Отказоустойчивый ЦОД за 5 дней

Александр Скороходов, Евгений Лагунцов
Илья Дрей, Максим Хаванкин, Михаил Окунев

1-5 февраля 2021

Обсуждение – в Telegram канале: https://t.me/CiscoDR

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: https://t.me/CiscoDR

Отказоустойчивый ЦОД за 5 дней

День 5: комплексные сценарии

Хаванкин Максим
mkhavank@cisco.com

5 февраля 2021

Обсуждение – в Telegram канале: https://t.me/CiscoDR

Содержание

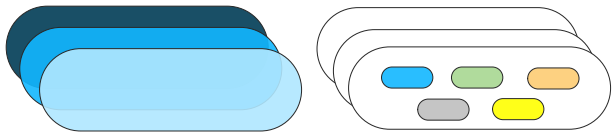
Живая демонстрация ACI Multi-Site и Hyperflex

- Архитектура современного ЦОД
- Сценарий № 1 - Распределенный ЦОД (High Availability - HA)
- Сценарий № 2 - Катастрофоустойчивый ЦОД (Disaster Recovery - DR)
- Сценарий № 3 - Гибридное облако (собственный ЦОД+публичное облако)

Архитектура современного ЦОД

Анализ архитектуры защищаемых от сбоев приложений

ЦОД-1



Классические монолитные приложения

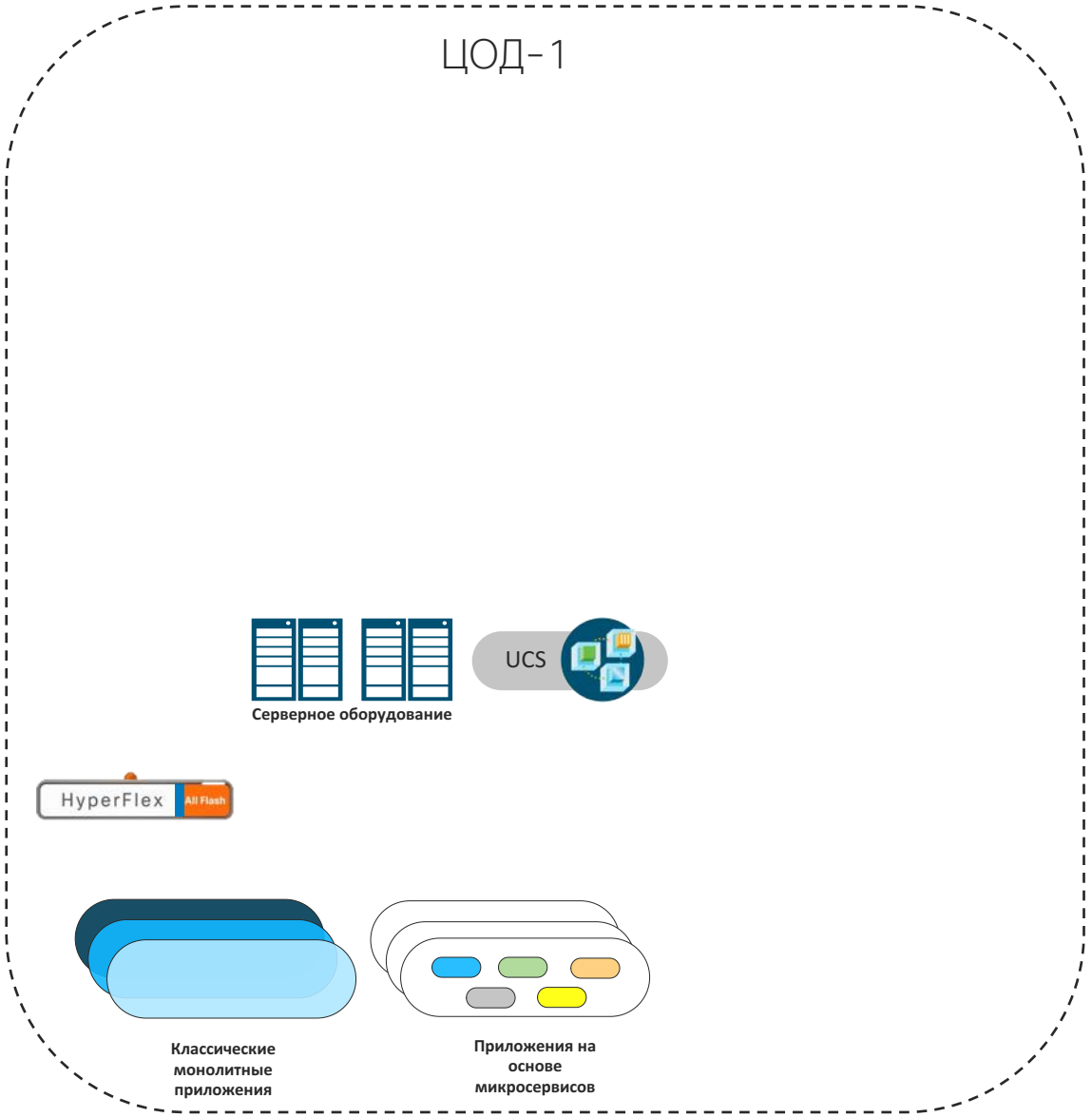
Приложения на основе микросервисов

Возможности по защите от сбоев встроенные в приложения?

Использование виртуальных машин?

Использование микросервисов на базе контейнеров?

Вычислительные платформы и гиперконвергенция

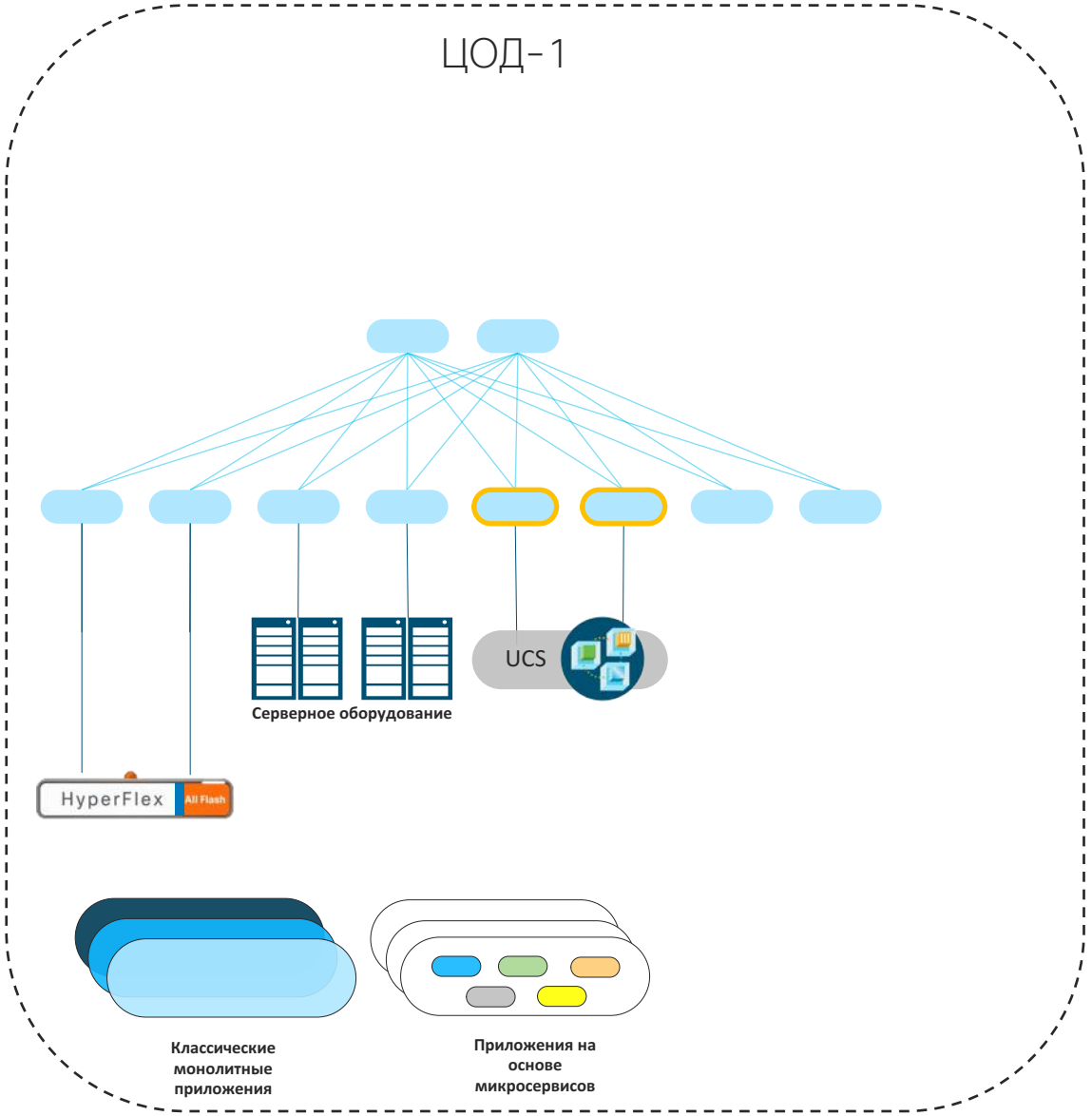


Какие вычислительные платформы используются?

Планируется ли для защиты приложений использовать возможности гиперконвергенции?

Какое количество vCenter планируется к использованию?

Сетевая фабрика

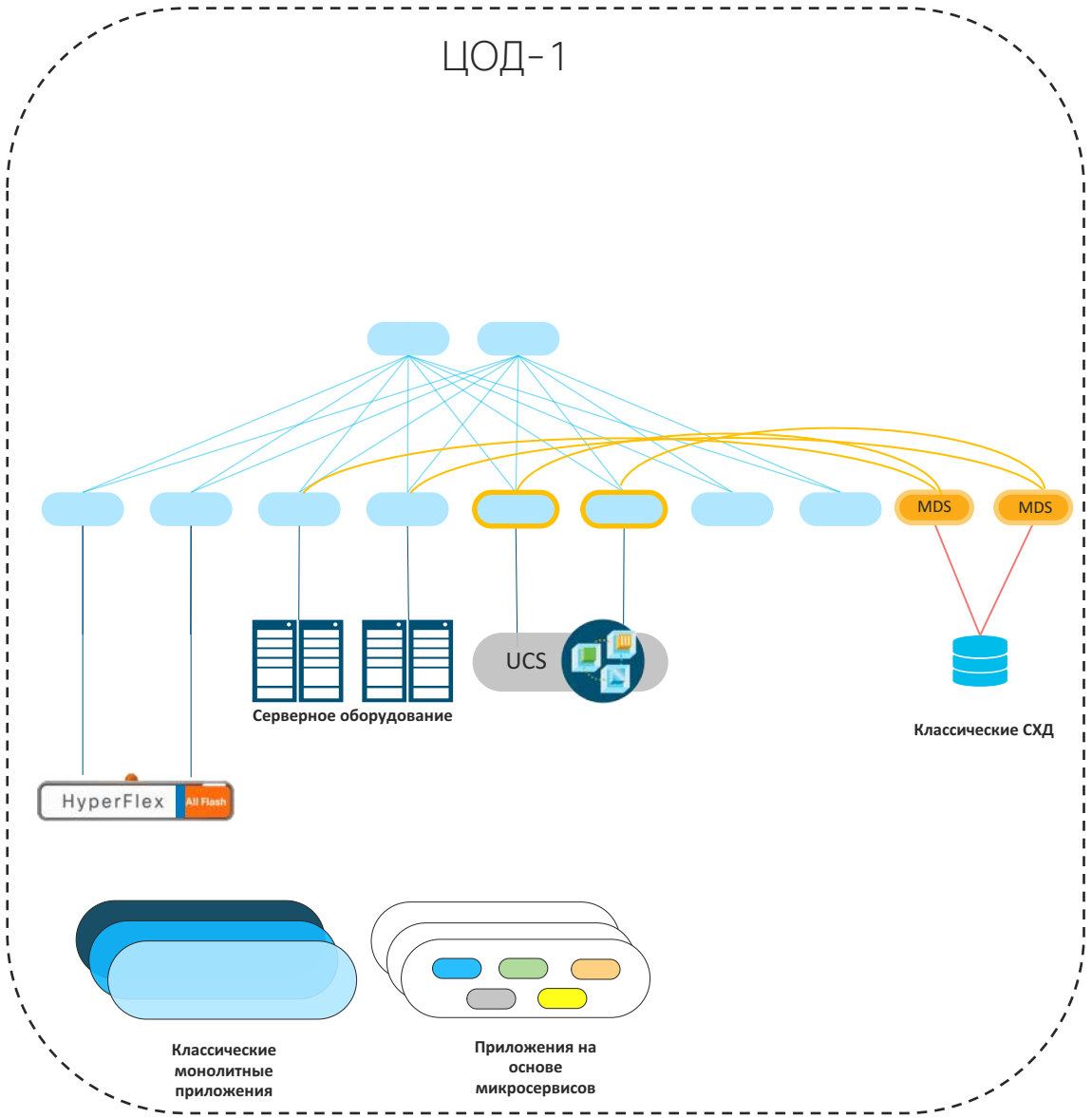


Какая архитектура используется на сети ЦОД?

Что требуется приложениям в части «растягивания» L2/L3?

Какие возможности для растягивания L3/L2 доступны?

Сеть хранения данных

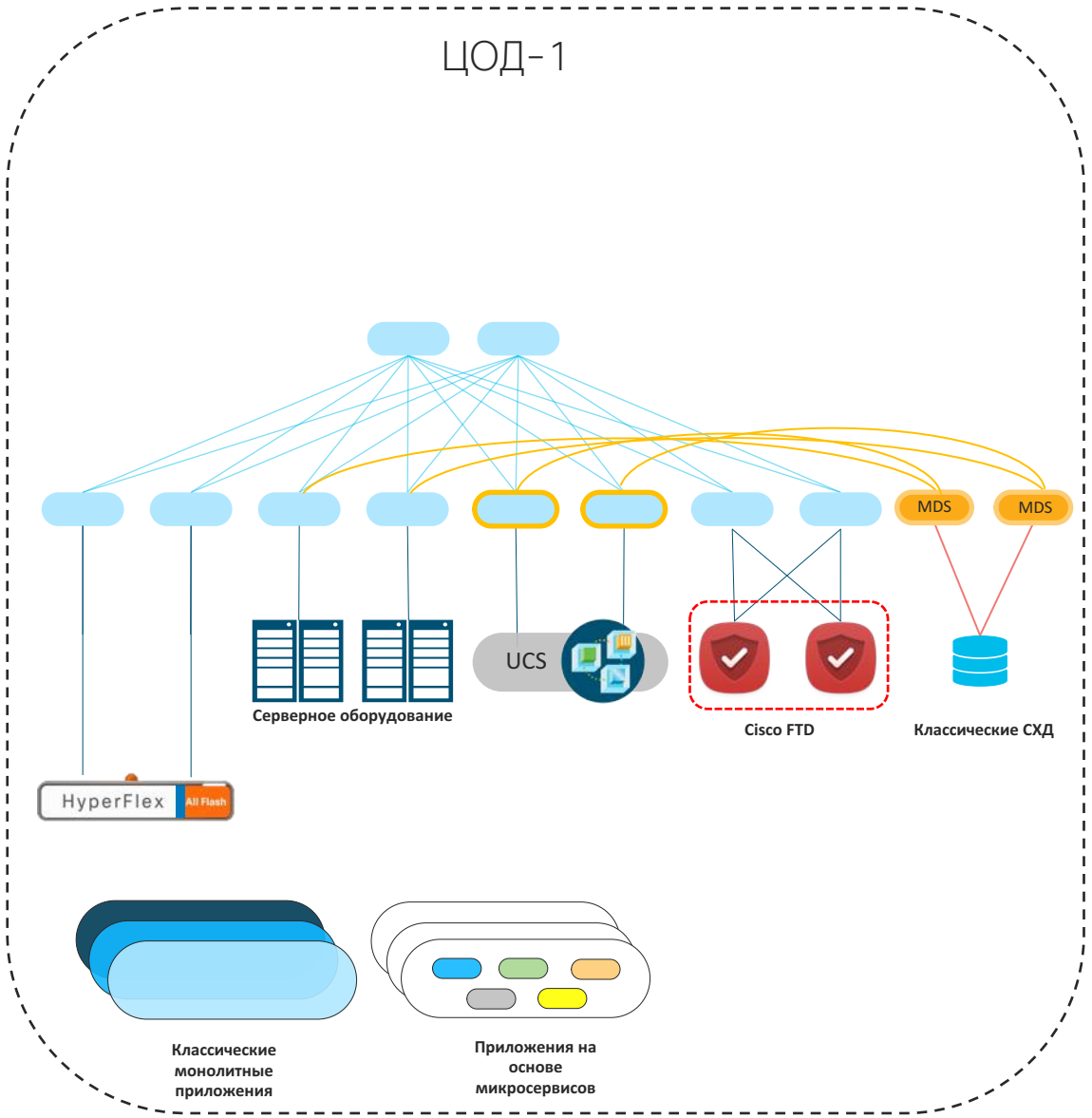


Используются ли классические СХД в существующем ЦОД?

Будет ли целевой дизайн HA/DR опираться на возможности СХД по репликации данных?

Какие возможности существующее SAN оборудование поддерживает в части DCI?

Межсетевые экраны на границе и внутри сети ЦОД (North-South + East West)



Используются ли в текущем дизайне МСЭ в ЦОД?

Как происходит встраивание МСЭ в путь передачи данных?

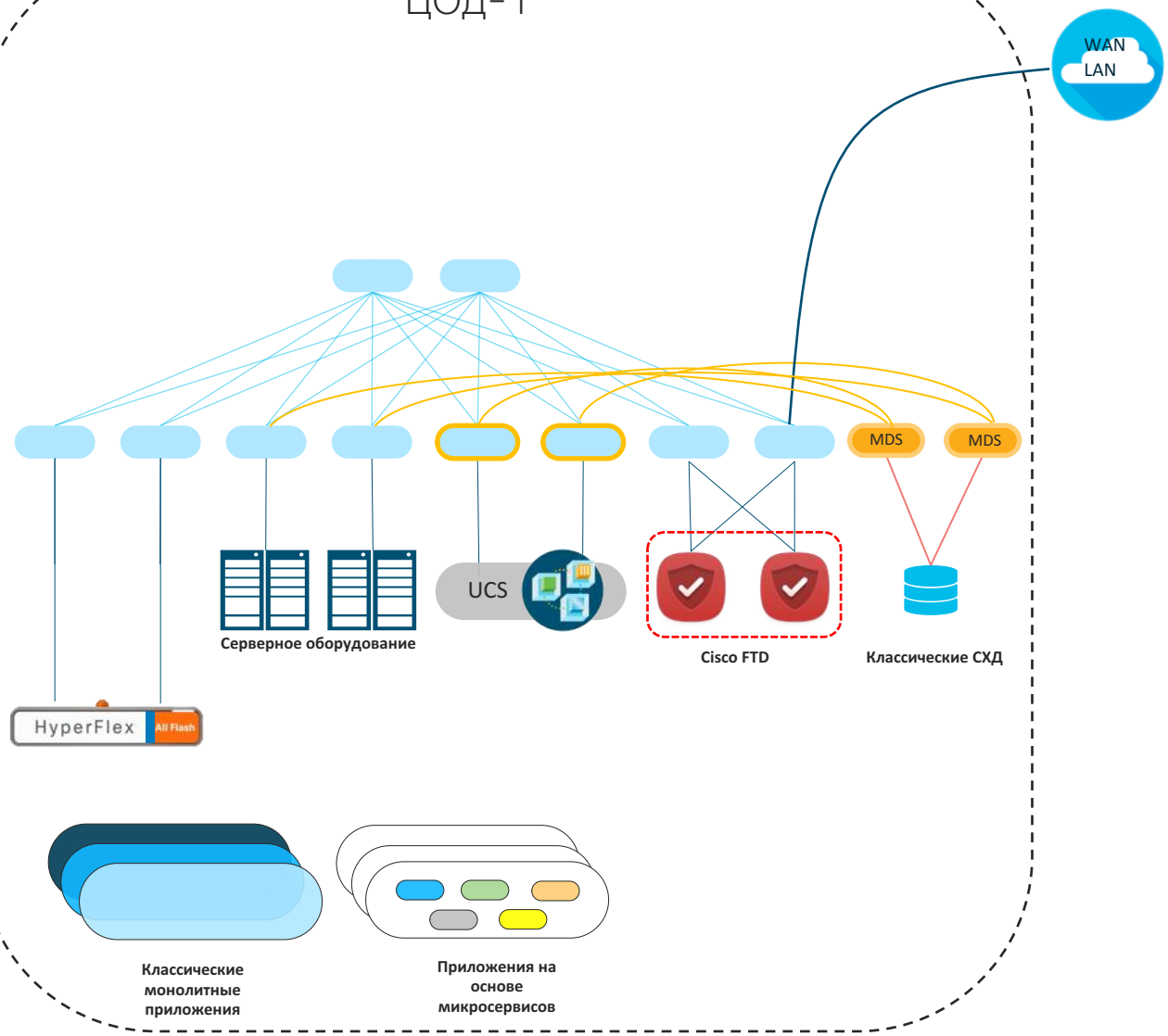
Какие возможности по упрощению встраивания МСЭ предлагает сетевая фабрика?

