущих версиях ESXi для решения проблемы зависания при рескане SAN администраторы использовали ряд дополнительных параметров. Указанные параметры автоматически устанавливаются в процессе обновлений (1-го для ESXi 4.1 и 3-го для ESXi 4.0).

В ESXi 5.0 появился новый статус Permanent Device Loss (PDL), когда хост считает, что дисковое устройство уже никогда не будет снова подключено, а ситуация APD рассматривается как временный сбой, после которого хост снова увидит пути и устройство. Для устройств PDL все незавершенные запросы ввода-вывода не сохраняются в очереди, а обрываются незамедлительно со статусом VMK_PERM_DEV_LOSS. Следовательно, не будет повторена ситуация, когда hostd будет заблокирован незавершенными I/O.

Статус устройства APD или PDL определяется с помощью SCSI Sense кодов. Существует перечень Sense кодов, которые однозначно указывают на то, что является PDL. В данном случае гипервизор полагается на массив, которому должно быть точно известно, о том, что логический том был удален безвозвратно.

Для того чтобы том был переведен в состояние PDL, все пути должны иметь соответствующий статус.

Со стороны виртуальных машин разница между PDL и APD не видна.

1.9 Поддержка протоколов

В vSphere 5 появились новшества, связанные с протоколами доступа к данным:

- значительно улучшен интерфейс управления iSCSI. Управление доступом к хранилищам с самым популярным интерфейсом было значительно улучшено, что на наш взгляд, может послужить дополнительным плюсом при выборе гипервизора;
- усовершенствован iSCSI Multipathing;
- количество общих томов NFSv3 увеличено до 256. Расширение поддержки устаревшего протокола вряд ли будет интересно для большинства пользователей;
- поддержка SATA 3. SATA 3 разъемы имеются практически во всех современных серверах с процессорами Intel Sandy Bridge. Добавление поддержки нового стандарта не стало неожиданностью;
- поддержка программного FCoE-инициатора. На наш взгляд, данное нововведение является основным с точки зрения поддержки интерфейсов. Транспорт набирает популярность и активно поддерживается вендорами из top 10.

1.10 Обнаружение SSD

Теперь гипервизор автоматически обнаруживает твердотельные накопители. Данная информация важна для корректной работы новой функциональности SWAP to Host cache, которая позволяет сохранять страницы виртуальной памяти на твердотельных носителях.

1.11 vSphere Replication

Новой функцией Site Recovery Manager 5.0 является поддержка Host Based Replication.

Поддерживается только асинхронная репликация, и она значительно уступает по производительности традиционной репликации на уровне СХД. Важным отличием является возможность репликации машин между массивами различных производителей и даже типов. Поэтому выполнять репликацию с FC СХД на NAS ничто не помешает.

Заключение

VMware выбрал эволюционный путь развития. Были улучшены технологии, полюбившиеся пользователям, исправлены ошибки, добавлена поддержка нового аппаратного обеспечения и протоколов. Несмотря на все старания маркетологов VMware, революционных изменений в vSphere мы не обнаружили. Новые функции гипервизора основаны на уже имеющихся и являются, в какой то мере, "пробой пера", из-за чего получили свою порцию критики пользователей. VMware нацеливается на более крупные внедрения, в связи с чем увеличивает масштабируемость решения и значительно упрощает управление виртуальной инфраструктурой.

*Сергей Платонов,
компания AVORAIID*