Сеть ЦОД на основе VXLAN/EVPN за 5 дней

День 1:
VXLAN/EVPN фабрика – основы

Александр Скороходов
Архитектор Cisco по технологиям
Апрель 2021

Программа спринта

День	Тема
5 апреля, понедельник	VXLAN/EVPN фабрика – основы
6 апреля вторник	VXLAN/EVPN фабрика – клиенты и внешние сегменты
7 апреля среда	VXLAN/EVPN фабрика – распределенные топологии
8 апреля четверг	VXLAN/EVPN фабрика – интеграция L4-L7 сервисов
9 апреля пятница	VXLAN/EVPN фабрика – эксплуатация и поддержка

Программа на сегодня

- Оверлеи, их преимущества и классификация
- VXLAN и его развитие
- Архитектура VXLAN фабрики с уровнем управления MP-BGP EVPN
- L2 и L3 сервисы в VXLAN/EVPN
- Целостное решение: физическая и логическая архитектура
- Дизайн MP-BGP
- Дизайн Underlay связности
- Управление VXLAN/EVPN фабрикой

Предмет обсуждения:
VXLAN фабрика с уровнем управления MP-BGP EVPN

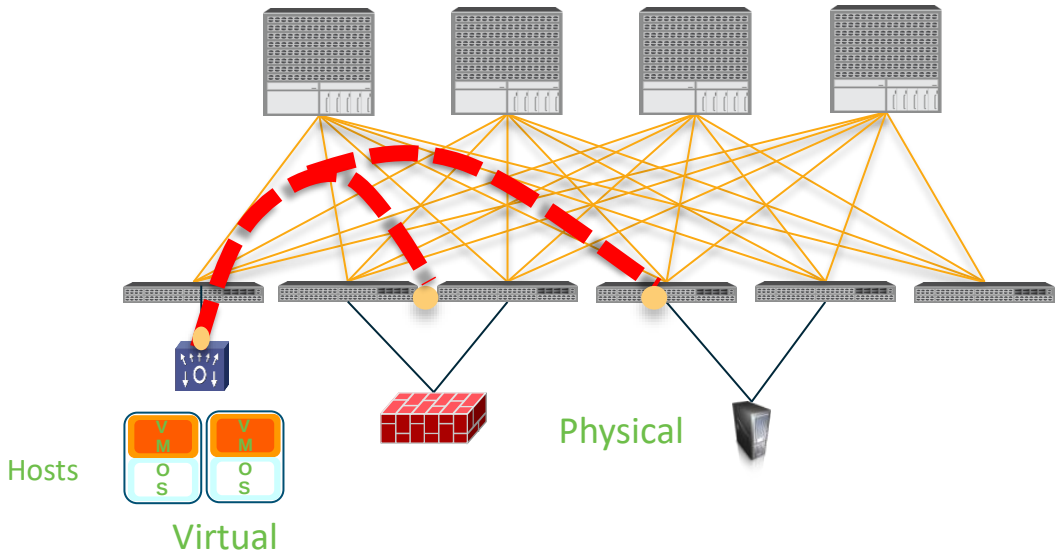
Посмотрим подробнее:

VXLAN	- оверлейная инкапсуляция
фабрика	- сетевая архитектура
с уровнем управления	- способ сигнализации
MP-BGP	- протокол сигнализации
EVPN	- адресное семейство для представления управляющей информации

Оверлеи, их преимущества и классификация

Требования к современным сетям ЦОД

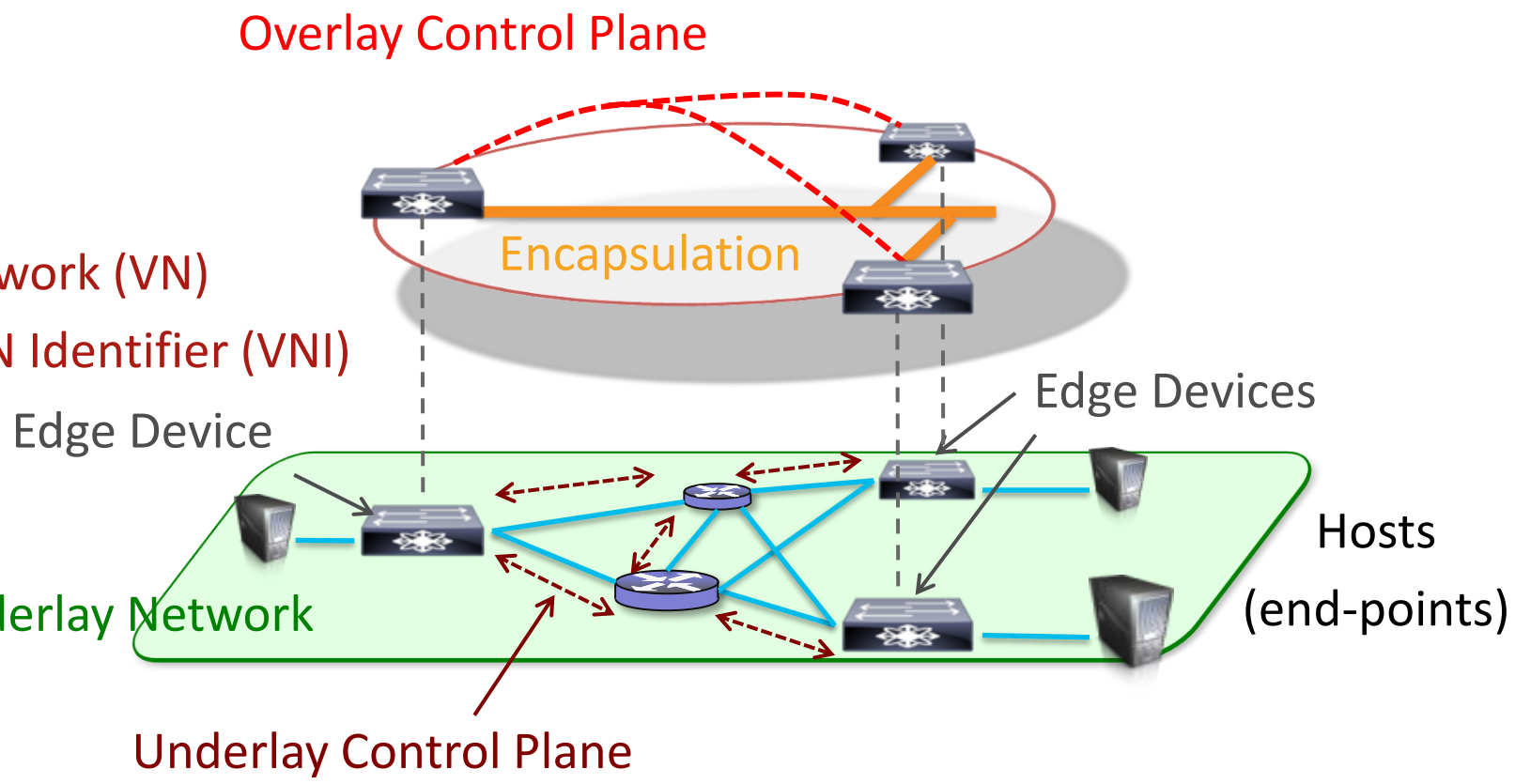
- Мобильность нагрузок
- Гибкость размещения нагрузок
- Сегментация – L2 и L3
- Масштабируемость – производительность и подключения
- Автоматизация и программируемость
- Связность на L2 и L3
- Физические и виртуальные подключения



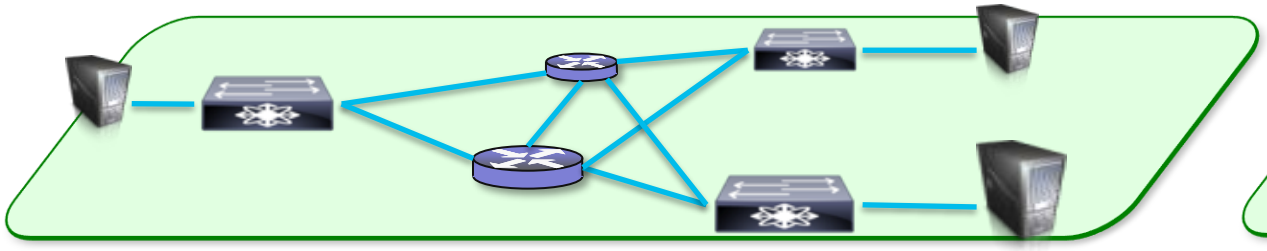
Что такое оверлейная сеть?

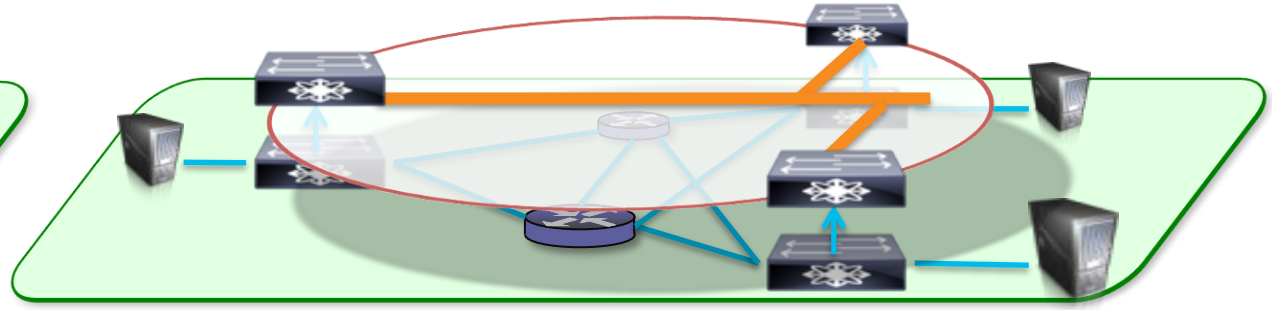
Сервис = Virtual Network (VN)

Идентификатор = VN Identifier (VNI)



Зачем нужны оверлейные сети?





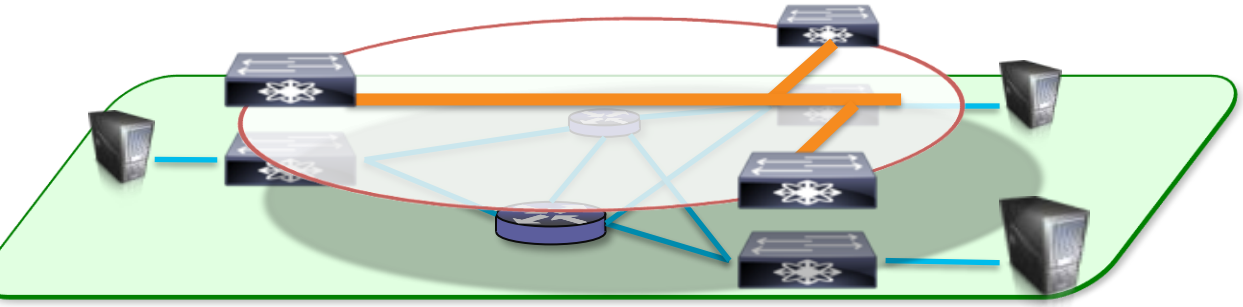
Эффективный транспорт

- Универсальная опорная сеть
- Резервирование и производительность
- Эффективное управление трафиком
- Управление, диагностика, программируемость

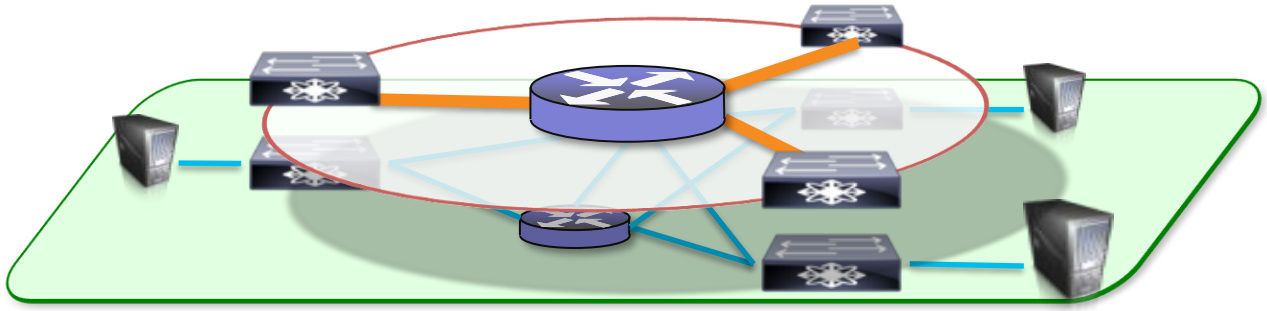
Гибкая наложенная сеть

- Мобильность подключенных адресов
- Нужный тип связности
- Сегментация
- Масштабирование – снижение сложности в ядре
- Гибкость, программируемость
- Передача дополнительных метаданных

Типы наложенных сетей: вид сервиса



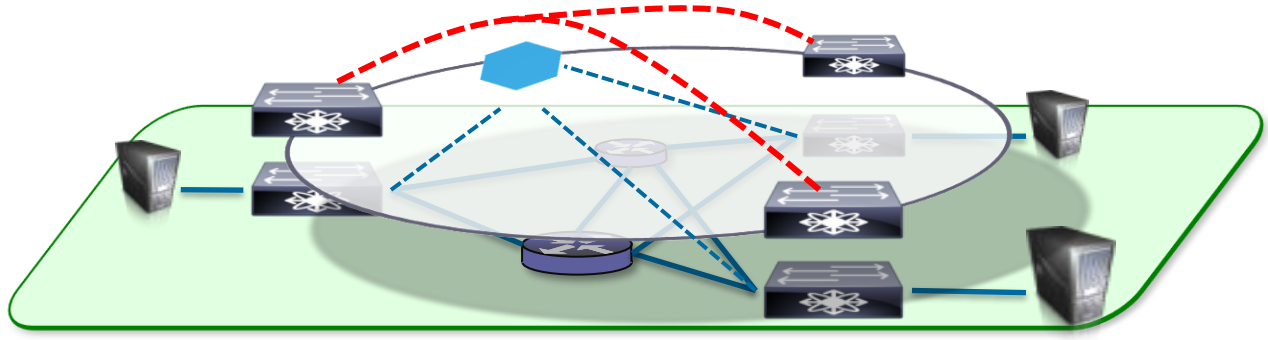
- **Layer 2**
- Эмуляция сегмента LAN
- Передача кадров Ethernet (IP и не-IP)
- Мобильность в подсети (L2 домене)
- Гибкость размещения нагрузок
- Поддержка (и риск!) L2 фладинга



- **Layer 3**
- Абстракция связности на основе IP
- Передача IP пакетов
- Мобильность адреса без растягивания L2
- Ограничение доменов сбоя

Сигнализация

- Обнаружение сервисов
  - Пограничные устройства должны обнаружить друг друга
- Анонс адресов и построение отображения
  - Соотнесение подключений с пограничными устройствами
- Управление туннелями
  - Построение и управление соединениями между пограничными устройствами

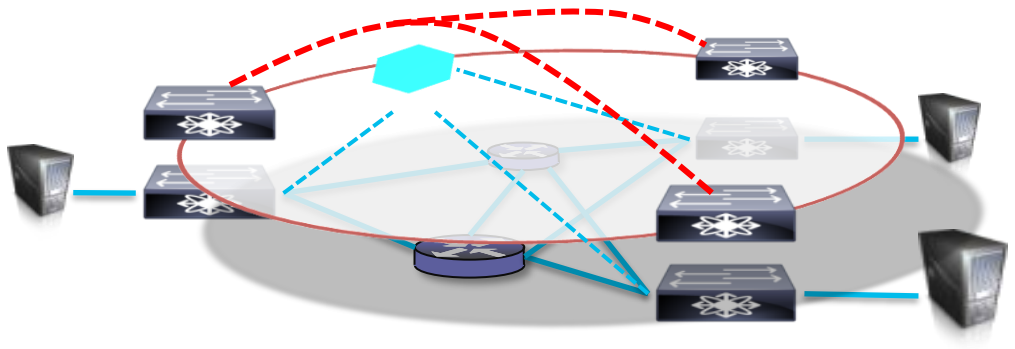


Сигнализация
Data Plane Learning

- Используется сбор информации от уровня коммутации
  - Примеры: «выучивание» адресов на традиционных коммутаторах, VXLAN Flood&Learn, FabricPath
- Обеспечивается:
  - Обнаружение адресов и построение соответствия
  - Возможно – управление туннелями
  - Не обеспечивается обнаружение сервисов
- Требует фладинга информации на уровне коммутации:
  - Multicast дерево
  - Репликация в unicast трафик на входе в оверлей

Сигнализация
Уровень управления

- Обеспечивает:
  - Обнаружение пограничных узлов
  - Анонс подключений
  - Управление туннелями
  - Могут обеспечиваться расширения для резервирования подключений и дополнительных сервисов



Протокол или контроллер:

- **Управляющий протокол** между пограничными узлами
  - BGP, IS-IS, LISP
- Центральная БД на **контроллере**
  - Распределённые виртуальные коммутаторы (OVS, N1Kv/VSM)

«Push» или «Pull»:

- **Распространение (Push)** информации до всех пограничных устройств
  - BGP, IS-IS, контроллеры
- **Запрос (Pull)** и кеширование на пограничном устройстве
  - LISP, DNS, контроллеры

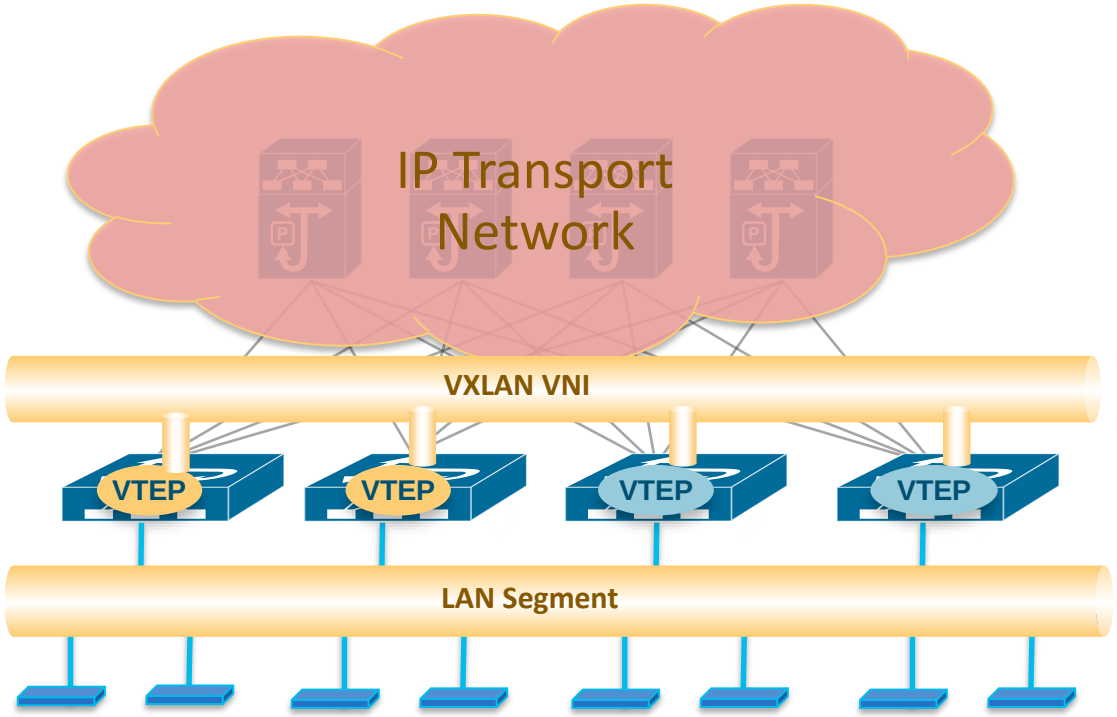
VXLAN и его развитие

Почему VXLAN?

