Jet Info

ИНФОРМАЦИОННЫЙ БЮЛЛЕТЕНЬ № 7 (122)/2003

# Методы построения систем хранения данных


КОРПОРАТИВНЫЕ СИСТЕМЫ

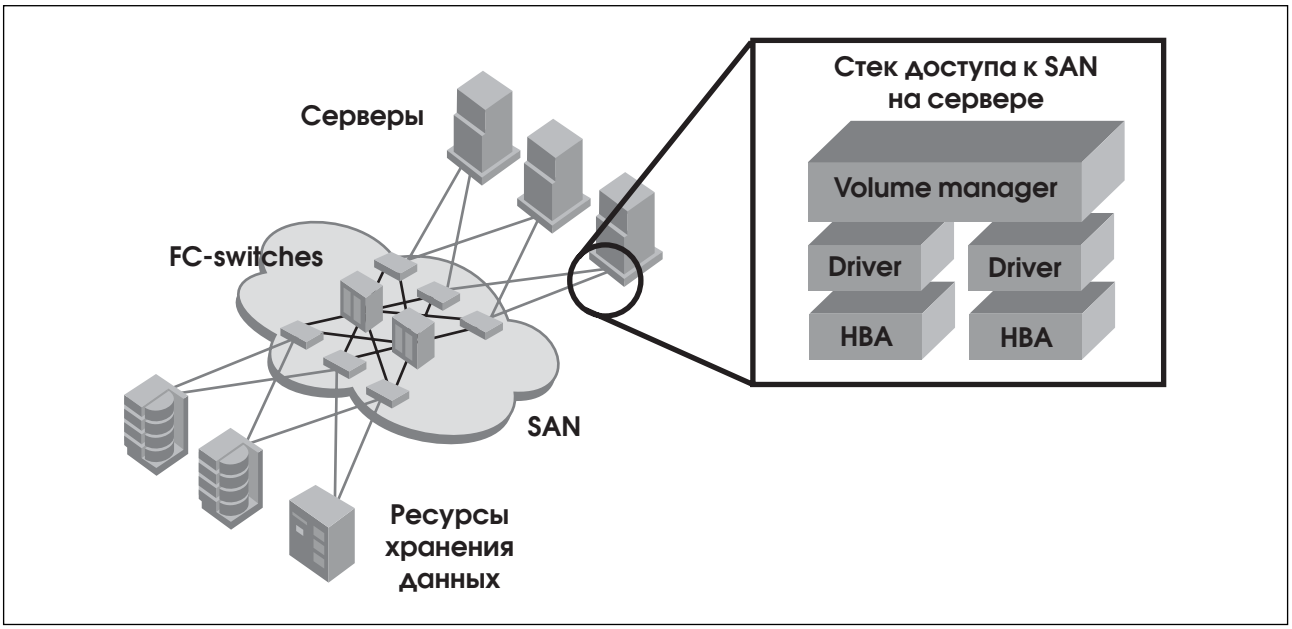


**Рис. 1. Инфраструктура системы хранения данных на основе SAN**

## Отсутствие доступа к данным равноценно отсутствию данных!

Доступ к данным невозможен как в случае выхода из строя каналов (доступа) или вычислительных средств, так и в случае отсутствия необходимой производительности для выполнения прикладных задач.

Выделение средств хранения данных в отдельную подсистему в рамках Вычислительного Комплекса позволит проектировщикам сконцентрироваться на решении проблем обеспечения надежного хранения и доступа к данным в рамках одной подсистемы. Кроме того, это создает предпосылки для оформления системы хранения данных (СХД) в организационно-техническую структуру, что является основой для аутсорсинга услуг по предоставлению средств хранения данных.

## Из чего состоит система хранения данных?

Для многих система хранения данных ассоциируется с устройствами хранения и, в первую очередь, с дисковыми массивами. Действительно, дисковые массивы сегодня являются основными устройствами хранения данных, однако, не стоит забывать, что обработка информации, формирование логической структуры ее хранения (дисковых томов и файловых систем) осуществляется на серверах. В процесс доступа к данным, (помимо процессоров и памяти сервера) вовлечены установленные в нем адаптеры (Host Bus Adapter — HBA), работающие по определенному протоколу, драйверы, обеспечивающие взаимодействие HBA с операционной системой, менеджер дисковых томов, файловая система и менеджер памяти операционной системы.

Если дисковый массив выполнен в виде отдельного устройства, то для его подключения к серверам используется определенная инфраструктура. В зависимости от протокола доступа (транспорта), реализованного в HBA и дисковом массиве, она может быть простой шиной (как в случае с протоколом SCSI), так и сетью (как в случае с протоколом Fibre Channel (FC)). Если это сеть, получившая название «сеть хранения данных» (Storage Area Network — SAN), то, как и положено сети, в ней используется активное оборудование — концентраторы и коммутаторы, работающие по протоколу FC, маршрутизаторы протокола FC в другие протоколы (обычно в SCSI). Таким образом, **помимо устройств хранения данных в состав СХД необходимо еще добавить инфраструктуру доступа, связывающую сервера с устройствами хранения.**

Отвечая на вопрос, где правильно провести черту, отделяющую систему хранения от серверного комплекса, предлагается рассматривать систему хранения данных как «черный ящик». Тогда, для подключения сервера к системе хранения, достаточно установить в сервер HBA с необходимым протоколом, подключить его к системе хранения и сервер сразу «увидит» свои данные — то есть по принципу «plug and play». Это идеальная ситуация, к которой IT-индустрия, возможно, придет в будущем. Сегодня границу, отделяющую систему хранения данных от серверов, надо проводить на самих серверах выше уровня менеджера дисковых томов. А почему именно так, можно убедиться на следующем примере: в системах, где требуется вы-

сокий уровень готовности, дисковый массив может считаться единой точкой отказа (Single Point Of Failure — SPOF). Для ликвидации SPOF обычно устанавливается второй массив, при этом данные зеркалируются на оба массива. Сегодня одним из самых распространенных средств зеркалирования является менеджер дисковых томов (например, VERITAS Volume Manager). Таким образом, менеджер дисковых томов вовлечен в процесс обеспечения отказоустойчивости системы хранения данных и становится её компонентом.

Сетевой инфраструктурой, объединяющей большое количество серверов и устройств хранения, необходимо управлять и, как минимум, отслеживать ее состояние. Сказанное не означает, что нет необходимости мониторинга состояния, например, двух серверов и одного массива, подключенного к ним напрямую. Однако, это можно реализовать подручными средствами — встроенными утилитами серверов, массива и операционной системы, бесплатными (freeware) утилитами или «самописными» скриптами. Каждое из устройств в СХД имеет несколько объектов, требующих управления и контроля состояния, например дисковые группы и тома у массивов, порты у массивов и коммутаторов, адаптеры в серверах. Как только число объектов управления в СХД начинает исчисляться десятками, управление такой конфигурацией при помощи "подручных" средств отнимает у администраторов слишком много времени и сил, и неизбежно приводит к ошибкам. Справиться с такой задачей можно только используя полномасштабную систему управления. Это справедливо для любых больших систем и для большой системы хранения данных, в частности. Внедрение системы управления становится особенно актуальным в тех случаях, когда система хранения данных выделена не только структурно и функционально, но и организационно.

