



# **Основы проектирования SAN**

Джош Джад

**Второе издание  
(русское, v.1.0)**

**Copyright © 2005 - 2008  
Brocade Communications Systems, Inc.**

*Все права защищены. Запрещается воспроизведение и передача любой части этой книги любым способом, в том числе электронным, механическим, магнитным, фотографическим, включая ксерокопирование, запись на любые системы хранения и извлечения информации, без письменного разрешения Brocade. Brocade не несет никакой ответственности за возможные нарушения патентов в этой книге. Издатель, автор и Brocade не несут ответственности за возможные ошибки и неточности, содержащиеся в этой книге, а также за убытки из-за использования содержащейся в ней информации. Материал книги может быть изменены без предварительного уведомления.*

**Brocade Bookshelf™**

Серия разработана Джошом Джадом

**Основы проектирования SAN**

Автор: Джош Джад

Редакторы: Дэниел Крюгер, Кент Хенсон и Джош Джад

Рецензенты: сотрудники отделов Маркетинга и Разработки Brocade

Обложка второго издания разработана Маркетингом Brocade

Русский перевод: Лев Левин

Правка русского текста: Николай Умнов, Павел Добринский, Brocade

**Издания**

“Advance Edition” -июнь 2005

“First Edition” – первое издание август 2005,

исправленное издание сентябрь 2005 и июль 2006

“Second Edition” – первое издание август 2007

исправленное издание март 2008

русское – декабрь 2008

*Издательство:*

**INFI<sup>∞</sup>ITY**  
PUBLISHING.COM

1094 New Dehaven St.  
West Conshohocken, PA 19428  
[info@InfinityPublishing.com](mailto:info@InfinityPublishing.com)  
[www.InfinityPublishing.com](http://www.InfinityPublishing.com)  
[www.BuyBooksOnTheWeb.com](http://www.BuyBooksOnTheWeb.com)  
Toll-free: (877) BUY-BOOK  
Local Phone: (610) 941-9999  
Fax: (610) 941-9959

# **Юридическая информация**

---

**Copyright © 2005 - 2008 Brocade Communications Systems, Inc.**

Brocade, Fabric OS, File Lifecycle Manager, MyView и StorageX являются торговыми марками, а символ Brocade B-wing, DCX и SAN Health – торговыми марками Brocade Communications Systems, Inc., в Соединенных Штатах и/или других странах. Все другие бренды, названия продуктов и сервисов являются или могут являться торговыми марками и марками сервиса соответствующих компаний и используются только для идентификации их продуктов и сервисов. Названия продуктов в данной книге могут отличаться от действующих – текущий список продуктов см.: <http://www.brocade.com/products-solutions/products/index.page>

**Примечание:** Использование этой книги означает согласие на следующие условия. Эта книга поставляется “**КАК ЕСТЬ (AS IS)**” исключительно как для информирования без каких-либо прямых или косвенных гарантий относительно какого-либо оборудования, функциональности, сервисов которые предлагаются или будут предлагаться Brocade. Brocade оставляет за собой право делать изменения в этом документе в любое время без предварительного уведомления и не несет никакой ответственности за его использование. Этот информационный документ описывает функции, которые могут быть недоступны в настоящий момент либо не будут доступны никогда. Узнать о доступности функций и продуктов можно в торговом представительстве Brocade. На экспорт технических данных, содержащихся в настоящем документе, может потребоваться разрешение от государственных органов Соединенных Штатов

## **Центральный офис Brocade в мире**

San Jose, CA USA  
T: +1 408 333 8000  
[info@brocade.com](mailto:info@brocade.com)

## **Brocade Европа, Ближний Восток и Африка**

Geneva, Switzerland  
T: +41 22 799 56 40  
[emea-info@brocade.com](mailto:emea-info@brocade.com)

## **Brocade Россия и СНГ**

Москва  
T: +7 985 762-5486  
[russia@brocade.com](mailto:russia@brocade.com)

## **Brocade азиатско-тихоокеанский регион**

Singapore  
T: +65 6538 4700  
[apac-info@brocade.com](mailto:apac-info@brocade.com)

## **Благодарности**

---

Особые благодарности Майку Клейко (Mike Klayko) и Тому Бьёччи (Tom Buiocchi) за поддержку на уровне руководства компании.

Концепции материала по некоторым темам основаны на презентациях и технических статьях (white-papers), подготовленных службой обучения Brocade и/или маркетингом Brocade. Различные источники Brocade использовались для сверки ссылок и иллюстраций, в том числе ряд технических статей, подготовленных Brocade и классической McDATA (до покупки компанией Brocade), презентации и документы с описаниями решений.

Джед Блесс (Jed Bleess), Джим Хойзер (Jim Heuser), Лиза Гесс (Lisa Guess), Стив Въен (Steve Wynne) и Мартин Скейген (Martin Skagen) предоставили свои отзывы о рисунках, а Мартин Скейген и Саймон Гордон (Simon Gordon) предоставили материал для приложений. Историческая информация была получена от Эй Джей Касаменто (AJ Casamento). Материал был адаптирован из технических статей Brocade, написанных Томом Кларком (Tom Clark) и другими авторами.

Появление этой книги было бы невозможно без её рецензирования, которое выполнили Томас Кэррол (Thomas Carroll), Дерек Гранат (Derek Granath), Мартин Скейген, Тод Эйнк (Todd Einck), Майкл О'Коннор (Michael O'Connor), Майк Шмит (Mike Schmitt), Марио Бландини (Mario Blandini), Кент Хансен (Kent Hansen), Роберт Снивели (Robert Snively) и Сью Вилсон (Sue Wilson).

Наконец, автор благодарит всех сотрудников отдела Разработки Brocade за их упорную и часто бескорыстную работу, без которой не имело бы смысла приниматься за написание книги о проектировании SAN.

## **Об авторе**

---

Джош Джад (Josh Judd) занимает в Brocade должность ведущего инженера (Principal Engineer) в отделе Технического маркетинга. Сам он говорит, что является «Главным занудой» в этом отделе. Помимо написания различных материалов он отвечает за поддержку выработки стратегических планов развития, определение спецификации новых продуктов и сотрудничает с системными инженерами (SE), OEM-производителями и конечными пользователями по всему миру.

В разное время он отвечал за выбор, инсталляцию и развертывание компьютеров Net Ware, Windows и UNIX и у него скопился запас дипломов и сертификатов по компьютерным дисциплинам (правда, он не помнит, где они сейчас хранятся). В результате, он обладает опытом проектирования, развертывания и управления сетевым оборудованием всех основных вендоров. Одно время отвечал за проектирование ИТ-инфраструктуры с нуля и до международного масштаба и при выполнении этого проекта накопил значительный опыт в сфере технологий хранения данных.

Он поступил на работу в Brocade более десяти лет назад и начал с должности старшего технического IT-консультанта (contrib utor), отвечающего за сетевую инфраструктуру, инфраструктуру серверов и ПК. Он стал первым сотрудником Brocade, получившим звание “Senior SAN Architect,” т.е. стал первым в мире архитектором FC SAN с полным рабочим днем.

Сейчас Джош живет в Калифорнии, но ему приходится много путешествовать по миру и эти строки он редактировал когда поезд Париж-Лондон проезжал туннель под Ла-Маншем.

## **Об этой книге**

---

### **О чем эта книга?**

Эта книга содержит основную информацию о сетях хранения данных (SAN) и описания конкретных продуктов Brocade. Ей можно пользоваться как общим справочником по этим темам, так и как пособием для развертывания решений Brocade. Как следует из ее названия, основное внимание в книге уделяется не просто перечислению примеров и успешных внедрений, но общим принципам проектирования SAN, следуя которым читатель сможет разработать конкретные конфигурации.

Книга написана в формате учебника и ее можно читать подряд с начала и до конца, однако многие читатели захотят использовать её и как справочник, поэтому книга состоит из двух частей, рассчитанных на разные категории читателей либо на разные задачи, которые стоят перед читателем.

В первой части дается обзор технологии SAN, излагается базовая концепция Fibre Channel и дается введение в концепции управления жизненным циклом информации (Information Lifecycle Management) и ресурсных вычислений (Utility Computing). Каждая глава в этой части книги посвящена одной теме, поэтому ее можно изучать отдельно от остальных. Особый акцент делается на выборе продуктов для разных типов SAN,

поэтому прочитавший первую часть ИТ-специалист получит основные сведения, которые необходимы для того, чтобы проектировать SAN.

Во второй части содержится практическая информация об использовании продуктов Brocade в SAN, включая рекомендации по проектированию сетей, внедрению и управлению, хотя основной акцент делается на проектировании. В ней собраны методические рекомендации и правила, поэтому ей можно пользоваться и как справочным разделом книги, и как учебником. Как и раньше, основное внимание уделяется материалам, которые обычно незначительно меняются при выходе новых версий.

Заключительная часть состоит из приложений с часто задаваемыми вопросами (FAQ), информацией о стандартах, детальным описанием архитектуры продуктов, информацией о конкретном оборудовании и программном обеспечении Brocade, а также словарь. Эта часть предназначена только для использования в качестве справочника.

Хотя у частей книги разные акценты, каждая из них содержит материал, относящийся к другим частям. Для того, чтобы читателю легче было найти подробные сведения по конкретной теме части книги связаны между собой системой ссылок. В каждую часть включено много диаграмм, таблиц и примечаний.

## **Что нового во втором издании?**

Это уже второе выходящее из печати издание *Основ проектирования SAN*. Некоторые разделы книги значительно переработаны и помимо стилистической правки в книге обновлена часть технической информации. Например, в этом издании:

- Добавлена информация о продуктах McDATA

- Добавлены описания новых продуктов Brocade для SAN
- Учтены отзывы, полученные от читателей и рецензентов Brocade
- Немного расширено содержание каждой части
- Изменено форматирование книги

## **На кого рассчитана эта книга?**

Эта книга окажет существенную помощь следующим категориям читателей:

- Любоому специалисту, собирающемуся сдавать экзамены BCSD
- ИТ-персоналу, отвечающему за развертывание SAN или подготовку такого проекта
- Системным инженерам, которые проектируют и внедряют SAN
- Персоналу OEM- производителей, занимающихся продажами или поддержкой SAN
- Аналитики, которым требуется глубокое понимание SAN
- Сетевым инженерам, которые хотят повысить свою квалификацию

## **Что такое сертификат BCSD и как его получить?**

BCSD расшифровывается как Brocade Certified SAN Designer ( сертифицированный Brocade разработчик SAN). Обладатель BCSD должен хорошо понимать процессы проектирования и внедрения законченного решения SAN.

Сертификат BCS D дает его обладателю ряд преимуществ при работе в индустрии хранения – он может поместить символ BCSD на своей визитке и маркетинговых материалах, как доказательство

своей высокой квалификации упомянуть этот сертификат в своем резюме, получить скидки на обучение на курсах Brocade Education Services и доступ к базе знаний через Brocade Connect. Этот сертификат пользуется авторитетом в индустрии хранения и его обладателю легче найти работу в условиях сегодняшней острой конкуренции на рынке труда (многие партнеры Brocade при приеме на работу требуют от кандидата наличия сертификата BCSD и/или BCFP). Подробно о процедурах получения этого сертификата можно узнать на web-странице Brocade Education Services:

<http://www.brocade.com/education/index.page>

## Где найти дополнительную информацию?

Имеется несколько книг, которые можно использовать как введение в SAN. Например, *Практическое Построение Сетей Хранения Данных (Practical Storage Area Networking)*, написанная Дэном Поллаком (Dan Pollack). *Многопротокольная Маршрутизация для SAN (Multiprotocol Routing for SANs)* Джоша Джада может служить справочным пособием для пользователей, имеющих общее представление о SAN, но которые хотели бы получить больше информации о маршрутизации FC-FC, iSCSI или FCIP. Читателям, интересующимся формирующейся сейчас технологией File Area Network (FAN), рекомендуется книга *Представление Сетей Хранения Файлов (Introducing File Area Networks)* Джоша Джада. Недавно вышла книга Тома Кларка *Стратегия Защиты Данных (Strategies for Data Protection)*.

Информацию об этой и других книгах можно найти на [www.brocade.com](http://www.brocade.com) в разделе Bookshelf, страница <http://www.brocade.com/data-center-best-practices/bookshelf/Browse.page>.

Пользователи продуктов Brocade могут вступить в сообщество Brocade Connect ( см. раздел Connect на [www.brocade.com](http://www.brocade.com)) и получить доступ к форумам, документации и скриптам. Аналогичное онлайновое сообщество для партнеров Brocade и OEM-производителей доступно по адресу <http://partner.brocade.com>.

## Как сообщить об обнаруженных ошибках и неточностях?

Отзывы по-английски можно направить по адресу [bookshelf@brocade.com](mailto:bookshelf@brocade.com). В отзыве просьба указать название книги, издание, дату публикации и, по возможности, номер страницы и параграфа, к которым относится каждый комментарий. Мы не можем отвечать на все отзывы по электронной почте, но они будут обязательно учтены при подготовке нового издания.

## Предисловие к русскому переводу

Перевод на русский язык выполнен для Частей 1 и 2, а также для Словаря терминов из Части 3. Часть 3, за исключением Словаря, оставлена на английском языке.

Это первый опыт перевода книги Brocade на русский язык. Встречаются два противоположных мнения: 1) переводить на русский язык такие книги нельзя, так как с переводом теряется смысл; 2) переводить книгу нужно, так как многие конечные пользователи не обладают достаточным знанием английского языка. Мы решили что перевод нужен и были бы очень благодарны за Вашу обратную связь – что понравилось/не понравилось, хорошее ли качество перевода? Особенно полезными будут предложения по поводу перевода терминов. Ваши отзывы, пожелания и вопросы просьба направлять по адресу [russia@brocade.com](mailto:russia@brocade.com) с темой письма BOOKSHELF. Просьба указывать название книги, номер издания и версию перевода. С нами также можно связаться по телефону +7 (985) 762-5486.

# **Содержание**

---

## **Краткое содержание**

ЮРИДИЧЕСКАЯ ИНФОРМАЦИЯ.....	III
БЛАГОДАРНОСТИ .....	IV
ОБ АВТОРЕ.....	V
ОБ ЭТОЙ КНИГЕ .....	VI
СОДЕРЖАНИЕ .....	XI

## **ПЕРВАЯ ЧАСТЬ .....1**

1: Основы SAN .....	3
2: Решения SAN.....	61
3: UC и ILM .....	85
4: Обзор проектирования SAN .....	121

## **ВТОРАЯ ЧАСТЬ .....147**

5: Планирование проекта .....	149
6: Планирование топологии .....	175
7: Планирование масштабируемости.....	207
8: Планирование производительности .....	227
9: Планирование доступности .....	296
10: Планирование безопасности.....	325
11: Проектирование территориально-распределенных SAN .	339
12: Планирование внедрения .....	373

## **ТРЕТЬЯ ЧАСТЬ.....395**

Приложение А: Базовые материалы .....	397
Приложение В: Расширенные материалы.....	493
Приложение С: Тест .....	537
Приложение D: Часто задаваемые вопросы .....	550
СЛОВАРЬ .....	561

# Подробное содержание

ЮРИДИЧЕСКАЯ ИНФОРМАЦИЯ.....	III
БЛАГОДАРНОСТИ .....	IV
Об авторе.....	V
Об этой книге .....	VI
<i>О чем эта книга? .....</i>	vi
<i>Что нового во втором издании? .....</i>	vii
<i>На кого рассчитана эта книга? .....</i>	viii
<i>Что такое сертификат BCSD и как его получить?.....</i>	viii
<i>Где найти дополнительную информацию? .....</i>	ix
<i>Как сообщить об обнаруженных ошибках и неточностях?.....</i>	x
<i>Предисловие к русскому переводу.....</i>	x
СОДЕРЖАНИЕ .....	XI
<i>Краткое содержание .....</i>	xi
<i>Подробное содержание .....</i>	xii
<i>Список иллюстраций .....</i>	xviii
<i>Список таблиц.....</i>	xxi
<b>ПЕРВАЯ ЧАСТЬ .....</b>	<b>1</b>
1: Основы SAN .....	3
<i>Сети хранения данных .....</i>	3
<i>История SAN.....</i>	6
<i>Продукты SAN .....</i>	11
Концентраторы, коммутаторы и маршрутизаторы SAN .....	12
HBA и NIC.....	17
JBOD и SBOD .....	18
RAID-массивы .....	20
Ленточные приводы и библиотеки.....	23
Мосты и шлюзы между протоколами .....	26
Программное обеспечение дублирования каналов .....	28
Менеджеры томов и виртуализаторы .....	31
<i>Протоколы SAN .....</i>	33
SCSI.....	35
Fibre Channel .....	36
ATM и SONET/SDH .....	47
IP и Ethernet .....	48
iSCSI.....	51
iFCP .....	57
FCIP .....	58
2: РЕШЕНИЯ SAN .....	61
<i>Консолидация хранения .....</i>	61
<i>Кластеры высокой доступности .....</i>	66
<i>Параллельные и последовательные вычисления.....</i>	69

<i>Консолидация ленточных накопителей / резервное копирование без использования LAN.....</i>	72
<i>Улучшение производительности.....</i>	77
<i>Восстановление после аварий/ Непрерывность бизнеса.....</i>	79
<i>Миграция данных .....</i>	81
<b>3: UC и ILM .....</b>	<b>85</b>
<i>Классификация UC и ILM.....</i>	86
<i>Utility Computing.....</i>	88
Преимущества Utility Computing.....	93
Проблемы внедрения Utility Computing .....	95
Текущее состояние Utility Computing .....	98
<i>Управление жизненным циклом информации .....</i>	<b>101</b>
Преимущества ILM .....	106
Проблемы внедрения ILM .....	108
ILM на практике .....	110
<i>SAN: на пересечении UC и ILM .....</i>	<b>112</b>
План поэтапного внедрения ILM и UC.....	113
Выбор внедряемых приложений на основе SAN.....	117
Проектирование подключений SAN .....	118
<b>4: ОБЗОР ПРОЕКТИРОВАНИЯ SAN .....</b>	<b>121</b>
<i>Совместимость .....</i>	122
<i>Сетевые топологии .....</i>	128
<i>Надежность, доступность и обслуживаемость (RAS) .....</i>	129
Надежность .....	129
Доступность .....	134
Обслуживаемость .....	136
<i>Производительность.....</i>	138
<i>Масштабируемость .....</i>	139
<i>Совокупная стоимость решения.....</i>	140
<i>Увеличение расстояний .....</i>	141
<i>Внедрение и другие работы .....</i>	142
<b>ВТОРАЯ ЧАСТЬ .....</b>	<b>147</b>
<b>5: ПЛАНИРОВАНИЕ ПРОЕКТА .....</b>	<b>149</b>
<i>Обзор процесса планирования SAN.....</i>	150
<i>Документирование проекта построения SAN .....</i>	152
<i>Определение участников проекта.....</i>	153
Выбор менеджера проекта и архитектора SAN .....	153
Организация технической группы.....	154
Определение круга лиц, принимающих решение.....	154
Идентификация пользователей SAN.....	155
<i>Сбор требований.....</i>	<b>156</b>
Определение проблем бизнеса.....	157
Определение требований бизнеса .....	159
Определение технических требований .....	162
Разработка расширенной технической спецификации .....	165
Оценка стоимости проекта .....	170
<i>Обоснование проекта (ROI или TCO) .....</i>	<b>170</b>

<i>Детальный проект SAN и план внедрения.....</i>	171
<b>6: ПЛАНИРОВАНИЕ ТОПОЛОГИИ .....</b>	175
<i>Топология, ориентированная на системы хранения.....</i>	176
<i>Топология каскада.....</i>	177
<i>Топология кольца.....</i>	180
<i>Топология mesh.....</i>	182
<i>Топология центр/периферия .....</i>	185
Оптимизация производительности фабрики СЕ.....	187
Масштабируемость топологии СЕ.....	189
Гибридные топологии СЕ.....	192
<i>Meta SAN центра/периферии.....</i>	195
<i>Топология встроенных коммутаторов.....</i>	197
<i>Топологии для больших расстояний .....</i>	205
<b>7: ПЛАНИРОВАНИЕ МАСШТАБИРУЕМОСТИ .....</b>	207
<i>Аксиомы масштабируемости .....</i>	207
Проектирование крупных решений с мощными компонентами .....	207
О пользе локализации .....	208
Использование линков между фабриками - не всегда оптимальное решение .....	208
Маршрутизаторы часто решают проблему .....	208
Встроенные коммутаторы должны использовать Access Gateway .....	209
<i>Определение требований к масштабируемости .....</i>	209
Требования к соединениям.....	210
Число портов хостов и устройств хранения.....	211
Число портов ISL и IFL .....	212
Учет территориальной распределенности.....	213
<i>Обеспечение максимальной масштабируемости .....</i>	215
Характеристики протокола .....	215
Управляемая масштабируемость .....	217
Изоляция сбоев .....	218
Матрицы поддержки продуктов разных вендоров .....	219
Сервисы сетей хранения .....	220
Масштабируемые топологии .....	225
<b>8: ПЛАНИРОВАНИЕ ПРОИЗВОДИТЕЛЬНОСТИ .....</b>	227
<i>Обзор факторов, влияющих на производительность .....</i>	229
Оконечные устройства.....	229
Протоколы SAN.....	231
Скорости линков.....	232
Переподписка и переполнение канала .....	234
Блокирование (HoLB) .....	238
Коэффициенты потерь и ошибок .....	239
Запаздывание и задержки .....	240
<i>Определение требований к производительности .....</i>	243
Типичные подходы к определению требований .....	244
Использование пропускной способности .....	246
Время отклика.....	247
<i>Обеспечение необходимых ISL и IFL .....</i>	249
<i>Локализация трафика .....</i>	258
Уровни локализации .....	260
Локализация внутри коммутаторов.....	263

Локализация и LSAN .....	266
Новые возможности локализации: UC и ILM .....	268
<b>Многоуровневые CE SAN .....</b>	<b>268</b>
<b>Балансировка линков.....</b>	<b>272</b>
Динамическая балансировка нагрузки: Балансировка маршрутов FSPF .....	275
Расширенный транкинг: Балансировка нагрузки на уровне пакетов .....	277
Динамический выбор пути: транкинг на уровне Exchange .....	283
Резюме балансировки линков .....	291
<b>Преимущества буферных кредитов для производительности .....</b>	<b>291</b>
<b>9: ПЛАНИРОВАНИЕ ДОСТУПНОСТИ .....</b>	<b>296</b>
<b>    Обзор теории SAN HA .....</b>	<b>296</b>
Единая точка отказа .....	297
Стек высокой доступности.....	299
<b>    Резервирование при проектировании SAN.....</b>	<b>303</b>
Нерезервированная неотказоустойчивая фабрика SAN или Meta SAN.....	304
Нерезервированная отказоустойчивая фабрика SAN или Meta SAN.....	304
Резервированные неотказоустойчивые фабрики SAN или Meta SAN .....	304
Резервированные отказоустойчивые фабрики SAN или Meta SAN .....	305
<b>    Узлы с двойным подключением и Multipathing .....</b>	<b>306</b>
<b>    Отказоустойчивые фабрики .....</b>	<b>308</b>
<b>    Резервированные фабрики.....</b>	<b>310</b>
<b>    Проектирование резервированной Meta SAN .....</b>	<b>315</b>
Отказоустойчивые Meta SAN .....	315
Резервированные Meta SAN.....	317
Параллельные резервированные фабрики BB Meta SAN .....	318
<b>    Изоляция сбоев и LSAN .....</b>	<b>320</b>
<b>    Асимметричные SAN .....</b>	<b>321</b>
<b>    Стратегии подключения устройств .....</b>	<b>322</b>
<b>10: ПЛАНИРОВАНИЕ БЕЗОПАСНОСТИ.....</b>	<b>325</b>
<b>    Обзор безопасности .....</b>	<b>325</b>
<b>    Безопасность на физическом уровне .....</b>	<b>326</b>
<b>    Безопасность сетевого управления.....</b>	<b>328</b>
<b>    Безопасность с блокированием портов.....</b>	<b>329</b>
<b>    Безопасность на уровне доступа к устройствам (зонирование) .....</b>	<b>330</b>
Типы зонирования .....	331
Разработка плана зонирования .....	333
<b>    Secure Fabric Operating System (SFOS).....</b>	<b>336</b>
<b>11: ПРОЕКТИРОВАНИЕ ТЕРРИТОРИАЛЬНО-РАСПРЕДЕЛЕННЫХ SAN .</b>	<b>339</b>
<b>    Определение требований.....</b>	<b>340</b>
Общие принципы учета расстояний.....	342
Общие принципы учета требований миграции данных.....	343
Обзор факторов, влияющих на DR.....	344
<b>    Кредиты FC Buffer-to-Buffer.....</b>	<b>346</b>
<b>    Режимы LD передачи на большие расстояния.....</b>	<b>351</b>
<b>    Скорости и технологии MAN/WAN.....</b>	<b>352</b>
<b>    “Ограничивающая” архитектура маршрутизации BC/DR....</b>	<b>359</b>

<i>FastWrite и Tape Pipelining</i> .....	360
<i>Лезвия 10Gbit и решения DR/BC</i> .....	364
<i>Ограничения расстояния для оптоволокна</i> .....	369
<b>12: ПЛАНИРОВАНИЕ ВНЕДРЕНИЯ</b> .....	<b>373</b>
<i>Расположение и монтаж стоек</i> .....	373
<i>Питание и ИБП</i> .....	377
<i>Выбор кабелей и оптических модулей, управление ими</i> .....	381
Оптические кабели и модули .....	381
Управление кабелями .....	385
<i>Настройка коммутаторов</i> .....	387
<i>Поэтапное внедрение и тестирование</i> .....	389
<i>Запуск в эксплуатацию</i> .....	389
Запуск в эксплуатацию с нуля .....	389
Модернизация инсталлированного оборудования .....	390
<i>Повседневное управление</i> .....	392
<i>Планирование устранения сбоев и неисправностей</i> .....	393
<b>ТРЕТЬЯ ЧАСТЬ.....</b>	<b>395</b>
<b>ПРИЛОЖЕНИЕ А: БАЗОВЫЕ МАТЕРИАЛЫ .....</b>	<b>397</b>
<i>Поставляемые платформы Brocade</i> .....	397
FC коммутатор Brocade 200E.....	398
Коммутатор Brocade 4100 .....	400
Коммутатор Brocade 5000 .....	403
Коммутатор Brocade 4900 .....	404
Директор Brocade 48000 .....	405
Многопротокольный маршрутизатор Brocade AP7420.....	410
Многопротокольный маршрутизатор Brocade 7500.....	414
Платформа для приложений Brocade 7600.....	415
Многопротокольное лезвие-маршрутизатор FR4-18i .....	415
Платформа для приложений лезвие FA4-18 .....	416
Лезвие FC10-6 10Gbit Fibre Channel.....	417
Лезвие FC4-16IP iSCSI to Fibre Channel .....	418
Встроенные платформы.....	419
Brocade iSCSI шлюз .....	424
<i>Платформы классической McDATA</i> .....	425
Директор Brocade Mi10k .....	426
Директор Brocade M6140 .....	427
Периферийные коммутаторы Brocade M4400 и M4700.....	427
Маршрутизаторы Brocade M1620 и M2640 .....	428
Шлюз Brocade Edge M3000 .....	428
Шлюз Brocade USD-X .....	429
<i>Инсталлированная база платформ Brocade</i> .....	430
Коммутаторы SilkWorm 1xx0 FC .....	430
Коммутаторы SilkWorm 2xx0 FC .....	432
Коммутаторы SilkWorm 3200 / 3800 .....	435
Коммутаторы SilkWorm 3250 / 3850 FC .....	436
SilkWorm 3900 и 12000 .....	437
Директор SilkWorm 24000 .....	440
Встроенные продукты.....	443
<i>Лицензируемые функции Brocade</i> .....	444

Модель лицензирования Brocade .....	444
Подключение Fabric Node (F_Port).....	445
Подключение Loop Node (FL_Port) (QL/FA) .....	445
Фабрики из нескольких коммутаторов (E_Port) .....	448
Виртуальные каналы .....	449
Буферные кредиты .....	451
Шлюз доступа .....	452
ПО Value Line .....	453
Виртуальные фабрики / административные домены .....	454
FCIP FastWrite и Tape Pipelining .....	456
FC FastWrite .....	459
Горячая загрузка и активация кода .....	460
Advanced ISL Trunking (Frame-Level) .....	460
Динамический Выбор Пути (Exchange-Level) .....	461
Зонирование .....	461
Fabric OS CLI .....	463
WEBTOOLS .....	463
Fabric Manager.....	464
SAN Health .....	464
Fabric Watch .....	466
Advanced Performance Monitoring .....	466
Extended Fabrics .....	467
Remote Switch .....	467
FICON / CUP .....	467
Маршрутизация Fibre Channel .....	468
FCIP .....	469
Secure Fabric OS .....	470
<i>Расчет возврата инвестиций ROI .....</i>	<i>471</i>
Цели анализа ROI .....	472
Анализ шаг 1: идентификация узлов и приложений .....	473
Анализ шаг 2: выбор сценариев.....	474
Анализ шаг 3: определение преимуществ сценариев .....	476
Анализ шаг 4: Определение сопряженных расходов .....	486
Анализ шаг 5: подсчет ROI .....	487
<i>Оборудование Ethernet и IP сетей .....</i>	<i>488</i>
Краевые коммутаторы и концентраторы Ethernet L2.....	488
Маршрутизаторы IP WAN.....	489
Конверторы Gigabit Ethernet медь-оптика .....	491
<b>ПРИЛОЖЕНИЕ В: РАСШИРЕННЫЕ МАТЕРИАЛЫ.....</b>	<b>493</b>
<i>Протоколы маршрутизации.....</i>	<i>493</i>
FSPF: маршрутизация внутри фабрики .....	494
FCRP: маршрутизация между фабриками .....	495
<i>FCR форматы заголовка фрейма.....</i>	<i>498</i>
<i>Механизм реализации зонирования.....</i>	<i>498</i>
“Программное зонирование” – реализация SNS .....	498
“Полное аппаратное зонирование” – фильтрация каждого фрейма .....	499
“Сессионное аппаратное зонирование” – ловушки команд .....	500
<i>Протоколы и стандарты FC .....</i>	<i>501</i>
<i>Brocade ASICs .....</i>	<i>502</i>
Эволюция ASIC .....	503
Stitch и Flannel.....	504
Loom.....	504
Bloom и Bloom-II .....	505

Condor .....	506
Goldeneye.....	508
Egret .....	509
FiGeRo / Cello.....	510
<b>Многоуровневые внутренние архитектуры .....</b>	<b>511</b>
SilkWorm 12000 и 3900 “XY” архитектура .....	513
Архитектура Brocade 24000 и 48000 “CE” .....	520
<b>Скорости соединений .....</b>	<b>523</b>
Форматы кодирования .....	524
1Gbit FC .....	525
2Gbit FC .....	525
4Gbit FC (Frame Trunked или Native) .....	525
8Gbit FC (Frame Trunked или Native) .....	533
10Gbit FC .....	534
32Gbit FC (Frame Trunked) .....	535
256Gbit FC (Frame или Exchange Trunked) .....	535
1Gbit iSCSI и FCIP .....	535
10Gbit iSCSI и FCIP .....	536
<b>ПРИЛОЖЕНИЕ С: ТЕСТ .....</b>	<b>537</b>
<b>Вопросы для самопроверки .....</b>	<b>537</b>
<b>Ответы.....</b>	<b>547</b>
<b>ПРИЛОЖЕНИЕ D: ЧАСТО ЗАДАВАЕМЫЕ ВОПРОСЫ.....</b>	<b>550</b>
<b>СЛОВАРЬ .....</b>	<b>561</b>

## Список иллюстраций

Рис. 1 – Уровни Fibre Channel .....	6
Рис. 2 – Архитектура DAS (топология точка-точка).....	8
Рис. 3 – Архитектура SAN (топология коммутируемой фабрики).....	10
Рис. 4 - Использование моста между протоколами.....	26
Рис. 5 – Использование шлюза протоколов .....	27
Рис. 6 – Пример использования ПО резервирования каналов.....	30
Рис. 7 – Передача данных между уровнями FC .....	37
Рис. 8 – Пакет Fibre Channel .....	40
Рис. 9 - Части Meta SAN .....	44
Рис. 10 – Сравнение заголовков iSCSI и FC .....	52
Рис. 11 – Эффективность заголовков iSCSI и FC .....	53
Рис. 12 – Дополнительная загрузка центральных процессоров при использовании FC и iSCSI .....	54
Рис. 13 – Пример физической топологии FCIP .....	60
Рис. 14 – Архитектура DAS – до консолидации хранения .....	62
Рис. 15 - White Space в подсистемах DAS .....	63
Рис. 16 - White Space после миграции с DAS на SAN .....	65
Рис. 17 – Три отдельных кластера .....	67
Рис. 18 – Кластер на основе SAN .....	68
Рис. 19 – Конвейер для редактирования видео .....	71
Рис. 20 – Резервное копирование через LAN.....	74
Рис. 21 – Резервное копирование без использования LAN.....	77
Рис. 22 - Пример архитектуры Business Continuance SAN.....	81

Рис. 23 – Необходимые для Utility Computing подключения .....	90
Рис. 24 – Уровни архитектуры ЦОДа UC.....	91
Рис. 25 – Требования ILM к соединениям.....	102
Рис. 26 – Архитектура ЦОДа ILM .....	103
Рис. 27 - SAN на пересечении UC и ILM .....	113
Рис. 28 – Каскадная топология. Каскады из четырех и шести коммутаторов .....	178
Рис. 29 – Топология кольца .....	180
Рис. 30 – Топология mesh .....	183
Рис. 31 – Топология СЕ .....	185
Рис. 32 – Простая отказоустойчивая фабрика центр / периферия .....	185
Рис. 33 – Масштабируемость в зависимости от расположения устройств.	191
Рис. 34 – Неотказоустойчивая фабрика СЕ.....	193
Рис. 35 – Асимметричная резервированная фабрика А/В.....	194
Рис. 36 – Резервированные фабрики с системой НА и не-НА .....	195
Рис. 37 – Общая схема блоков фабрики СЕ .....	196
Рис. 38 - Общая схема блоков СЕ Meta SAN .....	196
Рис. 39 - Обычная СЕ Meta SAN с фабриками СЕ на периферии .....	196
Рис. 40 – Общая архитектура SAN со встроенными коммутаторами. ....	198
Рис. 41 – Подключение E_Port встроенных коммутаторов .....	200
Рис. 42 – Подключение встроенных коммутаторов с помощью NPIV связи. ....	202
Рис. 43 – Детальная схема работы NPIV .....	204
Рис. 44 – Расширение фабрик СЕ.....	206
Рис. 45 – Коэффициент переподписки ISL равен 3:1 .....	250
Рис. 46 – Использование локальности .....	258
Рис. 47 – Уровни локализации .....	261
Рис. 48 – Двухуровневая фабрика СЕ.....	269
Рис. 49 – Трехуровневая фабрика СЕ .....	270
Рис. 50 – Потоки трафика в фабрике следующего поколения.....	272
Рис. 51 – Концепция транкинга на уровне пакетов .....	278
Рис. 52 – Транкинг на уровне пакетов плюс DLS.....	281
Рис. 53 – Транкинг на уровне пакетов вместе с DPS.....	285
Рис. 54 - DPS в смешанной фабрике.....	286
Рис. 55 – Балансировка DPS в большом кольце Fiber .....	287
Рис. 56 – Резервирование по горизонтали .....	299
Рис. 57 – Уровни НА .....	300
Рис. 58 - Сравнение отказоустойчивой и неотказоустойчивой фабрики....	309
Рис. 59 – Сравнение резервированных фабрик и разделов.....	313
Рис. 60 - Отказоустойчивая Meta SAN .....	317
Рис. 61 - Резервированные Meta SAN.....	317
Рис. 62 – Отказоустойчивая Meta SAN с резервированными фабриками ВВ .....	319
Рис. 63 – Резервированная Meta SAN + резервированные ВВ .....	319
Рис. 64 – Вариант резервированных Meta SAN .....	322
Рис. 65 – SCSI Write без FastWrite (быстрой записи) .....	362
Рис. 66 – SCSI Write с поддержкой FastWrite (быстрой записи) .....	363
Рис. 67 – крупное маршрутизируемое решение 10Gbit DR/BC .....	365

Рис. 68 – Зависимость расстояния от оптики, скорости и типа кабеля (таблица дополнена с учетом новейших модулей SFP - прим. переводчика).....	369
Рис. 69 – Неправильная организация охлаждения в стойке .....	376
Рис. 70 – Самая неудачная организация питания .....	378
Рис. 71 – Лучше, но все равно неудачная организация питания.....	378
Рис. 72 – Приемлемая организация питания.....	379
Рис. 73 – Рекомендуемая организация питания .....	380
Figure 74 - Brocade 200E .....	398
Figure 75 - Brocade 4100.....	401
Figure 76 - Brocade 5000.....	403
Figure 77 - Brocade 4900.....	404
Figure 78 - Brocade 48000 Director.....	405
Figure 79 - FC16 Port Blade for Brocade 48000 .....	408
Figure 80 - Brocade AP7420.....	412
Figure 81 - Brocade 7500 Multiprotocol Router .....	415
Figure 82 - Brocade 7600.....	415
Figure 83 - FR4-18i Routing Blade .....	416
Figure 84 – FA4-18 Application Blade .....	417
Figure 85 - FC10-6 10Gbit FC Blade .....	418
Figure 86 - FC4-16IP iSCSI Blade.....	419
Figure 87 - Brocade 4020 Embedded Switch .....	421
Figure 88 - Brocade 4016 Embedded Switch .....	422
Figure 89 - Brocade 4018 Embedded Switch .....	422
Figure 90 - Brocade 4024 Embedded Switch .....	423
Figure 91 - Brocade 4012 Embedded Switch .....	424
Figure 92 - Brocade iSCSI Gateway .....	424
Figure 93 - SilkWorm II (1600) FC Fabric Switch .....	431
Figure 94 - SilkWorm Express (800) FC Fabric Switch.....	431
Figure 95 - SilkWorm 1xx0 Daughter Card .....	431
Figure 96 - SilkWorm 2010/2040/2050.....	433
Figure 97 - SilkWorm 2210/2240/2250.....	434
Figure 98 - SilkWorm 2400.....	434
Figure 99 - SilkWorm 2800.....	435
Figure 100 - SilkWorm 3200.....	436
Figure 101 - SilkWorm 3800.....	436
Figure 102 - SilkWorm 3250.....	437
Figure 103 - SilkWorm 3850.....	437
Figure 104 - SilkWorm 3900.....	438
Figure 105 - SilkWorm 12000 Director .....	439
Figure 106 - SilkWorm 24000 Director .....	441
Figure 107 - SilkWorm 3016 Embedded Switch.....	443
Figure 108 - SilkWorm 3014 Embedded Switch.....	444
Figure 109 - VCs Partition ISLs into Logical Sub-Channels.....	450
Figure 110 - Foundry EdgeIron 24 GigE Edge Switch.....	489
Figure 111 - Tasman Networks WAN Router .....	490
Figure 112 - Foundry Modular Router .....	490
Figure 113 - WAN Router Usage Example .....	490
Figure 114 - Copper to Optical Converter .....	491

Figure 115 - SilkWorm 12000 Port Blades .....	513
Figure 116 - SilkWorm 12000 ASIC-to-Quad Relationships.....	514
Figure 117 - SilkWorm 12000 Intra-Blade CCMA Links .....	515
Figure 118 - SilkWorm 12000 CCMA Abstraction.....	515
Figure 119 - SilkWorm 12000 64-Port CCMA Matrix .....	517
Figure 120 - Full-Mesh Traffic Patterns.....	519
Figure 121 - Top-Level “CE” CCMA Blade Interconnect .....	521

## **Список таблиц**

Таблица 1 – Классификация среды UC и ILM .....	87
Таблица 2 – Уровни локальности.....	262
Таблица 3 – Режимы передачи на большие расстояния.....	351
Таблица 4 – Технологии и скорости MAN/WAN .....	358

# Первая часть

## Обзор проектирования SAN

### Темы

- Обзор основ SAN
- Ориентированные на бизнес решения SAN
- Модель ресурсных вычислений
- Управление жизненным циклом информации
- Общие принципы проектирования SAN

# 1

## 1: Основы SAN

В этой главе излагаются фундаментальные понятия, на основе которых построено остальное содержание книги, в том числе определение сетей хранения данных Storage Area Network (SAN), обсуждение их преимуществ и некоторых протоколов и продуктов, которые используются для построения SAN. Многие слушатели курсов по проектированию SAN уже знают эти концепции, и эта глава поможет им освежить свои знания.

### **Сети хранения данных**

SAN - это сети, предназначенные для обеспечения подключения хостов к устройствам хранения (дискам, RAID-массивам и ленточным библиотекам) и обмена данными между ними (обычно на уровне блоков). Если сеть соответствует такому определению, то она является SAN. Теоретически этому определению удовлетворяют различные технологии, включая такие стандартные сетевые протоколы, как IP/Ethernet, но на практике для использования SAN она должна соответствовать ряду дополнительных требований.

Такие сетевые протоколы, как IP, разрабатывались для приложений, для которых приемлемы потери пакетов, ошибки передачи и возникновение узких мест производительности, поэтому IP не может полностью исключить такие события. В то же время для SAN

недопустимы потери или порча данных и низкая производительность сети в течение долгого времени. Хотя потери данных в LAN допустимы для приложений и операционных систем, те же самые приложения разрабатывались исходя из допущения, что их система хранения данных работает быстро и надежно (попробуйте отсоединить от Windows-сервера кабель LAN и посмотрите, сможет ли он работать. А затем попробуйте вытащить из сервера его жесткий диск C: ...) Это означает, что на практике от SAN требуется максимум производительности и надежности.

В этой книге рассматриваются многие общие вопросы технологии SAN независимо от протокола, но тем не менее, основной акцент сделан на SAN, использующих Fibre Channel (FC), поскольку это самый популярный протокол для SAN, специально разработанный с учетом требований к производительности и надежности SAN.

Таким образом, термину SAN можно дать следующее более точное определение:

*SAN – это высокопроизводительные и очень надежные сети, предназначенные прежде всего для обеспечения связи и обмена данными на уровне блоков между хостами и любыми устройствами хранения (дисками, RAID-массивами и ленточными библиотеками). Они чаще всего включают в себя коммутаторы Fibre Channel, маршрутизаторы, мосты, хосты и устройства хранения.*

Fibre Channel SAN предназначены прежде всего для передачи трафика между хостами и устройствами хранения, но по ним может передаваться и другой трафик, например, трафик между хостами и трафик между устройствами хранения.

Например, SAN можно построить только для того, чтобы хосты могли соединяться со своими устройствами хранения, используя SCSI over Fibre Channel (отображение SCSI на FC часто называется “Fibre Channel Protocol” или *FCP*.) Однако по той же самой инфраструктуре одно устройство хранения может напрямую соединиться с другими для бессерверного резервного копирования или репликации томов. Кроме того, хосты могут соединяться между собой используя Internet Protocol<sup>1</sup> over Fibre Channel (IP/FC). Хосты могут использовать такие протоколы, гарантирующие отсутствие больших задержек, как FC-VI for DMA, для взаимодействия процессов внутри узлов кластера.

Для обеспечения этих характеристик в Fibre Channel используется многоуровневый подход, где протоколы верхнего уровня идут поверх сети FC. Такое отображение протоколов показано на Рис. 1.

---

<sup>1</sup> IP – это стандартный коммуникационный протокол Internet и IP/Ethernet – это стандарт де-факто для сетей передачи *данных* в корпоративной ИТ-инфраструктуре. Тем не менее, IP/Ethernet практически не используется в сетях хранения. Поэтому в SAN имеет смысл использовать IP/FC в добавлении к FCP, поскольку в этом случае IP *не несет трафика хранения* – IP-пакеты просто идут по тем же проводам, что и трафик хранения, но IP не передает этот трафик.

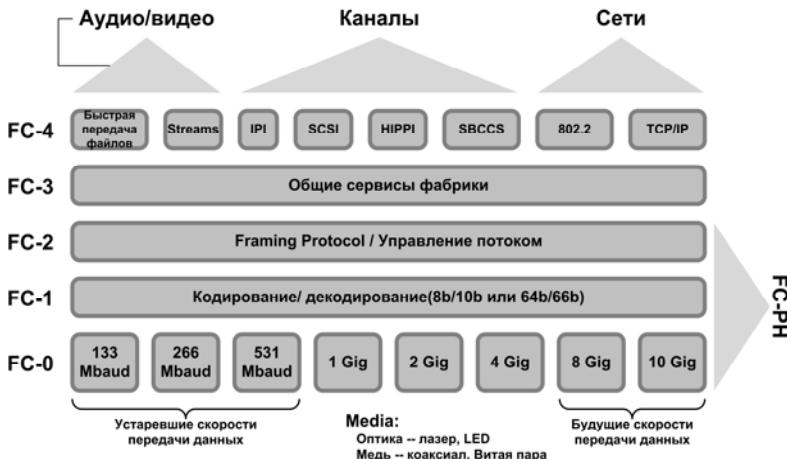


Рис. 1 – Уровни Fibre Channel

Все большее предприятий сегодня строят у себя SAN, поскольку сети данных обладают рядом важных преимуществ. SAN очень удобны для консолидации ресурсов хранения в единый централизованный пул, что невозможно обеспечить при прямом подключении устройств хранения. Такая консолидация позволяет внедрять новые поколения методов управления данными, например, ресурсные вычисления и управление жизненным циклом информации. SAN - эффективное решение для восстановления после аварий, кластеров высокой готовности (HA), улучшения производительности приложений и высокоскоростного резервного копирования/восстановления. Эти решения SAN рассматриваются в Главе 2.

## История SAN

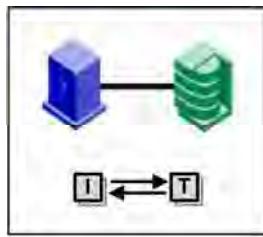
В 1980- е годы устройства хранения обычно подключались к серверам напрямую используя шину SCSI, однако из-за быстрого роста требований к хранению в 1990- е годы этот метод прямого подключения устройств хранения (Direct Attached Storage, DAS) стал неэффективным. Дело в том, что при

масштабировании DAS возникал целый ряд проблем, например:

- Для модернизации устройств хранения DAS нужно отсоединить шину SCSI и даже разобрать хост, поэтому расширение емкости DAS приводило к перебоям в работе приложений, часто на длительное время, что создавало серьезный риск для бизнеса.
- Максимальное расстояние между хостом и устройствами хранения было небольшим – часто их нельзя было установить в разных частях центра обработки данных. Это ограничение не позволяло обеспечить восстановление после аварий Disaster Recovery (DR).
- Для решений кластеров высокой доступности (High Availability, HA) требуется, чтобы каждый узел кластера имел доступ к устройствам хранения других узлов – это необходимо для переключения приложения узла в случае его сбоя на другие узлы. Однако SCSI может поддерживать не более двух инициаторов нашине, причем большинство продуктов SCSI поддерживают только один инициатор.
- ИТ-администраторы часто имели избыток емкости в подсистеме хранения одного хоста, и критическую нехватку емкости в другой подсистеме. При этом не было возможности предоставить избыточную емкость хосту, нуждающемуся в ней. Эта ситуация называется проблемой «неиспользуемой емкости (white space utilization)» дисковых или ленточных подсистем хранения. Чем больше емкости не используется, тем ниже эффективность хранения.
- Производительность шины SCSI не соответствовала быстро растущим мощностям

центральных процессоров и нагрузки приложений.

Самая простая архитектура DAS – это соединение точка-точка (point-to-point) одного хоста и одного диска (см. Рис. 2).



### Точка-точка

Два и только два устройства напрямую подключены между собой без промежуточных коммутаторов, маршрутизаторов или концентраторов.

Рис. 2 – Архитектура DAS (топология точка-точка)

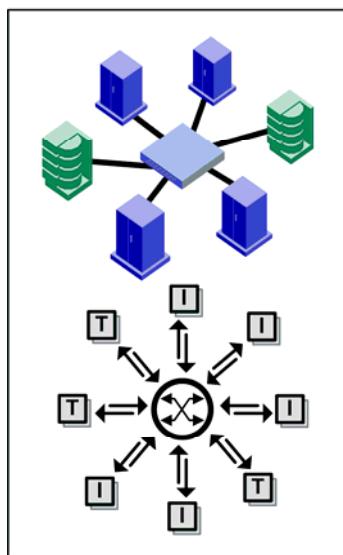
Другие архитектуры DAS (например, цепочка дисков на шине SCSI) не обеспечивают существенного улучшения подключений по сравнению с точка-точка. Например, один адаптер SCSI может обслуживать от семи и до четырнадцати дисков, но для многих современных приложений требуется намного больше накопителей на один адаптер: большинство серверов сегодня способны поддерживать двести и даже пятьсот дисков. (Подключение к шине по архитектуре DAS и его ограничение обсуждаются далее в разделе “Консолидация хранения” начиная со страницы 61.)

SAN революционизировали развертывание и управление устройствами хранения. Подключение устройств хранения к высокоскоростной сети позволило впервые преодолеть ограничения DAS, например:

- Дополнительную емкость можно установить без даже минимального нарушения работы хостов.
- SAN не только охватывают весь центр обработки данных и способны обеспечить построение катастрофоустойчивых решений (Disaster Recovery, DR), защищающих от аварий

регионального масштаба, но могут соединять разные континенты для построения DR для защиты от крупных катастроф.

- Решается проблема неэффективного использования емкости, поскольку любой хост может обращаться к любому устройству хранения и если ему нужна дополнительная емкость, то он сможет получить ее от любого устройства, у которого есть свободная емкость.
- Ранее кластеры НА использовались только для самых требовательных приложений, но благодаря SAN сфера их применения существенно расширилась.
- Производительность SAN на базе Fibre Channel превосходит требования самых «тяжелых» приложений. FC стал поддерживать 1Gbit еще до ратификации стандарта Gigabit Ethernet и сейчас поддерживает 2Gbit, 4Gbit, 8Gbit и 10Gbit, что обеспечивает разные опции производительности в зависимости от требований приложений. При использовании транкинга поставляемые сегодня платформы FC способны обеспечить полосу пропускания одного канала до 256Gbit.



## Коммутируемая фабрика

В фабрике используются 3-байтовые адреса, поэтому теоретически она может соединять свыше 16 млн. устройств. На практике максимальное число устройств меньше, но все равно на несколько порядков больше, чем для петли.

Кроме того, реализуется неблокирующая передача данных поскольку коммутаторы выделяют полосу пропускания для каждого порта и нет общей для всех портов полосы пропускания. Коммутируемая фабрика FC – самая распространенная топология SAN.

Рис. 3 – Архитектура SAN (топология коммутируемой фабрики)

Существует несколько способов построения FC SAN. Когда эта технология только вышла на рынок, некоторые вендоры пытались продвигать сети Fibre Channel Arbitrated Loop (FC-AL, петля). ( См. “ Fibre Channel” на странице 36, где обсуждается FC-AL и другие опции протоколов FC.) Однако петли не получили широкого распространения из-за связанных с ними ограничений по производительности, надежности и масштабируемости и самым популярным методом построения SAN оказались коммутируемые фабрики Fibre Channel. На Рис. 3 показана типичная архитектура фабрики.

Неудивительно, что первыми сети хранения стали применять пользователи, которым нужен максимальный уровень доступности, производительности и гибкости конфигураций, т.е. крупные предприятия и государственные организации. Внедрение инфраструктуры SAN позволило им в десять и более раз сократить расходы на внедрение. Эта экономия может

быть оценена в показателях «жесткой» окупаемости инвестиций (ROI) и при крупных инсталляциях может сэкономить миллионы долларов<sup>2</sup>.

Вслед за этой категорией пользователей и остальные секторы ИТ-рынка начали внедрять SAN и сегодня все больше ИТ-отделов компаний уже используют SAN либо выполняют ее развертывание. Весь рынок SAN и особенно сектор Fibre Channel быстро растет начиная с 1997 года.

Очевидно, что применение SAN очень выгодно для большого числа пользователей. В отличие от начала прошлого десятилетия, сегодня мало кто из ИТ-менеджеров может вкладывать деньги в технологии без обоснования этих инвестиций. Для выделения денег на SAN из ограниченного ИТ-бюджета нужно доказать окупаемость (ROI) решений на основе SAN, чему и посвящена Глава 2.

## Продукты SAN

Архитектор SAN должен хорошо разбираться в разных продуктах SAN, из которых и строятся решения SAN. В этом разделе дается обзор основных категорий продуктов SAN без детального описания продуктов конкретных вендоров, поскольку на высокотехнологическом рынке SAN характеристики продуктов меняются чаще, чем обновляется содержание этой книги. Для получения полной последней информации нужно обратиться к соответствующему торговому представителю вендора.

---

<sup>2</sup> “Жесткие” выгоды – это выгоды, которые можно оценить в денежных показателях как результат экономии или как результат увеличения доходов организации. Расчет ROI для SAN описывается в “Главе 4: Обзор проектирования SAN”.

## *Концентраторы, коммутаторы и маршрутизаторы SAN*

Сетевой концентратор (hub) или коммутатор обеспечивает соединение своих портов так, что любой порт может “разговаривать” с любым другим или всеми остальными портами<sup>3</sup>. Коммутаторы и концентраторы – это обычно устройства Уровня 2 (“Layer 2”, L2) в терминологии IP. Например, в сети IP/Ethernet нет коммутаторов работают на уровне Ethernet – втором уровне в многоуровневой модели Internet Protocol.

В отличие от концентраторов, коммутаторы не используют архитектуру «общей полосы пропускания» (“shared bandwidth”). Если в концентраторе два порта “разговаривают” между собой, то остальные порты в это время не могут разговаривать. Полосу пропускания между всеми узлами может загрузить только один порт, поскольку концентратор работает по принципу первым пришел – первым обслужен<sup>4</sup>. Если пара устройств через концентратор переговаривается на полной скорости, то в это время другие устройства не могут переговариваться. С другой стороны коммутатор позволяет устройствам соединяться независимо от трафика через любую другую пару портов, поскольку ни один ввод/вывод не может полностью захватить полосу пропускания коммутатора. Это преимущество стало одной из причин успеха коммутаторов Fibre Channel на рынке SAN, которые

---

<sup>3</sup> Во многих коммутаторах используется механизм ограничения доступа (например, зонирование FC), который может блокировать произвольное (any-to-any) соединение между портами этих устройств, но даже в этом случае *потенциально* можно организовать соединение между разными группами портов.

<sup>4</sup> Концентратор можно рассматривать как мост поскольку он реплицирует трафик на выходе для всех портов. В результате каждый узел может “подслушать” переговоры между любыми другими узлами, относящимися к тому же сегменту сети, соединенному мостом.

быстро вытеснили концентраторы FC-AL – при использовании концентраторов хост на какое-то время может лишиться доступа к устройствам хранения, что недопустимо для SAN.

Хотя допускается, что в некоторых случаях трафик между одной парой портов коммутатора влияет на трафик между другими портами, но в коммутаторе невозможна ситуация, когда один поток трафика захватывает всю полосу пропускания и другие потоки не могут проходить через коммутатор (как это происходит в концентраторе). Первая из описанных ситуаций – это пример перегруженности (congestion) в сети, разработанной в расчете на переподписку (oversubscription); вторая – пример блокирования. (Смотри “Глава 8: Планирование производительности” (стр. 227), где подробнее объясняется эта терминология).

Например, внутри многих коммутаторов с матрицей коммутации (crossbar) есть узкие места, из-за которых эти устройства работают так, как если бы имели ограниченную или разделяемую полосу пропускания. Архитектура матрицы коммутации использует централизованный “планировщик”, неспособный одновременно поддерживать соединения между всеми портами<sup>5</sup> и поэтому особенно неэффективный при обслуживании трафика “от одного ко многим (many to one)”, который часто используется при консолидации хранения. Если одновременно работают много портов, то планировщик станет главным узким местом

---

<sup>5</sup> Поясним это утверждение на примере. Возьмем 256-портовый crossbar-коммутатор. Если во все 256-портов подключить кабели от генератора сетевого трафика (например, SmartBits) и при этом все порты будут переговариваться между собой по схеме “полной сети” (full mesh), то чем больше портов будет в сетке, тем медленнее будет идти трафик между ними.

коммутатора и производительность упадет у всех активных в один и тот же момент времени портов. Однако, в правильно спроектированном коммутаторе этот сценарий не приведет к полной блокировке трафика между какой-либо группой портов: он может *замедлиться*, но в любом случае будет *идти*. При этом устройства не смогут получить нужную им полосу пропускания (как и при использовании концентратора), но некорректно называть «коммутатором» устройство, которым трафик может блокироваться.

На коммутаторах Fibre Channel также выполняется пакет программного обеспечения и протоколов, известных под общим названием “сервисы фабрики” (“fabric services”), без которых невозможна корректная работа SAN. Например, если на каждом коммутаторе не выполнялось программное обеспечение сервера имен, то все устройства (хосты и дисковые массивы) пришлось бы конфигурировать вручную, из-за чего была бы ограничена масштабируемость SAN. В SAN на основе других технологий (не FC) также используются аналогичные сервисы, хотя их внедрение отличается от SAN на основе FC. Например, хотя протокол iSCSI предусматривает аналогичные сервисы, их не поддерживают коммутаторы Ethernet, из-за чего сети iSCSI нужно конфигурировать вручную либо приобретать выделенные серверы для выполнения этих сервисов или использовать специальные коммутаторы iSCSI, которые обычно стоят дороже аналогичных моделей FC, хотя значительно проигрывают им по производительности. Более эффективны коммутаторы сетей хранения, в которых есть встроенная поддержка сервисов, что сегодня означает коммутаторы Fibre Channel (подробнее об этом в “Сервисы фабрики” на стр. 44.).

Маршрутизатор схож с коммутатором в том, что он должен обеспечить маршрут передачи данных между своими любыми портами, но маршрутизатор

работает на более высоком уровне стека протоколов. Если коммутатор сети передачи данных работает на уровне Ethernet (L2), то маршрутизатор сети передачи данных работает на уровне IP (L3), что позволяет ему автоматически и полуавтоматически соединять сетевые сегменты в иерархическую структуру, а не в одноуровневую. Маршрутизаторы SAN также работают на более высоком уровне.

Исторически маршрутизаторы работали намного медленнее коммутаторов и поэтому не использовались в высокопроизводительных фабриках. Стоит отметить, что многие IP-маршрутизаторы реализованы на уровне программного обеспечения, а не аппаратуры. В современных IP-сетях обычно используются коммутаторы Уровня 3 (L3), комбинирующие аппаратное ускорение коммутации L2 с интеллектуальными функциями маршрутизации L3 в единой интегрированной платформе. За последние несколько лет были выпущены маршрутизаторы для рынка SAN: лезвия Brocade AP7420, 7500 и FR4-18i – примеры маршрутизаторов с аппаратным ускорением на более высоком уровне для туннелей Fibre Channel и FCIP.

Сети хранения выдвигают очень жесткие требования к надежности и производительности коммутаторов и маршрутизаторов – сеть должна передавать каждый пакет без потерь или задержек, за исключением крайне редко возникающих особых условий, и в любом случае сохранять исходный порядок пакетов. Дело в том, что узлы и приложения, подключенные к SAN, предназначены для прямого подключения к устройствам хранения, поэтому доставка с нарушением порядка пакетов или их потери должны быть полностью исключены. Любая “рабочоспособная” SAN должна обеспечить такое использование устройств хранения,

при котором любой хост видит их как подключенные напрямую и тогда протокол SCSI в SAN будет работать точно так же, как в DAS.

В то время, когда были написаны самые популярные сегодня приложения, операционные системы и драйверы, хост с его диском соединяли только несколько метров кабеля SCSI без каких-либо промежуточных устройств. При такой архитектуре риск потери данных был крайне мал, как и того, что скорость передачи данных будет невелика при использовании данного подключения. Поскольку такие проблемы были полностью исключены, то разработчики не тратили время на создание надежных механизмов коррекции и исправления ошибок и поэтому в драйверах SCSI не предусмотрена защита от сбоев сети. Они быстро работают и способны использовать недорогое оборудование, однако если всё же возникнет ошибка или узкое место производительности, то они повлияют и на более высокие уровни. Например, при потере пакета и/или падении производительности большинство ленточных устройств прервут операцию резервного копирования, либо в лучшем случае перейдет в старт/стопный режим, при котором скорость записи резко уменьшается. Чтобы не возникло подобных ситуаций следует использовать коммутаторы и маршрутизаторы, соответствующие требованиям SAN к надежности и производительности.

В стандарты Fibre Channel изначально были заложены эти требования и все поставляемые Brocade платформы проектируются с их учетом: узлы в SAN не смогут правильно обработать ошибку и поэтому нужно полностью исключить вероятность ошибки. Независимо от того, какая технология используется в инфраструктуре SAN, сама инфраструктура должна быть лучшей в своем классе для обеспечения надежного и предсказуемого поведения приложений, т.е.

полного отсутствия потерь пакетов при нормальной работе и исключения риска нарушения порядка их доставки, эффективной балансировки нагрузки на всех портах для исключения перегрузки и блокировки трафика.

Более подробно о проектировании характеристик производительности рассказывается в “Глава 8: Планирование производительности” на стр. 227.

## **HBA и NIC**

АдAPTERЫ Fibre Channel Host Bus Adapters (HBA) и их драйверы обеспечивают интерфейс между фабрикой Fibre Channel и ОС хоста. Современные HBA работают на скорости 2, 4 либо 8 Gbit и благодаря мощному аппаратному ускорению позволяют обеспечить эти скорости без использования ресурсов центрального процессора хоста, которые нужны для обслуживания приложений<sup>6</sup>. Brocade предлагает самые эффективные по стоимости и надежные HBA из доступных сегодня на рынке для подключения хостов Unix/Linux и Windows к сетям хранения Brocade FC SAN.

В решениях iSCSI вместо HBA используются сетевые карты Net work Interface Cards (NIC). Хотя возможно использовать специализированные iSCSI HBA с аппаратным ускорением, применение таких адаптеров ведет к удорожанию решений iSCSI и в результате эта технология лишается своего основного преимущества

---

<sup>6</sup> Упрощенно можно считать, что каждый 1 GHz генерирует 1Gbit данных приложений. Если не использовать аппаратные ускорители, то потребуется вдвое больше Гигагерц на один Гбит/сек поскольку процессор помимо данных приложений должен генерировать и заголовки пакетов. Таким образом, половину мощности хоста придется расходовать на обработку протоколов вместо обслуживания приложений. На момент написания этой книги этот недостаток относился только к iSCSI, поскольку во всех Fiber Channel HBA используется аппаратное ускорение.

как дешевой альтернативы FC. iSCSI HBA обычно стоят примерно столько же, сколько Fibre Channel HBA, но работают на 50-75% медленнее. Экономичность iSCSI – это компромисс между производительностью, зрелостью технологий, надежностью и стоимостью и поэтому в тех редких случаях, когда iSCSI используется в промышленной эксплуатации, он почти всегда внедряется с помощью программных драйверов, работающих на обычных сетевых картах Gigabit Ethernet NIC.

Стоит отметить, что Brocade, как и другие вендоры, выпускает iSCSI NIC с аппаратным ускорением и шлюзы FC-iSCSI, применение которых имеет смысл для определенных задач.

## ***JBOD и SBOD***

Во многих случаях JBOD (“Just a Bunch of Disks - Просто набор дисков”) - это самый простой тип систем хранения, которые можно подключить к SAN. Этот “набор” дисков установлен в шкафу, обеспечивающем питание и охлаждение, и подключается к фабрике через “тупую” интерфейсную карту. (“Интеллектуальность” реализуется с помощью микрокода каждого диска, а не контроллера JBOD.) Внутри JBOD'ов используется FC-AL и их объединительная панель – это концентратор FC-AL.

Коммутаторы фабрики используют интерфейсы FL\_Port для подключения к JBOD'ам. Этим интерфейсам может потребоваться “невидимая” логика для трансляции сетевых адресов в зависимости от того, поддерживают ли диски JBOD публичную или частную петлю. Коммутатор фабрики представляет каждый диск из JBOD как отдельный объект для сервера имен

фабрики, т.е. диски в JBOD в их физическом виде без присущего для RAID управления томами<sup>7</sup>. Обычно JBOD'ы используются вместе с менеджерами томов, которые работают на хостах.

JBOD можно рассматривать как самый сложный тип систем хранения, подключаемых к SAN, поскольку при их использовании вместо централизованного управления всеми томами с единой консоли необходимо на каждом хосте управлять его томами, и из-за проблем, связанных с FC-AL.

Часть из этих проблем решаются с помощью SBOD. (“Switched Bunch of Disks – коммутируемому набору дисков.”) SB OD похожи на JBOD, но вместо встроенного концентратора FC-AL в них используется встроенный коммутатор, что повышает производительность, а также улучшает управление и устранение сбоев.

JBOD и SBOD практически никогда не используются как первичные системы хранения для критически-важных приложений, для которых лучше подходят RAID-массивы, но они эффективны для многих других задач, например, как загрузочные диски в решениях “boot over SAN” (загрузка через SAN) или как диски для восстановления “с нуля” в некоторых системах защиты от катастроф либо как системы хранения нижнего уровня для Ресурсных Вычислений (Utility Computing) или Управления жизненным циклом информации.

---

<sup>7</sup>

Этот также означает, что JBOD'ы больше ограничивают масштабируемость фабрики, чем RAID-массивы. Например, RAID-контроллер может отображать 256 LUN-ов на одну запись SNS, а при использовании томов JBOD для этого потребуется 256 записей сервера имен.

## RAID-массивы

“RAID” расшифровывается как “Redundant Array of Independent Disks (резервированный массив независимых дисков)”<sup>8</sup>. Подсистемы RAID – это набор физических дисков, которые “спрятаны” за одним или несколькими интерфейсами RAID-контроллеров<sup>9</sup>. Контроллеры предоставляют хостам логические тома, которые необязательно соответствуют физическим дискам. В результате хост “не видит” физические диски RAID-массива, а только логические тома.

RAID-контроллеры подключаются к коммутаторам фабрики с помощью портов N\_Ports или NL\_Ports. (Эти коммутаторы используют интерфейсы F\_Port или FL\_Port соответственно.) Они объединяют физические диски в логические тома, например, с помощью простой конкатенацией дисков так, что из многих маленьких дисков формируются большие тома, либо применяя сложные схемы, обеспечивающие резервирование и улучшение производительности. Обычно RAID-массивы используются для формирования одного или нескольких логических томов с использованием следующих схем:

**RAID 0:** Несколько дисков объединяются в незащищенный от сбоев том с использованием алгоритма чередования (striping). Хотя эта схема похожа на конкатенацию, она улучшает производительность и эффективность использования дисков. Основной недостаток striping – это отсутствие возможности восстановления группы дисков RAID 0 при выходе из строя даже одного диска, который приведет к потере

---

<sup>8</sup> Раньше “I” расшифровывалась как “Inexpensive (недорогие)”, но сейчас большинство RAID-систем трудно назвать “недорогими”, поэтому эта буква расшифровывается как “независимые”.

<sup>9</sup> Для резервирования в SAN обычно используется два или более интерфейсов.

всех данных тома. В большинстве решений RAID 0 данные по дискам распределяются по методу round-robin (кругового обслуживания).

**RAID 1** (зеркалирование дисков): Группа RAID 1 состоит из двух и более<sup>10</sup> физических дисков, содержащих точные дубликаты одних и тех же данных. Настройку и синхронизацию данных на дисках RAID-1 выполняет RAID-контроллер.

**RAID 5:** Как и в RAID 0 данные”распределяются” по нескольким<sup>11</sup> физическим дискам, на RAID 5 для резервирования используется механизм проверки *четности*. Имеется несколько разных алгоритмов организации томов RAID 5, но все они гарантируют, что при выходе из строя одного диска тома его данные не будут потеряны, хотя производительность RAID-массива упадет до тех пор,<sup>12</sup> пока неисправный диск не будет заменен на новый<sup>13</sup>. Выход из строя второго диска в RAID-группе приведет к потере данных, поэтому неисправный диск нужно заменить как можно быстрее. Тома RAID 5 даже в обычном режиме (когда нет неисправного диска) работают медленнее, чем другие варианты RAID из-за дополнительной нагрузки, связанной с расчетом и записью битов четности при выполнении каждой операции SCSI<sup>13</sup>.

---

<sup>10</sup> “больше двух дисков” иногда используются для резервного копирования. Третий зеркальный диск отсоединяется от основного тома, с него делаются резервные копии и затем он снова подсоединяется к тому. Такой подход позволяет получить резервные копии данных по состоянию на определенный момент времени. Однако после внедрения технологии мгновенных снимков его популярность резко упала.

<sup>11</sup> Для тома RAID 5 требуется не менее трех дисков.

<sup>12</sup> Диски горячего резерва сокращают до минимума время замены неисправного диска в томе RAID 5.

<sup>13</sup> Во многих RAID-контроллерах используется механизм кэширования. У контроллера есть своя память RAM (обычно с резервным

**RAID 1+0; 5+1:** Во многих массивах использует комбинация нескольких уровней RAID, например, можно сконфигурировать несколько дисков в зеркальные пары (RAID 1) и затем объединить эти зеркальные пары с чередованием в массив RAID 0. Таким образом, достигается сочетание оптимизации производительности RAID 0 с улучшением надежности RAID 1. В зависимости от конкретной реализации такой подход называется RAID 0+1 либо RAID 1+0. Хотя такие решения дороже, чем RAID 5, поскольку требуют больше физической емкости дисков на единицу доступной емкости, они улучшают скорость и надежность. В некоторых случаях можно зеркаливать между собой целые массивы RAID, т.е. каждый массив состоит из одного или нескольких томов RAID 5, которые зеркалируются между массивами. Этот подход популярен для дорогих решений обеспечения непрерывности бизнеса (business continuance, BC), поскольку для защиты от крупных катастроф массивы можно установить в разных городах. В этом случае важно обеспечить резервирование на уровне отдельной площадки чтобы переключение приложений BC между площадками требовалось только в случае выхода из строя всей площадки в результате крупной аварии.

В большинстве аппаратных подсистем RAID можно сконфигурировать один или несколько дисков как диски “горячего резерва (hot spare)”, которые будут задействованы в случае выхода из строя основного диска из тома RAID 1 или RAID 5. Это позволяет свести к минимуму то время, когда данные тома могут быть потеряны в случае сбоя еще одного диска, а производительность тома падает. Отметим, что диски

---

аккумулятором), куда хост записывает данные или считывает их вместо доступа напрямую к физическому диску. Кэширование при записи существенно улучшает производительность томов RAID 5.

горячего резерва имеет смысл использовать для защиты RAID 0 или томов, полученных конкатенацией дисков, только при использовании конфигурации RAID 0+1, т.е. зеркалированием.

RAID-массивы – это самый распространенный тип систем хранения, подключаемых к SAN, поскольку они обладают высокой гибкостью конфигураций, производительностью корпоративного класса и поддерживают различные опции обеспечения высокой доступности.

### *Ленточные приводы и библиотеки*

Для защиты данных приложений от потери в случае выхода из строя SAN, подключенных к ней устройств хранения или всего центра обработки данных, нужно регулярно проводить резервное копирование и хранить резервные копии в другом месте. Обычно этот сервис обеспечивается с помощью ленточных накопителей, которые обеспечивают оптимальную стоимость резервного копирования, поэтому большинство технологий резервного копирования разрабатывались в расчете на ленту.

Лента *плохо подходит* для случайного доступа к данным, к которым часто происходит обращение, поэтому не используется в основных “онлайновых” системах хранения, а только как носитель резервных копий, которые транспортируются в специальные хранилища. Хотя почти во всех SAN есть ленточные системы, они нигде не применяются для онлайнного хранения.

В небольших SAN для резервного копирования может использоваться один ленточный привод, подключенный к сети через мост, а в крупных SAN

может быть десятки ленточных библиотек размером с небольшую комнату.

Сначала ленточные устройства подключали к фабрике с помощью моста SCSI to Fibre Channel (стр. 26) - у ленточного накопителя был стандартный интерфейс SCSI (например, SCSI-2) и мост преобразовывал и пересыпал его на порты FC N\_Port и NL\_Port. Современные ленточные устройства оборудуются интерфейсом FC либо встроенным мостом, который делает преобразование прозрачным для пользователей.

Ленточные решения – один из главных стимулов внедрения 4 и 8 Gbit Fibre Channel. На момент написания этой книги скорость ленточных технологий уже превысила возможности интерфейса 2 Gbit. Когда ленточный привод подключается к сети, неспособной обеспечить его работу на максимальной скорости, то он не получает из сети данные достаточно быстро для того, чтобы лента перематывалась с максимальной скоростью. В результате, привод либо переходит в замедленный, так называемый “старт-стопный” режим<sup>14</sup>, либо не удается успешно завершить резервное копирование. Решением этой проблемы является перевод SAN на интерфейс 4 и 8 Gbit<sup>15</sup>.

---

<sup>14</sup> “Старт-стопный” режим означает, что привод записывает блок данных на ленту, затем ждет поступления нового блока данных и останавливает ленту, затем когда новый блок поступит, снова начинает перематывать ленту, ждет следующего блока и останавливает ленту и т.д. Поскольку старты и стопы ленты занимают *большое* время, сравнительно говоря, то это гораздо хуже чем просто медленный интерфейс на диске. Обычно считается, что если решение резервного копирования так себя ведет, то оно скорее «неисправно», чем «медленно».

<sup>15</sup> Это одна из причин того, что сейчас практически нет ленточных систем на базе iSCSI, ведь – если даже 2Gbit FC с аппаратным ускорением не может обеспечить необходимую скорость для современных ленточных накопителей, то не имеет никакого смысла применять интерфейс iSCSI, работающий на скорости ниже 1Gbit.

### Сокращение окна резервного копирования с помощью 4Gbit FC

Архитекторы SAN, разрабатывающие решение для резервного копирования на ленту, должны учитывать “окно резервного копирования”.

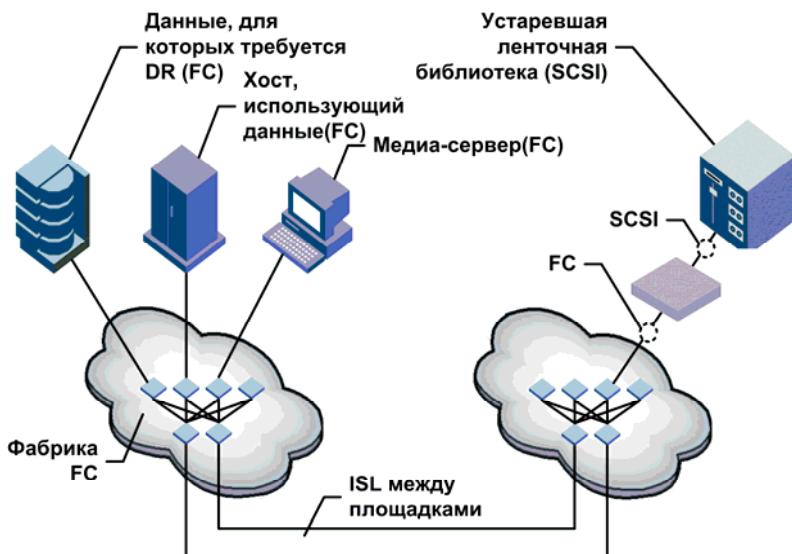
Для непротиворечивости резервная копия должна точно соответствовать состоянию тома данных на определенный момент времени (point in time), однако на ленту невозможно мгновенно записать резервные копии данных. Имеется несколько способов решения этой проблемы, самый простой из которых – это остановить приложения на время резервного копирования, но бизнес большинства современных предприятий не допускает длительных перерывов в работе критичных приложений. Более удобно делать резервную копию с зеркального диска, либо использовать получаемые с помощью специального программного обеспечения “мгновенные” снимки, однако применение обоих этих методов ведет к определенному падению производительности во время резервного копирования. Кроме того, при больших объемах данных ежедневное / инкрементальное резервное копирование на ленту может продолжаться более суток, если применять традиционные подходы без использования SAN. В любом случае, время выполнения резервного копирования называется окном резервного копирования (backup window) и его максимальная продолжительность – это то время, в течение которого допустимо временное падение производительности приложений из-за выполнения резервного копирования.

Для уменьшения до минимума окна резервного копирования архитекторы SAN стараются применять новые и самые производительные ленточные технологии. Хотя применение для резервного копирования технологии 1Gbit Fibre Channel может

снизить расходы, использование технологии 4 и 8 Gbit существенно уменьшил окно резервного копирования и во многих случаях это дает экономию расходов, существенно превышающую стоимость самой инфраструктуры.

### *Мосты и шлюзы между протоколами*

Иногда архитекторам SAN требуется соединить сегменты сети, в которых используются разные протоколы. Эта задача решается с помощью мостов или шлюзов. Каждый из этих двух подходов имеет свои особенности.



**Рис. 4 - Использование моста между протоколами**

Например, рассмотрим случай, когда компания, в которой строится первая Fibre Channel SAN, не может сразу перевести все свои системы хранения на Fibre Channel ( см. Рис. 4 ) и нужно сразу внедрить катастрофоустойчивое решение, хотя из-за ограничений бюджета нельзя провести модернизацию всех ленточных библиотек. В этом случае решением может стать

использование мостов SCSI to FC. Каждый такой мост преобразует интерфейс SCSI ленточной библиотеки в Fibre Channel, что обеспечивает доступ катастрофоустойчивого решения на базе SAN к ленточной библиотеке без модернизации последней. Любой пакет, который приходит из фабрики по интерфейсу FC, преобразуется в SCSI и пересыпается по интерфейсу SCSI ленточной библиотеки и наоборот. В мостах данные обычно идут потоками между интерфейсами без таких высокоуровневых функций, как контроль доступа, маршрутизация на сетевом уровне и расширенные функции устранения сбоев.

Мосты удобны для подключения к сети отдельных устройств, использующих другой протокол, но они неэффективны если к фабрике FC корпоративного класса нужно подключить большое число дешевых устройств iSCSI. В этом случае также требуется преобразование протоколов, но к фабрике подключается не отдельное устройство, а вся IP-сеть. В таких ситуациях требуются многопротокольные шлюзы (см. Рис. 5). Хотя преобразование протоколов в шлюзах мало отличается от мостов, первые обладают более высокой функциональностью. Платформа Brocade iSCSI Gateway и «лезвие» FC4-16IP для директора Brocade 48000 могут рассматриваться как примеры такого продукта.

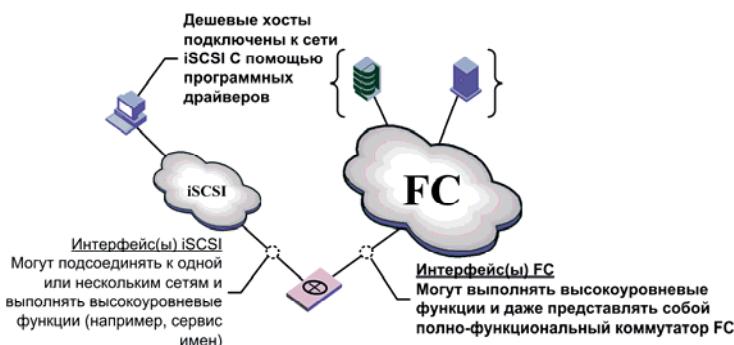


Рис. 5 – Использование шлюза протоколов

## *Программное обеспечение дублирования каналов*

Программное обеспечение дублирования каналов (Multipathing software) позволяет хосту для доступа к одному логическому тому использовать разные физические маршруты. Обычно в хосте устанавливается несколько НВА, подключенных к разным фабрикам, и система хранения подключается ко всем этим фабрикам<sup>16</sup>. Без этого невозможно обеспечить высокую готовность, поэтому эти технологии применяются практически во всех крупных SAN. (См. “Сетевые топологии” на стр. 128 и всю “Главу 9: Планирование доступности” со стр. 296.)

Программное обеспечение дублирования каналов должно поддерживать подключение резервированных НВА и портов систем хранения к резервированным SAN – без этого невозможно внедрить методические указания (best-practices) для критически-важных приложений НА, предусматривающих дублирование всех компонентов. В НА решении *все* должно быть более чем в одном экземпляре (включая адAPTERЫ НВА, шасси директора, операционные системы и даже сеть), иначе не гарантируется высокая доступность.

Это основное правило проектирования SAN часто не соблюдается на практике. Когда мы говорим, что «все» должно быть более чем в одном экземпляре, мы действительно имеем в виду *все*. Например, шасси директора разделяется на две фабрики, либо разделяется зонами, VSAN-ами или другими похожими

---

<sup>16</sup>

Для резервирования каналы должны быть изолированы между собой на всех уровня – оборудования, программного обеспечения и везде в сетевой инфраструктуре. Даже такие аппаратные средства управления, как зонирование и LSAN, должны работать на полностью физически изолированных коммутаторах, иначе не будет обеспечено резервирование.

механизмами, а затем НВ А-адAPTERЫ А и В подключаются к разным частям одного шасси. В таком случае при выходе из строя всего шасси будет нарушена работа всей SAN. Это может произойти, например, если помещение, где оно установлено, будет залито водой из-за ошибочного срабатывания системы пожаротушения или при установке в шасси лезвия объединительная панель (backplane) директора сгорит или один из контактов сломается. Хотя вероятность таких событий мала, их нельзя полностью исключить. Более вероятны выход из строя шасси в результате ошибочных действий операторов или ошибок в ОС. Есть вендоры, утверждающие, что его шасси – это система НА, в их операционной системе не может быть ошибок, а пользователь их продуктов не может сделать ошибку и не может произойти сбоев из-за внешних факторов. Некоторые вендоры называют свое шасси «пуленепробиваемым», но что будет, если кто-нибудь выстрелит по нему из винтовки. Разумеется, я не рекомендую проводить такой эксперимент, но попробуйте *представить* его последствия.

Правильно решение – использовать полностью разделенные оборудование и программное обеспечение для частей А и В сети. На Рис. 6 показан программный стек хоста с драйверами резервирования каналов, которые обращаются к одному LUN по двум изолированным между собой фабрикам. Смотри “Резервирование при проектировании SAN” начиная со стр. 303, где подробнее рассматриваются модель “A/B” и другие факторы, которые нужно учитывать при проектировании решений НА.

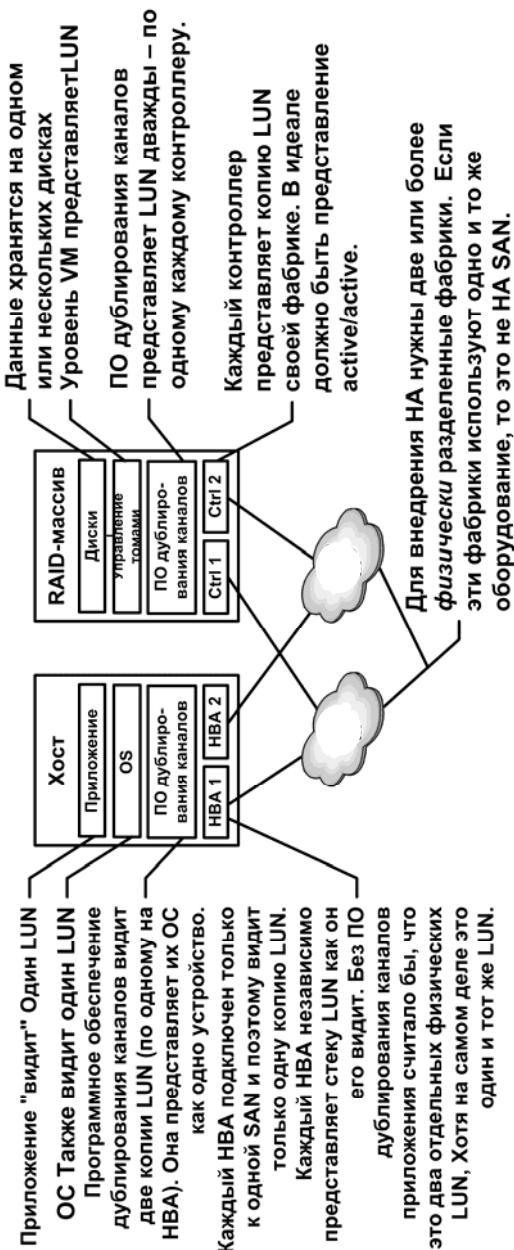


Рис. 6 – Пример использования ПО резервирования каналов

## *Менеджеры томов и виртуализаторы*

Программное обеспечение менеджеров томов (Volume Management, VM) и виртуализаторы имеют много общего с RAID- массивами. RAID- массив абстрагирует физические диски и представляет их как LUN-ы, характеристики которых независимы от параметров дисков. Каждый LUN может быть меньше или больше физического диска, быстрее или медленнее, более или менее надежен. С помощью RAID- контроллера администратор может добиться компромисса между стоимостью, доступностью и производительностью. (См. “RAID- массивы” на стр. 20.) RAID- массивы – это предпочтительное решение для большинства задач управления томами, поскольку они обеспечивают высокую производительность и надежность, хотя они достаточно дорогие, не обладают гибкостью, используют «закрытые» архитектуры и неспособны обеспечить зеркалирование и репликацию между разными массивами.

Программное обеспечение Volume Management обычно работает на серверах, а не на RAID- контроллерах, но реализует схожие функции. Менеджер томов делает группу дисков «видимой» для хоста и абстрагирует ее для того, чтобы реализовать похожие на RAID функции для систем не-RAID (например, JBOD) либо для реализации дополнительного уровня абстрагирования «поверх» RAID для зеркалирования, репликации или миграции данных между массивами. Решения VM на базе хоста обычно дешевле, чем аппаратные RAID- массивы, более гибки и эффективны при работе с системами хранения разных вендоров. Большинство из них также реализуют уровень резервирования каналов. К их недостаткам следует отнести низкую производительность и сложное управление ими в крупных инсталляциях.

Сравнительно недавно на рынок вышло еще одно решение – виртуализаторы на уровне сети хранения. Термин «виртуализация» появился еще в 1960-х годах, когда мэнфреймы начали «разделять» свои ресурсы между разными программами. Яркими примерами продуктов, которые в виртуальном виде представляют физические ресурсы, являются виртуальная память, виртуальные ленточные устройства и менеджеры логических томов. Сегодня термин «виртуализация хранения» часто используется для описания разных функциональных возможностей управляемых систем хранения.

Brocade определяет виртуализацию хранения с точки зрения реализуемых с ее помощью сервисов, в том числе:

- Управление томами – Этот сервис позволяет динамически перемещать тома и менять их размер, в результате эффективнее распределяя емкость.
- Миграция данных – Прозрачное перемещение томов данных между разнородными устройствами. Благодаря этому сервису приложения смогут работать даже когда их данные перемещаются с одного устройства хранения на другое.
- Репликация данных – Этот сервис, похожий на миграцию данных, позволяет создавать и постоянно обновлять независимые копии важных данных.
- Усовершенствование резервного копирования – Данные можно восстановить по состоянию на произвольный момент времени, пользуясь технологиями виртуализации, которые ведут журнал всех изменений конкретного тома и затем

применяют эти изменения к копии, которая была получена раньше.

Виртуализаторы располагаются между хостами и системами хранения, поэтому они могут выполнять такие операции RAID-to-RAID, как зеркалирование, репликация и миграция. В некоторых виртуализаторах (например, Brocade 7600 и лезвие FA-18) эти функции реализуются аппаратно, что обеспечивает высокую производительность, которую нельзя получить при использовании решений на уровне хостов.

По-видимому, в будущем будет использоваться комбинация всех трех подходов к управлению томами. Выполняющееся на уровне хоста программное обеспечение лучше подходит для резервирования каналов, RAID-массивы будут по-прежнему применяться для организации высокопроизводительных и защищенных LUN, а сеть – для миграции и репликации данных между массивами, загрузки через SAN и управления отображением томов для решений управления жизненным циклом информации (ILM) и Ресурсных вычислений (Utility Computing, UC).

## Протоколы SAN

Все продукты, которые были рассмотрены в предыдущих разделах, используют протоколы, а чаще комбинацию нескольких протоколов. Архитекторы SAN должны учитывать особенности каждого протокола SAN при выборе оборудования, планирования производительности, доступности, надежности и расширяемости в будущем.

Под *протоколами* понимаются “правила поведения компьютеров и сетевых устройств, позволяющие им обмениваться данными через сеть”. Если в сети устройства используют *разные протоколы*, то они не

смогут обмениваться данными. Например, человек, который говорит только по-английски, не сможет вести сложные философские дебаты с оппонентом, который говорит только на суахили. На самом деле проблемы могут возникнуть даже в том случае, если один из собеседников говорит на американском варианте английского, а другой – на британском (либо один использует парижский диалект французского языка, а другой – канадский, либо при диалоге испанца и мексиканца). Аналогичным образом сетевые устройства должны “говорить” на одном языке (например, английском) и использовать один и тот же диалект (например, американский вариант английского). Это означает, что требуется соблюдаемое всей индустрией соглашение об официальных стандартах и их применении де-факто<sup>17</sup>.

Протоколы применяются на всех уровнях обмена данными, начиная от физической среды и кабелей, до уровня приложений. При “общении” двух устройств задействуются протоколы разных уровней. Группа протоколов разного уровня называется *стеком протокола*. Если какой-то протокол из стека работает неправильно, то не будет работать и связь между устройствами.

В этом подразделе описаны протоколы, применяемые сегодня в индустрии хранения. Основной акцент делается на таких сетевых протоколах, как Fibre Channel, и меньше внимания уделяется протоколам уровня

---

<sup>17</sup>

Применение протоколов на практике часто отличается от стандартов (также как на практике не соблюдаются правила грамматики английского языка). Дело в том, что стандарты создаются до того, как продукты, поддерживающие их, выйдут на рынок, поэтому при применении этих продуктов могут возникнуть проблемы, о которых не могли знать авторы стандартов, поэтому чтобы продукт работал на практике иногда нужно отклониться от стандартов.

приложений, например, хсору для бессерверного резервного копирования. Это не означает, что протоколы уровня приложений – второстепенные, просто данная книга посвящена проектированию сетей, поэтому она больше ориентирована на сетевые технологии.

## ***SCSI***

Протокол Small Computer Systems Interconnect (SCSI) является основной технологией для построения современных инфраструктур хранения. Первоначально он был разработан как протокол для прямого подключения устройств хранения Direct Attached Storage (DAS) с короткой шиной, напрямую связывающей контроллер SCSI с одним и только одним хостом. В отличие от технологии point-to-point ( см. Рис. 2), SCSI позволяет подключить более одного узла хранения, но *не намного* больше чем один. Дополнительно, остается ограничение на один инициатор. Все это не позволяло строить сколько-нибудь значимые решения. Дальнейшие усовершенствования SCSI увеличили число поддерживаемых узлов нашине и ее длину, а в некоторых случаях даже позволили использовать второй инициатор, однако принципиально новые возможности были реализованы только после появления оптимизированных для сетей хранения протоколов, в том числе Fibre Channel.

Современные решения SAN инкапсулируют протокол SCSI на другой транспорт, обычно Fibre Channel. Такое отображение позволяет использовать стандартизацию SCSI (как официальные стандарты, так и стандарты де-факто для их применения на практике), поэтому вендоры, разрабатывающие решения для подключения к SAN хостов и устройств хранения, ведут проектирование и тестирование на основе стандартов SCSI. Это также снижает риск внедрения таких протоколов SAN, как Fibre Channel, ведь основные

протоколы сети тщательно протестированы и обладают высокой стабильностью, поскольку протокол FC в значительной степени унаследовал коды от протоколов SCSI, которые успешно используются на практике уже несколько десятилетий.

## ***Fibre Channel***

Основной протокол, используемый сегодня для SAN, - это Fibre Channel. К настоящему времени Brocade и другие вендоры поставили многие миллионы портов инфраструктуры Fibre Channel для производственных применений. Fibre Channel – это единственный протокол, каждый уровень которого спроектирован в расчете на сети хранения, поэтому FC чаще всего используется в SAN как наиболее технологически совершенный.

### Стек протоколов Fibre Channel

Fibre Channel состоит из набора протоколов начиная от физического уровня и до уровня приложений. Этот набор протоколов появился в 1994 году<sup>18</sup> и быстро стал стандартом, с которым сравнивали другие протоколы SAN. Fibre Channel используется на практике более десяти лет и успешно применяется для критически-важных приложений – на его долю приходится более 99% всех инсталляций SAN. Доля всех остальных технологий SAN, включая IP SAN, не превышает доли процента инсталляций SAN.

Сначала подсистемы Fibre Channel работали на скорости 250Mbits, но современные сети FC могут использовать скорости 1Gbit, 2Gbits, 4Gbits, 8Gbits или

---

<sup>18</sup> Стандарт FC-PH определил 250Mbit, 500Mbit и 1Gbit FC в 1994 году. Более подробную информацию о стандартах FC можно найти на web-сайте INCITS T11: <http://www.t11.org>.

10Gbits<sup>19</sup>. Кроме определения поведения продуктов и сервисов Fibre Channel стандарты FC определяют инкапсуляцию протоколов высокого уровня (например, SCSI или IP) для их передачи по FC-сетям и инкапсуляцию Fibre Channel для передачи по другим протоколам (например, ATM, SONET/SDH или IP). На Рис. 1 (стр. 6) показан стек протоколов Fibre Channel. Отметим, что здесь протоколы распределены по разным уровням. На Рис. 7 с помощью многоуровневости объясняется, как поток данных идет между уровнями Fibre Channel между приложениями на хосте и энергонезависимыми носителями (non-volatile media) в RAID-массиве.

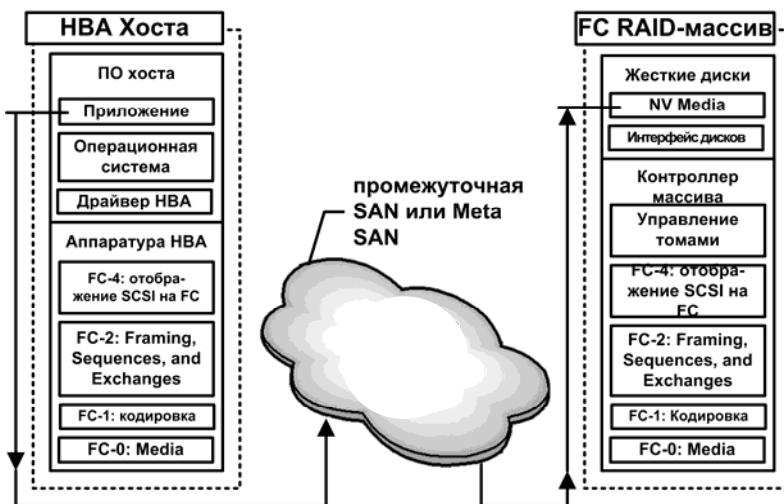


Рис. 7 – Передача данных между уровнями FC

19

Недавно FCIA утвердила разработку 8Gbit FC, который намного более эффективен по затратам, чем 10Gbit и может использоваться в одной инфраструктуре с 2Gbit и 4Gbit.



## Заметки на полях

*Насколько велик пакет Fibre Channel? Ответ на этот вопрос зависит от того, что понимается под словом “велик”. Пакеты FC используют заголовки фиксированной длины, но поддерживают переменную длину полезной информации пакета, поэтому их размер может быть в диапазоне от 60 байт до 2 килобайт. “Размер” пакета FC можно определить и исходя из скорости света в оптическом канале. Требуется примерно один буферный кредит на 2 км кабеля для того, чтобы заполнить линк 1Gbit FC 2-килобайтными пакетами. Это означает, что при передаче 2-килобайтный пакет будет распределен по такой длине кабеля, поэтому можно утверждать, что размер пакета 1Gbit FC – 2 км. Аналогичным образом при использовании 2Gbit FC один кредит нужен на каждый километр и для 2Gbit “размер” пакета FC равен 1 км, а при 4Gbit длина пакета FC – 500 м. По мере роста скорости уменьшается физический размер пакета, но логический размер упакованных в него данных остается постоянным. Это имеет значение при проектировании SAN, где используется связь на очень большие расстояния.*

В этом примере приложение на хосте генерирует блок данных, который записывается на массив. Оно «говорит» об этом ОС, которая в свою очередь передает эту информацию драйверу НВА. Обычно драйвер получает указатель на область памяти, куда записаны данные, а не сами данные (это улучшает эффективность). Драйвер НВ А "говорит" аппаратуре НВА где взять данные и куда их записать, в результате перемещение данных между памятью и сетью выполняет

только НВА без использования ресурсов центрального процессора<sup>20</sup>. НВА считывает блок данных из ОЗУ основной системы, инкапсулирует данные через уровни FC и пересыпает в сеть поток пакетов. На другом конце сети RAID-массив выполняет аналогичные операции для записи данных на нужный диск (или диски). В этом случае данные никогда не приходят к приложению в "сыром" виде - они отображаются с помощью аппаратно реализованного менеджера томов (например, RAID 5), который определяет, на какие физические диски нужно записать конкретные блоки данных из потока.

Как опция, промежуточная сеть может содержать линки MAN или WAN и в таком случае FC может инкапсулироваться через IP, SONET/SDH или ATM. Это не заменяет многоуровневую архитектуру FC, а только дополняет её. Аналогичным образом, тот же НВ А, который передает SCSI по Fibre Channel в данном примере может передавать IP по FC на уровне FC-4. При этом трафики IP и SCSI будут обрабатываться параллельно<sup>21</sup>. В большинстве приложений SAN использует «родной» Fibre Channel для передачи SCSI (FCP), что позволяет получить очень эффективный стек протоколов.

Когда данные попадают в SAN, то они состоят из потока *пакетов*. На Рис. 8 показана структура пакетов FC на высоком уровне.

---

<sup>20</sup> Это – главное отличие между Fibre Channel и дешевыми решениями iSCSI: для экономии в iSCSI используются стандартные NIC, в которых нет аппаратного ускорения, разгружающего центральный процессор.

<sup>21</sup> Многие из ранних приложений для коммутаторов фабрик использовали IP over FC. Оказалось, что Fibre Channel лучше передает IP, чем Ethernet.

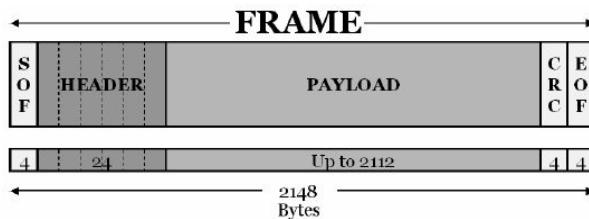


Рис. 8 – Пакет Fibre Channel

Как видно из этой диаграммы, пакет начинается с разделителя начала пакета, короткого заголовка и 2 килобайт содержания пакета, где обычно записаны данные SCSI. Разумеется, большинство современных файловых систем используют блоки данных больше 2 килобайт, поэтому Fibre Channel использует механизм группировки до 65 килобайт данных в *последовательность (sequence)*. В предыдущем примере было показано, как НВА извлекает блок данных непосредственно из ОЗУ и упаковывает их в пакеты. Последовательность пакетов является базовой единицей при передаче данных с использованием аппаратного ускорения, поэтому более сотни мегабайт данных можно передать через Fibre Channel НВА и при этом загрузка центрального процессора будет такая же, как при передаче только одного пакета Ethernet без применения аппаратного ускорения. Fibre Channel позволяет даже объединять последовательности в один *обмен (exchange)*. Обычно каждая операция SCSI ( чтения или записи) отображается в один exchange ID (смотри Раздел iSCSI начиная со стр. 51.)

### Линки ISL и IFL

Линк Inter-Switch Link (ISL) соединяет два коммутатора в фабрику Fibre Channel. ISL крайне важны для проектирования SAN - если они не используются, то SAN состоит из изолированных между собой коммутаторов, что существенно ограничивает

возможности подключения, поскольку тогда SAN, строго говоря, не является полноценной сетью.

Для формирования ISL между двумя коммутаторами Brocade, для которых приобретены лицензии фабрики, требуется только соединить их кабелем. Brocade поддерживает U\_Port - стандарт автоматического согласования типов портов, поэтому порты коммутаторов на обоих концах линка автоматически превращаются в E\_Port и формируется ISL.

Имеются стандартные сервисы и протоколы ISL для обеспечения “наименьшего общего кратного” сети и у каждого ведущего поставщика оборудования SAN есть уникальный набор расширений для реализации более сложных функций. Например, Brocade поддерживает режим “native” ISL, улучшающий масштабируемость, безопасность, RAS и управляемость. Обычно эти усовершенствования реализуются на уровне сервисов фабрики, хотя один из вендоров для этого применяет отличный от стандартного заголовок пакетов<sup>22</sup>.

Линки Inte r-Fabric Link (IFL) похожи на ISL. В них используются те же стандарты, что и в ISL, и обычно поддерживаются те же фирменные усовершенствования вендоров. На стороне фабрики в IFL линк по-прежнему формируются с помощью E\_Port. Разница в том, что IFL соединяет коммутатор FC и маршрутизатор FC-FC, а не два коммутатора. В соответствии со своим названием IFL обеспечивает поток данных между фабриками для формирования Meta SAN ( см. ниже), а ISL – передачу трафика и сервисов между коммутаторами в одной фабрике.

---

<sup>22</sup>

Тэги VSAN.

Для катастрофоустойчивых решений можно удлинять ISL и IFL на большие расстояния. Теоретически, они могут охватывать весь земной шар, но на практике из-за ограниченной мощности лазера максимальное расстояние не превышает ста километров. Максимальное расстояние можно увеличить вдвое с помощью мультиплексоров active wave division, например, шасси DWDM, а при расстояниях свыше 200 км используются шлюзы протоколов ATM, SONET/SDH или FCIP. ( эти протоколы рассматриваются дальше в этой главе).

### Сравнение фабрики с SAN и Meta SAN

Можно связать много узлов Fibre Channel с помощью одного или нескольких коммутаторов/директоров FC . Если используются несколько коммутаторов, то они соединяются через Inter-Switch Links (ISL ). В этом случае вся сеть называется *фабрикой*, а иногда просто SAN. Под этими терминами понимаются все коммутаторы и программное обеспечение сервисов фабрики, а также в зависимости от контекста – узлы и их программное обеспечение для управления хранением данных<sup>23</sup>.

Теоретически, архитектура фабрик Fibre Channel поддерживает миллионы устройств, поскольку в них используются трехбайтовые адреса<sup>24</sup>, однако на практике самые большие фабрики содержат несколько

---

<sup>23</sup> В этом контексте “узел” – это адаптер FC на конечном устройстве сети, например, хоста или массива хранения. Fibre Channel определяет поведение узла в стандартах N\_Port и NL\_Port. (“N” означает “node”, т.е. узел).

<sup>24</sup> Некоторые адреса зарезервированы для сервисов фабрики, а некоторые диапазоны адресов недоступны для небольших коммутаторов. Однако, даже с учетом ограничений, адресное пространство можно считать бесконечным если сравнивать его даже с самыми большими фабриками, которые сегодня используются или проектируются.

тысяч устройств<sup>25</sup> (причины этого ограничения масштабирования рассмотрены далее в разделе “Сервисы фабрики”). Помимо масштабируемости, имеются характеристики доступности и управляемости, которые относятся ко всей фабрике. В определенных случаях это является плюсом (например, управление всеми зонами фабрики с одной консоли упрощает текущее администрирование), но в других может быть и минусом (например, если подразделению компании нужна «собственная» фабрика, включая все оборудование и программное обеспечение). Даже в последнем случае могут потребоваться выборочные возможности подключения этих фабрик (см. Рис. 9).

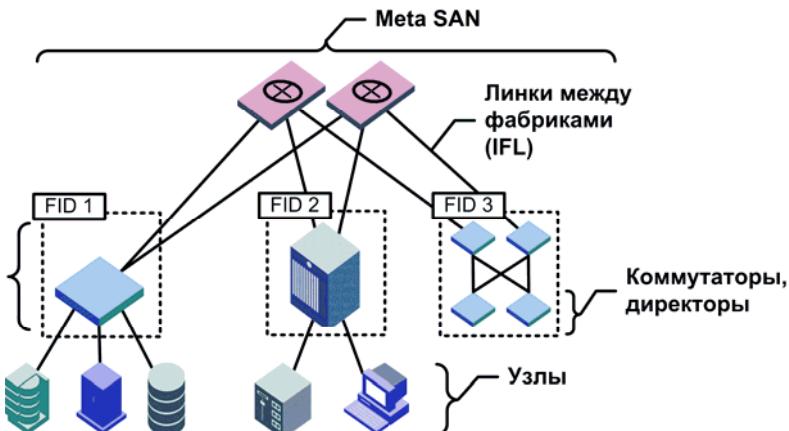
Как видно из этого рисунка, можно объединить несколько островов SAN (т.е. изолированных фабрик) вместе с помощью маршрутизаторов FC-FC<sup>26</sup> (стр. 12), соединенных с помощью Inter-Fabric Links (IFL). В данном случае полученная большая сеть называется *Meta SAN*, поскольку она реализует уровень иерархии выше традиционной SAN. В Meta SAN каждая фабрика идентифицируется с помощью уникального байт *Fabric Identifier* (FID). Для соединения устройств из разных фабрик Meta SAN создаются Logical Storage Area Networks (LSAN), представляющие собой зоны, охватывающие несколько фабрик. Маршрутизаторы FC-FC обеспечивают подключение и могут использоваться для увеличения архитектурной и практической масштабируемости на несколько порядков, поскольку они добавляют еще один байт (“Fabric ID” или *FID*) к адресному пространству фабрики и решают проблему масштабирования control-plane, которая до сих пор

---

<sup>25</sup> У некоторых клиентов Brocade в фабрике более 4 тысяч портов.

<sup>26</sup> Маршрутизация SAN подробно описана в книге *Multiprotocol Routing for SANs*.

ограничивала максимальный размер фабрики. Тем не менее, у каждой фабрики сохраняется собственная копия сервисов фабрики и она может управляться отдельно от остальных фабрик Meta SAN и их устройств.



**Рис. 9 - Части Meta SAN**

### Сервисы фабрики

Фабрика Fibre Channel реализует группу программных функций, которые называются “сервисами фабрики”. Эти сервисы используют стандартные протоколы и общепринятые методы внедрения для предоставления таких функций в масштабе всей фабрики, как сервис имен. В продуктах Brocade сервисы фабрики разработаны по принципу plug-and-play, что упрощает управление ими. Сеть хранения использует эти сервисы практически при любой операции. Ниже перечислены некоторые из самых важных функций, которые реализуются с помощью сервисов фабрики:

- Domain Address Manager для назначения ID домену
- Domain Controller для подключения к узлу (FLOGI)

- FSPF для маршрутизации между коммутаторами разных фабрик
- FCRP для маршрутизации внутри фабрики LSAN в Meta SAN
- Name Server для хранения информации о подключенных устройствах
- Зонирование для контроля доступа на уровне портов и WWN

Архитекторы SAN должны много внимания уделить проектированию архитектуры сервисов фабрики, так от нее зависят все аспекты проекта SAN, в том числе совместимость, взаимодействие, надежность, доступность, обслуживаемость и управляемость.

Например, сервисы фабрики обычно выполняются на центральных процессорах, встроенных в коммутаторы фабрики Fibre Channel, но в продуктах iSCSI таких процессоров нет, поэтому для внедрения решения iSCSI нужно предусмотреть бюджет на приобретение внешних серверов и программного обеспечения к ним, а также затраты на оплату консультантов, которые будут инсталлировать и настраивать эти системы, оплату отдельного контракта на техобслуживание вместе с затратами, связанными с усложнением администрирования. Если SAN строится на основе решений Fibre Channel, то эти дополнительные расходы отсутствуют.