Сценарий № 1 - Распределенный ЦОД (High Availability - HA)

Стратегия резервирования приложений

HA

ЦОД-1

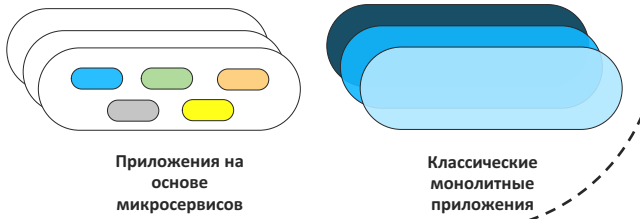


ЦОД-2

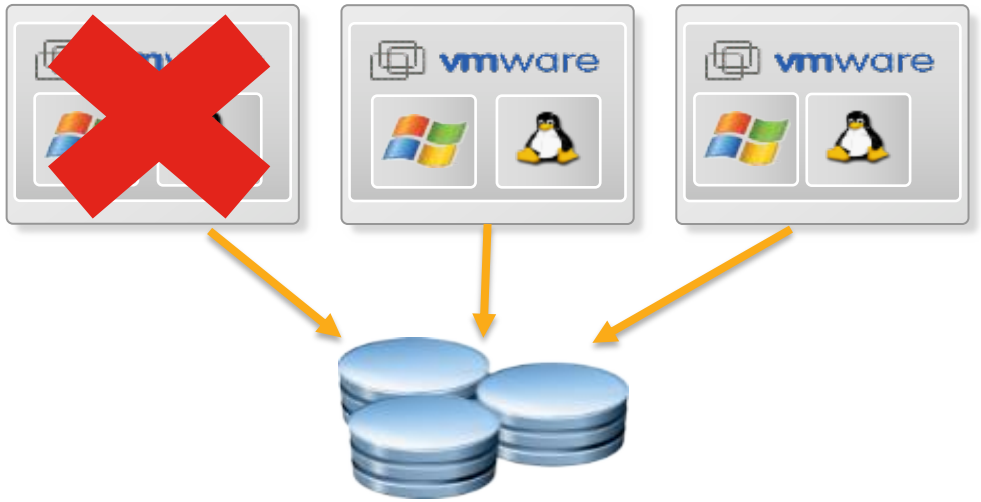
Опора на возможности системы виртуализации и управления контейнерами и микросервисами

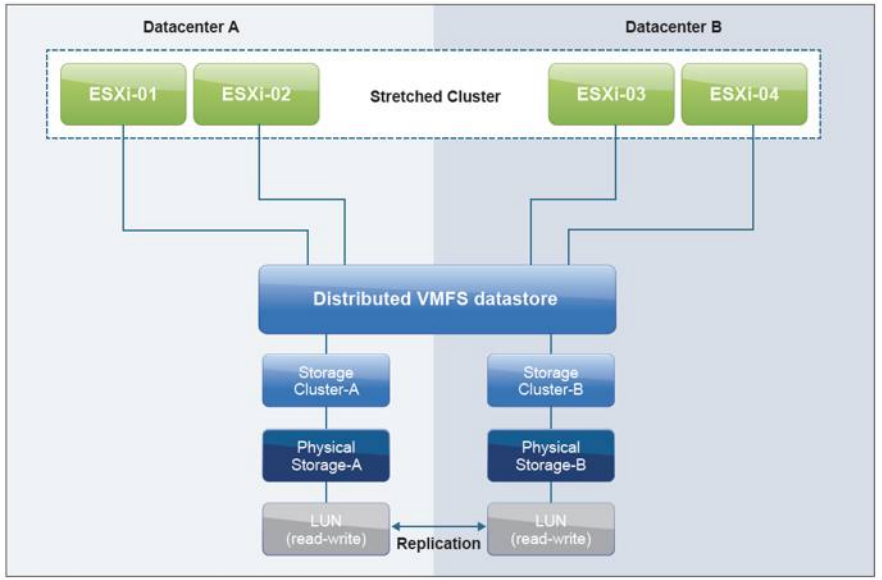
Бизнес-критичные приложения могут иметь собственные средства репликации и восстановления

Решения по оркестрации поведения системы в случае сбоя, как правило, не используются



Защита на уровне виртуализации
HA сценарии на примере VMware vSphere и HA

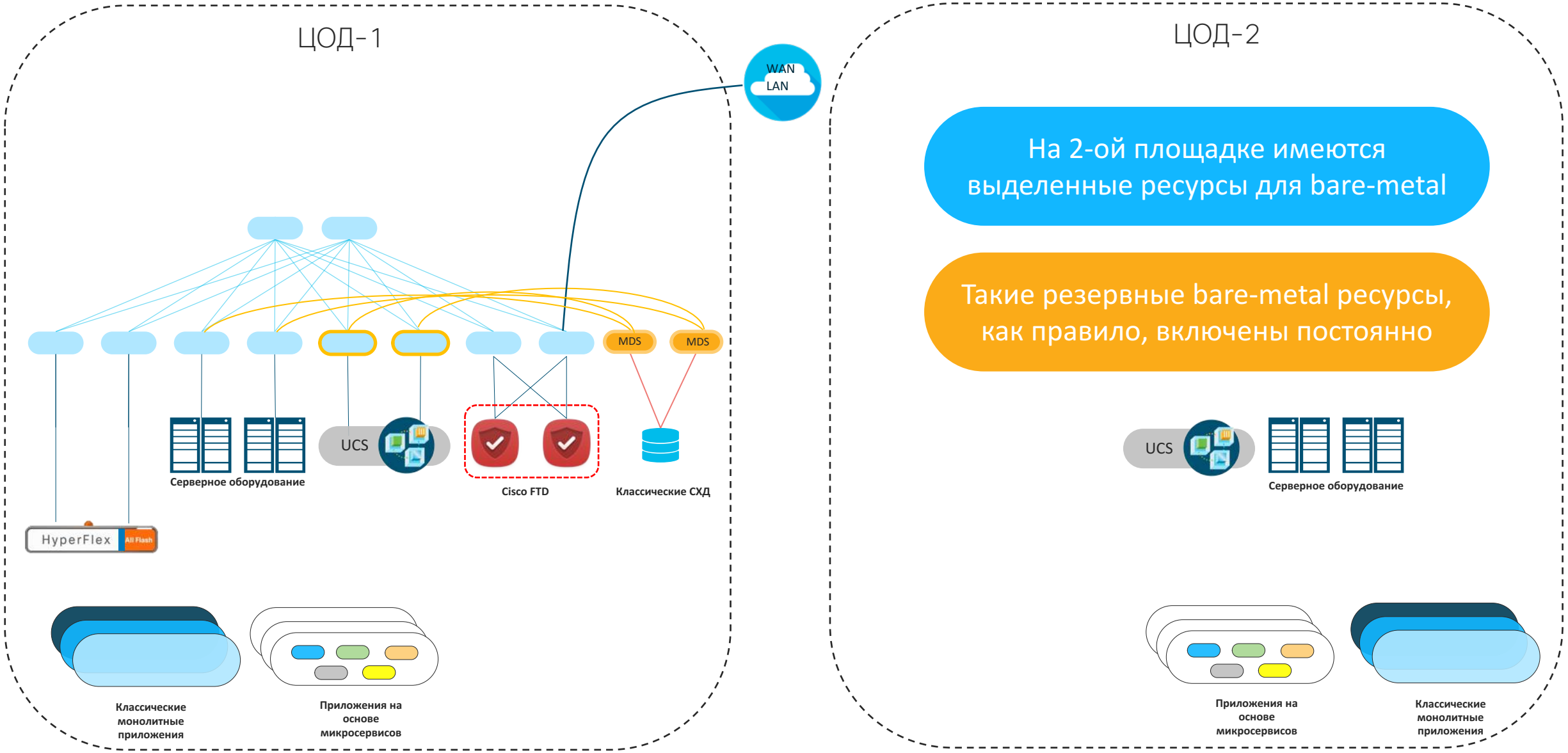




- Автоматический перезапуск VM в пределах кластера при отказе или недоступности хоста
- Необходимо разделяемое хранилище
- Возможна реализация HA между ЦОД, при «растягивании» хранилища за счет синхронной репликации СХД или гиперконвергентными системами

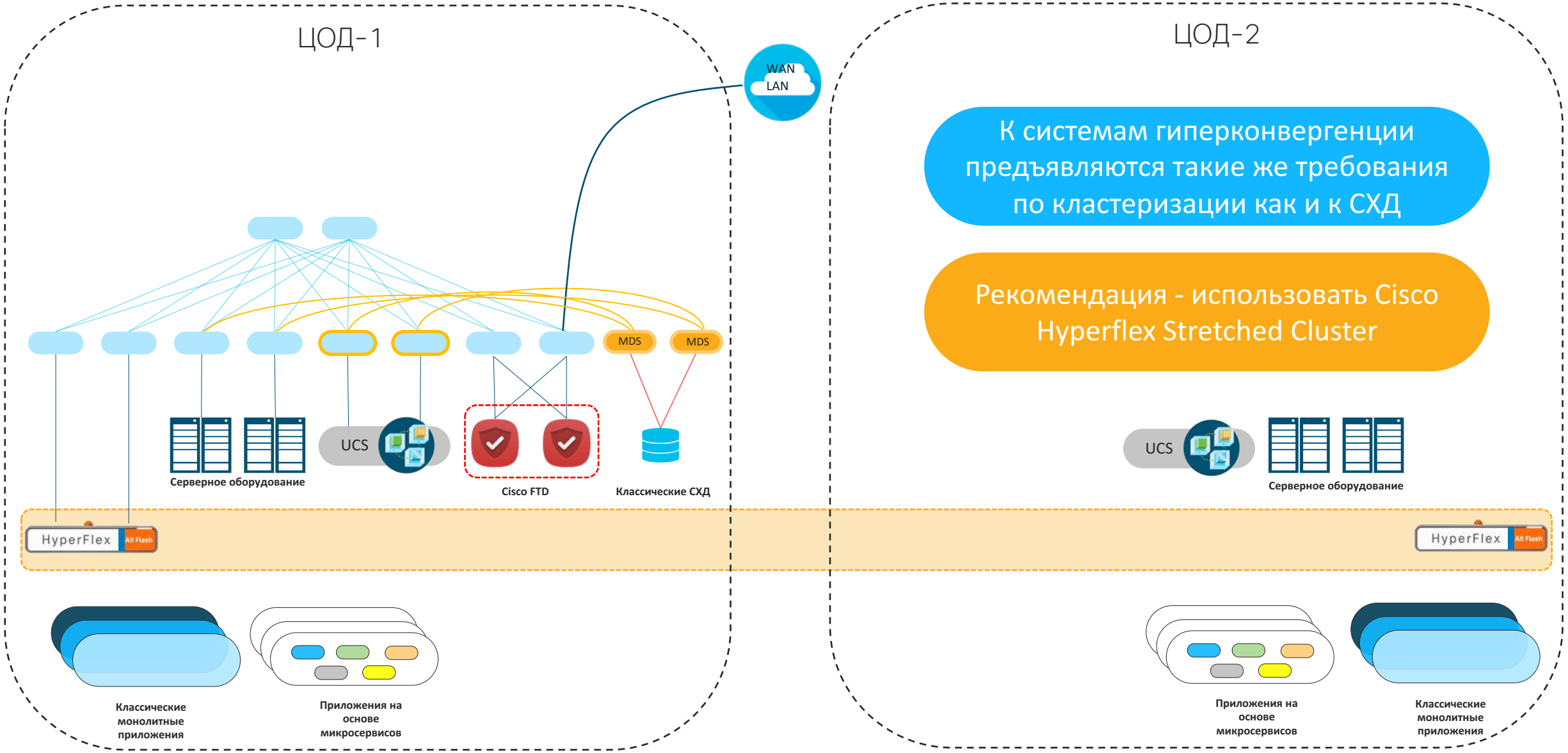
Стратегия резервирования для bare-metal систем

HA



Стратегия резервирования для систем гиперконвергенции

HA

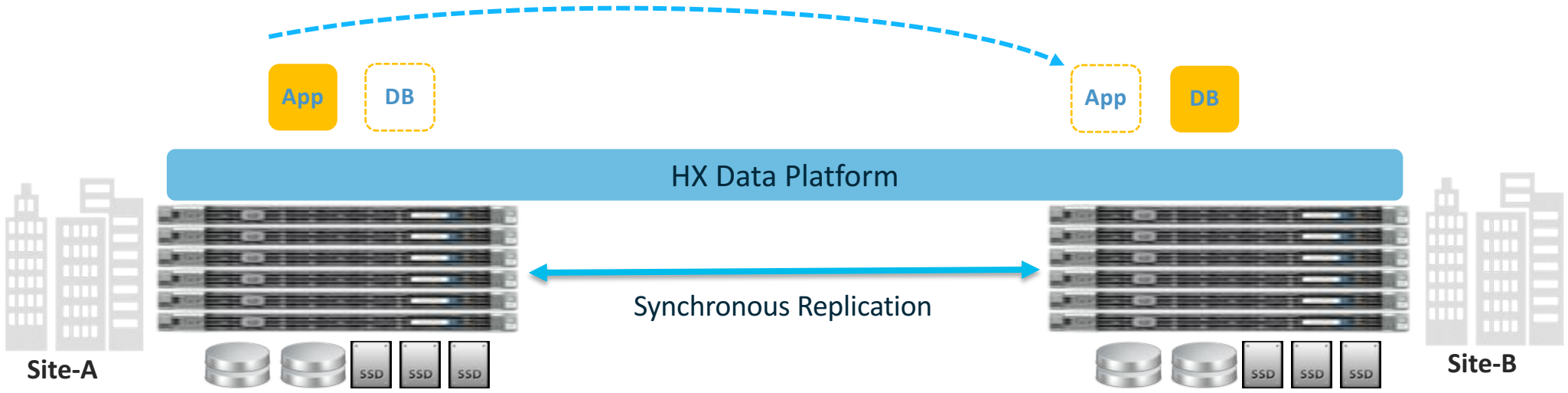


HyperFlex Stretched Cluster– «растянутый кластер»

HA

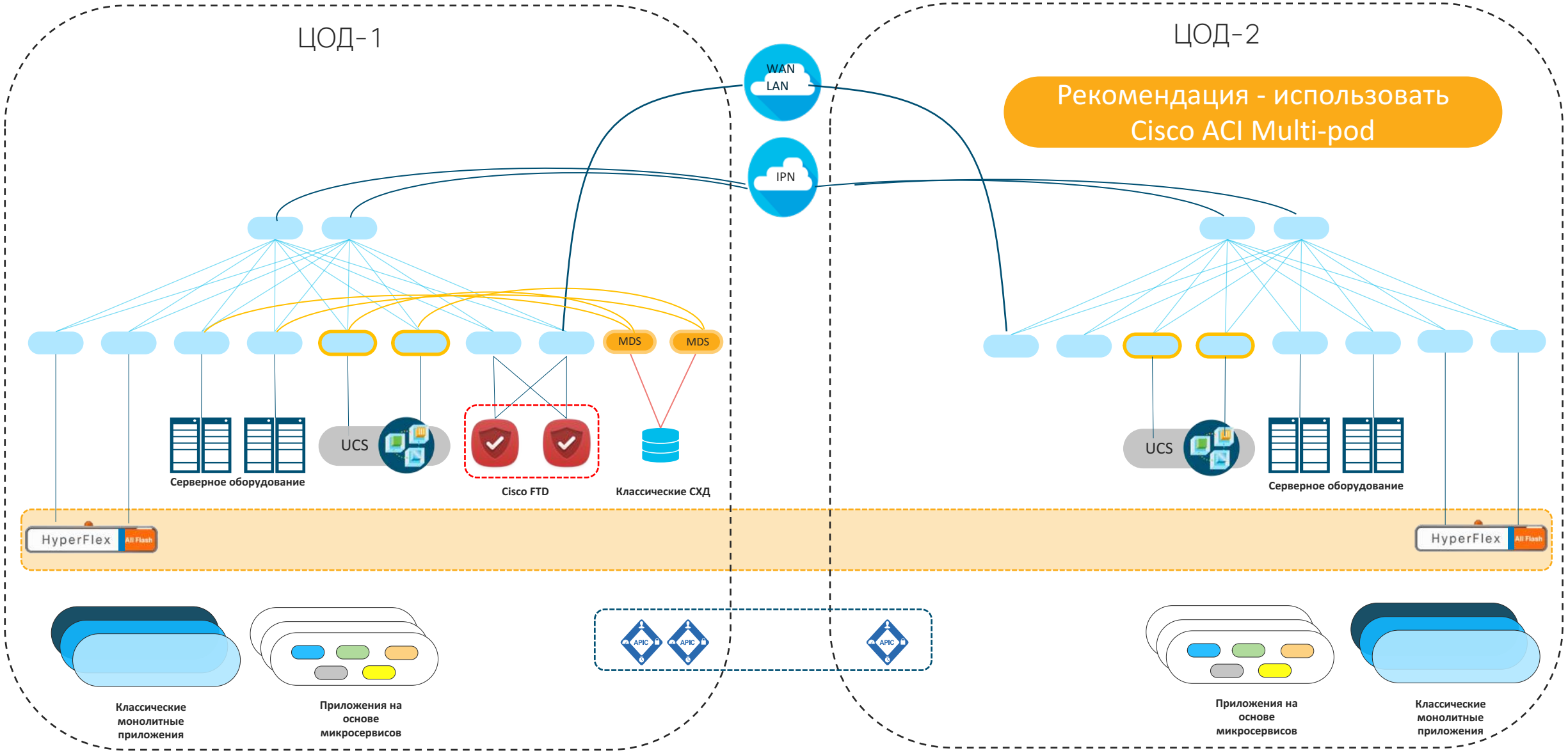
Распределенная между двумя площадками система Active-Active для поддержки бизнес-критичных задач, требующих высокую доступность (RTO = время рестарта VM) и отсутствие потери данных («нулевой» RPO)

- 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



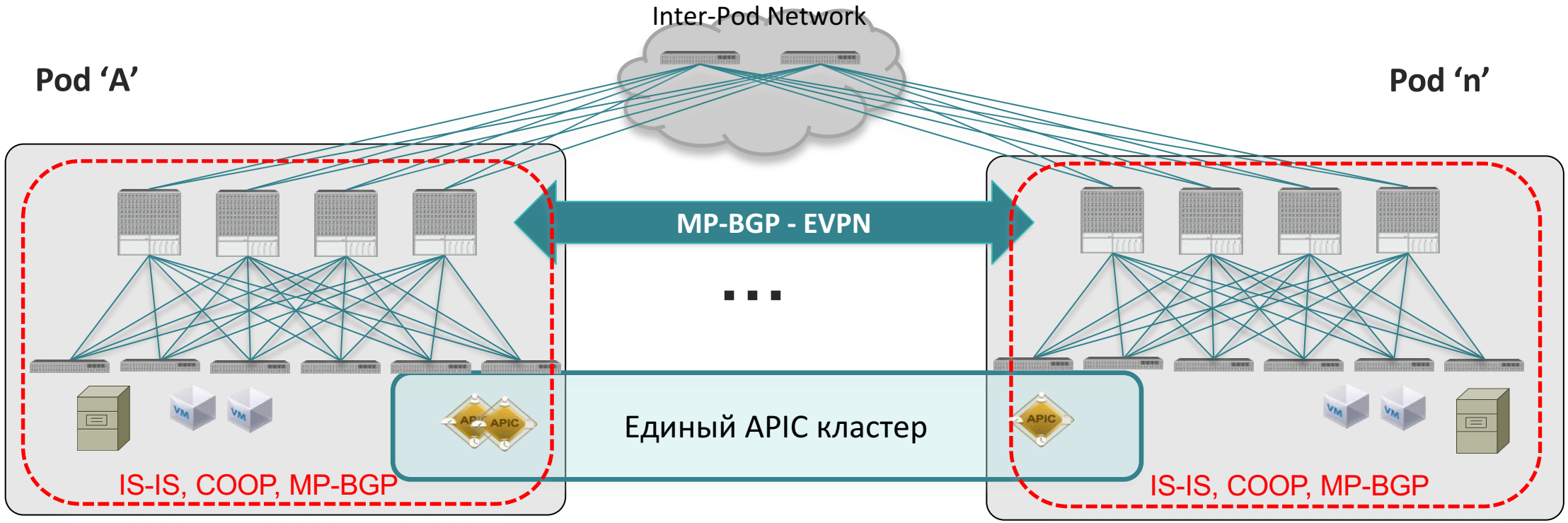
Сеть передачи данных для нескольких ЦОД

HA



ACI Multi-Pod

HA



- 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
- Сквозное применение политик

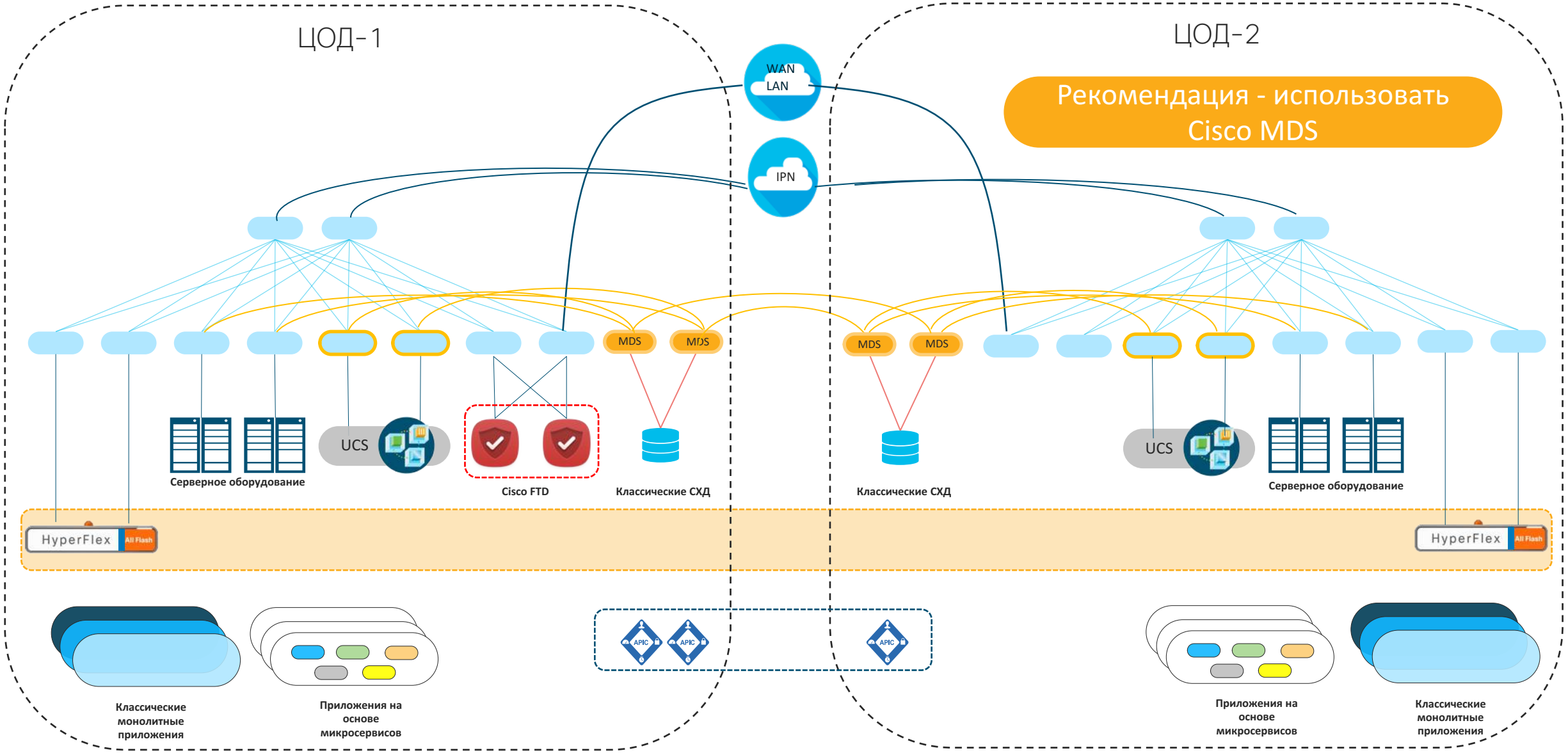
Cisco ACI Multi-Pod для HA?

HA

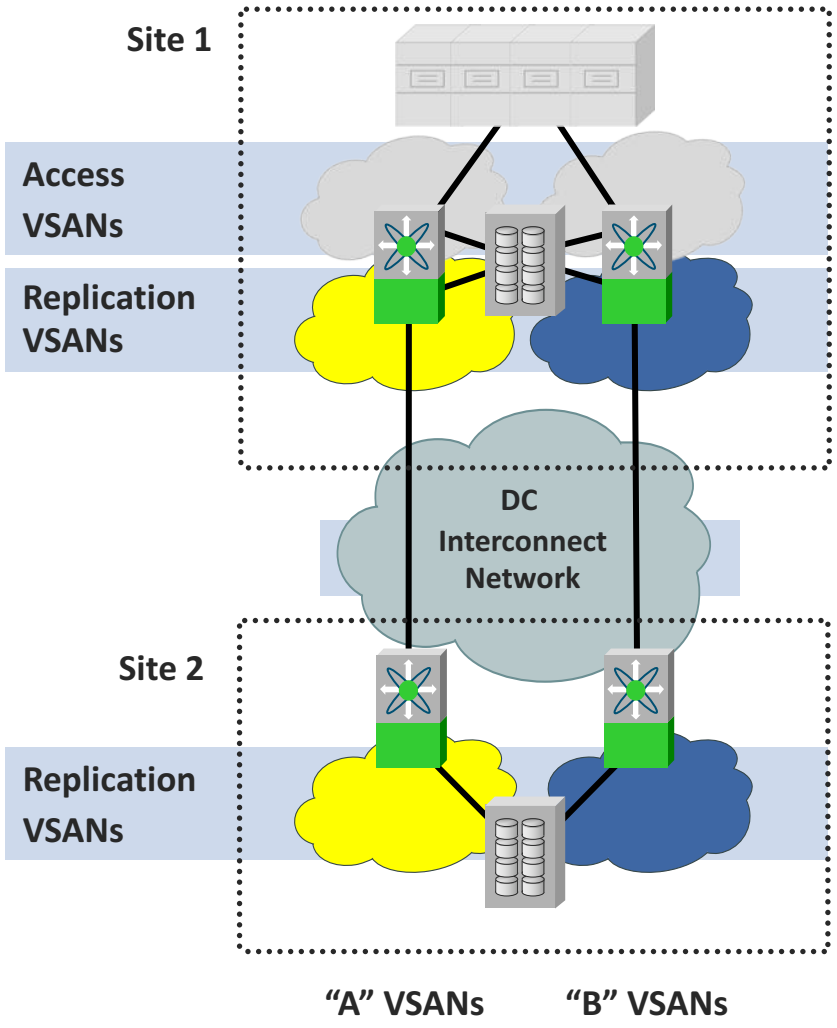
- Единый домен управления и политик для всех POD
  - Возможность использовать единый VMM домен во всех Pod (один vDS на стороне среды виртуализации, одна ACI CNI интеграция)
- Доступность самых последних функций ACI

