



Отказоустойчивый ЦОД за 5 дней

День 2: Вычислительные мощности и сеть хранения и сеть хранения

Обсуждение – в Telegram канале:





Отказоустойчивый ЦОД за 5 дней

День 2: Вычислительные мощности и сеть хранения (Часть 1)

Евгений Лагунцов
Технический консультант

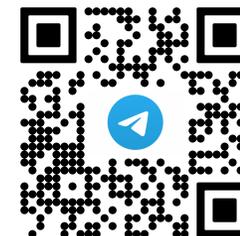
Обсуждение – в Telegram канале:



Программа спринта по отказоустойчивым ЦОД

День	Тема
1 февраля понедельник	Отказоустойчивый ЦОД: основные задачи и проблемы.
2 февраля вторник	Вычислительные мощности и сеть хранения
3 февраля среда	Гиперконвергентные системы и резервное копирование
4 февраля четверг	Сеть и сервисные устройства
5 февраля пятница	Комплексные сценарии

Обсуждение – в Telegram канале:



Защита информационных систем при аварии ЦОД, основные подходы

Остановимся на двух важных моментах

Что требуется для восстановления сервиса при потере ЦОД?

1. Репликация данных в другой ЦОД
2. Переключение сервиса на другой ЦОД
3. а также много всего другого, о чем мы поговорим на других сессиях (физическая связность, связь LAN, связь SAN, сетевые сервисы, балансировка или перенаправление нагрузки и т.д.)

Варианты переключения

Средствами приложения:

- например, Oracle RAC

Кластерные системы, интегрированные в ОС или как дополнительное ПО

- например, Windows Failover Cluster, SLES HAE, Veritas

Платформы виртуализации

- например, растянутый кластер и авто-перезапуск VM

Вручную или дополнительными средствами автоматизации

- например, скрипты

Варианты репликации

Средствами приложения:

- например, SAP HANA System Replication

Средствами дисковых массивов

- Синхронная или асинхронная, или «растянутый» том

Гиперконвергентными платформами

- Синхронная или асинхронная, на уровне VM

Дополнительным ПО

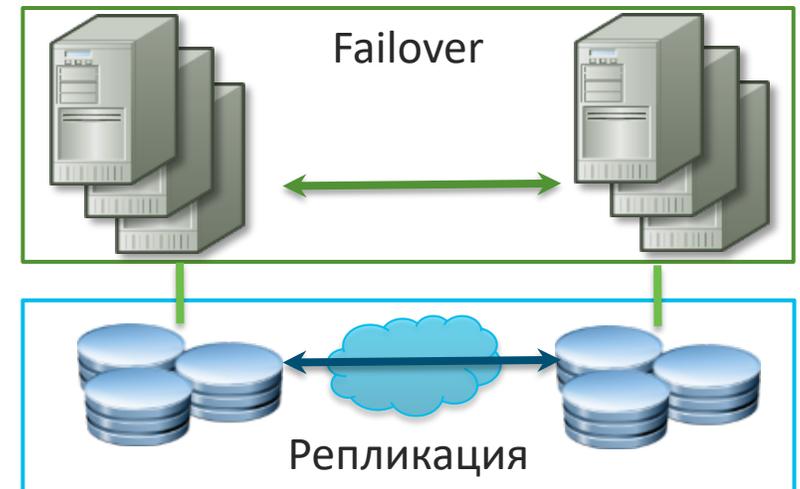
- На уровне VM, томов, файловых систем и т.д.

«Типовые» уровни защиты

Комбинаций вариантов репликации и переключения может быть много.

Подходы, наиболее часто встречающиеся в проектах и на которых хотелось бы остановиться:

- Уровень приложений + кластерное ПО
- Растянутый кластер VM
- Асинхронная репликация VM
- bare-metal



Защита на уровне приложений

- Репликация встроенными средствами бизнес-критичных приложений и СУБД к
 - Oracle, MS SQL, SAP HANA и др.
- Возможна репликация (как правило синхронная) средствами СХД
- Автоматизация восстановления обеспечивается самим ПО или средствами кластеризации, с учетом специфики приложения
 - Windows Failover Cluster, SLES HAE, и др., например, Veritas
- Минимальное время простоя, минимальная потеря данных, максимальная стоимость
- Отдельное внимание требуется к интеграции средств репликации и средств автоматизации восстановления



Растянутый кластер виртуализации

- Используется синхронная репликация, например средствами гиперконвергентной платформы HyperFlex
- Восстановление работоспособности – автоматически средствами HA гипервизора
- Потеря данных близка к нулю, время простоя – на перезапуск VM и инициализацию приложения
- Стоимость – средняя (как правило), диски и лицензии
- Применимо только для VM
- Хотя и растянутый, но ОДИН кластер со всеми вытекающими

DR решения на уровне VM

- Асинхронная репликация, например средствами гиперконвергентной платформы HyperFlex
- Восстановление вручную или с применением дополнительного ПО, например SRM
- Потеря данных может быть существенной (в зависимости от конфигурации), восстановление – перезапуск VM, как правило неавтоматизируемый
- Минимальная стоимость (если не считать SRM, только дисковое пространство), минимальная сложность
- Применимо только для VM

DR решения на уровне физического оборудования

- Репликация как правило средствами дисковых массивов (синхронная или асинхронная)
- Восстановление за счет запуска на резервной площадке идентичного «логического сервера» вручную или скриптами
- Время на переключение – несколько минут на переконфигурацию сервера, плюс старт ОС и приложений
- Минимальная стоимость (но требуется функционал репликации и диски), небольшая сложность
- Применима для любых ОС и гипервизоров
- Используется специфика Cisco UCS – сервисные профили

Демо (начало)

Конфигурация демо

- Одна UCS система – «продуктивная»:
 - блейд-сервер
 - boot-from-SAN
 - 2 vNIC, 2 vHBA
 - Выключен HyperThreading
- Вторая UCS система – «непродуктивная»:
 - Стоечный
 - Загрузка с HDD
 - 10 vNIC, нет vHBA
 - Включен HyperThreading
- Общая сеть SAN

Сценарий демо

- «Потеря» продуктивной системы
- Деассоциация «непродуктивного» сервера с физического сервера
- Ассоциация «продуктивного» сервисного профиля с освобожденным сервером
- Включение системы

Построение DR решений на базе Cisco UCS и сервисных профилей

Архитектура Cisco UCS

Универсальная фабрика

- Высокая производительность 25G/40G/100G Ethernet
- Подключение к фабрике серверов, VM, контейнеров
- Любые протоколы LAN/SAN/HPC/Управление

Простота масштабирования

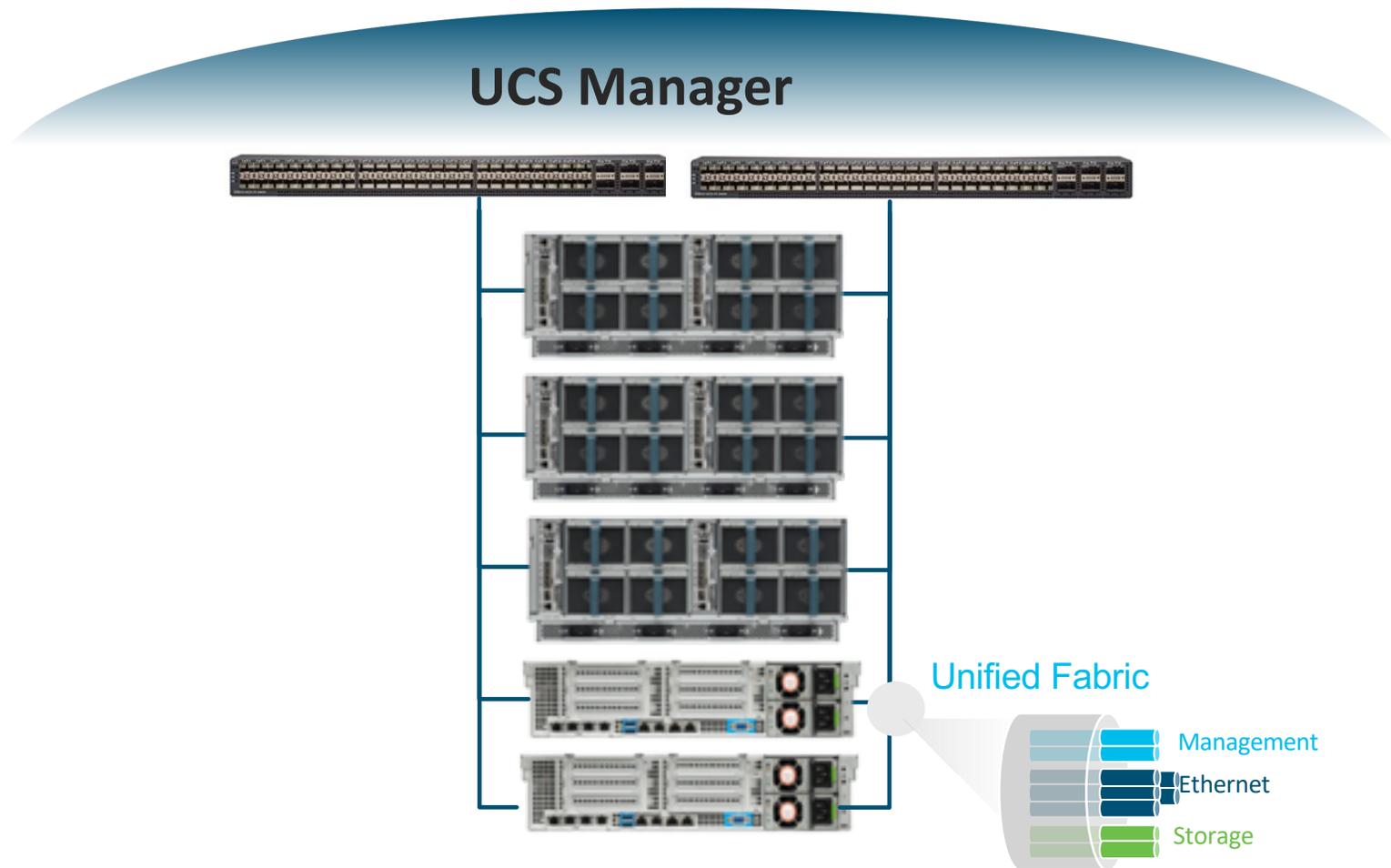
- Применение сервисных шаблонов на сотни и тысячи серверов
- Форм-фактор не имеет значения (Blades, Rack или гиперконвергентный)
- Настройки сервера по требованию