VXLAN обеспечивает сеть с сегментацией, мобильностью и масштабированием

- Используются «стандарты» (RFC)
- Используется IP сеть с Layer-3 ECMP – надёжная отработанная технология
- «Пространство имен» сегментов до 16 миллионов
- Развёртывание сегментов без изменений на промежуточных устройствах
- Поддержка физическими и виртуальными коммутаторами
- Поддержка разными производителями

Оверлейные сети с использованием VXLAN



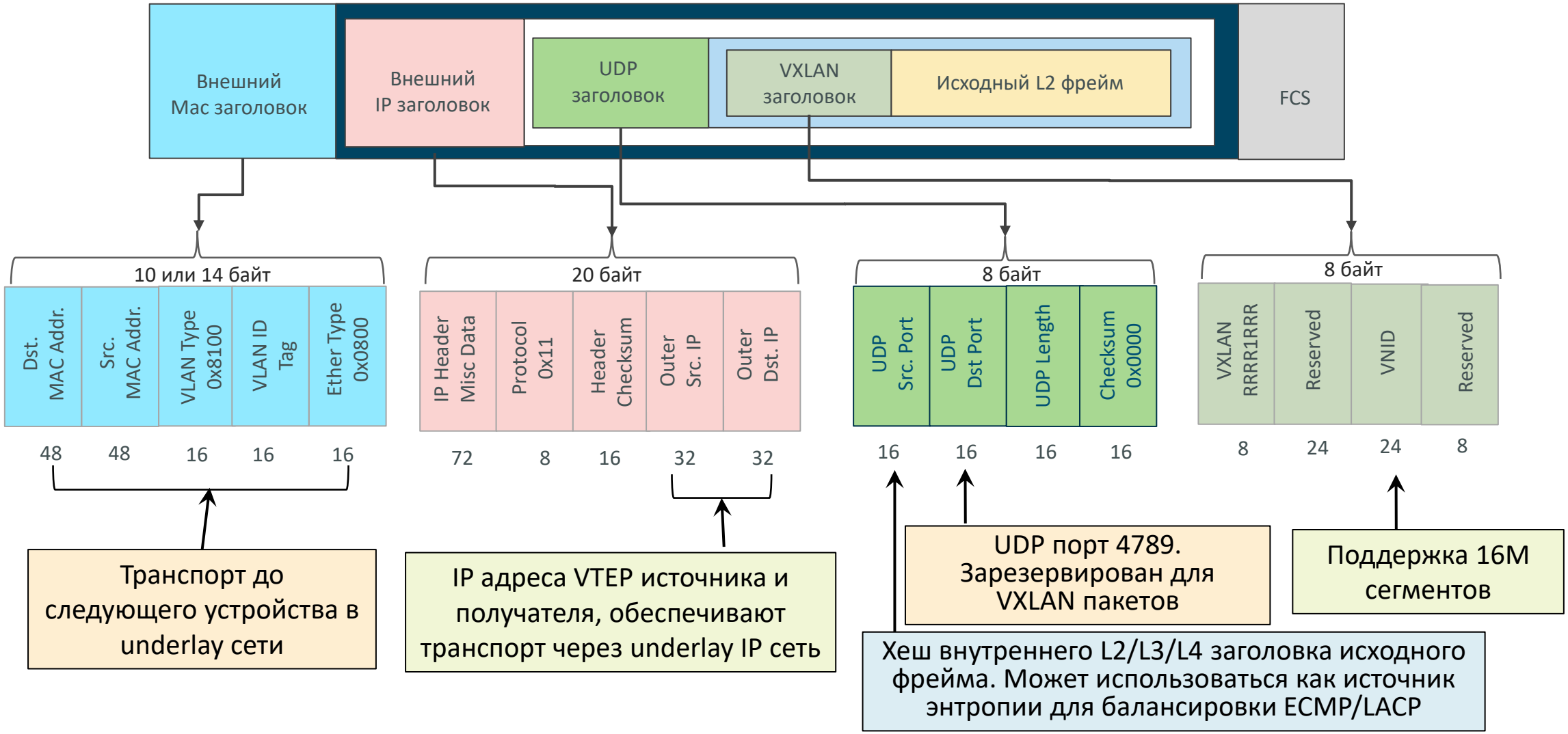
Underlay:

- Маршрутизация IP – проверенная, надёжная, масштабируемая
- ECMP – использование всех путей

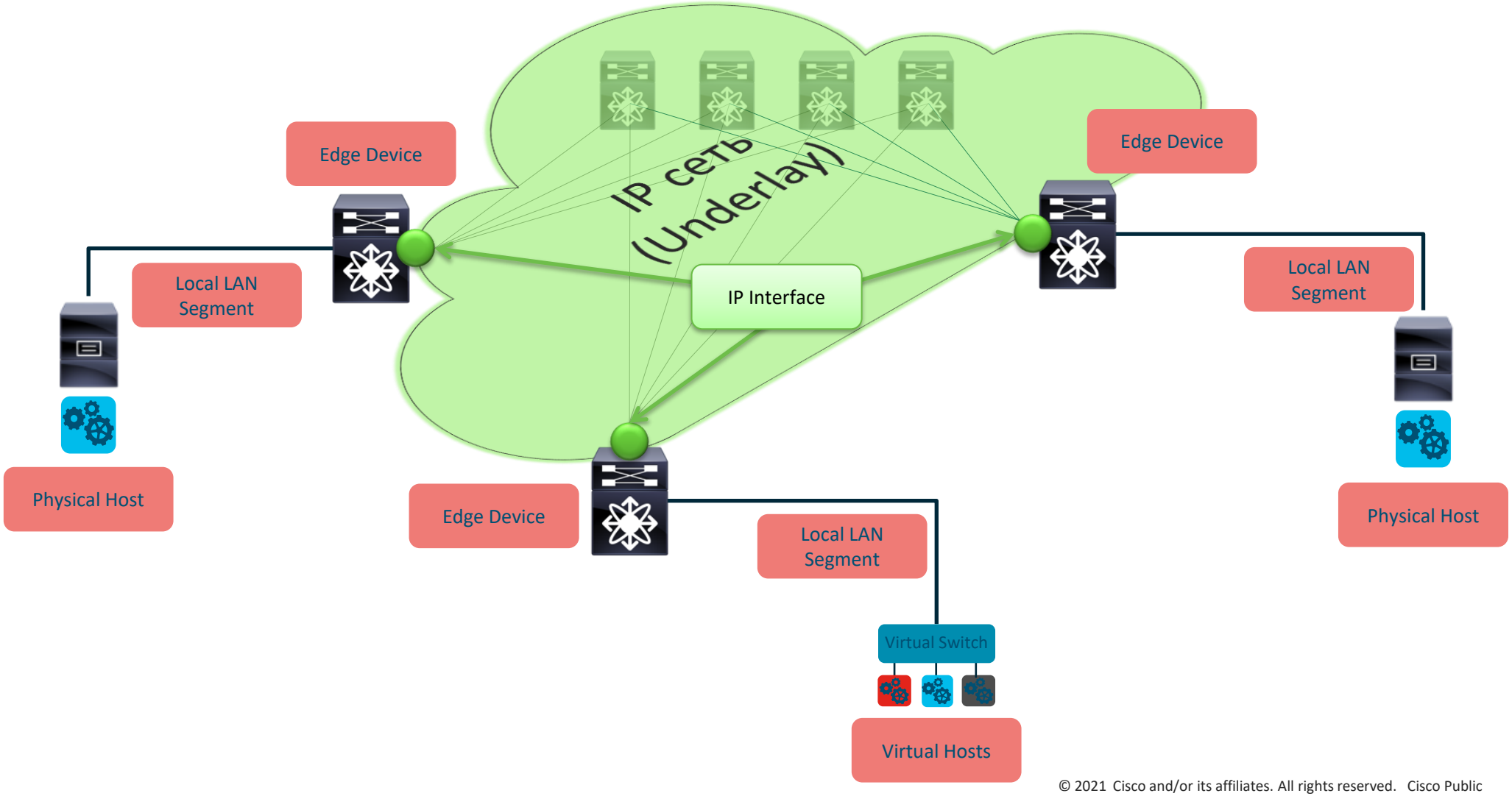
Overlay:

- Опора на стандарты
- Растягивание и мобильность на L2
- Расширенное пространство идентификаторов сегментов
- Масштабируемая сеть
- Multi-Tenancy

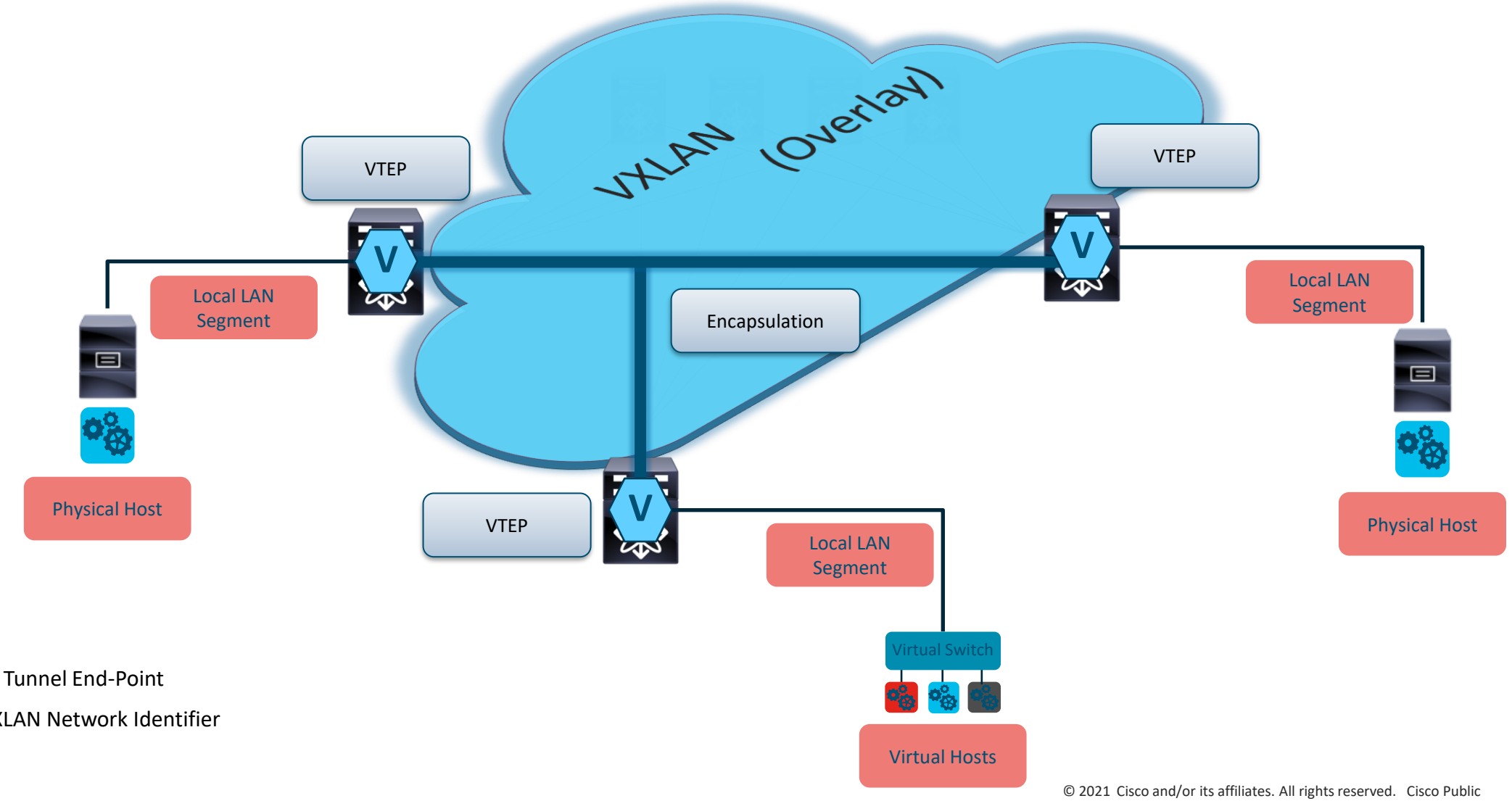
Формат пакета VXLAN
MAC-in-UDP



Функционирование VXLAN



Функционирование VXLAN

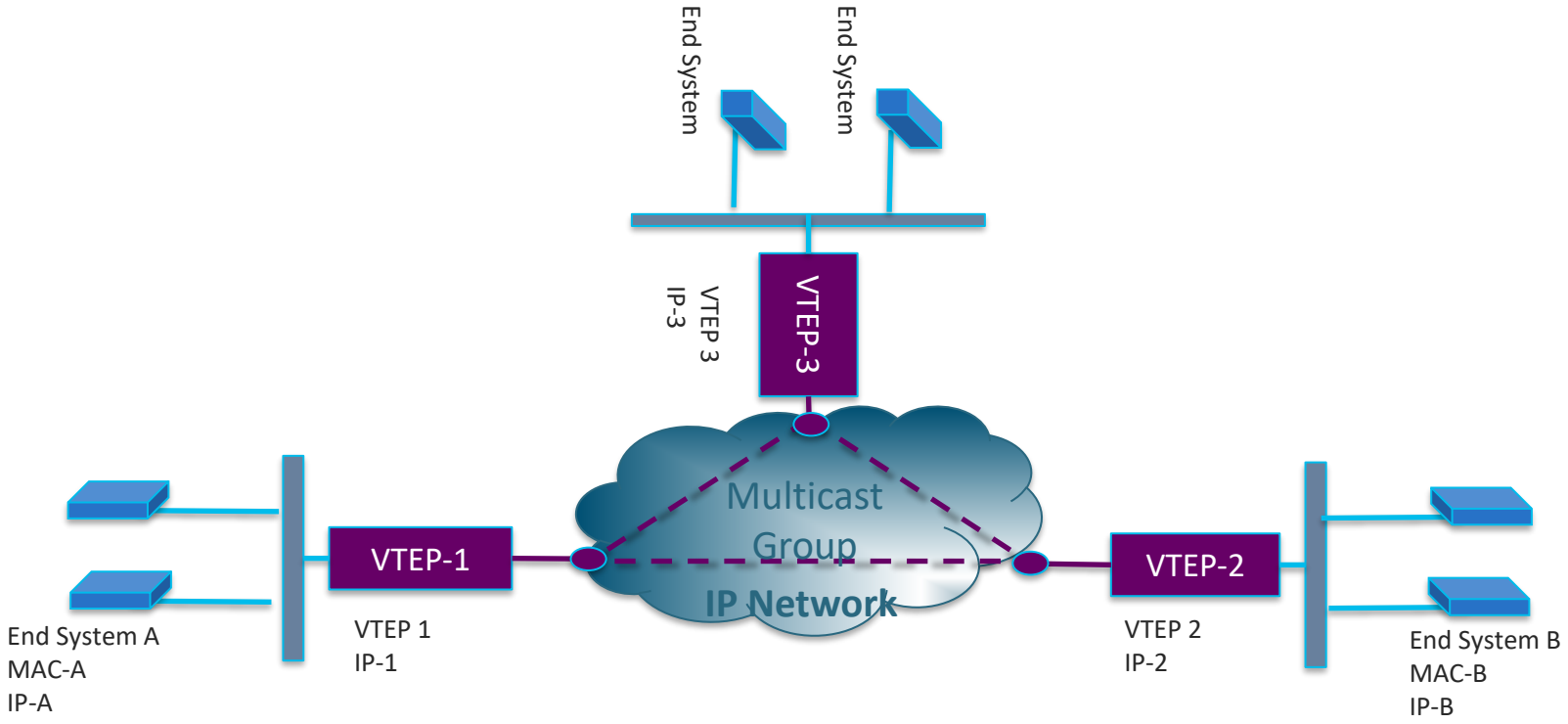


VTEP – VXLAN Tunnel End-Point

VNI/VNID – VXLAN Network Identifier

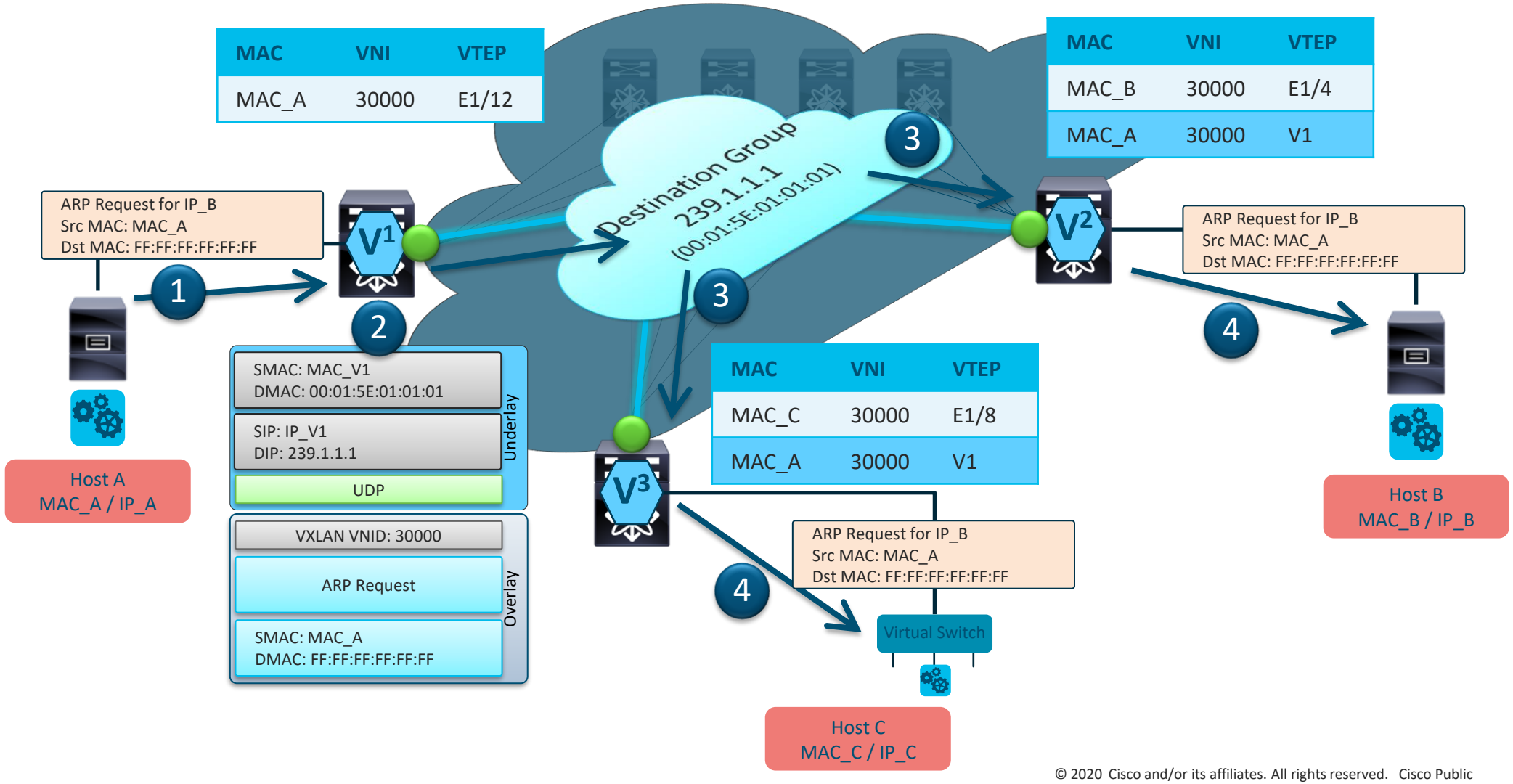
VXLAN Flood&Learn на основе mulicast
Первоначальная реализация VXLAN

- Без уровня управления
- Flood-&-learn с использованием передаваемых данных
- Multicast транспорт для BUM трафика (Broadcast, Unknown Unicast и Multicast).



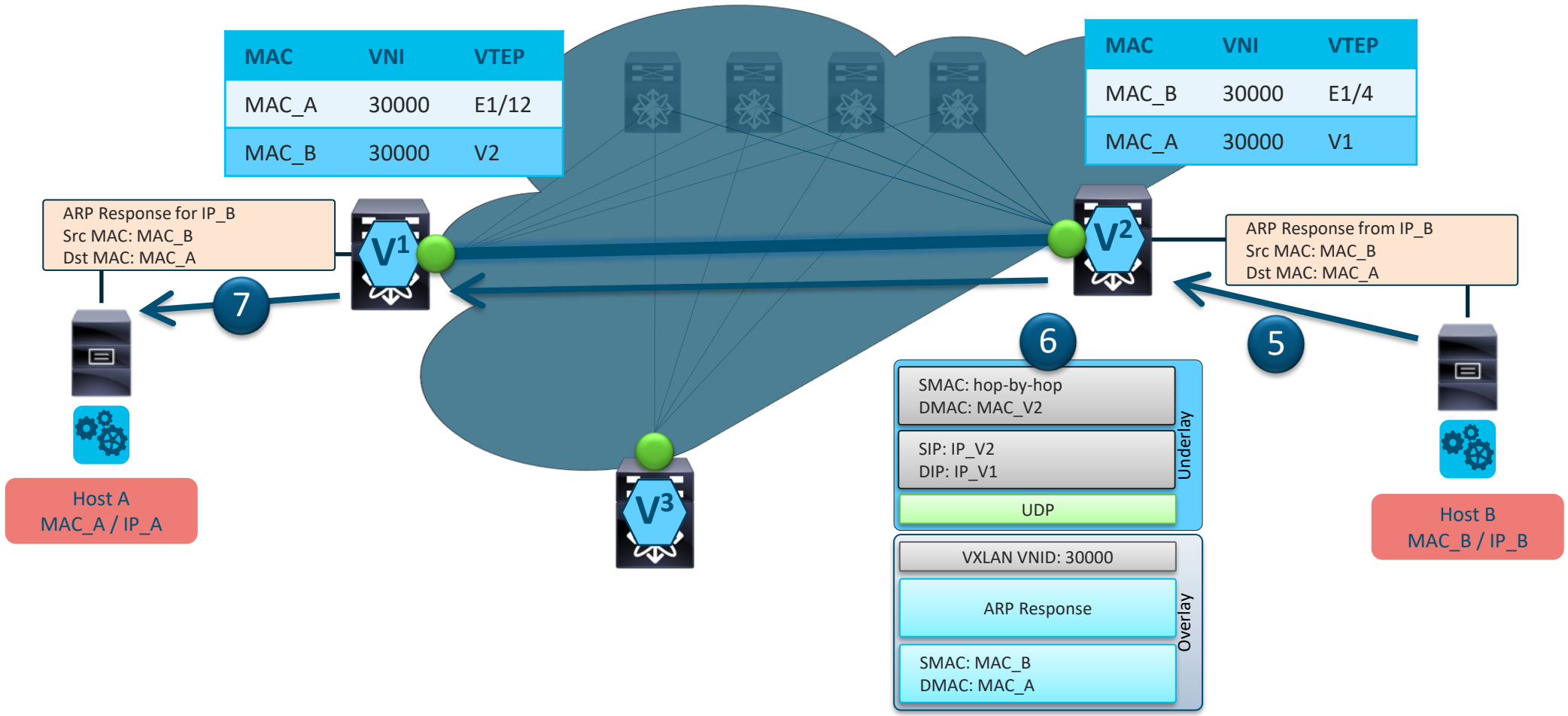
VTEP: выучивание адресов

VXLAN Flood&Learn



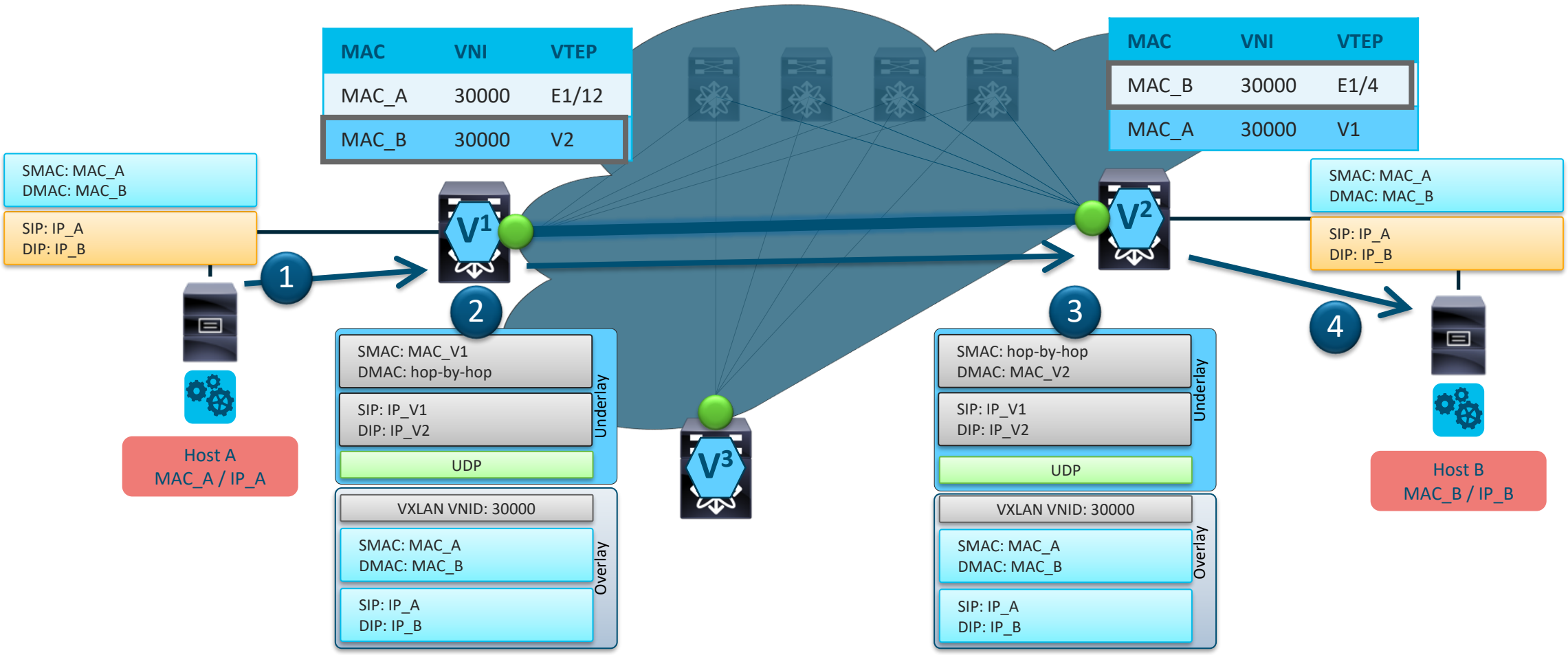
VTEP: выучивание адресов

VXLAN Flood&Learn



Передача данных

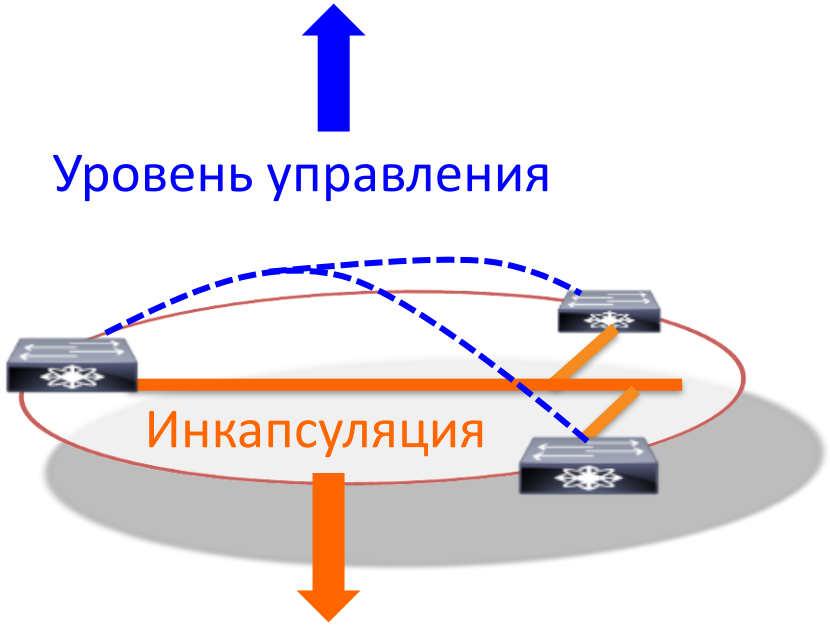
VXLAN Flood&Learn



VXLAN транспорт
И его развитие

Выучивание на уровне коммутации

- Flood & Learn в мультикастовой группе, в которой участвуют все устройства
- Уровень управления отсутствует