Сеть хранения данных для нескольких ЦОД

HA



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

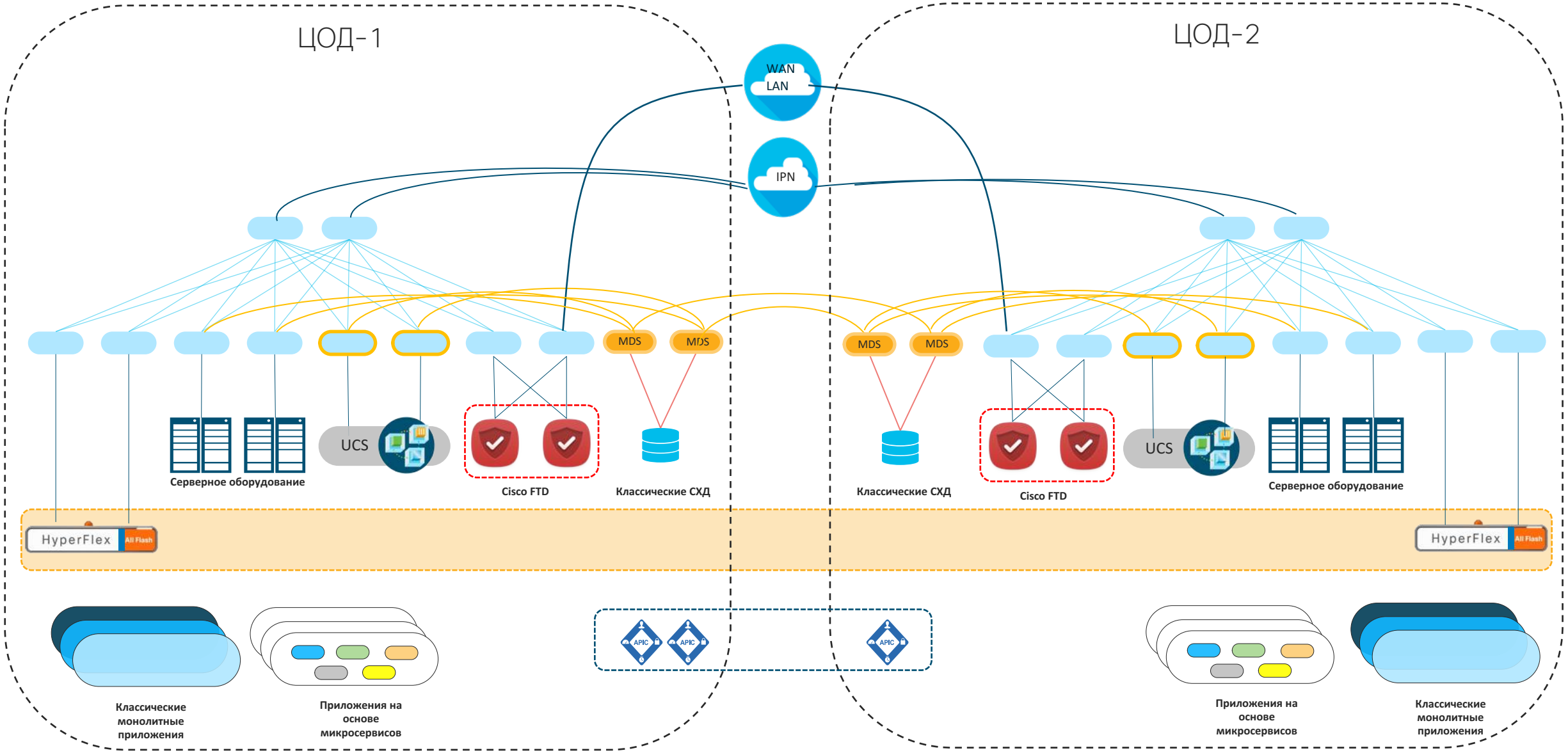
Что ограничивает расстояние?
Буферные кредиты

- 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

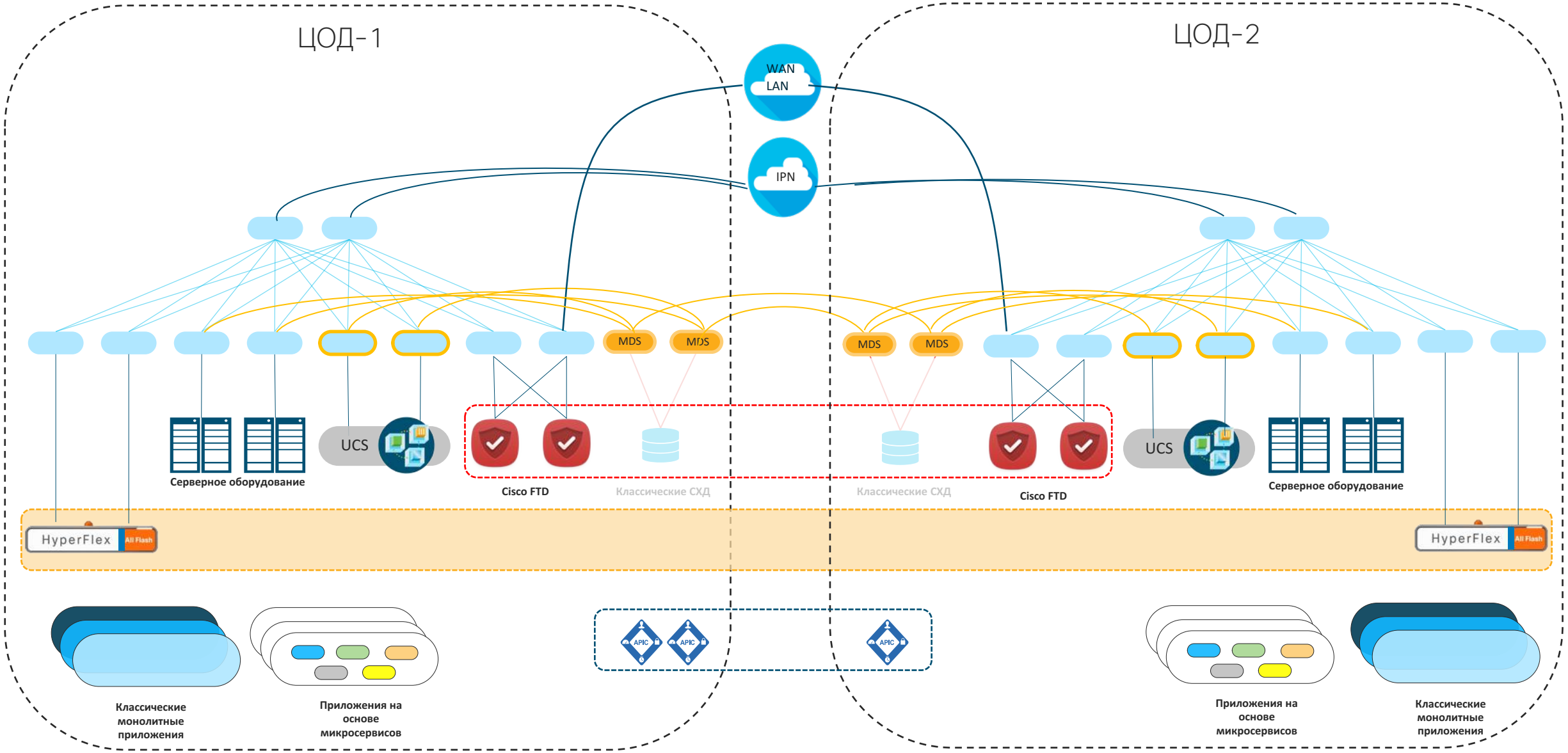
Межсетевые экраны на границе и внутри сети ЦОД (North-South + East West)

HA



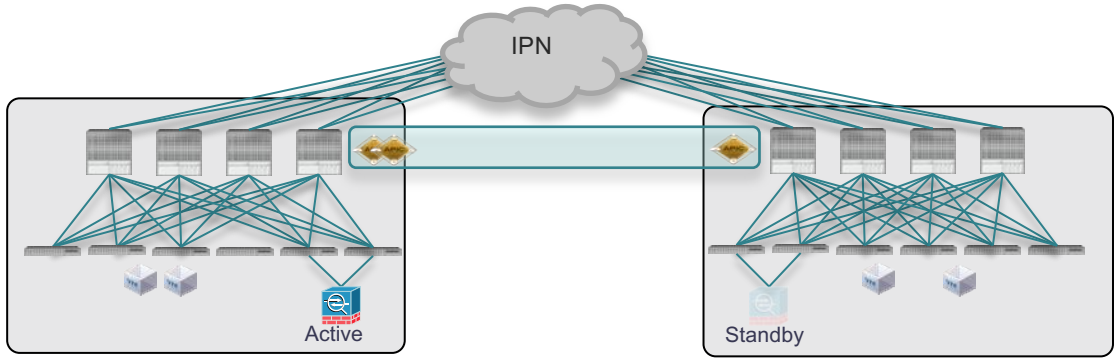
Межсетевые экраны на границе и внутри сети ЦОД (North-South + East West)

HA

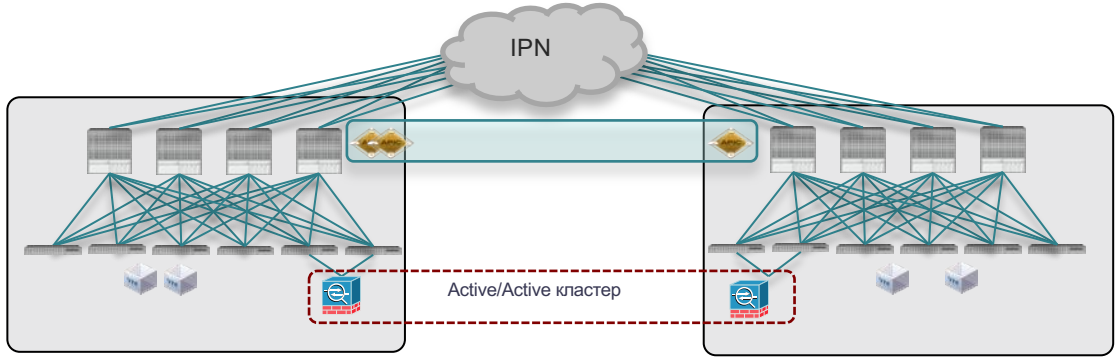


ACI Multi-Pod: Режимы интеграции МСЭ

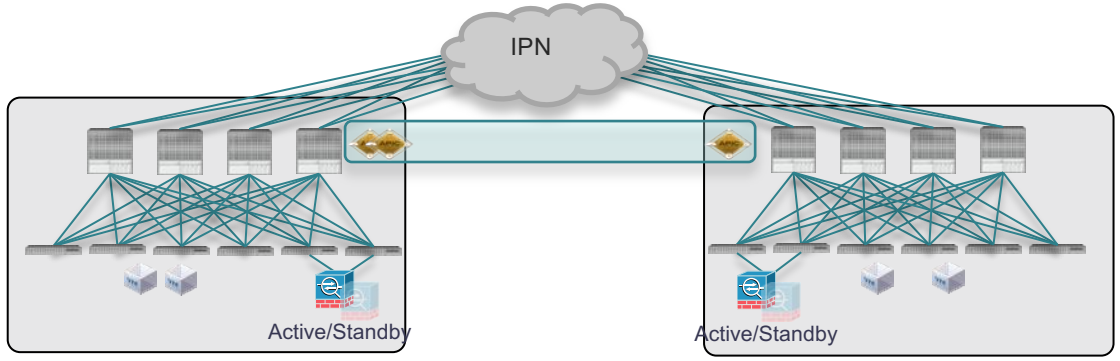
HA



- Одна Active/Standby пара «растянутая» между Pod
- Отсутствуют проблемы с ассиметричными потоками



- Active/Active кластер «растянутый» между Pod (один логический FW)
- Требует возможности подключения к фабрике одного и того же MAC/IP в нескольких POD одновременно
- Поддерживается начиная с версии ACI 3.2(4d) с обязательным использованием сервисного графа и PBR



- Независимые пары Active/Standby развернутые в каждом Pod
- Требуется использовать сервисный граф и симметричный PBR

Сервисная цепочка – лучше вместе с PBR
Организация передачи пакетов

- Для организации коммутации и маршрутизации пакетов через сервисное устройство требуется соответствующий дизайн сегментов сети, например VRF sandwich
- Функция PBR изменяет правила коммутации и маршрутизации и автоматизирует вписывание сервисного устройства в путь передачи данных

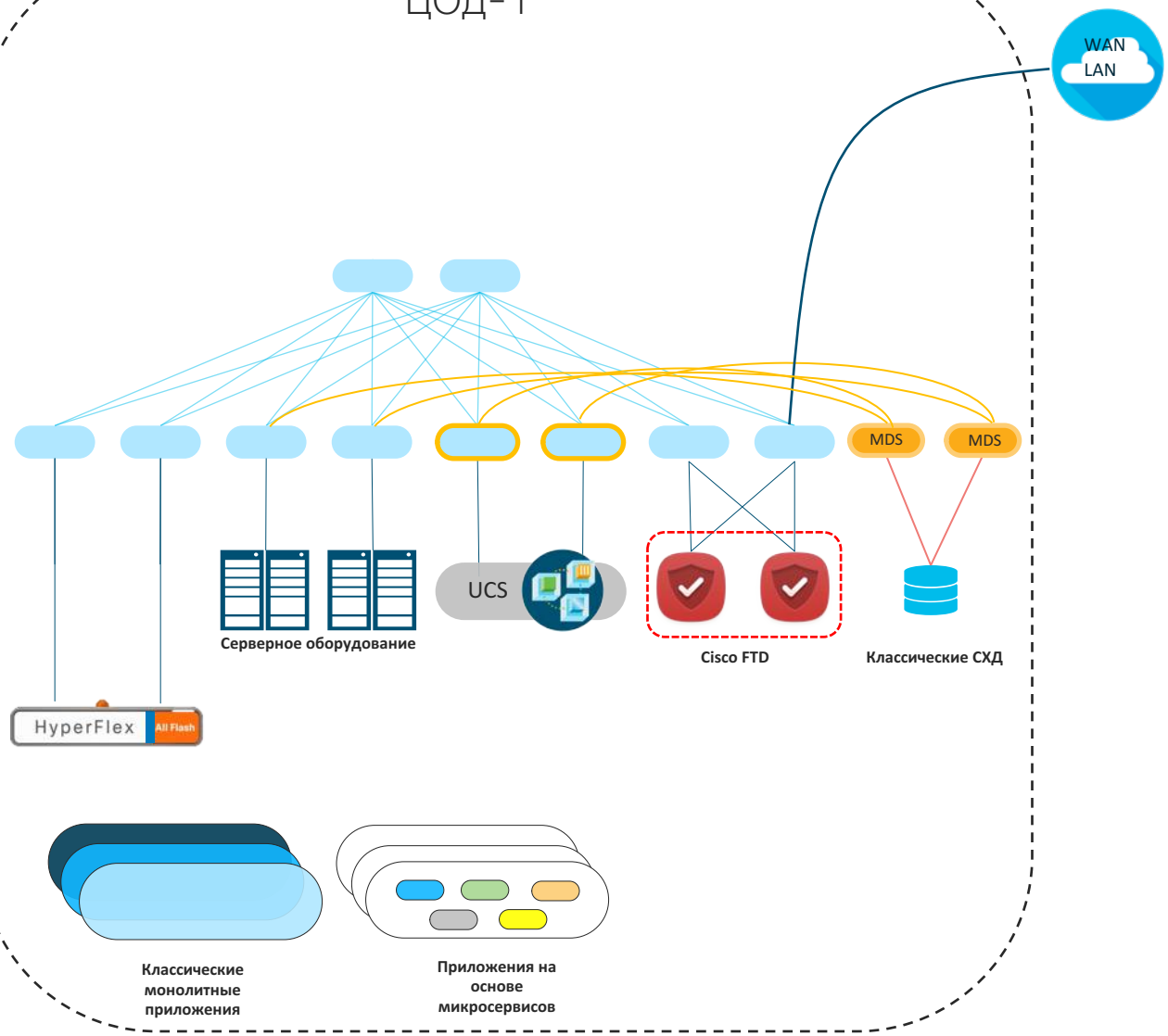
Сервисная цепочка	Сервисная цепочка с функцией PBR
Дизайн маршрутизации и коммутации вручную для вписывания L4/L7 устройства в путь передачи	Автоматизация вписывания
Дизайн с VRF Sandwich	Автоматическое вписывание в одном VRF

Сценарий № 2 - Катастрофоустойчивый ЦОД (Disaster Recovery - DR)

Стратегия резервирования приложений

DR

ЦОД-1

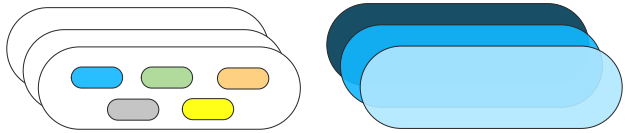


ЦОД-2

Опора на возможности систем хранения данных, гиперконвергенции и резервного копирования

Бизнес-критичные приложения могут иметь собственные средства репликации и восстановления

