Jet Info

ИНФОРМАЦИОННЫЙ БЮЛЛЕТЕНЬ № 1-2 (200)/2010

Современные виртуализированные системы хранения данных

ВЫЧИСЛИТЕЛЬНЫЕ КОМПЛЕКСЫ

Jet Info
ИНФОРМАЦИОННЫЙ БЮЛЛЕТЕНЬ
Издается с 1995 года

Редакция:

Дмитриев В.Ю.
vlad@jet.msk.su

Некрасова Н.А.
nekrasova@jet.msk.su

Слободчикова Т.А.
slobodchikova@jet.msk.su

Шедова Е.Л.
eshedova@jet.msk.su

Верстка:
Кулешова Ю.В.

Корректура:
Андрошко О.Ю.

Над номером работали:
Артемов С.В.
Викторов К.Г.
Голубев Д.Л.

Издатель:
Компания «Инфосистемы Джет»

Контакты:
тел. (495) 411 76 01
http://www.jetinfo.ru

СОДЕРЖАНИЕ

От редакции

Первый номер 2010 года посвящен теме современных виртуализированных систем хранения данных. Начать с обзора таких систем мы решили не случайно. К этому нас подтолкнула тенденция на рынке, которая сегодня становится все более заметной — построение системы хранения данных без потери в производительности, но по более низкой цене. В изменившихся экономических условиях компании все больше концентрируются не столько на развитии своих ИТ-инфраструктур, сколько на их оптимизации и повышении эффективности. На рынке востребованы максимально недорогие, но при этом обладающие достаточно высокой надежностью и масштабируемостью системы хранения данных.

И что самое интересное, сейчас действительно возможно получить максимум производительности от своих систем по сравнительно небольшой цене! Потому как на смену привычному ранее экстенсивному наращиванию объемов хранения приходят новые технологии и подходы: виртуализация СХД, дедупликация, многоуровневое хранение информации. Они становятся мощным средством преобразования инфраструктуры, которое обеспечивает ускорение развертывания приложений и постоянную доступность систем/ приложений/ данных, помогает устранить трудности, гарантирует сокращение расходов и повышение гибкости всей среды. И при этом не теряется общая производительность вычислительного комплекса. Весьма неплохая альтернатива дорогостоящим системам, не так ли?!

К тому же, стоит отметить, что спрос на такие решения только растет, а значит можно ожидать интересных идей от производителей и, возможно, новых красивых проектов, за реализацией которых будет следить наш «Jet Info» и своевременно информировать вас о последних веяниях мира высоких технологий.

Выгодных вам проектов!

С уважением, редакция JI

Трое в лодке

В связи с переходом корпоративных заказчиков на полностью виртуализованные ЦОД компании Cisco, NetApp (NASDAQ: NTAP) и VMware (NYSE: VMW) предлагают новое решение в области построения вычислительных комплексов — архитектуры, которые помогают сделать виртуализованные ЦОД более эффективными, динамичными и защищенными.

Не так давно вендорами была представлена всеобъемлющая защищенная многопользовательская архитектура, которая позволяет повышать информационную защиту облачных систем путем изоляции ИТ-ресурсов и приложений разных клиентов, структурных подразделений или отделов, использующих общую ИТ-инфраструктуру. В рамках сотрудничества компаний Cisco, NetApp и VMware реализуется модель совместной поддержки этих апробированных и предварительно испытанных архитектур, чтобы помочь заказчикам оперативно создавать единые виртуализованные инфраструктуры.

(подготовлено по материалам: http://www.netapp.com/ru/company/news/news-rel-20100126-cisco-vmware-ru.html)

Больше, лучше, масштабнее...

Blue Coat Systems, Inc. — мировой лидер в области производства решений для обеспечения безопасности и ускорения работы бизнес-приложений в корпоративных сетях — по итогам 2009 года признал компанию «Инфосистемы Джет» лидером продаж своих решений в России. Компания, обладающая высшим партнерским статусом (Blue Coat Systems Premier Partner), заняла первое место как по объему продаж, так и по количеству успешно завершенных проектов

«Решения Blue Coat очень востребованы, — говорит Кирилл Викторов, заместитель директора по развитию бизнеса Центра информационной безопасности компании «Инфосистемы Джет», — и кризис их популярность только усилил. В текущих рыночных условиях наши заказчики хотят быть уверены, что их корпоративный доступ в Интернет — безопасный, а ИТ-инфраструктура используется по максимуму. С таким инструментом, как Blue Coat, стало возможным обеспечить бизнес высочайшим уровнем контроля производительности и безопасности».

Компания «Инфосистемы Джет» первой начала продавать решения Blue Coat в России и выполняет полный цикл работ от проектирования решений на базе технологий вендора до их внедрения и технического сопровождения. А в 2008 году был выполнен первый в нашей стране территориально-распределенный проект на основе решений вендора, который является самым масштабным и на сегодняшний день.

ТЕМА НОМЕРА

Современные виртуализированные системы хранения данных

Сергей Артемов,
инженер по направлению Вычислительные комплексы,
компания «Инфосистемы Джет»

Тема виртуализации ресурсов современного вычислительного комплекса в последнее время является чрезвычайно популярной и освещается широко и разносторонне в различных изданиях по ИТ-тематике. Но большая часть рассматриваемых технологий и методов использования виртуализации, как правило, относится к серверным ресурсам: это аппаратная виртуализация серверов или организация множества виртуальных ОС на одном физическом сервере.

Системы хранения данных на этом фоне выглядят более консервативно — внутренняя архитектура множества популярных дисковых массивов (как уровня midrange, так и enterprise-уровня) принципиально почти не отличается от тех моделей, которые использовались до широкого распространения виртуализации в серверном мире. Конечно, архитектурные изменения дисковых массивов производятся, но, как правило, это эволюционные доработки — более производительные внутренние шины, более производительные внешние интерфейсы, больше возможностей по масштабированию решения и т.п. Но применение технологий виртуализации непосредственно в самих дисковых массивах позволяет получить тот же набор преимуществ, который предоставляет виртуализация серверов. В первую очередь, это более эффективное и экономное использование ресурсов, появление новых возможностей и уменьшение затрат на поддержание инфраструктуры СХД.

Данная статья посвящена обзору массивов, отличающихся от классических моделей, наиболее часто используемых при построении систем хранения данных в России. Их отличает принципиально иной уровень виртуализации данных внутри массива, что позволяет им предоставлять ряд сервисов, не реализуемых (или плохо реализуемых) в классических массивах. Кроме этого, большая часть массивов архитектурно реализована как массив из так называемого «commodity hardware» — широко распространенных и недорогих компонент. Использование commodity hardware позволяет получить более дешевые решения (по сравнению с массивом из компонент собственной разработки), не теряя при этом в производительности или функциональности системы.

В статье рассмотрены дисковые массивы компаний 3PAR, IBM, SUN и Compellent, их архитектурные особенности и программное обеспечение, а также возможности применения подобных массивов в проектных решениях.

Массивы компании 3PAR

Компания 3PAR (ранее известная как 3PARdata) основана в 1999 году тремя инженерами, ранее работавшими над серверами 3000 и 6000 серии в SUN Microsystems: Jeff Price, Ashok Singhal и Robert Rogers. Собственно, название компании это первые буквы фамилии или имени основателей (P — Price, A — Ashok и R — Rogers).

Массивы 3PAR являются пионерами технологии Thin Provisioning и являются, пожалуй, наиболее производительными массивами класса Midrange.

Архитектура и принципы работы массивов 3PAR

Thin Provisioning

Одной из основных отличительных особенностей массивов 3PAR является виртуализация дискового пространства внутри массива. Все диски в нем разбиваются на небольшие блоки (так называемые chunklets) размером 256 Мбайт. На основе chunklets создаются логические диски с заданными характеристиками (такими как тип RAID, ширина колонки в RAID и т.п.), причем пользователь не выделяет chunklets самостоятельно — этим занимается ПО массива, равномерно распределяя chunklets по дискам в пуле. Из полученных логических дисков администратор формирует логические тома, которые предоставляются потребителям дисковых ресурсов (см. рис. 1).

Теперь непосредственно о Thin Provisioning. Так как данные на дисках массив размещает самостоятельно, это позволяет ему выделять «надувные» тома, физически занимающие меньше места, чем заявлено в размерах тома. Основное отличие от прежних решений на эту же тему в том, что как только в этом томе кончится свободное **физическое** пространство, оно будет добавлено из свободных chunklets, совершенно прозрачно, без ущерба производительности массива и без многочисленных технических ограничений.

Это умение является одной из основных (и, пожалуй, главной) особенностей и позволяет массивам 3PAR получить несколько значительных преимуществ перед конкурирующими решениями:

1. Thin Provisioning позволяет значительно повысить эффективность использования дискового пространства массива. Так как дисковое пространство, как правило, покупается «на вырост» со сроком на несколько лет, это приводит к тому, что в среднестатистическом массиве используется менее 50% свободной емкости (см. рис. 2).

