ИПМ им.М.В.Келдыша РАН • Электронная библиотека

Препринты ИПМ • Препринт № 14 за 2008 г.

Горбунов-Посадов М.М.,
Корягин А.Н.

Современные системы хранения данных старшего класса

Рекомендуемая форма библиографической ссылки: Горбунов-Посадов М.М., Корягин А.Н. Современные системы хранения данных старшего класса // Препринты ИПМ им. М.В.Келдыша. 2008. № 14. 16 с. URL: http://library.keldysh.ru/preprint.asp?id=2008-14

Ордена Ленина
ИНСТИТУТ ПРИКЛАДНОЙ МАТЕМАТИКИ
имени М.В.Келдыша
Российской академии наук

М.М.Горбунов-Посадов, А.Н.Корягин

СОВРЕМЕННЫЕ СИСТЕМЫ ХРАНЕНИЯ ДАННЫХ СТАРШЕГО КЛАССА

Москва 2008

Рассматриваются требования, предъявляемые к современным системам хранения данных старшего класса и характерные архитектурные решения, отвечающие этим требованиям.

High-End Class of Modern Data Storage Systems

The requirements for high-end modern data storage systems are considered. The typical principles to satisfy these requirements are analyzed.

Оглавление

1. Требования к современным системам

К современным системам хранения данных старшего класса предъявляются следующие требования.

- Надежность
- Запас производительности
- Экономическая эффективность:
  - многоуровневое хранилище,
  - масштабируемость, гибкость конфигурирования,
  - простота и эффективность управления
- Локальные и удаленные репликации

В настоящее время системы хранения старшего класса предлагают три основных производителя: EMC, Hitachi и IBM (распределение рынка между ними представлено на рис. 1). На примере продукции этих производителей проиллюстрируем, в какой мере реализуются перечисленные выше требования.

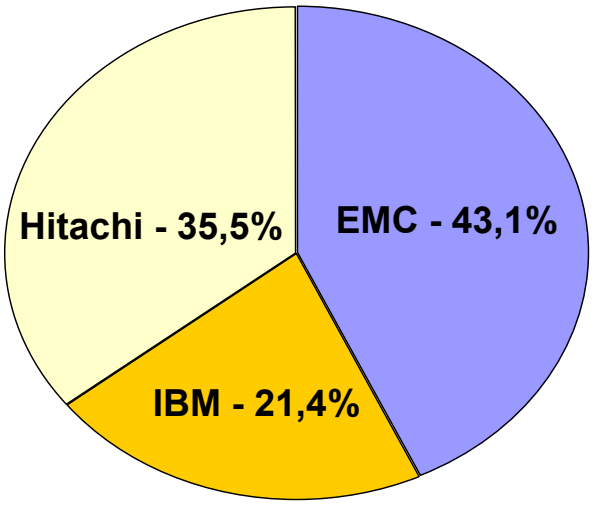

Рис. 1. Распределение долей рынка между производителями систем хранения данных старшего класса

2. Надежность

В основе систем старшего класса, как правило, лежит архитектура MPP (Massive Parallel Processing). Это означает, что на всех уровнях системы хранения, будь то front-end, cache или back end, параллельно работает достаточное количество взаимозаменяемых однородных компонентов. В случае отказа любого из таких компонентов нагрузка перераспределяется между оставшимися взаимозаменяемыми

однородными компонентами, и таким образом существенного снижения производительности не происходит. Именно отсутствие компонентов, выход из строя одного из которых приводит к отказу или же к катастрофическому падению производительности, отличает системы хранения старшего класса.

Наряду с взаимозаменяемыми компонентами обеспечению надежности системы хранения служат следующие механизмы.

- Быстрый сброс cache памяти в случае отказа внешнего электропитания (Cache Vault)
- Глобальные диски горячей замены (Global Hot Spare Drives)
- Упреждающее резервирование дисков (Proactive Hot Sparing)
- Изоляция диска (Ultra Point) в случае сбоя диска
- Постоянная проверка целостности данных в фоновом режиме
- Возможность переноса томов данных внутри системы без остановки приложений (Virtual LUN)
- Использование технологии RAID6, обеспечивающей защиту от одновременного выхода из строя двух дисков

3. Запас производительности

Современная система хранения данных старшего класса должна, разумеется, быть не просто масштабируемой, а иметь весьма внушительные возможности для наращивания производительности. Вот, например, каких предельных значений может достигать система Symmetrix DMX4 (Symmetrix DMX4 занимает лидирующие позиции среди систем рассматриваемого класса, поэтому здесь и далее в примерах, как правило, приводятся именно ее параметры).