Минимальная стоимость и простота



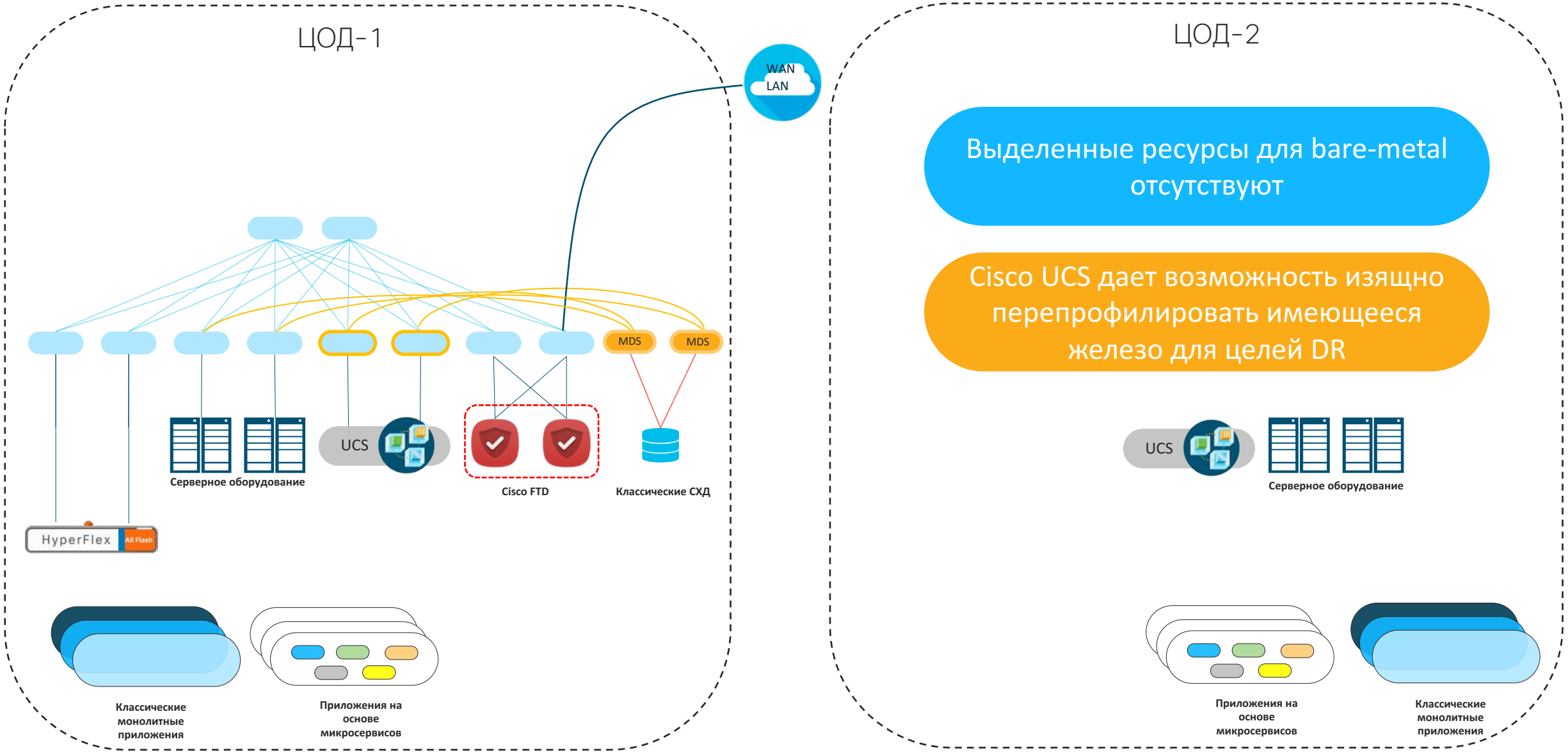
DR

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

Стратегия резервирования для bare-metal систем между ЦОД

DR



DR

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

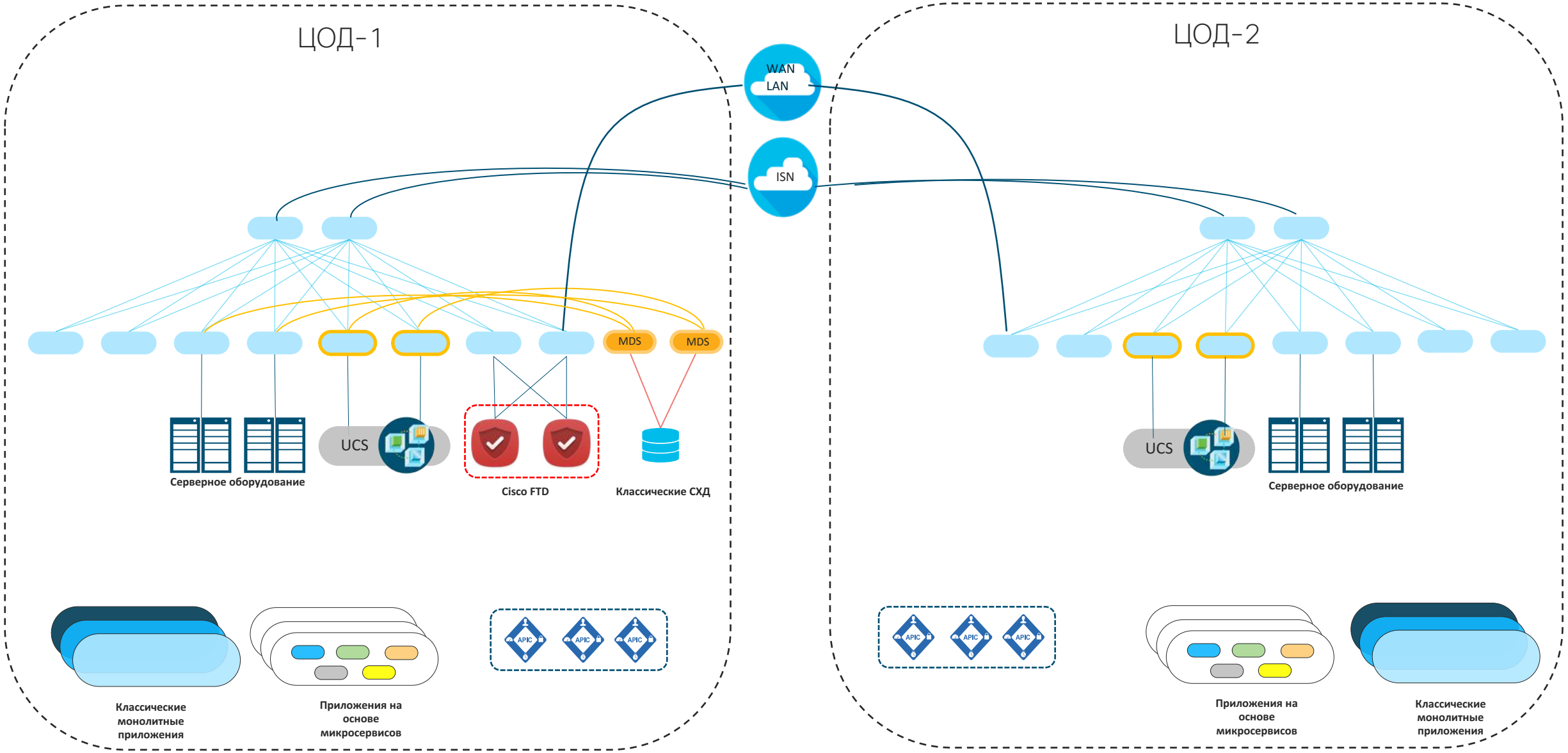
DR

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

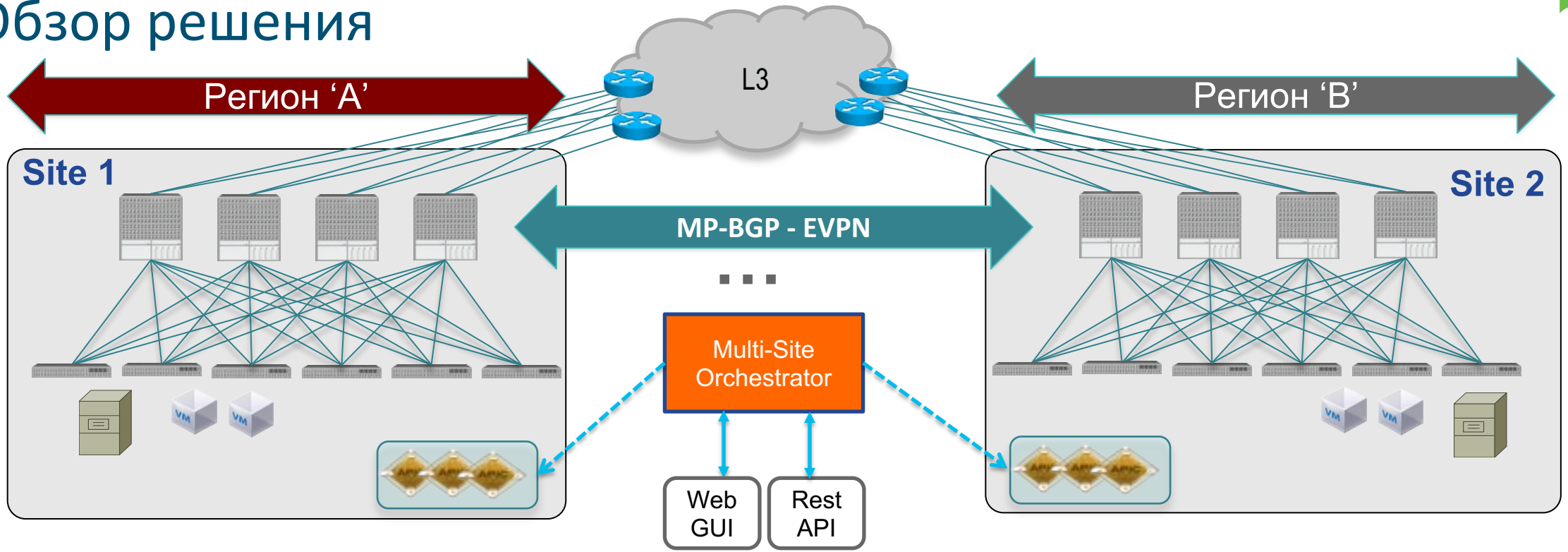
Сеть передачи данных для нескольких ЦОД

DR



ACI Multi-Site
Обзор решения

DR

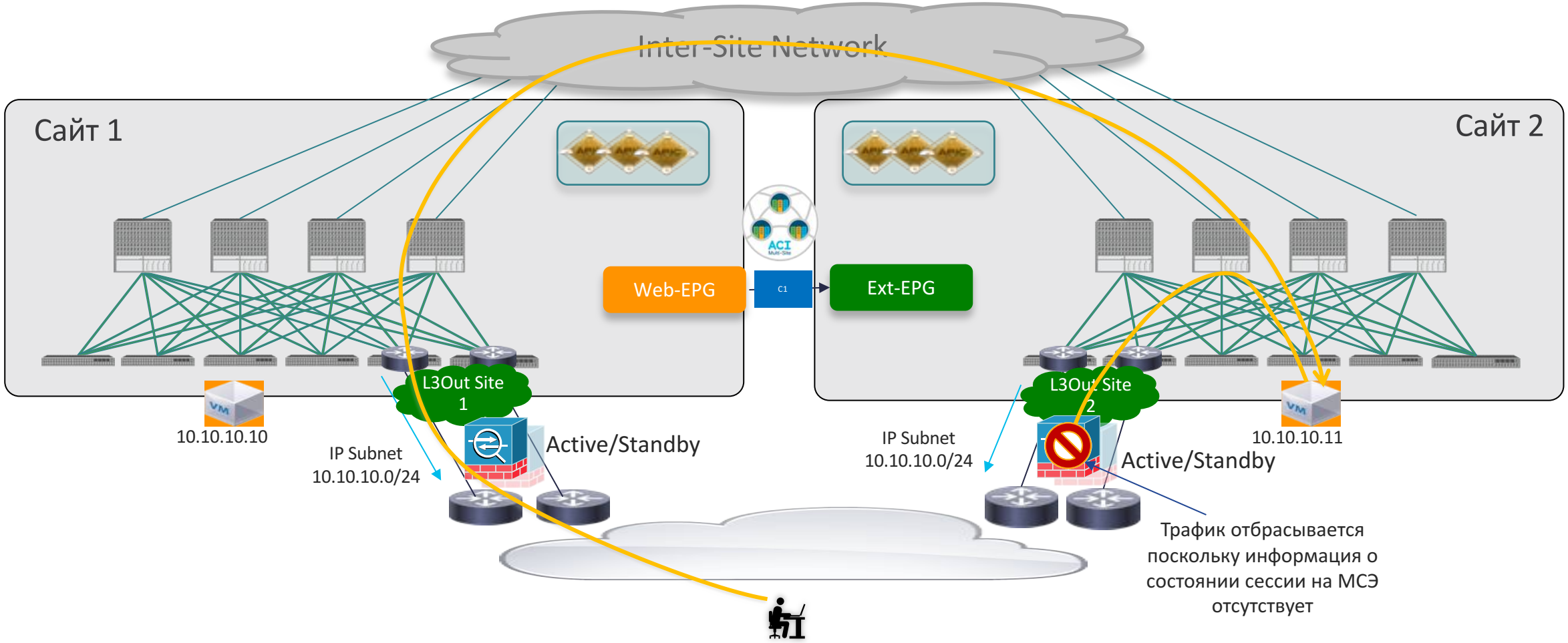


- Отдельные ACI фабрики с независимыми кластерами APIC
- Каждая фабрика рассматривается как отдельный «регион» и «зона доступности»
- Ограничение зоны вносимых изменений
- Использование для сценариев DR и Active/Active
- «Нет ограничений» по расстоянию

- 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

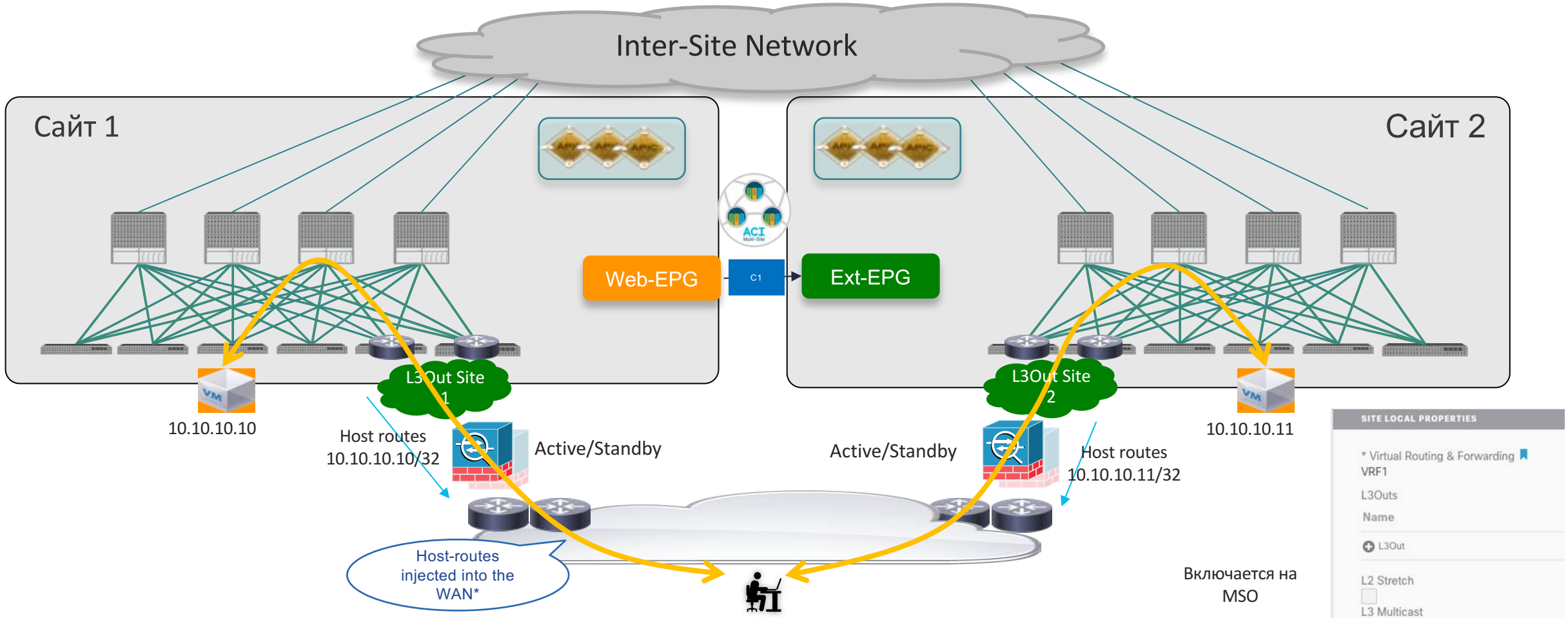
ACI Multi-Site и L3Out

По умолчанию для выхода наружу используется локальный L3Outs для исходящего трафика



Решение проблемы с ассиметричной маршрутизацией

Использование анонсирования Host-Routes

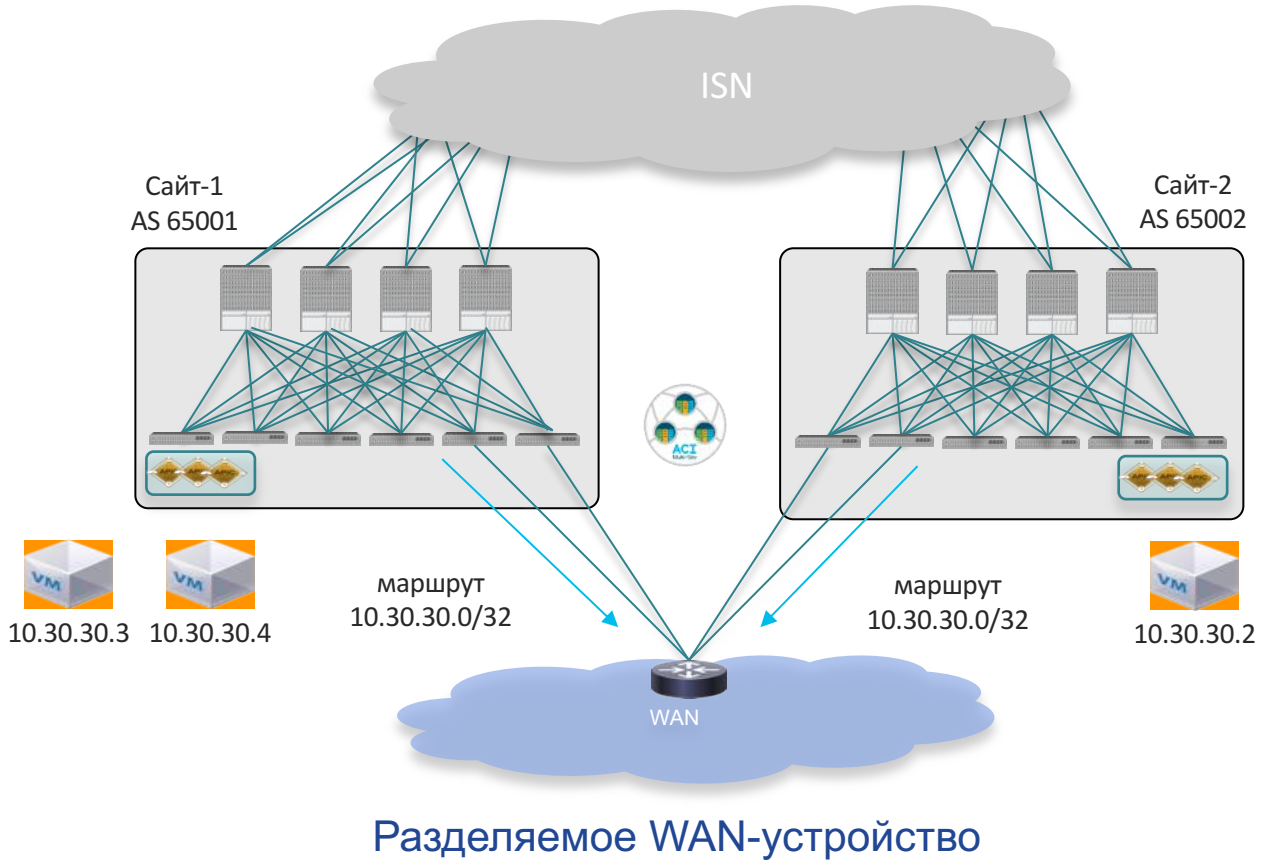


- Оптимизация для входящего трафика требует настройки host-routes на L3Out
  - Поддерживается на ACI Border Leaf начиная с релиза ACI 4.0(1)
  - Так же поддерживается на GOLF L3Out (включается на уровне VRF)

Настройка L3out в ACI Multi-Site

0 Созданы L3out и настроены BGP-стыки между фабриками и WAN-устройством

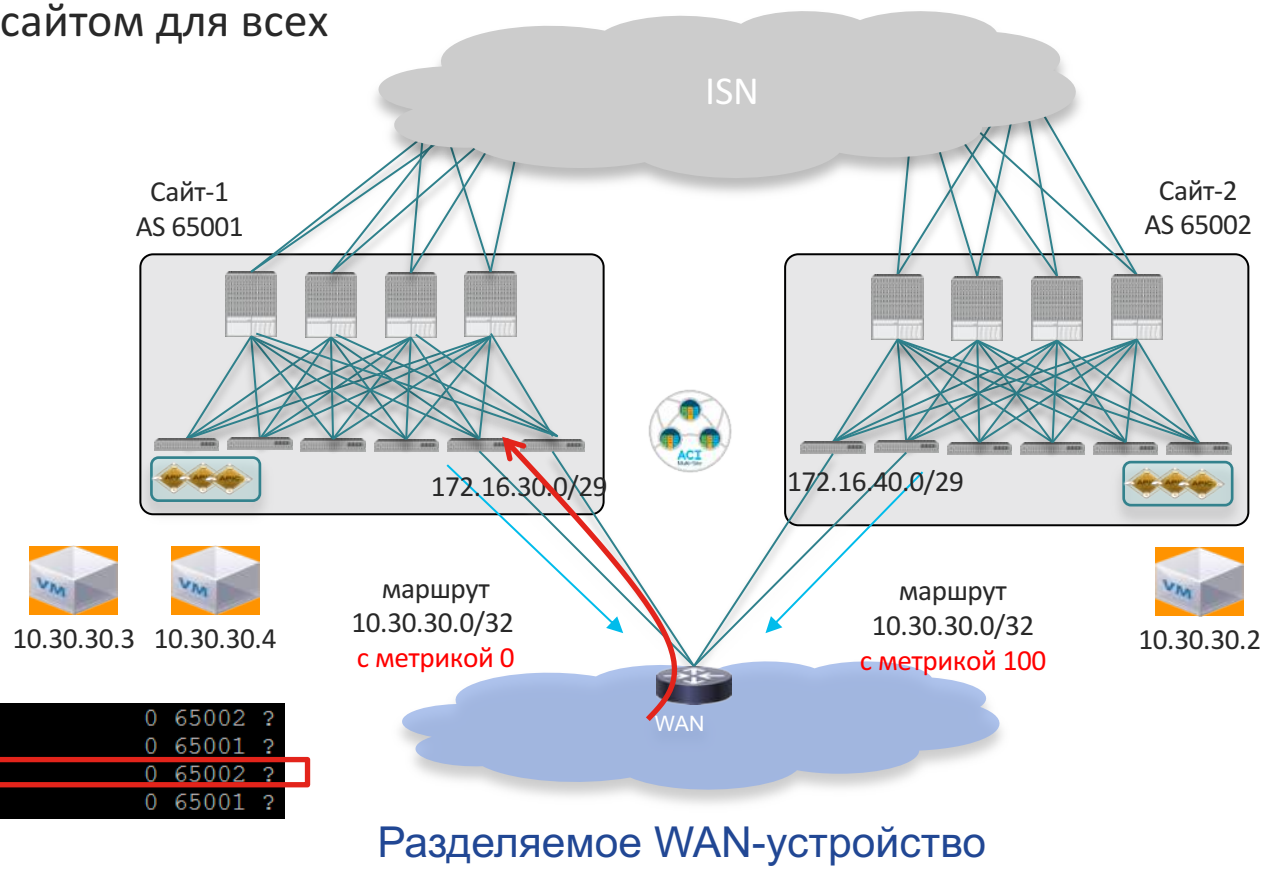
1 Исходное состояние



Настройка L3out в ACI Multi-Site
Манипуляция BGP-атрибутами

Сайт-1 должен стать «домашним» сайтом для всех приложений из VLAN-30

2 Настройка BGP-атрибутов для манипуляции входящим трафиком




```
*m   10.30.30.0/24     172.16.30.1              0             0 65002 ?
*                      172.16.40.1            100             0 65001 ?
*>                     172.16.30.2              0             0 65002 ?
*                      172.16.40.2            100             0 65001 ?
```

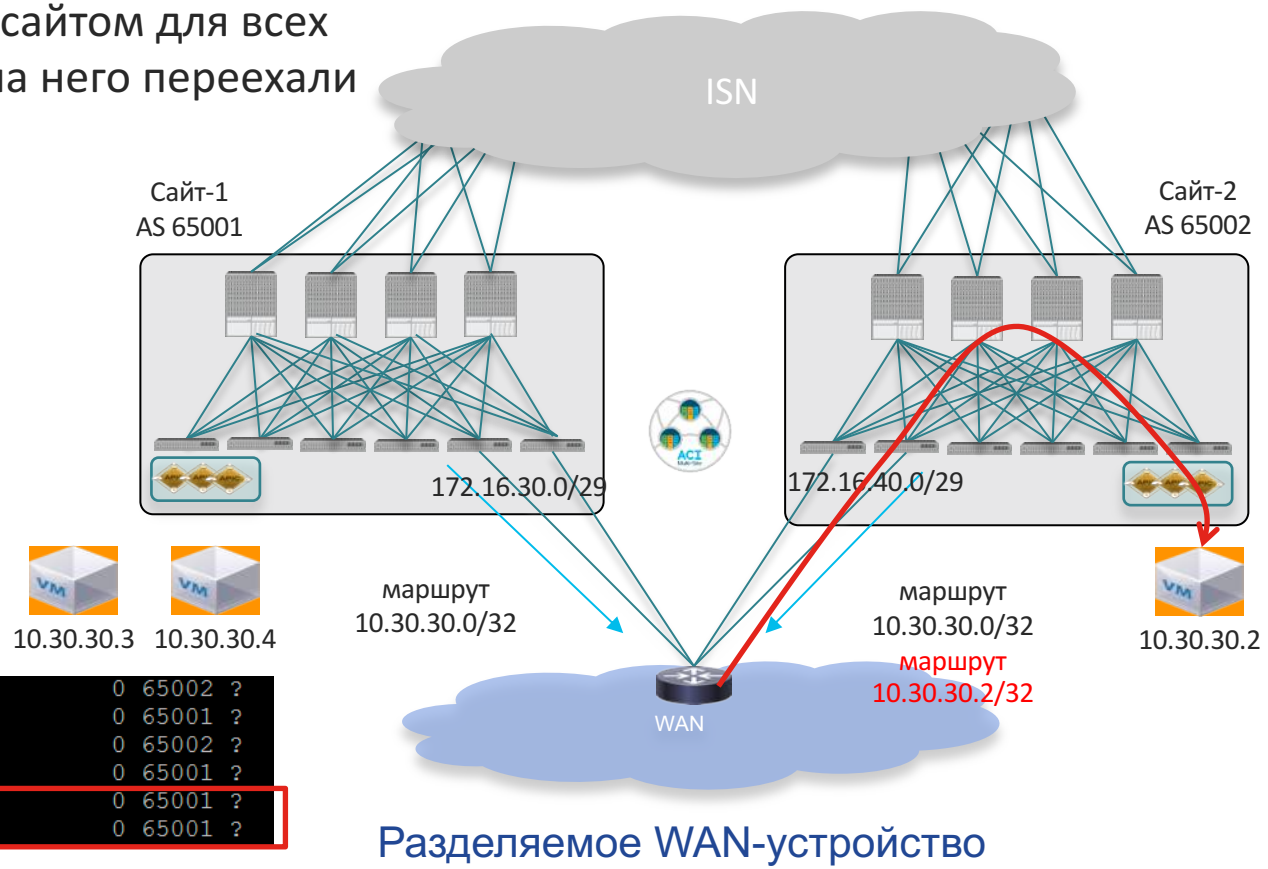
```
B        10.30.30.0/24 [20/0] via 172.16.30.2, 00:02:05
                       [20/0] via 172.16.30.1, 00:02:05
```

Настройка L3out в ACI Multi-Site
анонсирование /32 маршрутов

Сайт-2 должен стать «домашним» сайтом для всех приложений из VLAN-30 которые на него переехали

4

Настройка анонсирования /32 маршрутов для BD VLAN-30 при помощи MSO




```
*m   10.30.30.0/24     172.16.30.1              0             0 65002 ?
*                      172.16.40.1            100             0 65001 ?
*>                     172.16.30.2              0             0 65002 ?
*                      172.16.40.2            100             0 65001 ?
*>   10.30.30.2/32     172.16.40.1            100             0 65001 ?
*m                     172.16.40.2            100             0 65001 ?
```

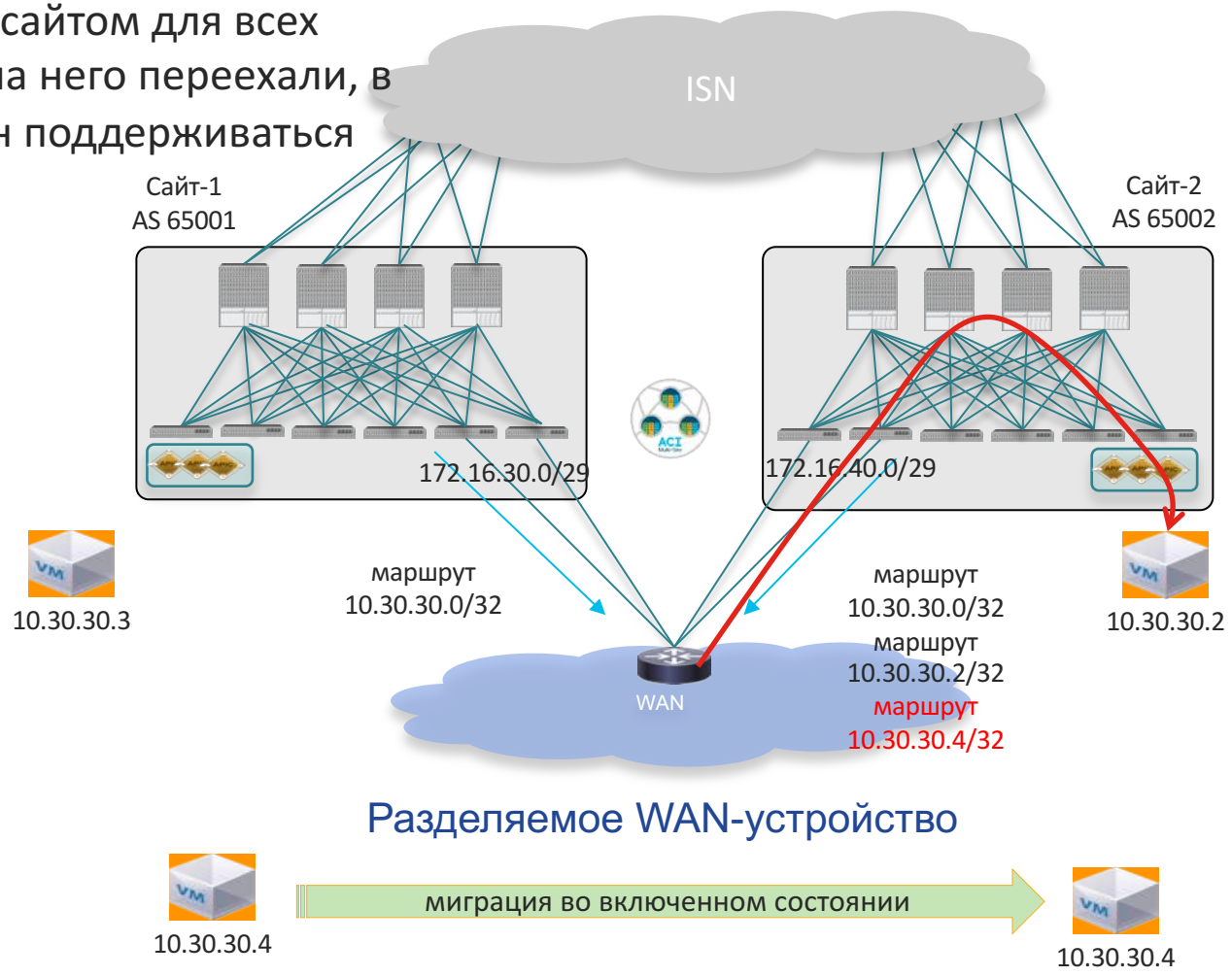
```
B        10.30.30.0/24 [20/0] via 172.16.30.2, 00:02:05
                       [20/0] via 172.16.30.1, 00:02:05
B        10.30.30.2/32 [20/100] via 172.16.40.2, 00:00:25
                       [20/100] via 172.16.40.1, 00:00:25
```

Настройка L3out в ACI Multi-Site
поддержка Cross vCenter Migration

Сайт-2 должен стать «домашним» сайтом для всех приложений из VLAN-30 которые на него переехали, в качестве способа переезда должен поддерживаться Cross vCenter Migration