Единое управление

- Единая точка управления жизненным циклом
- Управление «один ко многим» (шаблоны, политики, профили, пулы)
- Управление из облака через Cisco Intersight

100% программируемость

- Абстрагирование аппаратной части от операционной плоскости
- Управление на основе политик
- API



Интегрированная фабрика

Высокая производительность

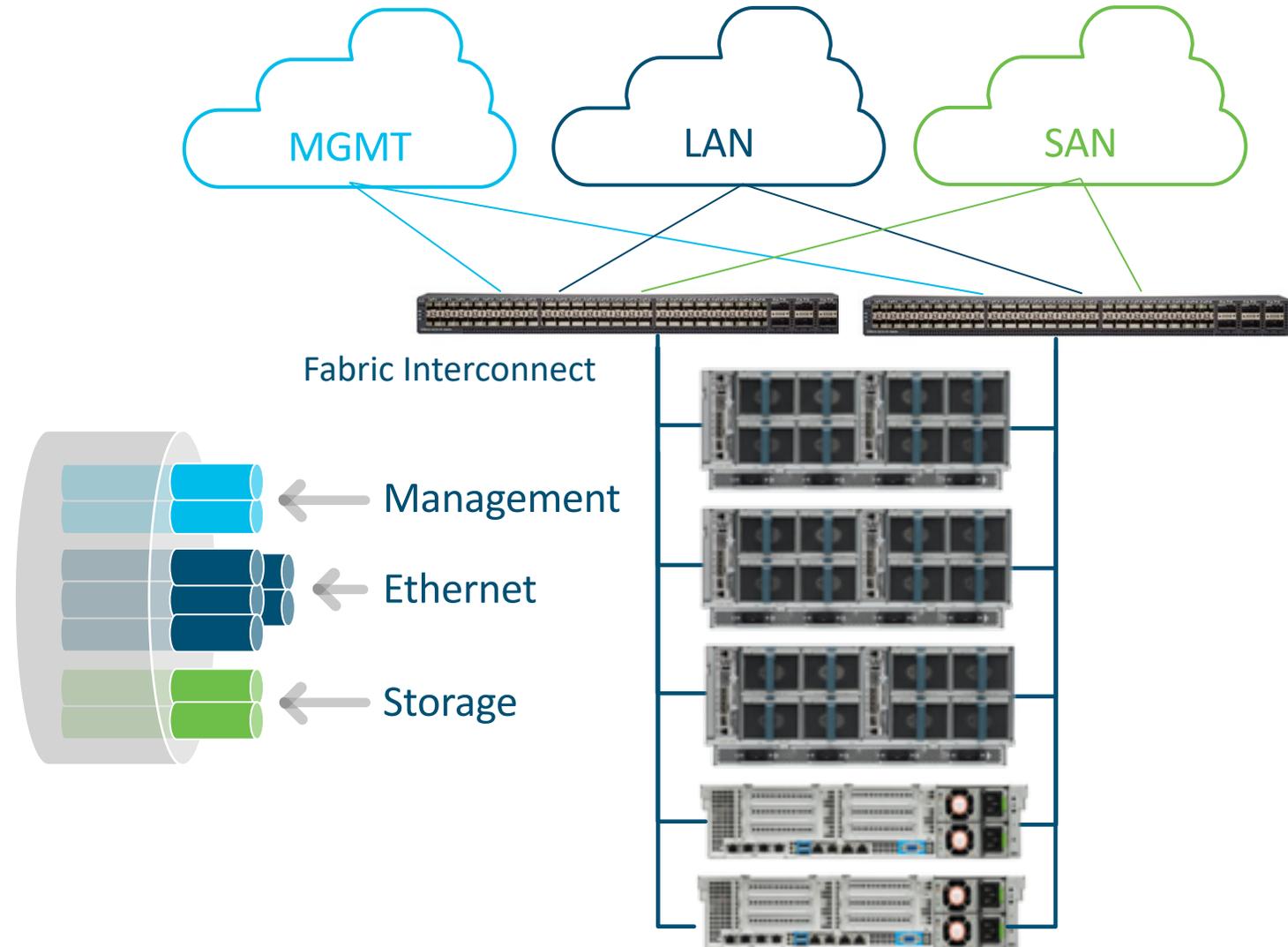
- 10G/25G/40G/100G Ethernet/FCoE
- 400 Гбит/с на шасси
- 16G/32G FibreChannel

Единая фабрика

- Универсальное соединение для LAN/SAN/HPC/MGMT
- Нужные тип и количество сетевых адаптеров для серверов

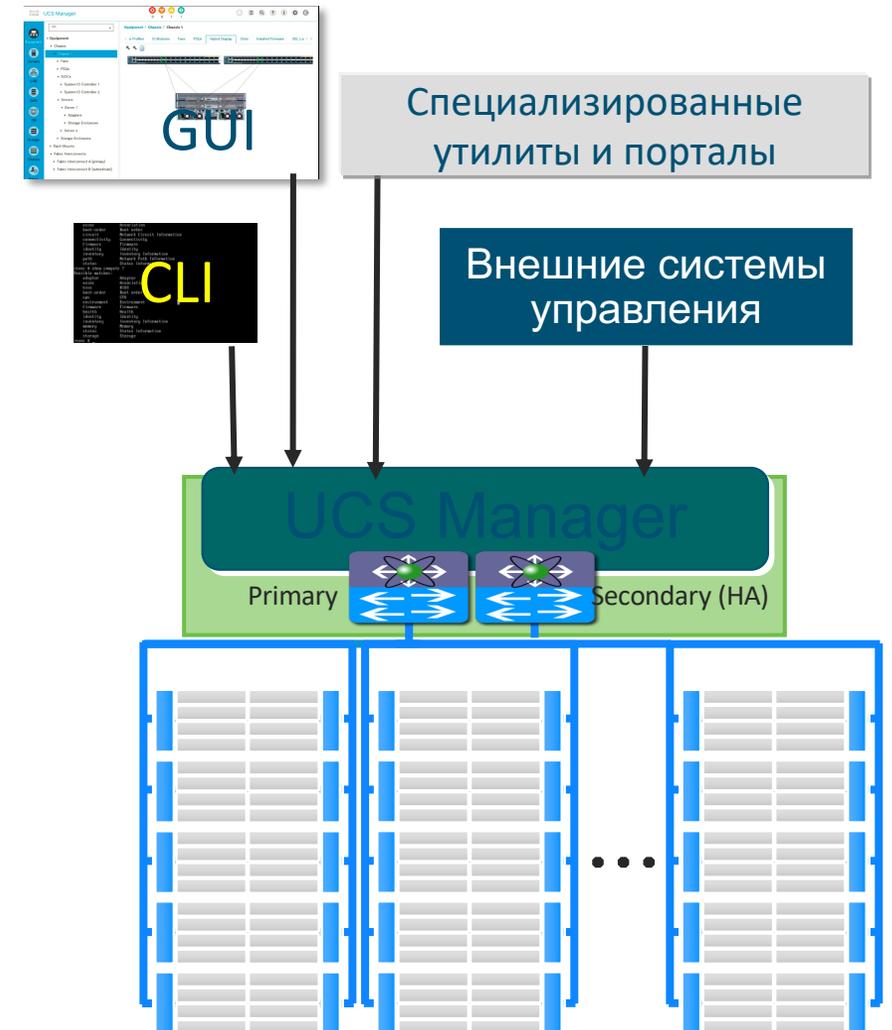
Простота масштабирования

- Добавление соединений вместо внедрения новой фабрики
- Внешние подключения и внутренние ресурсы масштабируются независимо