Система хранения данных должна включать следующие подсистемы и компоненты:

- **Устройства хранения данных: дисковые массивы и ленточные библиотеки.** Современные высокопроизводительные дисковые массивы используют технологию Fibre Channel для подключения к ним серверов и для доступа к дискам внутри массива. Они могут масштабироваться до десятков терабайт дискового пространства и обладают встроенным интеллектом для выполнения специальных функций, таких как: виртуализация дискового пространства, разграничение доступа к дисковому пространству, создание Point-In-Time (PIT) копий данных[1] и репликация данных между массивами. К устройствам хранения данных также относятся всевозможные библиотеки - ленточные, магнитооптические и CD/DVD, которые в данной статье рассматриваться не будут.
- **Инфраструктуру доступа серверов к устройствам хранения данных**. В настоящее время, как правило, инфраструктура доступа серверов к устройствам хранения данных создается на основе технологии SAN. SAN является высокопроизводительной информационной сетью, ориентированной на быструю передачу больших объемов данных.
  *В основе концепции SAN лежит возможность соединения любого из серверов с любым устройством хранения данных, работающим по протоколу Fibre Channel. Сеть хранения данных образуют: волоконно-оптические соединения, Fibre Channel Host Bus Adapters (FC-HBA) и FC-коммутаторы, в настоящее время обеспечивающие скорость передачи 200 МБайт/с и удаленность между соединяемыми объектами до нескольких десятков километров. В случае, если расстояние между объектами превышает возможности FC-оборудования или нет достаточного количества «тёмной» оптики[2], связь между объектами можно обеспечить используя технологию уплотненного спектрального мультиплексирования DWDM или инкапсулировав FibreChannel в другой транспортный протокол, например в TCP/IP. Технология DWDM (Dense Wavelength Division Multiplexing) позволяет оптимальным образом применять оптоволоконные ресурсы и передавать не только трафик Fibre Channel, но*

[1] Определение понятия Point-In-Time копии данных (PIT-копия, иногда встречается сокращение P-I-T-копия) следует из его названия — это копия данных, сделанная на определенный момент времени, и состояние данных «заморожено» в момент создания копии. Иногда путают PIT-копии с «моментальными снимками» (SnapShot), которые в действительности являются только одним из методов создания PIT-копий. К другим методам создания PIT-копий относятся методы клонирования (clone) данных.

[2] «Темная» оптика — это технический жаргон, обозначающий оптическую магистраль (кабель) на пути следования которой не установлены никакие активные устройства. Отсутствие таких устройств подразумевает, что по кабелю не передается никаких сигналов. Для оптики таким сигналом является свет, т.е. в оптический кабель не светит ни какое устройство. Отсюда и происхождение термина. Без применения дополнительных устройств, например FC-ATM конвертеров, FC-коммутаторы не могут предавать пакеты по магистрали, где присутствуют другие активные устройства.

*также Ethernet и другие протоколы по одним и тем же оптическим каналам одновременно. При этом расстояния между соединяемыми объектами могут составлять сотни и даже тысячи километров. Подробнее о SAN можно прочитать в [1].*

- **Систему резервного копирования и архивирования данных.** Система предназначена для создания резервных копий и восстановления данных. Система резервного копирования позволяет защитить данные от разрушения не только в случае сбоев или выхода из строя аппаратуры, но и в результате ошибок программных средств и пользователей.
  Выполнение резервного копирования является одним из необходимых методов обеспечения непрерывности бизнеса. Создание *централизованной* системы резервного копирования позволяет сократить совокупную стоимость владения IT-инфраструктурой за счет оптимального использования устройств резервного копирования и сокращения расходов на администрирование (по сравнению с децентрализованной системой).
- **Программное обеспечение управления хранением данных.** Программное обеспечение предназначено для решения задач управления хранением данных, например, для разметки дисковых томов или повышения производительности доступа к данным прикладного ПО. Например, встроенное в массивы Hitachi Lightning 9900V программное обеспечение Cruise Control собирает статистику по интенсивности работы с данными, и исходя из нее принимает решение о перемещении данных на диски, производительность которых соответствует скорости обращения к данным.
- **Систему управления.** Система предназначена для мониторинга и управления уровнем качества сервиса хранения данных. Она тесно интегрируется с системой управления ВК. Основой системы управления СХД являются средства управления аппаратными ресурсами сети хранения данных. Их интеграция с другими системами дает возможность контролировать ресурсы СХД и управлять ими на всех уровнях, от дисков в массиве до файловой системы сервера.

Среди подсистем СХД система резервного копирования заслуживает особого внимания. Как следует из определения, создание системы резервного копирования является одним из средств обеспечения надежного хранения данных, о которых поговорим ниже. Однако, систему резервного копирования необходимо включить в СХД как отдельную подсистему не только по этой причине. Объем данных, измеряемый единицами и десятками терабайтов, требует все больше времени на процедуру резервного копирования. Классические средства резервного копирования по ЛВС не успевают выполнить эту процедуру и уложиться в отведенное временное «окно», которое сокращается с приближением режима работы информационной системы к «24x7» (например, в системах обслуживающих регионы из центра). Решением указанной проблемы является использование SAN для передачи данных резервного копирования, а также применения средств современных дисковых массивов для создания PIT-копий. В этом случае потребуется тесная интеграция системы резервного копирования с SAN и дисковыми массивами.

## Какие задачи стоят перед системой хранения данных и как они решаются?

Система хранения данных предназначена для организации надежного хранения данных, а также отказоустойчивого, высокопроизводительного доступа серверов к устройствам хранения. Существующие в настоящее время методы по обеспечению надежного хранения данных и отказоустойчивого доступа к ним можно охарактеризовать одним словом — **дублирование**.

Так, для защиты от отказов отдельных дисков используются технологии RAID, которые (кроме RAID-0) применяют дублирование данных, хранимых на дисках. Уровень RAID-5 хотя и не создает копий блоков данных, но все же сохраняет избыточную информацию, что тоже можно считать дублированием. Для защиты от логического разрушения данных (разрушение целостности базы данных или файловой системы), вызванных сбоями в оборудовании, ошибками в программном обеспечении или неверными действиями обслуживающего персонала, применяется резервное копирование, которое тоже является дублированием данных. Для защиты от потери данных вследствие выхода из строя устройств хранения по причине техногенной или природной катастрофы, данные дублируются в резервный центр.

Отказоустойчивость доступа серверов к данным достигается дублированием путей доступа.

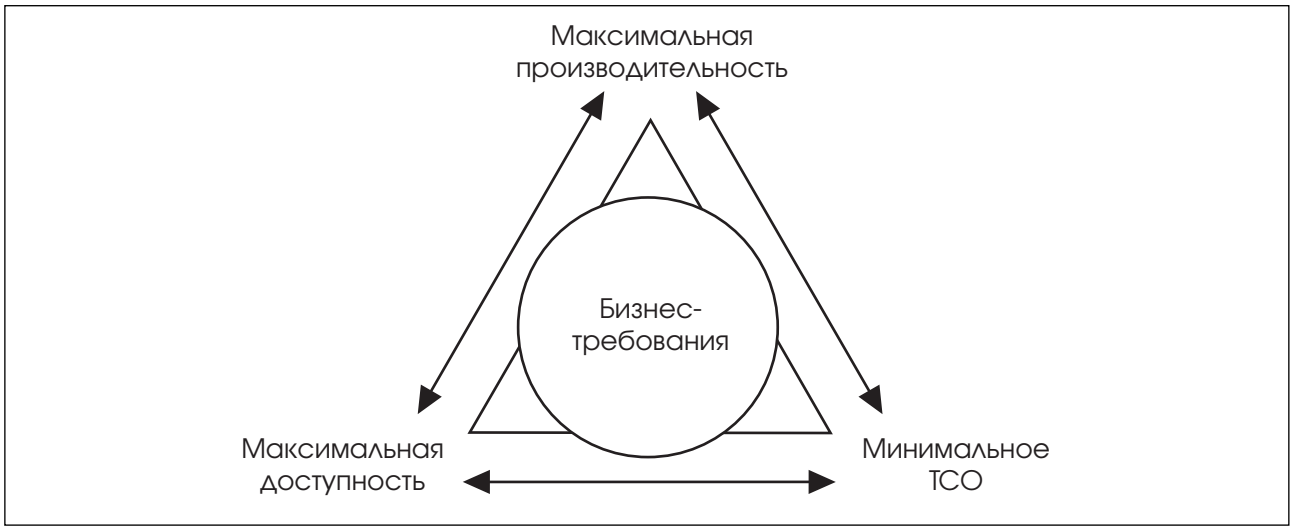


**Рис. 2. Конфликт целей**

Применительно к SAN дублирование заключается в следующем: сеть SAN строится как две физически независимые сети, идентичные по функциональности и конфигурации. В каждый из серверов, включенных в SAN, устанавливается как минимум по два FC-HBA. Первый из FC-HBA подключается к одной "половинке" SAN, а второй — к другой. Отказ оборудования, изменение конфигурации или регламентные работы на одной из частей SAN не влияют на работу другой. В дисковом массиве отказоустойчивость доступа к данным обеспечивается дублированием RAID-контроллеров, блоков питания, интерфейсов к дискам и к серверам. Для защиты от потери данных зеркалируют участки кэш-памяти, участвующие в операции записи, а электропитание кэш-памяти резервируют батареями. Пути доступа серверов к дисковому массиву тоже дублируются. Внешние интерфейсы дискового массива, включенного в SAN, подключаются к обоим ее "половинкам". Для переключения с вышедшего из строя пути доступа на резервный , а также для равномерного распределения нагрузки между всеми путями, на серверах устанавливается специальное программное обеспечение, поставляемое либо производителем массива (EMC CLARiiON — PowerPath, HP EVA — AutoPath, HDS — HDLM), либо третьим производителем (VERITAS Volume Manager).