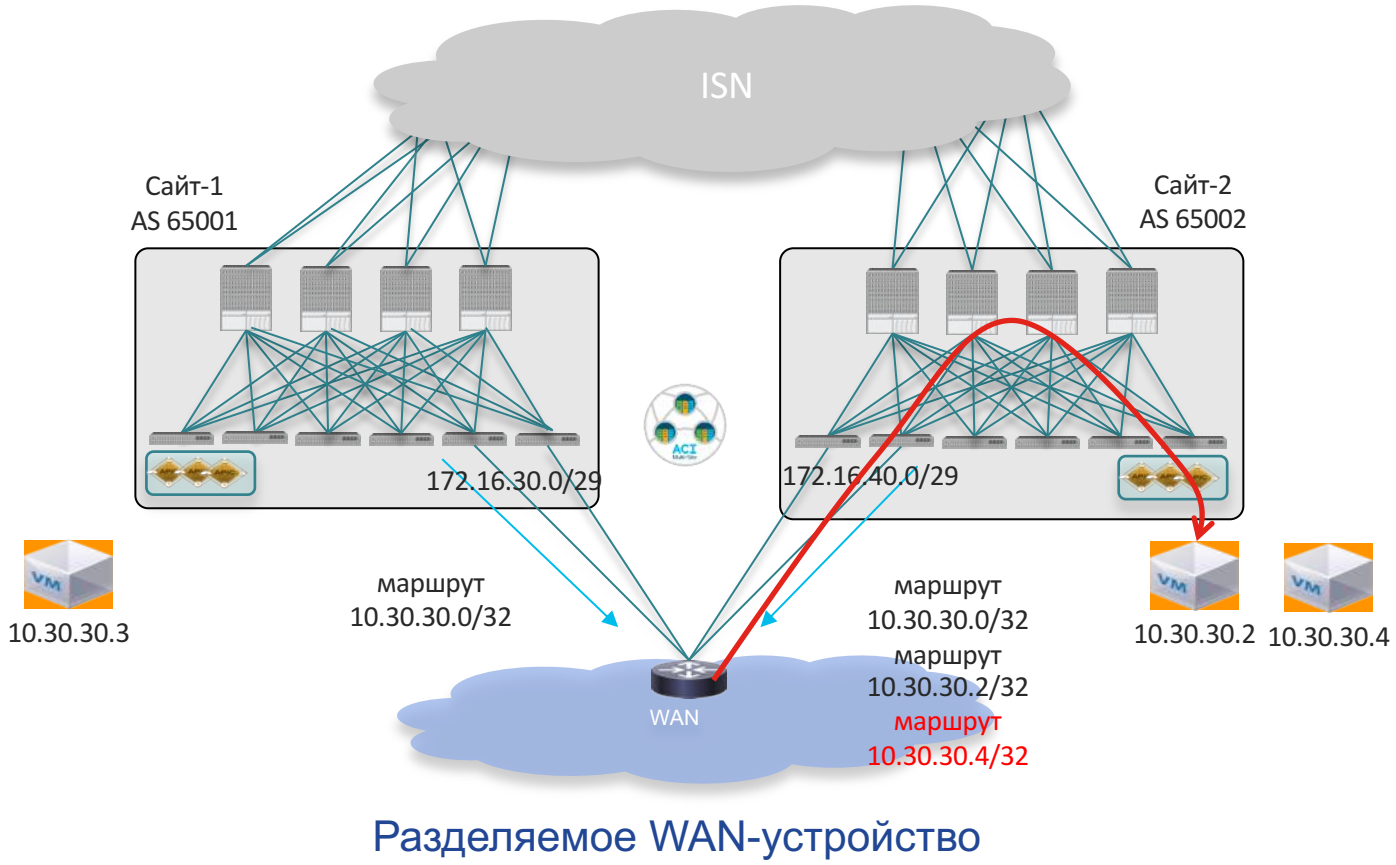
4

Миграция ВМ 10.30.30.4 с сайта-1 на сайт-2 при помощи Cross vCenter Migration



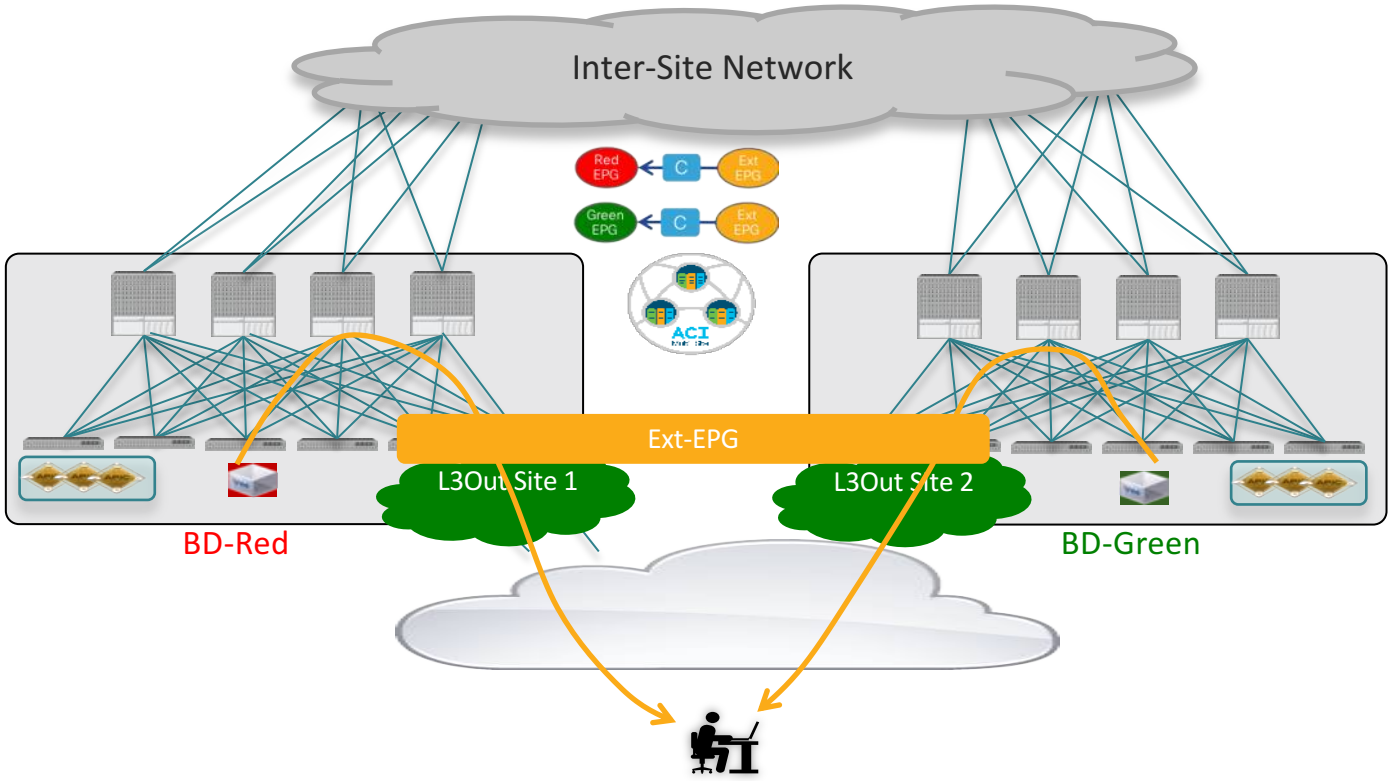
Настройка L3out в ACI Multi-Site
Промежуточные итоги демонстрации

- Удобство манипуляции BGP-атрибутами при помощи политик
- Включение анонса /32 за одну операцию, при необходимости
- Встроенная поддержка Cross vCenter Migration



ACI Multi-Site и L3Out

Растягивать или нет Ext-EPG между сайтами?

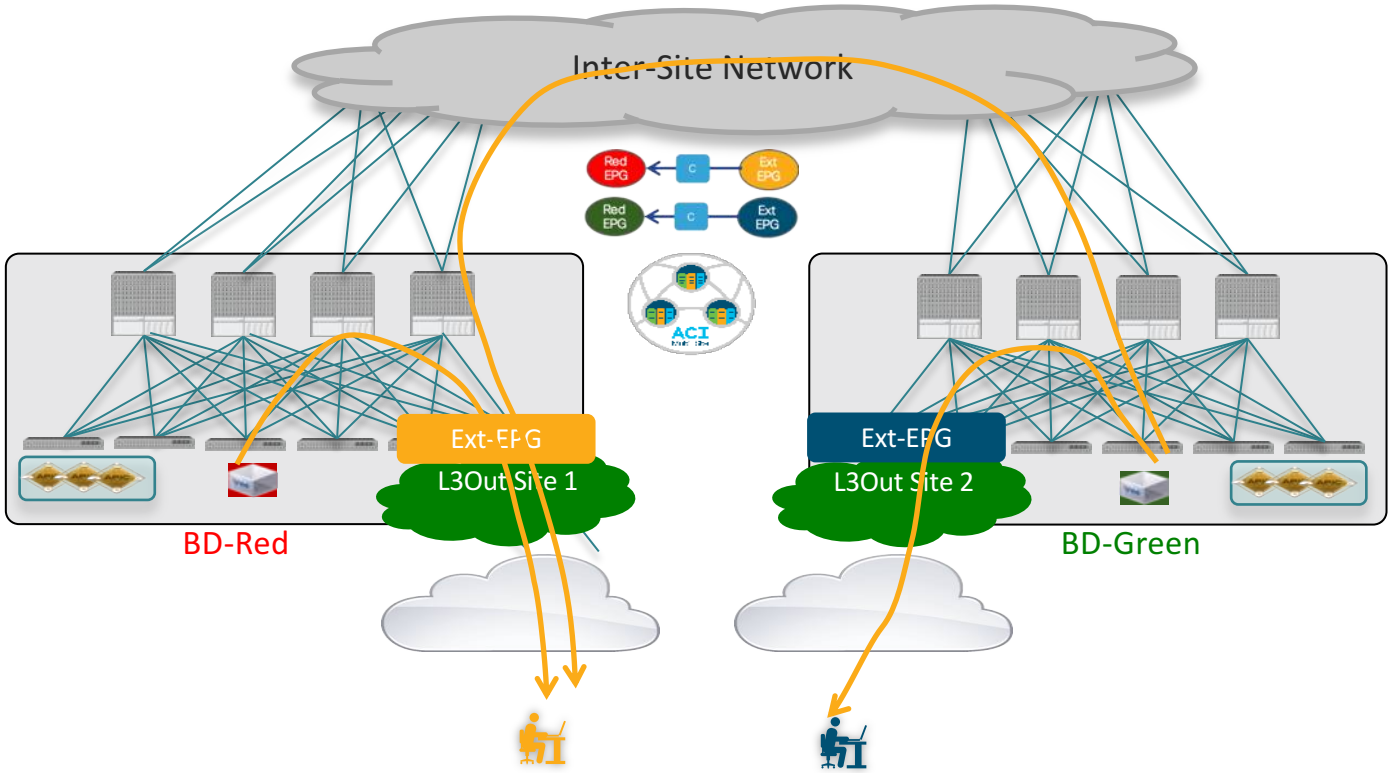


- Ext-EPG можно определить внутри шаблона который будет ассоциирован с несколькими сайтами (stretched object)
  - Ext-EPG **должен** быть потом связан с локальным L3Out на уровне сайта (“site level”)
  - L3Out-ы остаются независимыми объектами и настраиваются индивидуально
- Этот метод рекомендуется когда подключение фабрик происходит к общему набору внешних ресурсов, например к сети WAN или LAN
  - Упрощение настройки политик фильтрации между фабрикой и внешними потребителями
  - Возможность манипулировать префиксами при помощи route-map на каждом L3Out остается

Рекомендация – использовать один растянутый ext-EPG для упрощения

ACI Multi-Site и L3Out

Растягивать или нет Ext-EPG между сайтами?



Рекомендация – использовать один растянутый ext-EPG для упрощения

- Разные Ext-EPG могут быть определены в шаблонах которые привязаны к разным сайтам (non stretched objects)

  Каждый Ext-EPG **может** быть связан с локальным L3Out на “global” или “site level” уровнях шаблона конфигурации

- Предполагает что внешние домены к которым подключается фабрика разные

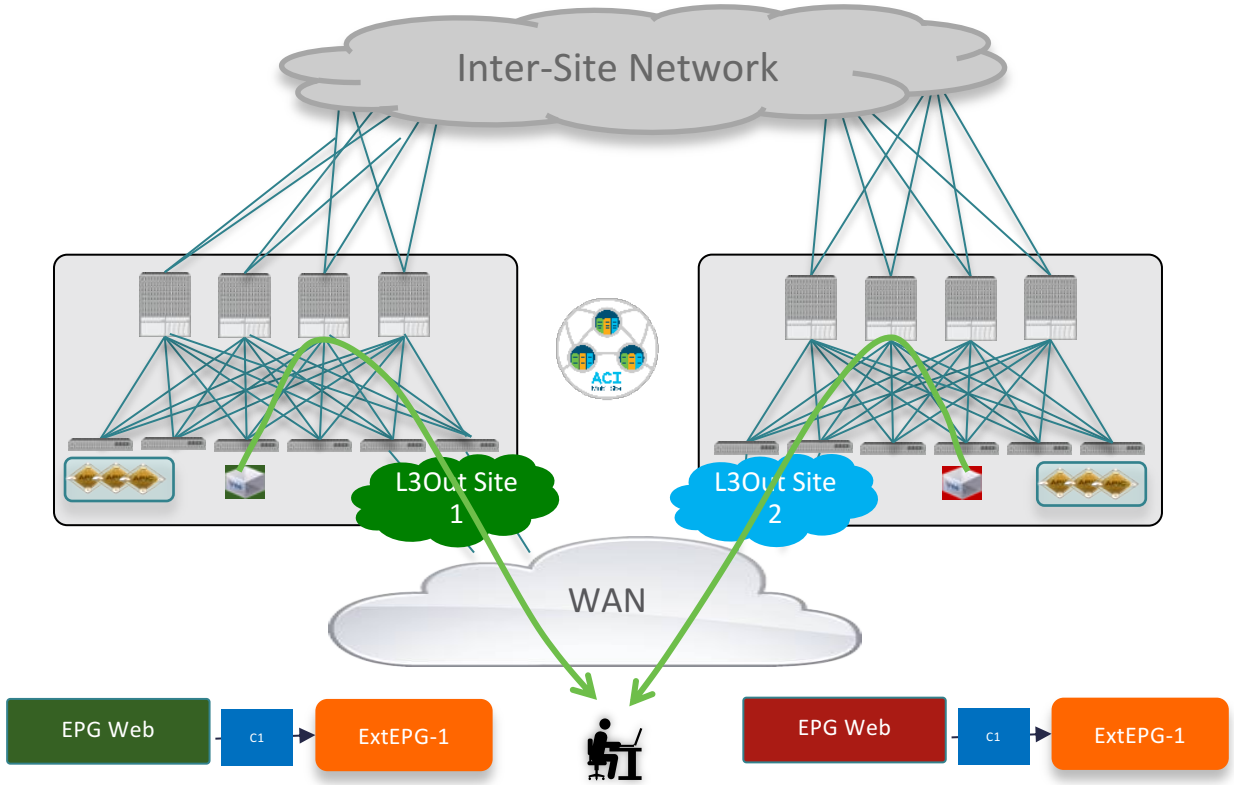
  это бывает редко !

- Можно использовать сеть 0.0.0.0/0 для классификации внешних объектов
- Может потребоваться настройка Intersite L3Out

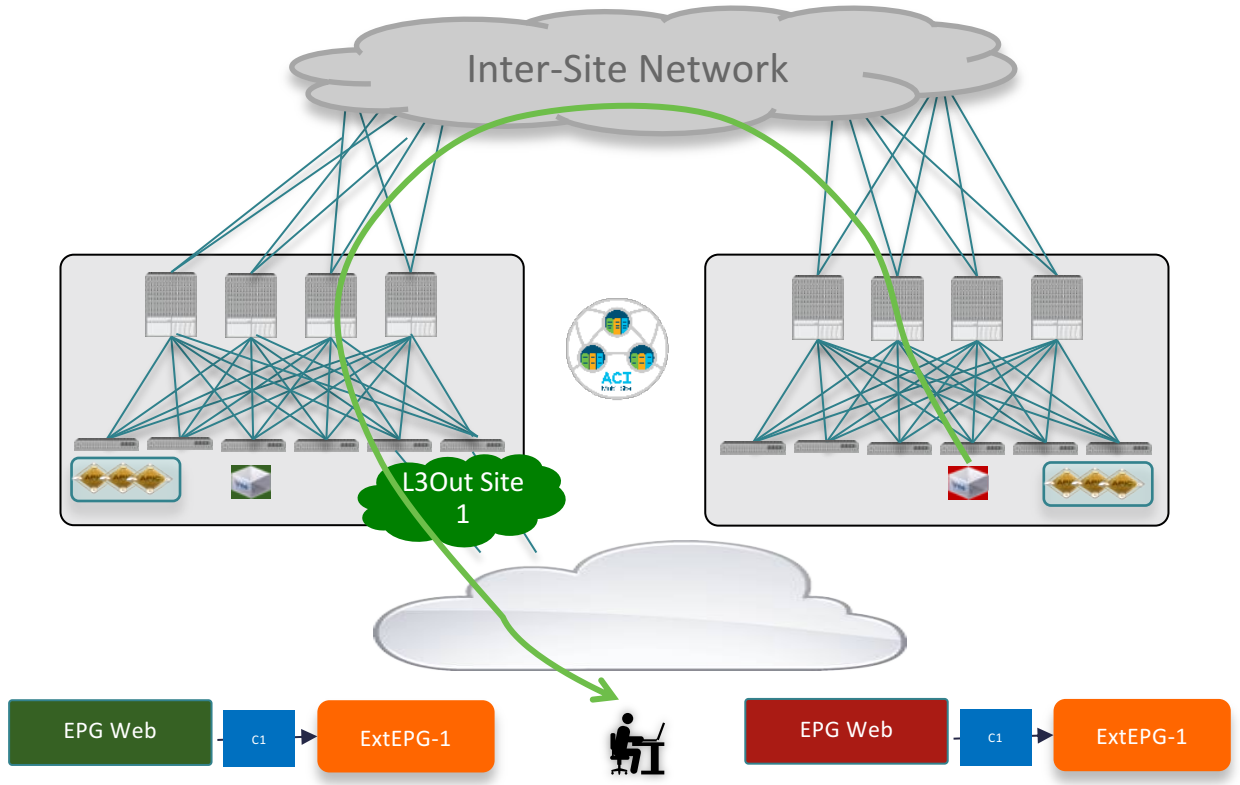
Multi-Site и L3Out

Различные возможности по организации внешних покдлючений

Независимые L3out

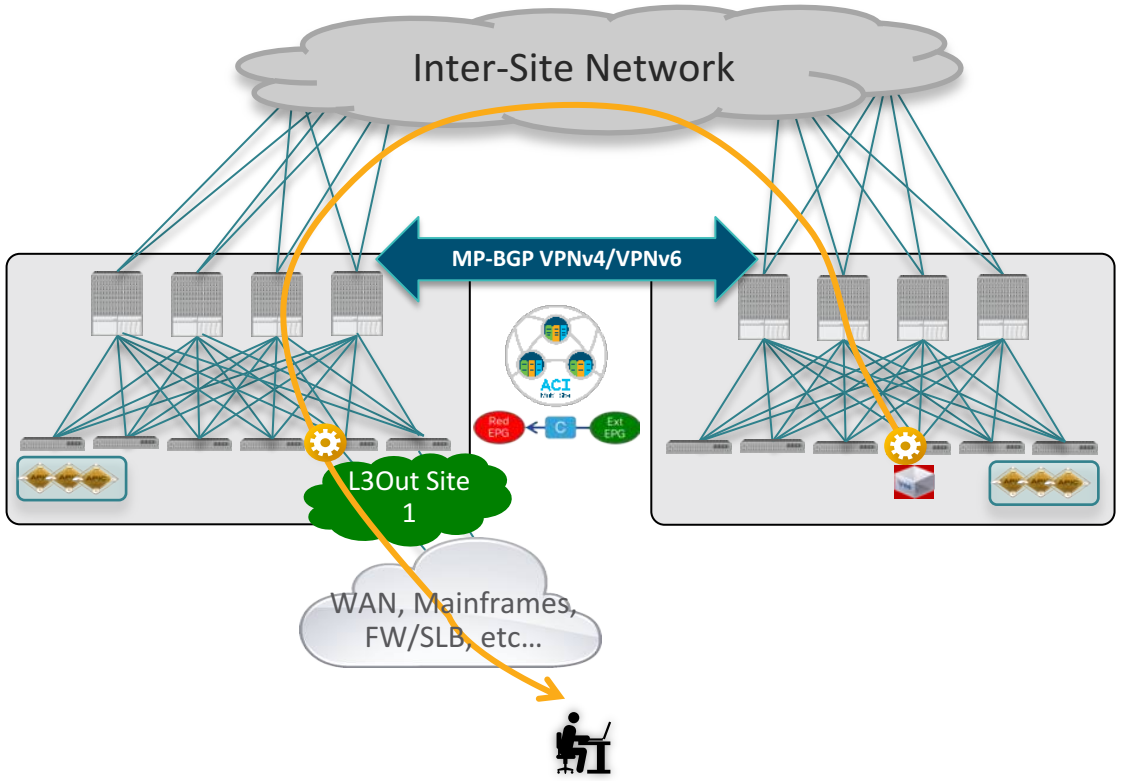


Использование L3out другого сайта



ACI Multi-Site и L3Out

Поддержка Intersite L3Out

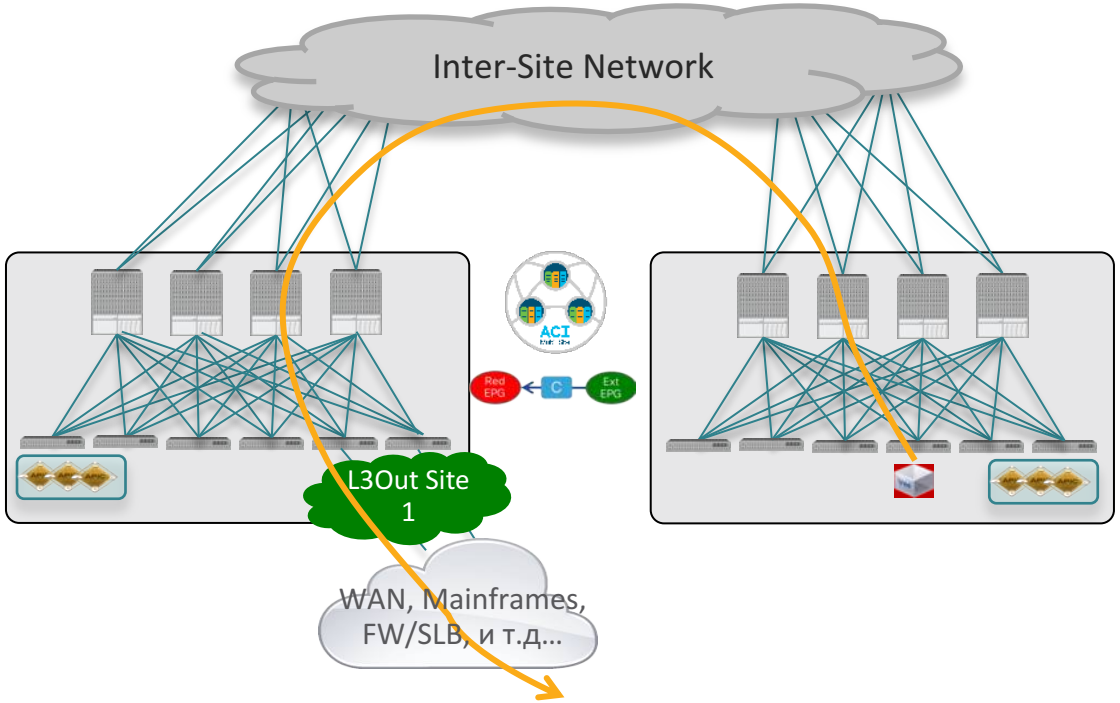


- Начиная с релиза ACI 4.2(1) возможно взаимодействие конечных хостов одной фабрики со внешними ресурсами (WAN, Mainframe, FW/SLB) которые доступны через L3Out другой фабрики
- Внешние префиксы передаются между спайнами при помощи MP-BGP VPNV4/VPNv6 сессий
- При этом трафик инкапсулируется напрямую в TEP на удаленных BL узлах
  - На узлах BL нужно настроить подсеть, которая будет маршрутизироваться поверх ISN сети
- Это же решение используется для превращения нескольких фабрик в транзитные (L3Out to L3Out)