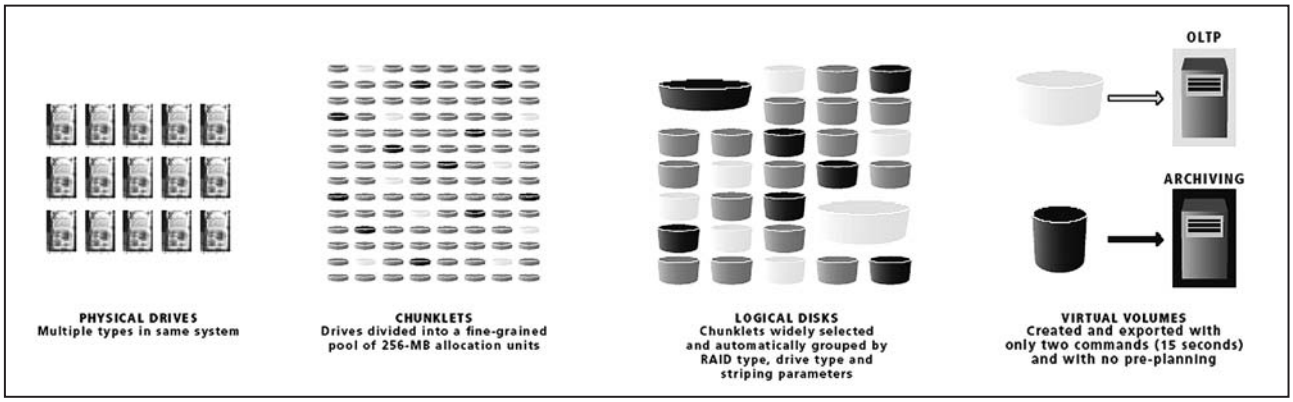

Рис. 1. Виртуализация дискового пространства в массивах 3PAR

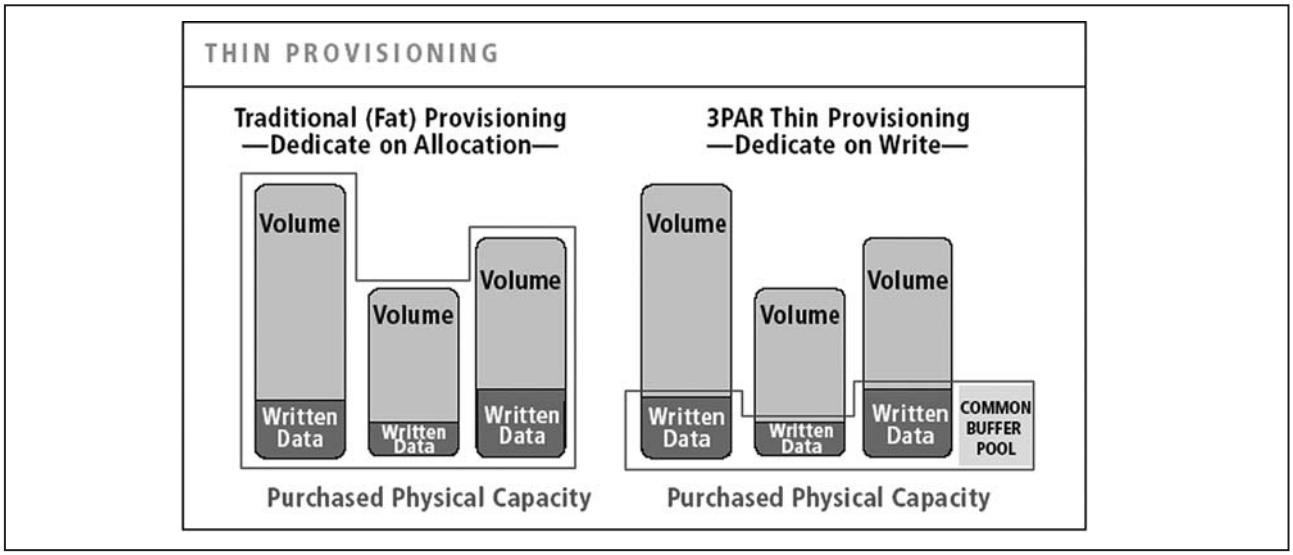

Рис. 2. 3PAR Thin Provisioning

2. Thin Provisioning **существенно** сокращает административные расходы на конфигурацию самого массива, а также на масштабирование дисковых ресурсов в дальнейшем. В классических массивах правильное выделение дискового пространства невозможно без четкого понимания его внутренней архитектуры и учета ее особенностей при разбиении массива. Незнание (или нежелание понять) архитектуру массива является типичным источником ошибок при его конфигурации, в результате которых одни компоненты перегружены, а другие — недогружены. В 3PAR таковой административной задачи не стоит — массив выбирает chunklets для томов самостоятельно, балансируя при этом нагрузку на backend. Со стороны администратора все выглядит очень просто — есть общий пул, из которого формируются тома требуемого типа (RAID-5, RAID-10 и т.п.). Увеличение дисковой емкости с Thin Provisioning также получается на порядок проще по сравнению с привычными решениями: нет необходимости создавать новые LUN, zoning, настраивать приложение для использования дополнительных файловых систем[1] (см. рис. 3).
3. И наконец, виртуализация дисков в массиве позволяет 3PAR динамически менять структуру RAID «на ходу». Например, если производительность уже используемого тома на RAID-5 формата 8D+1P неудовлетворительна, то можно динамически перестроить том RAID-5 формата 3D+1P или совсем переделать том в RAID-10. Главной особенностью является то, что при этом не производится никакой миграции данных, только пересчет четности (который обсчитывается специализированным ASIC). Разумеется, подобные трюки можно выполнять только в пределах одного пула, а точнее однотипных

REDUCED WORKLOAD: FOLLOW-ON PROVISIONING

STEP	Traditional (fat) Provisioning	3PAR Thin Provisioning	DESCRIPTION
1			Find an application that's running out of space or performance
2			Determine how to add storage (extend a LUN, make new LUN)
3			Determine how much storage to add
4			Determine where in pool to obtain additional storage from (taking performance and availability issues into account)
5			Allocate storage from pool
6			Set protection (LUN masks) of storage server so app can see it
7			Set switch zoning if necessary so app can see storage server
8			Add new storage to app LUN or form new app LUN
9			Configure O/S file system to handle bigger/new LUN
10			Configure application to utilize bigger/new LUN
11			Configure Backup system to back up bigger/new LUN
12			Replenish pool, if necessary (buy disks, build new RAIDsets)

Traditional Effort Required — Dramatically Simplified Effort — No Effort Required

Steps courtesy of Riche Lary, Tutelary, LLC

Рис. 3. Затраты административных ресурсов при Thin Provisioning

1 Разумеется, это верно только в том случае, если логический размер выделенного тома не требует изменений, а требуется только увеличение физического пространства в массиве.

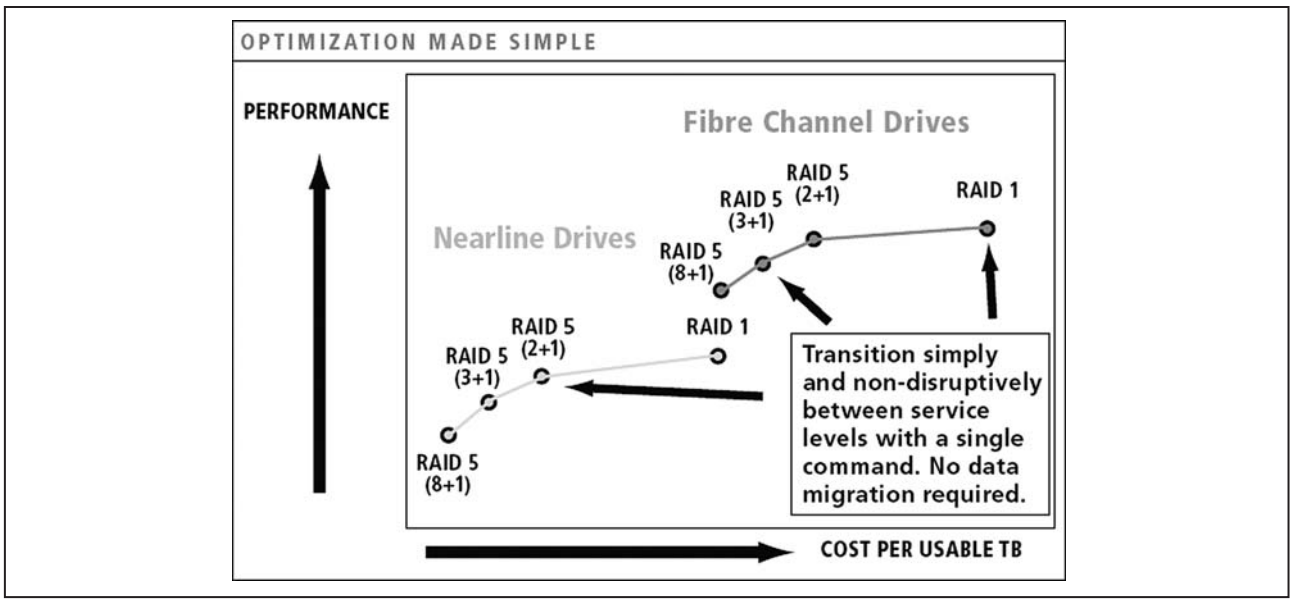

Рис. 4. Модификация структуры RAID в массивах 3PAR

дисков. Волшебным образом переместить том с SATA на FC, при этом поменяв структуру RAID, конечно же, нельзя (см. рис. 4).

Несмотря на привлекательность технологии Thin Provisioning, необходимо отметить, что не всегда ее применение позволяет получить экономию дискового пространства. Основной проблемой являются приложения, резервирующие для себя место на дисках и заполняющие его (свободное место) в дальнейшем. Как правило это различные системы управления базами данных, например, Oracle. При выделении ресурсов СУБД администраторы базы данных заранее создают табличные пространства требуемого размера, а также самостоятельно занимаются масштабированием по мере логического заполнения свободного места. Разумеется, при этом Thin Provisioning не приносит никакой пользы, так как с точки зрения массива все свободное место занято табличными пространствами (хотя с точки зрения Oracle они заполнены только частично).

Единственным решением в этом случае является применение опции Oracle Auto Extend (автоматическое увеличение размеров табличных пространств средствами Oracle). Но применение этой опции связано с возможными потерями в производительности СУБД, поэтому этот вариант должен согласовываться с администраторами Oracle, что приемлемо не во всех проектах.

Производительность и масштабируемость

Кроме нетипичной логической организации данных, массивы 3PAR отличаются от «классических» систем класса midrange тем, что прекрасно масштабируются по front-end. В отличие от привычной схемы с двумя контроллерами и общей шиной старшие модели, 3PAR поддерживают до восьми контроллеров, причем они объединены в пассивный full mesh backplane с хорошей пропускной способностью — 1.6 Гбайт/с между контроллерами (см. рис. 5). Наличие высокоскоростного backplane и большого количества контроллеров позволяет 3PAR очень выгодно отличаться от других midrange-систем. Производительность старшей модели 3PAR T800 по результатам тестов SPC-1 http://www.storageperformance.org конкурирует вовсе не с midrange-системами, а с high-end решениями, такими как массивы HDS USP-V (и что самое интересное 3PAR показывает более удачные результаты). Данные по количеству операций ввода/вывода в секунду по тестам SPC-1 представлены на рис. 6, точками на графиках обозначена загрузка массива в процентах, 10, 50, 80, 90,95 и 100 процентов соответственно.