Для вендора SAN разработка программного обеспечения, которое реализует сервисы фабрики, - самая сложная задача при проектировании любого шлюза или маршрутизатора SAN, независимо от того, использует ли SAN сервисы фабрики Fibre Channel или аналогичное программное обеспечение iSCSI. Хотя в этой книге нет детального описания сервисов iSCSI,

отметим, что для каждого сервиса фабрики FC есть аналогичная функция iSCSI.<sup>27</sup> <sup>28</sup>

С точки зрения доступности и масштабируемости сервисы фабрики являются едиными для всей фабрики. Для маршрутизации FC-FC это означает, что в Meta SAN у каждой «границной» фабрики есть одна полностью изолированная копия сервисов<sup>29</sup>.

Сервисы фабрики создают серьезную проблему при тестировании на совместимость. Например, чтобы любой узел N\_Port или NL\_Port мог подключиться к фабрике требуется совместимость сервиса подключению к фабрике (FLOGI) и имен (S NS). Brocade потратила десятилетие на обеспечение совместимости своих продуктов с различными установленными у клиентов устройствами FC и ни один другой вендор не способен обеспечить такое тестирование и интеграцию решений (см. “Совместимость” на стр. 122).

---

<sup>27</sup> Точнее, имеется *теоретическая* возможность создания аналога iSCSI для каждого сервиса фабрики, однако пока вендоры iSCSI реализовали только малую часть из сервисов фабрики FC.

<sup>28</sup> Для внедрения SAN необходимо глубокое понимание ее сервисов. При внедрении iSCSI одних знаний IP недостаточно. Архитектор SAN должен хорошо разбираться в сервисах, реализуемых на уровне хоста, систем хранения и SAN, а не только в сетевых протоколах. Ему потребуется экспертиза в области устройства томов RAID, драйверов программного обеспечения дублирования каналов, сервисов имен, совместимости драйверов и т.п. Глубокое понимание форматов пакетов и структуры заголовков не так важно при построении SAN, независимо от того, какой протокол в ней используется.

<sup>29</sup> Как и VSAN сети Meta SAN обеспечивают соединение между фабриками, но при этом изолируют их между собой логически и физически, поэтому маршрутизаторы Brocade FC улучшают доступность, масштабируемость и совместимость, а VSAN не способны на это.

### Топологии FC без фабрики

Кроме фабрики существуют еще два варианта Fibre Channel: point-to-point и arbitrated loop (петли с арбитражем, FC-AL).

Point-to-point – это метод подключения напрямую устройств хранения (DAS). Подключение порта RAID-массива напрямую к НВА хоста может выполняться по схеме FC point-to-point и в этом случае получается не SAN, а длинная шина, которая поддерживает только один инициатор и один получатель. В этой книге point-to-point не рассматривается подробно.

Fibre Channel Arbitrated Loop всегда используется в JBOD (стр. 18) и иногда в НВА (стр. 17), ленточных приводах (стр. 23) и контроллерах RAID- массивов (стр. 18).

Поддержка FC-AL необходима для коммутаторов Fibre Channel, поскольку иначе нельзя было бы соединяться с устройствами используя эту топологию. Тем не менее, по возможности надо использовать устройства N\_Port вместо NL\_Port. В петле используется семибитные адреса и поэтому в ней может быть немногим более сотни устройств. На практике не удается достичь даже этого уровня. Для решения этой проблемы Brocade использует функцию *phantom logic* в интегральных схемах ASIC коммутатора, выполняющие преобразование сетевых адресов Network Address Translation (NAT) между петлями FC-AL и фабриками, что на несколько порядков улучшает эффективность использования устройств FC-AL.

### **ATM и SONET/SDH**

В разделе “Стек протоколов Fibre Channel” (стр. 36) уже упоминалось, что протокол Fiber Channel может передавать трафик различных протоколов (например,

SCSI или IP) и в свою очередь его можно инкапсулировать в другие протоколы, например, можно передавать пакеты Fibre Channel через сети ATM и/или SONET/SDH. Это используется, когда нужно обеспечить высокую производительность и надежность в кампусных сетях, MAN или WAN, где нельзя использовать «родной» Fibre Channel.

ATM расшифровывается как Asynchronous Transfer Mode. Это транспорт на основе коммутации ячеек, который используется в небольших крупных сетях от LAN и до WAN. ATM передает короткие блоки данных фиксированной длины. Когда большие блоки данных, например, пакеты FC, передаются по сети ATM, они разбиваются на ячейки и на выходе из сети снова собираются.

SONET расшифровывается как Synchronous Optical Networks. В Европе и Азии эта технология называется Synchronous Digital Hierarchy (SDH), поэтому для ее обозначения часто используется сокращение SONET/SDH. Этот протокол вносит минимальную задержку, способен поддерживать полную полосу пропускания FC, отличается высокой надежностью и может работать на достаточно большой расстоянии.

У решений ATM и SONET/SDH производительность и надежность лучше, чем у решений IP SAN, но и стоят они дороже.

## ***IP и Ethernet***

Internet Protocol (IP) – это сетевой стандарт Internet и де-факто стандарт для корпоративных локальных сетей, обслуживающих такие низкопроизводительные приложения, как электронная почта и web- серверы на базе настольных ПК. В большинстве LAN трафик IP передается через Ethernet. Протоколы верхнего уровня (например, HTTP и FTP) обрабатываются поверх

IP и обычно между ними и IP располагается TCP для обнаружения ошибок. Адрес IPv4 состоит из четырех байтов, обычно представленных в десятичном виде и разделенных точками. Пример стандартного формата IP-адреса -192.168.1.1.

IP имеет множество преимуществ если он используется в тех приложениях, для которых он был первоначально разработан. Например, разработчики IP с самого начала старались обеспечить поддержку таких сильно распределенных и слабо связанных решений, как Internet, и поэтому архитектура этого протокола оптимизирована именно для таких сред<sup>30</sup>. Однако не возможно создать универсальный продукт или технологию, которая подходит для любых задач, поэтому прежде чем принимать IP для конкретного приложения необходимо понять, насколько он соответствует требованиям этого приложения.

Коммутируемая и/или маршрутизируемая IP-инфраструктура редко используется в критических к производительности приложениях и в задачах, когда потеря или искажение данных не допустимы, поскольку при разработке протокола IP не учитывались эти требования. Например, протоколы маршрутизации IP рассчитаны на обслуживание сетей с миллионами узлов, где временные каналы между узлами строятся по

---

<sup>30</sup> Спецификация IP требует слабой связанности подсетей IP как самого важного критерия архитектуры. Любое конкретное соединение считается расширяемым до тех пор, пока сеть сохраняет работоспособность. С другой стороны, Fibre Channel был разработан прежде всего для поддержки критически-важных подсистем хранения - на первом плане была не масштабируемость, а обеспечение очень высокой по сравнению с IP скорости и надежности. Появление маршрутизаторов FC-FC позволило получить лучшее от этих двух технологий – масштабируемую иерархическую архитектуру сети вместе с производительностью и надежностью Fibre Channel.

псевдослучайному алгоритму и не требуется обеспечить высокую производительность или надежность на время существования таких каналов. IP-сети могут терять пакеты или нарушать порядок при их доставке если коммутатор или маршрутизатор перегружены, причем не предусмотрены инструменты исправления таких ошибок. Если создать многоуровневый стек протоколов, например, [ Xcopy over SCSI over iSCSI over IPsec over TCP over IP over Ethernet over 1000baseT ], то обработка всех этих протоколов создаст существенную дополнительную нагрузку и в результате полоса пропускания резко сократится. Также такой стек сильно загрузит центральные процессоры и память серверов (если только на каждом узле будет установлено дорогое оборудование). Переконфигурирование при обрыве линка в больших сетях может выполняться несколько минут даже при использовании самого лучшего оборудования и протоколов.

Для Intel это означает ограниченную производительность и периодический обрыв соединения, из-за чего пользователям приходится снова загружать в браузер web-страницу. Эти недостатки считаются приемлемыми для тех приложений, которые обычно развертываются в IP-сетях.

С другой стороны, к SAN подключается меньше устройств (в самых больших – десятки тысяч), но все подключения структурированы, сохраняются длительное время и для них требуются наивысшие производительность и надежность. Неспособность IP обеспечить выполнение этих требований хорошо известна в индустрии и она же стала основным стимулом развития Fibre Channel и других технологий SAN. Для критически-важной базы данных недопустим повторный запрос при потере данных в сети и поэтому инфраструктуры SAN на несколько порядков надежнее

IP-сетей.

Стоит отметить, что *возможно* использование IP в инфраструктуре SAN и Brocade предлагает разные решения IP SAN, например, F CIP и iSCSI, которые описаны далее. Однако это решения подходят только для небольшого числа приложений и поскольку Brocade предлагает такие оптимизированные для SAN технологии, как FC, то заказчикам следует рассматривать решения IP SAN (даже от Brocade) лишь как запасной вариант.

### **iSCSI**

iSCSI – это медленно развивающийся протокол для передачи SCSI по IP-сетям. Его концепция похожа на отображение FC-4, которое уже обеспечивает FCP для Fibre Channel.

Brocade поставляет как продукты FC, так и iSCSI, поэтому заказчики могут использовать любую из этих двух технологий, но они при выборе должны учитывать существенные различия между iSCSI и FC.

Например, у пакетов iSCSI накладные расходы протокола больше, чем у пакетов FC. В отличие от Ethernet, IP, TCP и IPSec протокол FC с самого начала разрабатывался как эффективный транспорт для хранения данных. Для инициации обмена данными с получателем по iSCSI хост должен построить заголовок, пример которого показан на Рис. 10.

Разумеется, у некоторых пакетов iSCSI длина заголовка может быть меньше, чем в данном примере, но у других пакетов он может быть даже длиннее, а у всех пакетов Fibre Channel заголовок будет одинаково коротким. Из сравнения на Рис. 11 (стр. 53) видно, как комбинация заголовка iSCSI и стандартного для Ethernet

приводит к еще большей неэффективности по сравнению с FC.

Если только вся сеть iSCSI состоит из дорогих мощных коммутаторов и маршрутизаторов, для обмена данными между конечными устройствами iSCSI нужно использовать наименьший общий кратный размер пакета (MTU), поэтому единственный вариант – это большие (jumbo) пакеты Ethernet. Все вендоры iSCSI рекомендуют использовать пакеты jumbo и коммутаторы корпоративного класса для получения максимальной производительности. На самом деле iSCSI настолько неэффективен, что эксперты считают, что только 10Gbit Ethernet может решить проблемы iSCSI, поскольку реальная производительность iSCSI по 10 Gbit Ethe gnet будет “на уровне” SCSI по 4Gbit Fibre Channel.

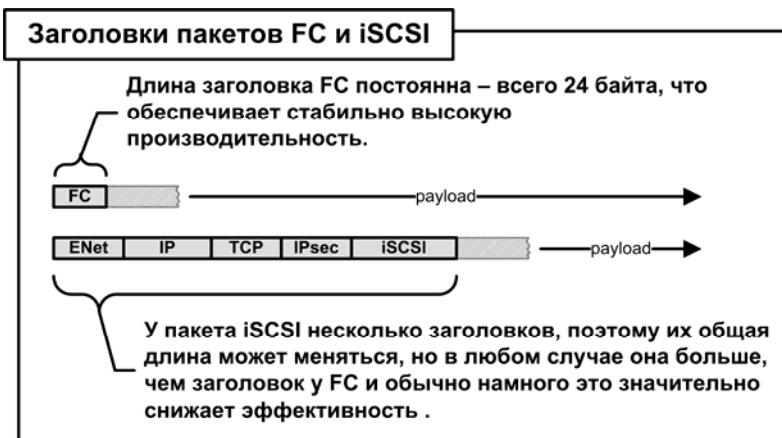


Рис. 10 – Сравнение заголовков iSCSI и FC

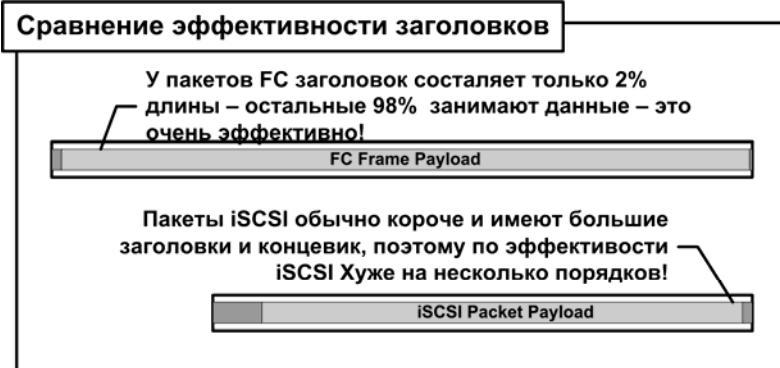


Рис. 11 – Эффективность заголовков iSCSI и FC

Однако при этом стоимость инфраструктуры iSCSI начинает превышать стоимость сети Fibre Channel, хотя скорость немного меньше и по-прежнему требуются дорогие внешние серверы для работы таких сервисов хранения, как сервер имен. Даже дорогие коммутаторы 10Gbit Ethernet корпоративного класса неспособны полностью устраниć проблему, поскольку причина низкой эффективности – это не только размер пакета и скорость. iSCSI существенно больше загружает процессоры хоста поскольку большой заголовок не только уменьшает полосу пропускания и увеличивает задержки, но его нужно еще и сгенерировать и интерпретировать. В большинстве случаев из-за этого требуется дорогой контроллер системы хранения и расходуются циклы центрального процессора сервера, что снижает скорость работы приложений, он которые обслуживает.

**Сравнение загруженности процессора**

АдAPTERЫ Fibre Channel объединяют в одну последовательность до 65 536 пакетов, что обеспечивает высокую эффективность использования ресурсов центральных процессоров хоста.

Даже при использовании пакетов "jumbo" iSCSI может передавать в одном пакете только 8 кбайт данных, а FC с помощью аппаратного ускорения способна передавать более 100MB в одной последовательности, поэтому использование CPU программным iSCSI в 17 тыс. раз менее эффективно!

Рис. 12 – Дополнительная загрузка центральных процессоров при использовании FC и iSCSI

Как уже говорилось в разделе “Fibre Channel” (стр. 36), пакеты FC объединяются в последовательности и вся последовательность может передаваться за один такт центрального процессора, после чего он продолжает обслуживать приложения хоста, а все перемещение данных выполняет НВА. На Рис. 12 (стр. 54) приведено сравнение передачи НВА-адаптерами последовательности пакетов и программной реализации iSCSI, при которой на каждый пакет тратится по крайней мере один такт процессора. На самом деле на Рис. 12 даже не соблюден масштаб, иначе на нем не уместился бы jumbo-пакет iSCSI. Разумеется, есть определенные задачи ввода/вывода, для которых этот недостаток iSCSI не так важен (например, если SAN используется только для обслуживания файловых систем с очень маленьким размером блока, обращение к файлам носит случайный характер и у всех хостов процессоры не загружены полностью) и в таких ситуациях имеет смысл использовать iSCSI. Однако при другом типе трафика этот недостаток будет сильно влиять на работу SAN и если в файловой системе используется средние или большие блоки или обращение к файлам происходит постоянно (например, при резервном копировании,

миграции данных и синхронизации томов), то оптимальный вариантом всегда будет FC.

Повысить производительность iSCSI можно, если на каждый хост установить специализированные HBA-адаптеры iSCSI, которые освобождают процессоры хоста от необходимости генерировать заголовки пакетов<sup>31</sup>, однако применение этих устройств увеличивает стоимость и сложность внедрения решений iSCSI, которые в результате теряет свое ценовое преимущество. В то же время одногигабитные iSCSI HBA с аппаратным ускорением работают более чем на 75% медленнее, чем имеющие такую же цену FC HBA.

Кроме транспорта протокол iSCSI определяет сервисы присваивания имен, обнаружения, контроля доступа и безопасности, что также аналогично существующим сервисам Fibre Channel, хотя сервисы iSCSI только начинают разрабатываться. У большинства сервисов Fibre Channel есть свои аналоги iSCSI, но они не такие зрелые и развертываются отдельно от сетевых коммутаторов, что ведет к дополнительным расходам на iSCSI, связанным с затратами рабочего времени и оборудования для развертывания, и усложняет текущее управление, поскольку в коммутаторах FC эти сервисы являются встроенными.

У читателя может возникнуть вопрос «а стоило ли разрабатывать iSCSI если уже существует более совершенное техническое решение, которое проверено на практике и широко используется на рынке?». Многие аналитики сейчас предсказывают, что сфера применения iSCSI будет «зажата» между Fibre Channel в

---

<sup>31</sup> Отметим, что это не просто карты для аппаратного ускорения TCP – они обрабатывают заголовки и выполняют операции PDU, освобождая таким образом процессорные ресурсы хоста.

корпоративном секторе и Serial ATA в сегменте начального уровня. После появления Fibre Channel over Ethernet (F CoE) можно уверенно утверждать, что дни iSCSI сочтены. Пока можно найти только один тип приложений, для которых имеет смысл использовать iSCSI – дешевые подключения начального уровня, например, работающие на настольных ПК пользователей web-серверы. Для таких приложений главное – это цена, но весьма скромные требования к производительности и надежности: некоторым серверам начального уровня не требуется высокая производительность сети хранения или центрального процессора и если они зависнут или их данные будут испорчены, то данные можно легко восстановить по резервной копии. Клиенты с такими приложениями для подключения обычно используют файловые системы NFS или CIFS. iSCSI подойдет только для некритичных приложений, которым нужен доступ на уровне блоков вместо доступа на уровне файлов с помощью сетевой файловой системы.

Многие аналитики считают, что в конце концов iSCSI превратится в протокол для Network Attached Storage (NAS) и не будет использоваться для SAN, тем более что пока всерьез iSCSI применяется только в секторе NAS. Для нишевых применений iSCSI в SAN компания Brocade разработала платформу Brocade iS CSI Gateway, лезвие iSCSI для директора 48000 и iSCSI HBA.

Недавно индустрия хранения начала разрабатывать новый стандарт Fibre Channel over Datacenter Ethernet (FCoE). Этот стандарт позволит строить сети хранения с помощью дешевого оборудования Ethernet (это и является основным ценовым преимуществом iSCSI), но в то же время обладает присущими для Fibre Channel сервисами и надежностью. По-видимому, эта технология станет могильщиком оставляя никаких перспектив для SAN на базе iSCSI в долговременной перспективе. iSCSI пока имеет шансы стать реальной альтернативой

протоколам NAS, но эта технология не может конкурировать с FCoE или «родным FC» в секторе SAN.



## Заметки на полях

*Сначала iSCSI была привлекательна из-за своей дешевизны для SAN начального уровня, но сокращение цен на FC ликвидировало это преимущество iSCSI. В то же время в этом секторе начинает использоваться технология Serial ATA и в результате продукты iSCSI уже не могут оправдать прежние ожидания из-за своей низкой функциональности, надежности, производительности, а также проблем совместимости и дополнительных расходов, необходимых для внедрения этой технологии. Как отмечала недавно одна специализированная служба новостей по хранению “из-за появления недорогих продуктов FC маркетологам iSCSI придется серьезно скорректировать свои презентации”*

*Brocade предлагает iSCSI для пользователей, которым требуется сократить совокупную стоимость владения и не предъявляющих высоких требований к производительности и надежности. Однако в большинстве современных SAN используется FC и эта ситуация сохранится и в будущем. Из-за своей низкой окупаемости продукты iSCSI любого вендора (в том числе и Brocade) следует применять только в том случае, когда по экономическим причинам они предпочтительнее FC.*

## **iFCP**

iFCP разрабатывался для замены фабрик Fibre Channel на IP-сети. Этот протокол унаследовал ограничение Fibre Channel, связанное с отсутствием в этом протоколе механизма подключения узлов – все узлы должны быть

Fibre Channel. В полностью “родном” решении iFCP все устройства FC подключаются к дорогому порталу преобразования протоколов FC-to-iFCP, а не к более простому коммутатору пакетов FC. Перед передачей каждого пакета выполняется преобразование протоколов, затем он передается на другой порта, где обратно преобразовывается в Fibre Channel, причем эти преобразования выполняются и даже если между портами нет оборудования IP- сети. Это двойное преобразование каждого пакета удорожает решения iFCP и снижает их производительность.

На практике iFCP используются только в географически-распределенных IP- сетях, которые сами по себе являются “дорогими и медленными”, что делает эту технологию прямым конкурентом FCIP, которая рассматривается далее). Раньше у iFCP было важное преимущество по сравнению с FCIP – эта технология позволяла изолировать граничные фабрики от нестабильности WAN, однако теперь это способна обеспечить FCIP с помощью сервиса FC-FC Routing Service.

Хотя iFCP ратифицирована как стандарт, на практике она применяется редко и только один вендор решился выпустить на рынок решение iFCP, однако после поглощения этого вендора никто больше не предлагает iFCP и, по-видимому, вскоре этот протокол останется только на бумаге.

## ***FCIP***

Fibre Channel over Internet Protocol (FCIP) – это механизм соединения портов Fibre Channel E\_Port через инфраструктуру IP. В то же время возможно сконфигурировать линки FCIP по схеме point-to-point без использования между ними промежуточного оборудования IP- сети и в этом случае физическую топологию можно рассматривать как

использующую линки Fibre Channel ISL в качестве туннелей. Однако такую архитектуру более эффективнее (как по затратам, так и производительности) можно построить с помощью «родных» каналов Fibre Channel, поэтому на практике FCIP внедряется с помощью IP-коммутаторов и/или маршрутизаторов между шлюзами FCIP (см. Рис. 13)<sup>32</sup>.

В этом примере две площадки соединены через IP WAN и на каждой установлены директор Fibre Channel (например, Brocade 24000 или 48000) и шлюз FCIP (например, Brocade 7500 или лезвие FR4-18i). Шлюз подключен к опционному коммутатору LAN, который соединен с маршрутизатором IP WAN. После конфигурирования решения хост с одной площадки получит доступ к ресурсам хранения другой площадке так, как если коммутаторы были напрямую соединены через FC ISL и между ними не было IP-сети.

---

<sup>32</sup>

На практике использование линков point-to-point FCIP имеет смысл только в определенных ситуациях, например, если у клиента есть шасси DWDM с картами Gigabit Ethernet, то он может захотеть соединить порты FCIP с шасси, а не с «родным» FC. Однако, потребуется несколько (больше двух) линков FCIP (more than two) для получения производительность, эквивалентной одному «родному» линку FC и даже в этих случаях многие сценарии использования SAN не поддерживаются. Шасси DWDM потребуется много портов GE, много частот и соединения со многими внешними портами шлюзов FCIP, причем дополнительные расходы будут намного больше затрат добавления в шасси DWDM родной карты FC.

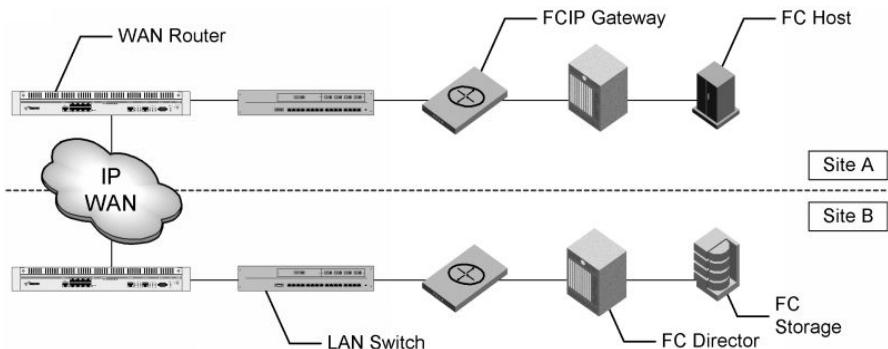


Рис. 13 – Пример физической топологии FCIP

FCIP – это общепринятый стандартный протокол для расширения SAN в случаях, когда имеется IP- сеть, но *нет возможности* применить более надежные и скоростные технологии, такие, как ATM, SONET/SDH, темное волокно и WDM. Как и большинство вендоров инфраструктуры SAN, Brocade предлагает продукты FCIP и разработанные совместно со своими партнерами различные решения FCIP. Однако FCIP менее надежна и работает медленнее других технологий увеличения дистанции подключения SAN, поэтому во многих случаях эту технологию нельзя применять и ее нельзя рассматривать как универсальное. В книге *Multiprotocol Routing for SANs* более подробно описаны технология FCIP и решения на ее основе от Brocade.

# 2

## **2: Решения SAN**

В этой главе представлены несколько наиболее популярных решений, для внедрения которых используется SAN. При знакомстве с их описанием следует учитывать, что это только примеры возможных решений SAN и стратегические инвестиции в сетевую инфраструктуру открывают возможности для новых решений на базе SAN. Даже если SAN была построена только в расчете на одно конкретное приложение, то обычно впоследствии и другие приложения начинают использовать возможности SAN.

### **Консолидация хранения**

Консолидация хранения – это процесс объединения различных распределенных ресурсов в несколько централизованных ресурсов. Этот подход улучшает текущее управление и дает прямую экономию расходов. Основной эффект от консолидации хранения на уровне «железа» связан с более эффективным использованием ресурсов хранения - в консолидированной среде меньше незадействованного дискового пространства и поэтому требуется меньше дисковых массивов, которые нужно приобрести и обслуживать. Консолидация хранения также улучшает эффективность работы администраторов.

В среде DAS<sup>33</sup> у каждого хоста должна быть своя система хранения. Даже это не встроенная, а внешняя система, другие хосты не смогут ее использовать и ее нельзя размещать на расстоянии от хоста (см. Рис. 14).

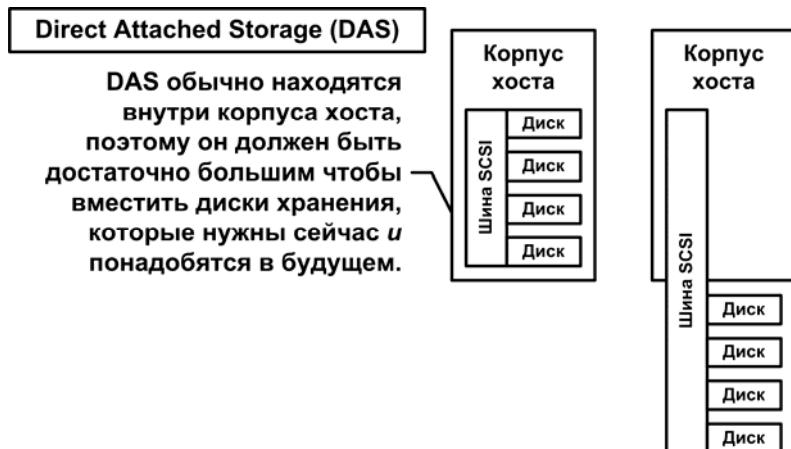


Рис. 14 – Архитектура DAS – до консолидации хранения

Из-за трудности точного прогнозирования роста потребности в емкости в будущем каждое устройство хранения DAS должно иметь существенный запас неиспользуемой емкости (так называемое “white space”). Обычно крайне нежелательно часть отключать приложения, однако этого нельзя избежать при использовании DAS поскольку новые массивы хранения нельзя подключать в онлайновом режиме. На Рис. 15 показаны несколько подсистем DAS, у каждой из которых есть своя собственная незадействованная емкость white space и разные уровни эффективности использования.

Как видно из схемы, у каждого хоста есть выделенная подсистема хранения (частично заполненный цилиндр). Уровень заполненности

<sup>33</sup>

Directly Attached Storage

соответствует эффективности использования подсистемы хранения. Стоит отметить, что суммарная незадействованная емкость (среднее значение для всех хостов) примерно равна использованной емкости. В этом примере усредненный коэффициент использования ресурсов хранения равен 50%, т.е. половина инвестиций в системы DAS не дают никакой отдачи: это white space только занимает место в центре обработки данных, потребляет электроэнергию и выделяет тепло, создавая дополнительные расходы на охлаждение и постепенно выходит из строя<sup>34</sup>.

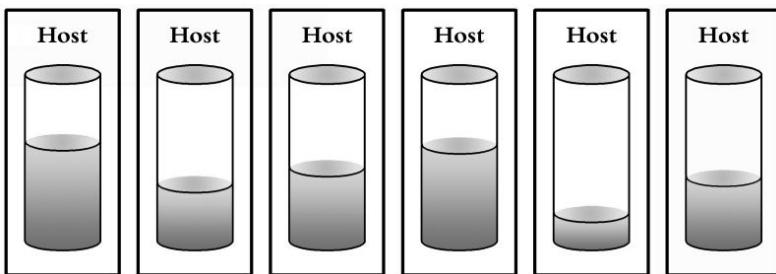


Рис. 15 - White Space в подсистемах DAS

Основная причина такого большого white space в DAS – это невозможность предоставить хосту, которому не хватает дискового пространства выделенной ему подсистемы хранения, часть емкости системы хранения другого хоста, поэтому каждому хосту нужен свой пул незадействованной емкости, размер которого

<sup>34</sup>

Хотя коэффициент использования 50% может показаться заниженным, в подготовленном Merrill Lynch и McKinsey в 2001 году отчете “Storage Report” говорится, что в среднем в центре обработки данных 70% емкости – это white space, т.е. та емкость, от инвестиций в которой нет отдачи. В то же время согласно этому же отчету коэффициент использования в SAN составляет от 80% до 90%: “В крупных компаниях это экономит 40-66% стоимости подсистем хранения”. Эти оценки демонстрируют существенную и быструю окупаемость инвестиций в построение SAN.

определяется исходя из самого неблагоприятного сценария. Вероятность того, что хосту действительно потребуется эта емкость и он без нее не сможет продолжать работу, очень мала, но такую ситуацию нельзя полностью исключить. Кроме того, устройства хранения можно заказывать только с определенной гранулярностью емкости, и если компания использует 100-гигабайтные диски, а конкретному приложению достаточно всего лишь 1 Гбайт емкости, то у хоста, на котором работает это приложение, white space будут равно 99%, А если соседнему хосту потребуется 101 Гбайт, то придется установить два диска. Таким образом, двум этим хостам будет выделено 300 Гбайт, из которых оба они используют только 102 Гбайт и white space се равно почти 2/3.

При использовании SAN основную часть white space можно сосредоточить в центральном пуле и тогда любой хост, которому не хватает емкости, может использовать свободное дисковое пространство любого устройства хранения. Например, если бы описанные в предыдущем примере два хоста были бы подключены к SAN, то второй хост (к которому нужны 101 Гбайт) мог бы получить недостающий гигабайт от диска первого хоста, который почти не заполнен и обоим хостам хватило бы двух дисков, а коэффициент использования дискового пространства был бы больше 2/3 ( т.е. white space сократилось бы до одной трети). В SAN любой хост, которому срочно потребуется дополнительная емкость, может сразу получить ее из центрального пула. Хотя по-прежнему требуется определенное white space, но его размер будет намного меньше, а это значит, что будет больше отдачи от инвестиций в хранение. На Рис. 16 показана конфигурация с шестью хостами из Рис. 15 после миграции с DAS на SAN.

При переходе на SAN не только повысился коэффициент использования и в результате

снизились расходы, но за счет консолидации также уменьшилось число устройств в хранения (в данном примере вместе шести небольших дисковых подсистем хосты используют один большой массив). Это уменьшает энергопотребление, нагрузку на систему кондиционирования центра обработки данных, затраты на поддержку, износ оборудования и затраты рабочего времени администратора, которому нужно обслуживать меньше систем.

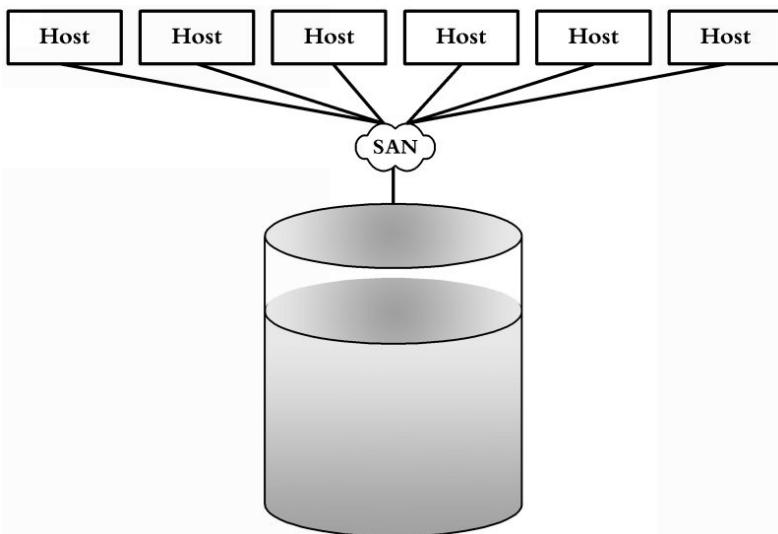


Рис. 16 - White Space после миграции с DAS на SAN

Разумеется, далеко не всегда можно или желательно заменить *все* подсистемы хранения на *одну* большую систему и в корпоративном центре обработки данных уже после консолидации могут использоваться десятки и даже сотни массивов. Однако в любом случае при консолидации уменьшается число подсистем по сравнению с DAS и увеличивается коэффициент использования.

Сети хранения сегодня чаще всего внедряют для консолидация хранения, поскольку она очень выгодна и

эффект от ее внедрения легко выражается в количественных показателях. Архитекторы SAN проектируют консолидацию как интегральную часть большинства решений SAN даже если сеть хранения строится для других задач. Например, если SAN проектируется для кластера высокой доступности (см. следующий раздел), то она может также использоваться для консолидации хранения данных узлов кластера.

## Кластеры высокой доступности

Кластеры высокой доступности High Availability (HA) позволяют выполнять приложение на нескольких серверах (*узлах кластера*) так, что если один сервер кластера выйдет из строя, приложение продолжит работу на другом сервере. Кластеризация стала популярной в связи с глобализацией бизнеса, которая многократно увеличила убытки от перебоев в работе приложений.

Существует несколько способов построения кластеров HA. Например, в одних кластерах приложения постоянно работают на всех узлах и производительность деградирует когда один из узлов выходит из строя поскольку другой узел вынужден обслуживать два приложения. В кластерах другого типа часть узлов постоянно находятся в состоянии активного резерва и становятся активными только когда один из основных узлов выходит из строя. Однако независимо от типа кластеров HA узел сможет «подхватить» приложение с другого узла при отказе последнего только если у него есть в доступе к тому *набору данных*, которых использовал отказавший узел, т.е. требуется общий для всех узлов ресурс хранения, таким образом, для кластера нужна SAN.

До массового внедрения SAN для общего хранения использовались устройства хранения с несколькими

инициаторами, например, массив с дисками SCSI, одновременно поддерживающий соединение с двумя серверами. Пример построения кластера из трех серверов без использования SAN показан на Рис. 17.

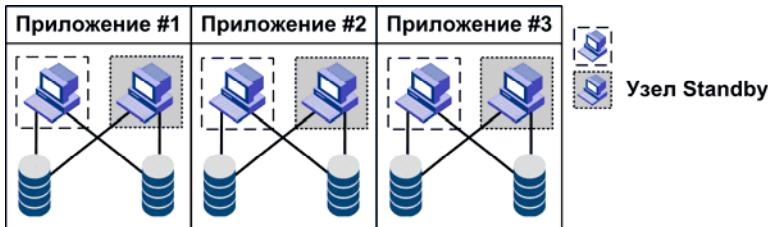


Рис. 17 – Три отдельных кластера

На рынке было доступно сравнительно мало массивов SCSI с несколькими инициаторами по сравнению с общим числом предлагаемых систем хранения, что существенно ограничивало выбор и увеличивало стоимость кластера. Кроме того, большинство этих продуктов поддерживали только два узла, поэтому нельзя было построить большой кластер НА и гибко менять конфигурацию кластера. Из-за этих ограничений кластеры обычно внедрялись под одно конкретное приложение. На Рис. 18 показаны три приложения, каждое со своей выделенной системой хранения и двумя хостами – основным сервером, на котором приложение работает при нормальных условиях, и резервным сервером, на который переходит приложение при сбое основного сервера. Все для этих трех приложений требуется шесть хостов, что означает высокую стоимость, сложность администрирования и отсутствие гибкости, поэтому для многих ИТ-отделов эффект от НА-кластера не оправдывал расходы на его внедрение и обслуживание. SAN полностью изменили эту ситуацию и применение кластеров на основе SAN сократили стоимость, сложность внедрения и текущее администрирование, поэтому для многих заказчиков кластеры НА стали доступным решением. Теперь можно

сконфигурировать кластер так, чтобы любой узел мог «подхватывать» нагрузку с любого другого узла. Те же три приложения теперь можно защитить с помощью одного резервного сервера – этот подход похож на RAID 5, поскольку если выйдет из строя только один узел, то это не приведет к падению производительности.

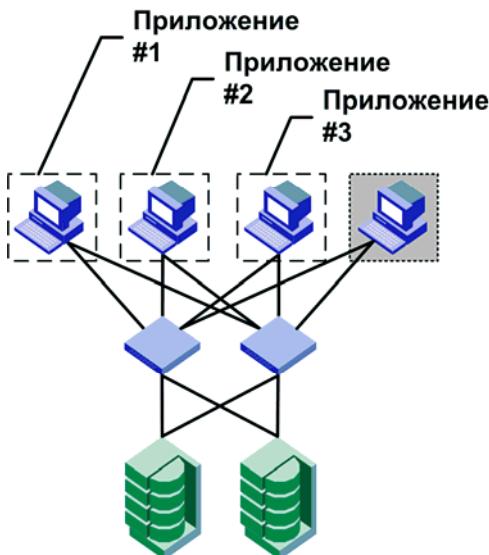


Рис. 18 – Кластер на основе SAN

Как видно из Рис. 18, шесть подсистем хранения из Рис. 17 заменены на два зеркальированных между собой дисковых массива, что уменьшает процент неиспользованного пространства и упрощает управление за счет консолидации хранения и сокращения числа обслуживаемых устройств. Даже с учетом коммутаторов, которые добавлены в конфигурацию, в кластере на основе SAN существенно меньше устройств, что означает меньше затрат на приобретение, конфигурирование и текущее обслуживание. В больших инсталляциях кластеры на основе SAN можно для дополнительной гибкости комбинировать с загрузкой через SAN (см. “Глава 3: UC и ILM“, стр. 85.)

В некоторых случаях кластеры могут применяться и не только для получения высокой доступности и в следующем разделе объясняется, как, кластер с хранением данных SAN может использоваться для увеличения процессорной мощности.



### Заметки на полях

*Инфраструктура SAN – это стратегические инвестиции, но часто SAN дают эффект и в краткосрочной перспективе. Серверы могут исчерпать все ресурсы хранения, которые им выделены для обслуживания критичных приложений, и при этом не выдать никакого предупреждения об этом либо предупреждение затеряется в потоке других сообщений, а администраторы не успеют установить новую систему хранения и мигрировать на нее данные до того, как бизнес-приложения остановятся из-за отсутствия дополнительной емкости. SAN позволяет быстро и эффективно удовлетворять краткосрочные потребности в ресурсах хранения, используя диски с любой системы, подключенной к SAN и обычно этот процесс происходит в онлайновом режиме. SAN позволяет менять конфигурации с помощью нескольких операций, что устраняет целый класс повторяющихся проблем. Эти тактические и логистические преимущества SAN нужно учитывать при рассмотрении возможностей описанных в этой книге решений.*

## **Параллельные и последовательные вычисления**

SAN помогает внедрить различные решения для параллельной и последовательной обработки данных

решений, начиная от систем уровня рабочей группы и до суперкомпьютеров.

Например, небольшой студии, выполняющей редактирование цифрового видео, может потребоваться обрабатывать одну и ту же видеозапись параллельно на разных рабочих станциях, на каждой из которых выполняется свое программное обеспечение и работают пользователи с разным уровнем квалификации, либо на рабочих станциях на видео накладываются различные фильтры для получения спецэффектов. Ясно, что для этого необходимо, чтобы все рабочие станции имели доступ к одним и тем же файлам с видео. Хотя файлы между рабочими станциями можно передавать с помощью FTP или даже электронной почты, однако на практике это не имеет смысла из-за размеров файлов с видеозаписями. Можно реализовать совместный доступ к файлам через сеть с помощью протоколов файлового уровня NFS или CIFS, но они не способны обеспечить тот уровень производительности, который необходим для редактирования цифрового видео и требуют существенных затрат. SAN можно сконфигурировать с общими файловыми системами, что позволит нескольким рабочим станциям одновременно обращаться к файлам без выполнения неэффективного преобразования протоколов или использования медленных методов организации совместного использования файлов.

На Рис. 19 представлен один из возможных подходов к применению SAN для редактирования цифрового видео.

В этом примере восемь узлов (рабочих станций) последовательно обрабатывают картинку, например, первый узел может преобразовывать файлы с необработанным цифровым видео в последовательность файлов с необработанными изображениями и

соответствующие звуковые файлы, второй с помощью фильтров уменьшает зернистость или выполняет балансировку цветов, третий устранит шумы в звуковых файлах или пустые картинки. Хотя в этом примере каждый кадр последовательно обрабатывается разными вычислительными узлами, все они работают параллельно.

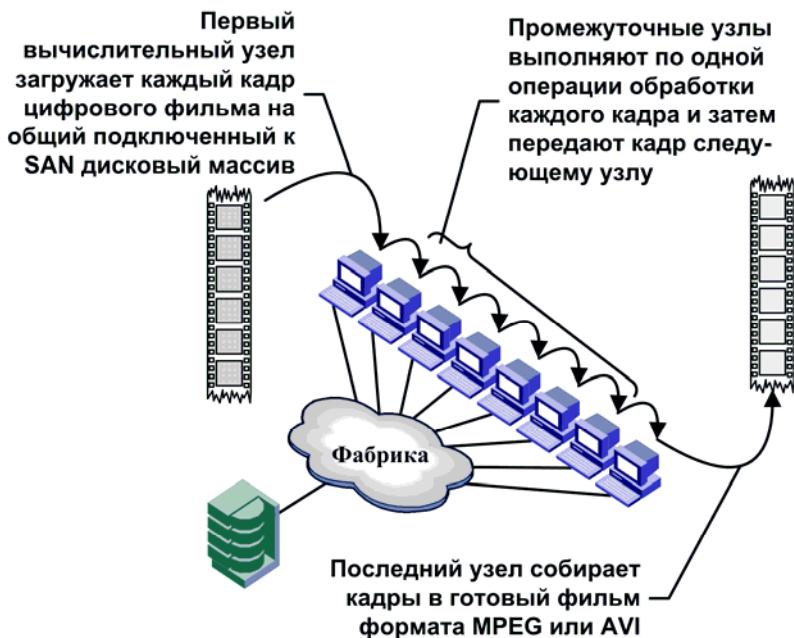


Рис. 19 – Конвейер для редактирования видео

Аналогичный подход используется не только при выполнении связанной с интенсивными вычислениями кадров цифрового видео для рекламы, фильмов и телевизионного шоу, но и для приложений data mining, научных исследований и систем принятия решений в бизнесе. Во многих случаях применяемое решение основано на фирменных механизмах балансировки нагрузки.

Как и для кластеров НА, для кластеров параллельных вычислений очень эффективно консолидация дисков. Разумеется, консолидация хранения возможна не только для дисков и RAID-массивов – консолидация ленточных накопителей также дает существенную экономию в долговременной перспективе.

## **Консолидация ленточных накопителей / резервное копирование без использования LAN**

Системные администраторы хорошо понимают выгоду централизации ресурсов и, как мы уже говорили, централизация дисков улучшает эффективность их использования и упрощает управление. Если говорить о резервном копировании, то иметь на каждом хосте выделенное ленточное устройство неэффективно с точки зрения коэффициента использования емкости лент, рабочего времени администраторов, которое уходит на каждодневную ротацию лент, периодические затраты на хранение лент на удаленной площадке, износа оборудования и простоями приложений во время модернизации оборудования и программного обеспечения резервного копирования. Естественно, администраторы были заинтересованы в решении, которое бы позволило производить централизованное резервное копирование данных. В начале 1990-х годов это можно было реализовать только с помощью локальной сети LAN (см. Рис. 20).

К сожалению, такой эффект не может сильно упростить работу администратора, но в то же время во многих случаях создает проблемы с производительностью. Хотя он позволяет в определенной степени обеспечить консолидацию, его применение снижает скорость резервного копирования и продукционных приложений, данные которых и должно

защитить резервное копирование, а также ведет к падению производительности всей локальной сети. Надежность IP-сетей недостаточна для резервного копирования и потеря пакета может привести к аварийному завершению всей операции резервного копирования и поскольку многие (если не все) пакеты резервного копирования неспособны возобновить свою работу с той точки, где произошло прерывание, то резервное копирование в такой ситуации надо выполнять заново. Локальные IP-сети также неспособны обеспечить полную загрузку ленточного накопителя при одновременном поступлении данных от нескольких источников – дело в том, что для архитектуры коммутаторов IP и стандартных сетей характерна переподписка (over-subscription), которая хорошо подходит для локальных сетей с нестабильным трафиком, но не позволяет обеспечить стабильную максимальную скорость в течение тех нескольких часов, когда выполняется резервное копирование. Таким образом, этот подход не дает оптимального решения проблемы резервного копирования, но создает много дополнительных проблем.

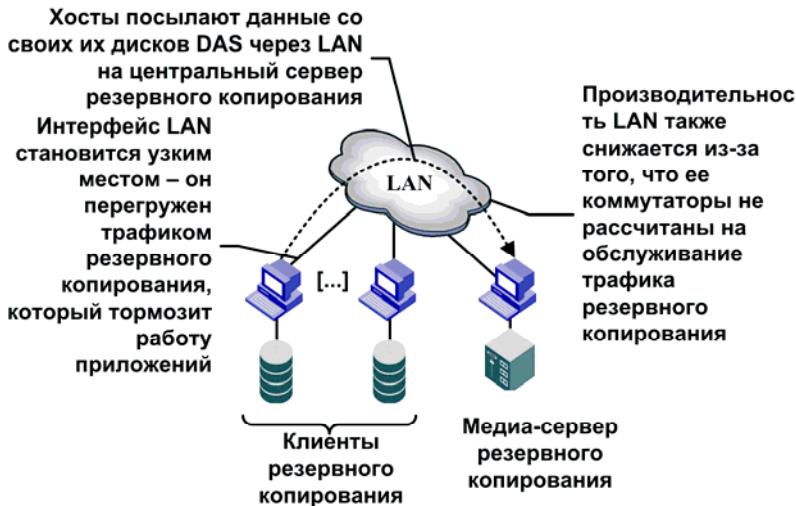


Рис. 20 – Резервное копирование через LAN

Несмотря на свои недостатки резервное копирование по локальной сети было одно время достаточно популярно, однако из-за новых тенденций большинство предприятий перестали применять этот подход. Дело в том, что экспоненциальный рост объемов данных привел к резкому увеличению необходимой для резервного копирования полосы пропускания и времени, в течение которого LAN нельзя использовать из-за резервного копирования, и в то же время нагрузка на локальную сеть, которую создают *другие* приложения, постоянно увеличивается. Поскольку обработка стека TCP/IP сильно загружает центральный процессор и оперативную память, то резервное копирование через сеть занимало часть процессорной мощности и ОЗУ хостов, хотя эти ресурсы нужны для обслуживания приложений, требования к производительности которых постоянно растут. Наконец, поскольку невозможно выполнять резервное копирование на уровне блоков данных между хостами и центральным сервером резервного копирования, поэтому на хостах и сервере нужно запускать фирменный протокол поверх TCP/IP, что

означает «привязывание пользователей» к закрытым решениям и появление новых проблем производительности сети и хостов.

Кроме того, окна резервного копирования (время, в течение которого нужно завершить резервное копирование) постоянно сокращаются поскольку из-за глобализации во многих компаниях бизнес-приложения работают в режиме 7x24<sup>35</sup>. Поскольку требования к производительности резервного копирования и продукционных приложения растут быстрее, чем выполняется модернизация LAN, то очевидно, что резервное копирование через локальную сеть ведет к *увеличению* окон резервного копирования.

Часть этих проблем можно решить с помощью создания выделенной для резервного копирования локальной сети IP/Ethernet LAN и эти сети стали первой реализацией сетей хранения. Однако как Ethernet, так и IP слишком ненадежны с точки зрения хранения и потеря одного пакета приводит к аварийному завершению резервного копирования, а обработка стеков IP уменьшает процессорные ресурсы, доступные для продукционных приложений. Ситуация усложняется из-за того, что многие поставщики оборудования Ethernet плохо понимают особенности технологий хранения и рекомендуют использовать VLAN для отделения трафика резервного копирования от трафика LAN вместо физически изолированных между собой сетей. Такой подход не решает проблем резервного копирования по локальной сети, поскольку оба типа трафика передаются через неприспособленные для трафика резервного копирования сетевые карты NIC и коммутаторы LAN. Так как использующие этот метод

---

<sup>35</sup>

Семь дней в неделю и 24 часа в сутки.

приложения не могут работать на уровне блоков, то из-за необходимости обрабатывать на процессорах хоста закрытый протокол резервного копирования верхнего уровня еще больше снижается производительность и сокращается гибкость.

Окончательным решением этой проблемы является перемещение трафика резервного копирования из IP LAN в FC SAN. На Рис. 21 видно насколько этот подход эффективнее, чем показанный на Рис. 20. Он полностью устраняет недостатки ввода/вывода резервного копирования через LAN. Хотя физически выделенные LAN для резервного копирования можно считать первыми сетями хранения, первое по-настоящему *работающее решение* удалось получить только с помощью Fibre Channel. В отличие от IP и Ethernet, FC гарантирует быструю доставку с сохранением порядка пакетов и низкой вероятностью ошибок, поэтому резервное копирование не нужно запускать заново из-за потери пакета. Кроме того, обработка протокола FC создает минимальную нагрузку на ресурсы хоста, поскольку этот протокол разрабатывался как «легкий» (light weight). Основную обработку этого протокола берут на себя адAPTERы FC Host Bus Adapter (HBA). Наконец, FC работал в 10 раз быстрее, чем Fast Ethernet, который был стандартом для LAN в то время и освобождал центральные процессоры от обработки протоколов, поэтому в результате многие сервера получили большую полосу пропускания<sup>36</sup>.

---

<sup>36</sup>

Ethernet вскоре «догнал» FC за счет использования нижних уровней Fibre Channel (FC-0 и FC-1) и применения более высоких уровней Ethernet (802.2 LLC и 802.3 CSMA/CD). Этот стандарт получил название “Gigabit Ethernet” (GE). Однако практически сразу же FC удвоил свою скорость до 2Gbit и через несколько лет снова удвоил до скорость уже до 4Gbit в то время как GE по-прежнему оставался на 1Gbit. Хосты с адAPTERами GE вынуждены тратить столько процессорных ресурсов на обработку протокола, что не могли полностью загрузить даже сеть 1Gbit, а

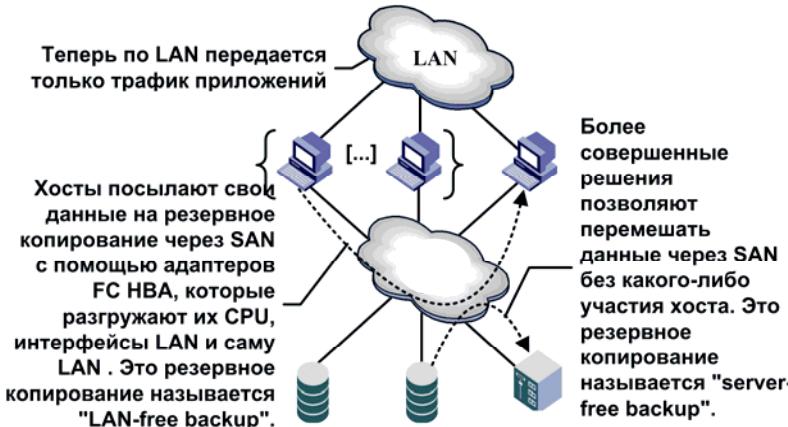


Рис. 21 – Резервное копирование без использования LAN

На самом деле, переход с IP/Ethernet на Fibre Channel давал так много преимуществ, что многие компании построили у себя SAN только для улучшения работы приложений.

## Улучшение производительности

Хотя и другие технологии используются для построения сетей хранения, в подавляющем большинстве случаев архитекторы SAN выбирают Fibre Channel, поскольку она обладает целым рядом преимуществ, в том числе с точки зрения производительности. Сейчас почти все выпускаемые устройства FC поддерживают 2Gbit и уже появились продукты с поддержкой 4Gbit и 8Gbit. Поскольку НВА-адAPTERЫ Fibre Channel берут на себя часть нагрузки с

аппаратное ускорение FC в НВА позволило приложениям работать быстрее. В результате вендорам серверов и систем хранения уже стало не хватать 2Gbit и индустрия FC перешла на интерфейс 4Gbit, который стоил за один порт примерно столько же, сколько 2Gbit. Brocade также поставляет лезвия 10Gbit для директоров (используется для связи на большие расстояния) и в будущем планирует выпуск продуктов 8Gbit.

центрального процессора, то в отличие от решений IP/Ethernet у многих хостов улучшается производительность. Однако оптимизация производительности подсистем хранения во многих случаях дает еще и ряд дополнительных преимуществ, например, когда не удается выполнить резервное копирование в отведенное для этой операции окно или база данных On-Line Transaction Processing (OLTP) не успевает обрабатывать запросы или вычислительные приложения работают слишком медленно. Часто проблемы с низкой производительностью систем хранения или процессорных мощностей удается решить с помощью переходом от DAS к Fibre Channel SAN.

В определенных случаях может отсутствовать проблема с производительностью, но всё равно желательно добиться ее повышения. Например, если *пока* производительность удовлетворяет потребности приложений, но в будущем создаваемая ими нагрузка может увеличиться, либо для бизнеса крайне важна высокая производительность приложения. В подобных случаях имеет смысл заранее построить SAN.

Высокая производительность и надежность Fibre Channel очень эффективны при синхронной и асинхронной репликации данных и резервного копирования/восстановления данных, поскольку для этих приложений требуется большая полоса пропускания в течение длительного времени. Кроме того, часто в этих приложениях используются территориально-распределенные конфигурации, которые не может поддерживать DAS.



## Заметки на полях

Все вендоры SAN утверждают, что их решения обеспечивают более высокую производительность по сравнению с DAS, однако при неправильном внедрении технологий SAN производительность может оказаться меньше, чем у DAS.

Например, если iSCSI работает на хосте используя программный стек протоколов (типичная ситуация для iSCSI), то он отбирает часто процессорных ресурсов хоста у других приложений. Это одна из основных проблем, из-за которой практически вся индустрия отказалась от передачи трафика резервного копирования по LAN. Решения SCSI DAS поддерживают скорости, намного превышающие максимальную для iSCSI скорость 100MB/sec и не создают дополнительную нагрузку, связанную с обработкой стека iSCSI. (см. Рис. 10 и Рис. 11 , стр. 52.) В этой ситуации DAS работает намного быстрее iSCSI (обычно на два или три порядка) и стоит существенно дешевле. В то же время Fibre Channel SAN практически всегда работают быстрее DAS.

## Восстановление после аварий/ Непрерывность бизнеса

После недавних глобальных потрясений многие корпорации и государственные организации стали внедрять решения для восстановления после аварий (disaster recovery, DR) и обеспечения непрерывности бизнеса (business continuity, BC). ( Решения BC также называются “Business Continuity and Availability” (обеспечения непрерывности и доступности бизнеса, BC&A.) В одних компания решения DR и BC внедряются чтобы повысить привлекательность

предприятия для инвесторов, в других использование этих решений диктуется требованиями государственных нормативов и правил. В любом случае эти решения должны обеспечить быстрое и надежное постоянное перемещение больших блоков данных на большие расстояния. SAN идеально подходят для организаций, которым нужно внедрить решения DR и BC и существуют различные опции для построения территориально-распределенных SAN для поддержки таких решений.

Например, каналы Fibre Channel можно удлинить на сотни километров, используя специальные SFP с длинноволновым лазером и одномодовым оптическим кабелем либо с помощью мультиплексоров с функцией повторителей сигнала, например, DWDM. Кроме того, трафик FC может передаваться с помощью таких надежных протоколов WAN, как SONET/SDH и ATM. Как и Fibre Channel, в эти протоколы уже при их разработке были заложены функции высокой производительности и доступности для стабильности передачи данных<sup>37</sup>.

На Рис. 22 показан пример распределенной SAN для обеспечения Business Continuance.

---

<sup>37</sup>

то есть постоянно обеспечивают высокую скорость передачи данных.

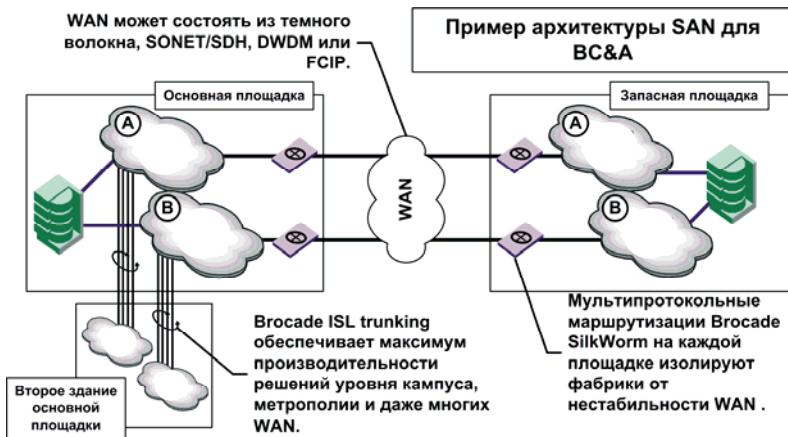


Рис. 22 - Пример архитектуры Business Continuity SAN

В этом примере штаб-квартира корпорации занимает два здания кампуса, каждое из которых имеет резервированную фабрику А/В. ( она обсуждается в “Глава 9: Планирование доступности” на стр. 296.) Из штаб квартиры данные нужно реплицировать на удаленный резервный ЦОД (business continuity site) . Сначала для соединения зданий штаб-квартиры в пределах кампуса они соединяются транками Brocad e ISL. ( описание транкинга дается в “Главе 8: Планирование производительности”, стр. 227). Эти две площадки должны быть территориально разнесены, чтобы катастрофа в районе, где расположена штаб-квартира, не затронула резервный ЦОД. Корпорация выбирает технологию для передачи трафика FC в зависимости от расстояний, выделенного бюджета и требований к производительности. В большинстве случаев используется DW DM, C WDM, SONET/SDH, ATM или простое темное волокно.

## Миграция данных

Во многих организациях за миграцию данных отвечает выделенный специалист. Миграция данных и

серверов может потребоваться в связи со следующими задачами бизнеса:

- Срок аренды дисковых массивов заканчивается или периодически нужно проводить их модернизацию
- Серверы устарели или их переводят на обслуживание других приложений
- Тома полностью заполнены и их надо переместить
- Объединение или перемещение ЦОДов после слияния, поглощения либо внутренней реструктуризации
- С течение времени меняется характер использования приложения

Во всех этих случаях нужно перемещать между массивами большие объемы данных либо изменить распределение систем хранения между хостами. Обычно миграция данных происходит в пределах одного ЦОДа, но иногда данные перемещаются между разными площадками. Например, миграция может выполняться в одном ЦОДе между устройствами, подключенными к одному коммутатору либо между двумя ЦОДами, расстояние между которыми превышает тысячу миль и в последнем случае данные могут проходить через различные промежуточные сетевые устройства и даже сети с разными сетевыми протоколами. В обоих случаях SAN упрощает миграцию.

Если в компании используется только напрямую подключенные к серверам устройства хранения, то для миграции данных между этими устройствами потребует больших затрат рабочего времени и длительного простоя приложений, а если миграция происходит между разными ЦОДами, то нужно будет переместить массивы между площадками либо записать резервные

копии данных на ленту, переместить эти ленты на второй ЦОД и там считать восстановить с ленты данные. Этот процесс крайне трудоемок и сопряжен с существенным риском потери данных.

При использовании SAN затраты и риски миграции данных снижаются и процесс упрощается настолько, что не требует даже краткосрочной остановки приложений<sup>38</sup>.

---

<sup>38</sup>

Например, при миграции данных между массивами можно использовать зеркалирование на основе хостов или сети для получения мгновенной синхронной копии тома и затем отключить основной экземпляр тома без остановки приложений. Однако для других операций остановка приложений является обязательной. Например, при перемещении между хостами *приложения* практически всегда его надо остановить прежде чем запустить на новом хосте. Даже в этом случае SAN упрощает процесс миграции и в результате риск и перебои в работе будут минимальными.

# 3

## 3: UC и ILM

Utility Computing (UC) и Information Lifecycle Management (ILM) крайне важны для архитекторов SAN, проблема, однако, заключается в том, что это – не конкретные продукты, технологии и решения, поэтому их часто рассматривают всего лишь как общие концепции, например, аналитики говорят о “тенденциях внедрения UC и ILM в индустрии”, но при этом не дают четкого определения этим терминам. Действительно, UC и ILM – это тенденции индустрии, однако из констатации этого факта архитектору трудно понять, что обозначает эти два термина и как они могут повлиять на его SAN.

UC и ILM – это набор процессов, которые могут внедряться на основе конкретной архитектуры построения центра обработки данных либо общим подходом к проектированию и управлению. На самом деле определенные архитектуры и подходы настолько помогают внедрению UC и ILM, что к ним можно применить эти термины, т.е. утверждать, что “у нашего ЦОДа архитектура ILM”.

Разумеется, ILM и UC можно внедрить с помощью продуктов и технологий, в том числе и SAN. В свою очередь ILM и UC помогают внедрить те решения SAN, о которых шла речь в предыдущей главе. В этой главе архитекторам SAN более конкретно объясняется, что

такое ILM и UC, как они связаны с SAN и общей архитектурой ЦОДа. В ней рассматриваются преимущества UC и ILM, те проблемы, которые могут помешать в полной мере реализовать их потенциал и текущее состояния технологий для их внедрения, а также рассказывается, как SAN неизбежно помогают использовать решения UC и ILM.

## Классификация UC и ILM

Прежде чем переходить к детальному рассмотрению UC и ILM важно понять, что их нельзя рассматривать как универсальные решения, способные устраниить все проблемы. В этом разделе речь идет о взаимосвязи ILM и UC и об их классификации на основе степени внедрения.

Хотя обе тенденции различаются как в применении на практике, так и по своим результатам, но у них есть и важное сходство. Например, с точки зрения окупаемости инвестиций они обеспечивают более эффективное использование ресурсов оборудования и персонала. Для них также можно разработать поэтапный план внедрения, они реализуются с помощью сетевой инфраструктуры более нижнего уровня и могут обеспечивать разную степень автоматизации.

Последний пункт является ключом к пониманию истинного смысла этих терминов. Как уже говорилось выше, UC и ILM – это процессы. Процессы могут внедряться в разном масштабе и вручную, либо автоматизировано. На самом верхнем уровне среди ILM и UC можно классифицировать с помощью Таблица 1.

Таблица 1 – Классификация среды UC и ILM

Уровень автоматизации процессов	Масштаб внедрения решения
Ручные (“бумажные”) процессы	Прототип (тестовая лаборатория)
Полуавтоматизированные процессы	Ограниченнное внедрение для производственных систем
Полностью автоматизированные процессы	Внедрение по всему предприятию

В первой колонке классификация выполняется на основе степени автоматизации процессов. В среде ILM или UC может полностью отсутствовать автоматизация, и процессы могут быть просто записаны на бумаге и внедряться IT- специалистами. Такой подход применяется сегодня чаще всего, однако вендоры выпускают продукты, которые автоматизируют некоторые процедуры, т.е. среда становится полуавтоматизированной. Однако пока нет среды, которая бы полностью автоматизирована и все попытки добиться полной автоматизации не оправдывают потраченных на них усилий.

Вторая колонка показывает, насколько решение ILM или UC управляет средой предприятия. Конечная цель ILM и UC – внедрение единого решения для управления всем предприятием. Однако как и в случае с автоматизацией, сейчас практически невозможно и не имеет смысла для IT -отдела добиться этого уровня внедрения. Сегодня эти решения сначала внедряются для тестирования и затем для управления ограниченным сегментом производственной среды.

Сейчас лучше всего ориентироваться на вторую строку таблицы. Например, компания может внедрить полуавтоматизированное решения для применения плана восстановления после катастроф, при котором реплицируются и создаются резервные копии только тех данных, которые критичны для бизнеса. Процессы ILM помогут определить, какие данные нуждаются в защите (т.е. провести классификацию данных), и что с ними нужно сделать (например, создать резервные копии и/или реплицировать). Кроме того, в этих процессах могут использоваться как ручные операции, например, пользователи вручную задают серверы, данные с которых нужно классифицировать, так и автоматизированные, например, построение конкретных решений резервного копирования и репликации на основе классификации данных.

## Utility Computing

Utility Computing (UC, ресурсные вычисления) – это концепция, согласно которой такие ресурсы, как процессорная мощность, оперативная память и емкость устройств хранения могут выделяться пользователям аналогично тому, как коммунальные службы обеспечивают потребителей электричеством и водой. В идеальной среде UC пользователь подключает дешевый терминал к любой «розетке» сети и получает доступ к любым объемам компьютерных ресурсов так, как если бы они находились внутри терминала. UC можно рассматривать как виртуализацию компьютерных ресурсов – все ресурсы объединены в виртуальное «облако», из которого любое приложение или пользователь получит нужные ресурсы без каких-либо дополнительных усилий и пользователь должен “платить” только за те мощности, которые он действительно использовал.

Существует три основных стимула для внедрения UC:

1. Необходимость сократить затраты на IT (как на развертывание, так и текущее обслуживание).
2. Необходимость более эффективно развертывать приложения в условиях ограниченной площади ЦОДа.
3. Потребность улучшить производительность приложений и их доступность.

На практике Utility Com putting внедряется не на уровне отдельных процессоров, а на уровне вычислительных узлов, например, блейд-сервера в шасси для лезвий, партиции (раздела) в большом SMP-сервере или обычном автономном сервере. В любом случае сетевая архитектура должна обеспечить для всех клиентов доступ к любому приложению на любом вычислительном узле (см. Рис. 23).

Иногда UC называют grid-вычислениями, поскольку она концептуально похожа на электросеть, к которой можно через электрическую розетку подключить любой электроприбор, например, холодильник<sup>39</sup>. Обычно холодильник работает так, как будто генератор электроэнергии находится внутри него, хотя на самом деле электричество вырабатывается на гидроэлектростанции, от которой его отделяют сотни или тысячи миль.

Utility Computing - это *не* конкретное оборудование или программное обеспечение, однако оборудование или программное обеспечение *помогает внедрить* решения UC. Например, среда UC должна предоставить любому

---

<sup>39</sup> Существуют и другие варианты названия UC, зависящие от того, как авторы хотят позиционировать эти решения, в том числе “адаптивное предприятие” и “автономные вычисления”. Чтобы не возникало путаницы мы будем использовать термин Utility Computing и его сокращение UC.

приложению любые объемы процессорной мощности, оперативной памяти и емкости дисков, что на практике означает, что в обозримом будущем приложения смогут перемещаться между вычислительными узлами, а для этого узлы должны быть соединены через общую LAN так, чтобы пользователи могли обращаться к любому серверу, на котором работает их приложение, и через общую SAN на стороне back end для того, чтобы аппаратная платформа каждого сервера могла обращаться к любому блоку данных в зависимости от того, какое сейчас приложение работает на этом узле.

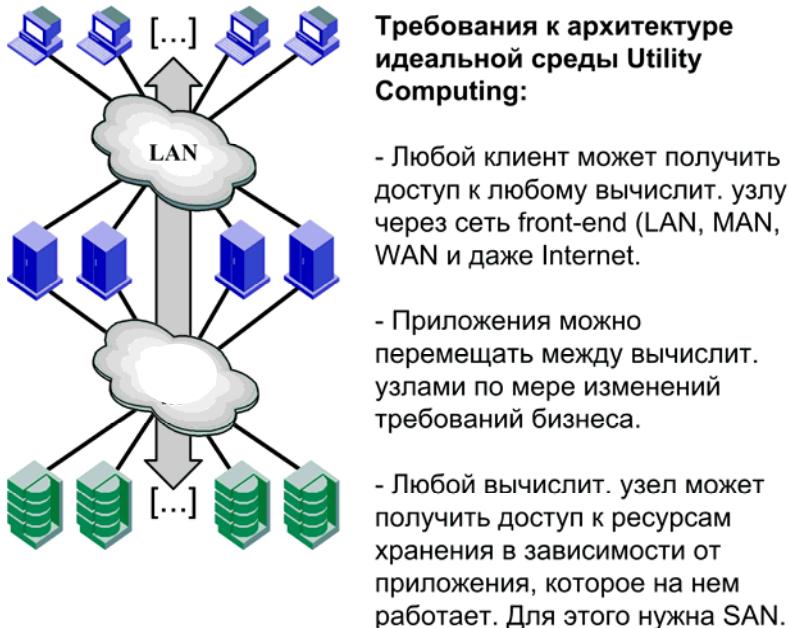


Рис. 23 – Необходимые для Utility Computing подключения

Эта архитектура с двумя сетями является составной частью внедрения UC и часто называется общей архитектурой ЦОДа Utility Computing. На двух сторонах Рис. 24 показаны пять разных уровней идеализированной архитектуры ЦОДа UC и то, как клиенты через эти уровни обращаются к своим данным.

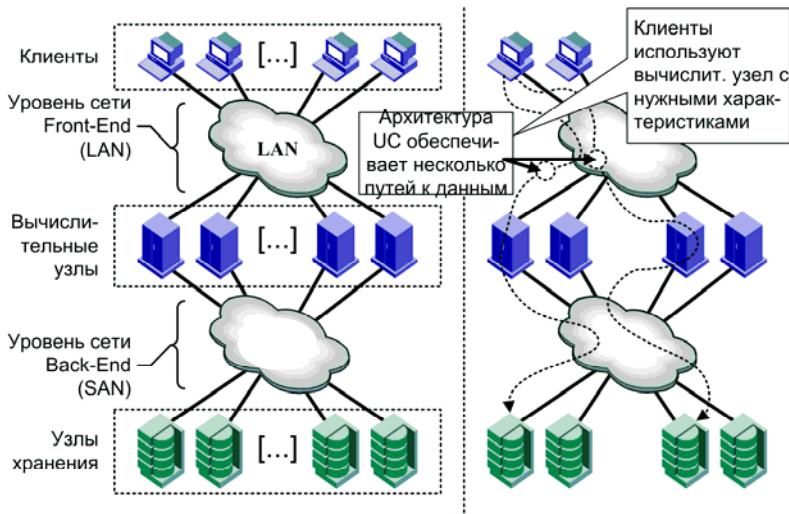


Рис. 24 – Уровни архитектуры ЦОДа UC

Первый уровень этой пятиуровневой архитектуры UC состоит из клиентов, которых нужно обслуживать. На следующем уровне клиенты подключаются к LAN, MAN, WAN или комбинации этих сетей. Этот уровень обеспечивает доступ any-to-any между всеми узлами на уровне клиентов и вычислительными узлами третьего уровня. У вычислительных узлов могут быть разные характеристики, например, мощность процессоров или параметры НА, что делает их подходящими для разных задач, но можно использовать и полностью идентичные вычислительные узлы и распределять между ними клиентов с помощью механизма балансировки нагрузки, тогда узлы, обращающиеся к конкретному набору данных, могут меняться каждый день и даже каждую секунду (такой сценарий используется в решениях с балансировкой нагрузки через сеть, например, в “фермах” web-серверов). Также требуется решение для управления перемещением приложений между вычислительными узлами и гарантирующее, что клиент обращается к тому узлу, на котором работает его

приложение. После того как клиенты получили доступ к одному или нескольким вычислительным узлам с нужными им приложениями и ресурсами они должны попасть на четвертый уровень SAN чтобы получить доступ к пятому уровню, где на устройствах хранения физически размещены их данные. Данные передаются клиентам в обратном порядке – сначала через SAN к приложению, затем через сеть front-end на клиентскую систему.

Кроме пятиуровневой архитектуры для решений UC нужны политики и процедуры, которые определяют, где будет приложение в конкретный момент времени, и решение для перемещения по мере необходимости приложений между физическими вычислительными узлами. Это решение может работать в ручном режиме, быть частично или полностью автоматизированным. В последних двух случаях нужны инструменты для управления.

Легко понять, что для архитектуры UC необходима SAN. Если не было бы SAN, то только один вычислительный узел мог бы обращаться к определенному набору данных и клиенты были бы “привязаны” к каналу доступа к этим данным. Они не могли бы масштабировать или изменять свои вычислительные ресурсы независимо от ресурсов хранения. С помощью SAN компьютерные ресурсы можно распределять для обслуживания данных на основе различных критериев, например, сколько нужно процессоров для обслуживания данных приложений или на основе характеристик доступности или других правил в зависимости от конкретной ситуации.

Независимо от технического подхода к внедрению UC в ЦОДе сам процесс будет состоять из следующих этапов:

1. Обследование среды и определение потребностей UC
2. Определение политик перемещения компьютерных ресурсов
3. Выбор технологий для автоматизации перемещений
4. Разработка процессов перемещения на основе выбранных технологий
5. Внедрение архитектуры ЦОДа UC (Рис. 24)
6. Внедрение технологий для внедрения UC
7. Перенос приложений и ресурсов в систему UC

### ***Преимущества Utility Computing***

В предыдущем разделе описывалась возможная архитектура UC и упоминались несколько преимуществ, которые UC реализует на высоком уровне. Однако плюсы внедрения utility computing в современном предприятии не ограничиваются только оптимизацией использования процессоров. В принципе, внедрение UC может быть *исключительно* выгодно, поэтому на эту модель переходят многие организации.

Применение модели управления UC позволяет системным администраторам практически независимо управлять оборудованием и программным обеспечением. Существуют различные причины для того, чтобы системному администратору потребовалось перемещать приложения между серверами помимо оптимизации использования процессоров.

Например, некоторые ИТ-отделы берут оборудование в аренду и если срок аренды истек, а приложение по-прежнему используется, то его надо перевести на другое оборудование. Обычно для этого надо было установить новый сервер, инсталлировать на нем приложение и протестировать его, запланировать остановку сервиса, переместить данные и затем запустить новый сервер в промышленную эксплуатацию. Такой процесс занимает

достаточно много времени и связан со значительными рисками для бизнеса. В ЦОДе Utility Com putting, теоретически, системный администратор может перемещать приложение одним нажатием клавиши поскольку все исполняемые коды приложений, образы ОС и данные пользователей находятся на подключенных к SAN устройствах хранения. Не нужно перемещать никакие *данные* - мигрируют только *приложения*.

Такой подход позволяет с помощью UC внедрить стратегию управления ресурсами приложений (Application Resource Management, ARM), которую можно рассматривать как частный случай Utility Computing. В большинстве организаций приложения переведены на режим работы 7x24, что неизбежно привело к разрастанию парка оборудования и увеличению затрат на текущее обслуживание. Кроме затрат на приобретение оборудования увеличение числа серверов и систем хранения усложняет управление средой – развертывание, конфигурирование, внесение изменений и мониторинг большого числа этих “контейнеров с приложениями” - это крайне сложный процесс. Технологии ARM обеспечивает более эффективное управление приложениями за счет оптимизации:

- Времени, которое уходит на развертывание, выделение ресурсов и тестирования нового приложений или контейнера с приложением.
- Скоординированного мониторинга и управления ресурсов приложений в крупном ЦОДе.
- Соответствия компьютерных ресурсов меняющимся требованиям приложений.
- Времени восстановления после сбоев.

Эта стратегия еще более важна для тех ЦОДов, которые переходят на архитектуру блейд-серверов. В этом случае главное - не оптимизация процессорных

ресурсов, а упрощение управления (т.е. экономия расходов достигается за счет снижения затрат на управление одним приложением и повышения эффективности других операций текущего обслуживания). Администраторам нужна возможность загрузиться с любого лезвия с любыми “персональными данными” с помощью централизованной консоли управления, что позволит при необходимости перенести любое приложение на любой блейд-сервер. Этот подход упрощает модернизацию оборудования и замену вышедших из строя компонентов.

UC помогает внедрить архитектуру высокой доступности и стратегии обеспечения непрерывности бизнеса. Если выйдет из строя весь ЦОД, то программное обеспечение и процессы, которые использовались при внедрении UC, например, для улучшения эффективности управления, можно применить и для восстановления после аварии. Если можно легко перемещать приложения между хостами одного ЦОДа, то следующим шагом будем перемещение между ЦОДами, которые отстоят друг от друга достаточно далеко и катастрофа, из-за которой выйдет из строя один ЦОД, не затронет второй.

### ***Проблемы внедрения Utility Computing***

Концепция utility computing достаточно проста: сделать все аппаратные ресурсы прозрачно доступными по требованию для всех приложений и пользователей. В теории, процесс внедрения тоже несложен – нужно определить политики выделения компьютерных ресурсов (т.е. определить, как вычислительный узел будет обслуживать конкретное приложение) и определить и применить процедуры на основе этих политик для перемещения приложений по мере необходимости.

Однако, на практике внедрение UC намного сложнее.

Вернемся к аналогии с электрораспределительными сетями. Для человека, который втыкает вилку от холодильника в электрическую розетку, концепция использования этих сетей очень проста – нужно подключить холодильник к электросети, включить его и ждать, пока продукты охладятся. Производство, доставка электроэнергии и выставление счетов за ее потребление производится автоматически и эти процессы выполняются прозрачно для конечного потребителя.

С другой стороны, с точки зрения инженера, проектирующего гидроэлектростанцию, диспетчера электрораспределительной сети или бухгалтера, который налаживает систему выставления счетов за электроэнергию, эти процессы намного сложнее. Массовые отключения электричества в Калифорнии в 2001 году и в северо-восточных штатах США два года спустя продемонстрировали конечным потребителям, насколько сложна система генерации и распределения электроэнергии.

Для реализации концепции UC на практике нужно программное обеспечение и, возможно, оборудование, которые их вендоры должны спроектировать, собрать и протестировать. Чтобы UC функционировала как электрораспределительная сеть требуется напряженная работа компаний, продвигающих технологии для внедрения UC.

Автоматизированная система UC должна перемещать приложения по мере необходимости, для чего она должна быть полностью интегрирована с ними, поскольку у каждого приложения разные требования и процедуры запуска и остановки. Система должна распознавать перемещения приложения или образа ОС и вести учет распределения между ними аппаратных

ресурсов. Она должна предоставить менеджерам систем UC инструменты для определения политик, средства контроля для таких задач, как определение приоритетных приложений, которым в первую очередь выделяются ресурсы, и выставление счетов пользователям за те ресурсы, которые они потребляли. Она должна интегрироваться с сетью fr ont-end для правильного перенаправления сетевого трафика после перемещения приложений, подключаться к SAN для загрузки образов ОС на вычислительные узлы и поддержания их разделов данных.

Таковы задачи, которые сейчас стоят перед разработчиками программного обеспечения и оборудования, пытающимися внедрить UC на практике. В зависимости от применяемого подхода программного обеспечения будущих ЦОДов UC должно инсталлироваться на каждом хосте и, возможно, также на сетевых коммутаторах и маршрутизаторах, т.е. оно должно быть полностью совместимо со всем оборудованием, прикладным программным обеспечением и версиями ОС разных вендоров. А для того, чтобы получить полную отдачу от UC , управляющее программное обеспечение должно иметь высочайшую функциональность и надежность.

Даже если выполнены все перечисленные требования к полноценному решению UC, остается еще целый ряд проблем. Надо убедить конечных пользователей перейти на новую модель работы, установить на системы новое программное обеспечение, заново обучить администраторов и внедрить пакеты мониторинга работы сети. Таким образом, потребуется много времени для внедрения всего решения даже при условии, что имеются все его компоненты.

## *Текущее состояния Utility Computing*

Проблемы настолько сложны, что до сих пор никто из вендоров не смог предложить решение, полностью реализующее концепцию UC, однако, как будет показано в следующем разделе, уже доступны некоторые важные компоненты решения, также как имеется сетевая инфраструктура, необходимая для построения ИТ-среды, в которой можно будет внедрять приложения UC по мере их появления.

Сегодня многие вендоры предлагают под разными торговыми марками решения, которые рекламируются как utility com putting, однако ни одно из них нельзя считать законченным. Эти решения полезны, но часто клиенты не могут определить, что на самом деле предлагает ему вендор сегодня и что обещает реализовать в будущем.

Например, многие производители сегодня выпускают блейд-серверы. Клиенты могут купить устанавливаемое в стойке шасси, обеспечивающее питание, охлаждение и некоторую сетевую инфраструктуру. Все эти ресурсы доступны хостам, т.е. установленным в шасси лезвиям, поэтому хостам требуется меньше кабелей, они занимают меньше места в стойке, потребляют меньше энергии и требуют меньше охлаждения, а также благодаря интеграции в шасси коммутаторов FC и Ethernet оптимизируется развертывание LAN и SAN.

Для Utility Com putting шасси блейд-серверов также должно поддерживать один или несколько интерфейсов управления для того, чтобы контролировать распределение между лезвиями приложений и образов ОС. Эта функция, очевидно, относится к классу UC поскольку упрощает администрирование, реализована во многих шасси. В шасси блейд-серверов также интегрированы фабрики Brocade Fibre Channel, обеспечивающие подключение к сети back-end,

необходимой для UC, существенно сокращая расходы на развертывание и управление. Аналогичный функционал реализован и для сетей front-end. Обе эти функции уменьшают количество кабелей, место в стойке, которое нужно выделить для сетевого оборудования и общие расходы на развертывание и текущее обслуживание.

Таким образом, внедрение шасси блейд-серверов помогает перейти на модель Utility Computing и дает преимущества в краткосрочной перспективе, хотя и не решает *всех* проблем, которые должна устраниТЬ UC. Например, установка одного шасси блейд-серверов не означает превращение всего ЦОД в среду UC и не позволяет перемещать приложения прозрачно и автоматизировать этот процесс, а также точно перераспределять ресурсы. Например, блейд-серверы не могут решить такую задачу, как перенос 1GHz мощности процессора от приложения *x* к приложению *y*. Когда потребитель подключает электроприбор к электрической розетке, он получает столько мощности, сколько необходимо этому прибору, а при использовании блейд-серверов минимальная единица мощности – это энергопотребление одного лезвия. Если бы электрическая сеть работала по тому же принципу, то для любого электроприбора выделялось бы не менее одного киловатта-час.

Очевидно, что блейд-серверы нельзя назвать полным воплощением UC, но сегодня они уже решают реальные проблемы и помогают подготовить ЦОД к внедрению других технологий UC в будущем.

Другим примером технологии UC, которую уже сегодня можно внедрить для решения реальных проблем, являются коммутаторы приложений на базе SAN. В краткосрочной перспективе эти устройства существенно упрощают администрирование загрузочных образов и профилей приложений.

Например, для упрощения процессов Brocade предлагает продукт Application Resource Manager (ARM), обеспечивающий управление и перемещение данных с помощью инфраструктуры Brocade SAN для автоматизации типичных задач выделения ресурсов. Пользуясь им IT-менеджеры могут динамически активизировать серверы, конфигурировать образы своих операционных систем и приложений с помощью программ-мастеров и шаблонов центральной консоли, а также автоматизировать управление сложными взаимозависимостями между оборудованием, программным обеспечением и SAN. Определения серверов можно сконфигурировать только один раз и затем копировать или перемещать их «на лету» когда это необходимо с центральной консоли. Для этого все образы серверных ОС и приложений «живут» в SAN и ARM координирует их соответствие аппаратным платформам.

Индустроля постепенно переходит к этому типу упрощенной архитектуры с загрузкой через FC, которая продвигает блейд-серверы и стандартные серверы ближе к полной реализации UC за счет полной независимости конфигурации отдельных устройств от аппаратуры этих устройств. Если нужно заменить вышедшее из строя лезвие в шасси блейд-серверов или переместить приложение на компьютерный узел с другими характеристиками производительности, эти операции выполняются одним нажатием кнопки на центральной консоли с помощью коммутаторов приложений на базе SAN. Эти устройства решают часть проблем развертывания законченного решения UC, устраняя необходимость в инсталляции специального программного обеспечения на каждом компьютерном узле. Как и блейд-серверы, они не решают *всех* проблем, которые должно решить внедрение UC, но они обеспечивают эволюцию ЦОДа в правильном

направлении и дают эффект в краткосрочной перспективе.

## Управление жизненным циклом информации

Так же как в случае с Utility Computing, существует несколько определений ILM<sup>40</sup>, которые разные компании из индустрии SAN используют в зависимости от того, что конкретная компания пытается продать. В этой книге используется официальное определение, предложенное независимой ассоциацией Storage Networking Industry Association (SNIA), которое не «привязано» к решениям одного вендора. Это определение гласит “ILM – это политики, процессы, практики и инструменты, используемые для того, чтобы добиться соответствия между ценностью информации бизнеса и эффективности по затратам IT-инфраструктуры начиная от момента генерации информации и до ее окончательного уничтожения”. Таким образом, ILM помогает организациям определить, где хранится информация в конкретный момент времени.

Есть два основных стимула внедрения ILM:

1. Необходимость сократить затраты на IT, включая как первоначальные инвестиции на внедрение, так и текущие расходы на обслуживание.
2. Необходимость в выполнении меняющихся требований законодательства и внедрения best-practices неизменяемого сохранения данных.

---

<sup>40</sup>

ILM иногда называть и «Управлением жизненным циклом данных» (Data Lifecycle Management, DLM) чтобы подчеркнуть разницу между информацией (реальной intelligence, которая хранится и передается) и данными (двоичным представлением этой информации). Однако в этой книге используется термин ILM, который чаще всего используется в индустрии хранения.

Сетевая архитектура ILM должна обеспечить для любого вычислительного узла доступ к данным на любом устройстве хранения, поскольку данные этого узла могут в любой момент времени перемещаться между устройствами хранения (см. Рис. 25).

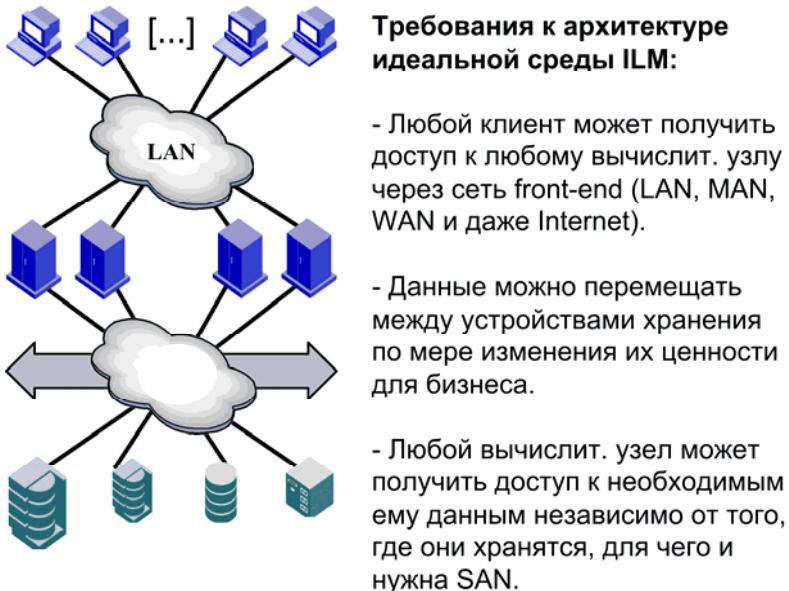


Рис. 25 – Требования ILM к соединениям

Информация на устройствах хранения имеет много характеристика, в том числе права доступа, производительность и доступность, история изменений, текущий размер и прогнозируемый рост размера. Многие из этих характеристик меняются с течением времени. ILM можно рассматривать как набор методов корректировки этих характеристик хранимых данных по мере изменения их ценности в течение времени, начиная с момента их первоначальной генерации и до того момента, когда они больше не нужны и поэтому удаляются. Применение этих методов позволяет менеджерам ЦОДов экономить за счет использования дисковых массивов корпоративного класса только для

хранения важных данных и по мере их постепенного обесценивания переводить их на более дешевые системы хранения.

Для этого требуется ЦОД с архитектурой, аналогичной ЦОДу Utility Computing. Клиентам нужно через сеть front-end обращаться к вычислительным ресурсам, как это и было раньше, но теперь перемещаются не приложения между вычислительными узлами, а данные этих приложений через сеть back-end, т.е. все приложения на всех вычислительных узлах должны иметь доступ ко всем устройствам хранения (хотя бы потенциальный), поскольку их данные могут переместиться на другое устройство хранения по мере изменения их ценности с течением времени. На Рис. 26 показана пятиуровневая архитектура ЦОДа ILM вместе с процессом миграции данных между разными носителями в течение их жизненного цикла.

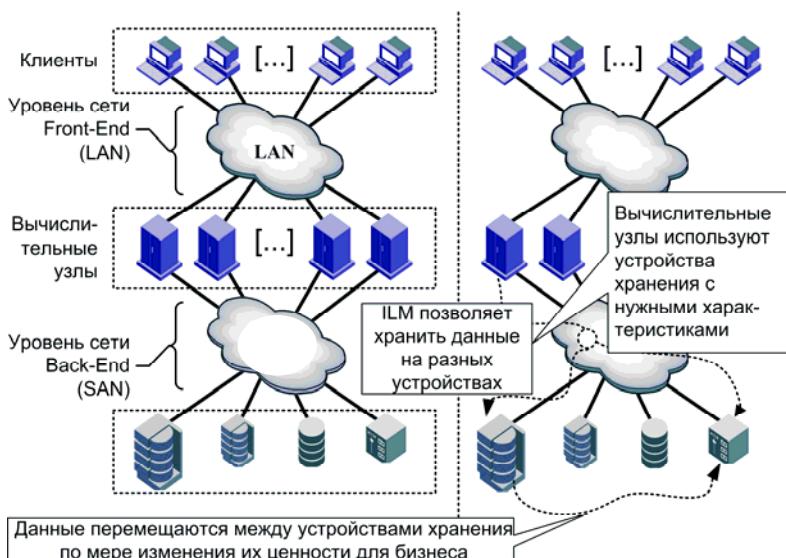


Рис. 26 – Архитектура ЦОДа ILM

Клиенты подключены к традиционной сети, обеспечивающей доступ между всеми клиентами на первом уровне и всеми вычислительными узлами третьего уровня (any-to-any ). Вычислительные узлы могут обращаться к своим данным через сеть back-end (т.е. SAN), которые располагаются на разных классах устройств хранения. Каждое устройство хранения может иметь разные характеристики, например, НА или производительность. В течение своего жизненного цикла данные могут храниться на разных устройствах. Новые данные имеют высокую ценность для бизнеса и их лучше хранить на RAID-массиве корпоративного класса. Когда они устаревают, но имеют определенную ценность для бизнеса, их можно переместить на более дешевые RAID- массивы или системы JBOD. К концу жизненного цикла они могут быть перемещены на ленту для долговременного хранения в архиве. Для реализации этого подхода нужно программное обеспечение, которое перемещает данные между уровнями хранения и гарантирует, что вычислительные узлы получат доступ к своим данным на правильном запоминающем устройстве.



## Заметки на полях

Сейчас *ILM* шумно рекламируется, но сама эта концепция не нова и можно сказать, что сегодня любая *IT*-среда в той или иной форме использует *ILM*. Например, если пользователь удаляет на своем ПК временные файлы, то эта процедура - процесс *ILM*. Пользователь изучает свои файлы и решает, какие временные файлы уже не нужны (хранятся большие определенного времени), ищет их и уничтожает чтобы освободить место на диске. Если бы он вместо уничтожения перемещал устаревшие ненужные файлы на *CD-R* или архивировал на ленту, то и эти процедуры можно было бы рассматривать как процессы *ILM*, которые многие *IT*-специалисты периодически используют в своей работе. Нет принципиальных отличий от внедрения *ILM* в крупном ЦОДе от этих широко используемых процессов, просто решение *ILM* более сложно и носит всеобъемлющий характер.

Кроме уровней сетевой архитектуры, решениям *ILM* нужны политики и процедуры чтобы определить, где данные должны храниться на определенном этапе их жизненного цикла, и технологические решения для перемещения данных между физическими устройствами хранения по мере изменения свойств данных. Такое технологическое решение может реализовываться только вручную либо частично, либо полностью автоматизировано. В последнем случае требуется набор средств управления.

Теперь становится ясно, почему SAN необходима для решений *ILM*. Без SAN каждый вычислительный узел имеет доступ к конкретному набору данных только до тех пор, пока он напрямую подключен к нужному дисковому массиву, что серьезно затрудняет

обеспечение хранения данных в соответствии с их ценностью. При использовании SAN данные приложений на вычислительных узлах могут располагаться на любом доступном устройстве хранения в зависимости от различных критериев, например, наличия на устройстве свободной емкости или таких характеристик доступности, как защита с помощью RAID либо любой другой используемой политики ILM.

Независимо от технологических решений, применяемых для внедрения ILM, скорей всего внедрение будет состоять из следующих шагов:

1. Обследование среды для определения потребностей ILM
2. Определение политик перемещения данных
3. Выбор технологий для автоматизации перемещения данных
4. Разработка конкретных процессов для внедрения этих технологий перемещения данных
5. Внедрение архитектуры ЦОДа ILM (Рис. 26)
6. Внедрение этих технологий
7. Перенос наборов данных в систему ILM

## *Преимущества ILM*

В предыдущем разделе объяснялось, как ILM помогает сэкономить деньги за счет оптимизации выбора устройства хранения в зависимости от таких характеристик данных, как ценность для бизнеса всей компании. Дорогие дисковые массивы хранения с самыми высокими показателями производительности, надежности и доступности используются для хранения самых ценных данных, а остальные данные хранятся на более дешевых системах. Процессы и инструменты ILM помогают организациям добиться соответствия характеристик устройств хранения и их стоимости требованиям к конкретному набору данных.

Например, когда были созданы файлы с текстом этой книги, они были очень важны для проекта и если бы сломалась система, на которой хранились эти файлы, то не удалось бы выпустить книгу к назначенному сроку. Из-за такой важности данных на этой стадии их жизненного цикла они хранились в нескольких местах, причем сохранялись и предыдущие версии каждого файла. Для хранения использовались системы высокой доступности и каждый день выполнялось резервное копирование данных. После публикации книги ее исходные файлы стали некритичны, поэтому предыдущие версии файлов как и дубликаты окончательной версии файлов были удалены. Сохраняется только одна копия окончательной версии, причем для этого используется недорогая система хранения. Разумеется, на случай потери этой копии еще несколько копий сохраняется в архиве на ленте.

Однако на этом не заканчиваются преимущества ILM. Экономия на стоимости хранения – это всего лишь одна из причин, по которым ИТ-отделы должны перемещать данные между разными устройствами. Применяемые для внедрения ILM технологии должны обеспечить механизм на основе правил для перемещения данных, а также зеркалирования, репликации и миграции данных.

Например, многие ИТ-отделы берут в аренду системы хранения и когда срок аренды заканчивается хранящиеся на системе данные надо переместить на другой массив. Если в рамках внедрения ILM на основе виртуализации SAN компания построила абстрагированный уровень хранения, то перемещение данных между массивами можно выполнить с помощью тех же технологий и процессов. Даже если система хранения принадлежит самой компании, то когда-нибудь у нее не останется свободной емкости и/или она

устареет. Способность технологий для внедрения ILM перемещать данные помогает независимо от исходной причины перевода устройства хранения на другой уровень. В скором будущем Brocade и ее партнеры предложат технологии для миграции данных.

Аналогичным образом решения ILM могут реплицировать данные между массивами и технология ILM в принципе может использоваться в решениях для непрерывности бизнеса. На самом деле, перемещение данных настолько полезно, что многие администраторы хранения крупных ЦОДов внедряют SAN прежде всего для обеспечения этих процессов. Внедрение решений для перемещения данных на основе SAN – это следующий логичный шаг, который также продвигает к внедрению законченного решения ILM в будущем.

### ***Проблемы внедрения ILM***

Очевидно, что концепции и архитектуры ILM и UC очень похожи, но они также имеют и схожие проблемы внедрения. Концепция ILM достаточно проста – все ресурсы хранения должны быть прозрачно доступны по требованию для всех вычислительных узлов и нужно внедрить политики и механизмы, которые гарантируют, что все данные «живут» на «правильных» устройствах хранения. В теории процесс внедрения достаточно прост – нужно определить политики для выделения ресурсов хранения (т.е. определения, какое устройство хранения будет хранить определенные наборы данных), а затем на основе этих политик определить и внедрить процедуры для перемещения данных между устройствами хранения по мере необходимости.

Как и в случае UC, на практике внедрение ILM намного сложнее, чем в теории.

Автоматизированная система ILM должна перемещать данные по мере необходимости,

поэтому для построения законченной системы ILM нужно решить проблему перемещения данных между устройствами хранения в онлайновом режиме без прерывания работы хостов, которые обращаются к этим данным. Разумеется, не каждый день нужно останавливать приложение из-за того, что истек срок аренды массива, на котором хранятся данные этого приложения, но миграция данных выполняется постоянно, причем она может выполняться и для отдельных файлов каждую минуту, поэтому даже самые терпеливые пользователи требуют прозрачности этого процесса.

Реализация ILM должна уметь копировать данные на новое место, которые в течение этого процесса могут активно модифицироваться, проверить корректность этой новой копии и затем вывести старую копию из активного сервиса. Для того, чтобы эти процедуры не нарушили работу приложений нужны специальные механизмы, которые внедряются на сервере приложений или в SAN, т.е. решение для перемещения данных располагается между приложением и устройствами хранения, причем само приложение «не видит», что его данные теперь хранятся на другом устройстве.

Реализация такого механизма на уровне серверов требует управления многочисленными взаимозависимостями оборудования и программного обеспечения, а также связанного с существенными затратами и перебоями процесса внедрения. В результате обычно используется подход на базе SAN и в идеальном ЦОДе ILM сама SAN возьмет на себя все операции перемещения данных и перенаправления запросов приложений после таких перемещений.

Кроме того, решения ILM должны предоставить менеджерам возможность задания политик, которые определяют когда и куда перемещаются наборы данных.

Очевидно, что для успеха система ILM должна обладать удобной, надежной и мощной функциональностью, однако на практике реализовать такую функциональность с помощью программного обеспечения намного труднее, чем полагают многие потребители, и здесь заключается еще одна серьезная проблема для компаний, разрабатывающих программное обеспечение и оборудование для внедрения ILM.

Даже если требования к внедрению «идеального» решения ILM выполнены, сам процесс внедрения представляет проблему. Даже при использовании виртуализации на базе SAN первоначальное внедрение ILM может в определенной степени нарушить нормальный режим работы. Кроме того, нужно провести дополнительное обучение администраторов, развернуть пакеты сетевого мониторинга и т.д.

### ***ILM на практике***

Теоретический потенциал ILM пока не реализован на практике, но определенные компоненты уже доступны сегодня. Например, уже возможно построить сетевую инфраструктуру для архитектуры ILM, что позволяет подготовить ЦОД к внедрению новых решений ILM по мере их появления. Помимо внедрения ILM эта архитектура обеспечивает и другие преимущества, в том числе и повышение эффективности работы *всего* ЦОДа. Кроме того, на рынке уже есть *специализированные* продукты для внедрения ILM.

Например, можно считать, что любая технология, обеспечивающая перемещение данных, помогает и внедрить ILM. Выполняющееся на хосте программное обеспечение для зеркалирования может использоваться для подключения к хосту нового дискового массива, синхронизации данных на нем, проверки новой копии и отключения исходной копии. Это надежный метод миграции данных, которую во многих случаях

можно выполнить без нарушения работы пользователей системы.

Однако применение такого программного обеспечения еще не означает полного внедрения ILM. Полнфункциональная система ILM сможет выполнять такие операции автоматически на основе заданных пользователем политик и в идеале не будет при этом использовать ОС хоста или «привязана» к конкретному оборудованию. Кроме того, решения для зеркалирования на уровне хоста при перемещении данных создают дополнительную нагрузку на процессоры системы и в результате снижается производительность основных приложений хоста.

Другой пример – программное обеспечение репликации на уровне массивов, которое предлагают многие вендоры. Однако это программное обеспечение используют специфичные для оборудования конкретного вендора функции, поэтому обычно не способно копировать данные между массивами разных производителей. Даже в тех случаях, когда такое копирование возможно, его нельзя выполнения без нарушения работы массива, с которого происходит копирование, и хотя существует решение этой проблемы в виде специальных программ, которые нужно установить на всех хостах, при их использовании, как и в случае с зеркалированием на уровне хостов, падает производительность основных приложений. Кроме того, решения ILM должны состоять из нескольких уровней хранения, в том числе и из дешевых массивов, которые обычно не поддерживают программное обеспечение репликации на уровне массивов.

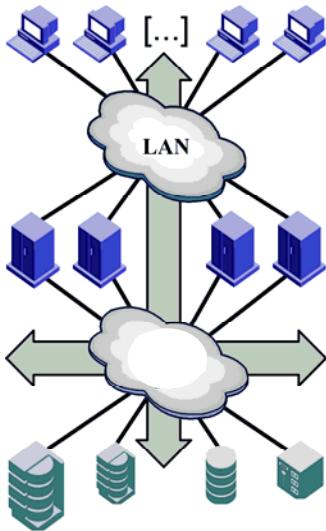
Решения для репликации на уровне хоста и массивов будут применяться в ЦОДах и в будущем, но по своей архитектуре они не подходят для «идеального» решения ILM – требуется решение на базе SAN. Компоненты и

приложения для виртуализации SAN уже доступны конечным пользователям и реализуют частичную автоматизацию ILM на базе SAN.

В ближайшем будущем эти продукты ускорят и упростят администрирование миграции данных, а также некоторые из проблем внедрения законченного решения ILM поскольку они избавляют от необходимости инсталлировать на каждом хосте специальное программное обеспечение. Виртуализаторы на базе SAN не следует рассматривать как *универсальное* решение для внедрения ILM, но они обеспечивают преимущества в кратко- и долгосрочной перспективе и поэтому их должны рассматривать как один из возможных вариантов клиенты, которые хотят внедрить у себя ILM.

## SAN: на пересечении UC и ILM

Хотя окончательные версии ILM и UC пока не вышли на рынок (и не выйдут в ближайшее время), но зато доступны полезные компоненты для их внедрения. Чтобы воспользоваться преимуществом существующих решений UC и подготовиться к внедрению будущих решений необходимо, чтобы каждый вычислительный узел имел доступ к набору данных любого приложения или сервиса, которые могут запускаться на нем. Чтобы вычислительный узел выиграл от решения ILM он должен одновременно иметь доступ к устройствам хранения всех уровней вместе с возможностью прозрачного перемещения между ними данных, что возможно только при использовании сети. Таким образом, SAN нужна для обеспечения мобильности данных и вычислений, а значит SAN – это *пересечение* этих двух важных тенденций (см. Рис. 27).



**SAN – это пересечение архитектур ILM и UC:**

- **UC:** Приложения при необходимости можно перемещать между вычислительными узлами.
- **ILM:** Данные при необходимости можно перемещать между устройствами хранения.
- **SAN:** Любой вычислительный узел может получить доступ к данным приложения, которое на нем работает, независимо от того, на каком ресурсе хранения они сейчас размещены. Таким образом, SAN поддерживает мобильность данных и вычислений для ILM и UC.

**Рис. 27 - SAN на пересечении UC и ILM**

Индустрія IT переходить к моделі, при якій для підтримання конкурентоспроможності требуються мобільність даних і гібке розподілення процесорних мощностей. Як було показано вище, SAN необхідна для обсягування цієї мобільності, тому стратегіческі інвестиції в інфраструктуру SAN допомагають внедрити оба цих підходів.

### ***План поэтапного внедрения ILM и UC***

Внедрение ILM и UC с помощью SAN рекомендуется выполнять поэтапно.

Первый этап – это оценка. Нужно определить, какие проблемы нужно разрешить или какие усовершенствования должны быть получены с помощью решений ILM и/или UC. Возможно, стоит провести аудит имеющейся среды IT чтобы выявить какие ресурсы имеются на настоящий момент и как они используются. Затем нужно проанализировать,

насколько улучшится эффективность операций после усовершенствования мобильности данных и/или вычислительных ресурсов.

При оценке возможных планов внедрения и их стоимости нужно учитывать существующую сетевую инфраструктуру. Если в компании уже построена SAN, то следует выяснить обеспечивает ли ее архитектура подключения any-to-any для всех узлов хранения и вычислительных узлов, обладает ли она достаточными характеристиками НА, способна ли каждая сеть справиться с той дополнительной нагрузкой, которую создают обычно решения ILM и UC? ( эти концепции подробно рассматриваются в следующих главах.)

Если имеется SAN, то следует рассмотреть возможность подключения к ней всех серверов, которым не хватает дисковой емкости. Если же используются несколько небольших SAN, то имеет смысл попробовать консолидировать их, используя FC- маршрутизаторы с LSAN. Наконец, можно рассмотреть возможность применения более мощных сетевых устройств, например, коммутаторов 4Gbit или 8Gbit и маршрутизаторов с поддержкой агрегирования каналов.

После завершения обследования необходимо подготовить набор правил, определяющих перемещение данных (IL M) и приложений (UC), в том числе, куда будет происходить перемещение и при каких условиях. На этом этапе надо заниматься выбором технологических решений для перемещения, а сосредоточиться на бизнес-процессах, для которых необходимо перемещение данных. Например, отчет может выглядеть следующим образом:

*Наша компания тратит слишком много денег на новые дисковые массивы корпоративного класса, причем согласно результатам последнего обследования, более 50% хранящихся на наших*

*массивах корпоративного класса данных либо вообще не используются, либо обращение к ним происходит, крайне редко, поэтому им не требуется производительность корпоративного класса или функции НА. Необходимы средства для переноса этих второстепенных данных на более дешевые устройства хранения по мере того, как они теряют ценность для бизнеса. Для этого требуется автоматизированное решение, иначе сэкономленные на стоимости оборудования деньги пойдут на оплату труда администраторов, задачи которых серьезно усложняются. Внедрение автоматизированного решения ILM сэкономит нашей компании \$x в месяц за счет сокращения затрат на текущее администрирование и \$y в год за счет экономии затрат на приобретение дисковых массивов.*

Отметим, что хотя в отчете речь идет об улучшении текущих технологий управления хранением, начинается и заканчивается он с экономического эффекта от внедрения предлагаемых решений. В “Глава 5: Планирование проекта” ( со стр. 149) даются другие примеры того, как надо изучать требования бизнеса и как описать их в плане проекта по построению SAN.

К концу этого процесса уже можно определить, что будет перемещаться (данные или компьютерные ресурсы), почему нужно перемещение (деньги, com pliance и т.п.), откуда и куда будут перемещаться данные или вычислительные ресурсы (между узлами) и когда будет происходить перемещение. После этого можно считать, что определены бизнес-ориентированные политики ILM и/или UC.

Следующий шаг – решение о поэтапном внедрении автоматизированной реализации этих политик. Этот план, который должен быть представлен руководству компании, состоит из кратко-, средне- и долгосрочных

целей проекта с указанием затрат на реализацию каждого этапа проекта и преимуществ от его выполнения. Например, можно внедрять решение ILM следующим образом:

**Этап I:** Внедрить базовую сетевую архитектуру и вручную частично внедрить процессы ILM для ограниченного числа узлов (только для бизнес-критичных серверов). Перемещение данных ограничивается ночным резервным копированием через сеть DR (Disaster Recovery). Хотя процессы будут выполняться в основном вручную, они помогают выполнить требования законодательства по более надежной защите данных компании.

**Этап II:** Расширить число узлов, охваченных ILM, добавить репликацию и частично внедрить автоматизацию. В результате ILM будет работать на узлах с бизнес-критичными системами и системами, при простое которых убытки не менее \$x в час. Резервное копирование и репликация через сеть DR будут автоматизированы, что обеспечит соответствие требованиям закона (compliance) и сократит расходы на администрирование. Эта модель также поддерживает перемещение данных между уровнями хранения в ручном режиме.