Необходимую производительность доступа серверов к данным можно обеспечить созданием выделенной высокоскоростной транспортной инфраструктуры между серверами и устройствами хранения данных (дисковым массивом и ленточными библиотеками). Для создания такой инфраструктуры в настоящее время наилучшим решением является SAN. Использование современных дисковых массивов с достаточным объемом кэш-памяти и производительной, не имеющей «узких мест» внутренней архитектурой обмена информацией между контроллерами и дисками, позволяет осуществлять быстрый доступ к данным. Оптимальное размещение данных (disk layout[3]) по дискам различной емкости и производительности, с нужным уровнем RAID в зависимости от классов приложений (СУБД, файловые сервисы и т.д.), является еще одним способом увеличения скорости доступа к данным.

Необходимо заметить, что оптимизация настроек программных средств, как самих приложений, так и операционной системы, дает существенно больший прирост производительности системы, чем использование более мощной аппаратуры. Обусловлено это в первую очередь тем, что оптимизация настроек устраняет «узкие места» (bottleneck) на путях следования потоков данных, тогда как новая аппаратура делает «горлышко бутылки» чуть шире и только (хотя иногда и этого достаточно для решения проблем быстродействия). В решении задачи оптимизации может помочь применение специального ПО, в котором реализованы функции, учитывающие особенности взаимодействия аппаратуры, операционной системы и прикладного ПО. Примером такого ПО служит опция Quick I/O файловой системы VxFS. Опция Quick I/O лицензируется в составе пакета VERITAS DataBase Edition (DBE) for ORACLE. Указанная опция позволяет СУБД ORACLE использовать Kernel Asynchronous IO (KAIO) для доступа к файлам данных, что существенно повышает производитель-

[3] Disk layout - это схема распределения данных приложения по дискам. Она учитывает в какие уровни RAID организованы диски, число и размеры разделов на дисках, какие файловые системы используются и для хранения каких типов данных они предназначены.

ность операций ввода-вывода СУБД. Подробнее об этом можно прочитать в [4].

Помимо достижения требуемых показателей производительности, отказоустойчивости и надежности хранения данных в СХД, заказчики также стремятся сократить совокупную стоимость владения системой (Total cost of ownership — TCO). Внедрение системы управления позволяет сократить расходы на администрирование СХД и спланировать расходы на её модернизацию. Консолидация технических средств также способствует сокращению расходов на эксплуатацию СХД.

## Какие задачи надо решить проектировщику в процессе создания системы хранения данных?

В процессе создания СХД необходимо достичь оптимального соотношения производительности, доступности (надежного хранения и отказоустойчивого доступа) и совокупной стоимости владения.

Одним из наиболее часто используемых методов обеспечения высокой доступности СХД является дублирование, которое повышает стоимость системы. Если не учитывать бизнес-требований заказчика к доступности системы, то система становится неоправданно дорогой. Погоня за ненужной производительностью также приведет к использованию дорогих технических средств. Помимо высоких показателей производительности, доступности и низкой стоимости нужно также обеспечить требуемую функциональность — объем хранения и число подключаемых серверов.

К сожалению, заказчики не всегда могут описать требования по производительности в количественных характеристиках, принятых для систем хранения — пропускной способности (Мбайт/с) или производительности (операций ввода-вывода в секунду — I/O per second (IOPS)). Поэтому, чтобы определить если не количественные характеристики, то хотя бы характер нагрузки, проектировщику необходимо выяснить, работу каких приложений должна обеспечивать СХД.

Различные классы приложений создают разную нагрузку на дисковую подсистему. Например, для СУБД характерны виды нагрузок, зависящие от классов задач — транзакционные системы (Online Transaction Processing (OLTP)) и аналитические системы (Decision Support Systems (DSS)). Для задач класса OLTP характерным является большой поток запросов на ввод-вывод небольших порций данных. Для задач класса DSS, напротив, характерно небольшое число запросов на чтение больших порций информации.

От того, какую нагрузку дает приложение, зависит выбор способа распределения данных по дискам и определение объема кэш-памяти дискового массива. Так для OLTP-задач наличие кэш-памяти в дисковом массиве может сыграть существенную роль для повышения производительности ввода-вывода. Напротив, в задачах класса DSS происходит считывание больших объемов данных, что практически исключает их повторное попадание в кэш-память (в отличие от OLTP-задач). Таким образом, кеширование считываемых данных при обработке DSS-задач не всегда увеличивает их производительность.

К типам нагрузки на СХД, производимыми задачами класса OLTP, DSS и файловыми сервисами можно отнести другие известные типы приложений. Так, почтовый сервис, построенный на базе MS Exchange или Lotus Domino, даёт сходную нагрузку на СХД, что и OLTP, поскольку указанные продукты построены на базе СУБД. Почтовый же сервис, основанный на sendmail, производит нагрузку на СХД, характерную для файловой системы с частым изменением метаданных. Средства резервного копирования выполняют последовательное чтение данных, подлежащих резервному копированию, что делает характер их операций схожим с DSS-задачами.

Существует еще один, не упомянутый ранее, тип нагрузки, характерный для процессов журналирования. Примером могут служить записи журналов транзакций СУБД или журналов событий операционных систем. К этому классу также можно отнести задачи загрузки данных в БД или хранилище данных (Data Warehouse). Нагрузка, осуществляемая на СХД этим классом задач, аналогична нагрузке DSS только для операций записи. Здесь наличие кэш-памяти (на запись) в дисковом массиве не увеличивает производительности записи данных. Данный тезис необходимо пояснить. Обычно применение кэш-памяти на запись уменьшает время, которое сервер тратит на операцию записи и ожидание её завершения. Но при записи большого объема информации (загрузка данных в БД) или при записи данных, которые пишутся на новое место, таких как журнал транзакций, существует высокая вероятность возникновения ситуации, когда кэш-память на запись уже будет полностью занята, а новый записываемый блок не соответствует ни од-