Интегрированное управление

- Единая точка управления
- Единое управление всей системой (вычислительная и LAN/SAN инфраструктура)
- неотъемлемая и интегрированная часть
- отказоустойчивость
- До 160 серверов в одном домене (на одной паре FI)
- Конфигурируются общие политики, шаблоны, профили – многократно используемые, тиражируемые элементы



100% программируемость

Открытый и задокументированный API

- Быстрая адаптация под различные сценарии эксплуатации и типы приложений
- Автоматизация
- Интеграция с современным инструментарием разработчиков

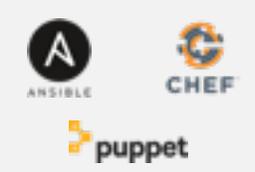
Гетерогенные типы приложений



Интеграция



Кастомизация



Единая плоскость управления

Конфигурация на основе политик

Открытый API



UCS B-Series

Серверы



UCS C-Series



Конвергенция



Hyperflex Systems

Гиперконвергенция



UCS S-Series

Scale Out

UCS: подход к управлению

- Система Cisco UCS построена на «абстракции» оборудования
- Сервисный профиль полностью описывает характеристики сервера, включая:
 - Количество адаптеров, тип (NIC и HBA), идентификаторы (MAC и WWN)
 - Настройки сетевых подключений (VLAN, VSAN, QoS);
 - Порядок загрузки ОС и используемые загрузочные устройства;
 - Версии прошивок, настройки BIOS, и т.д.
- Сервисный профиль может быть растиражирован
- Через один сервисный шаблон можно управлять десятками серверов одновременно
- Сервисный профиль может быть «перемещен» на другой сервер, в другую систему или в другой ЦОД



DR с применением UCS

Возможности

- Восстановление физического сервера на той же площадке
- Восстановление физического сервера на другой площадке (с изменением WWN и загрузочного LUN и прочих параметров или без изменения)
- Восстановление всей UCS системы целиком из полной резервной копии конфигурации



DR с применением UCS

Сценарии

- Как правило, до аварии «резервная» UCS система используется для задач тестирования и разработки
- Использование в качестве дополнительного уровня защиты: для доступности приложения используется кластеризация, сервисные профили обеспечивают восстановление защищенности



Как именно «переносить» профиль

- Создавать вручную "с нуля"
- С помощью простого скрипта через API интерфейс
- Экспорт логической конфигурации продуктивной UCS системы, импорт ее в непродуктивную систему
 - Это обычный XML файл, при необходимости могут быть сделаны корректировки (WWN, MAC, VLAN и т.д.)
- Полный бэкап конфигурации продуктивной системы, восстановление всей непродуктивной системы из этого бэкапа
 - Восстановление с полного бэкапа приведет к остановке всех непродуктивных сервисов

На что необходимо обратить внимание

- Как устроена репликация между ЦОД – возможно необходимо вручную делать LUN доступным на запись на резервной площадке
- Возможно необходимо изменить конфигурации (WWN, MAC, VLAN, загрузочный том и т.д.)
- Если конфигурации не менялись – нужно гарантировать что «умерший» сервер не «восстанет» и не будет обращаться в ту же сеть с теми же идентификаторами

Выводы по демо

- Мы восстановили работоспособность bare-metal сервера на другой UCS-системе, условно на другой площадке
- В процессе восстановления непродуктивный сервер полностью «превратился» в продуктивный:
 - Изменилось число и настройки NIC и сетевые подключения
 - Добавились адаптеры FC
 - Изменилось загрузочное устройство, с HDD на SAN
 - Настройки BIOS
 - (могло поменяться все что угодно, включая версии прошивок, тип ОС и т.д.)
- При том что продуктивный сервер был блейдом, а непродуктивный - стоечный



DR с помощью сервисных профилей

- Простое и эффективное средство для восстановления продуктивных или резервных систем в случае аварии
- Восстановление может быть ручным, может быть автоматизированным скриптом
- При восстановлении могут быть изменены идентификаторы и конфигурации для корректной работы в другом ЦОД
- Восстановление может быть проведено как для bare-metal ОС, так и для гипервизоров
- Может комбинироваться с более высокоуровневыми средствами защиты (кластеризация или HA) для восстановления уровня защищенности



Отказоустойчивый ЦОД за 5 дней

День 2: Отказоустойчивые ЦОД: сеть хранения (Часть 2)

Александр Скороходов
Технический консультант

Связь сетей хранения данных отказоустойчивых ЦОД

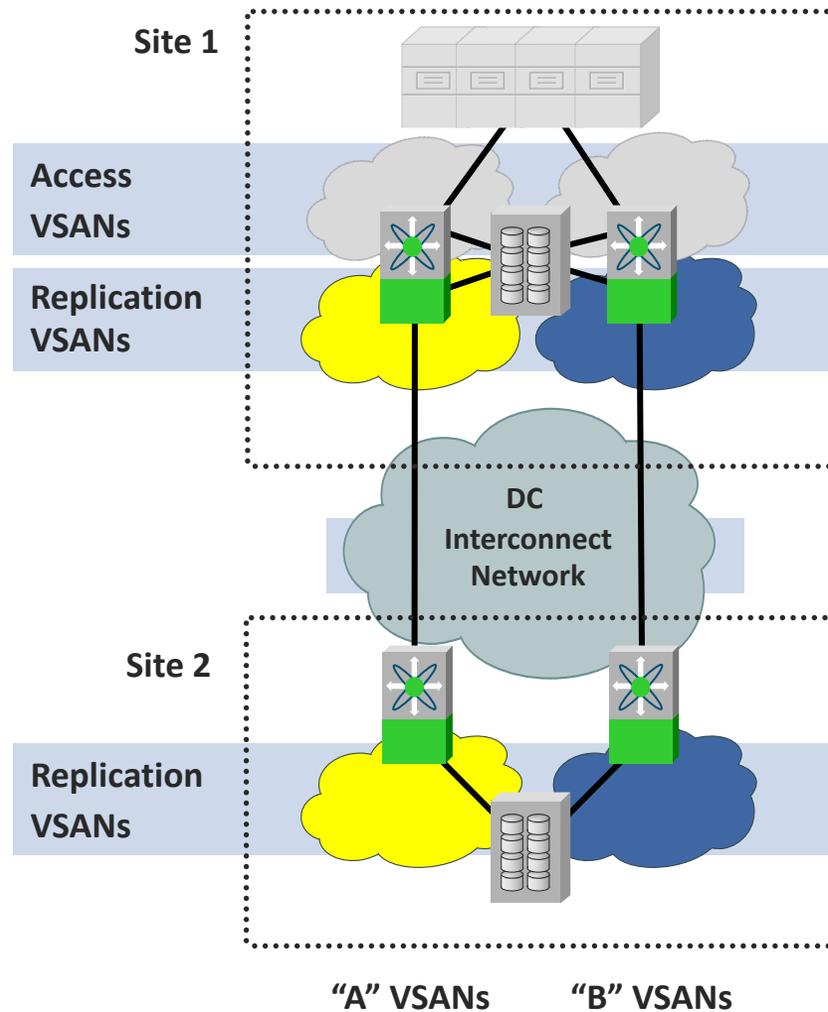
Резервирование СХД

Синхронная и асинхронная репликация

- **Синхронная репликация данных:** Приложение получает подтверждение I/O после его выполнения на обеих сторонах (zero RPO)
 - «Метро» расстояния
- **Асинхронная репликация данных:** Приложение получает подтверждение I/O после его выполнения на основном (локальном) диске, в то время как его копирование на удалённый массив продолжается
 - «Неограниченные» расстояния



Типичный отказоустойчивый дизайн для связи SAN



Две локальные SAN фабрики (A/B) в каждом ЦОД для отказоустойчивого подключения серверов и систем хранения данных