Так же можно отметить очень высокую плотность размещения дисков в старших моделях 3PAR T400 и T800. Диски размешаются в дисковых шасси высотой 4 RU, при этом одно дисковое шасси содержит до 40 дисков (для сравнения обычная дисковая полка высотой 3RU содержит 15-16 дисков). Но подобная плотность паковки дисков несет с собой проблемы в обслуживании массива. Диски в шасси вставляются магазинами по 4 диска, поэтому для замены одного диска в 3PAR необходимо запустить процедуру миграции данных с магазина, и только после ее окончания можно вытащить магазин и заменить диск. Процедура миграции занимает достаточно много

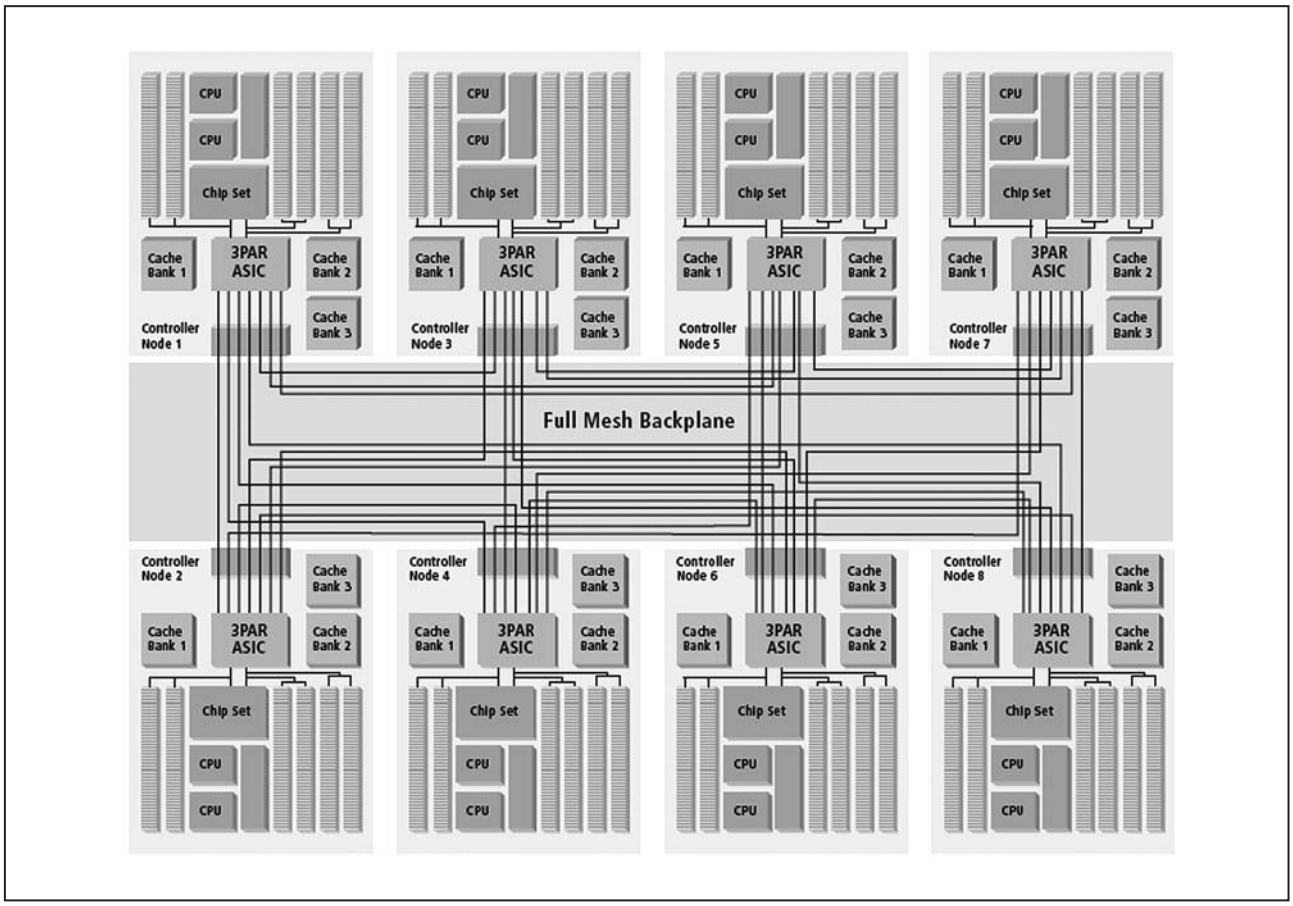

Рис. 5. Архитектура массивов 3PAR

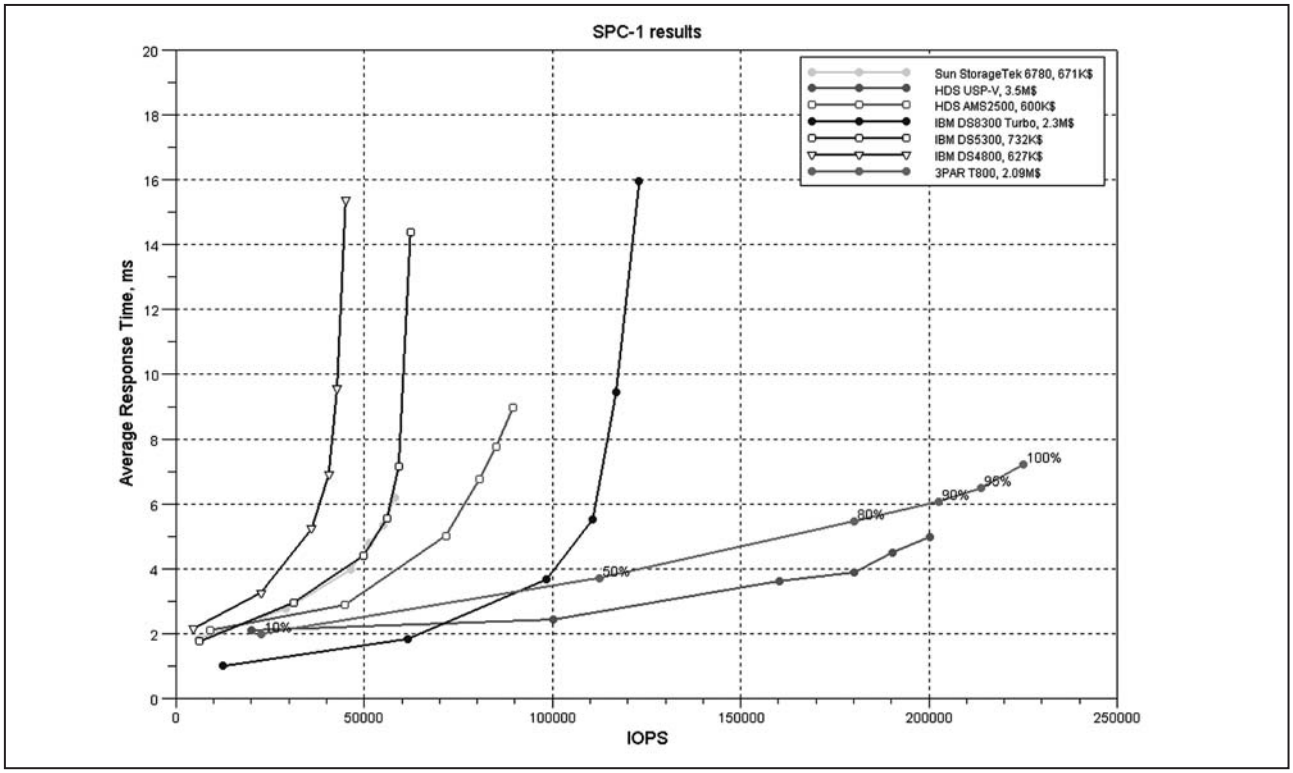

Рис. 6. Результаты теста SPC-1 для различных дисковых массивов

времени — от получаса до 2 часов. В принципе можно извлечь магазин без миграции, но если при этом в массиве откажет еще один диск, то данные будут потеряны.

На младших моделях массивов 3PAR F-класса (F200 и F400) паковка дисков более привычная — 16 дисков в 3RU шасси, и диски можно извлекать по одному.

Где можно применять массивы 3PAR

Несколько областей, где применение массивов 3PAR имеет видимые преимущества по сравнению с классическим midrange storage:

1. Как СХД для приложений, требования которых к производительности дисковой подсистемы находятся «на границе» между midrange и high-end оборудования (60000 — 80000 IOPS для нагрузки типа OLTP). В этом случае старшие модели 3PAR обеспечивают запас производительности, который позволит масштабировать массив, не меняя при этом саму модель[2].
2. Как СХД для приложений, гарантирующих пользователям фиксированный объем дискового пространства (например, почтовый сервер, домашние директории пользователей, WEB-хостинг и т.п.). В подобных проектах выделенное место расходуется далеко не сразу, и, зачастую, никогда не используется полностью. В этом случае Thin Provisioning в 3PAR позволяет значительно сократить затраты на дисковые ресурсы.
3. Как основное хранилище данных для проектов с VMWare. В этом случае применение Thin Provisioning весьма эффективно при выделении дисковых ресурсов под виртуальные машины. Место под них всегда выделяют «с запасом», который в 90% случаев никогда не будет использован. При использовании Thin Provisioning вполне реально получить большее количество виртуальных машин за те же деньги (см. рис. 7).
4. Как СХД проектов по Outsource и проектов с неясным прогнозом по требованиям с СХД. Так же как и в предыдущем случае 3PAR с Thin Provisioning позволит сократить или «растянуть» расходы на дисковую подсистему. В случае с outsource проектами 3PAR позволит сразу предоставить договорную диско-

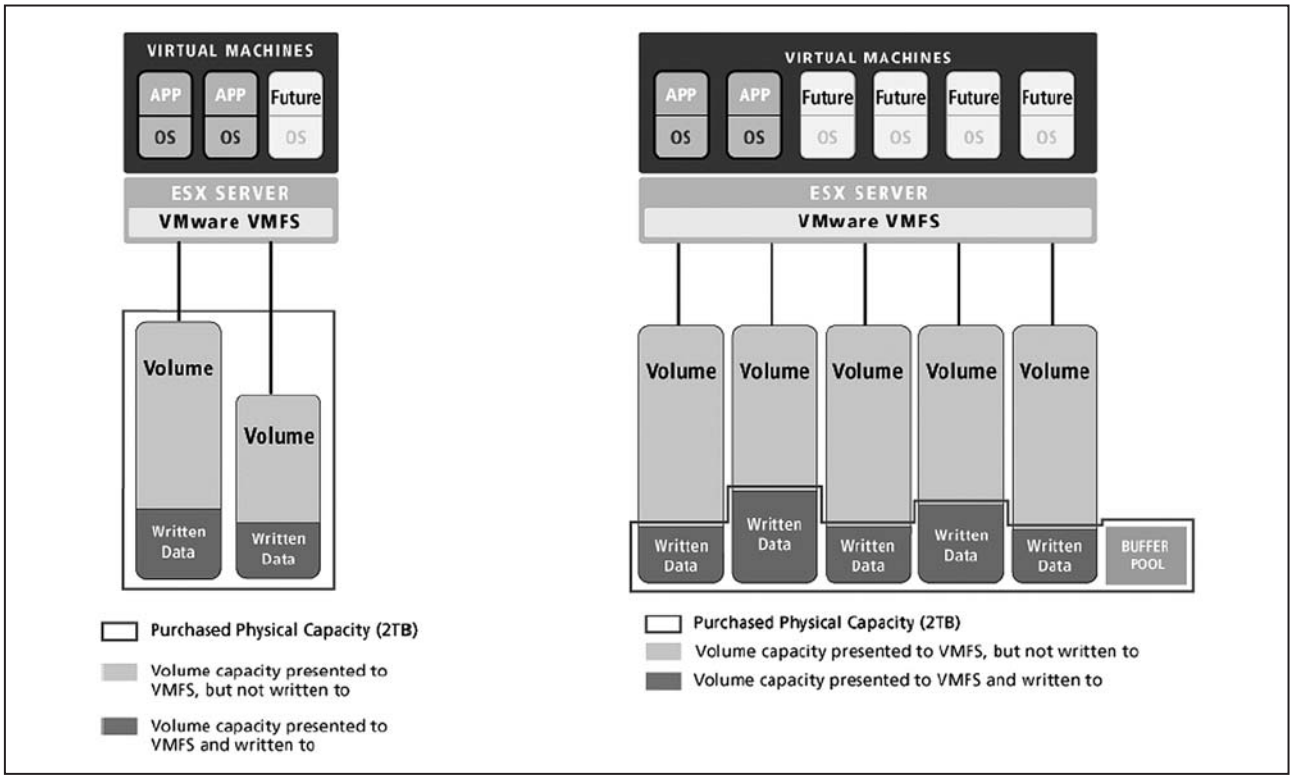

Рис. 7. Применение Thin Provisioning с VMware ESX Server

2 Необходимо учитывать, что массивы 3PAR не поддерживают подключение к mainframe по протоколам FICON/ESCON. Если есть требования на совместимость СХД с mainframe, 3PAR применять нельзя.

