Dynamic Storage Tiering – уровневое хранение от Symantec

В мае с.г. компания Symantec представила версию 5.0 Storage Foundation – значительно расширенную и дополненную новыми компонентами. Одно из наиболее заметных расширений – продукт Dynamic Storage Tiering (DST – в более ранних версиях продвигался как Quality of Storage Services), предназначенный для построения уровневых решений хранения на основе VxFS файловой системы и совершенно прозрачный для пользователей, приложений, БД, а также политик резервного копирования/восстановления.

Введение

Каждый из вендоров продвигает собственную концепцию динамического уровневого управления хранением информацией/данными на основе политик/правил, или уже в более распространенной терминологии – ILM-концепцию. SN уже достаточно много писал о ILM (Information Lifecycle Management – управление жизненным циклом информации). В самом широком толковании ILM это не только управление информацией/данными, но и всем, что с этим связано, включая приложения, серверы и др. В представлении Symantec это более ограниченный спектр решений на основе DST, дающих возможность динамически управляемого с использованием политик многоуровневого хранения файлов/блоков данных на базе свойств многотомной файловой системы Veritas – VxFS.

Одни из основных преимуществ при реализации решений Symantec по уровневому хранению это полная их прозрачность для пользователей/приложений/БД/политик резервного копирования&восстановления, а также универсальность самой технологии, которую можно использовать как с файловыми системами, базами данных, так и для решения различных прикладных задач с целью приведения в соответствие затрат на хранение данных с их бизнес-требованиями.

Необходимо отметить и то, что хотя последнее обстоятельство (выравнивание стоимости хранения в соответствии с ценностью данных) и является основным позиционированием уровневого хранения, нельзя забывать о том, что оптимизация хранения становится намного значимей в решениях, в которых онлайновые данные в целях повышения доступности реплицируются. В этих случаях экономия от уровневого хранения может многократно увеличиваться, а в тех случаях, когда решение строится на удаленном реплицировании, поддержание заданного объема тома/файла может оказаться решающим фактором обеспечения доступности.

Уровневое хранение и основные технологии его поддержания

В общем случае IT-инфраструктура может содержать очень большое количество уровней хранения, которые, например, различаются по:

1) *степени защищенности данных.* Уровни строятся на основе: сложной системы защиты на базе показателей RPO&RTO/зеркалированных томов/защиты на уровне RAID/защиты только средствами самого HDD;
2) *уровню производительности.* Уровни различаются производительностью: высокой/средней/низкой для случайного доступа; высокой/средней/низкой – для потокового доступа/для смешанного доступа и др.;
3) *уровню защиты данных от изменений.* Уровни могут: допускать изменения данных/не допускать изменения/представлять смешанный тип двух первых;
4) *способу доступа к данным.* Уровни могут отличаться по способу доступа к данным: файловый/блоковый или, например, типу подключения: DAS/NAS/SAN;
5) *сроку хранения* – 1 день/1 месяц/1 год/10 лет/вечно;
6) *типу системы хранения и физическому типу доступа* – флэш/HDD/лента/MO/CD/DVD и прямой/последовательный/смешанный.

Кроме этого, можно добавить еще ряд признаков и многочисленные смешанные варианты. Также непреложным фактом является изменяющаяся ценность данных в течение их жизненного цикла, а также их возрастающий объем.

Чтобы управлять всем этим многообразием, прежде всего, с целью повышения эффективности использования/доступ-

ности и др. в общем случае (для систем, управляемых только на основе общедоступных программных средств; в случае использования специализированных аппаратных/программных средств число возможных технологий значительно шире) используются 2 основных механизма управления иерархией хранения: *специализированный* (для конкретного приложения/применения) и *общий* (или HSM). В *первом* случае создаются множественные файловые системы, например, на различных типах запоминающих устройств, между которыми и организуется миграция файлов в соответствии с бизнес-требованиями. Во *втором* случае – при перемещении файла остается ссылка на местоположение файла и при обращении к нему он возвращается на первоначальное местонахождение.
К преимуществам первого механизма относятся:

- *точность:* при "заказных" сценариях, которые выполняются периодически, организация может выполнить фактически любое требование для управления перемещением файла. Например, если это последовательность команд ОС, она может выполняться скриптом, который запускается автоматически от авторизованных аккаунтов;
- *оперативность использования:* приложения и утилиты обращаются к перемещенным файлам непосредственно – без скрытых накладных расходов (как в случае с HSM). Воздействие на приложения (с учетом затрат на обработку сценария и доступной полосы пропускания) может быть смягчено за счет управления временем выполнения сценария.