- Использование multipathing функций на хостах и СХД

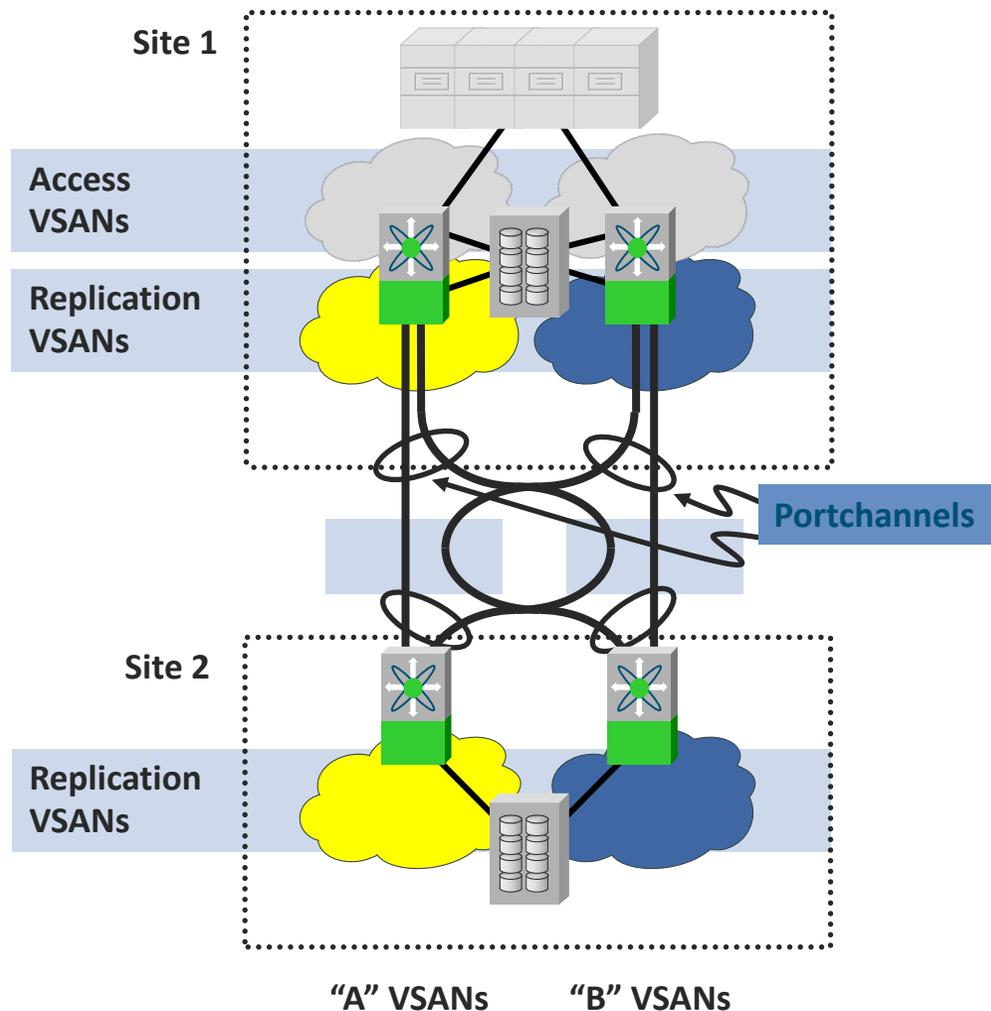
Фабрики для связи SAN, типично, изолированы от фабрик для подключения серверов

- Требования могут зависеть от технологии репликации
- Физическая или логическая (VSAN) изоляция

Соображения резервирования для репликации:

- Стандартный подход с дублированием фабрик (A/B) на участке между ЦОД
- «Клиентская защита» — массивы обеспечивают защиту от отказа в любой из фабрик
- Может дополняться «сетевой защитой» с помощью агрегированных каналов и/или защиты в оптическом транспорте

Связь SAN: резервирование соединений



Агрегированные каналы (portchannel) повышают доступность при связи SAN фабрик

- Выглядят как единый логический линк
- Включает до 16 физических интерфейсов
- Нет привязки к конкретным группам портов или модулям коммутатора

Использование интерфейсов, идущих по разным трассам, WDM транкам и т.д.

- Максимальная независимость

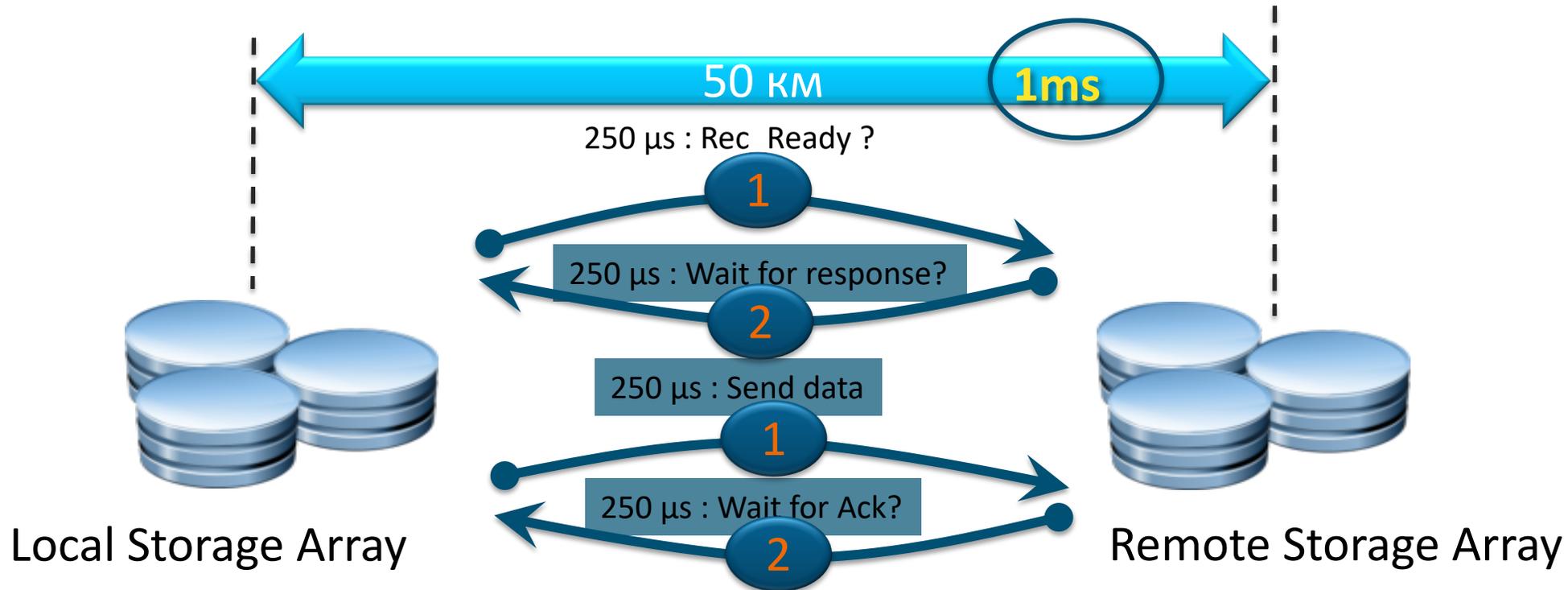
Балансировка трафика между интерфейсами в агрегированном канале:

- Балансировка SRCID/DESTID или SRCID/DESTID/OXID
- Трафик одной операции ввода-вывода всегда идёт по одному пути – нет риска искажения порядка фреймов
- Некоторые протоколы репликации требуют балансировки SRCID/DESTID

Что ограничивает расстояние?

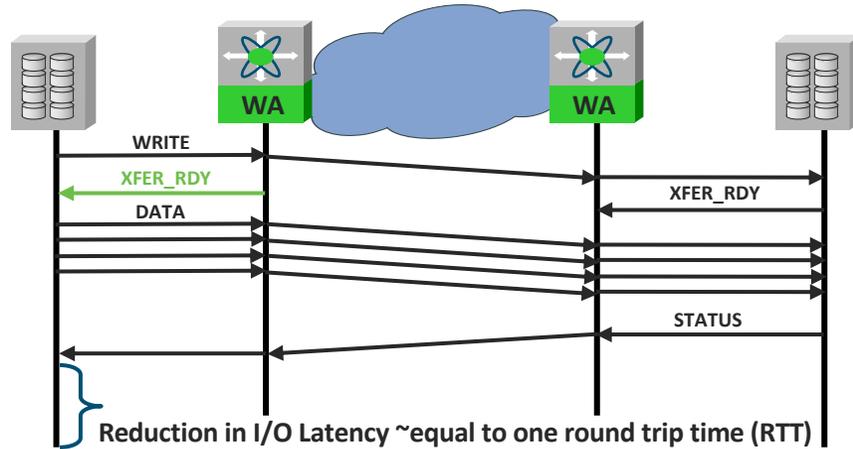
Требования синхронной репликации

- SCSI протокол (FC) требует два round trip на операцию
- Вносимая задержка операции $20\mu\text{s}/\text{км}$, $100\text{ км} = 2\text{ мс}$
- В зависимости от приложения синхронную репликацию, как правило ограничивают 50-100 км
- I/O Acceleration «убирает» один round-trip

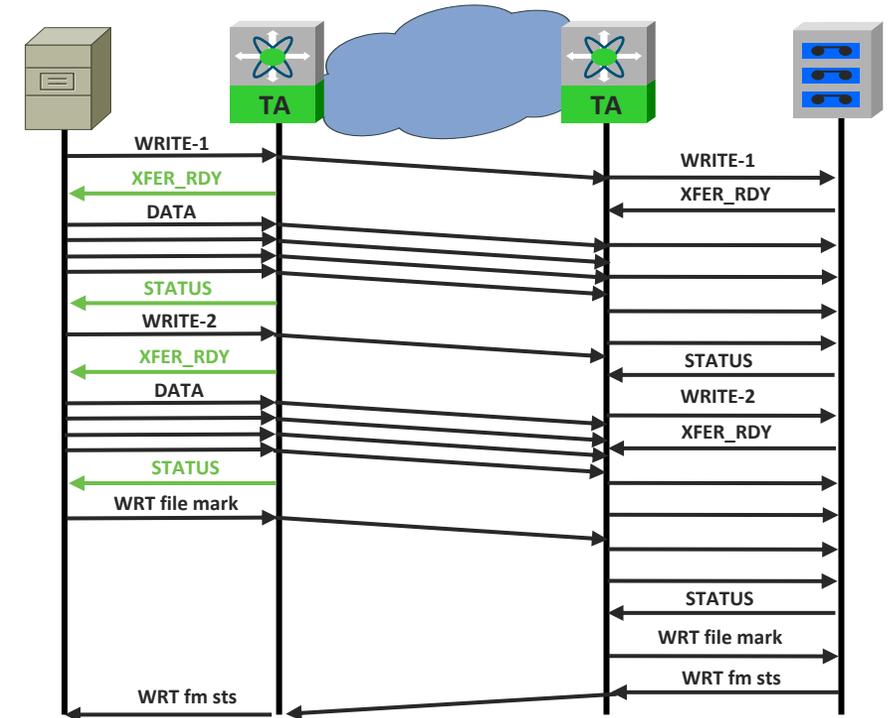


Работа ускорения ввода/вывода

Write Acceleration (WA)