**Этап III:** ILM внедряется на остальных узлах ЦОДа, расширяется автоматизация процессов и число поддерживаемых вариантов перемещения данных, включая автоматическую миграцию на экономичные уровни хранения для сокращения единовременных затрат и экономии на системах хранения корпоративного класса.

## *Выбор внедряемых приложений на основе SAN*

Как было показано выше, первый логичный шаг – это внедрение высокоуровневой сетевой архитектуры и попытка вручную обслуживать процессы, которые работают на некоторых хостах (узлах) в течение одного – двух месяцев для того, чтобы убедиться в правильности определения процессов до того, как будет внедрена автоматизация. Выполнение этого шага предусматривает внедрение SAN и основных технологических компонентов для перемещения данных.

Например, можно построить SAN в минимальной конфигурации (НВА, коммутаторы, маршрутизаторы и программное обеспечение) для загрузки через SAN если целью является внедрение решения UC. Когда потребуется переместить приложение на новый узел, то если на новой платформе использует другая аппаратная архитектура, то можно вручную отредактировать конфигурацию хранящегося в сети загрузочного образа с учетом специфики нового оборудования. Хотя процесс выполняется вручную, он все равно является решением UC, которое попадает в первую строку Таблица 1 (стр. 87).

Затем пользователи могут решить, что настало время улучшить процессы с помощью автоматизации и распространить их на другие узлы. На этом этапе обычно внедряется «контроллер конфигураций загрузочных образов» на базе SAN или другое аналогичное решение. Если по каким-то причинам такое решение нельзя внедрить сразу, то лучшее заранее выбрать общий подход – это гарантирует, что не потребуется менять общую архитектуру SAN для построения полноценного UC. Следует проанализировать предлагаемые разными вендорами технологии и выбрать из них ту, которая оптимальна для бизнеса в долговременной перспективе.

## Проектирование подключений SAN

На последнем этапе этот процесс будет распространен на весь ЦОД. Хотя не стоит заранее проектировать это решение на первом этапе проекта, лучше предусмотреть возможность расширения SAN до этого масштаба. Обычно для этого в каждой фабрике нужно использовать лучшие в своем классе высокопроизводительные коммутаторы 4/8Gbit Fibre Channel и иерархическую архитектуру с маршрутизацией (LSAN), охватывающую фабрики, которая хорошо масштабируется по мере внедрения новых компонентов. При выборе сетевой архитектуры нужно учитывать следующие факторы:

1. Может ли любой сервер получить доступ к любому устройству хранения? Даже если используются локальные устройства хранения, то следует учитывать принцип *энтропии операций*, согласно которому любое решение должно адаптироваться к изменению требований. Единственный способ подготовиться к дополнительным задачам, которые могут возникнуть в будущем, - это использование подключений по схеме any-to-any. В крупной инфраструктуре невозможно либо нежелательно подключить все устройства к одной большой фабрике, но такие новые технологии как маршрутизаторы FC для организации LSAN способны решить эту проблему.
2. Обеспечивает ли решение достаточное *число портов* для масштабирования на все узлы ЦОДа? Ценность сети увеличивается экспоненциально по мере роста числа узлов и поэтому решение эффективное ILM или UC будет быстро развиваться.
3. Можно ли масштабировать *производительность* сети для поддержки всех планируемых перемещений данных, например, одновременной перезагрузки всех серверов через сеть, восстановления после крупной

аварии или миграции данных одновременно с нескольких массивов?

Если проект сети соответствует всем этим критериям, то он хорошо подходит для внедрения решений ILM или UC, которые появятся в будущем, а также обеспечивает отдачу в краткосрочной перспективе. См. “Главу 5: **Планирование проекта**” и следующие главы, где подробно описано построение масштабируемой и высокопроизводительной SAN с подключением any-to-any.

# 4

## 4: Обзор проектирования SAN

В этой главе рассказывается об основных особенностях проектирования SAN. В ней дается описание разных вариантов построения вместе с рекомендациями по выбору оптимального варианта. Многие из рассматриваемых тем достаточно сложны, поэтому им посвящены отдельные главы этой книги, а в этой главе дается только введение в такие темы и ссылка на соответствующую главу.

Проектирование SAN, как и остальных типов ИТ-инфраструктуры, включает сбалансированный учет разных требований и, если они конфликтуют между собой, то нахождение компромисса. Например, требование внедрить сеть с минимальными затратами может противоречить требованию обеспечить максимум производительности и доступности сети, поэтому архитектор SAN часто вынужден идти на компромисс. В этой главе объясняется, как можно разрешить подобные конфликты.

Автор этой книги неставил своей целью разработать четкие законы проектирования SAN – у каждой сети есть свои специфические требования, поэтому не может быть точных правил обеспечения баланса разных требований, которые можно было применить в любой ситуации. Цель книги – объяснить архитектору SAN, какие аспекты нужно учитывать при проектировании и дать рекомендации на основе принятых в индустрии

методических указаний и практики использования (best practices).

В начале рассматриваются общие аспекты проектирования SAN, включая выбор используемых протоколов, расстояний, топологии и производительности. Эти факторы относятся к любой сети, поэтому их нужно определить с самого начала проектирования. После этого объясняются более узкие темы, например, технологии для построения территориально-распределенных сетей и баланс между стоимостью и производительностью.

## **Совместимость**

Первый вопрос, который нужно решить при проектировании SAN, - это совместимость каждого компонента инфраструктуры сети хранения с остальной инфраструктурой и с каждым хостом и устройством хранения, с которым этот компонент может взаимодействовать. Если устройства несовместимы, то сеть просто не будет работать и ее проектирование теряет смысл.

Если устройства совместимы между собой как на аппаратном, так и программном уровне, то их можно соединить напрямую или через сеть и они смогут обмениваться данными («разговаривать»).

Например, драйвер HBA- адаптера должен быть совместим с операционной системой хоста и если драйвер работает только под Windows, а хост работает под Solaris, то они несовместимы на программном уровне, хотя могут быть совместимы на аппаратном уровне. Также требуется и совместимость оборудования HBA и хоста: если HBA рассчитан на шину PCI, то его нельзя установить в слот SBUS.

Когда Fibre Channel разрабатывался в середине 1990-х, то, как и при разработке любого нового протокола, нужно было решить целый комплекс проблем совместимости, начиная с уровня формата пакетов и до уровня приложений. Brocade и другие компании из индустрии хранения приложили много усилий разрешению этих проблем и, хотя некоторые из них по-прежнему актуальны, большинство продуктов Fibre Channel совместимы между собой на всех уровнях. Остающиеся проблемы связаны с обеспечением совместимости сервисов верхних уровней – например, драйверы дублирования каналов могут поддерживаться только определенными устройствами хранения и должны использовать определенный микрокод контроллера HBA и RAID.

В своей работе архитектор SAN редко сталкивается с вопросами совместимости, поскольку большинство из проблем совместимости успешно решены индустрией FC, а немногие оставшиеся также постепенно устраняются. Однако другие технологии SAN (например, iSCSI) пока не продвинулись так далеко, поэтому при выборе технологии для построения SAN нужно учитывать и ее уровень зрелости, например, как давно протокол используется в крупномасштабных критически-важных сетях и сколько портов уже установлены в *промышленную эксплуатацию*? Также нужно учитывать совместимость с продуктами других вендоров на *всех* уровнях, не только на уровне пакетов, но и сервисов хранения, приложений хостов и дисковых массивов. Кроме того, при выборе решения следует выяснить, предлагает ли его вендор поддержку для всей инфраструктуры SAN.

Для SAN требуется совместимость всех аппаратных и программных компонентов, начиная от приложений и до устройств хранения. Например, HBA может быть

совместим как с устройством хранения, так и с коммутаторами, которые связывают его с этим устройством, но не поддерживать такие функции SAN, как фирменные теги VL AN. Итак, нужно проанализировать совместимость:

- Протоколы (форматы пакетов)
- Используют ли все устройства в цепочке один и тот же протокол?
- Например, используют ли все устройства стандартный формат пакетов FC? Если используется нестандартный заголовок пакетов FC, совместимость устройств невозможно обеспечить.
- Используют ли все устройства стандартные пакеты по *одному и тому же алгоритму*? Если правильные пакеты пересылаются между двумя устройствами, то их вендоры должны совместно обеспечить единую *интерпретацию* стандартов. Это требование относится ко всем обсуждаемым далее сервисам.
- Совместимость между узлами и коммутатором (сервисы, протоколы)
- Можно ли узел (хост или устройство хранения) успешно подключить к фабрике? Для этого нужна поддержка описанных в стандартах процедур и совместимость на уровне процедуры внедрения.
- Узел должен корректно взаимодействовать со всеми сервисами фабрики и процессами, прежде всего fabric login (FLOGI) и сервером имен (SNS).
- Как ведут себя узлы при сбоях? Нужно оценить все компоненты цепочки – кабели, волокно, коммутатор и даже фабрику целиком. С учетом каждого компонента нужно решить, как узел себя поведет если компонент полностью выйдет из

строя или будет работать нестабильно (во многих случаях нестабильные компоненты причиняют больше вреда, чем вышедшие из строя<sup>41</sup>.)

- Совместимость между узлами (приложений, сервисов, протоколов)
- Сегодня основные проблемы совместимости SAN связаны с уровнем приложений. Сможет ли драйвер дублирования каналов взаимодействовать с микрокодом контроллера RAID? Сможет ли драйвер HB A одновременно работать с JBOD и ленточной библиотекой?<sup>42</sup>
- Совместимость коммутатора с другими коммутаторами и маршрутизаторами (сервисы и протоколы)
- У коммутаторов SAN есть свои специфические проблемы совместимости, которые отсутствуют в коммутаторах для других сетей. Коммутатор SAN можно сравнить с коммутатором Ethernet L3, который также обслуживает сервисы DNS, DHCP, WINS, NIS, LDAP и т.п. Кроме того, сервисы всех коммутаторов должны быть кластеризованы и самоконфигурироваться для большинства возможных сценариев. У других сетевых

---

<sup>41</sup> Brocade проводит тщательное тестирование узлов чтобы обеспечить их работоспособность даже при самых маловероятных отказах. Этот процесс называется “OEM qualification”. Тщательное тестирование необходимо, поскольку убытки при сбое SAN будут очень большими.

<sup>42</sup> Важно понять, что это – основная нерешенная проблема совместимости SAN, о которой нужно помнить при чтении маркетинговых материалов вендоров о новых технологиях. Некоторые вендоры обещают, что iSCSI решит все проблемы совместимости потому что это «просто IP». Нерешенные проблемы совместимости SAN связаны не с уровнем пакетов, поэтому переход с FC на IP никак не повлияет на совместимость. На самом деле никто пока даже не начал заниматься проблемами совместимости iSCSI на уровне приложений, поэтому развертывание iSCSI сегодня создает больше проблем совместимости, чем сети FC.

технологий нет таких требований, поэтому их архитектура принципиально отличается от SAN.

- Все современные коммутаторы FC, не использующие VSAN, используют совместимый формат пакетов, однако протоколы определяют такие детали, как *когда* посыпать конкретный пакет и сколько ждать его подтверждения.<sup>43</sup> Для соединения коммутаторов разных вендоров требуется, чтобы сервисы верхних уровней (сервер имен, FSPF и зонирование) были в явной форме сертифицированы *всеми компаниями, участвующими в поддержке* как совместимые для внедряемой конфигурации.
- Даже если сеть из коммутаторов разных вендоров успешно работает, то установка на одном из них обновления микрокода может нарушить совместимость.

Без совместимости на всех уровнях решение просто не будет работать. Даже если обеспечена совместимость на уровне транспортов, несовместимость на уровне драйвера резервирования каналов не позволит внедрить это решение. Brocade обладает огромным опытом в обеспечении совместимости, которого нет у других вендоров, поэтому риск несовместимости в сети, построенной с помощью коммутаторов и маршрутизаторов, будет меньше, чем при использовании других подходов. Однако, само по себе использование продуктов Brocade еще не гарантирует полной совместимости узлов и требуется проверить, сможет ли

---

<sup>43</sup> Например, даже небольшая разница в синхронизации кадров может нарушить работу решения. Если одно устройство ждет подтверждения две секунды, а другое – четыре секунды, то между ними не будет стабильной связи либо они вообще не смогут «разговаривать».

определенный хост работать с определенным дисковым массивом и т.п.



## Заметки на полях

*Успех внедрения также зависит как от соблюдения стандартов, так и учета деталей.*

*Например, хотя микрочип, используемый во многих 1Gbit FC HBA, хорошо работает, и совместим со стандартами FC, он изначально был разработан для приложений point-to-point и обладает характеристиками, из-за которых его производительность в фабрике деградирует при подключении к коммутатору, который спроектирован в соответствии с популярной интерпретацией стандартов.*

*Поскольку компания Brocade с самого начала работала в индустрии хранения, то ее проектировщики ASIC учитывали эти особенности и поэтому реализовали для ASIC режим hardware assist для компенсации этого явления. Если коммутатор Brocade обнаружит, что в HBA используется этот чип, то он соответствующим образом интерпретирует стандарт. Если же коммутатор «разговаривает» с HBA другого типа, то используется другая более популярная интерпретация. Благодаря этому устройство Brocade обеспечивает максимум производительности при работе с HBA первого и второго поколения.*

*В то же время, вендоры, недавно вышедшие на рынок хранения, не обладают такими знаниями и поэтому их продукты не способны взаимодействовать со всеми установленными НВА. Даже сегодня большинство других вендоров даже не начали сертифицировать свои продукты на совместимость с этими широко используемыми чипсетами. Это еще один пример преимущества проектирования SAN с использованием продуктов Brocade.*

## Сетевые топологии

В зависимости от контекста можно использовать несколько разных определений «топологии SAN», например, если речь идет о типе портов коммутатора Fibre Channel можно сказать «Порт #1 использует топологию *петли* поскольку он работает в режиме FL\_Port». Однако если речь идет о проектировании SAN, то под топологией понимается геометрия расположения коммутаторов, маршрутизаторов и других инфраструктурных элементов, из которых состоит сеть хранения.

На схеме организации сети инфраструктурное оборудование образует геометрические фигуры и топологии получают название в зависимости от этих форм.

Имеется бесконечно много возможных технологий и надежная архитектура сервисов фабрики позволяет построить фабрики любой сложности. К счастью, на практике можно ограничиться простыми решениями, поэтому для построения SAN используются несколько технологий в комбинации или в модифицированном виде в зависимости от конкретного внедрения. Эти топологии популярны поскольку обеспечивают прекрасную масштабируемость, производительность, доступность и управляемость сети. На самом деле

выбор топологии так сильно определяет свойства SAN, что многие архитекторы считают его главным в процессе проектирования.

Наиболее часто используемые топологии SAN:

- Каскад
- Кольцо
- Сетка (mesh)
- Центр/периферия (Core / Edge, CE)

Более подробно топологии рассматриваются в “Глава 6: Планирование топологии”.

## **Надежность, доступность и обслуживаемость (RAS)**

Эти три темы известны как аббревиатура RAS. Компоненты RAS существенно отличаются между собой и каждый из них влияет на общую функциональность продукта, сети и решения. Архитекторы SAN должны учитывать параметры RAS при выборе компонентов и общей сетевой инфраструктуры.

### **Надежность**

Надежность – это параметр, определяющий, сколько времени компонент будет работать по отношению ко времени, которое он будет простоявать из-за сбоев. Параметры надежности имеются как у аппаратных компонентов, так и программных, а также у сети и всего решения.

Один из способов оценить надежность – это определить, как часто его должен обслуживать специалист по сервису. «Событием надежности (reliability event t)» считается любая ситуация когда требуется сервисное обслуживание, даже если во время

обслуживания компонент продолжает работать. Например, рассмотрим коммутатор SAN с резервированными и заменяемыми в горячем режиме источниками питания. Если один из них сломается, то коммутатор будет по-прежнему работать во время замены неисправного источника питания, но эта ситуация все равно считается событием надежности.

При выборе компонентов архитекторы SAN должны учитывать надежность по двум причинам:

1. Чем больше надежность, тем меньше затраты на текущую поддержку, поскольку при выходе из строя любого компонента происходит обращение в службу поддержки и в результате возникают расходы на замену оборудования, оплату работ специалистов службы поддержки и убытки из-за упущеных возможностей, поскольку персонал, ремонтирующий систему, мог бы быть использован для других работ, которые бы принесли компании дополнительную прибыль.
2. В некоторых случаях низкая надежность компонентов приводит к выходу из строя всей системы, т.е «событие надежности» приводит к «событию доступности» (см. следующий раздел).

Надежность обычно измеряется в среднем времени между сбоями и среднем времени восстановления (Mean Time Between Failures, MTBF и Mean Time To Repair, MTTR.) Эти показатели обозначают, с какой периодичностью компонент может выйти из строя и сколько времени занимает, в среднем, его ремонт. При проектировании SAN следует выбирать компоненты с большим MTBF и низким MTTR.<sup>44</sup>

---

<sup>44</sup>

Помимо времени, необходимого для ремонта компонента, нужно учитывать и саму процедуру ремонта. Например, один вендор SAN

Значение MTBF и MTTR может зависеть от нескольких факторов.

Например, MTBF пропорционально общей зрелости продукта. Независимо от того, насколько тщательно компания разрабатывала свое оборудование и программное обеспечение, неизбежно, что в первой версии продукта будут определенные ошибки, которые проявят себя только через какое-то время после начала эксплуатации и тогда их можно будет устраниить. (Из-за этого многие сетевые администраторы не хотят переходить на новые версии ОС или оборудования до выпуска по крайней мера первого пакета «заплаток».) Надежность продукта прямо пропорционально зависит от того, как долго его производитель работает на данном рынке. Многие архитекторы SAN выбирают инфраструктуру Brocade поскольку Brocade уже более десяти лет поставляет коммутаторы SAN для критически-важных приложений и ни один другой вендор не обладает таким опытом. Поэтому зрелость, как оборудования, так и программного обеспечения Brocade *намного выше*, чем у конкурентов, и что на практике означает более высокую надежность.

Другой фактор – это интеграция компонентов. Чем меньше в системе компонентов, тем меньше риск ее отказа. В директорах особенно много отдельных компонентов в каждом лезвии. Оценить ожидаемую надежность продукта можно по дизайну его оборудования и числу компонентов. Для повышения надежности Brocade потратила много времени и усилий на уменьшение числа компонентов всех своих продуктов

---

предлагает «директор» с активными компонентами backplane. Для замены неисправной backplane нужно разобрать весь директор. Не имеет значения, сколько по времени выполняется такая операция – она просто недопустима для продуктов класса директор.

и, например, лезвие Brocade 48000 или материнская плата Brocade 4100 содержат меньше компонентов, чем аналогичные продукты конкурентов (вся основная логика коммутации 16-портового лезвия реализована на одном микрочипе!), поэтому вероятность отказа у них меньше. Трудно себе представить более плотную интеграцию компонентов.



## Заметки на полях

*Мы уже говорили о возможной несовместимости заголовков пакетов. Хотя эта проблема была в основном разрешена еще в прошлое десятилетие при ратификации протокола Fibre Channel, упоминать о ней стоит по двум причинам:*

*(1) Хотя для FC эта проблема давно потеряла свою актуальность, для технологий IP SAN она еще не решена. Только один вендор выпускает устройство iFCP, которое к тому же несовместимо ни с одним другим продуктом, а в устройствах iSCSI может использоваться любой из 20 предлагаемых вариантов этого стандарта.*

(2) Один недавно вышедший на рынок SAN вендор представил функцию VSAN, использующую нестандартный формат пакетов FC и потому несовместимый со всеми существующими узлами, коммутаторами и маршрутизаторами. Этот вендор требует от пользователей отключить теги VSAN для всех портов, к которым подключено оборудование других фирм, в том числе хосты и устройства хранения. Таким образом, вендор для решения проблемы несовместимости просто предлагает не использовать несовместимую с его продуктом функциональность. За исключением этого вендора в современной индустрии в FC полностью решена проблема совместимости заголовков пакетов, поэтому сейчас имеет смысл сосредоточиться на проблемах верхних уровней.

Связь между интеграцией компонентов и надежностью очевидна, но из зависимости общей надежности от зрелости и интеграции всей системы пользователь может сделать некорректные выводы. Например, у Brocade SilkWorm 3800 источники питания можно заменять в горячем режиме, а SilkWorm 3850 – нет. Большинство людей решат, что у 3800 лучше характеристики RAS, хотя на самом деле заменяемые в горячем режиме источники питания увеличивают число компонентов системы и значительно усложняют материнскую плату у 3800, которая сама по себе является нерезервированным компонентом. Упрощенный и более тесно интегрированный дизайн 3850 улучшает надежность материнской платы до такой степени, что весь коммутатор (включая оба источника питания) имеет более высокий показатель MTBF, чем одна материнская плата 3800. Риск неисправности коммутатора 3800 из-за отказа материнской платы выше, чем вероятность неисправности коммутатора 3850 из-за отказа материнской платы или источника питания. При

оценке характеристик RAS продукта нужно прежде всего учитывать *общий* показатель MTBF всех компонентов.

## **Доступность**

Доступность системы определяет, сколько времени она способна выполнять такие функции высокого уровня, как обслуживание приложений для конечных пользователей. Как и надежность, этот показатель зависит как от оборудования, так и программного обеспечения. Однако доступность не всегда напрямую зависит от сбоев компонентов.

Например, большинство хостов для подключения к SAN используют резервированные интерфейсы с программным обеспечением дублирования каналов для защиты от отказов. ( См. “Программное обеспечение дублирования каналов” на странице 28 и раздел после “Резервированные фабрики” на страницах 310 - 321.) Если в одном НВА произойдет сбой SFP, то в системе возникнет «событие надежности», поскольку SF Р нуждается в сервисном обслуживании. Чем больше SFP, тем больше вероятность отказов SFP. Однако при сбое SFP система и ее приложения сохраняют доступность – хост по-прежнему может обращаться к данным через другой НВА и обслуживать свои приложения. Через какое-то время необходимо заменить SFP, но даже эта процедура не приведет к простою при условии, что НВА правильно спроектирован.

Обеспечение доступности приложений считается крайне важным при проектировании SAN, поскольку проблемы доступности влияют не только на работу ИТ и других служб управления системами, но и на работу конечных пользователей. В предыдущем примере сбой SFP привел к «событию надежности» и потребовалась замена неисправного компонента, но при этом приложение продолжало работать. Если

возникнет сбой нерезервированного компонента (например, операционной системы хоста или материнской платы), то возникнет «событие доступности» и не только потребуется замена компонента, но и будет нарушена работа приложения. Большинство архитекторов SAN при выборе между более частыми «событиями надежности» и «событиями доступности» выбирают первый вариант.

Это важно, поскольку часто меры по *повышению* доступности ведут к *снижению* надежности. Если в каждой системе установлен только один НВ А, то статистически будет требоваться меньше сервисного обслуживания – чем меньше компонентов, тем реже возникают сбои. Однако в таком случае «событие надежности» вызовет и «событие доступности».

Обычно для измерения этого показателя используют «девяшки доступности», например, говорят, у системы доступность – пять девяток, т.е. система доступна не менее 99.999% времени. Иначе говоря, система может быть недоступна не более 0.0001% времени, что составляет около пяти минут в год.

Для обеспечения пяти девяток доступности приложений SAN требуется применять физически изолированные резервированные фабрики А/В и тогда выход из строя всей фабрики не нарушит доступность приложений. Поскольку доступность обычно важнее надежности, то полностью резервированная архитектура стала best-practice в индустрии.

Подробнее вопросы обеспечения доступности рассматриваются в “Глава 9: Планирование доступности” (страница 296).

## ***Обслуживаемость***

Показатель обслуживаемости (Serviceability) определяет, насколько легко выполняется сервисное обслуживание продукта или системы. Такая оценка носит субъективный характер. Как и два остальных показателя RAS, обслуживаемость может зависеть от оборудования, программного обеспечения, решений и даже общей архитектуры сети.

Имеются две причины, по которым обслуживаемость должна учитываться при проектировании SAN:

1. Чем лучше обслуживаемость, тем обычно меньше затраты на текущее обслуживание продукта. Например, если продукт сложный в эксплуатации, то надо потратить дополнительные деньги на обучение персонала или привлечение консультантов из другой фирмы.
2. Время работы продукта зависит от его обслуживаемости. Чем лучше обслуживаемость, тем быстрее можно восстановить его работоспособность. Кроме того, вероятность неправильного действия администратора возрастает по мере усложнения продукта.

MTTR можно рассматривать не только как характеристику надежности, но и обслуживаемости. Если компонент сломается, то за какое время можно будет его заменить? Чем меньше MTTR, тем быстрее выполняется ремонт, а значит, и лучше обслуживаемость.

К сожалению, остальные и более важные аспекты обслуживаемости не поддаются количественной оценке. Например, во многих продуктах есть встроенные средства диагностики, однако оценка обслуживаемости по числу таких средств будет некорректной, поскольку она не учитывает их функциональность, от которой

напрямую зависит удобство обслуживания системы.

Многие архитекторы для сравнения разных продуктов используют таблицы функциональности средств диагностики, причем в этой таблице каждому средству присваивается свой весовой коэффициент. Например, ping часто используется, поэтому этой утилите присваивается весовой коэффициент два, а crash dump используется редко и у нее весовой коэффициент равен единице. Для сравнения двух продуктов вычисляется сумма баллов средств диагностики и тот, у которого эта сумма больше, считается обладающим более удобными средствами диагностики. Кроме того, при оценке также может учитываться степень зрелости средства диагностики и насколько полно реализованы его возможности.

На самом деле, эти две характеристики важнее для обслуживания на практике, а не для количественной оценке средств диагностики. Лучше использовать несколько надежных средств диагностики, чем большое число утилит, которые работают нестабильно. Хотя ни один вендор не предлагает идеальные по своей зрелости и полноте реализации возможностей средства диагностики, у продуктов Brocade эти показатели намного выше, чем у ее конкурентов.

К сожалению, подробный анализ обслуживаемости при выборе компонентов требует от архитектора существенных усилий, а поскольку обслуживаемость редко стоит на первом месте при выборе продуктов, то многие архитекторы SAN применяют упрощенный подход. Например, они составляют список десяти главных функций обслуживаемости (таких, как отсутствие активных компонентов в backplane шасси) и учитывают, как долго вендор работает на рынке SAN, а также используют другие критерии для выбора компонентов.

## Производительность

При оценке производительности SAN надо учитывать такие характеристики, как протоколы, скорости линков, загруженность, блокирование и задержки. Часто нужно найти компромисс этих и других характеристик производительности и стоимости.

Например, удлинение SAN с помощью любого IP-шлюза Brocade или другого вендора неизбежно снизит производительность по сравнению с «родным» удлинением FC, FC over DWDM или FC over SONET/SDH. Линк 1Gbit Ethernet просто неспособен передавать трафик с дополнительной служебной информацией IP SAN, с той же скоростью, что линки 4Gbit FC, по которым передаются пакеты более эффективного протокола Fibre Channel. Хотя на практике решения Brocade FCIP работают быстрее, чем решения конкурентов, все равно эта платформа не может обработать большие пакеты IP так же эффективно, как обрабатываются пакеты FC. (Разница между этими двумя протоколами объясняется в «Протоколы SAN» на странице 33). Однако внедрение решений IP SAN могут стоить меньше, поэтому при выборе архитектор SAN должен учитывать как производительность и надежность FC, так и возможную экономию за счет применения IP. Большинство архитекторов предпочитают первый вариант решения, который к тому же проще внедряется, обладает более эффективными средствами устранения сбоев и расширения сети в будущем, но в отдельных случаях оптимальным может оказаться и вариант IP SAN.

Таким образом, при выборе возможных вариантов построения SAN необходимо учитывать производительность. Следует выбирать решения, которые не только соответствуют текущим требованиям к производительности, но и прогнозируемому

росту этих требований в будущем. Нагрузки на сеть обычно растут со временем и выбор протоколов и топологии SAN должен производиться с учетом различных сценариев роста требований к производительности.

Производительность SAN подробно рассматривается в “Главе 8: Планирование производительности” (страница 227).

## Масштабируемость

Для сетей хранения термин «масштабируемость» имеет несколько значений, например, он определяет, сколько данных можно хранить в определенном RAID-массиве (тогда говорят “*Этот* RAID-массив масштабируется лучше, чем *тот* потому что в нем можно установить больше дисков”).

Однако при проектировании инфраструктуры SAN под масштабируемостью обычно понимается максимальное число портов, которое способна поддерживать сеть без фундаментальной перестройки (тогда говорят “*Эта* модель сети лучше масштабируется, потому что в ней используются более мощные коммутаторы и поэтому она поддерживает больше портов при заданном числе доменов”).

Теоретически одна фабрика FC может масштабироваться до более 16 миллионов устройств, что более чем достаточно даже для самых требовательных заказчиков. Однако на практике невозможно масштабировать фабрику до такого числа устройств из-за многих ограничений, включая ограничения сервисов фабрики, программного обеспечения управления SAN, матриц совместимости и процессов управления SAN. (смотри “Сравнение фабрики с SAN и Meta SAN” на странице 42).

Таким образом, SAN нужно проектировать так, чтобы она могла масштабироваться до самых больших размеров, которые прогнозируются при ее развитии в обозримом будущем, а не ориентироваться на текущие требования. В противном случае SAN может быстро исчерпать свои возможности и вскоре после начала эксплуатации придется провести ее перестройку. Архитекторам SAN необходимо тщательно изучить вопросы масштабируемости при проектировании общей архитектуры сети.

Дополнительная информация по этой темедается в “Главе 7: Планирование масштабируемости” на странице 207.

## **Совокупная стоимость решения**

Давно прошли времена, когда у ИТ-подразделений был неограниченный бюджет, поэтому внедрение SAN должно быть эффективным по затратам. Архитектору SAN следует ориентироваться на совокупную стоимость решения, а не ограничиваться несколькими статьями расходов.

Например, при проектировании SAN можно сэкономить на стоимости оборудования и не применять решения НА, но в результате убытки от неработоспособности сети многократно превысят эту экономию. Поэтому прежде чем принять решение о развертывании SAN без резервирования, нужно просчитать не только экономию средств, но и убытки в долговременной перспективе от простоя подключенных к сети систем. Рассмотрим, что произойдет, если нерезервированная SAN используется для поддержки решения восстановления после катастроф и сама сеть выйдет из строя в результате аварии – ясно, что тогда не удастся восстановить бизнес-сервисы. Любой подобный инцидент в крупномасштабной инфраструктуре

приведет к таким убыткам, которые во много раз перекрывают расходы по построению резервированной SAN.

Иногда, кажется, что можно добиться экономии в кратковременной перспективе если отказаться от использования высокопроизводительных технологий. Оборотная сторона такого решения – снижение производительности пользователей, использующих подключенные к SAN системы, необходимость в частой модернизации инфраструктуры SAN и риск того, что в SAN нельзя будет внедрить решения нового поколения, например ILM и UC ( см. стр. 85), которым требуется высокая производительность.

В большинстве случаев краткосрочная экономия на инфраструктуре ведет к существенному увеличению совокупной стоимости владения SAN. Даже если для проектировщиков стоимость стоит на первом месте, то они должны учитывать *все* составляющие стоимости SAN, а не только несколько статей расходов. Это надо помнить при оценке сетевых топологий и общей архитектуры решений НА.

## Увеличение расстояний

После недавних глобальных потрясениях многие корпорации и государственные организации стали внедрять решения для восстановления после аварий и обеспечения непрерывности бизнеса. В одних компания эти решения внедряются чтобы повысить привлекательность предприятия для инвесторов, в других использование этих решений необходимо для выполнения требований государственных нормативов и правил. Кроме того, в результате объединений, поглощений и консолидации ЦОДов возникает потребность в репликации между площадками. Эти тенденции привели к спросу на решения для соединения

сетей FC SAN, расположенных в разных географических регионах.

Чаще всего в качестве решения используются высокопроизводительные и надежные технологии (темное волокно, *x*WDM и шлюзы SONET/SDH). Однако в некоторых случаях такие решения нельзя применить, либо они слишком дорогие для заказчика. Если при этом доступна высокоскоростная IP- сеть с надежными сервисами, то можно рассмотреть как вариант и применение технологии IP SAN.

Дополнительную информацию по этой теме можно найти в “Глава 11: Планирование территориальной разнесенности” (стр. 339).

## **Внедрение и другие работы**

Обычно после завершения проектирования SAN миссия архитектора считается выполненной и если речь идет о крупномасштабной инсталляции, то внедрением и управлением проектом занимаются другие люди. Тем не менее, архитектор может участвовать и в последующих работах, поэтому он должен спланировать инсталляцию SAN и ее текущее обслуживание, включая мониторинг сети и выполнение изменения конфигурации самой SAN и подключенных к ней хостов и устройств хранения.

Масштаб задач внедрения и управления SAN сильно зависит от размера самой SAN.

В небольших SAN, построенных из компонентов Brocade, благодаря зрелости продуктов Brocade и их функциональности plug-and-play потребность в управлении сетью возникает редко. Сама Brocade периодически проводит обследования площадок пользователей и обнаружила, что во многих коммутаторах небольших SAN даже несконфигурирован IP-адрес для управления, т.е. эти коммутаторы

работают без какого-либо вмешательства администратора, который использует лишь простой интерфейс администрирования зон WEBTOOLS.

Однако, в крупных компаниях процессы инсталляции и управления намного сложнее и здесь от архитектора зависит удобство использования SAN в кратко- и долгосрочной перспективе.

Например, обычно сложнее обслуживать SAN, состоящую из одной очень большой фабрики, чем SAN из двух резервированных фабрик (стр. 310) Архитектору может казаться, что управлять одной фабрикой проще, потому что такие изменения, как модификация зон выполняются только один раз. Однако современные средства управления SAN, такие, как Brocade Fabric Manager, позволяют управлять несколькими фабриками с одной консоли.

Кроме того, две физически изолированные фабрики позволяют проводить такие изменения, как обновление микрокода, сначала в одной фабрике, затем в другой, что существенно снижает риски, связанные с изменением инфраструктуры<sup>45</sup>. Разделение SAN на две фабрики сокращает в два раза размер фабрики, что обычно улучшает надежность и скорость работы утилит

---

<sup>45</sup> Это один из основных недостатков коммутаторов VSAN. Если с помощью программного обеспечения VSAN полно связанные сети разбиваются на виртуальные сети хранения A и B, как рекомендует упомянутый вендор, то изменение микрокода повлияет на все VSAN. Аналогичным образом, атака «отказ в обслуживании», из-за которой выйдет из строя шасси коммутатора, нарушит также и работу всех VSAN. Из этого можно сделать вывод, что вендор VSAN просто не понимает требований НА к проектированию SAN, что неудивительно, если учитывать отсутствие у него предыдущего опыта работы с технологиями сетей хранения. Разумеется, Brocade предлагает продукты, которые могут поддерживать эту модель и архитектор может использовать аппаратное зонирование вместо отдельных фабрик либо применять два домена директора SilkWorm, хотя компания никогда не рекомендовала прибегать к таким стратегиям.

управления, а также масштабируемость уровня управления<sup>46</sup>.

В больших инсталляциях выбор общей архитектуры SAN также влияет на масштабируемость даже при использовании подхода с фабриками “A/ B”. Если архитектор выбрал одну пару больших фабрик, то устранение потенциальных сбоев, вызванных текущими изменениями, будет крайне затруднено. Но если используется большое число небольших несвязанных между собой фабрик, то проблемой станет недостаток подключений между фабриками. Эти проблемы вызвали большой спрос на функции LSAN, реализованные в Brocade AP7420, 7500 и лезвии FR4-18i, поскольку архитектура LSAN позволяет реализовать связь между фабриками без построения одного большого плоского региона управления (см. стр. 315)

Архитектор также участвует в выборе стратегии зонирования. Если использовать короткие и понятные имена для зон и их псевдонимов, то это значительно упростит работу администратора SAN ( зонирование рассматривается в “Глава 10: Планирование безопасности ”, страница 325.)

Выбор компонентов SAN в определенной степени влияет и на управляемость. Архитектор может выбрать зрелые продукты от известного вендора либо

---

<sup>46</sup> Это еще один недостаток коммутаторов с VSAN. Все сервисы VSAN фабрики работают в директоре на одном процессорном модуле СР. Масштабируемость пропорциональна числу доменов и портов, которые обслуживает SAN с помощью определенного набора компьютерных ресурсов. Если есть одна плоская фабрика с x доменами и у портами, обслуживаемыми СР, то есть возможность для масштабирования. При той же конфигурации и использовании VSAN для разбиения будет по крайней мере 2x доменов, но число портов по-прежнему будет у и их будет обслуживать все то же оборудование. То есть, на практике VSAN ухудшает масштабируемость сети ...

«революционные» технологии от вендора, который недавно стал работать на рынке SAN. Каждый из этих подходов имеет свои плюсы.

Новые игроки на рынке предлагают функции с расширенными количественными показателями, которые еще только частично реализованы и даже не прошли полного тестирования, поэтому у этих вендоров широкая стратегия управления, которая, однако, не отличается глубиной проработки. Часто новые игроки предлагают свое оборудование бесплатно, что снижает стоимость приобретения.

Вендоры со зрелой технологией более консервативны относительно реализации новых функций, поскольку у них есть большая инсталлированная база и поэтому применение новых функций создает риски. Однако у них функции более надежны и стабильны, так как они больше времени разрабатывают программное обеспечение. Зрелые вендоры крайне редко дают свое оборудование бесплатно, но следует учитывать, что данные, которые хранятся в SAN, намного дороже самой SAN, поэтому затраты на оборудование вполне можно обосновать.

От архитектора SAN во многом зависит и выбор критериев при оценке функций управления, которые можно рассматривать с точки зрения качества и количества. Часто надо найти компромисс между затратами в долговременной перспективе (проблемы управления, ошибки и потери данных) и единовременными затратами на приобретение оборудования.

Наконец, архитектор обычно участвует и в выборе пакета управления. Применение утилиты Brocade Fabric Manager упрощает координирование текущего управления несколькими фабриками, а использование

бесплатного программного обеспечения Brocade SAN Health и ПО SAN Health Professional существенно помогает внедрить проактивное управление – этот инструмент автоматически сравнивает состояние SAN с обновляемыми best-practices и включает в себя автоматизированные функции обслуживания, например, поиска неиспользуемых зон.

Дополнительную информацию по этой теме можно найти в “Глава 12: Планирование внедрения” на странице 373.

## **Вторая часть**

### **Планирование проекта**

#### **Темы**

- Теория и практика проектирования SAN
- Советы по монтажу и конфигурированию
- Советы по текущему управлению
- Теория и практика проектирования SAN

# 5

## 5: Планирование проекта

Любой проект внедрения SAN можно разбить на несколько этапов. Сначала выполняется сбор требований для определения того, что должна обеспечить SAN, затем проектируется SAN, после чего закупается оборудование. Новое оборудование сначала развертывается в тестовой среде, выполняется его проверка, а затем SAN запускается в эксплуатацию. Наконец, наступает этап обслуживания. Через какое-то время часть компонентов SAN устареет и они выводятся из эксплуатации или переводятся на выполнение второстепенных функций.

Эти этапы называют “жизненным циклом” SAN. Под этим термином также понимают процедуры, в ходе которых архитектор проводит изменения в существующей SAN (добавляет в нее новые компоненты или заменяет старые). Любой проект внедрения SAN состоит из этих этапов, однако самыми главными являются два – сбор требований к SAN и проектирование ее архитектуры. Принятие самых важных решений происходит именно на этих этапах и поэтому им уделяется основное внимание в этой главе.

От правильного выполнения первых двух этапов зависит успех последующих этапов проекта. Если допущены ошибки при определении требований бизнеса к SAN или при проектировании архитектуры, то даже

безупречное выполнение всех остальных этапов не сможет обеспечить успех всего проекта.

Для правильного проектирования архитектуры нужно использовать структурированный подход к процессу планированию. Процесс может быть основан на имеющихся формализованных политиках и процедурах ИТ-департамента, либо на процедурах, специально разработанных для внедрения SAN. В любом случае, его результатом должна стать разработка продуманного плана проекта до начала закупок оборудования. Разработка этого плана требует тщательного исследования и анализа. В этой главе объясняется, как следует организовать эффективный процесс планирования SAN, какие данные нужно собрать и как они интерпретируются, а также даются рекомендации по использованию этих данных для финансового обоснования проекта.

## **Обзор процесса планирования SAN**

Существует несколько методов эффективного планирования SAN, которые позволяют добиться удовлетворения требований бизнеса. В этой главе, как пример, рассматривается один из этих методов, но это не означает, что нельзя использовать другие методы. Планирование SAN показано на примере крупного предприятия (при построении небольшой SAN процесс планирования может оказаться проще, чем в данном примере). Процесс планирования делится на пять этапов:

**Этап I:** Сбор требований

**Этап II:** Разработка технических спецификаций

**Этап III:** Оценка стоимости проекта

**Этап IV:** Анализ окупаемости инвестиций (ROI) и

совокупной стоимости владения (TCO) (при необходимости)

### Этап V: Детальная архитектура SAN и план внедрения

На первом этапе архитектор SAN проводит интервью со всеми, кто как-то связан с проектом, в том числе, системных администраторов, администраторов хранения и сетевых администраторов, ИТ-менеджеров, владельцев приложений, ключевых конечных пользователей и владельцев функций бизнес-процессов, связанных с системами, которые нужно подключить к SAN.

На втором этапе анализируются данные, собранные на Этапе I и определяется (хотя бы в общих чертах), какую технологию нужно использовать для удовлетворения требований бизнеса. Необязательно уже на этом этапе точно определять, какой порт какого коммутатора будет использован для подключения данного хоста, однако архитектор должен проанализировать данные и решить, сколько всего портов нужно установить и какой будет общая топология сети.

Полученные данные будут использованы на Этапе III для подготовки списка необходимого оборудования и компонентов. После того, как архитектор выяснит, как пакеты программного обеспечения будут использоваться, сколько потребуется кабелей, Fibre Channel адаптеров, портов коммутаторов, маршрутизаторов и шлюзов, он сможет относительно точно оценить стоимость внедрения.

После получения относительно точной оценки стоимости можно рассчитать, окупится ли внедрение SAN. Обычно анализ окупаемости инвестиций (Return on Investment, ROI) выполняется очень просто. “Внедрение

SAN для защиты от катастроф обойдется нам в x долларов. Для работы на европейском рынке наша компания в соответствии с местным законодательством должна внедрить катастрофоустойчивое решение и если мы этого не сделаем, то нам придется уйти с этого рынка и тогда убытки из-за потери прибыли в 100 тысяч раз превысят стоимость проекта.” В этом случае для расчета ROI не потребуется много времени, но в других случаях обоснование может оказаться намного сложнее. В этой главе приводится простой расчет. Также для обоснования внедрения SAN можно провести анализ Совокупной стоимости владения (TCO). Этот тип анализа дает очень *убедительное* обоснование проекта, поскольку он показывает экономию средств за счет упрощения управления. Считается, что внедрение SAN сокращает TCO управления данными более чем на 50%. Однако анализ TCO намного сложнее расчета ROI и выходит за рамки этой книги.

В любом случае, после обоснования проекта архитектор подготавливает детальное описание архитектуры и план внедрения, который может включать план зон и подключений портов, процедуры монтажа, тестирования и запуска SAN в промышленную эксплуатацию, детальные схемы топологии с указанием всех соединений.

## **Документирование проекта построения SAN**

В процессе планирования нужно документировать собранные данные и их интерпретацию, необходимое оборудование и соответствующие расходы и предлагаемую архитектуру. Обычно этот набор документов служит планом проекта и его использует менеджер проекта SAN для того, чтобы построить SAN в соответствии с выделенным бюджетом и сроками. У многих ИТ-департаментов уже есть формы и процессы для построения планов проекта – их можно

применять и для проекта SAN. Если же их нет, то можно просто записать все действия, которые описываются далее в этой главе, и зафиксировать их на бумаге и в электронной форме.

Когда план проекта SAN будет готов, то менеджер проекта будет иметь всю информацию, которая необходима для покупки всего требуемого оборудования и составления графика его монтажа. Далее будет несложно провести распределение ресурсов и составление графика выполнения в соответствии с бюджетом проекта, его основными этапами и целями, а также обосновать это перед руководством ИТ-департамента.

Если проект предусматривает расширение или изменение существующей SAN, то рекомендуется с помощью бесплатного ПО Brocade SAN Health подготовить первоначальный набор документов.

## **Определение участников проекта**

До сбора требований клиентов нужно подобрать команду, определить лиц, которые будут принимать основные решения и конечных пользователей SAN.

### ***Выбор менеджера проекта и архитектора SAN***

Менеджер проекта SAN отвечает за координацию всех работ по построению SAN и обычно строит свою деятельность на основе плана проекта SAN. Архитектор отвечает за «трансляцию» требований бизнеса в технические требования и подготовку на их основе детальной архитектуры. Работа менеджера по координации включает регулярное проведение совещаний, определение задач и инициацию процесса принятия решений относительно этих задач. Во многих случаях он должен обеспечить взаимодействие между

нужным персоналом и информировать всех участников о плане и ходе его выполнения.

В некоторых случаях менеджер является и архитектором SAN и поэтому отвечает также за техническую сторону проекта, а в других случаях архитектор SAN подчиняется ему и отчитывается перед ним и тогда они оба несут общую ответственность за успех проекта.

### ***Организация технической группы***

Небольшую SAN может построить один человек, однако внедрение SAN, состоящей из нескольких коммутаторов обычно требует усилий команды специалистов. По возможности, этот персонал надо привлечь еще на этапе проектирования – это гарантирует учет всех технических особенностей и отсутствие неприятных сюрпризов на этапе внедрения. Полезно, чтобы эти специалисты почувствовали, что от них зависит успех проекта. В зависимости от функций SAN и масштаба внедрения при крупных внедрениях техническая группа может состоять из следующих специалистов:

- Администраторы SAN
- Системные администраторы
- Администраторы систем хранения
- Администраторы IP-сети
- Администраторы баз данных
- Специалисты по приложениям

### ***Определение круга лиц, принимающих решение***

Обычно три группы лиц принимают решение о внедрении SAN: руководители, которые дают разрешение на закупки, специалисты, отвечающие за внедрение и/или управление SAN, и люди, которые будут использовать приложения, работающие на

подключенных к SAN серверах. Как и в случае с технической группой, этих людей надо привлечь к проекту как можно раньше. Если они понимают эффективность SAN для бизнеса и имеют возможность высказать свои пожелания относительно ее архитектуры, то скорей всего процесс утверждения пройдет достаточно легко. Круг лиц, принимающих ключевые решения при внедрении в крупных масштабах, обычно включает:

- CEO, СТО, СИО или CFO (исполнительный директор, технический директор, ИТ-директор и финансовый директор)
- Бухгалтерия
- Менеджеры по закупкам
- вице-президент по ИТ, директоры и менеджеры
- Члены технической группы
- Руководители, являющиеся пользователями бизнес-приложений
- Владельцы бизнес-процессов

### *Идентификация пользователей SAN*

“Клиентом SAN” может быть любой пользователь системы, каким-то образом подключенной к SAN. Если следовать этому определению, то для SAN, обслуживающей кластер web-серверов, клиентом будет любой пользователь Internet, обращающийся к этим web-серверам. Полезно уже на этапе проектирования определить круг клиентов SAN, поскольку от этого может зависеть выбор архитектуры SAN. У кластера, обеспечивающего web-хостинг популярного сайта Internet, более высокие требования к доступности и производительности, чем SAN уровня департамента, которая обслуживает рабочую группу в небольшой компании.

Однако, использовать такое общее определение часто не имеет смысла, поскольку на *этом* этапе главное – выяснить, кто будет участвовать в процессе проектирования. Разумнее будет задействовать в процессе только тех пользователей, которые сильно заинтересованы в SAN в силу своих должностных обязанностей или от которых зависит принятие решений о проекте. На практике это означает привлечение ограниченного числа “тяжелых пользователей”, которые будут представлять клиентов на этапе проектирования.

## Сбор требований

Внедрение технологии ради технологии давно стало непозволительной роскошью. Аргументацию “SAN – это замечательно, поэтому нам нужно построить свою SAN” никто не будет воспринимать всерьез. Хотя SAN – это действительно замечательные технологии (особенно на базе решений Brocade), должна быть более конкретная причина для затрат времени и бюджета ИТ-департамента на выполнение этого проекта. Если *нет* требований бизнеса, для удовлетворения которых нужно построить SAN, то бюджет ИТ лучше потратить на другие технологии, например, программное обеспечение FAN. Сбор требований нужен не только для выбивания денег на проект, но и определения целей внедрения SAN, которое необходимо для правильной разработки ее архитектуры, поэтому собранная на этом этапе информация должна быть максимально подробной и точной.

После определения участников проекта надо провести интервью с каждым из них и выяснить, что по его мнению должна обеспечить SAN и как она должна работать. Такие интервью лучше проводить лично (в крайнем случае - по телефону). Участников проекта надо попросить рассказать о проблемах, с которыми они сталкиваются сейчас, и их требованиях к SAN.

Лучше всего опросить их всех по одному, собрать все данные и затем провести круглый стол для обзора требований и только затем перейти к следующему этапу.

### ***Определение проблем бизнеса***

Перечисление проблем бизнеса в плане проекта SAN должно объяснить, почему с точки зрения бизнеса нужно построить SAN. Например, для коммерческой организации это перечисление показывает, как внедрение SAN увеличит прибыль, а для некоммерческой организации оно объясняет, как SAN поможет выполнить ее миссию. Это перечисление проблем формируется по результатам интервью участников проекта, которые были выбраны на предыдущем этапе.

В каждом интервью нужно выяснить, какие проблемы бизнеса по мнению конкретного участника проекта должна решить SAN с тем, чтобы гарантировать устранение максимального числа проблем. Кроме того, нужно выяснить не только текущие потребности, но и требования, которые могут возникнуть в будущем. Считается, что SAN должна быть спроектирована с учетом *прогноза* роста требований в течение трех и более лет.

Не стоит задавать вопросы о конкретных технических деталях внедрения, например, какую топологию должна иметь SAN, поскольку это не связано напрямую с бизнесом. Например, в определенных случаях участника технической группы можно спросить: “За какое время должно выполняться резервное копирование?”. Это вполне понятный вопрос для системного администратора, поскольку у него могут быть проблемы из-за того, что он не успевает завершить эту операцию в ночную смену и в результате уменьшается время работы основных бизнес-

приложений. С другой стороны вопрос “На какой скорости должен работать каждый ISL” некорректен поскольку не относится напрямую к потребностям бизнеса. Скорость ISL не всегда непосредственно влияет на скорость резервного копирования<sup>47</sup>. Правильное определение некоторых проблем бизнеса может потребовать проведения дополнительных интервью и совещаний, поскольку многие участники проекта не могут или не хотят отделять технические вопросы от проблем бизнеса.

В большинстве случаев потребность бизнеса в построении SAN можно сформулировать в одном – двух приложениях, включая формулировку проблемы. Вот несколько примеров проблем бизнеса, для которых подходящим решением будет SAN:

- “Слишком долго выполняется резервное копирование. Нужно сократить это время на 50%”.
- “Мы тратим слишком много денег на системы хранения, но они используются неэффективно. Вместо закупки новых массивов нужно повысить эффективность использования уже установленных”.

---

<sup>47</sup>

Например, сравним применение ISL 4Gbit и 10Gbit. Если на вопрос о скорости ISL пользователь сказал, что каждый ISL должен работать на 10Gbit, то, скорей всего, он исходит из желания обеспечить перемещение конкретного объема данных за конкретное время, а не из реальной потребности в 10-гигабитных ISL. Он может не знать, что линки 4Gbit можно объединить в транк и получить хорошо сбалансированный канал 32Gbit, который будет работать намного быстрее, чем 10Gbit. Также возможно с помощью Dynamic Path Selection сформировать из восьми таких каналов сбалансированный канал 256Gbit. Выбирая 10-гигабитную технологию ISL, пользователь отбрасывает более быстрые и дешевые решения высокой доступности. Если архитектор спросит, сколько данных нужно передавать и за какое время, то он может сам выработать техническое решение для реальных проблем бизнеса.

- “Мы тратим слишком много денег на администрирование хранения. Нужно уменьшить численность обслуживающего персонала несмотря на рост объемов данных, которым он управляет”.
- “Законодательство требует, чтобы мы внедрили решения высокой доступности и непрерывности бизнеса”.
- “Последнее время просто случаются все чаще. Необходимо обеспечить функционирование приложений в режиме 24 x 7 так, чтобы резервное копирование не мешало работе пользователей”.

### *Определение требований бизнеса*

Требования бизнеса немного отличаются от проблем бизнеса. Проблема – это вопросы, которые нужно решить, а требования частично определяют сам способ решения. Приведенные выше проблемы бизнеса обычно можно транслировать в ориентированные на бизнес требования к SAN. Например, одну из приведенных выше в качестве примера проблем:

*Мы тратим слишком много денег на системы хранения, но они используются неэффективно. Вместо закупки новых массивов нужно повысить эффективность использования уже имеющихся.*

можно так транслировать в ориентированные на бизнес требования к SAN:

*SAN должна обеспечивать доступ большего числа хостов к любому дисковому массиву и за счет этого улучшить эффективность использования систем хранения на x процентов, что позволит нам сэкономить на покупке новых массивов.*

Такая формулировка проблемы поможет показать, как SAN связана с целями бизнеса. В плане проекта SAN

можно сначала указать проблему бизнеса, а затем требования бизнеса. Например, менеджер проекта SAN может сделать следующую запись в плане проекта после интервью участника проекта по имени Джо, который заявил:

### ***Примечания об интервью Джо***

*Джо считает, что мы тратим слишком много денег на системы хранения, но они используются неэффективно. Некоторые дисковые массивы заполнены только на 20% и тем не менее мы вынуждены покупать дополнительные массивы поскольку серверы, которым не хватает емкости, не могут использовать те массивы, на которых есть свободная емкость. Он говорит, что нужно решение, которое позволит более эффективно использовать установленную емкость вместо покупки новых массивов, поэтому SAN должна обеспечивать доступ большего числа хостов к любому дисковому массиву и за счет этого улучшить эффективность использования систем хранения. Он не знает точную сумму денег, которые мы сейчас тратим на массивы, и я попытаюсь разобраться в этом вопросе. Я считаю, что внедрение SAN будет оправдано если оно обеспечит повышение эффективности использования емкости не менее чем до 80% и за счет этого позволит избавиться от необходимости покупки новых массивов. На основе результатов этого интервью я так сформулировал требования к SAN:*

**Требования:** Обеспечить эффективность использования ресурсов хранения не менее чем 80%, что даст экономию x по покупке новых массивов в течение у лет.

Типичные требования бизнеса к SAN:

- SAN должна обеспечить непрерывность бизнеса в случае катастроф (пожаров, наводнений, ураганов и землятресений).
- Внедрение SAN должно сократить расходы на оплату труда ИТ-персонала, отвечающего за управление хранением на x долларов в год за у лет.
- Построение SAN должно обеспечить выполнение полного резервного копирования не более чем за x часов, причем эта процедура не должна нарушать нормальную работу основных бизнес-приложений что даст у долларов за счет исключения простоев производства.

Все эти заявления носят общий характер. В первом из них нет указаний на то, *в течение какого времени* допускается нарушение работы бизнеса из-за катастрофы и *какие убытки* понесет компания из-за таких простоев. Означает ли непрерывность что не допускается простой в течение одной секунды (минуты, дня)? Какие будут убытки, если не удастся обеспечить восстановление непрерывности за установленное время – один доллар за день или миллион долларов за час? Могут ли быть *человеческие жертвы* из-за простоя системы? (такой вопрос вполне обоснован когда речь идет о военных системах, аварийных службах и медицинских учреждениях). От ответа на любой из этих вопросов может зависеть выбор разных технических решений и структуры бюджета проекта построения SAN, поэтому обязательно нужно собрать *конкретные* требования бизнеса с указанием того, что и когда должно происходить и *каковы будут убытки или ущерб для миссии* если требования не будут выполнены.

## ***Определение технических требований***

После определения требований бизнеса к SAN нужно транслировать их в технические требования, включая сбор информации о технологиях, которые точно будут задействованы в проекте (например, уже установленные приложения и устройства, которые нужно подключить к SAN) и выбор новых технологий, которые нужно внедрить (например, число портов коммутаторов, которые сначала будут подключены и прогнозируемые темпы увеличения числа портов).

### **Идентификация имеющегося оборудования и компонентов**

Необходимо провести обследование уже установленных устройств, которые нужно подключить к SAN. Следует заметить, что новое оборудование существенно сокращает затраты на поддержку (т.е. затраты на поддержку старого оборудования больше, чем стоимость его замены при внедрении SAN)

Если проект SAN предусматривает модернизацию существующей SAN, то инструмент Brocade SAN Health поможет провести обследование имеющегося оборудования. Очень важно правильно выяснить типы оборудования коммутаторов и версии их микропрограмм.

Вопросы о существующем оборудовании нужно включить в интервью участников технической группы, в том числе следующие: есть ли документация с именами хостов, названиями приложений и версиями, установленным оборудованием. Также надо запросить список нового оборудования и программного обеспечения, которое уже выбрано и/или закуплено.

После составления общего списка оборудования и компонентов надо получить подробные сведения о них чтобы при монтаже не возникло неприятных сюрпризов. Ниже дан примерный список вопросов о спецификации

оборудования и компонентов, которые может задать менеджер проекта SAN:

- Какое оборудование (сервера, НВА, устройства хранения, коммутаторы, маршрутизаторы и мосты) установлены к настоящему времени?
- Как сейчас используется оборудование (хосты, НВА, устройства хранения, коммутаторы и т.п.)?
- Какой микрокод установлен на каждом компоненте?
- Какое программное обеспечение установлено на каждом компоненте?
- Есть ли проблемы совместимости этих устройств?
- Какие устройства должны быть связаны напрямую, а не через фабрику? (От этого будет зависеть конфигурация зон.)
- Где расположено оборудование и как организован физический и логический доступ к нему?
- Габариты каждого компонента и расположены ли они в стойке?

#### Идентификация критичных для бизнеса приложений

При проведении интервью нужно определить, какие приложения являются критически-важными для бизнеса и как они используются. Следует понять, что произойдет, если эти системы остановятся. Все связанные с критически-важными приложениями компоненты должны иметь двойное подключение к *физическому* изолированным фабрикам как того требуют принципы проектирования HA SAN. Использование одного шасси директора и разделение его на две зоны, виртуальные фабрики, административные домены, маршрутизация FC, VSAN и аналогичные технологии *не способны обеспечить решение HA*.

Обследование центра обработки данных

Необходимо обследовать помещения ЦОДа и выяснить ограничения пространства, питания, охлаждения и физического доступа, которые должны учитываться при проектировании SAN. Необходимая модернизация или модификация должна быть учтена в бюджете и графике проекта. Если же такие работы не могут быть проведены, то это накладывает серьезные ограничения на проектирование SAN, поскольку выбранное оборудование должно соответствовать возможностям ЦОДа. SAN-Директоры Brocade являются самыми эффективными по энергопотреблению и охлаждению, но даже они нуждаются в ресурсах.

Результатом обследования ЦОДа должно быть получение следующей информации:

- Имя и фамилия человека, который отвечает за ЦОД
- Достаточны ли ресурсы следующей инфраструктуры для внедрения SAN?
  - Питание переменного тока
    - Число доступных электросетей
    - Номинальная мощность каждой электросети
    - Напряжение и тип розеток
    - Защита ИБП
  - Охлаждение
    - Общая мощность
    - Воздушное охлаждение оборудования SAN
  - Сетевые кабели
    - Инфраструктура оптических и медных кабелей
    - Для оптических кабелей – тип коннектора (например SC, ST, LC), диаметр (например 9, 50, 62.5 микрон)

и тип волокна (например SMF, MMF)

- Также рекомендуется провести тесты надежности кабельных соединений и ослабления сигнала, особенно если соединение проходит через несколько патч-панелей или развернуто на большое расстояние.

○ Свободное место в стойках

- Если место в стандартных стойках с 4 опорами или в телекоммуникационных стойках с 2 опорами?

- Нужно записать объем, доступный для размещения оборудования, а не просто общее пространство

○ Грузоподъемность лифта (если монтаж будет производиться не на первом этаже)

○ Есть ли лифт для установки директоров в стойке

### *Разработка расширенной технической спецификации*

После того, как задокументированы требования бизнеса вместе с существующими техническими ограничениями, можно переходить к подготовке технической спецификации SAN, т.е. транслировать требования бизнеса в технологии SAN, которые удовлетворяют эти требования.

Важно понять, что следующие материалы основаны на концепции, которая еще не была полностью объяснена в предыдущих главах. Например, на этом этапе может потребоваться выбрать общую архитектуру SAN высокой доступности, но процедура оценки различных вариантов обеспечения высокой доступности описывает в следующих главах, как и оценка требований к производительности. В этой главе описан только

общий процесс, а его детальное рассмотрение дается в следующих главах книги.

При выполнении этого процесса нужно учитывать требования к SAN, которые возникнут в будущем, поэтому, как уже говорилось выше, обычно рекомендуется разрабатывать SAN в расчете на следующие три-пять лет.

Первый шаг в разработке технической спецификации – это определить физическое местоположение *центров обработки данных*, в которых будет установлено подключенное к SAN оборудование. Например, если SAN должна обеспечить защиту от катастроф, то компоненты должны быть распределены между основным и резервным ЦОДом. В кампусе оборудование может быть распределено между разными зданиями – это самый высокий уровень архитектуры SAN – конечные точки топологии MAN/WAN. Часто архитекторы разрабатывают несколько диаграмм архитектуры для их проекта и начинают с общей диаграммы, показывающей связи между разными географическими частями SAN.

Затем нужно идентифицировать средства, с помощью которых можно соединить между собой эти площадки. Если уже имеется сетевая инфраструктура и ее можно использовать для построения SAN, то надо задокументировать ее характеристики, в том числе полосу пропускания, задержки, коэффициент потери пакетов, доступность и т.п. Во многих случаях требуется вернуться к этому этапу и заново выяснить, сможет ли SAN использовать существующую сетевую инфраструктуру. Если же такая сетевая инфраструктуру *отсутствует* или она не может обеспечить построение SAN, то надо составить список доступных опций. Более подробно этот процесс описан в Главе 11 “Планирование территориальной разнесенности” (стр. 339).

Затем нужно определить, какая полоса пропускания нужна чтобы данные передавались по линиям связи между площадками. Эта оценка позднее будет уточнена, но сейчас она основана на собранных ранее данных, например, требованиям к скорости резервного копирования через SAN. Надо сравнить все характеристики уже имеющейся сети или проектируемой сети с требованиями к пропускной способности и доступности SAN. Эти характеристики должны соответствовать требованиям для SAN, однако многие сети WAN и MAN (особенно IP-сети) не обладают достаточной для передачи трафика систем хранения производительностью и надежностью. Поскольку в большинстве сетей с течением времени нагрузка возрастает, то при оценке нужно учитывать рост трафика SAN и других сетевых приложений, которые используют либо будут использовать ту же инфраструктуру.

Эти шаги помогут менеджеру проекта SAN вместе с архитектором SAN подготовить схему общей архитектуры SAN, на которой будут показаны все сетевые соединения масштаба кампуса, MAN и WAN, в том числе площадки, расстояния между ними, требования к пропускной способности каналов между всеми площадками и информация о сетях, которые могут их соединить. Эта общая схема – вид на SAN «с высоты 3000 метров».

После ее подготовки нужно принять решение относительно модели высокой доступности SAN. Если SAN должна обеспечить высокую готовность критически-важных приложений и/или масштабируемость, то следует использовать резервированную фабрику – этот подход широко применяется многие годы. Можно построить одну SAN с “высоконадежными” директорами и разбить их на

отдельные фабрики, но тогда само шасси будет точкой единичного отказа (например, если это шасси будет залито водой в результате срабатывания расположенной выше системы пожаротушения, то директор выйдет из строя независимо от того, насколько надежно его программное обеспечение). Эти вопросы обсуждаются в “Главе 9: Планирование доступности” (стр. 296).

Нужно выписать, сколько портов устройств (хостов и устройств хранения) будет расположено на каждой площадке и распределить порты коммутатора между этими устройствами. Обязательно надо включить все порты, а не *устройства*. Любое устройство может иметь один, два или несколько портов, подключенных к SAN. Также нужно определить, к какой из резервированных фабрик будут подключены порты, и какой уровень доступности нужен порту устройства от порта коммутатора на другом конце кабеля. Устройства с высокими требованиями к доступности нужно подключать к директорам Brocade 48000 или бэкбону DCX, а менее критичные устройства - к лезвиям директоров с переподпиской или к коммутаторам Brocade. С учетом этих требований архитектор может составить список портов, которые должна иметь фабрика с указанием типа подключенных к ним устройств.

Затем к этому списку надо добавить примерное число новых устройств, которые будут подключены к SAN в обозримом будущем. Это “обозримое будущее” может быть от следующего месяца и до следующих нескольких лет и зависит от того, как часто среда может претерпевать существенные изменения и какая имеется информация для прогнозирования. Если SAN должна сохранять стабильность в течение длительного времени, то лучше выделить порты для новых устройств, что упростит их подключение в будущем. Если же недостаточно данных для прогнозирования роста, то лучше ориентироваться на быстрый рост сети

чтобы в будущем не возникло проблем с масштабируемостью.

Кроме портов для текущих и будущих хостов и устройств хранения в SAN могут потребоваться порты для ISL и IFL (связи между коммутаторами и фабриками). Для SAN, состоящей из сотен и тысяч портов, рекомендуется выделить 10% - 15% портов для этих линков в “типичной” среде клиента. Если же основная часть трафика SAN носит локальный характер<sup>48</sup>, то потребуется меньше линков. Обычно в таких случаях для поддержки отказоустойчивой топологии нужны минимум два линка на коммутатор. В случаях, когда нет локального трафика или для SAN нужно обеспечить высокую производительность, может потребоваться больше линков. У некоторых клиентов до 50% портов выделены для ISL, а у других ISL не используются. В любом случае, нужно учитывать ISL и IFL при расчете числа портов.

К этому этапу архитектор должен иметь достаточно информации чтобы подготовить более детализированный предварительный эскиз SAN, где показаны площадки, соединения между ними, общее число портов, которое требуется на каждой площадке, требования к доступности и производительности этих портов и обычно распределение портов между фабриками (A/B) для обеспечения высокой доступности.

Если фабрики получаются “большими” (не менее тысячи портов), то нужно разбить их на несколько фабрик меньшего масштаба, соединенных через

---

<sup>48</sup> Т.е. устройства хранения расположены рядом со своими хостами. Такой подход улучшает производительность и надежность поскольку трафик не должен проходить через несколько сетей, однако усложняет планирование.

маршрутизаторы с помощью LSAN. Такой подход рекомендуется применять если используются линки масштаба metro или глобальных сетей или за разные части SAN отвечают разные администраторы или некоторые ISL организованы через ненадежную IP-сеть. Нужно модифицировать предварительный эскиз с учетом подключений отдельных фабрик через FC-маршрутизаторы (если это необходимо)<sup>49</sup>.

### ***Оценка стоимости проекта***

После составления технической спецификации SAN нужно оценить стоимость проекта SAN. Хотя на этом этапе подготовлена только предварительная схема SAN, по ней можно составить достаточно точный список всех компонентов и ресурсов, необходимых для построения SAN, включая кабели, системы хранения данных, коммутаторы, маршрутизаторы, HBA, стойки, патч-панели, каналы связи, проектируемое энергопотребление и т.п. В этот список нужно включить оплату услуг специалистов консалтинговых и сервисных компаний, которые могут понадобиться при развертывании. На основе этого списка составляются запросы на коммерческое предложение от вендоров, а эти коммерческие предложения служат основной для расчета стоимости внедрения.

### ***Обоснование проекта (ROI или TCO)***

Архитекторы проекта должны иметь представление о его стоимости и экономическом эффекте. Внедрение SAN будет оправдано в том случае, если первая сумма окажется меньше второй. При подсчете экономического

---

<sup>49</sup>

В других разделах книги описываются использование LSAN и маршрутизаторов. Более подробную информацию можно найти в книге Джоша Джада *Multiprotocol Routing for SANs*.

эффекта SAN надо учитывать как явную выгоду, так и скрытую. В индустрии используются несколько методов для расчета обоснования, из которых самыми популярными являются анализы “Окупаемость инвестиций” (ROI) и “Совокупной стоимости владения” (TCO).

В большинстве случаев анализ ROI или TCO не учитывает все детали – ИТ-персонал должен только доказать руководству, что SAN даст реальный экономический эффект поскольку вся ИТ-индустрия использует модель SAN для проектирования ЦОДов. Приведем такую аналогию – на заре сетевых технологий обоснование внедрения LAN требовало больших усилий, но теперь все знают, что без LAN невозможна работа ИТ, поэтому никто не занимается расчетами ROI или TCO для LAN. SAN также становится обязательным элементом ИТ.

Поскольку SAN сейчас считается обязательной для ЦОДа, то часто обоснование ее внедрения проводится в устной форме или с помощью одного листка бумаги с расчетами. Если же требуется более детальный анализ, то он относится к *выбору* типа SAN, а не к обоснованию пользы от ее внедрения. Если же нужно провести анализ именно для обоснования внедрения SAN, то можно использовать несколько структурированных методов.

## Детальный проект SAN и план внедрения

После утверждения проекта и заказа оборудования архитектор и менеджер проекта SAN должны разработать подробную схему SAN и поэтапный план запуска в эксплуатацию.

На этой схеме нужно показать, где точно должен быть расположен каждый коммутатор, как он должен быть сконфигурирован, как организовано зонирование,

какие порты коммутатора или маршрутизатора должны быть соединены между собой и подключены к узлам (хостам и устройствам хранения). Также полезно на схеме SAN указать прогнозируемый рост SAN. Если какие-то порты или слоты шасси для лезвий зарезервированы на будущее для ISL, то лучше их также показать на схеме чтобы потом не возникло проблем. Также нужно решить вопрос об управлении каждым компонентом SAN. Brocade предлагает различные инструменты для управления коммутаторами и маршрутизаторами. Например, в большинстве сетей с маршрутизацией рекомендуется использовать Brocade Fabric Manager. В любой сети хранения можно использовать Brocade SAN Health для текущего проактивного мониторинга и аудита. На момент написания книги этот инструмент можно было бесплатно загрузить с Web-сайта Brocade.

В большинстве крупных организаций уже formalизованы процессы запуска в эксплуатацию, управления изменениями или контроля над изменениями и план внедрения SAN можно легко «вписать» в этот процесс. В любом случае типичная структура плана развертывания сети имеет следующую структуру:

- 1.Запустить SAN в опытную эксплуатацию и протестировать ее перед запуском.
- 2.После успешного тестирования функциональности запустить SAN в промышленную эксплуатацию. В зависимости от сложности проекта это может потребовать только изменения параметров DNS либо планируемый простой приложения. Главное сократить до минимума риски и простое приложение. Следуйте правилу №1 - “Не допускаются изменения, которые мешают работе любого производственного приложения”. (Правило №2 гласит: “Смотри Правило №1”).

Составьте список всех этапов ввода в

эксплуатацию и связанных с ними рисков, и план восстановления исходных условий на случай если один или несколько рисков приведут к нештатной ситуации.

3. Во время запуска в эксплуатацию нужно снова проверить всю функциональность. Будьте готовы использовать план аварийного восстановления исходных условий на случай, если какая-то функция не работает и не удается сразу исправить ошибку.

4. После внедрения нужно регулярно проводить мониторинг SAN с помощью SAN Health и других средств диагностики и проверять работу сети. Также надо отслеживать обращения в службу поддержки, относящиеся к приложениям, использующим SAN, и проверять, что пользователи этих приложений могут работать эффективно.

Это первый этап в процессе после опытной эксплуатации, о которой говорилось ранее. Его можно опустить если проект небольшой по своим масштабам или приложения, на которые он влияют, не являются критичными. Многие крупные организации применяют специальную тестовую среду для своих приложений, но для небольших компаний это слишком дорого. В качестве примера рассмотрим банк, у которого группа серверов обслуживает базу данных для сети банкоматов в разных странах. Непосредственное изменение в работе этих серверов создает слишком большой риск, поэтому в банке для тестирования изменений базы данных используется выделенная группа серверов. Убытки банка только из-за одной проблемы в несколько раз превысят стоимость серверов для тестирования, поэтому можно легко обосновать необходимость в их покупке. Эти серверы могут часто подключаться к постоянной тестовой SAN, которая похожа по своим характеристикам на основную SAN. Используя маршрутизаторы Brocade можно перемещать данные из основной SAN в тестовую через LSAN для того, чтобы

тестовые серверы использовали новейшие данные, тогда после успешного выполнения тестов изменения можно применить и в производственной среде. В этом случае план проекта SAN должен предусматривать сначала развертывание в тестовой среде, тестирование и запуск в промышленную эксплуатацию.

Применение такого консервативного контроля за изменениями не является специфичным только для SAN – его следует применять при любом изменении ИТ-инфраструктуры если оно может повлиять на уже развернутые бизнес-приложения. В небольших проектах или проектах, затрагивающих только новые приложения, можно использовать более простой процесс запуска, а крупные проекты, затрагивающие критически-важные приложения, требуют более сложного процесса контроля изменений.

# 6

## 6: Планирование топологии

У любой проектируемой сети есть своя топология, т.е. инфраструктура коммутаторов, маршрутизаторов и другого сетевого оборудования. При составлении схемы сети эти элементы образуют различные геометрические фигуры, которые обычно легко идентифицируются и во многом определяют свойства сети. Такие параметры фабрики SAN или Meta SAN, как производительность, доступность и масштабируемость могут зависеть, прежде всего от топологии, поэтому архитектор должен разобраться в доступных опциях и как они влияют на SAN.

Надежность Brocade Fabric Operating System позволяет архитекторам SAN создавать фабрики со сложными топологиями с высокой масштабируемостью и большим радиусом сети. Однако то, что возможно построить сложную топологию еще не означает, что следует всегда стремиться к ней. Следуйте принципу бритвы Оккама<sup>50</sup> при конфигурировании сети – старайтесь спроектировать сеть так, чтобы она была меньше по размерам и проще по архитектуре.

---

50

Уильям Оккам – это средневековый европейский философ, который сказал примерно следующее: “Не должно множить сущее без необходимости”, поэтому если схема сети начинает напоминать тарелку спагетти, то лучше ее переделать ...

Чтобы не усложнять сеть, большинство архитекторов используют только несколько топологий. Обычно это: топология ориентированная на системы хранения, каскадная, кольцо, mesh (связь всех со всеми) и центр/периферия (core/edge, CE). Практически любая проблема архитектуры SAN может быть решена с помощью одной из этих топологий или их комбинации. Например, можно объединить ядра четырех сетей CE в mesh. Такой гибрид CE и mesh очень удобен для катастрофоустойчивых решений.

В этой главе кратко рассматриваются первые три топологии – они определяются и описываются их характеристики. Остальная часть главы (и на самом деле, остальные части книги) практически полностью посвящены топологии CE и ее вариантам, поскольку она чаще всего используется для построения больших сетей.

## **Топология, ориентированная на системы хранения**

Такой подход нельзя считать такой же полноценной “сетевой топологией”, как остальные варианты. Его основа – использование только многопортовых дисковых массивов и построение сети хранения из директоров и коммутаторов, подключенных к портам массива, но *не между собой*, поэтому такая архитектура не предусматривает использование ISL и IF L. Система хранения данных может направлять LUNы на любые фабрики, перемещая их между портами таким образом, что любой сервер подключенный к любой фабрике может обратиться к любому LUN.

Эта топология имеет несколько преимуществ – она достаточно надежна, быстро работает и простая, однако не обеспечивает масштабируемость и гибкость. Например, если в будущем понадобиться подключить 2-портовый RAID- массив или JBOD, то такая

сетевая модель не позволит это сделать, как и масштабироваться больше того числа портов, которыми оборудован массив. Также невозможно для хостов в этих фабриках напрямую обращаться к централизованной катастрофоустойчивой SAN или SAN, к которой подключены ленточные системы. Хостам потребуется три и больше НВА для подключения к этим сервисам.

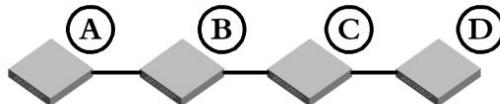
Разумеется, эти ограничения не означают, что такой подход неправильный, просто при его применении нужно учитывать как плюсы, так и минусы (как при любом проекте построения сети). Если же используется именно этот подход, то рекомендуется соединение фабрик между собой маршрутизаторами для обеспечения подключений за пределы фабрики.

## Топология каскада

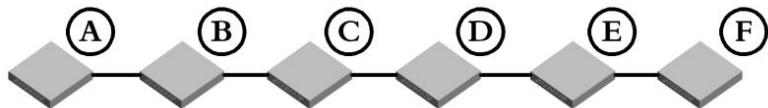
Каскадная сеть состоит из группы коммутаторов, соединенных цепочкой. Каждый коммутатор в середине цепочки соединен с коммутаторами, которые расположены слева и справа от него, а коммутаторы на обоих концах цепочки соединены только с одним коммутатором внутри цепочки. Можно использовать один или несколько ISL между соседними коммутаторами. На Рис. 28 показаны два примера каскадных сетей.

Такая архитектура не обладает масштабируемостью, не обеспечивает высокой производительности и не отличается высокой надежностью. Для масштабирования SAN нужно добавить коммутаторы к на концах цепочки, что увеличивает число переходов между концами сети и создает риск запаздывания пакетов и перегруженности сети. Другими словами, производительность сети *падает* по мере *добавления* в нее новых устройств и коммутаторов. Это полностью

противоречит целям построения SAN, которая должна улучшить производительность.



**Four-Switch Cascade**



**Six-Switch Cascade**

**Рис. 28 – Каскадная топология. Каскады из четырех и шести коммутаторов**

Эту проблему можно решить, добавляя ISL между коммутаторами по мере подключения к фабрике новых коммутаторов. Например, если сеть A → D на Рис. 28 расширена до A → F, то системный администратор может добавить ISL между A и B, B и C, и C и D при подключении к фабрике коммутаторов E и F.

Однако такое добавление ISL внутрь цепочки при подключении новых коммутаторов связано с большими расходами и ограничивает масштабируемость, поэтому на практике каскадные сети никогда не применяются поскольку по мере роста их производительность падает.

Снова рассмотрим Рис. 28. В SAN, показанной в верхней части этой иллюстрации, коммутаторы A и D могут «переговариваться» между собой, используя для этого ISL AB, BC или CD. Если же коммутаторы E и F добавлены на концах цепочки, то между собой смогут передавать данные коммутаторы B и E, C и F. Отметим,

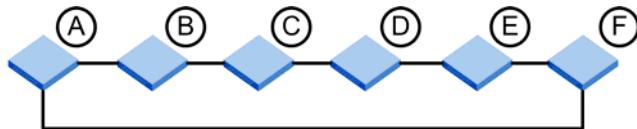
что у всех этих трех «разговоров» разные начальная и конечная точка, поэтому они должны быть независимыми между собой. Однако разговор B-E использует те же два ISL, как уже существующий A-D: оба идут между BC и CD. Аналогичным образом разговор C-F использует CD и DE, накладываясь на два других. Таким образом, в каскадной сети независимые разговоры конкурируют за ограниченную полосу пропускания.

Кроме того, если выйдет из строя коммутатор в середине сети, то сеть распадется на отдельные сегменты, т.е. этот коммутатор является точкой единичного отказа. Если выйдет из строя коммутатор D или ISL C между D, то это нарушит не только разговор A-D, но и два остальных. Добавление внутренних ISL не способно решить эту проблему.

Из-за этих недостатков архитекторам не следует использовать каскадное подключение если важны производительность, доступность и масштабируемость. Тем не менее, каскады остаются базовой топологией, которую можно просто и с небольшими затратами внедрить, поэтому они годятся для самых маленьких фабрик, прежде всего для сетей только из двух коммутаторов. В таких небольших фабриках перечисленные выше проблемы не могут возникнуть поскольку архитектура с двумя коммутаторами является самой упрощенной версией каскадной топологии. Ее можно также использовать когда трафик носит в основном локальный характер (стр. 258) и идет внутри отдельных коммутаторов, а ISL используются для управления и/или второстепенных и редко запускаемых приложений.

## Топология кольца

Кольцо похоже на каскад, но в нем конечные точки соединены между собой. На Рис. 29 показана топология кольца SAN с шестью коммутаторами.



Кольцо из 6 коммутаторов

Рис. 29 – Топология кольца

У колец важное преимущество по сравнению с каскадами – они реализуют альтернативный маршрут и если выйдет из строя коммутатор или линк в кольце, то трафик будет передаваться по кольцу в обратном направлении. Обычно устройства, подключенные к коммутатору A, обращаются к устройствам, подключенными к коммутатору E через F. При отказе коммутатора F трафик пойдет через коммутаторы C, B и D<sup>51</sup>. Если сравнить ситуацию с тремя ”разговорами” из предыдущего раздела, которые прерываются при отказе коммутатора D, то в данном случае два из них не прерываются благодаря использованию альтернативного маршрута.

Помимо улучшения доступности, кольца улучшают производительность. В использующей механизм FSPF (Fabric Shortest Path First) сети Fibre Channel трафик всегда идет между коммутаторами в кольце по самому короткому из доступных маршрутов. Например, устройство, подключенное к коммутатору A, будет разговаривать с устройством, подключенным к коммутатору F, через ISL AF, а не через маршрут,

<sup>51</sup>

При этом кольцо в превращается в каскад.

который связывает коммутаторы В и Е. В каскадной сети этот линк отсутствует и поэтому нельзя выбрать кратчайший маршрут, поэтому в среднем у колец меньше хопов, чем у каскадов. Хотя во многих отношениях кольца работают лучше каскадов, у обоих топологий есть общие проблемы, например, при добавлении в кольцо коммутатора увеличивается число хопов, что ухудшает производительность и надежность. Хопами (hop) называют расстояние, равное числу последовательных связей между коммутаторами или фабриками.

На самом деле в определенных отношениях кольца даже хуже каскада. Например, установка нового коммутатора приводит к необходимости временно отключить часть кольца. Если потребуется установить седьмой коммутатор в сеть, показанную на Рис. 29 и сохранить топологию кольца, то придется отключить один ISL, что неизбежно повлияет на все операции ввода/вывода, которые идут через этот линк. Например, если коммутатор устанавливается на дальней стороне F, то нужно временно отсоединить ISL между А и F и нарушить весь идущий через него трафик.

Как и каскадные сети, кольца подходят для небольших проектов, где трафик в основном носит локальный характер. Если же в SAN больше четырех доменов (коммутаторов), то следует использовать другую топологию. Кольца также можно использовать для некоторых приложений MAN/ WAN, где топология уже существующей MAN/ WAN определяет топологию внедряемой на ее основе SAN.



## Заметки на полях

*Топология кольца была популярна на начальном этапе развития сетей передачи данных. Протоколы ARCNet, FDDI и Token Ring были основаны на топологии кольца. Однако затем стали очевидны серьезные недостатки этой топологии и даже эти три протокола были адаптированы к топологии звезды. (Топология звезды является основой сетей центр/периферия.)*

## Топология mesh

Топология mesh предусматривает соединение каждого коммутатора со всеми остальными коммутаторами сети используя по крайне мере один ISL на соединение<sup>52</sup>. На Рис. 30 показана mesh из шести коммутаторов. Mesh решает многие проблемы каскадов и колец – расстояние до любого коммутатора составляет только один хоп, поэтому добавление новых коммутаторов в mesh не увеличивает число хопов между коммутаторами. Установка дополнительных коммутаторов не нарушает работу сети поскольку они устанавливаются не между уже соединенными коммутаторами. Можно использовать альтернативные маршруты, причем в отличие от каскада, их несколько, поэтому mesh обладает высокой отказоустойчивостью.

---

<sup>52</sup>

На самом деле, это определение соответствует топологии full mesh (все со всеми). В частичном mesh часть ISL отсутствует. Однако в частичном mesh трудно спланировать производительность и масштабируемость, поэтому такую топологию не рекомендуется использовать.

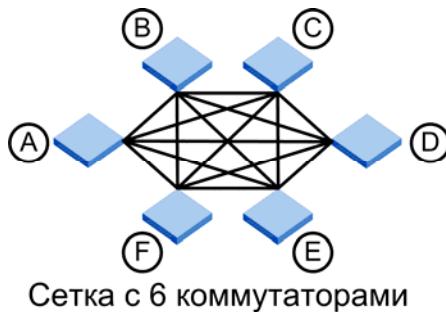


Рис. 30 – Топология mesh

К сожалению, у топологии mesh есть и свои существенные недостатки.

- Стоимость. Даже в небольшой mesh слишком много ISL по отношению к числу полезных портов и по мере расширения сети число доступных портов коммутаторов сокращается.
- Из-за ограничений, указанных в предыдущем пункте, mesh нельзя масштабировать – при добавлении нового коммутатора расходуется по одному порту на всех остальных и возникает проблема экспоненциального роста числа ISL. Для соединения через ISL коммутаторов в mesh из двух коммутаторов требуется в сумме два порта, трех коммутаторов – шесть портов, а четырех коммутаторов – двенадцать, пяти коммутаторов – двадцать портов и т.д. В результате mesh *быстро* становится *крайне невыгодной* по стоимости и ее расширение уже не имеет смысла после того, когда половина портов коммутаторов выделено для ISL. Например, из 16-портовых коммутаторов можно создать mesh, состоящий максимум из восьми доменов. Если в такую сеть добавить еще один коммутатор, то число свободных портов коммутаторов сети уменьшится.
- Производительность. Хотя у каждого коммутатора половина портов может использоваться для ISL,

между любыми двумя доменами есть только один маршрут. Для передачи трафика от коммутатора А к коммутатору В на Рис. 30 можно использовать только ISL AB. Если подключенные к коммутатору А пять устройств обращаются к устройствам, подключенными к коммутатору В, то четыре остальных линка коммутатора А никогда не будут использоваться<sup>53</sup>. Можно построить дополнительные ISL между коммутаторами А и В, но это еще больше ухудшит масштабируемость и если другие ISL никогда не используются, то непонятно, зачем они нужны. Чтобы все ISL в mesh использовались, трафик должен идти между несколькими точками и быть равномерно распределенным. Это последнее требование редко встречается в реальных SAN.

Из-за этих ограничений имеет смысл использовать full mesh только в фабрике с четырьмя доменами. Однако, если в каждом домене 384-портовый директор, то получится достаточно большая фабрика. Mesh также часто используется при построении небольших MAN/WAN, в которых каждая *площадка* является точкой mesh. Часто используется гибридная топология - CE внутри каждой площадки и mesh или кольцо для соединения площадок. Такой подход применим когда имеется небольшое число площадок, а большие MAN/WAN обычно используют частичный mesh или варианты CE.

---

<sup>53</sup> Если только не нарушится работа ISL AB. В этом случае все остальные четыре маршрута из А в В будут равны по длине и нагрузка распределится между ними. Такая абсурдная ситуация (обрыв линка приводит к улучшению производительности) говорит о неэффективности частичного mesh, из-за чего топология редко применяется на практике.

## Топология центр/периферия

Топология центр/периферия (core-to-edge, CE) – это эволюция популярной в мире сетей передачи топологии “звезды”. На Рис. 31 и Рис. 32 показана отказоустойчивая фабрика (стр. 308) CE Fibre Channel. Коммутаторы нижнего уровня фабрики – это “коммутаторы центра”, а соединенные через них коммутаторы называются “границочными”. На Рис. 31 границочными коммутаторами являются А - D, а коммутаторами центра – Е и F.

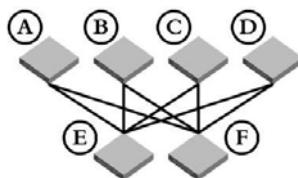


Рис. 31 – Топология СЕ

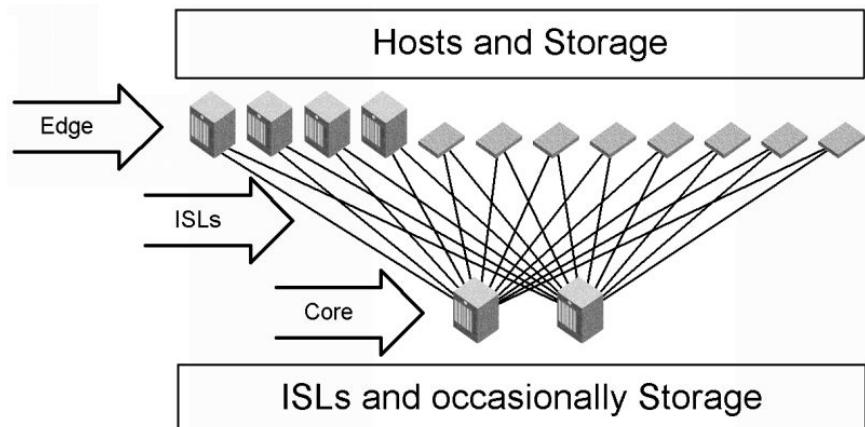


Рис. 32 – Простая отказоустойчивая фабрика центр / периферия

Топология СЕ стала основной для архитектуры SAN по нескольким причинам:

- Она хорошо протестирована, поскольку большинство применяемых фабрик, а также тестовые лаборатории Brocade и другие

лаборатории, занимающиеся тестированием SAN, используют топологию core / edge в той или иной форме.

- Она хорошо сбалансирована – ее симметричность обеспечивает балансировку нагрузки и резервирование. Трафик между граничными коммутаторами можно распределять между всеми центральными коммутаторами.
- Детерминированность. Скорость передачи данных между двумя граничными коммутаторами никак не влияет на скорость между любыми двумя другими коммутаторами. Например, для передачи трафика между А и В на Рис. 31 никогда не будут использоваться те же ISL, по которым идет трафик между С и D.
- Экономичность – есть опции для разных соотношений цены и производительности. Если пользователям нужно обеспечить высокую производительность конкретного граничного коммутатора, то он может добавить ISL центр/периферия только к этому коммутатору без дополнительных расходов в расширение SAN.
- Легко адаптировать и модифицировать сеть поскольку каждый центральный коммутатор можно дублировать и граничные коммутаторы взаимозаменяемы.
- Простая для понимания, документирования и устранения сбоев. В отличие от частичного mesh в топологии СЕ легко разобраться.
- Простое масштабирование без нарушения работы сети при подключении новых граничных коммутаторов к свободным портам центральных коммутаторов либо постепенной заменой центральных коммутаторов на модели с большим числом портов.

Хотя другие топологии используются в небольших

SAN, практически все крупномасштабные внедрения основаны на вариантах СЕ.

### ***Оптимизация производительности фабрики СЕ***

Одно из отличий фабрики СЕ от традиционной “звезды” – это то, что обычно сети СЕ имеют два и больше коммутаторов ядра для улучшения отказоустойчивости и производительности, а у сети с топологией “звезда” в центре только один коммутатор или концентратор. В сети Ethernet несколько коммутаторов в центре звезды обычно работают как tandem active/passive с использованием протокола Spanning Tree Protocol (STP). STP не обеспечивает никаких преимуществ с точки зрения производительности и при сбоях восстановление с помощью этого протокола занимает несколько минут, поэтому резервирование не дает существенного улучшения надежности.



### **Заметки на полях**

*Все топологии могут деградировать и превращаться в топологии другого типа. Например, full mesh с двумя коммутаторами является также и каскадом с двумя коммутаторами и кольцом с двумя коммутаторами. Full mesh с тремя коммутаторами является кольцом с тремя коммутаторами и треугольником. Если в кольце с четырьмя коммутаторами разорвать один ISL, то он превратится в каскад с четырьмя коммутаторами, а если выйдет из строя один из коммутаторов, то получиться каскад с тремя коммутаторами. Практически любую топологию можно назвать частичным mesh.*

*Классификация топологий полезна при определении возможного поведения сетей, однако отсутствуют четкие границы между разными типами сетей. Из приведенных далее в книге схем видно, что сети могут включать в себя разные топологии либо менять свою топологию при выходе из строя отдельных элементов.*

С другой стороны, фабрики используют FSPF, что обеспечивает балансировку нагрузки active /active (стр. 272) между двумя маршрутами с одинаковым числом хопов. Все маршруты между граничными коммутаторами имеют одинаковое число хопов, а именно два, поэтому фабрика может использовать все ISL, а не только часть из них (как в случае полной сетки). FSPF также обеспечивает быстрое восстановление после сбоев.

В результате сети СЕ, построенные на базе платформ Brocade, с точки зрения производительности являются "самообслуживаемыми", однако если необходимо получить максимальную производительность, то желательно провести дополнительный тюнинг. Оптимизация осуществляется несколькими способами, которые подробно описаны в "Глава 8: Планирование производительности" (стр. 227 и далее).

Есть два основные варианта оптимизации на основе подключения узлов:

С точки зрения производительности лучший вариант - локализация "горячих" подключений внутри граничных коммутаторов, в результате которой основная часть трафика не должна проходить через ISL.

Второй подход основан на связывании хостов и устройств хранения. Он гарантирует балансировку трафика между доступными ISL. Хотя связывание не дает такой высокой производительности, как

локализация, этот подход проще внедрить, он облегчает визуализацию и управление.

Многие клиенты выбирают средний путь – они связывают большинство соединений, но локализуют критически-важные хосты.

### **Масштабируемость топологии СЕ**

В отказоустойчивой фабрике СЕ два и более коммутаторов центра соединяют много граничных коммутаторов. Обычно свободные порты коммутаторов центра зарезервированы только для ISL и IFL для обеспечения максимальной масштабируемости и узлы подключаются к граничным коммутаторам. Порты, к которым подключаются эти узлы, обычно называются “граничными портами” или “пользовательскими портами” чтобы отличать их от портов, которые используются для ISL/IFL.

На Рис.33 показано, как на масштабируемость влияет добавление узла в фабрику СЕ. На обоих схемах SAN построена с помощью 16- портовых коммутаторов и соотношение числа хостов и устройств хранения составляет около 7:1, а коэффициент переподписки (oversubscription) ISL - 7:1<sup>54</sup>. Однако на левой схеме устройства хранения подключены к границе, а на правой – к ядру, поэтому показанная слева архитектура может масштабироваться до 224 устройств пока у коммутаторов

---

<sup>54</sup>

Здесь может стоять любое соотношение если оно используется в обоих вариантах. Например, если для обоих примеров переподписка ISL была бы 3:1, то результаты были бы другими, однако *соотношение* осталось бы прежним. Аналогичным образом применение более мощных коммутаторов улучшило бы масштабируемость, но расположение устройств будет по-прежнему влиять на максимальное число.

ядра остаются свободные порты для ISL<sup>55</sup>, а архитектура с подключением к ядру – только до 24 узлов хранения при максимальном числе портов SAN равным 80<sup>56</sup>.

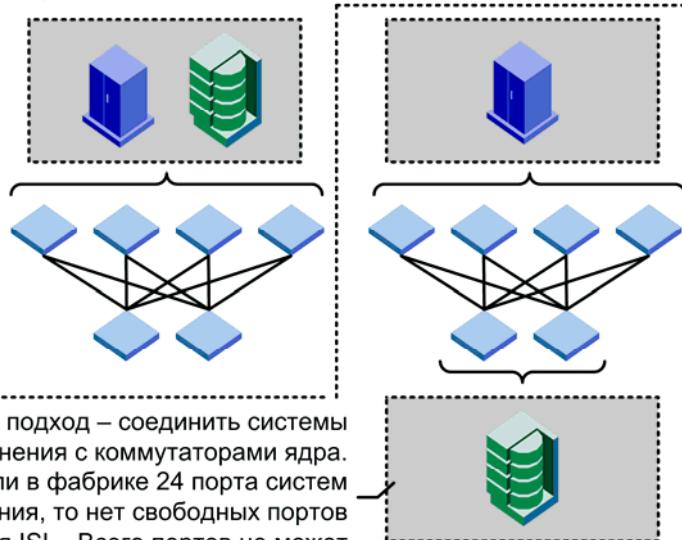
Этот пример доказывает ограничение масштабируемости при подключении к ядру. Если подключать устройство хранения к порту коммутатора ядра, то общее число узлов увеличивается всего на одно *устройство хранения*, а если подключать к этому порту линк ISL, то масштабируемость SAN увеличивается на целый *коммутатор*. Чем мощнее коммутатор границы и больше коэффициент переподписки ISL, тем значительнее этот эффект.

---

<sup>55</sup> 16 коммутаторов периферии по 12 свободных портов дают 192-портовую фабрику. При соотношении fan-out 7:1 это дает 24 устройства хранения и 168 хостов.

<sup>56</sup> У 11 коммутаторов периферии есть 66 пользовательских порта на периферии для хостов и 10 пользовательских портов в центре для устройств хранения. Соотношение 66:10 – это максимально близкое к 7:1, которое может обеспечить эта архитектура.

Один подход – подсоединить все устройства к граничным коммутаторам. Если SAN построена из 16-портовых коммутаторов с одним ISL от каждой граничного к каждому коммутатору ядра, то на каждом коммутаторе ядра свободно 12 портов для подключения дополнительных граничных коммутаторов. Всего фабрика может масштабироваться до 224 портов.



Другой подход – соединить системы хранения с коммутаторами ядра. Если в фабрике 24 порта систем хранения, то нет свободных портов для ISL. Всего портов не может быть больше 80.

**Рис. 33 – Масштабируемость в зависимости от расположения устройств**

Если нужно подключение именно к центру, то есть эффективное решение – узлы можно подключать к центру, если центр состоит из директоров с большим числом портов, например, Brocade 48000 или DCX Backbone. В этом случае архитектурные особенности этой топологии не будут существенно ограничивать масштабируемость, которая будет определяться прежде всего масштабируемостью сервисов фабрики.

Прежде чем выбирать стратегию нужно оценить, какую она обеспечит производительность и

масштабируемость. Если хосты равномерно распределяют ввод/вывод между портами хранения, а последние равномерно распределены между центральными коммутаторами и/или есть только один центральный коммутатор, то производительность обычно выше при подключении устройств хранения к *центру*. Если же нет такой балансировки ввода/вывода (что является типичной ситуацией) или центральных коммутаторов несколько, то при подключении устройств хранения к периферии производительность будет выше. В случае сомнения относительно выбора следует подключить все узлы к коммутаторам периферии и использовать коммутаторы центра только для ISL/IFL.



## Заметки на полях

*В коммутаторах Brocade 200E, 4100, 4900, 5000, 7500, 7600, 48000 и последующих применяется функция Dynamic Path Selection (DPS), которая значительно улучшает балансировку нагрузки в сетях СЕ. (стр. 272) Транкинг на основе фреймов работает только на портах, принадлежащих одной микросхеме ASIC, поэтому он не может распределять трафик между разными коммутаторами ядра. DPS может работать на портах, соединяющих разные ASIC и балансировка нагрузки между ядрами выполняется на аппаратном уровне.*

## *Гибридные топологии СЕ*

Реализовать СЕ на практике можно разными путями.

Например, можно по-разному комбинировать коммутаторы с разным числом портов и характеристиками НА. Одни архитекторы используют директоры с большим числом портов в ядре, а на границе – небольшие коммутаторы для fan-out, другие архитекторы устанавливаются директоры и на

границе, а трети применяют только коммутаторы с небольшим числом портов.

Также можно варьировать характеристики НА фабрики. Многие архитекторы SAN развертывают полностью резервированную фабрику (стр. 310) с хостами и устройствами хранения, которые одновременно подключены по крайней мере к двум изолированным сетям. Некоторые архитекторы считают, что обеспечивают достаточное резервирование и поэтому не нужно резервировать ядра в каждой фабрике. При таком подходе каждая фабрика может иметь только один коммутатор ядра, как показано на Рис. 34. Он является единой точкой отказа, однако если вся фабрика откажет, то приложения смогут обращаться к своим устройствам хранения через другую фабрику.

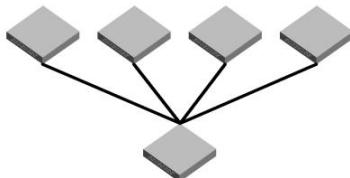


Рис. 34 – Неотказоустойчивая фабрика СЕ

Стоит отметить, что такой подход не считается оптимальным для НА SAN – это компромисс, при котором для снижения стоимости идут на увеличение рисков. Как объясняется в Главе 9: Планирование доступности”, если несколько хостов могут одновременно стать недоступны, то это увеличивает вероятность сбоя.

Некоторые архитекторы делают ставку только на отказоустойчивость для фабрики А, но не фабрики В. Этот вариант лучше, чем две неотказоустойчивые фабрики, особенно если сконфигурировать драйверы multipathing на предпочтительный путь через А. Аналогичным образом, если локальность высокая и/или

у разных узлов разные требования к НА, то архитекторы могут вообще не использовать СЕ для фабрики В. На Рис. 35 показан пример такой архитектуры.

В этом примере одним устройствам требуется доступ НА к SAN, а другим - нет. Все устройства подключены к масштабируемой фабрике СЕ, “фабрике А”, но те из них, которым *требуется* НА, имеют резервное подключение к физически изолированной фабрике В. Поскольку таких устройств меньше, чем устройств без требований НА, то фабрика В меньше фабрики А и состоит только из одного коммутатора.

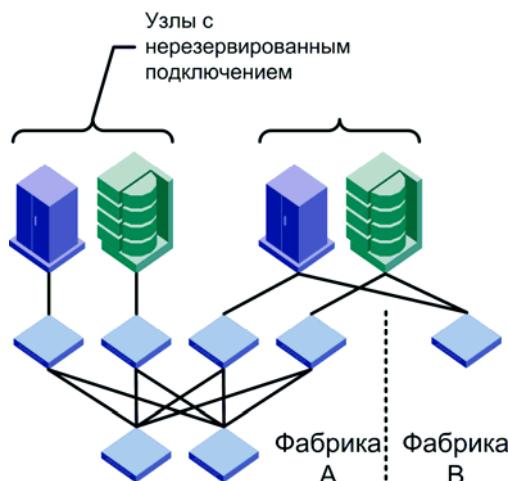


Рис. 35 – Асимметричная резервированная фабрика А/В

Другой вариант решения – создать две одинаковые по масштабу фабрики СЕ, между которыми распределены нерезервированно-подключенные устройства. Этот вариант чаще применяется на практике и обеспечивает более высокую масштабируемость SAN (см. Рис. 36).

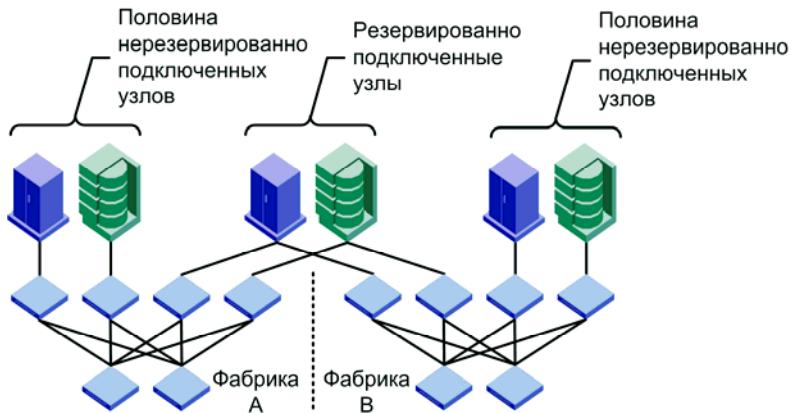


Рис. 36 – Резервированные фабрики с системой НА и не-НА

Таким образом, есть разные варианты применения топологии СЕ с обеспечением необходимых характеристик производительности, надежности и т.п. Архитекторы SAN, выполняющие крупномасштабный проект, обычно начинают с архитектуры СЕ и затем модифицируют ее в зависимости от конкретного внедрения.

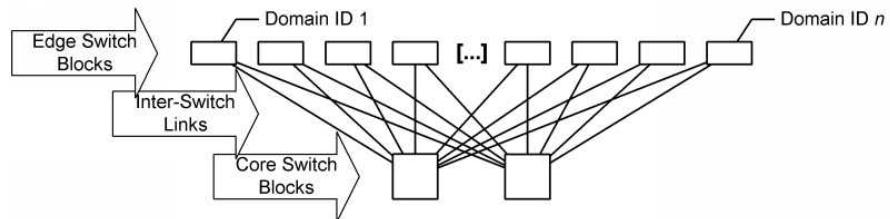
## Meta SAN центра/периферии

Как было объяснено в предыдущем разделе, топология СЕ очень выгодна для SAN, состоящей из фабрик Fibre Channel. Этот подход с некоторыми изменениями можно применять и для SAN на базе iSCSI, хотя ограничения протокола Ethernet снижают эффект. Все преимущества будут реализованы если с помощью функции F C-to-FC router (FCR) Brocade AP7420, 7500 или лезвия FR4-18i создать Meta SAN<sup>57</sup>, в которой LSAN охватывают эти фабрики. В результате,

<sup>57</sup>

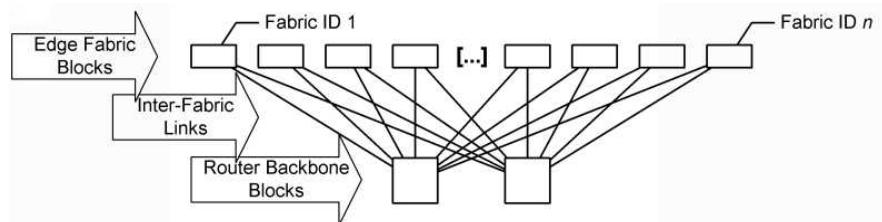
См. “Сравнение фабрики с SAN и Meta SAN” на стр. 42 где подробнее рассматривается эта категория сетей.

маршрутизируемые Meta SAN обычно создаются на базе топологии CE.



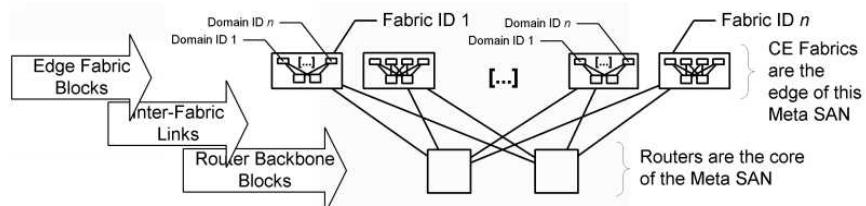
**Рис. 37 – Общая схема блоков фабрики CE**

Рис. 37 показывает общую диаграмму блоков фабрики CE, а Рис. 38 – аналогичную диаграмму CE Meta SAN. Отметим, что блоки периферийных фабрик на второй диаграмме могут содержать любые корректно построенные фабрики, в том числе и фабрики CE, поэтому CE Meta SAN может содержать  $n$  фабрик CE или других фабрик.



**Рис. 38 - Общая схема блоков CE Meta SAN**

В схеме Рис. 38 как опция может быть  $n$  копий Рис. 37 (см. Рис. 39).



**Рис. 39 - Обычная CE Meta SAN с фабриками CE на периферии**

Обычно любую топологию, которую можно использовать в фабрике, можно использовать и в Meta SAN. Более подробно проектирование Meta SAN описано в книге *Multiprotocol Routing for SANs*, выпущенной в серии Brocade *SAN Administrator's Bookshelf*.

## Топология встроенных коммутаторов

Brocade предлагает ряд платформ коммутации, которые встроены или устанавливаются непосредственно в шасси блейд-серверов или системы хранения.

Общая архитектура SAN со встроенными коммутаторами в шасси блейд-серверов мало отличается от любой другой архитектуры SAN. На Рис. 40 показан пример такой архитектуры.

В этом примере каждое шасси блейд-серверов оборудовано двумя встроенными коммутаторами. Для резервирования каждый коммутатор подключен к отдельным фабрикам. Каждый блейд-сервер оборудован двухпортовым контроллером НВА, подключенным к каждому встроенному коммутатору через backplane шасси. В результате каждый блейд-сервер получает доступ к устройствам хранения каждой фабрики как если бы хост с двумя НВА был подключен к паре граничных коммутаторов из разных фабрик.

