



Отказоустойчивый ЦОД за 5 дней

День 1: Отказоустойчивый ЦОД:
основные задачи и проблемы

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:



Типы распределённых ЦОД

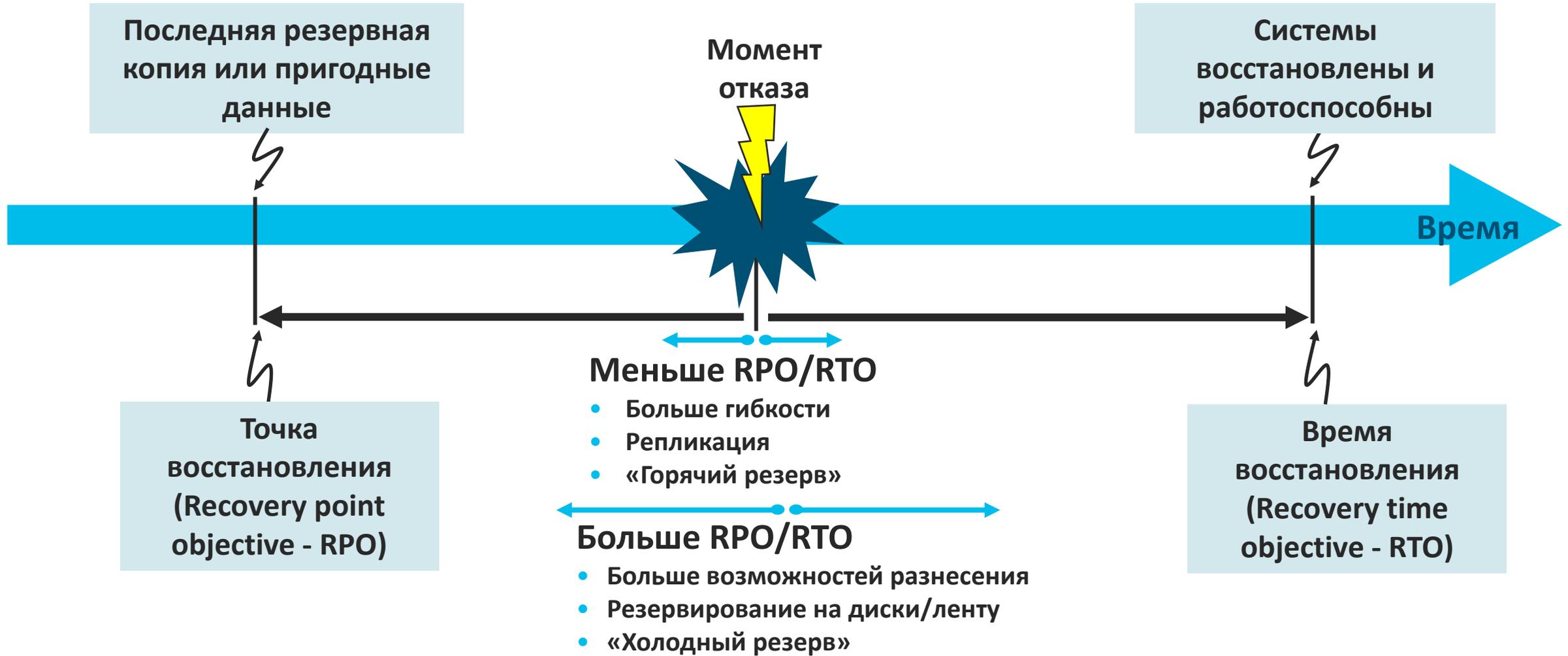
Распределённые ЦОД

Цели создания

- **Катастрофоустойчивость**
- **Непрерывность обработки**
- Мобильность нагрузок
- Миграция систем
- Нарращивание производительности/ёмкости
- Распределённые сервисы
- Географически локализованные сервисы

Отказоустойчивость и катастрофоустойчивость

Точка восстановления и время восстановления



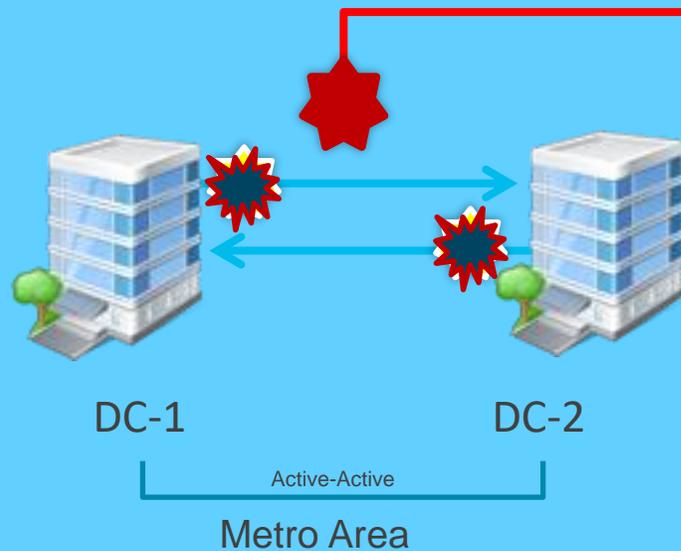
«Нулевые» RPO/RTO – система непрерывной доступности

Непрерывность или восстановление после сбоев?

Стратегия организации отказоустойчивости определяет набор технологий

Предотвращение сбоев:
непрерывность обработки
(disaster avoidance)

ЦОД-ы на «небольшом» расстоянии друг от друга, живая миграция, стратегия Active-Active



HA

Восстановление после сбоя:
катастрофоустойчивость
(disaster recovery)

Потеря основного/основных ЦОД влечет за собой восстановление на DR-площадке или в облаке



DR

Влияние расстояния

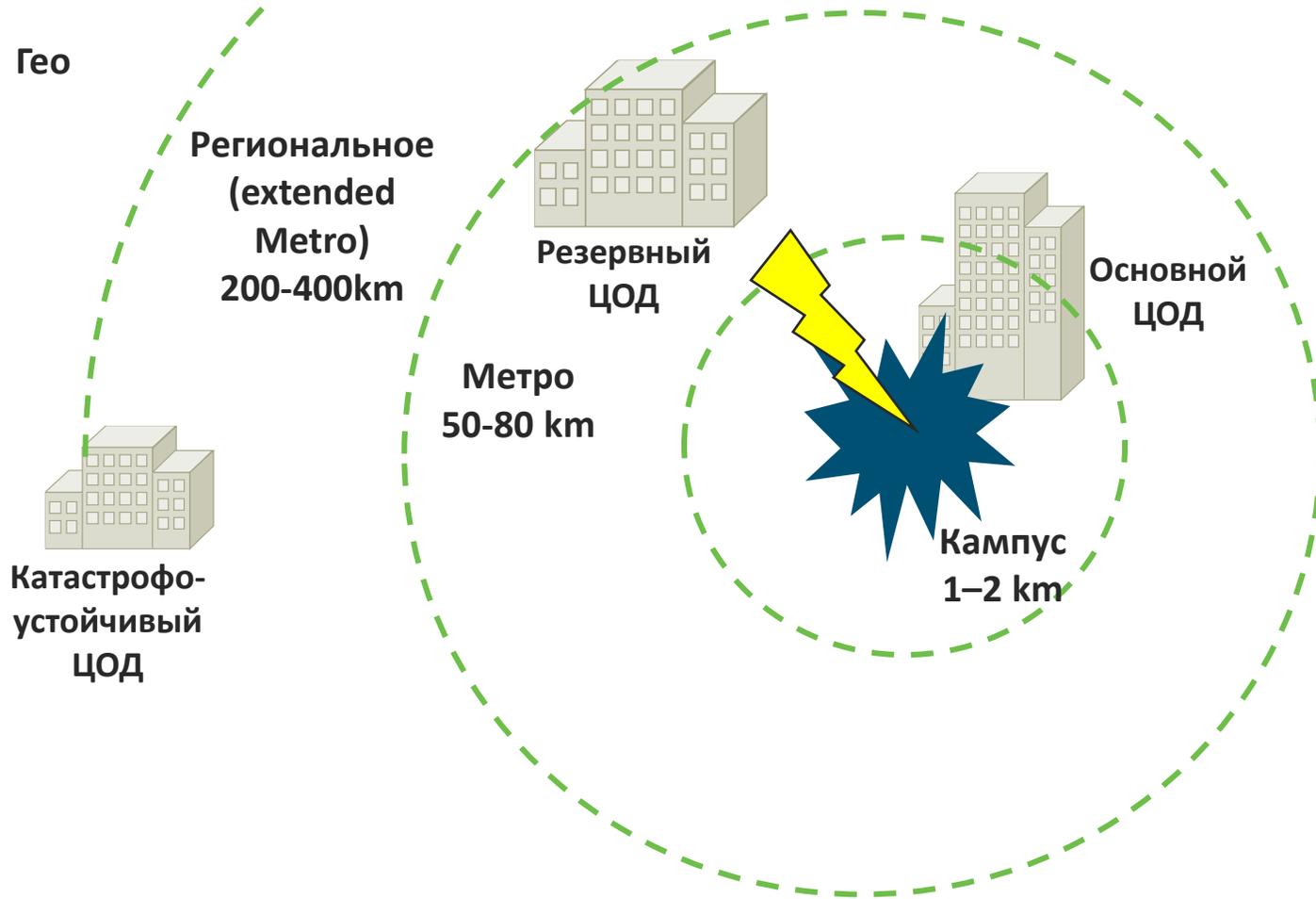
Задержка (latency)

- Скорость света в вакууме $\sim 300,000$ км/с
- Скорость света в оптоволокне: $\sim 200,000$ км/с
- Задержка сигнала: ~ 5 мкс/км, RTT ~ 10 мкс/км, ~ 1 мс на 100 км
- Для сравнения:
 - Среднее время доступа на (быстром) шпиндельном диске: $\sim 2-3$ мс
 - Среднее время доступа на SSD диске: десятки-сотни мкс
 - Максимальная задержка, допускаемая VMWare для vMotion: 10 мс RTT (до 150 мс для long-distance Vmotion)

Распределённые ЦОД

Классификация по расстоянию

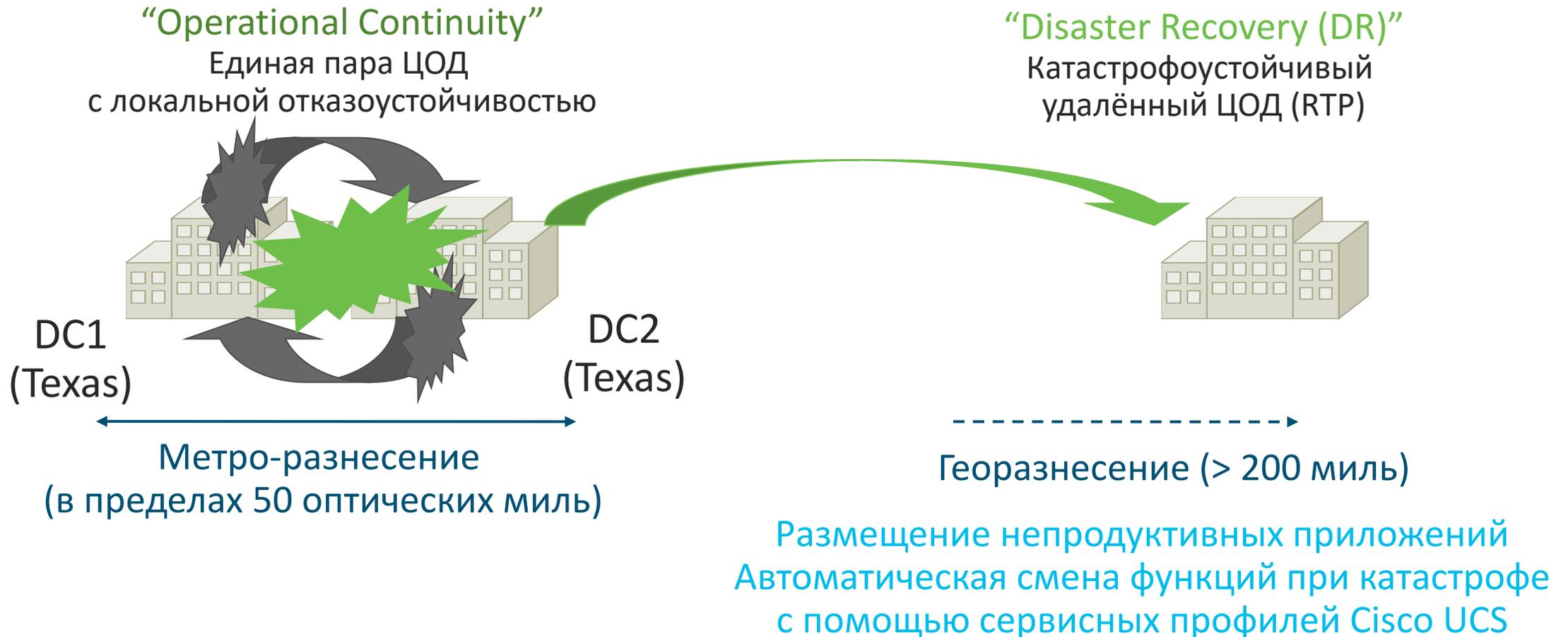
- Расстояние – ключевой фактор
- Ближе:
 - Выше производительность
 - Синхронная репликация
 - Проще коммуникации
- Дальше:
 - Катастрофоустойчивость
 - Распределение сервисов
- Нужен компромисс – или сочетание!