Tape Acceleration (TA)



- Ускорение синхронной репликации и резервирования на ленту: аналогичные подходы
- На работу с лентой дополнительно влияют особенности физического носителя и ограничения буферизации
- Write Acceleration имитирует только Transfer Ready, Tape Acceleration имитирует Command Status

Поддерживается на:

- Cisco MDS 9700 24/10 SAN Extension Module
- Cisco MDS 9250i Multilayer Fabric Switch

Что ограничивает расстояние?

Оптика: «дальнобойные» трансиверы Fibre Channel для Cisco MDS

Скорость	Трансивер	Длина волны	Расстояние / бюджет	Вендор
8 G	DS-SFP-FC8G-LW	1310	10 км	Cisco
	DS-SFP-FC8G-ER	1550	40 км	Cisco
	DS-CWDM8Gxxxx	CWDM 1470-1610	24 дБ	Cisco
	DS-8G-ZR	1550	70 км	Smartoptics
	DS-8G-ZR-CXX	CWDM 1470-1610	23 дБ	Smartoptics
	DS-8G-ZR-Dxxxx	DWDM 1529.55-1561.42	23 дБ	Smartoptics
16 G	DS-SFP-FC16G-LW	1310	10 км	Cisco
	DS-SFP-FC16GELW	1310	25 км	Cisco
	DS-16G-ER	1550	40 км	Smartoptics
	DS-16G-ER-Cxx	CWDM 1470-1550	13 дБ	Smartoptics
	DS-16G-ER-Dxxxx	DWDM 1529.55-1561.42	13 дБ	Smartoptics
32G	DS-SFP-FC32G-LW	1310	10 км	Cisco
	DS-32G-IR-Dxxxx	DWDM 1529.55-1561.42	7 дБ	Smartoptics

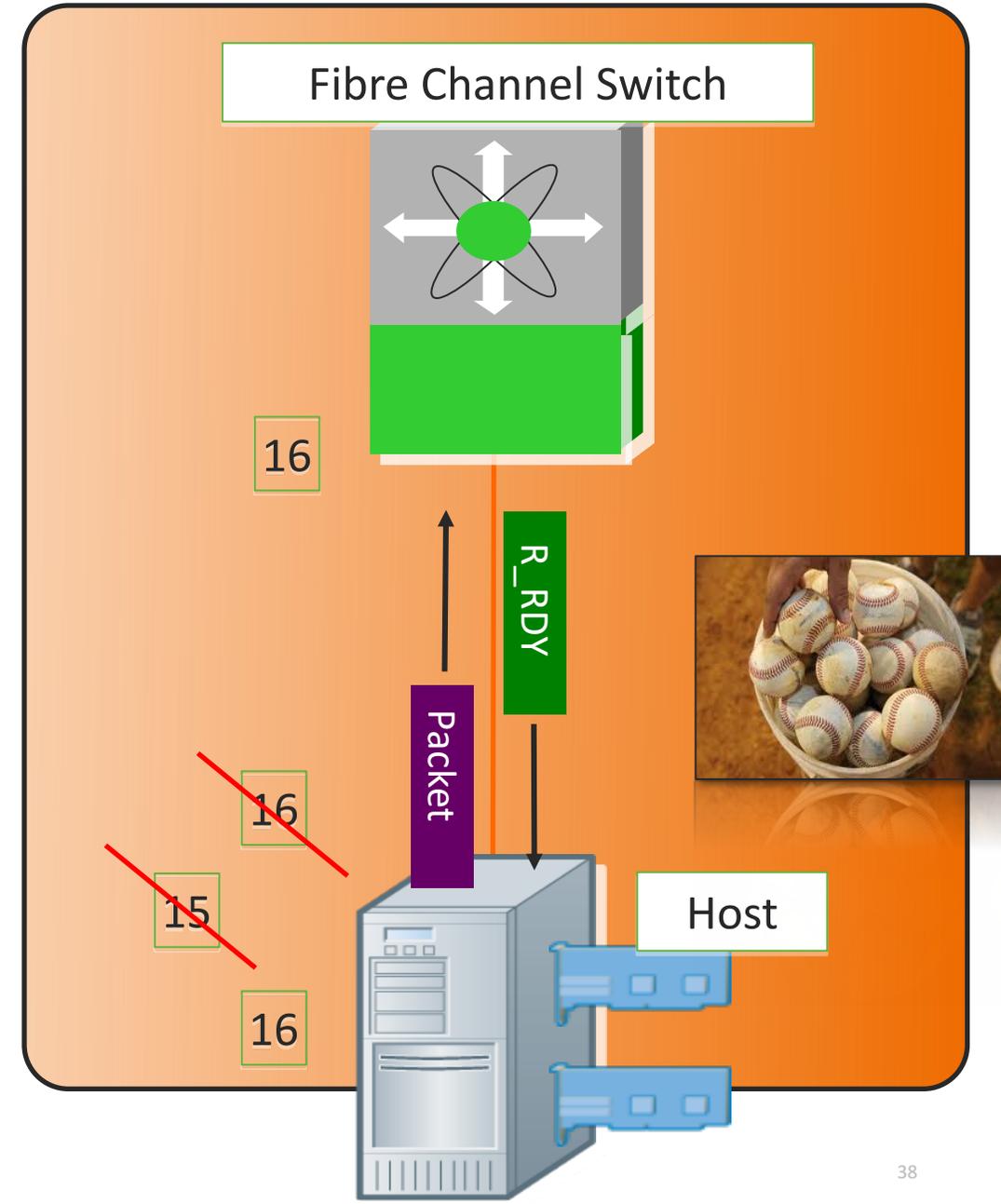
https://www.cisco.com/c/en/us/products/collateral/storage-networking/mds-9000-series-multilayer-switches/product_data_sheet09186a00801bc698.html

<https://www.smartoptics.com/products/cisco-collection/>

Управление потоком в Fibre Channel

Buffer to Buffer Credits

- Buffer to Buffer credits (B2B Credits) используются, чтобы обеспечить отсутствие потерь FC фреймов из-за отсутствия буферной ёмкости на приёмной стороне
- Управление потоком на каждом участке (хост-коммутатор или коммутатор-коммутатор)
- Число B2B Credits согласовывается при поднятии линка
- Оставшееся число B2B Credits – число FC фреймов, которые разрешается передать в данном направлении
 - Не зависит от размера фрейма
 - Уменьшается с передачей каждого фрейма
 - Если кредитов не осталось, передача останавливается
- Кредиты добавляются приходом фреймов Receiver Ready (R_RDY) от получателя



Что ограничивает расстояние?