- Максимальное число дисков: 2400
- Объем cache памяти: 512 Гб - полный объем, 256 Гб - полезный объем
- Внутренняя пропускная способность: 128 Гб/с

Помимо этого, для эффективной совместной работы нескольких приложений предоставляются специальные средства управления распределением cache памяти и приоритетами выполнения.

Особая забота производителей — эффективный cache. Как правило, удается добиться положения, когда попадания в cache составляют 80% и более.

Произвольное количество физических дисков может быть представлено как мета-том, что решает проблему ограниченной производительности и емкости отдельного тома данных.

4. Экономическая эффективность

Многоуровневое хранилище. Применение иерархического хранилища существенно снижает общую стоимость владения системой хранения данных, каждый тип данных получает адекватный уровень обслуживания. При этом процессорные ресурсы и ресурсы cache памяти можно динамически распределять между уровнями хранения.

Масштабируемость, гибкость конфигурирования. Как диски, так и память и ресурсы ввода-вывода должны иметь возможность наращиваться независимо друг от друга. Все виды модернизации системы должны, как правило, происходить без прерывания работы приложений.

Простота и эффективность управления. Функции администратора должны быть сведены к минимуму. Физическое размещение данных и оптимизация производительности практически полностью перекладываются на систему хранения данных.

5. Локальные и удаленные репликации

В последнее время все чаще применяются катастрофоустойчивые решения с репликацией данных между двумя или тремя площадками средствами системы хранения данных. Здесь к системе хранения предъявляются следующие требования.

- Интеграция с распределенными кластерами
- Поддержка механизмов резервирования без остановки приложений
- Немедленная доступность данных при переключении на аппаратуру резервной площадки и обратно
- Поддержка групп консистентности, т. е. объединений нескольких томов, содержащих связанные данные. Если по какой-либо причине репликация одного из томов прекращается, то прекращается и запись во все остальные тома данной группы консистентности на удалённой площадке
- Немедленное восстановление производственных данных из полных и логических реплик данных
- Инкрементальный режим обновления реплик и восстановления из реплик
- Одновременное применение локальной и удаленной репликации

- Полный контроль за производительностью системы, размещенной на нескольких площадках
- Перемещение данных между различными системами хранения данных

Наиболее убедительным представляется симметричное решение, поддерживающее репликацию данных между тремя площадками, причем при выходе из строя аппаратуры одной из площадок производится автоматическая инкрементальная синхронизация между оставшимися двумя.

6. Архитектура системы хранения старшего класса

В качестве характерного примера архитектуры системы хранения старшего класса вновь рассмотрим систему EMC Symmetrix DMX4. В ней можно выделить три уровня (рис. 2):

- процессоры ввода-вывода,
- коммутаторы,
- cache память.