вую емкость приложению, докупая фактическую емкость по мере необходимости. И если требования к дисковому пространству клиентом были сформулированы «с запасом», то закупки фактического дискового пространства может и не потребоваться (или потребоваться сильно позже). В этом случае экономия на дисковых ресурсах существенно повышает прибыльность проекта. Схожая ситуация с проектами, где прогнозы по требованиям к дисковым ресурсам весьма приблизительны. Применение Thin Provisioning позволяет придерживаться стратегии «выделить по максимуму — купить по минимуму». В данном случае при любом развитии ситуации проблем не возникает. Если места было выделено недостаточно, массив 3PAR позволяет добавить дисковые ресурсы, не производя работ по масштабированию дискового пространства (так как размер томов изначально максимален). Если же прогноз по требованиям дискового пространства был завышен — налицо экономия бюджета проекта, так как ненужные диски не были закуплены.

5. При жестких требованиях к занимаемому месту в серверной. Из-за высокой плотности паковки дисков в старших моделях массивов 3PAR (Inserv Storage T400 или Inserv Storage T800) требования по площади могут быть заметно ниже по сравнению с другими массивами. Но необходимо учитывать, что вместе с этим требования к прочности пола в серверной сильно возрастают, а также усложняются процедуры обслуживания массива (замена вышедших из строя дисков).

Массив IBM XIV

IBM XIV Storage — это система хранения данных, купленная IBM в 2008 году. Идеологом XIV Storage является небезызвестный специалист по системам хранения данных Моше Янай (Moshe Yanai), ранее работавший в EMC и являвшийся одним из разработчиков линейки массивов EMC Symmetrix.

Основной парадигмой XIV Storage является попытка спроектировать массив, свободный от так называемых «bottleneck» или синдрома «бутылочного горлышка» — то есть ситуации, при которой недостаточная производительность одного из компонент массива приводит к общей деградации производительности всей системы хранения данных. Причем основными составляющими массива должны быть недорогие и широко распространенные компоненты — для обеспечения недорогого масштабирования мощности и общего снижения стоимости решения. Для этого архитектура XIV Storage использует принципы массового параллелизма при предоставлении ресурсов и виртуализации дискового пространства.

Архитектура и принципы работы XIV Storage

По мнению архитекторов XIV Storage, одним из основных недостатков современных массивов является их монолитная и закрытая архитектура, которая, тем не менее, весьма схожа для всех и состоит из ряда типовых блоков: внешних интерфейсов, кэш-памяти, RAID-контроллеров, дисков и внутреннего интерконнекта массива (см. рис. 8). Подобный подход к архитектуре сложился давно и обладает несколькими фундаментальными недостатками:

- Производительность решения является производной от быстродействия отдельных компонент массива. То есть рост производительности систем является экстенсивным — за счет увеличения скорости работы дисков, контроллера и т.д. Это, во-первых, дорого, так как требует постоянной разработки и внедрения в производство новых, более производительных компонент. Во-вторых, ни-

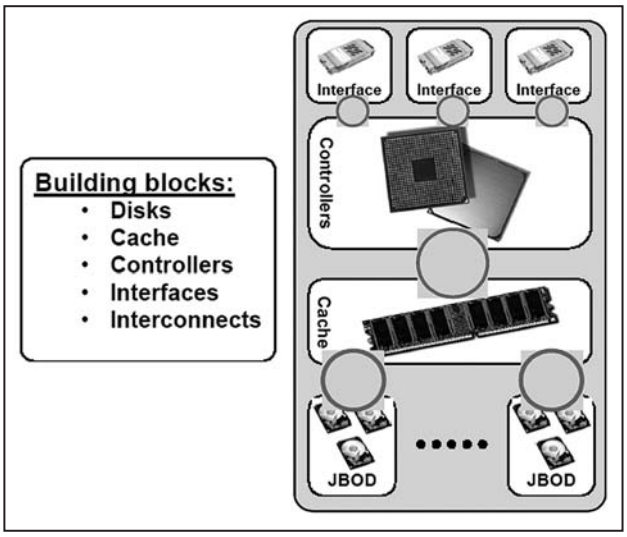

Рис. 8. Монолитная архитектура массива

как не гарантирует защиты от того, что производительность массива при определенных ситуациях (плохое проектирование при предоставлении дисковых ресурсов, выход из строя отдельных компонент или просто неудачное стечение обстоятельств) не будет ограничена производительностью одной из его компонент (например, одного из дисков).

- Подобные системы масштабируются в ограниченных рамках, заданных производителем решения. Так как система закрытая, увеличивать мощность решения можно только до определенного предела, после чего необходимо покупать СХД более высокого класса. Более того, масштабирование зачастую можно проводить только несимметрично — например, есть возможность расширения дисковой емкости без возможности увеличения числа контроллеров или их процессорной мощности и т.п.
- Несмотря на то, что функционально все типовые блоки массива похожи друг на друга, они, как правило, являются закрытыми (недоступными для производства сторонними компаниями) компонентами, а, следовательно, сами массивы являются дорогими решениями.

Проектировщики XIV Storage попытались создать решение, свободное от вышеописанных недостатков. Архитектурно XIV Storage представляет собой так называемый «grid storage», то есть массив, состоящий из множества типовых однообразных модулей, объединенных общим интерконнектом (см. рис. 9).

Модуль является его ключевым компонентом:он обеспечивает процессорные и дисковые ресурсы массива, кэш-память, а также интерфейс взаимодействия с серверами — потребителями дисковых ресурсов. Функционально модули в XIV Storage делятся на два типа:

- Data Module. Предназначены для хранения данных в XIV Storage;
- Interface Module. Предназначены для хранения данных и взаимодействия с серверами.

Каждый модуль представляет собой 2U сервер на базе процессора Intel QuadCore Xeon 2.33ГГЦ с 8ГБ RAM и 12x1 ТБ 7200 RPM SATA-дисками. Модель сервера для Data Module и Interface Module одинакова, они различаются лишь наличием FC-адаптеров и дополнительных Ethernet-адаптеров для Interface Module. В качестве операционной системы используется ОС Linux.

Все модули XIV Storage объединены через внутреннюю шину массива, построенную на базе 1Gb Ethernet. Локальная конфигурация модуля (внутренний адрес, имя модуля, адреса для доставки серверных сообщений и т.п.) хранится

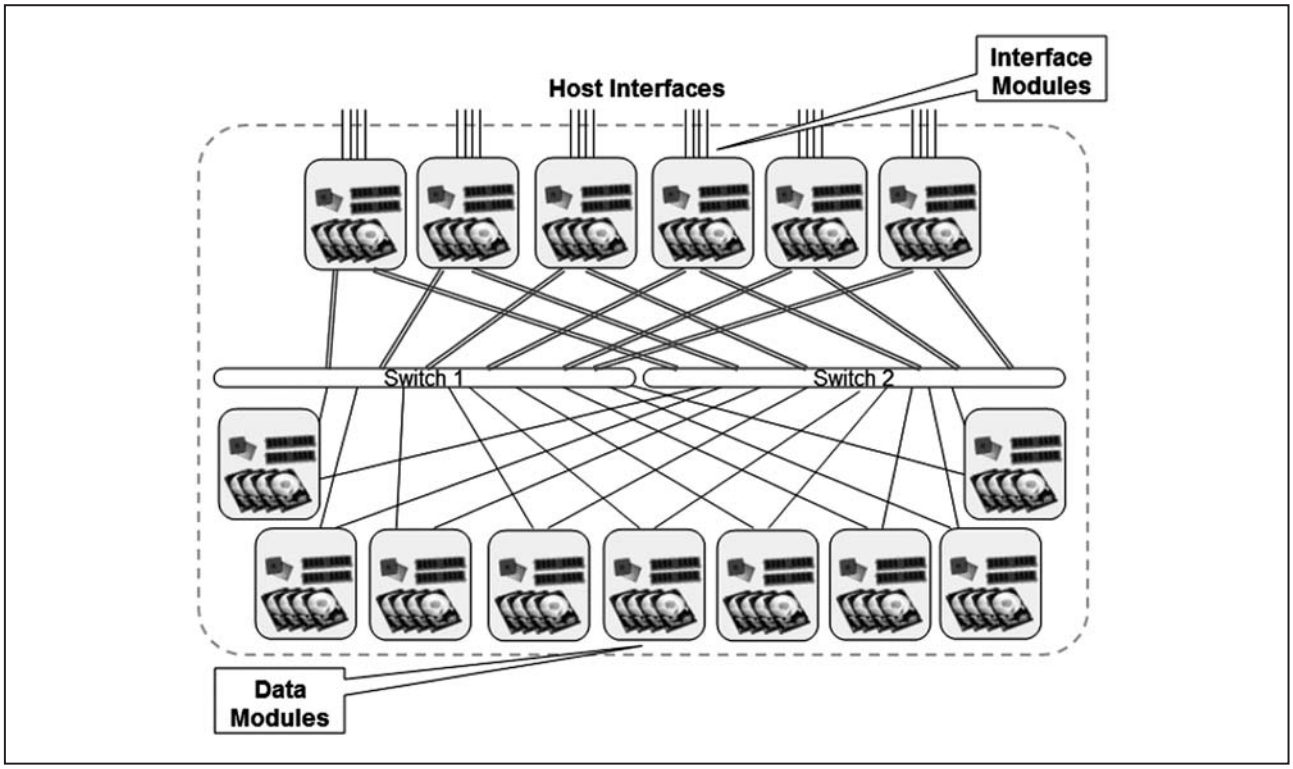

Рис. 9. Архитектура массива XIV Storage

на съемной флэш-карте на сервере, что позволяет оперативно заменять модуль в XIV Storage без переконфигурации.