К недостаткам специализированных скриптов относятся:

- *необходимость административных усилий* – каждая политика требует индивидуального создания и поддержания. Простое изменение политики, например, перемещения неактивных транзакций после 45, а не 30 дней, может затрагивать как саму БД, так и процедуры, приложения и сценарии, которые перемещают и оперируют перемещением данных, что требует каждый раз полной верификации;
- *процедурная сложность* – когда файлы перемещены на альтернативные файловые системы, они больше не доступны приложениям и процессам управления, разработанным для работы с первоначальным пространством имен. Приложения и операционные процедуры должны быть изменены, чтобы использовать перемещенные файлы, а модификации – проверены после каждого изменения политики.

Технологии на основе HSM не имеют ни одного из недостатков специализированных скриптов. Политики HSM стандартизованы (например, переместить файлы, к которым не обращались в течение "x" дней, на уровень "y"), и сканеры файловой системы иерархической семантической модели, выполненные в правильных списках (графиках). Ссылки для перемещенных файлов остаются в метаданных файловой системы, так что приложения и операционные процедуры могут в основном работать без модификации, если они не чувствительны ко времени доступа файла. Программные менеджеры резервного копирования обычно "понимают" HSM-архитектуру и перемещение файлов не вызывает неправильных или избыточных действий. Однако ряд свойств HSM-технологий ограничивает их применение:

- *задержка по доступу к перемещенному файлу* – при обращении пользователя или приложения к перенесенному файлу он должен быть переслан обратно в пространство имен первичной файловой системы, прежде чем запрос может быть удовлетворен;
- *негибкие политики* – ограниченность политик, используемых в HSM-технологиях, условия в которых формируются на фиксированном количестве признаков;
- *необходимость резервирования пространства для перемещения файлов.*

Создание иерархических систем хранения на основе файловой системы Veritas и свойств Dynamic Storage Tiering позволяет во многом использовать преимущества двух рассмотренных выше подходов, устраняя при этом свойственные им недостатки.

Построение иерархических систем хранения на основе файловой системы VxFS и DST

Чтобы правильно понимать преимущества и позиционирование подобных решений, необходимо представлять базовую функциональность, которая обеспечивает их реализацию.

Во-первых, данный класс решений ориентирован в основном на дисковое онлайновое многоуровневое хранение. *Во-вторых*, их основу составляет многотомная файловая система – VxFS, которая, представляя единое пространство имен для хранения файлов, может занимать 2 и более виртуальных тома (VxVM), идентифицируемых с разными уровнями хранения и полностью прозрачных для пользователей и приложений. При этом необходимо учитывать то, что DST может работать только с файлами, но не с блоками данных, и при оптимизации работы СУБД операции осуществляются только с целыми файлами.

Сделаем небольшое отступление в терминологию DST.

Тома, на которых создается файловая система, называются набором томов (volume set). Тома в наборе томов конфигурируются из дисков или LUN'ов на дисковых массивах, которые принадлежат одной Veritas Volume Manager (VxVM) дисковой группе. Тома могут иметь различные типы (например, со стрипованием/на основе RAID-5/с зеркалированием и т.д) и поддерживать различные протоколы – FC, SATA, параллельный SCSI JBOD и т.д. Все файлы в файловой системе – часть одного и того же пространства имен, процедура обращения к которым и управления которыми такая, как если бы они все размещались на одном томе.

Администраторы многотомных VxFS файловых систем могут управлять местоположением файлов в пределах наборов томов, определяя политики размещения файла (которые управляют как первичным размещением файлов, так и условиями, при которых файлы перемещаются между томами). VxFS-политики размещения файлов состоят из правил, которые управляют местоположением файлов, определенным администратором, в подмножестве набора томов файловой системы. Эти подмножества называют классами размещения. Класс размещения обычно идентифицируется с уровнем хранения (рис. 1).