Пример: собственные ЦОД Cisco

Metro-Virtual DC (MVDC)

Непрерывность обработки, катастрофоустойчивость и оптимальное использование



Отказоустойчивые и катастрофоустойчивые ЦОД

Отличия подходов

	Отказоустойчивые ЦОД HA	Катастрофоустойчивые ЦОД DR
Приоритетная цель	Возможность «нулевых» RTO/RPO	«Отсутствие» единых источников простоя
Расстояние разнесения	Ограничено репликацией и кластеризацией	Максимально приемлемое
Кластеры	«Сильно связанные»	«Слабо связанные»
Аналоги в облаке (AWS)	Availability Zone	Region

Распределённые ЦОД

Технологические элементы

- Кластеризация на уровне прикладных систем, СУБД
- Кластеризация на уровне виртуализации и контейнерной оркестрации
- «Первичная сеть»
 - Тёмная оптика
 - CWDM/DWDM
 - IP магистраль
- Связь сетей передачи данных
 - L2 связность / L3 связность
 - Подключение к транспортной сети
 - Оптимальный путь исходящего/входящего трафика
 - Интеграция L4/L7 сервисов
- Связь ресурсов хранения данных
 - Репликация данных
 - Доступ к удалённым ресурсам хранения
 - Резервное копирование на удалённые сайты
- Кластеризация МСЭ
- ...

Отказоустойчивые и катастрофоустойчивые ЦОД

Типичные «строительные блоки»

	Отказоустойчивые ЦОД	HA	Катастрофоустойчивые ЦОД	DR
Связь ACI фабрик	ACI Multi-Pod или ACI Multi-Site		ACI Multi-Site	
Связь VXLAN/EVPN фабрик	VXLAN/EVPN Multi-Site (+DCNM, опционально - MSO)		VXLAN/EVPN Multi-Site (+MSO)	
Хранение	Синхронная репликация (+резервное копирование!)		Асинхронная репликация и/или резервное копирование	
Связь SAN	«Тёмная оптика» или xWDM		DWDM или FCIP, может отсутствовать	
Виртуализация (VMWare)	VMWare HA		VMWare SRM	
vCenter сервер	Общий или отдельные		Отдельные	
Контейнерные среды	Растянутые кластеры K8S (или KubeFed)		KubeFed	
Сервисы: МСЭ и т.д.	Кластер (Failover пара или A/A)		Независимые кластеры	

Распределённые ЦОД

Продукты, архитектуры и решения Cisco

- Транспортная сеть
 - Оптические системы DWDM передачи Cisco NCS 1K/2K/ONS 15216
 - Опорные маршрутизаторы для IP магистралей: Cisco ASR5K/9K, Nexus 9K
- Связь сетей передачи данных
 - ACI Multi-Pod и Multi-Site для связи сетей ЦОД на основе ACI
 - VXLAN/EVPN Multi-Site для связи фабрик VXLAN/EVPN
 - LISP и механизмы в ACI и VXLAN/EVPN для оптимизации пути
- Распределённые сервисы безопасности
 - Межсетевые экраны Cisco Firepower с поддержкой кластеризации
- Связь сетей хранения данных
 - SAN коммутаторы Cisco MDS 9000 для транспорта FC, FCoE и связи по FCIP
- Вычислительные средства и гиперконвергенция
 - Вычислительные системы Cisco UCS с поддержкой смены «ролей» при катастрофе
 - Репликация и кластеризация в гиперконвергентном решении Cisco HyperFlex
 - Совместные с партнёрами решения для резервного копирования

Варианты организации транспортной сети

Транспортная («первичная») сеть

Типовые варианты

Тёмная оптика

- Ограничения расстояния:
 - затухание
 - дисперсия
- Зависит от трансиверов
 - И их поддержки на оборудовании
- «Клиентская защита»
 - Механизмами конкретных протоколов
- Опция – xDWM трансиверы без DWDM/CWDM системы



DWDM/CWDM

- Возможность усиления и компенсации дисперсии
 - Больше расстояние
- Больше сервисов на оптоволокно
- Клиентские xWDM трансиверы или транспондеры/мукспондеры
- Поддержка Ethernet и FC
 - Зависит от системы или трансиверов
 - Проверяйте ограничения и особенности
- Возможны механизмы «оптической» защиты



Опорная сеть

- L3 (предпочтительно) или L2
- Собственная или сервис от провайдера
- «Нет» ограничений на расстояние
- Проверяйте ограничения и особенности
 - Протоколы маршрутизации
 - Multicast
 - MTU
 - QoS
 - Переподписка / SLA



Транспортная («первичная») сеть

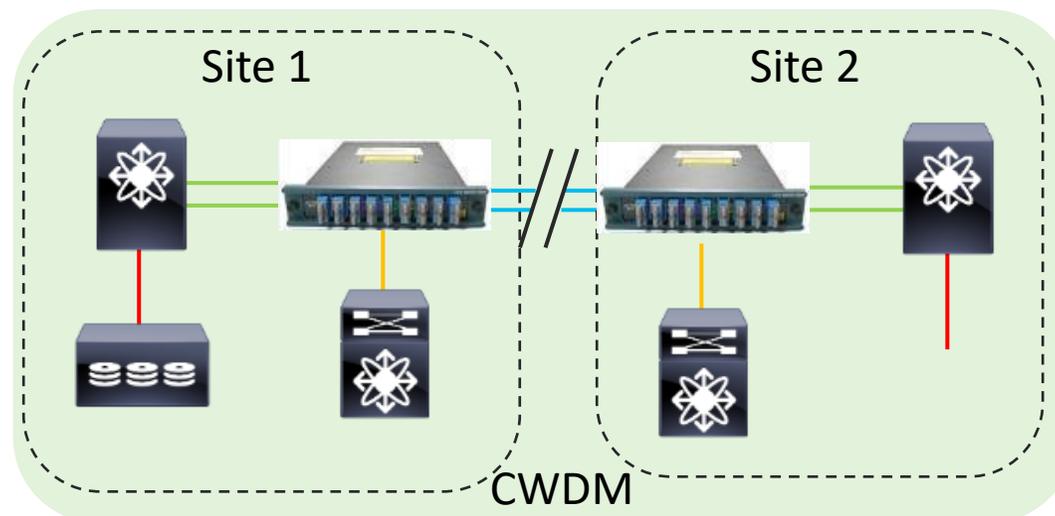
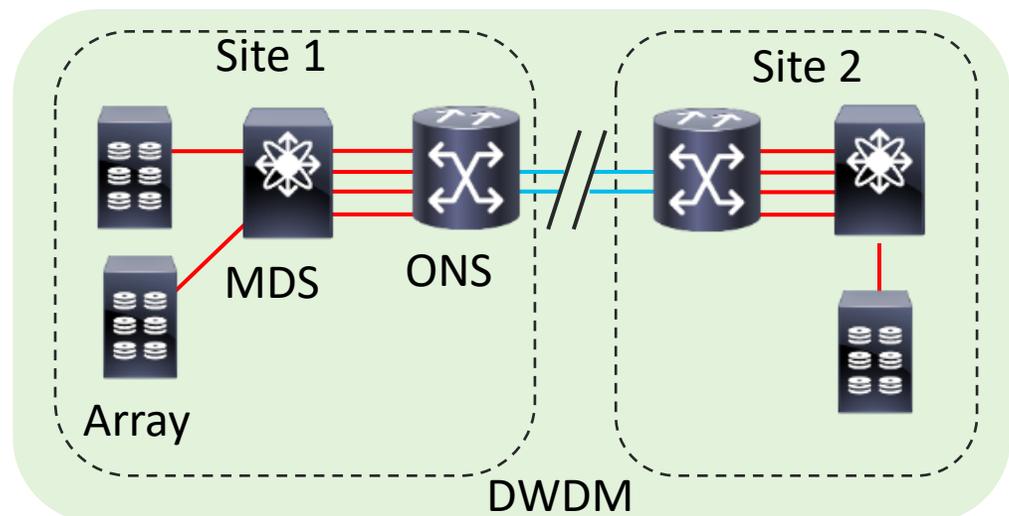
Тёмная оптика: варианты «дальнобойных» стыков для оборудования Cisco

- Ethernet:
 - 10GBASE-ZR: 80 км
 - 40GBASE-ER4: 40 км
 - 100GBASE-ER4: 40 км
 - 400GBASE-LR8: 10 км
 - 400GBASE-ZR+ (скоро): 80+ км (при использовании EDFA)
- Fibre Channel:
 - 8G-ER: 40 км
 - DS-8G-ZR (SmartOptics): 70 км
 - FC16GELW: 25 км
 - DS-16G-ER(SmartOptics): 40 км
 - FC32G-LW: 10 км

Транспортная («первичная») сеть

DWDM и CWDM

	DWDM	CWDM
Применение	Большие расстояния	Metro
Оптические усилители	Обычно – EDFA	Очень редко
Число каналов	До 80 и более	До 8
Разнесение каналов	0.4 nm	20 nm
Расстояния	До 3000 км	До 80 км
Спектр (типично)	1530nm - 1560nm	1270nm - 1610nm
Фильтры	Интеллектуальные	Пассивные
Оптическая защита	Возможна	Нет



Оптические DWDM системы Cisco

Решения для задач ЦОД

Семейство NCS 1000

Система, оптимизированная
для связи ЦОД



Семейство NCS 2000

Гибкое решение
операторского класса



Защита вычислительных и гиперконвергентных систем

Подробнее – в день 2, 3

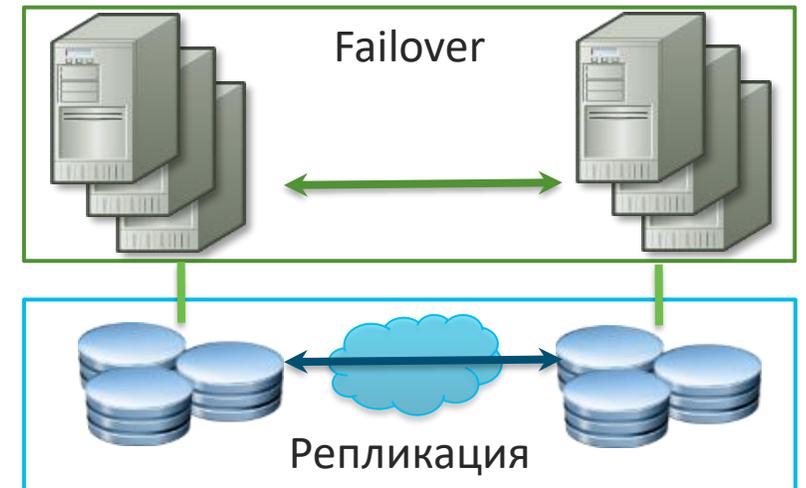
Разнесение вычислительных систем по разным ЦОД

Вопрос №1. Как осуществляется репликация:

- Средствами приложений;
- Средствами дополнительного ПО;
- Средствами дисковых массивов;

Вопрос №2. Как осуществляется переключение:

- Средствами приложений;
- Средствами платформы виртуализации;
- Средствами дополнительного или встроенного в ОС ПО кластеризации;
- Вручную, или средствами автоматизации



Защита на уровне приложений



- Бизнес-критичные приложения и СУБД как правило имеют собственные средства репликации данных
 - Oracle, MS SQL, SAP HANA и др.
- Репликация (чаще всего синхронная) может осуществляться СХД
- Автоматизация восстановления обеспечивается самим ПО или средствами кластеризации, с учетом специфики приложения
 - Windows Failover Cluster, SLES HAE, и др., например, Veritas
- Отдельное внимание требуется к интеграции средств репликации и средств автоматизации восстановления



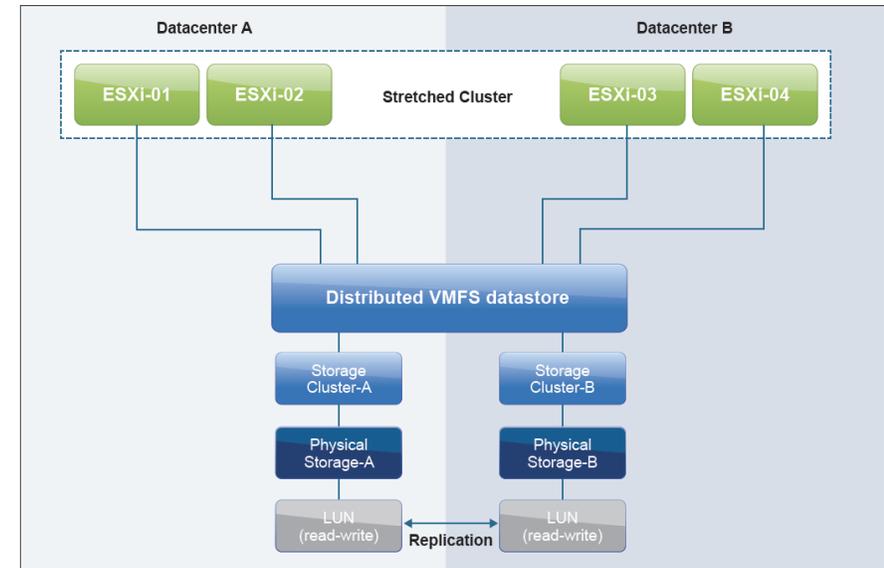
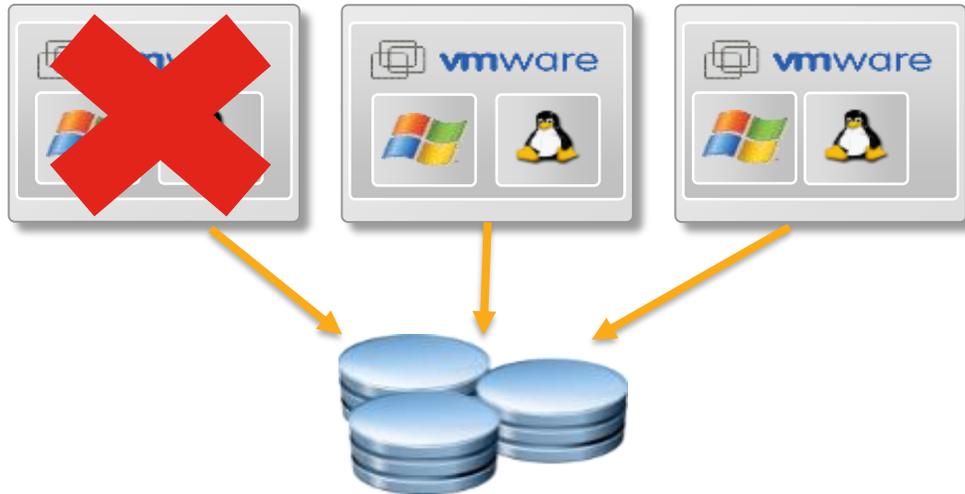