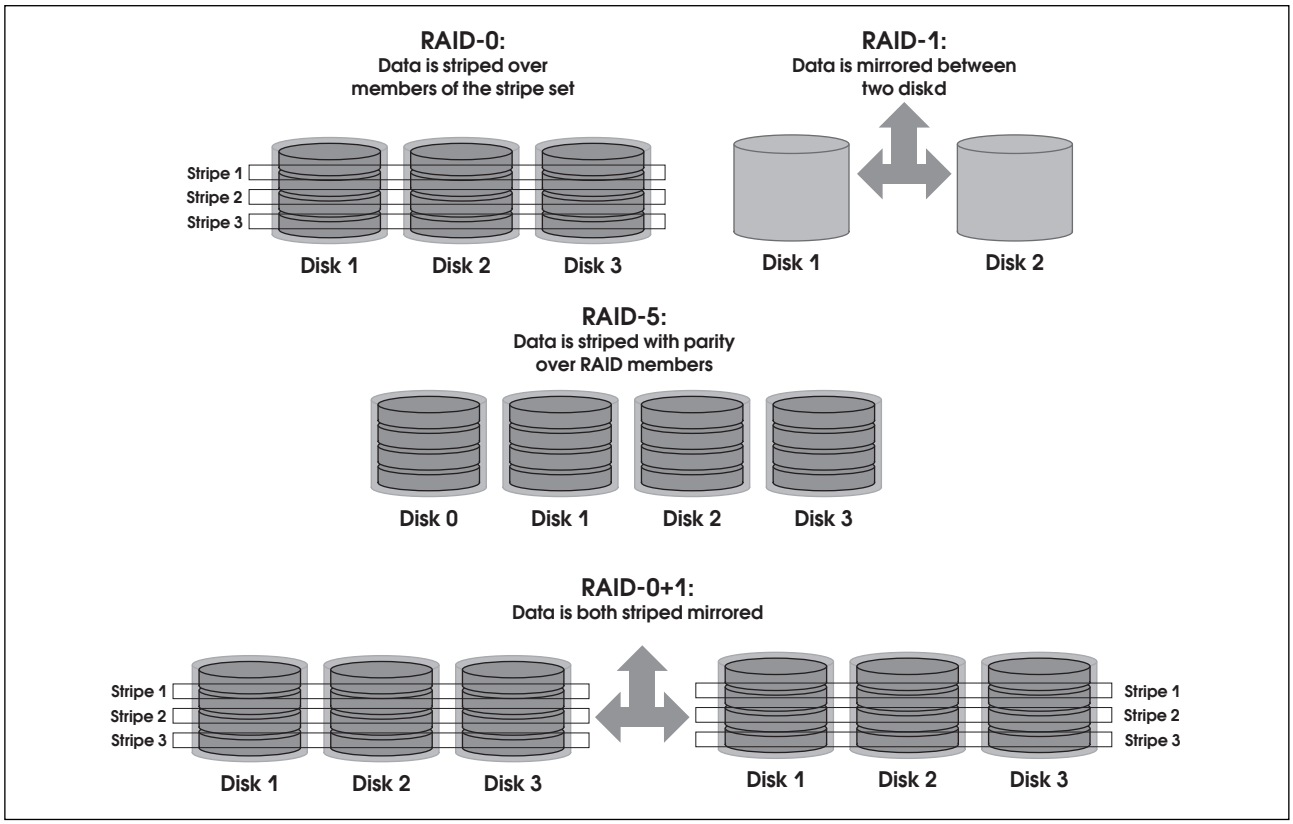


**Рис. 3. Уровни RAID**

ному из уже присутствующих в кэш-памяти. В этом случае массиву придется освободить несколько блоков кэш-памяти, записав их содержимое на диски, прежде чем начать обработку поступившей операции.

Определить классы задач, которые будет обслуживать СХД, необходимо не только для выработки политики работы с кэш, но также для распределения данных по дискам (disk layout). Для обеспечения надежного хранения информации диски организуются в уровни RAID 1, 0+1/1+0 или 5. Самым экономичным с точки зрения использования дополнительного (дублирующего) дискового пространства является уровень RAID 5. Однако производительность RAID 5 ниже, чем у RAID 1 или 0+1 при частых случайных изменениях данных, характерных для OLTP-задач, и высока при считывании данных, что характерно для DSS-задач.

Также разные уровни RAID обеспечивают различные уровни отказоустойчивости дисковой памяти к сбоям отдельных дисков. Так RAID 5 не спасает от двух последовательных отказов дисков, впрочем, как и RAID 0+1, если это диски разных половин «зеркала». Наиболее отказоустойчивым является уровень RAID 1+0 (попарное «зеркалирование» дисков и затем их «striping[4]»), поэтому его рекомендуется применять для хранения критичных данных, например, журналов транзакций СУБД. Практика показывает, что для хранения файловых систем и данных DSS-задач достаточно использовать RAID 5. Однако, если файловая система часто изменяется как, например, в почтовых системах sendmail, то имеет смысл использовать журналированную файловую систему или файловую систему с отдельно хранимыми метаданными, например файловую систему Sun QFS. Тогда для хранения журналов или отдельных метаданных лучше применять RAID 1 или 1+0.

Подробнее о влиянии кэш-памяти и disk layout на производительность ввода-вывода задач класса OLTP и DSS можно прочитать в [3].

Для «больших» систем актуальной становится оптимизация настроек СХД, направленная на повышение быстродействия для достижения заданного уровня сервиса. Под «большой» понимается такая система, в которой обрабатывается значительный

[4] Striping — метод размещения данных на дисках, при котором последовательно идущие блоки данных, составляющие логический том, записываются поочередно на каждый физический диск, входящий в дисковую группу. Таким образом достигается большая производительность, поскольку операции чтения и записи на диски могут производиться параллельно по сравнению с вариантом, когда все блоки логического тома записывается на один диск. Подробнее о striping и его влиянии на производительность ввода-вывода можно прочитать в [3].

объем данных — единицы и десятки терабайт, и/или к СХД подключены десятки серверов. Для небольших систем проблемы с быстродействием можно решить применением более производительной аппаратуры. В «больших» системах такой подход может оказаться неприемлемым либо в связи с тем, что потребуется очень дорогая аппаратура, либо в связи с тем, что уже достигнут предел аппаратной производительности. Единственным решением в данном случае является оптимизация. Для оптимизации производительности СХД желательно иметь возможность управлять настройками на всем пути следования данных от процессора к дискам и обратно. Для СУБД ORACLE такая оптимизация заключается в возможности использовать KAIO, а также возможности отключить кэш файловой системы для файлов данных ORACLE (поскольку СУБД ORACLE кэширует данные в собственной области памяти SGA). Для этой цели можно использовать упомянутый ранее пакет VERITAS DBE. Если в системе эксплуатируются OLTP- и DSS-задачи (что является типичным для большинства систем), то для оптимизации производительности дискового массива желательно иметь возможность управлять настройками кэш-памяти для каждого логического диска (LUN) в отдельности. Это необходимо для того, чтобы для тех дисков, где расположены данные OLTP-задачи, использовать кэш (и желательно большого объема), а для дисков с данными DSS-задачи использование кэш-памяти отключить. Однако, если для OLTP- и DSS-задач используются одни и те же таблицы данных, то такие настройки теряют смысл до тех пор, пока не будет решен вопрос о физическом разнесении данных задач по разным дискам, а выполнение самих задач перенесено на разные серверы. Это можно реализовать средствами СУБД, например, с помощью экспорта данных в файл и импорта их в другую базу. Если объем данных велик и синхронизацию баз данных OLTP- и DSS-задач надо проводить достаточно часто, этот вариант может оказаться неэффективным. Здесь может помочь технология создания PIT-копий данных, реализованная в большинстве современных дисковых массивов.

Выше говорилось, что СХД является важным звеном в обеспечении заданного уровня сервисов, предоставляемых информационной системой. Уровень сервиса зависит не только от производительности СХД, вопросы обеспечения которой только что обсуждались, но также от готовности и надежного хранения данных, другими словами, от доступности СХД. Исходя из бизнес-требований к информационной системе, необходимо определить режимы её работы (24x7, 12x5 и т.п.), степень критичности данных в зависимости от степени критичности сервисов, использующих эти данные, требования к готовности и надежности данных в зависимости от степени их критичности и режимов работы системы.

*Рассмотрим для примера работу информационной системы коммерческого банка. В банке эксплуатируется Автоматизированная Банковская Система (АБС), обслуживающая финансовые транзакции клиентов банка (OLTP-задача). Режим работы банка 8:00-20:00. Банк имеет несколько филиалов в ряде регионов России, которые работают с АБС головного офиса в терминальном режиме. Рабочие часы АБС составляют 6:00-22:00. Допустимое время простоя АБС — не более 1 часа. Допустимо потерять данные АБС не более чем за 0,5 часа, поскольку за этот период они могут быть повторно введены в систему с бумажных носителей (фактически это обусловлено бизнес-требованием по времени прохождения финансовой транзакции). Также в банке эксплуатируются аналитические задачи (DSS) на основе данных из АБС. Рабочее время аналитиков 9:00-18:00. Допустимое время простоя сервисов аналитических задач не более 4-х часов. Загрузка данных из АБС в аналитическую базу (синхронизация) происходит по закрытии «опердня» в 23:00. Таким образом, отставание аналитических данных от АБС составляет текущий «опердень». В случае потери данных аналитических задач они должны быть восстановлены на момент последнего закрытого «опердня». Срок восстановления аналитических данных зависит от того, в какой момент случился сбой в системе. Если сбой произошел утром, то аналитические данные должны быть восстановлены не позднее, чем за 4 часа. Если сбой случился после обеда или вечером, то восстановление должно быть завершено к утру следующего дня, плюс в аналитическую базу должны быть загружены данные последнего «опердня». Для выполнения OLTP- и DSS-задач используется СУБД ORACLE 8i.*

*Анализируя бизнес-требования к информационной системе из приведенного примера, получается, что СХД банка должна обеспечивать работу двух типов задач — OLTP и DSS. Данные OLTP-задачи являются критичными для системы в период с 6:00 до 23:00 (учитывается загрузка данных из АБС в аналитическую базу) и должны обеспечивать высокую готовность (простой не более 1 часа). Требования по надежности также высоки — потери не более получаса работы. Напротив, данные DSS-задачи не столь критичны и требования по готовности не столь высоки, но должна быть обеспечена высокая надежность — потери не допустимы.*

В предыдущем разделе были указаны методы обеспечения доступности СХД: дублирование аппаратных компонентов и дублирование данных, включающее применение различных уровней RAID, резервное копирование и репликацию в резервный центр.