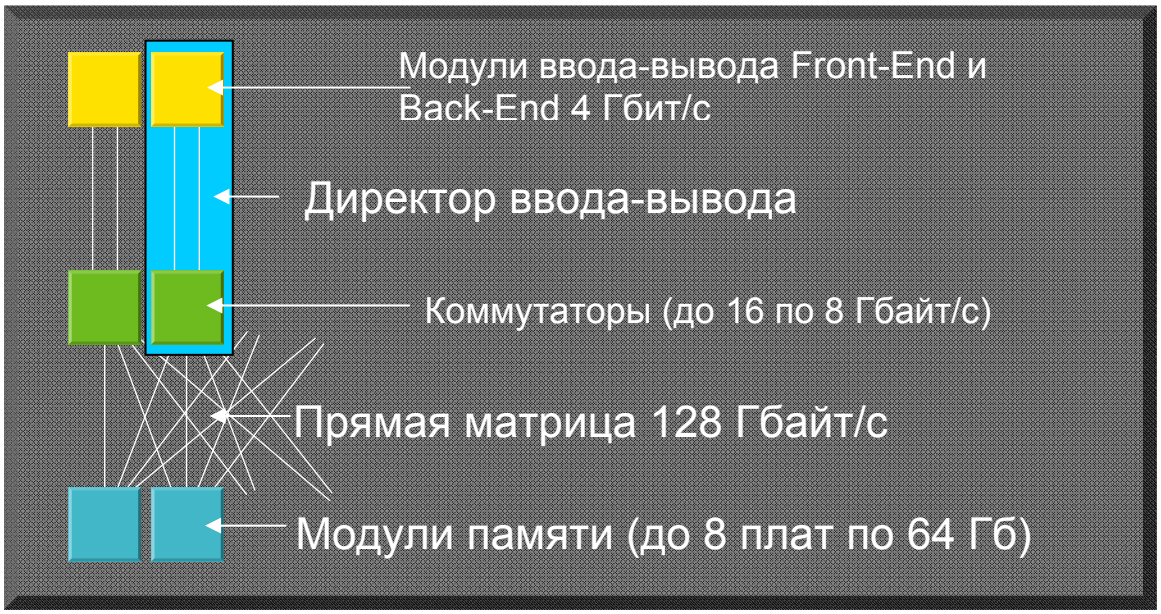

Рис. 2. Архитектура системы хранения EMC Symmetrix DMX4

Как уже упоминалось, на всех уровнях системы параллельно работает достаточное количество однородных взаимозаменяемых компонентов. В случае отказа любого из таких компонентов производится перераспределение нагрузки между оставшимися компонентами, и таким образом существенного снижения производительности не происходит.

Модули ввода-вывода связаны с коммутаторами, которые в свою очередь связаны с платами памяти по схеме каждый с каждым.

Cache память является ключевым элементом данной архитектуры. Устройство cache памяти рассматривается в следующем разделе.

В системе 8 портов ввода-вывода, 8 процессоров PowerPC 1,3Ггц и коммутатор объединены в едином конструктиве и образуют элемент, называемый директором ввода-вывода. Часть директоров ввода-вывода отвечает за Front-End – обмен данными с подключенными к системе серверами. Остальные директора выделены под Back-End и отвечают за управление дисковой подсистемой.

Благодаря высокому уровню параллелизма обеспечивается высокая производительность и масштабируемость системы. В максимальной конфигурации система включает 130 процессоров PowerPC 1,3 Ггц, соединенные через прямую матрицу с 512 Гбайт cache памяти, расположенными на 8 платах. Максимальное количество дисков в системе составляет 2400.

7. Организация cache памяти

Как уже было сказано, физически cache память в системе в системе Symmetrix DMX располагается на специальных платах памяти - директорах. Помимо этого, в пределах одного директора память разделена на четыре независимых, параллельно работающих области, соединенные с основными каналами матрицы также по принципу каждый с каждым (рис. 3).

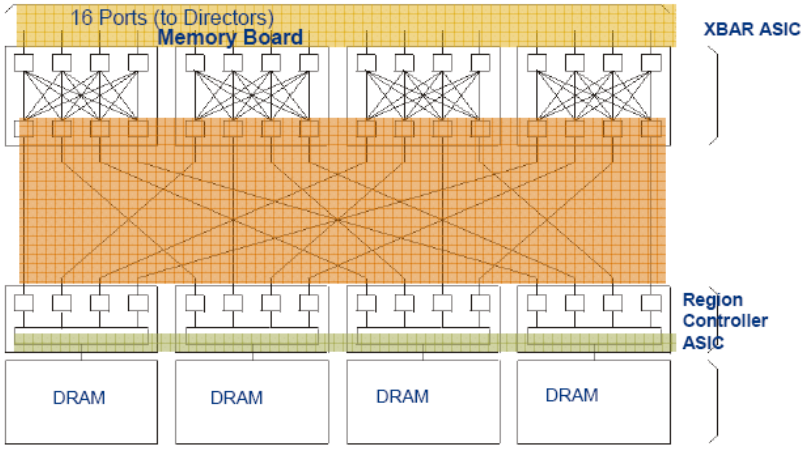

Рис. 3. Физическая организация памяти в пределах одного директора

Благодаря архитектуре прямой матрицы каждый процессор на каждом директоре ввода-вывода может получить доступ к любой области

cache памяти. Далее, уже логически, на уровне ОС память разбита на множество независимых областей, называемых тегами (tag). Все это вместе позволяет добиться такого высокого уровня параллельности выполнения операций с памятью как 32 независимые очереди запросов на каждый гигабайт памяти.