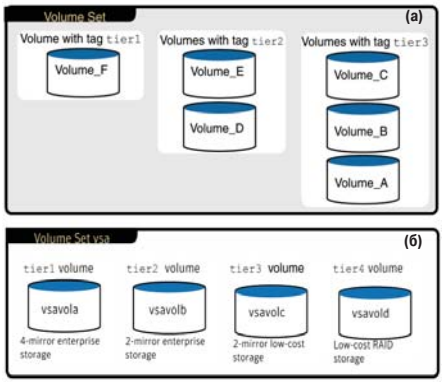

Рис. 1. Два примера организации системы томов:
(а) – система состоит из трех классов размещения: tier1, tier2 и tier3. Соответственно на уровне tier1 размещаются несколько критичных файлов, на уровне tier2 – большое число файлов средней важности, на уровне tier3 – большое число неиспользуемых файлов;
(б) – классы классифицируются по степени защищенности данных и соответственно стоимости.

Администраторы записывают политики размещения файлов на XML-языке, согласно Document Type Description (DTD), поставляемому вместе с VxFS. Консоль управления Storage Foundation включает мастера, которые помогают создавать 4 наиболее популярных типа политик в соответствии с задаваемыми пользователем параметрами. Политики размещения файлов жестко не связаны с определенными файловыми системами. Администратор назначает политику на файловую систему, делая ее активной политикой для данной файловой системы. Файловая система может иметь только одну активную политику одномоментно.

В версии 5.0 Storage Foundation для DST сделано значительное расширение, в частности, введена поддержка баз данных: DB2, Oracle и Sybase. Также, кроме уже отмеченных, можно выделить еще ряд преимуществ, которые получает пользователь при использовании многотомной файловой системы в сравнении с традиционными:

- возможность использования статистики по файлам/директориям при организации уровневых хранилищ;

– возможность визуализации собранной статистики;

– возможность использования специальных (но стандартных) опций, например, для балансировки нагрузки.

К преимуществам DST и Storage Foundation в целом следует отнести и то, что они работают на Red Hat, SuSe, Solaris, HP-UX, AIX одинаково. Обеспечивая универсальную технологию управления информацией независимо от ОС и моделей дисковых массивов.

Примеры построения систем хранения на основе DST

Пример 1. Управление уровневым хранением на основе показателей активности использования файлов.

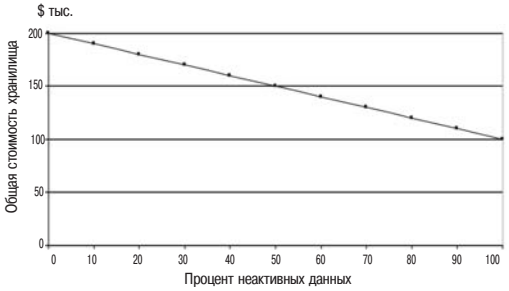

Рис. 2. Изменение стоимости двухуровнего 10 Тбайт хранилища в зависимости от % неактивных данных (при "правильном" хранении).

В большинстве систем бóльшая часть файлов или не используется, или интенсивность обращения к ним крайне низка. Вследствие этого, размещение всех данных на одном уровне приводит к неоправданным издержкам. Так, если при общем размере хранилища 10 Тбайт, состоящем из двух уровней хранения (с единичной стоимостью хранения $20/Гбайт и $10/Гбайт, соответственно, для 1-го и 2-го уровня), то при "правильном" хранении стоимость хранилища будет варьироваться от $20 тыс. до $10 тыс. в зависимости от процента неактивных данных (рис. 2).

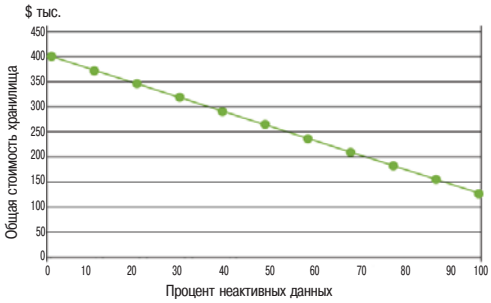

Рис. 3. Изменение стоимости 10 Тбайт двухуровнего (со 100% и 25% избыточностью 1-го и 2-го уровней соответственно) хранилища в зависимости от % неактивных данных (при "правильном" хранении) в сравн. с рис. 1.