Выучивание на уровне протокола

- Оконечные узлы анонсируются управляющим протоколом между VTEP с помощью MP-BGP/EVPN

Варианты распространения BUM трафика

- Multicast или репликация на источнике через unicast

Использование для связи ЦОД

- VXLAN/EVPN Multi-Site

Развитие VXLAN

Независимость от multicast	Внешние подключения	Уровень управления	IP сервисы
• Репликация на входе (Ingress Replication) позволяет использовать unicast транспорт – если нужно! • Уровень управления позволяет автоматически обнаруживать адреса VTEP	• Резервирование подключения к VTEP (VPC/MLAG) • VXLAN шлюзы во внешние сети – MPLS VPN, Segment Routing (SR-MPLS, SRv6), … • «Сшивание» VXLAN туннелей для связи ЦОД: VXLAN Multi-Site	• MAC и IP адреса нагрзок выучиваются VTEP • Анонс ассоциации L2 и L3 адресов с VTEP с использованием управляющего протокола • Сокращение фладинга • Оптимальное распространение ARP	• Маршрутизация между VXLAN • Распределённый шлюз по умолчанию • Изоляция организаций (Multi-Tenancy) на L2 и L3

Архитектура VXLAN фабрики с уровнем управления MP-BGP EVPN

Архитектура VXLAN EVPN фабрики
Базовые составляющие решения

Оверлеи
- Layer-2
- Layer-3
- Layer-2 и Layer-3

VXLAN инкапсуляция

Underlay транспортная сеть

Плоскость управления на основе BGP
- Peer-Discovery
- Route Learning and Distribution
  - Local Learning
  - Remote Learning

Плоскость передачи данных
- Overlay Layer-2/Layer-3 Unicast Traffic
- Overlay Broadcast, Unknown Unicast, Multicast traffic (BUM traffic) forwarding
  - Ingress Replication (Unicast)
  - Multicast

Основные элементы

VXLAN

- Инкапсуляция на основе стандарта RFC 7348
- Инкапсуляция в UDP
- Независимость от транспортной сети
- Layer-3 транспорт (Underlay)
- Расширенное «пространство имён»
- 24-битное поле (VNID) - ~16M идентификаторов
- Allows Segmentations

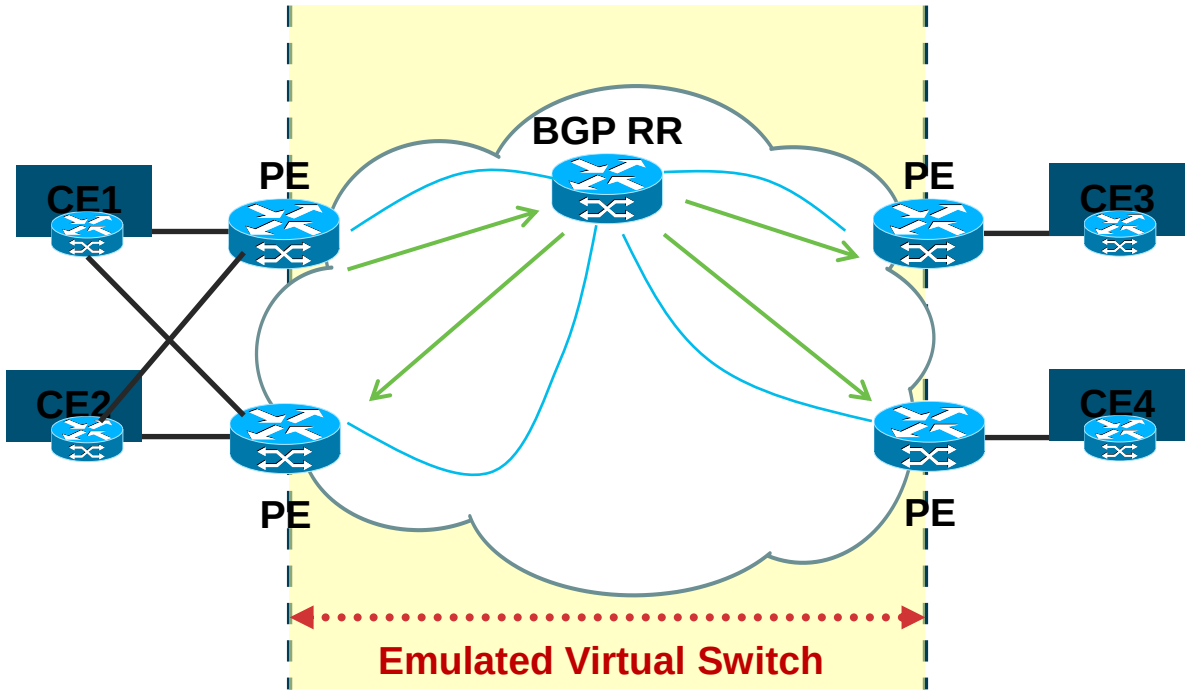
EVPN

- Уровень управления на основе стандартов
- RFC 8365 (и RFC 7432)
- Использует Multiprotocol BGP
- Поддерживает разные виды транспорта
- VXLAN (EVPN-Overlay), MPLS, Provider Backbone (PBB)
- Много поддерживаемых сценариев
- Bridging, MAC Mobility, First-Hop & Prefix Routing, Multi-Tenancy (VPN)

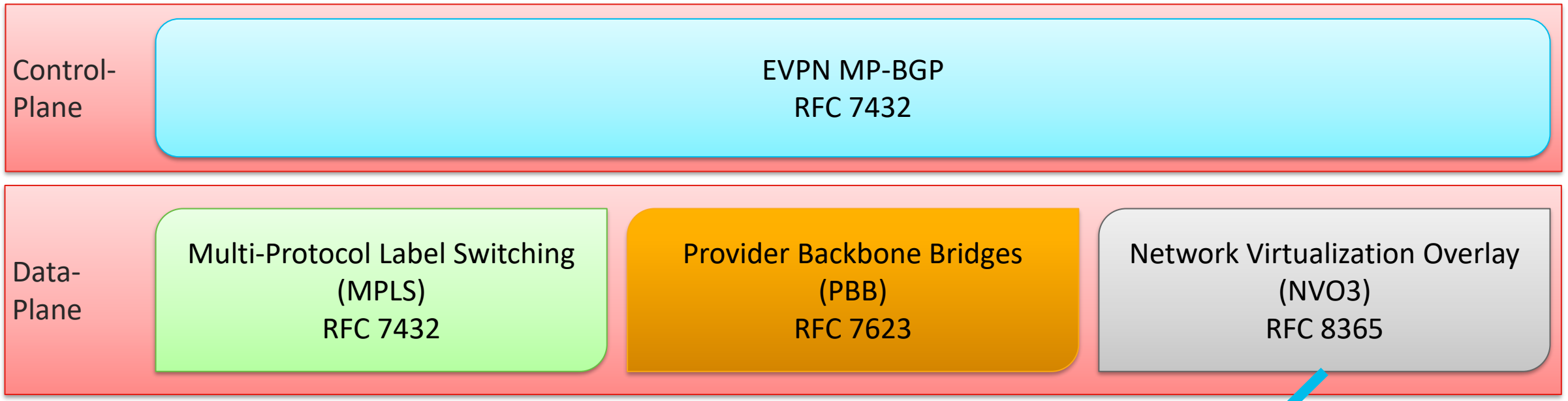
EVPN (Ethernet VPN)

- Уровень управления на основе Multi-Protocol BGP (MP-BGP) с использованием EVPN NLRI (Network Layer Reachability Information)
- Выполнение коммутации на VTEPs для Layer-2 (MAC) и Layer-3 (IP)
- Обнаружение: BGP, используя механизмы MPLS VPN (RT)
- Сигнализация: BGP
- Выучивание хостов: Control plane (BGP)

BGP advertisement:
L2VPN/EVPN Addr = **CE1.MAC**
BGP Next-Hop = PE1
Route Target = 100:1
Label=42



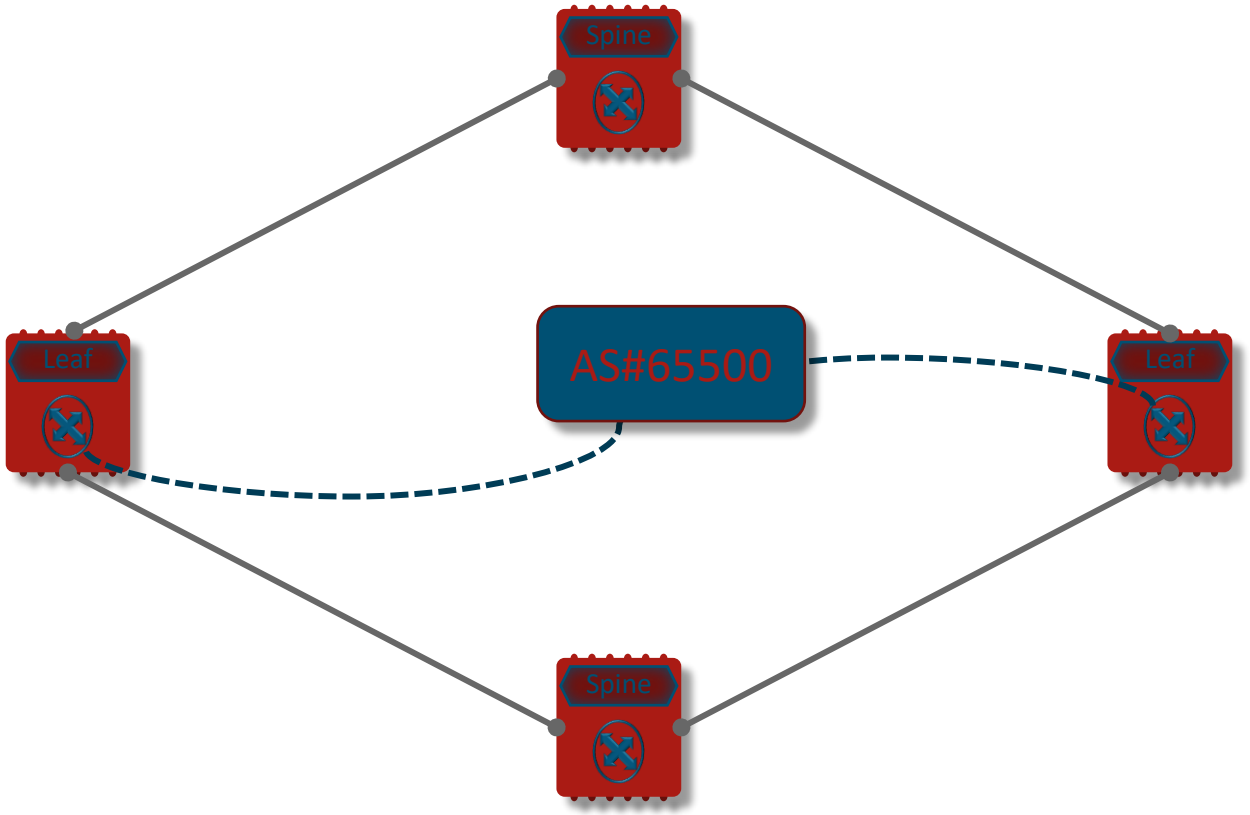
EVPN – Ethernet VPN



- EVPN с использованием NVO туннелей (VXLAN, NVGRE, MPLSoGRE) для сетевых фабрик ЦОД
- Построение L2 и L3 оверлеев поверх IP сетей
- Поддержка различными производителями

Multiprotocol BGP (MP-BGP)

Универсальный транспорт для распространения информации хостах и сетях подключенных к фабрике



- Расширение для Border Gateway Protocol (BGP)
  - RFC 4760
- Multiprotocol BGP (MP-BGP)
- Адресные семейства
  - В рамках одной BGP сессии может переносится информация различных Address-Families (например VPNv4/v6, MVPN, L2VPN, EVPN)

Важные элементы реализации MP-BGP
Параметры VRF

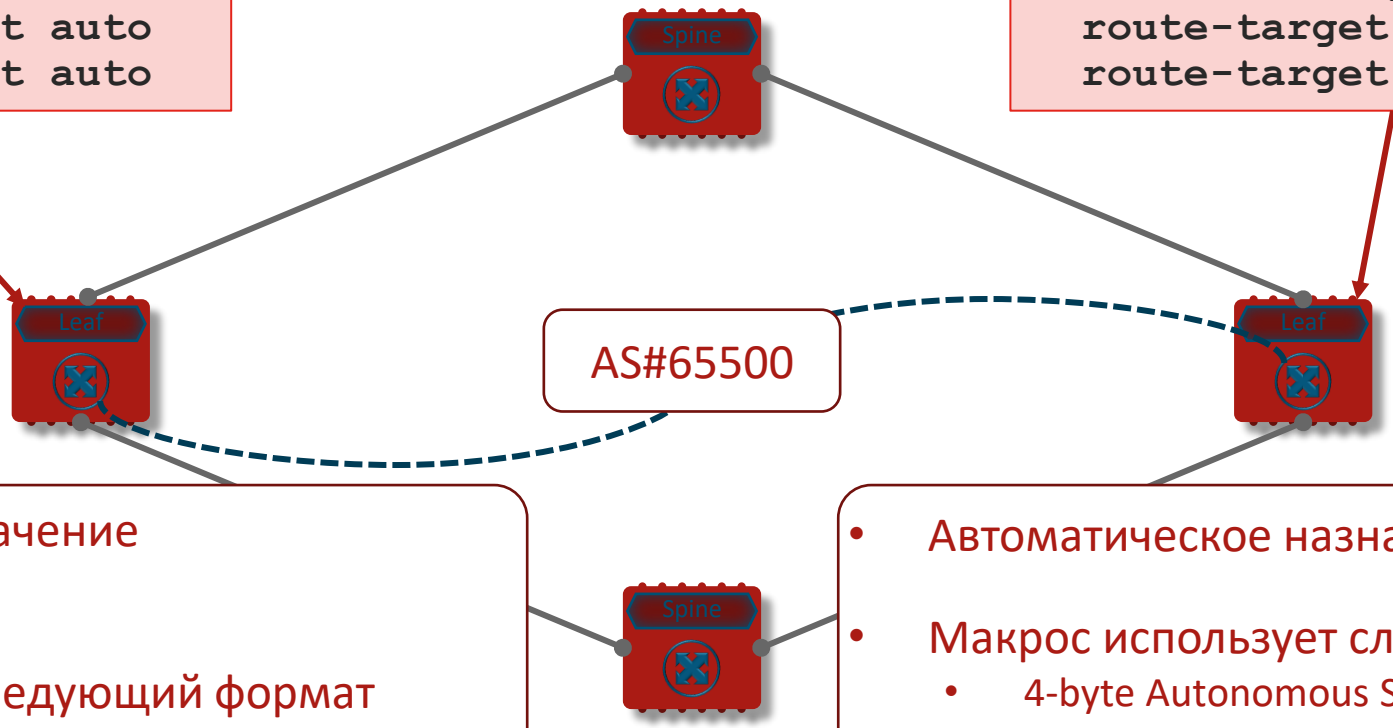
- **Route Distinguisher (RD):**
  - 8-байтное поле, параметр VRF. Делает префиксы уникальными между клиентами/VRF
  - **VPNv4 address**: RD + VPN IP префикс

- **Route Target (RT):**
  - 8-байтное поле, параметр VRF. Уникальные значения для определения правил импорта/экспорта маршрутов VPNv4
  - Анонсируется MP-BGP вместе с префиксами

Multiprotocol BGP (MP-BGP)
Автоматизация для RD, RT

```
vrf context A
 rd auto
 address-family ipv4 unicast
  route-target import auto
  route-target export auto
```

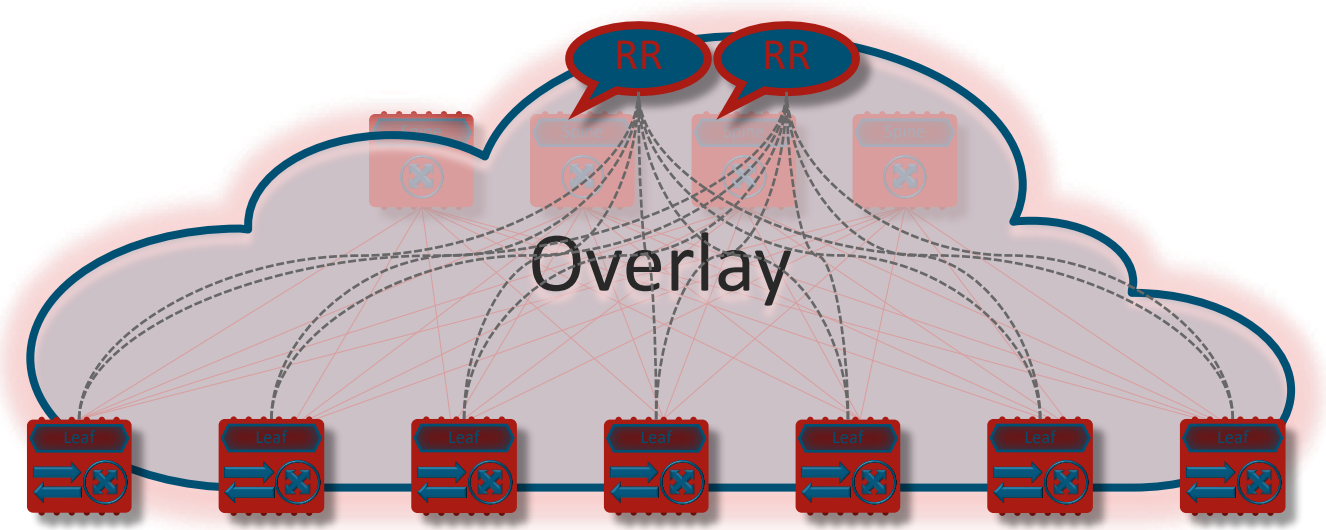
```
vrf context A
 rd auto
 address-family ipv4 unicast
  route-target import auto
  route-target export auto
```



- Автоматическое назначение Route Distinguisher
- Макрос использует следующий формат
  - 4-byte Router ID (RID)
  - 4-byte VRF ID (internal number)
  - Пример для RD:
    - 10.10.10.101:3

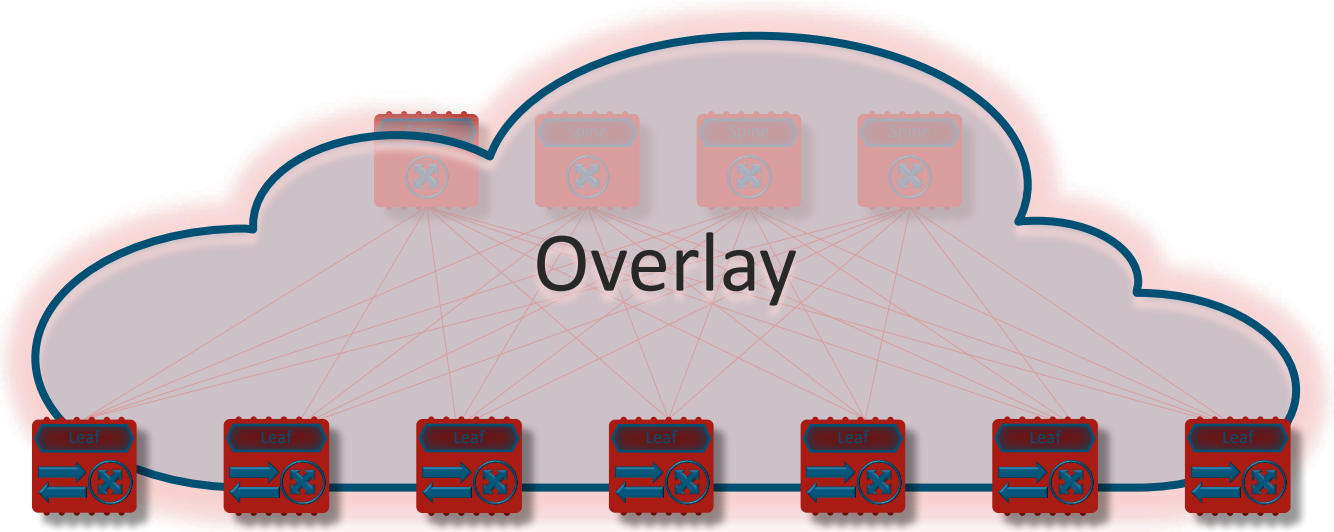
- Автоматическое назначение Route Target
- Макрос использует следующий формат
  - 4-byte Autonomous System
  - 4-byte VNI
  - Пример для RT:
    - Import, Export или Both
    - 65500:5000