Защита на уровне виртуализации

- «Растянутые» кластеры
 - Используется синхронная репликация средствами СХД или платформ гиперконвергенции
 - Восстановление работоспособности – средствами HA гипервизора
- Disaster Recovery решения
 - Как правило, асинхронная репликация средствами СХД или дополнительного ПО
 - Восстановление или вручную или дополнительным ПО

Защита на уровне виртуализации

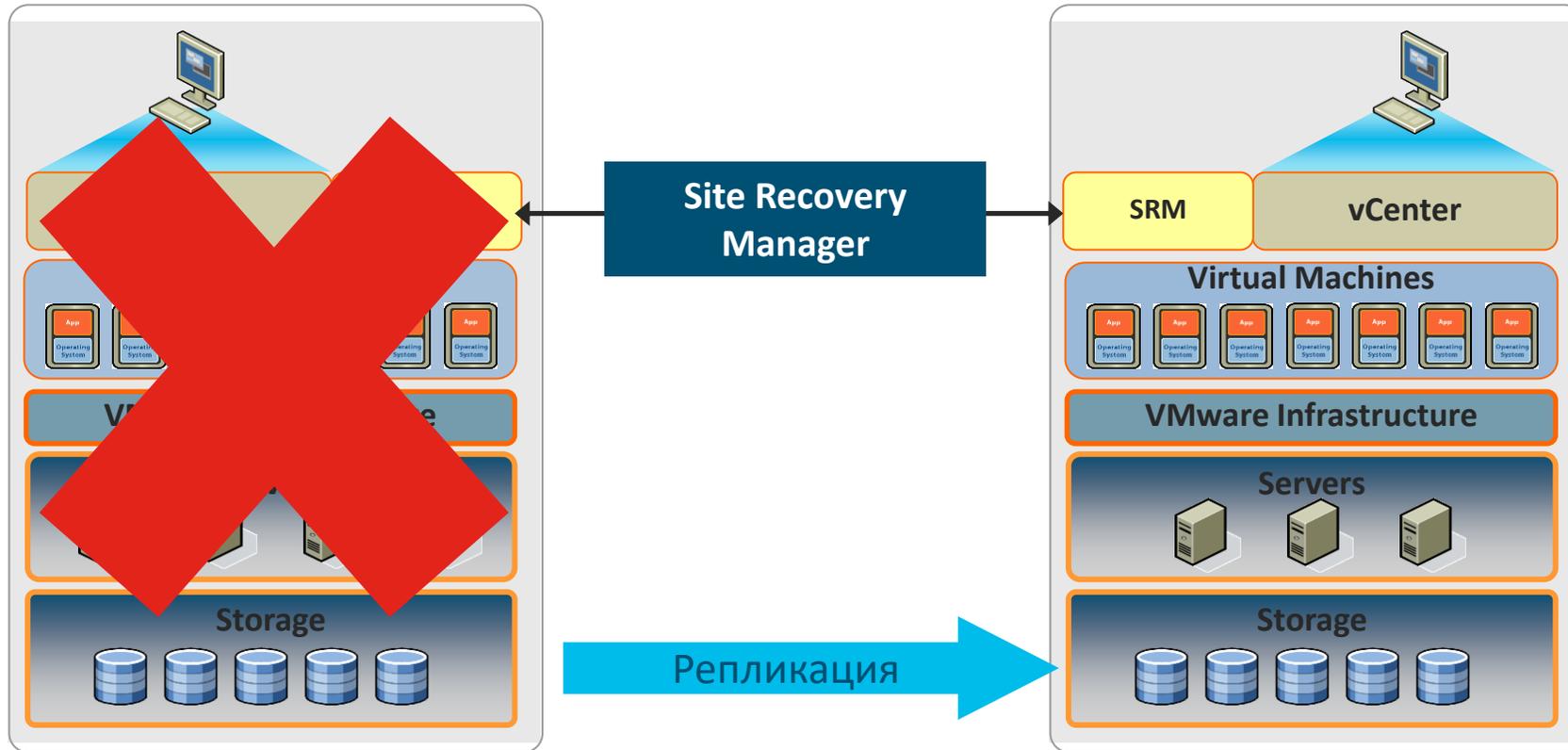
HA сценарии на примере VMware vSphere и HA



- Автоматический перезапуск VM в пределах кластера при отказе или недоступности хоста
- Необходимо разделяемое хранилище
- Возможна реализация HA между ЦОД, при «растягивании» хранилища за счет синхронной репликации СХД или гиперконвергентными системами

Защита на уровне виртуализации

DR сценарии на примере VMware vSphere и SRM

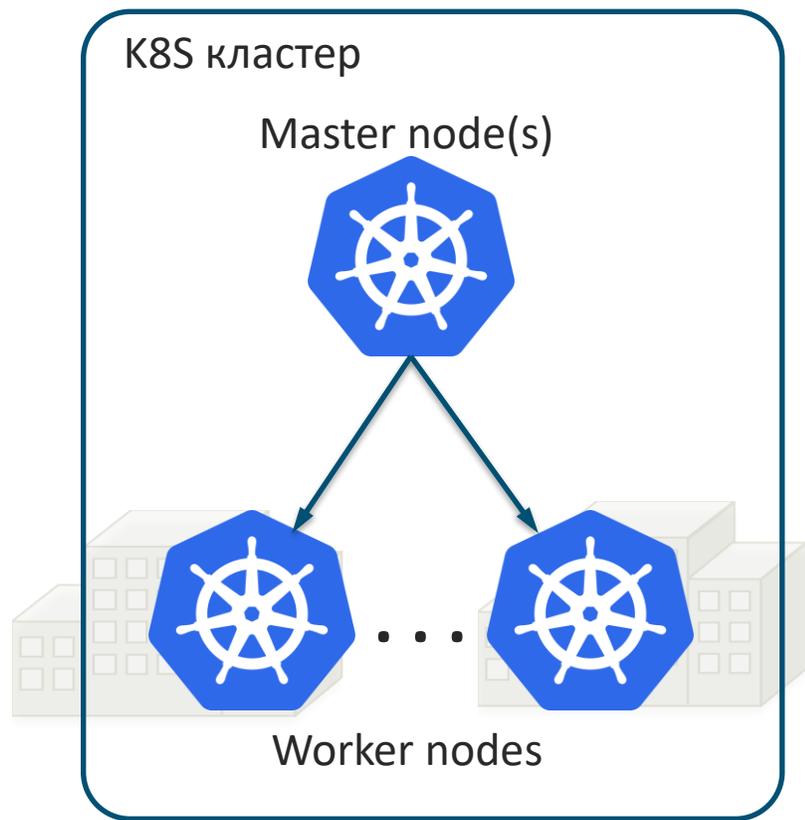


- Ручной или автоматический запуск систем в резервном ЦОД при аварии основного
- Поддерживаются сценарии (планы) восстановления с кастомизацией среды
- Не требует «растягивания» сетей – возможна смена адресов
- Требуется репликация – средствами СХД или HCI или дополнительного ПО

Внедрение Kubernetes в отказоустойчивых ЦОД

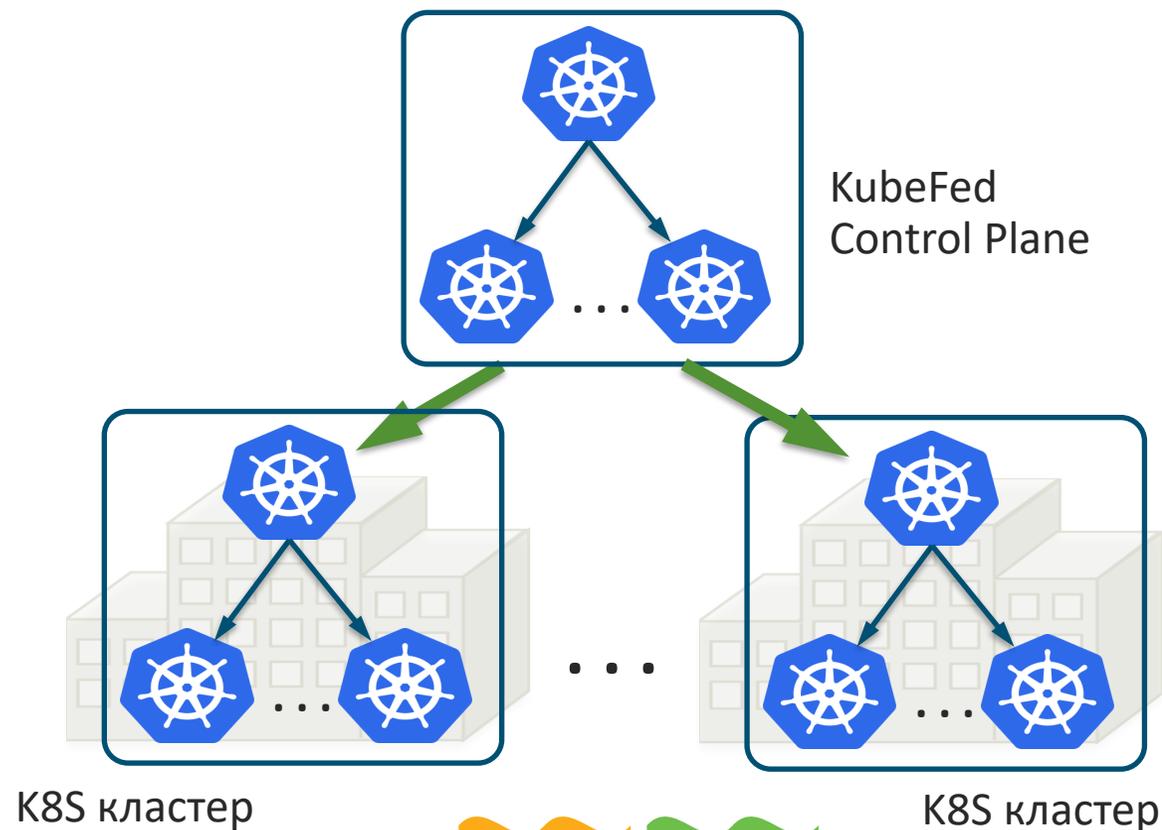
Кластеры и KubeFed

Единый кластер



HA

Федерация кластеров



HA DR

Защита на уровне оборудования

Возможности Cisco UCS для реализации DR

- 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 «перемещен» на другой сервер, в другую систему или в другой ЦОД
- Как правило, используется загрузка с SAN и репликация средствами СХД



DR с применением UCS

Сценарии

- Восстановление физического сервера на той же площадке
- Восстановление физического сервера на другой площадке (с изменением WWN и загрузочного LUN или без)
- Восстановление всей UCS системы целиком из полной резервной копии конфигурации
- Использование в качестве дополнительного уровня защиты: для доступности приложения используется кластеризация, сервисные профили обеспечивают восстановление защищенности
- Как правило, до аварии «резервная» UCS система используется для задач тестирования и разработки





Защита данных в HyperFlex

Файловая система Enterprise уровня

Целостность данных и надежность

- Контрольные суммы для защиты от ошибки носителя данных
- Увеличение срока службы SSD дисков на уровне архитектуры
- Снэпшоты без снижения производительности

Высокая доступность

- Полностью распределенная архитектура для быстрого восстановления
- Быстрая синхронизация и эффективное перераспределение данных
- Полный апгрейд без останова системы