В реальных системах разные уровни хранения отличаются разной степенью защищенности/доступности данных. Так, к примеру, 1-й уровень обычно зеркалируется (100%-ное добавление объема хранения), а для 2-го уровня можно использовать и RAID-5 (25%-ная избыточность). В результате линейная зависимость (см. рис. 2) получается гораздо более "крутой" (рис. 3). При расширении этого решения до уровня катастрофоустойчивого оптимизация хранения становится насущной проблемой не только из-за фактора стоимости хранения, но также из-за стоимости передачи избыточных данных.

Как уже было сказано, средствами DST удается избежать всех недостатков HSM-систем, строя минимальными усилиями самые гибкие политики перемещения файлов (с недостаточной активностью) с использованием одних механизмов для сбора статистики и перемещения.

Пример 2. Упрощение масштабируемости томов и балансировки нагрузки в решениях для СУБД ORACLE на основе опции DST – LBFS.

В "классических" традиционных системах для повышения производительности при записи/считывании ORACLE-данных используется метод стрипования (деления) файлов, на части. Это решается на контроллерах (организацией дисков по типу RAID-0) или средствами ORACLE, путем распределения тома между несколькими физическими дисками. По мере роста данных требуется добавление новых дисков. Если не используются специальные средства, процедура реорганизации RAID/тома (на большее число дисков) занимает и много времени, и много сил. С помощью опции Load Balanced File System (LBFS), поставляемой в составе Veritas Storage Foundation for Oracle, эта проблема решается достаточно просто.

LBFS создается на многотомной файловой системе (на множестве томов), в которой каждый том "привязан" к отдельной физической дисковой группе (а отдельные тома не стрипуются на разные физические диски, повышение доступности может обеспечиваться, например за счет зеркалирования). Файловая система на LBFS имеет специальную политику размещения, названную "политикой балансировки", при которой все файлы делятся на "куски" ("chunks", по умолчанию – 1 Мбайт), которые, в свою очередь, размещаются на отдельных томах. Опция LBFS предоставляет такие же возможности по стрипованию файлов, но значительно упрощает масштабирование.

Пример 3. Оптимизация хранения вспомогательных файлов СУБД ORACLE.

Для поддержания высокой доступности производственных БД ORACLE используются файлы регистрации (logs, примерами таких файлов могут служить файлы, создаваемые на основе Veritas Storage Foundation for Oracle Flashback technology), которые хранят все изменения в БД, произошедшие за определенный период времени. В случае возникновения сбоев в системе или выявления ее неправильной работы, восстановление БД производится на основании файлов регистрации. Обычно, как в целях поддержания высокой скорости записи logs, так и для быстрого восстановления БД файлы регистрации пишут на те же устройства хранения, на которых хранится и сама БД. С течением времени вероятность обращения к более ранним версиям файлов регистрации уменьшается, поэтому их целесообразно перемещать на устройства с меньшей стоимостью хранения.

Например, для OLTP Oracle БД с тысячами активных сессий, которая должна быть доступна в режиме 24x7 с вероятностью 99% (требуемое время восстановления – 15 мин.) для хранения logs может быть организовано 3 уровня. Первоначально создаваемые файлы регистрации в течение 7 дней хранятся на тех же высокопроизводительных устройствах, что и производственная БД (например, EMC Symmetrix), далее они перемещаются на системы хранения среднего класса (EMC Clarion, medium-том) и по истечении еще 7 дней – на медленные JBOD диски (old-том), а после 30 дней с момента создания удаляются.

Следующее правило перемещает файлы из каталога Flashback, к которым не было обращений в течение 2-х дней, на medium-том:

 # /opt/VRTS/bin/dbdst_file_move -S PROD -o flashback -c MEDIUM:2

Второе правило перемещает в архив файлы с medium-тома, к которым не было обращений в течение 7 дней, и файлы с old-тома, к которым не обращались в течение 15 дней:

 # /opt./VRTS/bin/dbdst_ file_move -S PROD -o archivel -c MEDIUM:7 -c OLD:15

Добавляя к этим правилам несколько установочных процедур, можно организовать решение, которое при минимуме усилий позволяет (не затрагивая производственную БД) своевременно и в соответствии с текущей активностью и бизнес-требованиями "очищать" дорогостоящее хранилище от вспомогательных файлов.

Пример 4. Сезонная оптимизация хранения БД.