В системе Symmetrix DMX cache память является глобальной, т. е. разбиение на области чтения и записи отсутствует. В зависимости от нагрузки в предельном случае 100% памяти может быть выделено на чтение и до 80% на запись.

Принципиально важным является максимальный размер cache памяти на запись. Объем cache памяти на запись напрямую влияет на способность системы поглощать пиковые нагрузки.

Реальная эффективность работы cache памяти является существенно более важной характеристикой, чем размер в гигабайтах. Под эффективностью понимается процент попаданий в cache память в реальных приложениях. Известно, что в системах Symmetrix процент попаданий в базах данных (основное применение рассматриваемых систем) практически всегда превышает 60%. Это говорит о высокой эффективности используемых алгоритмов.

8. Многоуровневое хранилище

Инфраструктуру хранения предпочтительно строить с использованием стратегии многоуровневого хранилища. Эта стратегия подразумевает хранение данных на различных по производительности, доступности, функциональности, а следовательно, и стоимости системах, в зависимости от требований, предъявляемых к конкретному типу данных на текущем этапе их жизненного цикла.

Данные на каждом уровне хранения получают адекватный уровень обслуживания, при этом цена за единицу емкости второго и третьего уровней в разы в разы меньше, чем уровня 1, где размещаются данные наиболее критичных к производительности приложений. Таким образом, с одной стороны для всех данных сохраняются степень защиты и централизованное управление на уровне системы старшего класса, а с другой — обеспечивается существенная экономия средств. Так, например, в системе могут присутствовать три уровня

1. Уровень 1 — высочайшая транзакционная производительность: база данных основного приложения (управление предприятием) и почтовая система.
2. Уровень 2 — существенная транзакционная и потоковая производительность: аналитические приложения.

3. Уровень 3 — достаточная потоковая производительность: резервные копии основных приложений для обеспечения быстрого восстановления производственных данных.

В обеспечение многоуровневого хранилища входят

- Независимые каналы ввода-вывода для пулов физических дисков, относящихся к различным уровням хранения
- Автоматическое выравнивание и оптимизация производительности дисковой подсистемы в пределах каждого уровня хранения
- Динамическое разделение ресурсов cache памяти и процессоров ввода-вывода в зависимости от потребностей каждого уровня хранения
- Поддержка недорогих дисков

9. Логические разделы cache памяти

Цель разбиения системы хранения на разделы – обеспечение гарантированного уровня обслуживания для приоритетных приложений. Оборотной стороной такого разделения является снижение загрузки системы, усложнение управления и, как следствие, существенное повышение общей стоимости владения.

Решением данной проблемы является введение динамики и специализированного управления в технологию разделов. Cache память динамически разделяется на логические разделы, при этом ресурсы системы, выделяемые каждому разделу в конкретный момент времени, зависят одновременно от:

- реальных потребностей конкретных приложений в данный момент времени,
- приоритетности приложений, относящихся к данному разделу.

Таким образом, достигаются сразу две цели:

- гарантированные ресурсы системы, выделяемые приоритетным приложениям,
- высокая общая загрузка системы в целом, оптимизация общей стоимости владения системой

Каждому разделу устанавливается минимальный и максимальный объем потребляемой памяти. В этих пределах границы разделов

динамически изменяются в зависимости от текущих потребностей и приоритетности приложений.

10. Управление томами данных. Виртуализация физического пространства

Управление томами данных (создание, размещение, расширение, перемещение) – одна из основных задач администратора системы хранения данных. Большинство эксплуатационных действий в системе так или иначе связано с управлением томами данных. От реализации управления томами зависят такие ключевые показатели системы как:

- Обеспечение непрерывного доступа к данным, несмотря на изменяющиеся требования к уровню обслуживания приложений
- Снижение риска недоступности данных из-за ошибок администрирования
- Общие затраты на администрирование системы.

На начальном этапе эксплуатации система должна обеспечивать быстрый и удобный способ создания томов данных. Виртуализация емкости позволяет администратору абстрагироваться от конкретного расположения данных на физических дисках.