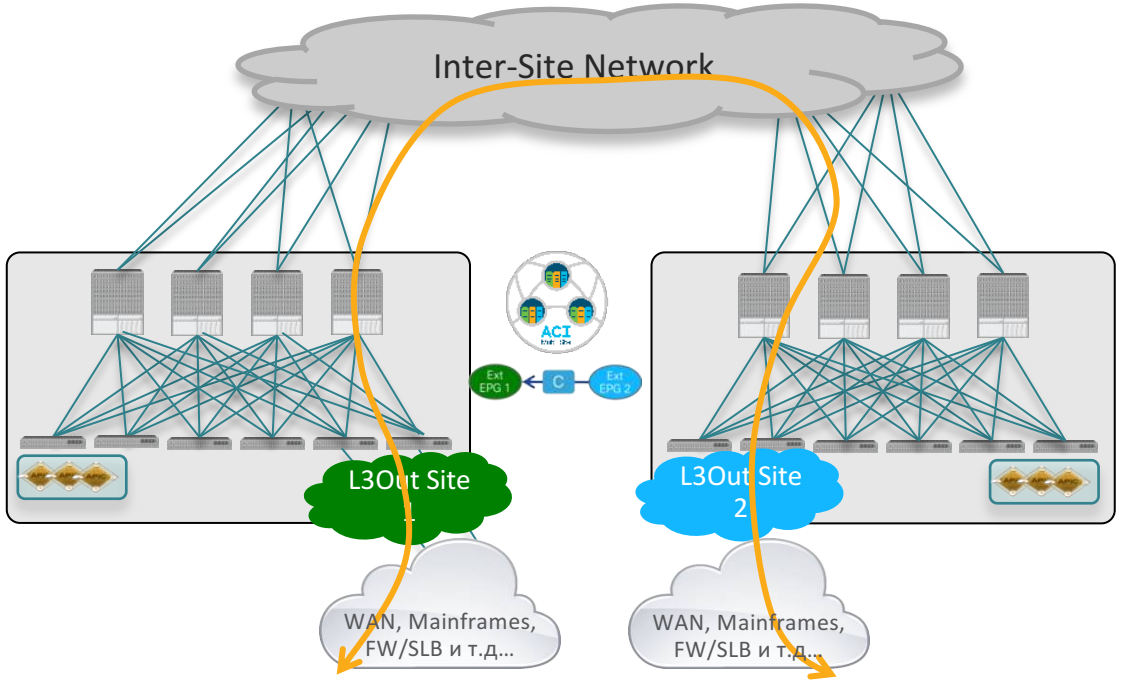
ACI Multi-Site и Intersite L3Out

поддерживаемые сценарии

ACI 4.2(1) Release



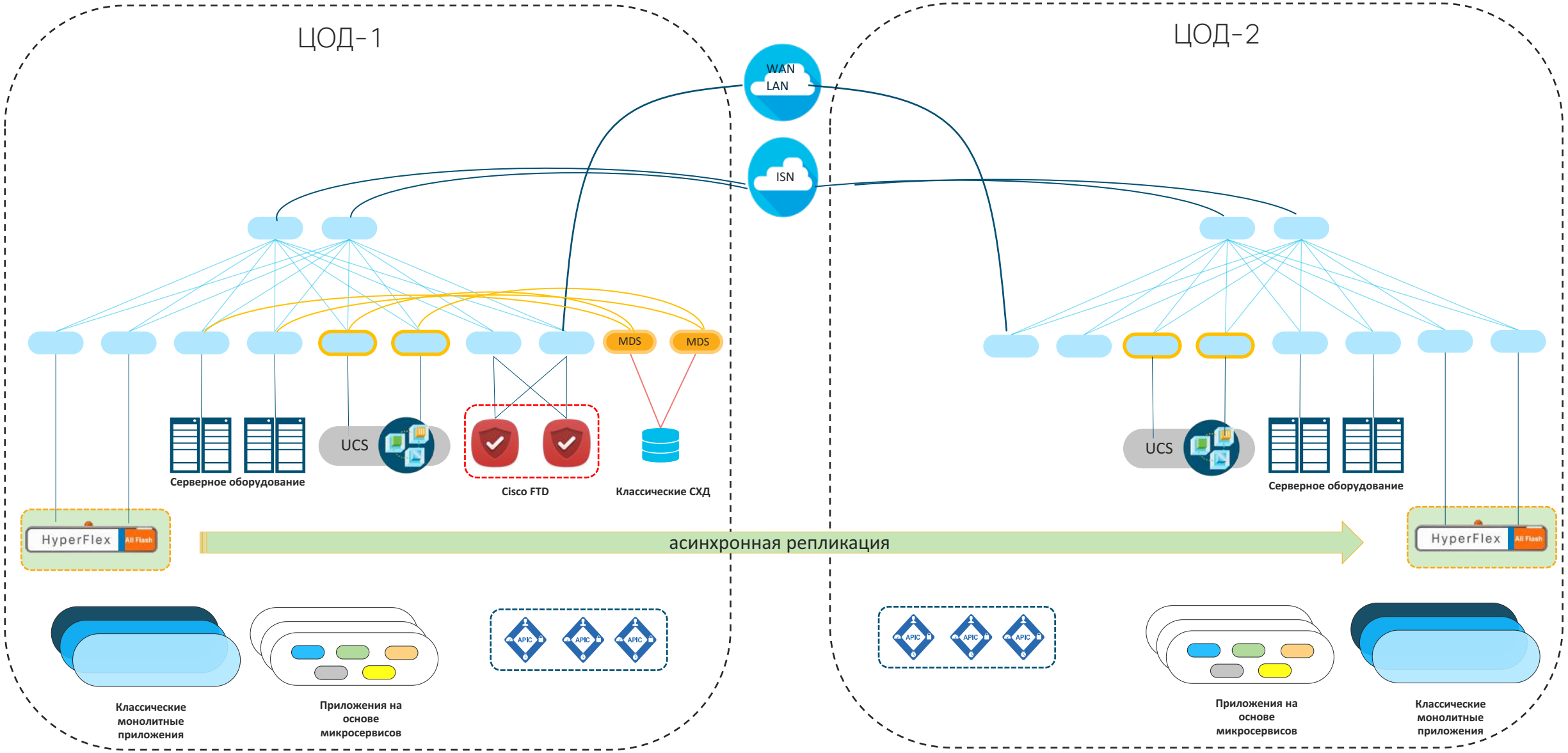
- Endpoint to remote L3Out communication (intra-VRF)
- Endpoint to remote L3Out communication (inter-VRF)



- Inter-site transit routing (intra-VRF)
- Inter-site transit routing (inter-VRF)

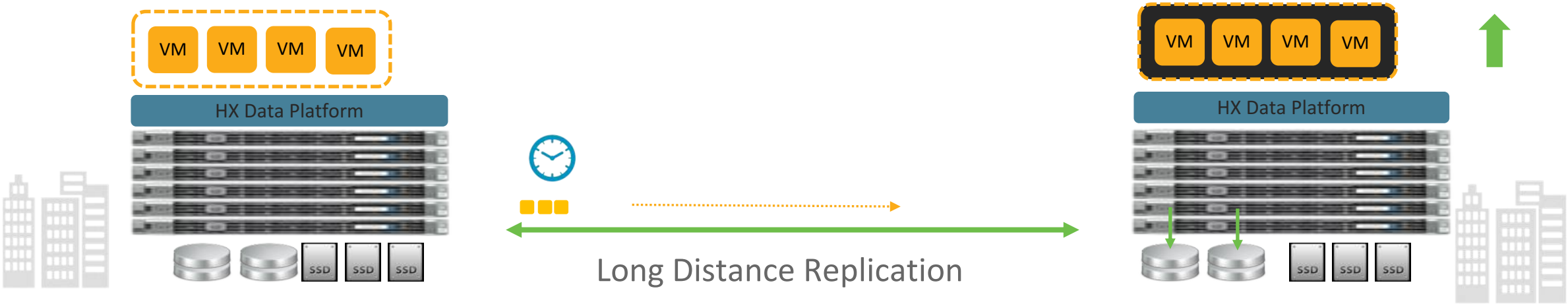
Стратегия резервирования для систем гиперконвергенции

DR



Асинхронная репликация систем Hyperflex
Disaster Recovery встроенными средствами

DR



Автоматизированные сценарии восстановления

Тест DR

- Готовность к DR
- Кастомизация параметров теста DR

Плановая миграция

- Миграция ВМ между кластерами и ЦОД
- Защита после миграции

Отработка отказа

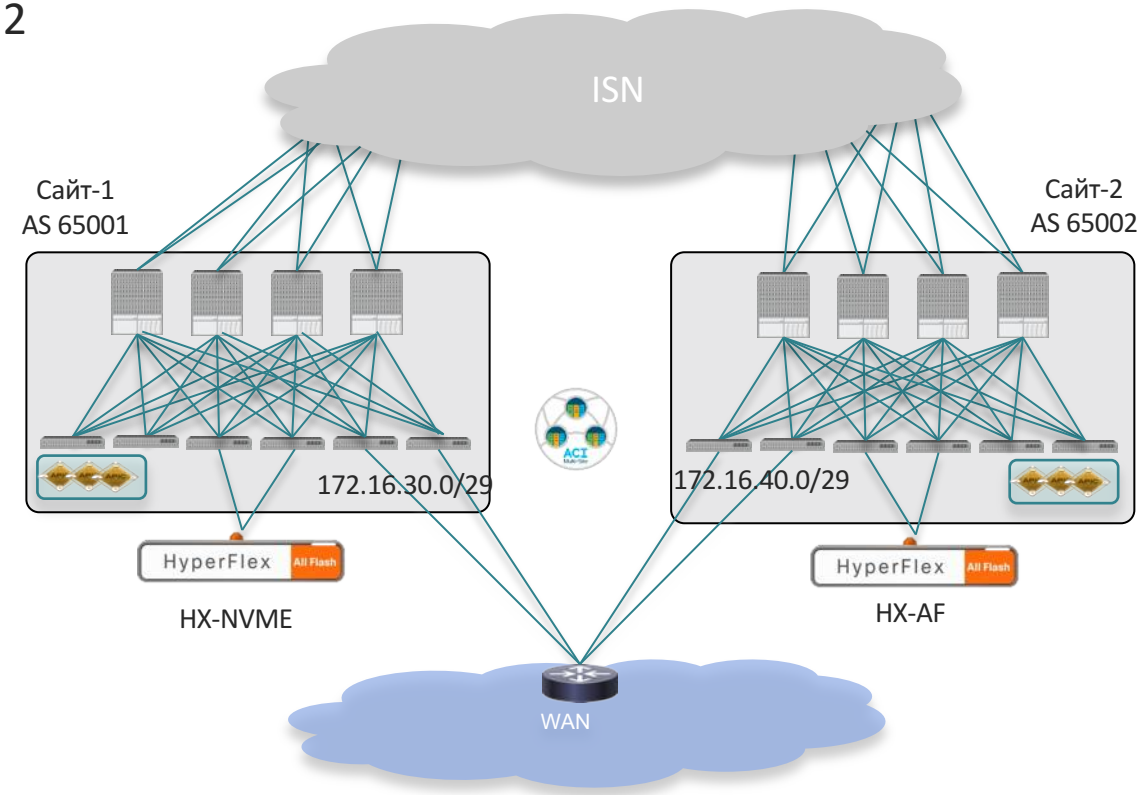
- Восстановление ВМ после аварии
- Защита после восстановления

Асинхронная репликация между Cisco Hyperflex миграция нагрузки между сайтами

Переезд ВМ 10.30.30.3 с сайта 1 на сайт 2

0 Асинхронная репликация между двумя системами Heprflex настроена ранее

1 Миграция ВМ 10.30.30.3 с сайта-1 на сайт-2 при помощи Hyperflex

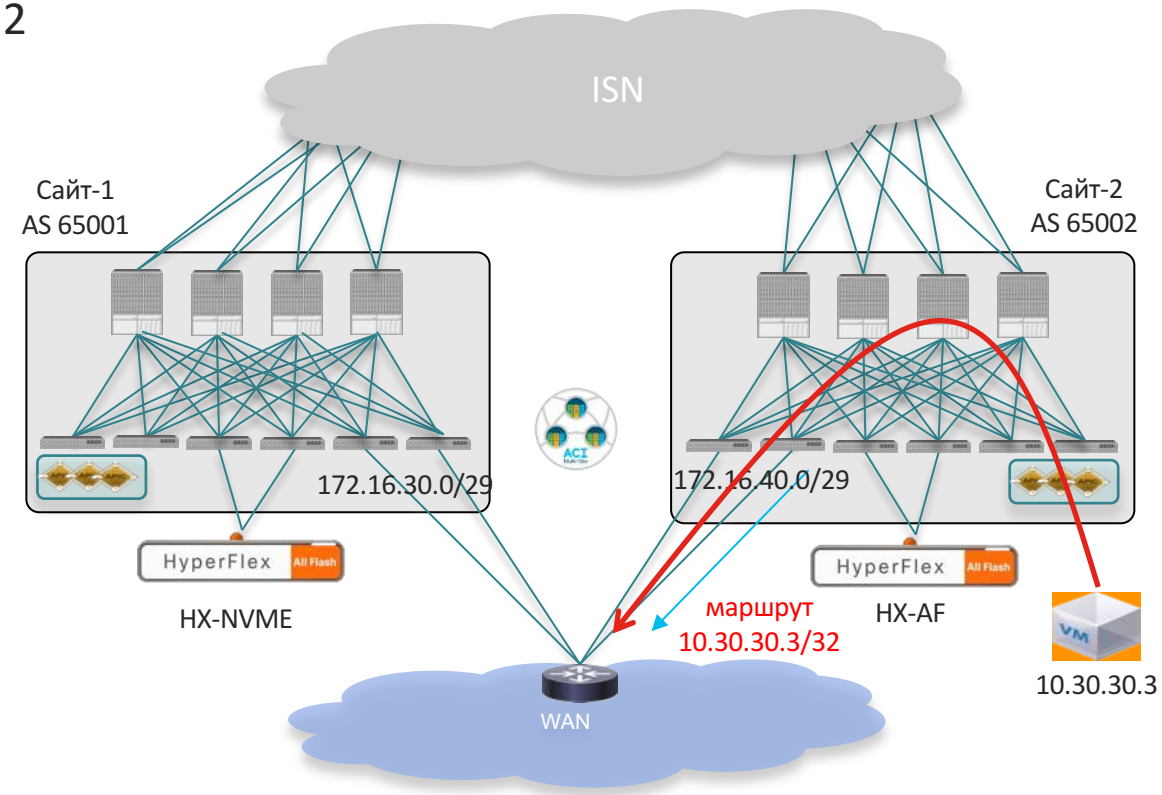


Асинхронная репликация между Cisco Hyperflex миграция нагрузки между сайтами

Переезд ВМ 10.30.30.3 с сайта 1 на сайт 2

2

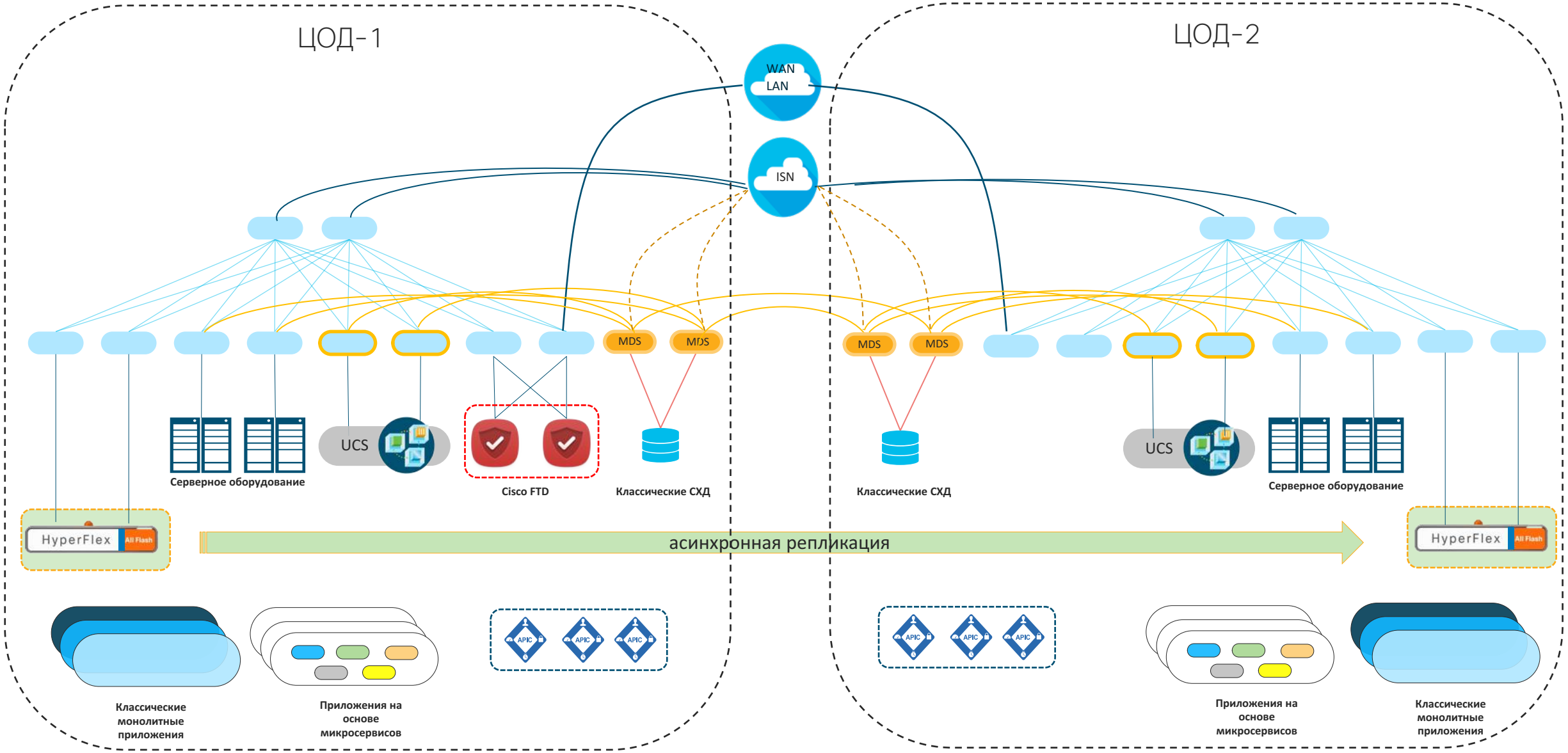
ACI фабрика автоматически обеспечивает анонс /32 после восстановления ВМ



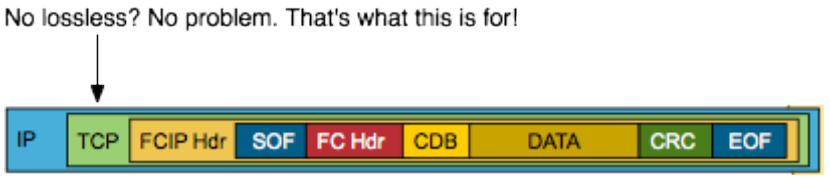
Разделяемое WAN-устройство

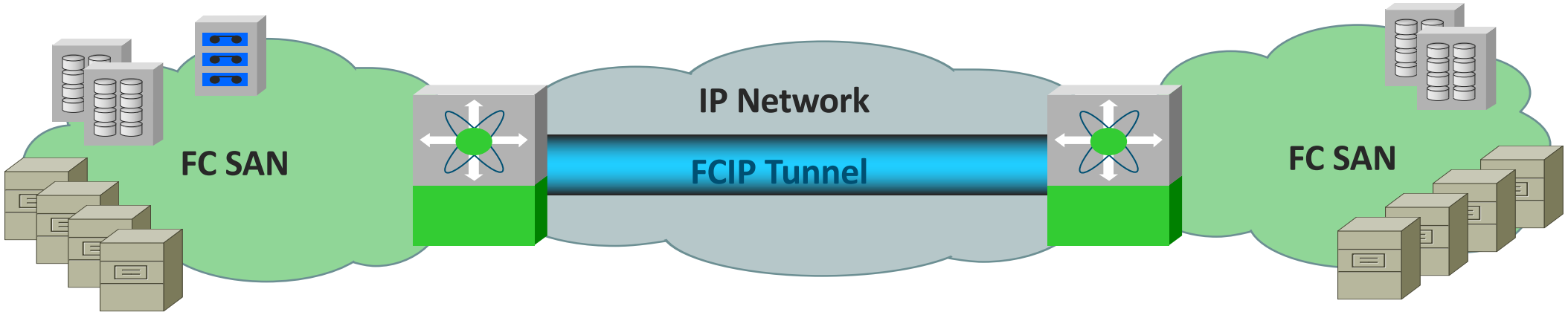
Сеть хранения данных для нескольких ЦОД

DR



FCIP: Fibre Channel over IP



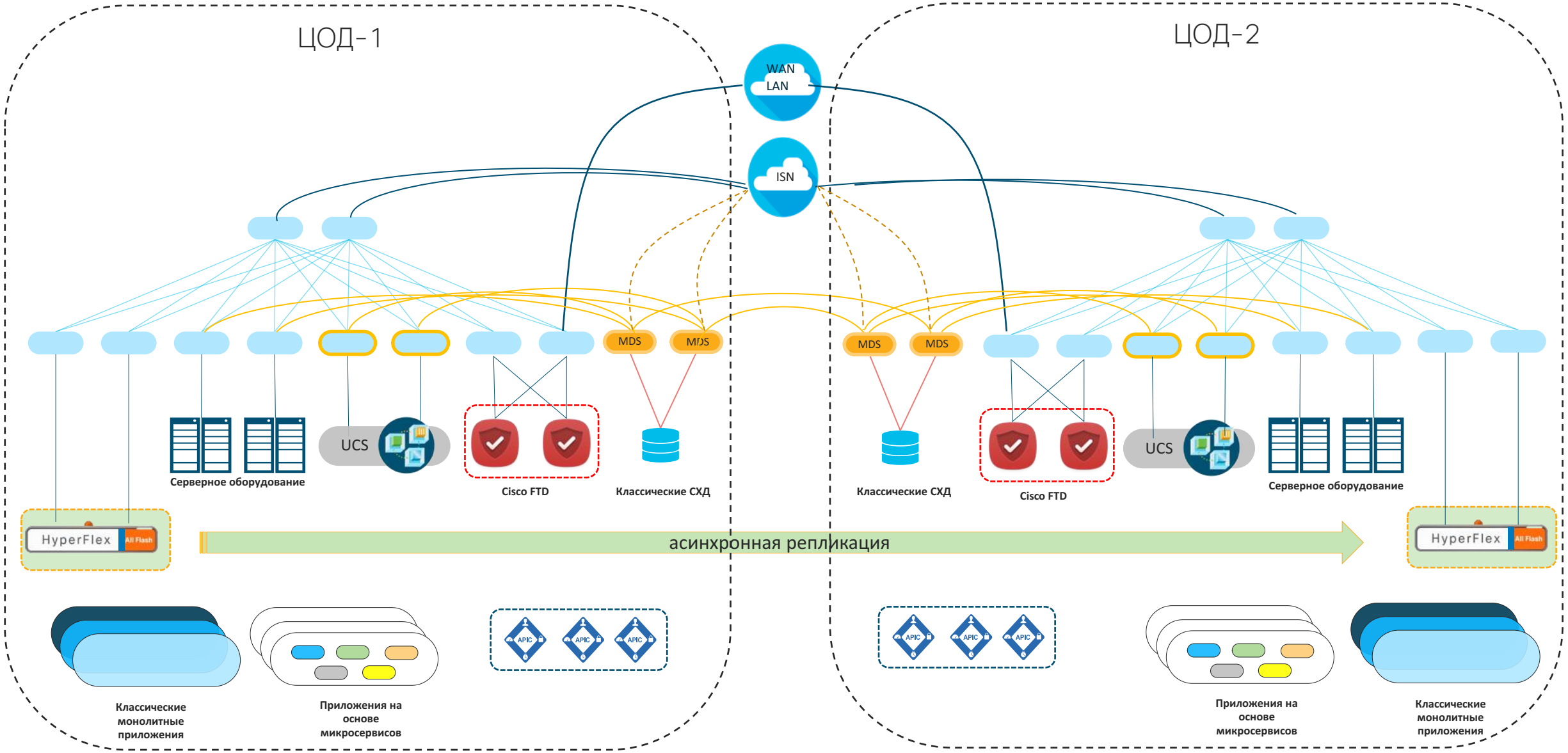


FCIP: IETF стандарт для связи Fibre Channel SAN через IP (RFCs 3821 и 3643)

- Соединение «точка-точка» (туннель) между двумя FCIP устройствами
- Используется TCP – могут использоваться механизмы 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 >= 2300 байт