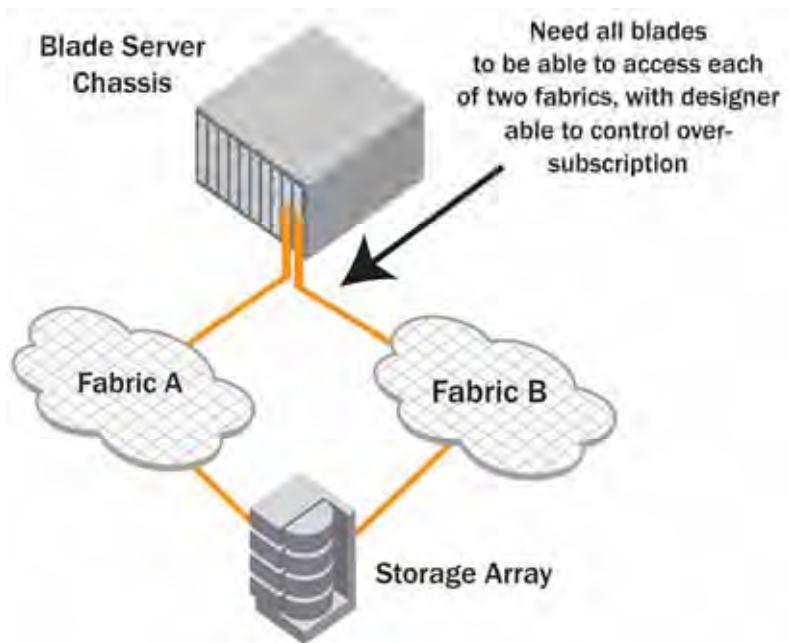


Рис. 40 – Общая архитектура SAN со встроенными коммутаторами.

В некоторых компаниях (особенно из малого и среднего бизнеса) конечный пользователь даже не догадывается, что используются коммутаторы. Например, если у заказчика есть шасси блейд-серверов со встроенным коммутатором, то можно напрямую подключить порты коммутатора к устройствам хранения. В этом случае фабрики А и В будут полностью интегрированы в шасси блейд-серверов. Если заказчик использует маскирование LUN для устройств хранения, то нет смысла напрямую управлять коммутаторами. В таких случаях конечному пользователю даже не потребуется назначать IP-адреса для управления коммутаторами.

Однако при использовании этих продуктов в корпоративных SAN необходимо определить их как коммутаторы и соответствующим образом

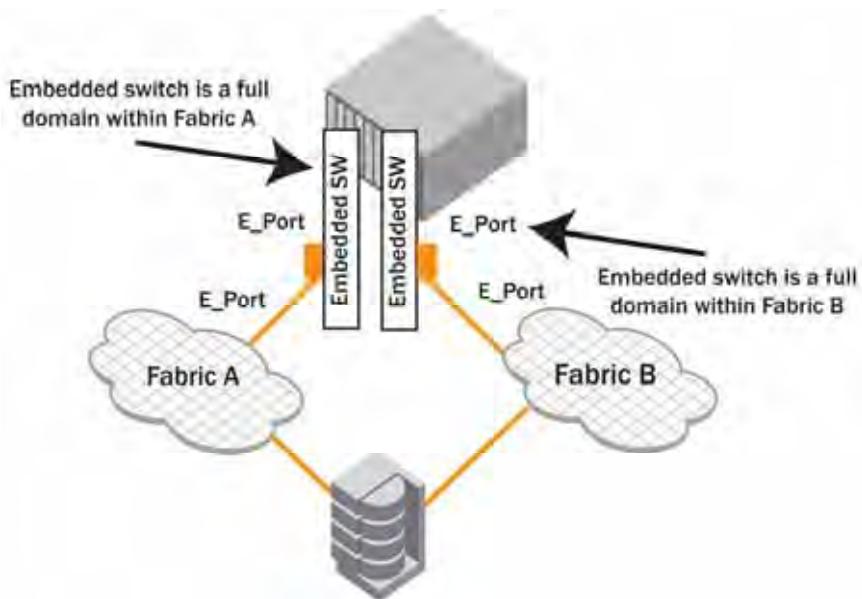
построить и управлять фабриками. В средних по размеру фабриках (состоящих из около 10 доменов) встроенные коммутаторы можно рассматривать как периферийные в простой архитектуре СЕ. В этом случае порты встроенных коммутаторов можно использовать как ISL, соединяющие один или несколько центральных коммутаторов.

В более крупных инсталляциях такое использование встроенных коммутаторов ограничивает масштабируемость. Например, если фабрика в будущем будет поддерживать 100 шасси блейд-серверов (каждое с 10 лезвиями), то для хостов потребуется 1000 записей в сервере имен плюс еще записи об устройствах хранения, что не создает проблем с поддержкой. Однако если фабрика будет включать в себя более 100 доменов коммутаторов, то управление резко усложнится.

На Рис. 41 показано, как шасси блейд-серверов с двумя встроенными коммутаторами обычно подключается к паре внешних фабрик. В этом примере каждый встроенный коммутатор образует один или несколько ISL в своей фабрике через подключение E\_Port. Коммутатор полностью используется в сервисах своей фабрики (как и любой другой коммутатор) и поэтому напрямую влияет на масштабируемость домена коммутатора.

На первый взгляд использование оптических модулей pass-through устранит домены, но на самом деле не решит проблему, а только переведет проблему масштабируемости в другую плоскость, поскольку НВА, через которые выполняется подключение, все равно надо подсоединить к коммутаторам. При этом коммутаторы будут расположены вне шасси, т.е. займут дополнительное место в стойке, а также увеличат расход электроэнергии, охлаждения и кабелей. На самом деле, появится новая проблема из-за того, что управление

кабелями при использовании оптических модулей pass-through – очень сложная процедура.



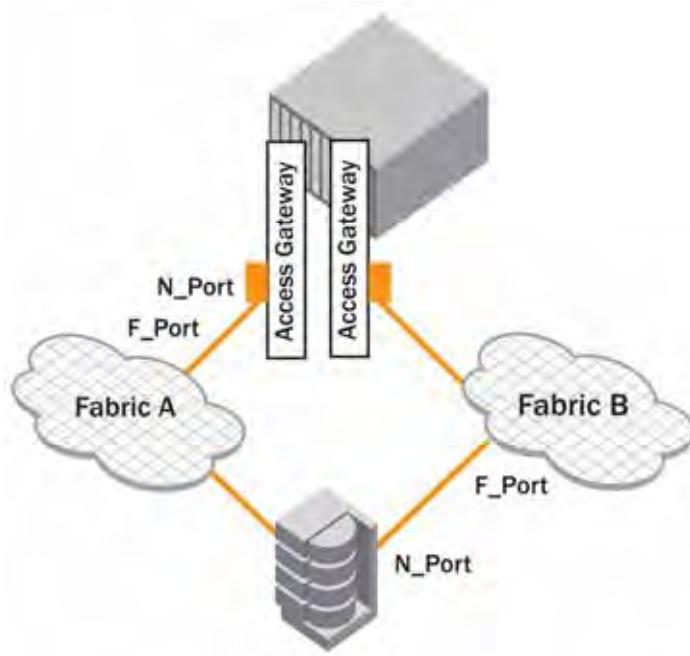
**Рис. 41 – Подключение E\_Port встроенных коммутаторов**

Имеется два способа решения этой проблемы с использованием корректной архитектуры: (1) Разбить фабрику на несколько фабрик и, как опция, соединить эти фабрики маршрутизаторами FC, и (2) использовать продукт Brocade Access Gateway для того, чтобы встроенные коммутаторы в шасси блейд-серверов выглядели бы как отдельные хосты, *а не* домены коммутаторов.

В некоторых случаях в зависимости от требуемых подключений хорошо подойдет вариант с маршрутизаторами, однако если требуется подключение любого хоста к любому порту систем хранения, которое не ограничивается рамками фабрики, то этот вариант не даст существенного улучшения масштабируемости,

поскольку маршрутизатор должен “спроектировать” все периферийные фабрики со встроенными коммутаторами шасси блейд-серверов в одну централизованную фабрику хранения. На самом деле, это может даже привести к *снижению* масштабируемости поскольку у маршрутизаторов могут возникнуть проблемы масштабируемости из-за числа граничных фабрик в дополнение к проблеме масштабируемости централизованной фабрики хранения. Маршрутизаторы великолепно подходят для масштабируемости и изоляции сбоев, но только при условии *использования локализации трафика на уровне фабрики*, что редко удается обеспечить для встроенных коммутаторов.

К счастью, у Brocade есть продукт, решающий эту проблему. Access Gateway был выпущен вместе с Fabric OS 5.2.1b. С помощью NPIV (N\_Port Id Virtualization) он делает встроенные коммутаторы невидимыми для фабрики, но в то же время обеспечивает легкое подключение к SAN хостов-лезвий.



**Рис. 42 – Подключение встроенных коммутаторов с помощью NPIV связи.**

Рис. 41 и Рис. 42 очень похожи. Основное отличие Рис. 42 – это использование портов F\_Ports вместо E\_Ports. Внедрение Access Gateway практически не отличается от построения простой CE, которое было описано выше – нужно соединить каждый “коммутатор” Access Gateway непосредственно к ядру фабрики как если бы Access Gateway был граничным коммутатором в сети CE.

Ключевое отличие Access Gateway в том, что он не подключается через E\_Ports и не обеспечивает сервисы фабрики. Эти сервисы предоставляет коммутатор фабрики, подключенный к каждому Access Gateway, а Access Gateway подключается к коммутатору через соединение, которые выглядят как соединение с помощью НВА. Запросы на сервисы от встроенных

HBAs (например, к сервису *Na me Server*) через Access Gateway передаются коммутатору фабрики, поэтому нужно запускать полный набор протоколов фабрики switch-to-switch для обеспечения взаимодействия Access Gateway с фабрикой. Более подробно соединение NPIV показано на Рис. 43.

Эта архитектура позволяет развертывать много дополнительных серверов без добавления доменов и связанной с этим перестройкой (rebuild) фабрики, что происходит всякий раз когда у коммутатора включается питание, коммутатор добавляется или извлекается из фабрики. Иными словами, она обеспечивает все необходимые соединения без запуска сервисов, обычно ограничивающих масштабирование. Кроме того, она снимает большинство проблем несовместимости коммутаторов, которые раньше не позволяли смешивать устройства FC разных вендоров.

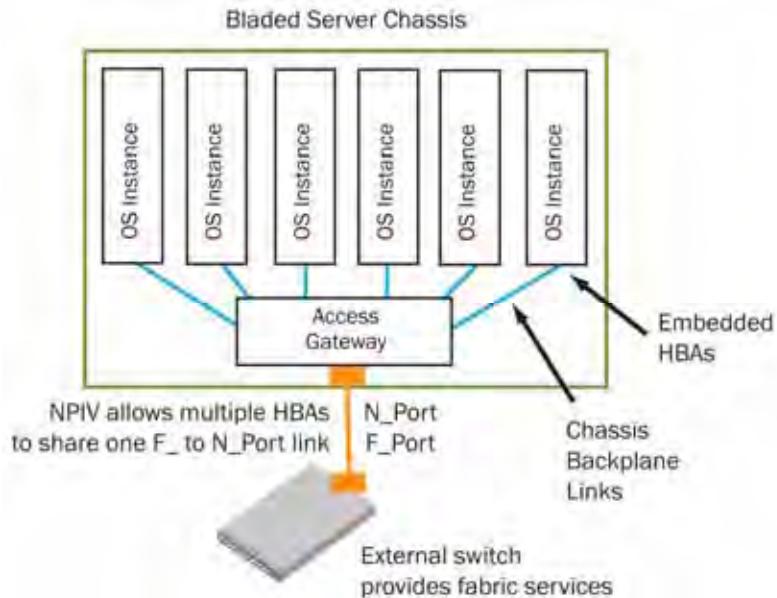


Рис. 43 – Детальная схема работы NPIV

Архитектор SAN может легко применять эту функцию на практике. Единственное ограничение на момент написания этой книги связано с тем, что функция Access Gateway поддерживалась только в платформах, использующих ASIC Goldeneye, например, встроенные продукты 4Gbit и коммутатор Brocade 200E.

Разумеется, платформу Access Gateway можно соединять с другими коммутаторами, использующими другие AS IC. Например, Brocade 200E может функционировать как Access Gateway и подключаться к директорию Brocade 48000 или классической платформе McDATA, которая работает как коммутатор фабрики, но обратное утверждение будет неверным, поэтому функция Access Gateway должна работать на платформе Goldeneye, а коммутатор фабрики, к которому она подключена, должен использовать версию кода, поддерживающую NPIV.

## Топологии для больших расстояний

Самой распространенной вариацией фабрик СЕ является сеть из нескольких территориально-распределенных площадок. На этих площадках может быть одна или несколько фабрик СЕ, которые могут быть соединены между собой, например, для защиты от катастроф. В этом случае центры фабрик соединяются в full mesh, частичный mesh или кольцо в зависимости от числа площадок, топологии WA N и требований к скорости канала между площадками (см. Рис. 44).

В этом примере соединены две удаленные площадки. На каждой есть по паре фабрик СЕ с резервированием “A/B”. Оба коммутатора центра фабрики А СЕ на площадке 1 подключены к обоим коммутаторам центра фабрики А площадки 2. Эти фабрики образуют единую фабрику если только они не изолированы с помощью маршрутизаторов FC-to-FC (это предпочтительный способ соединения территориально-распределенных площадок). В любом случае полученная сеть будет рассматриваться как вариант СЕ и характеристики производительности *внутри* каждой площадки будут считаться соответствующими обычной сети СЕ.

Наиболее надежная и самая производительная технология для линков между площадками – это “родной” Fibre Channel. При его использовании линки будут просто очень длинными FC ISL. Для гарантирования высокой производительности таких линков Brocade разработала продукт под названием “Extended Fabric”, оптимизирующий управление буферными кредитами (buffer-to-buffer) этих линков. Лицензия на Extended Fabrics обычно включается в базовую цену и часто поставляется многими вендорами вместе с коммутаторами Brocade.

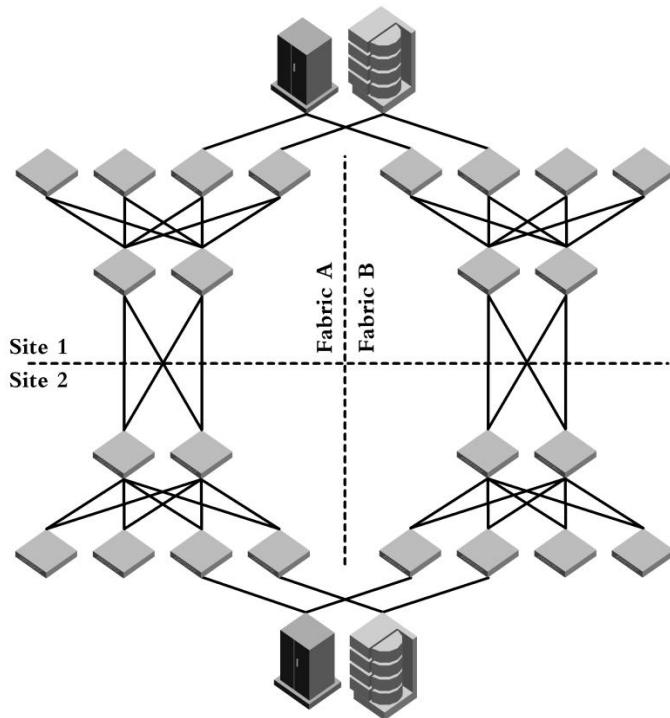


Рис. 44 – Расширение фабрик СЕ

Дополнительную информацию можно найти в "Глава 11: Планирование расстояния" (стр. 339 и далее).

# 7

## 7: Планирование масштабируемости

Рассматривать масштабируемость инфраструктуры SAN можно как с точки зрения ограничений масштабируемости, так и требований масштабируемости. Сначала лучше сосредоточиться на требованиях и затем с учетом ограничений определить, как лучше удовлетворить эти требования.

### **Аксиомы масштабируемости**

Прежде чем обсуждать определение требований к масштабируемости SAN и факторы, от которых она зависит масштабируемость, мы перечислим основные принципы, которые должен использовать архитектор SAN при проектировании крупномасштабного решения

### ***Проектирование крупных решений с мощными компонентами***

Намного проще спроектировать 3000- портовую фабрику с помощью десяти 384- портовых директоров, чем с помощью пятидесяти 64- портовых коммутаторов или ста 32- портовых. Даже если в сети менее тысячи портов, проектировать удобнее с использованием более мощных компонентов. Если фабрика состоит из нескольких сотен портов, установленных на одной площадке, то намного легче ее проектировать, внедрить

и управлять с помощью директоров, а не коммутаторов. Это даст экономию средств за счет уменьшения затрат на лицензирование программного обеспечения, кабели ISL, SFP E\_Port портов, питание, охлаждение, место в стойках и т.д. Если сравнить 300-портовую фабрику из десяти 32-портовых коммутаторов и такую же фабрику из одного Brocade 48000, то совокупная стоимость владения директора будет намного ниже.

### ***О пользе локализации***

Если удастся частично локализовать ввод/вывод (стр. 258), то можно сократить число требуемых ISL и разбить архитектуру на несколько легко управляемых фабрик. Хотя полную локализацию трудно обеспечить и она не оправдывает усилий на ее реализацию, локализация ввода/вывода самых критически-важных приложений намного проще и значительно улучшает их работу.

### ***Использование линков между фабриками - не всегда оптимальное решение***

Иногда архитекторы чтобы упростить управление соединяют отдельные сегменты фабрики с помощью ISL. Теоретически это упрощает операции с зонами и некоторые приложения управления могут «видеть» всю фабрику через подключение к любому коммутатору. Однако, если эти ISL не будут задействованы для ввода/вывода, то лучше не идти этим путем и не добавлять в SAN эти ISL. Более эффективным решением будет применение специализированных приложений управления, способных обслуживать несколько фабрик, например, Brocade Fabric Manager и SAN Health, а не строить новые линки, механизм которых ориентирован на передачу данных.

### ***Маршрутизаторы часто решают проблему***

Если в фабрике много локального трафика, но тем не

## Проектирование С7: Планирование масштабируемости

менее ее требуется соединить с другими фабрики, то лучше попробовать использовать для этого соединения маршрутизаторы FC. Маршрутизаторы обеспечивают выборочное соединение каналов передачи данных без объединения сервисов фабрики. Например, если у основного ЦОДа 100 узлов и 10 из них нужно подключить к резервному ЦОДу, то соединение ЦОДов через маршрутизаторы обеспечивает лучшую масштабируемость, чем объединение фабрик. Однако, надо помнить, что применение маршрутизаторов имеет смысл при высокой локальности трафика на уровне фабрики. Если же каждый хост фабрики должен иметь доступ к каждому устройству хранения другой фабрики, то маршрутизаторы будут неэффективны.

### ***Встроенные коммутаторы должны использовать Access Gateway***

При проектировании больших фабрик обычно используются такие мощные системы, как Brocade 48000, но при работе с шасси блейд-серверов лучше применять встроенные коммутаторы, а не оптические модули pass-through. К сожалению, при подключении большого числа встроенных коммутаторов к фабрике число доменов резко возрастает. Решением в этой ситуации может быть использование Brocade Access Gateway (стр. 197), позволяющего подключать встроенные коммутаторы к фабрике как НВ А, а не как коммутаторы.

### **Определение требований к масштабируемости**

При определении необходимого уровня масштабируемости архитектор S AN должен учитывать четыре аспекта – модель подключения, число портов

хостов и устройств хранения, число портов, выделенных под линки, и разнесенность.

### ***Требования к соединениям***

В некоторых случаях нужно, чтобы каждый порт в сети имел постоянную возможность подключиться к другому порту, например, такое требование часто возникает в IP-сетях. Некоторые архитекторы SAN пытаются использовать эту модель, однако в большинстве SAN требования к соединениям намного ниже. Если допускается несколько ограничений на подключения, то намного проще обеспечить высокую масштабируемость.

Как пример, рассмотрим инфраструктуру с 10 тысячами двухпортовых хостов и 5 тысяч устройств хранения. Теоретически, для подключения потребуется 25 тысяч портов, что на практике невозможно обеспечить. Однако существует несколько способов ограничить требования к подключениям без ущерба к функциональности. Разделение SAN на резервированные SAN A/B наполовину сокращает требования к масштабируемости фабрики, а также улучшает масштабируемость. Хотя при этом порт из фабрики B не может получить доступ к порту фабрики A, но это подключение не является необходимым – если устройство хранения подключено как к A, так и B, то хост сможет получить доступ к данным, записанным на этом устройстве, по двум маршрутам. Тем не менее 12,5 тысяч портов – это слишком много для любой фабрики и поэтому нужен дальнейший анализ.

Очень редко возникает необходимость в обеспечении подключении каждого хоста к любому другому хосту SAN, поскольку в SAN главное – это подключение хостов к устройствам хранения, т.е. максимум,

## Проектирование С7: Планирование масштабируемости

что требуется обеспечить – это подключение любого хоста к LUN любого устройства хранения, что упрощает масштабирование по нескольким причинам.

Например, если устройства хранения – это RAID-массивы корпоративного класса с 20 портами у каждого, то их можно распределить между несколькими разными фабриками. Вместо двух фабрик с 12,5 тысячами портов можно построить 20 фабрик по 1250 портов в каждом и соединять каждый массив по одному порту. В этом случае получится 10 фабрик А и 10 фабрик В и все эти фабрики будут иметь вполне допустимое число портов. Если требуется передавать какой-то трафик между разными фабриками А или В, то можно с помощью маршрутизатора и LSAN создать Meta SAN A и B.

Таким образом, для максимальной масштабируемости очень большой SAN лучше разбить ее на небольшие фабрики. Сначал надо применить резервированную модель А/В и затем разбить фабрики по их функциям, местонахождению, административным группам, либо распределяя между ними порты устройств хранения, как в предыдущем примере.

### ***Число портов хостов и устройств хранения***

Сколько потребуется свободных портов для устройств когда начнет работать SAN? Ответ на этот вопрос легко получить во время интервью, однако спрогнозировать рост числа портов намного сложнее, хотя без этой оценки нельзя правильно спроектировать SAN. Геометрия SAN должна масштабироваться для поддержки максимального числа устройств, которые, как ожидаются, потребуется подключить в будущем, иначе придется перестраивать запущенную в эксплуатацию SAN, что, разумеется, крайне нежелательно.

При прогнозировании возможных темпов роста подключений SAN лучше давать оценку с запасом, поскольку если портов не хватит, то придется перестраивать SAN. Обязательно нужно учитывать *порты*, а не *устройства* потому что у большинства подключенных к SAN хостов и устройств хранения два и более портов. Затем следует разработать общую архитектуру SAN на основе *равномерного распределения* портов и после уменьшить ее с учетом *исходных требований*.

### **Число портов ISL и IFL**

Даже если сначала в SAN будет только один коммутатор, скорей всего через какое-то время у него не останется свободных портов и потребуются ISL или IFL, поэтому необходимо оценить число ISL. Для этого нужно ответить на два вопроса:

- Сколько соединено коммутаторов, маршрутизаторов и шлюзов и какая топология соединений? От ответа на этот вопрос зависит число ISL и IFL.
- Какая требуется производительность для этих соединений? Может понадобиться несколько параллельных линий связи между сетевыми элементами.

Как и при определении числа портов хостов и устройств хранения необходимо учитывать не только текущие требования, но и рост в будущем. Надо выделить порты в зависимости от первоначальных требований к подключениям и производительности линий связи между коммутаторами и фабриками и оставить часть портов свободными с учетом роста в будущем. Обычно, по мере роста числа портов хостов и устройств хранения увеличивается и число коммутаторов, равно как и требования к производительности, из-за чего могут

## Проектирование С7: Планирование масштабируемости

потребоваться дополнительные параллельные линии связи.

Лучше завысить требования к производительности поскольку на работу конечных пользователей не может повлиять низкая загрузка ISL (в отличие от низкой скорости сети). Если планируется внедрение решений ILM или UC, то требования к производительности SAN резко возрастают. К счастью, коммутаторы Brocade 4Gbit – это самые мощные продукты в индустрии SAN и обеспечивают масштабирование от 1Gbit до 256Gbit/s одного сетевого канала.

Рекомендуется зарезервировать несколько портов для дополнительных ISL или IFL если в будущем ожидается рост потребности SAN к производительности и числу портов устройств. Определите число дополнительных портов ISL и IFL, которые будут зарезервированы исходя из необходимого числа коммутаторов и роста требований к производительности, а также выделения линков для специальных будущих проектов (например, построение катастрофоустойчивой инфраструктуры или интеграции с другими фабриками). Как и в случае с определением числа портов хостов и устройств хранения, сначала следует спроектировать общую архитектуру SAN исходя из агрессивных оценок и затем уменьшить ее с учетом первоначальных требований.

### *Учет территориальной распределенности*

Каковы географические характеристики предлагаемого решения? Для построения территориально-распределенных сетей (даже кампусных) требуется дополнительные коммутаторы и/или маршрутизаторы, поскольку коммутаторы желательно располагать на той же площадке, где находятся устройства, которые они соединяют.

Например, у заказчика установлено шесть серверов и два порта устройств хранения, которые равномерно разделены между двумя ЦОДами в пределах одного кампуса. В таком случае может подойти 8- портовый коммутатор SilkWorm 3250 или Brocade 200E с лицензией на 8 портов, однако у такого подхода есть свои минусы.

Прежде всего, нулевой порт надо зарезервировать для расширения, иначе если потребуется добавить еще один хост, то придется перестраивать SAN. В случае использования 200E эта проблема решается покупкой еще одной лицензии программного обеспечения для 8 портов.

Более важное ограничение связано с тем, что для половины устройств физически коммутатор будет установлен в другом ЦОДе. Если расстояния небольшие (как в этом примере), то можно проложить оптические кабели от коммутатора напрямую к трем удаленными хостами и одному порту устройства хранения. Это потребует четырех кабелей и восьми волокон и во многих случаях стоимость оптических кабелей будет равна и даже больше стоимости самого коммутатора.

Также можно установить второй SilkWorm 3850 и тогда у каждого ЦОДа будет свой коммутатор. Тогда для соединения ЦОДов потребуется только один кабель для ISL. Такой подход не только упростит кампусную сеть, но и решит проблему свободных портов для масштабируемости – у каждого коммутатора будет три свободных порта для подключения новых устройств и/или ISL.

В большинстве территориально-распределенных SAN требуется намного больше дополнительных коммутаторов чем в этом примере. Если SAN внедряется для защиты от катастроф, то линии связи между площадками могут иметь длину несколько сот

## Проектирование С7: Планирование масштабируемости

километров. Разумеется, при таких расстояниях обеспечить соединение каждого хоста с каждым устройством хранения нецелесообразно, поэтому нужно на каждой площадке установить отдельную группу коммутаторов и архитектор SAN должен анализировать требования к масштабируемости каждой площадки по отдельности.

### **Обеспечение максимальной масштабируемости**

После того как определены первоначальные и будущие требования SAN к масштабируемости, требуется спроектировать соответствующую архитектуру. При этом нужно учитывать все факторы, ограничивающие масштабируемость SAN и применять решения, позволяющие снимать такие ограничения без ухудшения функциональности.

Эти ограничения относятся к пяти категориям – управляемости, локализации сбоев, матрицам поддержки продуктов разных вендоров, сервисам сетей хранения и самим протоколам.

### ***Характеристики протокола***

К счастью, при внедрении решений Brocade можно не учитывать архитектурные ограничения, поскольку адресное пространство фабрики рассчитано на миллионы устройств, однако у других вендоров максимальные характеристики намного ниже, поэтому стоит рассмотреть это вопрос подробнее.

У каждой фабрики FC есть свое трехбайтное адресное пространство. В первый байт записывает домен коммутатора в фабрике, во второй – порт коммутатора, а в третий – адрес конкретного устройства, которое подключено к порту коммутатора (эта информация

требуется когда порт подключен к коммутатору FC-AL или устройству с архитектурой петли (например, JBOD).

Если бы все это адресное пространство было доступно, то можно было бы использовать адреса 16.7 миллионов устройств ( $256^3$ ). Однако стандарты фабрик FC требуют резервирования некоторых адресов для сервисов фабрики, например, сервера имен, а некоторые адреса нельзя использовать из-за ограничений протокола FC-AL, поэтому стандарт обеспечивает 7.7 миллионов доступных для использования адресов (239 доменов на 256 портах на 127 адресах FC-AL).

Поскольку невозможно построить стабильную петлю FC-AL, в которой более двух десятков устройств, доступное адресное пространство фабрики сокращается еще до 1.2 миллиона устройств. Если не использовать адреса в петле, то максимальный размер адресного пространства будет более 60 тысяч портов.

Даже последнее число является достаточно большим по современным меркам и пока ни одному заказчику не потребовалась фабрика с такими возможностями масштабирования<sup>58</sup>. Таким образом, ограничения применения на практике адресного пространства протокола FC не влияют на архитектуру фабрик Brocade.

К сожалению, не все вендоры корректно реализуют адресное пространство. Например, у одного вендора адресное пространство поддерживает не более 31 домена и если у каждого коммутатора 256 свободных портов, то в фабрике будет свыше 7000 портов. Хотя это число - только десятая часть от теоретической мощности Bro-

---

<sup>58</sup> Хотя у одного клиента Brocade проект сети рассчитан на масштабирование свыше 600 тысяч портов, а другого – свыше 100 тысяч ... оба используют несколько фабрик.

## Проектирование С7: Планирование масштабируемости

cade, в большинстве случаев этого будет более чем достаточно.

На практике адресное пространство *не должно быть ограничением* масштабируемости SAN при использовании оборудования любого вендора. Теоретически, возможно построить фабрику Brocade, к который подключено более миллиона устройств, однако на практике из-за влияния других факторов максимальный размер фабрики будет намного меньше.

Итак, при использовании фабрик Brocade не нужно учитывать ограничения масштабируемости адресного пространства, хотя при использовании оборудования других вендоров этот фактор может играть роль. Кроме того, нужно помнить, что на практике пока не используется все теоретически доступное адресное пространство.

### **Управляемая масштабируемость**

Помимо разработки схемы и последующего построения SAN необходимо обеспечить постоянное управление фабрикой и сам процесс управления может ввести дополнительные ограничения на масштабируемость.

В определенной степени эти ограничения носят технологический характер – программного обеспечение для управления SAN, как и любое программное обеспечение, имеет пределы масштабируемости из-за ограниченного объема оперативной памяти и числа процессоров компьютерной платформы, на которой оно работает. Хотя предельные значения масштабируемости управляющего приложения могут быть достаточно высокими, их все равно следует учитывать.

Ограничения могут быть связаны с человеческим фактором и определениями процессоров. Если много

разных приложений подключены к одной фабрике, то различные группы пользователей и ИТ-администраторов могут участвовать в процессе изменения управления этой фабрики и трудно выбрать время для установки нового коммутатора или обновления версии микрокода. Во многих инфраструктурах даже те изменения, которые не нарушают нормальную работу, должны проводиться под контролем, для устранения рисков и координации таких изменений с различными пользователями и администраторами. Чем больше фабрика, тем больше людей затрагивает процесс изменений.

Вне зависимости от того, связаны ли ограничения с технологиями, человеческим фактором или процессами, любая организация рано или поздно столкнется с проблемами масштабирования из-за управления, поэтому архитектор SAN должен предусмотреть управление окончательным решением и масштабируемость для пользователей, процессов и технологий.

## ***Изоляция сбоев***

Любой одиночный элемент – это единичная точка отказа, включая сети и программное обеспечение сетевых сервисов. Этот принцип справедлив как для SAN, так и для любой другой системы любой степени сложности, поэтому архитекторы SAN, которые хотят обеспечить доступность, должны учитывать, что одиночная фабрика может отказать как одиночный элемент. Также как одиночный элемент может выйти из строя Ethernet в результате широковещательного шторма.

Сетевая индустрия давно осознала, что сегменты Ethernet на практике имеют ограничение масштабируемости из-за необходимости изоляции сбоев. Даже если было возможно построить один сегмент Ethernet с тысячами устройств, то это не имело бы

## Проектирование С7: Планирование масштабируемости

смысла, поскольку независимо от того, насколько аккуратно построена сеть и выполняется управление ею, равно или поздно возникнет какая-то внештатная ситуация и весь сегмент потеряет работоспособность.

Для решения этой проблемы сетевые архитекторы ограничивают размер сегмента, чтобы ограничить влияние потенциальных проблем. Сетевая индустрия предпочитает строить сегменты Ethernet, размер которых не превышает одной IP-подсети класса C, т.е. не более 250 устройств, хотя для организаций, где более 250 компьютеров, это создает дополнительную проблему при подключении.

Для достижения баланса подключения и изоляции сбоев нужно применять иерархическую модель сети на базе маршрутизаторов. Такой подход может применяться и для сетей хранения при соединении фабрик маршрутизаторами FC, например, Brocade 7500.

### *Матрицы поддержки продуктов разных вендоров*

Даже если сеть нормально работает, то это еще не означает, что производитель используемого оборудования обеспечит официальную поддержку его конфигурации. Brocade может протестировать и объявить о поддержке определенного числа портов в фабрике в новом релизе Fabric Operating System , но не все OEM-партнеры Brocade могут сразу же поддерживать это число портов. Важно понять не только технологии, которые разработала Brocade, но и отношения с поставщиком поддержки, для чего нужно получить новейшую версию матрицы поддержки перед подготовкой окончательной версии архитектуры и проверить ее на соответствие параметрам масштабируемости.

Крайне важно вместе с инженерами поставщика поддержки выявить те элементы архитектуры, которые

могут не попадать в матрицу поддерживаемых конфигураций. Архитектору нужно разобраться (а) почему конфигурация не попадает в матрицу и (б) может ли поставщик все же обеспечить ее поддержку. Во многих случаях планируемая конфигурация отсутствует в списке конфигураций, которые были протестированы поставщиком сервисов поддержки, но в то же время быть похожа на одну из них, тогда он согласится на поддержку. Если проектируемая SAN не попадает в матрицу поддержки вендора, то следует получить от него письменное подтверждение поддержки, чтобы в будущем не возникло конфликтов из-за разной трактовки того, на что предоставляется поддержка.

## *Сервисы сетей хранения*

Основной фактор, ограничивающий масштабируемость SAN, не связан с нижними уровнями стека протоколов или других рассмотренных выше критериев. Как было показано в разделе “Характеристики протокола” (стр. 215), формат пакетов Fibre Channel обеспечивает масштабирование адресного пространства, которое более чем достаточно для любой фабрики. То же относится и к нижним уровням – кабелям и физической среде. Даже ограничения масштабируемости из-за процессов управления можно снять если применить зафиксированные на бумаги политики и процедуры, либо в крайнем случае распределением управления SAN по нескольким административным регионам.

Самый проблематичный ограничивающий фактор – это проектирование сервисов, поддерживающих функции SAN, например, сервисы фабрики в решении Fibre Channel или аналогичные сервисы в сети iSCSI. В обоих случаях возникает одна и та же проблема масштабируемости.

## Проектирование С7: Планирование масштабируемости

С точки зрения проектирования развертывания коммутаторов, SAN сервисы – это самый сложный компонент проектирования и тестирования, поскольку они взаимозависимы и корректная работа одного сервиса зависит от других сервисов.<sup>59</sup>

Например, в Brocade Zoning включен протокол switch-to-switch zone transfer protocol и метод программирования зон на уровне аппаратуры каждого порта. Это накладывает серьезные технические ограничения на размер таблицы зон, но напрямую не связано с максимальным числом портов в фабрике, поэтому может показаться, что проектирование сервисов зон не влияет на масштабируемость.

Однако сервисы зонирования каждого коммутатора должны “разговаривать” с сервисами для координации контроля доступа к оборудованию. Вряд ли стоит разрешать серверу имен “говорить” узлу о других узлах, которые не имеют доступ через зонирование<sup>60</sup>, однако именно такая ситуация может возникнуть если работа сервисов не будет скординирована. У сервисов имен есть ограничения на масштабируемость числа портов в фабрике, поскольку у каждого коммутатора есть ограниченное число ресурсов процессора и ОЗУ. Кроме того, сервисы NS должны оперативно обрабатывать запросы даже если к ним обращаются все устройства фабрики, а время реакции процессов может быть

---

<sup>59</sup> Это не зависит от протокола. Взаимозависимость сервисов вызвана связями между функциями сервисов, а не низкоуровневых протоколов, которые используют сервисы. Сервисы iSCSI ведут себя аналогичным образом.

<sup>60</sup> Кроме создания рисков безопасности это приведет к тому, что все хосты попытаются получить доступ ко всем устройствам фабрики, что будет блокировано механизмом зонирования. В результате каждый хост будет бесконечно пытаться получить доступ и будет напрасно загружать процессоры хоста и коммутаторов.

ограничено скорость отклика процессов зонирования. Таким образом, процессы зонирования *могут влиять* на стабильность фабрики из-за взаимосвязи с другими сервисами.

На самом деле из-за таких взаимосвязей сервисы не удается масштабировать линейно, т.е. нагрузка на процессор коммутатора возрастает экспоненциально по мере добавления в фабрику портов и доменов и поэтому повышение частоты процессора не даст пропорциональное увеличение масштабируемости фабрики и удвоение частоты может улучшить масштабируемость фабрики только на 0.01%.

Разумеется, от архитектора SAN не требуется точное знание реализации процессов сервисов фабрики и их взаимодействия. Однако для масштабирования фабрики *производитель коммутаторов* должен приложить существенные усилия и потратить значительно времени на настройку внутреннего поведения процессов коммутатора, осуществляющих каждый сервис и их взаимодействия с другими сервисами. Если все устройства фабрики одновременно запрашивают сервер имен фабрики (такая ситуация периодически возникает), то процессы NS должны быстро ответить на эти запросы, для чего они должны быть скоординированы с процессами зонирования, а также и с большинством других сервисов SAN. По мере роста размера фабрики сложность взаимодействия процессов растет нелинейно. Удвоение размера фабрики увеличивает в намного больше раз нагрузку на процессоры коммутаторов в зависимости от того, как спроектированы сервисы, поэтому не следует принимать на веру рекламу тех вендоров, которые утверждают, что более мощные процессоры компенсируют отсутствие опыта в точной настройке сервисов фабрики.

К счастью, Brocade занимается такой настройкой уже

## Проектирование С7: Планирование масштабируемости

более десяти лет при поставках коммутаторов FC для критически-важных сетей хранения данных и ее продукты являются признанными лидерами по реальной поддержке масштабируемости. От архитектора SAN не требуется понимание того, как устроены сервисы хранения, но он должен учитывать эти аспекты при выборе оборудования SAN, поскольку если у Brocade было время чтобы научиться обеспечить масштабирование коммутаторов, другие вендоры пока не имеют такого опыта.

Именно пренебрежение вопросами масштабируемости привело к тому, что в последние годы ряд производителей вынуждены были уйти с рынка SAN. Один производитель пытался решить проблему масштабирования разделением каждой сети на изолированные сети VSAN со своей копией сервисов, однако все эти сервисы работали *на одних и тех же* процессорах, поэтому при росте размера сети нагрузка на процессоры росла так же быстро, как при использовании одной фабрики. Добавление коммутаторов в новую VSAN давало такой же эффект, как добавление их в единую сеть. На самом деле, применение нескольких VSAN может *увеличить* нагрузку на процессоры при том же количестве портов потому что каждый коммутатор представляет один или несколько отдельных доменов для каждой VSAN. В результате, масштабируемость только ухудшилась из-за того, что вендор не понимал принципов работы сервисов SAN. Brocade предлагает аналогичную функцию под названием Virtual Fabrics, но позиционирует ее не как решение проблем масштабируемости, а лишь для улучшения управляемости.

С точки зрения масштабируемости предлагаемый Brocade подход LSAN изолирует сервисы на процессорах отдельных коммутаторов, поэтому

добавление дополнительных фабрик в сеть не будет ухудшать масштабируемость. Каждая добавляемая в SAN фабрика имеет собственные процессорные ресурсы. Маршрутизируемая архитектура Meta SAN предотвращает непосредственное взаимодействие сервисов разных фабрик, поэтому устраняет проблемы нелинейной масштабируемости. ( См. “ Сравнение фабрики с SAN и Meta SAN” на стр. 42.)

До сих пор LSAN остается единственным из предлагаемых вендором подходов, который позволяет на практике решить проблему масштабирования сервисов сетей хранения. Его можно рассматривать как аналог подхода, который применяется с помощью IP-маршрутизаторов для масштабирования сегментов Ethernet, однако он оптимизирован для более строгих требований к производительности и надежности сетей хранения. Таким образом, архитекторы SAN могут с помощью этого подхода успешно построить сервисы для очень больших сетей хранения, а другие подходы не способны устраниить проблему масштабирования сервисов и даже ее усложняют.

Интересно попытаться провести аналогию между масштабированием фабрики FC и подсети Ethernet, хотя в IP-сетях *нет аналога* проблемы масштабирования сервисов. Вендоры, накопившие опыт при создании крупномасштабных решений Ethernet с IP-маршрутизаторами, не обладают опытом создания масштабируемых SAN и поэтому делают такие же ошибки, как уже упоминавшийся вендор VSAN. Чтобы такая же проблема существовала для коммутатора IP Layer 3 VL\_AN, он должен обслуживать сервисы DNS, DHCP, WI NS, NIS+, NTP, LDAP, iSNS, RADIUS и многие другие в *дополнение* к протоколам STP, RIP и OSPF. Ни один производитель оборудования для IP-сетей никогда не пытался создать такой продукт, однако аналогичный функционал используется в

## Проектирование С7: Планирование масштабируемости

коммутаторах FC SAN от Brocade уже более десяти лет. Он усложняет проектирование и построение коммутаторов фабрики FC, но зато пользователи получают продукты, которые проще в развертывании и управлении.

Решение проблемы масштабируемости достаточно простое – прежде всего, архитектор SAN должен выбрать оборудование с архитектурой хорошо масштабируемых сервисов, т.е. вендора, который давно занимается разработкой сервисов SAN, а затем построить резервированную архитектуру с физически изолированными фабриками А/В (сервисы взаимодействуют между собой внутри фабрики и хотя это создает проблему масштабируемости, их взаимодействие происходит только в пределах фабрики А или В). После этого следует разработать сетевую архитектуру каждой половинки (А и В), поддерживающие иерархическое масштабирование с помощью маршрутизаторов FC для обслуживания взаимодействия сервисов. Если сеть использует шасси блейд-серверов, то следует использовать функцию Access Gateway. Такой многоэтапный подход позволяет построить сеть хранения с числом портов, которое несколько лет назад казалось недостижимым.

### ***Масштабируемые топологии***

Для максимальной масштабируемости SAN лучше выбрать топологию, которая помогает провести крупномасштабное внедрение. Теоретически, можно построить фабрику с топологией частичный mesh с 100 тысячами портов, однако на практике такая фабрика (a) не работает, (b) даже если бы работала, то ее нельзя было бы поддерживать (c) была бы слишком сложной для устранения сбоев даже если бы ее можно было поддерживать и (d) имела бы совершенно неконтролируемые характеристики производительности,

т.е с любой стороны ее построение не имеет никакого смысла.

Лучшая стратегия построения крупномасштабной SAN – это сначала разбить ее архитектуру на отдельные части, так называемые “единицы управления”. Такое разбиение можно провести по функциональности, приложениям или географии и другим подходящим для конкретной сети критериям. Главное – каждая единица управления должна иметь как можно меньше связей с другими единицами управления, т.е. внутри нее должна быть высокая степень локальности.

Если не требуется никакое подключение единиц управляемости (например, в случае резервированных фабрик А/В), то их можно сконфигурировать как независимые SAN. Если же нужно подключить хосты из одной единицы управления и LUN устройств хранения из другой, то можно использовать стратегию ориентированную на системы хранения (стр. 176) либо архитектуру с маршрутизацией. Последний вариант очень популярен для внедрения DR и ВС. Для отдельных резервных фабрик обычно подсоединение выполняется с помощью маршрутизаторов или хостов с дополнительным НВ А. Таким образом, резервные серверы можно подсоединить к каждой соответствующей “дисковой” фабрике и “ленточной” фабрике через различные адаптеры.

В результате, будет построена архитектура с несколькими фабриками и определен метод соединения фабрик между собой. Затем можно спроектировать отдельные фабрики небольшого радиуса, используя директоры или топологию центр/периферия.

## 8: Планирование производительности

При проектировании SAN надо особое внимание уделять планированию производительности даже в том случае, когда у обслуживаемых приложений скромные требования к производительности.

Планирование производительности требуется при проектировании любой сети, но для сетей хранения оно имеет особую важность – даже сеть хранения начального уровня без высокой производительности и надежности просто *не сможет работать* и это является одной из причин, по которой в качестве транспорта SAN используется Fibre Channel.

Низкие показатели производительности и надежности часто допустимы для приложений, которые поддерживают сетевые технологии IP и Ethernet. Протоколы верхнего уровня таких сетей разрабатывались в расчете на возможность потери пакетов, недостаточной пропускной способности, задержек и нарушения порядка доставки пакетов. Такие приложения IP-сетей как web-браузеры начинали свою историю с версий, рассчитанных на медленные модемные соединения со скоростью 9600 baud. Низкая стабильность IP-сети не будет дестабилизировать соединение – например, если соединение между web-

клиентом и web- сервером оборвется, то пользователь просто кликнет по кнопке “обновить” и попытается снова загрузить страницу. Такая ситуация *не допустима* для SAN ( нельзя нажать “обновить” в ERP-системе если LUN устройств хранения вдруг отключатся...)

SAN поддерживает приложения, первоначально разработанные в расчете на использование выделенной шины SCSI и поэтому не допускающие даже возможность возникновения проблем с производительностью и надежностью, которые часто исключает их архитектура, например, если у шины SCSI только один инициатор, то не может быть проблем с пропускной способностью. Поскольку узлы SAN рассчитаны на максимальную надежность, то если SAN станет нестабильной или медленной, то в результате могут быть испорчены данные и придется восстанавливать их по резервным копиям или даже вводить в действие план аварийного восстановления. В критически-важном сервере баз данных не допускается команда “обновить” ради устранения сбоев в сетевой архитектуре.

Сказанное выше еще не означает, что SAN – это «рискованная» технология . Fibre Channel SAN уже доказали свою надежность при использовании в критически-важных инфраструктурах в течение многих лет. Просто я хотел подчеркнуть, что производительность нужно учитывать всегда, даже если приложениям *не требуется* высокая производительность.

На следующих страницах описывается, как определить требования к производительности и некоторые инструменты, которые архитектор SA N

## Проектирование С8: Планирование производительности

может использовать для удовлетворения этих требований.

### **Обзор факторов, влияющих на производительность**

В этом разделе рассматриваются основные факторы, от которых может зависеть производительность SAN, в том числе краткое описание каждого фактора и его влияния на производительность SAN, а также рекомендации по устранению ограничений производительности, связанных с конкретным фактором.

#### ***Оконечные устройства***

Наиболее распространенное ограничение производительности SAN вызвано не самой SAN, а подключенным к ней устройствами. Архитектору следует выяснить как внешние, так и внутренние характеристики производительности подключенных к SAN устройств для того, чтобы они не стали узким местом в сети.

Например, можно установить два интерфейса 4Gbit Fibre Channel на ПК с процессором 800MHz. Каждый интерфейс способен обрабатывать трафик 8Gbits (в режиме full duplex), поэтому теоретически хост может передавать по SAN трафик 16Gbits, однако ПК с 800-мегагерцевым процессором не способен обеспечить даже часть этой производительности – его процессор просто не может генерировать данные с такой скоростью. Ограниченный объем ОЗУ хоста и скорость его шины также влияют на производительностью SAN.

На момент написания этой книги *большинство* хостов, подключенных к SAN, не могли генерировать

трафик, который бы полностью заполнил полосу пропускания четырехгигабитного Fibre Channel, не говоря уже о нескольких интерфейсах FC. Сказанное не означает, что не имеет смысла использовать адаптеры 4Gbit FC HBA, наоборот – они гарантируют, что Fibre Channel S AN никогда не станет ограничением для производительности приложений.

Аналогичным образом интерфейсы Fibre Channel дисков, ленточных устройств и RAID- массивов теоретически способны генерировать трафик 1Gbit, 2Gbit или 4Gbit, но на практике сами эти устройства не могут обеспечить такие скорости. Если двухгигабитный интерфейс SAN подключен к диску, у которого максимальная скорость 200Mbit, то узким местом будет диск, а не сама SAN.

При использовании медленных хостов FC HBA 1Gbit или 2Gbit обеспечат оптимальное соотношение цены и производительности. Может показаться привлекательной идея запускать на таких хостах программное обеспечение iSCSI, но надо учитывать, что у этих хостов могут быть сильно загружены процессоры и применение стека приведет к падению производительности хоста и его нестабильности.

Медленные устройства хранения лучше модернизировать, например, увеличить объем кэш-памяти RAID- контроллера либо применить в массиве JBOD чередование по нескольким дискам с помощью менеджера томов, который работает на уровне хоста или SAN, что увеличит производительность до уровня аппаратного RAID- контроллера. Однако лучшее решение – это с самого начала использовать мощные устройства хранения.

## Проектирование С8: Планирование производительности

### **Протоколы SAN**

Следующий фактор, влияющий на производительность – это выбор протокола SAN, т.е. первого компонента механизма для передачи трафика между граничными устройствами. Если в самих граничных устройствах нет узких мест, то проблемы производительности могут возникнуть из-за этого протокола.

Сети Fibre Channel можно сконфигурировать для разных требований производительности, начиная от самых простых инсталляций и до обеспечения построения фабрики из нескольких тысяч узлов с соединениями каждого с каждым на полной скорости в обе стороны (full-duplex). Brocade 5000 – прекрасный пример высокопроизводительного коммутатора Fibre Channel – несмотря на небольшой форм-фактор (1U) и агрессивную цену, Brocade 5000 обеспечивает внутреннюю пропускную способность 256Gbits, что намного превышает требования современных приложений SAN.

Такие протоколы IP SAN, как iSCSI, подходят (в том числе и по цене), если не требуется высокая производительность. В недавнем номере одного известного сетевого журнала говорилось, что если правильно выбрать оборудование и программное обеспечение для iSCSI, то можно получить реальную производительность 4Gbits при использовании 10 Gbit Ethernet. В данном случае “правильное” оборудование будет дорогим и намного дороже оборудования 4gbit FC (подробно технические причины оставания производительности iSCSI от FC рассмотрены в разделе “Протоколы SAN” Главы 1: Основы SAN”, стр. 33).

Главное, о чем должен помнить архитектор SAN, - это то, что ценовое преимущество iSCSI теряется если пытаться обеспечить даже сколько-нибудь существенную долю производительности FC.

Таким образом, архитектор SAN должен учитывать присущие протоколу характеристики производительности поскольку от этого зависит остальной выбор. Если выбрать iSCSI, то выбор скоростей будет очень ограничен и производительность узла будет зависеть от выбора программных стеков iSCSI.

Разумеется, от протокола зависит не только производительность. Нужно учитывать и зрелость самого протокола, надежность, успех на рынке и зрелость доступных для этого протокола сервисов и приложений для управления. Все эти факторы указывают на предпочтительность внедрения Fibre Channel и объясняют, почему FC используется в подавляющем большинстве SAN. Однако важна и цена решения. Если от SAN не требуется высокая производительность и надежность, то можно использовать и iSCSI, обеспечивающий более высокую эффективность по цене, чем Fibre Channel. iSCSI следует рассматривать как замену протоколов NAS (например, CIFS или NFS), а FC следует использовать для всех остальных задач.

### ***Скорости линков***

Затем следует рассмотреть скорости линий связи и решить на какой скорости должна работать сеть - 1Gbit, 2Gbit, 4Gbit, 8Gbit, 10Gb it, 16Gbit, 32Gbit, 256Gbit или какой-то другой. На самом деле правильным было бы назвать этот параметр “скоростями пути”, поскольку

## Проектирование С8: Планирование производительности

коммутаторы Brocade поддерживают пути с равномерным распределением 256Gbit с использованием аппаратных методов объединения линков и тот же физический порт Brocade может поддерживать линки 1Gbit. Ни один другой вендор SAN не может обеспечить этого и поскольку внедрения iSCSI не поддерживают большинство этих опций, то перед приобретением оборудования, которое будет подключено к SAN, надо изучить вопрос о скорости линий связи.

При этом нужно учитывать несколько факторов. Очевидно, что крайне важно соотношение цены и производительности, однако нужно учитывать совокупную стоимость подключения. Например, интерфейсы 10Gbit могут использовать одномодовую оптику, а интерфейсы 4Gbit - многомодовую (см. стр. 381 и 384). Даже если стоимость пересылки данных для этих скоростей будет одинакова, то стоимость кабелей между интерфейсами будет совершенно разной. Именно поэтому переход на 4Gbit произошел так быстро, а 10Gbit все еще используется редко. Однако при построении территориально-распределенных сетей требуются одинаково дорогие кабели и оптика как для 4Gbit, так и для 10Gbit, и в этом случае 10Gbit может оказаться предпочтительным поскольку требует меньше линий связи для достижения определенной производительности (см. пример “Лезвия 10Gbit и решения DR/BC” на стр. 364.)

Обычно лучше рассматривать максимально большое число опций. Даже если сначала SAN требуется производительность только 1Gbit, то нужно помнить, что индустрия хранения быстро движется в сторону более высоких скоростей, в том числе и за счет таких

тенденций, как ILM и UC ( 85) и роста общего объема данных.

Лучше всего выбрать инфраструктуру, которая может масштабироваться по мере изменения потребностей бизнеса, например, можно использовать порты Brocade 48000 для линий связи 1Gbit и без дополнительных затрат увеличить их производительность до 4Gbit, а затем на еще более быстрый интерфейс, приобретя новые лезвия. В противном случае для повышения производительности потребуется полностью заменить оборудование. Платформы Brocade сегодня при использовании транкинга обеспечивают еще более высокие скорости. По мере роста среды и добавления в сеть новых коммутаторов транкинг Brocade ISL и Dynamic Path Selection (DPS) могут соединять элементы для достижения полосы пропускания 256Gbit без дополнительных затрат.

Менее совершенные продукты требуют полного обновления инфраструктуры для достижения этого уровня производительности и даже смены технологий SAN. Например, если сначала построена сеть на основе iSCSI, а затем для повышения производительности потребуется перейти на Fibre Channel, то будут потеряны инвестиции во всю инфраструктуру iSCSI SAN. Чтобы не оказаться в подобной ситуации архитектор SAN должен выбирать компоненты сети с гибкими характеристиками производительности, способные масштабироваться по мере роста потребностей бизнеса.

### ***Переподписка и переполнение канала***

Переподписка (over-subscription) – это ситуация, когда ресурс не может полностью поддерживать столько устройств, сколько *могут* потребовать доступ к нему.

## Проектирование С8: Планирование производительности

Как аналогию можно рассмотреть случай, когда авиакомпания продает больше билетов на рейс, чем есть мест в самолете. Если никто из пассажиров не опаздывает, то несколько пассажиров придется отправить следующим рейсом, но если кто-то из пассажиров не успеет на рейс или откажется от поездки, то все остальные улетят вовремя и даже не будут знать о переподписке<sup>61</sup>, поэтому неправильно утверждать, что только из-за переподписки пассажир прилетел позже.

Такая же ситуация возникает в SAN, когда несколько хостов или устройств хранения могут использовать один и тот же ISL. Тогда возникнет переподписка на ISL, однако это не обязательно скажется на производительности. Например, два хоста используют одну и ту же линию связи, но один только ночью, а другой – только днем. Хотя теоретически они могут одновременно использовать линию связи, на практике этого не происходит.

Если же линия связи с переподпиской *действительно* используется сразу двумя хостами, то возникнет переполнение канала и часть трафика ставится в очередь и передается с задержкой и в результате сокращается эффективная полоса пропускания между конечными устройствами. Для большинства сетей переполнение канала не означает полное нарушение связи, но временное падение ее скорости.

Итак, переподписка не вызывает автоматически переполнение канала, но создает риски переполнения. Скорость пути в сочетании с общей архитектурой сети

---

<sup>61</sup>

Авиакомпании называют такую ситуацию “over sold”, но принцип остается тем же.

создают потенциал для переполнения канала (либо предотвращают его), однако переполнение наступает только в зависимости от характера ввода/вывода подключенных устройств.



## Заметки на полях

*Линия связи с переподпиской – это линк, в котором несколько устройств могут конкурировать за полосу пропускания. Традиционные сети передачи данных, например Internet, давно проектируются в расчете на высокий уровень переподписки. Переполнение канала происходит, когда несколько устройств действительно начинают конкурировать за полосу пропускания. Главное для управления полосой пропускания – это определить требования к производительности и соответствующим образом спроектировать SAN. Если все подключенные к коммутатору приложения способны генерировать только трафик 4Gbits/сес и два линка ISL 4Gbit/sec соединяют коммутатор с остальной SAN, то переполнения не произойдет, хотя теоретически может возникнуть переподпись на эти линки.*

SAN обычно не способны поддерживать высокий уровень переподписки (как у Internet), но большинство построенных SAN может работать при небольшом уровне переподписки линий связи за счет таких характеристик, как поддержка скачков и провалов трафика, совместного использования ресурсов, локальной передачи данных и использования устройств, способных работать на небольшой части доступной полосы пропускания.

Более того, проектирование с переподпиской часто экономит средства на внедрение SAN. Если 16-портовые

## Проектирование С8: Планирование производительности

границные коммутаторы соединены в топологию центр/периферия, то архитектор SAN может использовать по два ISL на коммутатор, подключив тем самым по 14 устройств на коммутатор. Это дешевле, чем соединение только восьми устройств и использование остальных восьми портов коммутатора для ISL.

Архитекторы SAN обычно идут на допущение переполнения, но стараются уменьшить его влияние в зависимости от требований к производительности приложений, использующих SAN. Если критичен уровень производительности и полоса пропускания, то для снижения уровня переполнения чаще всего применяется локализация трафика. Основные методы снижения переполнения:

**Локализация** – близкое соединение портов отправителя и получателя, при котором трафик между ними не будет вообще или крайне редко будет идти через ISL, где обычно и происходит переполнение.

**Использование больших коммутаторов** – такие директоры, как Brocade 48000 обладают максимальной производительностью и масштабируемостью - до 384 устройств 4Gbit можно подсоединить к одному Brocade 48000. Обычно не менее 10 хостов приходится на один интерфейс устройства хранения и при этих условиях можно сконфигурировать порты 384- портового 48000 так, что не происходит переполнения. Можно использовать локальные соединения внутри лезвий. "Горячие" устройства можно подсоединять, например, к 16- или 32-портовым лезвиям для обеспечения высокой производительности.

**Использование быстрых линий связи** – 4Gbit FC ISL обеспечивают больше полосы пропускания, чем,

например 1Gbit FC или Ethernet. Переход на более высокую скорость также уменьшает вероятность переполнения. Для высокопроизводительных решений DR и BC, можно использовать ISL 10Gbit.

**Использование транкинга для расширения связей между коммутаторами** – в один транк 32Gbit можно объединить до восьми 4Gbit IS L и между восемью такими транками можно балансировать трафик с помощью Dynamic Path Selection, в итоге получив канал 256Gbit. На сегодняшний день нет приложений, которым требуется такая полоса пропускания ISL, поэтому при использовании транкинга переполнением можно пренебречь.

### ***Блокирование (HoLB)***

Блокирование связано с очередями. Если коммутатор, маршрутизатор или директор неправильно спроектированы, то трафик между двумя устройствами может полностью блокировать трафик между остальными парами устройств.

Блокирование часто путают с переполнением. Как и переполнение, блокирование возникает когда несколько устройств конкурируют за ограниченный ресурс и в результате у этих устройств может упасть производительность. Однако у блокирования и переполнения разные причины, как и некоторые из следствий.

Блокирование правильнее назвать “Head of Line Blocking” ( блокированием в начале линии, HoLB) потому что оно связано со сложной проблемой очередей, а не просто нехваткой пропускной способности. Если коммутатор неправильно спроектирован, то пакет в начале очереди может “застрять” и блокировать все

## Проектирование С8: Планирование производительности

остальные пакеты, хотя большая часть полосы пропускания при этом может быть свободна. Блокирование не просто *снижает* производительность других устройств, пытающихся использовать этот линк, но *полностью нарушает* передачу трафика через него на длительное время.

У всех продуктов Brocade полностью исключен такой тип ошибки, однако у продуктов других вендоров SAN она возникает. Пользователи должны учитывать вероятность блокирования HoLB в crossbar-коммутаторах и концентраторах. Хотя некоторые поставщики этих устройств утверждают, что их алгоритм “виртуальных очередей” решает эту проблему, однако эти заявления не подтверждены независимыми тестами.

### ***Коэффициенты потерь и ошибок***

В любой сети может произойти потеря или порча отдельных фрагментов данных. Если пакет будет испорчен при передаче через сеть FC (что случается крайне редко), то коммутаторы FC всегда обнаруживают это с помощью проверки CRC и отбрасывают испорченный пакет, поэтому с точки зрения использующего SAN узла, потеря и порча данных не отличаются. Поскольку изначально приложения разрабатывались в расчете на системы хранения, где не может быть порчи и потери данных (например, подключенные напрямую диски SCSI), для них недопустимым является сколько-нибудь ощутимый коэффициент ошибок.

При использовании настоящих Fibre Chan nel SAN коэффициент ошибок и потерь можно считать равным нулю (если только в сети не возникла какая-то

неисправность). При нормальном режиме работы сети потеря даже одного пакета должна расследоваться. (Т.е. потеря пакетов допускается при определенных действиях по реконфигурированию фабрики или при инициализации линка, но при нормальной работе все пакеты должны доставляться.)

У других протоколов выше коэффициент ошибок и потери пакетов, поэтому архитекторы SAN, проектирующие территориально-распределенную инфраструктуру через IP WAN, должны учитывать эту статистику.

### *Запаздывание и задержки*

Другой фактор – это запаздывание. В контексте SAN запаздывание – это время, которое тратит коммутатор или маршрутизатор на обработку пакета перед тем, как передать его дальше. Сразу же после того, как пакет приходит в один порт коммутатора Brocade, он начинает выходить из другого порта коммутатора еще до того, как коммутатор обработает его полностью. Такой механизм называется маршрутизацией “cut-through” и представляет собой самый быстрый в теории способ передачи пакетов. Другие вендоры вместо этого метода используют механизм коммутации “store and forward”, который не позволяет передавать пакет дальше пока он не будет полностью обработан, что приводит к значительному запаздыванию при передаче трафика.

С точки зрения приложения запаздывание коммутатора – это только одна из составляющих общего запаздывания. То время, которое уходит на передачу пакета от отправителя и до получателя, считается запаздыванием пути. Разумеется, запаздывание пути имеет значение, но приложение зависит от того, сколько

## Проектирование С8: Планирование производительности

времени проходит от того момента, когда оно решило послать ввод/вывод на НВА, и до того момента, когда оно получило ответ. Это время включает сумму запаздываний всех коммутаторов, через которые проходят пакеты, и, как вариант, запаздывание внутри устройств хранения на другом конце.

Общее запаздывание также учитывает запаздывание из-за применяемого в коммутаторе механизма контроля переполнения, который в неудачно спроектированных коммутаторах и директорах пересыпает пакеты обратно порту хоста или устройства хранения. В результате пакеты не передаются (пусть даже и медленно) через фабрику, а остаются внутри хоста, который вынужден ждать, пока их сможет принять фабрика. Таким образом, запаздывание перекладывается с фабрики на хост и приложение все равно «тормозится».

Иногда одни пакеты на своем пути проходят только один коммутатор, а другие пакеты проходят один или несколько ISL или IFL между коммутаторами. Число ISL и IFL, через которые нужно пройти, называется числом переходов (хопов, *hop count*) между устройствами. В большинстве случаев это число практически не влияет на производительность приложения и задержки при прохождении одного или нескольких Brocade ISL намного меньше по сравнению с задержками других каналов передачи данных. Например, для “быстрого” дискового устройства время доступа измеряется миллисекундами. Каждый хоп фабрики Brocade дает около двух микросекунд, или даже до 700 наносекунд. Таким образом, в большой фабрике, где от одного устройства к другому надо пройти семь хопов, задержка маршрута будет 14 микросекунд, что на несколько порядков меньше задержки доступа к дискам через

линки. Для большинства конфигураций задержка из-за числа хопов не влияет на работу сети FC как с точки зрения задержки коммутатора, так и оптического кабеля.

Как мы видим, число хопов в SAN мало влияет на задержку маршрута и основная причина, по которой стоит добиваться уменьшения числа хопов, это переподписка – чем больше ISL проходит пакет, тем больше вероятность того, что он попадет в переполненный ISL. Помимо ухудшения пропускной способности, переполнение обычно приводит к задержкам из-за того, что пакеты ждут в буфере коммутатора освобождения переполненного линка. Таким образом, переполненный линк не только работает в режиме store-and-forward, но и ведет к образованию длинной очереди пакетов. Уменьшение числа хопов снижает риск того, что пакет попадет в переполненный линк и поэтому придет с задержкой. Разумеется, лучше всего для уменьшения переполнения снизить число хопов до нуля, поэтому в тех случаях, когда на первом месте стоит производительность, лучше использовать локальную коммутацию.

Запаздывание маршрута ощущается при использовании соединений на большие расстояния. Для линков Fibre Channel время, которое пакет идет по кабелю между двумя портами, пренебрежимо мало и зависит от скорости света, который в оптическом кабеле проходит километр примерно за пять микросекунд. При связи на небольшие расстояния приложение не ощущает этого запаздывания. При использовании для связи на большие расстояние темной оптики или оборудования xWDM, запаздывание может ощущаться на уровне линка, но (a) оно очень мало и намного меньше времени доступа к дискам и (b) невозможно передавать пакеты быстрее скорости света.

## Проектирование С8: Планирование производительности

Для поддержания максимально возможной производительности связи на больших расстояниях Brocade разработала продукт Brocade Extended Fabrics, который обеспечивает производительность на уровне всей полосы пропускания на расстояниях в сотни километров, причем при понижении скорости передачи возможна связь на еще большие расстояния. Brocade также разработала продукт FastW rite для дальнейшей оптимизации связи на большие расстояния. Таким образом, FC SAN на базе решений Brocade всегда обеспечивают максимальную скорость передачи данных между площадками.

Технологии IP SAN работают совершенно иначе. Например, для расширения SAN на большие расстояния используется FCIP и появляется задержка из-за того, что на одном конце FC-пакеты инкапсулируются в IP, а на другом – извлекаются. Кроме того, сама природа IP WAN привносит существенное запаздывание при соединении конечных точек, которое на несколько порядков больше, чем у родного FC, поэтому при проектировании территориально-распределенных SAN нужно обязательно учитывать запаздывание.

## **Определение требований к производительности**

Все сети и сетевые продукты имеют свои предельные возможности, поэтому главное, чтобы они не стали ограничениями производительности *приложений* и первый шаг при решении этой задачи – определение реальных характеристик производительности.

## ***Типичные подходы к определению требований***

Есть несколько подходов к определению требуемой для приложения производительности.

Некоторые архитекторы считают, что для завершения этапа проектирования нужно знать детальные характеристики производительности, что часто требует тщательного изучения работы приложения, внутренней архитектуры серверной платформы и тестирования.

Этот подход обеспечивает самую точную оценку требований к производительности еще до внедрения, однако требует много затрат усилий и рабочего времени и, в результате, заказчик позже получает те преимущества, ради которых он и внедряет SAN.

Другие архитекторы считают, что можно обойтись приблизительными оценками и уже после запуска в эксплуатацию точно настроить сеть. Например, можно запросить у администраторов приложений, сколько данных в секунду требуется предоставить приложению, какова пиковая скорость передачи данных, когда возникают эти пиковые нагрузки и что произойдет, если производительность SAN окажется меньше требуемого для приложения уровня. Специалисты, использующие этот метод, при проектировании ISL обычно пользуются эмпирическими правилами и информацией, которую им сообщили системные администраторы. Например, можно начать с соотношения подсоединений хостов и ISL 7:1 и затем по результатам интервью увеличить число ISL.

Архитектор рассматривает числа, полученные из интервью и на основе эмпирических правил, как

## Проектирование С8: Планирование производительности

установленные факты при проектировании ISL и другой сетевой инфраструктуры.

Это значительно более эффективный способ принятия решений, ускоряющий внедрение SAN. Однако есть риск, что спроектированная SAN будет недостаточно мощной если системные администраторы неправильно оценивают требования своих приложений. Зная, что информация о производительности строго говоря носит оценочный характер, архитектор должен предусмотреть запас ресурсов чтобы можно было добавить ISL или сделать другие модификации если этого потребуется после запуска SAN в эксплуатацию.

Часто архитекторы проектируют сеть с запасом для устранения риска проблем производительности – они делают общую оценку требований к производительности и на ее основе дают пессимистичный прогноз требований к производительности. Такие архитекторы оценивают, какая производительность ISL *может* потребоваться и предусматривают больше линков, чем нужно для обеспечения требуемой пропускной способности.

Не следует называть такой подход “плохим” или “для ленивых” – это просто признание того, что невозможно получить *всю информацию* о требованиях к производительности в будущем при проектировании сети. Всегда какой-то информации будет не хватать, поэтому детальный анализ не даст более точный ответ и не оправдает затрат времени и денег. Выделение производительности с запасом под неучтенные потребности – это реалистический подход. Его единственный серьезный недостаток – увеличение первоначальной стоимости внедрения из-за

необходимости построения дополнительных ISL, которое первое время будут не задействованы.

Три этих подхода (тщательный анализ, исходя из эмпирических правил и проектирование с запасом) чаще всего применяются для определения требований к производительности. Для всех них требуется информация об использовании полосы пропускания и время реакции.

### ***Использование пропускной способности***

Использование пропускной способности определяется тем, сколько блоков данных нужно передать приложению через SAN в единицу времени за определенный период. Этот параметр сильно зависит от производительности SAN, скорости линка (или маршрута), коэффициента переподписки и переполнения.

При расчете требований хоста к полосе пропускания важно понять объемы данных, пересылаемых за *единицу* времени и за *период* времени. Например, приложению может потребоваться переслать 50 гигабайт в час чтобы выполнить ночное резервное копирование в отведенное для этой операции окно, т.е. использование полосы пропускания составит примерно 100 Мбит/сек. Этому же приложению может понадобиться переслать 50 гигабайт за 1 минуту в часы максимальной загруженности для выполнения операций Data Mining для системы поддержки принятия решений, т.е. скорость должна составить более 6 Гбит/сек и нужен будет достаточно быстрый интерфейс LAN с iSCSI. Для требований Data Mining может понадобиться установка нескольких 4Gbit FC HBA и балансировка между ними нагрузки. Архитектор SAN должен предусмотреть ресурсы для

## Проектирование С8: Планирование производительности

обеспечения полосы пропускания, которая требуется при развертывании или в течение обозримого жизненного цикла приложения.

Проектирование SAN для обеспечения требуемой приложению полосы пропускания с применением “детального анализа” означает подробное изучение каждого аппаратного и программного компонента, который есть между приложением и данными, включая внутренние характеристики хоста и устройства хранения, скорость их интерфейсов SAN, типы коммутаторов между ними и структуру ISL и IFL фабрики. При изучении структуры ISL надо для каждого маршрута выяснить, как приложения могут его использовать и суммировав их требования к производительности, получить общую возможную нагрузку сети. Требуется обеспечить поддержку требуемой скорости передачи данных даже в случае сбоя линка.

Разумеется, такой анализ достаточно сложен и выполняется долго, поэтому многие архитекторы предпочитают просто заложить в свою сеть характеристики производительности, которые превышают параметры, полученные на основе ее базовых требований и эмпирических правил. В этом случае архитектор оценивает требования приложения к производительности, и затем, исходя из предположения, что хосты и дисковые массивы соответствуют этим требованиям, на основе эмпирических правил вычисляет требования сетевых структур ISL и IFL.

### ***Время отклика***

Требованиями приложений к времени отклика определяется промежуток, который начинается с

момента посылки приложением пакета в сеть и моментом получения ответа на него. Этот показатель сильно зависит от выбранного протокола, запаздывания и задержки в сети, коэффициента ошибок и потерь, производительностью граничных устройств.

Все приложения в определенной степени зависят от времени отклика и если оно достаточно большое и хост не получает ответа на запрос ввода/вывода, то после тайм-аута он повторит операцию. Такие повторные запросы ухудшают производительность приложений. Увеличение продолжительности тайм-аута может уменьшить число повторных попыток, но оно также приводит к ухудшению производительности, поэтому лучше использовать протокол с низкой задержкой, практически нулевыми коэффициентами потерь пакетов и ошибок, сеть с очень маленьким переполнением и граничные устройства, достаточно мощные для оперативного ответа на запросы.

Требуемое время реакции зависит от настроек операционной системы хоста (например, времени тайм-аута НВА и контроллера устройств хранения) и от природы самого приложения. Например, синхронные приложения в отличие от асинхронных очень чувствительны ко времени отклика.

Важно помнить, что *не обязательно* соединять коммутаторы центра с другими коммутаторами центра или периферийные коммутаторы к другим периферийным коммутаторам фабрики СЕ – это типичная ошибка, которую делают архитекторы IP-сетей, плохо разбирающиеся в сетях хранения. Хотя такое соединение допустимо, оно не дает никакого эффекта, но увеличивает стоимость, поскольку трафик в обоих случаях не идет через горизонтальные ISL либо

## Проектирование С8: Планирование производительности

использует эти линки *вместо* CE ISL. На самом деле горизонтальные ISL могут *снизить* производительность также, как она падает в топологии full mesh. Если у граничного коммутатора есть по одному 4Gbit ISL для соединения с двумя коммутаторами ядра, то его суммарная полоса пропускания будет 8Gbit, которую он может предоставить другому коммутатору. Если один ISL соединяет напрямую этот коммутатор с другим граничным коммутатором, то полоса пропускания между ними будет только 4Gbit. Соединение горизонтальных линков уменьшает число переходов между коммутаторами, но также сокращает полосу пропускания. Так как горизонтальный путь с одним переходом короче пути CE с двумя переходами, то 4-гигабитные линки CE ISL никогда не будут использоваться. Число переходов для производительности не так важно, как пропускная способность и сокращение числа переходов на 50% не оправдает такого же уменьшения полосы пропускания.

## **Обеспечение необходимых ISL и IFL**

Обеспечение линков в любой сети – это достижение соответствия ее пропускной способности нагрузке, которую создает приложение, т.е. число линков между коммутаторами и фабриками определяется объемом трафика, который будет по ним передаваться. В этом разделе объясняется, как на основе требований к производительности спроектировать структуру ISL и IFL.

Определить необходимое число ISL и IFL можно разными способами. На следующих страницах рассматриваются три из них – исходя из пиковой нагрузки, средних значений и эмпирических правил. Все

эти способы позволяют выяснить, сколько линков нужно на *одно периферийное устройство*. Соотношение числа линков и устройства обычно называются “коэффициентом переподписки” сети.

Когда все порты SAN работают на одной скорости, то коэффициент переподписки линка – это соотношение числа входных портов устройств к числу линков, через которые может передаваться трафик. На Рис. 45 коэффициент переподписки верхнего левого коммутатора – двенадцать портов устройств на четыре ISL, т.е. три к одному.

Коэффициент переподписки обычно записывается в числовой форме и для данного примера можно сказать, что у сети “коэффициент переподписки 3:1”. В этой сети 12 хостов подключены к верхнему левому коммутатору и четыре ISL подключены к центру сети, т.е. по три хоста на один ISL. Если все три хоста попытаются использовать ISL на полной скорости и в установленвшемся режиме, то даже если они будут обращаться к разным устройствам хранения, то каждый хост получит только треть от потенциально доступной полосы пропускания.

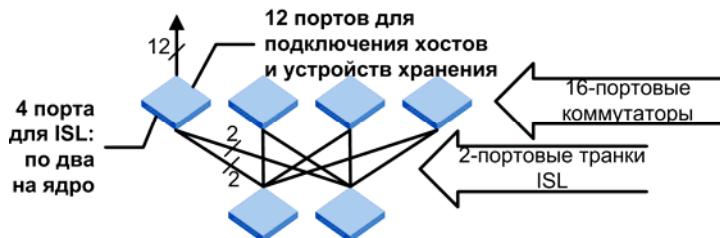


Рис. 45 – Коэффициент переподписки ISL равен 3:1

Основная формула расчета переподписки “Переподписка ISL = число узлов к числу ISL” или

## Проектирование С8: Планирование производительности

$I_o = N_n : N_i$ . Это соотношение обычно приводится к наименьшему общему кратному  $N_i = 1$ . ( Например, вместо 12:4 пишут 3:1.)

Подключенные к SAN устройства (хосты, устройства хранения и ISL) могут работать на разных скоростях, что надо учитывать при расчете переподписки ISL. Для определения коэффициента переподписки ISL нужно усреднить скорость входных портов и разделить этот результат на скорость выходных портов. Если у каждого хоста на Рис. 45 одногигабайтный HBA, а ISL работают на 4Gbit, то у ISL недоподписка будет 3:4 поскольку на входе полоса пропускания будет 12x1Gbit со стороны хостов, а на стороне ISL - 4x4Gbit. В подобных случаях пользуются приведенным выше коэффициентом или сокращают до “1” значение  $N_n$ , а не  $N_i$ , например, записывают этот коэффициент как 1:1.3.

Можно обеспечить разные соотношения в сети CE, варьируя число сконфигурированных ISL. Например, если периферийный коммутатор в сети CE имеет 16 портов, из которых 14 используются устройствами и два - ISL, то на один ISL приходится семь устройств. У такой сети коэффициент переподписки семь к одному или 7:1. Обычно от степени локальности и требований к производительности соотношение составляет от 1:1 и до 63:1.

Архитектор должен определить коэффициент переподписки, обеспечивающий достаточную для всех приложений полосу пропускания. Может показаться, что лучше использовать соотношение 1:1 для всех SAN, т.е. у каждого устройства будет свой ISL и тогда переполнение будет невозможно, однако использование

такого большого числа линков слишком дорого и ограничивает масштабируемость, поэтому лучше обеспечить ровно столько линков, сколько нужно исходя из требований приложений к производительности при начальном внедрении и в обозримом будущем.

Главное в выборе правильного соотношения – определить свойства и потребности приложений с учетом коэффициента переподписки портов хранения.

Любое устройство, подключенное к SAN по интерфейсу 4Gbit, *теоретически* может обеспечить эту скорость, но на практике работает значительно медленнее, поэтому лучше конфигурировать ISL на основе эмпирических правил, обеспечив соответствие соотношения ISL и устройств, подключенных к граничному коммутатору, соотношению инициаторов и получателей в фабрике. Такой подход проще понять и применить и он решает большинство основных проблем проектирования ядра/границы SAN.

Если SAN строится для поддержки большого числа устройств, способных на полной скорости генерировать трафик ввода/вывода, то следует применять принцип локальности (стр258). Если используется локальность, то структуры ISL и IF L могут поддерживать большую переподпись независимо от требований к производительности. Для упрощения проектирования можно использовать в качестве коэффициента переподписки ISL соотношение хостов и устройств хранения.

Если локализацию нельзя применить, а к SAN подключены очень быстрые устройства, то допустимый коэффициент переподписки будет уменьшаться по мере роста требований к производительности. В этом случае нужно провести дополнительный анализ для выяснения

## Проектирование С8: Планирование производительности

оптимального соотношения с помощью одного из следующих методов.

### Расчет исходя для пиковых нагрузок

В некоторых случаях имеет смысл проектировать сеть исходя из *пиковых* нагрузок. Такой подход наиболее дорогой и его применять следует в тех случаях, когда низкая скорость сети также недопустима, как отказ сети. Например, если при обслуживании приложений, связанных с загрузкой и обработкой спутниковых снимков, сеть работает медленно, то поступающие со спутников данные будут потеряны. Другими примерами приложений, для которых недопустима низкая скорость сети, являются крупномасштабные операции Data Mining, параллельные суперкомпьютерные кластеры, рабочие группы редактирования цифрового видео и студии телевидения.

Для определения нагрузки в пиковом режиме всех IFL отдельной фабрики надо определить число узлов в ней, которые могут использовать эти линки одновременно, и умножить это число на их максимальную производительность, которая вычисляется на основе типа интерфейса и внутренних ограничений. Если имеются исторические данные о производительности этих узлов, то надо сложить максимальные уровни загрузки каждого, который были зафиксированы в прошлом. Если же таких данных нет, то нужно суммировать максимальные скорости интерфейсов (если нет внутренних «узких мест», например, медленного процессора или устаревшей шины PCI).

Например, проектируемая периферийная фабрика может объединять 100 серверов и 15 портов устройств

хранения. В любое время большая часть трафика может быть локализована внутри фабрики, но два порта устройств хранения будут зеркальированы на другую фабрику для защиты от катастроф и два хоста будут использовать ISL для резервного копирования за пределы площадки. Если реальные данные о производительности недоступны и все порты - это 2Gbit Fibre Channel, то пиковая загрузка соединяющих площадки ISL будет равна (2x порты устройств хранения + 2x хоста) x 2Gbit = 8Gbits/sec пропускной способности. Для передачи этого трафика потребуется четыре линка 2Gbit либо два 4Gbit.

Если приложению требуется еще более консервативный подход, то требуется учитывать дополнительные факторы, связанные с временными требованиями, сбоями линков и другими незапланированными событиями, а также временной несбалансированностью линков. В таком случае может понадобиться сконфигурировать шесть и даже восемь линков с удаленной фабрикой.

Главное в этом подходе – сконфигурировать каждый сетевой маршрут так, чтобы он справился с максимально возможным для него объемом трафика.

#### Расчет исходя из средних значений

В большинстве сетей не требуется конфигурировать линки в расчете на пиковые нагрузки, поэтому использование этого метода только ведет к росту расходов. Разные потоки трафика очень редко одновременно достигают пиковых значений в установившемся режиме, а если такое совпадение продолжается короткое время, то это почти не влияет на бизнес-процессы. В таких случаях расчет делается исходя из *средних* показателей загруженности .

## Проектирование С8: Планирование производительности

Вернемся к предыдущему примеру. Предположим, что два зеркальзованных порта устройств хранения мало загружены и между ними в среднем передается 0.5Gbit (например, скорость резервного копирования ограничена быстродействием ленточных приводов и не превышает 1Gbit между двумя хостами либо резервное копирование выполняется ночью, когда дисковые массивы мало используются). В этом случае средняя загруженность будет менее 1Gbit.

Однако этот показатель можно рассматривать и с другой точки зрения. В течение резервного копирования связанный с этой операцией трафик идет в установившемся режиме, поэтому полоса пропускания должна быть не меньше требований самого «тяжелого» приложения. Возможно, удастся обеспечить ее только одним линком, соединяющим площадки, однако чаще всего таких линков потребуется несколько.

При консервативном подходе для резервирования и обеспечения роста трафика в будущем конфигурируются два линка, однако конфигурирование восьми линков как при использовании метода пиковых нагрузок не имеет смысла.

### Расчет на основе эмпирических правил

Чаще всего архитектор заранее не знает точно, как будет использоваться проектируемая сеть и попытка определить это приведет только к напрасным тратам времени и денег, тем более что после начала эксплуатации сети возникают новые задачи. Можно определить набор линков, который обеспечит по крайней мере заданную пропускную способность (например, с помощью описанного выше метода на основе средних значений), но трудно гарантировать, что

заданная пропускная способность не будет превышена. Однако конфигурирование в расчете на нулевую локальность (т.е. исходя из предположения, что все линки фабрики одновременно могут работать в пиковом режиме) ведет к удорожанию сети и почти всегда неточно. Лучше идти от средней и даже минимальной загрузки и определять дополнительные требования, пользуясь эмпирическими правилами.

Один из подходов – использовать соотношение имеющихся во всей сети хостов и устройств хранения. Этот метод используется для конфигурирования IS\_L в традиционной фабрике центр/периферия и дает меньшее число линков, чем при расчете исходя из пиковых нагрузок. Метод является более реалистичным, поскольку обычно весь ввод/выводов хостов идет на порты устройств хранения. Хотя соотношение хостов и устройств хранения – это переменная величина, большинство вендоров систем хранения рекомендуют значение 6:1 или 7:1 как для этого соотношения, так и для переподписки ISL внутри фабрики.

Однако локализация в граничной фабрике Meta SAN намного больше, чем на уровне коммутаторов внутри фабрики, поэтому в большинстве случаев вполне можно сократить число IF \_L. Разумные коэффициенты переподписки IFL могут быть намного больше, чем у ISL даже если неизвестны конкретные данные о производительности и конфигурациях LSAN.

Эмпирические правила конфигурирования IFL можно сформулировать следующим образом:

1. Сначала нужно сконфигурировать все известные нагрузки IFL как базовое значение.

## Проектирование С8: Планирование производительности

2. Затем добавить по крайней мере один дополнительный IFL для резервирования и незапланированных задач.
3. Если неизвестен характер трафика, то это число нужно увеличить на соотношение инициатор/получатель в Meta SAN, разделенное на прогнозируемый коэффициент локализации. Если в Meta SAN есть 1000 хостов и 100 портов устройств хранения, то соотношение 10:1. Если в граничной фабрике 100 хостов, то потребуется 10 IFL и это число нужно сократить с учетом локализации фабрики . Если он равен 90%, то потребуется только один дополнительный IFL кроме тех IFL, которые были определены на шагах 1 и 2.

Таким образом, число линков IFL вычисляется по формуле:

```
Число IFL = ( известный_трафик + 1 ) +
(число_хостов_фабрики / соотношение
хостов:портов_устройств_хранения ) * ( 1 -
процент_локальности ) )
```

разумеется, это не является аксиомой и если все устройства хранения Meta SAN сосредоточены в одной фабрике, то большинство IFL должны быть сконфигурированы для соединения этой фабрики. Это многоуровневый подход (стр. 268), который обычно не рекомендуется использовать для маршрутизируемых Meta SAN.

Важно помнить, что эмпирические правила не могут полностью компенсировать недостаток знаний о SAN и если у архитектора есть причины увеличить или уменьшить переподписку, то не имеет смысла ориентироваться на соотношение хостов и портов устройств хранения.

## Локализация трафика

Оптимальную производительность любой сети невозможно обеспечить без оптимизации подключений устройств на основе информации о характере трафика. Если характер трафика хорошо известен, то можно оптимизировать трафик, расположив максимально близко те порты, которые часто «переговариваются» между собой или их «переговоры» важны для приложений. Такой подход называется *локализацией*.

Например, в сети СЕ можно подключить хосты и основные для них порты устройств хранения к одному граничному коммутатору. В результате путь ввод/вывод будет максимально коротким – трафик будет идти только внутри коммутатора и не будет использовать ISL. На Рис. 46 показано, как в фабрике СЕ передается локализованный и нелокализованный трафик.

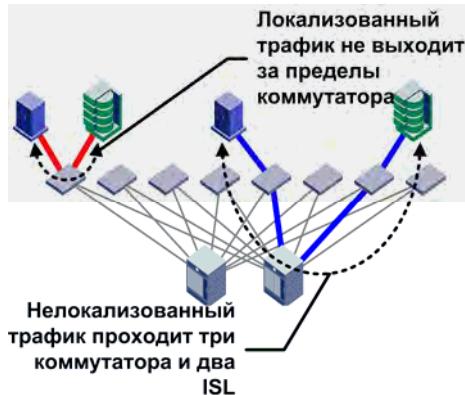


Рис. 46 – Использование локальности

Хотя этот подход давно применяется в сетях передачи данных, SAN по своей природе лучше подходят для такой локализации. Например, в сетях передачи данных обычно возможно соединение каждого

## Проектирование С8: Планирование производительности

с каждым (any-to-any) и оно должно учитываться при проектировании сети. Однако, практически во всех SAN хосты обмениваются трафиком только с портами систем хранения – а не с другими хостами – причем известно, какое конкретное устройство хранения в основном использует конкретный хост.

Можно локализовать только часть трафика SAN, т.е. оба показанных на рисунке потока трафика будут идти внутри одной фабрики, но определенные проценты трафика SAN будут локализованы. Если локализовать весь трафик, то локализация SAN будет 100%. Если же нет никакой локализации, то говорят, что она равна 0%.

Подавляющее большинство используемых сегодня сетей хранения не настолько чувствительны к производительности, что им требуется или даже желательна локализация на уровне 100%, и обычно архитекторы стараются локализовать только критично-важные устройства, для которых необходима высокая производительность.

Локализация улучшает не только производительность, но и RAS, поскольку при локализации части трафика сбой в сети меньше влияет на работу приложений и в результате повышается их доступность. Локализация также означает, что в сети нужно меньше ISL и поэтому уменьшается число ее компонентов, что сокращает расходы и повышает такие показатели надежности, как MTBF. Кроме того, локализация упрощает устранение сбоев поскольку для диагностики проблемы с локализованным трафиком нужно изучить меньше компонентов.

Локализация может быть заложена в архитектуру сети и тщательно поддерживаться в течение жизненного

цикла SAN. Другой вариант, применяемых во многих SAN корпоративного класса – это изменение процента локализации с течением времени. Например, первоначально SAN может быть построена для перехода от Direct Attached Storage (DAS). У DAS локализация равна 100%, поэтому при переходе от DAS к SAN сохраняется часть первоначальной локализации. По мере развития SAN и добавления, перемещения, перепрофилирования новых узлов и отключения старых, процент локального трафика будет снижаться.

## Уровни локализации

Локализация – это не технология «все-или-ничего». Определенный процент локализации может быть в граничной фабрике Meta SAN<sup>62</sup>, в коммутаторе этой фабрики и в лезвии или группе портов ASIC этого коммутатора. Локализацию можно представить как несколько уровней начиная от порта-источника трафика (см. Рис. 47), причем уровень с наивысшими показателями RAS и производительности находится ближе к центру, и пользуясь этой моделью локальности легче спроектировать SAN.

Обычно, чем выше требования к производительности и RAS приложения, тем больше должна быть локализация. Наивысшие значения этих двух показателей обеспечиваются при локализации небольшой группы портов единственного коммутатора в одной фабрике. К сожалению, реализация такого подхода к обеспечению высокой локальности требует больших затрат времени, уменьшает эффект от внедрения сети и во многих случаях просто невозможна,

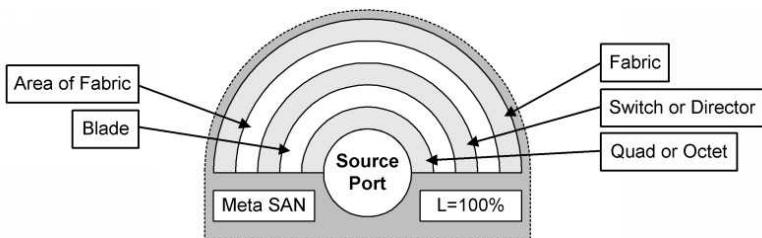
---

<sup>62</sup>

См. “Сравнение фабрики с SAN и Meta SAN” на стр. 42.

## Проектирование С8: Планирование производительности

поэтому такой тип локализации используется для подключения хостов, обслуживающих критически важные приложения, к их основным дисковым массивам, а также в тех случаях, когда SAN используется для таких массивно-параллельных высокопроизводительных приложений, как параллельные суперкомпьютеры, кластер Da ta Mining или загрузка и анализ спутниковых снимков в реальном времени.



**Рис. 47 – Уровни локализации**

В архитектуре с маршрутизацией локализация возникает, когда граничные фабрики подключены к одному маршрутизатору. Она также может возникнуть между площадками в MAN или WAN. Локальность уровня Meta SAN всегда будет дополняться до 100-процентов, а у более низких уровней локализована только часть трафика.

Если весь трафик идет через ISL и IFL, то локализация будет 0%, а если никакой трафик не идет по этим линкам - то 100%. Типичные коэффициенты локализации приведены в Таблица 2.

Таблица 2 – Уровни локальности

<u>Уровень локализации</u>	<u>%</u>	<u>Локализация</u>	<u>Суммарная локализация</u>
ASIC		<b>5%</b>	<b>5%</b>
Лезвие		<b>10%</b>	<b>15%</b>
Коммутатор /директор		<b>40%</b>	<b>55%</b>
Часть фабрики		<b>25%</b>	<b>80%</b>
Фабрика		<b>10%</b>	<b>90%</b>
Зона ВВ		<b>5%</b>	<b>95%</b>
Meta SAN		<b>5%</b>	<b>100%</b>

Во втором столбце таблицы указана локализация только для текущего уровня, а в третьем - его локализация *плюс* локализация предыдущих уровней. Например, если трафик идет между двумя портами одного квартета или октета портов, то он также локализован внутри лезвия, директора, части фабрики, фабрики, зоны backbone и Meta SAN. Трафик, локализованный внутри лезвия, но не внутри одного октета портов, показан во втором столбце. Трафик, локализованный внутри лезвия *включая* локализацию внутри одного октета портов (один ASIC), показан в третьем столбце.

## Проектирование С8: Планирование производительности

Самая близкая к порту-источнику локализация (например, на уровне ASIC) обычно дает небольшой процент поскольку ее трудно спроектировать и поддерживать. Локализация на уровне группы портов и лезвия оправдывает усилия по ее реализации только для самых требовательных к производительности приложений. Локализация на уровне коммутатора легко проектируется и хорошо окупает затраты. Если фабрика сложная, то можно сконцентрировать трафик внутри зоны, окруженнной линками ISL, и тогда только небольшой процент трафика будет идти от каждого к каждому (any-to-any) внутри фабрики.

На самом деле, в граничных фабриках в Meta SAN существует так много «встроенной» локализации, что очень мало трафика фактически выходит за пределы граничной фабрики. За исключением отдельных случаев, локализации легко добиться и она обеспечивает серьезные преимущества на уровне коммутаторов, фабрики и backbone, поэтому администратору не надо проектировать ее – она происходит сама по себе. График распределения локализации по разным уровням часто похож на контур колокола и основная часть трафика локализуется еще до того, как он покинет фабрику. Этот факт следует учитывать при конфигурировании IFL.

### ***Локализация внутри коммутаторов***

Возможно локализовать трафик на разных уровнях даже внутри одного коммутатора, хотя в этом нет особой необходимости и мало кто из клиентов применяет такую оптимизацию производительности. Brocade поддерживает эту функцию для обеспечения максимума производительности и доступности. Есть два принципа

локализации, которые действуют на этом уровне: локализация коммутации внутри ASIC для повышения производительности и локализация внутри лезвий для ограничения взаимосвязей внутри директора.

Все коммутаторы и директоры Brocade<sup>63</sup> используют очень быстрые заказные ASIC с центральной памятью. В зависимости от используемого продукта архитектура памяти коммутатора может быть одноуровневой или многоуровневой. Во втором случае коммутатор оборудован несколькими основными ASIC коммутации, которые соединены между собой внутри коммутатора через линки центральной памяти.

Например, у Brocade 48000 есть восемь слотов для лезвий с FC-портами. Каждое из лезвий с портами может иметь один или несколько ASIC в зависимости от своих функций и числа портов. Все венды, строящие крупномасштабные продукты подобные директору Brocade 48000, используют свои варианты этой архитектуры, но ни один вендор не предлагает единый ASIC, который одновременно может работать с несколькими лезвиями портов. Типичный подход – это разместить ASIC одного типа на лезвиях портов и ASIC другого типа на центральных лезвиях директора, соединяющих между собой блейды с портами. Отличие Brocade в том, что используемая в ее директорах многоуровневая архитектура позволяет ASIC лезвий портов коммутировать локально без необходимости по backplane обращаться к лезвиям ядра. Любой порт 16-портового лезвия Brocade 48000 может коммутироваться локально с любым другим портом того же лезвия. У 32-

---

<sup>63</sup>

Кроме некоторых продуктов, которые были включены в продуктовую линейку после приобретения McDATA.

## Проектирование С8: Планирование производительности

портового лезвия две 16- портовые группы локализации, а у 48-портового две 24-портовые.

В SilkWorm 12000 можно коммутировать локально внутри 4- портовой группы, которая называется квартет (*quad*). Например, если хост подключен к порту 1 лезвия 1, а устройство хранения – к порту 2 лезвия 1, то они в одном квартете и трафик между ними не идет через backplane и поэтому имеет оптимальную производительность. С другой стороны, в большом масштабе не имеет смысла удерживать ввод/вывод внутри 4- портовой группы, поэтому локализация на уровне квартета применяется редко. В SilkWorm 3900 и 24000 локальность поддерживается на уровне 8- портовых групп *октетов*, которые проще использовать. Другие коммутаторы, например, Brocade 3250, 3850 и 4100 применяют одноуровневую архитектуру, поэтому локализованным считается весь коммутатор. Для архитекторов относительно несложно локализовать основной поток трафика в пределах 32 портового коммутатора Brocade 5000 или 24- портовых групп 48- портового лезвия Brocade 48000.

Локализация улучшает производительность, но необходимо рассматривать ее в перспективе. В примере выше передача между локализованными портами дает задержку примерно от 0.7  $\mu$ s до 0.8  $\mu$ s (700-800 наносекунд, что считает достаточно быстрым). Передача через backplane Brocade 48000 даже при самых неблагоприятных условиях дает задержку 2.1  $\mu$ s - 2.4  $\mu$ s. Типичное время доступа для “быстрых” дисковых подсистем на несколько порядков больше задержки “медленных” коммутаторов, поэтому обычно архитекторы мало внимания уделяют задержкам при

передаче внутри коммутатора если речь идет об улучшении производительности.

Однако может быть и другая причина кроме производительности для использования локализации внутри коммутатора. Если трафик локализован внутри лезвия, то он меньше восприимчив к сбоям других лезвий. Если выйдет из строя центральное лезвие директора, то трафик перенаправляется на другое лезвие, хотя при таком переключении могут быть потеряны некоторые пакеты. Для современных SAN- устройств хранения или НВА это не является проблемой, хотя и и может вызвать задержку. На локализованный трафик этот сбой вообще не повлияет. То же справедливо для локальности на уровне коммутатора внутри фабрики и другом уровне – чем больше локализован трафик, тем меньше взаимозависимостей, способных нарушить его передачу.

Как мы уже говорили, большинству архитекторов не нужен такой уровень доступности и производительности и поэтому они используют локализацию только уровня коммутатора, директора и фабрики. Локализация внутри коммутатора редко применяется, но иногда может понадобиться.

## ***Локализация и LSAN***

Как было объяснено выше, локализация внутри одного коммутатора или директора дает мало эффекта, но требует больших затрат. Более эффективным является локализация в больших масштабах, особенно на уровне фабрики в Meta SAN. Трафик между инициатором и получателем можно локализовать, разместив оба устройства в одной фабрике даже если несколько фабрик соединены через коммутаторы. При такой локализации

## Проектирование С8: Планирование производительности

улучшается масштабируемость SAN потому что сервисы фабрики физически изолированы (стр. 220) и сбои будут локализованы в отдельных регионах.

Обычно локализация фабрики в Meta SAN действует сама по себе. Обычно архитекторы стараются подключить переговаривающиеся между собой устройства к одной фабрике. Кроме того, большинство построенных к сегодняшнему дню Meta SAN образовались в результате консолидации фабрик “островков SAN”, у которых локализация на уровне фабрики равна 100%. До того, как Brocade представила функцию Fibre Channel маршрутизации (routing), которая позволила создавать Meta SAN, не было способов избавиться от «островков» SAN, поэтому локализация сначала была 100- процентной. Архитекторы, внедряющие Meta SAN в таких средах, сохраняют часть локализации после подключения маршрутизатора.

Существует один фактор, о котором разработчикам следует помнить применительно к локальности Meta SAN – это архитектура зон LSAN. LSAN (Logical Storage Area Network) – это зона, охватывающая несколько фабрик в Meta SAN. В пределах созданных зон LSAN разные фабрики Meta SAN начинают взаимодействовать – обмениваться трафиком. Взаимодействие, правда, ограничено, но лучше держать фабрики вообще разъединенными кроме тех случаев, когда ввод-вывод действительно необходим. Поэтому создавайте зоны с префиксом LSAN только в тех случаях, когда трафику действительно необходимо ходить между фабриками – т.е. когда он не локализован.

## **Новые возможности локализации: UC и ILM**

Технологии нового поколения, связанные с Utility Computing и Information Lifecycle Management (стр. 85) обещают дать много преимуществ администраторам и пользователям SAN, однако у UC и ILM есть серьезный недостаток – из-за этих тенденций локализация трафика теряет смысл. Главное в ILM и UC – это виртуализация инфраструктуры серверов и хранения, благодаря которой связи между приложениями, операционными системами и физическим оборудованием можно динамически менять по мере необходимости и поэтому физическое расположение порта хоста около портов хранения не обязательно даст эффект, поскольку политики ILM и UC не нуждаются в ней. Этот вопрос подробнее обсуждается в следующем разделе, а пока мы отметим, что локализация трафика будет терять популярность и эффективность по мере внедрения ILM и UC, которые ведут к распространению многоуровневых SAN.

## **Многоуровневые CE SAN**

“Многоуровневость” – это соединение хостов к одной группе (или *уровню, tier*) коммутаторов и устройств хранения к другой группе. Использование выделенных коммутаторов для хостов и устройств хранения улучшает управление за счет упрощения планирования мощностей и диаграммы сети. Обычно многоуровневую модель применяют в архитектуре центр/периферия. Либо уровень хранения располагается на коммутаторах центра (двууровневая модель) или на отдельной группе периферийных коммутаторов (трехуровневая модель). На Рис. 48 показана фабрика CE с подключением устройств по двухуровневой модели – один уровень для

## Проектирование С8: Планирование производительности

подключения хостов, а другой для устройств хранения. Эти два уровня напрямую связаны между собой.

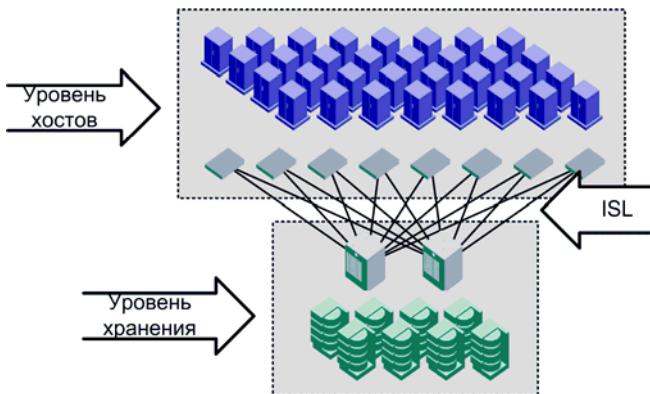


Рис. 48 – Двухуровневая фабрика СЕ

Также можно использовать третий уровень коммутаторов между уровнями хостов и устройств хранения. На Рис. 49 показан пример трехуровневой архитектуры. Двухуровневая архитектура ограничивает масштабируемость, поскольку нужно порты устройств хранения напрямую подключать к коммутаторам ядра (см. “Масштабируемость топологии СЕ” на стр. 189.). Во многих случаях она, в отличие от трехуровневой архитектуры, негативно влияет на производительность.

Снова рассмотрим Рис. 48, где имеется два коммутатора уровня хранения. Если устройство хранения видно как LUN только одному из этих коммутаторов, то каждый хост сможет обращаться к этому LUN только через один ISL хотя есть по два ISL у каждого коммутатора уровня хостов. Если такая ситуация возникнет в показанной на Рис. 49 трехуровневой архитектуре, то каждый хост сможет воспользоваться обоими ISL, поскольку с точки зрения LUN у них одинаковый вес пути. Разработанные Brocade

функции DLS и DPS равномерно распределяют ввод/вывод между ISL (р 272), гарантируя улучшение общей производительности SAN. Если ввод/вывод равномерно распределен между всеми LUN самими хостами, то балансировка на уровне ISL будет не нужна и негативным эффектом двухуровневой архитектуры можно пренебречь. Однако в большинстве случаев в двухуровневой SAN возникает перегруженность некоторых ISL хотя в это время другие ISL могут быть полностью свободны.

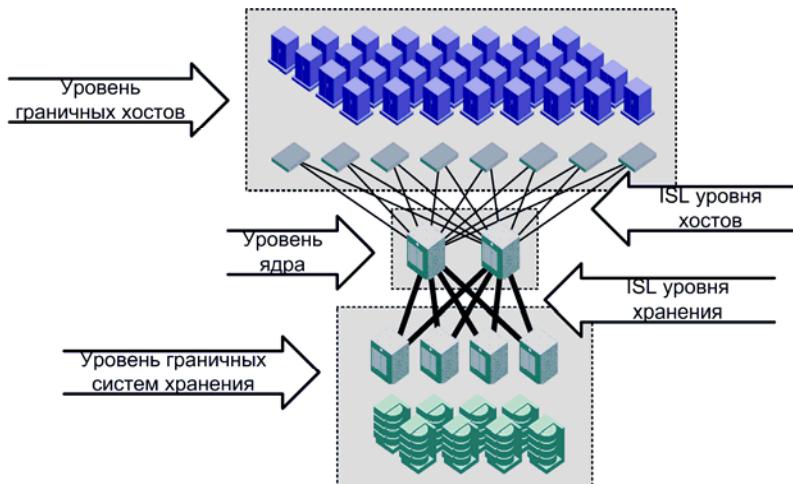


Рис. 49 – Трехуровневая фабрика СЕ

Есть преимущества многоуровневости с точки зрения управления, но следует помнить, что эта практика является полной противоположностью локализации с точки зрения производительности и RAS - каждый пакет должен пройти по крайней мере через один ISL. С точки зрения легкости развертывания и администрирования многоуровневая SAN обычно предпочтительнее и трехуровневая модель почти всегда эффективнее

## Проектирование С8: Планирование производительности

двууровневой. Однако для производительности лучшим вариантом будет локализация.

Исключением из этого анализа производительности является ситуация, когда в SAN используется новая технология, например data m over на базе фабрики, виртуализаторы или платформы приложений, которые не выигрывают от локализации. Архитекторы, планирующие внедрить архитектуру автоматизации Utility Com putting и/или Information Lifecycle Management должны рассматривать многоуровневые SAN как для улучшения управления, так и производительности. Дело в том, что внедрение нового поколения продуктов инфраструктуры SAN с “уровнем приложения”, как ожидается, приведет к централизации в большинстве случаев и трафик должен будет проходить через централизованное устройство даже если конечные точки относятся к одному граничному коммутатору. На Рис. 50 показан пример трафика в SAN следующего поколения.

В этом примере трафик между хостами и устройствами хранения идет через один или несколько виртуализаторов, в роли которых могут выступать коммутаторы Brocade 7600 или лезвия директора, например, Brocade FA18, но в любом случае они скорей всего будут централизованы хотя бы ради сокращения расходов. Даже если некоторые порты устройств хранения подключены к тем же граничным коммутатором, что и их хости, трафик должен проходит через структуру ISL к виртуализатору, поэтому применение локализации будет неэффективно.

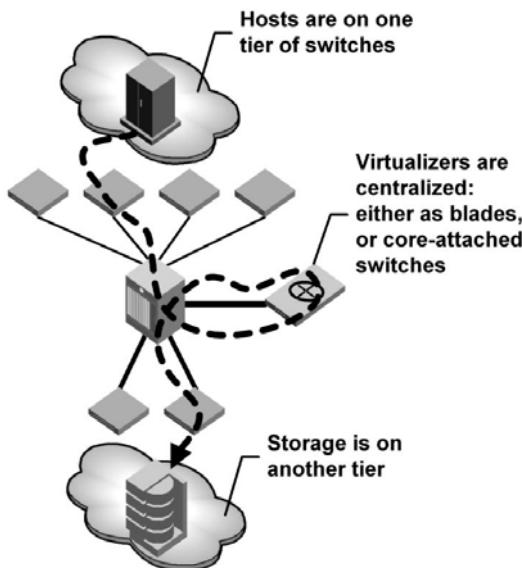


Рис. 50 – Потоки трафика в фабрике следующего поколения

## Балансировка линков

Даже в сетях с переподпиской может возникать переполнение, когда одни пути будут перегружены, а другие свободны. Такая ситуация похожа на двухуровневую архитектуру когда запросы хостов к LUN не сбалансиированы. Другими словами, в сети возникает узкое место производительности хотя она обладает достаточной полосой пропускания для передачи всего трафика без ограничений. Это связано с тем, что в сети не применяются интеллектуальные функции балансировки нагрузки между доступными маршрутами. Неиспользуемый путь может применяться для резервирования, но для производительности он бесполезен.