Впоследствии, в процессе эксплуатации, по мере изменения требований к конкретным томам данных все более критичной становится гибкость системы. Важно, чтобы такие задачи как увеличение емкости и/или производительности томов выполнялись в широких пределах и без остановки работы приложений. То же самое относится и к задаче переноса томов между уровнями хранения. Если СХД не предоставляет такой гибкости, то перечисленные задачи означают остановку приложений и ручное копирование томов данных в новое расположение, что требует существенных усилий администратора и чревато ошибками.

11. Виртуализация физических дисков

При организации системы хранения данных предпочтительным является решение, где администратор не занимается физическим размещением томов данных на дисках. Диски эквивалентного объема и скорости представляют собой единый пул. При этом тома данных собираются администратором из однотипных квантов (рис. 4).

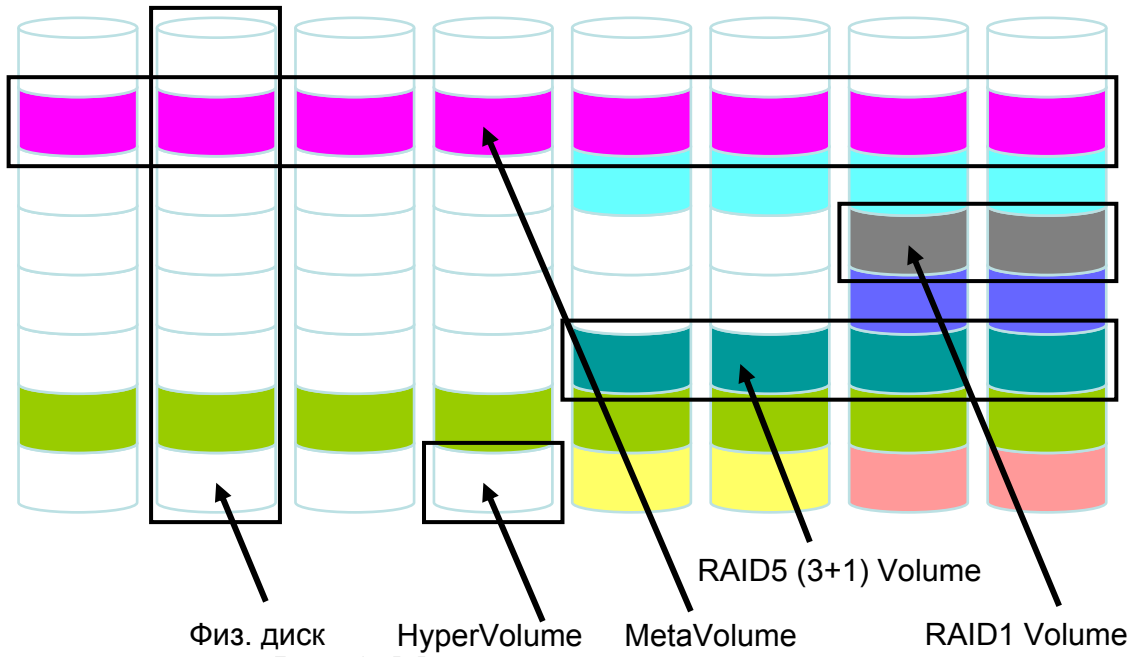

Рис. 4. Модель виртуализации дисков

Из гипер-томов собираются тома (volume), в зависимости от типа защиты. Например, RAID1 том состоит из двух гипер-томов, а RAID5 (3+1) том состоит из четырех гипер-томов. Совсем не обязательно, чтобы тома состояли из последовательных гипер-томов, как на рис. 4, – это сделано лишь для наглядности изображения. Фоновый процесс оптимизации производительности реформирует гипер-тома таким образом, чтобы распределение нагрузки на все диски было равномерным в пределах заданного пула физических дисков.

При необходимости администратор может собрать мета-том из нескольких томов. При этом возможно как простое сращивание томов, так и чередование данных по всем членам мета-тома. В последнем случае все запросы к данному тому равномерно распределены по большому количеству физических дисков.

Итак, модель виртуализации физической емкости обладает следующими преимуществами:

- Абстрагирование от физических носителей при управлении системой. Перед администратором не стоит задача физического размещения данных на диске.
- Возможность увеличить как емкость, так и производительность любого тома в широких пределах без остановки обслуживания приложений.

- Перенос томов между уровнями хранения без остановки работы приложений
- Автоматическое выравнивание производительности по всему пулу физических дисков.