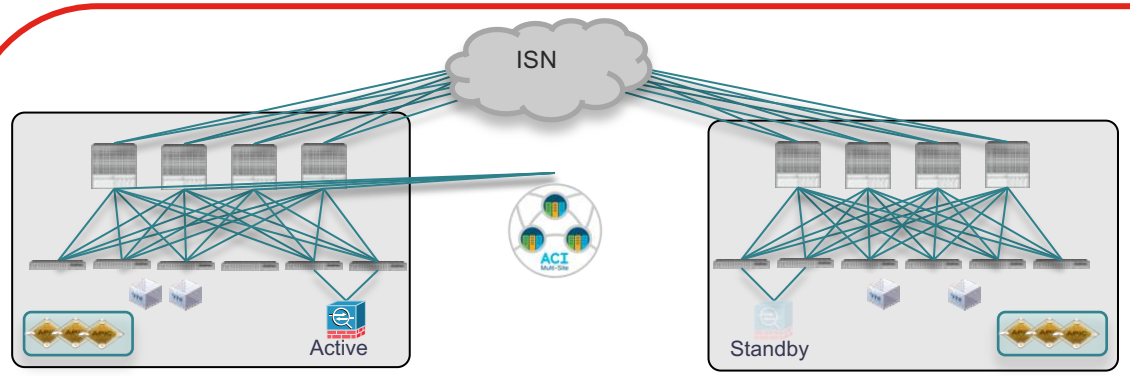
Межсетевые экраны на границе и внутри сети ЦОД (North-South + East West)

DR

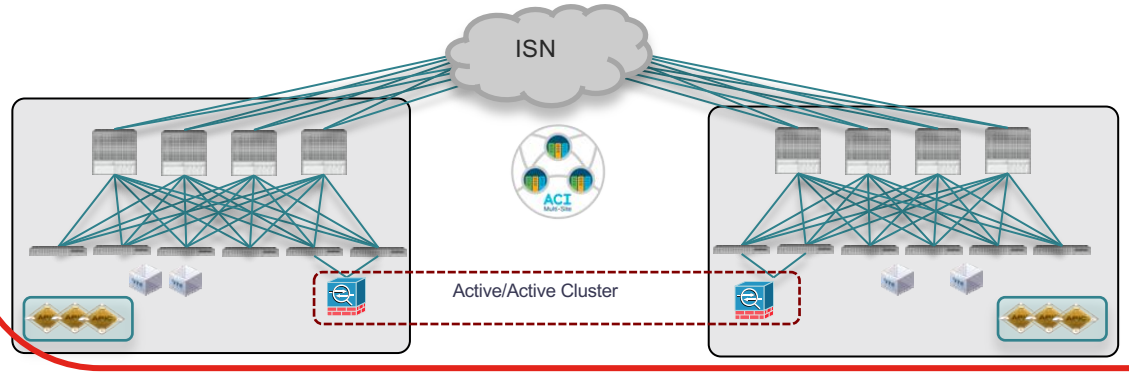


Multi-Site и режимы интеграции L4/L7 устройств

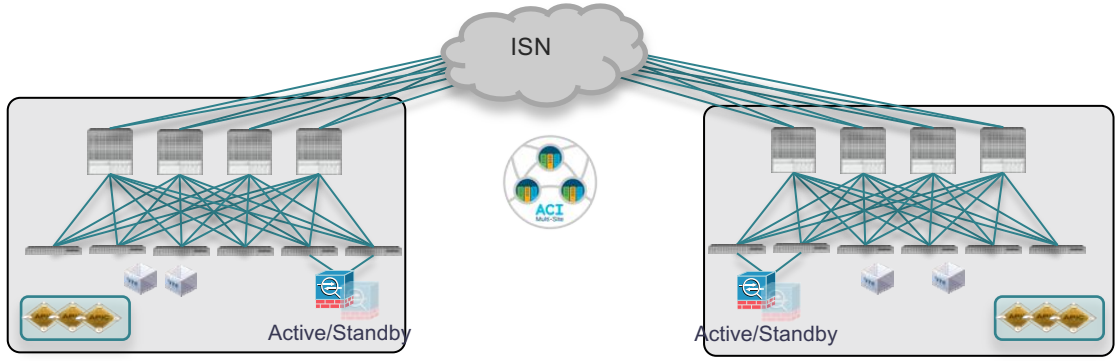
Поддерживается в ACI Multi-Pod



- Одна пара Active/Standby устройств «растянутая» между сайтами
- **Поддерживается если МСЭ работают в L2 режиме или в L3 режиме, но используются как шлюзы по умолчанию для нагрузки**



- Active/Active кластер МСЭ растянутый между сайтами (один логический FW)
- Requires the ability of discovering the same MAC/IP info in separate sites at the same time
- **Не поддерживается и не рекомендуется в Multi-Site**



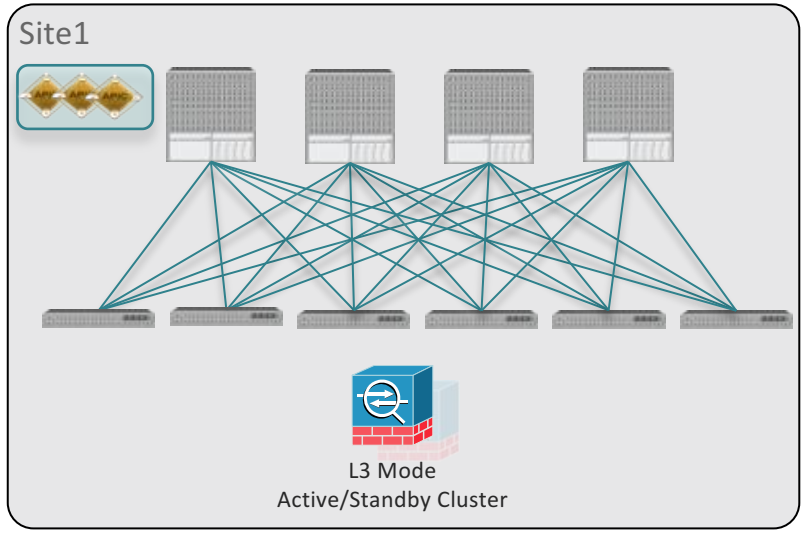
- **Рекомендуется для режима ACI Multi-Site**
- Вариант 1: с версии ACI 3.0 для режима N-S если FW подключается в L3 режиме к фабрике → обязательно использовать оптимизацию для вх. трафика
- Вариант 2: с версии ACI 3.2 с использованием сервисного графа и Policy Based Redirection (PBR)

Независимая пара Active/Standby на каждом сайте Вариант 2: использование сервисного графа и PBR

- Зависимость от версий ПО и железа:
  - Поддержка с релиза ACI 3.2(1)
  - Обязательно использовать EX/FX коммутаторы (для compute и service узлов)
- PBR политика применяемая на leaf коммутаторе может перенаправлять трафик на сервисное устройство расположенное на локальном сайте
- Требуется настраивать независимые сервисные узлы на каждом сайте
  - Поддерживаются различные варианты дизайна: Active/Standby пара на сайт, Active/Active кластер на сайт, несколько активных Active узлов на сайт

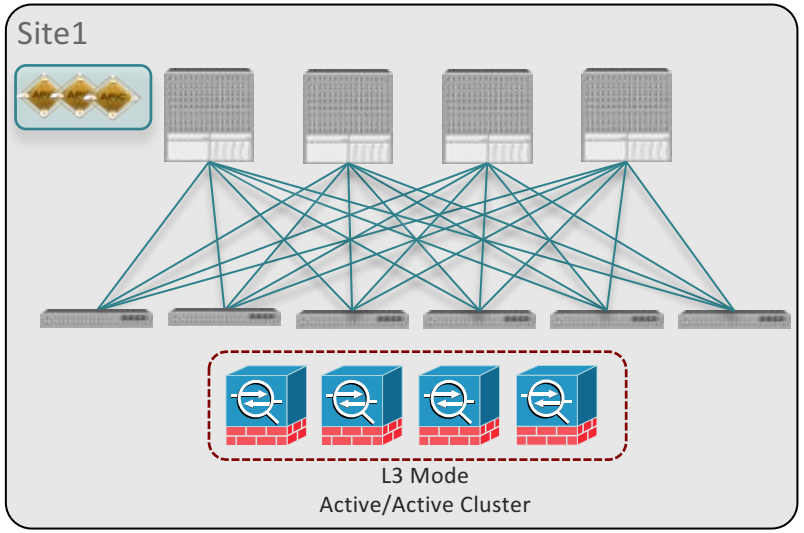
Использование сервисного графа и PBR
Варианты резервирования сервисных устройств

Active/Standby кластер



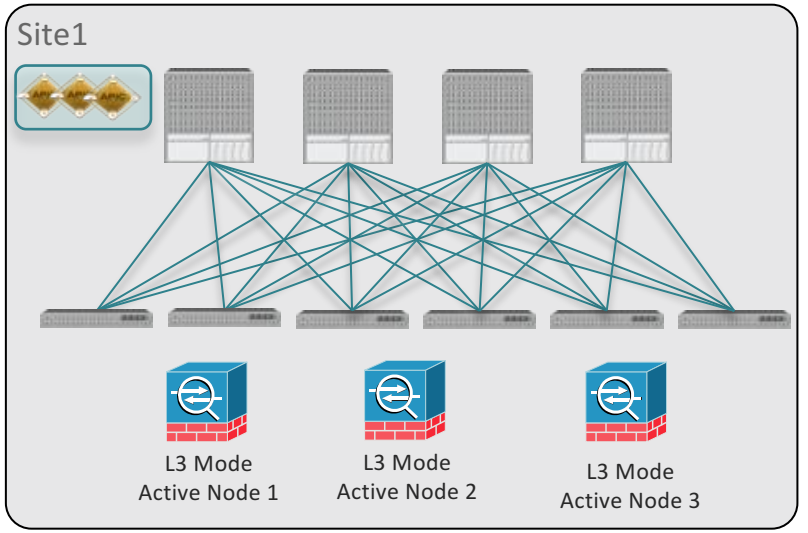
- Active/Standby пара представляет собой одну запись с парой MAC/IP в PBR политике

Active/Active кластер



- Active/Active кластер представляет собой одну запись с парой MAC/IP в PBR политике
- Режим Spanned Ether-Channel поддерживается на Cisco ASA/FTD

Независимые устройства, каждое в режиме Active



- Каждый Active узел представляет собой уникальную запись в парой MAC/IP в PBR политике
- Симметричный PBR обеспечивает для каждого потока обработку на одном Active узле во всех направлениях

Использование сервисных графов и PBR
Шаги по настройке

Шаги по настройке выполняются с использованием двух интерфейсов APIC и Multi-Site Orchestrator

Настройка APIC (Выполняется на каждом сайте)

1. Создать Logical Device
2. Создать PBR policy

Замечание: предполагается что EPG, BD, Контракты и т.д. уже настроены

Настройка MSO

- Уровень общего для сайтов шаблона (Template level)
  1. Создать Service Graph
  2. Ассоциировать Service Graph с контрактом
  3. Для сценария East-West, настроить IP подсеть внутри Provider EPG
- Для каждого сайта
  1. Выбрать Logical Device(s) созданный на APIC
  2. Выбрать интерфейс и PBR политику

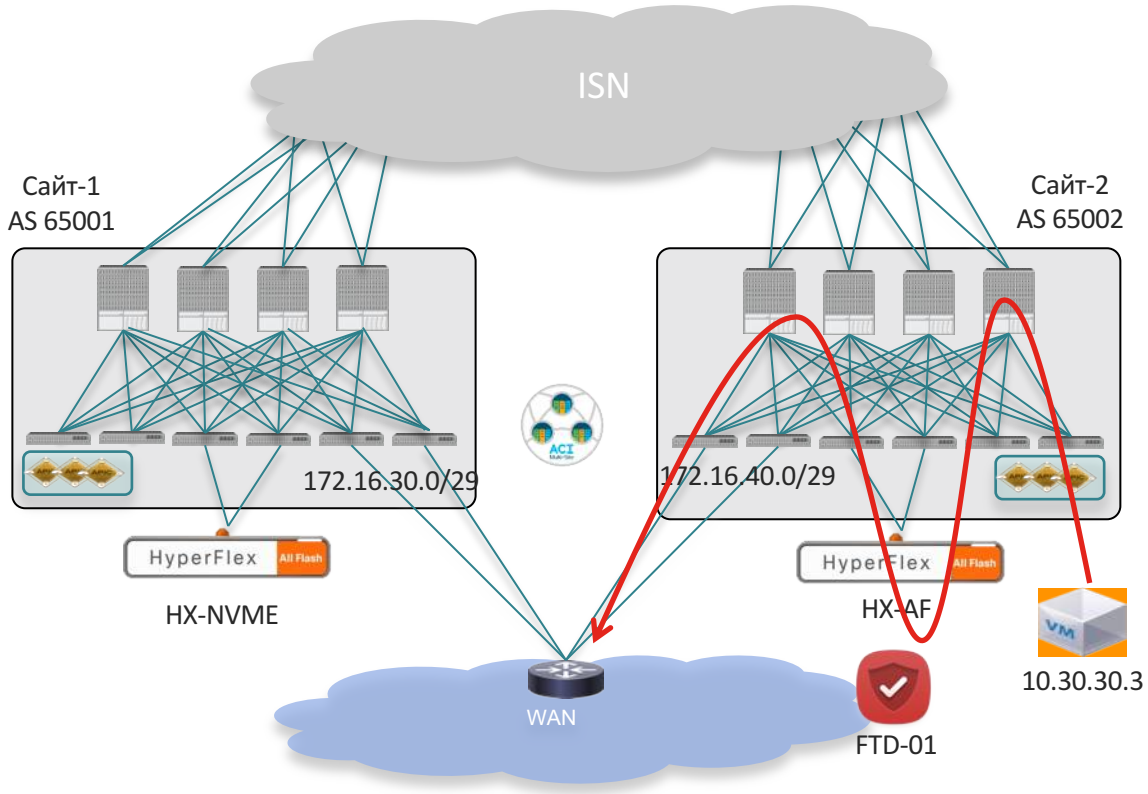
Встраивание МСЭ в путь передачи данных
Использование PBR