VXLAN/EVPN – распространение информации о хостах и сетях при помощи MP-BGP EVPN



- Распространение информации о хостах подключенных к L2/L3 сегментам отделено от Underlay протокола
- Используется MultiProtocol-BGP (MP-BGP) на VTEP для распространения информации о достижимости Host/Subnet или внешних маршрутов
- Route-Reflectors (RR) для масштабирования

VXLAN/EVPN – распространение информации о хостах и сетях при помощи MP-BGP EVPN

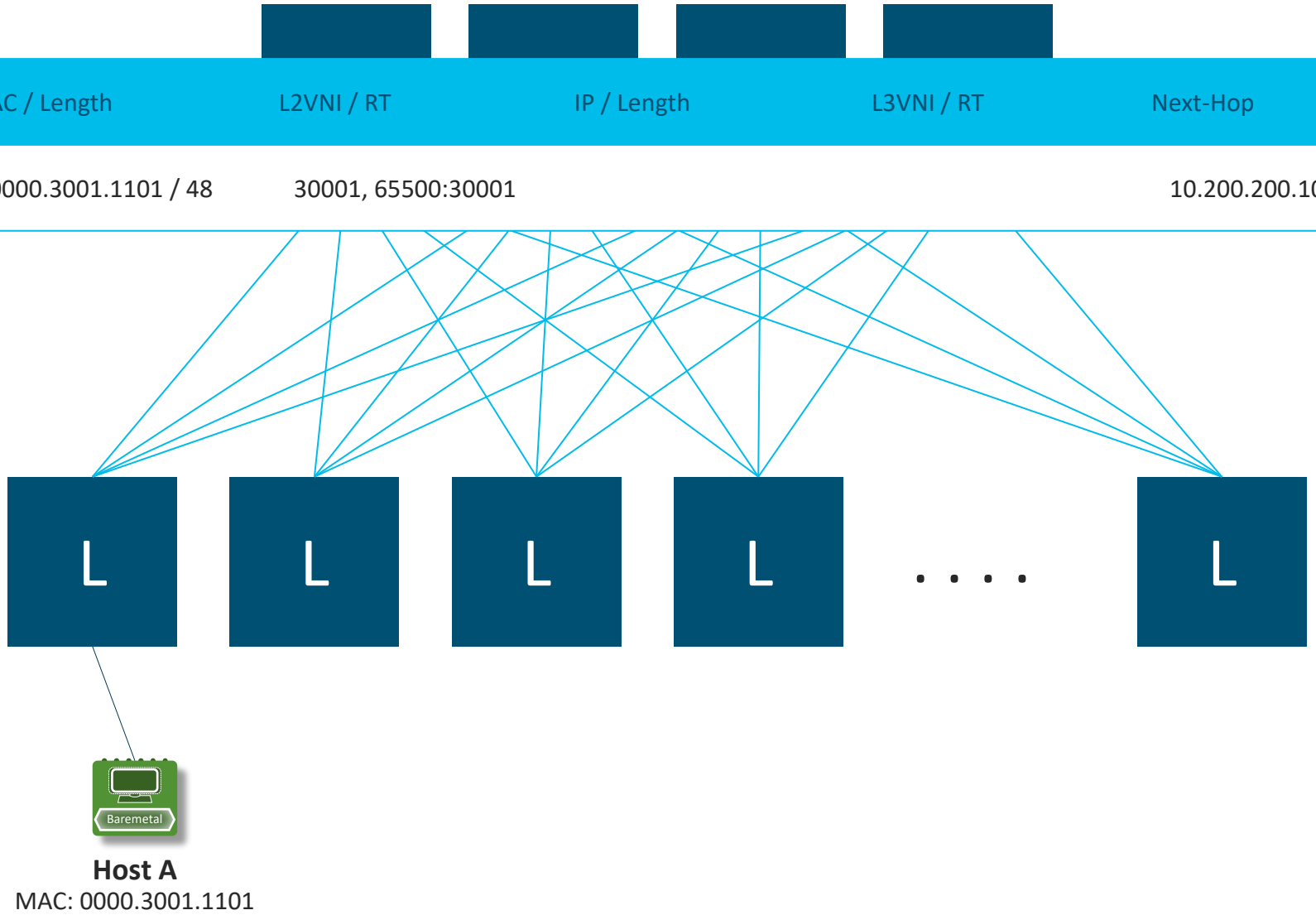


- BGP EVPN NLRI*
- Host MAC (Route Type 2)
  Только MAC, один VNI, один Route Target
- Host MAC+IP (Route Type 2)
  MAC и IP, два VNI, два Route Target, Router MAC
- Internal and External Subnet Prefixes (Route Type 5)
  IP префикс, один VNI, один Route Target

*NLRI: Network Layer Reachability Information (BGP Update Format)

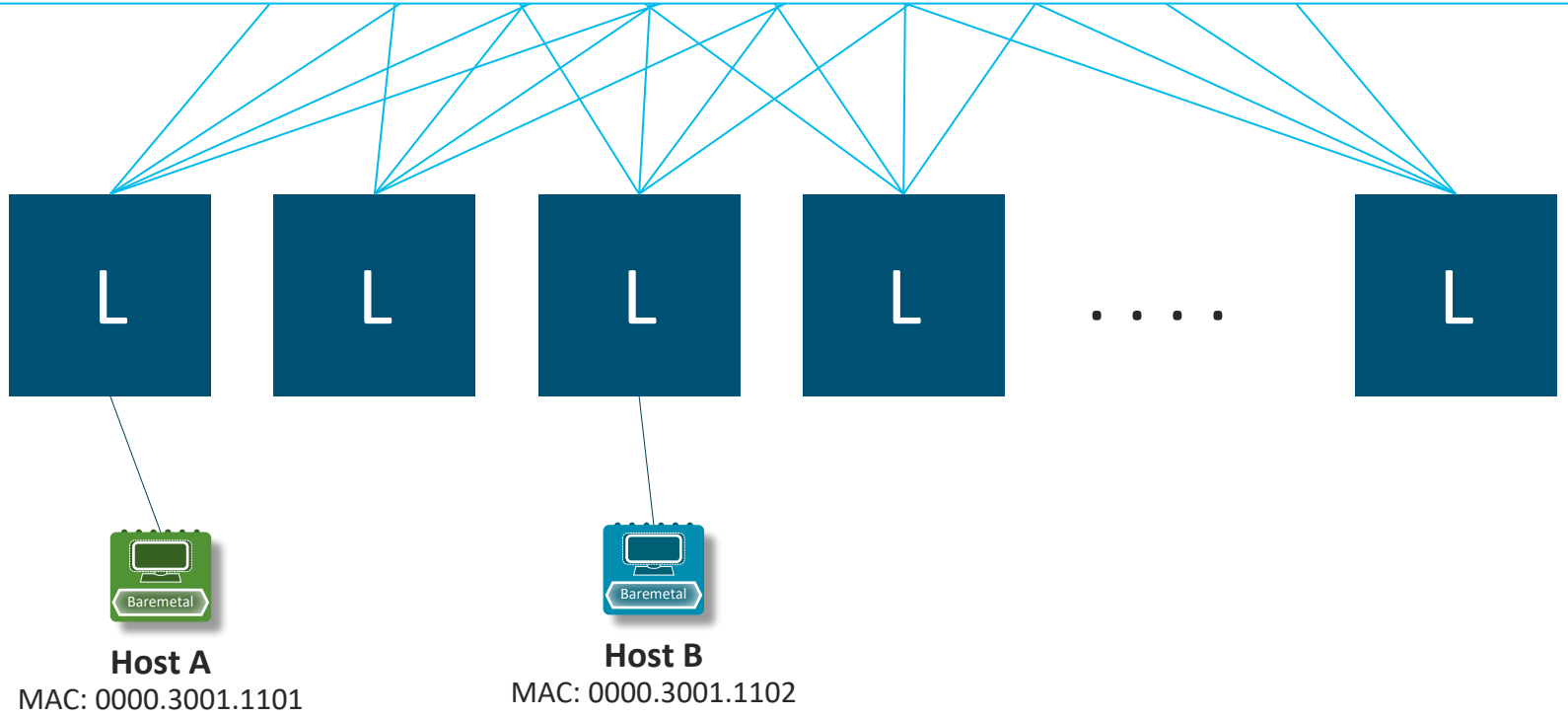
Анонс хостов

Type	MAC / Length	L2VNI / RT	IP / Length	L3VNI / RT	Next-Hop	Seq.
2	0000.3001.1101 / 48	30001, 65500:30001			10.200.200.101	



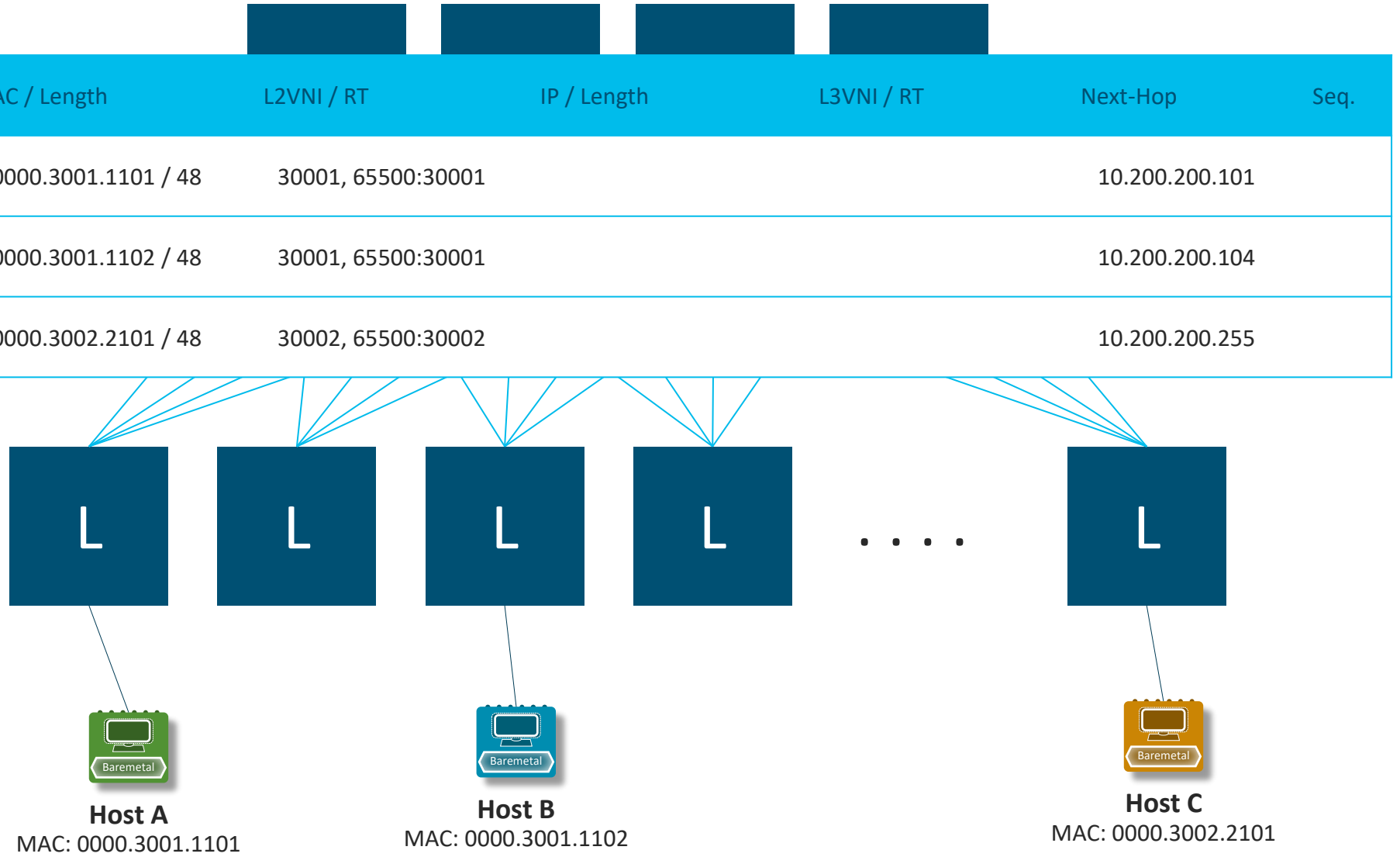
Анонс хостов

Type	MAC / Length	L2VNI / RT	IP / Length	L3VNI / RT	Next-Hop	Seq.
2	0000.3001.1101 / 48	30001, 65500:30001			10.200.200.101	
2	0000.3001.1102 / 48	30001, 65500:30001			10.200.200.104	

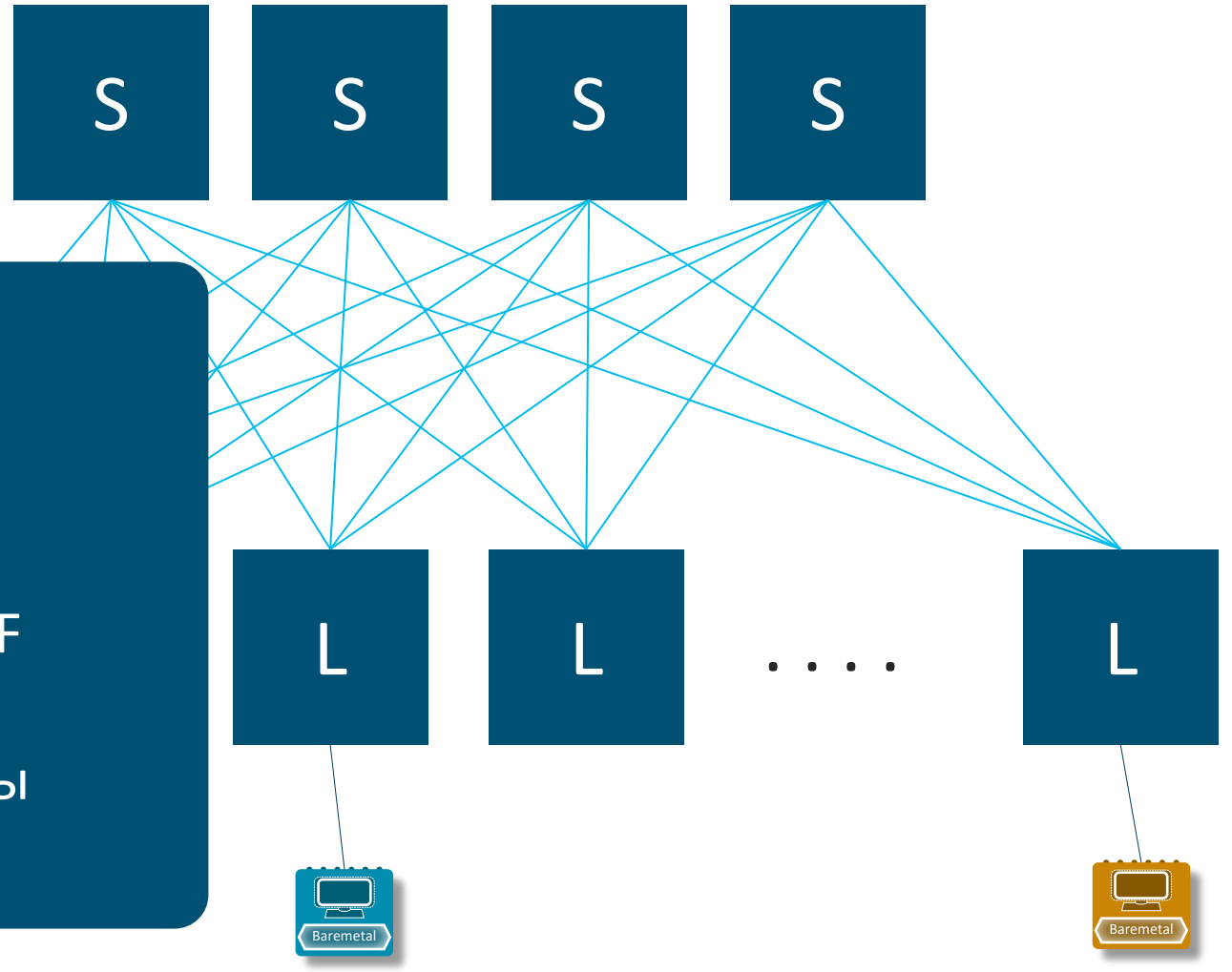


Анонс хостов

Type	MAC / Length	L2VNI / RT	IP / Length	L3VNI / RT	Next-Hop	Seq.
2	0000.3001.1101 / 48	30001, 65500:30001			10.200.200.101	
2	0000.3001.1102 / 48	30001, 65500:30001			10.200.200.104	
2	0000.3002.2101 / 48	30002, 65500:30002			10.200.200.255	



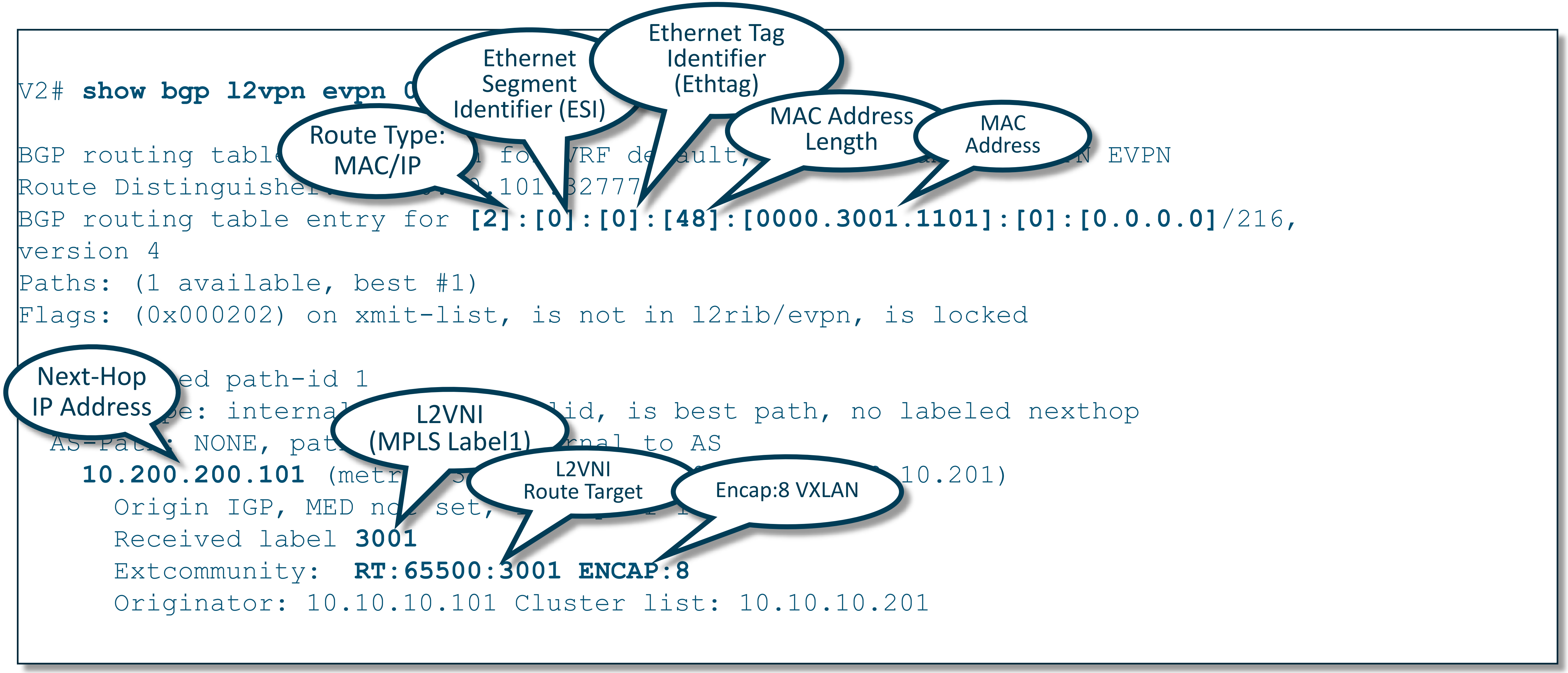
Анонс хостов



Host MAC (Route Type 2)

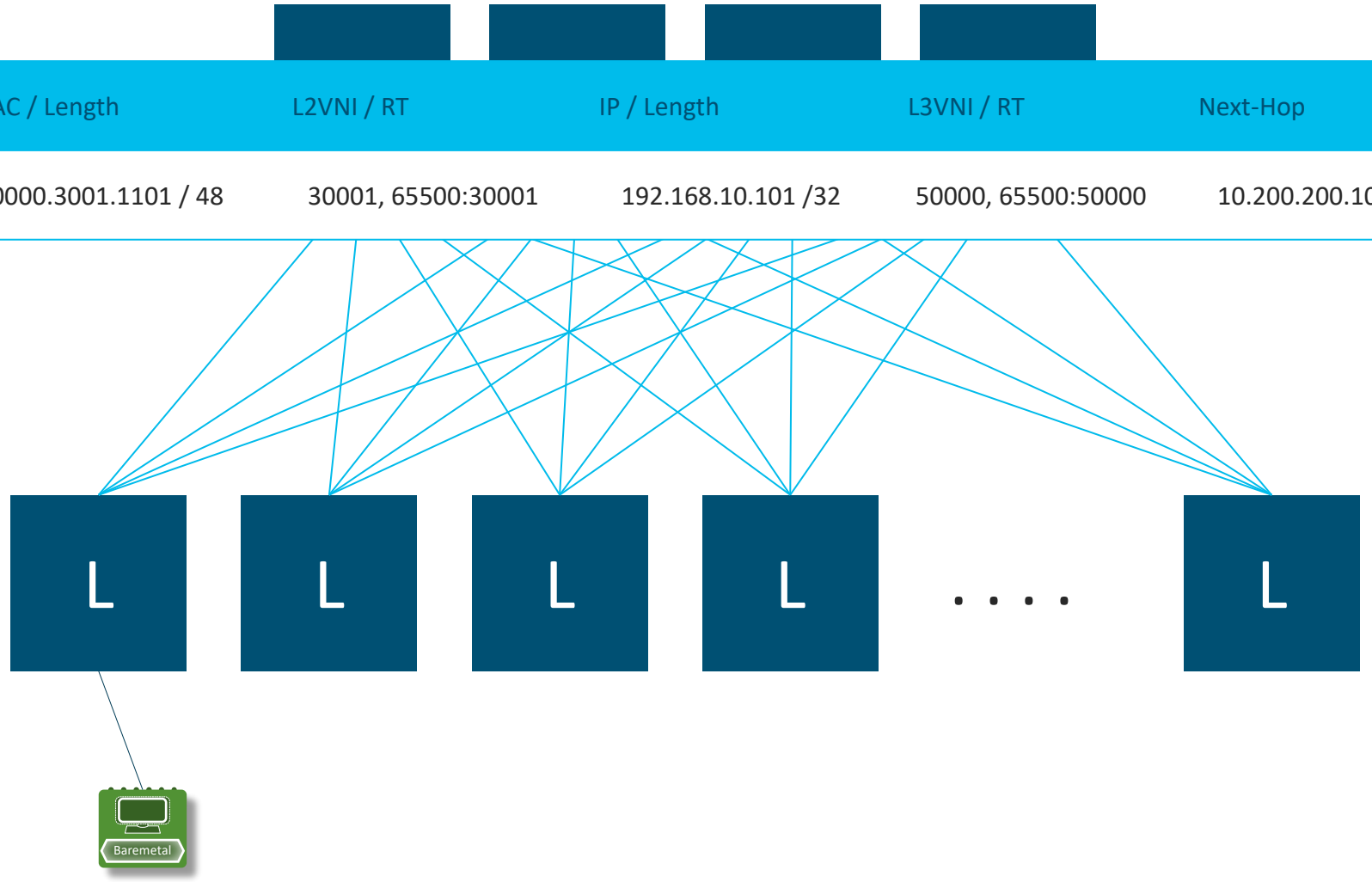
- MAC
- MPLS Label1 (L2VNI*)
- Route Target for MAC-VRF