Том не привязывается к какой-либо группе дисков. Его кванты могут физически располагаться в любом месте в пределах пула. Кванты автоматически переносятся таким образом, чтобы общая загрузка дисков была равномерной, а время отклика минимальным. На рис. 5 изображен типичный результат работы такого оптимизатора. Состав и емкость томов остались неизменными, как и на рис. 4 Но теперь физические диски загружены существенно более равномерно. Отметим, что желательно учитывать даже такие нюансы, как перенос наиболее загруженных томов ближе к средней области диска. Разумеется, процессы оптимизации происходят прозрачно для приложений.

Рис. 5. Результат оптимизации производительности дисковой подсистемы

В результате не только обеспечиваются более равномерная загрузка дисков и лучшее время отклика, но и сокращается нагрузка на администратора.

12. Миграция томов между уровнями хранения

Задача перемещения томов внутри системы возникает в целом ряде случаев. Например, при добавлении новых, более производительных дисков в систему. Или же при повышении или снижении требований к производительности конкретного тома данных.

В случае изменения требований к какому-либо приложению или тому тома переносятся между уровнями хранения без прерывания работы. Все настройки тома сохраняются. Если том был защищен локальным или

удаленным зеркалированием, то при миграции тома внутри системы настройки пар зеркал также сохраняются.

Тем самым обеспечивается дополнительная гибкость при управлении, снижается риск человеческой ошибки при миграциях данных и в целом ряде случаев исключаются нежелательные остановки работы приложений.

13. Локальная репликация

Создание копий производственных данных имеет целый ряд важных применений. Это — снижение до минимума окна резервного копирования, наличие локальной копии данных для практически мгновенного восстановления тома на определенный момент времени, задачи тестирования различных изменений в программном обеспечении и многие другие задачи. Во всех случаях используется копия производственных данных на определенный момент времени.

Копии данных бывают полные и логические.

Полная копия (клон) – это отдельный том такого же размера, как и исходный том, на который определенным образом копируются данные.

Логический снимок (снапшот) представляет собой набор ссылок на исходные данные. При этом измененные данные с момента создания логического снимка копируются в заранее выделенную область массива. Таким образом, логический снимок позволяет экономить дисковое пространство. Однако использование логического снимка для чтения данных приводит к снижению производительности основного тома.

Эффективность создания копии данных сильно зависит от степени гибкости программного обеспечения локального копирования. Различия между существующими средствами копирования проявляются на первый взгляд в мелочах, которые, однако, кардинальным образом влияют на эффективность работы администратора, время восстановления в случае повреждения данных и т. д. В частности, чрезвычайно полезными оказываются следующие технологии

- Возможность создать полную копию данных (клон) мгновенно. Копирование данных происходит в фоновом режиме.
- Инкрементная синхронизация данных копии и источника в обоих направлениях для всех копий производственного тома. Немедленная доступность данных при синхронизации в обоих направлениях.
- Согласованное (консистентное) копирование групп зависимых томов

14. Удаленная репликация и зеркалирование данных

Основным применением удаленной репликации и зеркалирования данных является защита от крупномасштабных аварий, делающих невозможным продолжение работы основного центра обработки данных (в дальнейшем – площадки) заказчика. Наличие зеркала всех данных на резервной площадке позволяет продолжить работу несмотря на аварию. Все современные системы хранения данных реализуют в той или иной степени удаленное копирование и зеркалирование.

Существенно, чтобы том-цель на удаленной системе являлся именно аппаратным зеркалом, а не просто копией тома-источника на удаленной системе. В таком случае чтение данных возможно с обеих систем прозрачно для приложения. Следствиями такой архитектуры являются следующие важные преимущества:

- Немедленная доступность данных для приложений при любого рода переключениях и синхронизациях между площадками в обе стороны. Причем речь идет как о плановых, так и об аварийных переключениях. Данные будут копироваться в фоновом режиме, а запросы к еще не скопированным данным будут прозрачно перенаправлены на удаленную систему.
- Возможность считывать данные напрямую с резервной системы при повреждении данных на основной системе без переключения на удаленный сайт.