Для защиты от перебоев подачи питания стойка с XIV Storage комплектуется тремя источниками бесперебойного питания (UPS). Система UPS предусматривает нормальную его работу в случае выхода из строя одного из трех модулей UPS. При потере питания массива емкости UPS хватает для того, чтобы перенести данные кэш-памяти XIV Storage (а фактически оперативной памяти серверов) на жесткие диски и корректно завершить работу массива.

Виртуализация дискового пространства в XIV Storage

Виртуализация дискового пространства в XIV Storage похожа на реализацию в массивах 3PAR (см. раздел «Thin Provisioning»), но имеет несколько отличий:

- Дисковое пространство разделено на кусочки еще более мелкого размера (в XIV Storage это называется «partition», (см. рис. 10)). Размер одной partition равен 1МБ, в то время как размер одного chunklet в 3PAR составляет 256 MB. Однако LUN в XIV Storage должны быть кратны примерно 17 ГБ (десятичных 17 ГБ, как написано в документации), несмотря на то, что каждый LUN собирается из chunklets размером 1 МБ. Число близкое к 17 ГБ выбрано, чтобы каждый LUN был равномерно распределен по всем дискам массива.
- В XIV Storage полностью отсутствует понятие дисков разного типа и разных RAID. Массив самостоятельно распределяет данные по всем дискам системы, используя собственный алгоритм. Подобную организацию данных XIV Storage называет RAID-X. По своей сути RAID-X не является RAID в классическом понятии. Данные (а именно chunklet по 1 МБ) распределяются таким образом, чтобы система могла перенести полную потерю одного модуля или диска XIV Storage. То есть каждый chunklet дублируется на диске в другом модуле. В случае потери диска происходит копирование тех chunklet, у которых отсутствует копия, на свободное место на дисках XIV Storage. Так как chunklet равномерно распределены по всем дискам массива, подобная операция занимает относительно небольшой промежуток времени (20-30 минут). После этого система снова функционирует в нормальном режиме, при потере еще одного диска производится такая же процедура.

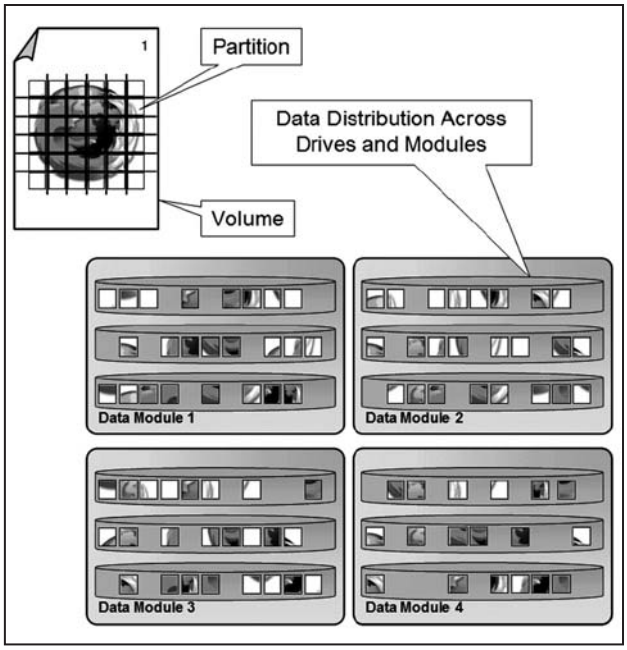

Рис. 10. Виртуализация данных в XIV Storage

Так же как и в случае с массивами 3PAR, виртуализация дисковых ресурсов позволяет XIV Storage реализовать следующие особенности:

- **Thin Provisioning.** Принципы работы этой технологии полностью аналогичны с описанной ранее для 3PAR (см. раздел «Thin Provisioning»).
- **Массовый параллелизм в обработке запросов.** Этот принцип работы массива являлся одной из основных идей при конструировании его архитектуры. Алгоритм разбиения данных на partition работает таким образом, что данные равномерно распределяются по всем дискам массива, тем самым гарантируя, что все компоненты при его работе нагружены одинаково. При добавлении новых ресурсов XIV Storage перераспределяет данные так, чтобы сохранялся принцип равномерной загрузки всех дисков в массиве. При этом сама архитектура массива обеспечивает постоянную и прогнозируемую производительность решения, без провалов в производительности по тем или иным причинам.

Моментальные снимки данных

Разбиение данных на partition такого маленького размера позволяет XIV Storage создавать моментальные снимки данных (далее snapshot), похожие на те, которые используются в NAS компании

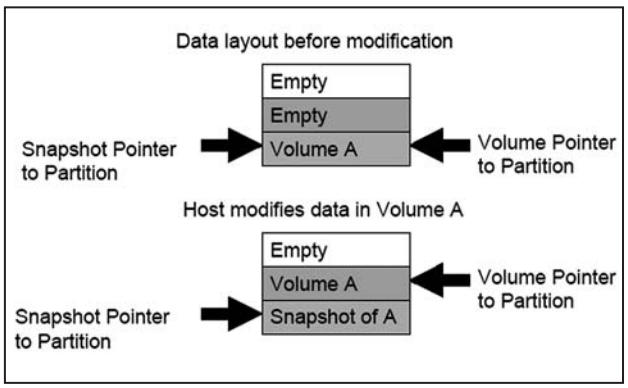

Рис. 11. Redirect on write snapshots в XIV Storage

NetApp. Их ключевой особенностью является то, что при создании snapshot используется не принцип «copy on write», а принцип «redirect on write».

При создании классического «copy on write» snapshot изменяемые данные предварительно копируются ПО массива в свободное дисковое пространство или выделенный пул. Таким образом, «цена» каждой операции записи, которая изменяет данные, возрастает в три раза — вместо одной операции массив вынужден сделать три: прочитать изменяемые данные, скопировать их в выделенный пул и после этого записать измененные данные. Разумеется, это происходит только при первом изменении данных (при повторном изменении копировать данные уже не надо), но тем не менее приводит к серьезной потере производительности массива при широком использовании моментальных снимков данных.

XIV Storage при работе со snapshot использует несколько другую технологию: измененные данные просто записываются в свободные chunklet, а XIV Storage просто модифицирует таблицу указателей в соответствии с изменившейся конфигурацией (см. рис. 11).

Подобная технология очень похожа на ту, что применяется в продуктах компании NetApp и имеет те же преимущества — создание и поддержание моментальных снимков в XIV Storage слабо влияет на производительность массива.

Где можно применять массивы XIV Storage

Наиболее логично использовать системы XIV Storage в тех же областях, где и системы 3PAR (см. раздел «Где можно применять массивы 3PAR», за исключением ситуаций с ограничением по серверным площадям). Но текущая реализация, несмотря на архитектурную оригинальность и привлекательность, имеет ряд недостатков, из-за которых применение XIV Storage «прямо сейчас» нежелательно. На то есть несколько причин, о которых далее и пойдет речь.

- RAID-X имеет по крайней мере две проблемы. Во-первых, система уязвима к потере двух дисков подряд. Если в момент перестроения RAID-X после потери одного диска из строя выходит еще один диск, данные будут с большой вероятностью потеряны. Конечно, вероятность одновременной потери двух дисков очень небольшая и подобная угроза существует и в классическом массиве. Но все же есть нюансы, из-за которых подобная ситуация в XIV Storage выглядит более угрожающе. Все диски в XIV Storage — только SATA, для которых в обычных массивах можно использовать RAID-6, который как раз и защищает от потери двух дисков. В XIV Storage такой возможности нет. Кроме этого, в обычном RAID количество дисков как правило невелико (5-10 дисков), и одновременный выход из строя нескольких дисков в массиве не означает того, что это случится именно в одном RAID. В RAID-X же объединены между собой все диски массива — 180 штук, и одновременная потеря нескольких из них — это всегда критическая ситуация[3]. Во-вторых, выход из строя нескольких дисков подряд может привести к потере данных в одном RAID, но данные в других RAID-группах останутся неповрежденными. В RAID-X, в силу его глобального характера, одновременная потеря двух или более дисков **означает полную потерю данных** всего массива.
- Преимущества Thin Provisioning в XIV Storage несколько меньше (по сравнению с тем же 3PAR) из-за того, что в нем нет альтернативы RAID-X. Понятно, что экономия дисковых емкостей в связке RAID-5 + Thin Provisioning выглядит гораздо убедительнее, чем RAIX-X (фактически «зеркало») + Thin Provisioning.
- Из всех массивов, существующих на данный момент, XIV Storage имеет самые высокие требования по электропитанию и охлаждению на единицу емкости. Это логично вытекает из его архитектуры — по сути это большое объединение серверов (иначе называемых модулями).

3 За исключением того варианта, когда все диски находятся в одном модуле.

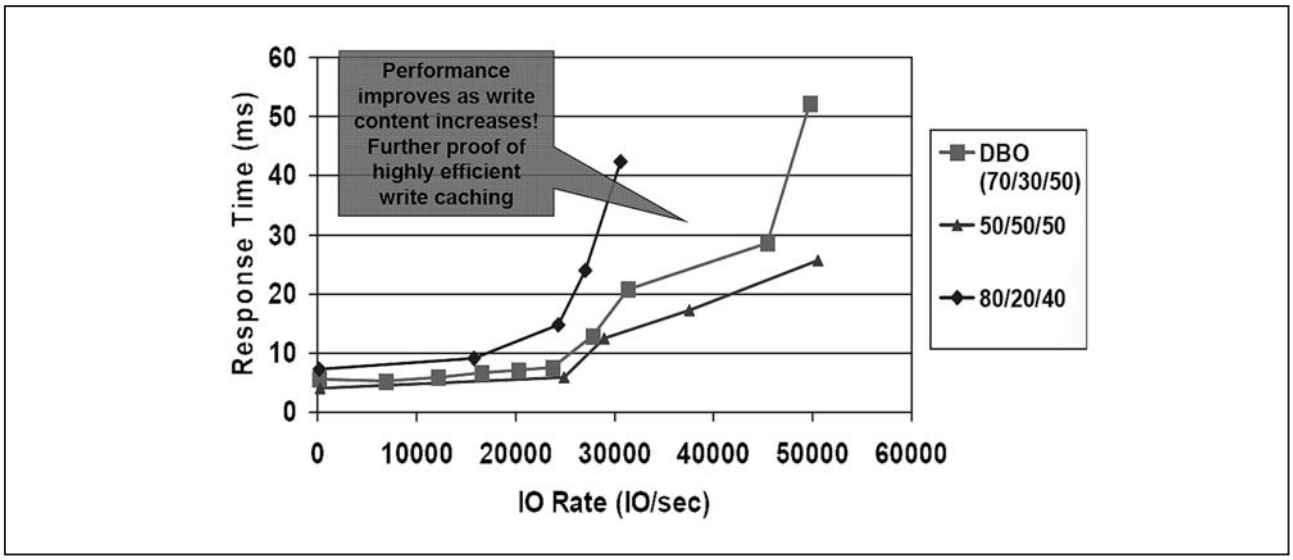

Рис. 12. Производительность XIV Storage для OLTP-нагрузки

- Производительность массива не так высока, как заявлялось при его представлении. Более-менее объективных тестов типа SPC-1 для XIV Storage на сегодняшний момент нет. Существуют внутренние тесты IBM, результаты которых приведены в документе «IBM XIV Storage Performance Review» от 19 марта 2009 года. Результаты поведения XIV Storage в максимальной конфигурации под OLTP-нагрузкой весьма скромные (см. рис. 12) и соответствуют производительности не самого мощного массива класса midrange (см. рис. 6, стр.9).
- ПО в XIV Storage на сегодняшний момент не поддерживает асинхронную репликацию между массивами. Также отсутствует интеграция ПО массива с распространенными системами резервного копирования, например, Symantec NetBackup.