Асинхронная репликация между площадками



Партнерские решения для резервного копирования

COHESITY

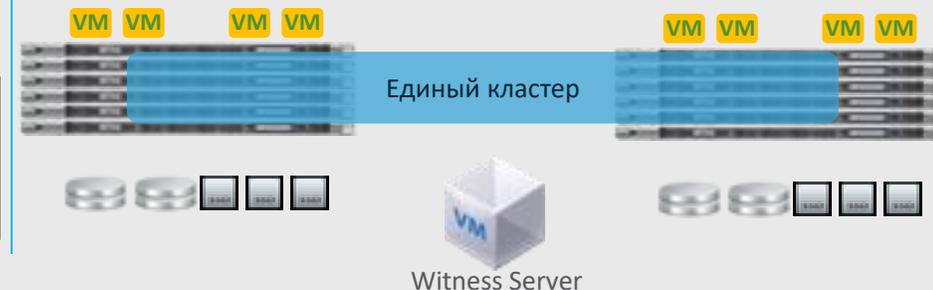
COMMAVAULT

VEEAM

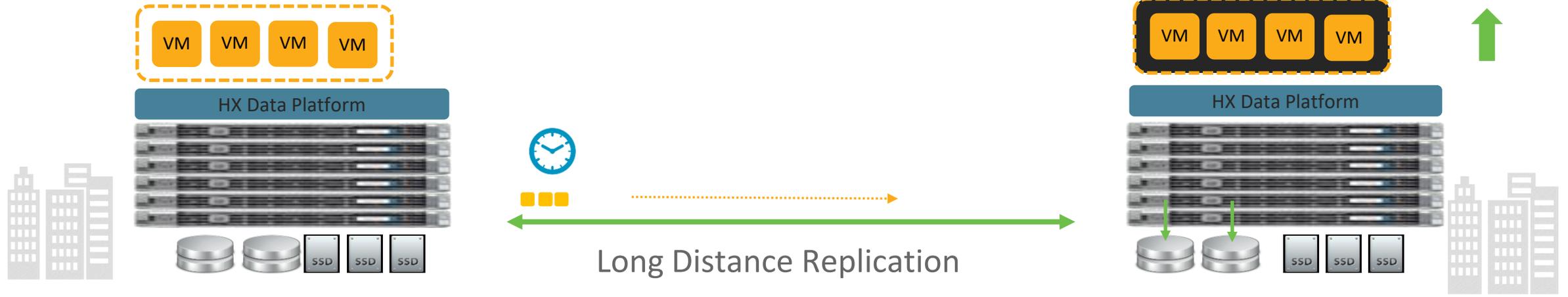
Logical Availability Zones



Синхронная репликация



Асинхронная репликация систем Hyperflex Disaster Recovery встроенными средствами



Автоматизированные сценарии восстановления



Тест DR

- Готовность к DR
- Кастомизация параметров теста DR



Плановая миграция

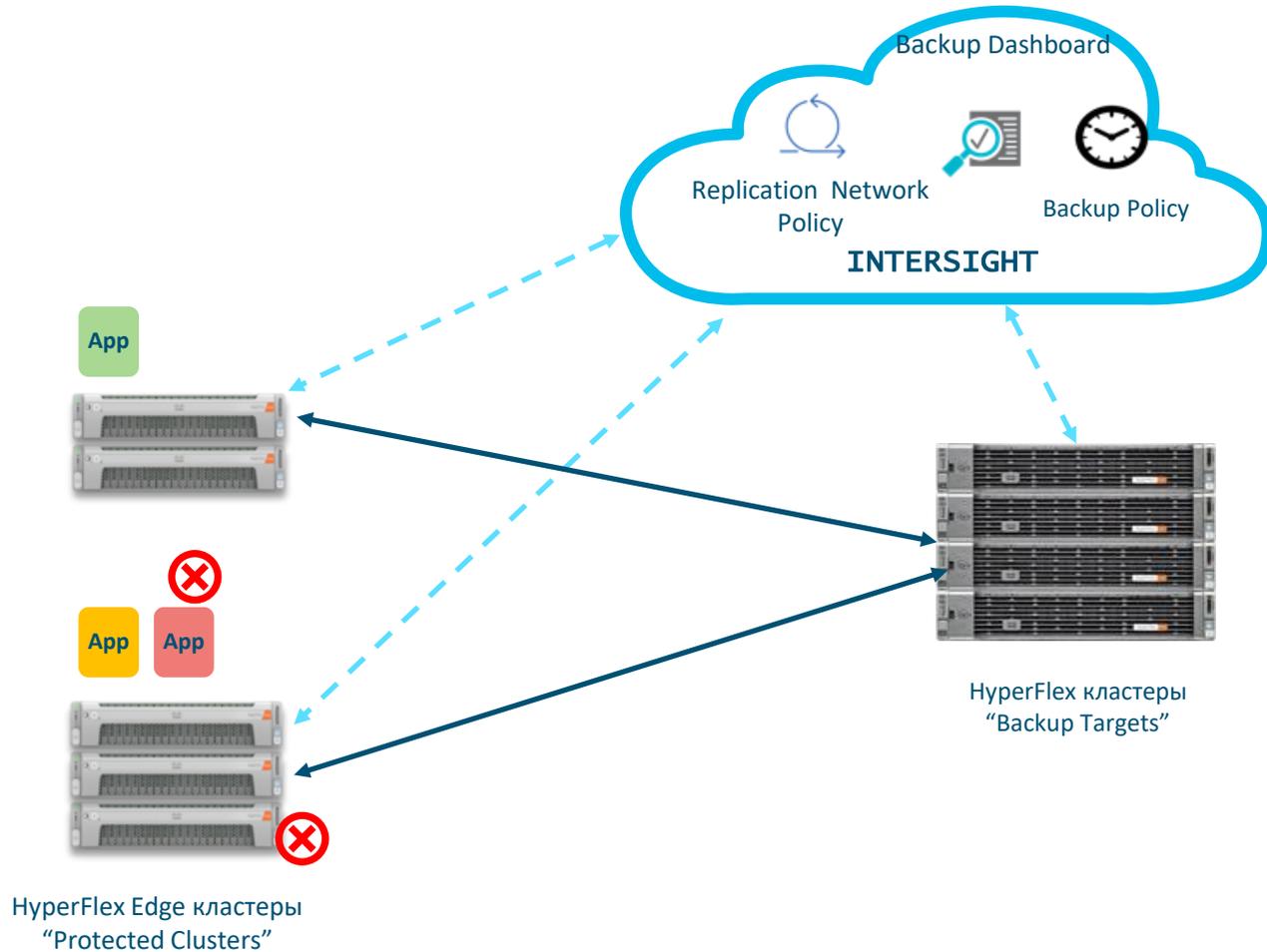
- Миграция VM между кластерами и ЦОД
- Защита после миграции



Отработка отказа

- Восстановление VM после аварии
- Защита после восстановления

Асинхронная репликация HyperFlex Edge N:1



Сценарии

- #1 Защита VM на NX Edge от человеческого фактора и программных ошибок → VM восстанавливаются из локальных копий в тот же Edge кластер

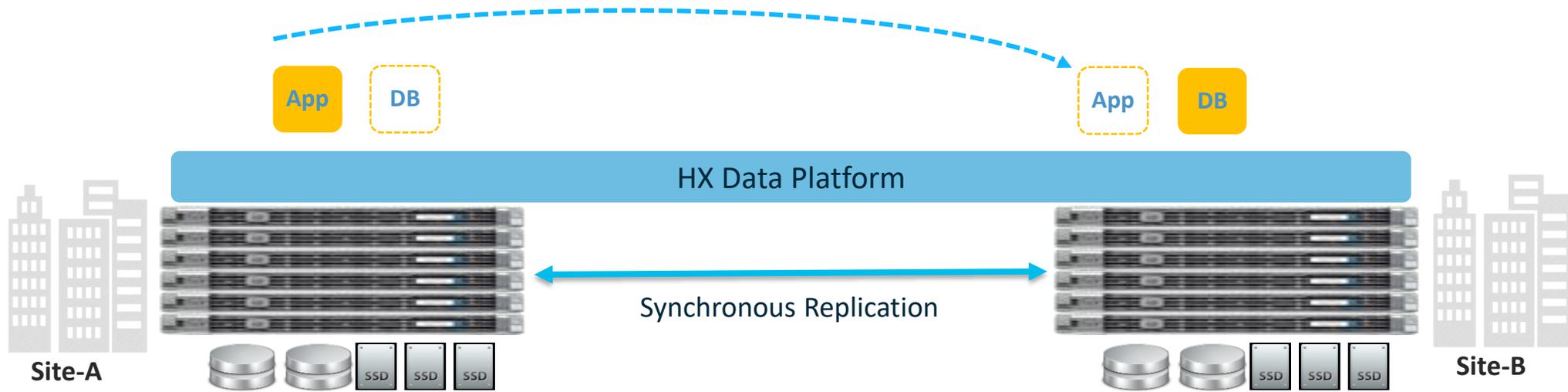
- #2 Защита VM на NX Edge от сбоя кластера или всей площадки → VM восстанавливаются из копий в другой Edge кластер

- #3 Миграция VM с NX Edge на другой NX Edge кластер в другом ЦОД → VM мигрируют на другой Edge кластер

HyperFlex Stretched Cluster– «растянутый кластер»

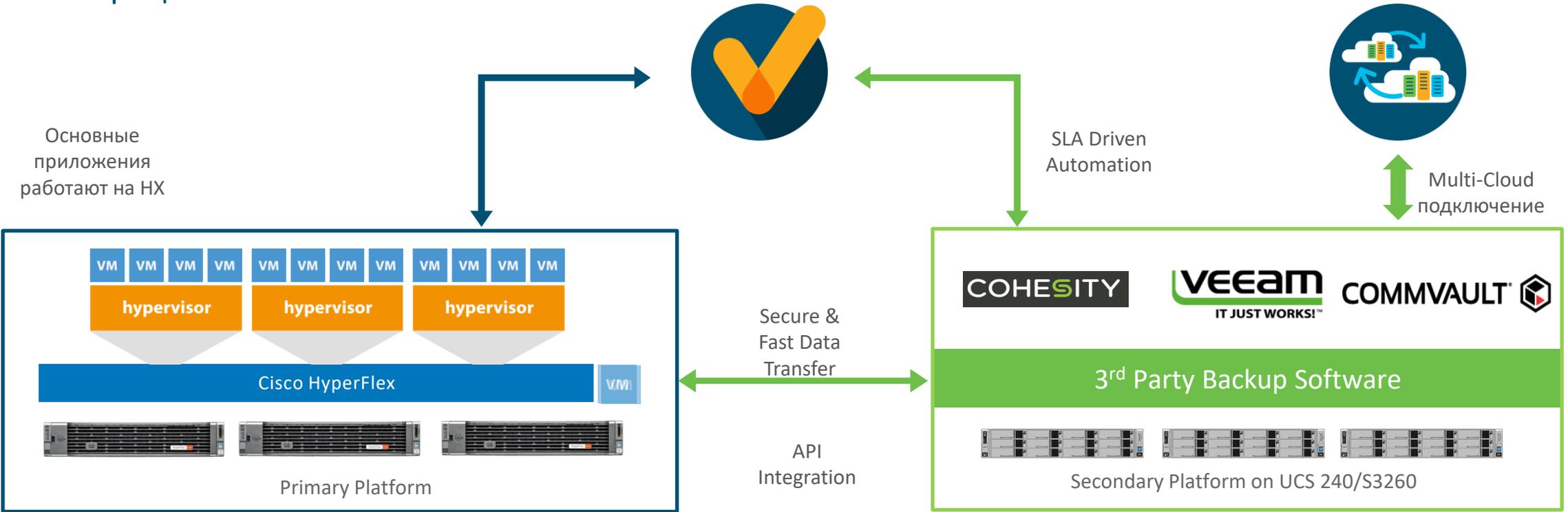
Распределенная между двумя площадками система Active-Active для поддержки бизнес-критичных задач, требующих высокую доступность (RTO = время рестарта VM) и отсутствие потери данных («нулевой» RPO)

- Вторая площадка может быть «через дорогу» или за сотни километров (5мс RTT)
- Требуется третья площадка для арбитра (небольшая служебная VM)
- «Обычный» кластер VMware – не требуется средств оркестрации DR



Резервное копирование Hyperflex

Интеграция со снимками НХ

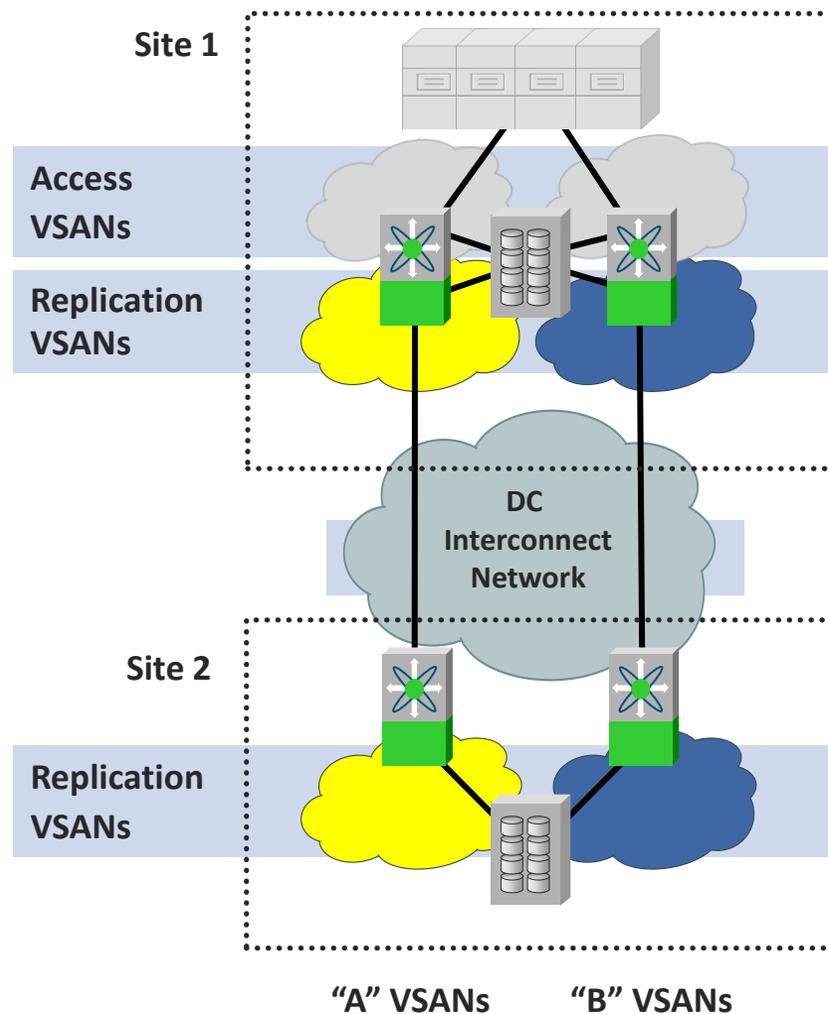


Просто	Преимущества UCS: быстрое развертывание, легкое обслуживание, единая архитектура для первичных и вторичных систем
Надежно	Значительно эффективнее чем redo log, 24x7 Backup & Restore, Long distance DR, DR to Cloud
Быстро	Высокая производительность IO, небольшие окна backup, быстрое восстановление
Оптимизировано	Дедупликация и компрессия первичных и вторичных данных