MAC атрибуты обязательны



Анонс хостов

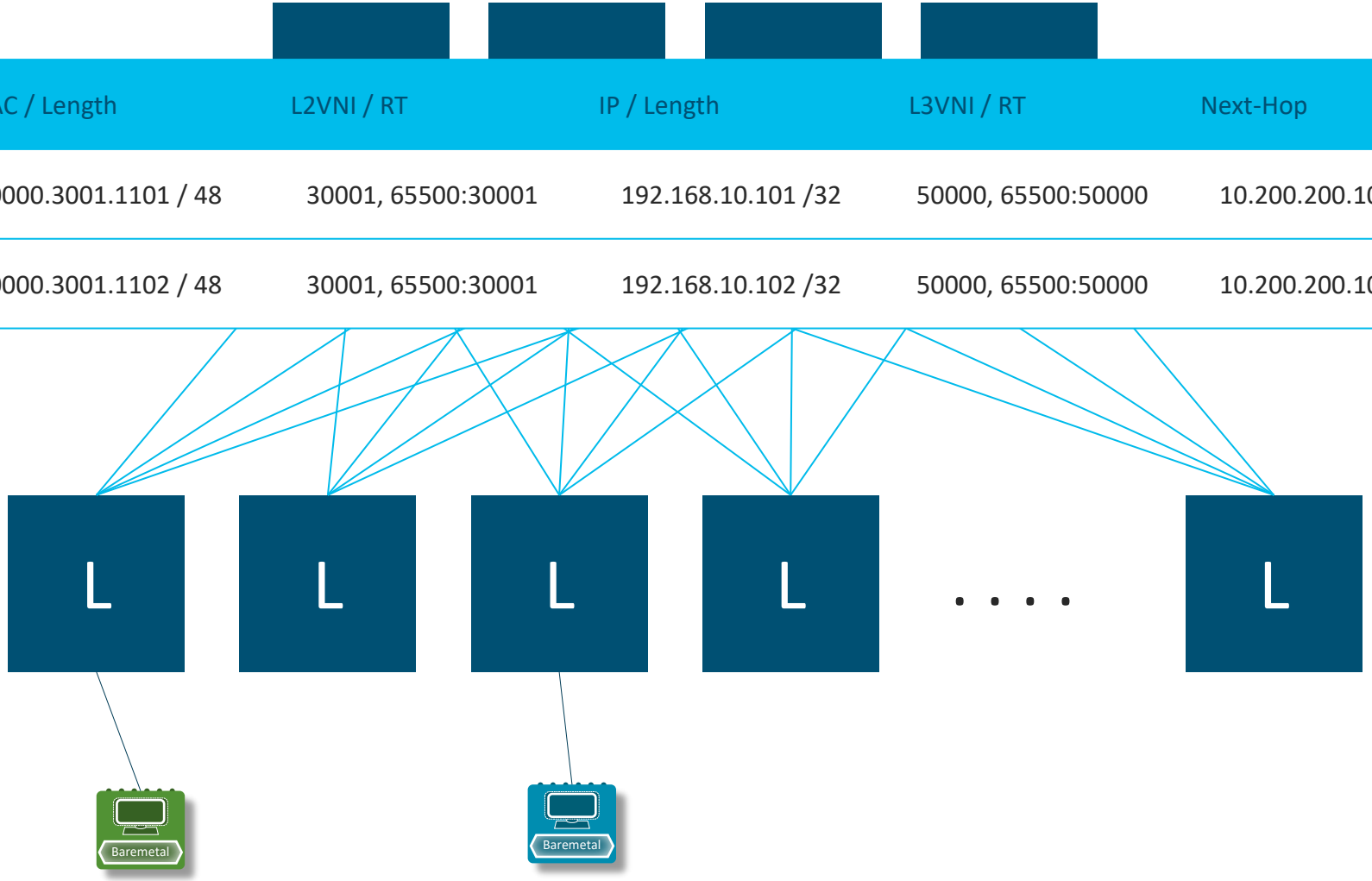
Type	MAC / Length	L2VNI / RT	IP / Length	L3VNI / RT	Next-Hop	Seq.
2	0000.3001.1101 / 48	30001, 65500:30001	192.168.10.101 /32	50000, 65500:50000	10.200.200.101	2



Host A
MAC: 0000.3001.1101
IP: 192.168.10.101

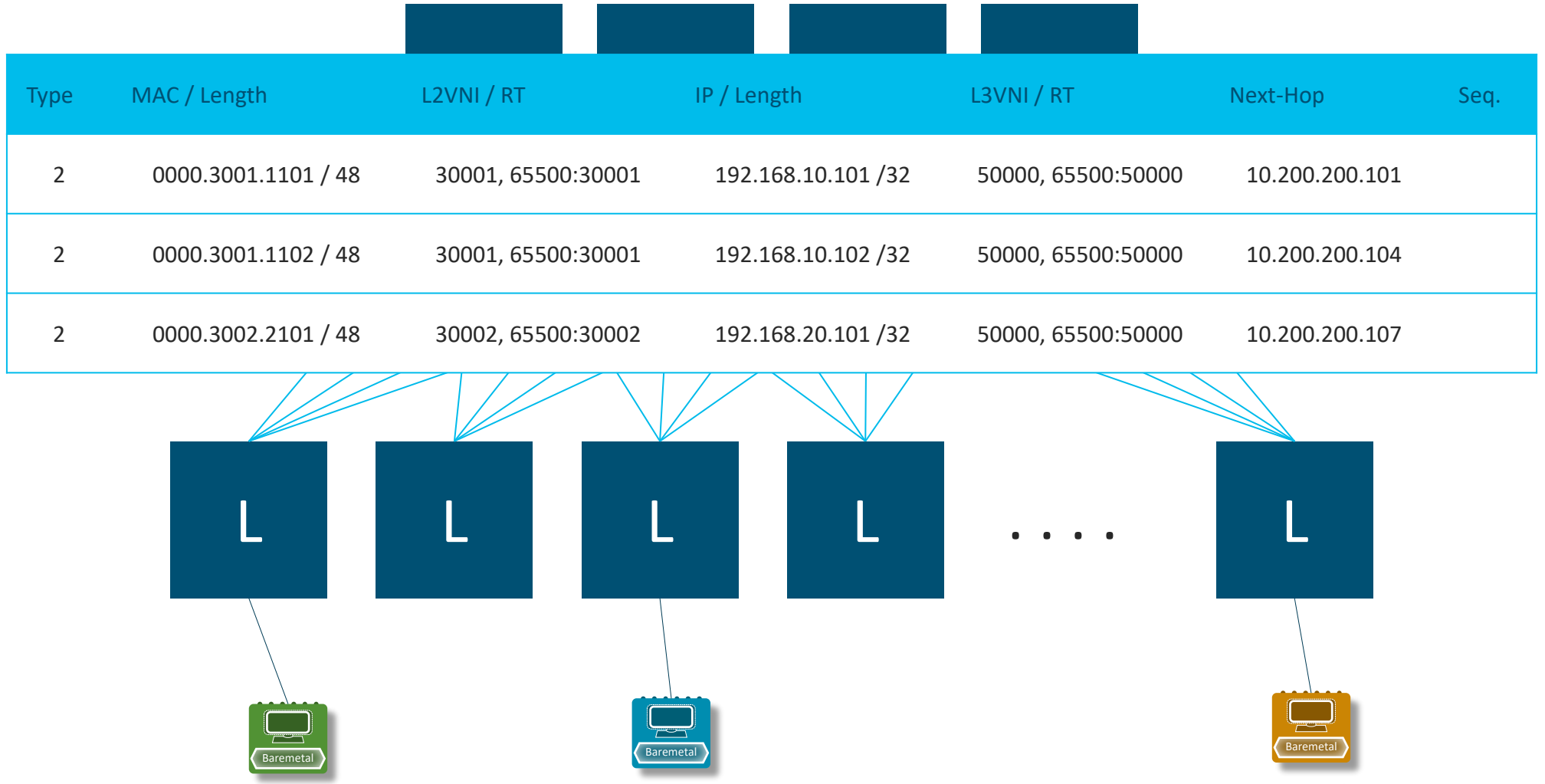
Анонс хостов

Type	MAC / Length	L2VNI / RT	IP / Length	L3VNI / RT	Next-Hop	Seq.
2	0000.3001.1101 / 48	30001, 65500:30001	192.168.10.101 /32	50000, 65500:50000	10.200.200.101	
2	0000.3001.1102 / 48	30001, 65500:30001	192.168.10.102 /32	50000, 65500:50000	10.200.200.104	



Анонс хостов

Type	MAC / Length	L2VNI / RT	IP / Length	L3VNI / RT	Next-Hop	Seq.
2	0000.3001.1101 / 48	30001, 65500:30001	192.168.10.101 /32	50000, 65500:50000	10.200.200.101	
2	0000.3001.1102 / 48	30001, 65500:30001	192.168.10.102 /32	50000, 65500:50000	10.200.200.104	
2	0000.3002.2101 / 48	30002, 65500:30002	192.168.20.101 /32	50000, 65500:50000	10.200.200.107	



Анонс хостов

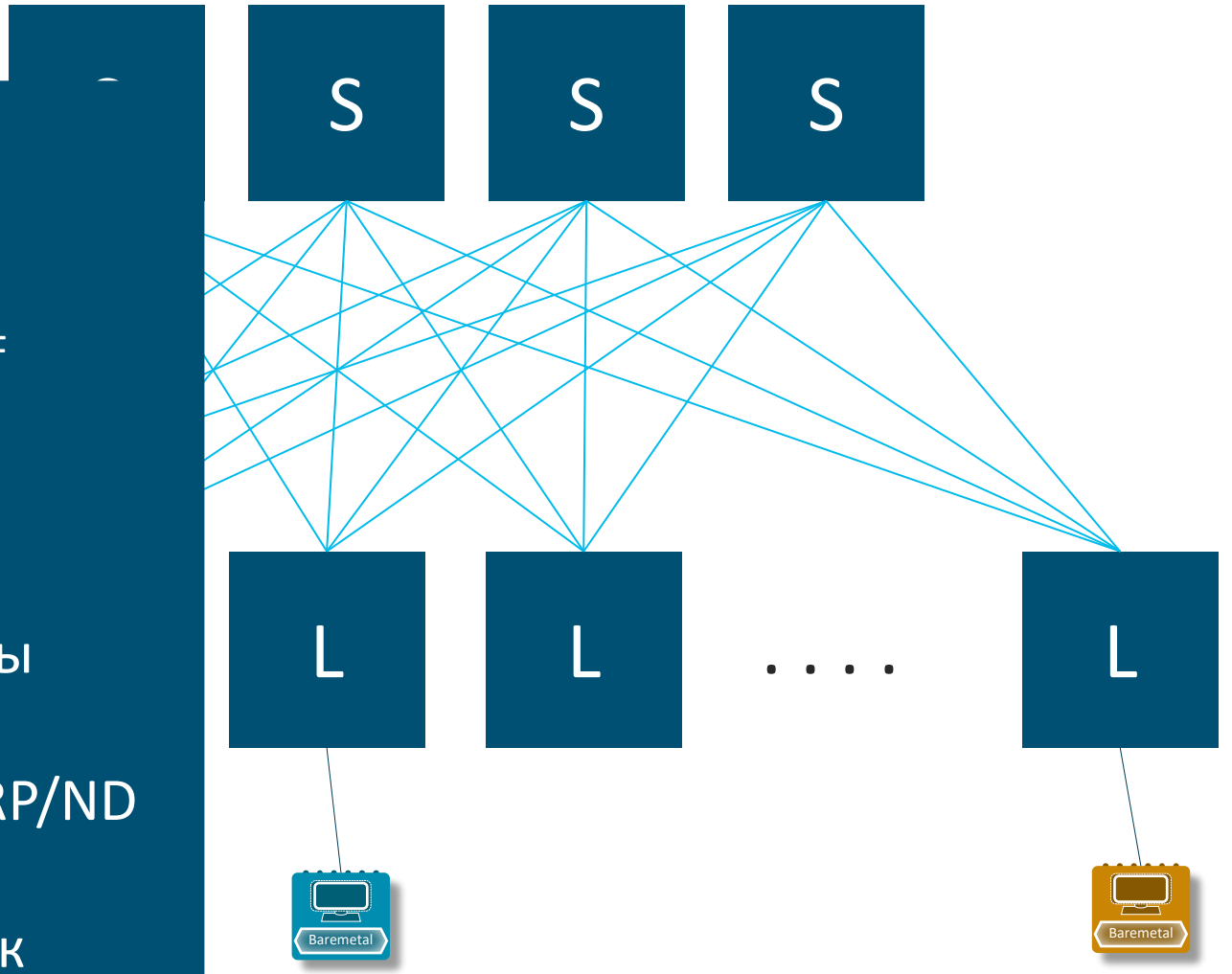
Host MAC (Route Type 2)

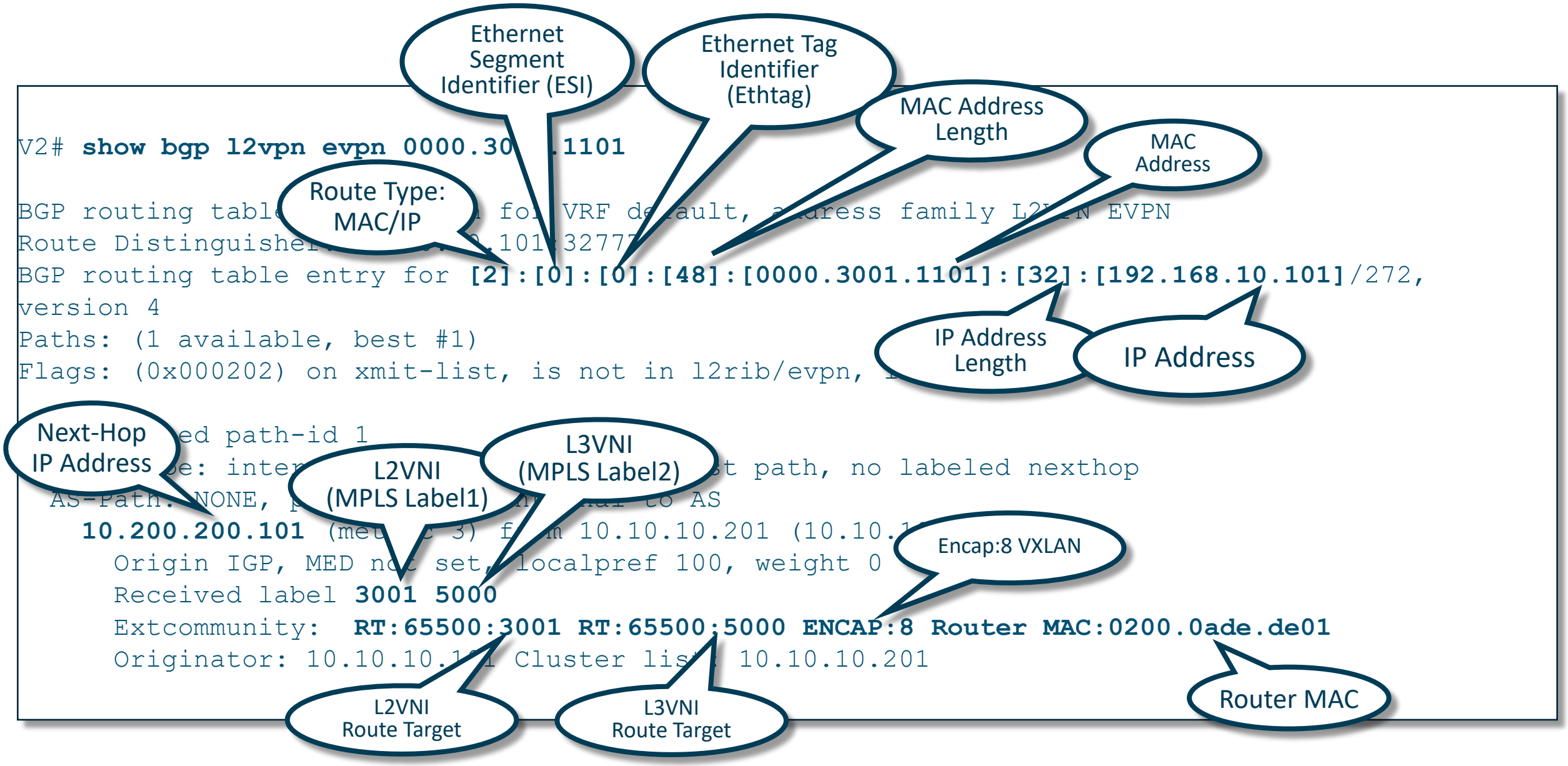
- MAC and IP
- MPLS Label1 (L2VNI)
- Route Target for MAC-VRF
- MPLS Label2 (L3VNI*)
- Route Target for IP-VRF
- Router MAC

IP атрибуты не обязательны

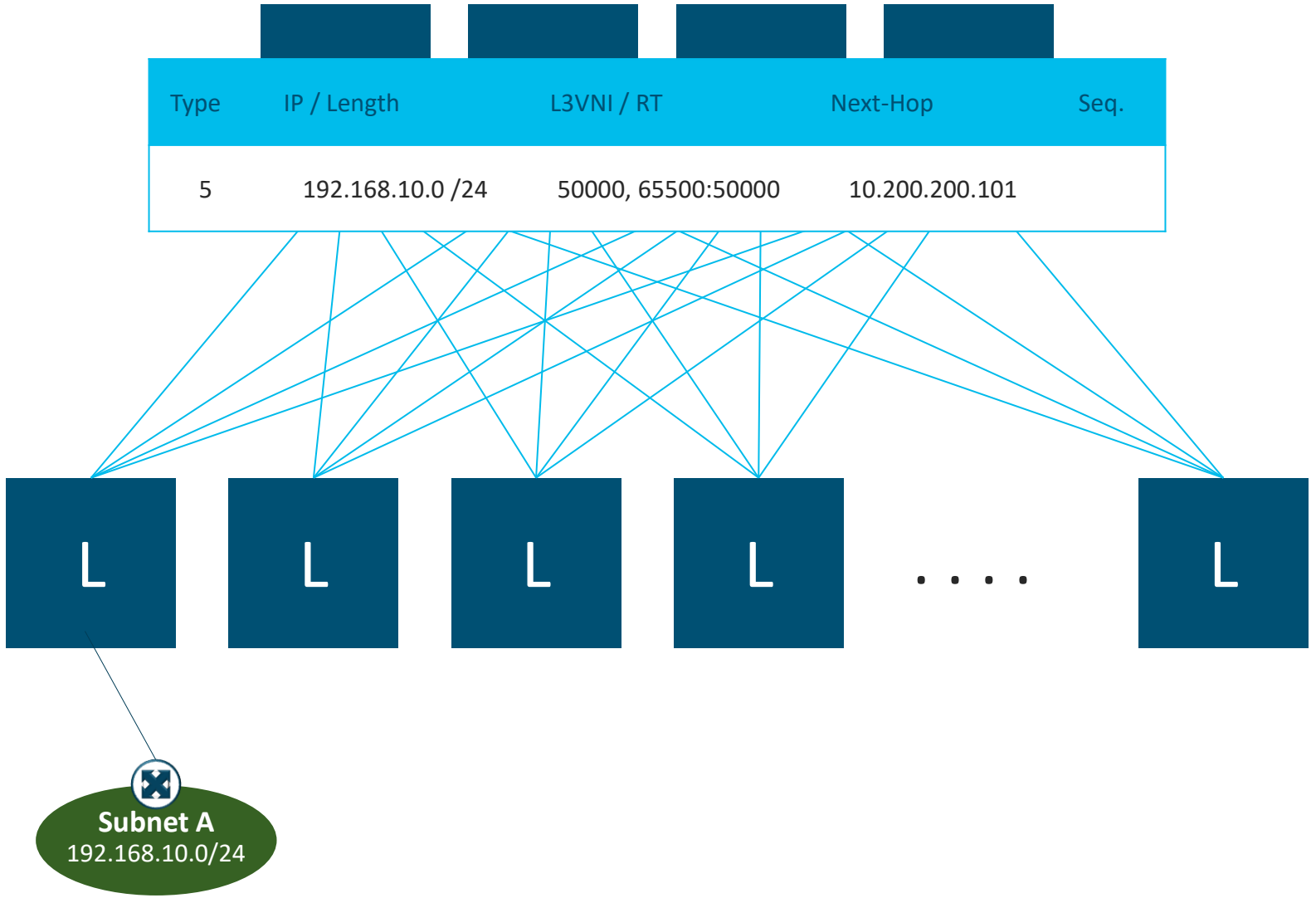
Формируется на основе ARP/ND

Передаётся в дополнение к анонсу «только MAC»