В том виде, в каком XIV Storage существует сейчас, наиболее оптимально использовать его для хранения данных различных сред разработки. Наличие «недорогих» моментальных снимков данных позволяет разрабатывать и тестировать несколько версий приложений, не выделяя для каждой версии собственное дисковое пространство. Еще один из вариантов — это хранение данных для HPC-систем, где, как правило, нет высоких требований к надежности хранения данных и жестких требований к скорости работы дисковой подсистемы.

Но исходя из всего вышеперечисленного, логично подождать выхода следующей версии XIV Storage, которая будет более производительна (ожидается что backplane на основе 1 Гбит Ethernet будет заменен на Infiniband) и надежна (решена проблема с одновременным выходом из строя двух дисков в разных модулях). Сейчас это скорее интересная рабочая концепция массива из недорогих стандартных компонент, чем зрелое промышленное решение.

Массивы компании Compellent

Компания Compellent основана в марте 2002 инженерами Phil Soran, John Guider и Larry Aszmann (ранее бывшим CTO Xiotech).

Массивы Compellent Storage Systems также относятся к классу виртуализированных СХД и обладают теми же преимуществами, что и ранее описанные решения (например, Thin Provisioning). Отличительной особенностью решений Compellent является встроенная в массив технология иерархического управления данными или HSM (Hierarchical Storage Management). Массивы Compellent самостоятельно проводят динамическую оптимизацию размещения данных, мигрируя наиболее востребованные блоки данных на быстрые дисковые ресурсы. Те же блоки данных, обращения к которым производятся менее интенсивно, перемещаются на менее производительные RAID или более медленные диски. Так как информация о интенсивности обращений к тому или иному блоку данных содержится в его атрибутах, для организации HSM не требуется какого-либо дополнительного ПО, эта задача выполняется массивом самостоятельно.

Архитектура и принципы работы Compellent Storage Systems

Архитектурно массивы Compellent выполнены по вполне традиционной для midrange массивов технологии: два контроллера, объединенные в отказоустойчивый кластер и полки расширения с FC или SATA-дисками.

Контроллер представляет собой обычный 3 RU сервер на базе процессора Intel Xeon 3.2 ГГЦ. Контроллеры полностью самостоятельны, не объединены в один бокс и могут быть разнесены на расстояние до 300 метров друг от друга. Контролеры объединены в HA-кластер и имеют зеркалированный кэш достаточно скромного объема — 2.25 или 3.5 ГБ в зависимости от модели. В случае отключения питания данные кэша поддерживаются от батарейки, допустимая длительность работы составляет 72 часа. В качестве управляющей ОС Compellent использует ECos.

Сами контроллеры дисков не имеют, к ним по FC-AL подключатся полки расширения. Полки расширения бывают двух типов: SBOD FS (SBOD — Switched Bunch of Disks, полка со встроенным FC коммутатором) и JBOD SATA. В максимальной конфигурации массив имеет 9 Back-End FC петель, к каждой петле допускается подключить до 7 SBOD FC или до 5 JBOD SATA полок. Массив поддерживает RAID-0, RAID-5 и RAID-10.

Compellent не предоставляет в открытом доступе подробной информации о hardware или software компонентах массивов, поэтому точно сказать о моделях сервера и устройстве полок расширения затруднительно.

Dynamic Block Architecture

Основным отличием массивов Compellent от других массивов является так называемая Dynamic Block Architecture. Так же как и в ранее описанных решениях, дисковое пространство в Compellent разбивается на блоки. Но кроме собственно данных, каждый блок в массиве содержит и метаданные, относящиеся к этому блоку (см. рис. 13):

- время записи;
- время доступа;
- частота доступа к блоку;
- том, к которому принадлежит этот блок;
- тип диска, на котором лежит этот блок;
- тип RAID, которому принадлежит этот блок.

При каждом обращении к блоку или его изменении массив соответствующим образом обновляет метаданные этого блока.

Automated Tiered Storage

Именно наличие технологии Dynamic Block Architecture позволяет массивам Compellent реализовать свою основную особенность — автоматическую иерархическую структуру хранения данных или Automated Tiered Storage. Основной принцип работы Automated Tiered Storage (далее ATS) заключается в следующем: массив автоматически мигрирует данные по заданным пользователям правилам, основываясь на информации из метаданных блоков. Фактически администратор массива определяет несколько иерархических уровней хранения данных (по типу дисков, RPM дисков, типу RAID), а массив автоматически поддерживает заданную иерархию, мигрируя наиболее интенсивно запрашиваемые блоки данных на самые быстрые устройства хранения и наоборот.

Так как LUN в Compellent не имеет жесткой привязки к конкретным блокам данных, все действия, связанные с миграцией данных, происходят прозрачно для приложения и без участия дополнительного ПО.

Thin Provisioning и моментальные снимки данных

Эти технологии в Compellent очень похожи на конкурентов и подробно их описывать не имеет смысла. Необходимо только отметить, что моментальные снимки данных в Compellent используют технологию «redirect on write» (см. раздел «Моментальные снимки данных», стр. 14), что делает их несколько привлекательнее по сравнению с моментальными снимками в 3PAR. Кроме того,

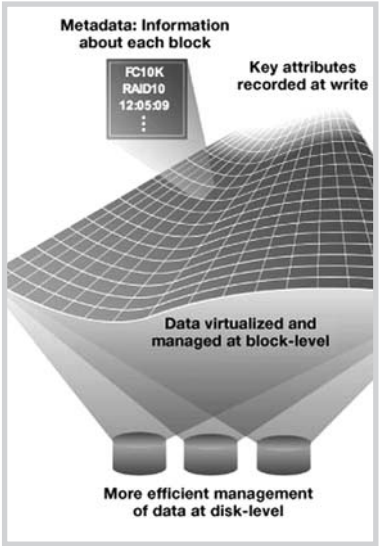

Рис. 13. Dynamic Block Architecture в Compellent

Compellent «умеет» убирать пустые блоки при импортировании томов с обычных массивов (тот же 3PAR анонсировал эту возможность относительно недавно).

Где можно применять массивы Compellent

Никаких официальных данных по производительности решений от Compellent нет (за исключением статьи Open Magazine от 2006 года), тестов SPC-1 для моделей этих серий тем более, поэтому сказать что-то определенное про эти массивы трудно.

Так как массив использует только два контроллера (причем младшая модель имеет single-core CPU), производительность явно не очень высокая. К тому же сама технология Automated Tiered Storage должна вносить свой вклад в понижение производительности — при любом обращении к блоку данных должны обновляться его метаданные, что, конечно, не ускоряет работу. Кроме этого, в массиве постоянно происходит фоновый процесс миграции данных — из-за того же ATS. То есть Compellent — явно не высокопроизводительное и ограниченно масштабируемое решение (на возможную емкость массива в 1008 дисков двух контроллеров недостаточно). Несколько настораживает и инсталляционная база — с 2002 года своего существования Compellent поставила массив 1300 клиентам[4], что за 7 лет не очень много.

Та ниша, куда явно «напрашиваются» эти массивы — хранилища данных, где применение HSM желательно, но зачастую невыгодно из-за накладных расходов на покупку, внедрение и поддержку подобных решений. Например, СХД для электронной почты, домашние директории пользователей, хранилища документации и т.п.

Sun Unified Storage

Унифицированные системы хранения Sun Storage 7000 — серия NAS-массивов, основанная на программном обеспечении Sun OpenStorage (проект FISHworks, задуманный как потенциальный «убийца» NetApp). Упор делался на открытость и доступность решения, и в основу Sun Unified Storage легли операционная система OpenSolaris, сопутствующая ей файловая система ZFS и службы сетевого доступа. Дополнив ОС средствами для управления и мониторинга, специфичными для массива, а также расширениями, модифицирующими поведение ZFS, компания SUN получила программное решение, которое, в противоположность традиционным массивам, дает возможность клиенту самостоятельно выбирать серверную платформу. Программный продукт Open Storage, как и любой открытый продукт, можно скачать и установить на собственные серверы, и Sun предлагает несколько готовых конфигураций на основе собственных стандартных компонент. Пока массивы поддерживают только сетевые протоколы доступа (NFS, CIFS, HTTPS/WebDAV, FTP/SFTP/FTPS, iSCSI), поддержку FC-Target Sun обещает добавить в скором будущем.

Архитектура и принципы работы Unified Storage

Как уже говорилось, в Open Storage использует файловую систему ZFS и стандартные коммуни-

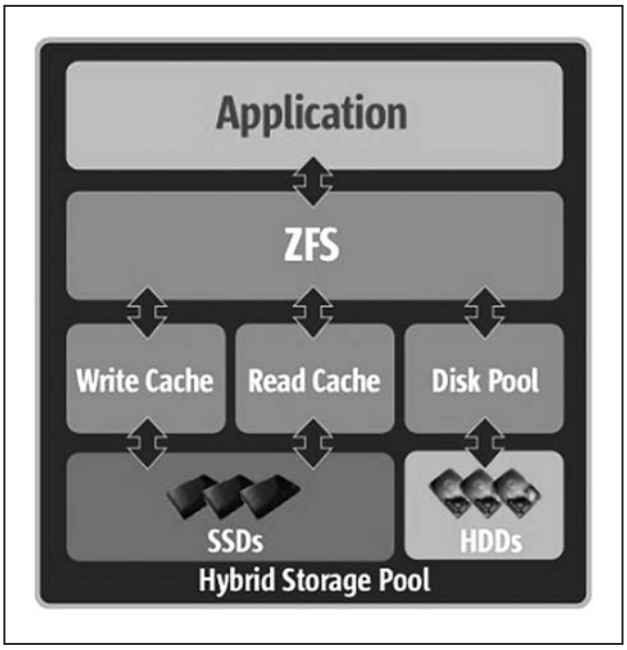

Рис. 14. Архитектура Unified Storage

4 По данным сайта www.compellent.com

кации OpenSolaris. ZFS объединяет в себе менеджер томов и файловую систему, собирая диски в пулы, позволяя формировать тома RAID-0, RAID-1 и RAID-Z (RAID с переменным числом колонок и избыточной четностью), создавать моментальные снимки и проводить репликацию данных. Для выпуска полноценного массива перед разработчиками стояли три задачи: добавление энергонезависимой кеш-памяти, реализация кластерной конфигурации и разработка средств управления и мониторинга. Модернизация кэширования данных пошла по пути внедрения многоуровневой системы хранения, что привело к вытеснению кэша на твердотельные носители (Solid State Disk, далее SSD) и созданию гибридных пулов.