Асинхронная репликация реализует поддержание копии данных с отставанием по времени. Такой метод позволяет предотвратить задержки, связанные с передачей данных на удаленную площадку при существенных расстояниях. Вместе с тем, удаленная копия с отставанием по времени – одна из наиболее широко распространенных технологий защиты от катастроф. При повреждении основной копии данных у заказчика есть консистентная копия на удаленной площадке. Используя эту копию и журнал изменений можно восстановить данные на последний момент времени до начала повреждения. Асинхронную репликацию можно использовать для обновления point-in-time копий на удаленной площадке. При этом point-in-time копия немедленно может использоваться приложениями, не дожидаясь окончания синхронизации. Очень важным моментом является восстановление после сбоя. Если на удаленной площадке велась работа приложений, то для возвращения исходной конфигурации необходимо только копирование изменений, а не полная ресинхронизация.

В случае необходимости возможно переключение между синхронным и асинхронным режимами без остановки репликации.

15. Катастрофоустойчивые решения с применением трех площадок

Синхронная репликация данных между двумя близкорасположенными площадками позволяет защититься от локальных катастроф масштаба здания или района города. Для защиты от более существенных катастроф, таких, например, как обесточивание целого города или области, все чаще применяются решения с репликацией данных между тремя площадками.

При этом на ближнюю резервную площадку (до 200 км) данные передаются с основной площадки в синхронном режиме, а на дальнюю резервную площадку — в асинхронном режиме. Типичная реализация такого проекта изображена на рис. 6. При крахе на первой площадке можно восстановить работу на второй или третьей площадке в зависимости от масштабов катастрофы.

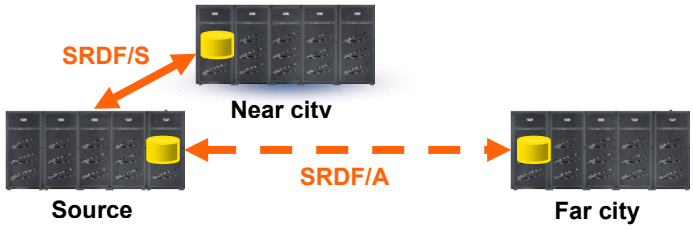

Рис. 6. Concurrent SRDF/S–SRDF/A

Основной недостаток данного подхода заключается в следующем. При крахе на первой площадке и восстановлении на второй теряется синхронизация между данными второй и третьей площадок. Для установления синхронизации необходимо полное копирование данных между уцелевшими второй и третьей площадками. При этом в течение процесса копирования данных система не защищена от потери второй площадки.

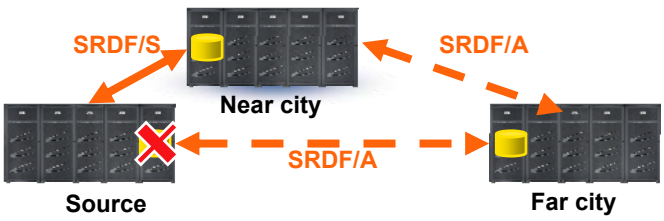

Рис. 7. SRDF/STAR

Положение исправляется, если постоянная синхронизация данных и отслеживание изменений происходят между всеми тремя площадками (рис. 7). В этом случае при крахе первой площадки и восстановлении

работы на второй происходит автоматическая инкрементная синхронизация данных между второй и третьей площадками.

Такое решение обеспечивает следующие преимущества перед одновременной репликацией на две резервных площадки либо каскадной репликацией.

- Возможность немедленно переключиться на любую из трех площадок
- Автоматическая инкрементная синхронизация данных между оставшимися двумя площадками вне зависимости от того, какая из трех площадок вышла из строя

Заключение

На сегодняшний день системы хранения данных старшего класса работают во многих организациях по всему миру. В основном это организации, операционная деятельность которых предъявляет чрезвычайно высокие требования к доступности и производительности приложений, работающих с данными. Таковыми являются финансовые организации, для которых важны как производительность (поскольку задержки в функционировании основных систем могут стоить миллионы долларов), так и надежность (поскольку потеря важной финансовой информации может повлечь за собой еще более тяжелые последствия). Еще один потребитель систем хранения старшего класса — телекоммуникационные компании, где необходимо реализовать биллинг огромного количества информации в пределах заданного времени и, следуя нормативным актам регулирующих органов, надежно сохранять эту информацию длительный период времени.

055(02)2

Подписано в печать 25.03.2008 г. Заказ 21.
Формат бумаги 60x90 1/16. Тираж 59 экз.

Отпечатано в Институте прикладной математики РАН
Москва, Миусская пл., 4