Распространение информации о подсетях

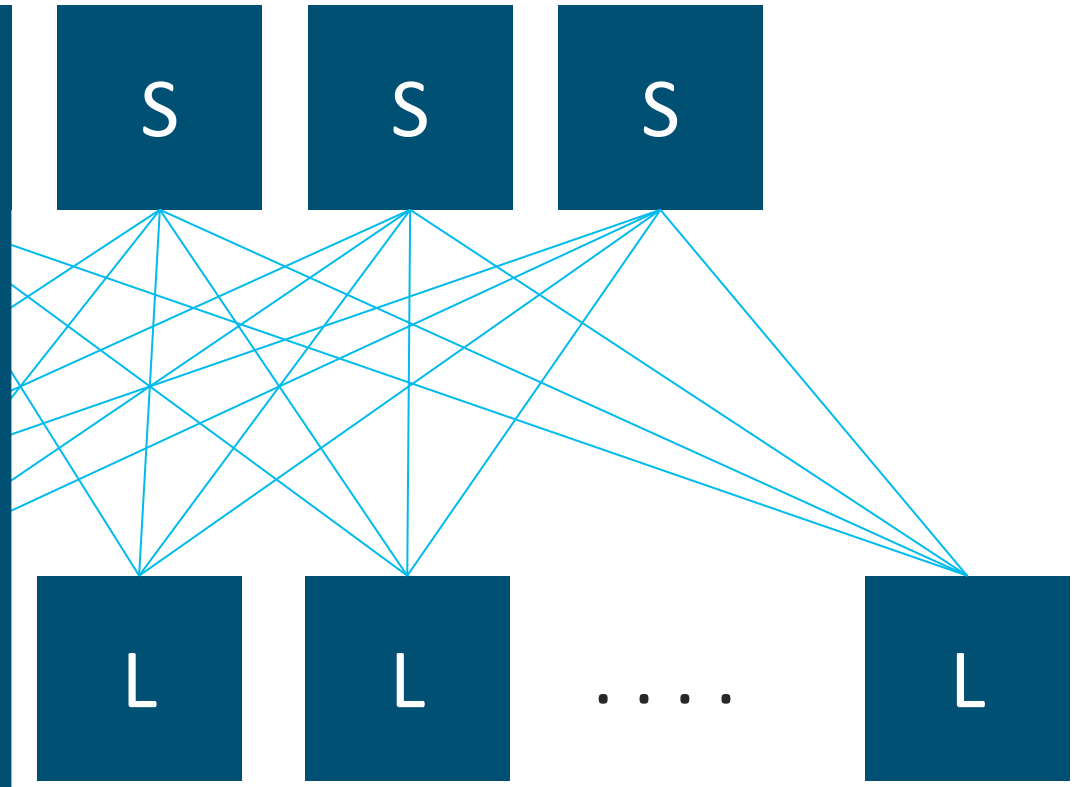


Распространение информации о подсетях

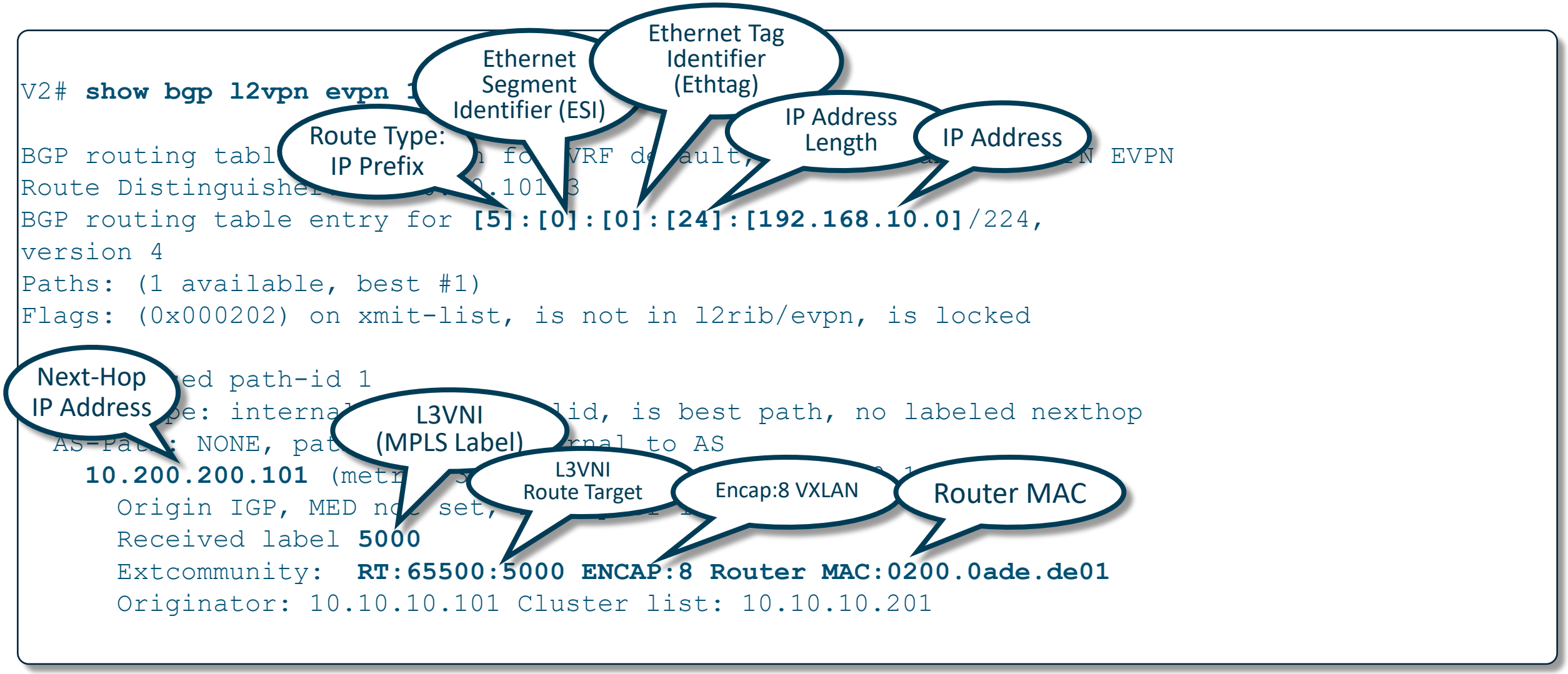
Internal and External Subnet Prefixes (Route Type 5)

- IP Prefix
- MPLS Label (L3VNI)
- Route Target for IP-VRF
- Router MAC

Формируется на основе информации из протоколов маршрутизации (или static/connected маршрутов)



Subnet A
192.168.10.0/24



VXLAN/EVPN и VXLAN Flood&Learn Сравнение

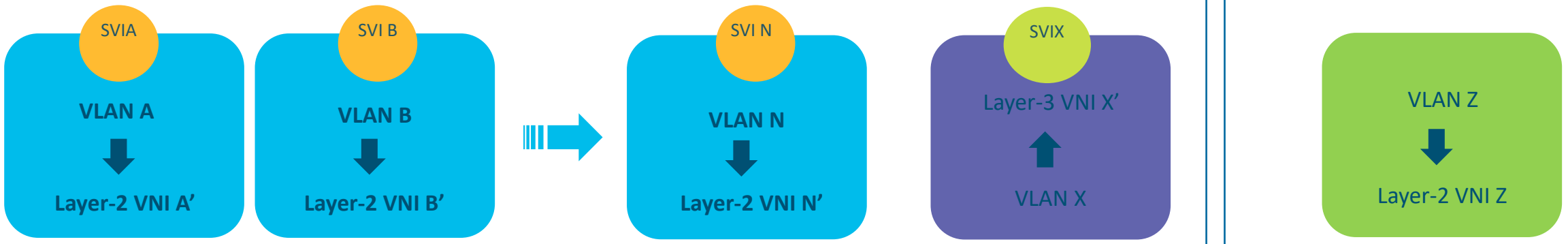
	VXLAN/EVPN	VXLAN Flood&Learn
Инкапсуляция	VXLAN (MAC-in-UDP)	VXLAN (MAC-in-UDP)
Требования к транспорту	IP	IP
Распространение информации о подключенных узлах	MP-BGP EVPN	Flood&Learn
Доставка BUM* трафика	Multicast (PIM) или репликация через unicast	Multicast (PIM) или репликация через unicast
Уровень управления в транспортной сети	Любой протокол машрутизации (OSPF, IS-IS, eBGP)	Любой протокол машрутизации (OSPF, IS-IS, eBGP)
Тип сервиса	L2 и L3	L2
Идентификатор пограничного узла	VTEP IP	VTEP IP
Аутентификация	MP-BGP	Нет
Стандарт	RFC 7348 + RFC 8365	RFC 7348

*BUM: Broadcast, Unknown Unicast, Multicast

L2 и L3 сервисы в VXLAN/EVPN

Модель сервисов в VXLAN EVPN

Tenant A (VRF A)



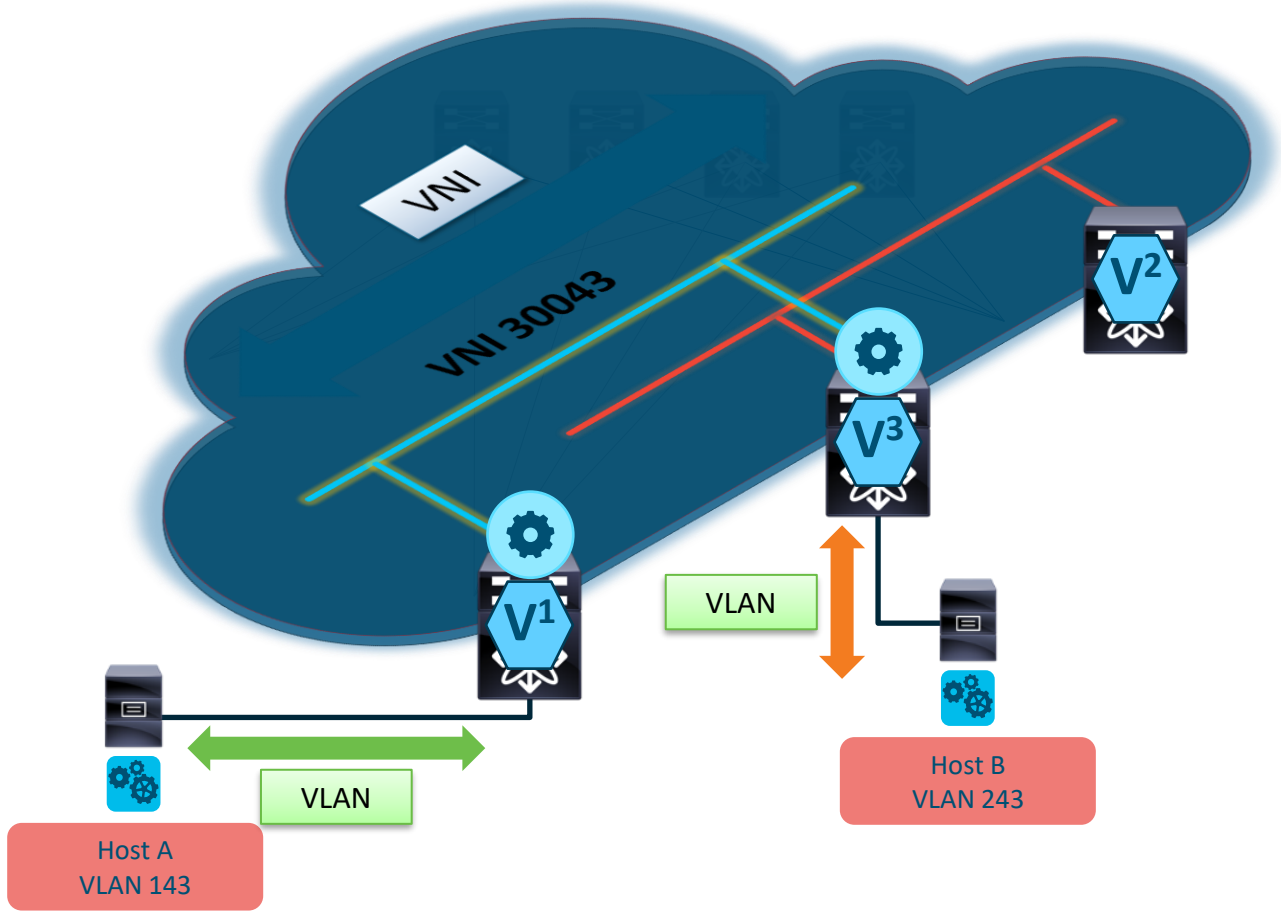
- Один локальный VLAN отображается на VTEP в один L2 VNI => сегмент, относящийся к VRF
- К VRF может относиться несколько сегментов (L2 VNI)
- Внутри L2 VNI – L2 коммутация
- Между L2 VNI – маршрутизация

- Один L3 VNI на VRF для маршрутизации между VTEP
- Требует назначения отдельного VLAN X

- Возможно также использование «чистых» L2 VNI без маршрутизации на фабрике (без SVI)
- Отображение локального VLAN на L2 VNI

Layer-2 Multi-Tenancy

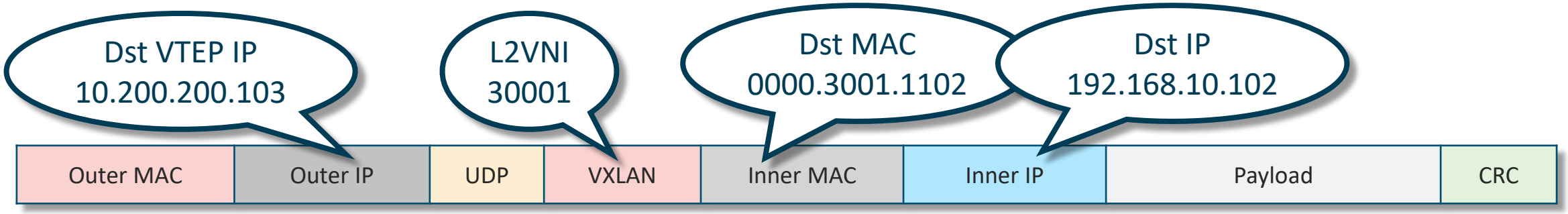
- VNI обеспечивают изоляцию на L2 и L3 в VXLAN сети
  - Принятые фреймы отображаются в конкретный VNI
  - Соответствие VLAN-VNI возможно (в завимости от версии ПО и платформы):
    - На уровне коммутатора
    - На уровне порта
- VLANы имеют локальное значение, VNI - глобальное



VXLAN и BGP EVPN – L2 взаимодействие

Control-Plane (BGP EVPN)

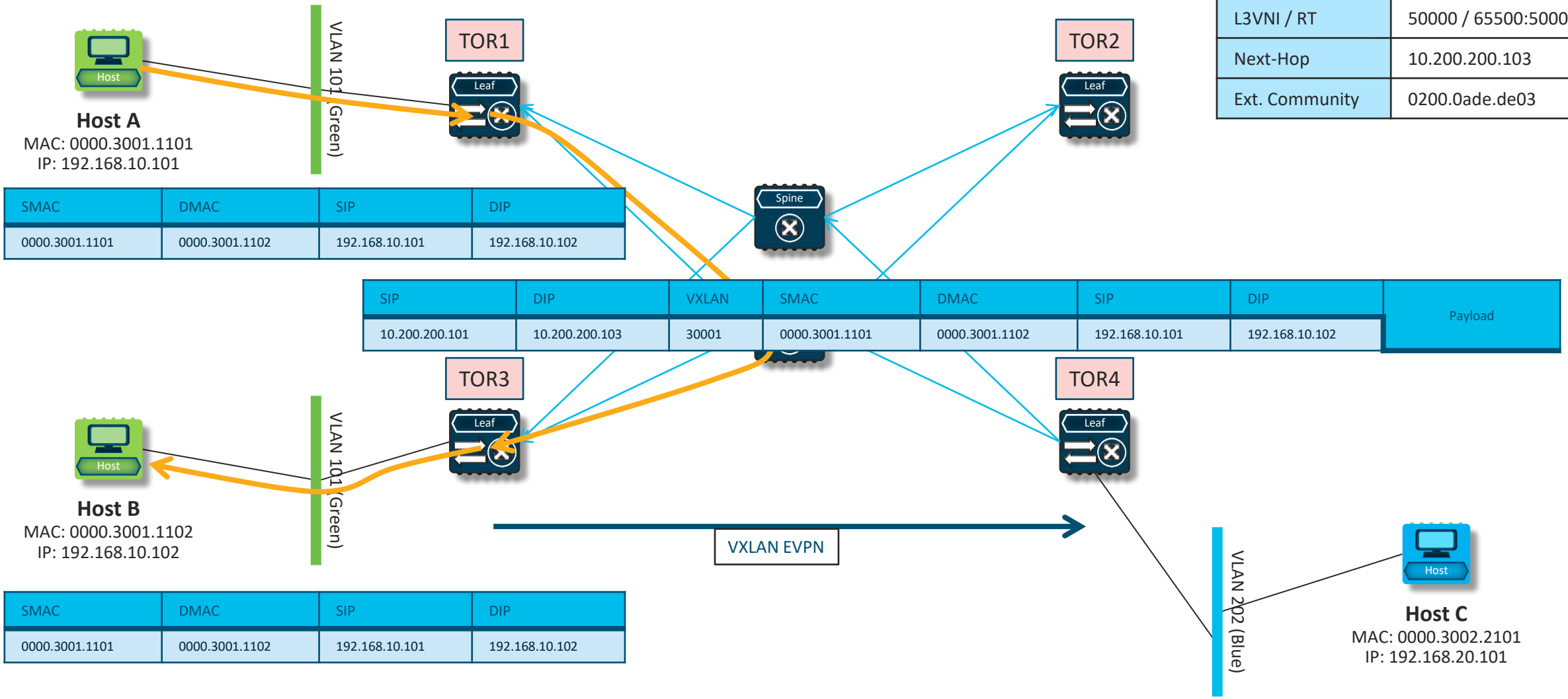
Type	MAC / Length	L2VNI / RT	IP / Length	L3VNI / RT	Next-Hop	Seq.
2	0000.3001.1102/48	30001 65500:30001	192.168.10.102/32	50000 65500:50000	10.200.200.103	



Data-Plane (VXLAN)

Packet Walk – Bridging

EVPN Control-Plane	
Type	2
MAC / Length	0000.3001.1102 / 48
L2VNI / RT	30001 / 65500:30001
IP / Length	192.168.10.102 / 32
L3VNI / RT	50000 / 65500:5000
Next-Hop	10.200.200.103
Ext. Community	0200.0ade.de03

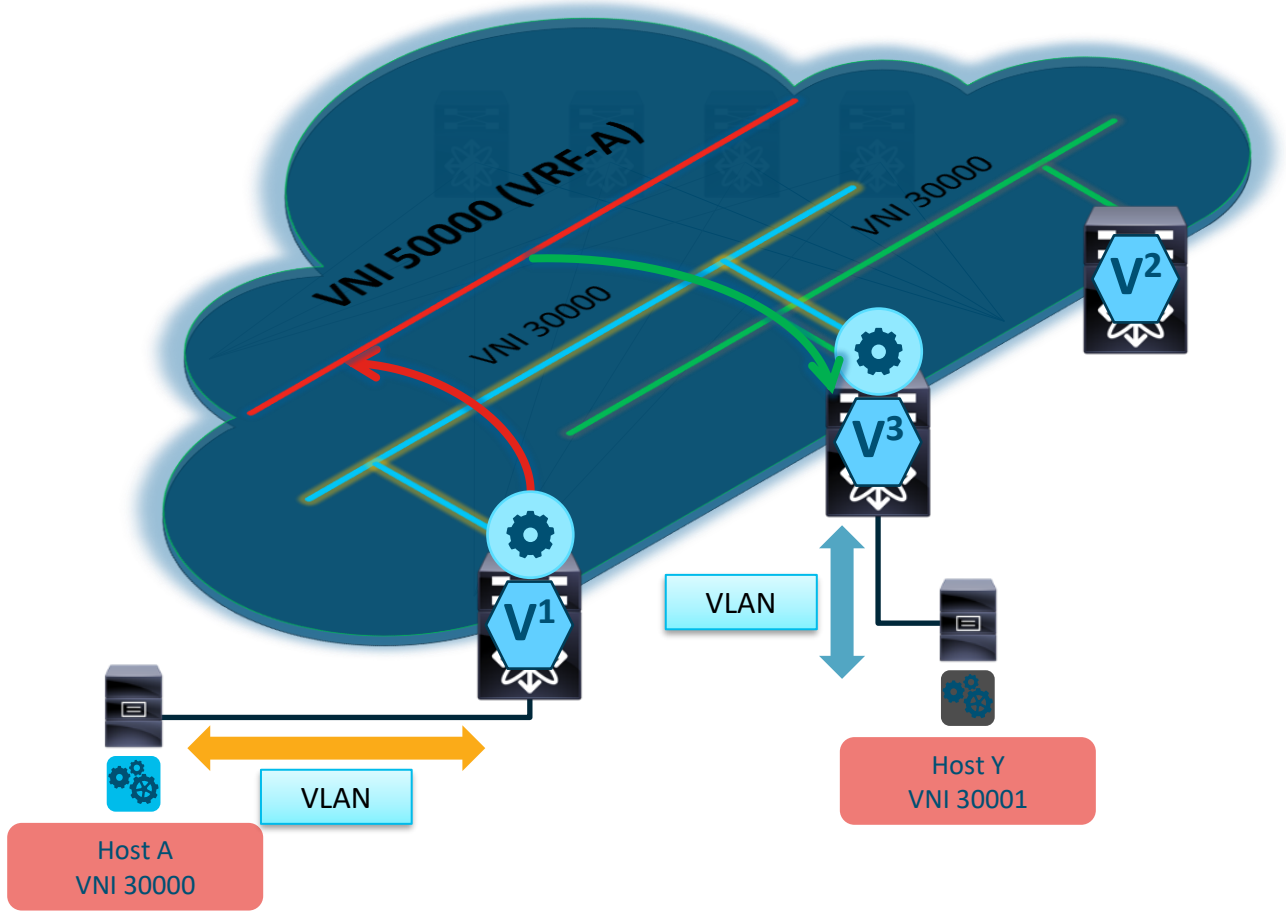


VXLAN/EVPN ≠ Ingress Replication!