Как это ни странно, но многие сетевые технологии не используют никакой механизм балансировки линков.

## Проектирование С8: Планирование производительности

Например, протокол Spanning Tree Protocol (STP), являющийся стандартом для сетей Ethernet, можно сконфигурировать только на резервирование путей активный/пассивный (active/passive)<sup>64</sup> и ввод/вывод идет по резервным линкам только в случае сбоя основного. Тем не менее, при сбое у STP может уйти несколько минут для повторного расчета топологии и активизации резервных ликов и все это время по сети не будет идти трафик. Пользователи сетей хранения считают, что должна быть исключена недоступность пути даже в течение пары секунд.

Это одна из причин, почему IP/Ethernet считался принципиально непригодным для сетей хранения *всеми* основными вендорами в 1990-ые и в результате были разработаны более интеллектуальные сетевые протоколы, например, Fibre Channel. FC способен быстро восстановить сеть при сбое линка и балансировать трафик между разными путями, причем FC поддерживает разные варианты этой процедуры.

Доступные опции зависят от платформы, поскольку часть из них реализуются на аппаратном уровне. Все платформы Brocade поддерживают активную балансировку маршрута от порта-источника (source-port) с помощью FSPF (Fabric Shortest Path First). Эта функция называется Dynamic Load Sharing (DLS). В дополнение к DLS, все поставляемые сегодня коммутаторы и директории фабрики поддерживают транкинг на уровне пакетов между ASIC. В зависимости от архитектуры один транк может объединять от четырех до восьми

---

<sup>64</sup>

Существуют более совершенные альтернативы STP, но они редко применяются. На время написания этой книги все коммутаторы Ethernet поддерживали STP.

линков. Эта функция также называется Advanced IS L Trunking. Новейшие коммутаторы<sup>65</sup> и Multiprotocol Router поддерживают транкинг на уровне обменов (exchange), который называется Dynamic Path Selection (DPS). Для коммутаторов для использования второго варианта балансировки нужно приобрести дополнительную лицензию.

Все три опции улучшают доступность и производительность. Если сконфигурированы несколько линков, то сеть автоматически обработает сбой линка вместо одновременного выполнения failover для всех хостов. Хотя производительность в то время, пока идет устранение проблемы, может снизиться, но сеть все равно продолжает работать (улучшение доступности еще нагляднее проявляется для хостов, не использующих программное обеспечение дублирования каналов.) То же происходит и при добавлении линков – при транкинге на уровне фреймов линки добавляются без прерывания потоков ввода/вывода. Если нет такого мощного механизма балансировки линков, то новый линк не будет задействован либо будет нарушен поток ввода/вывода.

Отметим, что коммутаторы с поддержкой транкинга могут использовать все порты как транки либо часть портов как транки, а часть для подключения узлов. При использовании в центре архитектуры СЕ, коммутаторы часто целиком состоят из групп транков. В периферийных коммутаторах обычно 75% - 90% портов используются для узлов, а остальные для ISL и транков.

Также важно отметить, что разные методы балансировки не являются взаимоисключающими и в

---

<sup>65</sup>

Например, платформы на базе Condor.

## Проектирование С8: Планирование производительности

коммутаторе с их поддержкой на аппаратном уровне они могут сочетаться в любой комбинации. Например, можно использовать транкинг на уровне пакетов для создания нескольких групп транков и DLS либо DPS на верхнем уровне для балансировки трафика между этими группами.

### *Динамическая балансировка нагрузки: Балансировка маршрутов FSPF*

Все фабрики Fibre Channel поддерживают протокол FSPF<sup>66</sup> (Fabric Shortest Path First – кратчайший путь фабрики идет первым). Он является частью базовой операционной системы пока есть функции фабрики и E\_Port. FSPF рассчитывает топологию фабрики и определяет вес путей для любой конечной точки. Во многих популярных сетевых топологиях, например, центр/периферия (стр. 185) может быть несколько путей с одинаковым весом между источником и периферийным коммутатором получателя. Выбор пути производится по отдельным портам на коммутаторе- отправителе. По умолчанию FSPF пытается распределить соединения с разными портами по доступным путям на уровне портов источника. Опция коммутаторов Brocade позволяет FSPF перераспределять ресурсы при любом событии в фабрике. Эта функция называется Dynamic Load Sharing (DLS) потому что позволяет динамически менять маршруты не нарушая порядка доставки пакетов.

---

<sup>66</sup> Первоначально разработанная Brocade функция FSFP стала стандартной функцией фабрик всех вендоров.

DLS по возможности лучше (“best effort”) выполняет операции распределения ввода/вывода за счет балансировки маршрутов от портов источника. Однако, некоторые порты передают больше трафика, чем другие, и DLS не может заранее предсказать, какой маршрут будет перегружен. Кроме того, характер трафика меняется со временем, поэтому узкие места могут возникнуть позднее и изменение маршрута в реальном времени приведет к нарушению порядка доставки пакетов<sup>67</sup>. Балансировка числа маршрутов, выделенных конкретному пути, отличается от балансировки потоков ввода/вывода, поэтому DLS не способна точно и равномерно балансировать трафик. Именно поэтому функция называется *load sharing* (*разделение нагрузки*), а не *load balancing* (*балансировка нагрузки*).

Функция DLS удобна, и поскольку она бесплатна и работает автоматически, то в разных формах она используется практически во всех фабриках из нескольких коммутаторов Brocade. Однако DLS не решает всех проблем производительности и нужен более эффективный метод балансировки нагрузки, для которого нужна поддержка на аппаратном уровне, поскольку выбор пути должен производиться для отдельных фреймов и выполнение этой операции с помощью программного обеспечения привело бы к падению производительности control plane.

---

<sup>67</sup>

Один вендор коммутаторов Fibre Channel嘗試 использовать аналогичный метод для изменения маршрута без приостановки ввода/вывода, однако это привело к массовым нарушениям порядка доставки пакетов, что противоречит стандартам FC и, что более важно для практики, приводит к сбоям многих приложений.

## Проектирование С8: Планирование производительности

### ***Расширеный транкинг: Балансировка нагрузки на уровне пакетов***

#### Внедрение транкинга на уровне пакетов

Транкинг позволяет балансируировать трафик между разными ISL при сохранении порядка доставки. Brocade поддерживает две формы транкинга – на уровне пакетов и обменов (exchange). Первый метод балансирует трафик так, что каждый последующий пакет может идти по другому физическому ISL, а коммутатор-получатель гарантирует, что пакеты перенаправляются в исходном порядке. На Рис. 51 показан транк на уровне пакетов между двумя коммутаторами SilkWorm 3850. чтобы это работало, требуются мощные интеллектуальные функции обоих коммутаторов.

На уровне программного обеспечения коммутаторы должны понять, что можно сформировать группу транка, запрограммировать эту группу на уровне оборудования, отображать и управлять группой линков как единым логическим объектом и оптимально управлять такими низкоуровневыми параметрами, как буферные кредиты (buffer-to-buffer credits) и виртуальные каналы. Программное обеспечение управления должно правильно представлять группу транка и максимально прозрачно для пользователей.

На аппаратном уровне коммутаторы по обоим концам транка должны обрабатывать разделение и заново собирать несколько мультигигабитных потоков на скорости физической среды без потери даже одного пакета или его доставки с нарушением порядка. Задача усложняется из-за того, что кабели между ISL всегда имеют разную длину. В группе транков это создает *смещение (skew)* во времени, когда каждый линк

доставляет пакеты, т.е. на ASIC передаются пакеты с нарушением порядка если только не предприняты специальные меры для компенсации смещения<sup>68</sup>.



**Рис. 51 – Концепция транкинга на уровне пакетов**

У ASIC есть предельно допустимая величина смещения, но она достаточно большая и на практике ее можно не учитывать. Однако есть и другие ограничения смещения, например, при конфигурировании одного

<sup>68</sup>

Имеются и другие подходы к проблеме смещения, но они снижают производительность. Можно снабжать пакеты метками, но это приводит к нарушению стандартов и несовместимости в VSAN.

## Проектирование С8: Планирование производительности

линка в транке для передачи трафика по часовой стрелке в кольце DWDM масштаба metro, а другого линка – для передачи трафика в обратном направлении. Если разница в длине кабеля равна нескольким метрам, то смещением можно пренебречь, однако если она составляется несколько десятков километров, то транк нельзя сформировать и коммутатор построит два отдельных ISL и будет балансировать нагрузку между ними с помощью DLS или DPS.

### Преимущества транкинга на уровне пакетов

Главное преимущество транкинга на уровне пакетов – это оптимальная производительность за счет того, что происходит суммирование пропускной способности линков транка. Кроме того, улучшается доступность за счет добавления в транк линков без нарушения работы сети и снижения влияния сбоев линков.

### Ограничения транкинга на уровне пакетов

В платформах на базе Bloom<sup>69</sup> можно скомбинировать несколько групп из двух – четырех линков 2Gbit и получить сбалансированные каналы по 4Gbit - 8Gbit. На Рис. 52 показан канал между двумя периферийными коммутаторами 3850 и двумя центральными коммутаторами 24000, используя группы 2-портовых транков.

Эта конфигурация работает аналогичным образом и с другими коммутаторами ядра, например, Brocade 4100, 4900, 5000 или 48000. Данные пример иллюстрирует основные ограничения транкинга на уровне пакетов –

---

<sup>69</sup> Например, SilkWorm 3200, 3250, 3600, 3800, 3850, 3900, 12000 и 24000.

все порты транка должны быть в одной группе портов ASIC на каждом конце линка, потому что пары ASIC гарантируют доставку с сохранением порядка пакетов (см. Рис. 51).

В коммутаторах на базе Bloom группы портов строятся из группы соседних 4 портов, называемых квартет (*quad*). Например, в SilkWorm 3250, есть два квартета: порты 0–3 и порты 4–7. В коммутаторах на базе Condor<sup>70</sup>, группы транкинга строятся из 8 соседних портов и такие группы называются *октетами*<sup>71</sup>. В коммутаторе Brocade 4100 четыре октета: порты 0–7, 8–15, 16–23 и 24–31. Если попытаться построить транк на уровне пакетов не из групп портов, то вместо транка получится несколько взаимозависимых ISL.

В коммутаторах на базе Condor ASIC можно построить несколько групп, содержащих до восьми линков 4Gbit. Результат будет очень похож на Рис. 51, за исключением увеличения в 4 раза производительности одного транка: вместо формирования нескольких каналов 8Gbit Condor может создать несколько каналов 32Gbit. (64 Gbit в режиме full-duplex). При соединении коммутаторов Condor с коммутаторами Bloom используется подход наименьшего общего частного – каждая группа транков может быть ограничена до 4x 2Gbit вместо 8x 4Gbit.

Хотя транк на уровне пакетов всегда работает быстрее, чем DLS, он ограничивает опции конфигурирования из-за необходимости использовать

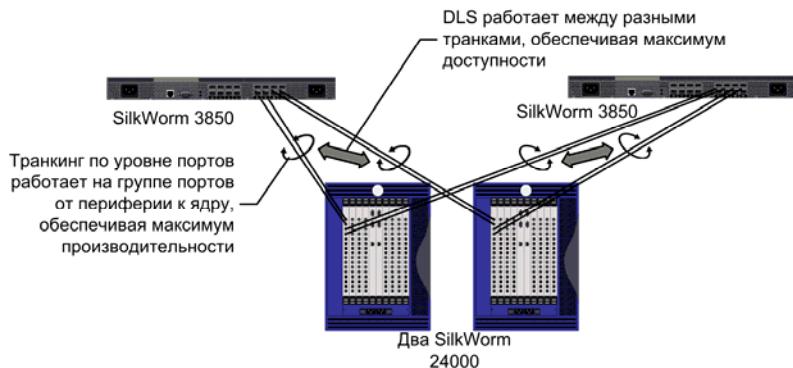
---

<sup>70</sup> Например, SilkWorm 4100, 4900, 5000 и 48000.

<sup>71</sup> Октеты также используются в SilkWorm 3900 и 24000 для локализации коммутации (см. “Локализация трафика”, стр. 258 и далее).

## Проектирование C8: Планирование производительности

только линки. Решение состоит в комбинировании транкинга на уровне пакетов с одним из других методов, как показано на Рис. 52, когда транкинг на уровне фреймов работает внутри групп портов, а DLS между транками. Хотя на каждом граничном коммутаторе все 4 ISL находятся внутри группы транков, эти четыре линка не могут сформировать транк, поскольку для отказоустойчивости они идут к разным группам портов разных центральных коммутаторах.



**Рис. 52 – Транкинг на уровне пакетов плюс DLS**

Также можно сконфигурировать несколько транков внутри группы портов. Например, на Silk Worm 3850 можно сконфигурировать один транк на портах 12-13, а второй – на портах 14-15. Эти транки могут идти к разным центральным коммутаторам в сети CE или к разным лезвиям директора.

При проектировании территориально-распределенной SAN важно выбрать технологию, обеспечивающую необходимую производительность, поэтому важно понять взаимодействие транкинга на уровне пакетов с Extended Fabrics.

У каждой ASIC есть определенное число буферных кредитов buffer-to-buffer. У коммутатора Bloom-II оно достаточно для работы всех четырех портов в quad для любого режима при коротких расстояниях<sup>72</sup>, но для больших расстояний нужно ограничить функциональность некоторых портов. Например, Bloom-II может поддерживать 4-портовый транк на расстоянии 25 км, но для 50 км поддерживается только 2-портовый транк, а два остальных порта можно использовать только для подключения узлов<sup>73</sup>. При расстоянии 100 км можно сконфигурировать только один порт, поэтому нельзя сформировать транк. Можно сконфигурировать несколько 100-километровых линков на коммутаторе Bloom-II до одного на quad, однако они будут в разных группах портов, поэтому требуется DLS для балансировки между ними, но не транкинга.

У коммутаторов с Condor ASIC более гибкая поддержка транкинга на большие расстояния. В Condor буферы общие для всей микросхемы, а не ограничены quad или октетами. Например, можно сконфигурировать транки из 8 портов 4Gbit на 50 км (группа транкинга 32Gbit) или транки из 4 портов 4Gbit на 100 км (группа транкинга 16Gbit). В некоторых случаях требуется сконфигурировать транки из линков 2Gbit, например, если транки проходят через DWDM, не поддерживающую 4Gbit. В этом случае можно построить 100-километровый транк из 8 портов 2Gbit.

---

<sup>72</sup> Под “короткими” расстояниями здесь понимаются расстояния до 25 км без деградации производительности либо более 25 км если не требуется производительность на порт 2Gbit.

<sup>73</sup> Для транкинга на большие расстояния может потребоваться минимальная версия микрокода.

## Проектирование С8: Планирование производительности

### ***Динамический выбор пути: транкинг на уровне Exchange***

Dynamic Path Selection (DPS) – новый метод, применяемый на всех платформах 4Gbit. На момент написания этой книги к ним относились коммутаторы Brocade 4100, 4900 и 200E, директор Brocade 48000, маршрутизатор Brocade 7500 и почти все последние модели встроенных продуктов.

#### Реализация транкинга на уровне Exchange

При методе DPS происходит чередование обменов FC (exchanges) между путями с одинаковым весом. Отправитель помещает в заголовок каждого пакета FC идентификатор обмена. При нормальном режиме работы этот идентификатор остается постоянным во время операций SCSI. Когда поддерживающая DPS платформа получает пакет, она рассматривает все пути с одинаковым весом и рассчитывает исходящий порт с помощью формул, учитывающих PID-идентификаторы отправителя (SID), получателя (DID) и идентификатор обмена (OXID). Формула дает один и тот же путь для идентичного набора [ SID, DID, OXID ].

На самом деле DPS «расщепляет» ввод/вывод на уровне SCSI<sup>74</sup>. Для конкретного сеанса между хостом и портом устройства хранения одна команда SCSI идет по одному пути, а вторая – по другому. Все пакеты в одном обмене будут идти с соблюдением порядка, поскольку передаются по одному пути. Теоретически, возможно

---

<sup>74</sup>

Это происходит *почти* все время, но есть некоторые устройства, которые не отображают операции SCSI на OXID и такие протоколы, как FICON, также не поддерживают такое расщепление.

нарушение порядка при *разных* операциях SCSI, но при тестировании разных устройств такая ситуация никогда не возникала. Если два хоста пишут на два разных устройства хранения через одну сеть, то нарушение порядка пакетов от разных серверов (т.е. пакеты от одного сервера идут быстрее, чем от другого) не имеет принципиального значения.

Эта особенность создает незначительную разницу в производительности по сравнению с транкингом на уровне пакетов. В результате транкинг на уровне `eh_change` работает быстрее аналогичных функций любого вендора *за исключением* функции транкинга на уровне пакетов Brocade, а поскольку DPS можно комбинировать с транкингом на уровне пакетов, то, как объясняется в следующем разделе, возможно добиться максимума производительности и доступности.

#### Преимущества транкинга на уровне Exchange

Хотя в некоторых случаях производительность может немного снижаться, метод DPS обеспечивает несколько преимуществ.

В отличие от транкинга на уровне пакетов для DPS необязателен транкинг внутри одной группы портов ASIC. Как и DL\_S, DP\_S можно сконфигурировать на разных центральных коммутаторах в сети CE или разных лезвиях директора. Однако в отличие от DLS, метод DPS действительно балансирует трафик, а не пытается с помощью `best_effort` балансировать маршруты от источника. На Рис. 53 показан пример такой балансировки трафика между четырьмя Brocade 4100 в архитектуре центр/периферия.

## Проектирование С8: Планирование производительности



Транк на уровне пакетов с помощью коммутаторов SilkWorm 4100 может объединять до 8 линков по 4Gbit. В этом примере четыре линка 4Gbit объединены в канал 16Gbit Между каждым граничным коммутатором и двумя коммутаторами ядра. DPS Балансирует два транка в единый путь 32Gbit между граничными коммутаторами.

**Рис. 53 – Транкинг на уровне пакетов вместе с DPS**

DPS, в отличие от транкинга на уровне пакетов, может балансировать трафик между разными коммутаторами ядра. Кроме того, DPS *не исключает* возможности транкинга на уровне пакетов. Можно балансировать несколько групп портов пользуясь методом на уровне пакетов и затем балансировать между полученными группами транков, используя метод на уровне обменов. Это обеспечивает оптимальный баланс производительности (транкинг на базе пакетов работает быстрее) и доступности (DPS обеспечивает гибкость по настоящему сбалансированных топологий НА), что также показано на Рис. 53.

Затем DPS может балансировать ввод/вывод от поддерживающей эту функцию платформы к другой платформе, которая *может не поддерживать эту функцию*. Выбор пути делает передающий коммутатор и получатель не должен брать на себя обеспечение доставки с сохранением порядка. Это обеспечивает обратную совместимость с ранее инсталлированными коммутаторами и улучшает производительность даже если не все коммутаторы фабрики используют новейшие технологии. Нужно учитывать, что выбор пути в сети СЕ делает периферийный коммутатор, поэтому если Brocade

4100 установлен в сети CE, где центр может не поддерживать DPS, то пользователи все равно могут воспользоваться преимуществами этой функции. Любой трафик, посылаемый *от* коммутатора с функцией DPS, будет сбалансирован независимо от того, поддерживают ли центральный или периферийный коммутатор получателя эту функцию. Трафик, идущий *от* периферийного коммутатора без DPS, будет использовать DL\_S, независимо от того, идет трафик к коммутатору с DPS или нет (Рис. 54)

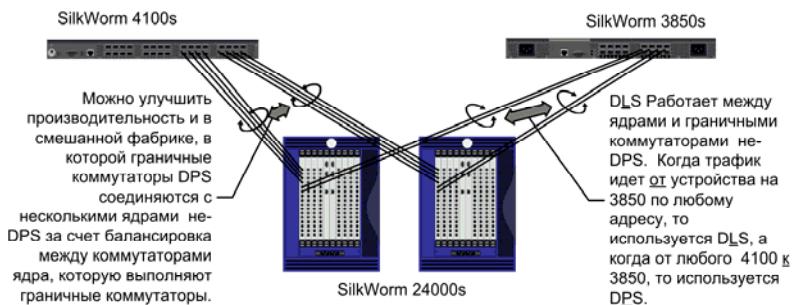
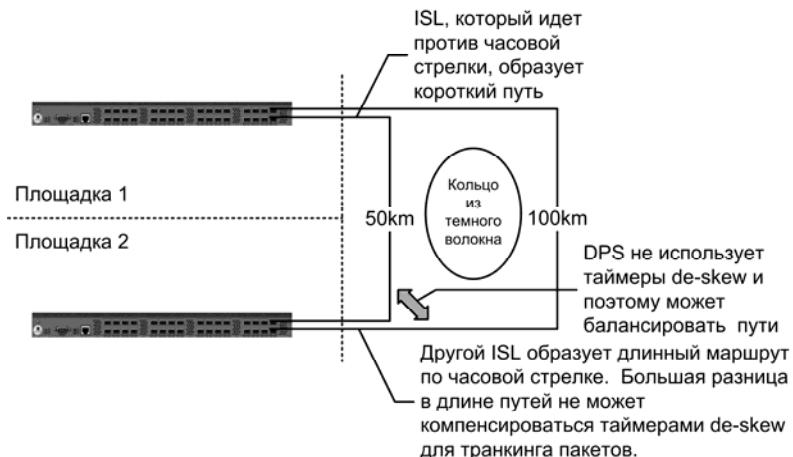


Рис. 54 - DPS в смешанной фабрике

Наконец, DPS может балансировать ввод/вывод в территориально-распределенных конфигурациях, которые не поддерживают транкинг на уровне пакетов (Рис. 55.) Например, если между двумя площадками сконфигурированы два линка с разными путями, то из-за смещения не удастся сформировать транк на уровне пакетов. DPS может использовать эти линки, поскольку метод не использует временные смещения (skew) для доставки с сохранением порядка пакетов.

## Проектирование C8: Планирование производительности



**Рис. 55 – Балансировка DPS в большом кольце Fiber**

### Балансировка производительности на уровне обменов и на уровне пакетов

Расширенные функции транкинга ISL работают на уровне пакетов, а DPS – на уровне обменов (exchange). В большинстве фабрик FC обмен – это аналог операций SCSI. Как уже отмечалось ранее в “Главе 1: Основы SAN”, обмен может состоять из большого числа пакетов, поэтому транкинг на уровне пакетов будет “перебалансируться” чаще, чем DPS.

Разницей в гранулярности обычно можно пренебречь, но она может привести к небольшой разнице в производительности между транкингом DPS и на уровне пакетов. Для того, чтобы разобраться в этом явлении необходимо понять, как узлы Fibre Channel формируют пакеты из данных SCSI.

Пакеты Fibre Channel могут иметь разный размер, начиная от 60 байт и до примерно 2 кбайт. Размеры блоков SCSI могут быть намного больше 2 кбайт,

поэтому для передачи одной команды SCSI может потребоваться несколько пакетов. Эти группы пакетов комбинируются в FC обмен.

Например, чтение блока SCSI размером 8 кбайт может потребовать пересылки нескольких пакетов FC – одного от отправителя для посылки команды SCSI read получателю, четыре пакета по 2 кбайт от получателя для доставки данных и подтверждение от инициатора о получении данных. Обычно пакеты команд небольшие и в этом примере будет два маленьких пакета и четыре больших. Все известные инициаторы будут использовать FC exchange ID для идентификации всех фреймов этой группы, чтобы показать, что они относятся к одной операции SCSI. Последующие операции SCSI могут состоять из десяти пакетов и могут иметь те же конечные точки, но у них будет другой exchange ID.

Две операции read будут состоять из двух обменов, разделенные всего на 20 пакетов. Использующие DPS коммутаторы посылают первые 10 пакетов по одному пути, а вторые 10 – по другому пути. Если только одна пара инициатор/получатель использует сеть только для того, чтобы выполнять по одной операции SCSI, то только один путь будет использоваться в один момент времени. Это даже не балансировка ввода/вывода. А коммутаторы с транкингом на уровне пакетов ровно распределяют пакеты через все доступные пути к группе транков, поэтому даже при одном сеансе связи нагрузка будет распределена равномерно.

Это не *реальная* проблема производительности, а только пример работы этой функции. Если только одна пара инициатор/получатель использует сеть и они выполняют за один раз только одну операцию SCSI, то

## Проектирование С8: Планирование производительности

им понадобится только один путь в сети для обеспечения полной производительности.



### Заметки на полях

Необходимо помнить, что даже в самой неблагоприятной ситуации производительность DPS будет аналогична или лучшая, чем у SAN без этой функции.

Например, некоторые другие вендоры предлагают только продукты с версией DLS, которая никогда не может работать быстрее, чем DPS. Один из этих вендоров попытался сделать свою DLS-подобную функцию более динамичной, но это привело к массовым нарушениям порядка доставки и не улучшило производительность. При нарушении порядка доставки НВА полностью прекращается ввод/вывод при каждой активизации функции “*performance enhancing*”. Другие конкурирующие продукты предлагают примитивную версию DPS и не поддерживают транкинг на уровне пакетов, т.е. DPS работает быстрее, чем конкуренты Brocade, но медленнее транков Brocade на уровне пакетов.

Таким образом, DPS не медленный метод, но он работает медленнее транкинга на уровне пакетов.

Более интересно рассмотрение случая, когда несколько потоков одновременно пересекают сеть. Если сеть пересекают два потока, то в теории возможно синхронизировать их и передавать по одному пути. Если это удается обеспечить в течении большого времени, то

результат будет такой же, как если бы никакие методы балансировки не применялись.

Однако это всего лишь теория и на практике такая синхронизация может поддерживаться не более секунды либо недостижима, особенно если число потоков больше двух.

Это явление называется *переходным* (*transient*) эффектом производительности и обычно не влияет на пропускную способность и работу приложений. Оно только приводит к скачкам на графиках производительности, которые можно получить с помощью SAN Health или Advanced Performance Monitoring. Чем больше трафика идет по сети, тем больше требуется его балансировки и тем меньше скачков будет на этих графиках. Таким образом, эффективность DPS увеличивается по мере роста объемов трафика.

Также стоит отметить, что комбинирование DPS с транкингом на уровне пакетов может даже полностью устраниить пики на этих графиках. Даже если потоки между собой несинхронизированы, они будут передаваться в группе транков на уровне пакетов, а не по одному линку. Например, если идут два потока, то они получат необходимую полосу пропускания даже при использовании 2- портовой группы транков на уровне пакетов. Узкое место в сети возникнет только если имеется сильная синхронизация в течение длительного времени и все потоки идут по одному пути. Вероятность такого совпадения можно считать нулевой.

Таким образом, разница в производительности транков на уровне пакетов и обменов носит только теоретический характер и даже в самой быстрой сети переходные эффекты можно устраниить комбинированием методов транкинга и поскольку у

## Проектирование С8: Планирование производительности

транкинга на уровне обменов есть ряд важных преимуществ, то теоретически возможная разница в производительности не имеет особого значения. Brocade – это единственный вендор, применяющий оба этих метода и пользователи могут выбрать тот, который для них оптимален.

### ***Резюме балансировки линков***

Все сети, у которых есть несколько ISL, выигрывают от балансировки линков. Самая эффективная высокопроизводительная конфигурация высокой доступности сочетает два метода балансировки - на уровне пакетов для наивысшей производительности и DLS либо DPS для высокой доступности.

Даже если для SAN сначала не требуется максимальная производительность, ее можно обеспечить по мере необходимости с помощью транкинга. Загруженность сети всегда растет до пределов возможностей сети и какая бы большая не была у сети пропускная способность, через какое-то время ее будет недостаточно. В “Главе 12: Планирование производительности” (стр. 373 и далее) даются рекомендации по проектированию SAN, обеспечивающие возможность в будущем роста и изменения сети.

## **Преимущества буферных кредитов для производительности**

Коммутаторы Brocade обрабатывают буферные кредиты (BB) прозрачно для конечных пользователей и эти кредиты нужно учитывать только в территориально-распределенных родных линках FC. Однако в линках на малые расстояния у Brocade они обрабатываются почти

автоматически. В этом разделе рассматриваются особенности обработки буферных кредитов в продуктах Brocade. Хотя информация вряд ли потребуется при проектировании SAN, она представляет определенный интерес для архитекторов.

### Объединение буферов в пул

Трафик в SAN носит “взрывной” характер, т.е. короткое время передаются большие объемы данных. Устройству можно выделить только ограниченное число кредитов, которые могут быть полностью израсходованы если такие «всплески передачи» затянутся и трафик остановится и будет ждать пока устройство-получатель ответит командой R\_RDY.

Для решения этой проблемы Brocade применяет в своих ASIC механизм объединения буферов в пул. Пул буферов – это набор кредитов BB, которые предоставляются по запросу любому порту ASIC, которому они нужны по принципу «первым пришел – первым обслужен». Когда F-port или N-port регистрируются в фабрике, то им выделяется определенное число кредитов (в коммутаторах Brocade это число обычно равно 16). Когда устройство запрашивает кредиты после начального процесса FLOGI/PLOGI, коммутатор Brocad e начинает использовать пул буферов и когда этот пул будет исчерпан, коммутатор начнет использовать 16 кредитов буфера, которые ему были выделены первоначально.

Единственная тонкость при использовании этого подхода заключается в том, что при применении тестового оборудования Fibre Channel, например, генераторов трафика, для измерения задержки при нагрузке, общий пул (например, 64 кредита) будет сразу же доступен генератору (как и любому конечному

## Проектирование С8: Планирование производительности

устройству на практике) и после исчерпания пула будут предоставлены остальные 16 кредитов. Генератор трафика посыпает данные коммутатора, заполняя все доступные буферы и в результате пул буфера должен освободиться, на что уходит определенное время. Таким образом, в этом случае на самом деле измеряется глубина пула буферов, а не задержки в коммутаторе: явная задержка увеличивается потому что в очереди может быть, например, 80 пакетов и пакет на конце очереди буфера должен ждать своей очереди отправления на порт получателя пока не будут доставлены другие пакеты. Ошибка в данном случае связана не со скоростью, с которой пакеты передаются внутри коммутатора, а с тем, что коммутатор слишком рано ставит пакеты в очередь.

Если бы описанный тест выполнялся с коммутатором, у которого мало буферов или он не поддерживает пул буферов, то генератор трафика получил бы мало кредитов (возможно два). Если только два буфера должны освободиться, то на это уйдет меньшее времени чем когда предоставлены 80 кредитов. Меньшее число буферов у коммутатора означает, что компьютер не сможет посылать коммутатору пакеты с одинаковой скоростью и коммутатор будет отправлять их обратно серверу.

Главное – это то, что с точки зрения *приложения* полное время доставки любого пакета будет по крайней мере такое же при использовании пула как если бы коммутатор просто давал команду генератору трафика приостановить отправление новых пакетов. Если измерять общее время выполнения ввода/вывода, то коммутатор, у которого больше буферов, работает по крайней мере не медленнее коммутатора, у которого

буферов меньше (при условии, что фактическая задержка (а не задержка при нагрузке) у них мало отличается. Коммутатор с меньшим числом буфером будет задерживать пакеты еще до того, как они до него дошли, на время, которое по крайней мере не меньше, чем у коммутатора с поддержкой пула буферов.

Это можно продемонстрировать следующим образом. Пакеты попадают в буфер только когда порт на выходе не может отправлять пакеты с той же скоростью, с которой они попадают в очередь из порта на входе. Чтобы воспроизвести такую ситуацию попробуйте довести скорость потока ввода/вывода до уровня, который чуть ниже теоретического максимума. Например, генератор работает на 99.9% максимальной скорости. В этом случае у Brocade ASICs запаздывание менее 2 мксек. Если же довести скорость до 100% от максимальной, то очередь начнет заполняться поскольку отправитель будет чуть отставать от передатчика и *фактическое* запаздывание будет повышаться пропорционально глубине очереди: если у коммутатора маленькая очередь, то и задержка будет маленькой, а у коммутаторов с большими очередями задержка будет намного больше. Но если замерять время, необходимое для выполнения всей операции ввода/вывода, например, копирования 2 Гбайт данных, то коммутатор с большим пулем буферов будет работать по крайней мере не медленнее, чем коммутатор с маленьким пулем.

### Выводы

Замеры запаздывания могут оказаться некорректными если они проводятся компаниями, которые заинтересованы в результатах тестирования или просто не понимают механизм работы буферов Fibre Channel, поэтому заказчики и партнеры должны

## Проектирование С8: Планирование производительности

тщательно изучить отчет о тестировании. Тесты запаздываний при полной загрузке создают переполнение, которое приводит к заполнению очереди. Это не влияет на производительность приложений, но приводит к некорректным результатам тестов на запаздывание. Следует критически подходить к результатам любых тестов, в которых измеряется «запаздывание при нагрузке», но нет подробного описания мер, которые были предприняты для того, чтобы действительно измерялось запаздывание, а не глубина очереди.

# 9

## 9: Планирование доступности

“Доступность” – это время, которое система и приложение доступно для использования. В этой главе доступность SAN рассматривается с точки зрения доступности *приложений*: как отдельные компоненты SAN и общая архитектура сети могут повлиять на работу конечного пользователя приложений. В этом контексте, если сбой компоненты или элемента сети *не влияет* на приложения, то говорят о «надежности» и «обслуживаемости», а не о доступности.

### **Обзор теории SAN HA**

С точки зрения архитектуры сетей хранения следует рассматривать характеристики доступности каждого подключенного устройства, каждого компонента инфраструктуры SAN и самой сети. Доступность любой системы или приложения – это доступность его самого слабого звена, поэтому для оценки доступности архитектор SAN должен изучить все звенья. Для построения подключенной к SAN компьютерной системы высокой доступности (Highly Available, HA) недостаточно иметь кластер НА. Нужно обеспечить доступность на уровне всей системы, например, использовать двойные НВА, программное обеспечение для альтернативных путей (multipathing), системы хранения высокой доступности с несколькими портами и программное обеспечение кластеризации.

Однако, главное для архитектора SAN – это доступность подключенных к SAN *приложений*. Для конечного пользователя приложения неважно, где произошел сбой (например, в SFP) – им нужно, чтобы приложение работало. Архитектор SAN должен выяснить, как сбой каждого компонента сети повлияет на доступность приложений.

### **Единая точка отказа**

Первый принцип теории НА для SAN и любых других систем гласит: “*если что-то есть только в одном экземпляре, то это не НА.*”

Любой компонент SAN (включая все программное обеспечение, оборудование и саму фабрику), который полностью не дублирован, считается единой точкой отказа. Высокой доступностью обладает только тот компонент, который дублирован и не связан тесно со своим дубликатом (на аппаратном или программном уровне), поскольку событие, приводящее к сбою основного компонента, легко может вывести из строя и тесно связанный с ним дубликат.

Под тесной связью понимаются как логические, так и физические связи. Даже система с архитектурой НА является единой точкой отказа, поскольку все ее компоненты находятся в одном месте. Директоры Brocade имеют резервированные источники питания, вентиляторы, процессоры, образы операционных систем, компоненты коммутации и т.п. Все активные компоненты директоров Brocade задублированы и применяются программные механизмы для быстрого переключения между этими компонентами при сбоях. Если одновременно выйдут из строя процессор и коммутирующее лезвие, директор ни на секунду не прервет пересылку пакетов. Однако, если этажом выше сработает система пожаротушения, то все шасси выйдет

из строя. Если ЦОД будет уничтожен в результате землетрясения или урагана, то вентиляторы не помогут.

Аналогичным образом атака Denial of Service (DoS) может вывести из строя *любую* систему, даже такую надежную, как директор Brocade. Разумеется, продукты Brocade защищены от всех *известных* типов атак, однако природа атак DoS такова, что постоянно разрабатываются новые категории рисков DoS. Например, сам администратора SAN может вызвать атаку DoS если сделает серьезную ошибку при вводе команды. Это вовсе не специфика коммутаторов Brocade или фабрик FC – это теория высокой доступности компьютерных систем. Единственный способ защиты от таких проблем – построение укрепленной «стены» между резервированными компонентами. Она должна надежно изолировать эти компоненты, а не просто применять механизм логических разделов, который сам может отказать.

Следуя этим рекомендациям, архитектор SAN может подключить кластеры НА серверов к таким компонентам с высокой степенью резервирования, как директоры Brocade, в надежную архитектуру фабрики, например, отказоустойчивую СЕ, со второй отказоустойчивой фабрикой для резервирования на случай нештатных ситуаций в первой фабрике, и воспроизвести всю конфигурацию во втором ЦОДе (ведь сама площадка основного ЦОДа – это единая точка отказа).

Разумеется, такой подход требует больших инвестиций на покупку оборудования, его монтаж и оплату труда обслуживающего персонала, которые могут позволить себе не все заказчики. Поэтому следующий логичный шаг – определить, как области резервирования можно для экономии средств исключить без серьезного ущерба для доступности приложений. Для этой задачи нужно изучить взаимосвязи внутри

стека НА.

### *Стек высокой доступности*

Резервирование компонентов может производиться с помощью горизонтальных и вертикальных связей.

Если у директора два источника питания и он может работать от одного, то источники питания резервированы между собой по горизонтали, поскольку они в сети на одном уровне. Если же используются два директора и у хоста резервированы подключения НВА к ним (стр. 17, 28 и 306) или два хоста образуют кластер НА, то связь также идет по горизонтали (см. Рис. 56).

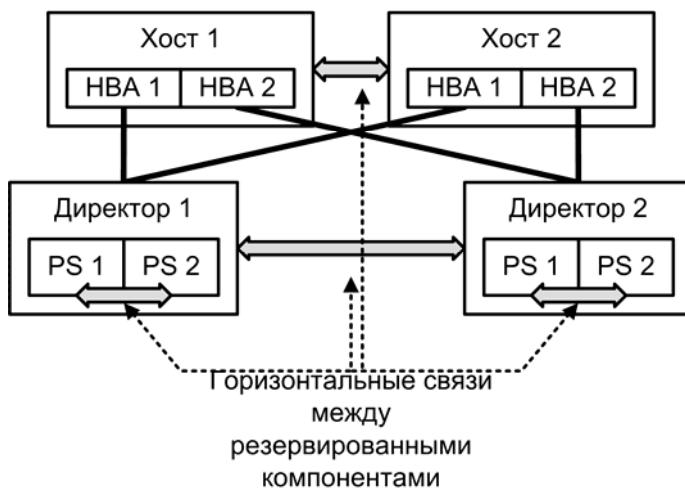


Рис. 56 – Резервирование по горизонтали

Компоненты SAN также могут иметь вертикальные связи. Можно считать, что источники питания находятся ниже директоров, а те в свою очередь ниже программного обеспечения кластера НА. Вертикальные связи *необязательно* относятся к НА. Два связанных по горизонтали источника питания одного директора означают применение стратегии НА, но вертикальная связь между одним директором и источником питания

не означает их взаимное резервирование. Резервированные компоненты, которые по вертикали находятся выше резервированных источников питания, означают НА. Если же резервированные компоненты одного уровня используют один тот же компонент нижнего уровня, то сбой последнего приведет к сбою и тех компонентов, которые находятся над ним.

В показанной конфигурации есть два директора, резервированные по горизонтали, и если один из них целиком выйдет из строя, то не возникнет перебоев в работе хоста и его приложений. Это делает необязательным резервирование питания отдельного директора, ведь если произойдет сбой нерезервированного источника питания, то директор не сможет работать, но работа приложений не нарушится.

Значит ли это, что можно сэкономить расходы за счет отказа от резервирования источников питания? Чтобы ответить на этот вопрос рассмотрим подробнее стек НА из Рис. 56, представленный на Рис. 57.

Компонент	НА?
Приложение	○
ЦОД	✗
Кластер серверов	○
НВА	○
Фабрика	○
Коммутатор/директо	○
PS [вент]	○
СР [ядро]	○
Защита ИБП	✗
Электросеть	✗

Рис. 57 – Уровни НА

Как видно из схемы стека НА<sup>75</sup>, если выйдет из строя ИБП, электрическая сеть или ЦОД, то скорей всего приложение не сможет работать. Источники питания (PS), вентиляторы, центральные процессоры (CP) и коммутационные платы (core cards) директоров Brocade относятся к одному уровню резервирования.

Чтобы определить, действительно ли нужно резервировать конкретный уровень, надо выяснить, как его сбой повлияет на более высокие уровни. Например, отсутствие резервирования электросети одновременно повлияет на оба директора и в результате нарушится работа более высоких уровней независимо от их архитектуры НА.

У источников питания есть определенные отличия от этого механизма – если они нерезервированы, но есть резервирование на уровне коммутаторов или директоров, то это *не* означает, что у SAN только один PS, поскольку каждый директор имеет по крайней мере один PS – иначе директор просто не сможет работать.

Сбой на уровне нерезервированных PS приведет к сбою только одного директора. Если такая ситуация возникнет в структуре, показанной на Рис. 56, то выйдет из строя вся фабрика, включая и подключенные к ней НВА. Однако резервированные НВА подключены к отдельным фабрикам и поэтому отказ не распространится на более высокие уровни. Отказ будет перехвачен драйвером *multipathing*, который поддерживает резервированные НВА. Но что это означает в действительности?

---

<sup>75</sup>

На примере явно не показаны источники бесперебойного питания, электрические сети и ЦОДы, поэтому стек считается нерезервированным на этих уровнях.

Говоря упрощенно, приложение будет простоявать несколько секунд и даже минут пока тайм-аут идет по стеку драйвера прежде чем выполнится переключение (failover). При переключении на альтернативный путь возникает риск того, что неправильно настроена конфигурация или возник одновременный сбой компонентов в резервированном стеке фабрики. В таком случае администратору потребуется починить SAN и переключить хост обратно на основную фабрику после замены источника питания.

Ситуация усложняется из-за того, что почти во всех SAN к неисправному директору подключено несколько серверов, поэтому сбой одного источника питания может в разной степени повлиять на работу сотен приложений. Это усиливает риск сбоя (например, из-за ошибки конфигурирования multipathing) и усложняет восстановление после сбоя.

Наконец, чем больше масштаб сбоя, тем выше риск того, что НА не обеспечит защиту от множественного сбоя. При сбое SFP порта “Директора 2”, через который подключен дисковый массив, и одновременном сбое источника питания “Директора 1”, приложения, использующие порты массива, теряют доступ к своим данным по обоим маршрутам.

Таким образом, чем выше проблема поднимается в стеке НА до того, как ее локализуют, тем сложнее понадобится решение и тем выше риск нарушения работы приложения, несмотря на корректную работу механизма НА. Кластеризацию серверов и multipathing труднее спроектировать и внедрить, чем резервирование источников питания. Чем выше уровень стека, где возникла ошибка, тем больше риск, что механизм резервирования не сработает и потребуется вмешательство администратора для устранения проблемы. Если это возможно, то лучше изолировать

проблему на уровне источников питания, даже если возможно изолировать ее и на более высоком уровне.

Второй принцип теории НА гласит: “*Ошибка нужно локализовать пока она не распространилась на верхние уровни стека*”.

Архитекторы используют резервирование на нескольких вертикальных уровнях (например, источников питания и фабрики) для значительного уменьшения риска того, что ошибка приведет к сбою приложения, и упрощения восстановления после проблем, которые возникли на нижних уровнях стека. Все директоры Brocade поставляются с резервированными источниками питания, вентиляторами, процессорными модулями.

## Резервирование при проектировании SAN

Если сбой в работе SAN может привести к нарушению работы критически-важных систем, то следует применять резервированную архитектуру сети. Кроме того, резервирование рекомендуется использовать если SAN обслуживает большое число некритически-важных систем. Для определения, нужно ли резервирование сети, надо выяснить, что произойдет если все подключенные к SAN устройства вдруг прекратят работу. Если это приведет к нарушению бизнес-процессов в организации, то требуется резервированная архитектура SAN, и чем больше будут потери организации, тем большая степень резервирования должна быть обеспечена.

Это утверждение можно сформулировать как еще один основной принцип теории SAN НА: “*Используемая модель резервирования должна соответствовать критичности проектируемых систем*”. Чем важнее

система, тем надежнее должна быть модель резервирования.

В традиционной архитектуре SAN есть четыре основные категории доступности, которые перечислены в порядке увеличения доступности:

### ***Нерезервированная неотказоустойчивая фабрика SAN или Meta SAN***

Все коммутаторы подключены в одну фабрику, которая содержит по крайней мере одну единую точку отказа. Примером такой категории SAN является каскадная топология, показанная на Рис. 28 (стр. 178). При этом подходе уровень доступности минимален.

### ***Нерезервированная отказоустойчивая фабрика SAN или Meta SAN***

Все коммутаторы подключены в одну фабрику, но отсутствует аппаратная единая точка отказа, из-за которой может произойти распад фабрики на сегменты. Примерами этого подхода могут служить топологии “кольцо”, “mesh” и “центр/периферия” (стр. 180 - 185) и он обсуждается в разделе “Отказоустойчивые фабрики” этой главы. Этот подход защищает от сбоев на уровне оборудования, но сбои на уровне сервисов фабрике приведут к сбою всех подключенных систем, поэтому он не может рассматриваться как метод обеспечения НА.

### ***Резервированные неотказоустойчивые фабрики SAN или Meta SAN***

В неотказоустойчивой SAN с двумя фабриками половина коммутаторов соединены в первую фабрику, а другая половина – в отдельную от первой вторую фабрику. (Этот метод часто называют “модель резервирования А/В” поскольку фабрики Meta SAN обычно обозначаются этими буквами, хотя некоторые

архитекторы называют их «красной» и «синей».) в каждой фабрике есть хотя бы одна единая точка отказа. Этот подход можно использовать в комбинации с хостами и устройствами хранения с двойным подключением и драйверами multipathing, что обеспечивает продолжение работы приложений даже при отказе всей фабрики или выполнении ее модернизации. Пример такой SAN показан в разделе “Резервированные фабрики” (стр. 310 и далее). Этот подход обеспечивает HA, однако может оказаться неэффективным если из-за одиночного незначительного сбоя произойдет крупномасштабное переключение с помощью multipathing, как на предыдущем примере.

### *Резервированные отказоустойчивые фабрики SAN или Meta SAN*

Это идеальная архитектура SAN для обеспечения высокой доступности. В отказоустойчивой SAN с двумя фабриками половина коммутаторов соединены в фабрику А, а другая половина – в отдельную фабрику В (как и при использовании предыдущего подхода), однако у фабрик нет единой точки отказа, из-за которой может произойти их распад на сегменты. Этот подход можно использовать в комбинации с хостами и устройствами хранения с двойным подключением, чтобы приложение продолжало работать даже при сбое всей фабрики в результате ошибки оператора, катастрофы или дефекта компонентов. Только этот подход может обеспечить максимум доступности. Другим важным преимуществом такого подхода является возможность отключения части SAN для модернизации или обслуживания без нарушения работы второй фабрики. Пример SAN этого типа приведен в разделе “Резервированные фабрики” (стр. 310 и далее).

## Узлы с двойным подключением и Multipathing

В контексте SAN узлы с двойным подключением – это устройства, у которых к сети подключены несколько портов, например, у хоста может быть несколько НВА-адаптеров (см. Рис. 56), а у RAID- массивов – две или несколько карт контроллера, каждая из которых обычно оборудована несколькими портами.

Когда эти порты подключены к разным коммутаторам отказоустойчивой фабрики или, что еще надежнее, к разным фабрикам резервированной SAN, то их узлы будут по-прежнему доступны для приложений даже при серьезном сбое. При использовании резервированной фабрики сбой всей фабрики не повлияет на приложение.

Однако, для этого требуется программное обеспечение *multipathing*. Без *multipathing* резервированные НВА будут показывать операционной системе соответствующую фабрику. Если каждый НВА будет «видеть» один и тот же LUN, то операционная система будет видеть его как разные устройства хранения, что создаст проблемы при доступе к LUN приложений вплоть до порчи данных. Программное обеспечение *multipathing* работает между драйвером НВА и операционной системой и заставляет ОС показывать приложению только одну копию каждого LUN. Программное обеспечение также с помощью тайм-аутов обнаруживает сбои и выполняет быстрое переключение путей.

Архитектор SAN должен тщательно разобраться в том, как в его решении *multipathing* работает механизм тайм-аутов и переключения. Многие драйверы *multipathing* позволяют системному администратору проводить тюнинг для ускорения или замедления

переключений при отказе, но лучше оставить значения драйвера по умолчанию если нет особых причин изменить их. Если архитектор SAN хорошо понимает, как быстро должно происходить переключение, то системный администратор может при необходимости произвести дополнительный тюнинг. Увеличение частоты повторных попыток или тайм-аутов может повысить производительность при сбое в большой фабрике, но также может привести и к переключению когда SAN работает нормально.

Некоторые решения multipathing применяют подход active/standby, при котором только один HBA или порт контроллера обрабатывает весь ввод/вывод, если только нет сбоя в пути. В других решениях используется подход active/active: ввод/вывод балансируется между портами обычно с помощью операций SCSI или на уровне Fibre Channel exchange аналогично DPS (п. 272). При оценке решений multipathing архитектор SAN должен учитывать два момента:

1. Active/active при нормальном режиме операций работает лучше, чем active/standby, а при сбое оба подхода работают одинаково.
2. Производительность решений Active/standby не зависит от того, произошел сбой или нет. При сбое у решения active/active может упасть производительность приложений.

Выбор варианта зависит от характера приложений, которые обслуживают подключенные к SAN хосты, и требований пользователей этих приложений. Чаще всего для приложения нужно постоянно обеспечивать максимальную производительность и поэтому используется active/active. Однако иногда приложению нужен точно заданный уровень производительности и тогда лучше подойдет active/standby. Сконфигурированный активный или резервный путь

должен на 100% соответствовать требованиям приложения.

Другое отличие между решениями multipathing – это использование «открытых систем» или фирменных (proprietary) технологий. В первом случае можно использовать серверную платформу одного вендора и систему хранения другого вендора. Использование фирменных технологий может обеспечивать больше функций и более высокий уровень надежности, поскольку они тщательно протестированы при обслуживании конкретных конфигураций, поэтому у таких решений multipathing меньше риск сбоя и проще восстановление после сбоя.

Независимо от используемого драйвера multipathing надо соблюдать следующий принцип: “*Нужно всегда использовать программное обеспечение multipathing для управления резервированными HBA и подключением устройств хранения в HA SAN.*”

## Отказоустойчивые фабрики

Согласно Британской Энциклопедии (Encyclopedia Britannica<sup>76</sup>) “resilient ( отказоустойчивый)” означает «способный восстанавливаться или адаптироваться к неприятностям и изменениям.” Отказоустойчивая фабрика способна выдержать сбои ISL или отдельного коммутатора без распада на отдельные сегменты и позволяет без отключения сети модернизировать коммутаторы ядра. Отказоустойчивость достигается за счет обеспечения топологией отказоустойчивой фабрики по крайнем мере двух маршрутов между любыми двумя ее коммутаторами (см. на Рис. 58 - сравнение отказоустойчивой и неотказоустойчивой фабрики).

---

<sup>76</sup>

Encyclopedia Britannica 2004 Ultimate Reference Suite DVD.

Самоизлечение фабрики с помощью маршрутов в обход отказавших сегментов обеспечивается разработанным Brocade протоколом Fabric Shortest Path First (FSPF). Сейчас этот протокол, первоначально использовавшийся только в продуктах Brocade, утвержден в качестве стандарта для маршрутизации в фабриках FC и применяется всеми вендорами коммутаторов и маршрутизаторов FC.

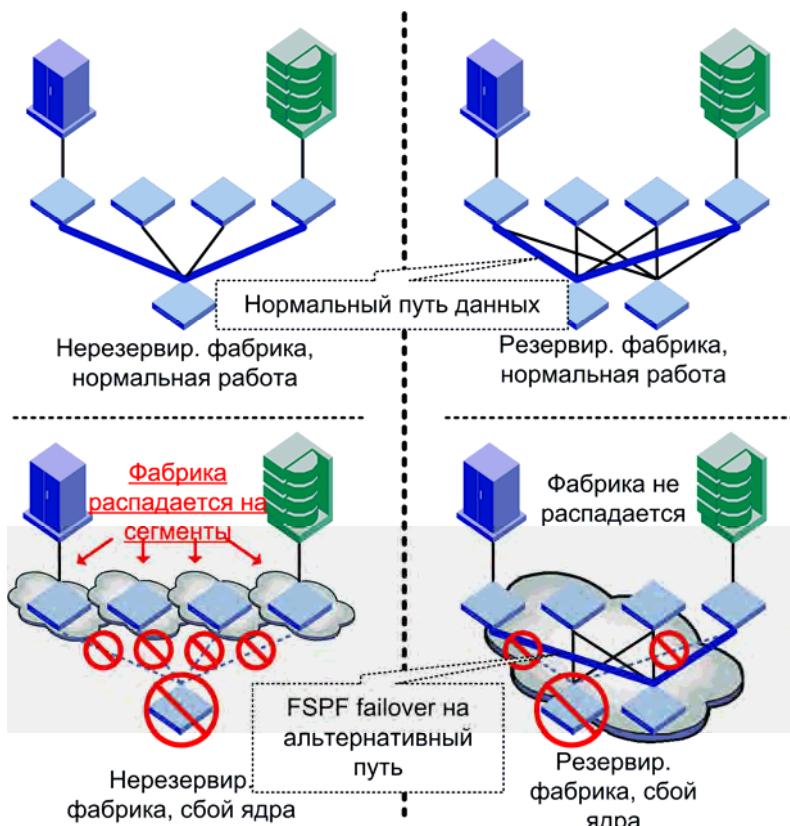


Рис. 58 - Сравнение отказоустойчивой и неотказоустойчивой фабрики

Одиночный директор Brocade считается отказоустойчивым. Аналогично тому, как фабрика на Рис. 58 - сохраняет работоспособность несмотря на сбои, директор Brocade продолжает работу даже после сбоя лезвия. Сама внутренняя архитектура директора Brocade 48000 похожа на отказоустойчивую фабрику и применяемые в нем программное обеспечение быстрого переключения для НА и другие аналогичные механизмы работают быстрее, чем такие механизмы фабрики, как FSPF.

Это не означает, что невозможен полный отказ отказоустойчивой фабрики или директора. В их аппаратной части нет точки одиночного отказа и сервисы фабрики Brocade обладают высокой надежностью. Однако все коммутаторы фабрики и лезвия директора “тесно связаны” на уровне передачи данных и команд, поэтому теоретически вся отказоустойчивая система может отказать из-за атаки denial of service, ошибки оператора, серьезных ошибок в работе узлов и таких внешних событий, как например срабатывание неисправной системы пожаротушения. Для защиты от таких проблем всегда рекомендуется резервированная фабрика, особенно в комбинации с отказоустойчивостью.

Соответствующий принцип НА гласит: “*Решение НА SAN должно использовать отказоустойчивую архитектуру, но сама отказоустойчивая сеть может стать единой точкой отказа*”.

## Резервированные фабрики

Британская энциклопедия определяет “резервированный” (redundant) как “использующий дубликат для предотвращения сбоя всей системы (например, космического корабля) из-за сбоя одного компонента”. В нашем случае “система” – это вся SAN.

Резервированные фабрики являются “компонентами” этой системы, которые “дублируют” функции друг друга так, чтобы сбой одной фабрики не привел к отказу всей “системы” SAN.

Отказоустойчивые фабрики пытаются восстановиться при сбоях в фабрике, а резервированные фабрики предотвращают сбой всей SAN. Да, отказоустойчивые фабрики надежны, однако никакая одиночная фабрика не может рассматриваться как настоящее решение НА, поскольку она сама может потерять работоспособность из-за аварии, ошибки оператора или программного обеспечения.

Для защиты от таких ошибок необходим еще один уровень доступности: резервированные фабрики “A/B” SAN, также известные как “dual-fabric” SAN. Эта архитектура предусматривает использование двух полностью физически изолированных фабрик (также, как в кластере НА используются два физически изолированных сервера). Дублирование компонентов и программное обеспечение переключения при сбоях – это самые эффективные средства обеспечения высокой доступности серверных платформ. Аналогичным образом, архитектура с несколькими фабриками – лучший способ обеспечить высокую доступность SAN. Резервированные фабрики улучшают не только доступность, но и масштабируемость и использование двух фабрик вдвое увеличивает максимальный размер SAN.

Без физической изоляции решение не является резервированным – оно только отказоустойчиво и в нем может возникнуть сбой также, как и в отказоустойчивой фабрике. Сегментация директора с помощью разделов, зон или программного обеспечения VSAN не увеличит доступность SAN. Рис. 59 – Сравнение резервированных фабрик и разделов показывает правильно

спроектированное решение НА и не правильно спроектированную SAN с единой точкой отказа.

Сбой всего директора VSAN может произойти по разным причинам. Например, все VSAN работают с образом операционной системы и ошибка в коде ОС или оператора при ее обновлении приведет к сбою всего шасси и одновременно нарушит оба пути.

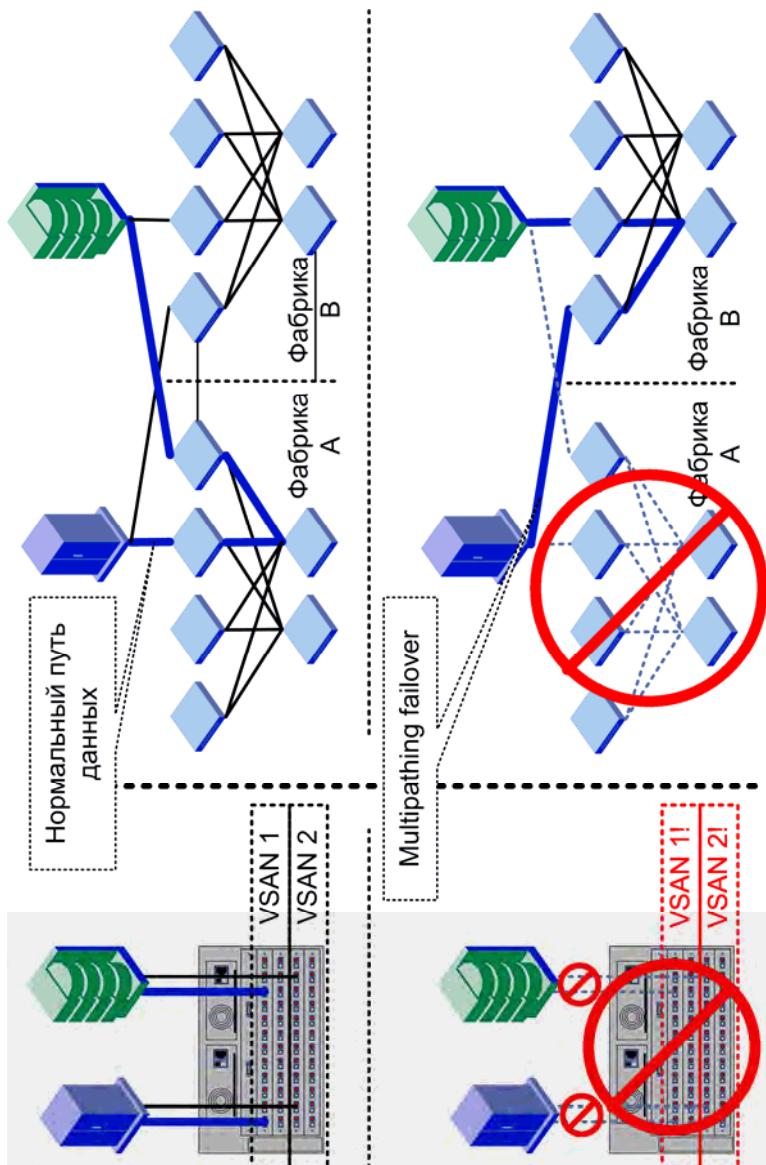


Рис. 59 – Сравнение резервированных фабрик и разделов

Это не значит, что схемы разделов, такие, как VSAN, не имеют перспектив. Brocade разработала несколько таких механизмов и поддерживает их применение для

решения многих проблем администрирования SAN, в том числе безопасности, совместимости и управляемости. Механизмы сегментации Brocade включают зонирование, Virtual Fabrics и поддержку нескольких доменов в некоторых директориях Brocade. Например, зонирование внутри фабрики предотвращает взаимодействие между собой разных НВ А, поэтому проникнувший в один хост хакер не сможет попасть на другой хост через SAN. Один НВА не сможет случайно запустить атаку DoS на другой адаптер, а также можно независимо управлять дисковыми подсистемами разных хостов. Все эти функции полезны и применение зонирования приветствуется при проектировании SAN. Такие механизмы сегментирования, как зонирование и Virtual Fabrics, эффективны, но не для обеспечения доступности.

Применение полностью резервированной фабрики позволяет обеспечить непрерывную работу узлов даже при сбое всей фабрики или ее отключения для обслуживания. При описании характеристик доступности на первом месте стоит доступность пути. Если отдельный линк оборвется, но путь данных будет работать по-прежнему, то не произойдет сбой приложений, использующих SAN. Хотя возможно некоторое ухудшение производительности серверов приложений, но главное, что они продолжают работать.

Чтобы резервированная архитектура работала правильно, две и более фабрики нужно использовать с несколькими НВА, несколькими RAID-контроллерами и программным обеспечением multipathing, обслуживающих устройства SAN, которым нужна максимальная доступность. На Рис. 59 – Сравнение резервированных фабрик и показана способность резервированной фабрики выдержать крупномасштабный сбой, а также неэффективность в

этой ситуации любой схемы сегментирования.

Отметим, что резервирование фабрики не исключает применение отказоустойчивости и лучше комбинировать эти два подхода. На Рис. 59 – Сравнение резервированных фабрик и каждая фабрика отказоустойчива и они резервированы между собой.

Итак, можно сформулировать еще один принцип: “Решения HA SAN должны использовать полностью резервированную архитектуру ‘A/B’ и предпочтительно, что каждая фабрика была отказоустойчивой”.

## Проектирование резервированной Meta SAN

Предпочтительная архитектура для Meta SAN мало отличается от предпочтительной архитектуры фабрики. Как и фабрики, Meta SAN обычно строятся по разным вариантам топологии центр/периферия и в них, как и в фабриках, для высокой доступности используется отказоустойчивость и/или резервирование.

Разумеется, могут быть разные вариации описываемых ниже подходов. Одна часть Meta SAN может использовать одну модель, а вторая (менее критичная) часть может использовать другую модель. Модель резервирования следует адаптировать к требованиям бизнес-приложений и она должна использоваться на всем пути от сервера приложения и до его устройств хранения.

### Отказоустойчивые Meta SAN

Главный признак отказоустойчивой SAN – отсутствие единой точки отказа. У каждого линка есть один или несколько альтернативных путей, у каждого центрального коммутатора и маршрутизатора - дублер. Хотя возможен сбой периферийного коммутатора,

который повлияет на работу всех подключенных к нему узлов, однако важные узлы всегда подключены по крайней мере к двум периферийным коммутаторам.

В отказоустойчивой Meta SAN каждая граничная фабрика, которая экспортирует устройства, имеет не менее двух маршрутизаторов, обеспечивающих путь к любой другой граничной фабрике. Обычно узлы подключают к резервированным фабрикам A/B при использовании этой модели. На Рис. 60 - "Отказоустойчивая Meta SAN" показан пример такой конфигурации.

Это относительно надежная архитектура сети, позволяющая маршрутизировать между резервированными фабриками A/B для сложных моделей переключения при отказах, что невозможно реализовать без маршрутизаторов. Она дает серверам с одиночным подключением некоторую функциональность multipathing на уровне сети, когда порты устройств хранения подключены к разными фабрикам Meta SAN. Наконец, такой подход удобен когда ленточные устройства имеют одиночные подключения к одной или нескольким фабрикам резервного копирования, а у серверов по два подключения.

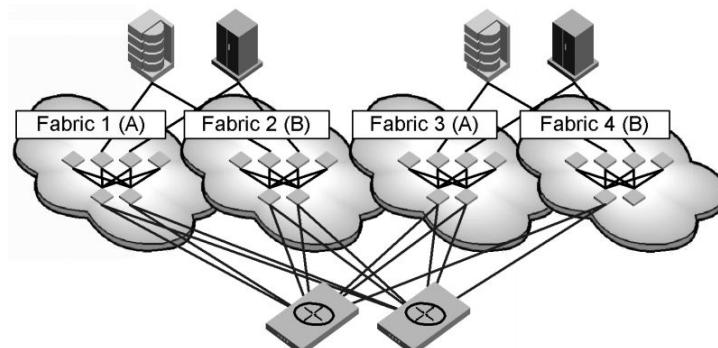


Рис. 60 - Отказоустойчивая Meta SAN

Однако это не полностью резервированное решение – выход из строя маршрутизатора может (хотя бы в теории) одновременно повлиять на обе фабрики, поэтому для критически-важных систем нужна полностью резервированная Meta SAN.

### *Резервированные Meta SAN*

На Рис. 61 - Резервированные Meta SAN показана полностью резервированная Meta SAN. В резервированной фабрике есть две полностью разделенные фабрики SAN, а в резервированной *Meta SAN* – две полностью изолированные (A/B) *Meta SAN*. Критичные сервера и устройства хранения имеют по два подключения – по одному на каждую *Meta SAN*. Это обеспечивает максимально возможную изоляцию ошибок, например, работающий с ошибками НВА в одной *Meta SAN* не может влиять на узлы другой *Meta SAN*.

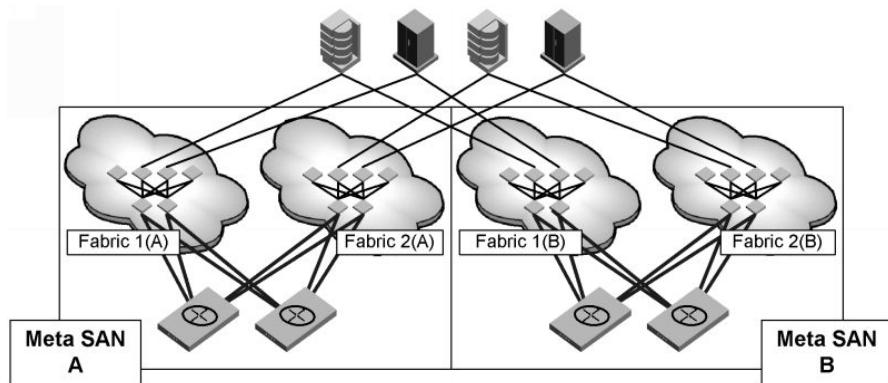


Рис. 61 - Резервированные Meta SAN

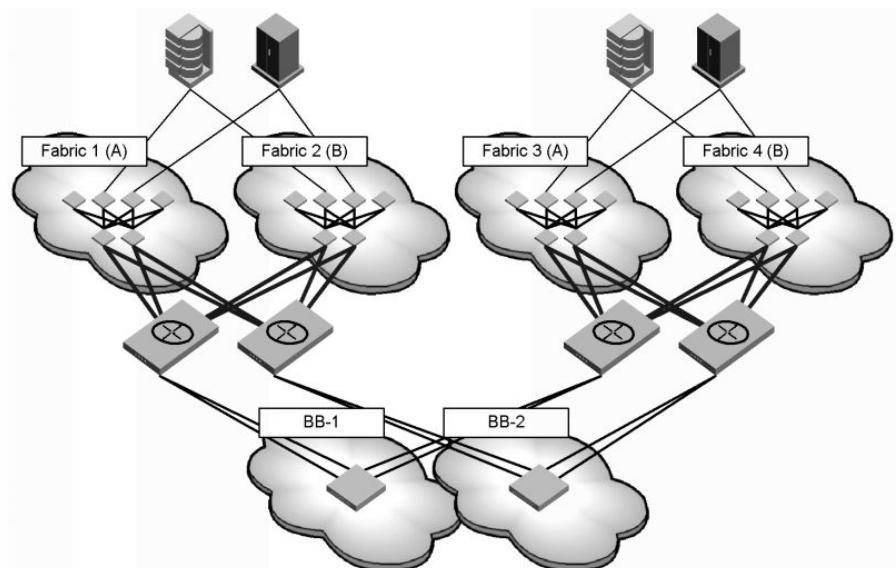
## *Параллельные резервированные фабрики BB Meta SAN*

Когда требуется соединить большое число маршрутизаторов для масштабируемости или связи на большие расстояния, то используется механизм “фабрика backbone” или “фабрика BB”, при котором подсоединение маршрутизаторов выполняется с помощью E\_Ports и они образуют фабрику через обычные механизмы switch-to -switch. Программное обеспечение более высокого уровня Fibre Channel Router Protocol (FCRP) позволяет им автоматически координировать все операции Meta SAN.

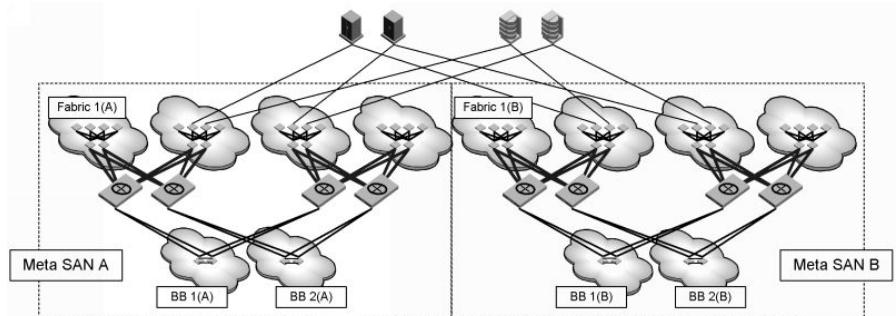
Для обеспечения отказоустойчивости маршрутизаторы размещают так, чтобы они использовали отдельные backbone. Хотя работа при этом мало отличается от одиночной backbone, но улучшается доступность и масштабируемость. Вся фабрика backbone может выйти из строя и это не приведет к обрыву соединений в какой-либо LSAN. Архитектура с резервированной backbone показана на Рис. 62. отметим, что каждый маршрутизатор имеет много подключений к фабрике BB для достижения разных показателей производительности и надежности, но каждый маршрутизатор может быть подключен только к *одной* фабрике backbone и нельзя подключить один и тот же маршрутизатор и к BB-1, и к BB-2 – это приведет распаду фабрики на сегменты или объединению двух backbone в одну фабрику, т.е. к потере резервирования. Поэтому для резервирования backbone нужно несколько маршрутизаторов.

Meta SAN с резервированной фабрикой backbone можно комбинировать с другими моделями резервирования, которые обсуждались в этой книге. Например, на Рис. 61 - Резервированные Meta SAN показано использование этого метода в

сокращенной форме вместе с полностью резервированными Meta SAN. На Рис. 63 – Резервированная Meta SAN + резервированные ВВ показан расширенный вариант этого случая.



**Рис. 62 – Отказоустойчивая Meta SAN с резервированными фабриками ВВ**



**Рис. 63 – Резервированная Meta SAN + резервированные ВВ**

Наконец, в большинстве территориально-распределенных конфигураций используются резервированные фабрики backbone. В таких внедрениях FCIP, как показанная на Рис.63, резервированные фабрики backbone туннелируются через IP -сеть, хотя это не показано на схеме. Это пример интеграции FC-FC Routing Service с FCIP Tunneling Service, которая является оптимальной для резервированных WAN.

Принцип проектирования высокодоступных маршрутизируемых SAN гласит: “*Используйте для Meta SAN ту же модель обеспечения доступности, которая подходит для аналогичной большой фабрики*”.

## Изоляция сбоев и LSAN

Для достижения максимума масштабируемости (тысячи портов в сети) лучше использовать много относительно небольших фабрик (сотни портов), соединенных через маршрутизаторы Fibre Channel с помощью LSAN (см. также “Обеспечение максимальной масштабируемости” на стр. 215 и далее.) При этом маршрутизатор может изолировать сбои и не дать им распространяться между граничными фабриками. Таким образом, меньше устройств будут затронуты сбоем. Кроме того, маленькие фабрики стабильнее больших, поэтому вероятность возникновения в них сбоев меньше.

Этот механизм особенно удобен при больших расстояниях. Всегда рекомендуют использовать LSAN в территориально-распределенных решениях независимо от того, применяются ли родные FC ISL, интегрированный шлюз FCIP в Brocade Multiprotocol Router или решение других фирм для больших расстояний. Он максимально надежно предотвращает влияние нестабильности WAN на конечные точки WAN. Если WAN работает нестабильно, то это затронет только реально использующие ее устройства (разумеется, этого

можно избежать если применить полностью резервированные WAN).

Для изоляции сбоев в Meta SAN нужно применять локализацию (стр. 258) при проектировании LSAN. Простой экспорт всех устройств всем другим фабрикам ведет к потере большинства преимуществ НА и преимуществ масштабирования маршрутизатора.

При проектировании Meta SAN нужно следовать следующему принципу: “*Meta SAN должны использовать локализацию в максимальной степени для изоляции сбоев и масштабируемости*”.

## **Асимметричные SAN**

При развертывании резервированных SAN или Meta SAN симметричность не всегда обязательна. Например, при построении резервированной фабрики SAN фабрика “A” может состоять из отказоустойчиво соединенной топологии СЕ, а вторая фабрика – из изолированного коммутатора (см. Рис. 35, стр. 194.)

Есть несколько вариантов такого подхода. Например, крупномасштабная Meta SAN может быть несимметричной – например, Meta SAN A будет большой и сложной и к ней будут подключены узлы, а Meta SAN B – маленькой и к ней подключены только несколько критически-важных узлов (аналогично Рис. 35). Либо наличие двух Meta SAN снизит требования к отказоустойчивости каждой Meta SAN, как и в резервированной, но неотказоустойчивой фабрике. В этом случае вторые центральные коммутаторы могут быть удалены из каждой граничной фабрики, либо могут не устанавливаться вторые маршрутизаторы и не дублироваться соединения IFL. Этот вариант показан на Рис. 64.

Лучше закладывать резервирование на каждом уровне насколько это позволяет бюджет. Можно отказаться от резервирования центральных коммутаторов в граничных фабриках, но это еще не значит, что оно не имеет значения. Как и многие другие решения при проектировании сети решение о закладывании резервирования в каждый уровень зависит от компромисса между ценой, производительностью и RAS.

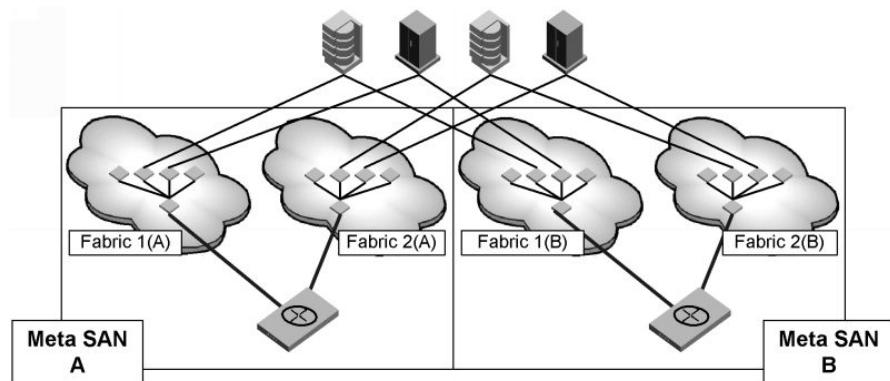


Рис. 64 – Вариант резервированных Meta SAN

Когда на первом месте стоит сокращение затрат, используйте следующий принцип: “*жертвуите отказоустойчивостью (resiliency) до того, как пожертвовали резервированием (redundancy)*”. Лучше развернуть две неотказоустойчивые фабрики “A/B”, чем одну отказоустойчивую, при том что сервера и дисковые массивы имеют двойное подключение.

## Стратегии подключения устройств

Если SAN должна быть ограничена одним отказоустойчивым объектом (одним директором, фабрикой или Meta SAN), то для достижения наивысшей

отказоустойчивости нужно следовать следующим простым правилам подключения устройств.

Лучше всего применять локализацию там, где это возможно. В неотказоустойчивой SAN сбой коммутатора ядра приведет к распаду фабрики на сегменты, что обычно не может нарушить связь между локализованными устройствами, поэтому следует по возможности использовать локализацию.

Если из-за характера трафика трудно или нельзя применять локализацию (например, из-за его нестабильности), то лучше распределить устройства по определенному алгоритму. При подсоединении их к такому директорию, как Brocade 48000, лучше распределять подключения серверов и устройств хранения между лезвиями и тогда при неисправности лезвия с портами фабрика не потеряет сразу все свои устройства хранения или все сервера. Аналогичным образом при распределении подключений серверов и устройств хранения между периферийными коммутаторами гарантирует, что сбой одного коммутатора не приведет к потере всех соединений.

# 10

## 10: Планирование безопасности

Большинство компаний понимает ценность данных, которые обслуживает SAN, и поэтому предпринимает меры для защиты их конфиденциальности, целостности и доступности. Брешь в системе безопасности может легко привести к утечке конфиденциальной информации, финансовым потерям и другому ущербу для бизнеса. В этой главе обсуждаются основные меры и инструменты обеспечения безопасности SAN.

### **Обзор безопасности**

С момента выхода Secure Fabric OS в 2001 году Brocade остается лидером в области безопасности SAN. Secure Fabric OS была разработана на основе многолетнего опыта внедрений SAN разных размеров и архитектур и предназначена для обеспечения самых строгих требований к безопасности приложений. Secure Fabric OS впервые реализовала списки контроля доступа Access Control List (ACL) в индустрии Fibre Channel и первой обеспечила для Fibre Channel аутентификацию на базе механизма PKI, который затем был заменен основанным на стандартах DH-CHAP.

Разумеется, модель безопасности SAN продолжала развиваться после 2001 года. Все функции первой версии Secure Fabric OS теперь заменены более совершенными и гибкими функциями базовой Fabric OS, начиная с

версии 5.3.0, на которую клиенты Brocade могут перейти без покупки дополнительных лицензий.

Эта книга не является заменой руководства по продуктам, поэтому в ней нет описаний конкретных команд и детальных рекомендаций по конфигурированию, которые меняются по мере выхода новых версий кода. Вместо этого она содержит полезную для архитектора SAN информацию, которая основана на общих принципах и поэтому остается актуальной. В данной главе дается такая информация о безопасности.

Нет полностью идентичных двух ИТ-инфраструктур и поэтому у каждой SAN система безопасности имеет свою специфику, однако есть базовые принципы безопасности, которые должны применяться во всех SAN, включая первоначальные ограничения доступа к SAN и контроль над управлением изменениями. Чаще всего при построении системы безопасности SAN нужно решить вопросы:

- Физического доступа
- Доступа для управления
- Блокировки портов
- Политики зонирования

В этой главе даются общие рекомендации по решению этих четырех вопросов.

## **Безопасность на физическом уровне**

Эксперты по безопасности считают, что полную безопасность ИТ-инфраструктуры может обеспечить только защита на физическом уровне. Например, если

злоумышленник получит физический доступ к SAN, то он сможет инициировать атаку denial of service attack простым отключением силового кабеля. Более изощренные атаки могут привести к доступу злоумышленника к данным, которые хранятся в SAN, и даже манипуляции ими.

Эта книга – не учебник по общим мерам безопасности ЦОДа и, кроме того, в любом современном ЦОДе уже действуют политики безопасности для серверов и устройств хранения. Инфраструктура SAN, обеспечивающая доступ к данным от серверов к устройствам хранения, должна быть физически защищена в той же степени и с помощью тех же методов, которые используются для самых защищенных устройств, подключенных к SAN. Если к SAN подключен сервер, установленный в открытой комнате, а другой сервер установлен в охраняемом ЦОДе, то инфраструктурное оборудование должно располагаться в охраняемом ЦОДе. По крайнем мере, все коммутаторы, через которые идет трафик от самых защищенных устройств, должны быть защищены так же надежно, как эти устройства.

Если нет возможности установить все элементы инфраструктуры в физически защищенной зоне, то нужно использовать шифрование данных при их передаче по незащищенным сегментам, например, шифрование необходимо использовать в решениях для удаленного доступа, поскольку хакер может подключиться к линии передачи данных. Однако только шифрование на одном конце линии и их дешифрование на другом может оказаться недостаточным. Если компания работает с очень ценностными данными, то лучше шифровать их до того, как они уйдут из сервера-источника, и в зашифрованном виде записать их диск или ленту на другом конце линии – таким образом,

данные будут защищены и при передаче, и при хранении. На рынке имеется несколько систем для шифрования файлов, которые можно применять в этом случае.

Крайне важно задокументировать физическую конфигурацию сети, включая размещение всех коммутаторов, маршрутизаторов и кабелей, и затем оценить, насколько безопасно размещение каждого компонента. Надо включить в рассмотрение физические аспекты интерфейсов сетевого управления и рабочих станций. Все физические и логические аспекты безопасности SAN нужно документировать в соответствии с корпоративной политикой безопасности хранения данных.

## **Безопасность сетевого управления**

Второй по важности после физической безопасности инфраструктуры передачи данных SAN идет безопасность сетевого управления out-of-band. Если злоумышленник взломает интерфейс сетевого управления, то будут бесполезны и все другие обсуждаемые далее меры безопасности, поэтому надо обеспечить надежную защиту сетевого управления.

Лучше начинать с физической безопасности сетевого управления (как и других элементов инфраструктуры SAN). Если хакер сможет подключиться к кабелю, соединяющему станцию управления с SAN, то он сможет взломать сеть даже если трафик управления шифруется, поэтому физическая безопасность интерфейса управления должна быть на том же уровне, что у серверов и компонентов инфраструктуры. Лучше всего LAN, с помощью которой осуществляется управление SAN, физически изолировать от остальной локальной сети площадки. У VLAN уровень физической безопасности хуже, чем у физически изолированной сети

(хотя некоторые менеджеры по продажам утверждают обратное). Если необходимо подсоединить остальную инфраструктуру LAN к сети управления SAN, то лучше сделать это через межсетевой экран.

Даже если сеть управления физически защищена, лучше также применять защиту и на логическом уровне. Все программные пакеты управления должны использовать шифрование при подключении к оборудованию SAN. Например, Brocade поддерживает secure shell (ssh) как альтернативу telnet.

После установления соединения с интерфейсом управления пользователь должен ввести свое имя и пароль. У всех коммутаторов и маршрутизаторов Brocade имеется несколько имен и паролей по умолчанию. Лучше изменить эти значения при начале промышленной эксплуатации устройства. Также следует изменить или отключить SMTP. У каждого администратора должен быть свой идентификатор – это позволит легко удалить учетную запись уволившегося администратора и контролировать доступ к функциям управления. Virtual Fabrics – это удобная функция если одна физическая SAN разбивается на несколько доменов управления, обеспечивающая точный контроль за доступом на основе ролей. Пароли для администраторов следует выбирать следуя правилам “strong password” и периодически менять.

## **Безопасность с блокированием портов**

Еще один аспект безопасности SAN – это планирование и внедрение системы безопасности. Как это ни странно, архитекторы SAN часто пренебрегают этим аспектом. Блокирование портов означает постоянное отключение тех портов, к которым не подключены устройства, блокирование для части портов возможности стать E\_Ports и блокирование WWN

устройств для определенных портов. Все текущие модели продуктов Brocade поддерживают эти функции как часть базовой ОС с помощью команд CLI и опций GUI. Для дополнительной защиты серверов и коммутаторов можно применять аутентификацию DGH-CHAP для строгой аутентификации при добавлении новых коммутаторов и серверов в фабрику. Это самая надежная защита от подключения методом WWN-spoofing.

## Безопасность на уровне доступа к устройствам (зонирование)

Наконец, нужно обеспечить безопасный доступ между подключенными к фабрике устройствами. С помощью зонирования администратор может разбить фабрику на зоны, в которые объединены взаимодействующие между собой устройства, и применить лучшие в своем классе механизмы блокировки связи между устройствами из разных зон. Зонирование похоже на VLAN<sup>77</sup> из мира IP-сетей, но у него значительно лучше конфигурирование и оно не использует закрытые механизмы тегов, снижающие пропускную способность. Зоны могут перекрываться и поддерживать как очень простые, так и сложные схемы передачи трафика. Зонирование можно рассматривать как комбинирование лучших аспектов технологии VLAN с мощными функциями безопасности, обычно используемых только в межсетевых экранах корпоративного класса.

Зоны удобны для создания барьеров между разными операционными средами, создания внутри фабрики

---

<sup>77</sup> В документации Brocade 1990-х годов зоны назывались “виртуальными SAN” задолго до того, как другие вендоры стали использовать этот термин для обозначения фирменного функционала.

нескольких функциональных групп или тестовой среды и/или областей, изолированных от остальной фабрики. Если в фабрике есть зоны, то никакой сервер или устройство хранения не смогут использовать фабрику до тех пор, пока оно явно не включено хотя бы в одну зону авторизованным администратором фабрики.

Зонирование – это ресурс всей фабрики и конфигурации зон автоматически передаются на все ее коммутаторы. Конфигурацией зон можно управлять с помощью интерфейса WEBTOOLS GUI, Fabric Manager или инструментов управления SAN от третьих фирм (через API). Администраторы могут использовать эти интерфейсы управления по отдельности или в комбинации. Фабрика обеспечивает максимум резервирования и надежности, поскольку каждый коммутатор хранит информацию о зонах и может ее передать другому коммутатору, добавленному в фабрику.

### **Типы зонирования**

Самое простое зонирование – это зонирование на базе портов, т.е. физические порты коммутатора распределяются между разными зонами, например, задается такое правило: только устройства, подключенные к порту 1 коммутатора 1 могут разговаривать с устройствами на порте 2 коммутатора 3. Когда технология Fibre Channel только вышла на рынок, Brocade поддерживала аппаратную реализацию зонирования только на уровне портов, поэтому этот механизм получил название «жесткого зонирования (hard zoning)». Все современные коммутаторы поддерживают аппаратную реализацию зонирования.

Сейчас более предпочтительным методом стало зонирование на базе WWN, которое ограничивает доступ по его WWN вместо идентификатора порта. Это более

гибкий метод, поскольку любые узлы сети всегда остаются в назначеннной им зоне и он поддерживает такие устройства на основе петли как JBOD. Однако у него есть один недостаток – если заменить устройство, то у него изменится WWN, хотя адрес порта останется прежним, поэтому потребуется вносить обновление в конфигурацию зоны. Разумеется, переключение устройства от одного порта к другому при использовании зонирования WWN выполняется проще, поэтому этот недостаток не считается принципиальным.

Как уже говорилось выше, зонирование бывает жесткое и мягкое. При мягком зонировании разбивка на зоны выполняется только с помощью операционной системы коммутатора - обычно сервер имен фабрики предоставляет конечным узлам только ограниченную информацию. Узлы в фабрике информируются о существовании других узлов только с помощью сервера имен, и если этот сервис ограничивает его ответы на запросы в соответствии с данными зон, то на практике никакое устройство не может обращаться к устройствам из других зон (однако пакеты могут передаваться между разными зонами). Этот подход работает эффективно до тех пор, пока зоны не надо будет изменить, пока оборудование не запивает в кэш таблицы сервера имен или пока пользователю не потребуется гарантировать, что никакие пакеты случайно или предумышленно не могут выйти за пределы определенных зон.

Хотя все современные коммутаторы Brocade используют мягкое зонирование, также всегда могут применять ту или иную форму жесткого зонирования. Жесткое зонирование использует специализированные микросхемы в ASIC в каждом коммутаторе для проверки передающихся по фабрике пакетов. Обычно применение зонирования выполняется на уровне порта получателя.

Если же такого разрешения не будет, то пакет будет отброшен.

## *Разработка плана зонирования*

Хотя текущим управлением зонированием обычно занимается администратор SAN, архитектор SAN должен продумать, как построить зоны, поскольку от этого зависит, какие «переговоры» между устройствами будут разрешены. Он также должен еще на фазе проектирования проанализировать характер трафика с точки зрения производительности, что упростит разработку плана зон.

Здесь важен творческий подход, поскольку не может быть только одной «правильной» конфигурации зон для данной конфигурации SAN. Далее описывается один из таких возможных подходов, но лучше следовать рекомендациям по зонированию от вендора фабрики (если таковые имеются).

Сначала нужно составить список всех серверов и устройств хранения, которые нужно объединить в зоны (обычно к этому моменту уже имеется список подключенных к SAN устройств). В зависимости от привычных для архитектора SAN утилит этот список может быть в виде электронной таблицы, текстового файла ASCII или другом формате. Список должен включать имя, физическое расположение WWN, ID порта и функцию каждого устройства.

Затем надо оценить потребности каждого сервера в хранении и если ему выделено отдельное устройство, то указать это в списке, а если нет, то сколько устройств хранения и какого типа потребуется серверу? Затем надо найти ближайшее устройство хранения, которое соответствует требованиям сервера и выделить серверу

емкость из этого дискового массива. В списке надо указать, что серверу требуется доступ к этому массиву. Надо предусмотреть сценарии переключения при сбоях, несколько портов под управление программного обеспечения multipathing и доступ к серверам резервного копирования и лентам.

После того как у архитектора появится понимание, какому порту сервера нужен доступ к какому порту устройства хранения, надо записать эти сведения.

На этом этапе можно создать общий план зонирования, где отмечен необходимый доступ. Например, архитектор может сконфигурировать одну зону из одного порта устройства хранения и всех серверов, которые обращаются к нему. При этом один сервер может обращаться к другому серверу из этой же зоны, что в SAN обычно не требуется и даже нежелательно. Поэтому лучше иметь по одной зоне на пару устройств и использовать перекрывание зон. В перекрывающихся зонах адаптеры НВА совместно используют один или несколько портов хранения, но сами адаптеры изолированы между собой. Зоны лучше строить из тесно связанных между собой устройств, если это только не приводит к излишнему усложнению управления. В идеале фабрика состоит из большого числа зон с двумя портами или двумя WWN и определяет связь между одним инициатором и одним получателем (такие зоны иногда называют point-to-point). Если инициатор сконфигурирован на доступ к нескольким портам устройств хранения, то следует включить его в несколько зон вместо того, чтобы объединять несколько портов в одну зону устройств хранения – тогда будет возможен только тот доступ, который необходим.

Однако в очень больших фабриках применение зон point-to-point неэффективно, поскольку оно означает

создание большого числа зон и необходимость многочисленных изменений в конфигурации фабрики при добавлении и перемещении устройств. Более популярный метод – это зоны с одним инициатором, при котором достигается баланс между необходимостью ограничить число зон и упрощением управления. Для каждого инициатора в фабрике создается отдельная зона, которая также содержит все порты получатели пакетов от этого инициатора.

Кроме обеспечения повышенной безопасности этот подход помогает оптимизировать сервисы фабрики. Зоны оптимизируют сервисы фабрики, такие как распространение RS CN и ответ сервера имен, и ограничивают излишнее обнаружение устройств. Правильная конфигурация зон необходима для больших фабрик (даже если для них безопасность некритична) и методы point-to-point и «одного инициатора» повышают до максимума плюсы зонирования.

Архитектору также следует позаботиться об удобной системе имен для зон. Для улучшения масштабируемости и управляемости лучше всего использовать короткие и понятные имена. Разумеется, трудно придумать короткое, но в то же время понятное имя, поэтому рекомендуется сокращать имена так, чтобы они не потеряли своего значения.

Зоны должны содержать только псевдонимы устройств, но не их PID или WWN для того, чтобы можно было определить устройство только один раз и использовать его в нескольких зонах без переопределений. Если устройство вышло из строя, было перемещено или модернизировано, то нужно изменить его псевдоним, после чего это изменение отразится на всех зонах, в которых был прописан старый псевдоним. Как имена зон, псевдонимы должны быть по возможности короткими и понятными чтобы их сразу же

мог идентифицировать в конфигурации зон администратор SAN.

Стоит рассмотреть возможность использования зон на уровне свичей в комбинации с другими методами, такими как маскирование HBA или LUN контроллера хранения. У зонирования несколько преимуществ, включая управление из одной точки, жесткое зонирование на уровне ASIC, возможность создания виртуальных SAN и ограничение распространения RSCN. Однако, зонирование на уровне узлов имеет свои преимущества и может использоваться для дополнительной безопасности SAN.

## **Secure Fabric Operating System (SFOS)**

В связи с частым применением SAN в критически-важных крупномасштабных средах возникла потребность в обеспечении дополнительной безопасности на уровне фабрики в дополнение к функциям безопасности базовой операционной системы. Brocade Secure Fabric Operating System (SFOS), лицензия на которую предлагается как опция, обеспечивает надежный контроль над тем, как могут подключаться к фабрике коммутаторы, маршрутизаторы и устройства, а также предоставляет дополнительную безопасность каналов, используемых для управления. Функции SFOS сначала были реализованы в версиях 2.6.1, 3.1 и 4.1 ОС для коммутаторов Brocade.

По мере наступления зрелости рынка SAN некоторые функции SFOS оказались слишком сложными, но другие настолько удобными, что их используют все клиенты Brocade и было принято решение перевести функции SFOS на уровень базовой операционной системы. Эти функции делятся на четыре категории:

1. Fabric Configuration Server (FCS) обеспечивает

централизованное управление конфигурациями и политиками фабрики.

2. IP Filters реализуют дополнительный уровень гранулярности при задании того, какие устройства могут получать доступ к коммутаторам SAN с помощью каких приложений. Эта функция обладает параметрами, похожими на межсетевой экран, и используется для контроля доступа к IP-интерфейсу управления.

3. Switch Connection Control (SCC) улучшает взаимную идентификацию коммутаторов за счет использования цифровых сертификатов, а также блокирования превращений части портов в E\_Ports.

4. Device Connection Control (DCC) позволяет подключаться к фабрике через определенный порт или группу портов только отдельным устройствам (по WWN).

Для управления этими функциями можно использовать Fabric Manager, интерфейс WEBT OOLS или Fabric OS CLI.

# 11

## 11: Проектирование территориально-распределенных SAN

Как уже отмечалось в “Главе 2: Решения SAN” и “Главе 4: Обзор проектирования SAN”, за последние годы резко выросла потребность в решениях Disaster Recovery (DR) и Business Continuity (BC), что вызвало существенный спрос на высокоскоростную и высоконадежную передачу данных на уровне блоков между системами хранения данных, установленных на разных площадках. Кроме DR и BC, имеются и экономические факторы, способствующие построению SAN, которые объединяют несколько площадок, в том числе требования к ИТ-подразделениям достигать большего, используя меньше ресурсов. Более эффективное использование систем хранения корпоративного класса также стало стимулом для роста SAN, а технологии обеспечения связи на больших расстояниях позволяют реализовать преимущества SAN для нескольких площадок. Те же экономические тенденции способствуют слиянию компаний и консолидации ЦОДов, для чего необходима миграция данных между площадками.

Для таких внедрений очевидным решением являются географически распределенные сети хранения FC. По возможности, следует использовать темное волокно,

xWDM или шлюзы SONET/SDH. Если эти опции нельзя применять или они слишком дорогие, то как альтернативу можно использовать FCIP.

Поскольку подробное описание технологий построения территориально-распределенных сетей можно найти в посвященных этой теме книгам, то эта глава дает общие рекомендации по определению требований к решению, выбору нужных технологий и продуктов для построения SAN, а также советы по планированию топологии WA N. Ее нужно рассматривать как набор рекомендаций для архитектора, а не детальный учебник по решениям для территориально-распределенных SAN.

## Определение требований

Как и для других сетей, проектирование территориально распределенной SAN следует начать с определения требований. Обычно архитектору нужно найти ответы на следующие вопросы:

- какое расстояние будет поддерживать каждый линк?
  - Некоторые технологии лучше всего подходят для определенных расстояний. Например, «родной» (native) FC по темной оптике (dark fiber) может работать только на ограниченных расстояниях из-за ограничений лазерной оптики<sup>78</sup>.

---

<sup>78</sup>

Ограничением передачи данных через «родные» FC-расширения (extension) являлись буферные кредиты (buffer-to-buffer credits), но технология буферов FC быстро опережает развитие оптики. Коммутаторы и маршрутизаторы Brocade могут поддерживать достаточно буферов для работы FC на полной скорости на несколько сот километров, но оптика обычно поддерживает расстояние менее 100 км.

## C11: Проектирование территориально-распредел. SAN

- о Преодоление этих ограничений возможно и даже рекомендуемо, но для этого могут потребоваться дополнительные технологии, например, повторители.
- о FCIP может поддерживать большие расстояния, но ведет к снижению пропускной способности.
- Насколько быстрой должна быть WAN?
  - о Это включает задержки и пропускную способность. Хотя для решений native FC задержки не являются проблемой, у некоторых других технологий WAN задержки могут оказать влияние на работу приложений.
  - о Также следует учесть производительность при сбоях. Например, если используются резервированные маршруты WAN, то как повлияет на работу приложений выход из строя основного маршрута?
- Каковые требования к надежности и доступности?
  - о Обычно территориально-распределенные SAN более устойчивы к сбоям, чем SAN масштаба ЦОДа, однако это не означает, что сбои допустимы. Нужно выяснить, что произойдет если WAN будет недоступна более суток.
  - о Надежность может повлиять на производительность. Чему равен ожидаемый коэффициент потери пакетов и ошибок на каждом линке?
- Допустимо ли, чтобы проблема WAN дестабилизировала фабрики конечных точек?
  - о В некоторых случаях ответ будет отрицательным, поэтому следует применять

надежную технологию WAN и изолировать ее с помощью маршрутизаторов от конечных точек.

о Чем более критично это требование, тем больше внимания следует уделять использованию таких транспортов, как FC, xWDM или SONET/SDH.

- Каковы требования безопасности для WAN?

о Если через WAN передаются важные данные, то их нужно шифровать. Лучше делать шифрование на уровне хоста и оставаться в зашифрованном виде на устройстве хранения.

о Может ли WAN стать целью для атаки management-plane? Обычно это справедливо для решений IP SAN, но возможно и при использовании других технологий.

В следующих разделах описаны факторы, которые нужно учитывать при проектировании территориально-распределенной SAN.

### *Общие принципы учета расстояний*

- Доступность ресурсов WAN: какой тип инфраструктуры уже доступен?

о Если уже есть ресурсы WAN, то соответствуют ли они требованиям DR? Как это повлияет на текущее использование сети?

о В 1990-ые годы многие организации перестроили свою систему резервного копирования, чтобы вывести трафик резервного копирования из локальной сети, поскольку он не давал другим приложениям использовать локальную сеть. Сочетание трафика DR в WAN с другими приложениями может привести к таким же результатам.

## C11: Проектирование территориально-распредел. SAN

- о Если имеющийся трафик WAN считается важным, то лучше построить отдельную WAN для DR. Наилучший вариант - внедрение решения native FC. Если же этот вариант исключен, то следует использовать SONET/SDH, а, в крайнем случае можно применять FCIP.
- Бюджет и ценность данных: внедрение эффективного решения потребует существенных инвестиций.
  - о Чем более эффективно решение, тем оно дороже. Как его стоимость соотносится с убытками от потери данных и недоступности приложения в результате аварии?
  - о Если решение относительно недорого внедрить, то каковы будут расходы на его поддержание? Решение FCIP можно дешевле внедрить, если использовать существующую IP-инфраструктуру, но его текущее управление и обслуживание может потребовать больше расходов.

### *Общие принципы учета требований миграции данных*

Обычно миграция данных планируется в расчете на то, что во время этой процедуры данные не будут использоваться приложениями. Если же это условие не выполняется, то используется удаленная репликация локальной зеркальной копии и поэтому на миграцию данных меньше влияют сбои, проблемы с производительностью и ошибки, чем на другие решения для территориально-распределенных SAN. Главные критерии при оценке решения для миграции данных – это как долго миграция идет и насколько хорошо продукционная среда изолирована от процесса миграции. Если время для миграции ограничено, то с точки зрения производительности лучше использовать native FC и аналогичные самые мощные решения, а если

такого ограничения нет, то можно применять и FCIP. В обоих случаях рекомендуется изолировать с помощью маршрутизаторов миграцию от производственной среды.

## *Обзор факторов, влияющих на DR*

При планировании территориально-распределенных решений, которые должны обеспечить восстановление после аварий и непрерывность бизнеса нужно учитывать следующие моменты:

- Критичность приложений: какие приложения и данные нуждаются в защите?
  - Критичные серверы приложений и массивы хранения должны быть подключены к резервированным фабрикам “A/B” в соответствии с лучшими практиками.
  - Резервированная фабрика на основной площадке должна быть отделена от WAN и резервной площадки с помощью маршрутизаторов SAN, и передаваться по WAN через LSAN должны только те тома, которые действительно используются в решении DR. Если этого не сделать, то авария на одной площадке или в самой WAN может нарушить работу фабрики другой площадки.
- Допустимое время восстановления: максимальное время, в течение которого приложение может быть недоступным.
  - Определите лучший механизм быстрого восстановления для каждого сценария. Если данные на основной площадке будут испорчены, то можно будет скопировать данные с резервной площадки через WAN и восстановить приложение на резервной площадки либо использовать сервер горячего резерва на

## C11: Проектирование территориально-распред. SAN

резервной площадке?

- о Если сценарий предусматривает копирование данные через WAN, то нужно рассчитать, сколько времени займет эта процедура с учетом пропускной способности каналов между площадками.
- Допустимая потеря данных: сколько данных может быть сгенерировано в основном ЦОДе до того, как их передадут на удаленную площадку?
  - о В случае аварии или небольшого сбоя в основном ЦОДе все данные, которые не были скопированы на удаленную площадку, будут потеряны.
  - о Можно применять как синхронное, так и асинхронное решение. Если для приложения должна быть полностью исключена потеря данных, то требуется синхронное решение с высокопроизводительным транспортом SAN между площадками – обычно native Fibre Channel, xWDM, или FC over SONET/SDH.
  - о ПРИМЕЧАНИЕ: В некоторых продуктах для территориально-распределенных SAN используется механизм ускорения записи (write acceleration), при котором выдается подтверждение завершения записи на хосты основной площадки до того, как данные дойдут до устройств хранения на другом конце. Практически все программные продукты для репликации на уровне хостов и устройств хранения поддерживают этот механизм, уменьшающий риск потери данных, поэтому он не нужен на уровне сетевого оборудования. Если требуется ускорение записи, то используйте его

реализацию на уровне хоста или устройства хранения.

- Расположение: расстояние между ЦОДами должно быть достаточно большим чтобы хотя бы один из них уцелел в случае крупной аварии.
  - От расстояния между площадками может зависеть выбор технологии для WAN. Например, если площадки расположены в разных полушариях, то можно использовать только FCIP или WAFS, а если расстояние составляет несколько сот километров, то следует применять native FC или xWDM.
- Тестирование: нужно регулярно тестировать работоспособность стратегии непрерывности бизнеса, иначе она может не сработать когда возникнет авария.
  - До внедрения нужно решить, какое тестирование требуется, как и когда будут проводиться тесты.
  - Также важно оценить, как тесты повлияют на производственную среду. Маршрутизаторы Fibre Channel с помощью LSAN изолируют тесты на резервной площадке от постоянно выполняющихся операций на основной площадке.

## Кредиты FC Buffer-to-Buffer

Некоторые методы расширения фабрики (например, FCIP и SONE T/SDH) используют шлюзы, отображающие Fibre Channel на другой протокол для передачи пакетов между площадками. Несмотря на преимущества шлюзов они имеют высокую цену, сложны в управлении и обычно снижают производительность по сравнению с применением native Fibre Channel.

## C11: Проектирование территориально-распред. SAN

Также можно расширять линки Fibre Channel ISL или IFL на большие расстояния. Обычно такими решениями проще всего управлять и они дают самую большую производительность. Однако, для обеспечения максимальной производительности ISL на больших расстояниях архитектор SAN должен хорошо разбираться в механизме буферных кредитов (buffer-to-buffer, В В или В2В). Например, для портов, которые соединяют коммутаторы центральные коммутаторы двух площадок на Рисунке Рис. 44 (стр. 206) могут потребоваться дополнительные буферы для поддержки максимальной скорости между площадками 1 и 2 в зависимости от расстояния между площадками. В этом разделе объясняется механизм кредитов ВВ и его использование при связи на большом расстоянии.

Хотя сеть может отбрасывать пакеты при определенных типах сбоев (например, обрыве кабеля), стандарты Fibre Channel не разрешают сети отбрасывать пакеты при нормальной работе, поэтому порт-отправитель не может передавать пакеты, если порт-получатель не готов их принять.

Вполне возможно, что порт-получатель не может сразу же передать пакет дальше, поэтому ему нужен буфер для временного хранения пакетов. Все устройства в SAN имеют ограниченное число буферов, поэтому необходим механизм, с помощью которого они могут сообщить другим устройствам, что у них есть свободные буферы до того, как к ним будет послан пакет, т.е. отправитель должен знать, способен ли принять пакет получатель. Контроль потока (Flow control) используется для того, чтобы никакой порт в цепочке между отправителем и получателем не мог отбросить пакет.

Имеются различные способы реализации контроля потока в сети и в Fibre Channel используется механизм кредитов между буферами (buffer-to-buffer credits). Все

устройства Fibre Channel (коммутаторы, маршрутизаторы, HB A, JBOD-диски и контроллеры дисковых массивов) используют кредиты ВВ. Механизм кредитов буферов гарантирует, что пакеты не будут посыпаться до тех пор, пока отправитель не получит от получателя подтверждение, что у него освободились кредиты буферов.

Один кредит буфера равен одному пакету независимо от длины последнего. Длина пакетов Fibre Channel может быть от 60 байтов до 2148 байтов (~2k). Большинство пакетов имеют длину 2k, поэтому такие устройства, как НВА и дисковые массивы договариваются о длине пакетов 2k для передачи данных. Проведенные Brocade исследования показали, что не менее 95% всех пакетов имеют длину 2k, а более короткие пакеты используются в основном только для команд SCSI и служебного трафика FC класса F (обновление информации о зонах, RSCN, информация сервера имен и т.п.), поэтому на практике размер пакета Fibre Channel можно считать равным 2k.

Когда порт устройства первоначально подключается к фабрике, то устройство-получатель говорит отправителю, сколько у него свободных буферов, а отправитель сообщает ему, сколько ему буферных кредитов нужно. Когда узел договаривается о кредитах ВВ с коммутатором, то обычно он запрашивает от двух до 16 буферных кредитов. Коммутатор в ответ сообщает число буферных кредитов, которые может выделить порту. Это и есть число кредитов, которые можно потратить.

Когда пакеты передаются от узла к коммутатору, то они записываются в память коммутатора<sup>79</sup> и передающее

---

<sup>79</sup>

Пакеты сохраняются если коммутатор не может их сразу же переслать. В коммутаторах Brocade обычно происходит “cut-thru switching”,

## C11: Проектирование территориально-распред. SAN

устройство делает отметку о занятии буферных кредитов, которые оно использовало. Оно прекращает передачу если больше не осталось свободных кредитов. Между тем, получатель попытается освободить буферы, пересылая пакеты дальше, и оповестит отправителя если ему удастся освободить буферы. Таким образом, кредиты В2В возвращаются узлу максимально быстро и, если нет проблем в фабрике, устройство может постоянно передавать пакеты со скоростью физической линии. Если где-то (в фабрике или конечной точке-получателе) возникнет переполнение трафиком линии, то в конце концов все кредиты буферов будут исчерпаны и передающее устройство будут ждать освобождения сети и возобновит пересылку данных только когда у нее будет хотя бы один кредит.

Коммутаторы Brocade FC обычно прозрачно для пользователей обрабатывают буферные кредиты. Например, таким локальным портам, как F-port или FL-port не нужно много кредитов для поддержания скорости линии. По умолчанию, продукты Brocade 4G bit «рекламируют» (advertise) восемь кредитов на портах F и FL, что более чем достаточно для поддержания скорости линии на уровне 4Gbit в крупном ЦОДе без какого-либо участия пользователей.