*В приведенном примере ИС банка можно рекомендовать использовать RAID 1+0 для файлов данных и журналов транзакций OLTP-задачи, при этом необходимо расположить файлы данных и журналы транзакций на разных LUN. Такая схема позволит управлять производительностью (если используемый массив может управлять кэш-памятью для отдельных LUN) и обеспечит высокую надежность хранения данных. Для данных DSS-задачи рекомендуется использовать RAID 5. Этого вполне достаточно для надежного хранения данных DSS-задачи и производительности при чтении данных. Для отказоустойчивого доступа к данным в серверах АБС необходимо установить как минимум по 2 FC-HBA и подключить их к разным контроллерам дискового массива. При этом компоненты дискового массива должны быть продублированы, а участки кэш-памяти, используемые для операций записи, зазеркалированы и защищены от сбоев питания.*

О выборе массива определенного класса и с определенными характеристиками речь пойдет в следующем разделе. Но необходимо отметить, что не обязательно все компоненты массива должны быть дублированы, если за указанный допустимый период простоя они могут быть заменены, а данные при необходимости восстановлены с резервных копий.

Тип приложения влияет на то, как будет осуществляться резервное копирование. Например, для резервного копирования СУБД ORACLE средствами Recovery Manager (RMAN) рекомендуется использовать отдельный сервер (а, следовательно, и отдельный экземпляр базы данных и дисковое пространство для него), где будет размещен RMAN Recovery Catalog. Для резервного копирования файловых систем этого не требуется. Чтобы восстановить базу данных ORACLE необходимо иметь копии журналов транзакций, для чего рекомендуется активизировать в СУБД ORACLE режим архивирования журналов транзакций (ARCHIVELOG). Для архива журналов транзакций потребуется выделить дисковое пространство. Для его защиты от разрушения уровня RAID 5 будет достаточно. Какой тип резервного копирования использовать (полное или инкрементальное) зависит от того, за какое время можно будет скопировать базу данных и журналы транзакций с лент на диски, и укладывается ли полное резервное копирование в отведенное временное окно. Использование полного ежедневного резервного копирования позволяет восстановить базу быстрее, чем, например, применение полного еженедельного и каждодневного инкрементального. При расчете времени восстановления СУБД ORACLE рекомендуется учесть, что данные с ленты копируются на диски медленнее, чем записываются с дисков на ленты, поскольку надо записывать метаданные файловой системы [2]. Также надо учесть, что после копирования файлов данных и журналов транзакций с лент на диски СУБД ORACLE должна будет выполнить процедуру восстановления базы — RECOVERY, т.е. «накатить» все незавершенные транзакции из журналов.

*В приведенном примере ИС банка журналы транзакций необходимо будет копировать каждые полчаса. Если изменения в базе данных АБС не велики (за день осуществляется небольшое число финансовых транзакций), то процедура RECOVERY будет выполняться быстро (не более десятка минут). Если база данных и журналы транзакций могут быть скопированы с лент на диски менее чем за 50 минут, достаточно будет производить полное резервное копирование базы раз в сутки после закрытия «опердня». Но если эти условия не выполняются, то потребуется использовать более сложные технологии, такие как Storage Checkpoint, реализованные в VERITAS DBE, или средства создания PIT-копий.*

Как говориться, сбой сбою рознь. До этого момента обсуждались методы восстановления данных после «незначительных» сбоев — разрушения данных в результате отказов отдельных элементов аппаратуры (дисков, контроллеров и т.п.), ошибок ПО или действий персонала/пользователей системы. Сбой в работе системы может произойти из-за аварии, причиной которой может быть выход из стоя технического средства целиком (дискового массива или сервера) или, что еще хуже, техногенная (пожар, затопление) или природная (землетрясение, наводнение) катастрофа. От того, какие требования предъявляются к СХД по срокам восстановления после аварий, применяются различные схемы резервирования данных, что влияет на выбор технических и программных средств и, в конечном итоге, на стоимость решения.

*В примере с ИС банка, если требуется обеспечить восстановление работоспособности АБС в течение 1 часа после техногенной катастрофы (например, пожара) в помещении комплекса, то, в резервном центре необходимо по-*

*мимо серверов иметь резервный дисковый массив, дубликаты лент с резервными копиями и ленточную библиотеку достаточной мощности. Если объемы данных АБС и интенсивность их изменений не велики, то постоянное резервное копирование и регулярная перевозка дубликатов лент в резервный центр будет достаточным для защиты данных от аварии. Но в случае больших объемов данных (сотни гигабайт) и частых изменениях в базе (большом числе финансовых транзакций) указанная схема не эффективна, особенно, если учитывать, что резервное копирование дает существенную дополнительную нагрузку на аппаратный комплекс и, следовательно, может привести к недопустимому снижению его производительности. Альтернативным решением в приведенном примере может служить передача данных между массивами, расположенными на разных площадках, которая реализуется с помощью репликации или зеркалирования.*

Репликация может выполняться программным обеспечением, установленным на серверах (программная репликация), или встроенным в дисковые массивы программным обеспечением «прозрачно» для серверов (аппаратная репликация). Для организации репликации или зеркалирования необходимо создать высокопроизводительную инфраструктуру передачи данных, объединяющую обе площадки. Как правило, для этого применяют SAN [1].

При разработке проекта СХД для сохранения вложенных инвестиций необходимо учесть имеющиеся технические средства и наиболее оптимальным образом вписать их в новую систему. Надо также учесть наличие, навыки и опыт обслуживающего персонала. Если квалификации персонала недостаточно для обслуживания высокотехнологичной системы, то помимо организации обучения, тщательной проработки регламентов составления детальных инструкций и другой рабочей документации, необходимо внедрить систему управления СХД. Система управления не только облегчит работу, но еще и снизит вероятность ошибочных действий персонала, которые могут привести к потере данных.

Не секрет, что наличие плана развития системы облегчает её внедрение и снижает расходы на модернизацию, по сравнению с хаотичным развитием. С другой стороны, бизнес подвержен изменениям — меняются задачи, появляются новые пользователи и т.п. Поэтому, даже хорошо спроектированная СХД со временем перестает удовлетворять бизнес-требованиям.

Помочь в решении указных проблем может опять же внедрение все той же системы управления, которая позволит отследить изменение нагрузок на систему хранения, учесть утилизацию дискового пространства, спрогнозировать потребности в его наращивании и т.д.

Суммируя выше сказанное получаем, что для определения требуемых параметров СХД и поддержания их в заданных рамках необходимо четко определить классы, режимы работы и характеристики задач, работу которых должна обеспечить СХД, требуемый объем хранения данных и его возможный прирост, количество и платформы (UNIX, Windows и др.) подключаемых серверов. Иными словами, при создании СХД проектировщик должен исходить от задач и бизнес-требований заказчика.

# Как правильно выбирать дисковый массив?

Одной из главных составляющих системы хранения данных является дисковый массив. В процессе проектирования СХД неизбежно возникает вопрос, какой массив выбрать?