- VXLAN с уровнем управления BGP EVPN по-прежнему требует механизм репликации BUM (Broadcast/Unknown Unicast/Multicast) трафика
- Два возможных варианта
  - **Multicast** – эффективный способ отправки всем VTEP в данном L2 VNI (посылается одна копия)
  - **Unicast** (Ingress Replication) – отправка копии BUM пакета каждому из VTEP в данном L2 VNI

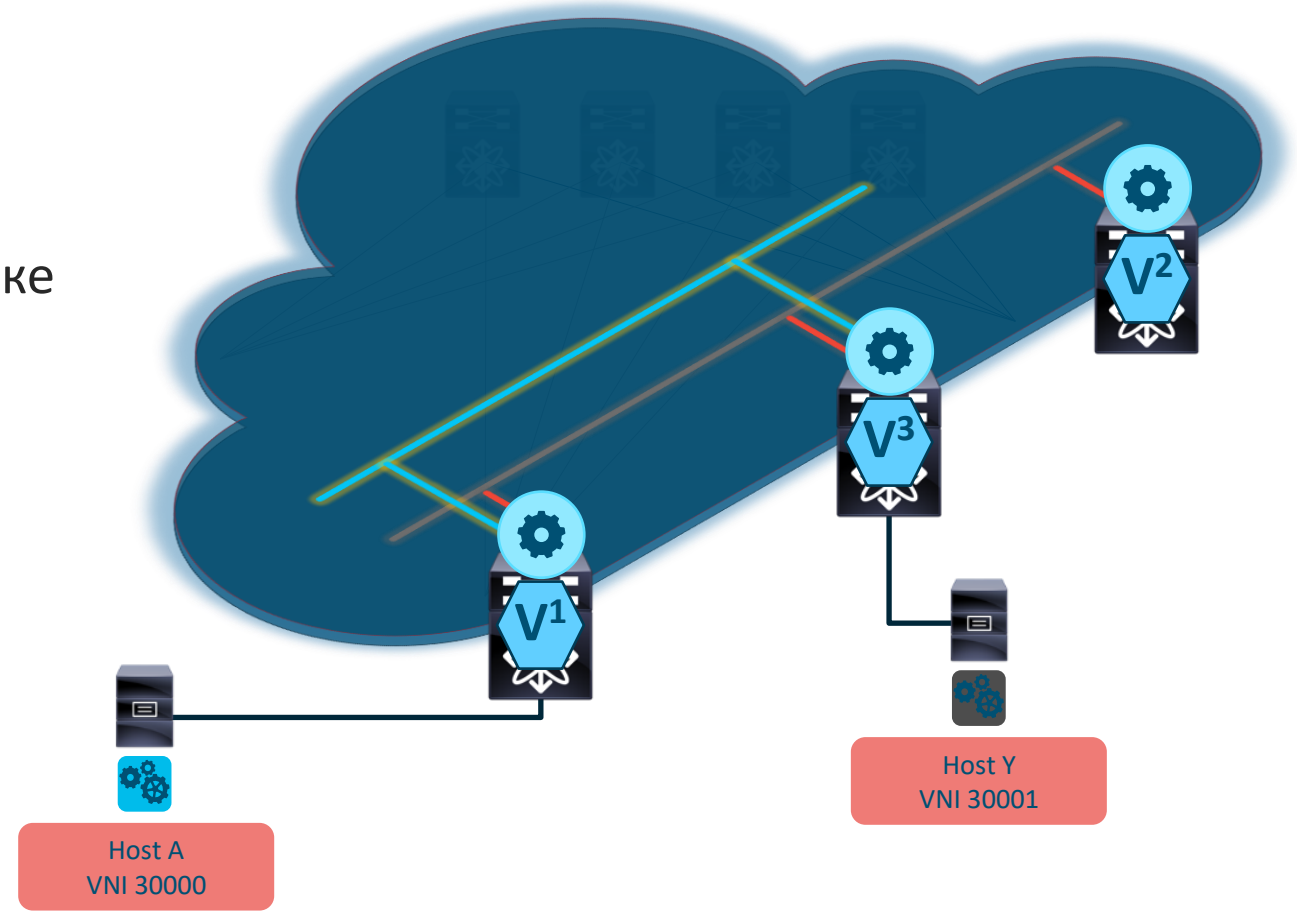
Layer-3 Multi-Tenancy

- VNI обеспечивают изоляцию на L2 и L3 в VXLAN сети
  - Принятые фреймы отображаются в VNI, соответствующий данному VRF
- Весь маршрутизируемый между VTEP трафик использует L3 VNI, соответствующий данному VRF («симметричный режим»)



Распределённый шлюз по умолчанию (Anycast Gateway)
Требует EVPN control plane

- Распределённая маршрутизация с Distributed Anycast Gateway (интегрированная маршрутизация и коммутация IRB)
  - Маршрутизация между VNI
  - Бриджинг внутри VNI
- Функции шлюза по умолчанию в любой точке подключения хостов
  - На всех VTEP коммутаторах один и тот же Gateway IP и MAC адрес
  - Шлюз всегда активен на всех устройствах
  - Никаких специальных протоколов или обмена hello-пакетами
- Лучше масштабирование
  - Только локальные ARP записи

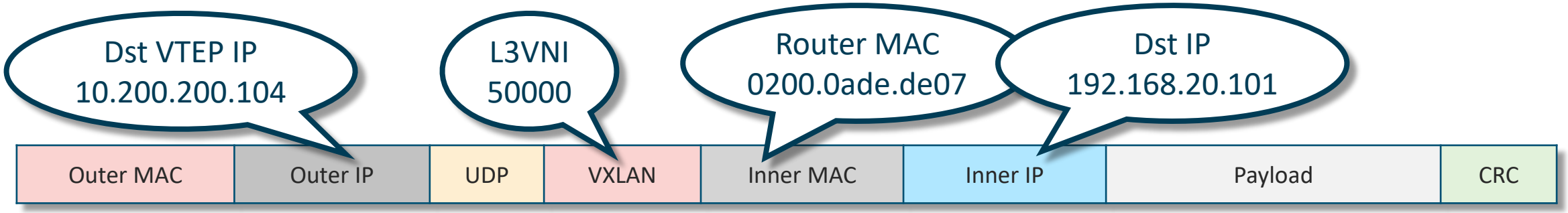


VXLAN и BGP EVPN – L3 взаимодействие

Control-Plane (BGP EVPN)

Extended Community
Router MAC
0200.0ade.de07

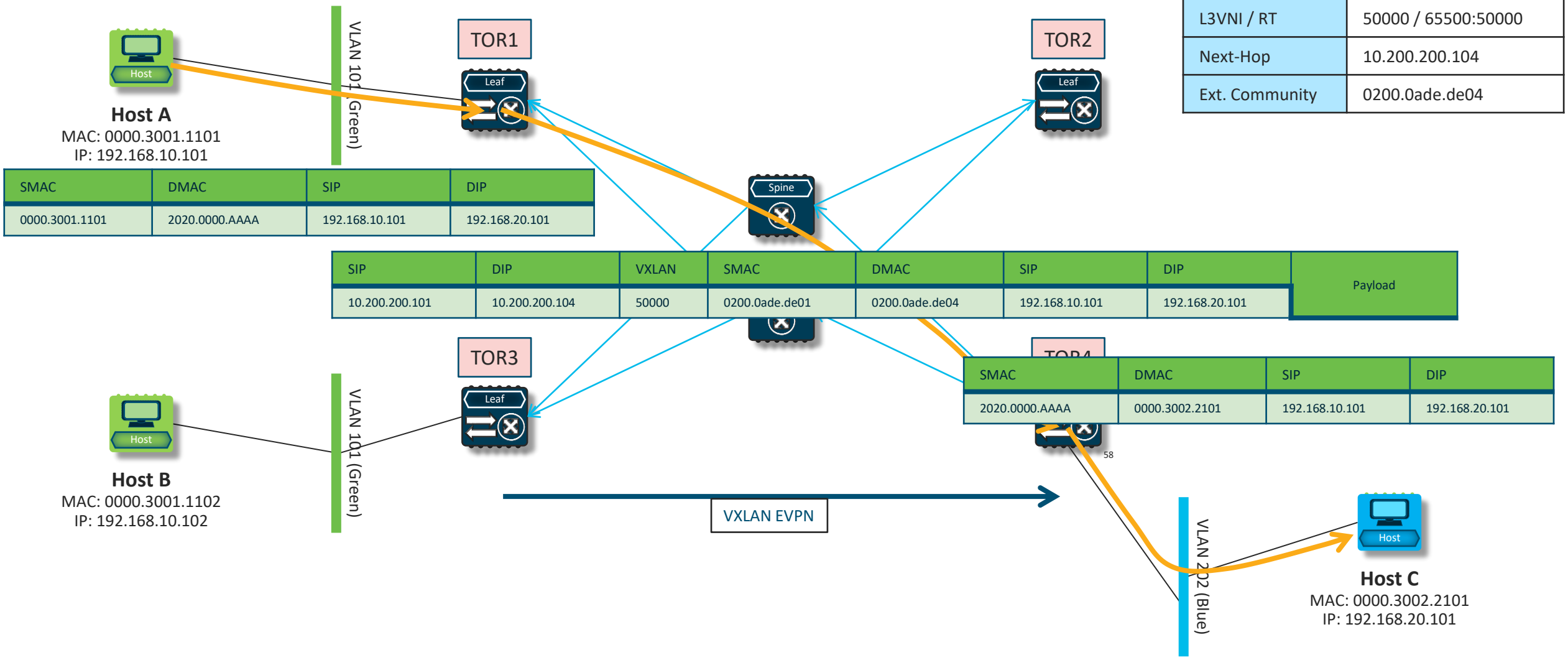
Type	MAC / Length	L2VNI / RT	IP / Length	L3VNI / RT	Next-Hop	Seq.
2	0000.3001.2101/48	30001 65500:30001	192.168.20.101/32	50000 65500:50000	10.200.200.104	



Data-Plane (VXLAN)

Packet Walk – Routing

EVPN Control-Plane	
Type	2
MAC / Length	0000.3002.2101 / 48
L2VNI / RT	30002 / 65500:30002
IP / Length	192.168.20.101 / 32
L3VNI / RT	50000 / 65500:50000
Next-Hop	10.200.200.104
Ext. Community	0200.0ade.de04

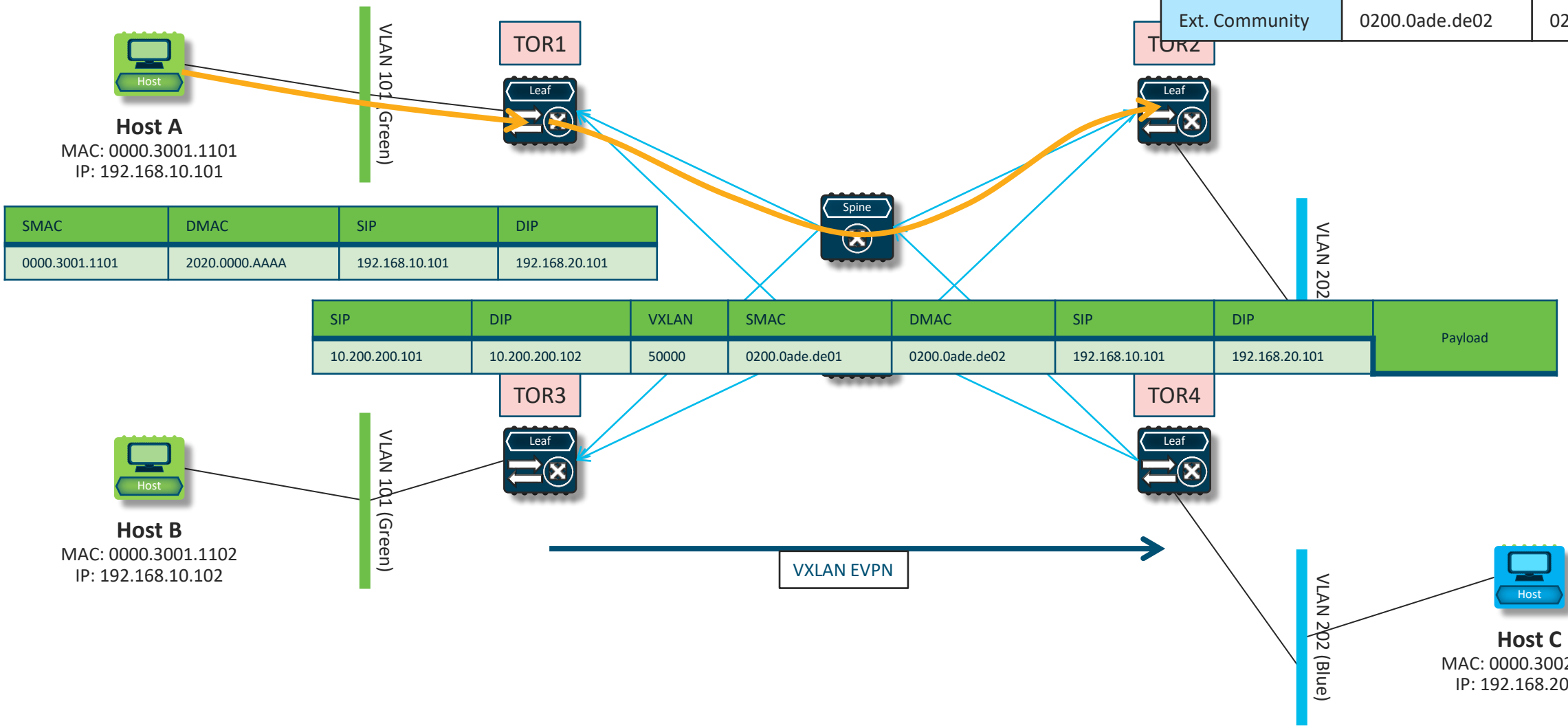


Что, если получатель – «молчаливый» хост??

Packet Walk – Routing

Обнаружение «молчаливого» хоста

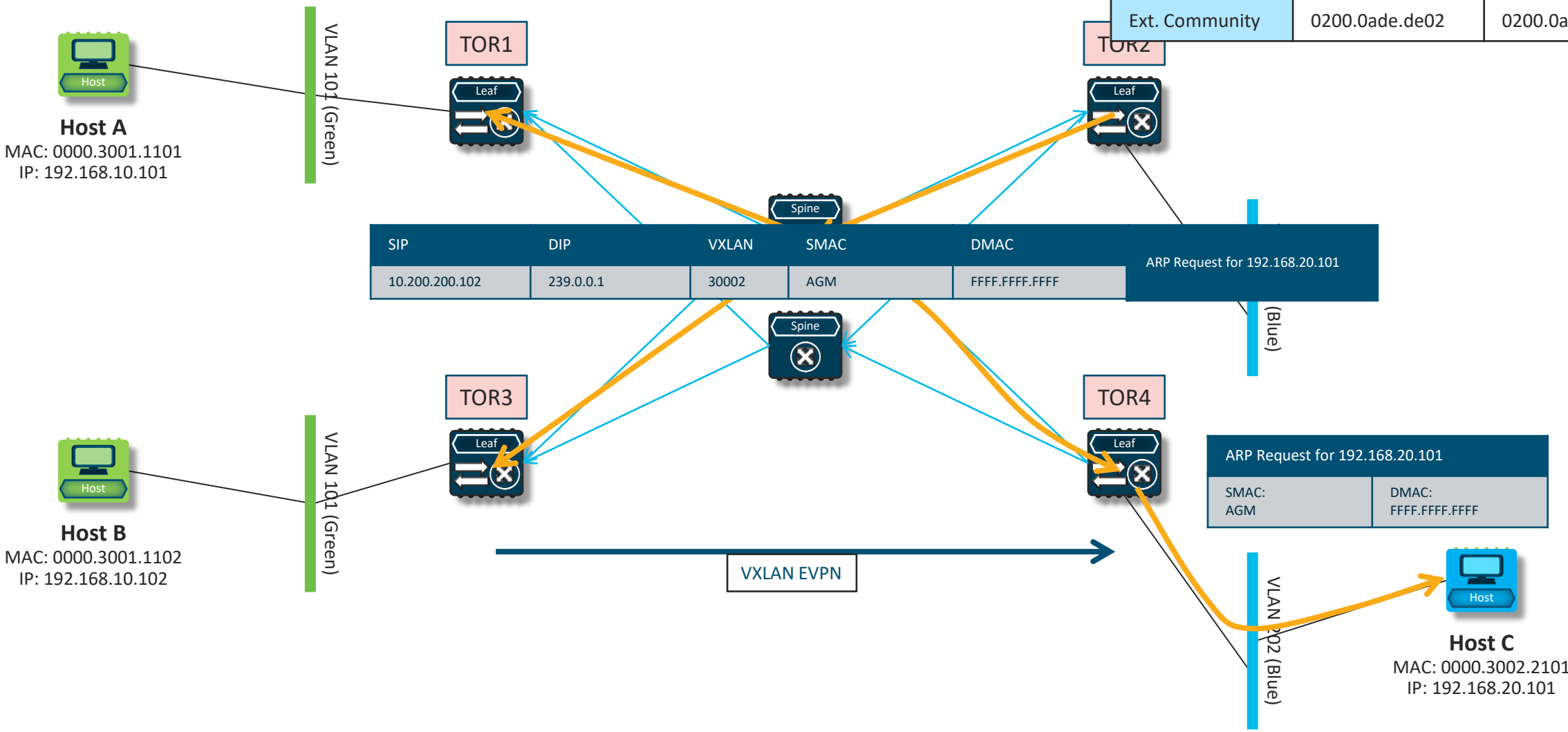
EVPN Control-Plane		
Type	5	
IP / Length	192.168.20.0 / 24	
L3VNI / RT	50000 / 65500:50000	
Next-Hop	10.200.200.102	10.200.200.104
Ext. Community	0200.0ade.de02	0200.0ade.de04



Packet Walk – Routing

Обнаружение «молчаливого» хоста

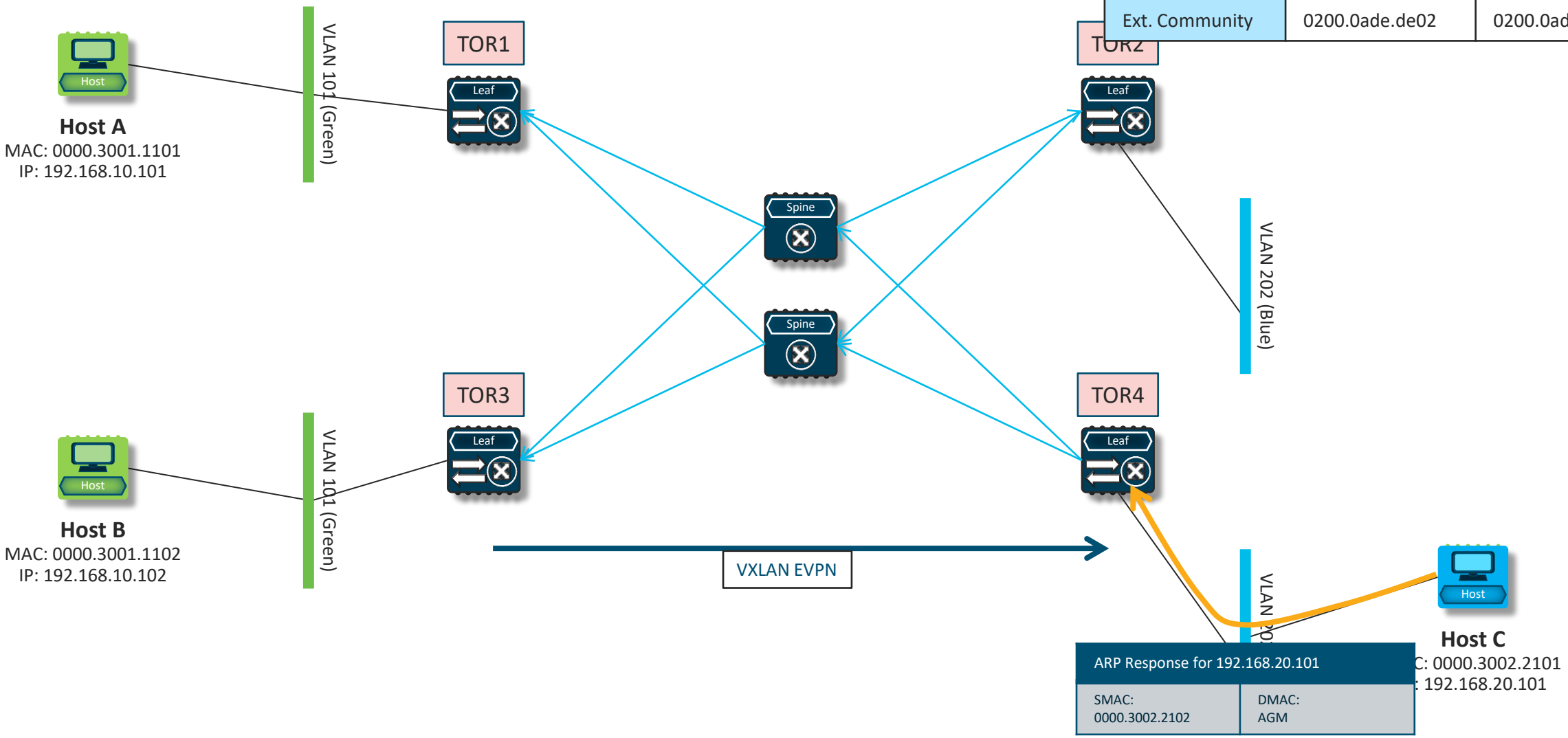
EVPN Control-Plane		
Type	5	
IP / Length	192.168.20.0 / 24	
L3VNI / RT	50000 / 65500:50000	
Next-Hop	10.200.200.102	10.200.200.104
Ext. Community	0200.0ade.de02	0200.0ade.de04



Packet Walk – Routing

Обнаружение «молчаливого» хоста

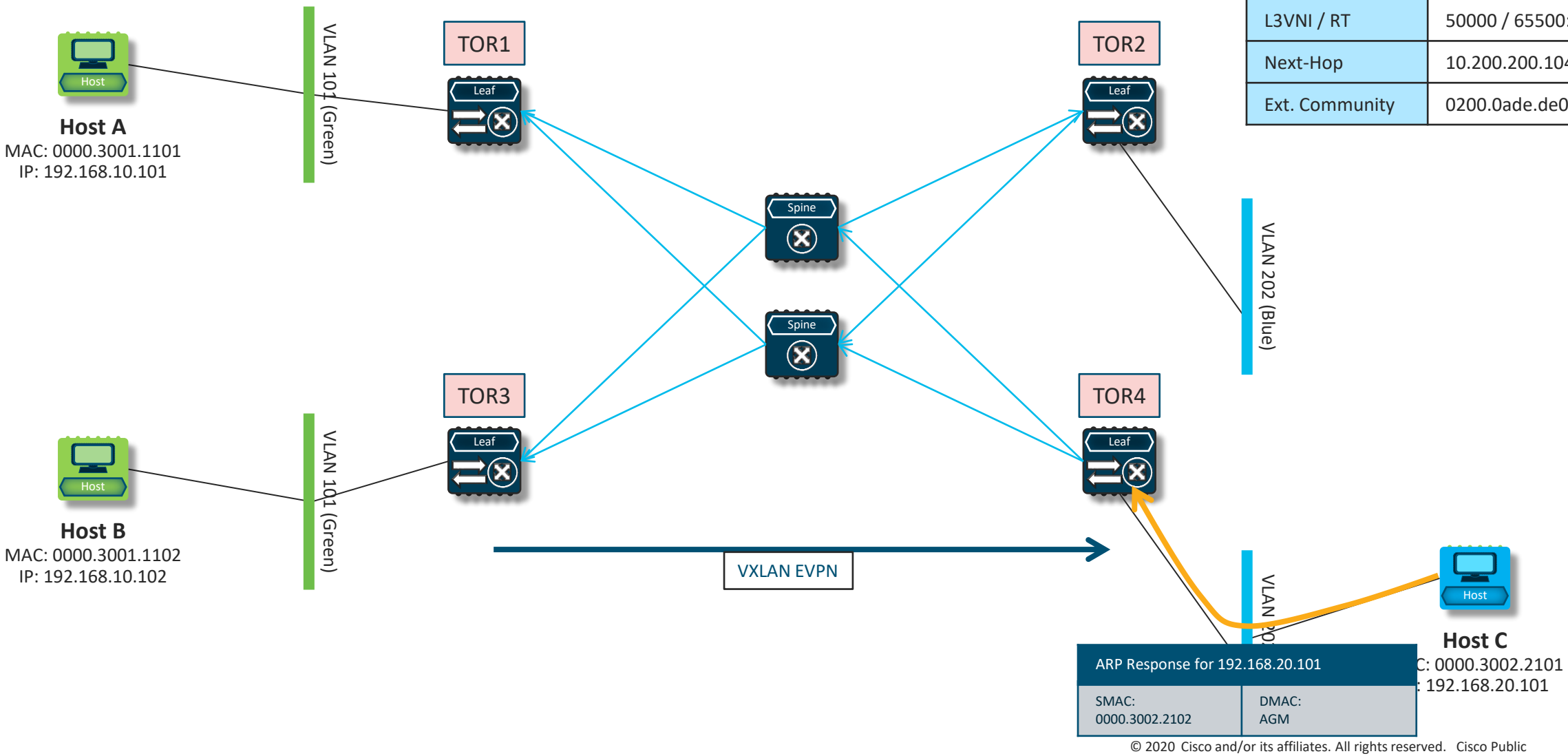
EVPN Control-Plane		
Type	5	
IP / Length	192.168.20.0 / 24	
L3VNI / RT	50000 / 65500:50000	
Next-Hop	10.200.200.102	10.200.200.104
Ext. Community	0200.0ade.de02	0200.0ade.de04