Однако при использовании Fibre Channel на больших расстояниях (например, свыше 500 метров) использование кредитов ВВ очень важно, особенно при очень больших расстояниях (10 – 500 км) при передаче

---

при котором пакет начинает пересылаться коммутатором дальше еще до того, как он полностью получен. В этом случае передающее устройство быстро получает обратно кредиты ВВ. Коммутаторы FC конкурентов используют устаревшую технологию “store and forward”, которая не обеспечивает быстрого освобождения буферов и поэтому потребляет больше кредитов при одинаковой пропускной способности.

по темному волокну, с помощью оборудования WD M или SONET/SDH, поскольку передача пакета по оптическому кабелю занимает определенное время. Разумеется, пакеты передаются со скоростью света и поэтому это время не может быть очень большим, но при достаточно большой длине кабеля и скорости линка одновременно по кабелю может передаваться сразу много пакетов. Например, по линку 10Gbit длиной 120 км одновременно могут передаваться несколько сотен пакетов. В этом случае передающий коммутатор должен быть уверен, что у получателя хватит памяти чтобы получить по крайней мере столько пакетов, сколько передается по линку в один момент времени поскольку этот коммутатор не получит кредит обратно сразу же по получению пакета. Архитектор SAN должен предусмотреть, чтобы было достаточно кредитов для заполнении ISL.

Грубо говоря, правило таково: требуется один кредит BB на один километр для работы на полной скорости 2Gbit. Для поддержания полной пропускной способности на расстоянии 30 км нужны 30 буферов, а если доступны только 15 кредитов, то пропускная способность упадет, например, до 1Gbit. С точки зрения длины, при одинаковом числе кредитов линк 1Gbit будет вдвое длиннее линка 2Gbit. Для линка 4Gbit требуется вдвое больше кредитов на километр, чем для линка 2Gbit. Для определения числа кредитов для линка native FC в зависимости от расстояния используется следующая формула:

$$\text{Число кредитов} = \text{километров} * (\text{Gigabits} / 2)$$

Стоит отметить, что линки FC смогут работать если число свободных буферов недостаточно, но передача будет идти медленнее. 100- километровый линк 2Gbit сможет работать и с только 75 кредитами буферов, но его максимальная пропускная способность будет

## C11: Проектирование территориально-распред. SAN

~1.5Gbit.

Важно помнить, что число передаваемых пакетов равно числу кредитов, т.е. если нужно обеспечить максимальную скорость для линка длиной 100 км (в данном случае 2Gbit), то нужно передавать 100 пакетов. В данном примере подразумевается, что длина пакетов - 2k, а если будут использоваться пакеты 1k, то потребуется вдвое больше кредитов. Хотя это может показаться странным, но эта особенность помогает лучше понять механизм кредитов буферов Fibre Channel.

Также стоит отметить, что если 30-километровый ISL работает на 2Gbit/se с и у него 100 кредитов, то его производительность не будет лучше по сравнению с тем, если бы ему были выделены только 30 кредитов, т.е. нельзя «разогнать» ISL выделением дополнительных кредитов.

### **Режимы LD передачи на большие расстояния**

Brocade предлагает несколько режимов передачи на большие расстояния (Long Distance, LD):

**Таблица 3 – Режимы передачи на большие расстояния**

Режим передачи	Расстояние	Требуется лицензия
L0 (локальный E-port)	<500 метров	Нет
LE 1	0 км	Нет
L0.5 25	км	Да
L1 100	км	Да
L2 200	км	Да
LD	Автоматическое обнаружение	Да
LS	Статическая конфигурация	Да

Есть простые правила выбора нужного режима.

Все локальные или “нормальные” порты E-port и EX-port, которые не используются для удаленной передачи, относятся к режиму L0, LE относится к расстояниям не более 10 км и не требует лицензии (на ограничение 10 км не влияет скорость передачи и если порты договорятся о скорости 4Gbit, то автоматически выделяются дополнительные кредиты для поддержки этой скорости на линии).

L0.5, L1 и L2 первоначально использовались для упрощения конфигурирования линков для больших расстояний, но потом стало ясно, что эти режимы не являются гибкими, поскольку они полностью статичны и кредиты автоматически выделяются в соответствии с режимом и скоростью – как и для LE они назначались в зависимости от скорости порта для поддержания заданного в этой таблице расстояния. Ожидается, что в будущем эти режимы преобразуются в LD или LS.

LD (dynamic distance discovery mode – режим динамического обнаружения расстояния) является самым удобным режимом. Он автоматически проверяет линк и с помощью сложного алгоритма вычисляет число необходимых кредитов исходя из расстояния и скоростей линка.

LS – это статически сконфигурированный режим, который появился в FOS 5.1. Это самый гибкий режим для более опытных пользователей, который позволяет обеспечить полное управление при специальных требованиях. Например, задано требование работы в линке длиной 110 км, на 2Gbit при длине фреймов 1k вместо обычных 2k. Это означает, что требуется вдвое больше буферных кредитов, а именно 220.

## Скорости и технологии MAN/WAN

## C11: Проектирование территориально-распред. SAN

В этом разделе рассматриваются и сравниваются некоторые популярные технологии построения территориально-распределенных SAN.

**FC Over Dark Fiber.** Трафик native Fibre Channel, который обычно идет через E\_Ports или EX\_Ports, передается по выделенной оптической линии, соединяющей площадки. Если темная оптика арендуется, то ее владелец не предоставляет сервисов на линии<sup>80</sup> и их должен обеспечить сам клиент. FC через темную оптику (over dark fiber) обеспечивает высокую надежность и производительность, поскольку при применении этого подхода до минимума снижается использование оборудования, которое может отказать, не требуется конверсия протоколов и полоса пропускания используется только приложениями, которые относятся к приложениям SAN. Эта технология меньше всего влияет на задержки и на практике этот подход имеет смысл применять когда расстояние не более нескольких сот километров<sup>81</sup> при использовании маршрутизаторов Brocade и большинства коммутаторов 4 и 8 Gbit, поскольку у этих устройств много буферных кредитов на порт.

- **FC Over xWDM.** Соединения native Fibre Channel осуществляются с подключением к аппаратуре уплотнения канала по длине волны (DWDM или CWDM, оба типа этих устройств обозначаются как xWDM.) Трафик передается по выделенной длине волны, но по

---

<sup>80</sup> Именно так темное волокно получило свое название – пока клиент не подключит свое оборудование на другом конце к кабелю провайдера, кабель остается темным (по нему не идет свет). Другие сервисы используют оборудование провайдера, которое передает свет лазера по кабелю.

<sup>81</sup> Могут понадобиться специализированные SFP и/или повторители сигналов.

оптике между площадками могут работать и другие сервисы, использующие другую длину волны. По производительности этот подход аналогичен темному волокну и не добавляет задержек кроме тех, что связаны со скоростью света или ограничением пропускной способности, и практически также надежен, как темное волокно (если не считать риск отказа самого устройства WDM). В то же время он поддерживает передачу на большие расстояния, чем темное волокно, если промежуточные WDM используются как повторители. Следует отметить, что в некоторых случаях коммутаторы или маршрутизаторы FC на обоих концах должны поддерживать большое число кредитов ВВ если сам WDM не обеспечивает этого.

- **FC Over SONET/SDH.** Можно передавать сигналы Fibre Channel через сети Synchronous Optical Networks (это сервис за пределами Америки обычно называется Synchronous Digital Hierarchy). При этом, в зависимости от провайдера, может использоваться сервис OC3, OC12 и даже native FC. Более медленные линии, например E3/T3, можно использовать для сокращения расходов на подключение. SONET/SDH добавляют минимальную задержку, часто способны обеспечить полную полосу пропускания FC, обладают высокой надежностью и могут охватывать большие расстояния при условии, что у оператора есть соответствующие мощности.
- **FC Over ATM.** Передача пакетов Fibre Channel по ATM была одним из первых методов построения территориально-распределенных SAN. Хотя популярность ATM снижается, такой подход хорошо проверен на практике. Он поддерживает высокую детерминированность, с точки зрения задержек работает лучше, чем FCIP и обладает высокой надежностью, однако маршрутизаторы и сервисы ATM обычно дорогие.

## C11: Проектирование территориально-распред. SAN

- **FC Over IP.** Основное преимущество FCIP – возможность относительно недорого подключаться к практически везде доступным сервисам MAN/WAN. В то же время, это самый ненадежный и медленный вариант из-за того, что обычно в IP WAN используется много хопов через оборудование разных операторов (поэтому среду передачи труднее конфигурировать) и в них заложена переподписка. В результате, коэффициенты переполнения и потери пакетов больше, чем у FC SAN. В теории, FCIP может использовать коммутируемые сервисы IP/Ethernet или выделенные соединения точка-точка (point-to-point) Gigabit Ethernet. На практике второй вариант не используется. Если есть соединения point-to-point, то лучше использовать native FC. Для большинства приложений работа с помощью FCIP невозможна без применения технологии ускорения записи.

Выбор конкретной технологии MAN/ WAN зависит от нескольких факторов:

- **Доступности сервиса.** Способен ли провайдер обеспечить сервис на каждой площадке? Например, если между площадками нет темного волокна, то придется использовать другую технологию как бы идеально темное волокно не подходило для требований сети. Обычно как альтернативу можно использовать FCIP поскольку IP-сервисы доступны практически везде, однако иногда трудно найти IP-сервисы с нужной полосой пропускания и соглашением об уровне сервиса (SLA).
- **Требования приложений к RAS.** Если у приложения повышенные требования к надежности, доступности и обслуживаемости, то можно применять любую технологию, способную обеспечить соответствующий SLA и использующую только технологию

корпоративного уровня. Однако практически всегда легче обеспечить RAS с помощью таких решений native Fibre Channel, как темное волокно и xWDM, поскольку при этом будет меньше компонентов от меньшего числа вендоров. Отсутствие преобразований протоколов для native FC означает меньше сложности, что уменьшает риск ошибок и упрощает диагностику и устранение проблем. Чем меньше аппаратных компонентов обеспечивают соединение, тем надежнее оно будет работать. С точки зрения RAS худшим вариантом является применение FCIP, а SONET/SDH и ATM являются компромиссными решениями.

- **Требования к производительности приложений.**

Многие приложения чувствительны к задержкам и частоте ошибок на устройствах хранения, в то время как другие приложения менее чувствительны.

Производительность на хостах при синхронном зеркалировании на расстоянии будет существенно деградирована, если только производительность WAN не лучшая в своем классе. В то же время, асинхронное зеркалирование приложений обычно менее чувствительно к задержкам и частоте ошибок. Синхронным приложениям лучше подходят SONET/SDH, ATM, xWDM, и темная оптика; асинхронные приложения могут использовать FCIP, но по-прежнему остается важным использование WAN с наивысшим уровнем обслуживания (SLA). Аналогично, различные приложения требуют больше или меньше пропускной способности. IP SAN-соединение через линию ISDN 128k не будет полезно для большинства заказчиков, поскольку оно не поддерживает достаточной пропускной способности. Лучшим решением по производительности *всегда* является native FC, передаваемый или по темной оптике или через xWDM для средних расстояний, и FC over SONET/SDH или ATM для больших расстояний.

## C11: Проектирование территориально-распредел. SAN

- **Расстояние между площадками.** Некоторые технологии (например, темное волокно и xWDM) можно использовать только в MAN и при небольших расстояниях WAN, а другие, такие, как FCIP или iSCSI, могут работать и на больших расстояниях, но они вносят задержку и риск потери пакетов. Для больших расстояний обычно лучше подходит SONET/SDH и ATM.
- **Стоимость решения.** Какова стоимость разных вариантов сервисов и сетевой инфраструктуры как с точки зрения первоначальных затрат, так и текущего обслуживания? Например, если для приложения оптимальный вариант - это SONET/SDH, но у заказчика есть только половина необходимого бюджета, то придется применить другое решение либо отложить проект. Внедрение FCIP стоит меньше, чем других вариантов благодаря широкой доступности и дешевизне IP-сети, но в долговременной перспективе оно может оказаться более дорогим из-за низкой производительности и надежности.
- Кроме выбора между native FC, ATM, SONET/SDH и IP часто нужно выбрать низкоуровневый протокол. В Таблице 4 перечислены некоторые технологии MAN/WAN вместе с их скоростями.

Таблица 4 – Технологии и скорости MAN/WAN

технология	скорость
ISDN BRI	128kbits
FracT1	$\leq 1.5\text{Mbps}$
ADSL <sup>82</sup>	$\leq 1.5\text{Mbps}$
ISDN PRI (NA)	1.5Mbps
DS1/T1 1.	5Mbps
ISDN PRI (E)	2Mbps
E1 2M	bps
Ethernet	10Mbps
E3 34M	bps
DS3/T3 45M	bps
Fast ENet	100Mbps
OC3	155Mbps
STM1 155M	bps
технология	скорость
OC12	622Mbps
STM4 622M	bps
Native GE (1)	1Gbps
Native FC (1)	1Gbps
Native FC (2)	2Gbps
OC48	2.5Gbps
STM16	2.5Gbps
Native FC (4)	4Gbps
Native FC (8)	8Gbps <sup>83</sup>
OC192	10Gbps
STM64	10Gbps
Native GE (10)	10Gbps
Native FC (10)	10Gbps

Подчеркнем, что скорость ниже 100Mbps может быть слишком низкой для многих *синхронных* приложений SAN и часто требуется пропускная способность на уровне нескольких гигабит. OC3/STM1 обычно удовлетворяет таким требованиям, а применение OC12/STM4 и больше может дать лучшие результаты.<sup>84</sup> Асинхронные приложения обычно способны работать на меньших скоростях и использовать менее надежный транспорт, например IP.

---

<sup>82</sup> Многие провайдеры предлагают многогигабитный ADSL, но в SAN эта технология практически не используется.

<sup>83</sup> На момент написания книги этот протокол не был еще широко распространен и не был эффективен по стоимости.

<sup>84</sup> Для заполнения этих каналов могут потребоваться несколько порталов FCIP на шлюзе.

## “Ограничивающая” архитектура маршрутизации BC/DR

При проектировании территориально-распределенных SAN обычно нужно изолировать площадки между собой чтобы сбой на одной не затронул и остальные, а также изолировать все площадки от нестабильности в MAN/ WAN. Также желательно изолировать сервисы фабрики каждой площадке чтобы свести к минимуму риск того, что ошибка администратора на одной площадке нарушит работу фабрики на другой. По этим причинам в настоящее время при построении территориально-распределенных SAN на каждой стороне MAN/WAN размещаются один или несколько маршрутизаторов и затем площадки выборочно соединяются с помощью зон LSAN.

Подробно архитектура маршрутизаторов FC рассмотрена в книге *Многопротокольная Маршрутизация для SAN (Multiprotocol Routing for SANs)* Джоша Джада. Здесь мы только упомянем, что почти всегда рекомендуется использовать интерфейсы EX\_Port маршрутизаторов с целью *ограничения (bracket)* MAN/WAN для наилучшей возможной изоляции фабрик на площадках от сбоев. Таким образом, обмен пакетами между устройством на Площадке1 (Site1) с устройством на Площадке2 (Site2) происходит примерно так: Site 1 Device → Fabric1 → Fabric1 E\_Port → Router1 EX\_Port → Router1 E\_Port → MAN/ WAN Transport (Backbone Fabric) → Router2 E\_Port → Router2 EX\_Port → Fabric2 E\_Port → Fabric2 → Site2 Device.

Разумеется, иногда лучше подойдет и другой подход, но «по умолчанию» архитектуре SAN лучше ограничивать MAN/WAN этим способом.

## FastWrite и Tape Pipelining

Запаздывание в некоторых случаях может оказывать влияние на работу приложений, существенно большее чем могло бы интуитивно казаться. Даже если запаздывание initial-to-target на несколько порядков меньше времени поиска данных на диске, оно может повлиять на производительность приложения из-за устройства протокола SCSI.<sup>85</sup>

Запаздывание port-to-port у коммутаторов Brocade составляет от сотен наносекунд до нескольких микросекунд и, теоретически, оно должно быть незаметно для приложений, но, к сожалению, запаздывание между коммутаторами из-за скорости света создает проблемы в MAN/WAN, которые нельзя устраниить более быстрым перемещением пакетов между коммутаторами. В данном случае запаздывание происходит не внутри коммутатора, а в оптическом кабеле между коммутаторами. В высокоскоростных решениях DWDM с каналами FC длиной более 100 км, по одному каналу могут одновременно передаваться многие сотни пакетов.

Это может ухудшить производительность из-за того, что протокол SCSI иногда должен работать в холостом режиме во время кругового движения пакетов (Round Trip Time). Когда инициатор хочет послать данные получателю, то сначала он запрашивает разрешение, ждет обратного ответа на этот запрос и только после его получения передает данные. При соединении с относительно большим запаздыванием это может привести к значительному времени ожидания, когда инициатор ждет от получателя обратного ответа на

---

<sup>85</sup>

Хотя определение протокола как FCP будет более точным, но обсуждаемые в этом разделе свойства унаследованы от SCSI.

## C11: Проектирование территориально-распред. SAN

запрос и в это время данные не передаются. В результате, резко снижается коэффициент использования полосы пропускания MAN/WAN ( см. Иллюстрацию 65 на стр. 362).

На горизонтальной оси этой диаграммы показано расстояние между инициатором и получателем при передаче данных через WAN, а на вертикальной оси - время. Сначала посыпается запрос на запись, который практически мгновенно пересыпается на локальный маршрутизатор, однако до удаленного получателя он доходит за определенное время и еще больше времени уходит на получение ответа на запрос. Этот интервал называется временем Round Trip Time (RTT) и до его окончания инициатор не может послать данные.

Для решения этой проблемы можно использовать интеллектуальную буферизацию с поддержкой SCSI в коммутаторах и маршрутизаторах SAN для сокращения времени ожидания запроса на команду записи. Это решение используется в нескольких продуктах Brocade, предназначенных для различных расстояний. Например, у Brocade в шлюзах FCIP используются функции Fast-Write и Tape Pipelining и аналогичные функции для FC ISL, проходящих через MAN. Архитектура данных сервисов описана в английской версии книги в приложениях “FCIP FastWrite and Tape Pipelining” и “FC FastWrite”. В результате, удается сократить один интервал RTT (см. Рис. 65 – SCSI Write без FastWrite).

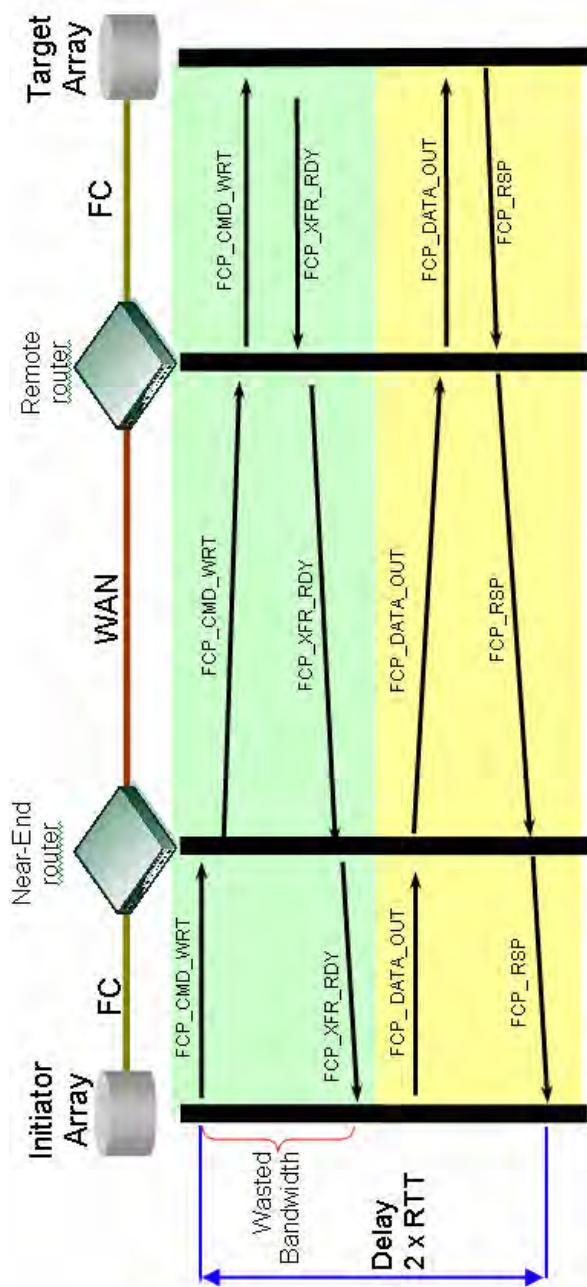


Рис. 65 – SCSI Write без FastWrite (быстрой записи)

## C11: Проектирование территориально-распред. SAN

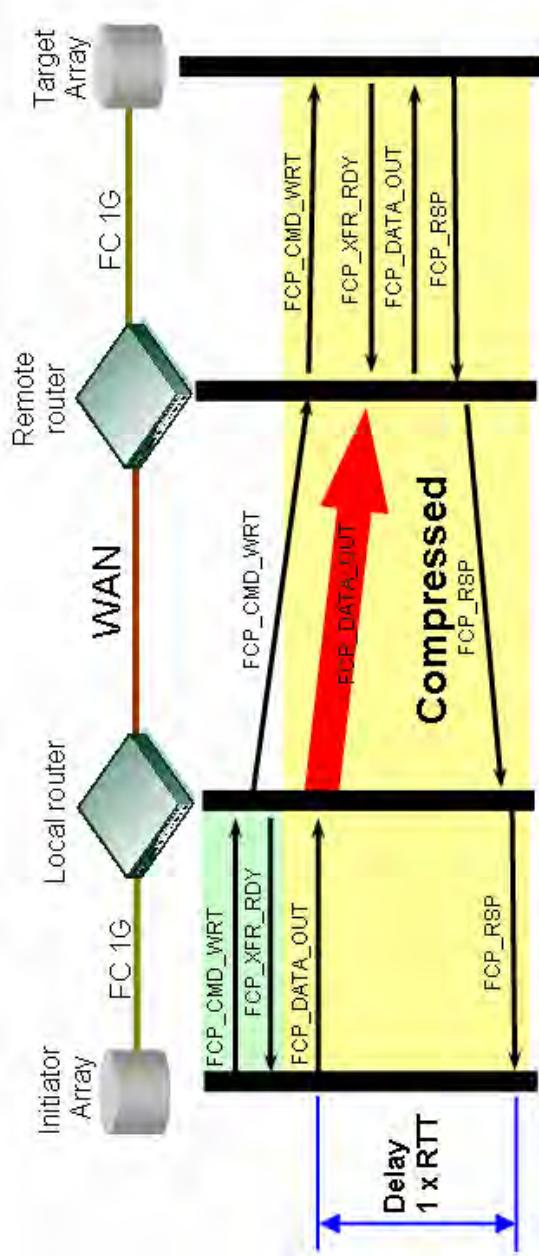


Рис. 66 – SCSI Write с поддержкой FastWrite (быстрой записи)

С точки зрения архитектора SAN важно знать, что сервисы ускорения FCIP внедряются централизованно на самих портах FCIP. Но поскольку для решений native FC требуется значительно более высокая производительность чем для FCI P, то FC FastWrite внедряется на уровне порта узла. Для этого узлы (инициаторы), работу которых требуется ускорить, должны напрямую подключаться к устройствам с необходимой аппаратной поддержкой, например, к лезвию-маршрутизатору FR4-18i. В следующем разделе приведен пример внедрения FC FastWrite.

## Лезвия 10Gbit и решения DR/BC

На крупных предприятиях для решений восстановления после аварий (DR) или обеспечения непрерывности бизнеса (BC) иногда требуется очень высокопроизводительная сетевая инфраструктура. Эта потребность может быть связана с очень высокой скоростью изменения данных, или с очень большим объемом данных или с очень коротким временем восстановления. В большинстве подобных инфраструктур соединения на большие расстояние организованы с помощью *x*WDMs или темной оптики, которые обеспечивают наивысшую пропускную способность. Этот сценарий – один из немногих, в которых технология 10Gbit лучше работает, чем транкинг интерфейсов 4Gbit, поскольку более высокие цены на лезвия и кабели 10Gbit полностью компенсируются за счет сокращения в три раза числа необходимых линков темной оптики между площадками или определенной длины волны WDM для одной и той же пропускной способности. Экономия затрат на текущее обслуживание может всего за один месяц компенсировать разницу в цене оборудования.

### C11: Проектирование территориально-распред. SAN

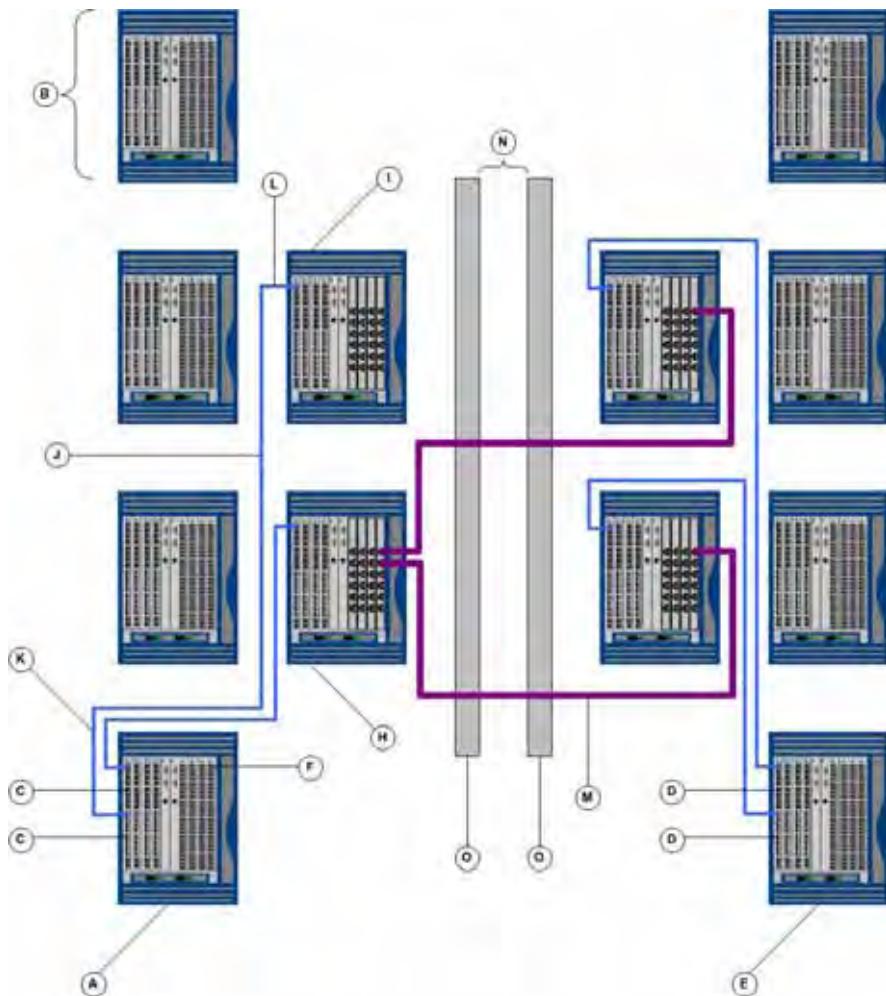


Рис. 67 – крупное маршрутизируемое решение 10Gbit DR/BC

В подобных случаях часто оптимальным решением является комбинация DWDMs с интерфейсами 10Gbit FC, директоров Brocade с лезвиями 10Gbit FC, лезвий FR4-18i для маршрутизации FC и дополнительных лезвий FR4-18i для ускорения записи. Иногда такие территориально-распределенные конфигурации получаются сложными. Пример такого решения DR/BC

показан на Рис. 67. Эта конфигурация достаточно сложна и нуждается в дополнительных пояснениях.

На диаграмме для наглядности показана только часть соединений между портами, например, только малая часть линков между фабриками, связывающими шасси периферийной фабрики и два шасси маршрутизатора на каждой площадке. Остальные соединения по своей структуре аналогичны и читатель может самостоятельно дорисовать их. Как это все работает:

На рисунке показана половина резервированной SAN, т.е. "Meta SAN A" (сторона "B" будет идентична). Конфигурация состоит из восьми однодоменных периферийных фабрик (по четыре на каждую площадку) и по одной фабрике магистрали на каждую Meta SAN, которые избыточно соединены через сеть DWDM. Она обеспечивает около 480Gbits полезной пропускной способности между площадками Meta SAN. Если используется дублирование каналов по схеме active/active, то это даст общую пропускную способность 960Gbits (1.9Tbits в полном дуплексе) при репликации или зеркализации между площадками.

Каждое периферийное шасси (одно из них отмечено (A) в левом нижнем углу рисунка) оборудовано четырьмя лезвиями FR4-18i для ускорения записи репликации между системами хранения или зеркализации портов, плюс сторона E\_Port IFL (K) и четыре 48- портовых лезвия для хостов (F). Разные периферийные шасси, такие, как (B), включают в себя разные периферийные фабрики, поэтому максимальный размер периферийной фабрики - 192 хоста и 64 порта устройств хранения. Можно подсоединить восемь портов устройств хранения для репликации или

## C11: Проектирование территориально-распред. SAN

зеркалирования<sup>86</sup> (C) к лезвиям FR4-18i в периферийном шасси так, чтобы четыре порта были подсоединенны к одной quad, затем объединить четыре порта в транк IFL, затем подключить еще четыре порта устройств хранения, которые образуют еще один транк IFL. Такой метод подключения дает максимальный выигрыш от локальной коммутации в директоре и позволяет максимально эффективно использовать дорогие отсеки для лезвий шасси.

Цель – ускорить репликацию на симметрично сконфигурированные устройства хранения на другом конце (D), который подключен к симметрично сконфигурированному шасси (E). Хосты на каждой фабрике (например, (F)) могут обращаться к тем же портам хранения, которые используются для репликации или разным портам устройств хранения этого же шасси или на другом шасси через LSAN.

Эта конфигурация очень похожа на резервированную архитектуру Центр/Периферия на каждой площадке с соединением центральных коммутаторов на большое расстояние для DR/BC. Однако, центры каждой площадки – это на самом деле резервированные маршрутизаторы в шасси, например, (H) и (I), а не коммутаторы или директоры второго уровня. Каждое периферийное шасси (например, (A) и (B)) соединяется с обоими центрами через несколько 4-портовых транков IFL (J). Всего восемь таких транков соединяют каждый периферийный коммутатор и пару центров: по четыре 4-портовых транка IFL между каждым периферийным и каждым центральным коммутатором. Эти IF L используют порты FR4 на обоих концах, хотя периферийные порты сконфигурированы как

---

<sup>86</sup>

Дальше мы называем это “репликацией”.

стандартные E\_Ports (K), а порты центра (L) - как EX\_Ports. Одна из причин для использования портов FR4 на периферии – это то, что они могут локально коммутировать реплицируемые порты, а также улучшают надежность и обеспечивают оптимальное использование полосы пропускания лезвий FR4-18i, для которой нет переподписки (не рекомендуется подсоединять высокопроизводительные порты устройств хранения к лезвиям с переподпиской). Трафик будет равномерно распределяться между транками IFL через DPS и внутри каждого пути с помощью транкинга на уровне пакетов.

Каждый центр одной площадки подсоединен к каждому центру другой площадки через матрицу 10-гигабитных ISL (M). Расстояние между площадками (N) около 120 км и соединение обеспечивает DWDM (O) на каждой площадке. Имеется двенадцать (12) 10Gbit ISL между каждой парой шасси и двадцать четыре (24) пути с одинаковой стоимостью между каждым центром и каждой периферийной фабрикой-получателем. Эти линки балансируются комбинацией DPS на Condor ASIC лезвий FR4-18i, аналогичная схема балансировки реализуется чипами маршрутизатора лезвий и внутренним механизмом балансировки backplane на уровне пакетов внутри шасси. В любом случае, такая матрица образует одну сбалансированную резервированную backbone фабрику.

Мы привели достаточно подробное описание сложного решения. Возможны различные варианты такой конфигурации, но анализ поддержки для нее трудно провести, поэтому, как и в случае с крупномасштабными конфигурациями, перед развертыванием лучше проконсультироваться у соответствующего партнера по поддержке и/или в Brocade Professional Services.

## Ограничения расстояния для оптоволокна

Для каждого типа оптики и кабелей имеются свои ограничения на расстояния, которые нужно учитывать при проектировании сети с каналами native FC. Также следует учитывать, что поддерживаемое расстояние включает в себя, как переменную, скорость линии – если остальные параметры равны, то чем больше скорость, тем меньше максимальное расстояние (см. Рис. 68).

Тип трансивера	Форм-фактор	Скорость, Gbps	Максимальное расстояние			
			многомодовый кабель			Одномодовый кабель
SW			62.5µm/200MHz (OM1)	50µm/500MHz (OM2)	50µm/2000MHz (OM3)	9µm
	SFP	1	300m	500m	860m	N/A
	SFP/SFP+	2	150m	300m	500m	N/A
	SFP/SFP+	4	70m	150m	380m	N/A
	SFP+	8	21m	50m	150m	N/A
LW	XFP	10	33m	82m	300m	N/A
	SFP	2	N/A	N/A	N/A	30km
	SFP	4	N/A	N/A	N/A	80km
	SFP+	8	N/A	N/A	N/A	25km
	XFP	10	N/A	N/A	N/A	10km

Рис. 68 – Зависимость расстояния от оптики, скорости и типа кабеля (таблица дополнена с учетом новейших модулей SFP - прим. переводчика)

В таблице показана зависимость поддерживаемого расстояния от комбинации типа трансивера, скорости

линка и типа кабеля. Помимо увеличения скорости технологий FC, поддерживаемое расстояние сокращалось. Если в конце 1990-х годов 1Gbit FC при использовании лазеров с короткими волнами (SW) поддерживал расстояние до 300 метров по кабелю Multi-Mode Fiber (MMF) 62.5  $\mu\text{m}$  OM1, который тогда использовался в большинстве ЦОДов, то при переходе на 2Gbit максимальное расстояние сократилось до 150 метров.

Однако, в то время в ЦОДах стали использовать более мощные кабели и при применении стандарта 50 $\mu\text{m}$  OM2 2- гигабитный FC снова смог поддерживать расстояние до 300 метров, а когда для 4Gb it FC стали использовать следующий стандарт OM3, то поддерживаемое расстояние увеличилось до 380 метров. Это расстояние достаточно для большинства приложений ЦОД, за исключением самых требовательных. Аналогичное увеличение поддерживаемого расстояния ожидается и при переходе на скорости 8Gbit и 10Gbit.

Для расширения линков native FC на расстояния между ЦОДами нужно использовать специальные типы кабелей и оптики. Вместо кабелей MMF для линков на большие расстояния используются кабели Single-Mode Fiber (SM F) 9  $\mu\text{m}$ . При использовании лазеров с длинными волнами (LW) они позволяют довести длину ISL до нескольких десятков километров. Некоторые вендоры предлагают лазеры со сверхдлинными волнами, которые увеличивают это расстояние во много раз. Если эта оптика доступна с немного другой длиной волны, то ее можно использовать вместе с технологией CWDM для передачи нескольких ISL по одному физическому кабелю.

Заказчикам, которым нужно передавать сигналы FC на расстояние свыше десятков километров лучше

## C11: Проектирование территориально-распред. SAN

всего использовать DW DM, тогда оптика и кабели фабрики FC должны поддерживать относительно короткое расстояние между коммутаторами/маршрутизаторами и локальными устройствами DWDM. При этом DWDM будут отвечать за передачу сигнала в удаленный ЦОД.

См. также “Оптические кабели и волокно” начиная со стр. 381.

# 12

## 12: Планирование внедрения

В контексте SAN планирование внедрения означает обеспечение удобной инсталляции и обслуживания. В этой главе нет детального руководства по инсталляции и обслуживанию, как и пошаговых инструкций по установке в стойке, поэтому оно может использоваться как замена руководств по оборудованию и программному обеспечению. Для больших инсталляций SAN проектирование и внедрение – это разные функции, которые обычно выполняют разные люди, а эта книга рассчитана на архитектора и поэтому в этой главе говорится только о некоторых особенностях, которые нужно учитывать при проектировании SAN, а также включить в документацию проекта SAN.

### **Расположение и монтаж стоек**

До выбора конкретного проекта следует убедиться, что подходящее место существует для всего нового оборудования SAN. А именно, должны иметься шкафы правильного типа и их расположение должно быть уместно.

Различные монтируемые в шкафы (rack) коммутаторы, маршрутизаторы, шлюзы, хосты и системы хранения имеют разную высоту, интеграционные потребности, вес и даже ширину. На самом деле, один вендор SAN- оборудования продает шасси директоров, которые по весу превосходят лимиты

лифтов большинства шкафов, что означает сложность или невозможность безопасной инсталляции их оборудования. По вполне понятным причинам следует убедиться, что шкафы, предназначенные для монтирования SAN- оборудования имеют те же характеристики, что и монтируемое в них оборудование. Это может потребовать покупку новых шкафов и лифтов (racks, lifts, cabinets) специально для SAN.

Также важно, чтобы направление потоков воздуха в устройствах, которые размещены в одной стойке или в соседних стойках, совпадало. У некоторых устройств воздух подается с передней части назад (front-to-back), а у других воздух идет в обратном направлении (back-to-front). Если требуется установить оба типа устройств в одной стойке, то устройства одного из двух типов должно быть размещено так, чтобы впереди была его задняя панель, чтобы все устройства вытягивали холодный воздух с одной стороны стойки и горячий воздух выходил с другой стороны. В противном случае горячий воздух от вентиляторов одних устройств будет попадать на забор вентиляторов других устройств и в результате возникнет перегрев оборудования.

К счастью, все продукты Brocade разработаны так, что они соответствуют обслуживаемому ими OEM- оборудованию, поэтому в Brocade SAN редко возникают проблемы с охлаждением. Далеко не у всех вендоров нет таких проблем, поэтому архитектуре следует учитывать это обстоятельство.

Например, один поставщик инфраструктуры SAN использует для охлаждения воздушные потоки, которые идут между боковыми сторонами стойки (*side-to-side*). Такая организация охлаждения была стандартной на ранних этапах развития оборудования для IP- сетей. Старое оборудование было не таким мощным и плотно упакованным, как современное, поэтому температура

внутри стойки была намного меньше. Кроме того, раньше в ЦОДах было много свободной площади для размещения стоек. Сегодня, в эпоху сверхскоростных сетей и ЦОДов с блейд-серверами трудно правильно установить оборудование с устаревшим боковым охлаждением – если его установить в двух соседних стойках, то горячий воздух от одной стойки будет подаваться на вентиляторы оборудования второй стойки. Единственный выход в такой ситуации – это заполнить стойку только наполовину, что означает неэффективное использование площади ЦОДа. На Рис. 69 показано, что происходит при использовании оборудования side-to-side и неправильной установкой оборудования с охлаждением front-to-back / back-to-front.



Рис. 69 – Неправильная организация охлаждения в стойке

Рекомендуется проектировать архитектуру SAN так, чтобы в ней не было единой точки отказа, особенно если строится резервированная сеть (стр. 303). Смысл резервирования фабрик – это устранение единых точек отказа, но если две фабрики будут установлены в одной стойке, то сама стойка будет единой точкой отказа. Лучше всего поместить отказоустойчивые центральные коммутаторы и резервированные фабрики в разных стойках и так организовать питание стоек, что сбой в электросети не мог бы нарушить работу обеих фабрик (см. далее). В идеале, резервированные компоненты должны отстоять по крайней мере на одну стойку, а резервированные фабрики должны стоять в разных комнатах. Степень изоляции фабрик зависит от

их размера, сложности подведения кабелей в разные зоны, ущерба от одновременного выхода из строя обеих фабрик и т.п. Например, надо сбалансировать риск одновременного сбоя, который может быть вызван срабатыванием системы пожаротушения этажом выше, и расходами на распределение компонентов по разным площадкам.

Таким образом, архитектор SAN должен выбрать стойки для монтажа оборудования, которые:

- По своим физическим характеристикам соответствуют оборудованию
- Поддерживают архитектуру НА для SAN
- Обеспечивают необходимое воздушное охлаждение оборудования

## Питание и ИБП

При проектировании SAN высокой доступности нужно учитывать и то, как эта сеть будет внедрена. Для НА SAN у ЦОДа должны быть подключения к разным источникам питания и резервированные компоненты подключены к разным ИБП и электросетям. На Рис. 70 - 73 показана неудачная архитектура питания и рекомендации по исправлению ситуации.

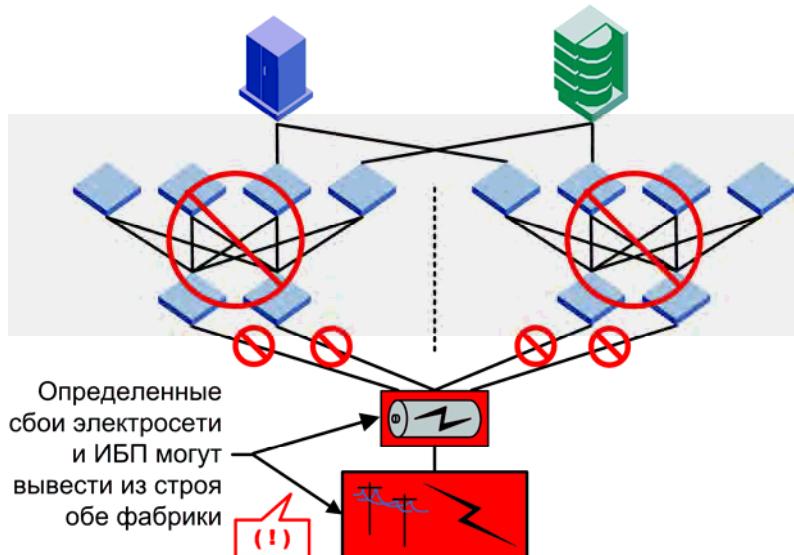


Рис. 70 – Самая неудачная организация питания

Любой сбой в такой системе выведет из строя обе фабрики.

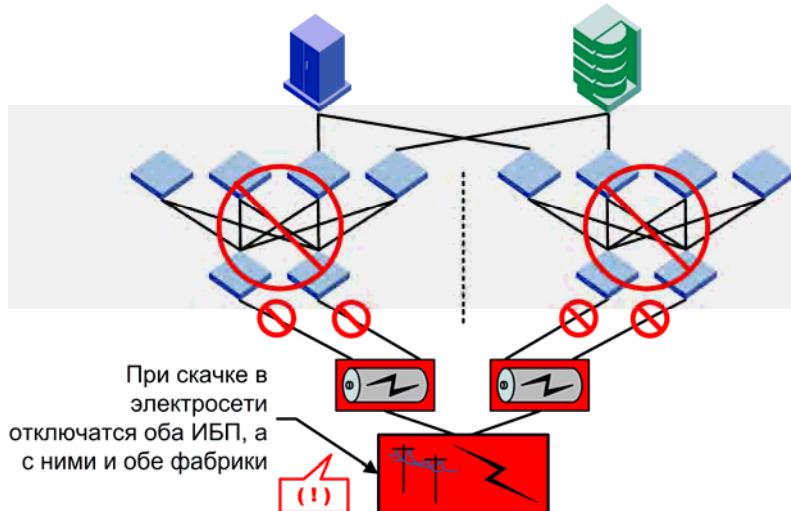


Рис. 71 – Лучше, но все равно неудачная организация питания.

Такая конфигурация намного надежнее, но

достаточно плохой скачок напряжения в электросети выведет из строя оба ИБП, а с ними и обе фабрики.

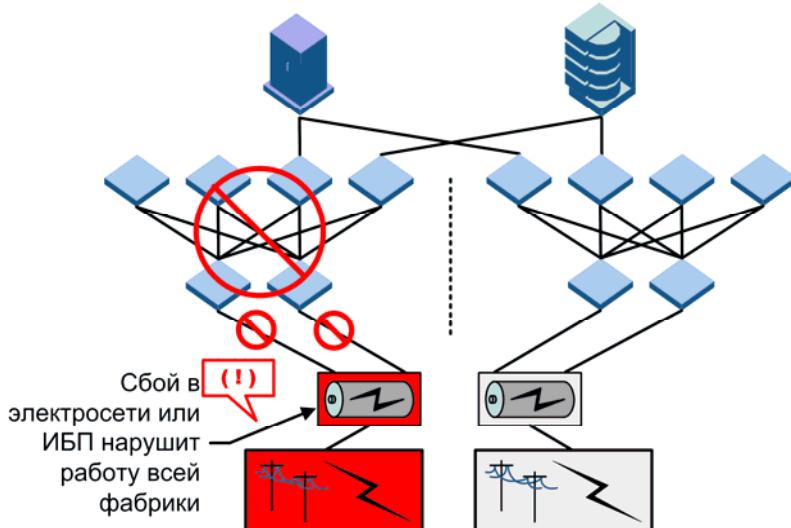


Рис. 72 – Приемлемая организация питания

Две фабрики не могут выйти из строя одновременно, но при сбое одной потребуется переключение всех путей передачи данных.

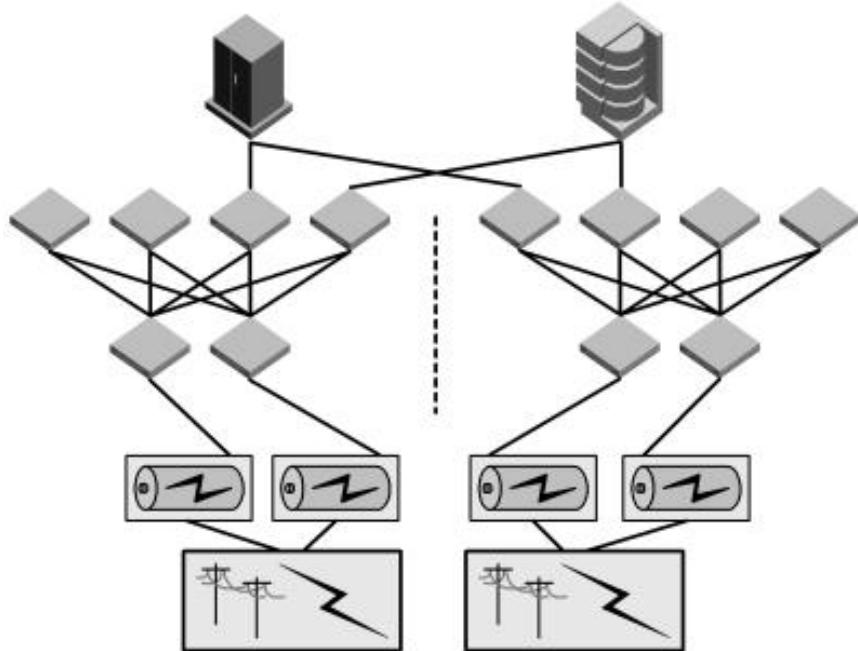


Рис. 73 – Рекомендуемая организация питания

Оптимальная архитектура организации питания похожа на архитектуру сетей НА, т.е. использует резервирование и отказоустойчивость. У фабрик “A/B” нет ни общих ИБП, ни электросетей, поэтому броски в электросети (например, из-за удара молнии) в худшем случае выведут из строя только одну фабрику (разумеется, если электросети действительно независимые, а не подключены к одному источнику).

Помимо резервирования питания архитектор SAN должен учесть и общее энергопотребление. Если заказчик интересуется проектированием в соответствии с лучшими «зелеными» практиками, то не желательно потреблять больше энергии, чем абсолютно необходимо. Такой подход оправдан и с экономической точки зрения – в течение жизненного цикла SAN расходы на электричество могут превысить первоначальные затраты

на оборудование и программное обеспечение, поэтому увеличение затрат на оборудование на 10% быстро окупится, если оно позволит сократить расходы на электричество на 50%.

Однако, экономия энергии не сводится к выбору "мощность или затраты на электричество" либо "затраты или защита окружающей среды". Во многих ЦОДах остро не хватает электроэнергии и питания, доступного для одной стойки. Если эффективно спроектировать энергопотребление SAN, то вместо строительства нового ЦОДа необходимое оборудование можно разместить в существующем ЦОДе.

В любом случае, многие архитекторы SAN рассматривают энергопотребление как один из главных критериев при выборе коммутаторов и маршрутизаторов SAN. Для небольшой SAN, где меньше 100 портов, имеет смысл использовать коммутаторы Brocade 4900 или 5000. Если в SANs более 100 портов на фабрику или ожидается, что она выйдет на этот уровень в следующие два года, самым эффективным является архитектура на основе директоров Brocade 48000.

## **Выбор кабелей и оптических модулей, управление ими**

В этом разделе обсуждаются некоторые вопросы внедрения SAN, которым часто не уделяется должное внимание, в том числе выбор кабелей и оптических модулей (медиа, тринсиверов SFP, GBIC), принципы управления кабелями.

### ***Оптические кабели и модули***

Большинство сетевых устройств хранения для связи на большие расстояния и высокой доступности используют оптические соединения вместо медных.

Важно подчеркнуть, что нельзя произвольно смешивать коммутаторы, маршрутизаторы, узлы, оптические модули и кабели. Например:

- Нельзя использовать GBIC в устройствах, рассчитанных на SFP. У них разный форм-фактор, поэтому GBIC просто не может встать в сокет SFP. Архитектор SAN должен проверить, что заказанные им оптические модули соответствует устройствам.
- Если установлена фабрика MMF для оптических кабелей, то соединения устройства, предназначенных для кабелей SMF, не будут работать корректно. Например, соединение устройства 10Gbit к кабельной фабрике ЦОДа, рассчитанной на скорости 1Gbit, 2Gbit и 4Gbit скорей всего не сможет работать. Нужно убедиться, что заказанные патч-кабели соответствуют остальной инфраструктуре.
- Кабели SC нельзя подсоединить к SFP без адаптеров. Аналогичным образом, кабели LC нельзя напрямую подключить к GBIC. Если патч-панель предназначена для одного формата, а оптические модули для другого, то нужно заказать правильные патч-кабели.
- Коммутаторы, маршрутизаторы и узлы поддерживают только определенные модели оптических модулей. Чтобы реализовать поддержку определенного оптического модуля<sup>87</sup> требуется разработка программного обеспечения и его тестирование. Прежде чем заказывать SFP или GBIC нужно запросить у поставщика поддержки матрицу поддержки оптических модулей.

---

<sup>87</sup>

Оптические модули (Трансиверы) преобразуют сигналы, передаваемые по медному проводу от материнской платы коммутатора, маршрутизатора или НВА в световые сигналы, которые идут по оптоволоконному кабелю.

- Оптические модули должны использовать правильный тип оптического кабеля. Если SFP рассчитан на работу с одномодовым кабелем, то он не сможет правильно работать с многомодовым.

Архитекторы SAN не выбирают тип оптических модулей (SFP или GBIC) и затем проектируют SAN. Тип модулей в общем случае определяется архитектурой сети. Однако, архитектор должен выяснить, с какими оптическими модулями будет поставляться заказанное оборудование, иначе могут возникнуть проблемы совместимости при инсталляции, а также проблемы с поддержкой расстояний и заказом правильного типа кабеля.

Таким образом, до оформления заказа на оборудование архитектор SAN должен убедиться, что заказываются все нужные оптические модули и кабели, иначе систему просто нельзя будет установить. Полезно знать определенные факты о кабелях и модулях, которые влияют на максимальное расстояние. Далее мы обсудим эти вопросы.

### Оптические модули SWL

Чаще всего используется коротковолновые оптические модули Short Wavelength Laser (SW L). Они могут быть SFP или GBIC. Это самые недорогие оптические модули и они используются для соединений коммутаторов, маршрутизаторов и узлов Gigabit Ethernet или Fibre Channel на относительно небольшие расстояния. Обычно они используются в пределах ЦОДа, между ЦОДами в одном кампусе и на больших расстояниях при использовании таких активных устройств увеличения расстояния, как например, повторители или DWDM. SW L-модули обычно используется вместе с кабелями MMF. Архитекторы

обычно используют SW L SFP или GBIC если им по каким-то причинам не нужен другой тип модулей.

### Оптические модули LWL и ELWL

Длинноволновые оптические модули Long Wavelength Laser (LWL) и Extended Long Wavelength Laser (ELWL) используется для передачи native FC через темную оптику или xWDM на больших расстояний. Как и в случае SWL, они могут использовать SFP или GBIC в устройствах Fibre Channel или Gigabit Ethernet. Как и следует из названия, они работают на более длинных волнах, чем SWL модули и поэтому без повторителей могут поддерживать связь на расстоянии до 100 км и даже больше. Для оптических модулей LWL и ELWL требуются кабели SMF. Архитекторы должны вместе с вендорами выбрать и заказать эти компоненты при удлинении FC за пределы расстояний, поддерживаемых модулями SWL. Крайне важно выяснить, полностью ли поддерживаются оптические модули. Из-за относительно небольших объемов есть мало опций LWL и ELWL, и они могут быть сертифицированы OEM - поставщиками и поставщиками поддержки, используя процессы, отличные от используемых при сертификации оптических модулей SWL.

### Кабели MMF и оптические модули

Многомодовое волокно Multi-Mode Fiber (MMF) используется для коротких расстояний. Оно значительно дешевле одномодового SMF и чаще всего используется в кабелях внутри ЦОДа или кампуса. Существует два типа кабелей, отличающиеся диаметром<sup>88</sup> - 50/125  $\mu\text{m}$  и 62.5/125 $\mu\text{m}$ . Коммутаторы и маршрутизаторы Brocade

---

<sup>88</sup>

Указываются два диаметра - первое чило относится к внутреннему диаметру, второе – к диаметру с оплеткой. Для лазера важно только первое число, поэтому 50/125 $\mu\text{m}$  MMF иногда сокращают до 50 $\mu\text{m}$ .

поддерживают оба типа. Кабели MMF обычно используются вместе с SW L GB IC и SFP . Если все другие параметры идентичны то большие расстояния поддерживаются кабелями 50/125  $\mu\text{m}$ . Кроме того, недавний стандарт “ОМ3” будет использоваться в больших ЦОДах.

Очень важно правильно выбрать тип (50/125  $\mu\text{m}$  или 62.5/125 $\mu\text{m}$ ) при развертывании решений MMF. Нужно убедиться, что оптический модуль предназначен для работы с нужным диаметром кабеля. Также рекомендуется использовать один и тот же тип кабеля по всему соединению, поскольку при использовании разных типов ослабляется сигнал, что может создать проблемы надежности и даже сбой линка.

### Кабели SMF и оптические модули

Одномодовое волокно Single-Mode Fiber может использоваться для коротких и больших расстояний, но из-за высокой стоимости используется только для линков на большие расстояния в комбинации с решениями LW L, ELW L и xWDM. Диаметр жил у кабелей SMF самый маленький – всего 9/125 $\mu\text{m}$ .

### **Управление кабелями**

Отсутствие продуманного управления кабелями создает серьезные проблемы для ИТ-отдела, например, если не выполняется требования к изгибу кабеля, то из-за этого может уменьшиться поддерживаемая длина линка. Управление кабелями сокращает расходы на обслуживание SAN и помогает продлить срок эксплуатации кабелей и коннекторов, а также упрощает идентификацию кабелей, что крайне важно для быстрого устранения сбоев и расширения SAN.

Строгие требования к непрерывности работы SAN не позволяют переключать кабели после ввода в

эксплуатацию сети хранения. Для предотвращения дополнительных расходов и простоев необходимо внедрить надежную стратегию управления кабелями сразу после запуска в эксплуатацию SAN, а для этого уже на этапе планирования подготовить детальные схемы и документацию. Внедрение продуманной стратегии управления кабелями дает следующие преимущества для SAN:

- Уменьшение сбоев из-за ошибок оператора
- Улучшение общей поддержки
- Улучшение диагностики сбоев
- Улучшение масшабируемости
- Быстрое устранение проблем

Эффективное управление кабелями не только упрощает обслуживание, но и улучшает внешний вид стоек с оборудованием. Аккуратно уложенные кабели обычно легче обслуживать. Например, кабели для ISL нужно аккуратно снабдить метками и расположить так, чтобы их нельзя было перепутать с кабелями хостов и устройств хранения данных. В некоторых ЦОДах для ISL используются кабели определенного цвета.

Когда архитектор SAN выбирает решение для управления кабелями в стойке, то необходимо учитывать, что требуется дополнительное место для того, чтобы можно было изогнуть оптоволоконный кабель, подключаемый к портам коммутатора. Если не соблюдать требования к минимальному изгибу оптоволоконного кабеля, который изготавливается из дорогой оптики, то это приведет к ослаблению сигнала и в результате могут быть потеряны некоторые пакеты и произойдет порча данных. Даже если не произойдет потеря пакетов, то слишком натянутый изгиб оптоволоконного кабеля вызовет сокращение поддерживаемого расстояния из-за ослабления сигнала.

Если кабель закрывает переднюю панель коммутатора, то усложнится процедура замены компонент (Field Rep laceable Un its, FRU). Правильное управление кабелями помогает избежать такой ситуации. Оно также упрощает идентификацию кабелей по меткам. Использование направляющих для кабелей устраниет эту проблему.

Единственный способ надежно подключить много кабелей к стойке, проложить их на площадке – это использовать соответствующие продукты для управления кабелями. Горизонтальные направляющие можно использовать для аккуратной прокладки кабеля в стойках, а вертикальные направляющие – для прокладки кабелей с боковой стороны стоек и между стойками.

Всегда лучше проектировать управление кабелями с запасом. Например, если требуется направляющая шириной два дюйма с одной стороны стойки, то лучше установить две четырехдюймовые направляющие с обеих сторон стойки. Хотя это ведет к дополнительным расходам, однако они меньше убытков, которые возникнут, если техник по ошибке отключит кабель в самый неподходящий момент.

## Настройка коммутаторов

До того, как соединять коммутаторы кабелями нужно настроить определенные параметры, в том числе IP-адреса для управления и имя коммутатора, которое должно совпадать с именем хоста, которое отображается на IP-адрес коммутатора. Описание процедуры настройки этих параметров можно найти в руководстве по эксплуатации коммутаторов, но мы дадим несколько советов по ее выполнению на практике.

Архитектуре SAN следует давать коммутаторам понятные имена, по которым будет легко

идентифицировать физический коммутатор при возникновении проблемы, например, где он находится логически в SAN (например, “Центр Фабрики А” или “Периферия Фабрики В”) и физически (если коммутаторы размещены в нескольких ЦОДах).

Также архитектор должен выбрать правильную версию Fabric OS до внедрения SAN (это необязательно будет последний релиз). Выбор Fabric OS может потребовать консультаций с несколькими ОЕ М-производителями устройств SAN и/или поставщиками поддержки поскольку уровни поддержки Fabric OS у разных вендоров могут отличаться.

У каждого коммутатора могут быть различные дополнительные функции, на которые нужно приобретать отдельные лицензии и надо заранее приобрести все нужные лицензии до начала инсталляции оборудования.

Также следует продумать назначение ID доменов. При разработке коммутаторов Brocade особое внимание уделяет удобству использования, и обычно эти и другие их параметры конфигурируются автоматически. Однако, в некоторых случаях, для упрощения управления нужно выставлять ID доменов вручную – например, центральным коммутаторам назначать маленькие ID, а для периферийных коммутаторов – более высокие и идущие подряд ID. В результате администратор сможет интуитивно определить место коммутатора в фабрике по ID домена и с помощью этого параметра провести анализ проблемы при диагностике сбоев. Также вручную следует устранять совпадение ID доменов разных фабрик, если эти фабрики в будущем могут быть объединены, и назначать одинаковые ID вручную для двух фабрик, которые сконфигурированы как пары “A/B”. ( Это также делает управление интуитивно-понятным, поскольку устройства будут иметь

одинаковые PID в обеих фабриках.)

## **Поэтапное внедрение и тестирование**

До ввода фабрики в промышленную эксплуатацию необходимо проверить, что SAN готова к работе. Для этого лучше всего построить тестовый стенд, на котором будет проверяться соответствие показателей производительности и доступности заданным критериям.

На тестовом стенде нужно смоделировать отказы и убедиться, что сама фабрики и периферийные устройства могут восстанавливаться после сбоев. На следующем этапе в SAN генерируется трафик ввода/вывода, примерно соответствующий реальному профилю трафика. Наконец, тестируется работа в самой неблагоприятной ситуации - по SAN идет объем трафика, близкий к реальному, и моделируются различные сбои.

При крупных инсталляциях архитектор обычно не отвечает за выполнение этих процессов, однако он может объяснить команде внедренцев, как смоделировать трафик ввода/вывода фабрики и какие сбои могут возникнуть.

## **Запуск в эксплуатацию**

Имеются два типа запуска в эксплуатацию – «с нуля», если в компании раньше не было SAN, и модернизация установленного оборудования, когда к существующей фабрике добавляются новые коммутаторы, маршрутизаторы и другое оборудование.

### **Запуск в эксплуатацию с нуля**

Это более простой сценарий ввода в эксплуатацию, поскольку отсутствует риск нарушения работы

продукционных приложений. Обычно его выполнение состоит из следующих шагов:

1. Разработать детальную схему и документацию SAN в соответствии с методикой из Главы 4 “4: Обзор проектирования SAN” (стр. 121 и далее).
2. Подготовить общий план ввода в эксплуатацию с описанием перечисленных ниже этапов и сроков их выполнения, а также бюджета. Если в вводе в эксплуатацию участвует третья сторона, то нужно учесть и профессиональные сервисы, а также бюджет на поддержку для всего оборудования и программного обеспечения.
3. Поиск и приобретение оборудования и программного обеспечения. Этот шаг может выполняться сразу или этапами.
4. Начало развертывания и тестирования SAN. Сначала строится инфраструктура центра, т.е. развертываются коммутаторы и маршрутизаторы, чтобы при вводе в эксплуатацию не пришлось менять общую инфраструктуру центра. Для тестирования ISL и IFL лучше использовать как локальные, так и удаленные соединения.
5. После того, как структура сети полностью построена и работает стабильно, можно начать промышленную эксплуатацию, не дожидаясь завершения инсталляции. Например, если проект сети предусматривает 60 периферийных фабрик, но построены только 30 фабрик, то можно начать эксплуатацию сети и остальные фабрики добавлять по мере необходимости.

### ***Модернизация инсталлированного оборудования***

Для добавления нового оборудования в существующую инфраструктуру может понадобиться только его подключение и включение, но эта процедура

может быть и очень сложной и потребовать построения специальной тестовой среды и выполнения многомесячных стрессовых тестов до запуска в промышленную эксплуатацию. Она зависит от применяемых методов контроля изменений, от того, насколько сетевой администратор уверен в новом оборудовании и потенциальных убытков компании при сбое во время ввода в эксплуатацию.

Если у заказчика среда жестко контролируется, то ввод в эксплуатацию может происходить по тому же сценарию, что и при вводе в эксплуатацию с нуля – новое решение строится в изолированной среде и производственные приложения переносятся на него только после того, как тестами будет доказана его стабильность.

У заказчиков с более мягким контролем, но все равно использующих процедуры контроля изменений, обычно уже построены фабрики А/В. В таком случае новые устройства нужно сначала установить только в одной из резервированных фабрик и выполнить с ним набор тестов и только потом перевести в промышленную эксплуатацию. После подключения устройств к фабрике “А” и проверки стабильности их работы нужно выполнять аналогичные подключения к фабрике “В”. Такой процесс в комбинации с подходом с резервированием/отказоустойчивостью “А/В” (стр. 303) гарантирует безопасность и простоту ввода в эксплуатацию.

Наконец, у некоторых заказчиков есть несколько фабрик с различными уровнями бизнес-критичности. В таком случае, новые устройства лучше сначала подключить к фабрикам самого нижнего уровня критичности и тогда потенциальные проблемы могут нарушить работу только некритичных приложений.

## Повседневное управление

Архитектор сети хранения при ее проектировании может упростить повседневное управление. Например, если у сети будет резервированная и отказоустойчивая архитектура, то упростится переход на новые технологии в будущем. Кроме того, архитектор должен предусмотреть организацию хранилища документов как части процесса проектирования и внедрения. Если конфигурация сети будет хорошо задокументирована и процедуры обслуживания заранее разработаны, то управление не потребует больших усилий.

Архитектор должен создать журнал конфигурации, в котором будет зафиксировано, как сконфигурирована SAN с объяснениями выбора конфигурации. В журнал будут заноситься все изменения и если после какого-то изменения возникнет нештатная ситуация, то администратор SAN по этим записям быстро определит, что изменилось в сети. Архитектор должен обеспечить процедуры, которые нужны, чтобы администраторы вносили в журнал любые изменения существующей конфигурации при добавлении или удалении коммутаторов.

Brocade предлагает инструменты, упрощающие создание и ведение такого журнала, например, у Fabric Manager есть функции для управления изменениями, а SAN Heath можно сконфигурировать на автоматический периодический аудит среды.

Эти инструменты упрощают и другие текущие операции, например, резервное копирование конфигураций коммутаторов и фабрики, например, ID доменов и базы данных зон. Fabric Manager также хорошо подходит для автоматизации таких операций как обновление микрокода и его активизация.

## Планирование устранения сбоев и неисправностей

Как и в других сетях, в SAN периодически могут возникать сбои и для их устранения нужно найти то оборудование, программное обеспечение или конфигурацию, которые вызвали сбой. Сам архитектор SAN редко занимается устранением сбоев, но может упростить этот процесс.

Например, обычно легче устранить сбой в маленькой сети, чем в большой. Если архитектор будет применять архитектуру Meta SAN, в которой каждая фабрика является небольшой подсетью, то сбои будут ограничены отдельными фабриками и их будет легче локализовать.

Локализация сильно упрощает устранение сбоев – если нет связи между портами хоста и устройства хранения, то причину этого сбоя легче выяснить когда трафик идет только между несколькими линками. Для бизнес-критичных приложений часто можно локализовать основной трафик ввода/вывода в одной микросхеме ASIC. Тогда трафик в директре не будет выходит с лезвия даже на backplane, что упростит локализацию проблемы.

Внутри каждой фабрики можно использовать простой или сложный дизайн и при прочих равных условиях простые решения работают лучше. Это означает применение архитектуры на основе директоров с сотнями портов и сетей СЕ для больших конфигураций. Часто при проектировании СЕ директоры применяются и в центре, и на периферии для максимальной масштабируемости и сокращения до минимума числа устройств в SAN, в которых может возникнуть сбой.

Затем архитектор может участвовать в выборе инструментов, которые будет использовать администратор. В каждом продукте есть встроенные утилиты и первый шаг (который часто оказывается и последним) при устранении сбоев – это запустить такую утилиту.<sup>89</sup> Таким полезным инструментом являются Brocade SAN Health и интерфейс Web Tools. Команда *portLogDump CLI* поможет устраниить проблемы на уровне FC. (хотя такие проблемы возникают редко, но в любом случае «запас карман не тянет».)

В некоторых случаях *может быть желательно* использовать анализатор протокола Fibre Channel. Продукты Brocade доказали свою надежность на практике использования в течение многих лет, поэтому крайне редко возникает проблема, для разрешения которой нужен анализ на уровне пакетов FC. Подобные проблемы редко возникали даже при использовании продуктов Brocade первого поколения, но с ними сталкиваются пользователи, использующие продукты вендоров, которые недавно вышли на рынок Fibre Channel и еще не успели полностью отладить свою продукцию. Большинство проблем с протоколом на практике возникают в экспериментальных (не выпускаемых серийно) устройствах и их диагностировать можно с помощью команды *portLog-Dump*. Анализатор FC за пределами лаборатории обычно требуется только когда сервисные инженеры выполняют на площадке заказчика отладку трудно устранимой проблемы, с которой не удается справиться с помощью других инструментов отладки. Но если у заказчика инсталлировано несколько тысяч портов FC, то такой анализатор лучше иметь на всякий случай.

---

<sup>89</sup>

Это первый шаг *после* того, как все подключено, включено питание и правильно соединены кабели. К сожалению, эти три условия часто не выполняются.

---



## **Третья часть**

### **Дополнительные материалы**

#### **Темы**

- Базовые материалы
  - Расширенные материалы
  - Тест
  - Часто задаваемые вопросы
  - Словарь
- 



## Приложение А:

## Базовые материалы

This chapter provides reference material for readers who may be less familiar with either Fibre Channel or IP/Ethernet technology, or advanced readers who just occasionally need to look up certain details. Topics covered include an overview of some of the more notable items in the Brocade hardware and software product lines, and some of the external devices that might be connected to SAN infrastructure equipment.

### **Поставляемые платформы Brocade<sup>90</sup>**

Brocade offers a full range of SAN infrastructure equipment, including switches and routers ranging from entry-level 8-port platforms up to 384-port enterprise-class fully-modular directors. The networking capabilities of the platforms allow solutions with up to about 10,000 ports in a single network today, with the potential to scale much higher in the future.<sup>91</sup> Brocade currently offers products with Fibre Channel, FICON, iSCSI, and FCIP. The Brocade Fabric Application Platforms deliver switching at all levels of the protocol stack up to and including the application layer.

---

<sup>90</sup> Shipping to OEMs for sale as of the date of first printing of this edition of this book. Check with the appropriate sales channel for product availability.

<sup>91</sup> Very large solutions generally require FC-FC routers as well as switches.

All currently shipping FC fabric switch platforms run a version of Brocade Fabric OS 5.x or higher.<sup>92</sup> The use of a common code base enables compatibility between switches and nodes, and consistent management between platforms. It also allows a common set of value-added software features. (See “Лицензируемые функции Brocade” on p444.)

### **FC коммутатор Brocade 200E**

The Brocade 200E (below) is the entry point into the Brocade FC product portfolio.



**Figure 74 - Brocade 200E**

This platform provides enterprise-class features, performance, and scalability, at an affordable price point for the entry market. Features include:

- Sixteen 4Gbit<sup>93</sup> non-blocking / uncongested interfaces to support the most performance-intensive applications: enterprise-class performance at an affordable price. It is the highest-performing 8-to-16-port SAN switch in the industry.

---

<sup>92</sup> Products brought into the Brocade family from the recent acquisition of McDATA are an exception to this rule. Brocade intends to converge these into a common director platform running Fabric OS in the future. Former McDATA customers are encouraged to discuss any concerns they may have regarding the roadmap with their local Brocade sales team.

<sup>93</sup> See also “4Gbit FC” on p525.

- Investment protection for existing SAN infrastructure to reduce deployment cost and complexity. This means forward and backward compatibility with other Brocade switches, routers, and directors at 1Gbit, 2Gbit, and 4Gbit.<sup>94</sup>
- Enterprise-class features and high-availability characteristics such as hot-swappable FRUs and hot code load and activation. The switch is ideal for mission-critical SAN environments too small or cost-sensitive to allow director deployments.
- Ports demand and via optional software license keys allows the switch to be used in configurations starting at stand-alone 16-port solutions, but it can also be used as a core in small to medium CE fabrics, and as an edge in medium to large solutions.

The Brocade 200E was intended to replace the Brocade 3250 and 3850 (p. 436). In many respects, these switches are similar. All have hot fixed fans and power supply(s). All support hot code load and activation. All are compatible with Fabric OS 5.0.1 and later. All three use SFP media.

However, the Brocade 200E also improves on the older switches in many ways. For example, the 200E uses more modern and highly integrated technology, resulting in a more reliable switch and lower power consumption.

---

<sup>94</sup> It is never possible for a technology company to perform regression testing for all firmware released on new products in all combinations with all firmware releases on all old products. This would result in a virtually infinite number of tests needing to be passed before any new products could be qualified for shipping. Since this is impractical, Brocade will periodically end support for very old platforms. For example, the SilkWorm 1000 series (which has not been shipping this century) has never been supported in combination with the Brocade 48000. Customers running products which have been at end of life for multiple years should explicitly check for compatibility before using them with newer platforms, and should consider upgrading in any case.

Most notably, the 200E is the first entry platform to use the forth-generation 4Gbit “Goldeneye” ASIC. (p502) In addition to the Brocade Fabric OS 5.x features available on other platforms, the Brocade 200E enables the next-generation features of the Goldeneye ASIC, including but not limited to:

- 4Gbit Fibre Channel interfaces
- Each port is an autosensing U\_Port interface, supporting F\_Port, FL\_Port, and E\_Port
- Auto-negotiates 4Gbit on ISLs and Trunks with other Goldeneye & Condor based switches.
- Capable of running all ports at 4Gbit line rate simultaneously. That is 128Gbits of cross-sectional bandwidth per switch.
- 4-way frame-based trunking, and DPS.
- Cut-through routing to minimize latency.
- Centralized pool of 288 buffer-to-buffer credits.
- Hardware offload support for node login. This improves control-plane scalability.
- Centralized hardware zone tables allow more flexible deployment scenarios. Up to 256 hardware zones are supported per ASIC.
- 8 VCs per E\_Port to support non-blocking (HoLB) operations in larger networks. This can be used for advanced QoS features in the future.

### ***Коммутатор Brocade 4100***

The Brocade 4100 Switch, shown in Figure 75, provides enterprise-class features, performance, and scalability.



**Figure 75 - Brocade 4100**

Some of its features include:

- 4Gbit non-blocking / uncongested interfaces to support the most performance-intensive applications, yielding enterprise-class performance at a midrange price. It is the highest-performing 16-to-32-port SAN switch in the industry.
- Investment protection for existing SAN infrastructure to reduce deployment cost and complexity. This means forward and backward compatibility with other Brocade switches, routers, and directors. All ports can operate at 1Gbit and 2Gbit, as well as 4Gbit, and that Fabric Services behaviors are consistent.
- Enterprise-class features and high-availability characteristics such as hot-swappable FRUs and hot code load and activation. The switch is ideal for mission-critical SAN environments too small or cost-sensitive to allow director deployments.
- Ports on demand via optional software license keys allows the switch to be used in configurations starting at stand-alone 16-port solutions, but it can also be used as a core in small to medium CE fabrics, and as an edge in medium to large solutions.

The Brocade 4100 replaced the Brocade 3900 (p436) in late 2004. In many respects, the two switches are very similar. Both provide up to 32 ports in high-density fixed configuration. Both have hot swappable fans and power supplies. Both support hot code load and activation. Both run Fabric OS. Both use SFP media.

However, the Brocade 4100 also improved on the 3900 in many ways. For example, the Brocade 3900 was 50% larger than the 4100, so the new platform supports higher density rack configurations. There were corner cases in which the Brocade 3900 could exhibit internal

traffic configuration.<sup>95</sup> (See “SilkWorm 12000 и 3900 “XY”” on p 513 for more information about the 3900 internal architecture.) The 4100 uses more modern and highly integrated technology, resulting in a more reliable switch and lower power consumption.

Most notably, the Brocade 4100 is the first platform to use the forth-generation 4Gbit “Condor” ASIC. (See “Condor” on p 506.) In addition to the Brocade Fabric OS features available on other platforms, the Brocade 4100 enables the next-generation features of the Condor ASIC, including but not limited to:

- 4Gbit Fibre Channel interfaces
- Each port is an autosensing U\_Port interface, supporting F\_Port, FL\_Port, and E\_Port
- Auto-negotiates 4Gbit on ISLs with 4Gbit switches.<sup>96</sup>
- Capable of running all ports at 4Gbit line rate simultaneously. That is 256Gbits of cross-sectional bandwidth per chip.
- 8-way frame-based trunking, and DPS.
- Cut-through routing to minimize latency
- Centralized pool of 1024 buffer-to-buffer credits
- Up to 255 buffers allocated to any given port
- Native FC connectivity up to 500 km

---

<sup>95</sup> Note that traffic patterns consisting of large percentages (e.g. 90%) of small (e.g. 64-byte) frames will have lower *throughput*. This is not caused by *congestion*. It is because the ratio of frame header and inter-frame gap to payload is less favorable with small frames. All networking technologies behave this way to some extent if they support variable frame sizes. Fortunately, there are no known bandwidth-sensitive applications that produce large percentages of small frames on all ports in a network simultaneously, which is the only scenario in which the switch would exhibit degraded performance. Typical SAN traffic patterns lean *much* more heavily towards 2k frames than towards 64-byte frames, and the average frame size is very close to 2k.

<sup>96</sup> At the time of this writing, there are few generally available 4Gbit nodes. The intent is for F\_Ports also to auto-negotiate as the 4Gbit node market develops in much the same way that 1Gbit/2Gbit is auto-negotiated today.

- Hardware offload support for node login. This improves control-plane scalability.
- Centralized hardware zone tables allow more flexible deployment scenarios. Up to 256 hardware zones are supported per ASIC.
- 16 VCs per E\_Port to support non-blocking (HoLB) operations in larger networks. This can be used for advanced QoS features in the future.

### ***Коммутатор Brocade 5000***

The Brocade 5000 fabric switch is shown in Figure 76. This platform provides enterprise-class features, performance, and scalability, delivering high value at an affordable price point. This product functionally replaces the Brocade 4100, and entirely replaces the M4700.

In many respects, these switches are very similar. All provide up to 32 ports in a high-density fixed configuration. All three have hot swappable fans and power supplies. Each can support hot code load and activation. Both the 4100 and 5000 run Fabric OS. Both use SFP media. One minor difference is that the 5000 has a combined FAN/Power Supply FRU, whereas the 4100 had separate FRUs for each of those parts. Since this has no impact whatsoever on availability, this is considered an academic difference.



**Figure 76 - Brocade 5000**

However, the 5000 is not simply a replacement for the 4100; it also improves on the 4100 in many ways. For example, the 4100 was twice as deep as the 5000. Because

of the shallower “rack footprint”, it is possible to mount the 5000 without a rail kit. In some configurations, the 5000 supports higher density rack configurations in that it can be mounted back to back in a cabinet, provided that the overall airflow is appropriate. That is, it can be mounted on the direct opposite side of a cabinet vs. other equipment, or even behind another Brocade 5000. The 5000 uses more modern and highly integrated technology, resulting in a more reliable switch and lower power consumption: it is about 20% more efficient than the 4100.

From a software feature set viewpoint, the 5000 is identical to the 4100 with the exception that, at the time of this writing, the 4100 does not have a near-term roadmap to support native interoperability with McDATA fabrics whereas the Brocade 5000 does have this.

### ***Коммутатор Brocade 4900***

The Brocade 4900 fabric switch is shown in Figure 77. This platform is essentially identical to the Brocade 4100 (p400) and 5000 in terms of features supported. The difference is that it has twice as many ports, and takes up 2u instead of 1u. (I.e. the port density is identical.) The ports on demand feature ranges from 32 to 48 to 64 ports. Like the Brocade 4100, the Brocade 4900 has sufficient internal bandwidth to support all ports at full-speed / full-duplex operation simultaneously in all traffic configurations. (I.e. is fully non-blocking and uncongested.)



**Figure 77 - Brocade 4900**

## Директор Brocade 48000

The Brocade 48000 (below) is a fully-modular 10-slot enterprise-class director, and can be populated with up to eight port-blades and two Control Processors (CPs).



**Figure 78 - Brocade 48000 Director**

This platform first shipped in mid 2005. It can be configured from 32 to 384 ports in a single domain using 16-, 32-, and 48-port 4Gbit FC blades. Using the “Virtual Fabrics” feature, it can be carved up into multiple virtual chassis. The platform has industry-leading performance and high availability characteristics. Each blade is hot-pluggable, as are the fans, WWN card, and power supplies. The chassis has redundant control processors (CPs) with redundant active-active uncongested and non-blocking switching elements, which run Fabric OS 5.0.1 or higher and support HCL/A. To support 48-port blades,

Fabric OS 5.2.0 or higher is required and some advanced function blades may require higher OS releases.

The Brocade 48000 is an evolution of the Brocade 12000 and 24000 design. The blades can even use the same chassis as its predecessors in some cases: the power supplies, fans, backplane, and sheet metal enclosure are generally compatible. As a result, it is possible to upgrade an existing 12000 chassis all the way to the 48000 in the field by replacing just the CP and port blades.<sup>97</sup> Similar procedures can work with the 12000 to 24000, or 24000 to 48000. Look between Figure 78 and Figure 105 (p439) and the similarity will be apparent.

There are also differences between the directors. Some of the differences are minor. For example, the 24000 and 48000 chassis and blade set has an improved rail glide system that makes blade insertion / extraction easier compared to the 12000. Larger ejector levers help by providing greater mechanical advantage. The 48000 also has a redesigned cable management system to accommodate using the larger number of ports.

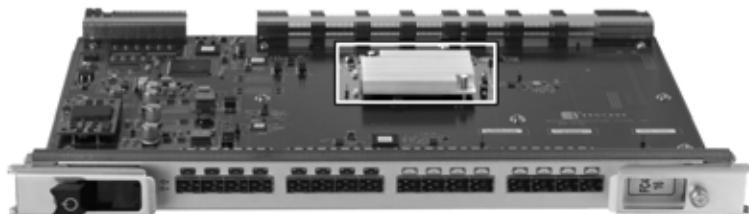
There are also much more important differences in the underlying technology. For example, the 24000 uses the 2Gbit “Bloom-II” ASIC, while the 48000 uses the 4Gbit “Condor” chipset. (See “Bloom и Bloom-II” p 505 and “Condor” on p 506.) The overall chassis power consumption and cooling requirements have been lowered drastically, with the result that ongoing operational costs

---

<sup>97</sup> As a practical matter, this is almost never done. It's virtually always easier, less risky, and even less expensive to deploy a new director vs. upgrading an existing chassis. Also, not all OEMs can support upgrading chassis for administrative reasons. For example, it may be that the chassis serial number is used to define the support contract for a platform, and it may not be administratively practical to change it from a 12000 to a 48000 in the support system, even if it is technologically possible from a hardware and software viewpoint. The bottom line is that field upgrades are rarely performed.

are reduced and MTBF is increased substantially as well. Further improvements in MTBF are achieved through component integration: fewer components means less frequent failures, and the Condor chipset is the most tightly integrated in the industry. Performance has been improved from the 12000 by changing the multistage chip layout from an “XY” topology to a “CE” arrangement. (See “**Многоуровневые внутренние архитектуры**” on page 511 for more information.) This allows the 48000 to present all of its ports in a single fully-internally-connected domain. The 12000, in contrast, presented two 64-port domains and required external ISLs if traffic was required to flow between the domains. In addition, the 48000 runs its internal links faster than the 12000 or 24000. Using the advanced trunking capabilities of Condor, the 48000 maintains an evenly balanced 1:1 relationship of front-end to back-end bandwidth on the 16-port 4Gbit blades. By taking advantage of local switching and high-port-count blades, it is not only possible, but actually *practical* to sustain 1.5 Tbits (3Tb) cross-sectional throughput in the chassis.

When making design trade-offs, availability is usually considered the most important factor. This is especially true for customers of director-class products. Of course, the 48000 has the usual director-class feature set, but it also has a more subtle characteristic related to reliability of port blades, which translates to availability of connections. The 48000 has the most efficient component integration of any FC director built to date.



**Figure 79 - FC16 Port Blade for Brocade 48000**

Note the highlighted section of the figure in the middle of the blade. This is the blade's "Condor" ASIC. It is the "brain" of the blade, containing the FC protocol logic, the serdes<sup>98</sup> functions, buffer memory, zoning enforcement memory and logic, performance counters, and so on. Having all of these functions in a single chip drastically reduces the complexity of the blade vs. competing approaches, which improves MTBF and lowers power and cooling requirements. Compare the single-ASIC approach of the FC16 to any other director in the industry, and the difference will be immediately apparent. It is certainly apparent when comparing the refinement of this blade to the port blade designs from its predecessors.

Perhaps the most important difference between the 48000 and its predecessors is that the Brocade 48000 is the "go forward" platform for the Brocade Enterprise roadmap. This means that purchasing a Brocade 48000 today is a strategic investment that will still have value for years to come. Brocade shipped the FR4-18i blade for FC

---

<sup>98</sup> A "Serializer / Deserializer" function, or serdes for short, is required by all FC switches to convert frames from a parallel mode (such as being held in a buffer) to a serial format suitable for transmission. In non-Brocade switches, serdes functions are generally on separate chips, which increases power draw, and lowers reliability.

routing and FCIP some time ago, and recently shipped several additional blades such as:

- iSCSI port blade
- 10Gbit FC blade
- Application Processor (AP) blade

The advanced feature blades are discussed in more detail later in this section.

The intention is to be able to populate the chassis with many different combinations of port blades.<sup>99</sup> For example, the system should support a configuration with a combination of e.g. 128 4Gbit fabric ports plus two LSAN router blades plus two iSCSI blades.

The Brocade 48000 Fibre Channel Director provides the following features today:

- 384 ports per chassis configured in 16-, 32-, or 48-port increments
- Current port blades support 1Gbit, 2Gbit, and 4Gbit Fibre Channel on a per-port basis
- FR4-18i router blade supports LSANs and FCIP
- FC4-16IP blade with FC and iSCSI support
- FA4-18 Application Blade with 16 virtualization ports
- FC10-6 10Gbit Fibre Channel blade
- Management access via 10/100Base-T RJ45 Ethernet ports and DB9 serial ports
- 14U rack mountable enclosure <30 inches deep. This allows up to 768 ports in a single rack.<sup>100</sup>

---

<sup>99</sup> It is possible that some combinational restrictions could apply, and support may vary between OEMs.

<sup>100</sup> Not all racks can support high-density configurations. The rack must be at least 42u high. There may need to be space between chassis for cable management. Power and cooling infrastructure, cable management, and structure of the floor must be sufficient. The organization supporting the SAN must often approve the installation of extremely high density deployments.

- High-availability features include hot-swappable FRUs for port blades, redundant power supplies and fans, and redundant CP blades
- Extensive diagnostics and monitoring for high Reliability, Availability, and Serviceability (RAS)
- Non-disruptive software upgrades (HCL/A)
- Non-blocking architecture enables 128 ports to operate at full 4Gbit line rate in full-duplex mode
- Forward and backward compatibility within fabrics with all Brocade 3000-series and later switches
- Brocade 12000s are upgradeable to 48000s
- Small Form-Factor Pluggable (SFP) optical transceivers allow any combination of supported Short and Long Wavelength Laser media (SWL, LWL, ELWL), as well as CWDM “colored laser” media
- Cables, blades, and PS are serviced from the cable side and fans from the non-cable side
- Air is pulled into the non-cable-side of the chassis and exits cable-side above the port and CP blades and through the power supplies to the right

### *Многопротокольный маршрутизатор Brocade AP7420*

Routers are used to connect different networks together, as opposed to bridging segments of the same network. In this context, “multiprotocol” means connecting networks using different protocols, generally at the lower levels of the stack.

For example, one network could use SCSI over Fibre Channel (i.e. FCP) and another could use SCSI over IP (e.g. iSCSI). In general usage, a router that can merely handle multiple Upper Layer Protocols (ULPs) but *not* different lower layer protocols is *not* considered multipro-

tocol.<sup>101</sup> For example, the ability to handle both SCSI/FC and FICON/FC would not qualify a product as a multiprotocol router, whereas handling both SCSI/FC and SCSI/IP might qualify for the term.

In the context of SANs, a multiprotocol router must connect Fibre Channel fabrics to each other and/or to some other networking protocols. Fibre Channel is mandatory since it is by far the leading protocol for use in SANs. Other protocols that a router may connect to include IP storage protocols such as FCIP and the emerging iSCSI standard.



### Side Note

*To learn more about SAN routing in general and the Brocade routers in particular, read the book Multiprotocol Routing for SANs, by Josh Judd.*

Brocade has created a multiprotocol SAN router provides which three functions critical to modern enterprise SAN deployments, and is designed to provide more in the future. At the time of this writing, the multiprotocol router software provides:

- FC-FC Routing Service for greater connectivity than traditional Fibre Channel SANs provide
- FCIP Tunneling Service for extending FC fabrics over distance using IP wide area networks

---

<sup>101</sup> For differing ULPs, a router may not need any special capabilities. For example, an IP/Ethernet router can handle both HTTP/IP/Ethernet and Telnet/IP/Ethernet without being “multiprotocol” per se: the ULP is transparent to the router. In some cases, a switch may need special “upper layer services” support for a ULP, such as CUP support on a FICON/FC switch. Even this does not qualify the switch as a multiprotocol router; it is simply an FC switch with enhanced FICON support.

- iSCSI Gateway Service for sharing Fibre Channel storage resources with iSCSI initiators

In addition to running these three services, the Brocade router is also a high-performance FC fabric switch.

The first platform the multiprotocol router software was delivered on was the Brocade AP7420 Multiprotocol Router (Figure 80). This platform first became generally available in early 2004. Multiprotocol router capabilities were added to the Brocade 48000 director in early 2006 via the FR4-18i blade, and the 4Gbit Brocade 7500 router shipped in the same timeframe. For most routing deployments, these two platforms have supplanted the 7420. For most application-layer deployments, the Brocade 7600 and FA-18 blade have replaced the 7420. However, this device is still useful in some cases.



Figure 80 - Brocade AP7420

Multiprotocol routing is a subset of the AP7420's capabilities: as well as performing its role as a multiprotocol router, it was designed to handle storage application processing requirements (a.k.a. "virtualization") for the full range of environments from small business to large-scale enterprises.

At only two RETMA units (2U) in height, the AP7420 allows deployment of fabric-based applications and multiprotocol routing using very little space. With "ports on demand" licensing, a single platform can be purchased with as few as eight ports and is scalable to sixteen ports.

with only the addition of a license key. Furthermore, its advanced networking capabilities allow scalability far beyond that level.

The AP7420 can make switching decisions using any protocol layers up to the very top of the protocol stack. This means that the platform hardware is able to function as a standard Fibre Channel fabric switch, an FC-FC router, or a virtualizer. Similarly, it could theoretically function as an Ethernet or IP switch from layer 2 to layer 4.<sup>102</sup> The platform has considerable flexibility, since every port has its own ASIC with multiple embedded CPUs and both Ethernet and Fibre Channel protocol support.

The Brocade AP7420 Fabric Application Platform provides the following features:

- 16 ports with software selectable modes, including auto-sensing 1Gbit/2Gbit FC, and 1Gbit Ethernet
- 2U rack mountable enclosure ~25 inches deep
- HA features including redundant hot-swappable power supply and fan FRUs
- Compatibility with all Brocade 2000-series and later Brocade switches within fabrics
- Management access via dual 10/100Base-T RJ45 Ethernet ports and one RJ45 serial port
- When in Fibre Channel mode:
  - Auto-sensing ports negotiate to the highest speed supported by attached devices
  - FC ports auto-negotiate as E\_Port or F\_Port. Any port may be configured as an FL\_Port to permit an

---

<sup>102</sup> Of course, as a practical matter, not *all* features and combinations of features will be supported in *software* just because the platform *hardware* is capable of delivering them. For combinations not explicitly called out in this book, discuss them with support and sales personnel.

- NL\_Port device to be attached, or as an EX\_Port for FC-FC routing
- Exchange-based ISL and IFL routing
- When in Gigabit Ethernet mode:
- Hardware acceleration through offloading TCP to port ASICs' ARM processors
- FCIP for deliver TCP/IP distance extension
- iSCSI initiator to FC target protocol mediation
- Per-port XPath ASICs for rapid data manipulation
- The XPath Fabric ASIC provides non-blocking connectivity between port ASICs
- SFP optical transceivers allow any combination of supported SWL, LWL, and ELWL media
- Latency is minimized through the use of storage application processors inside each port ASIC.
- Each port has LEDs to indicate behavior and status
- Air is pulled into the non-cable-side of the chassis, and exits cable-side above the SFPs

### *Многопротокольный маршрутизатор Brocade 7500*

The Brocade 7500 (Figure 81 - Brocade 7500) is a fixed-configuration 16-port 4Gbit FC router/switch with two additional ports for FCIP connectivity. The FCIP ports have the same capabilities as in the FC4-18i (p415). In addition to being able to perform all standard FC switching functions, it can route FC (i.e. to form IFLs) on all sixteen FC ports, and to route tunneled FC IFLs across the FCIP ports. This is a multiprotocol routing switch running Fabric OS 5.1 or above. In almost all cases, this is considered to be a replacement for the AP7420.



**Figure 81 - Brocade 7500**

Like the FC4-18i, the sixteen Fibre Channel ports will negotiate 1Gbit, 2Gbit, or 4Gbit speeds. The internal switching architecture is fully non-blocking and uncongested at full-speed / full-duplex. The internal switching fabric supports up to 256Gbits of bandwidth: more than enough to handle all ports at full speed.

### ***Платформа для приложений Brocade 7600***

The Brocade 7600 is a fixed-configuration 16-port Application Platform / Virtualization Switch. In addition to the 16 FC ports it has two additional 1000baseT ports for application management connectivity. This platform requires Fabric OS 5.3 or above.

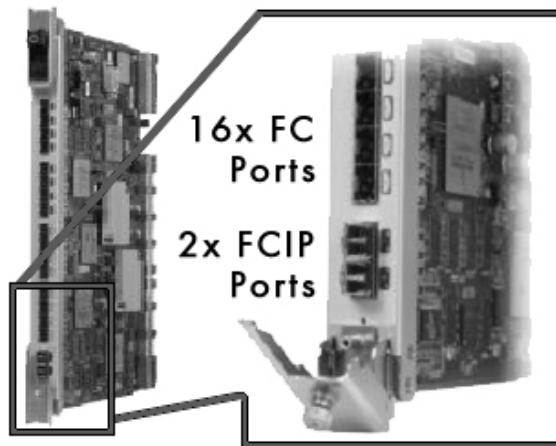


**Figure 82 - Brocade 7600**

The sixteen Fibre Channel ports that will negotiate 1Gbit, 2Gbit, or 4Gbit speeds. The internal architecture is fully non-blocking and uncongested at full-speed / full-duplex. The internal switching fabric supports all ports at full speed. In addition to the switching bandwidth it also has 128Gbit of virtualization bandwidth and the ability to support more than 1 million IOPS.

### ***Многопротокольное лезвие-маршрутизатор FR4-18i***

The FR4-18i (Figure 81) is a multiprotocol routing blade for the Brocade 48000 director (p405). There are sixteen FC ports on the blade, and two FCIP ports.



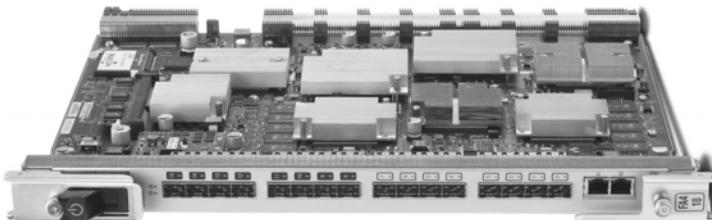
**Figure 83 - FR4-18i Routing Blade**

Each of the FC ports may be used for attachment of 1Gbit, 2Gbit, or 4Gbit FC devices such as hosts or storage, connection of FC Inter-Switch Links (ISLs), or for FC to FC routing via Inter Fabric Links (IFLs). It is also possible to use this blade to enable FC write-acceleration features generally applicable to DWDM or FCIP deployments in enterprise DR or BC solutions.

The FCIP ports may tunnel IFLs or ISLs. They support advanced distance extension features such as compression, encryption, and FastWrite acceleration, as well as having hardware acceleration for TCP headers to ensure top performance and standards compliance.

### **Платформа для приложений лезвие FA4-18**

The FA4-18 (Figure 84) is an Application / Virtualization blade for the Brocade 48000 director. There are sixteen (16) FC ports on the blade, and two (2) GE ports.

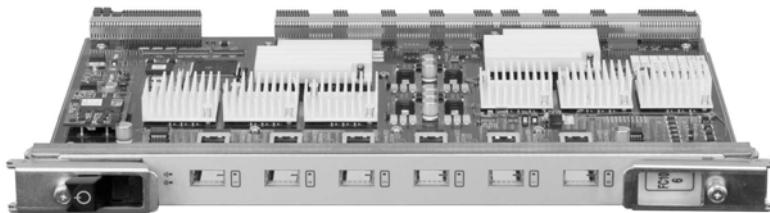


**Figure 84 – FA4-18 Application Blade**

Each of the FC ports may be used for attachment of 1Gbit, 2Gbit, or 4Gbit FC devices such as hosts or storage, or for FC ISLs. The internal architecture of the blade is fully non-blocking and uncongested at full-speed / full-duplex. The internal switching fabric supports up to 128Gbits of bandwidth: more than enough to handle all ports at full speed. In addition to the switching bandwidth it also has 128Gbit of virtualization bandwidth and the ability to support more than 1 million IOPS. The two 1000baseT ports are used to connect to external application management servers.

### ***Лезвие FC10-6 10Gbit Fibre Channel***

The FC10-6 (Figure 85) is a 10Gbit FC blade for the Brocade 48000 director. There are six (6) 10Gbit FC ports on the blade. Each of the ports may be used for attachment of 10Gbit FC ISLs. The internal architecture of the blade is fully non-blocking. There are two Condor ASICs to handle backplane connectivity, and six Egret ASICs (p509) to operate the 10Gbit ports. Each Egret has 720 buffer-to-buffer credits, which is sufficient to support a full-speed connection at 120km.



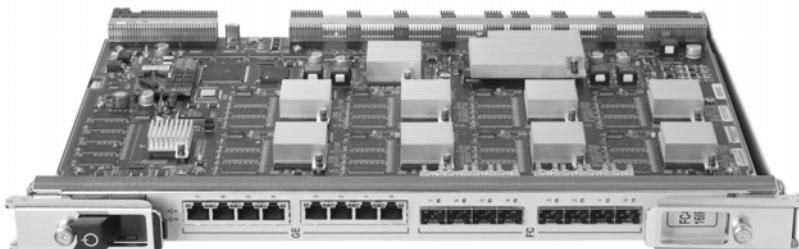
**Figure 85 - FC10-6 10Gbit FC Blade**

The 10Gbit FC ports are only supported for ISL connectivity at this time, as no 10Gbit Fibre Channel devices (hosts or storage) are widely available. Brocade does not expect 10Gbit devices to become widely available because 10Gbit is not cost efficient for nodes, and less expensive 8Gbit FC will be available before nodes could take full advantage of higher speeds in any case. 10Gbit Fibre Channel is targeted for MAN/ WAN deployments over xWDM or dark fiber networks.

### ***Лезвие FC4-16IP iSCSI to Fibre Channel***

The FC4-16IP (Figure 86) has a combination of 8 x 1Gbit iSCSI/Ethernet ports (GE) and 8 x 4Gbit Fibre Channel (FC) ports. It is designed for use in the Brocade 48000 director (p405).

The GE ports are used to connect to external iSCSI initiators directly, or (more typically) via external Ethernet switches for fan-in. The blade can support up to 64 iSCSI initiators per port (512 per blade).



**Figure 86 - FC4-16IP iSCSI Blade**

At the time of this writing, iSCSI initiators for Microsoft Windows, HP-UX, Solaris, Linux (RedHat and SuSE), and AIX are supported. The iSCSI initiators can take advantage of advanced features such as LUN masking and remapping. Additional features include Error Level Recovery 0, 1 and 2, iSCSI load balancing, CHAP support, and many other iSCSI protocol-specific features.

Each of the FC ports may be used for attachment of standard 1Gbit, 2Gbit, or 4Gbit FC devices such as hosts or storage, or connection of FC Inter Switch Links (ISLs). The internal architecture of the blade is fully non-blocking and uncongested at full-speed / full-duplex.

### ***Встроенные платформы***

In addition to stand-alone platforms, Brocade ASIC and software technology is used *within* products from a number of partners and OEMs. For example, Brocade FC switch ASICs are embedded into blade server products offered by some of the industry's top OEMs. This allows the connection of high density server blades into either existing fabrics or directly to storage. Brocade technology is also embedded within storage array controllers, providing a server fan-in capability integrated into the array. In effect, the OEM host or storage product contains some or all of the SAN internally, which tends to improve man-

ageability and reliability, and also lowers power, cooling, and rack space requirements.

Historically, connecting a large number of platforms with embedded switches to a larger SAN created scalability and manageability problems. If each storage device or blade-chassis also had one or more switch domains inside it, the size of the FC fabric could get out of hand quickly. Brocade developed the Access Gateway feature (p452) to eliminate this effect. Now, most embedded switches are capable of connecting to Brocade fabrics as F\_Ports instead of E\_Ports, so that they do not “show up” as switches in the fabric. Instead, they are projected as one or more nodes... which is actually what they really are, so that tends to work out well. To support Access Gateway, it is necessary to run appropriate code levels in the fabric as well as on the embedded switch, and OEM support is also required. Consult your local support organization to see if you can benefit from this feature.

#### Brocade 4020 Встроенный FC коммутатор

The Brocade 4020 was designed for the IBM eServer BladeCenter & the Intel Blade Server. It is powered by the “Golden eye” ASIC (p 502). The product is a single-stage central memory switch. It has a cross-sectional bandwidth sufficient to support all ports full-speed full-duplex at once in any traffic configuration. Fabric OS 5.0.2 or later is required.

The Brocade 4020 (Figure 87) has six outbound (to the SAN) and 14 inbound ports (one to each blade server), all are non-blocking and uncongested 4Gbit (8Gbit full-duplex) Fibre Channel fabric U\_Ports. This platform was introduced in 2006 by Brocade and IBM. The 4020 is available with software packages ranging from entry level (10-ports enabled) package up to the full enterprise-class Fabric OS 5.x feature set (as well as all 20 -ports

enabled). This allows the platform to be purchased with the right balance of cost vs. features for a wide range of customers, from small businesses to major enterprises. Regardless of licensed options, the 4020 has enterprise features such as HCL/A and the Fabric OS CLI.



**Figure 87 - Brocade 4020 Embedded Switch**

#### Brocade 4016 встроенный FC коммутатор

The Brocade 4016 was designed for the Dell Power-Edge blade server and for Fujitsu-Siemens PRIMERGY Server Blade. It is powered by the “Goldeneye” ASIC (p502). The product is a single-stage central memory switch. It has a cross-sectional bandwidth sufficient to support all ports full-speed full-duplex at once in any traffic configuration. Fabric OS 5.0.4 or later is required.

The Brocade 4016 (Figure 88) has six outbound ports (to the SAN), which are 4Gb bit, and 10 inbound ports (one to each blade server), which are 2Gb non-blocking and uncongested Fibre Channel fabric U\_Ports. This platform was introduced in 2006 by Brocade and Dell. The 4016 is available with software packages ranging from entry level (“12-ports enabled”) package up to the full enterprise-class Fabric OS 5.x feature set and all 16-ports enabled via Port-On-Demand. (See “Brocade Software” on p354.)



**Figure 88 - Brocade 4016 Embedded Switch**

Brocade 4018 Embedded FC Switch

The Brocade 4018 ( Figure 89) was designed for the Huawei Blade Server Chassis, and was introduced in 2006 by Brocade and Huawei. It is powered by the “Goldeneye” ASIC (p 502). The product is a single-stage central memory switch. It has bandwidth sufficient to support all ports full-speed full-duplex at once in any traffic configuration. Fabric OS 5.0.5 or later is required.



**Figure 89 - Brocade 4018 Embedded Switch**

The 4018 has four outbound ports (to the SAN) and 14 inbound ports (one to each blade server). All ports are non-blocking and uncongested 4Gbit (8Gbit full-duplex) Fibre Channel fabric U\_Ports. This board is typically factory installed, since - unlike other blade switches - it is a daughter board for an already existing controller module.

Brocade 4024 встроенный FC коммутатор

The Brocade 4024 was designed for the HP c-class BladeSystem. The Brocade 4024 is powered by the “Goldeneye” ASIC (p 502) and is a single-stage central memory switch. It has a cross-sectional bandwidth suffi-

cient to support all ports full-speed full-duplex at once. Fabric OS 5.0.5 or later is required.

The Brocade 4024 (Figure 90) has eight outbound ports (to the SAN) and 16 inbound ports (one to each blade server), all ports are non-blocking and uncongested 4Gbit (8Gbit full-duplex) Fibre Channel fabric U\_Ports. This platform was introduced in 2006 by Brocade and HP. The 4024 is available with software packages ranging from entry level (“12-port configuration”) up to the full enterprise-class Fabric OS 5.x feature set with all 24-ports enabled via Ports-On-Demand.



**Figure 90 - Brocade 4024 Embedded Switch**

#### Brocade 4012 встроенный FC коммутатор

The Brocade 4012 was introduced in 2005 by Brocade and HP. It represented the industry's first 4Gbit switch for embedded Blade Server market. The Brocade 4012 was specifically designed for the HP p-class Blade-System. It is powered by the “Goldeneye” ASIC. It has a cross-sectional bandwidth sufficient to support all ports full-speed full-duplex at once. Fabric OS 5.0.1 or later is required. The Brocade 4012 (Figure 91) has four outbound (to the SAN) and 8 inbound ports (one to each blade server), all outbound ports are non-blocking and uncongested 4Gbit (8Gbit full-duplex) and the inbound are all non-blocking and uncongested 2Gbit (4Gbit full-duplex) FC fabric U\_Ports.



Figure 91 - Brocade 4012 Embedded Switch

### ***Brocade iSCSI шлюз***

The Brocade iSCSI Gateway is an iSCSI-optimized product, designed to connect enterprise FC fabrics to low-cost “edge” servers. (Figure 92)



Figure 92 - Brocade iSCSI Gateway

Because this platform is smaller and offers fewer features than the FC4-16IP (p 419), it can be less expensive, and may be adequate for users who desire an entry point into the iSCSI bridging market. However, there are differences between the platforms besides cost and port count which must be considered when making a selection.

The iSCSI Gateway product is not capable of providing FC fabric switching. It has fewer features and lower performance than the bladed version. The Gigabit Ethernet interfaces on the iSCSI product are low-end copper, whereas the FC4-16IP uses more reliable optical ports capable of spanning greater distances. Be-

cause the iS CSI Gateway has RJ45 copper GE interfaces on the gateway itself, rather than just on the iSCSI hosts, users need to make sure that their IT networking group provides the correct interface.

This solution should be considered for customers who need a low cost entry point into the iSCSI bridging market above all else. Otherwise, a native Fibre Channel solution or the FC4-16IP will likely provide better results.

## Платформы классической McDATA

In 2007, Brocade purchased McDATA: one of its long-time rivals. However, this was not the first time that the two companies had enjoyed a partnership-style relationship. In fact, McDATA was one of Brocade's first customers, having purchased intellectual property from Brocade with which to implement its line of FC directors. Many McDATA installed-base platforms still run Brocade ASICs and code-chunks to this day. In addition, some of the companies that McDATA acquired prior to being purchased by Brocade had equivalently long-term partnerships with Brocade. For example, Brocade had a long-standing relationship with CNT in which CNT resold Brocade switches, and Brocade supported CNT for DR and BC solutions requiring certain distance extension methods.

Upon the close of the acquisition, Brocade announced end of sale for a subset of McDATA products in cases where they directly overlapped with Brocade offerings. For example, the McDATA "pizzabox" edge switches were superseded by the Brocade 5000. They had no value-added features beyond those available on the Brocade switches, so it was not necessary to continue to ship them for much longer after the close of the acquisition. Brocade announced that it intended to stop shipping these platforms at the end of 2007.

However, Brocade has a firm commitment to McDATA customers, and has not stopped shipping products such as the 140- or 256-port directors. It is expected that these platforms will converge with the Brocade director strategy at some point, but even when that happens they will be supported in Brocade networks via routed connections and compatible software releases for the foreseeable future. Also, Brocade intends to honor the support lifecycle commitments made by McDATA, which means that even products which Brocade no longer intends to actively sell are still being supported. Typically, support continues for five years after end of sale is announced.

This section discusses a few of the more notable classic McDATA products, and indicates how they may be integrated into a Brocade environment.

### ***Директор Brocade Mi10k***

The Brocade Mi10K offers up to 256 1-, 2-, and 4Gbit FC ports in a 14 U chassis. 10Gbit FC interfaces are also available for DR and BC solutions. It offers exceptional performance and availability. In some cases, it can even outperform the Brocade 48000, although in most deployments the 48000 has 50% more usable bandwidth<sup>103</sup> as well as 50% greater rack density, and much lower power and cooling requirements. Brocade is actively selling the Mi10k platform and has no immediate plans to stop doing so.