В предыдущем разделе обсуждались вопросы обеспечения производительности, доступности, масштабируемости, оптимизации и эксплуатации СХД. Исходя из этого, можно определить, какими свойствами должен обладать массив, чтобы обеспечить решение указанных задач. Требования наличия у массива определенных свойств или характеристик можно разделить на категории. Однако одно и то же свойство массива может попадать в разные категории, поскольку оно может быть использовано для решения разных задач.

Приведем часто встречающийся на практике, но не претендующий на полноту перечень требований, разбитый по категориям.

## Функциональные требования

- **Емкость «сырого» (raw),** т.е. без разметки на уровни RAID, дискового пространства массива должна составлять N ТБ. Если Вам встретилось требование к дисковому объему массива в такой формулировке, то это означает, что планирование disk layout еще не проводилось. В противном случае формулировка была бы иная: столько-то дисков такого-то объема и такой-то

скорости вращения, столько-то дисков другого объема и т.д.

- **Число LUNs, поддерживаемых дисковым массивом**. Данное требование можно четко сформулировать опять же только после планирования disk layout. Но число необходимых LUN можно «грубо» посчитать по числу серверов, подключаемых к дисковому массиву, с учетом выполняемых ими классов задач. Например, сервер ORACLE — 3 LUNs (данные, журналы, архив журналов), файл-сервер — 1 LUN, сервер sendmail — 2 LUNs (файлы и журнал файловой системы) и т.п.
- **Число подключаемых серверов и платформы подключаемых серверов**.
- **Возможность создания PIT-копии данных средствами массива**. Данная функциональность массива может потребоваться, если, например, принято проектное решение о загрузке данных из OLTP-задачи в DSS-задачу средствами массива. Функция создания PIT-копий может быть реализована различными методами — через «моментальный снимок» (SnapShot) (рис. 4) или через полное копирование данных (clone). Разница между этими методами заключается в том, что SnapShot экономит дисковое пространство, поскольку для его создания требуется всего лишь место для битовой карты и некоторого пула для сохранения старых значений измененных блоков. Напротив, clone требует того же (полезного) объема, что и копируемый LUN. Однако, если исходный LUN подвержен частым изменениям, то требуемый для поддержания SnapShot объем дискового пространства может существенно возрасти. Если с копией LUN, созданной с помощью SnapShot будет вестись интенсивная работа (большое число запросов на ввод-вывод), это может снизить производительность обмена данными с исходным LUN. Копия LUN с помощью SnapShot создается моментально (отсюда и название — «моментальный снимок»), поскольку процесс «копирования» заключается только в создании битовой карты. Для создания clone требуется определенное время, поскольку происходит полное копирование блоков данных. В этот момент нагрузка по вводу-выводу на копируемый LUN существенно возрастает. Существуют промежуточные способы создания PIT-копий, когда сначала создается SnapShot, а потом он постепенно преобразуется в clone. Проектировщик должен учесть все эти особенности методов создания PIT-копий и в требованиях четко указать какой метод планируется использовать.

## Требования к производительности

- **Дисковый массив должен обеспечивать производительность N IOPS, а агрегированная пропускная способность массива должна составлять M МБ/с**. Как уже отмечалось, получить такие цифры не просто. Если существует прототип системы или выбор дискового массива осуществляется для модернизации существующей СХД, то можно провести «натурные» замеры производительности и аппроксимировать их для ожидаемого роста нагрузки на СХД. Если система создается с «нуля», то можно попытаться получить эти цифры в качестве требований производителя прикладного ПО (что практически не реально) или обраться к производителям массивов, чтобы те провели определение требуемых параметров массива (sizing). Обычно производители имеют методики «грубой» оценки требуемой производительности. Но входными данными для этих методик, как правило, служат ожидаемое число транзакций и их «вес» (light, medium, heavy), которые тоже не всегда можно определить.
- **Специфические функции управления кэш-памятью массива**. Например, к таким функциям относятся:
  - возможность закрепления участка кэш-памяти за конкретным LUN (может пригодиться для размещения в кэш часто используемых служебных таблиц базы данных);
  - отключение использования кэш на запись и/или чтение для конкретного LUN (может потребоваться для DSS-задач);
  - обеспечение уровня сервиса в виде заданного уровня производительности (IOPS) или пропускной способности (МБ/с) для указанного сервера.

## Требования по отказоустойчивости и надежности хранения данных

- **Поддержка нужных уровней RAID**. Как правило, это уровни 1, 0+1,1+0 и 5.
- **Наличие дисков «горячей замены»** (hot-spare). Механизмы использования hot-spare дисков могут быть разные. Например, возможен вариант, когда в случае отказа диска данные из дисков затронутой RAID-группы копируются на hot-spare диск. Но также возможен вариант, когда нет специально выделенного hot-spare диска — все диски содержат данные, но при этом на всех дисках выделена резервная область, куда копируются данные с поврежден-

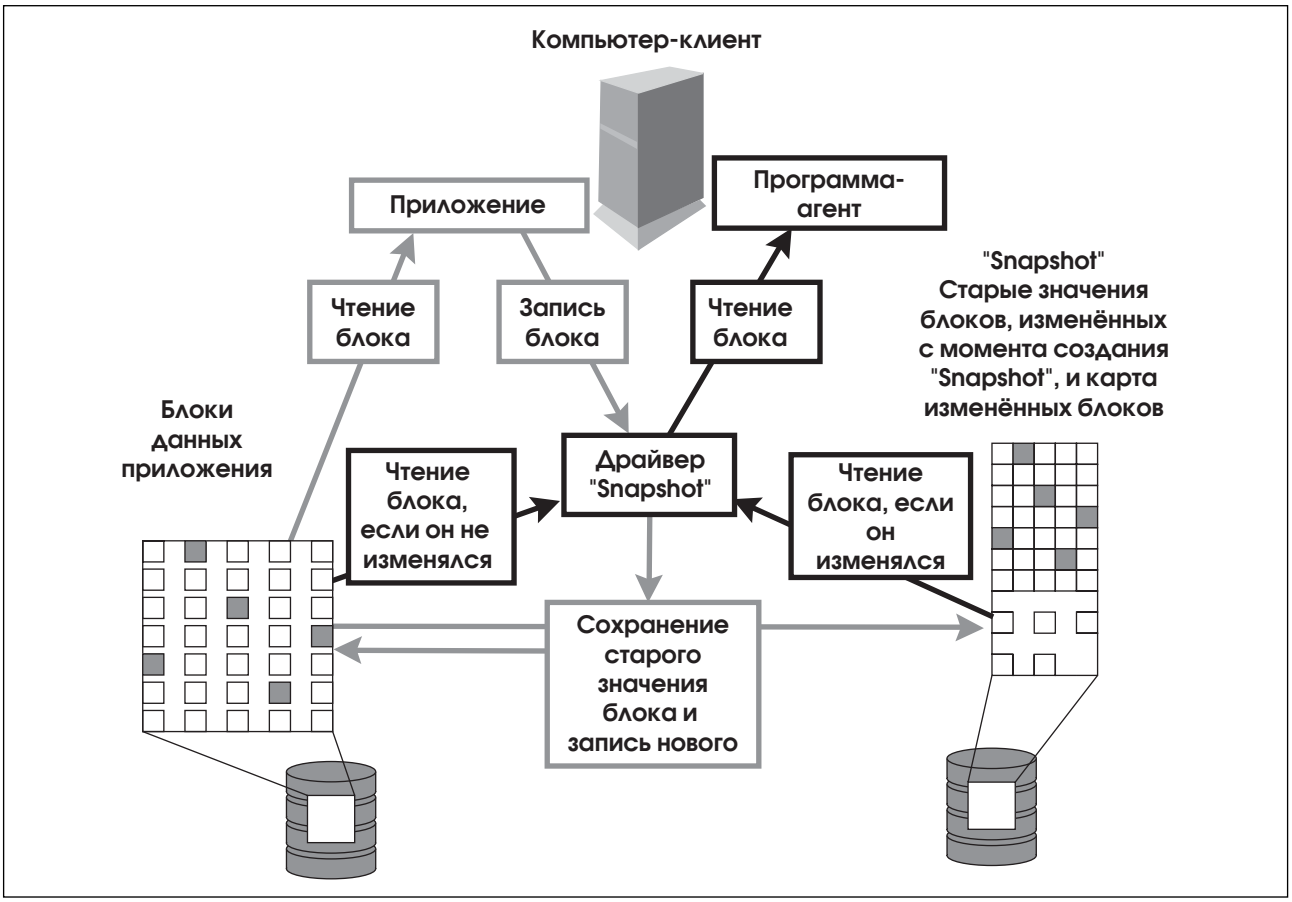


**Рис. 4. Схема работы SnapShot на примере Veritas Volume Manager**

ной RAID-группы. Определение требуемого метода опять же за проектировщиком.