Hybrid storage pools

Гибридный дисковый пул (Hybrid Storage Pool), добавленный к файловой системе ZFS, позволяет на стандартных компонентах, без применения заказных схем и адаптеров, реализовать управление дисковой подсистемой массива.

Технология объединяет SSD-накопители и классические жесткие диски в единый пул. Если классическая технология построения дисковых массивов предполагает использование оперативной памяти (кэша) в качестве быстрого временного хранилища данных, то гибридный дисковый пул добавляет между оперативной памятью и дисками еще один уровень кэширования (см. рис. 15).

ZFS располагает кэшем в оперативной памяти — Adaptive Replacement Cache (ARC), но расширение Readzilla, написанное специально в рамках проекта Open Storage, наиболее часто используемые блоки данных дублирует на быстродействующих SSD-накопителях (создает кэш второго уровня — L2ARC). Другое расширение — Writezilla — перемещает журнал транзакций ZFS (ZFS Instant Log, сокращенно — ZIL) из оперативной памяти на SSD-накопители, что избавляет от необходимости применять сложные аппаратные решения для дублирования (или поддержания питания) модулей оперативной памяти, содержащих несохраненные на диски данные.

Конечно, твердотельные диски обладают большим временем доступа, чем оперативная память, но этот недостаток должен компенсироваться значительно возросшими объемами кэша (сотни гигабайт) по сравнению с объемом традиционного кэша контроллеров дисковых массивов среднего класса (единицы гигабайт).

На рисунке рис. 16 показаны снимки экрана системы мониторинга массива, демонстрирующие показатели производительности массива (140 дисков, 128 ГБ оперативной памяти) до и после добавления 600 ГБ SSD кэша. Более темные участки графика означают более плотное распределение.

Утверждается, что путем совместного использования разнородных носителей, достигается практически в два раза большая производительность при в два раза меньшей стоимости. Гибридная архитектура также способствует повышению энергоэффективности. В частности, применение SSD-носителей в качестве кэша большого объема позволяет для основного хранилища использовать энергоэффективные жесткие диски с меньшей скоростью вращения (4000 об/мин вместо 10000 или 15000 об/мин).

Производительность и масштабируемость

Увеличение производительности в SUN Unified Storage должно обеспечиваться путем добавления оперативной памяти (а в старшей модели и процессоров в контроллер) и твердотельных накопителей в качестве кэш-памяти второго уровня. Ка-

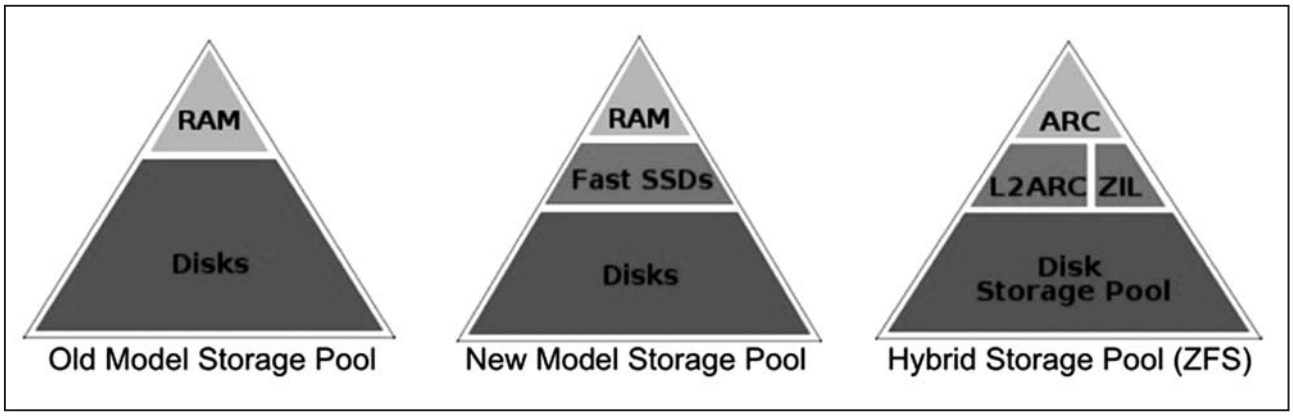

Рис. 15. Представление различных моделей кэширования

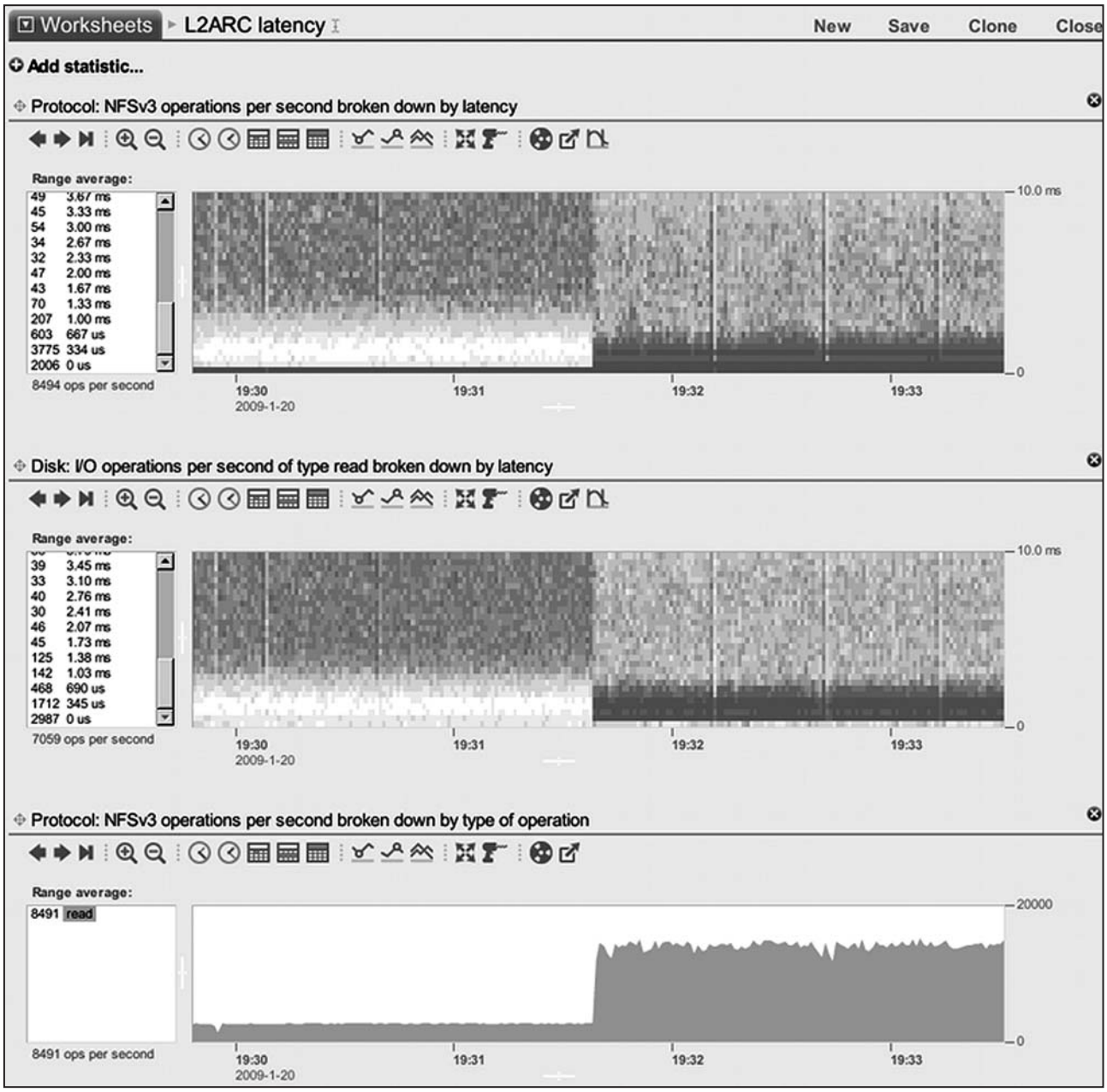

Рис. 16. Показатели производительности при добавлении L2ARC17

жется, что большим ограничением как гибкости, так и надежности, является то, что у контроллера может быть только один Z-пул. Судя по всему, массивы пока не рассчитаны на большие объемы.

Следующим этапом расширения является объединение двух контроллеров в кластер. В ZFS не предусмотрено одновременное использование дисков несколькими серверами, и параллельной обработки запросов нет (это находится в планах развития), поэтому каждый контроллер оперирует только своим Z-пулом. С одной стороны, это не позволяет балансировать нагрузку, но с другой — не требуется синхронизация кэша, и обмен данными между контроллерами минимальный. Массив, состоящий из двух контроллеров, по сути представляет собой HA-кластер. Первое поколение Sun Unified Storage обходилось обычными сетевыми интерфейсами, во втором добавилась специальная плата (CLUSTRON), содержащая два порта RS232 и один 1 Gbit Ethernet. Модуль CLUSTRON — единственный нестандартный компонент в массиве. Все три интерфейса используются только для проверки состояния соседнего контроллера, данные через платы CLUSTRON не передаются. При сбое одного из контроллеров, дисковый пул импортируется на оставшийся контроллер. Кэш каждого пула, находящийся на SSD-дисках, «перемещается» вместе с остальными дисками.