Данное решение может представлять интерес для торговых организаций, когда вся несезонная номенклатура товаров (файлы БД) перемещается на менее высокопроизводительный уровень хранения, сохраняя при этом всю функциональность для редких покупателей. При этом перемещение файлов несезонных товаров можно производить по мере снижения активности обращения к ним (без привязки к месяцам). Потребность в подобных решениях резко возрастает при наличии, например, удаленной репликации основного хранилища (не только для поддержания доступности, а например, в целях поддержания второго удаленного офиса).

Пример 5. Оптимизация хранения внешних файлов БД.

Многие базы данных содержат метаданные о большом числе внешних файлов, типа изображений, документов, экспериментальных данных и т.д. Поскольку внешние файлы имеют тенденцию быть большими, в ряде случаев может быть выгодно перемещать неактивные файлы на более низкие уровни хранения. Правило, представленное ниже, с определен-

ной периодичностью просматривает директорию external_file_dirs и перемещает файлы, к которым не было обращений в течение 2-х дней на том medium, а к которым не обращались в течение 30 дней на том old:

 # dbdst_file_move -S proddb -o external -f external_file_dirs -c MEDIUM:7 -c OLD:30

Это решение может использоваться с Oracle и DB2 базами данных.

Производительность DST

Вопрос о производительности возникает при выполнении процедур сканирования файлов на предмет выполнения каких-либо условий или сбора статистики. Как показали исследования, время сканирования напрямую зависит от числа файлов и процентного отношения размещения информации о файле в кэше при выполнении сканирования. Тестирование файловой системы с 10 млн файлов на сервере с 32 Гбайт ОП (Sun 880, 32GB, Fibre Storage), из которой 6 Гбайт использовались буферным кэшем, показало, что при неиспользовании буферного кэша просмотр занял около 27 мин. После того, как буферный кэш полностью заполнялся, время просмотра уменьшалось в 3 раза и составляло 9 мин.

Заключение

Введение Dynamic Storage Tiering с расширенной функциональностью в состав Storage Foundation 5.0 сделало возможным построение очень гибких, доступных и в то же время простых при создании решений для оптимизации хранения и управления данными, ценность которых будет возрастать при увеличении требований к IT-системам, прежде всего с точки зрения их надежности и доступности.

Осенью с.г. (как анонсировала Symantec) станет доступным инструментарий для оценки эффективности использования DST в различных применениях, что еще больше упростит его внедрение в различные секторы бизнеса и IT-инфраструктуры.

Seagate объявила о 10 новых продуктах

Июнь 2006 г. Компания Seagate анонсировала о доступности 9 своих новых дисковых накопителей и одной системы.

Хотя корпоративный рынок по-прежнему остается основным заказчиком решений для хранения данных в мире, по прогнозам, к 2009 г. структура общемирового рынка дисковых накопителей заметно изменится. Его объем в финансовом выражении составит $45 млрд (или в количественном – около 700 млн шт.). А заметную долю в нем займет один из самых быстрорастущих секторов – рынок HDD для бытовой электроники, доля которого в общемировом обороте составит около $14,3 млрд. Соответственно этим прогнозам – и распределение усилий Seagate по секторам. Среди объявлений – 5 продуктов для сектора бытовой электроники, 2 – для enterprise рынка и 3 – для использования в составе ноутбуков.

Для корпоративных нужд в дополнение к недавно анонсированной серии Cheetah 15K.5 компания выпустила еще 2 диска: Savvio 10K.2 и Barracuda ES:

Enterprise Storage	RPM	Макс.емк.	Best Fit	Дост-ть
Savvio 10K.2 (SAS, FC)	10000	147 GB	High I/O density Servers, Blades	CQ2 '06
Barracuda ES (SATA)	7200	750 GB	High Capacity Server/ Arrays	CQ2 '06
Cheetah 15K.5 (Ultra320 SCSI, SAS, FC)	15000	300 GB	High-performance Enterprise Servers/Arrays	CQ2 '06

Как сказано в пресс-релизе, *Savvio 10K.2 – самый надежный корпоративный диск в мире, а также и самый "прохладный".*