Связь сетей хранения данных

Подробнее – в день 2

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

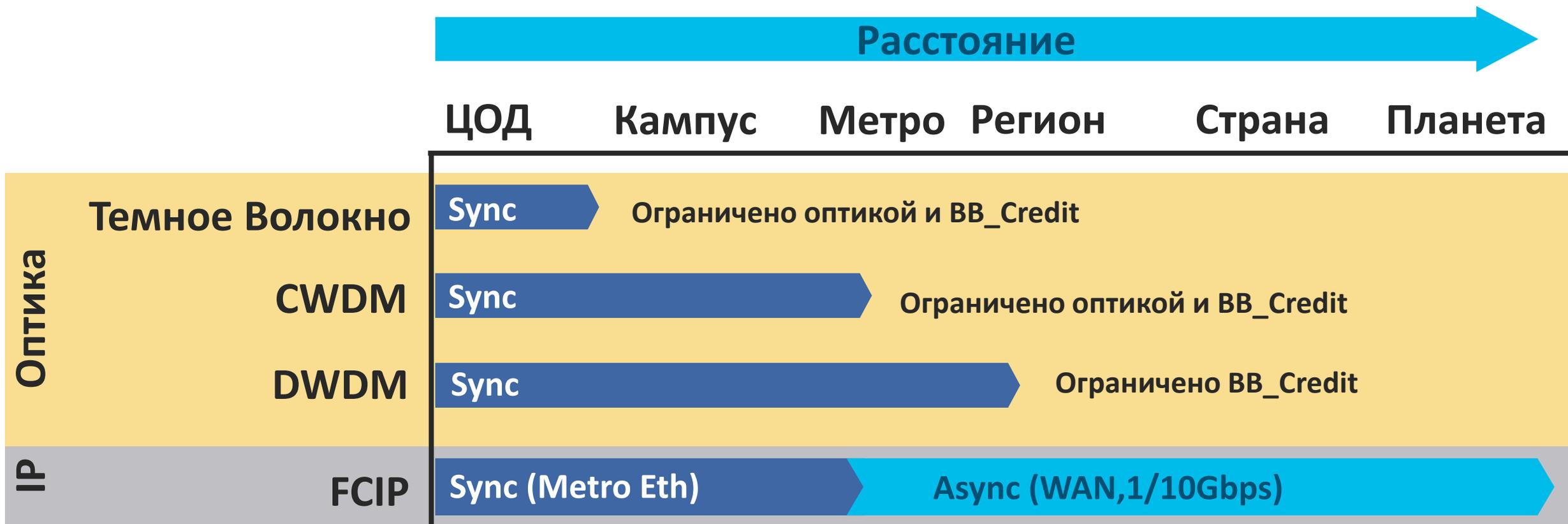
- Требования могут зависеть от технологии репликации
- Физическая или логическая (VSAN) изоляция

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

Влияние расстояния

Варианты объединения Fibre Channel SAN



Число B2B Credits для актуальных моделей MDS

И ограничения по расстоянию для связи по FC

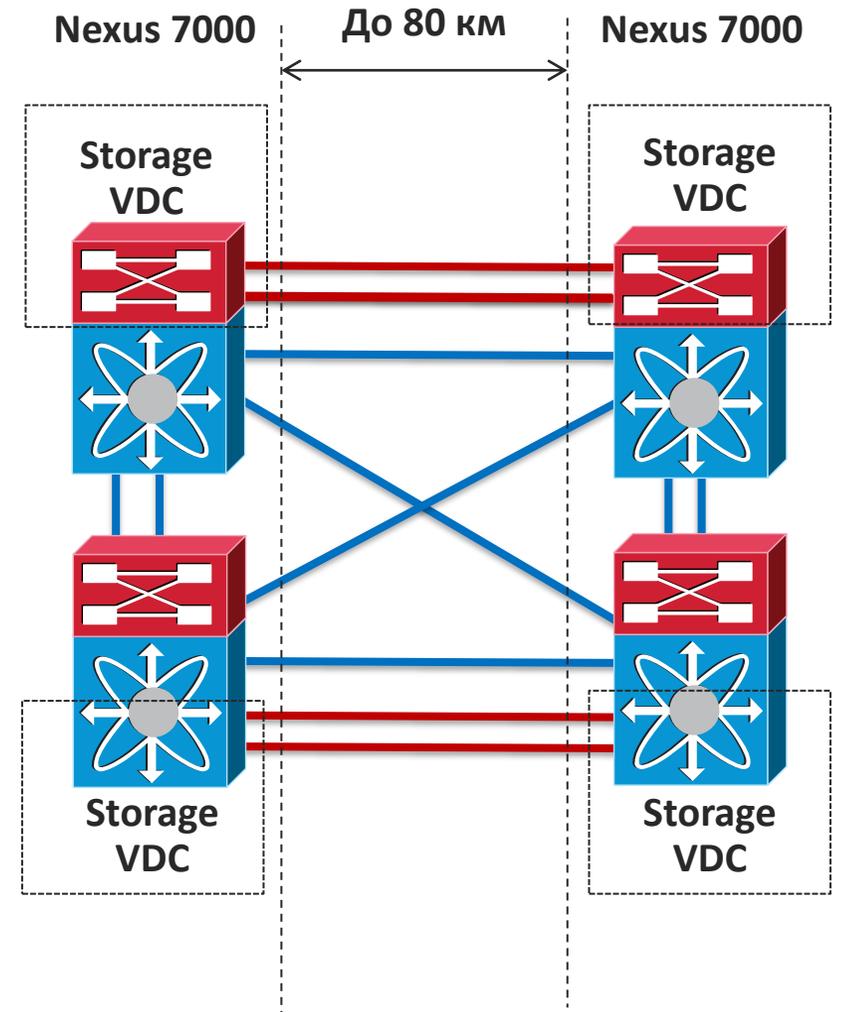
Модель коммутатора/ карты	Максимальное число B2B Credits на порт - по умолчанию	Максимальное число B2B Credits на порт - с лицензией Enterprise	Максимальное расстояние(км)*, примерно
MDS 9700 / DS-X9648-1536K9 (32G FC)	500	8191	510
MDS 9250i (16G FC)	253	253	31
MDS 9148S (16G FC)	253	253	31
MDS 9396S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

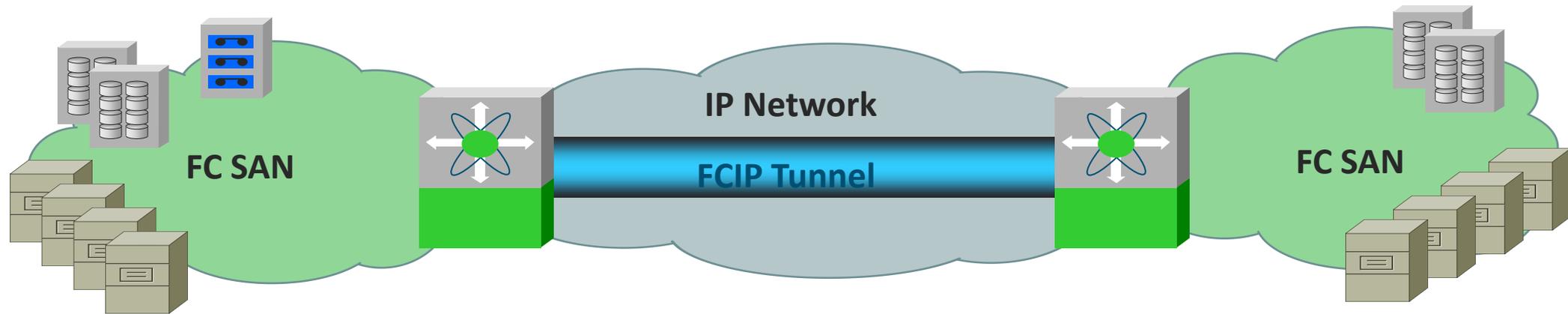
FCoE для связи SAN между ЦОД?

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:
 - 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



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 сжатия и оптимизации
- Создаётся единая FC фабрика (общий FSPF домен)
- Транспорт – IP сеть, в том числе и на большие расстояния (тысячи км)

Связь сетей передачи данных

Подробнее – в день 4

Разнесение ЦОД и передача данных

Полная картина



Связь сетей ЦОД

Вчера и сегодня

Было:

- L2 связь
 - OTV
 - «двухсторонний VPC»
 - L2 VPN (VPLS/EoMPLS pseudowire)
+ механизмы резервирования
- L3 связь
 - Отдельные линки
 - L3 VPN
 - L3 через VPC

Сейчас:

- Опора на технологии фабрик
- L2+L3 связь!
 - ACI Multi-Pod
 - ACI Multi-Site
 - VXLAN/EVPN Multi-Site
- Подключение legacy сетей ЦОД
 - ACI site (вариант – ACI mini)
 - ACI remote leaf
 - Пара VXLAN/EVPN VPC BGW