- **Защита участков кэш-памяти, обслуживающих операции записи**. За исключением тех случаев, когда отключен кэш на запись, сервер получает подтверждение завершения операции записи сразу после попадания данных в кэш-память еще до записи их на диск. Для обеспечения целостности данных обычно применяются следующие методы:
  - Зеркалирование участков кэш-памяти, обслуживающих операции записи.
  - Поддержка батареями кэш-памяти в течении N часов или сохранение ее содержимого на диски в случае отключения внешнего питания. Какой из указанных вариантов определить в требованиях — задача проектировщика.
- **Дублирование всех компонентов и отсутствие единой точки отказа (SPOF)**. Степень важности этого требования зависит от режима работы системы и требований к доступности сервисов. Однако, не надо забывать, что сам массив является SPOF, если он не задублирован другим массивом.
- **Возможность создания PIT-копий данных для использования их в системе резервного копирования**. В ряде систем, где обрабатываются большие объемы данных (терабайты), а сервисы должны быть доступны 24x7 при больших нагрузках, необходимо применять Serverless резервное копирование. Для этого используется механизм создания PIT-копий средствами дискового массива.

## Требования по обслуживаемости

- **Возможность замены компонентов массива «на ходу» без остановки системы**. Выполнение этого требования важно для систем, работающих в режиме 24x7.

## Требования по масштабируемости

- **Наращивание дискового пространства до N ТБ без замены ранее установленных дисков**. Такая формулировка позволяет «убить двух зайцев» — обеспечить требуемую функциональность СХД при росте объемов обрабатываемых данных и сохранить сделанные инвестиции. Здесь может быть добавлено требование: «без потери производительности». Архитектура массива (об этом речь пойдет ниже) может стать «узким местом» и привести к тому, что при очередном добавлении дисков производительность массива существенно снизится, что повлияет на уровень качества сервиса.

- **Расширение размера LUN путем добавления новых дисков без разрушения хранимых данных**. Это требование важно не только для систем, работающих в режиме 24x7, но также когда имеется дефицит квалифицированного персонала, способного осуществить расширение дискового пространства при отсутствии у массива данной функции. Желательно, чтобы операционная система, данные которой хранятся на расширяемом LUN, могла автоматически расширить свою файловую систему.
- **Увеличение числа подключаемых серверов до N**.
- **Увеличение объема кэш-памяти до N ГБ без замены ранее установленных модулей**.

## Требования по управляемости

- **Управление политикой использования кэш-памяти для различных LUN**. Может потребоваться при «тонкой» настройке массива.
- **Наличие средств сбора статистики о работе массива**.
- **Наличие встроенных средств оптимизации работы массива**. Это достаточно специфичное требование, однако, наличие таких средств может помочь, когда потребуется оптимизация, а квалифицированного персонала, способного её выполнить, не будет.
- **Интеграция средств управления массива с уже развернутой системой управления, например HP OpenView.**

Чтобы не сравнивать все существующие на рынке массивы, было бы удобно разбить их на классы. Тогда на основе полученных требований можно выбрать нужный класс и уже сравнивать массивы только этого класса. Классы массивов придумывать не надо, они уже определены самим рынком, это: начальный класс (low-end), средний класс (mid-range) и высший класс (high-end).

Массивы указанных классов отличаются, в первую очередь, не количественными характеристиками, а функциональностью и архитектурой. К функциональности low-end массивов можно отнести поддержку различных уровней RAID и возможность дублирования контроллеров (если это не JBOD). От массивов класса mid-range уже требуется поддержка LUN-masking и создание PIT-копий. А в массивах класса high-end в дополнение к указанным возможностям также реализованы аппаратная репликация, поддержка OS/390 (zOS) и управление

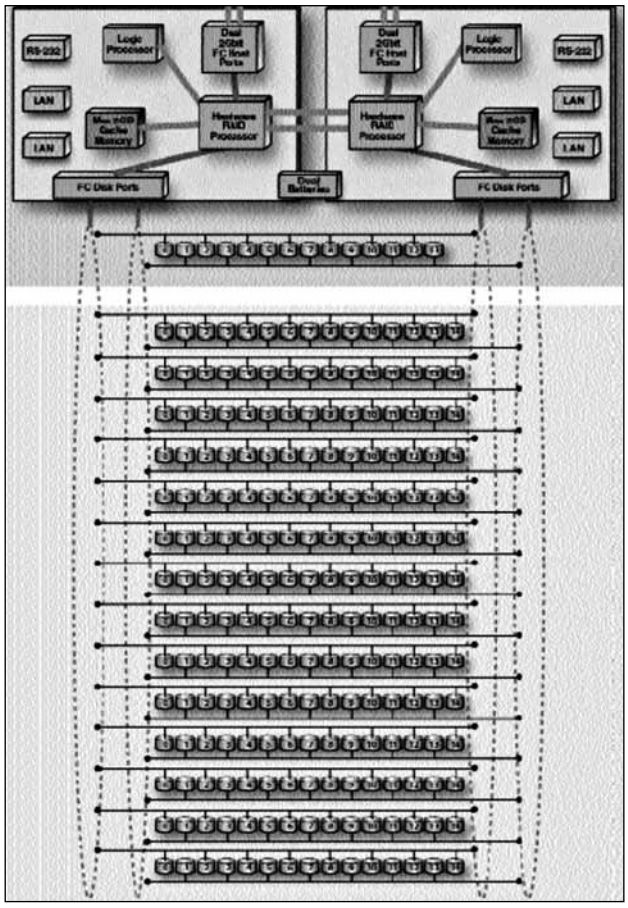

**Рис. 5. Типичная архитектура mid-range массива на примере HDS Thunder**

качеством сервиса (на уровне производительности в IOPS или пропускной способности в Мбайт/с).

Но все же основным критерием, по которому можно отнести массив к одному из классов mid-range или high-end, является архитектура. Многие производители заявляют, что mid-range массивы имеют модульную архитектуру, а high-end массивы — монолитную. Это не совсем верно. Модульная или монолитная «архитектура» говорит о конструктиве массива — собирается из отдельных блоков или шкафов. В действительности архитектуру всех mid-range массивов (и многих low-end) можно характеризовать как «двухконтроллерную с общей шиной». Смысл этого определения становится понятен, если взглянуть на рис. 5.

Для high-end массивов характерна коммутируемая или матричная архитектура[5] (рис. 6). Очевидно, что в данной архитектуре нет «узких мест», тогда как в mid-range архитектуре узкими местами являются: контроллер, поскольку каждый контроллер обслуживает свои RAID-группы (набор дисков, на которых реализован один уровень RAID), шина между контроллерами, ограниченное число FC-AL

[5] Иногда еще для high-end массивов говорят про «cache-centric» архитектуру, подчеркивая тем самым, что центральным звеном является кэш-память, к которой имеют доступ все контроллеры массива, тогда как в mid-range массивах кэш-память жестко привязана к определенному контроллеру.