Буферные кредиты

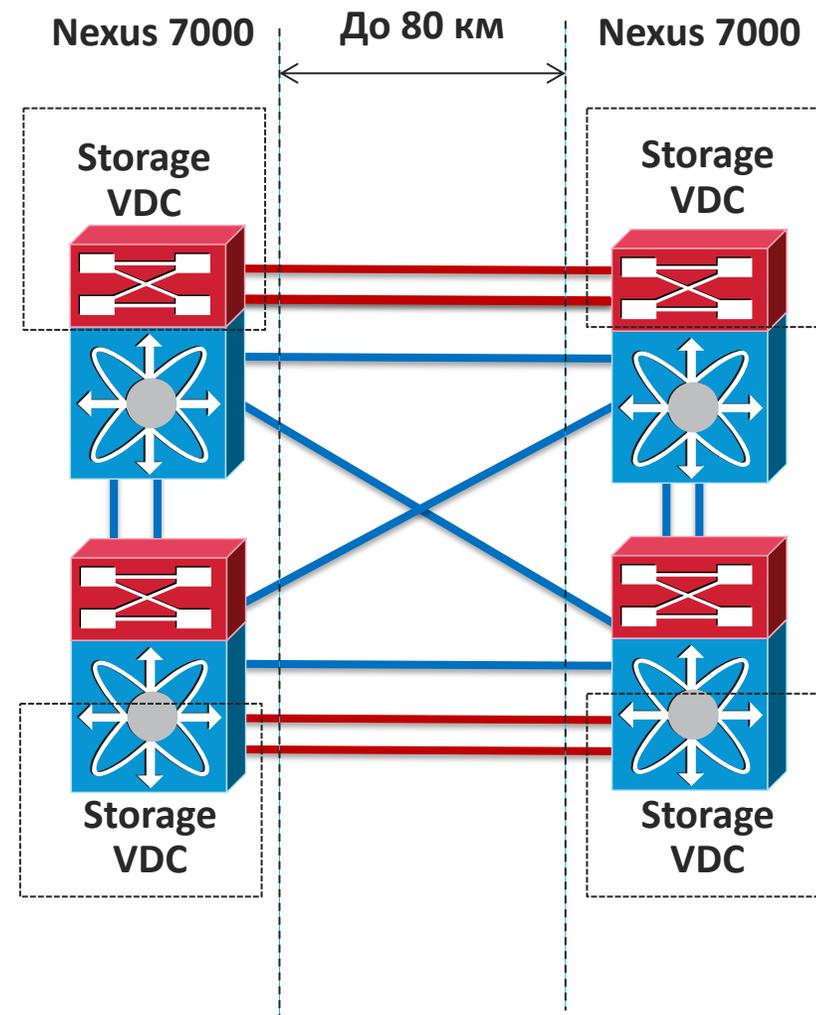
- BB_Credits (буферные кредиты) нужны, чтобы «заполнить» соединение фреймами FC
- Если BB_Credits не хватает для данного расстояния – снижается производительность, соединение простаивает: данные не будут передаваться, пока не вернётся R_RDY
- Правило для запоминания: на 2G на 1 км необходим (примерно) один BB_Credit (для фреймов максимального размера 2112 байт), на больших скоростях – пропорционально больше
- Число необходимых BB_Credits растёт при росте расстояния и скорости, и снижении размера фрейма
- Число доступных BB_Credits определяется оборудованием и его настройками
- Cisco MDS 9700/9396T обеспечивают до ~8200 кредитов на порт - более 500 км на 32G!

Frame Size	2 Gbps	4 Gbps	8 Gbps	10 Gbps	16 Gbps	32 Gbps
512 Bytes	4 BB/km	8 BB/km	16 BB/km	24 BB/km	32 BB/km	64 BB/km
1024 Bytes	2 BB/km	4 BB/km	8 BB/km	12 BB/km	16 BB/km	32 BB/km
2112 Bytes	1 BB/km	2 BB/km	4 BB/km	6 BB/km	8 BB/km	16 BB/km

FCoE для связи SAN между ЦОД?

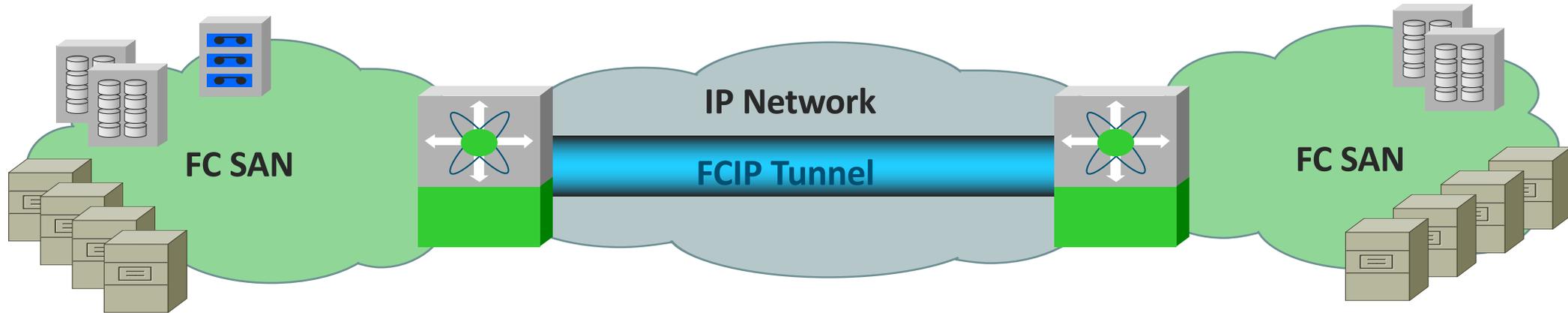
Да, если это поддерживает оборудование

- Управление потоком в FCoE опирается на посылку «пауз» (per-priority pause)
- Размер оставшихся буферов должен быть достаточен, чтобы поместить фреймы, передаваемые за время RTT данного линка
- Поддерживаемые расстояния для «дальнобойного» FCoE транспорта (10G):
 - Nexus 7000/7700 с F3 картами: до 80 км
 - MDS 9700: до 80 км
 - MDS 9250i: до 80 км
- **Использование отдельных соединений для LAN и SAN трафика**
 - Не пытайтесь «смешивать» LAN+SAN между ЦОД



FCIP: Fibre Channel over IP

No lossless? No problem. That's what this is for!

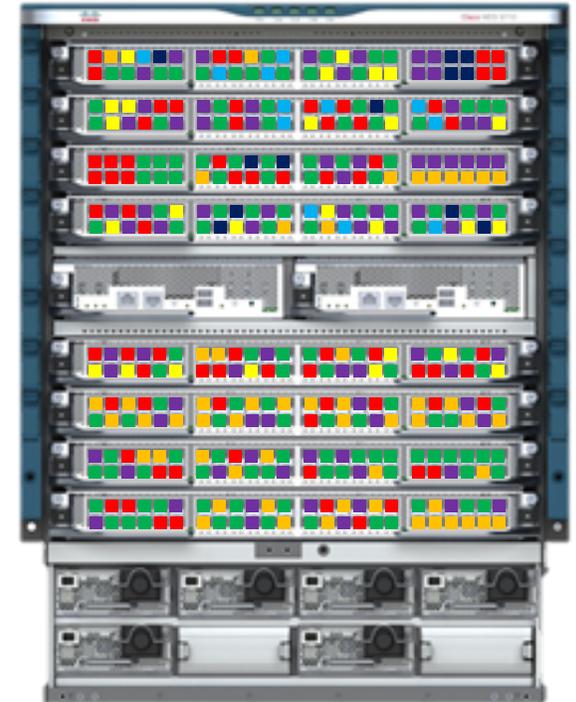


FCIP: IETF стандарт для связи Fibre Channel SAN через IP (RFCs 3821 и 3643)

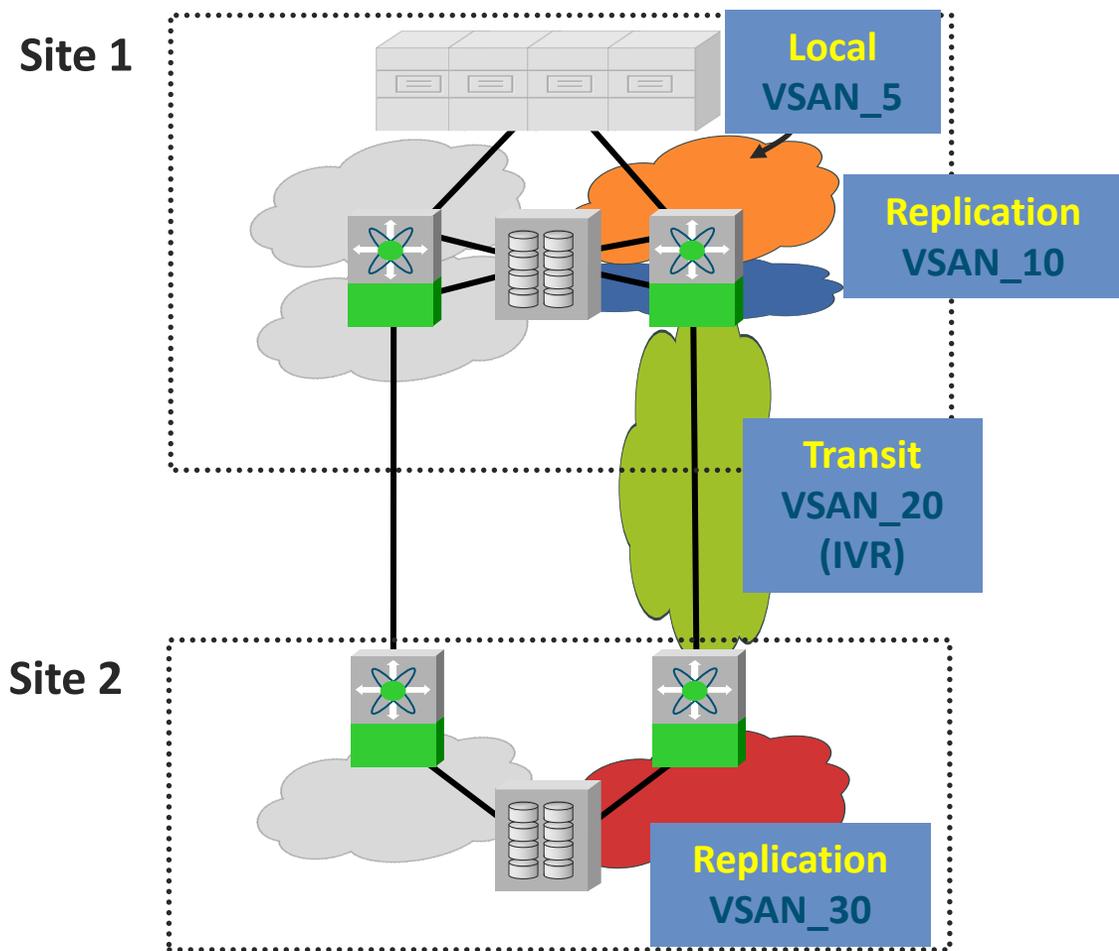
- Соединение «точка-точка» (туннель) между двумя FCIP устройствами
- Используется TCP – могут использоваться механизмы оптимизации и компрессии
- Создаётся единая FC фабрика (общий FSPF домен)
- Транспорт – IP сеть, в том числе и на большие расстояния (тысячи км)
- Рекомендуется (но не обязателен) MTU \geq 2300 байт