Связь сетей ЦОД с опорой на Cisco ACI

Базовые варианты

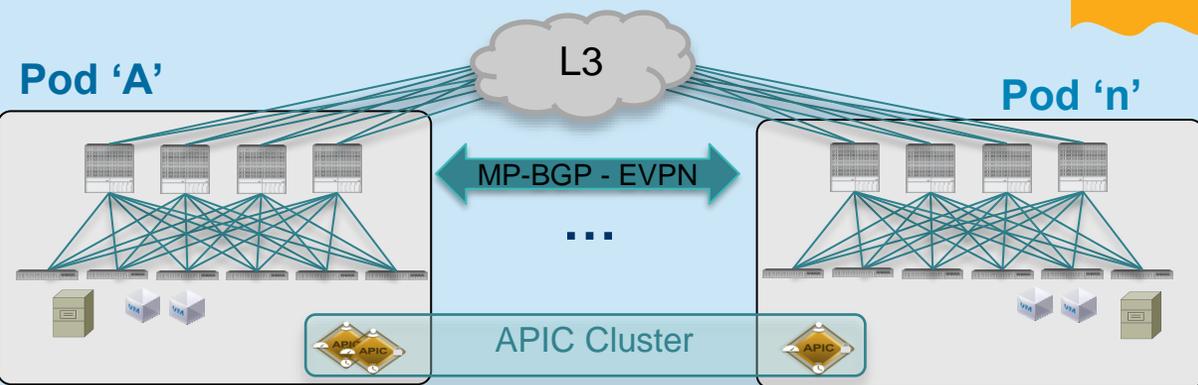
Единый APIC кластер/домен

Много APIC кластеров/доменов

Связь ЦОД на базе ACI

Multi-Pod

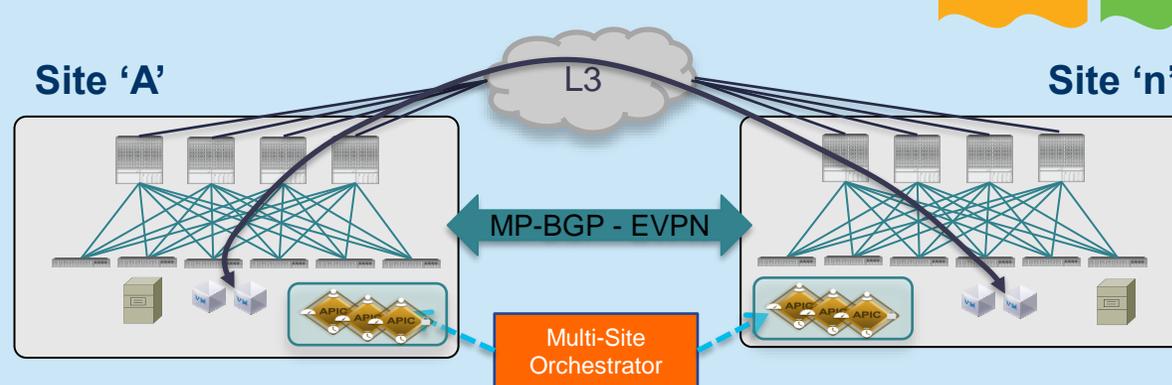
HA



Multi-Site

HA

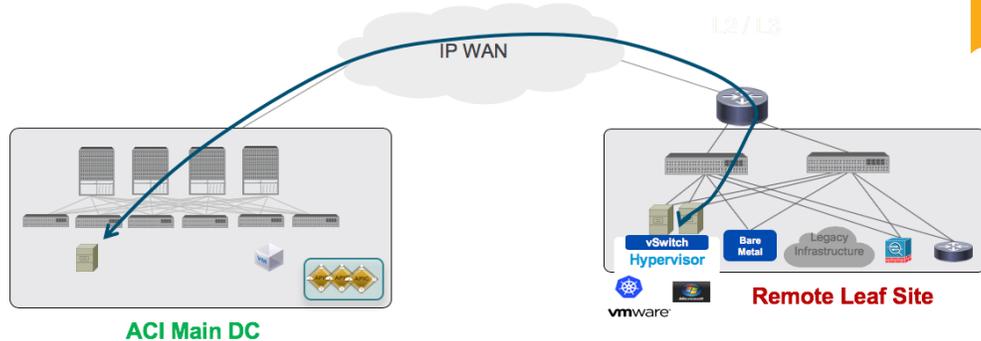
DR



Связь традиционных ЦОД и ACI

ACI Remote Leaf

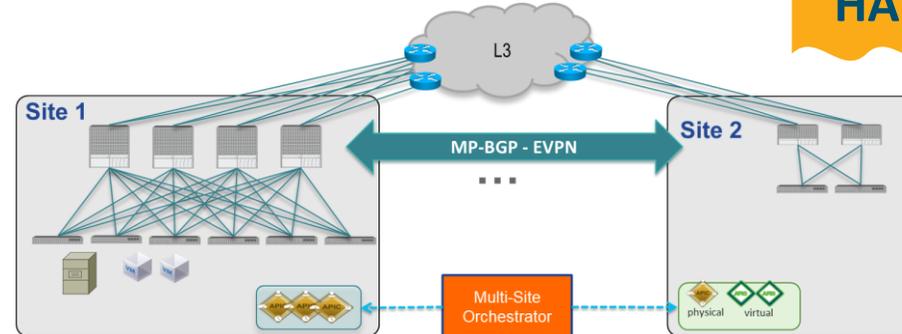
HA



ACI Mini + Multi-Site

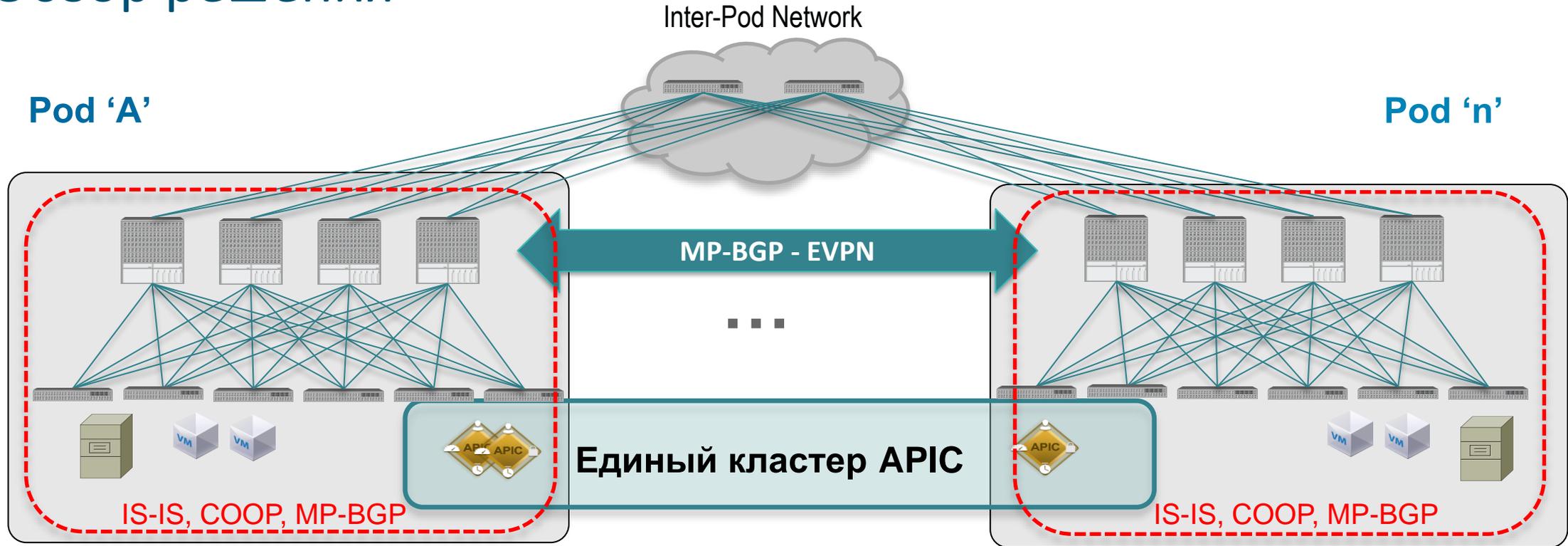
HA

DR



ACI Multi-Pod

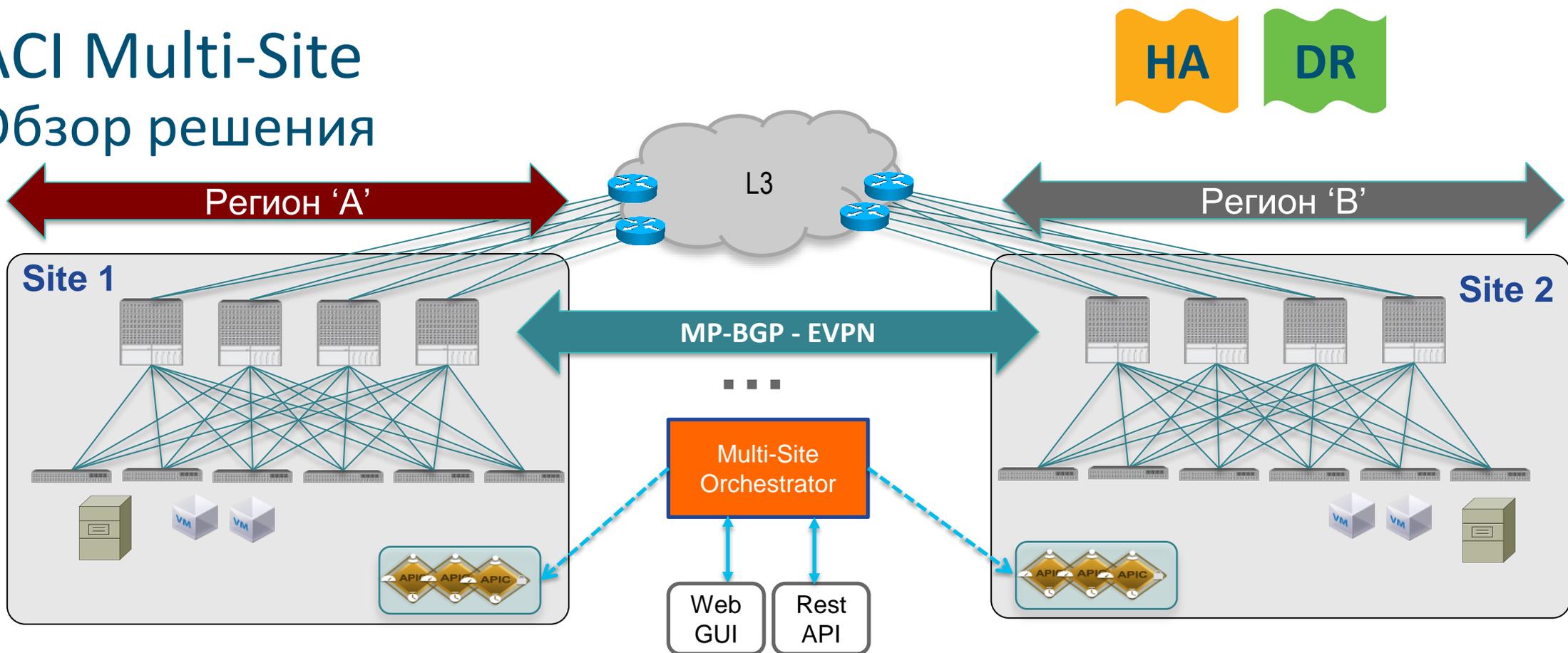
Обзор решения



- Несколько модулей (ACI Pod) связанных IP сетью (Inter-Pod network), каждый состоит из набора leaf и spine
- Управлением единым кластером APIC
- Единый домен управления и политик
- Изоляция доменов отказов протоколов control plane (IS-IS, COOP)
- Инкапсуляция VXLAN между модулями
- EVPN (MP-BGP) для плоскости управления
- Сквозное применение политик

ACI Multi-Site

Обзор решения

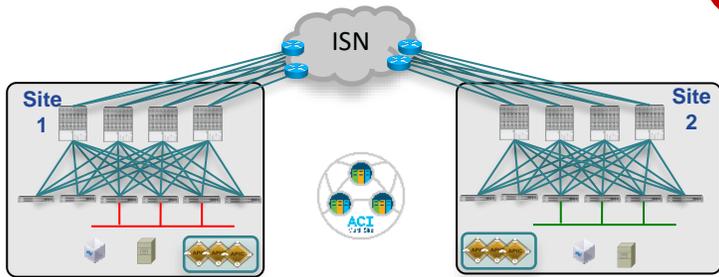


- Отдельные ACI фабрики с независимыми кластерами APIC
- Каждая фабрика рассматривается как отдельный «регион» и «зона доступности»
- Ограничение зоны вносимых изменений
- Использование для сценариев DR и Active/Active
- «Нет ограничений» по расстоянию
- Multi-Site контроллер настраивает сквозные конфигурации в нескольких кластерах APIC
- MP-BGP EVPN с VXLAN инкапсуляцией между сайтами
- Сквозное применение политик
- Поддержка Multi-Pod сайтов

ACI Multi-Site - организация сетевой связанности

Варианты настройки Bridge Domain (L2-сегмент)

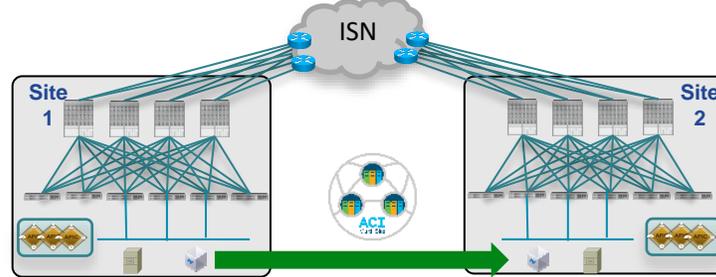
Только L3 связанность



- Bridge Domain и подсети не растягиваются между сайтами
- Layer 3 Intra-VRF или Inter-VRF взаимодействие (разделяемые сервисы между VRF/Tenant-ами)



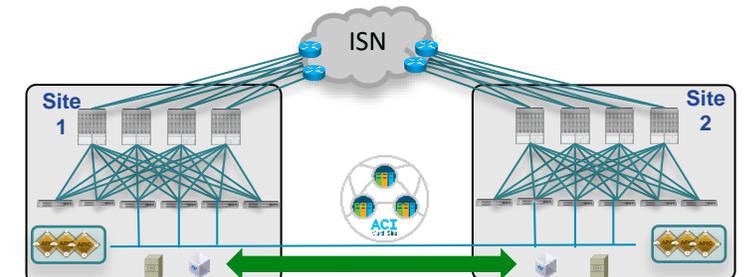
IP мобильность без передачи BUM трафика



- Одна и та же IP подсеть определена на разных сайтах
- Поддерживается IP Mobility ('холодная' и 'живая'* миграция VM) и взаимодействие между хостами внутри одной подсети
- Нет Layer 2 BUM флдинга между сайтами



Классическая L2 связанность



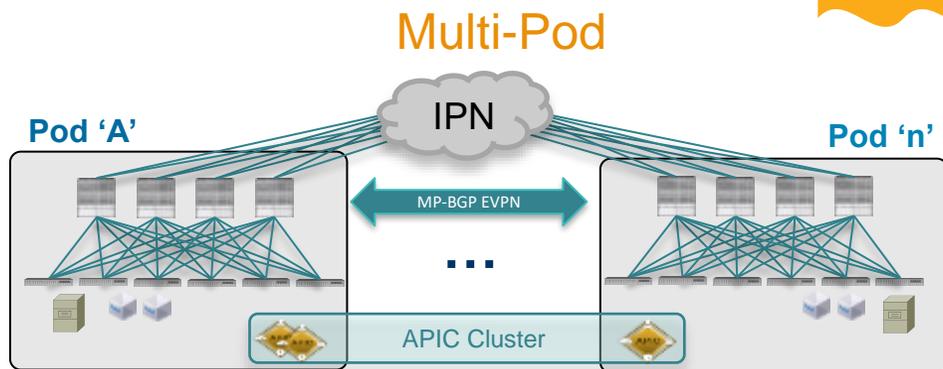
- Взаимодействие между сайтами с возможностями изоляции доменов сбой
- L2 растягивается между сайтами, 'живая'* VM миграция и кластеризация приложений
- Есть Layer 2 BUM флдинг между сайтами