Настроено в tenant dr

Вписывание FTD-01
в путь передачи данных на Сайте-2

0 Политики на APIC и MSO настроены заранее

1 Включаем PBR при помощи одной операции

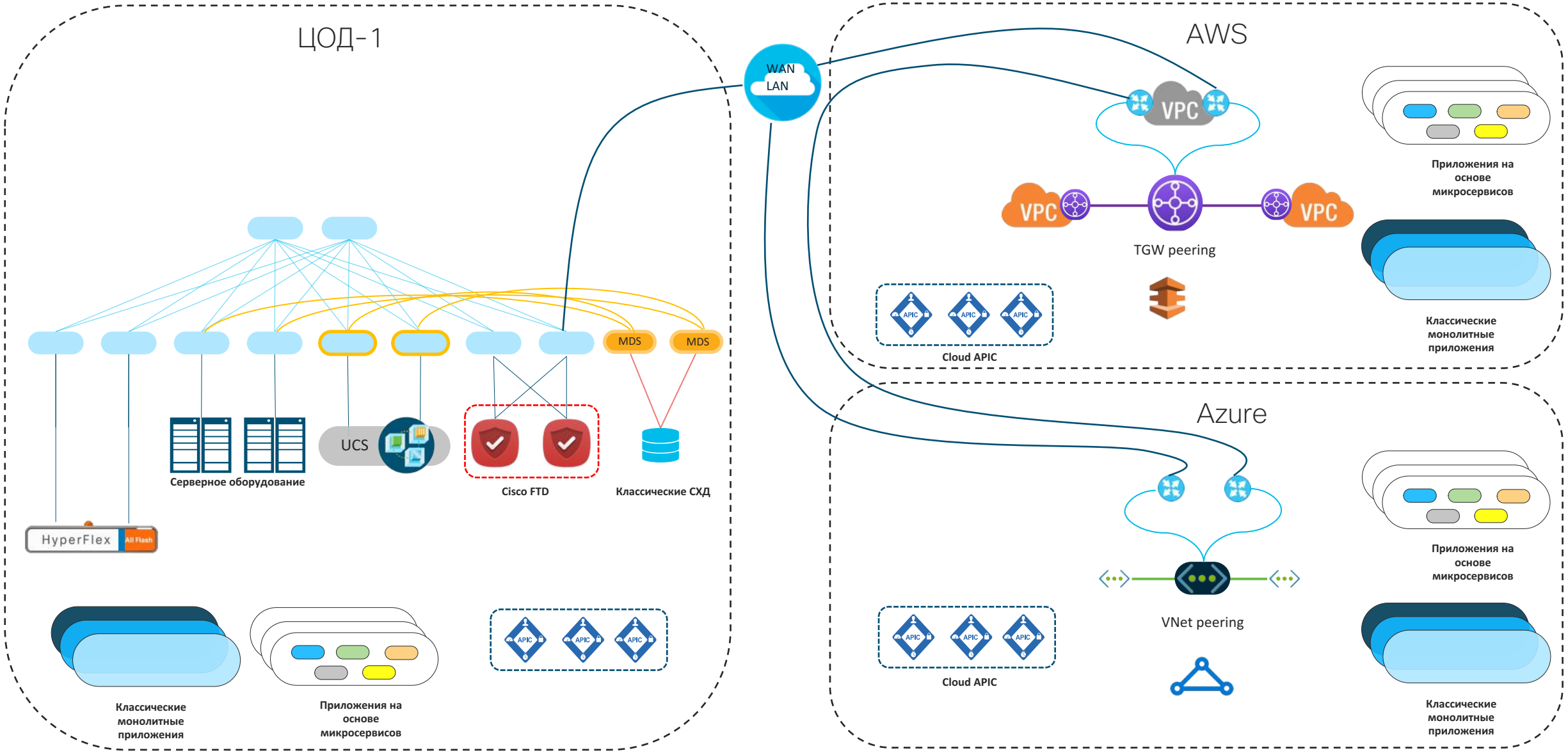


Разделяемое WAN-устройство

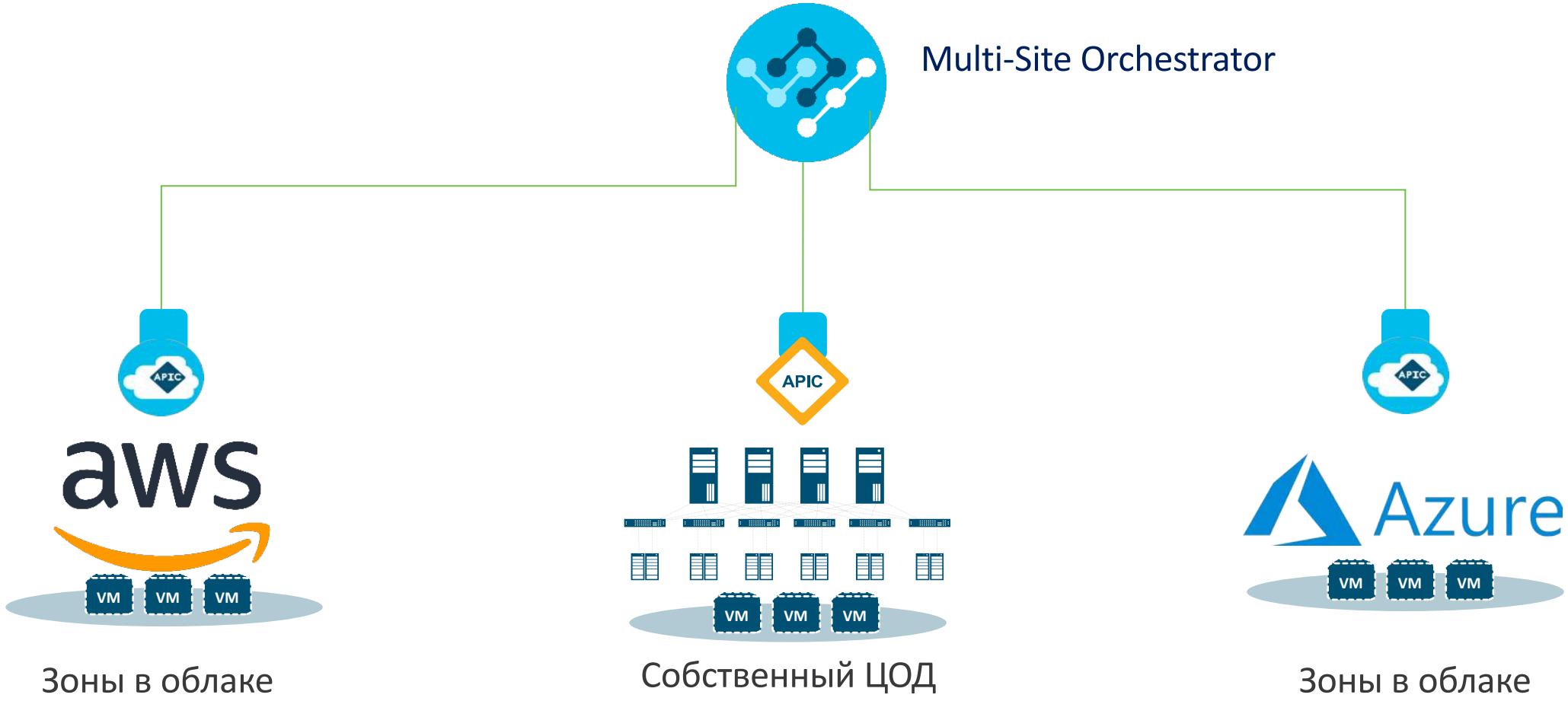
Сценарий № 3 - Гибридное облако

Стратегия использования публичного облака
Уникальная для каждой организации

Стратегия подключения публичных облаков



ACI Extensions для Multi-Cloud



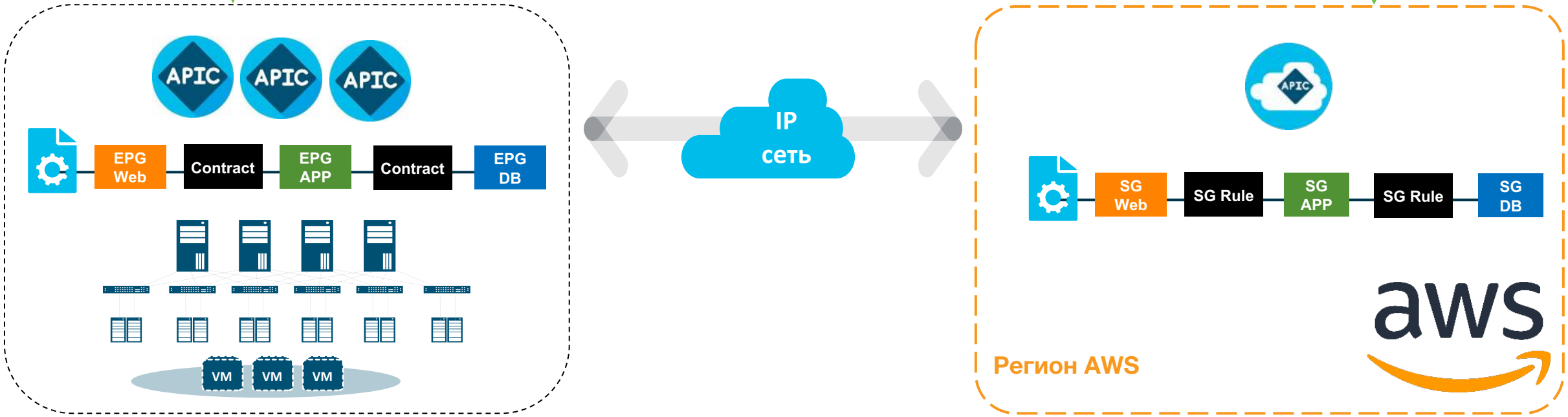
ACI 4.1

ACI Extension для AWS

Multi-Site

Собственный ЦОД

Публичное облако

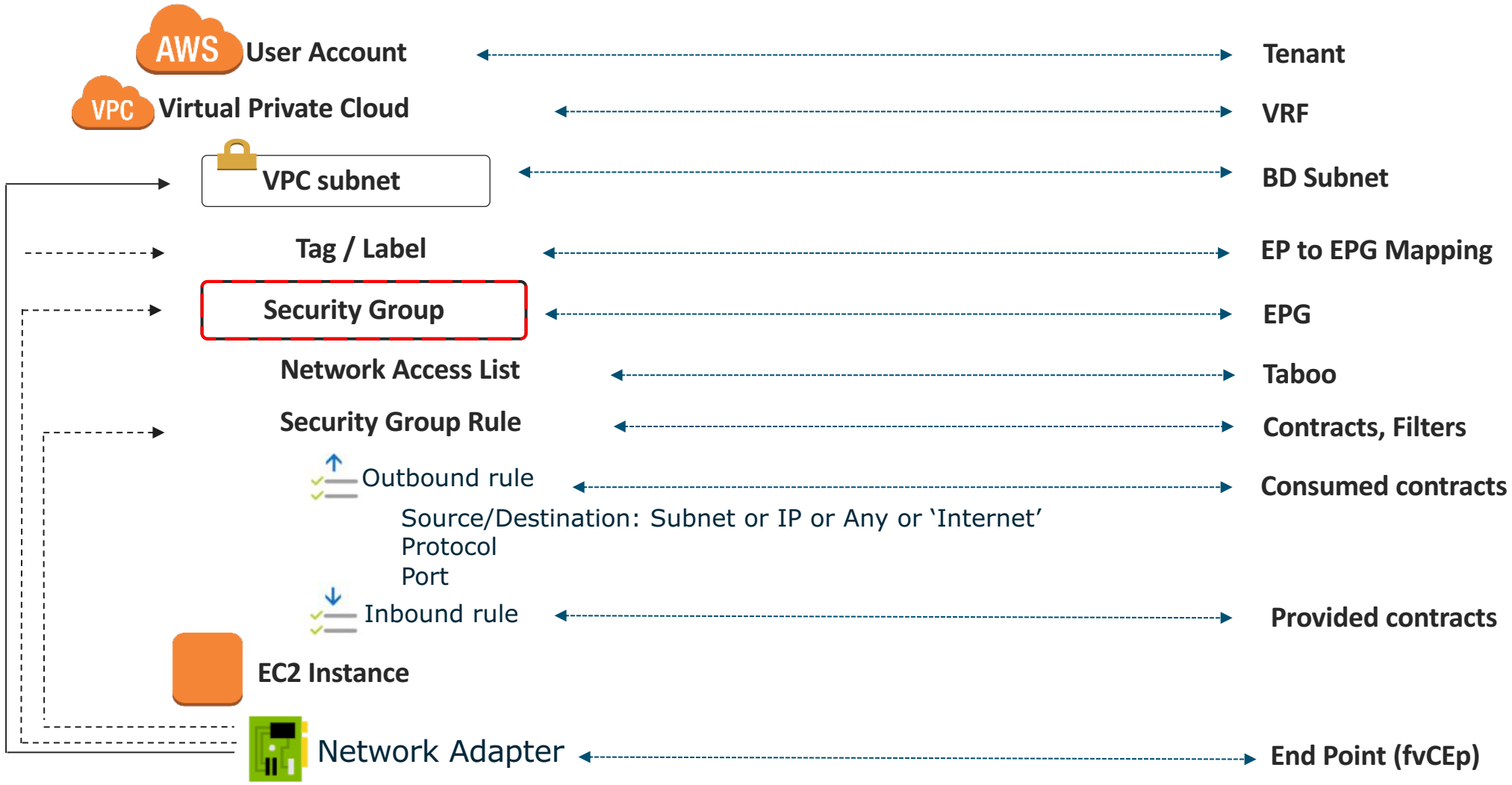


Консистентные политики для собственного ЦОД и облака

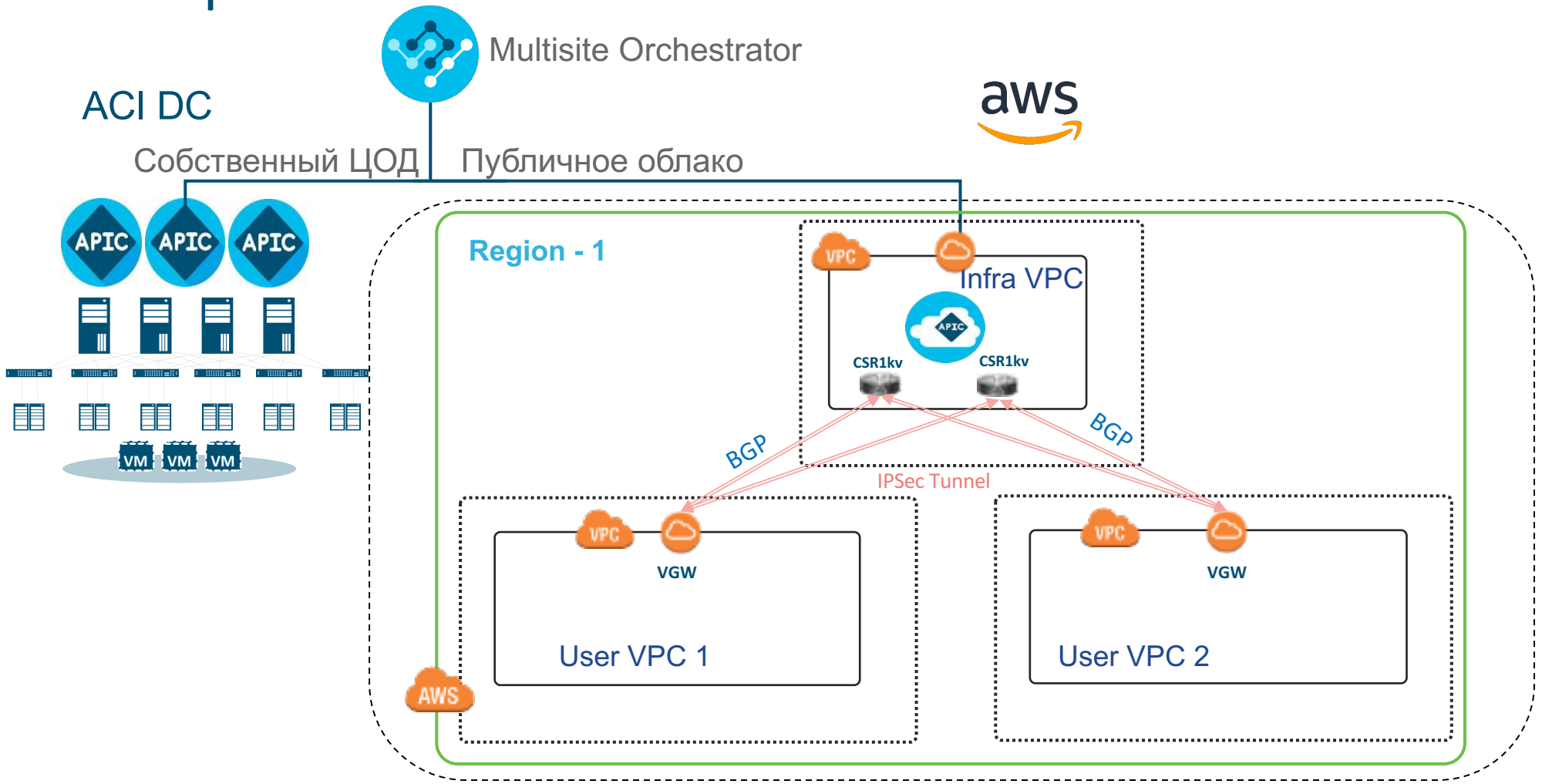
Автоматизация настройки взаимодействия

Упрощение операций и сквозной мониторинг

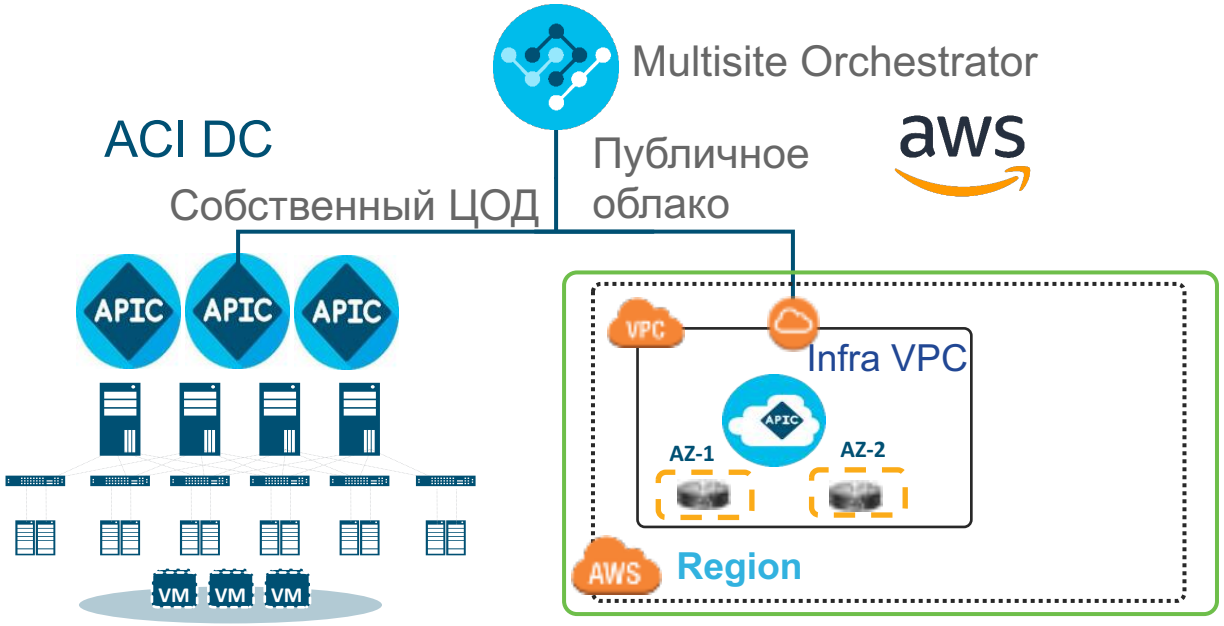
Отображение политик - AWS



Организация сетевой связанности



Организация сетевой связанности Infra VPC

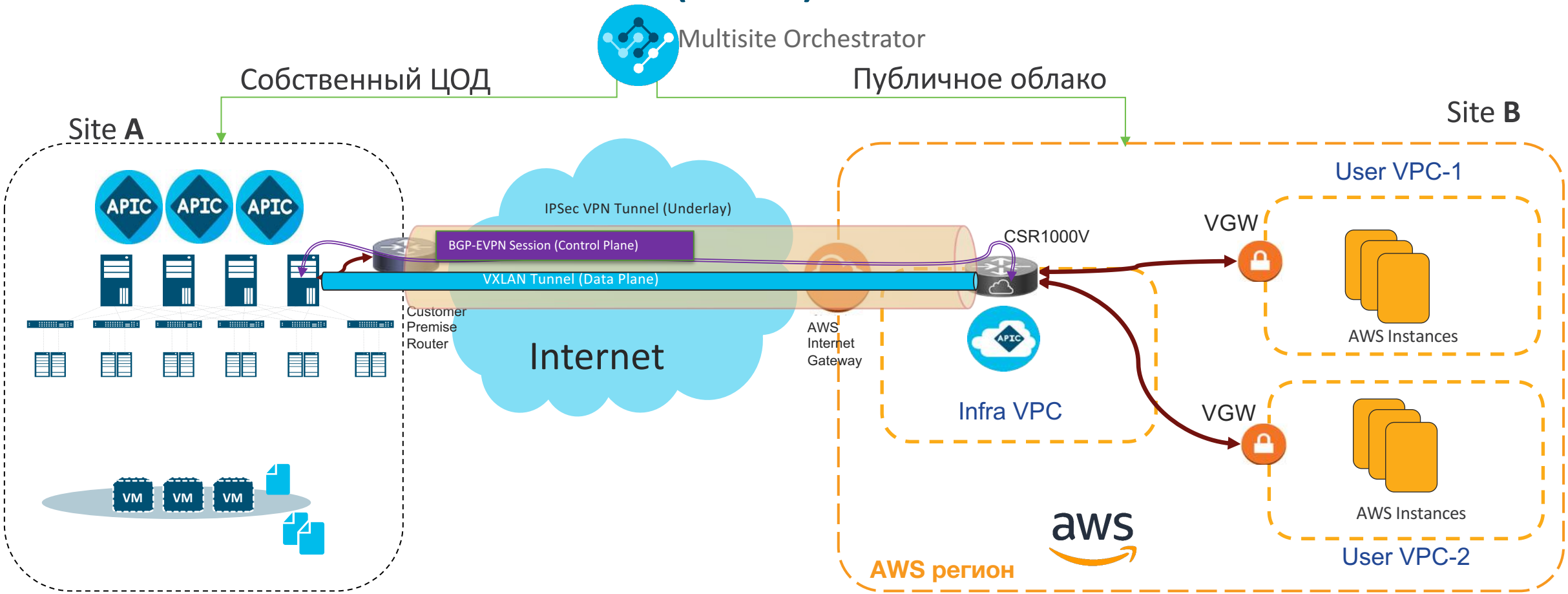


- Infra VPC используется в качестве транзита для подключения физической ACI фабрики к облаку AWS в определенном регионе и для объединения различных регионов AWS друг с другом
- Требуется эккаунт в AWS с правами администратора
- Необходимо выделить IP адреса для организации связанности перед началом настройки

Организация сетевой связанности User VPC

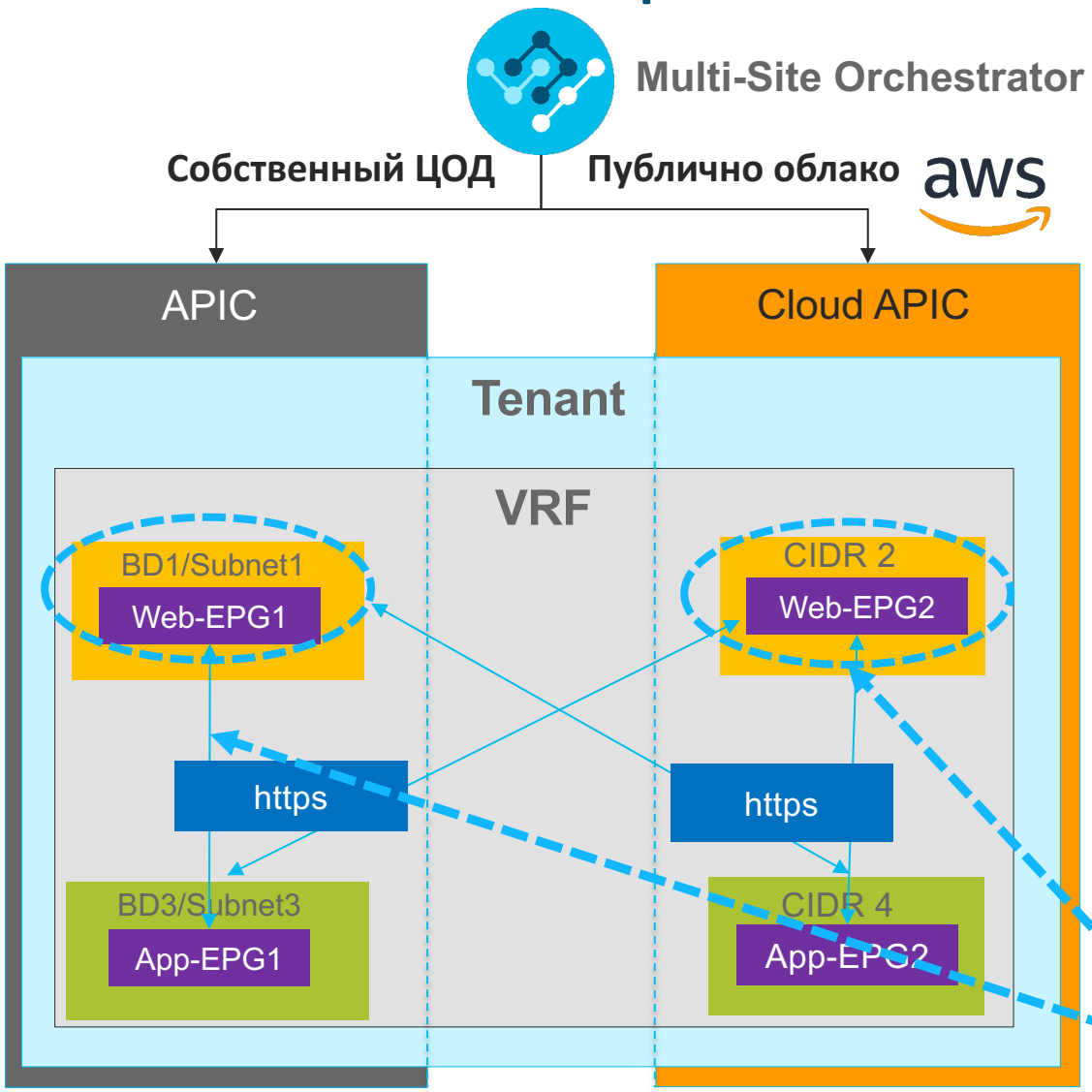
- User VPC создается при помощи Cloud APIC, внутри него трафик между компонентами приложения будет контролироваться
- Требуется эккаунт AWS, который будет использоваться как Tenant admin
- IP подсеть внутри User VPC должна быть уникальна
- User VPC обменивается данными с другим User VPC через Infra VPC

Virtual Private Network (VPN)



- Организация передачи данных при помощи VXLAN
- Обеспечение связанности при помощи BGP-EVPN
- Безопасное IPSec VPN соединение между маршрутизатором подключенном к ACI фабрике и CSR1kv

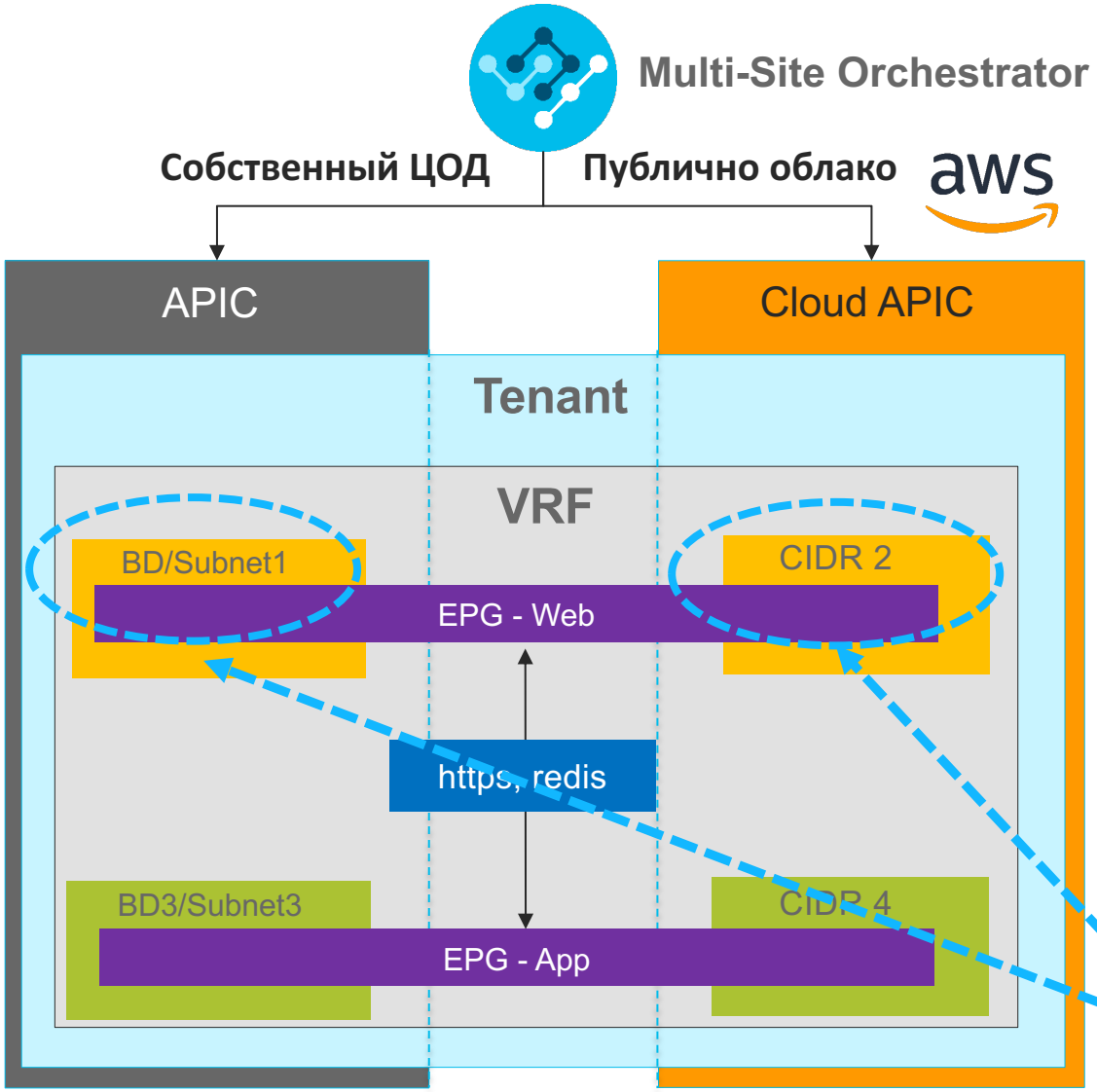
Растягивание приложений



- Растягивание tenant/vrf между собственным ЦОД и облаком
- Во время наибольшей нагрузки на приложение в облаке развертываются дополнительные ресурсы
- Консистентность политик фильтрации между собственным ЦОД и облаком
- Возможность резервирования между сайтами (active/disaster recovery)

Разные подсети, L2 не растягивается

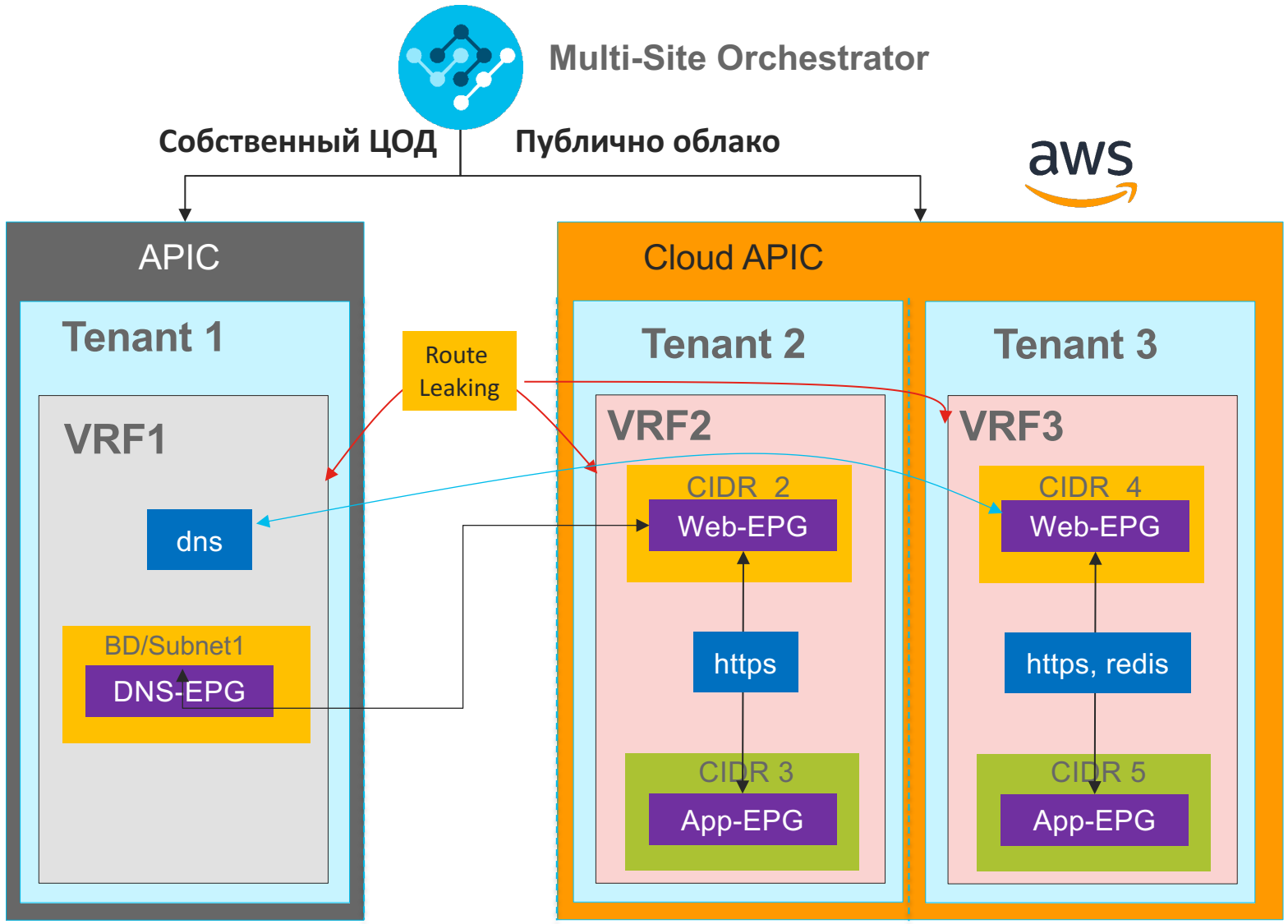
Растянутый EPG с консистентной сегментацией



- Web Tier и App Tier растянуты и для них реализована безопасная изоляция в рамках собственного ЦОД и облака
- Политика сегментации для ресурсов, подключенных к Web/App Tier независимо от места подключения

Разные подсети, L2 не растягивается

Разделяемые сервисы для Hybrid-Cloud



- Возможность использовать разделяемые сервисы между разными облаками
- Разделяемый сервис разворачивается в собственном ЦОД
- Организуется сетевая связанность (route leaking) между VRF


CISCO