VSAN-ы

- Virtual SAN (**VSAN**) – способ разделения портов физической фабрики на логические (виртуальные) фабрики (аналог VLAN в Ethernet)
- Логическое разделение с аппаратной изоляцией
- Стандартизованы ANSI T.11 как Virtual Fabrics
- Повышение эффективности использования оборудования, сокращение числа неиспользуемых портов и отдельных SAN коммутаторов
- События (например, RSCN) изолированы в пределах VSAN, что повышает доступность в работе других VSAN
- Сервисы FC (zone server, name server, login server...) и настройки фабрики функционируют на уровне VSAN
- Поддерживаются на всех моделях Cisco MDS 9000 без дополнительного лицензирования
- Поддерживается совместная передача многих VSAN (VSAN trunking, EISL): интерфейс типа TE
- Поддерживается (контролируемая) маршрутизация между VSAN: Inter-VSAN Routing (IVR)



Расширение SAN и Inter-VSAN Routing (IVR)



- Сбой на «транзитной» VSAN_20 (оборудование или кабель) не нарушит трафик в VSAN_10 или VSAN_30
- Нужно, если эти VSAN используются для локальной обработки
- Работает с любым транспортом («тёмная оптика», DWDM/CWDM, FCIP)

VSAN_5 - Site 1 Host Fabric

VSAN_10 - Site 1 Replication Fabric

VSAN_20 - Inter-site SAN Extension Fabric

VSAN_30 - Site 2 Replication Fabric

Семейство SAN коммутаторов Cisco MDS и его применение для связи SAN между ЦОД

Семейство Cisco MDS 9000

Cisco 16/32G Multilayer
Fabric Switch Series

Cisco Multiservice Fabric
Switch

Cisco 16/32G+ Multi-layer
Director Series

Порты Fibre Channel



MDS 9148S/T



MDS 9396S/T



MDS 9250i



MDS 9706



MDS 9710



MDS 9718

16/32G SAN разного масштаба с семейством Cisco MDS 9000
Готовность к 64G!

Основные принципы развития семейства MDS

- Сохранение инвестиций
 - Совместимость всех поколений оборудования в единой фабрике
 - **Совместимость всех поколений FCIP решений**
 - MDS 9500 – четыре поколения карт без необходимости замены шасси
 - MDS 9700 – переход на 32G без замены коммутационных фабрик, готовность к 64G
 - Совместимость карт 16G FC, 32G FC, 10G FCoE, 40G FCoE и 1/10/40G FCIP в одном директорном шасси
- Богатство функций на всей линейке
 - Поддержка VSAN и Inter-VSAN Routing даже на младших моделях
 - Интегрированная аппаратная FC телеметрия/аналитика на всех 32G моделях

Директоры Cisco MDS 9700



MDS 9710



MDS 9718



MDS 9706



48x 32G Line-rate FC
с поддержкой
аналитики



24/10 SAN Extension



48x 10G Line-rate FCoE



24x 40G Line-rate FCoE

16G/32G FC (64G ready), 10/40G FCoE, FCIP 1/10/40G –
в едином шасси без необходимости замены фабрик !

Линейная карта 48 портов 32G Fibre Channel



- 1.5Tbps полосы передачи
- Для MDS 9718 изменений не нужно
- MDS 9710/9706 нужно укомплектовать 6 фабриками

- 48 неблокируемых портов 4/8/16/32G FC
- Поддержка 32G / 16G трансиверов SFP+
- **До 8191 В2В кредитов на порт**
- Аппаратная FC аналитика
- Идентификация VM
- Расширенная диагностика
- Изоляция/деприоритезация Slow Drain трафика

- Интероперабельность с 16G FC и 10/40G FCoE модулями
- Поддержка FCR для IOA с FCIP модулем
- Работа с существующей СКС

Линейная карта MDS 9700 24/10 SAN Extension

Высокая плотность и производительность:

24 x 2/4/8/10/16G FC

8 x 1/10GE IP

2 x 40GE IP

80 Гбит/с FCIP производительности

FCIP функции:

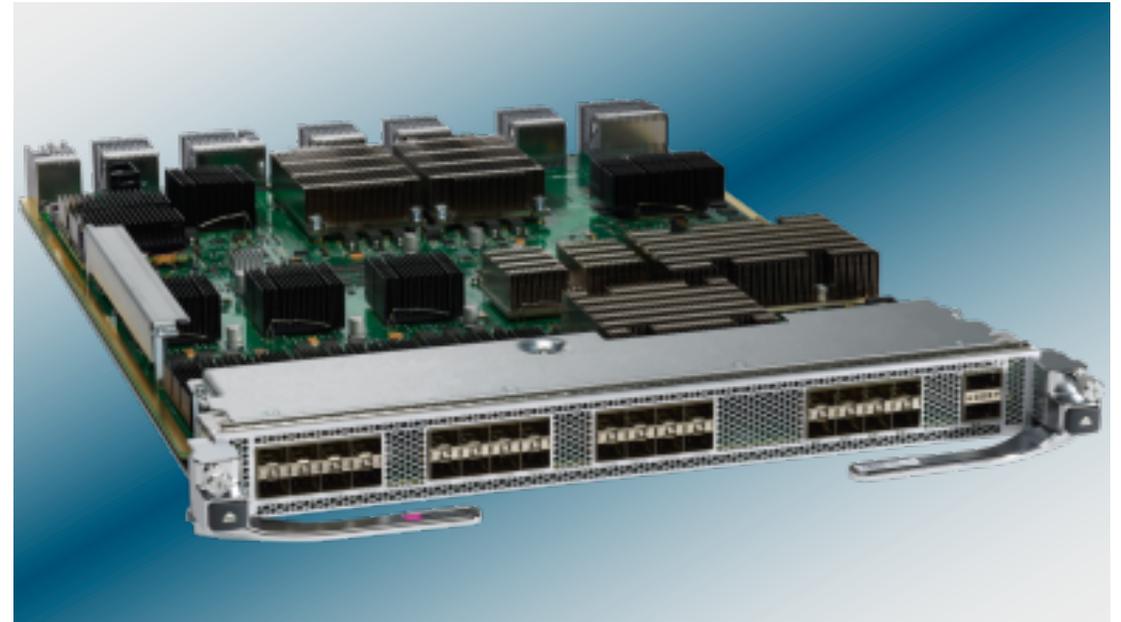
Compression, Write Acceleration,
Tape Acceleration, IPsec Encryption

FC функции :

FC Trustsec Encryption

Совместимость:

Интероперабельность со всем FCIP шлюзами Cisco MDS



Мультисервисный коммутатор Cisco MDS 9250i

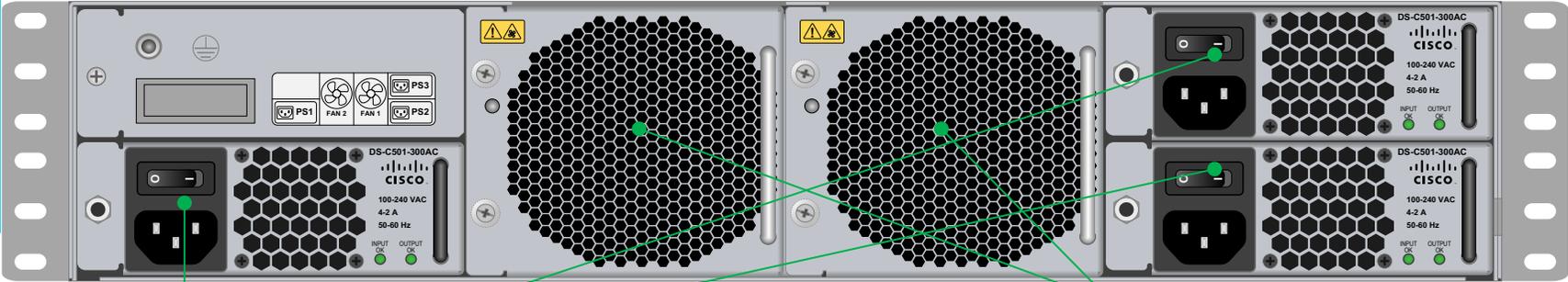
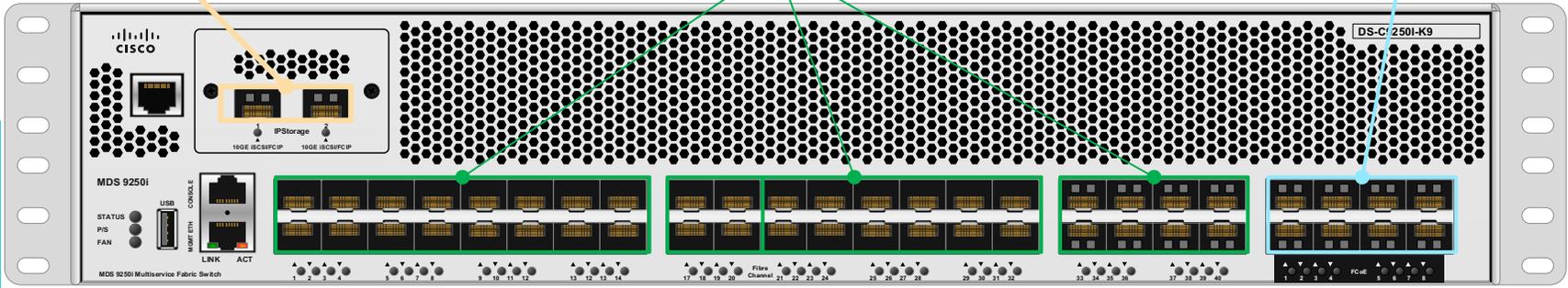
Универсальная сервисная платформа для SAN



2 порта 10G FCIP/iSCSI

40 портов FC 16G:
20 активно, 20 активируются лицензией

8 портов 10G FCOE



Hot-Swappable 2+1
Redundant Power Supplies

Port exhaust Fan
Redundant Hot-swappable fan tray
with integrated temperature &
power management

40 неблокируемых портов FC 16/8/4/2G с поддержкой FICON:
20 активно, 20 активируются лицензией
+
8 портов 10G FCOE
+
2 порта 10G FCIP/iSCSI

Automated provisioning
Quick Configuration Wizard

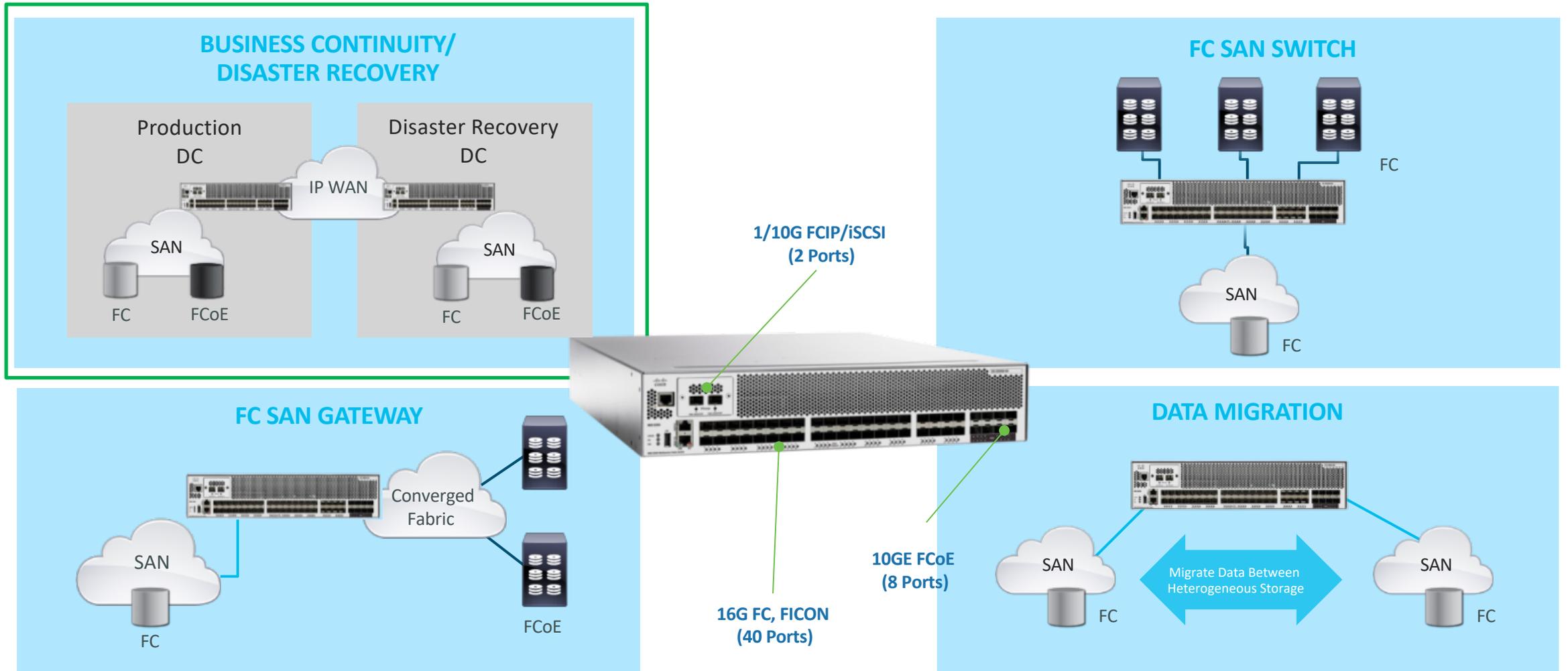
Intelligent Capabilities
VSAN, IVR, FC Redirect

Buffer-to-buffer credits
До 253 на порт

FCIP
Расширение SAN через MAN/WAN с высокой полосой пропускания

I/O ACCELERATOR
Ускорение резервного копирования и репликации

Мультисервисный коммутатор Cisco MDS 9250i



Одно устройство, много сценариев использования

Фиксированные SAN коммутаторы MDS 32G

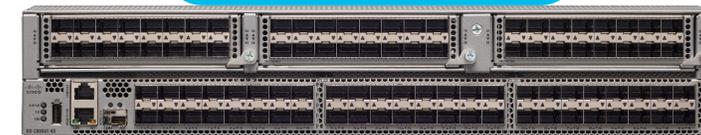
MDS 9132T



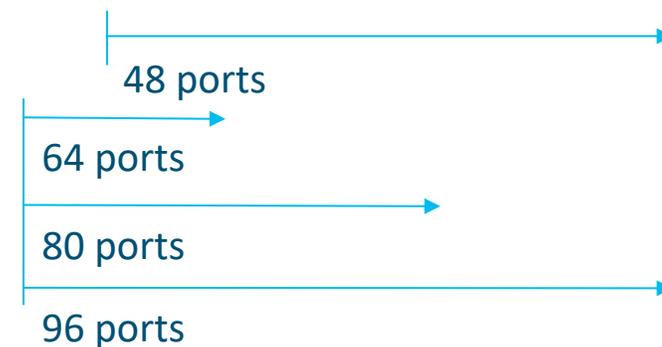
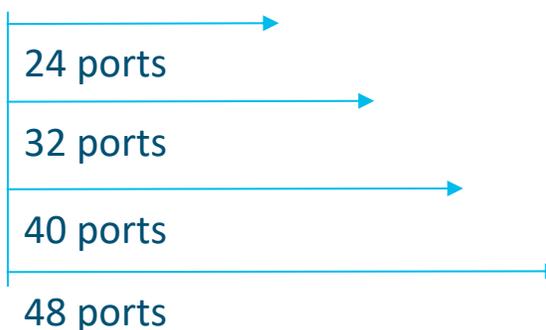
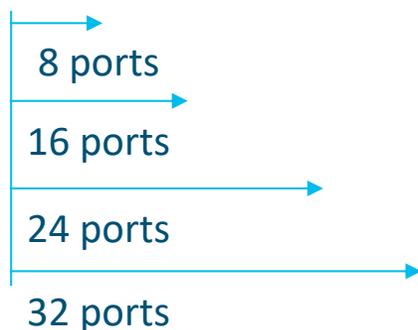
MDS 9148T



MDS 9396T



Варианты
по числу
портов



Неблокируемые
порты 32G

Пригодны для
«All Flash» и NVMe задач

Встроенная телеметрия
и аналитика

Функции
корпоративного класса

Поддержка
FC-NVMe

Защита инвестиций

Простая схема
лицензирования

Высокая надёжность

Число B2B Credits для актуальных моделей MDS

Модель коммутатора/ карты	Максимальное число B2B Credits на порт – без дополнительной лицензии	Максимальное число B2B Credits на порт - с лицензией Enterprise	Максимальное расстояние(км)*, примерно
MDS 9700 / DS-X9648-1536K9 (32G FC)	500	8191	510
MDS 9250i (16G FC)	253	253	31
MDS 9148S (16G FC)	253	253	31
MDS 9396S (16G FC)	500	4095	510
MDS 9132T / 9148T (32G FC)	500	8191	510
MDS 9396T (32G FC)	500	8191	510

* На номинальной скорости порта, для фреймов 2112 байт