ACI Multi-Pod и ACI Multi-Site

Основные отличия

HA



Простота эксплуатации

Меньше число APIC

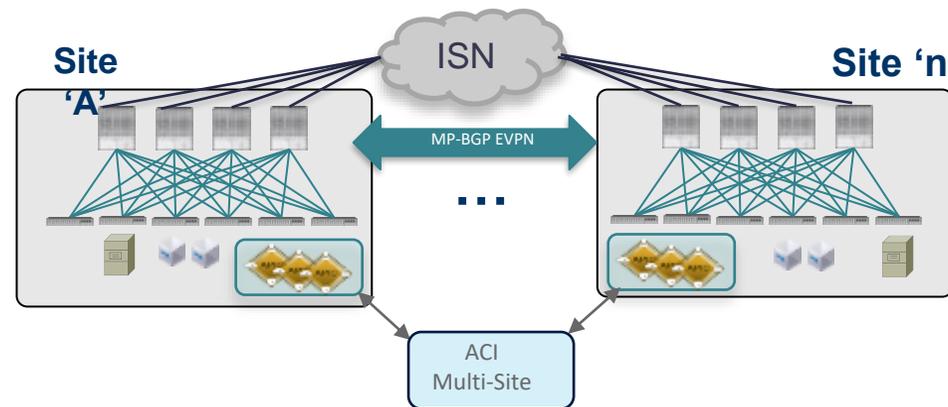
Сквозные функции между модулями

Единые VMM на все модули

HA

DR

Multi-Site



Изоляция доменов настройки

Более высокая допустимая задержка на транспорте

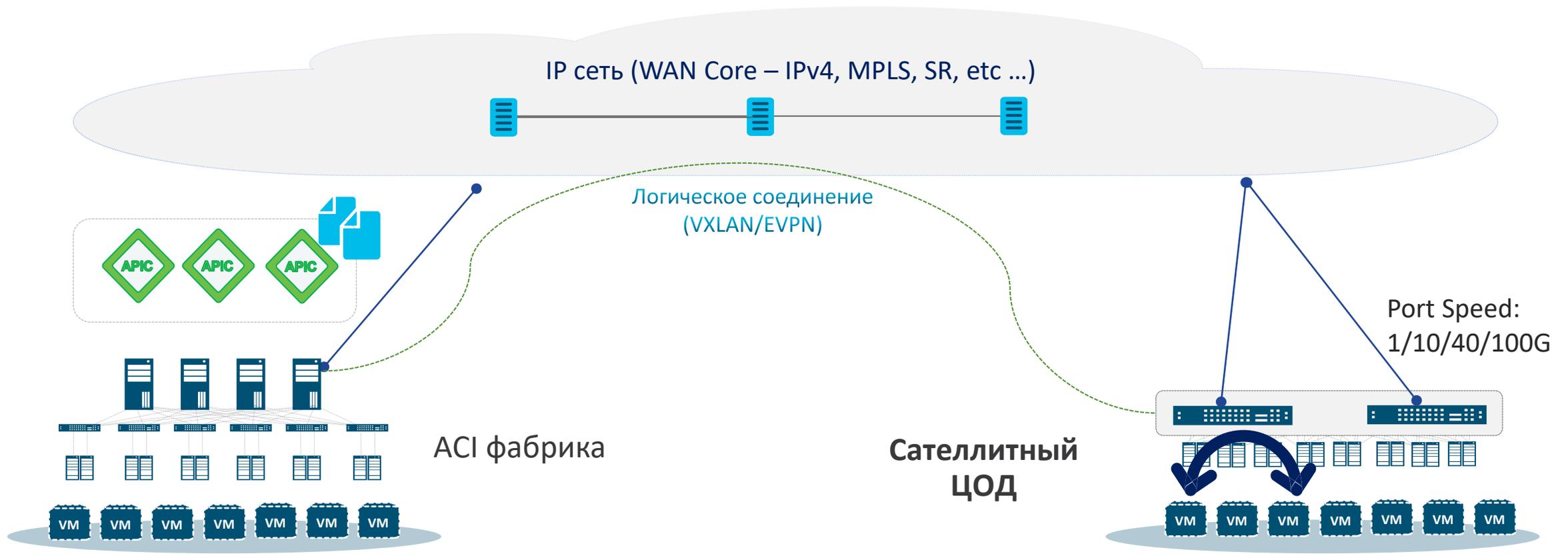
Выше число узлов

Не нужен multicast на транспорте

Требуется MSO и лицензии Advantage

«Сателлитные ЦОД»: ACI Remote Leaf

Вариант подключения традиционных сетей в Multi-Pod архитектуре



Автоматическое обнаружение вынесенных коммутаторов

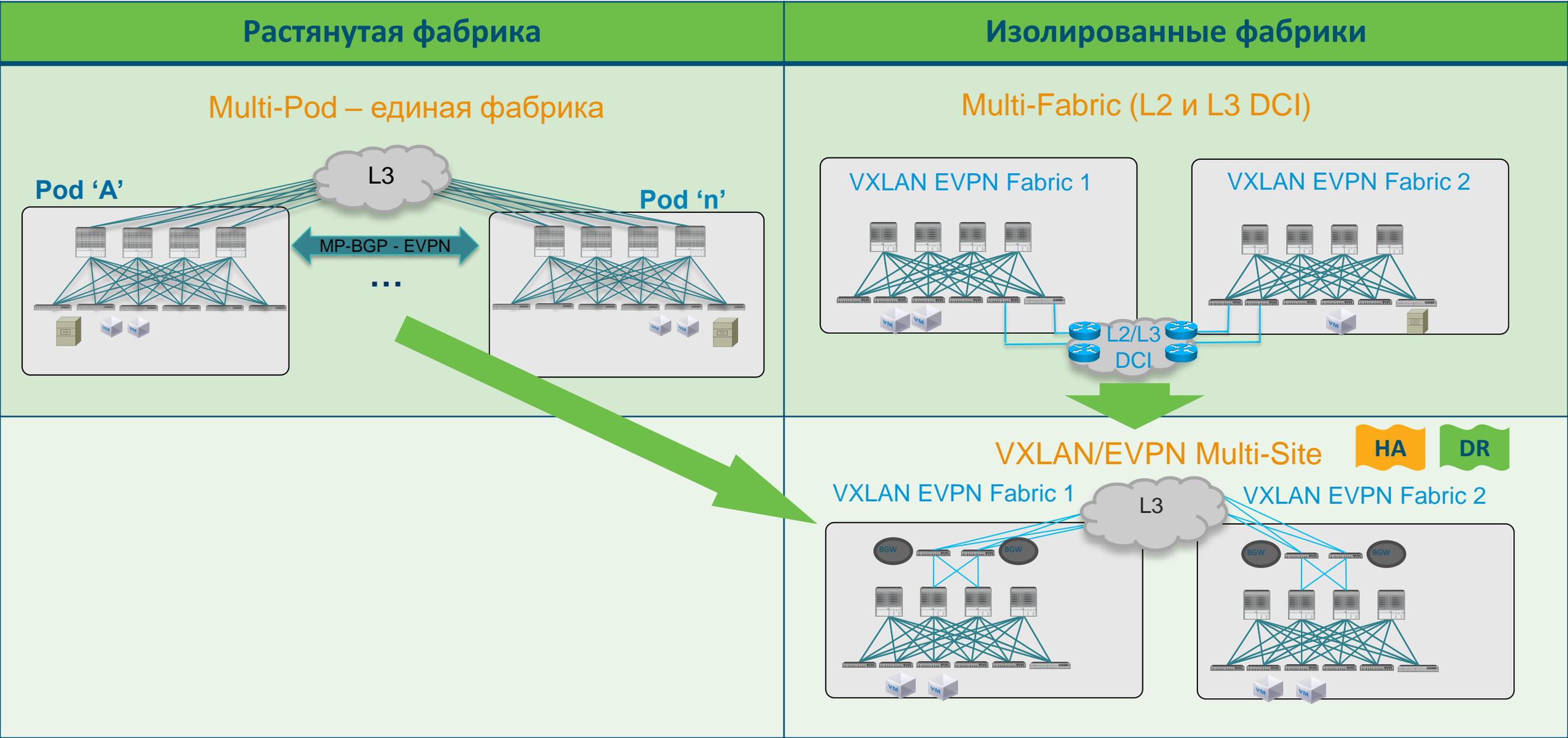
До 64 сателлитных ЦОД (пар Remote Leaf)

Продолжение EPG, BD, VRF, тенантов, контрактов

Диагностика и статистика

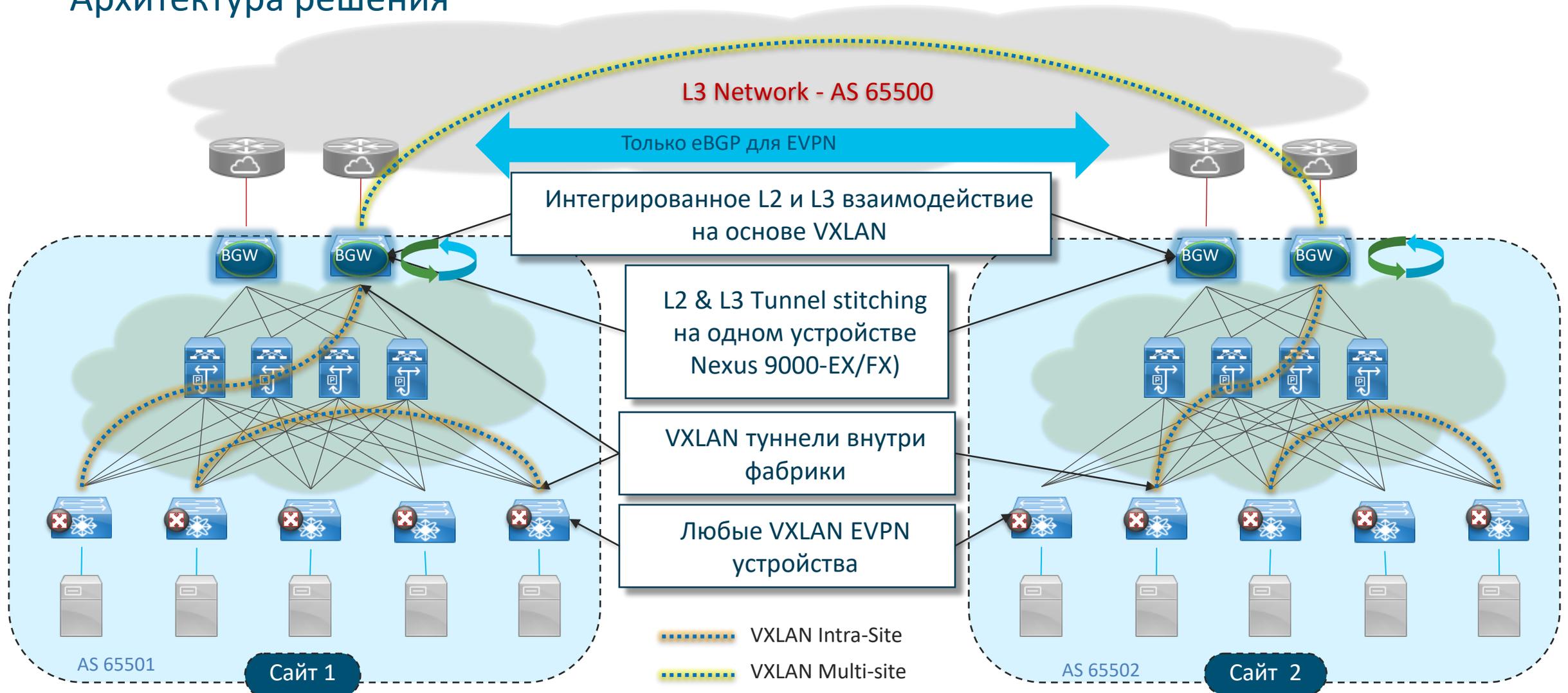
Связь сетевых фабрик ЦОД на основе VXLAN/EVPN