While this director is built using somewhat limited technology compared to the Brocade 48000, costs quite a

---

<sup>103</sup> The cases in which the Mi10k can outperform the 48000 are those in which little or no flow locality is achievable, and the host-to-storage port ratio is near 1:1. If either of those statements are false, then the 48000 will outperform the Mi10k by a considerable margin.

bit more, and requires considerably more power and cooling resources, for Classic McDATA customers who already have extensive Mi10k deployments, this is still the best option for transparently growing those environments. It is expected that Brocade will converge the applicable portions of the Mi10k feature-set with Brocade “native” director technology at some point in the future. In the meantime, the Mi10k is still being sold and supported, and can co-exist with Brocade-classic platforms using a number of strategies such as compatible firmware, routers, and storage-centric network topologies.

### ***Директор Brocade M6140***

The 140-port Brocade M6140 provides a high availability, high-performance, flexible building block for large SAN deployments. It is a single-stage, 140-port director designed supporting 1Gbit to 10Gbit FC interfaces. It can meet the connectivity demands of both open systems and mainframe FICON environments. Brocade is actively selling this platform and has no immediate plans to stop doing so.

While this director is built using somewhat outdated technology compared to the Brocade 48000, for Classic McDATA customers who already have extensive M6140 or 6064 deployments, the M6140 is still the best option for transparently growing those environments.

### ***Периферийные коммутаторы Brocade M4400 и M4700***

The M4400 has 16x 4G bit FC ports in a 1u / ½ rack-width form factor. The M4700 has 32x 4Gbit FC ports in 1u, and takes a full rack-width. These two platforms are still shipping at the time of this writing. Since the Brocade 5000 offers a superset of their capabilities, Brocade will stop selling the M4400 and M4700 at the end of 2007.

Support is expected to continue for five years after the final shipment date.

### ***Маршрутизаторы Brocade M1620 и M2640***

The M1620 has two GE ports for SAN extension, and two FC ports for local E\_Port connectivity. The platform can be deployed to support lower-end DR and BC environments. The M2640 has a similar architecture and use case, but with 12x FC ports and 4x GE ports.

These platforms used the now-defunct iFCP protocol for SAN extension. Since no other vendors ever implemented iFCP besides McDATA, and even McDATA had an FCIP roadmap, the iFCP protocol has actually been considered a dead end by the industry at large for several years. As a result, Brocade intends to stop selling the two platforms at the end of 2007 in favor of extension solutions using the Brocade 7500 router and FR4-18i blade, which support the FCIP protocol.

### ***Шлюз Brocade Edge M3000***

The Edge M3000 interconnects Fibre Channel SAN islands over an IP, ATM or SONET/SDH infrastructure. Brocade is actively selling this platform and has no immediate plans to stop doing so.

The M3000 enables many cost-effective, enterprise-strength data replication solutions, including both disk mirroring and remote tape backup/restore to maximize data availability and business continuity. Its any-to-any connectivity and multi-point SAN routing capability provide a flexible storage infrastructure for remote storage applications.

In most cases, the Edge M3000 has been superseded by the Brocade 7500 router and FR4-18i blade. However, in some cases the M3000 provides a superior fit. For example, depending upon the nature of the payload,

the M3000 can compress data by up to 20:1, dramatically reducing bandwidth costs. With this compression technology, customers can achieve gigabit per second throughput using existing 100Mb Ethernet infrastructure – but at a fraction of the cost. It also implements tape pipelining which can provide a considerable performance benefit for remote tape vaulting solutions.

Of course, not all customers have such highly compressible data, and equivalent features enabled or planned for the Brocade 7500 and FR4-18i may provide equivalent benefits, so the market for the M3000 is considered to be limited where compression and tape pipelining in particular and concerned. But the M3000 does have ATM and SONET/SDH connectivity advantages which are likely to keep it in the product portfolio for quite some time to come.

### ***Шлюз Brocade USD-X***

The USD-X is a high-performance platform that connects and extends mainframe and open-system storage-related data replication applications for both disk and tape, along with remote channel networking for a wide range of device types. Brocade is actively selling this platform and has no immediate plans to stop doing so.

While it is possible to use this platform in pure open-systems environments, the primary current use cases for this product are mixed and pure mainframe environments as other products solve the extension problem more cost-effectively for most open-systems customers.

This multi-protocol gateway and extension platform interconnects host-to-storage and storage-to-storage systems across the enterprise – regardless of distance – to create a high capacity, high performance storage network using the latest high speed interfaces. It supports Fibre Channel, FICON™, ESCON, Bus and Tag, or mixed en-

vironment systems. The intermediate WAN may be ATM, IP, DS3, or many other technologies.

## Инсталлированая база платформ Brocade

One of the advantages that Brocade has in the SAN marketplace is its large installed base. Brocade has millions of ports running in mission-critical production environments around the world, representing literally billions of hours of production operation to date. Brocade has a policy of prioritizing backwards compatibility with the installed base for new products.<sup>104</sup> This allows customers buying Brocade products to get a long useful life out of them, to achieve high ROI before needing to upgrade.

This subsection describes many of the platforms in the Brocade installed base. SAN designers may encounter any of these products, and must know their capabilities when designing solutions that involve them.

### *Коммутаторы SilkWorm 1xx0 FC*

The first platform-group that Brocade shipped was the Brocade 1xx0 series (Figure 93 and Figure 94). Shipped first in early 1997, this design was simply called the “SilkWorm switch” as there were no other Brocade platforms to differentiate between. Over time, other platforms were added. The first 16-port switch became known as the “SilkWorm I,” with its successor being the “SilkWorm II.” In early 1998, a lower cost 8-port “SilkWorm Express” platform was shipped based on the same

---

<sup>104</sup> Some restrictions apply, of course. For example, it may be necessary to run certain firmware versions and design solutions within scalability constraints to fit within a vendor support matrix. Also, it is not possible to continue support for an installed-base platform literally forever. The typical case is to continue support for five years after the last sale date of a product line.

architecture, but with half of the ports removed. By the time that the SilkWorm 2000 series shipped, Brocade had enough platforms that the first generation switches became known as the “SilkWorm 1xx0 series.”



Figure 93 - SilkWorm II (1600) FC Fabric Switch



Figure 94 - SilkWorm Express (800) FC Fabric Switch

These switches could be configured at the time of manufacture to support either FC-AL or FC fabric devices (Flannel or Stitch ASICs respectively, p 503) using combinations of 2-port daughter cards (Figure 95).

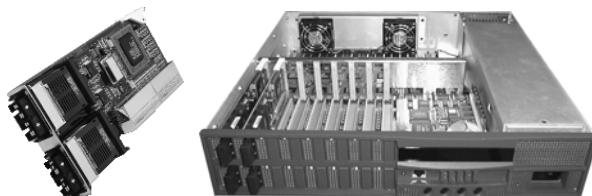


Figure 95 - SilkWorm 1xx0 Daughter Card

All SilkWorm 1xx0 switches ran Fabric OS 1.x. The product line consisted of 8- and 16-port FC fabric switches, with all ports running at 1Gbit. (8-port = SilkWorm Express and SilkWorm 800; 16-port = SilkWorm

II and SilkWorm 1600.) Ports could accept either optical or copper GBICs. Management tasks could be performed using buttons on the front panel on most models. All models had RJ45 IP/Ethernet and DB9 serial interfaces.

This Brocade platform group is considered to be entirely obsolete. The 1xx0 switches are simply not compatible with many of the new features released from Brocade over the past few years, and the hardware pre-dated some of the FC standards. Brocade recommends that SilkWorm 1xx0 series switches be upgraded to newer Brocade products and technologies in all cases.

### ***Коммутаторы SilkWorm 2xx0 FC***

The SilkWorm 2xx0 series consisted of several platforms all using the Loom ASIC (p. 504) and running Fabric OS 2.x. The first platform is in this group – the SilkWorm 2400 and 2800 – shipped in the middle of 1999. At the time of this writing, the SilkWorm 2xx0 platform group has reached the end of its supportable life. Most OEMs have declared these switches to be unsupported, and the rest are expected to do so by the end of the year. Users should consider 2xx0 switches to be obsolete, and should plan for upgrading in the near future.

Figure 96 through Figure 99 show the most popular 2xx0 series platforms. All of these products operated at 1Gbit Fibre Channel, and had a single-stage central memory architecture for non-blocking and uncongested operation. All of the switches in this series had an IP/Ethernet management port. Most had a DB9 serial port for initial configuration, emergency access, and out-of-band management, with the 2800 being the exception to that rule. (It had a push-button control panel and screen for initial configuration.)

The 2xx0 series has been superseded by other Brocade products. However, these switches are still widely

deployed. Brocade has found that the number of SilkWorm 2800 platforms still in production is close to the number that originally shipped: something on the order of a million ports in production. As a result, Brocade anticipates that many customers will need to perform 1Gbit to 4Gbit migrations over the next year, now that these switches have reached the end of their lifecycle.

### SilkWorm 20x0

The entry-level SilkWorm 20x0 (Figure 96) was a 1u 8-port switch, with seven fixed ports (GLMs) and one port with removable media (GBIC).

The platform could be purchased in three varieties, depending on the software keys that were loaded at the factory. The third digit in the platform product ID (20x\_0) indicated these software options, not any difference in hardware. The 2010 came with support only for QuickLoop, so only FL\_Ports could be attached, not F\_Port fabric devices or E\_Ports. The 2040 supported fabric nodes but only one E\_Port, and the 2050 had unlimited fabric support. Both the 2010 and 2040 provided customers with complete investment protection, as either could be upgraded to the full-fabric 2050 with license keys available through all channel partners. Power input was provided by a single fixed supply, and fans were fixed as well, so the entire platform was considered a FRU.



**Figure 96 - SilkWorm 2010/2040/2050**

SilkWorm 22x0

The 1.5u 16-port SilkWorm 22x0 (Figure 97) brought higher rack density to the entry-level switch market.



**Figure 97 - SilkWorm 2210/2240/2250**

It had a single fixed power supply, like the 20x0, and could be purchased with the same three software license variations. Also like the 20x0, the entire platform was considered a single FRU. However, all 16 media on the 22x0 were removable GBICs.

This platform was also used as the basic building block for the SilkWorm 6400, which consisted of a sheet metal enclosure containing six SilkWorm 2250 switches, configured and wired together at the factory to form a Core/Edge fabric, manageable as a single platform. That arrangement yielded sixty-four usable ports.

SilkWorm 2400

The SilkWorm 2400 (Figure 98) was targeted at the midrange segment. Like the 20x0, it was an 8-port switch, but had redundant hot-swappable power supplies and fans.



**Figure 98 - SilkWorm 2400**

SilkWorm 2800

The SilkWorm 2800 (Figure 99) was a 16-port switch like the 22x0, but had enterprise-class RAS features.

tures like the 2400. This was by far the most popular of the 2xx0 series. In many environments, the number of 2800 switches installed today still rivals the number of later platforms. This was the only platform in the series that did *not* have an externally-accessible serial port. Instead, the initial switch configuration could be performed using buttons and a screen built into the cable-side panel.



Figure 99 - SilkWorm 2800

### **Коммутаторы SilkWorm 3200 / 3800**

In 2001, the SilkWorm 2xx0 product family was superseded by the SilkWorm 3200 and 3800 switches. They were both powered by the Bloom ASIC (p505), which increased the port speed to 2GbE and added a range of new features including trunking, advanced performance monitoring, and more advanced zoning. Both platforms had IP/Ethernet and DB9 serial management interfaces, and both ran Fabric OS 3.x. Another major difference between these and prior Brocade platforms was that the SilkWorm 3200 and 3800 used SFPs, whereas all prior platforms had used GBICs.

At the time of this writing, the SilkWorm 3200 has been superceded by the SilkWorm 3250 (p 436), and the SilkWorm 3800 has been largely superceded by the SilkWorm 3850. (The SilkWorm 3800 is still shipping, but most users are expected to transition to the 3850 in the near future because of its many improvements.)

#### SilkWorm 3200

This platform had eight 2Gbit FC ports in a 1u enclosure. It was targeted at the entry market. Like its

predecessor, the SilkWorm 20x0, this switch had a single fixed power supply and fixed fans: the entire platform was considered a FRU.



**Figure 100 - SilkWorm 3200**

#### SilkWorm 3800

The SilkWorm 3800 was targeted at the midrange and enterprise markets. It had RAS features equivalent to the SilkWorm 2800.



**Figure 101 - SilkWorm 3800**

#### *Коммутаторы SilkWorm 3250 / 3850 FC*

These platforms represented the entry level of the Fibre Channel fabric switching market. They each had non-removable power supplies. Both were powered by the “Bloom-II” ASIC (p 503). The ASIC arrangement in both platforms yielded a single-stage central memory switch. They both had a cross-sectional bandwidth sufficient to support all ports full-speed full-duplex at once. Fabric OS 4.2 or later was required. The SilkWorm 3250 (Figure 102) had eight non-blocking and uncongested <sup>105</sup> 2Gbit

---

<sup>105</sup> There has been debate in the industry about the definition of “blocking.” When Brocade uses the word, it refers to Head of Line Blocking (HoLB). For example, the SilkWorm 24000 is not subject to HoLB because it uses virtual channels on the backplane. It is therefore “non-blocking.” All ports can run full-speed full-duplex at the same time, which is “uncongested operation.”

(4Gbit full-duplex) Fibre Channel fabric U\_Ports.<sup>106</sup> The SilkWorm 3850 (Figure 103) had sixteen ports.

These two platforms were introduced in 2004 to replace the popular Silk Worm 3200 and 3800 switches. Both were available with software packages ranging from the lowest entry level (“Value Line”) package up to the full enterprise-class Fabric OS 4.x feature set. (See “Brocade Software” on p444.) This allowed the platforms to be purchased with the right balance of cost vs. features for a wide range of customers, from small businesses to major enterprises. Regardless of licensed options, both switches had enterprise features such as hot (non-disruptive) code load and activation (HCL/A) and the Fabric OS CLI.



Figure 102 - SilkWorm 3250



Figure 103 - SilkWorm 3850

### *SilkWorm 3900 u 12000*

The SilkWorm 3900 (Figure 104) delivered 32 ports of 2Gbit Fibre Channel in a 1.5u rack-mountable enclosure.

---

<sup>106</sup> U\_Port interfaces automatically detect FC topology to become F\_Port, FL\_Port, or E\_Port as needed.



**Figure 104 - SilkWorm 3900**

First shipped in 2002, this platform was targeted at the midrange SAN market, but had many features appropriate for the enterprise market as well. In many ways, the SilkWorm 3900 was more like a small director than like a switch. Like the SilkWorm 12000, this platform had an “XY” topology CCMA multistage architecture. ( See “Многоуровневые внутренние архитектуры” on page 511 for more information.) Like the 12000, it supported FICON (a mainframe protocol), had redundant and hot swappable power and cooling FRUs, and ran Fabric OS 4.x with hot code load and activation.

Typical usage cases for the 3900 included stand-alone applications for small fabrics, edge deployments in small to large Core/Edge (CE) fabrics, and core deployments in small to medium CE fabrics.

The SilkWorm 12000 ( Figure 105) was Brocade’s first fully-modular 10-slot enterprise-class director. This system first shipped in 2002.



**Figure 105 - SilkWorm 12000 Director**

The chassis was rack-mountable in 14u, and could be populated with up to eight port-blades and two CPs. Overall, the chassis could be configured starting with 32 and going up to 128 2Gbit Fibre Channel ports. Each blade was hot-pluggable, as were the fans and power supplies. The redundant CPs ran Fabric OS 4.x and supported HCL/A. Typical usage cases for the 12000 included stand-alone applications, edge deployments in large CE fabrics, and core deployments in medium to large CE fabrics.

The backplane interconnected the port blades with each other to form two separate 64-port domains. The interconnection employed an “XY” topology CCM A multistage architecture, much like the SilkWorm 3900. The two 64-port domains were both controlled by the same redundant CP blades, and resided in the same chassis, but had no internal data path between them. They could be used separately in redundant fabrics, or could be used together in the same fabric by connecting them with ISLs.

At the time of this writing, the SilkWorm 3900 has been superseded by the SilkWorm 4100 (p 400), and the SilkWorm 12000 has been superseded, first by the SilkWorm 24000 (p 403), and then the 48000 (p 403). For the foreseeable future, the older platforms will continue to be supported in networks with more advanced platforms. In addition, the SilkWorm 12000 chassis can be upgraded in the field to become a SilkWorm 24000 or 48000.<sup>107</sup>

### *Директор SilkWorm 24000*

The SilkWorm 24000 (Figure 106) was a fully-modular 10-slot enterprise-class director, and could be populated with up to eight port-blades and two Control Processors (CPs). This platform first shipped in early 2004. It could be configured from 32 to 128 ports in a single domain using 16-port 2G bit Fibre Channel blades. The platform had industry-leading performance and high availability characteristics. Each blade was hot-pluggable, as were the fans and power supplies. The chassis had redundant control processors (CPs) with redundant active-active uncongested and non-blocking switching elements, which ran Fabric OS 4.2 or higher and supported HCL/A.

---

<sup>107</sup> Of course, not all OEMs support this procedure.



**Figure 106 - SilkWorm 24000 Director**

The SilkWorm 24000 was an evolution of the SilkWorm 12000 design. It could use the same chassis as the 12000: the power supplies, fans, backplane, and sheet metal enclosure were all compatible. As a result, it was possible to upgrade an existing 12000 chassis to the 24000 in the field by replacing just the CP and port blades. Look between Figure 106 and Figure 105 (p 439) and the similarity will be apparent. It can also support 16-port 4Gbit FC Brocade 48000 blades in some combinations with existing SilkWorm 24000 blades.

Even though the chassis were mechanically compatible, there were differences between the SilkWorm 24000 and the SilkWorm 12000.

Some of the differences were minor. For example, the 24000 chassis and blade set had an improved rail glide system that makes blade insertion / extraction easier. Larger ejector levers helped by providing greater mechanical advantage. The 24000 CP blades had a blue LED to indicate which CP was active.

There were also more important differences in the underlying technology. For example, the 24000 used the “Bloom-II” ASIC, while the 12000 used the original “Bloom” chipset. (See “Bloom и Bloom-II” p 505.) The overall chassis power consumption and cooling requirements were lowered by more than 60%, with the result that ongoing operational costs were reduced and MTBF increased by more than 25%. Further improvements in MTBF were achieved through component integration: fewer components means less frequent failures. Performance was improved by changing the multistage chip layout from an “XY” topology to a “CE” arrangement. (See “**Многоуровневые внутренние архитектуры**” on page 511 for more information.) This allowed the 24000 to present all of its ports in a single internally-connected domain. The 12000, in contrast, presented two 64-port domains and needed external ISLs if traffic was required to flow between the domains.

The SilkWorm 24000 Fibre Channel Director provided the following features:

- 128 ports per chassis in 16-port increments
- Port blades are 1Gbit/2Gbit Fibre Channel
- Management access via Ethernet and serial ports
- High-availability features include hot-swappable FRUs for port blades, redundant power supplies and fans, and redundant CP blades
- Extensive diagnostics and monitoring for high Reliability, Availability, and Serviceability (RAS)
- Non-disruptive software upgrades (HCL/A)
- 14U rack mountable enclosure allows up to 384 ports in a single rack.
- Non-blocking architecture allows all 128 ports to operate at line rate in full-duplex mode
- Forward and backward compatibility within fabrics with all Brocade 2000-series and later switches

- SilkWorm 12000s are upgradeable to 24000s
- Small Form-Factor Pluggable (SFP) optical transceivers allow any combination of supported Short and Long Wavelength Laser media (SWL, LWL, ELWL), as well as CWDM media
- Cables, blades, and PS are serviced from the cable side and fans from the non-cable side
- Air is pulled into the non-cable-side of the chassis and exits cable-side above the port and CP blades and through the power supplies to the right

### *Встроенные продукты*

#### SilkWorm 3016 встроенный FC коммутатор

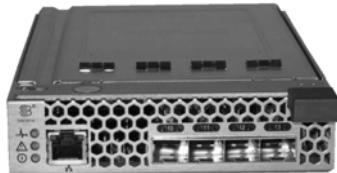
The SilkWorm 3016 was specifically designed for the IBM eServer BladeCenter. It was powered by the “Bloom-II” ASIC. It had a cross-sectional bandwidth sufficient to support all ports full-speed full-duplex at once. The SilkWorm 3016 (Figure 107) has two outbound ports (i.e. facing to the SAN) and 14 inbound ports (one to each blade server), all are non-blocking and uncongested 2Gbit (4Gbit full-duplex) Fibre Channel fabric U\_Ports. This platform was introduced in 2004 by Brocade and IBM. The 3016 was available with software packages ranging from entry level (“Value Line”) package up to the full enterprise-class Fabric OS 4.x feature set.



**Figure 107 - SilkWorm 3016 Embedded Switch**

SilkWorm 3014 встроенный FC коммутатор

The SilkWorm 3014 was specifically designed for the Dell PowerEdge blade server. It was powered by the “Bloom-II” ASIC. It had a cross-sectional bandwidth sufficient to support all ports full-speed full-duplex at once. The SilkWorm 3014 (Figure 108) had four outbound (to the SAN) and 10 inbound ports (one to each blade server), all were non-blocking and uncongested 2Gbit (4Gbit full-duplex) Fibre Channel fabric U\_Ports.



**Figure 108 - SilkWorm 3014 Embedded Switch**

This platform was introduced in late 2004 by Brocade and Dell. The 3014 was available with software packages ranging from entry level (“Value Line”) package up to the full enterprise-class Fabric OS 4.x feature set.

## **Лицензируемые функции Brocade**

Brocade adds value in its products with both hardware (i.e. ASICs) and software. This subsection describes some of the most popular software features Brocade offers. It only covers features developed internally by Brocade Engineering; it does not, for example, discuss third-party management tools which use one of the supported APIs.

### **Модель лицензирования Brocade**

Some features are basic components of the operating system and platform ASICs, such as support for nodes using N\_Port. (I.e. support for F\_Port on a switch.) These generally do not require purchasing a license key, but do add value. Some Brocade competitors (i.e. Ilo op-switch

vendors) do not offer products that support F\_Port, so even though it seems like this should be a basic building block of all switches, it is worth calling it out explicitly to show its value.

Other features, such as the FC-FC Routing Service, require much higher value enhancements to both ASIC and OS support. Routing features and more advanced fabric service options require the purchase and installation of license keys to enable them. On all platforms, the CLI<sup>108</sup> command *licenseShow* can be used to determine which keys are installed. If a desired feature is missing, work with the appropriate sales channel to purchase the key, and then use the *licenseAdd* command to install it on the switch or router.

### ***Подключение Fabric Node (F\_Port)***

At the time of this writing, most Fibre Channel nodes (e.g. host and storage devices) use the N\_Port topology. “Node Port” is a set of standards-defined behaviors that allow a node to access a fabric and its services most cleanly. In order to connect an N\_Port to a switch, the switch must support the corresponding “Fabric Port,” or F\_Port topology as defined in the standards. Every Brocade platform ever shipped supports F\_Port, although in a few of the older platforms (e.g. the SilkWorm 2010) this feature required purchasing a separate license key. This is the preferred method for connecting nodes into fabrics.

### ***Подключение Loop Node (FL\_Port) (QL/FA)***

Early in the evolution of Fibre Channel, there was debate about whether or not fabrics were necessary. Some

---

<sup>108</sup> There are also equivalent GUI commands in WEBTOOLS and Fabric Manager. CLI commands are generally used for examples because all platforms include the CLI as part of the base OS, while some do not include the GUI tools.

vendors believed that F\_C-AL hubs and “loop switches” provided sufficient connectivity. The argument went something like, “How many people will ever need more than a dozen or so devices in a SAN? Nobody!” It turned out that the real answer was, “Just about everybody,” so the vastly more scalable and flexible fabric switches rapidly eroded the hub market.

To accomplish this market transition gracefully, it was necessary for nodes designed for FC-AL hubs to attach to fabric switches. The Fibre Channel standards defined a switch port type to accomplish this: the FL\_Port. (“Fabric Loop” port.) This allowed, for example, HBA drivers written for hubs to present NL\_Ports (“Node Loop” ports) and plug into switch FL\_Ports. Brocade developed the Flannel ASIC (p503) to address this need. Platforms using Flannel needed to be configured with loop ports at the factory, but in subsequent products with more advanced ASICs, any port could support loop nodes.<sup>109</sup>

There are some important variables that affect how loop devices connect to a fabric:

- Does the loop device know how to talk to the name server, and does it know how to address devices using all three bytes of the fabric “PID” address? (Public vs. private loop.)
- If the device uses private loop, is it an initiator or a target? Private loop initiators need more help to use fabric services, i.e. the name server.
- Is there just one loop device directly attached to a switch port (like an NL\_Port HBA) or are there many loop devices on that port (like a JBOD)?

---

<sup>109</sup> Throughout the remainder of this subsection, the obsolete SilkWorm 1xx0 series will not be considered. E.g. statements about “all platforms” may actually refer to “all platforms except the SilkWorm 1xx0.”

Public loop support for a directly attached NL\_Port is the easiest case for a switch to handle. The switch ASIC needs to be able to support FC-AL “loop primitives,” which is the protocol used for loop initialization and control. All ports on all Brocade platforms today have the hardware and software to support this mode of operation as part of the base OS.

Public loop support for multiple nodes on a single switch port is slightly more complex. At the time of this writing, all platforms except the AP7420 Multiprotocol Router support this mode as part of the base OS. The major application for this is JBODs: it is currently possible to attach a JBOD directly to the AP7420, but JBODs can coexist in a fabric or Meta SAN with that platform.

Private loop storage devices require still more advanced ASIC functionality known as “phantom logic,” and corresponding software enhancements. This allows Network Address Translation (NAT) between the one-byte private loop and three-byte fabric address spaces. This needs ASIC hardware support because every frame needs to be rewritten without performance penalty. Trying to implement multi-gigabit NAT in software would not be practical. Brocade began to provide support for private loops with the Flannel ASIC.

Private loop technology has been declining rapidly, so Brocade had not prioritized phantom logic for future platforms. All ASICs through Bloom-II (p505) support this, but subsequent ASICs like FiGeRo (p510) and Condor (p506) do not. Platforms like the Multiprotocol Router and the Brocade 4100 cannot accept direct private storage attachment, but can co-exist seamlessly in networks with private storage attached to Loom, Bloom, and Bloom-II switches. Switches with private storage support include it as part of the base OS.

Private loop *initiators* (hosts) are the hardest case to solve. Not only do they require loop primitives and phantom logic, but they also require much more advanced fabric services enhancements.

An initiator normally queries the fabric name server for targets, and then sends IO to them. With *public initiators* talking to *private targets*, a switch can “notice” the IO from the initiator and automatically set up phantom logic NAT entries as needed. *Private initiators* do not know how to talk to the name server; they learn about available targets by probing their loop. They cannot send IO to a target until after NAT has been set up, so the automatic learning mechanism does not work.

The “Quick Loop / Fabric Assist” optionally licensed feature set is designed to address this need. Users explicitly define which devices a private host needs to access using zoning, and the switch creates the required NAT entries on that basis. QL/FA is supported as an optionally licensed feature on the SilkWorm 2xx0 series, and the SilkWorm 3200/3800 switches, i.e. all Fabric OS 2.x and 3.x platforms. QL/FA only applies to private initiators, not to any other usage case, and private initiators are the most rapidly declining segment of the SAN market. As a result, Brocade has not prioritized porting the feature to 4.x or beyond, except to support QL/FA on 2.x/3.x switches in the same fabric as 4.x switches. At the time of this writing, even that level of QL/FA support is essentially obsolete.

### **Фабрики из нескольких коммутаторов (E\_Port)**

The E\_Port (Expansion Port) protocol allows switches to be interconnected to form a larger fabric: a single region of connectivity built from multiple discrete

switching components.<sup>110</sup> This feature allows SAN solutions to be built using a “pay as you grow” approach, adding switches to a fabric as needed. It also allows much more flexible network designs, including support for geographical separation of components. Without this feature, the maximum scalability of a connectivity model would be limited to the number of ports on a single switch, and the maximum geographical radius of a network would be the distance supported by a node connected to that switch.

Today, the ability to network switches together to form a fabric seems commonplace, but when Brocade started selling switches for production use in 1997, it was a key differentiator. Most competitors could not do this at all, and the few that had the feature had many configuration constraints. Brocade was not just *a* pioneer in this space; Brocade was *the* pioneer. This is reflected in the fact that FSPF<sup>111</sup> was authored and given to the standards bodies by Brocade. Without this and other Brocade-authored protocols, it would not be possible much less commonplace to form multi-switch fabrics today.

### ***Виртуальные каналы***

A unique feature available in every Brocade 2Gbit and 4Gbit fabric switch, Brocade Virtual Channel (VC) technology represents an important breakthrough in the design of large SANs.<sup>112</sup> To ensure reliable ISL communications, VC technology logically partitions bandwidth within each

---

<sup>110</sup> This also requires the interaction of other fabric services, such as the name server and zoning database processes, but Brocade keys the feature off of E\_Port.

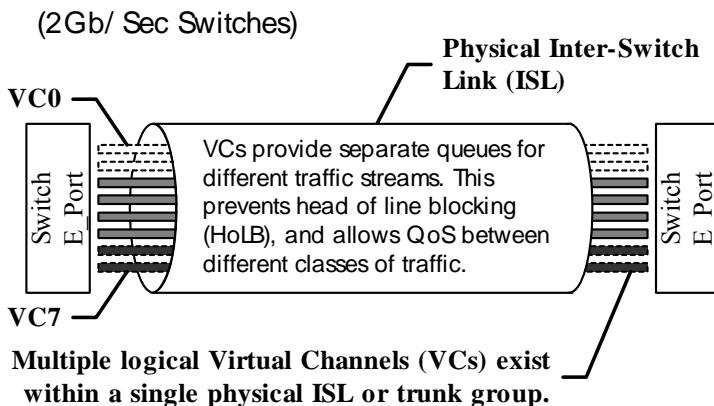
<sup>111</sup> The protocol used by all vendors to determine topology and path selection.

<sup>112</sup> Actually, even the SilkWorm 1xxx series of switches had a form of VC support, but it was quite different and not particularly relevant to SAN design today. But it is interesting to note that Brocade has already gone through four generations of VC development: it's a “well-baked” feature.

ISL into many different virtual channels as shown in Figure 109, and prioritizes traffic to optimize performance and prevent head of line blocking.

Fabric Operating System automatically manages VC configuration, eliminating the need to manually tune links for performance. This technology also works in conjunction with trunking to improve the efficiency of switch-to-switch communications, and simplify fabric design.

## Virtual Channels Mapped to ISLs



**Figure 109 - VCs Partition ISLs into Logical Sub-Channels**

In 2Gbit Brocade products, there were a total of 8 VCs (0-7) assigned to any link. This could be internal links, ISLs, or trunk groups. Each VC had its own independent flow control mechanisms and buffering scheme.

In Brocades 4Gbit products, the Virtual Channel infrastructure has been greatly enhanced, and some of the automatic assignment mechanisms have been improved. There are now 17 VCs assigned to any given internal link: one for class F traffic and sixteen for data. Each data VC now has 8 sub-lists or sub-Virtual Channels; each of those has its own credit mechanism and independent flow control. SID/DID pairs are assigned in a round-robin

fashion across all the VCs, but with these new enhancements, a better distribution is made. Of course, when connecting 4Gbit switches together with 2Gbit switches, the ISLs and trunk groups still use 8 VCs. This is done to avoid potential backwards compatibility issues.

In the near future, Brocade will be releasing a QoS feature which allows 4Gbit switches to use the increased VC capabilities to prioritize some flows above others in congested networks. As a practical matter, this feature is expected to apply almost exclusively to long distance connections in DR or BC solutions, since, for local-distance ISLs and IFLs, it is generally better to avoid congestion in the first place than it is to manage which devices are most harmed by congestion.

### ***Буферные кредиты***

Buffer-to-buffer (BB) credits are used by switch ports to determine how many frames can be sent to the recipient port, thus preventing a source device from sending more frames than can be received. The BB credit model is the standard method of controlling the flow of traffic within a Fibre Channel fabric.

Like VCs, BB credits are handled automatically by the Fabric Operating System in most cases. For extremely long distance links, it may be desirable to manually increase the number of credits on a port to maximize performance. (This may require an Extended Fabrics license.)

In the context of host or storage connections to a switch, the number of BB credits on a link will be negotiated between the device and the switch at initialization time. For ISL connections, each Virtual Channel will receive its own share of BB credits. In this case, credits are handled the same way whether the port is part of a trunk group or operating independently.

This topic is discussed in more detail under “” on page 346.

### ***Шлюз доступа***

Access Gateway uses the N\_Port ID Virtualization (NPIV) standard to present blade server FC connections as logical nodes to fabrics. This eliminates entire categories of traditional heterogeneous switch-to-switch interoperability challenges. Attaching through NPIV-enabled switches and directors, Access Gateway seamlessly connects server blades to Brocade, Cisco McDATA, or even to other vendors’ SAN fabrics.

Traditionally, when blade servers in chassis have been connected to SANs, each enclosure would add one or two more switch domains to the fabric, which had a potentially disastrous effect on scalability. Increasing the number of blade enclosures also meant additional switch domains to manage, increasing day-to-day SAN management burden. These additional domains created complexity and could sometimes disrupt fabric operations during the deployment process. Finally, fabrics with large numbers of switch domains created firmware version compatibility management challenges: sometimes it was impossible to find a firmware version which was supported by all devices in the fabric.

To address these challenges, Access Gateway presents blade server NPIV connections rather than switch domains to the fabric. This means that Access Gateway can support a much larger fabric, and that switch firmware on the Access Gateway does not interact with the other switches in the fabric *as a switch*. Rather, it interacts as a node, which greatly reduces firmware dependencies. Unlike FC pass-through solutions, it can do all of this without substantially increasing the number of switch ports required.

To enhance availability, Access Gateway can automatically and dynamically fail over the preferred I/O connectivity path in case one or more fabric connections fails. This approach helps ensure that I/O operations finish to completion, even during link failures. Moreover, Access Gateway can automatically fail back to the preferred fabric link after the connection is restored, helping to maximize bandwidth utilization.

### **PO Value Line**

The Value Line software license packages reduce the cost of acquiring and deploying an entry-level SAN, while allowing software-key upgrades to full enterprise-class functionality. Designed for small and medium sized organizations, the Value Line integrates innovative hardware and software features that make it easy to deploy, manage, and integrate into a wide range of IT environments. These powerful yet flexible capabilities enable organizations to start small and grow their storage networks in a scalable, non-disruptive, and efficient manner. This is especially beneficial for organizations that need to upgrade their existing SAN environment with minimal disruption. In addition, they simplify administration through embedded Brocade WEBTOOLS software.

The main thing that SAN designers need to be aware of is that a Value Line switch might not have full fabric capabilities. In exchange for substantially reduced acquisition cost, the buyer of a Value Line switch would give up features such as fabric scalability (number of domains supported) or number of E\_Ports allowed. When deploying a Value Line switch into a larger solution, it might therefore be necessary to upgrade its license key to a full fabric key.

## *Виртуальные фабрики / административные домены*

Virtual Fabrics allows the partitioning of one physical fabric into multiple logical fabrics that can be managed by separate Admin Domains by administrators. Virtual Fabrics are characterized by hierarchical management, granular and flexible security, and fast and easy reconfiguration to adapt to new infrastructure requirements. They allow IT administrators to manage separate corporate functions separately, use different permission levels for SAN administrators, provide storage for teams in remote offices without compromising local SAN security, and increase levels of data and fault isolation without increasing SAN cost and complexity. Once Fabric OS 5.2.0 or later is installed in the SAN, Virtual Fabrics can be implemented on the fly with no physical topology changes and no disruption.

The Administrative Domains feature is the key enabler for Virtual Fabrics technology. Admin Domains create partitions in the fabric. Admin Domain membership allows device resources in a fabric to be grouped together into separately managed logical groups. For example, a SAN administrator might have the Admin role within one or more Admin Domains, but be restricted to the Zone Admin role for other Admin Domains.

Although they are part of the same physical fabric, Virtual Fabrics are separate logical entities because they are isolated from each other via several mechanisms such as:

• Data isolation: Although data can pass from one Virtual Fabric to another using device sharing, and links can be shared among multiple Virtual Fabrics, no data can be unintentionally transferred even when Virtual Fabrics are not zoned.

Control isolation: Within Virtual Fabrics, fabric services are independent and are secured from unwanted interaction with other Virtual Fabric services. This includes zoning, RSCNs, and so on.

Management isolation: Switches in a Virtual Fabric provide independent management partitions. If a switch is a member of more than one Virtual Fabric, it has multiple, independent management entities. Administrators are authenticated to manage one or more Virtual Fabrics, but they cannot access management objects in other, unauthorized Virtual Fabrics.

Fault isolation: Data control or management failures in one Virtual Fabric will not impact any other Virtual Fabric services.

Admin Domains administrators can manage one or more Admin Domains while Virtual Fabric administrators have administrative permissions on all Admin Domains. Separate Admin Domains can be created for different operating systems (FICON®, Z-Series, and open systems, for example).

Devices can easily be shared among different Admin Domains without any special routing requirements. Admin Domain administrators can configure and manage their own zones; they can configure all rights and devices as long as they have the Admin role for that particular Admin Domain. The Admin Domain feature is backwards compatible with the millions of Brocade SAN ports already deployed, and no new hardware is required.

Implementing Virtual Fabrics is straight-forward, and fits into existing SAN management models. The management and best practices used today in a pre-Fabric OS 5.2.0 physical fabric with zoning can be implemented in the same way in a Fabric OS 5.2.0 fabric with Admin Domains and zoning.

## ***FCIP FastWrite и Tape Pipelining***

FCIP is a method of transparently tunneling FC ISLs between two geographically distant locations using IP as a transport. Storage is often sensitive to latency, and throughput is a great concern as well. Unfortunately, IP networks tend to have high latency and low throughput compared to native FC solutions. Tape Pipelining and FastWrite are features available on the Brocade 7500 router and FR4-18i blade that improve throughput and mitigate the negative affects of IP-related delay.

Tape Pipelining refers to writing to tape over a Wide Area Network (WAN) connection. FastWrite refers to Remote Data Replication (RDR) between two storage subsystems. Tape is serial in nature, meaning that data is steadily streamed byte by byte, one file at a time onto the tape from the perspective of the host writing the file. Disk data tends to be bursty and random in nature. Disk data can be written anywhere on the disk at any time. Because of these differences, tape and disk are handled differently by extension acceleration technologies.

Tape Pipelining accelerates the transport of streaming data by maintaining optimal utilization of the IP WAN. Tape traffic without an accelerator mechanism can result in periods of idle link time, becoming more inefficient as link delay increases.

When a host sends a write command, a Brocade 7500/FR4-18i sitting in the data path intercepts the command, and responds with a “transfer ready”. The router buffers the incoming data and starts sending that data over the WAN. The data is sent as fast as it can, limited only by the bandwidth of the link or the committed rate limit. On the heels of the write command is the write data that was enabled by the proxy target’s transfer-ready reply. After the remote target receives the command, it responds with a transfer ready. The remote router intercepts

that transfer ready, acts as a proxy initiator, and starts forwarding the data arriving over the WAN.

The host is on a high-speed FC network, and most often will have completed sending the data to the local router by this time. The local router returns an affirmative response. While the buffers are still transmitting data over the link, the host sends the next write command and the process is repeated on the host side until the host is ready to write a filemark. This process maintains a balance of data in the remote router's buffers, permitting a constant stream of data to arrive at the tape device.

On the target side, the transfer ready indicates the allowable amount of data that can be received, which is generally less than what the host sent. The transfer ready on the host side, from the proxy target, is for the entire quantity of data advertised in the write command. The transfer ready the proxy target responds with for the entire amount of data does not have to be the same as the transfer ready the tape device responds with, which may be for a smaller amount of data, that is, the amount that it was capable of accepting at that time. The proxy initiator parses out the data in sizes acceptable to the target per the transfer ready from the tape device. This may result in additional write commands and transfer readies on the tape side compared to the host side. Buffering on the remote side helps to facilitate this process.

The command to write the filemark is not intercepted by the routers and passes uninterrupted from end to end. When the filemark is complete, the target responds with the status. A status of OK indicates to the host that it can move on.

FastWrite works in somewhat different manner. FastWrite is an algorithm that reduces the number of round trips required to complete a SCSI write operation. FastWrite can maintain throughput levels over links that

have significant latency. The Remote Data Replication (RDR) application still experiences latency; but reduced throughput due to that latency is minimized.

There are two steps to a SCSI write:

1. The write command is sent across the WAN to the target. This is essentially asking permission of the storage array to send data. The target responds with an acceptance (FCP\_XFR\_RDY).
2. The initiator waits until it receives that response from the target before starting the second step, which is sending the actual data (FCP\_DATA\_OUT).

With the FastWrite algorithm, the local SAN router intercepts the originating write command and responds immediately requesting the initiator to send the entire data set. This happens in a couple of microseconds. The initiator starts to send the data, which is then buffered by the router. The buffer space in the router includes enough to keep the “pipe” full plus additional memory to compensate for links with up to 1% packet loss.<sup>113</sup> The Brocade 7500/FR4-18i has a continuous supply of data in its buffers that it can use to completely fill the WAN, driving optimized throughput.

The Brocade 7500/FR4-18i sends data across the link until the committed bandwidth has been consumed. The receiving router acts on behalf of the initiator and opens a write exchange with the target over the local fabric or direct connection. Often, this technology allows a write to complete in a single round trip, speeding up the process considerably and mitigating link latency by 50%.

---

<sup>113</sup> If a link has more than 1% packet loss or more, it means that there are serious network issues that must be resolved prior to a successful implementation of FastWrite.

There is no possibility of undetected data corruption with FastWrite because the final response (FCP\_RSP) is never spoofed, intercepted, or altered in any way. It is this final response that the receiving device sends to indicate that the entire data set has been successfully received and committed. The local router does not generate the final response in an effort to expedite the process, nor does it need to. If any single FC frame were to be corrupted or lost along the way, the target would detect the condition and not send the final response. If the final response is not received within a certain amount of time, the write sequence times out (REC\_TOV) and is retransmitted. In any case, the host initiator knows that the write was unsuccessful and recovers accordingly.

### ***FC FastWrite***

For native FC links or FC over *x*WDM, delay and congestion are typically one or more orders of magnitude better than with FCI\_P. However, the speed of light through glass still creates noticeable latency over long distance connections. As a result, it is possible for FC links over MAN/WAN distances to benefit from the same algorithms used in FCIP FastWrite. Brocade has added support for this feature to its 4Gbit router portfolio.

For example, it is possible to deploy FR4-18i blades into chassis at each side of a DR or BC solution, and attach storage ports directly to these blades. (This is illustrated in “” starting on page 364.) After configuring appropriate zoning policies, any replication or mirroring traffic between the storage ports will be accelerated using a similar mechanism to the one described in the previous section. This can sometimes result in massive increases in throughput, with the exact improvement depending on the distance, congestion of the network, block size, and the number of devices sharing the inter-site links.

## ***Горячая загрузка и активация кода***

Hot code load and activation supports the stringent availability requirements of mission-critical environments by enabling firmware upgrades to be downloaded and activated without disrupting other operations or disruption to data traffic in the SAN. The switch continues to route frames and provide full fabric services while new firmware is loaded onto its non-volatile storage. Once the download is complete, the new image is activated. During the activation process, the switch still continues to route frames, without losing even a single bit of data traffic.

## ***Advanced ISL Trunking (Frame-Level)***

Brocade ISL Trunking is ideal for optimizing performance and simplifying the management of a multi-switch SAN fabric containing Brocade switches. When two, three, or four adjacent ISLs are used to connect two Brocade 2Gbit FC switches, the switches automatically group the ISLs into a single logical ISL, or “trunk.” With 4Gbit switches, it is possible to trunk up to eight adjacent links. Traffic will be balanced across these links, while still guaranteeing in-order and on-time delivery.

ISL Trunking is designed to significantly reduce traffic congestion in storage networks. When up to eight 4Gbit ISLs are combined into a single logical ISL, the aggregated link has a total bandwidth of 32 Gbit/sec which can support a large number of simultaneous full-speed “conversations” between devices.

To balance workload across all of the ISLs in the trunk, each incoming frame is sent across the first available physical ISL in the trunk. As a result, transient workload peaks for one system or application are much less likely to impact the performance of other parts of the SAN fabric. Because the full bandwidth of each physical link is available, bandwidth is not wasted by ineffi-

ficient traffic routing. As a result, the entire fabric is utilized more efficiently.

### **Динамический Выбор Пути (Exchange-Level)**

Dynamic Path Selection (DPS) may also be referred to as exchange-level trunking. Like Advance ISL Trunking, DPS balances traffic across multiple ISLs. Unlike trunking, DPS does not require that the ISLs be adjacent. It uses the industry standard Fabric Shortest Path First (FSPF) algorithm to select the most efficient route for transferring data in multi-switch environments. Any paths which are deemed by FSPF to have equal cost will be evenly balanced by the DPS software and hardware. This is a particular advantage in core/edge networks with multiple core switches, since DPS can distribute load between different cores while Advanced ISL Trunking cannot do so.

DPS matches or outperforms all similar features from any vendor *except* for Brocade Advanced ISL Trunking. However, because DPS can be combined with frame-level trunking, organizations can achieve both maximum performance and availability.

### **Зонирование**

Brocade Zoning is a feature of all switch models. Using zoning, organizations can automatically or dynamically arrange fabric-connected devices into logical groups (zones) across the physical configuration of the fabric. It is functionally similar to VLANs from the IP networking world, though considerably more advanced in many ways. In fact, zones could be thought of as being a combination of VLAN controls plus firewall-like ACLs.

Providing secure access control over fabric resources, Zoning prevents unauthorized data access, simplifies heterogeneous storage management, segregates storage

traffic, maximizes storage capacity, and reduces provisioning time.

The need for this kind of access control relates to the “roots” of SAN technology: the SCSI DAS model. Storage devices directly attached to hosts (DAS) have no need for network-based access control features: access by other hosts is precluded by the limitations of the DAS architecture. In contrast, SANs allow a potentially large number of hosts to access all storage in the network, not just the systems that they are *intended* to access. If each host is allowed to access every storage array, the potential impact of user error, virus infection, or hacker attacks could be immense. To prevent unintended access, it is necessary to provide access control in the network and/or the storage devices themselves.

There are many mechanisms for solving the SAN-based access control problem. All of them have some form of management interface that allows the creation of an access control policy, and some mechanism for enforcing that policy. Brocade switches and routers use a set of methods collectively referred to as “Brocade Advanced Zoning.” Brocade Advanced Zoning requires a license key on all platforms, but all currently shipping platforms bundle this key with the base OS.

Using this key allows the creation of many zones within a fabric, each of which may be comprised of many “zone objects,” which are storage or host PIDs or WWNs. These objects can belong to zero, one, or many zones. This allows the creation of overlapping zones. Every switch in the fabric then enforces access control for its attached nodes. Zone objects are grouped into zones, and zones are grouped into zone configurations. A fabric can have any number of zone configurations. This provides a comprehensive and secure method for defining exactly which devices should or should not be allowed to com -

municate.

## ***Fabric OS CLI***

All Brocade switches provide a comprehensive Command Line Interface (CLI) which enables manual “lowest common denominator” control, as well as task automation through scripting mechanisms via the switch serial port or telnet interfaces.

## ***Wgd'Tqqm***

Brocade WEBTOOLS is web-browser-based Graphical User Interface (GUI) for element and network management of Brocade switches. WEBTOOLS uses a set of processes (e.g. httpd) and web pages that run on all Fabric OS switches in a fabric. Once a switch or router has an IP address configured, it is possible to manage most functions simply by pointing a Java-enabled web browser at that address.

This product simplifies management by enabling administrators to configure, monitor, and manage switch and fabric parameters from a single online access point. Organizations may configure and administer individual ports or switches as well as small SAN fabrics. User name and password login procedures protect against unauthorized actions by limiting access to configuration features. Web Tools provides administrative control point for Brocade Advanced Fabric Services, including Advanced Zoning, ISL Trunking, Advanced Performance Monitoring, Fabric Watch, and Fabric Manager integration. For instance, administrators can utilize time-saving zoning wizards to step them through the zoning process.

While this is technically a licensed feature, like zoning, WEBTOOLS is included with all currently shipping Brocade platforms.

## ***Fabric Manager***

Fabric Manager is a flexible and powerful tool that provides rapid access to critical SAN information and configuration functions. It allows administrators to efficiently configure, monitor, provision, and other perform daily management tasks for multiple fabrics or Meta SANs from a single location. Through this single-point SAN management architecture, Fabric Manager lowers the overall cost of SAN ownership. It is tightly integrated with other Brocade SAN management products, such as Web Tools and Fabric Watch, and enables third-party product integration through built-in menu functions and the Brocade SMI Agent. Organizations can use Fabric Manager in conjunction with other leading SAN and storage resource management applications as the drill-down element manager for a single or multiple Brocade fabrics, or use Fabric Manager as the primary SAN management interface.

## ***SAN Health***

SAN Health is a powerful tool that helps optimize a SAN and track its components in an automated fashion. The tool greatly increases SAN manager productivity, since it automates many mandatory recurring SAN management tasks. It simplifies the process of data collection for audits and change tracking, uses a client/server “expert systems” approach to identify potential issues, and can be run regularly to monitor fabrics over time. This is especially useful to SAN designers in three ways:

- When designing changes to existing environments, the tool can help to audit the target environment before finalizing a design
- In any design context, it can help to document a SAN after implementation
- It can be specified in the SAN project plan as an on-

going proactive maintenance and change-control tool to satisfy manageability requirements

The tool has two software components: a data capture application and a back-end report processing engine. SAN managers may run the data capture application as often as needed. After SAN Health finishes capturing diagnostic data, the back-end reporting process automatically generates a point-in-time snapshot of the SAN, including a Visio topology diagram and a detailed report on the SAN configuration. This report contains summary information about the entire SAN as well as specific details about fabrics, switches, and individual ports. Other useful items in the report include alerts, historical performance graphs, and any recommended changes based on continually updated best practices.

The SAN Health program is powerful and flexible. For example, it is possible to configure many different fabrics in a single audit set, and schedule them to run automatically on a recurring basis. These audits can run in “unattended mode”, with automatic e-mailing of captured data to a designated recipient.

The tool also has enhanced change-tracking features to show how a fabric has evolved over time, or to facilitate troubleshooting if something goes wrong. This can be an invaluable addition to the change-tracking process, both for most-mortem analysis *and* for proactive management. For instance, SAN Health can track traffic pattern changes in weekly or monthly increments. This can help to identify looming performance problems proactively, and take corrective action before end-users are affected.

SAN Health is currently available to SAN end-users and Brocade OEM and reseller channel partners. It can be used with Brocade install-base fabrics, and fabrics using equipment from selected other infrastructure vendors as

well. The tool is available for download on the public Brocade web site ([www.brocade.com/sanhealth](http://www.brocade.com/sanhealth)). For partners, Brocade also provides a co-branded version.

### ***Fabric Watch***

Brocade Fabric Watch provides advanced monitoring capabilities for Brocade products. Fabric Watch enables real-time proactive awareness of the health, performance and security of each switch, and automatically alerts network managers to problems in order to avoid costly failures. Monitoring fabric-wide events, ports, and environmental parameters permits early fault detection and isolation as well as performance measurement.

With Fabric Watch, SAN administrators can select custom fabric elements and alert thresholds or they can choose from a selection of preconfigured settings for gathering valuable health, performance and security metrics. In addition, it is easy to integrate Fabric Watch with enterprise systems management solutions.

By implementing Fabric Watch, storage and network managers can rapidly improve SAN availability and performance without installing new software or system administration tools.

### ***Advanced Performance Monitoring***

Brocade Advanced Performance Monitoring is a comprehensive tool for monitoring the performance of networked storage resources. It enables administrators to monitor both “transmit” and “receive” traffic from source devices to destination devices, enabling end-to-end visibility into the fabric. Using this tool, administrators can quickly identify bottlenecks and optimize fabric configuration resources to compensate.

## ***Extended Fabrics***

Extended Fabrics software enables native Fibre Channel ISLs to span extremely long distances. Extended Fabrics optimizes switch buffering (BB credits) to ensure the highest possible performance on these long-distance ISLs. When Extended Fabrics is installed on gateway switches, the ISLs (E\_Ports) are configured with a large pool of buffer credits. The enhanced switch buffers help ensure that data transfer can occur at full or near-full bandwidth to efficiently utilize the connection over the extended links. As a result, organizations can use Extended Fabrics to implement strategic applications such as wide area data replication, high-speed remote backup, cost-effective remote storage centralization, and business continuance strategies.

## ***Remote Switch***

Remote Switch is a now largely obsolete feature which enabled fabric connectivity of two switches over long distances by supporting external gateways to encapsulate Fibre Channel over ATM. Connecting SAN islands over Fibre Channel-to-ATM device enabled organizations to extend their solutions over a WAN. This type of configuration could be used for solutions such as remote disk mirroring and remote tape backup. While ATM extension may still be used, this method has largely been superseded by FC over SONET/SDH and native FC links using Extended Fabrics. For all such configurations, Brocade now supports an “Open E\_Port” mode to support for Gateway/Bridge devices. Customers may simply use *portCfgISLMMode* CLI command which is now part of the base OS: there is no need for a license anymore.

## ***FICON / CUP***

The Brocade directors and selected switches support the FICON protocol for mainframe environments, ena-

bling organizations to utilize a single platform for both open systems and mainframe storage networks. FICON-certified Brocade platforms support the ability to run both open systems Fibre Channel and FICON traffic on a port-by-port basis within a single platform. The Brocade FICON implementation also supports cascaded FICON fabrics at 1 and 2 Gbit/sec FICON speeds.

With Fabric OS version 4.4, Brocade fully supports CUP in-band management functions, which enable mainframe applications to perform configuration, management, monitoring, and error handling for Brocade directors and switches. CUP support also enables advanced fabric statistics reporting to facilitate more efficient network performance tuning.

### ***Маршрутизация Fibre Channel***

The Brocade FC-FC Routing Service provides connectivity between two or more fabrics without merging them. Any platform it is running on can be referred to as an FC router, or FCR for short. At the time of this writing, the feature is available on the Brocade AP7420, the Brocade 7500, and the FR4-18i blade.

The service allows the creation of Logical Storage Area Networks, or LSANs, which provide connectivity that can span fabrics. It is most useful to think of an LSAN in terms of zoning: an LSAN is a zone that spans fabrics. The fact that an FCR can connect autonomous fabrics without merging them has advantages in terms of change management, network management, scalability, reliability, availability, and serviceability to name just a few areas.

The customer needs for this product are similar to those that brought first routers and then Layer 3 switches to the data networking world. An FC router is to an FC fabric as an IP router is to an Ethernet subnet.

Early efforts were made to create large, flat Ethernet LANs without routers. These efforts hit a ceiling beyond which they could not grow effectively. In many cases, Ethernet broadcast storms would create reliability issues, or it would become impossible to resolve dependencies for change control. Perhaps merging Ethernet networks that grew independently would involve too much effort and risk. An analogous situation exists today with flat Fibre Channel fabrics. Using an FCR with LSANs solves that problem, while other proposed solutions – such as VSANs – just move the problem around in a shell-game effort to confuse users.

For more information about this feature, see the book *Multiprotocol Routing for SANs* by Josh Judd.

## **FCIP**

Fibre Channel over IP (Internet standard RFC 3821) is one of several mechanisms available to extend FC SANs across long distances. FCIP transparently tunnels FC ISLs across an intermediate IP network, making the entire IP MAN or WAN appear to be an ISL from the viewpoint of the fabric. This is available as a fully-integrated feature on the Brocade AP7420 Multiprotocol Router, the Brocade 7500 router, and the FR4-18i blade.

It is important to note that FCIP is neither the only nor always the best approach to distance extension. The major advantages of FCIP are cost and ubiquitous availability of IP MAN and WAN services. However, for users interested in reliability and performance, it is theoretically impossible for FCIP – or any other IP SAN technology for that matter – to match native FC solutions. Generally speaking, SAN designers prefer distance extension solutions in the following order:

1. Native FC over dark fiber or xWDM
2. FC over SONET/SDH

3. FC over ATM
4. FC over IP

Many of the shortcomings of FCIP can be mitigated – though not eliminated – by using FastWrite and/or Tape Pipelining. (p456) In fact, before the advent of FC Fast-Write, it was sometimes even possible to achieve better performance on a 1Gbit FCIP link than a 4Gbit FC link. FCIP should therefore almost always be used in combination with some form of write acceleration technology.

For more information about this feature, see the book *Multiprotocol Routing for SANs* by Josh Judd.

### ***Secure Fabric OS***

As organizations interconnect larger and larger SANs over longer distances and through existing networks, they have an ever greater need to effectively manage their security and policy requirements. To help these organizations improve security, Secure Fabric OS™, a comprehensive security solution for Brocade-based SAN fabrics, provides policy-based security protection for more predictable change management, assured configuration integrity, and reduced risk of downtime. Secure Fabric OS protected the network by using the strongest, enterprise-class security methods available. With its flexible design, Secure Fabric OS allowed organizations to customize SAN security in order to meet specific policy requirements. All Secure Fabric OS features have now been made available in the base OS for free as of Fabric OS 5.3.0. It is recommended that customers migrate to that solution as it provides additional features such as DH-CHAP to end devices (HBAs) and is also more scalable.

## Расчет возврата инвестиций ROI

This section provides guidance on ways to calculate the Return on Investment (ROI) for the SAN project. For a more comprehensive evaluation of the benefits of a SAN, it is better to perform a Total Cost of Ownership (TCO) analysis. However, TCO is harder to calculate, and ROI analysis may be sufficient in many cases, so this is usually where a designer would start.

In fact, even doing a detailed ROI analysis is not needed in most cases. This should be done only if the stakeholders responsible for signing off on the SAN budget have asked for it. For example, if the SAN is being deployed in order to meet a legal requirement for a disaster recovery solution, the implementation is mandatory, so analyzing the financial ROI could be meaningless. After all, if the legal requirement is not met, it could cause an organization-wide disaster, so most stakeholders would agree that the deployment is needed regardless of the financial ROI analysis. Many organizations also put in a SAN based on a total cost of ownership justification, which may not require ROI justification.

For installations which *do* require it, the ROI analysis method below will provide a useful guideline for how to approach the project. It is not intended to be viewed as a hard and fast procedure set, indicating the only “right” way of calculating ROI, but simply as a starting point. In many organizations, there is already an established methodology for ROI calculations, in which case the following guidelines can be mapped into the existing processes.

Some of the sources of SAN ROI include:

- Additional revenue or productivity gains generated during backups that - prior to the SAN - required taking systems off line.

- Similar gains generated through higher average system or application uptime
- Lower IT management costs and increased productivity generated through the centralization of resources.
- Significantly shorter process time for adding and re-configuring storage.
- Reduced capital spending through improved utilization of space on shared storage.

To perform an ROI analysis for a SAN, the following steps can be used:

- Identify the servers and applications which will participate in the SAN. (This should already have been done previously in the planning process. Refer to “Chapter 5: Планирование проекта ” starting on page 149.)
- Select ROI scenarios. These are the primary functions that the SAN is expected to serve, such as storage consolidation or backups.
- Determine the gross business-oriented benefits of this scenario. E.g. how much money will the company save by purchasing fewer storage arrays?
- Determine costs to achieve this benefit. (Again, this should already have been done in a previous step in the planning process.)
- Calculate the net benefits. Essentially, this means subtracting the costs from the benefits.

### ***Цели анализа ROI***

An ROI analysis can focus on specific themes which generally have business relevance. This will help IT organizations demonstrate the financial value of the SAN. The Brocade ROI model clarifies in non-technical terms the benefits of SANs, quantifying the financial benefits to demonstrate real-world ROI. Five key SAN benefit

themes which are often used for ROI analysis are:

- **Improved storage utilization:** SAN-enabled access to enterprise storage will result in economies of scale
- **Improved availability of information:** Enterprises are increasingly relying on information to control costs and improve their competitive advantage. SAN-enabling access to storage (where the information resides) will make that information more available by keeping the systems processing the information running longer. Backups (and restores) will finish quicker in SAN-enabled environments. The result is that mission critical information is at the disposal of the enterprise more of the time.
- **Improved availability of applications:** SAN solutions dramatically reduce application downtime – both scheduled and unscheduled. Global enterprises can profit from the extra availability.
- **More effective storage management:** SAN-based solutions are easier to manage because they tend to be centralized. Centralization translates to increased operational control and management efficiencies. These are directly related to cost reductions.
- **Foundation for disaster tolerance:** Certain elements of SAN-enabled solutions create the opportunity for improved disaster tolerance as a by-product of the architecture. Examples include remote backups, disk-to-disk-to-tape backups, data mirroring or replication, and inter-site applications failovers.

## *Анализ шаг 1: идентификация узлов и приложений*

The first step is to define important servers, their applications, and their associated storage. This should have been done during the requirements gathering phase of the SAN planning process. Then group them according to the role they play. For example, an organization might have

back-end database servers, front-end application servers, email servers, web servers, and servers hosting network file systems such as NFS or CIFS.

Using data from the inventory of existing equipment, define groups of servers performing similar tasks. For each server-group, define the average amount of direct-attached storage they currently have configured. Also define for each server-group how fast their storage capacity is growing and how much space they need to leave uncopied on storage arrays to grow into for a given year. (I.e. how much headroom each requires.) Also define the availability requirements for each server group, if you have not already done so.

### ***Анализ шаг 2: выбор сценариев***

In the beginning of this chapter we discussed the business requirements of the SAN. The requirements define a set of ROI scenarios. This next section illustrates how to process three common scenarios: Storage consolidation, backup and restore, and high availability clustering. (These and other scenarios are discussed in “Chapter 2: Решения SAN” starting on page 61.) In your own analysis, include *all* business-oriented benefits which the SAN will provide.

#### Консолидация хранения

The goal of this scenario is to migrate from traditional Directly Attached Storage (DAS) to SAN-based storage. Two benefits to consider are (1) reduced need for storage headroom (a.k.a. “white space”), and (2) reduced downtime associated with storage adds, moves, and changes. See “Консолидация хранения” starting on page 61 for a description of this scenario.

Резервное копирование/восстановление

This scenario addresses backup and restore savings opportunities based on performance. It is assumed that an existing enterprise network-based distributed backup/restore facility is already in place, e.g. sending backup data to a tape server via a LAN. If that is *not* true, then the ROI will be greater. See “Консолидация ленточных накопителей / резервное копирование без использования LAN” starting on page 72 for a description of this scenario.

Кластеры высокой готовности

High Availability (HA) clustering is a method of improving the availability of applications. Normally in HA configurations, a standby server stands by ready to “step in” for a failing production server. If the production server fails, the applications are transferred to the standby server through partially or totally automated means. In addition to protecting against failures, HA clusters can be used to reduce planned downtime for upgrades or changes to a server hardware platform. In this case, an administrator would manually trigger an application failover (usually called a “switchover” in this context) to the standby server, perform maintenance on the primary, and then manually move the application back once the maintenance was complete and verified.

Most HA configurations have a dedicated standby server for every production server they are protecting. One reason for this is the inability to attach more than two computers to external SCSI disk arrays. The resulting 1:1 ratio of primary to hot standby servers means a very costly HA facility, which – in practice – means that most applications are not included in HA clusters, and are therefore exposed to outages during failures or planned hardware maintenance operations. See “Кластеры

высокой ” starting on page 66 for a more comprehensive discussion of this topic.

### ***Анализ шаг 3: определение преимуществ сценариев***

Once you have decided which scenarios apply to your SAN by looking at the business problems which it will address, it is time to calculate the benefits of those scenarios. When calculating ROI, benefits are commonly divided into two types: hard benefits, and soft benefits.

“Hard” benefits include any benefits for which a specific monetary savings or revenue increase can be identified with a high degree of confidence. For example, it is often relatively easy to assign specific values to reduced capital expenditures, operational budget savings, and gains through some kinds of staff productivity increases.

“Soft” benefits include items for which specific monetary savings are more difficult to define. One typical example is opportunity costs. It may be difficult to assign an exact value to the opportunity cost of degraded performance, system downtime for repairs, lengthy backup windows, or lengthy data restoration times. The characterization of a benefit as “soft” does not imply that it is less important; just that it is harder to prove exactly how much money it is worth.

Remember while reading the remainder of this section that each of the benefits listed below can be classified as either hard or soft. Also remember that *costs* will be calculated in a subsequent step; this section is only about *benefits*.

#### Консолидация хранения

Benefits of storage consolidation can be calculated by evaluating the savings of eliminating unused white

space on storage (a.k.a. excess headroom), which is a “hard” benefit, and the savings obtained by the elimination of some of the downtime associated with upgrading server-attached storage, which is usually a “soft” benefit.

Headroom savings are *deferred* savings, which means that the organization will get benefits in the future, and will continue to get the benefits perpetually instead of merely having a one-time savings. If the overall storage capacity keeps expanding in an organization, so will the requirement for storage headroom. Of course, this is true of both SAN and DAS environments. The difference is that the demand for storage headroom will always be proportionally lower in a SAN. So as long as the need for storage grows over time, the benefits of the SAN will keep growing, too.

The benefit of reduced downtime includes the savings obtained by eliminating much of the downtime associated with upgrading storage. If an administrator adds a new storage array to a SAN, configuring servers to access it can be completely non-disruptive, and much of the configuration can be performed by management software. Adding storage in a DAS environment usually requires rebooting or even disassembling servers, which is costly in administrative time as well as causing an application outage.



### Side Note

*It is possible to achieve ROI through improved management of storage, or through economies of scale in purchasing power achieved by using few large arrays instead of many small units.*

Here is an example of how storage consolidation ROI might be discussed in the SAN project planning document:

**Комментарии по ROI консолидации хранения**

*In our current environment, we have 60% unused space on our storage arrays, on average. This ranges from 1% free space on some arrays, to over 95% free space on others. I estimate that we will need to spend \$x to purchase new arrays over the next year, if we continue to use directly attached storage. This is because the servers currently at the “1% free” end of the spectrum will need to grow their storage pool, but cannot access the arrays attached to the servers at the “95% free” end. I.e. we have plenty of free space, but no way to get the servers which need it to the arrays which have it. By putting in a SAN, we should be able to avoid all of the new array purchases this year, and for most of next year as well. This means that we will directly save more than \$x through implementing a SAN.*

*In addition, the SAN will increase the uptime of each server. Today, each time a server runs out of space, we need to schedule an outage to add another disk, controller, or array. In some cases, this is no problem, but in others, it is extremely disruptive to our business. For example, the manufacturing line relies on several of the servers which are currently almost out of space. It may be necessary to shut down the line to add more disk. Shutting down the line costs \$y per hour. Last year, we had to take four hours of manufacturing line outages for storage upgrades, and next year is projected to be even higher. Therefore we will save in excess of \$4y per year in downtime by putting in the SAN.*

**Total First Year Benefit:** \$x due to reduced array purchases because of white space optimization, plus \$4y from reduced downtime on the manufacturing line.

An ROI benefit expressed as a dialogue such as the one above will often be translated into another form to satisfy an accountant. This is often just a spread-

sheet, with little or no supporting text. However, it is usually not the responsibility of the SAN designer to do this translation. Rather, the technical team would normally provide this kind of dialogue to an accounting department member.

#### Резервное копирование/восстановление

The backup scenario contracts the backup window, thus reducing amount of time the servers are unavailable or have degraded performance because their data is being backed up. Shrinking the backup window creates savings for the organization through increased productivity, whether or not the applications need to be taken off line. Even if they are still online during the operation, performance is often degraded quite a bit. This is often a “soft” benefit, though it might be quantifiable for mission-critical applications.

In addition to speeding up backups, a SAN will speed up restore operations. A restore will occur when data is lost or corrupted, and in most cases, operations at the organization will be disrupted while waiting for this to complete. The ROI to an organization for improved restore time is the reduced opportunity cost of being unable to operate between the time of a data loss and the final restoration of data. Typically, the metrics for quantifying this will involve productivity decreases and lost revenue during the outage.

In many cases, it is easy to determine the cost of an outage to a system. The previous scenario gave the example of a manufacturing line, which had a defined cost of downtime. However, in that example, the SAN project manager had a good idea of how many outages could be avoided. By looking at historical growth for storage array data, it is possible to make defensible projections about future growth. This told the SAN project manager which arrays were likely to run out of space. It is harder to pre-

dict which systems will have corrupted filesystems, or in which cases user error will require a restoration. Avoidance of unplanned downtime has to be calculated based on statistical probabilities: what is the percentage chance that a restoration will need to happen on any given server? How long is that likely to take without a SAN? How long will it take *with* a SAN? Once you know how much time a SAN would save in restoring from a *hypothetical* downtime event, and how much per hour uptime of the system is worth, you multiply the savings times the probability of the event occurring to get the benefit of the shorter restore time.

This example calculates the savings realized through improved backup and restore performance alone. Another possibility is consolidating many small tape drives onto fewer larger libraries. This can create a significant economy of scale when buying new tape libraries, and can reduce management costs as well. Yet another way to achieve backup savings via a SAN is to consolidate white space on tapes, in much the same way that the previous scenario consolidated space on disk drives. Each tape in a backup set is only partially used. Depending on the backup software used, it may be possible to put backups from multiple servers onto a single tape, thus filling it more completely. This is generally *not* possible with DAS tape solutions. Over time, the savings achieved by using up fewer tapes could be significant.

For example, take the manufacturing line SAN again. That SAN might be performing backups as well as consolidating storage arrays. The SAN project manager might make an entry in the planning document like this:

### ***Комментарии по ROI резервного копирования на основе SAN***

*The manufacturing line has to run backups once a day.*

*When we do this, the server response time drops*

*by 50%, and as a result, the line runs 50% slower. That window currently lasts one hour. 50% performance degradation for one hour on the line costs at least \$x in lost revenue. The SAN will reduce that window to six minutes, or 90% of the window. In addition, the SAN-enabled software is more efficient, and will lower the performance impact to the application during the remaining window, though it will not be possible to quantify that until implementation time. This means that we will directly save more than **0.9 times \$x** through implementing a SAN.*

*In addition, using centralized tape libraries will allow us to compress white space out of backup tapes. Currently, our average tape utilization is 50%. With the SAN, our utilization will reach increase enough to use 10% fewer tapes. We currently spend \$y per month on tapes, so the SAN will save **0.1 times \$y** each month. Since our storage needs increase over time, this benefit will increase as the SAN ages.*

*Finally, the SAN will reduce downtime during data restorations. Last year, the manufacturing line had two hours of downtime for restores. If we assume that the same things will happen next year, the higher performance of SAN-enabled restorations will reduce the restoration time by 25% or more. Total downtime for the line costs \$z per hour, so the SAN will save an estimated **0.75 times 2 times \$z**. Another way to estimate the potential for needing to restore data is to look at the overall odds of a failure occurring. By taking the mean time between failures (MTBF) and mean time to repair (MTTR) of all components in the manufacturing systems into account, I estimate the probability being 50% that we will have four hours of downtime due to component failures. A 50/50 chance of four hours of downtime means that avoidance of the risk is worth 50% of the cost of the outage. This is **0.5 times 0.75 times 4 times***

$\$z$ , which reduces down to the same equation as the two hour estimate above.

**Total First Year Benefits:** 0.9 times  $\$x$  due to increased productivity on the manufacturing line from backup window reduction, plus 0.1 times 12 times  $\$y$  from reduced monthly tape consumption, plus an estimated 0.75 times 2 times  $\$z$  from hypothetical reduced restoration times.

In general, the cost of downtime will vary by server class. The example above showed only one class of server: the platform's running applications critical to the manufacturing line. In most large-scale SANs, there will be more than one class of server attached. For example, the SAN might connect both the manufacturing servers above, and also the corporation's email servers and file-servers. Each class of server should have its own separate valuation for uptime. Even servers for which "hard" uptime numbers are unavailable should be included in the ROI analysis as a "soft" benefit, with an indication that the exact financial value is unknown.

### Кластеры высокой готовности

SANs and complementary software products eliminate many of the restrictions traditionally associated with HA solutions, and allow solutions based on clusters of more than two servers. Larger clusters allow for a single platform to serve as a standby for more than one primary server. (I.e. SANs allow an  $n:1$  ratio of primary to hot standby servers.) This dramatically reduces the cost of protecting applications. This is illustrated in Рис. 17 and Рис. 18 starting on page 67.

In addition to achieving ROI through lowered cost of protecting mission-critical applications, SANs also expands the number of applications which can be cost-justified to participate in an HA solution. That means that

an organization can achieve ROI through increased uptime and associated productivity / revenue gains for the services which would otherwise not have been protected.

The benefits of HA clustering can be found by calculating the savings on both planned and unplanned downtime for all protected server classes, and the savings on equipment obtained by implementing  $n:1$  HA cluster instead of using a 1:1 primary/standby design. Additional savings can be calculated by accounting for the reduced maintenance cost for all protected server classes over a year. I.e. having fewer total platforms in the solution means buying maintenance contracts on fewer machines, and – statistically – reducing the number of repairs needed.

Once again, take the manufacturing line SAN as an example. There could be four critical applications required to support the line, one of which (application number “a4”) spans two platforms. An outage to either platform causes an outage to a4. The project manager might make an entry in the planning document like this:

### ***Комментарии по ROI кластеров высокой готовности на основе SAN***

*The manufacturing line has four critical applications: a1, a2, a3, and a4. The value of protecting these applications via an HA solution is increased uptime for the manufacturing operation. Previously, the value of uptime for the line was shown to be \$x per hour. Last year, we had two outages to the line caused by failures of these applications, which could have been avoided by clustering them. The total avoidable downtime to repair them was four hours. Assuming that the same events occurred over the next year, the avoided cost would be 4 times \$x. Using the MTBF and MTTR of all related components to calculate the statistical probability of a failure in the line shows that there is a 25% chance of*

*eight hours of avoidable downtime. **0.25 times 8 times \$x** reduces to **2 times \$x**, which is lower than the previous estimate. We will use midpoint for this analysis, and say that HA protection will likely save the company more than **3 times \$x** in downtime for each year of operation. This is a conservative estimate, particularly since our business is growing. This means that the number of servers requiring protection will increase, which will increase the likelihood of an avoidable failure and the cost of failures not avoided, so the benefit of clustering will increase substantially in subsequent years. In addition to unplanned failures, we had to take four hours of planned downtime last year, and expect the same for next year. Half of that would be avoidable with a cluster, so the total downtime reduction is **5 times \$x** for both planned and unplanned downtime.*

*There are two approaches to building this HA solution: we can dedicate a hot standby server for each application platform, or we can use a SAN to allow one standby platform to protect all of the production servers.*

*The a4 application spans two hosts, so there are a total of five servers which need to be protected. This will require five standby servers in the first method, or just one in the second. The difference is four extra platforms, vs. installing a SAN. Accounting for software package and operating system licenses, maintenance contracts, and projected staff time for performing maintenance, each extra server costs \$y, so the SAN will save **4 times \$y** on hardware, software, and maintenance.*

*This benefit will accelerate with time. The clustering package we propose to use allows up to z platforms to be protected by a single hot standby server, so we will include several lower-tier applications in the cluster as well. This will still leave room for projected increases in*

*the number of manufacturing line servers required for the next year, so we can add to the cluster without increasing its cost.*

**Total First Year Benefits:** 5 times \$x due to increased productivity on the manufacturing line from downtime reduction, plus 4 times \$y from reduced hardware, software, and maintenance cost. We would also receive “soft” benefits from having lower-tier applications protected by the cluster.

It is also worth mentioning that one SAN can support many clusters. The benefits of protecting the manufacturing line might easily justify the cost of the SAN by themselves, but whether they do or not, it would often be possible to connect other mission-critical hosts to the same SAN even if they are in a different cluster or even if they use a completely different kind of clustering software. While evaluating SAN ROI, look at all of the applications which could benefit from SAN attachment, whether or not they are the immediate focus of the project.

### Комбинированные решения

This brings up the topic of combined SAN solutions. Historically, almost all SANs were built as application-specific islands. However, today’s SANs are increasingly heterogeneous, with one SAN supporting not just different applications, but indeed supporting hardware and software from different vendors. The SAN used in the preceding examples could support a storage consolidation solution, a tape backup / restore solution, and an HA clustering solution. The benefits of the SAN would come from all three use cases, but the cost of the infrastructure would only need to be paid once. Always look for other applications which could benefit from SAN attachment even if one particular application is “driving” the project. To the extent that any can be identified, see if they can be

quantified as “hard” benefits. In most cases, even if the SAN is initially envisioned only to host one application, over time more and more uses will inevitably come to light. Even if there are no initial plans to include other uses, it is appropriate to include some discussion of this principle in the ROI analysis as a “soft” benefit.

#### *Анализ шаг 4: Определение сопряженных расходов*

The next step is to determine the costs required to achieve the benefits. As this cost determination is being done in the early stages the cost used will be preliminary estimates. If you are following the overall SAN project plan discussed in this chapter, you will already have a good idea about the costs of the project at this stage. If this is the case, utilize this information and proceed to the step five on page 487.

If you do *not* already have a cost estimate, you will need to make one. To create an estimate, the top-level SAN architecture must be defined. The architecture need not be correct in every detail: for a SAN of any complexity, it will have to be refined as the project progresses. It only needs to be sufficient for budgetary purposes, which means knowing more or less how many ports you will need to buy, and their HA characteristics.

Create an estimate of costs for each scenario. Treat each scenario independently and create discrete ROI calculations for each. This will allow you to determine the most effective strategy for justifying the SAN infrastructure. However, this means that the analysis will contain duplicated elements. For example, one switch port used for an ISL will support traffic from a backup solution, a storage consolidation solution, and an HA cluster solution. Therefore you should present an aggregate ROI as opposed to the sum of the individual ROI analysis numbers.

bers to show the real cost savings.

### ***Анализ шаг 5: подсчет ROI***

In step three, you showed the *gross* benefits of a SAN. I.e. you showed how much money the SAN would save or help to produce, but did not take into account the costs to achieve those benefits. In this step, you will produce an estimate of the *net* benefit that the SAN will deliver: the benefits minus the costs.

There are a number of ways to calculate ROI. Two of the most common methods are Internal Rate of Return (IRR) and Net Present Value (NPV). Here are commonly used “accountant” definitions of the two methods:

**IRR:** The discount rate to equate the project’s present value of inflows to present value of investment costs.

**NPV:** The sum of a project’s discounted net cash flows (present values including inflow and outflows, discounted at the project’s cost of capital).

What do you actually *do* to calculate either of those? One answer is, “get an accountant to do it.” In fact, most organizations have a preferred method for performing an ROI calculation, and have accounting departments which would insist on being the ones to perform the analysis in any case, so this is the answer that most SAN designers will use.

However, it is sometimes useful for the SAN project team to estimate the ROI of the project before discussing it with accounting. To do a rough ROI estimate, simply subtract any identified costs from any quantified benefits. In the example used throughout the previous sections, the manufacturing line would receive benefits from three different sources. Add all three up to get a total first-year figure. Then add up the costs of the project as estimated in previous steps. Subtract the second number from the first,

and that is how much “hard” benefit the SAN will provide in the first year of operations. An accountant would also need to take equipment depreciation into account, and might look at ROI over a longer timeframe, but this should at least give the SAN design team an idea of how the ROI analysis will come out.

The key to ROI is to be sure you have identified and accounted for all of the benefits. Many things in life tend to have hidden costs – such as the maintenance problems associated with buying a used car. However, some things also have hidden benefits – such as the reduction in administrative overhead inherently associated with implementing a SAN. As long as the ROI analysis includes all costs and all benefits – both hard and soft – it will give you a good idea about whether or not a SAN is right for your organization.

## Оборудование Ethernet и IP сетей

This section does not provide a comprehensive tutorial on Ethernet or IP equipment. Nor is it intended to supplement the manuals for those products. It is simply a high-level discussion of how such equipment relates to the Brocade AP7420 Multiprotocol Router, and other Brocade platforms.

### *Краевые коммутаторы и концентраторы Ethernet L2*

It is possible to use commodity 10/100baseT hubs and/or switches to attach to the Ethernet management ports of an FC switch or router. It is *not* recommended to use hubs for data links to iSCSI hosts or for FCIP connections, since performance on hubs is rarely sufficient for even minimal SAN functionality.

When connecting to iSCSI hosts, it is *possible* to use accelerated Gigabit Ethernet NICs with optical

transceivers to connect hosts directly to the router. However, this is not *recommended*: this approach has much higher cost and much lower performance than attaching the host to a Fibre Channel switch using a Fibre Channel HBA. The value proposition of iSCSI vs. Fibre Channel only works if the low-end hosts are attached via already existing software-driven NICs to a low-cost Ethernet edge switch. Many iSCSI hosts then share the same router interface. There are many vendors who supply Ethernet edge switches. Figure 110 shows an example from Foundry Networks. (<http://www.foundrynetworks.com>)



Figure 110 - Foundry EdgeIron 24 GigE Edge Switch

### Маршрутизаторы IP WAN

When connecting to a WAN in an FCIP solution, it is usually necessary to use one or more IP WAN routers. These devices generally have one or more Gigabit Ethernet LAN ports and one or more WAN interfaces, running protocols such as SONET/SDH, frame relay, or ATM. They almost always support one or more IP routing protocols like OSPF and RIP. Packet-by-packet path selection decisions are made at layer 3 (IP).

Figure 111 (p490) shows an IP WAN router from Tasman Networks. (<http://www.tasmannetworks.com>) There are many other vendors who supply IP WAN routers, such as Foundry Networks (Figure 112).

Make sure that the WAN router and service are both appropriate for the application. Two considerations to keep in mind when selecting a WAN router for SAN extension are performance and reliability. Most WAN

technologies were not intended for either the performance or reliability needs of SANs.

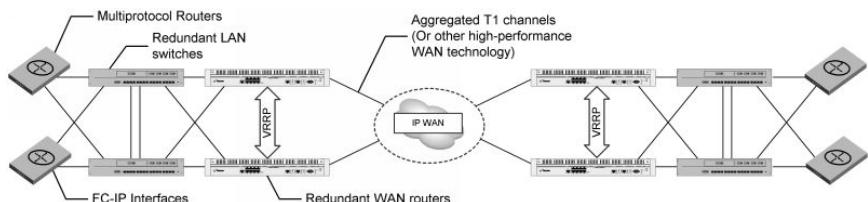


**Figure 111 - Tasman Networks WAN Router**



**Figure 112 - Foundry Modular Router**

Finally, for redundant deployments it is strongly desirable for a WAN router to support a method such as the IEEE standard VRRP. Such methods can allow redundantly deployed routers to fail over to each other and load balance WAN links while both are online. Figure 113 shows one way that an IP WAN router might be used in combination with the Multiprotocol Router.



**Figure 113 - WAN Router Usage Example**

In this example, there are two sites connected across a WAN using FCIP. The Multiprotocol Routers each have two FCIP interfaces attached to enterprise-class Ethernet

switches. These are connected redundantly to a pair of WAN routers, which are running VRRP.

### **Конверторы Gigabit Ethernet медь-оптика**

Some IT organizations supply Gigabit Ethernet connections using copper 1000baseT instead of 1000baseSX or LX. To connect copper Ethernet ports directly to optical FCIP or iSCSI ports - e.g. on a Brocade AP7420 - is not possible. One solution is to use a Gigabit Ethernet switch with both copper and optical ports, attaching the router to the optical ports and the IT network to the copper ports. A product such as the Foundry switch shown in Figure 110 (p 489) could be used in this manner. Alternatively, a media converter (sometimes called a “MIA”) can be used. There are a number of vendors who supply such converters. TC Communications is one example. ([www.tccomm.com](http://www.tccomm.com))



**Figure 114 - Copper to Optical Converter**

**B**

## **Приложение В: Расширенные материалы**

This chapter provides advanced material for readers who need the greatest possible in-depth understanding of Brocade products and the underlying technology. It is not necessary for the vast majority of Brocade users to have this information. It is provided for advanced users who are curious, for systems engineers who occasionally need to troubleshoot very complex problems, and for OEM personnel who need to work with Brocade on new product development.

### **Протоколы маршрутизации**

This subsection is intended to clarify the uses for the different routing protocols associated with the multiprotocol router, and how each works at a high level. Broadly, there are three categories of routing protocol used: intra-fabric routing, inter-fabric routing, and IP routing. The router uses different protocols for each of those functions.

To get from one end of a Meta SAN to another may require all three protocol groups acting in concert. For example, in a disaster tolerance solution, the router may connect to a production fabric with FSPF, use OSPF to connect to a WAN running other IP routing protocols, and run FCRP within the IP tunnel.

## **FSPF: маршрутизация внутри фабрики**

Fabric Shortest Path First (FSPF) is a routing protocol designed to select paths between different switches within the same fabric. It was authored by Brocade and subsequently became the FC standard intra-fabric routing mechanism.<sup>114 115</sup>

FSPF Version 1 was released in March of 1997. In May of 1998 Version 2 was released, and has completely replaced Version 1 in the installed base. It is a link-state path selection protocol. FSPF represents an evolution of the principles used in IP and other link-state protocols (such as PNNI for ATM), providing much faster convergence times and optimizations specific to the stringent requirements of storage networks.

The protocol tracks link states on all switches in a fabric. It associates a cost with each link and computes paths from each port on each switch to all the other switches in the fabric. Path selection involves adding the costs of all links traversed and choosing lowest cost path. The collection of link states (including cost) of all the switches in a fabric constitutes the topology database.

FSPF has four major components:

- The FSPF hello protocol, used to identify and to establish connectivity with neighbor switches. This also exchanges parameters and capabilities.
- The distributed fabric topology database and the protocols and mechanisms to keep the databases synchronized between switches throughout a fabric
- The path computation algorithm

---

<sup>114</sup> Much of the content in this subsection was adapted from “Fabric Shortest Path First (FSPF) v0.2” by Ezio Valdevit.

<sup>115</sup> This and other Fibre Channel standards can be found on the ANSI T11 website, <http://www.t11.org>.

- The routing table update mechanisms

The first two items must be implemented in a specific manner for interoperability between switches. The last two are allowed to be vendor-unique.

The Brocade implementation of FSPF allows user-settable static routes in addition to automatic configuration. Other options include Dynamic Load Sharing (DLS) and In-Order Delivery (IOD). These affect the behavior of a switch during route recalculation, as, for example, during a fabric reconfiguration.

This feature works in concert with Brocade frame-by-frame trunking mechanisms. Each trunk group balances traffic evenly on a frame-by-frame basis, while FSPF balances routes between different equal-cost trunk groups.

The Brocade Multiprotocol Router further enhances FSPF by providing an optionally licensed exchange-based dynamic routing method that balances traffic between equal cost routes on an OX\_ID basis. (OX\_ID is the field within a Fibre Channel frame that uniquely defines the exchange between a source and destination node.) While this method does not provide as even a balance as frame-by-frame trunking, it is more even than DLS.

### ***FCRP: маршрутизация между фабриками***

The Fibre Channel Router Protocol (FCRP) is used for routing between different fabrics. It was designed to select paths between different FC Routers on a backbone fabric, to coordinate the use of multiple domains and LSAN zoning information, and to ensure that exported devices are presented consistently by all routers with EX\_Ports into a given edge fabric. Like FSPF, this protocol was authored by Brocade. At the time of this writing it is in the process of being offered to the appropriate standards bodies. (T11)

Within FCRP, there are two sub-protocols: FCRP Edge and FCRP Backbone.

The FCRP Edge protocol first searches the edge fabric for other EX\_Ports. If it finds one or more, it communicates with them to determine what other fabrics (FIDs) the routers have access to, and to determine the overall Meta SAN topology. It checks the Meta SAN topology, looking for duplicate FIDs and other invalid configurations. Assuming that the topology is valid, the routers hold an election to determine ownership of xlate phantom domains for FIDs that they have in common.

For example, if several routers with EX\_Ports into the FID 1 fabric each have access to FID 5, one and only one of them will “own” the definition of network address translation to FID 1 from FID 5. This router will request a domain ID from the fabric controller for the xlate domain intended to represent FID 5, and will assign PIDs under that domain for any devices in LSANs going from FID 5 to FID 1. All of the other routers with FID 5 to FID 1 paths will coordinate with the owner router and will present the xlate domain in exactly the same way. If the owner router goes down or loses its path to FID 5, another election will be held, but the new owner must continue to present the translation in the same way as the previous owner. (In fact, all routers save all translation mappings to non-volatile memory and even export the mappings if their configurations are saved to a host.)

Note that the owner of the FID 5 to FID 1 mapping does *not* need to be the same as the owner of e.g. the FID 4 to FID 1 mapping. Each xlate domain could potentially have a different owner.

It is important to stress that the Fibre Channel standard FSPF protocol works in conjunction with FCRP. Existing Fibre Channel switches can use FSPF to coordinate with and determine paths to the phantom domains projected by the

router, but only because FCRP makes the phantom domain presentation consistent.

On the backbone fabric, FCRP operates using ILS 0x44. It has a similar but subtly different set of tasks. It still discovers all other FC Router's on the backbone fabric, but instead of operating between EX\_Ports it operates between domain controllers. For each other FCR found, a router will discover all of its NR\_Ports and the FIDs that they represent, each of which yields a path to a remote fabric. It will determine the FCRP cost of each path. Finally, it will transfer LSAN zoning and device state information to each other router.

When the initial inter-fabric route database creation is complete, routers will be consistently presenting EX\_Ports with xlate domains into all edge fabrics, each with phantom devices for the appropriate LSAN members. Into the backbone fabrics, routers will present one NR\_Port for each EX\_Port. This is another situation in which FCRP and FSPF work together: FCRP allows the NR\_Ports to be set up and their activities coordinated. Once traffic starts to flow across the backbone, it will flow between NR\_Ports. FSPF controls the path selection on the standard switches that make up the backbone.



### Side Note

*Not only FSPF and FCRP are complementary. On an FCIP connection in a Meta SAN, all routing protocol types plus layer 2 protocols like trunking and STP can apply to a single connection. STP works outside the tunnel on LANs between FCIP gateways and WAN routers, IP protocols like OSPF work through the WAN outside the tunnel, FSPF operates at the standard FC level inside the tunneled backbone fabric, and FCRP operates above FSPF but still within the tunnel.*

## FCR форматы заголовка фрейма

The FC-FC Routing Service defines two new frame headers: an encapsulation header and an Inter-Fabric Addressing (IFA) header. These are used to pass frames between N\_R\_Ports of routers on a backbone fabric. These extra headers are inserted by the ingress EX\_Port and interpreted and removed by the egress EX\_Port.

The format for these headers going to be submitted for review in the T11 FC Expansion Study Group and is subject to change. Since frame handling is performed by a programmable portion of the port ASIC on router platforms, header format changes can be accommodated without hardware changes.

The Inter-Fabric Addressing Header (IFA) provides routers with information used for routing and address translation. The encapsulation header is used to wrap the IFA header and data frame while it traverses a backbone fabric. This header is formatted exactly like a normal FC-FS standard header, so an encapsulated frame is indistinguishable from a standard frame to switches on the backbone. This ensures that the router is compatible with existing switches, unlike proprietary tagging schemes proposed by other vendors.

## Механизм реализации зонирования

This subsection discusses three different enforcement mechanisms used in zoning, including when each is used, and what the significance is in each case. For a high level discussion of zoning, see “Зонирование” on p461.

### “Программное зонирование” – реализация SNS

When an HBA logs into a Fibre Channel fabric, it queries the name server to determine the fabric addresses all storage

devices. The most basic form of zoning is to limit what the name server tells a host in response to this inquiry. Hosts cannot access storage devices without knowing their addresses, and the SNS<sup>116</sup> inquiry is the only way they should have of obtaining that information. If the name server simply does not tell a host about any storage devices other than the ones it is allowed to access, then it will never try to violate the access control policy.

SNS zoning works well unless the HBA driver is defective in a significant and specific way and/or the host is under control of a *very* skilled attacker. It does rely on each host to be a “good citizen” of the network, but in most cases this is a safe assumption.

SNS zoning is always used in Brocade SANs if zoning is enabled at all, but it is *always* supplemented by one or both of the two “hardware” methods below.<sup>117</sup>

### **“Полное аппаратное зонирование” – фильтрация каждого фрейма**

In the per-frame hardware zoning method, switches program a table in destination ASIC ports with all devices allowed to send traffic to that port. This is in addition to SNS zoning, not instead of it.

For example, if the access control policy for a fabric allows a host to “talk” to a storage device, then the ASIC to

---

<sup>116</sup> Both Fibre Channel and iSCSI support automatic device discovery through a name server. In Fibre Channel, the service is known as the “Storage Name Server,” or SNS. In iSCSI, it is known as the iSNS. This subsection discusses FC SNS zoning, but a similar mechanism works with the iSNS.

<sup>117</sup> The exceptions are the SilkWorm 1xx0 or 2xx0 series switches. The SilkWorm 1xx0 switches did not support hardware zoning at all, and the SilkWorm 2xx0 switches only supported hardware zoning for policies defined by PID, not by WWN.<sup>117</sup> All 200, 3xx0, 4xx0, 12000, 24000, and 48000 products support one or both hardware zoning methods in all usage cases. In other words, all Brocade switches shipped in this century.

which the storage is attached will be programmed with a table entry for that host. It will drop any frame that does not match an address in the table.<sup>118</sup> This method is very secure. Even if a host tries to access a device that the SAN does not tell it about (extremely rare but theoretically possible) hardware zoning will prevent frames from that host from reaching the storage port.

However, in very large configurations it is possible to exceed the table size for a destination port.<sup>119</sup> If this happens on a particular storage port, the per-frame hardware zoning method will usually still be in force on the host port, which is sufficient to prevent access. Even if all ports in a fabric were to exceed zoning table size limitations (*highly unlikely*) all now-shipping Brocade switches can fall back to the “Session Hardware Zoning” method.

Another limitation on hardware zoning is related to WWN zoning vs. “Domain, Port”, or *PID* zoning. In the older “Loom” switches, WWN zones were software enforced, and only PID zones would be enforced by hardware. With all currently shipping switches, full hardware enforcement is available whether using WWN or PID zoning definitions, but only for zones that contain WWNs *or* PIDs. If a single zone uses both WWNs *and* PIDs, that zone will use session hardware zoning.

### ***“Сессионное аппаратное зонирование” – ловушки команд***

If the fabric access control policy results in a zoning table larger than a destination ASIC can support, or if a zone contains both WWNs and PIDs, then some ports on the af-

---

<sup>118</sup> Note that there is no performance penalty for hard zoning with Brocade ASICs.

<sup>119</sup> Each generation of Brocade ASIC has improved the zoning subsystem, but it is never possible to support “infinitely large” tables within an ASIC.

fected chip(s) will use the second hardware zoning method. In addition to SNS enforcement, certain command frames (e.g. PLOGI) will be trapped by the port hardware and filtered by the platform control processor.

This is effectively like the previous method, except that hardware filtering is not done on *all data frames*, which is why it is called *session* hardware zoning. This works because Fibre Channel nodes require command frames to allow communication: data frames sent without command frames will be ignored by destination devices. For example, if a host cannot PLOGI into a storage device, the storage should not accept data from the host since PLOGI is needed to setup a session context in the storage controller.<sup>120</sup> Any frames that managed to get past both SNS zoning and hardware-based session command filtering should be dropped by the destination node.

Since this is based on a category of frame rather than a device address, there is no theoretical limit to the number of devices supportable with this method, short of the main system memory and CPU resources on the platform CP. Since the trap is implemented in hardware, it is still secure and efficient.

## Протоколы и стандарты FC

All Brocade products adhere to applicable standards wherever possible. In some cases, there may not be a ratified standard. For example, there is no standard for upper-level FC-FC routing protocols at this time, so Brocade created

---

<sup>120</sup> This is effective unless the storage device has a serious driver defect. That small chance is the main reason why Brocade implements “full” hardware zoning whenever possible, but as a practical matter the “command” version works fine. There has never been a reported case of an initiator accessing a storage device protected by “command” zoning, even in a lab environment in which experts were trying to achieve that effect.

FCRP in much the same way that Brocade created FSPF when there was a vacuum in the standards for switch to switch routing. Brocade has in fact either authored or co-authored essentially every standard used in the Fibre Channel marketplace. While Brocade tends to offer such protocols to the standards bodies, there is no guarantee that they will be adopted by competitors.

Some of the applicable standards include FC-SW-x, FC-FLA, FC-AL-x, FC-GS-4, FC-MI-2, FC-DA, FCP-x, FC-FS, and FC-PI-x.

For more information on these and other Fibre Channel standards, visit the ANSI T11 website, [www.t11.org](http://www.t11.org).



### **Side Note**

*Gigabit Ethernet was created by “bolting on” some of the existing Ethernet standards on top of 1Gbit FC layers. Few IP network engineers realize it, but all optical Gigabit Ethernet devices still use Fibre Channel technology today.*

## **Brocade ASICs**

Brocade adds value as a SAN infrastructure manufacturer by developing custom software and hardware. Much of the hardware value-add comes from the development of Application-Specific Integrated Circuits (ASICs) optimized for the stringent performance and reliability requirements of the SAN market.<sup>121</sup> Brocade has been building best-in-class custom silicon for SAN infrastructure equipment since 1995. This also enables greater software value-add, since custom

---

<sup>121</sup> ASICs are customized microchips designed to perform a particular function very well. Brocade uses ASICs developed in-house as opposed to using generic “off the shelf” technology designed to perform different tasks such as IP switching. Most other FC vendors use off the shelf technology.

silicon is required to enable many software features like hardware zoning, frame-filtering, performance monitoring, QoS, and trunking. This subsection discusses several<sup>122</sup> Brocade ASICs, and shows how their feature sets evolved over the years.

## Эволюция ASIC

Brocade takes an “evolution, not revolution” approach to ASIC engineering. This balances the need to add as much value as possible with the need to protect customer investments and de-risk new deployments. Each generation of Brocade ASICs builds upon the lessons learned and features developed in the previous generation, adding features and refinements while maintaining consistent low-level behaviors to ensure backward and forward compatibility with other Brocade products, as well as hosts and storage. Brocade has been developing ASICs for a decade now, with each generation becoming more feature-rich and reliable than the last.



### Side Note

*The ASIC names used in this subsection are the internal-use Brocade project codenames for the chips. Brocade codenames generally follow a theme for a group of products. There have been three different themes for ASICs to date: fabric-related, bird-related, and music-related. Platforms and software packages also have codenames, but their external “marketing” names are used throughout this book. This is not done with ASICs because Brocade does not have external-use names for ASICs.*

---

<sup>122</sup> Brocade has developed a number of ASICs that are not yet being shipped, and thus are not included in this work. Register on the *SAN Administrator’s Bookshelf* website to receive updated content as additional chips become generally available.

## ***Stitch u Flannel***

The first ASIC that Brocade developed was called Stitch. Development on Stitch began in 1995. It was initially introduced to the market in the SilkWorm 1xx0 series of Fibre Channel switches in 1997. (See “Коммутаторы SilkWorm 1xx0 FC” on p430.)

Stitch had a dual personality: it could act as either a 2-port front-end Fibre Channel fabric chip, or a back-end central memory switch. The SilkWorm 1xx0 motherboards had a set of back-end Stitch chips, and accepted 2-port daughter cards that each had one front-end Stitch. The ASIC could support F\_Port and E\_Port operations on those cards. However, it could not support FL\_Port.

To address that gap, Brocade developed the Flannel ASIC. Flannel could act as a front-end loop chip on a daughter board, but could only act as an FL\_Port. It was therefore necessary to configure a SilkWorm 1xx0 switch as the factory for some number of fabric ports and some number of loop ports. Once deployed, the customer would need to live with the choices made at the time the switch was ordered. Furthermore, there was no way to make device attachment entirely “auto-magic;” it could matter which port a user plugged a device into.

## ***Loom***

The second-generation Brocade ASIC, Loom, was designed to replace both Stitch and Flannel. The new ASIC lowered cost, improved reliability, and added key features. The first Loom-based products were introduced in 1999.

The port density of the chip was increased from 2-port to 4-port, and each Loom had the personalities of both Stitch and Flannel. Four Looms could be combined to form a single non-blocking and uncongested 16-port central memory switch. This substantially lowered the component count in

the SilkWorm 2xx0 series platforms, improving reliability as well as lowering cost. (See “Коммутаторы SilkWorm 2xx0 FC” on p432.)

Feature improvements were made in many areas, including PID-based hardware zoning, larger routing tables, and improved buffer management. Updated “phantom logic” was introduced to support private loop hosts. (The QL/FA feature.) Virtual channels were added to eliminate blocking on inter-switch links.

One of the most important features that Loom introduced was the U\_Port. All three port types (F, FL, and E) could exist on any interface, depending on what kind of device was attached to the other end of the link. Switches using Loom could auto-detect the port type of the remote device: a substantial advance in “plug and play” usability. Auto-detecting switch ports came to be known as a Universal Ports (U\_Ports) and the SilkWorm 2800 running the Loom ASIC was the first in the industry to support this feature.

Loom enjoyed remarkable success and longevity. Brocade shipped well over a million Loom ports, and still has a very high percentage of them active in the field, despite the length of time for which the chip has been shipping. Brocade has therefore continued to support backwards compatibility with Loom-based products in all subsequent ASICs and platforms.

### ***Bloom u Bloom-II***

Bloom was designed to replace Loom, again lowering cost, improving reliability, and adding features.

Bloom first appeared in 2001 in the SilkWorm 3800 switch. It had eight ports per ASIC, and two Bloom's could be combined to form a single non-blocking and uncongested 16-port central memory switch called a “Bloom ASIC-pair.” (One ASIC-pair is what powered the SilkWorm 3800, for

example.) Because this ASIC had more ports than its predecessor, Brocade named the chip by adding a “B” in front of “Loom” to indicate that it was Bigger than Loom.

Bloom also increased the port speed to 2Gbit, doubling performance vs. Loom. In addition, the new ASIC added better hardware enforced zoning (both PID- and WWN-based), frame-level trunking to load-balance groups of up to four ports, frame filtering, end-to-end performance monitoring, and enhanced buffer management to support longer distances on extended E\_Ports. The chip also had routing table support allowing many chips to be combined to form a 128-port single-domain director (SilkWorm 24000).

The Bloom-II ASIC has such minor changes to Bloom that it is considered a simple refinement, not a new generation. A new process was used in its design to shrink the size of each chip, lowering power and cooling requirements. Additional test interfaces were added to improve manufacturing yield and reliability. Buffer management was improved to allow longer distance links at full 2Gbit speed.

At the time of this writing, Bloom is still shipping in the SilkWorm 12000 port blade and the SilkWorm 3800 switch. It was also used in the SilkWorm 3200 and 3900 switches, and in a number of OEM embedded products. Bloom-II is still shipping in the SilkWorm 3250 and 3850 switches, and in the SilkWorm 24000 blade set. (See both “Поставляемые платформы Brocade” on p397 and “Инсталлированая база платформ Brocade” on p430.)

## ***Condor***

The fourth generation ASICs from Brocade have code-names related to birds. Condor is the fourth-generation Fibre Channel fabric ASIC, and the first of its generation to become generally available. It builds upon the previous three ASIC generations, adding significant features and improving reliability to an unprecedented degree. At the time of

this writing, Condor is shipping in the Brocade 4100 and 4900 switches, and Brocade 48000 director.

Like previous Brocade ASICs, Condor is a high-performance central memory switch, is non-blocking, and does not congest. It builds on top of the advanced features that Brocade added to Bloom-II.<sup>123</sup> However, Condor has many major enhancements as well, and is not simply a “Bloom-III.” It is truly a fourth-generation technology.

Condor has thirty-two ports on a single chip, with each port able to sustain up to 4G bits per second (8Gbits full-duplex) in all traffic configurations. Each chip has 256Gbits of cross-sectional bandwidth. It was designed to support single-domain director configurations much larger than the Bloom-II-based SilkWorm 24000, in which case the platform cross-sectional bandwidth will be *massively* higher. For example, if the Brocade 48000 is configured with 128 4Gbit Condor ports, its internal cross-sectional bandwidth is 1Tbit. The number of virtual channels per port has also been increased to allow non-blocking operation in larger products and networks.

The doubling in port speed is only the beginning of Condor’s performance enhancements. Frame-based trunking has been expanded to support 8-way trunks, yielding 32Gbits (64Gbits full-duplex) per trunk. Exchange-based load balancing (DPS) is possible between either trunked or non-trunked links. (See “Балансировка линков” starting on page 272.) Two Condor ASICs networked together with half of their ports could sustain 64Gb/s (128Gbits full-duplex) between them, and far more bandwidth could be sustained between Condor-based blades in directors. In fact, com-

---

<sup>123</sup> Except for private loop support. This is near end of life based on declining customer demand, so priority was given to other features. Private loop devices are almost entirely out of circulation already, and the little remaining demand can be met by using Bloom-based switches in the same network as Condor platforms.

ing multiple Condor ASICs running 4Gbit link s with fram e and exchange trunking can yield 256Gbit evenly balanced paths.

Condor also i mproves control-plane perfor mance. Each ASIC can offload the platfo rm CP from many node login tasks. When a Fibre Channel device attem pts to initialize its connection to the fabric, previous ASICs would forward all login-related fram es to the CP. Condor is capable of performing m uch of this without involving the CP, which improves s witch and fabric scal ability as well as response time for nodes.

The ASIC m emory system s have also been improved. Buffer management and hardwa re zoning tables are the primary benefi ciaries of this. A centralized buffer pool allows better long distance support.: any port can receive over 200 buffers out of the pool. Centra lized zoning m emory allows more flexible and scalable deploym ents using “full” hardware zoning. (See “ Механизм реализации зонирования” on p498 for m ore inform ation.)

### ***Goldeneye***

Goldeneye, like Condor, is part of the fourth-generation Fibre Channel fabric ASIC set from Brocade, and the second of its generation to become nerally av ailable. It build s upon the previous three ASIC generations, adding significant features and im proving reliability to an unprecedented de gree. At the time of this writing, Goldeneye is shipping in the embedded switches and Brocade 200E switch.

Like previous Brocade ASIC s, Goldeneye is a high performance central m emory switch, is non-blocking, and does not congest. It builds on top of the advanced f eatures that Brocade added to Bloom-II. However, Goldeney e has many major enhancem ents as well, and is not sim ply a

"Bloom-III." It is truly a fourth-generation technology.

Goldeneye has 24 ports on a single chip, with each port able to sustain up to 4Gbits per second (8Gbits full duplex) in all traffic configurations. Each chip has 192Gbits of cross-sectional bandwidth. It was designed to support highly dense products such as the embedded blade server switches.

The doubling in port speed is only the beginning of Goldeneye's performance enhancements: Frame-based trunking can support up to 4-way trunks, yielding 16Gbits (32Gbits full-duplex) per trunk. Exchange-based load balancing (DPS) is possible between either trunked or non-trunked links.

Goldeneye also improves control-plane performance. Each ASIC can offload the platform CP from many node login tasks. When a Fibre Channel device attempts to initialize its connection to the fabric, previous ASICs would forward all login-related frames to the CP. Goldeneye is capable of performing much of this without involving the CP, which improves switch and fabric scalability as well as response time for nodes.

The ASIC memory systems have also been improved. Buffer management and hardware zoning tables are the primary beneficiaries of this. A centralized buffer pool allows better long distance support: any port can receive over 200 buffers out of the pool. Centralized zoning memory allows more flexible and scalable deployments using "full" hardware zoning.

### **Egret**

Egret is a bridge chip which takes three internal 4Gbit FC ports on a blade, and converts them into a single external 10Gbit FC interface. At the time of this writing, it is used only on the FC10-6 blade (page 418), which has six Egret chips connected to two Condor ASICs. From a performance stand-

point, an Egret-Egret IS L can be thought of as functionally identical to a three-port by 4Gbit frame-level trunk.