Второе поколение дисков Savvio 10K.2 размером 2,5" обладает высокими показателями I/O: рекордным показателем среднего времени безотказной работы (1,6 млн часов) и объемом до 146 Гбайт. Новый диск, имея размер на 70% меньше по сравнению с 3,5" и потребляя электроэнергии на 15% меньше по сравнению с предыдущим поколением, обеспечивает беспроблемное интегрирование большего количества дисков в корпуса меньшего размера. Savvio 10K.2 также позиционируется как решение для создания новых типов накопительных систем, включая blade серверы, 1U серверы с уровнем RAID-5 и большие серверы-стойки с максимально большой плотностью работы. По оценкам IDC, ожидается, что совокупные темпы годового роста рынка жестких дисков корпоративного класса с маленьким форм-фактором достигнут 110% в 2005–2010 гг. Диски поставляются с интерфейсом 3Gb Serial Attached SCSI (SAS) и 4Gb Fibre Channel.

Barracuda ES был представлен как самый вместительный корпоративный диск в мире.

Barracuda ES 7200RPM намного повысил свою надежность и способен сохранять продуктивность при работе в близкостоящих мультидисковых системах благодаря технологии Rotational Vibration Feed Forward (RVFF). Разработанная таким образом, чтобы без труда добиваться лучших показателей за среднее время безотказной работы (Mean Time Between Failure (MTBF), новинка семейства Barracuda также обладает технологией Workload Management, которая защищает диск от перегрева, позволяя добиться особой надежности и долговечности дисков этого семейства.

Barracuda ES оснащен интерфейсом 3Gb Serial ATA (SATA), который включает организацию очереди команд (NCQ) и 8/16 Мбайт кэш. С данным интефейсом Barracuda ES доступен с объемом в 250, 400, 500 и 750 Гбайт. В комплектации с 2Gb FC – емкость 400 и 500 Гбайт.

В потребительском классе теперь доступны еще 3 продукта: карманный накопитель Seagate 8.0 GB Pocket Drive, внешний eSATA 500 Гбайт диск и 750 Гбайт диск Pushbutton. Pocket Drive легко умещается в кармане джинсов или на ладони и является удобным переносным носителем. Внешний диск eSATA External Hard Drives на 500 и 300 Гбайт предлагает возможность внешнего хранения с интерфейсом передачи данных до 300 Мбайт/с (или 3 Гбайт/с), что приблизительно в 5 раз быстрее того, что предлагали традиционные интерфейсы, такие как USB 2.0 и 1394a.

Карманный накопитель Seagate 8.0GB Pocket Drive в продаже с июня 2006 г. по розничной цене $149, внешний диск Seagate 750GB Pushbutton Hard Drive – по розничной цене $559. Внешние диски Seagate 300GB и 500GB eSATA External Hard Drives – в рознице по $269 и $399 соответственно.

Среди других новинок – ST18 *(дост. – 1 кв. 2007 г.)* – первый 1,8" диск, разработанный для карманных видео- и аудиоустройств; серия жестких дисков следующего поколения LD25 *(дост. – 1 кв. 2007 г.)* с объемом до 80 Гбайт, на одной пластине, специально разработанных для растущего сегмента новых технологических устройств (консоли для видеоигр, устройства для домашнего развлечения, мультимедийные компьютеры footprint, устройства set-top, телевизоры со встроенным цифровым видеорекодером – DVR и др. техника, используемая дома); серия DB35 – новейшие диски для цифровых видеорекодеров с объемом 750 Гбайт для записи развлекательных передач и фильмов, а также 3 новых диска для ноутбуков *(дост. – 1 кв. 2007 г.)*.

Диски серии ST18, использующие технологию перпендикулярной записи, при размере 1,8" представляют самый большой объем на одной пластине – 60 Гбайт. Эта серия позиционируется для портативных медиапроигрывателей, портативных навигационных систем (GPS), цифровых видеокамер и КПК и включает технологии:

- *G-Force Protection* защищает жесткий диск от удара или неаккуратного обращения, что повышает степень надежности диска до 1500 G, делая диск серии ST18 самым надежным 1,8" диском на рынке;
- *Run-On* позволяет пользоваться устройством с новым диском даже во время бега;
- *управления электропитанием,* которая позволяет продлить работу батареек в карманных устройствах.

Диски ST18 серии совместимы с новым интерфейсом CE-ATA для переносных устройств. Кроме того, серия ST18 поддерживает также интерфейс IDE, используемый во многих современных переносных устройствах.