VXLAN/EVPN Multi-Site – рекомендованный подход

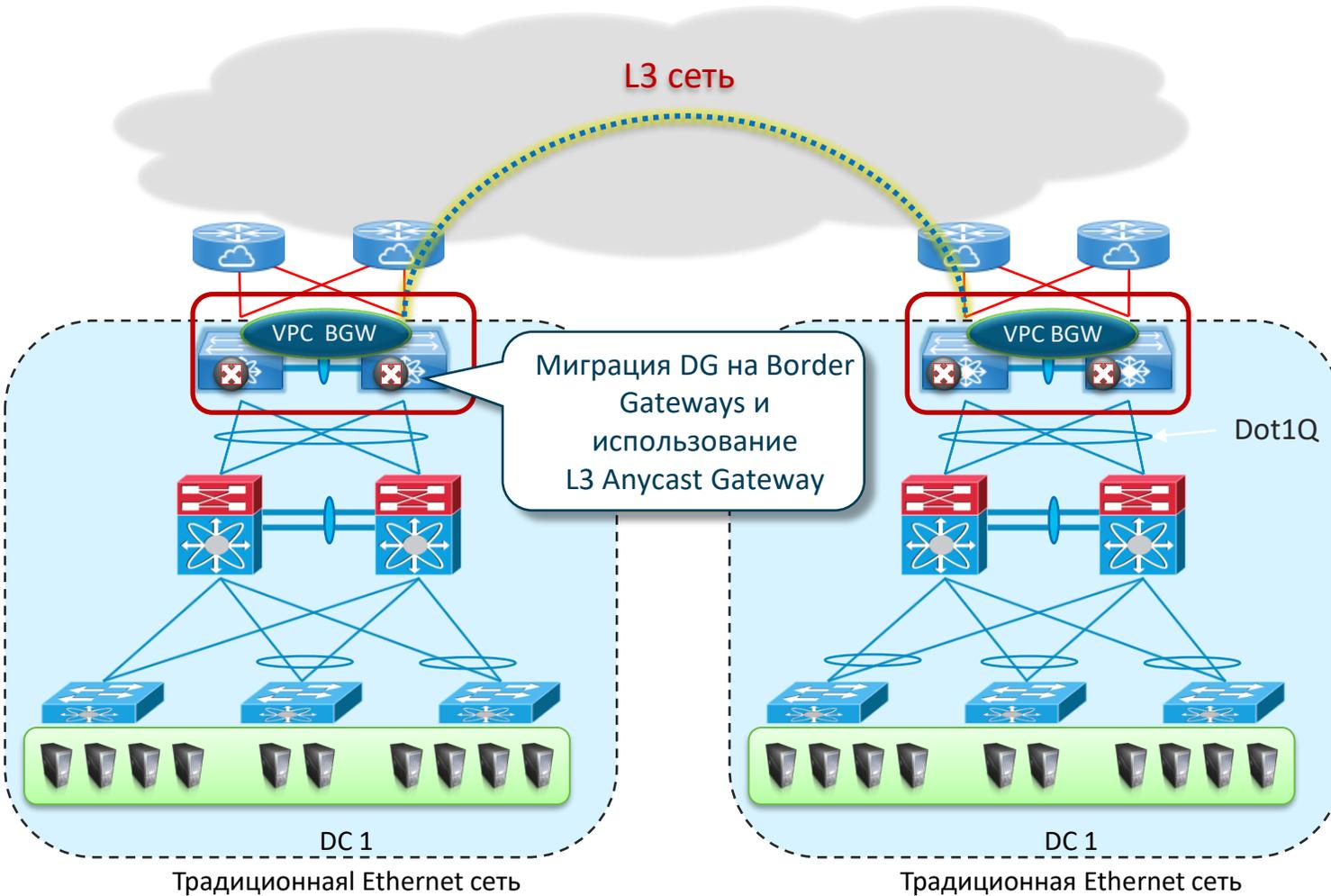


VXLAN EVPN Multi-Site

Архитектура решения



VXLAN EVPN Multi-Site для связи традиционных ЦОД

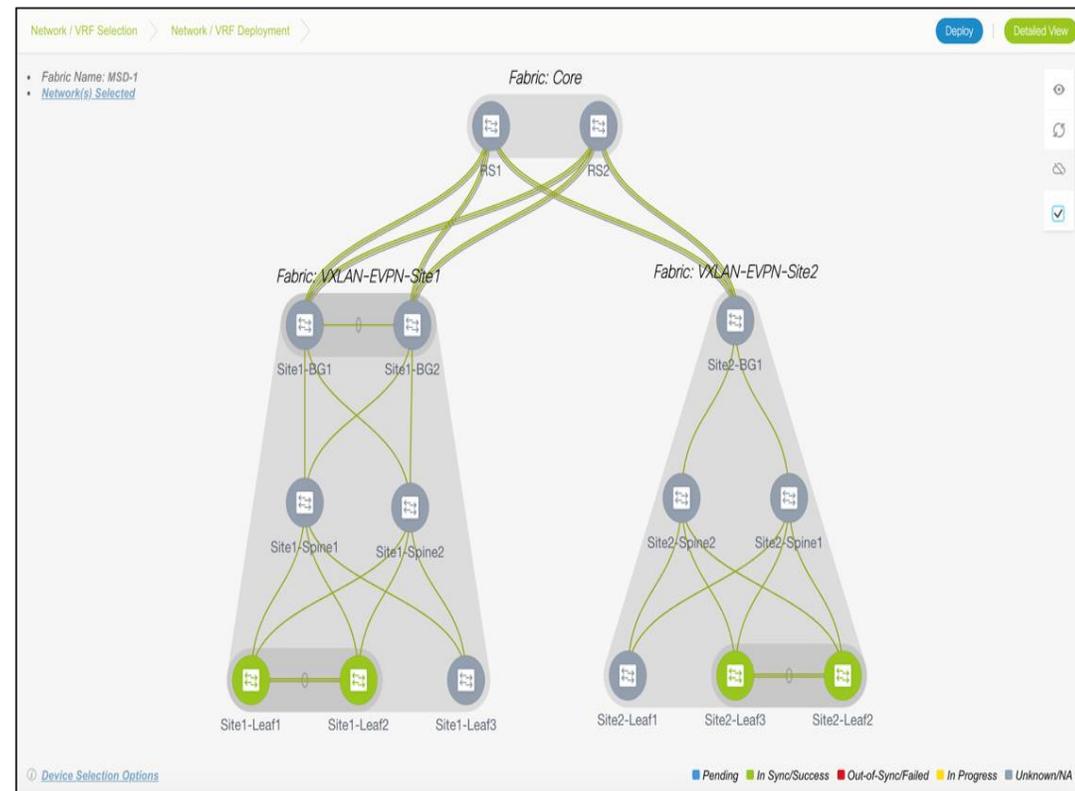
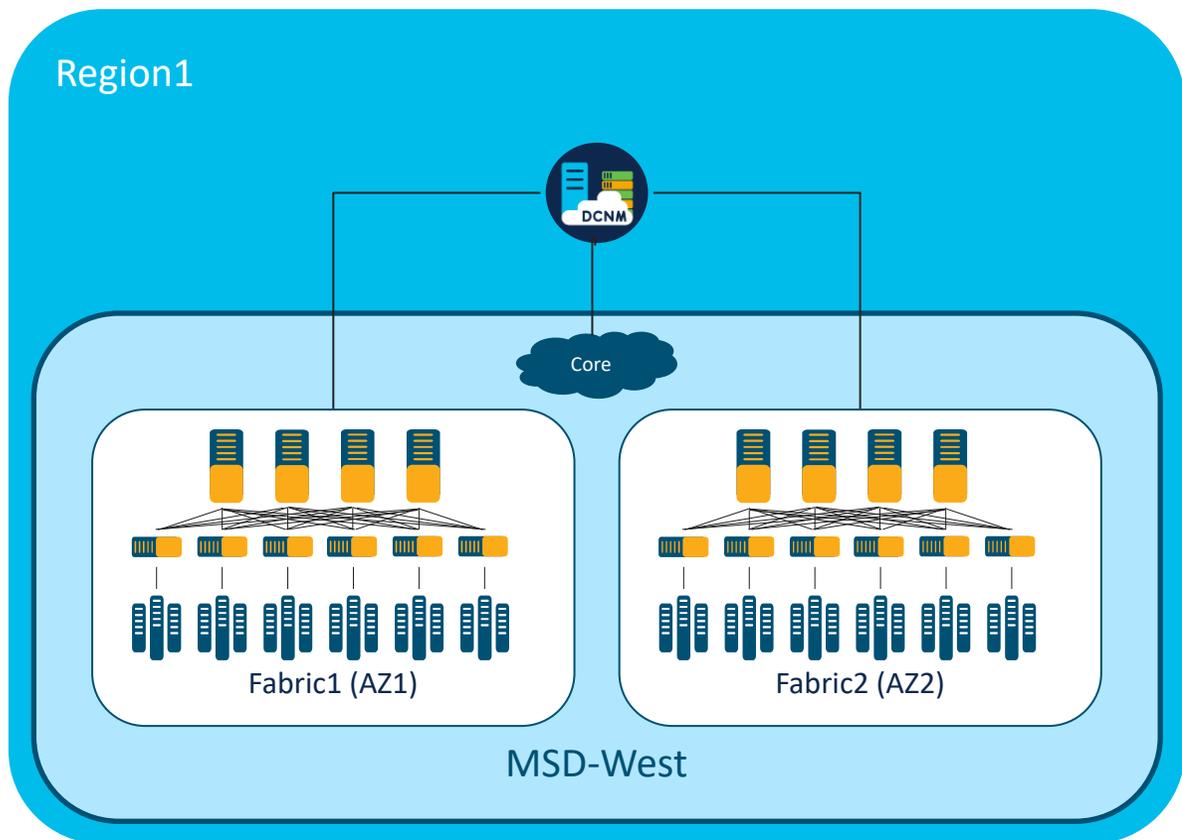


VXLAN EVPN VPC BGW

- Пара VPC BGW устройств на каждом сайте
- Миграция с классических архитектур на архитектуру фабрик
- Поддержка функций Anycast Gateway
- Контроль за BUM-трафиком (блокирование или ограничение)
- Control Plane with Selective Advertisement (L2 & L3)

Модель управления в VXLAN EVPN Multi-Site

Единый DCNM для управления несколькими фабриками

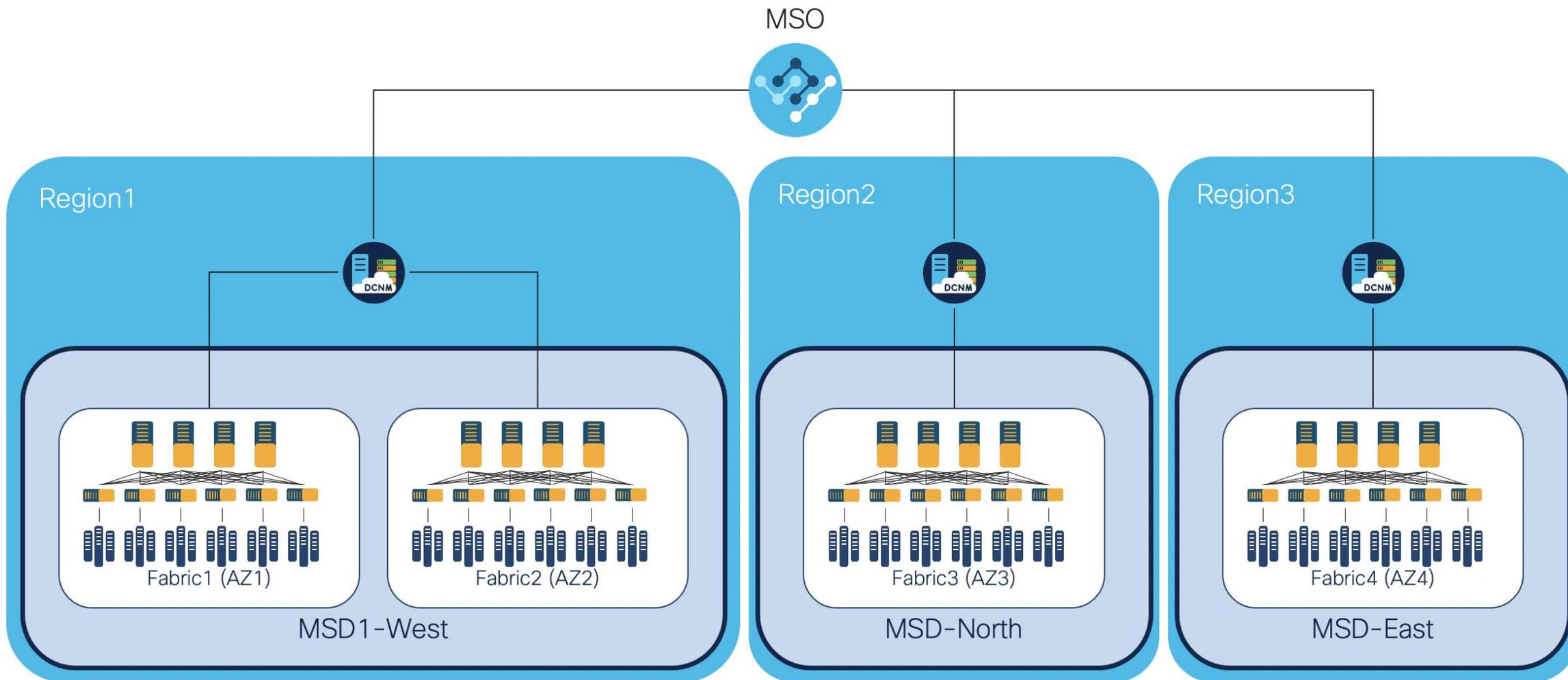


Аналог модели управления ACI Multi-Pod



Модель управления в VXLAN EVPN Multi-Site

MSO для управления несколькими регионами с DCNM в каждом



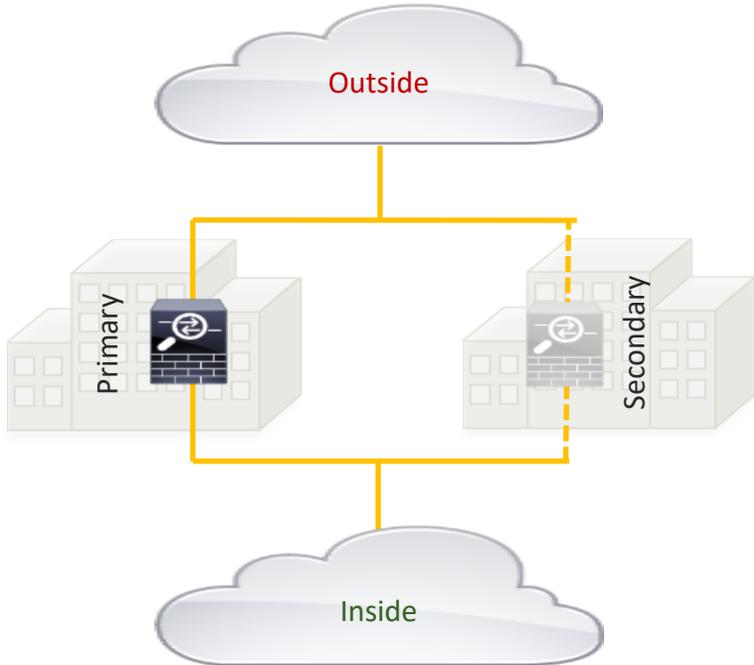
Аналог модели управления ACI Multi-Site



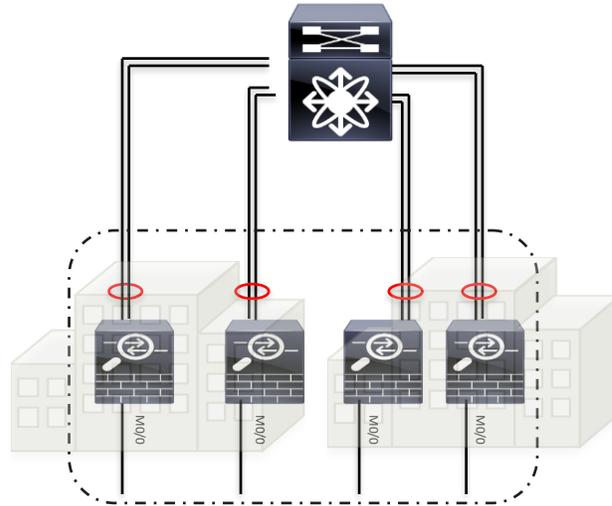
Сервисы L4/L7 в отказоустойчивых ЦОД

На примере МСЭ

Active/Standby (Failover пара)



Кластер Active/Active



Независимые МСЭ/пары