The are differences, however. The Egret app roach uses 1/3<sup>rd</sup> of the number of fiber optic strands or D WDM wavelengths, which can produce substantial cost-savings in some long distance solutions. On the other hand, 10Gbit FC requires more expensive XFP media, more complex and thus more expensive blades, and single-mode cables, which can increase cost massively for shorter-distance IS Ls. As a result, it is expected that Egret will only be used for DR and BC solutions. In addition to aggregating three interfaces into one, the Egret chip also contains its own buffer-to-buffer credit memory, allowing each and every 10Gbit port to support a full-speed connection over dark fiber or xWDM of up to 120km.

### ***FiGeRo / Cello***

The FiGeRo and Cello chips power the Brocade Multi-protocol Router (AP7420). Both ASICs were acquired when Brocade bought Rhapsody Networks. The platform consists of sixteen FiGeRo chips (one per port) interconnected via one Cello that acts as a cell switching fabric.

FiGeRo was codenamed to follow a music theme. (As in, “The Marriage of Figaro.”) The “Fi” and “Ge” components of the name refer to the fact that a FiGeRo ASIC can act as either a Fibre Channel port or a Gigabit Ethernet port. Cello got its name by being a cell switching ASIC.

Each FiGeRo ASIC has fixed gates to perform frame-level functions efficiently, and three embedded RISC processors plus external RAM to give each port exceptional flexibility for higher-level routing and application processing functions. Currently, the Multiprotocol Router running FiGeRo supports FC fabric switching, FC-to-FC routing, FCIP tunneling, and iSCSI bridging. More advanced fabric applications are being developed by Brocade and its part-

ners. In fact, at the time of this writing, several ILM and UC applications for this architecture are just beginning to ship.

Similar functionality is expected to be available throughout the Brocade product line by the end of 2005.

## Многоуровневые внутренние архитектуры

Modular switches like SAN directors always require internal connectivity between discrete components over a midplane or backplane. It is not possible or even desirable to have, for example, a single-ASIC director. Some of the major benefits of a bladed architecture are that customers can select different blade types for different applications, swap out old blades one at a time during upgrades, and have the overall system continue to operate even in the face of failures on some component. A single-chip solution would prevent all of these features and more from working. As a result, all such products from all vendors have some chips on port blades, some other chips on control processor blades, and (typically) some chips on back-end data-plane switching blades. The Brocade directors are no exception.

There are many different approaches that can provide the required chip-to-chip connectivity. It is possible to use shared memory, a cross bar, a cell switch, or a bus, to name just a few approaches that have been used in the networking industry. A director might have connectivity between front-end protocol blades via a crossbar using "off the shelf" commodity chips, or it might use native Fibre Channel connections between blades using SAN-optimized ASICs. High-speed packet switches for both Ethernet and Fibre Channel use shared memory designs for highest performance. Commodity Ethernet switches often use crossbars to lower research and development costs, thus increasing short-term profits for investors at the expense of long-term viability and customer satisfaction. It is also possible for more than one

option to be combined within the same chassis, which is often known as a *multistage* architecture.

Most Brocade products are single-stage central memory switches, often consisting of just one fully-integrated chip. However, some of the larger products use multistage designs to support the required scalability and modularity. All of internal-connectivity approaches from all vendors have an *internal topology*, a set of *performance characteristics*, and a set of *protocols*, much like a network.<sup>124</sup> The arrangement of the chips and traces on the backplane or midplane create the topology, and the chips connected to this topology have link speed and protocol properties. Indeed, it is possible to make many analogies between networks and internal director designs, no matter what connectivity method is used.

Brocade multistage switches use central memory ASICs with back-end connections based on the same protocol as the front-end ports. This avoids the performance overhead associated with protocol conversions that affect other designs like crossbars. The back-end connectivity is an enhanced Fibre Channel variant called the “Channeled Central Memory Architecture.” (CCMA) The connections between ASICs are therefore called CCMA Links. While these are enhanced beyond standard FC links in a number of ways, the payload and headers of frames carried by the CCMA Links use an unmodified, native Fibre Channel frame format. This allows the director to operate efficiently and reliably.

The use of CCMA links defines protocol characteristics, but there are variations in terms of other performance characteristics and topology depending on how CCMA connections are made. (I.e. the back-end topology of a director is the geometrical arrangement of the back-end ASIC-to-ASIC

---

<sup>124</sup> While the internal connectivity in a chassis does not work exactly the same way that an external network works, they do have enough in common that this provides a useful analogy.

links, much the same way as the topology of a SAN is the arrangement of ISL connection.) The remainder of this subsection discusses two variations on the Brocade CCMA multistage architecture in detail.

### *SilkWorm 12000 и 3900 “XY” архитектура*

The Brocade SilkWorm 12000 is a highly available Fibre Channel Director with two domains of 64 ports each, and the SilkWorm 3900 is a high-performance 32-port midrange switch. Both platforms can deliver full-duplex line-rate switching on all ports simultaneously using a non-blocking CCMA multistage internal architecture. This section discusses the details of how ASICs are interconnected inside the two products, and provides some analysis of how that structure performs.

#### Внутренние связи лезвия SilkWorm 12000

The SilkWorm 12000 chassis (Figure 105 p 439) is comprised of up to two 64-port domains, each of which may contain up to four 16-port cards. Each card is divided into four 4-port groups known as quads. Viewed from the front and the side, a blade is constructed as depicted in Figure 115.

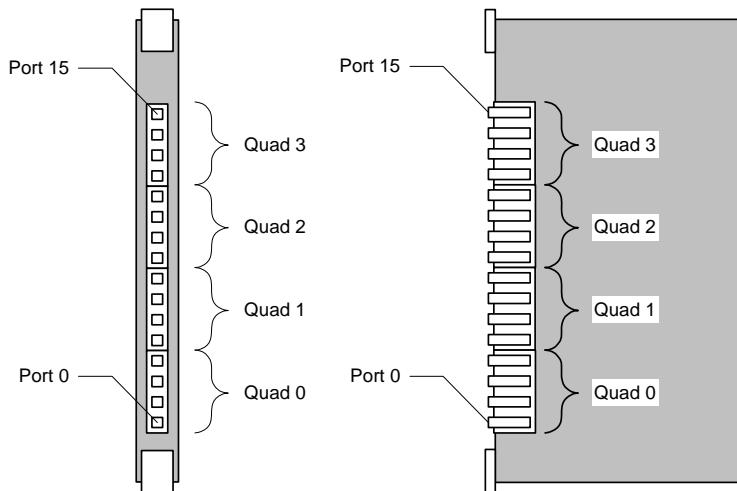


Figure 115 - SilkWorm 12000 Port Blades

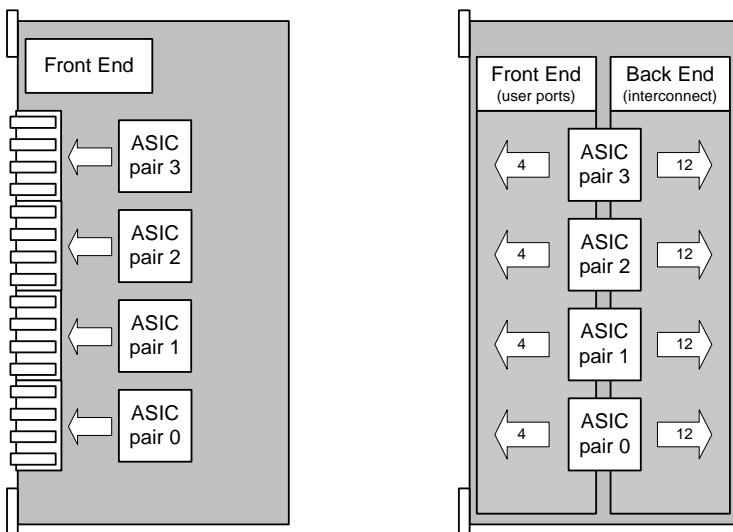
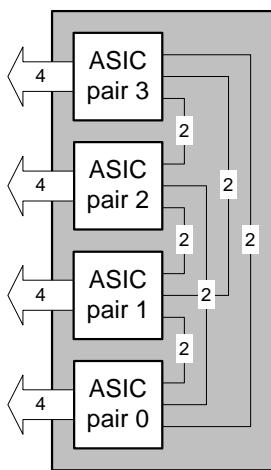


Figure 116 - SilkWorm 12000 ASIC-to-Quad Relationships

The SilkWorm 12000 uses a distributed switching architecture. Each quad is a self-contained 16-port central memory switching element, comprised of two ASICs. Four ports of each quad are exposed outside the chassis, and may be used to attach FC devices such as hosts and stor-

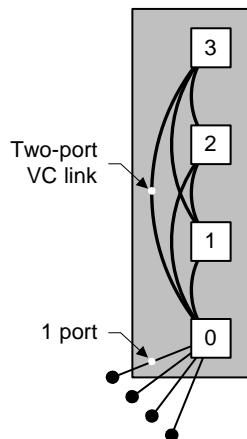
age arrays, or for Inter-Switch Links (ISLs) to other domains in the fabric. The remaining twelve ports are used internally, to interconnect the quads together, both within and between blades. This means that the SilkWorm 12000 actually has three ports of internal bandwidth for each port of external bandwidth: a 1:3 undersubscribed design. Viewed logically from the side, the ASIC-to-quad relationship on a blade can be viewed in either of the ways shown in Figure 116.

The interconnection mechanism used to tie the quads together involves connecting each quad directly to every other quad in the same row and column with one internal 4Gbit CCMA link. Each link uses two internal ports plus frame-level trunking to achieve 4Gb full-duplex bandwidth on its path. Three of the six links are vertical (within a blade) and three are horizontal (between blades). Within a blade, the connection pattern is as shown in Figure 117.



**Figure 117 - SilkWorm 12000 Intra-Blade CCMA Links**

Each of the four quads has four ports for front-end connections, and six ports (three 4Gbit VC links) going to the other quads within that blade. (Each of the lines with a “2” in the figure represents 2x2Gbits balanced with frame trunking.) Figure 118 provides a more abstract depiction of this.



**Figure 118 - SilkWorm 12000 CCMA Abstraction**

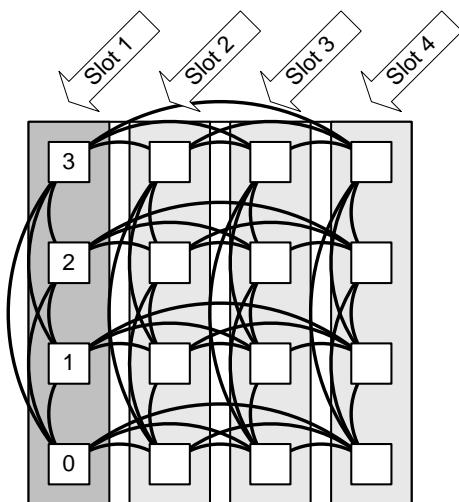
Each one curved vertical line rep resents a 4Gbit internal trunk. Each numbered box is a quad, which has four external connections, represented by the four “pins” attached to quad 0. The diagram represents one SilkW orm 12000 port blade.

In addition to the three vertical back-end 4Gbit CCMA links within the blade, each quad has three horizontal back - end 4Gbit links to the o ther three blades in the domain. The overall interconnection with in a SilkW orm 12000 64-port domain can be viewed like Figure 119.

This matrix connection method is known as the “XY” method, since the internal CCMA links follow a grid. The name comes from mathematics. The horizon tal connections are called “X” connections, since that is the v ariable traditionally used to represent the horizontal axis on a graph. The vertical connections are called “Y” links.

If the source and destination quads are in th e same row, the director will use one X-axis internal CCMA “hop” to get between them, since there is a direct connection available. This adds just 700 or so nanosec onds of latency. If they are in the same column, it will use one Y-axis hop. Look back at the figure. See how any two qua ds in the sam e row or col-

umn are directly connected? This shortest path” will always be used if it is available. If the source and destination are in different rows and columns, there is no direct connection. In that case, in the default switching configuration, the platform will route traffic between any two quads using an X-then-Y formula: first the frame will traverse a horizontal CCMA link to an intermediate ASIC, then it will take the vertical link to the destination ASIC.



**Figure 119 - SilkWorm 12000 64-Port CCMA Matrix**

#### Внутренние связи портов SilkWorm 3900

SilkWorm 3900 internal connections are similar to those in the SilkWorm 12000 port blade. The platform consists of four ASIC-pairs wired together in an XY topology. Since there are no other blades to connect to, all of the links are used to connect the ASIC-pairs into a square. Each quad has eight external ports, and eight internal ports. Like the 12000, traffic will take a direct path if it is available, and will take an X-then-Y path if moving diagonally.

Анализ производительности“ХҮ”

There are three ways to evaluate performance of a network product: theoretical analysis, empirical stress-testing, and real-world performance testing.

From a theoretical standpoint, both XY products have more than adequate performance. There is more bandwidth used to interconnect the quads together on a 12000 than there is input bandwidth on the front-end of the switch. This is referred to as an *under-subscribed* architecture: for each quad, there are fewer ports subscribed to the backend than there is bandwidth on the backend by a ratio of one-to-three, usually written 1:3. (Four front-end connections to twelve back-end ports reduces to a ratio of 1:3.) This is 8Gbits of front-end bandwidth feeding into 24Gbits of total back-end bandwidth per quad. The SilkWorm 3900 has a 1:1 subscription relationship: 16Gbits of input feeding into 16Gbits of back-end CCMA link capacity.



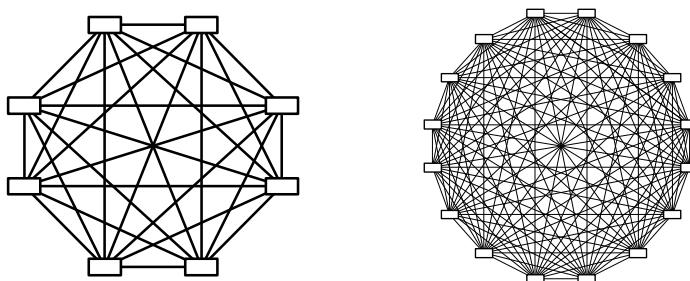
### Side Note

*For almost all users, all Brocade multistage platforms have “plug and play” performance, and the information in this section is only provided to satisfy curiosity. However, for advanced users who need to tune their applications for ultimate performance, the topology information below can be relevant. The rule of thumb is this: It is worth taking the time to understand the internal topology of a multistage product only if it is necessary to run all ports on the platform full-speed, full-duplex, for sustained periods, and there will be a business impact if even a few of the ports run slower than the theoretical maximum possible line rate.*

While the front-end ports cannot generally flood all of the back-end bandwidth on the SilkWorm 12000, it is theoretically possible for certain traffic patterns to exhibit congestion due to an *imbalanced usage* of this band-

width. To determine if theoretical limits of a platform can be exhibited in the real world, empirical testing can be performed. This has been done extensively by Brocade, by third parties such as networking magazines, major customers, and independent laboratories, and – of course – by other switch vendors. In every case, the conclusion was the same: the XY products produce uncongested operation in any real-world and most purely contrived traffic patterns. Even incredibly stressful traffic configurations such as a full mesh test will produce no congestion.

For example, it is possible to connect all 32 ports of a SilkWorm 3900 to a SmartBits™ traffic generator. Using their management tool, the SmartBits can be configured to send traffic flows from every port on the switch to every other port. This is known as a full mesh traffic pattern, and is generally acknowledged as one of the most stressful traffic configurations possible. Figure 120 illustrates an eight node full mesh and a sixteen node full mesh. Each box represents a port on the switch, and each line a pair of flows.



**Figure 120 - Full-Mesh Traffic Patterns**

Clearly, there are quite a few simultaneous traffic flows in these configurations. When testing the SilkWorm 3900 with a 32-port full mesh, far more connections are in play, and yet all 32 ports show full-speed, full-duplex performance. Similarly, the SilkWorm 12000 will perform at peak with a 64-port full mesh.

It seems unlikely based on this that any given environment would experience any internal performance bottlenecks related to the XY CCMA architecture. If that ever did happen, there are a number of options for tuning XY performance. For example, following Brocade's tradition of supporting localized switching, each group of four ports on the 12000 (quad) and eight ports (octet) on the 3900 can switch locally without even using the XY traces. This provides users who take advantage of known locality the opportunity to optimize performance still further.

### **Архитектура Brocade 24000 и 48000 "CE"**

The Brocade 24000 and 48000 chassis (Figure 106 p441 and Figure 78 p405 respectively) are functionally equivalent to that of the SilkWorm 12000. Both are CCMA multistage directors, though the products use different backplane traces. Both of the newer directors can exhibit uncongested operation both in theory *and* in empirical testing.

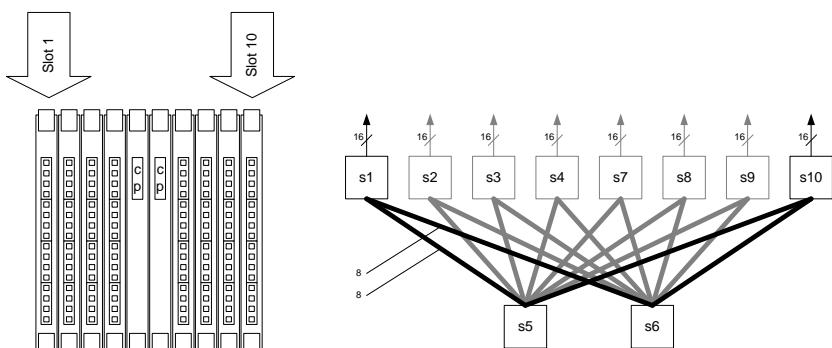
In the Brocade 24000, each port blade has two Bloom-II (p505) ASIC-pairs which expose eight ports to the user, and have equivalent bandwidth used for backplane CCMA links: any given octet has 16Gigabits (32G full-duplex) of possible external input, and the same bandwidth available to connect to any other octet. Local switching can be done within an 8-port group.

The Condor-based (p 506) Brocade 48000 has 16-, 32-, and 48-port blades. Local switching is possible within a 16-port group on the first two, and a 24-port group on the 48-port blade. In each case, the director has 64Gbits of internal bandwidth per slot (128Gbits full-duplex) *in addition* to the local switching bandwidth. This means that the 16-port blade has a 1:1 subscription ratio even if all external ports are all connected to 4Gbit devices and no traffic is localized. The larger blades also have 4Gbit interfaces, and are uncongested in most real-world scenarios. However, it is important to realize that the larger blades *can* exhibit internal con-

gestion if (a) traffic on enough ports is sustained at or near full speed, and (b) none of the flows are localized. Most environments have some degree of “burstyness” and/or some degree of locality, so the over-subscription of the two high-port-count blades is largely academic.

The characteristics of the two newer directors are similar to the SilkWorm 12000 in some respects, but radically different in others. This is because the two newer platforms use a Core/Edge (CE) ASIC layout instead of the XY layout. The CE layout is more symmetrical: all ports have equal access to all other ports. In addition, local switching is allowed within an octet rather than a quad on the 24000, which doubles the opportunity to tune connection patterns for absolute maximum performance if locality is known. The 48000 doubles that again for two blades, and triples it for the 48-port blade.

Figure 121 shows how the blade positions in the Brocade 24000 director are connected to each other. On the left is a somewhat abstract cable-side view of the director, showing the ten blade slots. Each of the port cards has four quads depicted. Quad boundaries are still relevant for things like ISL trunking. The top two and bottom two quads on each blade each form an octet for local switching.



**Figure 121 - Top-Level “CE” CCMA Blade Interconnect**

On the right is a high-level diagram of how the slots interact with each other over the backplane. Each thick line represents a set of eight 2Gbit CCMA links connecting the port blades with the CP blades. The CP blades contain the ASICs that switch between octets. Every port blade is connected to every CP blade, and the aggregate bandwidth of these CCMA links is equal to the aggregate bandwidth available on external ports. Each port blade has 16 2Gbit FC ports going outside the box, and  $2 \times 8 = 16$  2Gbit CCMA Links going to the backplane.

As this diagram illustrates, the internal connectivity looks similar to a resilient core/edge fabric design. This is no accident: the geometry of the core/edge design has been universally accepted as *the* best-practice for high-performance, highly scalable, high availability SAN designs, and is currently recommended by all vendors. By using the same geometry for the internal layout of its directors, Brocade has achieved the same benefits within the chassis that users have adopted for external connections. The “every port blade to every CP blade mesh” is what makes it a “CE” layout, and the 1:1 internal-to-external bandwidth ratio makes it a “fat-tree” or non-over-subscribed layout.

The Brocade 48000 has the same top-level connectivity diagram when populated with 16-port blades. The difference is that each “unit” represents a 2 Gbit connection in the 24000 and a 4Gbit connection in the 48000. So, for example, the “8 unit” link between s1 and s5 represents 16Gbits of aggregate bandwidth in the Brocade 24000, and 32Gbits in the Brocade 48000.

Of course, the two directors are not *really* Core/Edge networks of discrete switches, but thinking of them that way does provide a useful visualization. Because they are fully-integrated single-domain FC directors and not merely “networks in a can”, the two platforms also:

- Are easier to manage than the analogous network of individual switches.
- Take up less rack space than a network would use.
- Are easier to deploy and manage.
- Simplify the cable plant by eliminating the large number of ISLs and media required for a network.
- Are far more scalable, as they do not consist of a large number of independent domains.
- Are much less expensive, both in terms of its initial and ongoing costs.
- Have higher reliability due to having far fewer active components.
- Do not run switch-to-switch protocols internally.
- Are capable of achieving even greater performance due to internal routing optimizations.

When frames enter a port blade on either director, under normal working conditions it can select between either of the two CP blades to switch the traffic. This provides redundancy in case one CP blade should fail, and also allows full performance. For example, the Brocade 48000 uses frame-level and exchange-level trunking to balance IO between the two CPs in much the same way Condor-based switches can balance traffic in a core/edge fabric. The net result is that no empirical test has ever shown congestion within either director: testing from Brocade, independent laboratories, networking magazines, and other vendors alike have confirmed that these two platforms are simply the highest performing SAN products in the world today.

## Скорости соединений

Storage networks may operate at a variety of speeds. Fibre Channel standards define speeds including 1Gbit, 2Gbit,

4Gbit, 8Gbit, and 10Gbit.<sup>125</sup> Ethernet defines 10Mbit, 100Mbit, 1Gbit, and 10Gbit, though only 1Gbit and 10Gbit are relevant to storage networking.

This subsection discusses each link speed. More detail is provided for 4Gbit FC than for the other speeds, since it is the newest of the link rates from an implementation perspective. (Although it predates 10Gbit from a standards point of view.)

### **Форматы кодирования**

Each of the link speeds discussed in this section has an *encoding format*. Encoding is used on the signal to make it transition from zero to one more often, thus allowing the high vs. low signals to be distinguished from each other. If long periods were allowed to elapse between transitions, a link might not be able to tell the difference between minor signal variations (i.e. noise) and real 0/1 transition. It could begin treating noise as if it were data, which could cause link failures and even data corruption in extreme cases. Encoding formats ensure that this will not occur. As a side benefit, encoding provides an error detection method, somewhat like parity bits in a modem protocol.

There are many formulas that can be used to encode a signal. Some encoding formats are referred to by the number of bits on the link required to represent a certain number of data bits, such as “8b/10b.” The ratio indicates the amount of user data in a given data unit.

---

<sup>125</sup> FC-PH also defines 250Mbit “1/4 speed” and 500Mbit “1/2 speed” Fibre Channel interfaces. However, 1/4 speed has been obsolete for about a decade, and 1/2 speed was never implemented. It is also possible to run Fibre Channel at other speeds on intra-platform links. For example, the Condor ASIC is capable of forming 3Gbit FC connections to other Brocade ASICs, even though there is no standard defined for this.

8b/10b requires that ten bits be sent down the line to represent eight data bits. This affects throughput. 8b/10b is 20% “encoding overhead.”

In contrast, the “64b/66b” encoding format is only about 3% overhead, which means more payload can be moved for a given link speed. However, it also means that the link can be less effective at detecting errors, and could be subject to more frequent failures.

The bottom line is that encoding is necessary and present on all technologies discussed below. It is also necessary that devices on both ends of a connection use the same encoding format, i.e. 8b/10b or 64b/66b. It is not possible to have an 8b/10b device talk to an 64b/66b device natively; one or the other would need to be converted before communication would be possible. This caveat only applies to 10Gbit, since *all* other speeds use 8b/10b encoding.

## **1Gbit FC**

1Gbit Fibre Channel was defined in the FC-PH standard in 1994. All Brocade platforms ever shipped support this speed. It was considered the “sweet spot” in the industry for many years, and is still viable today for many customers. Links running at this speed use 8b/10b encoding, and can achieve a user-data throughput of just over 100Mbytes/sec. (200Mbytes full duplex.) Both copper and optical media are defined by the standard. 1Gbit interfaces most often use GBICs, although 2Gbit Fibre Channel SFPs also support this rate to maintain backwards compatibility.

## **2Gbit FC**

2Gbit Fibre Channel was defined in the FC-PH-2 standard in 1996, though no vendor implemented it for some time after that. All Brocade platforms more recent than the SilkWorm 2xx0 series support auto-negotiation between 1Gbit and 2Gbit FC. This is considered to be the “sweet

spot” in the industry today, although 4Gbit is expected to replace 2Gbit in 2005. Links running at this speed use 8b/10b encoding, and can achieve a user-data throughput of just over 200Mbytes/sec. (400Mbytes full duplex.) Both copper and optical media are defined by the standard. 2Gbit interfaces most often use SFPs.

### ***4Gbit FC (Frame Trunked или Native)***

For several years now, Brocade has offered frame-level trunking (p 460) on all 2Gbit products. This can be used to combine two 2Gbit interfaces into one evenly balanced 4Gbit channel.

Recently, Brocade introduced a *native* 4Gbit interface, in which each individual port can run at that speed. These ports still may be trunked to form even higher rate pipes. This allows node connections at 4Gbit as well as higher speeds and lower costs for ISL connections. Native 4Gbit is expected to become the “sweet spot” in the SAN industry for 2005 and beyond.

Like 2Gbit Fibre Channel, native 4Gbit was defined in the FC-PH-2 standard in 1996. The first Brocade platform to support this standard is the Brocade 4100. (p 400) It can support auto-negotiation between 1Gbit and 2Gbit FC on all ports for backwards-compatibility. While other 4Gbit vendors may not support trunking, on Brocade platforms up to eight 4Gbit links can be trunked to form a single 32Gbit channel (p 535), and multiple trunks can be balanced into a single 256Gbit pipe.

Links running at 4Gbit use the same 8b/10b encoding as existing 1Gbit/2Gbit infrastructure, and can achieve real-world payload throughput of over 400Mbytes/sec. (Over 800Mbytes in full-duplex mode.) 4Gbit interfaces use the same SFP standard and optical cabling as 1Gbit and 2Gbit interfaces, which allows 4Gbit products to be backwards-

compatible with installed base switches, routers, nodes, and data center cable plants.

Despite the fact that the 4Gbit standard was ratified at the same time as the 2Gbit standard, no 4Gbit products were built until 2004. There was a debate in the FC industry about whether or not to build 4Gbit products at all, or to go straight to 10Gbit. The debate ended when the Fibre Channel Industry Association voted to adopt 4Gbit, and all major FC vendors began to add 4Gbit products to their roadmaps. The factors that motivated the industry in this direction included both economic and technological trends.

#### Technical Drivers for Native 4Gbit FC

Two of the most critical questions in the 4Gbit vs. 10Gbit debate were whether or not higher than 2Gbit speeds were needed at all, and if so which of the candidates could be widely deployed in the most practical way.

Higher speeds were deemed desirable for several reasons. For example, some hosts and storage devices - e.g. large tape libraries - were running fast enough to saturate their 2Gbit interfaces. In some cases, this was causing a business impact for customers: if a backup device could stream data faster, then backup windows could be reduced and/or fewer tape devices could be purchased. Furthermore, running faster ISLs would mean needing fewer of them, thus saving cost on switches and cabling. For long distance applications running over xWDM or dark fiber, the reduction in number of links could have a substantial ongoing cost savings.

For these and many other reasons, the industry acknowledged that 2Gbit speeds were no longer sufficient for storage networks. The choice was to use 4Gbit or 10Gbit. It turned out that 4Gbit had substantial technical advantages related to deployment, and provided at least the same performance benefits as 10Gbit.

Hosts and storage devices that were exceeding their 2Gbit interface capacity were not doing so by a large amount. Some tape drives were designed to stream at between 3Gbit and 4Gbit, and some hosts could match these speeds, but only a handful of the highest-end systems in the world could exceed 4Gbit, and even these could not generally sustain 10Gbit streams. 4Gbit interfaces could be marketed at cost parity with 2Gbit, but 10Gbit interfaces demanded a massive price premium due to architectural differences in the interfaces, so there was no point in using the more expensive 10Gbit interface in a node that could not even saturate a 4Gbit interface. Actual performance on nodes would be identical whether using 4Gbit or 10Gbit, and 10Gbit cost more across the board.

The biggest barrier to wide deployment of 10Gbit was its innate incompatibility with existing infrastructure. It required different optical cables, used different media, and was not backwards compatible with 1Gb or 2Gbit. Needing to rip and replace all HBAs and storage controllers at once, not to mention an entire data center cable plant would not only be prohibitively expensive, but operationally impossible in the “always on” data centers that power today’s global businesses.

It became clear because of these factors that the optimal speed for nodes would be 4Gbit. However, there was still a case to be made for *ISLs* at 10Gbit.

Replacing the optical infrastructure would be less of a technical issue with backbone connections, because there are typically far fewer of them than there are node connections. Additionally, some high-end installations really do require their switch-to-switch connections to run faster than 4Gbit. Indeed, some networks require backbones to run at far higher than 10Gbit speeds. No matter how fast an individual interface can be made, there always seems to be an application that needs more bandwidth. Brocade decided to solve this

with trunking for 4Gbit interfaces, giving 4G bit networks performance parity with 10GbE (and indeed beyond) while still lowering costs and simplifying deployments.

Another technical factor to consider is network redundancy. Most users configure links in pairs, so that there will be no outage if one link should fail. With a single 10Gbit link, any component failure will result in an outage, which means that the minimum realistic configuration between two switches is 20Gbits (2x 10Gbit links). Relatively few applications require so much bandwidth between each pair of switches, and given the cost of 10Gbit interfaces, redundancy would be harder to justify to management when purchasing a SAN.

To fully appreciate this, consider the performance parity case. If three 4Gbit links are configured, and one fails, the channel is 33% degraded. For a network with the exact same performance requirement, a single 10Gbit link is needed, which is more expensive than the three 4Gbit interfaces and requires more expensive single-mode optical infrastructure. If that link fails, the network has an outage because 100% of bandwidth is lost, thus requiring a second expensive 10Gbit link to be provisioned, even though the additional performance is not required. If a 10Gbit proponent were to argue that two times the performance were really needed, the 4Gbit proponent could configure six 4Gbit links, which would still cost less, have higher availability, and perform identically.

All of this adds up to substantial technical advantages for 4Gbit above 10Gbit. Until mainstream nodes can saturate 4Gbit channels, this is likely to remain the mainstream interface speed for storage networks.

### Экономические предпосылки Native 4Gbit FC

In the final years of the 20<sup>th</sup> century, companies were buying technology for its own sake, regardless of proven value proposition. In the early 21<sup>st</sup> century, however, the

overall global economy downturn caused the high-tech industry to adapt: any new technology had to provide end-users with a proven Return on Investment (ROI) in order to be adopted, so technology companies began to reevaluate their value propositions before going to market with new products. Since 4Gbit interfaces could provide more real technical benefit than 10Gbit in most cases, it became a question of which technology could lower the total cost of ownership the most, thus providing the highest ROI.

When using 10Gbit interfaces, the lowest speed possible is on a link is, obviously, 10Gbit. If a network designer feels that less performance is needed, and that less cost would be appropriate, there is no way to install *part* of a 10Gbit pipe. With 4Gbit trunked interfaces, the granularity of configuration is much finer: a designer can start with one 4Gbit link and add more links as needed if real performance data justifies the added cost.

4Gbit interfaces use the same low-level technology and standards as 1Gbit and 2Gbit across the board: the encoding format is just one example. One way to think of a 4Gbit switch is that it is *like* running a 2Gbit switch with a higher clock rate. The net result is that 4Gbit products can be marketed at about the same price as the existing 2Gbit products. 10Gbit, on the other hand, is fundamentally different: it uses technology that requires different components, which are all much lower volume. This is true to such an extent that current price projections indicate that three 4Gbit links will cost quite a bit less than one 10Gbit link, so even deploying equal bandwidth is more economical with 4Gbit.

With 4Gbit, redundancy and performance can be decoupled to a greater extent than with 10Gbit: redundant configurations can start at 8Gb/s (2x 4Gbit) at a fraction the cost of a non-redundant 10Gbit link, and can scale up to trunked configurations supporting far more bandwidth than 10Gbit: Brocade 4Gbit ASICs support up to 2 56Gbit con-

figurations using frame-based plus exchange-based trunking algorithms.

Not only were 10Gbit interfaces more expensive, but the optical infrastructure users already installed for 1Gbit and 2Gbit would not work with 10Gbit devices. 10Gbit interfaces require expensive single-mode fiber, and the vast majority of data centers today are wired with multi-mode fiber. 4Gbit, on the other hand, could use the existing cable plant, and could support the same SFP interface used for 1Gbit and 2Gbit. This meant that media and cable plants could be designed to run at all three speeds, providing backwards compatibility, whereas 10Gbit installations would require forklift upgrades.

Since 4Gbit products cost less than 10Gbit even at performance parity, and installation would be less expensive as well, the economic debate came out firmly on the side of 4Gbit, just as had the technical discussion.

### Сроки появления Native 4Gbit

At every point in the price / performance / redundancy / reliability map, 4Gbit is more desirable than 10Gbit. All major Fibre Channel vendors have 4Gbit on their roadmaps, including switch, router, HBA, and storage manufacturers. The Fibre Channel Industry Association has officially backed this movement, and it is expected that most FC equipment shipping by the end of 2005 will run at this speed. Indeed, at the time of this writing, Brocade has already been shipping 4Gbit products since late 2004.

Even though the benefits are clear and numerous, 4Gbit will not fully penetrate the Fibre Channel market immediately. Like any new technology, 4Gbit FC is expected to follow a curve of adoption, with different market penetration extents and different end-user benefits at different points on the timeline.

During the early-adoption time, 2Gbit native switches will still be in high volume production. First, the 4Gbit tech-

nology will be available only in selected pizzab ox switches like the SilkWor m 4100. It is usual for director-class products to follow behind switches by at least several months, since modular platforms are by nature harder to engineer, test, and market. This is why the Brocade 48000 shipped later than the 4100. During the interim period, 4Gbit switches will be deployed in stand-alone configurations, as the cores and/or edges of small to medium CE networks, and as edge switches in larger SANs.

Once 4Gbit blades begin to ship in higher volume, Silk-Worm 24000 2Gbit directors at the edge of fabrics will simply have all new blades purchased with SilkWorm 48000 4Gbit chips. There is probably no real incentive for most users to throw out their existing 2Gbit blades, so it is likely that 4Gbit ports will simply sit along side the existing 2Gbit interfaces within existing chassis.<sup>126</sup> The new 4Gbit blades will replace 2Gbit ISLs going to the core. Directors at the core of large SANs will either have their blades upgraded (4Gbit blades purchased and old blades transferred to edge chassis) or in some cases the entire core chassis may be migrated to the edges of a fabric.

The time lag between edge switches and directors is not considered to be a problem: the industry does not believe that 2Gbit is by any means obsolete. Most customers do not immediately require 4Gbit interfaces, and many customers will be able to use their 2Gbit switches for years to come. In fact, it is likely that 2Gbit switches will still be shipping for all of 2005 and even into 2006: they will simply decline in volume over that time.

---

<sup>126</sup> Brocade will offer 4Gbit blades that can co-exist with SilkWorm 24000 2Gbit blades in the same chassis, but at least two other vendors require forklift chassis upgrades. Be sure to ask if a 2Gbit chassis purchased today will support 4Gbit and 10Gbit blades in the future, and if these can co-exist with existing blades in an existing chassis.

Some time after the first 4Gbit switches ship, node vendors will start to come out with 4Gbit interfaces. Most users will not have an immediate need for e.g. 4Gbit HBAs, so it is likely that only new installations will use this speed. (This is why backwards compatibility with 1Gb and 2Gb was so important: it will take years for the installed base to become purely 4Gbit.)

By the end of 2005, it is expected that all major vendors will ship 4 Gbit interfaces by default on products in every segment, and that the vast majority of green field deployments will use this speed almost exclusively.

### ***8Gbit FC (Frame Trunked или Native)***

Brocade offers 8Gbit FC trunks on all of its 2Gbit platforms today. 8Gbit trunks are created by striping data across four 2Gbit channels to form one 8Gbit pipe. It is also possible to trunk two native 4Gbit interfaces on products which support that link rate; this has the same effect. Trunking can be used to resolve or proactively prevent performance bottlenecks in the network, which is where high-speed links are most needed.

In the future, it is expected that storage controllers and some hosts will need higher speeds on the network interfaces as well, and trunking cannot easily be used to solve this challenge. Unfortunately, the theory that 10Gbit would be the next logical step for node interconnects has run into cost and technology problems, as discussed under “10Gbit FC” later. As a result, the FCIA announced that its members have ratified the extension of the Fibre Channel roadmap to include native 8Gbit speeds on a single interface.

This should allow each interface on a node or switch to support 1Gbit, 2Gbit, 4Gbit, or 8Gbit, all using the same media and cable types. The intent is to allow customers to preserve their existing infrastructure investments and avoid

costly “forklift” upgrades, which would be needed to support 10Gbit technology.

In fact, at the time of this writing, 8Gbit products are already in late stages of development, and so some additional details are now available about this technology. It is expected that 8Gbit products will sell for a premium above 4Gbit, and that they will of course require new SFP media to operate at that speed. In general, 8Gbit can operate over the same optical infrastructure as 4Gbit, but it is advisable to run some tests – e.g. for DB loss – to make sure that the cable plant is sufficiently reliable. For a given cable quality, 8Gbit may support a shorter distance than 4Gbit, in the same way that 2Gbit supported shorter distances than 1Gbit. Finally, it seems almost certain that 8Gbit capable media will *not* auto-negotiate all the way down to 1Gbit; they will support 2Gbit, 4Gbit, and 8Gbit negotiation. The SFP industry realized that it would be costly and complex to add 1Gbit support, and did not expect customers to pay a premium for 8Gbit media only to connect it to 1Gbit devices. There is a simple work around for this: if you intend to connect 1Gbit devices to an 8Gbit switch, use 1Gbit, 2Gbit, or 4Gbit SFPs to do so.

## **10Gbit FC**

10Gbit FC uses a different low-level encoding format (p524) than any of the other port speeds – 64b/66b instead of 8b/10b – so a 10Gbit FC link has the throughput of three 4Gbit links. 10Gbit can be thought of as equivalent to 12Gbit from a payload carrying standpoint. On the other hand, at the time of this writing, three 4Gbit links cost *much* less than one 10Gbit link, and have higher availability: if a 10Gbit link fails, the connection is 100% down, whereas if a 4Gbit link fails in a 3-port trunk, the link is just degraded.

Perhaps more to the point, 10Gbit has fundamentally different requirements vs. any of the other link speeds across the board. 1Gbit, 2Gbit, 4Gbit, and 8Gbit can all use

SFPs and multi-mode fiber, but 10Gbit uses XFPs and more expensive single-mode fiber. Most existing data center infrastructure is designed with multi-mode fiber, and virtually all existing SAN components are designed to receive 8b/10b format; substantial reengineering is required for 64b/66b both at the product and data center levels. This adds total cost of ownership burden far beyond the massive price premium that 10Gbit interfaces are currently demanding.

This has kept 10Gbit adoption slow. In fact, there is widespread speculation that 10Gbit FC will simply never be implemented in hosts or storage devices, and that the industry will bypass it by adopting 8Gbit and then 16Gbit or faster link speeds based on the 8b/10b encoding method. However, there is a case to be made in favor of 10Gbit links for DWDM extension, since these products already have 10Gbit interfaces today. Brocade has therefore developed a 10Gbit FC blade for the Brocade 48000 director to support these distance extension applications. See the sections “Директор Brocade 48000” on page 405 and “Лезвие FC10-6 10Gbit Fibre Channel” on page 417 for more information. The section “” starting on page 364 has an extended example of this use case.

### **32Gbit FC (*Frame Trunked*)**

All of the Condor-based platforms support 32Gbit FC trunks. These are evenly balanced paths, so that one 32Gbit trunk is truly equivalent to a single link operating at that speed. The major difference is that trunks are comprised of multiple physical interfaces, and therefore have an inherent element of redundancy built in: if one link fails in a 32Gbit trunk, the remaining seven links will still deliver 28Gbits of bandwidth: more than 87% of the original capacity will remain. A single physical 32Gbit link would have failed down to 0% in a similar scenario.

## ***256Gbit FC (Frame или Exchange Trunked)***

Up to eight 8-port fram e-level trunks can be balanced at the exchange level by DPS to for m a single 256Gbit path. In this cas e, a single link failure will still leave in excess of 98% of the aggregate capacity. This is m ost likely only applicable to large -scale CE ne tworks for med fr om Brocade 48000 directors at both the core and edge layers.

## ***1Gbit iSCSI и FCIP***

In theory, it should be po ssible to achieve about 1/4<sup>th</sup> the performance of a Fibre Channel link by using commodity Ethernet equipment instead of purpose-built storage network gear. If this were true, it might allow organizations to deploy their SANs at a lower cost, if p erformance were not a factor. As it turns out, neith er iSCSI nor FCIP can ach ieve nearly 1Gbit of real throughput on a 1Gbit interface. See “iSCSI” on page 51 for some of the reasons behind this.

## ***10Gbit iSCSI и FCIP***

Some industry comm entators make an argum ent which goes something like this:

*1Gbit iSCSI cannot meet requirements for performance in today’s SANs, much less meet requirements for future datacenter architectures involving ILM or UC. However, deploying 10Gbit interfaces with hardware iSCSI and TCP engines will allow 10Gbit iSCSI to almost match 4Gbit Fibre Channel performance. Therefore 10Gbit iSCSI shall have a market.*

On the one hand, Brocade does carry num erous iSCSI and FCIP products, and is inve sting substantial R&D money in im proving them . There are use cases for SAN technolo gies which do not require the pe rformance of Fibre Channel, and Brocade intends to support them .

On the other hand, just as with 10Gbit FC, this is *not* expected to form a substantial percentage of the overall SAN market, because arguments like the one above are unlikely to convince many users. It is currently possible to implement 3x 4Gbit FC ports for about the same price as a single *non-accelerated* optical 10Gbit Ethernet link, and iSCSI protocol acceleration typically adds up to an order of magnitude to the cost of an interface. With Fibre Channel maintaining that kind of lead in price/performance, and also having about a decade lead in maturity and market adoption, IP SAN interfaces are likely to remain a fringe market for the future.



## Приложение С: Тест

This study guide is divided into two sections: a set of questions, and a corresponding set of answers. After reading the main body of the book, go through the questions below, and on a separate sheet of paper, write your answers. If you cannot think of an answer, first try looking it up in the preceding chapters. If you cannot find the answer there, also try looking in “Приложение D:” starting on page 550.

Once you have completed the questions, double-check your answers by looking at the section “**Error! Reference source not found.**” on page 546. You can also use that section as a last resort if you cannot think of an answer and cannot find it by looking it up in the main body of the book or in the FAQ.

### Вопросы для самопроверки

5. Storage Area Networks (SANs) are primarily intended to provide \_\_\_\_\_ level connectivity between hosts and storage devices.
6. \_\_\_\_\_ is by far the most common technology used for SANs today.
7. The traditional \_\_\_\_\_ architecture failed to meet increasing storage performance and asset utilization requirements, which paved the way for SANs.
8. Existing network technologies like \_\_\_\_\_ were too slow and unreliable to support SANs, which prompted the SAN industry to invent the \_\_\_\_\_ protocol.

9. \_\_\_\_\_ is a SAN solution category which allows improved asset utilization through reduced white space on storage arrays.
10. \_\_\_\_\_ is the industry leader in SAN infrastructure, carrying FC, iSCSI, FCIP, virtualization, and SAN Management products.
11. \_\_\_\_\_ is a set of processes and procedures related to managing the way the business value of information changes over time.
12. Switches are distinguished from hubs in that switches do not have a \_\_\_\_\_ architecture.
13. When deploying a SAN to support mission-critical systems, industry best-practices mandate a \_\_\_\_\_ SAN architecture with redundant HBAs and multi-pathing software.
14. When communication between port-pairs in a switch or network of switches *impair* communication between other ports it is known as \_\_\_\_\_. This distinguished from *blocking* which actually *prevents* communication, and is a typical characteristic of crossbar switches.
15. In order to optimize compute resources such as CPU cycles, a \_\_\_\_\_ solution should be considered.
16. The last step in the SAN planning process is to create a more detailed \_\_\_\_\_ document and \_\_\_\_\_ plan.
17. The ILM and UC trends intersect in the \_\_\_\_\_.
18. To justify the cost of a SAN, the design team should compare the hard and soft benefits of the SAN to the costs as part of a \_\_\_\_\_ analysis.
19. When considering which protocol to use for a SAN, it is important to understand that the \_\_\_\_\_ protocol is vastly more efficient and mature than \_\_\_\_\_.
20. The first step in designing a SAN is to \_\_\_\_\_.

21. The \_\_\_\_\_ has the responsibility of coordinating the entire SAN effort and usually has the SAN project plan as a deliverable.
22. In order to optimize \_\_\_\_\_, it is best to move tape systems onto the SAN.
23. SAN-enabled \_\_\_\_\_ are a good way to increase application uptime by allowing a standby node to take over if a production node fails.
24. The mapping of SCSI over Fibre Channel is called \_\_\_\_\_, whereas the mapping of SCSI over IP is called \_\_\_\_\_.
25. Looking at Gigabit Ethernet and Fibre Channel from a maturity standpoint, one factor to consider is that \_\_\_\_\_ came first, and \_\_\_\_\_ was actually on top of the \_\_\_\_\_ protocol layers.
26. Originally invented by Brocade, \_\_\_\_\_ is now the industry-standard protocol for routing between FC switches in a fabric.
27. The time during which the backup runs is called the \_\_\_\_\_ and its maximum size is determined by the length of time that the business can tolerate the associated performance degradation or application outage.
28. \_\_\_\_\_ is the fundamental storage protocol that lies under both FC and IP SAN technologies.
29. To connect a host to a Fibre Channel fabric, a card called a \_\_\_\_\_ is required.
30. To achieve even a fraction of FC performance, iSCSI hosts require an expensive \_\_\_\_\_.
31. \_\_\_\_\_ are sets of processes and overall design and management philosophies, not specific products.
32. Currently shipping Fibre Channel products support the following link rates: \_\_\_\_\_.

33. The FC standards also provide for the following link rates: \_\_\_\_\_, some of which are obsolete and some of which are expected to ship in the future.
34. Two important concepts for SAN designers moving forward are \_\_\_\_\_, both of which are related to virtualizing resources, and neither of which are currently available in “feature complete” solutions.
35. In order for devices on a SAN to discover each other, they need to register with and inquire from the \_\_\_\_\_, which is built in to FC switches but generally requires external hardware in an iSCSI network.
36. \_\_\_\_\_ is a solution category related to moving data between storage subsystems e.g. when old systems are coming off of lease.
37. The Fibre Channel equivalent of an Ethernet hub uses the rather limited \_\_\_\_\_ protocol.
38. In order to achieve faster performance between switches than a single ISL can support, Brocade supports two link aggregation methods: \_\_\_\_\_ and \_\_\_\_\_.
39. Almost all companies use \_\_\_\_\_ or \_\_\_\_\_ instead of iSCSI when they want to support storage over IP.
40. Regulatory requirements and fiduciary duty to investors are increasingly driving IT departments to implement \_\_\_\_\_ solutions, which are facilitated by SANs mapped over a MAN or WAN.
41. \_\_\_\_\_ is a category of SAN solution used in most other SAN solutions, which results in more efficient utilization of storage assets.
42. \_\_\_\_\_ is the concept that resources such as CPU power, RAM, and storage capacity could be provided in a manner similar to an electric power grid.
43. In an HA cluster or UC solution, compute nodes need access to each other’s data sets to enable application

- mobility. This means building the cluster onto a \_\_\_\_.
44. JBODs and SBODs are almost never used as primary storage in mission-critical applications. Such needs are usually better met by \_\_\_\_ arrays.
45. \_\_\_\_\_ in the context of SANs are behaviors that devices must follow in order to communicate.
46. SANs have been used to connect multiple processing nodes to scale \_\_\_\_\_, either through parallel operations or sequential workflow optimization.
47. Running backups over \_\_\_\_ robs hosts of needed CPU power, whereas running them over \_\_\_\_ is even more efficient than DAS.
48. Using the FC protocol guarantees \_\_\_\_\_ and timely frame delivery with negligible error rates.
49. \_\_\_\_\_ pose the greatest challenge for compatibility testing within storage networks, regardless of protocol.
50. In a “formulaic” resilient CE fabric, \_\_\_\_ core switches interconnect many edge switches.
51. Fibre Channel SANs almost always outperform DAS, but \_\_\_\_\_ most often does not.
52. FC links can be extended across up to a hundred kilometers or so of dark fiber using long-wavelength \_\_\_\_.
53. \_\_\_\_ allows an organization to determine where data belongs at any point in time.
54. UC is being driven primarily by three factors: \_\_\_\_\_.
55. There are five phases to the SAN planning process for green field deployments: \_\_\_\_\_.
56. There are five layers to the UC and ILM data center architectures: \_\_\_\_\_.
57. The place where ILM and UC intersect is the \_\_\_\_.

58. Specific \_\_\_\_\_ requirements must be gathered to determine what the SAN is supposed to accomplish for the organization.
59. “Compatible” devices are capable of being \_\_\_\_\_.
60. If devices are not compatible, further analysis is \_\_\_\_\_ because the network will simply not function.
61. Designers should try to support *initial* performance requirements, and also \_\_\_\_\_.
62. \_\_\_\_\_ is a measure of how often service personnel need to “touch” a system.
63. \_\_\_\_\_ is a measure of how much time a system is able to perform its higher-level functions.
64. \_\_\_\_\_ is a somewhat subjective measure of, among other things, how easy it is to fix problems in a SAN.
65. \_\_\_\_\_ allows multiple fabrics to be controlled from a single management point.
66. \_\_\_\_\_ automatically checks the SAN against evolving best-practices and has automated “housekeeping” features such as looking for unused zones.
67. \_\_\_\_\_ refers to how large a network can become without needing to be fundamentally restructured.
68. The most common SAN topology is \_\_\_\_\_.
69. \_\_\_\_\_ allows native FC ISLs to cross very long distances while maintaining full performance.
70. The rule of thumb is that it takes one \_\_\_\_\_ per kilometer of distance for full-speed 2Gbit operation.
71. Performance in a network will \_\_\_\_\_ over time.
72. \_\_\_\_\_ are the most common performance limiting factor in a SAN.
73. The mechanism which carries traffic across a SAN between edge devices is known as the SAN \_\_\_\_\_. FC

and iSCSI are two examples.

74. \_\_\_\_\_ is a condition in which more devices *might* need a resource than that resource can serve.
75. \_\_\_\_\_ is a condition in which devices actually *are* trying to use a path beyond its capacity, so some of the traffic destined for that path must be delayed.
76. \_\_\_\_\_ refers to a queuing problem, not merely to contention for bandwidth on a link.
77. \_\_\_\_\_ is how long it takes to forward a frame.
78. \_\_\_\_\_ is often matched to the ratio of storage to hosts.
79. Using the \_\_\_\_\_ product will help to automate UC and other advanced solutions by managing the complex relationships between hosts, storage, operating systems, and applications.
80. \_\_\_\_\_ is the practice of optimizing traffic by putting ports that communicate “close” together.
81. \_\_\_\_\_ is the practice of connecting hosts to one group of switches, and storage to a different group.
82. \_\_\_\_\_ are two features which allow traffic to be balanced across ISLs while preserving in order delivery.
83. The process of taking a design from paper all the way through release to production is \_\_\_\_\_.
84. Avoid single points of failure when selecting racks for switches by \_\_\_\_\_.
85. The most effective access control mechanism for a SAN is \_\_\_\_\_, because it is enforced by both the Name Server and the ASIC.
86. It is important to \_\_\_\_\_ a SAN before releasing it to production to verify that all switches, routers, devices and applications are capable of recovering from faults.

87. Maintaining a \_\_\_\_\_ can help with tasks such as switch and fabric maintenance, troubleshooting, and recovery.
88. Users interested in clean, stable fabric environments should run \_\_\_\_\_ regularly.
89. It is possible to use the \_\_\_\_\_ product to optimize storage performance at branch offices.
90. When evaluating candidate SAN designs, it is appropriate to consider which of the following factors:
  - a. Compatibility
  - b. RAS
  - c. Scalability
  - d. Performance
  - e. Manageability
  - f. Total solution cost
  - g. All of the above
91. Any SAN design should meet or exceed all requirements, but most designers consider \_\_\_\_\_ to be the most important consideration when making trade-offs.
92. If a fabric has a single point of failure, and the SAN has only one fabric in it, then the overall architecture is considered to be \_\_\_\_\_.
93. Connecting a host to the same switch as its primary storage is an example of the use of \_\_\_\_\_.
94. ILM and UC are two trends which are likely to increase the use of \_\_\_\_\_ fabric topologies, in which hosts are connected to one group of switches and storage to a different group.
95. To maximize fabric scalability, compatibility, and reliability, when planning zoning for a fabric it is best to zone HBAs so that:
  - a. All HBAs accessing a given storage port are in the same zone.

- b. Hosts with a common OS type are all zoned together, and separated from all other OSs.
  - c. Each HBA is in its own dedicated zone.
  - d. All devices in the fabric are in one zone.
  - e. If possible, zoning should be avoided, since it is hard to manage.
96. If every switch in a fabric is directly connected to every other switch, this is an example of a \_\_\_\_\_ topology.
97. The most reliable way to connect fabrics across MAN or moderate WAN distances is by using \_\_\_\_\_ connections, either over dark fiber or xWDM equipment.
98. The FCIA has approved the \_\_\_\_\_ line rate, which has now replaced 2Gbit as the basic rate for FC fabrics.
99. Dividing a director into two or more partitions - using zoning, VSANs, or a similar scheme such as the dual-domain capability of a Brocade director - will make it into a highly available system. (True/False)
100. Some of the options available for increasing the performance of a fabric include \_\_\_\_\_.
101. It is necessary for a SAN designer or project manager to prepare and maintain proper \_\_\_\_\_ to ensure that future administrators will know what has been done and why various decisions were made.
102. The simplest fabric design is the \_\_\_\_\_ topology, but this is only suitable for very small deployments, due to its limited scalability, performance, and reliability.
103. Proper use of zoning will improve fabric services scalability and reliability through Brocade's automatic use of \_\_\_\_\_ scoping.
104. The maximum number of ports currently supported by Brocade inside a single-domain director is \_\_\_\_\_. The smallest switch offered by Brocade has \_\_\_\_\_ ports.

105. The single biggest factor in determining how vulnerable a SAN is to DoS attacks or failures is whether or not the SAN uses a \_\_\_\_\_ design.

## **Ответы**

- 106. block
- 107. Fibre Channel (FC)
- 108. Directly Attached Storage (DAS)
- 109. Ethernet and IP ; Fibre Channel
- 110. storage consolidation
- 111. Brocade
- 112. Information Lifecycle Management (ILM)
- 113. shared bandwidth
- 114. Redundant (A/B) fabrics
- 115. congestion
- 116. Utility Computing (UC)
- 117. SAN design ; implementation plan
- 118. Storage Area Network (SAN)
- 119. Return on Investment (ROI)
- 120. Fibre Channel ; iSCSI
- 121. gather business-oriented requirements
- 122. SAN Project Manager
- 123. Backup, restore, and LAN performance
- 124. HA clusters
- 125. FCP ; iSCSI
- 126. Fibre Channel ; Gigabit Ethernet ; FC-0 and FC-1
- 127. Fabric Shortest Path First (FSPF)
- 128. backup window
- 129. SCSI
- 130. Host Bus Adapter (HBA)
- 131. iSCSI hardware accelerated HBA
- 132. Utility Computing (UC) and Information Lifecycle Management (ILM)
- 133. 1Gbit, 2Gbit, 4Gbit
- 134. 133Mbaud, 266Mbaud, 531Mbaud, 8Gbit, 10Gbit
- 135. ILM and UC
- 136. Name Server
- 137. data migration
- 138. Fibre Channel Arbitrated Loop (FC-AL)
- 139. frame-level trunking ; Dynamic Path Selection (DPS)

- 140.NFS ; CIFS
- 141.Disaster Tolerance (DT), Disaster Recovery (DR), or Business Continuity and Availability (BC&A)
- 142.storage consolidation
- 143.UC
- 144.SAN
- 145.Redundant Array of Independent Disks (RAID)
- 146.Protocols
- 147.compute power
- 148.TCP/IP
- 149.On-time and in-order
- 150.Storage-related services, such as FC fabric services
- 151.two or more
- 152.iSCSI
- 153.SFPs, GBICs, or other similar laser media
- 154.ILM
- 155.Lowering capital costs, increasing management efficiency, and improving application performance
- 156.gathering requirements, developing technical specifications, estimating cost, performing an ROI analysis, and creating a detailed design and rollout plan
- 157.clients, LAN, compute nodes, SAN, storage
- 158.SAN
- 159.business-oriented
- 160.connected to each other directly or across a network
- 161.irrelevant
- 162.all anticipated future increases in performance demand
- 163.Reliability
- 164.Availability
- 165.Serviceability
- 166.Fabric Manager
- 167.SAN Health
- 168.Scalability
- 169.Core/Edge (CE)
- 170.Extended Fabrics
- 171.BB credit
- 172.increase

- 173. Hosts and storage devices
- 174. protocol
- 175. Over-subscription
- 176. Congestion
- 177. Blocking, or “Head of Line Blocking” (HoLB)
- 178. Latency
- 179. ISL over-subscription
- 180. Tapestry Application Resource Manager (ARM)
- 181. Locality
- 182. Tiering
- 183. Frame-level trunking and exchange-level Dynamic Path Selection (DPS)
- 184. SAN implementation
- 185. separating redundant fabrics into different rack and providing separate power grids and UPSs
- 186. hard zoning
- 187. stage and validate
- 188. configuration log
- 189. SAN Health
- 190. Tapestry Wide Area File Services (WAFS)
- 191. “G”; all of the above
- 192. Application availability
- 193. Non-resilient and non-redundant
- 194. Locality
- 195. Tiered
- 196. “C”; each HBA should have its own zone
- 197. full mesh
- 198. Native FC
- 199. 4Gbit
- 200. False – One of anything is not HA
- 201. adding ISLs or IFLs, increasing line rates, using trunking and/or DPS, localizing flows
- 202. SAN documentation
- 203. cascade
- 204. Registered State Change Notification,(RSCN)
- 205. 256; 8
- 206. redundant (A/B) fabric



## **Приложение D:**

### **Часто задаваемые вопросы**

---

**Q:** What SAN planning process does Brocade use?

**A:** There are five phases in the recommended SAN planning process: gather the requirements of the SAN through interviews, develop preliminary technical specifications, estimate the project cost, calculate ROI, and finally create a detailed SAN design and rollout plan.

**Q:** What is a SAN project plan?

**A:** The SAN Project Plan may be very similar to other IT project planning tools used within your company. The key items it should include are: notes and documents to support collected data such as interviews and device surveys; interpretations of the data; the design which emerges from the data; a list of required equipment and associated costs; a plan for implementing, testing, releasing to production, and managing the SAN.

**Q:** Generally, who is included on the project team?

**A:** The SAN Project Manager and SAN Designer are arguably the two most important roles. The project manager will coordinate the effort and the designer will translate business needs into technical requirements. It is not uncommon for both roles to be accomplished by the same person. The technical team will consist of SAN Administrators, System Administrators, Storage Administrators,

IP Network Administrators, Database Administrators and Application Specialists. The members of the team should have a strong interest in, or have decision making authority related to the project.

**Q:** What is the difference between a business requirement and a business problem?

**A:** A business *problem* is a statement about what needs to be “fixed” or at least improved to help the organization accomplish its mission. For example, “Backups are interfering with customer service.” A business *requirement* will state a direction for the solution to one or more business problems, and can be used as a guideline for choosing the appropriate solution. For example, “The SAN must complete the backup in no more than  $x$  hours, and remain online during the process. This will save \$ $y$  by increasing productivity.”

**Q:** What should be included in business requirements?

**A:** Be sure to gather *specific* business requirements, with each requirement statement including *what* needs to happen, *when* it needs to happen, and how much *money* or *mission impact* is involved if the requirement is not met. This answers what, when, and why. “How” is answered by a subsequent step. “Where” is generally self-evident.

**Q:** How do I develop technical specifications for a SAN?

**A:** The specification document will be created in the planning phase. A number of factors must be taken into consideration in addition to the business requirements statement. The locations of SAN equipment, the mechanisms for connecting the locations together, estimated bandwidth, uptime, and the number of attached devices must all be analyzed when creating the specifications document.

**Q:** How do I justify my project?

**A:** As part of the ROI analysis you will have to produce

an estimated net benefit. This is done by subtracting the estimated cost of equipment from the projected gross benefits. The projected benefit may include things like increased productivity, lower management costs, reduced capital spending, and revenue gains. This task may be best suited for your accounting department, or at least should be taken on in partnership with them.

**Q:** What is the most commonly used SAN technology?

**A:** Fibre Channel. Period.

**Q:** iSCSI is supposed to be cheaper, but there do not seem to be many real-world deployments. Why is it not being used extensively?

**A:** Although many vendors, including Brocade, offer iSCSI solutions, it is an immature and unreliable protocol with marginal ROI and many hidden costs. FC products have had price reductions which eroded the iSCSI value proposition, and serial ATA is available in the low end market. This is “squeezing” out iSCSI from both ends of the market, and its long-term viability is now in question.

**Q:** What is the difference between an ISL and an IFL?

**A:** An Inter-Switch Link, or ISL, is the connection between two FC switches in a fabric. An Inter-Fabric Link, or IFL, is the connection between an FC switch and an FC-FC router. LSANs cross IF Ls. An IFL allows traffic to flow between different fabrics in a Meta SAN, whereas an ISL allows traffic and services to flow between switches within a single fabric.

**Q:** How can SANs be extended over long distances?

**A:** There are many options to extend a FC network over long distances including SONET/SDH, xWDM, ATM, and native FC over dark fiber. With limited solutions, IP may also be an option. Both ATM and SONET/SDH solutions have very high performance and reliability

compared to IP SAN solutions, but also tend to cost more.

**Q:** What services do Fibre Channel switches provide?

**A:** Unlike IP SAN switches, all Brocade FC switches have a robust group of built-in services. Fabric services include a name services, management services, high-speed routing services, auto-discovery and configuration, and so on.

**Q:** What is driving the increased Fibre Channel speeds?

**A:** There are always increasing demands for performance in networking. One example is the need to reduce backup windows. Another is the increasing need for high-speed long-distance connections to support disaster recovery. ILM and UC architectures are also drivers.

**Q:** Will my SAN support HA clustering?

**A:** All modern clustering methods have one thing in common: in order for one node to be able to take over an application if another node fails, it needs to have access to the data set that the failed node was using just before the crash. As long as your SAN provides that connectivity, it should be a good basis for building HA clusters.

**Q:** What is SAN implementation?

**A:** This is the process of taking your “paper” design to physical setup, through staging and testing, all the way through release to production.

**Q:** I am designing dual fabrics, what are the implementation considerations?

**A:** The concept of dual fabrics is to avoid any single point of failure. For high-availability fabrics, ensure that you have separate power circuits available, and mount redundant devices into different racks.

**Q:** What is the difference between hard and soft zoning?

**A:** Hard zoning is enforced by ASICs, while soft zoning is enforced by the name server. All Brocade plat-

forms shipped since about the turn of the century support some form of hardware zoning in all usage cases. Older switches supported hardware zoning only when zones were defined by PID.

**Q:** How do I prepare my SAN to go into production after it has been cabled and configured?

**A:** Prior to transitioning your fabric to production, it is important to validate the SAN by establishing a profile and injecting faults into the fabric to verify that the fabric and the edge devices are capable of recovering.

**Q:** Will keeping a change management log be helpful?

**A:** A diligently maintained configuration log can help you with many tasks such as switch and fabric maintenance as well as troubleshooting and recovery.

**Q:** Zoning is backed up to every switch, but what about the rest of the configuration parameters?

**A:** The best-practice is to create a backup of each switch configuration on a host when implementing a new SAN, changing a switch configuration, or adding or replacing a switch in the SAN.

**Q:** With so many protocols available, which should be used in my SAN?

**A:** Fibre Channel is the dominant SAN transport because of the importance for even lower-tier storage networks to have high performance and reliability. Brocade supports other options, but FC should be the default choice unless there is a comprehensive business case showing why another option should be used, and proving that it will actually work properly.

**Q:** What are common performance limitations in a SAN?

**A:** SAN attached devices, the SAN protocol, and link speeds are usually the bottlenecks.

**Q:** What is the impact of protocol selection on the SAN?

**A:** It affects performance, reliability, scalability, manageability, cost, and indeed most other aspects of SAN design. The best approach is to use a protocol with a long and proven track record of production deployment.

**Q:** My SAN will initially be used as a low-end SAN but I would like to scale in the future, is Fibre Channel an appropriate choice?

**A:** Fibre Channel networks can be configured to meet any performance requirement. Also, Brocade SANs can be designed to scale and for investment protection.

**Q:** What are some of the cost issues should think about when designing ISLs and IFLs?

**A:** The cost to performance ratio is probably the most obvious, but some designers may forget to consider the total cost of a connection. This means the cost of cables and connectors. It also means the cost of downtime if redundant links are *not* used, and the cost of productivity of links are allowed to congest massively.

**Q:** What is over-subscription?

**A:** Over-subscription refers to a condition in which more devices might need to access a resource than that resource could fully support. In many instances, oversubscription is deliberately engineered into a SAN to reduce cost.

**Q:** Does over-subscription cause congestion?

**A:** No. However, it does create the potential for congestion. Congestion is a condition in which devices are actually trying to use a path beyond its capacity, so some of the traffic destined for that path must be queued and transmitted after a delay.

**Q:** What can I do to avoid congestion in my SAN?

**A:** The most common approaches for dealing with congestion include using locality, faster links such as 4Gbit

or 10Gbit interfaces, or using hardware trunking to broaden link speeds into higher path rates.

**Q:** Do Brocade switches have Head of Line Blocking?

**A:** No. Head of Line Blocking occurs on poorly designed switches. Brocade does not ship products which are capable of exhibiting this misbehavior. However, other SAN infrastructure vendors do.

**Q:** How do Brocade switches have such low latency?

**A:** Brocade uses “cut-through routing” which allows a frame to be transmitted out the destination switch port while it is still being received into the source port.

**Q:** How do I determine the amount of bandwidth will be required for any given path?

**A:** Analyze how much data each application will need to move over that path, and then apply one of several calculation methods. For example, it is possible to add up all application peak loads, or to take their average loads, or simply to apply a rule of thumb such as using the ratio of hosts to storage ports.

**Q:** In addition to increasing SAN performance what other benefits does locality provide?

**A:** Locality improves RAS as there are fewer links and therefore fewer total components in the network, thus reducing cost and improving reliability numbers like MTBF.

**Q:** Do Brocade switches offer load balancing?

**A:** Brocade switches have an option that allows FSPF to reallocate routes whenever a fabric event occurs. This feature is called Dynamic Load Sharing (DLS) because it allows routes to be reselect dynamically under conditions that can still guarantee in-order delivery. Also, Brocade platforms support one or more forms of hardware trunking.

**Q:** Does trunking work well over long distances?

**A:** Yes, although different trunking methods work over different distances, or work best in different ways.

**Q:** What factors affect compatibility?

**A:** Protocols, frame formats, node-to-node compatibility, node-to-switch storage service behaviors, switch-to-switch services exchange.

**Q:** How important is it to plan for future expansion?

**A:** Always consider performance and scalability requirements of the initial deployment, *and* all anticipated future increases in demand. Network requirements tend to increase rather than decrease over time, and so all SAN protocol and topology choices should be able to accommodate a wide range of scenarios.

**Q:** What can impact SAN performance?

**A:** Areas to consider when thinking about SAN performance include protocols, link rates, congestion, blocking, and latency.

**Q:** Should I be more concerned with congestion or blocking?

**A:** Congestion does not stop communication between endpoints entirely; it just slows it down somewhat for a period of time. Blocking, more properly called Head of Line Blocking (HoLB), can actually stop communication for an extended period of time and is therefore an area of concern. Brocade does not sell any product which exhibits HoLB and any such product should be avoided.

**Q:** How should I prioritize RAS?

**A:** Application availability is the most important consideration in SAN designs overall because an availability issue can have an impact at the end-user level. Reliability should be considered second because of the potential impact of a failed component to the SAN. Serviceability is

usually of least concern; however it should be considered.

**Q:** What SAN management tasks should be expected on a day to day basis?

**A:** Day-to-day management tasks generally include monitoring the health of the network, and performing adds, moves, and changes to the SAN itself and to the attached hosts and storage devices. Using Fabric Manager will simplify tasks associated with coordinating day-to-day management of multiple fabrics. SAN Health will vastly simplify proactive management, since it automatically checks the SAN against evolving best-practices and has automated “housekeeping” features such as looking for unused zones.

**Q:** When planning my SAN for scalability, what is the best approach?

**A:** To maximize the scalability of a SAN, it is always best to break it down into smaller fabrics. Use an A/B redundant model first, then split off other fabrics by function, geographical location, administrative groups, or by spreading storage ports.

**Q:** When planning for scalability, what limitations should be considered in the SAN design?

**A:** Limitations can be classified into five categories: manageability, fault containment, vendor support matrices, storage networking services, and the protocol itself.

**Q:** Which topologies are the most commonly used?

**A:** Just a few topologies are typically used as the basis for SANs, and these are combined or varied to fit the needs of specific deployments. The most common topologies for SANs include cascades, rings, meshes, and various core/edge designs.

**Q:** What is the best way to prevent denial of service attacks against a SAN?

**A:** It is never possible to make a system completely proof against deliberate or accidental DoS attacks. However, it is possible to make such events far less likely. Following security best-practices is a good start. Implementing sound management procedures helps, too. However, the single biggest factor in determining vulnerability to this form of attack is whether or not the SAN uses physically isolated redundant fabrics, with redundant HBA connections.

**Q:** What is the best long-distance method in a SAN?

**A:** Extended native Fibre Channel ISLs or IFLs over long distances are generally the easiest extension solutions to manage and have the highest performance. Long distance ISLs require that the SAN designer have an understanding of buffer to buffer credits (BB credits).

**Q:** What are buffer to buffer credits (BB credits)?

**A:** In order to prevent frames from dropping, no port can transmit frames unless the port to which it is directly communicating has the ability to receive them. It is possible that the receiving port will not be able to forward the frame immediately, in which case it will need to have a memory area reserved to hold the frame until it can be sent on its way. This memory area is called a buffer. All devices in a SAN have a limited number of buffers, and so they need a mechanism for telling other devices if they have free buffers before a frame is transmitted to them. This mechanism is the exchange of BB credits.

**Q:** How do BB credits impact long distance links?

**A:** When using FC over long distance links, BB credits become important. The rule of thumb is that it takes one credit per kilometer for full-speed 2Gbit operation. Given a fixed number of BB credits, a link can go twice as far at 1Gbit as with 2Gbit. With 4Gbit links, twice as many buffers per kilometer are required as with 2Gbit links. However, it is important to note that all currently

shipping Brocade platforms support more BB credits than are needed to go the maximum distance supported by today's optical components. Realistically, it is necessary to move to an DWDM architecture to go beyond a hundred kilometers or so, regardless of how many credits a switch can supply, and the leading DWDM vendors also provide a credit mechanism which supersedes that of the switches. Note that BB credits do not apply to FCIP or other protocol tunneled links in any significant way.

# G

## Словарь

**Access Gateway (Шлюз доступа)** Использует NPIV для соединения встроенного коммутатора с шасси блейд-серверов в фабрику по методу N\_Port, а не E\_Port

**AL\_PA (Arbitrated Loop Physical Address)** – адрес, используемый для идентификации устройства в петле arbitrated loop.

**American National Standards Institute** Смотри ANSI

**ANSI** American National Standards Institute – государственный институт по стандартам США. Комитет ANSI T11 занимается стандартами FC.

**AP (Application Platforms – Платформа приложений)** – обеспечивает такие основанные на фабрике приложения хранения, как зеркалирование, миграция данных, мгновенные снимки, виртуальные ленты и т.п.

**API (Application Programming Interfaces – Интерфейсы прикладного программирования)** – реализуют слой абстрагирования между сложными низкоуровневыми процессами и средой разработки приложений верхнего уровня. Они упрощают создание сложных приложений, предоставляя программистам отдельные «строительные блоки».

**Application Platform** Смотри AP

**Application Programming Interface** Смотри API

**Application-Specific Integrated Circuit** Смотри ASIC

**Application Resource Manager** (инфраструктура управления ресурсами приложения) Инфраструктура управления, включающая программное обеспечение и оборудование для реализации определенных функций ресурсных вычислений (Utility Computing) в среде Brocade SAN. Также называется Tapestry Application Resource Manager или Tapestry ARM.

**Arbitrated Loop** Совместно используемый транспорт Fibre Channel, теоретически поддерживающий до 126 устройств

**ARM** См. Application Resource Manager

**ASIC** (Application -Specific Integrated Circuit) – специализированные микросхемы, разработанные для выполнения конкретных функций

**Asynchronous Transfer Mode** см. ATM

**ATM** Asynchronous Transfer Mode – транспорт с коммутацией ячеек для передачи данных через сети CAN, MAN и WAN. ATM передает короткие блоки данных и имеет более высокую производительность и надежность, чем коммутация IP.

**Backbone Fabric** см. BB Fabric

**Bandwidth (пропускная способность)** Скорость, с которой линк или система могут передавать данные.

**BB\_Credit** Кредиты Buffer-to-buffer – механизм управления потоками, который определяет, сколько пакетов можно переслать получателю через определенный порт

**BB Fabric** (Backbone Fabric) Маршрутизация FCR позволяет соединять маршрутизаторы в опциональную Backbone Fabric для построения более масштабируемых и гибких Meta SAN. Маршрутизаторы подключаются к фабрике BB Fabric через порты E\_Ports.

**Bloom ASIC** третьего поколения для коммутаторов Brocade F C. Основан на двухчиповой 16- портовой архитектуре центральной памяти. Использовались в коммутаторах SilkWorm 3000 и 12000, а также во встроенных продуктах (например, RAI D-контроллерах), производимых OEM- партнерами Brocade. Все порты поддерживают 1Gbit и 2Gbit FC.

**Bloom-II** Усовершенствованная версия Bloom. Потребляет меньше энергии и выделяет меньше тепла, несколько улучшены функции управления буферами для территориально-распределенных сетей. Использовались в SilkWorm 3250, 3850, 24000 и встроенных продуктах OEM-партнеров Brocade.

**Broadcast (широковещательная передача)** Пакеты передаются все узлам фабрики.

**Bridge (мост)** Соединяет сегменты одной сети

**Brocade** Основанная в 1995 году, компания Brocade быстро стала ведущим поставщиком коммутаторов Fibre Channel. В настоящее время компания производит коммутаторы, директоры и многопротокольные маршрутизаторы.

**Buffer-to-Buffer Credits** См. BB\_Credit

**CAN Campus Area Networks** (кампусные сети) – обычно их размеры составляют около 1 километра или меньше. Отличаются от локальных сетей LAN, размеры которых обычно находятся в диапазоне

около 100 метров и, что важнее, сети CAN охватывают несколько зданий.

**Carrier Sense Multiple Access with Collision Detection**  
См. CSMA/CD

**Class Of Service** См. COS

**CLI** Command Line Interface (интерфейс командной строки) – текстовый способ управления устройствами. FCR использует интерфейс командной строки Brocade Fabric OS, что упрощает обучение администраторов.

**Coarse Wave Division Multiplexer** See CWDM

**Command Line Interface** См. CLI

**Condor** ASIC четвертого поколения для фабрик Brocade FC. Используют архитектуру центральной памяти с одним чипом и 32 портами. Использовались в коммутаторе Brocade 4100. Все порты поддерживают 1Gbit, 2Gbit и 4Gbit FC. Может использоваться вместе с Egret для 10Gbit FC.

**COS** Class Of Service (класс сервиса) обозначает качество связи, включая такие характеристики, как задержки и скорость передачи данных.

**CRC** Cyclic Redundancy Check (циклическая проверка избыточности) – механизм самотестирования для обнаружения и исправления ошибок. Для защиты от ошибок все Brocade A SIC выполняют проверку CRC для всех пакетов

**Credit (кредит)** Количественное представление максимального числа буферов-получателей, предоставляемых портом F/FL\_Port подключенному к нему порту N/NL\_Port для того, чтобы при передаче пакетов через N/NL\_Port не произошло переполнение F/FL\_Port.

**CSMA/CD** Carrier Sense Multiple Access with Collision Detection (множественный доступ с контролем несущей и обнаружением столкновений) – определяет, как будут вести себя сетевые контроллеры (NICs) когда два или более контроллера пытаются одновременно использовать общий сегмент

**CWDM** Coarse Wave Division Multiplexer - грубое волновое мультиплексирование - технология передачи данных, позволяющая одновременную передачу различных потоков данных по одной паре оптических волокон. См. Также WDM и DWDM.

**Cyclic Redundancy Check** См. CRC

**Dark Fiber (темная оптика)** Арендуемый волоконно-оптический кабель между площадками без каких-либо сервисов от провайдера – все сервисы обеспечивает клиент.

**DAS** Direct Attached Storage (подключение устройств хранения напрямую) – метод подключения устройства хранения непосредственно только к одному хосту. В корпоративный ЦОДах вместо DAS используются сети хранения, но DAS по-прежнему применяется в персональных компьютерах и хостах начального уровня, хотя появление недорогих Fibre Channel HBA скорей всего полностью исключит такое использование.

**Denial of Service** См. DoS

**Dense Wave Digital Multiplexer** См. DWDM

**Destination Fabric ID** См. DFID

**Destination Identifier** См. DID

**DID** Destination Identifier (идентификатор получателя) – трехбитовый адрес Fibre Channel для задания физического местоположения получателя

пакета – домен коммутатора, порт коммутатора или место в петле – если получатель находится в петле. DID равный 010100 обозначает домен 1, порт 1 и отсутствие петли). Обычно обозначается шестнадцатеричными числами.

### **Direct Attached Storage** См. DAS

**DLS** Dynam ic Load Sharing (динамическое распределение нагрузки) обеспечивает перерасчет маршрутов при выключении или включении портов

**Domain ID** Уникальный номер от 1 до 239, идентифицирующий коммутатор FC, порт маршрутизатора или адрес перенаправления фабрике

**DoS** Denial of Service – атака типа «отказ в обслуживании». Может намеренно вызываться хакером или вирусом либо возникает случайно. В любом случае приводит к нарушению нормальной работы. Лучший способ защитить SAN от DoS – это следовать лучшим практикам (best-practices) безопасности и использовать резервирование фабрик (A/B).

**DWDM** Dense W ave Digital M ultiplexer – плотное волновое мультиплексирование. См. также WDM и CWDM. Обеспечивает больше длин волн, чем CWDM.

### **Dynamic Load Sharing** См. DLS

**E\_D\_TOV** Error-Detect Tim e Out Value – максимальное время ожидания подтверждения получения пакета, по истечению которого считается, что возникла ошибка

**E\_Port** Expansion port (порт расширения), через который два коммутатора соединяются для формирования фабрики. Соединение портов E\_Port

образует ISL. Порты E\_Port также могут соединяться с портами EX\_Ports для формирования IFL.

**Edge Fabric (периферийная фабрика)** Фабрика Fibre Channel, соединенная посредством FCR через порты EX\_Port с центральной фабрикой (примерно также периферийные коммутаторы подключаются к центральным коммутаторам в фабрике Core-edge). Через нее хосты и устройства хранения подключаются в Meta SAN.

**Egret** ASIC Brocade, в которой три внутренние линки 4Gbit балансируются и преобразуются в один внешний линк 10Gbit.

**ELWL** Extended Long Wavelength Laser - Трансиверы с расширенными длинноволновыми лазерами ELWL могут использовать лазеры 1550 nm . Они используются для обеспечения соединения Fibre Channel на расстояния, которые не поддерживают даже лазеры LWL. Обычно эти трансиверы используют кабели SMF.

**Error-Detect Time Out Value** См. ED\_TOV

**Ethernet** Широко используемый стандарт сетей IEEE 802.3. Ethernet – это протокол LAN, который поддерживает скорость передачи данных до 10Mbps. Он использует CSMA/CD для доступа к общей среде передачи данных. Fast Ethernet поддерживает скорости передачи данных до 100 Mbps, а Gigabit Ethernet - до 1 Gbps. Сейчас появился стандарт для 10Gbps.

**EX\_Port** Enhanced E\_Port используется для подсоединения маршрутизатора к периферийной фабрике. С точки зрения коммутатора периферийной фабрики, порт EX\_Port практически не отличается от E\_Port. Он соответствует тем же стандартам Fibre Channel, что и другие порты Brocade E\_Port.

Однако, маршрутизатор терминирует порты EX\_Port, не позволяя различным фабрикам объединяться так, как они объединялись бы через обычные порты E\_Port. Каждый порт EX\_Port использует набор псевдо-доменов для представления удаленных фабрик, каждая с «подключенными» прокси-устройствами для представления устройств этих фабрик.

**Exchange** (Обмен) Механизм самого высокого уровня FC для обмена данными между портами N\_Port. Exchange состоит из одной или нескольких связанных между собой последовательностей передачи данных.

**Expansion Port (порт расширения)** См. E\_Port

**Exported Device** Узлы одной фабрики могут экспортироваться в другие фабрики с помощью маршрутизатора FC и механизма зон LSAN. Если нужно экспортовать один узел из одной фабрики в другую, то надо дать команду “Хост экспортируется из Fabric 1 и импортируется в Fabric 2.”

**Extended Long Wavelength Laser** См. ELWL

**F\_Port** (Fabric Port) Порт Фабрики на коммутаторе, к которому можно подключить такой порт N\_Port, как например HBA

**Fabric** Фабрика (1) Топология Fibre Channel, которая образуется при подключении портов N\_Port к портам коммутатора F\_Ports. (2) Один или несколько коммутаторов Fibre Channel в сетевой топологии ISL. (3) Набор ISL, соединяющих коммутаторы Fibre Channel и другие устройства (хосты и устройства хранения). (4) Программное обеспечение Fabric Services, состоящее из Storage Name Server, Management Server, FSPF routing и т.п.

**Fabric Identifier** См. FID

**Fabric Loop Port** См. FL\_Port

**Fabric Operating System** См. FOS

**Fabric Port** См. F\_Port

**Fabric Shortest Path First** См. FSPF

**FC** Fibre Channel - протокол построения сетей хранения данных SAN. В отличие IP и Ethernet протокол FC с самого начала разрабатывался для поддержки всех типов устройств хранения.

**FC-0** Физический уровень Fibre Channel

**FC-1** Уровень кодирования Fibre Channel. Это означает использование 8b/10b с 1G, 2G или 4G; 64b/66b используется с 10G.

**FC-2** Обработка пакетов, протокола, управление последовательностями/обменами и упорядоченными наборами (ordered sets) для Fibre Channel

**FC-3** Общие сервисы для Fibre Channel

**FC-4** Отображение таких протоколов верхнего уровня (ULP), как SCSI или IP, в FC

**FC-FC Routing Service** (также называется FCR service). Реализует иерархическое управление для фабрик Fibre Channel, позволяя создавать L SAN так, чтобы устройства из разных фабрик могли обмениваться данными, но сами фабрики при этом не объединялись. См. также FCR.

**FCIP Tunneling Service** FCIP – основанный на TCP/IP протокол туннелирования для прозрачного соединения территориально распределенных фабрики через IP-сеть. Это позволяет расширять SAN на расстояния, которые не поддерживают «родные»

линики FC. Этот сервис позволяет отображать E\_Port через прозрачный туннель FCIP на другой шлюз и коммутатор FCIP на дальнем конце. Порт одновременно является и E\_Port, и FCIP.

**FC-NAT** Fibre Channel Network Address Translation обеспечивает коммуникацию между устройствами фабрик даже если в фабриках совпадают имена устройств, аналогично механизму NAT сетей передачи данных.

**FCP** Протокол Fibre Channel для отображения SCSI в Fibre Channel. Сейчас, по-видимому, самый популярный протокол верхнего уровня для сетей хранения.

**FCR** Fibre Channel Routers (маршрутизаторы) – платформы, реализующие сервис маршрутизации FC-FC. На платформе FCR может также работать другое программное обеспечение, поэтому, в теории, одна платформа может одновременно работать и как FCR, и как туннель FCIP, и как шлюз iSCSI.

**FCRP** Fibre Channel Router Protocol - разработанный Brocade протокол для маршрутизации FCR между различными периферийными фабриками и, как опция, через фабрику backbone.

**Fibre Channel** См. FC

**Fibre Channel Router** См. FCR

**Fibre Channel Router Protocol** См. FCRP

**FID** Fabric ID – уникальный идентификатор фабрики в Meta SAN. См. также Global Header, SFID и DFID.

**Field Programmable Gate Array** См. FPGA

**Field Replaceable Unit** См. FRU

**Flannel** Первое поколение микросхем Brocade ASIC для преобразования из FC -AL в фабрику FC. Использовались в серии SilkWorm 1000 вместе с Stitch ASIC.

**FL\_Port** Порт Fabric loop, к которому подключается петля или устройство из петли. Это шлюз к фабрике для портов NL\_Ports петли loop.

**FOS** Brocade Fabric Operating System – это программное обеспечение, работающее на большинстве платформ Brocade. На момент написания этой книги последней версией была Fabric OS 4.x. (на момент перевода – 6.1). См. также XPath.

**FPGA** Микросхемы Field Programmable Gate Array похожи на ASIC, но логику FPGA можно программировать непосредственно у заказчика. Они обычно дороже ASIC и часто работают медленнее, но зато более гибкие.

**Frame** Единица данных, состоящая из ограничителя начала фрейма Start-of-Frame (SoF), заголовка, передаваемых данных, CRC (Cyclic Redundancy Check) и ограничителя конца фрейма End-of-Frame (EoF). Длина передаваемых данных может быть от 0 до 2112 байтов (при передаче через EX\_Port – не более 2048 байтов), а CRC состоит из 4 байтов.

**FRU** Field Replaceable Units – компоненты, которые могут заменять пользователи или сервисные инженеры

**FSPF** Fabric Shortest Path First (кратчайший путь в фабрике идет первым) – разработанный Brocade и принятый в качестве стандарта механизм маршрутизации между коммутаторами Fibre Channel в фабрике

**Full Duplex** (полный дуплекс) Одновременная передача и прием данных по одному каналу

**G\_Port** (Generic port) порт, выполняющий автоматическое согласование для поддержки E\_, F\_ или функциональности FL\_Port

**GBIC** Gigabit Interface Controller (или Converter) – съемные трансиверы оптика-медь, сейчас практически вытеснены трансиверами SFP.

**Generic Port** См. G\_Port

**Gigabit Interface Controller** См. GBIC

**Global Header** - это информация фабрики ВВ для идентификации устройств в контексте Meta SAN. Состоит из межфабричного заголовка с адресом (inter-fabric addressing header, IFA header) и обычного заголовка пакета FC-FS, содержащего SID и DID, соответствующим реальным PID устройства-отправителя и устройства-получателя. Невидима для коммутаторов в фабрике ВВ.

**HBA** Host Bus Adapter (HBA- адаптер) – интерфейс между шиной сервера или рабочей станции и Fibre Channel SAN

**Host Bus Adapter** См. HBA

**Hot Swappable** (заменяемый в горячем режиме) Компонент, который можно заменить без выключения системы

**IEEE** Institute of Electrical and Electronics Engineers – организация, утверждающая стандарты компьютерной индустрии

**IETF** Internet Engineering Task Force – рабочая группа, разрабатывающая протоколы для Internet

**iFCP** Internet Fibre Channel Protocol - стандарт, который предлагается использовать вместо FCIP для передачи трафика Fibre Channel через IP WAN. На время написания этой книги только один вендор внедрил этот протокол, а большинство используют FCIP.

**IFL** Inter-Fabric Link – соединения между маршрутизаторами и периферийными фабриками. Работают аналогично ISL. Могут соединять EX\_Port с E\_Port или EX\_Port с EX\_Port, хотя на практике применяются только второй вариант. См. также EX-IFL и EX<sup>2</sup>-IFL.

**ILM** Information Life cycle Management ( управление жизненным циклом информации) – концепция, согласно которой можно добиться соответствия информации и уровня хранения в зависимости от той ценности, которую представляет информация в данный момент

**In Order Delivery** См. IOD

**In-Band** Передача протокола управления или сервиса через транспорт Fibre Channel. FSPF и FCRP - это протоколы in-band.

**Initiator (инициатор)** Узел в сети Fibre Channel, который инициирует транзакцию на ленты или диски. См. также HBA.

**Information Lifecycle Management** См. ILM

**Institute of Electrical & Electronics Engineers** См. IEEE

**Inter-Fabric Link** См. IFL

**Internet Engineering Task Force** См. IETF

**Internet Fibre Channel Protocol** См. iFCP

**Internet Protocol** См. IP

**Internet Storage Name Server** См. iSNS

**Inter-Switch Link** См. ISL

**IOD In Order Deliver** – это параметр, который гарантирует, что пакеты будут доставлены получателю в том порядке, в каком отправлены. Если не удается сохранить порядок, то пакеты отбрасываются фабрикой.

**IP Internet Protocol** - это адресный компонент TCP/IP

**IPsec Internet Protocol Security** - набор протоколов, обеспечивающих безопасность на сетевом уровне. Часто используется при построении VPN для аутентификации узлов и/или шифрования данных.

**iSCSI Gateway Service** iSCSI – это инкапсуляция SCSI в транспорт IP.

**ISL Inter-Switch Link** – соединение двух коммутаторов через порты E\_Port

**iSNS Internet Storage Name Server** – это эквивалент в iSCSI для Fibre Channel SNS.

**JBOD Just a Bunch Of Disks** ( просто связка дисков); диски, обычно сконфигурированные в Arbitrated Loop внутри шасси

**Just a Bunch Of Disks** См. JBOD

**L\_Port Node Loop** порт, поддерживающий протокол FC\_AL

**LAN Local Area Network** ( локальная сеть) – сеть, в которой данные передаются на расстояние не больше 5 км.

**Latency** (запаздывание) Период времени, в течение

которого при передаче пакет задерживается сетевым устройством или идет по кабелю, соединяющему устройства. (второй тип запаздывания влияет только при передаче данных на большое расстояние).

**LED** Light Emitting Diode – светодиод, отображающий состояние устройства.

**Light Emitting Diode** См. LED

**Logical Storage Area Network** См. LSAN

**Loom** Микросхема FC ASIC второго поколения. Основана на дизайне центральной из четырех микросхем и 16 портов. Использовалась в коммутаторах серии SilkWorm 2000. Все порты - 1Gbit FC.

**LSAN** Логические SAN могут простираться между фабриками. Маршрут между устройствами в LSAN может лежать внутри фабрики или проходить через один или несколько маршрутизаторов FC и максимум через одну фабрику ВВ. Администрируются с помощью зон.

**LSAN Zone** Механизм для управления LSAN. Маршрутизатор FC, подключенный к двум фабрикам, будет образовывать соответствующие зоны LSAN в обеих фабриках. Если удастся успешно выполнить эту процедуру, то создаются домены-фантомы и соответствующие учетные записи FC-NAT в серверах имен фабрик. Зоны LSAN совместимы со стандартным механизмом зонирования. Их единственная особенность – зоны LSAN могут содержать имена WWN или псевдонимы WWN, начинающиеся с приставки LSAN\_.

**Local Area Network** См. LAN

**Long Wavelength Laser** См. LWL

**LUN** Logical Unit Number (логический номер устройства) – использует для идентификации разных устройств SCSI или томов, имеющих один SCSI ID. В Fibre Channel номера LUN используются для идентификации устройств или томов с одинаковым адресом WWN/PID.

**LWL** Long Wavelength Laser – трансивер, использующий лазер 1310nm . Такие трансиверы применяются для обеспечения передачи данных на большие расстояния по «родным» линкам FC. Обычно эти трансиверы используют кабели SMF.

**MAC** Media Access Control – это один из подуровней уровня OSI Data Link. Он отвечает за перемещение пакетов между NIC, используя общую среду передачи данных.

**MAN** Metropolitan Area Networks по своим масштабам занимает промежуточное положение между LAN и WAN. MAN может соединять несколько кампусов в одном городе или несколько соседних городов. Размер MAN может достигать нескольких десятков миль. MAN обычно полностью обслуживаются одним провайдером, а WAN – несколькими провайдерами услуг связи.

**Mean Time Between Failures** См. MTBF

**Mean Time To Repair** См. MTTR

**Media Access Control** См. MAC

**Meta SAN** Набор всех устройств, коммутаторов, периферийных и ВВ фабрик, LSAN и маршрутизаторов FC, образующих физически соединенную, но разделенную с помощью маршрутизаторов сеть хранения. LSAN соединяют с помощью FCR периферийные фабрики в Meta SAN и эти маршрутизаторы обеспечивают как

изоляцию, так и соединение периферийных фабрик. С точки зрения сетей передачи данных ее можно назвать «internetwork» или иногда просто «сеть».

### **Metropolitan Area Network** См. MAN

**MMF** Multimode Fiber – спецификация оптико-волоконного кабеля, позволяющего соединять устройства на расстоянии до 500 метров. Кабели MMF используют оптическое волокно 50 или 62.5 микрон. Обычно используются с трансиверами SWL.

**MTBF** Mean Time Between Failures - среднее время между сбоями любого компонента системы. Это показатель того, как часто систему нужно обслуживать. Сбой компонента системы не обязательно влечет нарушение доступности.

**MTTR** Mean Time To Repair - среднее время, которое уходит на ремонт неисправного компонента.

**Multicast** передача пакетов для части узлов фабрики (при unicast идет одному узлу, при broadcast – всем узлам). Часто используется в видеоприложениях.

### **Multimode Fiber** См. MMF.

**Multiprotocol** (многопротокольный) Устройство, способное использовать несколько протоколов. Например, если маршрутизатор оборудован интерфейсами Ethernet и Fibre Channel, то он считается многопротокольным.

**N\_Port** (Node Port) Порт хоста или устройства хранения Fibre Channel в фабрике или при подключении точка-точка.

### **Name Server/Service** См. SNS

**NAS** Network Attached Storage – общее название специализированный сетевых файл-серверов (обычно

использующих CIFS и/или NFS). Часто единственным отличием «файлера» NAS от, например, файл-сервера UNIX NFS является корпус устройства.

**Network Attached Storage** См. NAS

**Network Interface Card** См. NIC

**NIC** Network Interface Card – плата, соединяющая шину хоста с сетью. Аналогична НВА.

**NL\_Port** Порт Node Loop, поддерживающий протокол FC\_AL protocol

**Node Loop Port** См. L\_Port и NL\_Port

**NPIV** (N\_Port Id Virtualization) используется для агрегирования нескольких копий операционных систем в одном соединении N\_Port в фабрику, например, с помощью Access Gateway

**OEM** Original Equipment Manufacturers (производители оригинального оборудования) – компании, которые покупают продукты Brocade и интегрируют их с другими продуктами для хранения данных, например, дисковыми приводами, ленточными библиотеками и хостами, и продают их под своими торговыми марками.

**Open Shortest Path First** См. OSPF

**Original Equipment Manufacturer** См. OEM

**OSPF** Open Shortest Path First – это протокол динамической маршрутизации в IP сетях. С точки зрения протоколов маршрутизации IP это достаточно надежный протокол.

**PID** Port ID – трехбитовый физический адрес узла Fibre Channel в фабрике. PID делятся на три иерархические класса - Domain\_ID, Area\_ID и Port\_ID, которые Brocade использует для

идентификации домена коммутатора, порта коммутатора и FC-AL AL\_P А соответственно. Пример типичного PID: 010f00.

**Point-to-Point** Точка-точка выделенное соединение Fibre Channel между двумя устройствами, обычно между портами хоста и устройства хранения

**Port Identifier** См. PID

**Proxy Device** (прокси-устройство) также называется “xlate device” – представление того, как устройство видит фабрику, в которую это устройство экспортировано. PID прокси-устройств начинается с фантомного домена, представляющего фабрику.

**QoS** Quality of Service (качество сервиса) – обобщенный термин, описывающий механизм, гарантирующий соблюдение приоритетов, характеристики полосы пропускания, запаздывания, частоты ошибок и других характеристик канала передачи данных между узлами.

**Quality of Service** См. QoS

**R\_A\_TOV** Resource Allocation Time Out Value; максимальное время, в течение которого пакет может быть задержан в фабрике и все равно доставлен

**RAID** Redundant Array of Independent (прежний вариант Inexpensive) Disks – резервируемый массив независимых (недорогих) дисков. Набор дисков, который выглядит как один том. Используется несколько вариантов RAID, обеспечивающих разных уровень производительности хранения, масштабируемости и доступности, чаще всего с защитой от сбоев и/или высокой производительностью.

**RAS** Reliability Availability and Serviceability

(надежность, доступность и обслуживаемость) – общий показатель качества компонентов, устройства или сети. На RAS влияют показатели MTBF и MTTR компонентов, архитектура программного обеспечения и использование резервирования.

**Redundancy** (избыточность) Использование одного или нескольких резервных компонентов для обеспечения высокой доступности

**Redundant Array of Independent Disks** См. RAID

**Registered State Change Notification** См. RSCN

**Reliability Availability and Serviceability** См. RAS

**Resource Allocation Time Out Value** См. RA\_TOV

**RETMA** Radio Electronics Te levision Manufacturers Association – в контексте сетей хранения это спецификация стандартных стоек для ЦОДа. Типичное стоечное сетевое оборудование устанавливается в 19- дюймовом стойке стандарта RETMA и его высота определяет в единицах стойки RETMA (rack unit). Исторически сложилось, что один юнит считается равным около 1.75 дюйма

**Route** (1) Маршрут между двумя коммутаторами фабрики в контексте FSPF. (2) Маршрут между разными фабриками в Meta SAN в контексте FCRP.

**Router** (маршрутизатор) Устройство для связи двух и более разных сетей.

**RSCN** Registered S tate Chan ge Notifications – служебное сообщение, обеспечивающее оповещение узлов об изменениях в фабрике

**SAN** Storage Area Networks ( сеть хранения данных) связывает компьютеры с дисковыми массивами или ленточными библиотеками. Сейчас почти все SAN

построены на базе фабрик Fibre Channel.

**SAN Island** (остров SAN) Если одна SAN никак не связана с другими SAN, которые построены в компании, то она ее называют «островом» чтобы подчеркнуть ее изолированность. Острова можно объединить в большие фабрики или соединить с помощью маршрутизаторов FC-FC.

**SCR (State Change Registration)** используется устройствами для регистрации, необходимой для получения RSCN – сообщений.

**SCSI Small Computer System Interface** – первоначально это был набор протоколов для передачи больших блоков данных на расстояние 15 - 25 метров. Усовершенствованные версии этих протоколов - SCSI-2 - SCSI-3. При использовании вместо подключения напрямую сетевой модели, для передачи команд и пакетов SCSI используются такие протоколы, как FC и IP.

**SCSI Inquiry** Команда SCSI, на которую обычно устройство-получатель отвечает инициатору строкой, из которой можно узнать изготовителя этого устройства, модель устройства и версию микрокода. Ее использует сервер имен SNS для более полной идентификации устройств Fibre Channel. iSCSI Gate - way Service вставляет строки IP и IQN, пользуясь этим SNS полем.

**SDH** См. SONET/SDH

**Sequence** Последовательность взаимосвязанных пакетов, передаваемых между двумя портами N\_Port

**Serial** Последовательная передача бит данных по одной линии

**SFP Small Form Factor Pluggable** заменили GBIC в

качестве основного типа трансивера оптика-медь для оборудования Fibre Channel и Gigabit Ethernet, хотя в некоторых устройствах Gigabit Ethernet до сих применяются GBIC.

**SID** Source Identifiers трехбайтный физический адрес устройства отправителя пакета Fibre Channel, идентифицирующий домен коммутатора, порт коммутатора и место в петле (если устройство в петле). SID равный 010100 означает домен 1, порт 1 и отсутствие петли. Обычно обозначается шестнадцатеричными числами.

**SilkWorm** Зарегистрированная торговая марка семейства коммутаторов, директоров и маршрутизаторов Brocade. После ребрендинга в связи с приобретением McDATA этот бренд больше не используется для выпускаемых платформ Brocade.

**Simple Name Server** См. SNS

**Single Mode Fiber** См. SMF

**SMF** Single Mode Fiber – спецификация кабеля, обеспечивающую передачу данных на 10 км и больше. Кабели SMF сделаны из 9- микронного оптического волокна и обычно используются с трансиверами LWL или ELWL.

**Small Computer Systems Interface** См. SCSI

**SNS** Simple (или Storage) Name Server (или Service); Сервер Имен – сервис, предоставляемый коммутатором, который сохраняет имена, адреса и атрибуты объектов Fibre Channel. Также известен как сервис каталогов (directory service).

**SONET/SDH** Синхронные Оптические Сети (Synchronous Optical Networks) используются в MAN и WAN. Трафик FC может передаваться по

**SONET/SDH.** Обладает высокой производительностью и надежностью. В некоторых странах называется SDH (Synchronous Digital Hierarchy).

**Source Identifier** См. SID

**State Change Registration** См. SCR

**Stitch** Первое поколение ASIC для фабрик Brocade FC. Использовалось в серии коммутаторов SilkWorm 1000 вместе с Flannel.

**Устройства хранения** Запоминающее устройства с дисками или лентами

**Storage Area Network** См. SAN

**Storage Subsystem** См. Subsystem

**Storage Virtualization** См. Virtualization

**Subsystem** (подсистема) Синоним устройства хранения. Часто внешнее. В SAN может обслуживать несколько вычислительных узлов.

**SWL** (Short W avelength Laser) – трансиверы с коротковолновым лазером 850nm, передающие данные на небольшие расстояния. Наиболее часто встречающийся тип трансиверов.

**Synchronous Digital Hierarchy** См. SDH

**Synchronous Optical Networks** См. SONET/SDH

**T11** Комитет ANSI, разрабатывающий стандарты протоколов для перемещения данных к/от центральных компьютеров

**Tapestry** Торговая марка семейства продуктов Brocade высшего уровня, в том числе виртуализаторов. Использовалась до 2007 года. Далее использовалось

название File Area Networks (FAN)

**Target** Порт дискового массива или ленточного накопителя в SAN

**TCP/IP** Transmission Control Protocol over Internet Protocol – метод передачи данных в Internet

**TCP** Transmission Control Protocol - ориентированный на соединение протокол, преобразующий сообщения в пакеты, затем передающий пакеты по IP и собирающий их на другом конце. Обнаруживает ошибки и потерянные данные и при необходимости запускает повторную передачу.

**TCP Offload Engine** См. TOE

**Порт TCP** Адрес, по которому узел может обращаться к конкретному сервису. Протоколы верхнего уровня (например, HTTP) используют назначенные им определенные порты и web-серверы знают, что надо принимать пакеты через определенный порт, а web-клиенты посылают пакеты на этот порт.

**TOE** (TCP Offload Engines) используется в iSCSI NIC для разгрузки процессоров хоста. Даже самые мощные TOE NIC могут обеспечить не более половины скорости Fibre Channel HBA и обычно стоят дороже.

**Topology** (топология) Физическая, логическая или «фантомная» схема соединения устройств в сети

**Transceiver** (трансивер) Устройство, преобразующее сигналы из одной формы в другую для передачи и приема. Волоконно-оптические трансиверы преобразуют оптические сигналы в электрические.

**Transmission Control Protocol** См. TCP

**Tunneling** (туннелирование) Механизм связи двух однотипных сетей через промежуточную сеть другого типа

**UC Utility Com puting** – концепция Ресурсных Вычислений, согласно которой компьютерные ресурсы можно потреблять так же, как электроэнергию от электрической сети

**U\_Port** Universal Port – универсальные порты, которые могут работать как порты G/E/F/FL\_Port. Все коммутаторы Silkworm начиная с серии 2xxx используют Universal Port, что позволяет подключить любое устройство к любому порту. Выбор используемого типа порта происходит автоматически.

**ULP** Upper Level Protocols Протоколы более высокого по отношению к FC уровня, выполняемые поверх FC-4, например SCSI, IP и VI.

**Unicast** пересылка пакета между двумя точками. В отличие от broadcast и m ulticast может быть только один получатель.

**Universal Port** См. U\_Port

**Upper Level Protocol** См. ULP

**Utility Computing** См. UC

**Virtual Local Area Network** См. VLAN

**Virtual Private Network** См. VPN

**Virtual Router Redundancy Protocol** См. VRRP

**Virtual Storage Area Network** См. VSAN

**Virtualization** (виртуализация) Абстрагирование устройств хранения (дисков и лент) от конкретных физических устройств. Обычно выполняет функции, которые не могут реализовать традиционные RAID

системы, например, LUN, который охватывает несколько систем, LUN over-provisioning (размер LUN больше, чем его физическая емкость) и прозрачная для приложений репликация и миграция данных .

**VLAN** Virtual Local Area Networks – Виртуальные LAN позволяют разбить физические сети на небольшие сегменты. Благодаря этому механизму в сетях IP/Ethernet не возникает нестабильность при широковещательных штормах (broadcast storms). Аналогичная функциональность для Fibre Channel давно реализована с помощью зонирования.

**VPN** Virtual Private Network (виртуальные частные сети) используют шифрование для построения туннелей через общедоступные сети. Устройства в VPN работали так, как если бы они были подключены к физически изолированной отдельной сети.

**VRP** Virtual Router Redundancy Protocol - протокол, с помощью которого при выходе из строя одного маршрутизатора его функции начнет выполнять другой маршрутизатор. Этот протокол можно считать кластеризацией маршрутизаторов. Благодаря ему узлам IP-сети (например, порты многопротокольного маршрутизатора Multiprotocol Router в режиме iSCSI и FCIP) необязательно поддерживать такие протоколы, как OSPF или RIP.

**VSAN** (Virtual SAN) фирменный механизм, похожий на зонирование, но с ограниченными возможностями. Не следует путать с LSAN. VS AN разбивает сеть на сегменты, которые до этого были связаны между собой, а LSAN выборочно обеспечивает ранее отсутствовавшие соединения.

**WAN** ( Wide Area Network) – сети, охватывающие города, области, штаты и даже континенты. Из-за больших расстояний в них больше

запаздывание. В AN часто используются в сетях хранения для обеспечения катастрофоустойчивости.

**WAFS** Tapestry Wide Area File Services – решение для филиалов по удаленному доступу к файлам. Не доступно от Brocade с 2008 г.

**Wavelength Division Multiplexer** См. WDM

**WDM** Wavelength Division Multiplexers ; Мультиплексирование по длине волны - позволяет по одному оптическому кабелю передавать сигналы с разной длиной волны

**Wide Area Network** См. WAN

**World-Wide Name** См. WWN

**WWN** World-Wide Name – зарегистрированный 64-битный идентификатор узлов и портов в фабрике. Пример адреса WWN: 10:00:00:60:69:51:0e:8b.

**XPath** Похожая на Fabric OS операционная система для платформ Brocade. На момент написания этой книги использовалась только в AP7420 Multiprotocol Router.

**xWDM** См. DWDM и CWDM

**Zoning** Стандартный механизм контроля доступа для фабрики. Использует для проверки прав доступа PID или WWN.