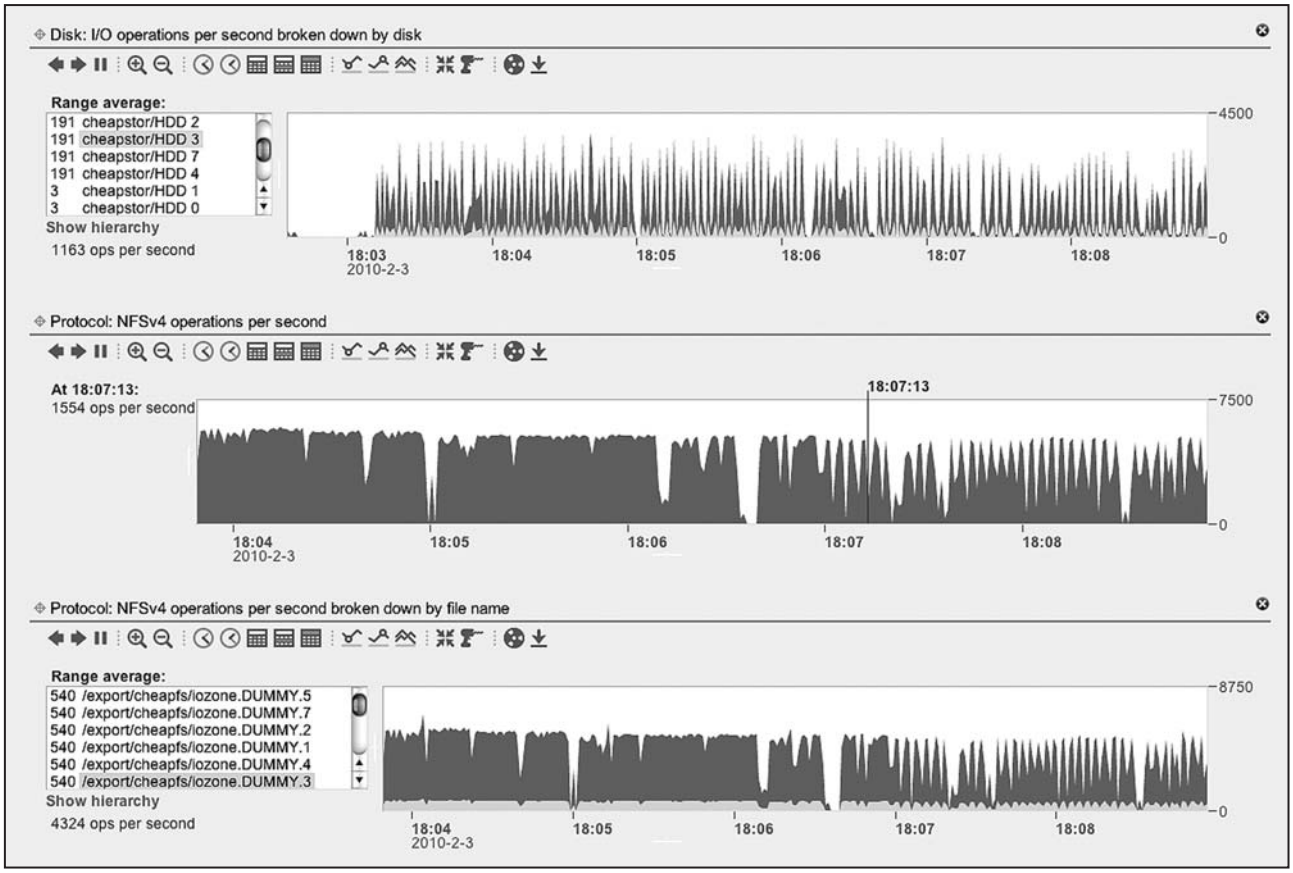

Рис. 17. Пример графика системы мониторинга

Thin Provisioning, моментальные снимки и репликация данных

Файловые системы, как это принято в ZFS, не имеют постоянного размера, физически они занимают столько же места, сколько занимают данные, а максимальный размер ограничен квотами или размером Z-пула. Моментальные снимки файловых систем, которые могут быть доступны и на запись, также не занимают постоянный объем: копируются только измененные блоки. Создавать снимки можно или по расписанию, или через равные промежутки времени, или вручную. Количество же копий ограничено лишь объемом доступного дискового пространства.

Возможна репликация файловых систем (или томов) на другой дисковый массив. Репликация может быть единовременной (запущенной вручную или по расписанию) и непрерывной асинхронной. И в том, и в другом случае данные передаются с использованием протокола ssh. Направление репликации может быть изменено.

Средства управления и мониторинга

Управление массивом может осуществляться как через web-приложение, так и из командной строки. Оба интерфейса обладают полным функционалом. Командная строка похожа на ILOM; web-интерфейс, если не считать того, что загрузка массива измеряется не цветовой шкалой, а погодными условиями (солнечно, облачно и т.д.) — не представляет из себя ничего особенного. Интерес представляет система мониторинга с возможностями аналитики и наблюдения в реальном времени. Массив с секундным интервалом собирает статистические данные и хранит их вечно. Под статистиками понимаются как простые (например, загрузка процессора или количество операций в секунду), так и рассортированные или агрегированные данные (например, количество операций по клиентам и по протоколам). Развернутое управление статистиками и их графическим отображением позволяет строить свои представления данных. Особо отмечается, что можно изучать производительность ввода/вывода с точностью до конечного файла в файловой системе.

Где можно применять массивы Sun Unified Storage

Массивы Sun Unified Storage, в первую очередь, привлекательны своей ценой, а выход на рынок уже второго поколения говорит о том, что «детские болезни частично вылечены» (первое поколение имело недостатки в плане совместимости и интеграции). В ближайшее время будет добавлен FC-target, и массивы станут хорошим антикризисным решением начального уровня. Большой плюс систем Sun Unified Storage в том, что все функции массива бесплатны, и не надо платить за возможность репликации, создание моментальных снимков, за добавление компрессии, связи с антивирусом, систему управления производительностью.

Во вторую очередь, архитектура с многоуровневым кэшем в ряде случаев может позволить при низкой цене решения достигнуть производительности массивов ценового диапазона классом выше.

Большой минус системы SUN Unified Storage — отсутствие интеграции с системами резервного копирования. Но с этой проблемой сталкиваются все новые решения в области СХД, и если массив оказывается удачным и популярным, поддержки в основных системах СРК долго ждать не приходится.

Заключение

Как видно из обзора, все рассмотренные выше дисковые массивы имеют как сильные, так и слабые стороны по сравнению с «классическими» и широко распространенными моделями. Но, тем не менее, тенденция на более высокий уровень виртуализации ресурсов внутри массива явно прослеживается. Большинство ведущих вендоров СХД применяют (или внедряют) технологии, которые предоставляет виртуализация СХД, например, Thin Provisioning. Без архитектурных изменений массива это выглядит не так эффектно, но, тем не менее, все чаще используется в реальных рабочих системах хранения данных.

Целесообразность применения вышеописанных моделей дисковых массивов в проектных решениях напрямую зависит от конкретных требований к характеристикам СХД. Если «плюсы» от использования того или иного массива перевешивают его потенциальные «минусы», то почему бы и нет? Точно сказать, когда и в каких случаях использовать вышеописанные массивы, конечно, нельзя (тем более, что наиболее выигрышная, по нашему мнению, область применения каждого массива была приведена в его описании), но можно сделать общие выводы по каждому из них.

Из рассмотренных выше моделей явно стоит поближе рассмотреть массивы компании 3PAR. Возможно их применение как альтернативы mid-range storage от HDS или EMC, причем альтернатива более масштабируемая и производительная. Также возможно применение этих массивов как компонентов антикризисных решений — за счет встроенной технологии Thin Provisioning.

IBM XIV Storage выглядит интересно, но хотелось бы дождаться следующей версии уже от IBM, и только после этого можно делать выводы о возможностях этих массивов. В текущем состоянии — это скорее рабочая модель, а не зрелое промышленно решение.

По массивам компании Compellent, к сожалению, слишком мало информации, чтобы делать какие-то определенные выводы. Если вы планируете использовать SSD-диски, а также требуете максимально эффективного использования дисковых ресурсов, то массивы Compellent выглядят весьма привлекательно.

Массивы Sun Unified Storage — вариант антикризисного NAS от SUN. Если применение NetApp невозможно по ценовым характеристикам, то Sun Unified Storage является интересной альтернативой.

Список используемой литературы

При подготовке материала использовалась следующая литература:

- «IBM XIV Storage System: Concepts, Architecture, and Usage», IBM Redbook;
- «Storage IBM XIV Storage System Theory of Operation», IBM Guide;
- «IBM XIV Storage System (Type: 2810) Model A14 (Gen 2) Introduction and Planning Guide for Customer Configuration», IBM Guide;
- «IBM XIV Storage Performance Review», IBM System Storage Technical Users Group Webinar, Lee LaFrese, March 19, 2009;
- «The 3PAR InSpire Architecture. Returning Simplicity to IT Infrastructures», 3PAR White Paper;
- «3PAR INSERV STORAGE SERVERS F200, F400 TECHNICAL SPECIFICATIONS», 3PAR Datasheet;
- «3PAR INSERV STORAGE SERVERS T400, T800 TECHNICAL SPECIFICATIONS», 3PAR Datasheet;
- «3PAR Thin Provisioning: Eliminating Allocated but Unused Storage and Accelerating ROI», 3PAR White Paper;
- «When is Thin not really Thin?», 3PAR Vendor Tutorial;
- «ESG Storage Innovations Series Focus on 3PAR», ESG Report;
- «Simplified Database Storage Management That Lowers Management Costs And Yields High Storage Utilization», Oracle & 3PAR White Paper;
- «3PAR InServ T-Class Hand-on-Lab», 3PAR Lab Materials SNW2008;
- «Designing a Persistent Hardware Architecture», Compellent White Paper;
- «Compellent: Intelligently Managing Inside the Volume», Tanega Group Technology Validation;
- «Technical Specifications», Compellent Datasheet;
- «Compellent — Harnessing SSD's Potential», ESG Report;
- «Unmatched Investment Protection In a Single Scalable Storage Platform», Compellent White Paper.

Перечень принятых сокращений

Сокращение	Полное наименование
ЛВС	Локальная вычислительная сеть
ПО	Программное обеспечение
СХД	Система хранения данных
СУБД	Система управления базами данных
FC	Fiber Channel
FC-AL	Fiber Channel Arbitrated Loop
HDD	Hard Disk Drive
RU	Rack Unit
RPM	Rotation Per Minute
HBA	Host Bus Adapter
LAN	Local Area Network
SAN	Storage Area Network
GE	Gigabit Ethernet
RAID	Redundant Array of Independent Disks
ATS	Automated Tiered Storage
HA	High Availability
SSD	Solid State Disk
ASIC	Application Specific Integrate Circuit
ARC	Adaptive Replacement Cache
SBOD	Switched Bunch of Disks
JBOD	Just Bunch of Disks
IOPS	Input/Output per Second
OLTP	Online Transaction Processing

Jet Info
ИНФОРМАЦИОННЫЙ БЮЛЛЕТЕНЬ
Издается с 1995 года
Издатель: компания «Инфосистемы Джет»

Главный редактор: Дмитриев В.Ю.
Редактор: Слободчикова Т.А.
Россия, 127015, Москва, Б. Новодмитровская, 14/1
тел. (495) 411 76 01
факс (495) 411 76 02
email: JetInfo@jet.msk.su http://www.jetinfo.ru

Подписной индекс по каталогу Роспечати 32555