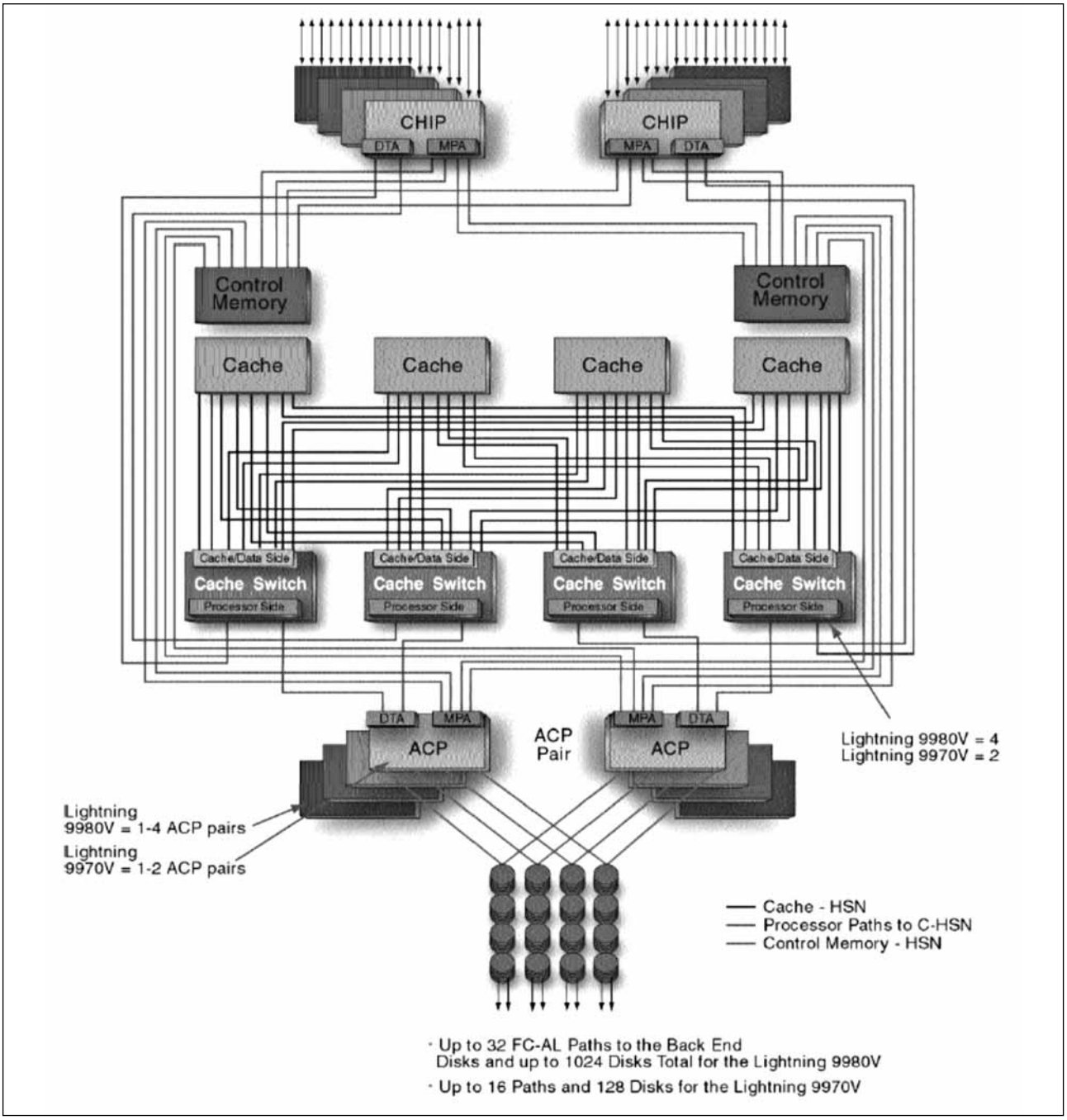


**Рис. 6. Коммутируемая архитектура high-end массива на примере HDS Lightning 9900V**

петель к дискам, расположение дисков RAID-группы «вдоль» одной петли FC-AL. В high-end массивах RAID-группы располагаются «поперек» FC-AL петель. Например, в high-end массивах Hitachi RAID-группа состоит из 4-х или восьми дисков, где каждый диск подключается к двум различным петлям от двух различных дисков контроллеров. Такая конфигурация позволяет выполнять операции записи-чтения со всех дисков RAID-группы параллельно, чего нельзя добиться в mid-range массивах, когда диски одной RAID-группы расположены вдоль одной петли и доступ к ним осуществляется по очереди.

Указанные отличия в архитектуре приводят к потери производительности при масштабировании mid-range массивов, чего не наблюдается у high-end массивов при добавлении новых дисков. Хотя современные mid-range массивы имеют высокие характеристики масштабируемости: позволяют устанавливать до двух-трех сотен дисков, распределяя их по нескольким FC-AL петлям, а также наращивать кэш-память до 8 ГБ, все же «узким ме-

стом» остается их архитектура, являющаяся ограничителем масштабируемости.

Если придерживаться указанной классификации, то только массивы HDS семейства 9900 и массивы EMC семейства Symmetrix можно отнести к классу high-end. Массивы HP EVA и IBM ESS 800 (Shark), которые позиционируются производителями как high-end массивы, имеют архитектуру, типичную для mid-range массивов.

## В заключение о виртуализации

Последнее время в маркетинговой литературе все чаще встречается понятие «виртуализации в системах хранения», которое определяется как скрытие от серверов физического расположения данных на дисках и представление всего дискового пространства как некоего общего пула блоков.

Этот пул уже, в свою очередь, «нарезается» на логические (виртуальные) диски (Logical Unit Number — LUN), которые «видны» серверам как физические. В действительности подобную организацию дисковой памяти давно уже позволяет создавать менеджер томов VERITAS Volume Manager. Данный тип виртуализации получил название «host-based» виртуализация.

Практически все современные дисковые массивы выполняют функцию создания из наборов физических дисков логических дисков (LUN), получившую название «disk array-based» виртуализация. Это легко определить на основании того факта, что ряд массивов поддерживают число LUNs больше, чем физических дисков, как, например, в недавно анонсированном массиве Sun StorEdge 3510. Вопрос заключается в том, насколько это удобно. Администраторы предпочитают иметь возможность управлять физическим размещением файлов данных СУБД ORACLE по физическим дискам для достижения оптимальной производительности и отказоустойчивости. Настройка СУБД под оптимальную производительность может стать проблематичной, если контроллер дискового массива не позволяет управлять размещением RAID-групп на конкретных дисках.

Кроме указанных двух типов виртуализации — «host-based» и «disk array-based», существует еще один тип — «SAN-based». В этом типе виртуализации скрытие от серверов физического расположения данных осуществляется либо с помощью специальных устройств, расположенных между FC-коммутаторами SAN («in-band» виртуализация), либо средствами самих FC-коммутаторов, считывающих информацию о конфигурации виртуального дискового пространства с внешнего устройства («out-off-band» виртуализация).

В настоящее время продуктов «SAN-based» виртуализации на рынке мало и говорить об их промышленном внедрении пока не приходится. Возможно, потребность в этом типе виртуализации появится тогда, когда объемы данных предприятий возрастут настолько, что для их оперативного хранения не будет хватать нескольких high-end дисковых массивов.

## Список литературы.


1 JetInfo № 9 (112)/2002. Сети хранения данных (SAN). Денис Голубев, Алексей Лобанов.

2 Backup and Restore Practices for the Enterprise. Stan Stringfellow, Miroslav Klivansky, Michael Barto, Michael Barton. Prentice Hall PTR. ISBN: 013089401X

3 Configuring and Tuning Databases on the Solaris Platform. Allan N. Packer, Sun Microsystems Press. ISBN: 0130834173

4 Configuring and Tuning Oracle Storage with VERITAS Database Edition™ for Oracle. Best Practices for Optimizing Performance and Availability for Oracle Databases on Solaris. http://eval.veritas.com/downloads/pro/dbed22_best_practices_paper.pdf



Jet Info
ИНФОРМАЦИОННЫЙ БЮЛЛЕТЕНЬ
Издается с 1995 года
Издатель: компания Джет Инфо Паблишер

Главный редактор: Дмитриев В.Ю. *(vlad@jet.msk.su)*
Технический редактор: Овчинникова Г.Ю. *(galya@jet.msk.su)*
Россия, 127015, Москва, Б. Новодмитровская, 14/1
тел. (095) 411 76 01
факс (095) 411 76 02
*email: JetInfo@jet.msk.su http://www.jetinfo.ru*

Подписной индекс по каталогу Роспечати 32555