Packet Walk – Routing

Обнаружение «молчаливого» хоста

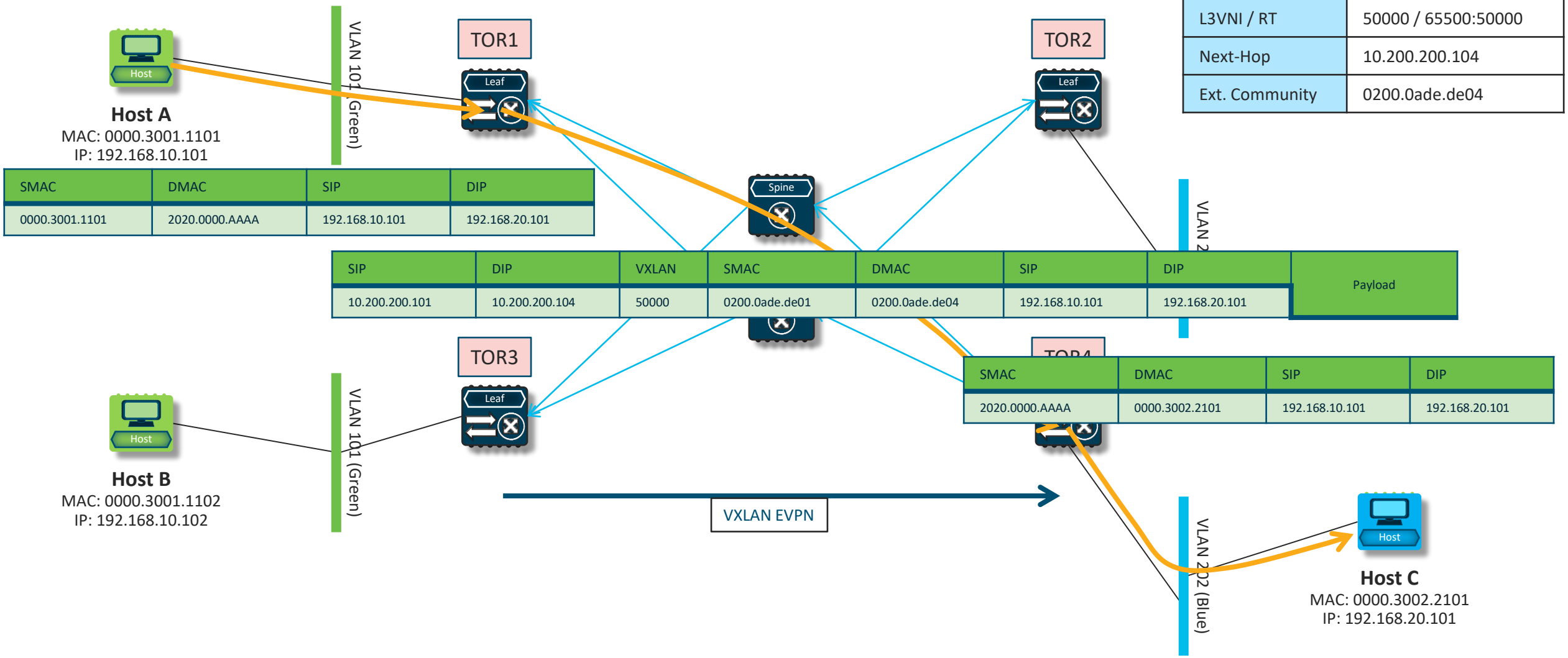
EVPN Control-Plane	
Type	2
MAC / Length	0000.3002.2101 / 48
L2VNI / RT	30002 / 65500:30002
IP / Length	192.168.20.101 / 32
L3VNI / RT	50000 / 65500:50000
Next-Hop	10.200.200.104
Ext. Community	0200.0ade.de04



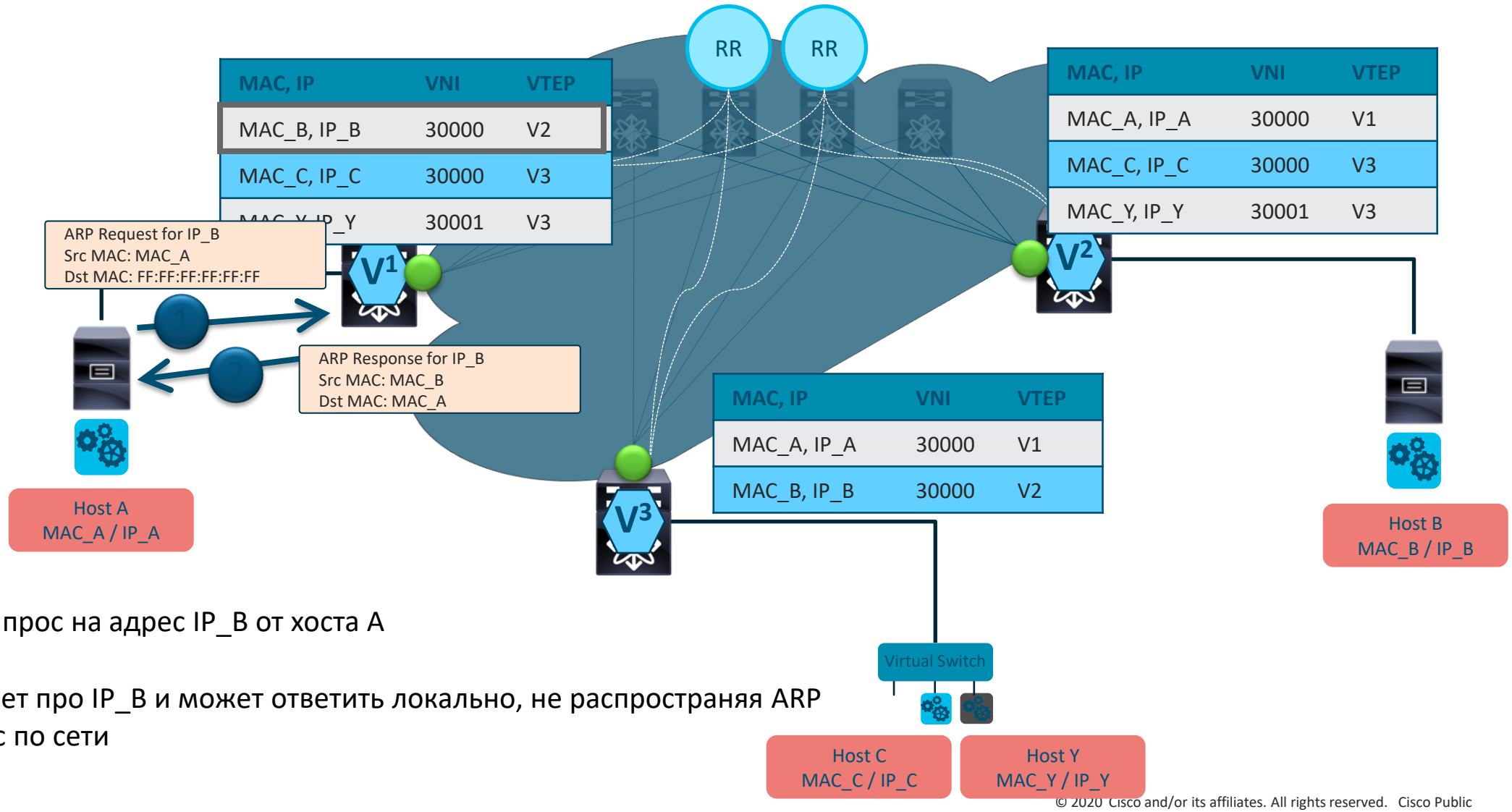
Packet Walk – Routing

«Молчаливый» хост обнаружен

EVPN Control-Plane	
Type	2
MAC / Length	0000.3002.2101 / 48
L2VNI / RT	30002 / 65500:30002
IP / Length	192.168.20.101 / 32
L3VNI / RT	50000 / 65500:50000
Next-Hop	10.200.200.104
Ext. Community	0200.0ade.de04



Подавление ARP
Дополнительная функция

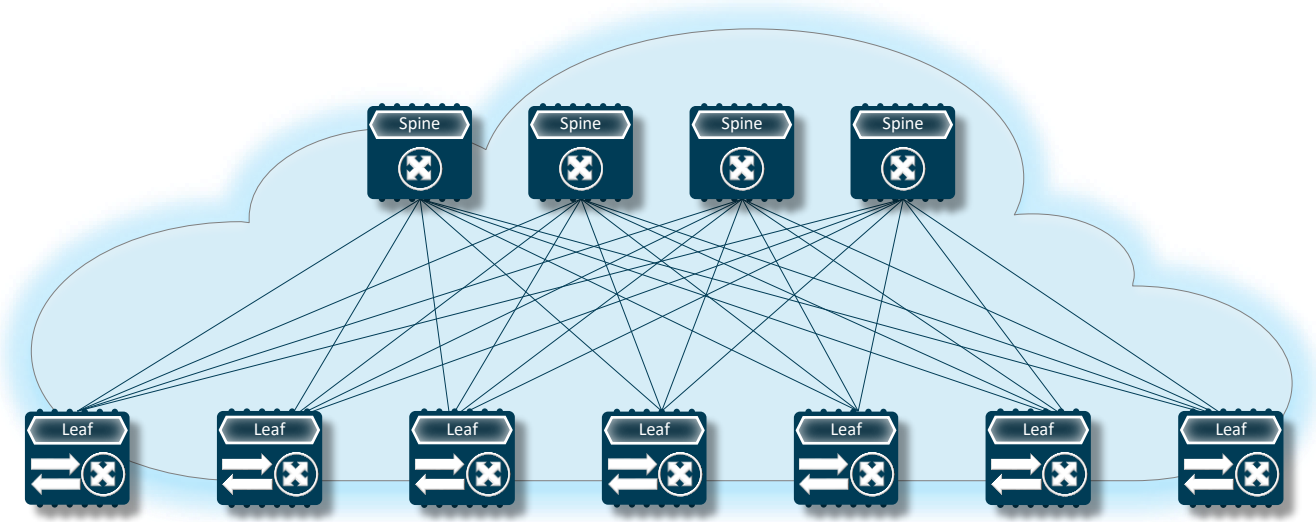


1. ARP запрос на адрес IP_B от хоста A
2. V1 знает про IP_B и может ответить локально, не распространяя ARP запрос по сети

Целостное решение: физическая и логическая архитектура

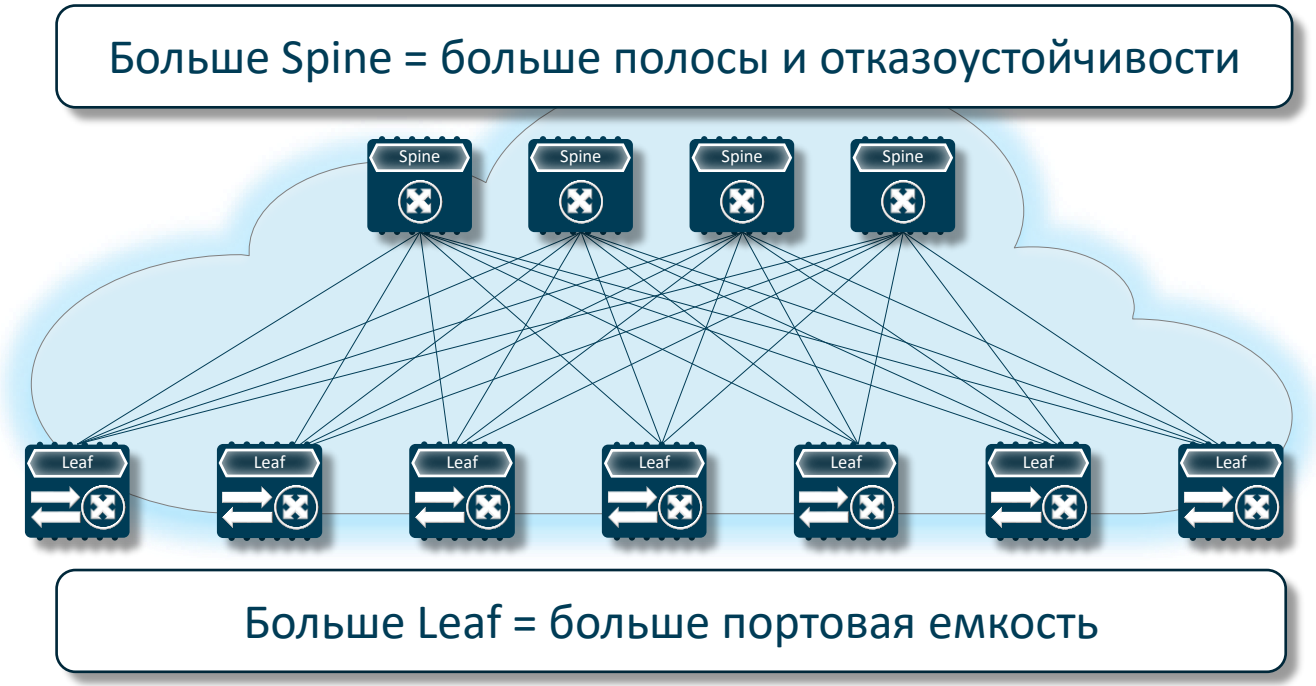
Типичная архитектура сетевой фабрики
Spine-Leaf, она же «сеть Клоза» (folded Clos)

- Использование ECMP для передачи данных: для Unicast или Multicast трафика
- Фиксированный «радиус» сети
- Детерминированная задержка
- Высокая отказоустойчивость



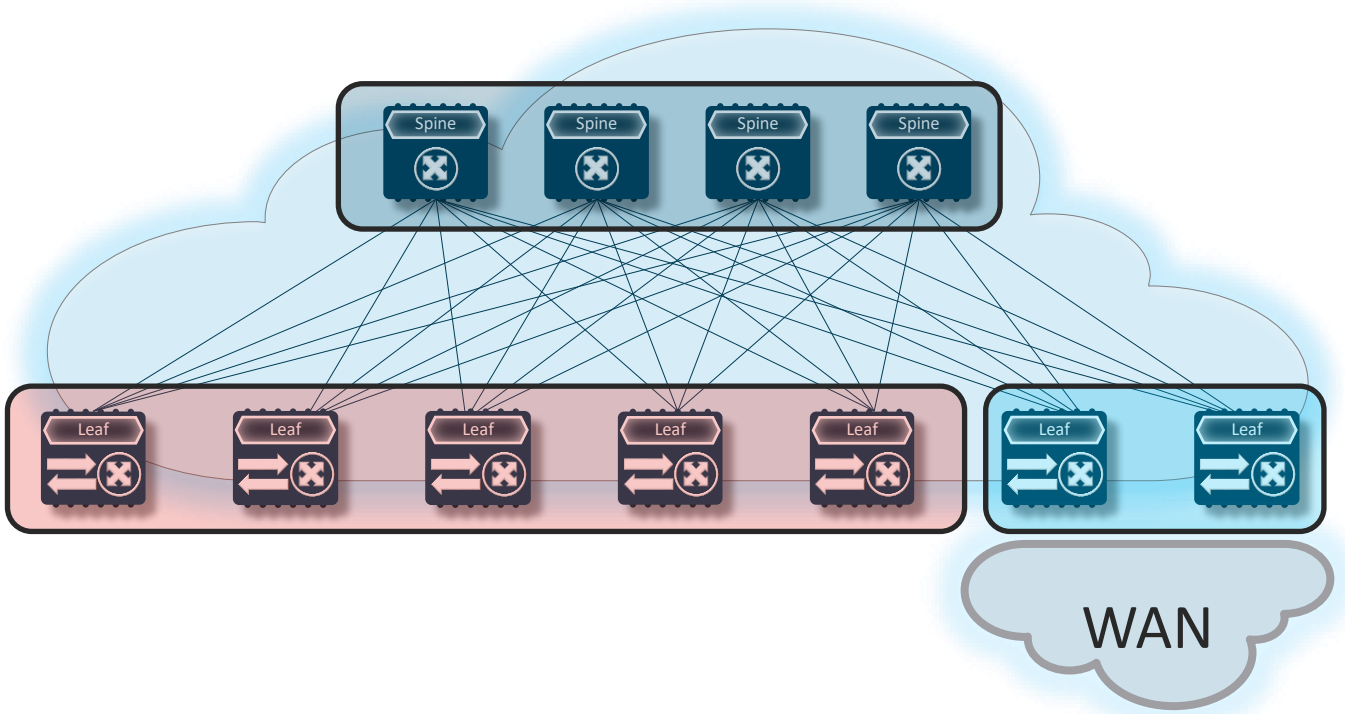
Сетевая фабрика
Масштабирование

- Leaf
  - Минимальная операционная единица
- Spines
  - Горизонтальный рост полосы
- Аплинки
  - Симметричные подключения
- SAYG: Scale as You Grow



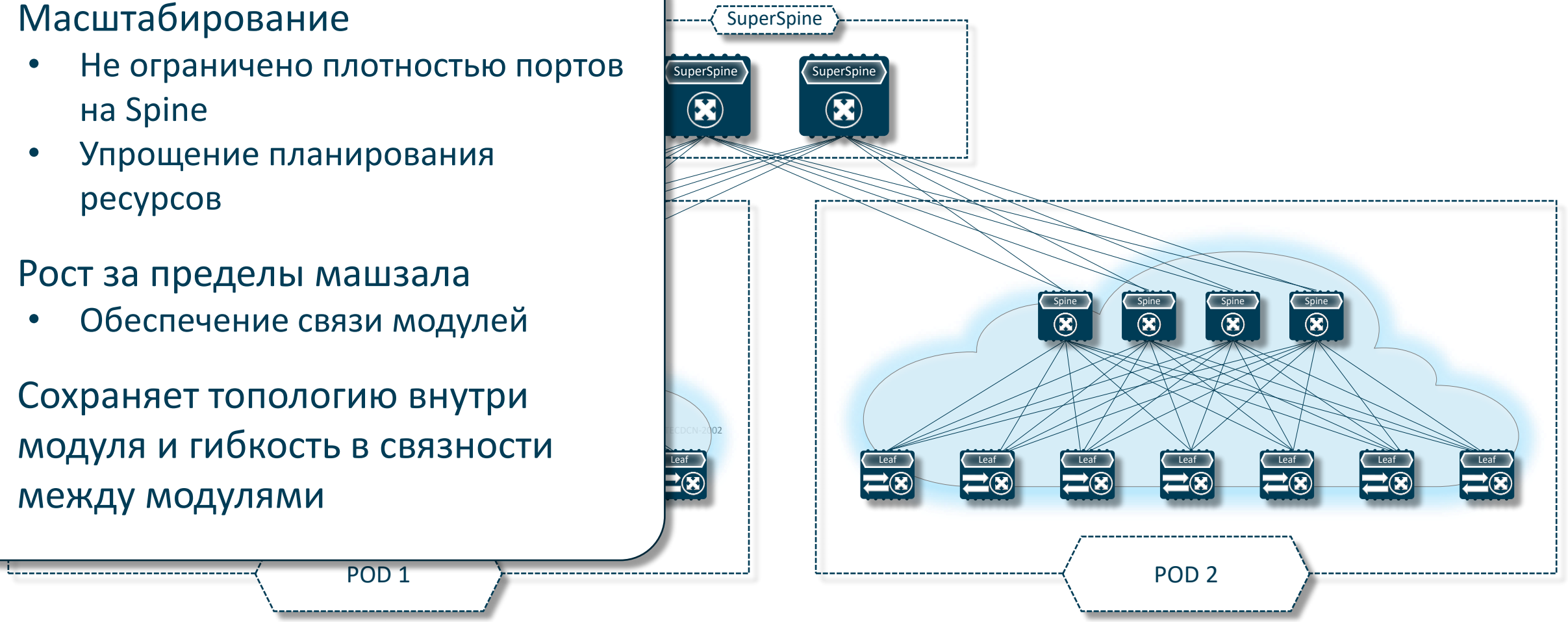
Сетевая фабрика
Внедрение VXLAN/EVPN

- **Spine**
  - Подключение всех Leafs and Border Leafs ко всем
- **Leaf (VTEP)**
  - Виртуальные машины
  - Физические сервера
  - FEX-ы
  - Другие коммутаторы
  - UCS FI
  - Блейд коммутаторы
- **Border Leaf (VTEP)**
  - Внешние подключения



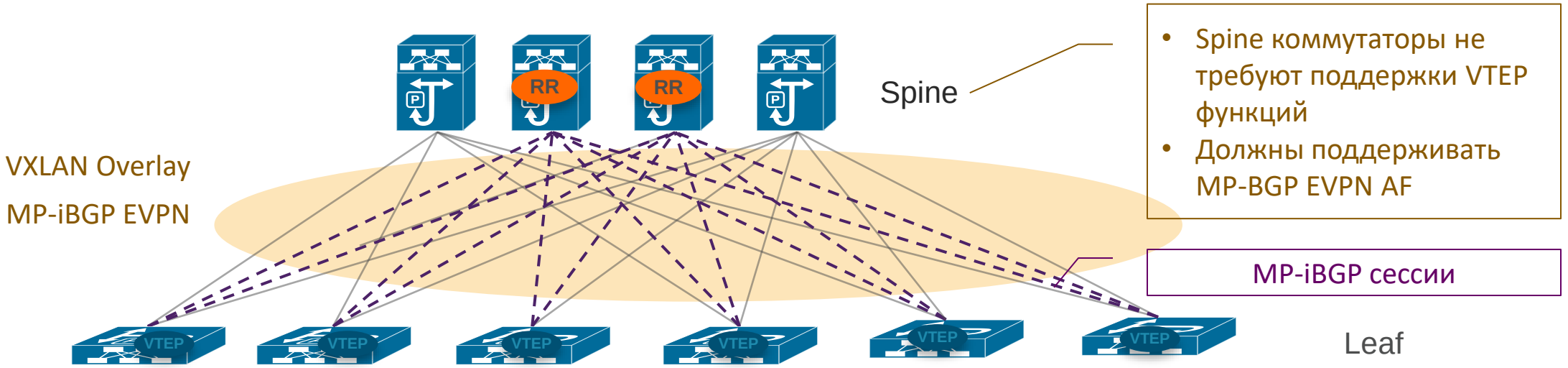
Super-Spine
Модульная Spine-Leaf архитектура

- Масштабирование
  - Не ограничено плотностью портов на Spine
  - Упрощение планирования ресурсов
- Рост за пределы машзала
  - Обеспечение связи модулей
- Сохраняет топологию внутри модуля и гибкость в связности между модулями



Дизайн MP-BGP

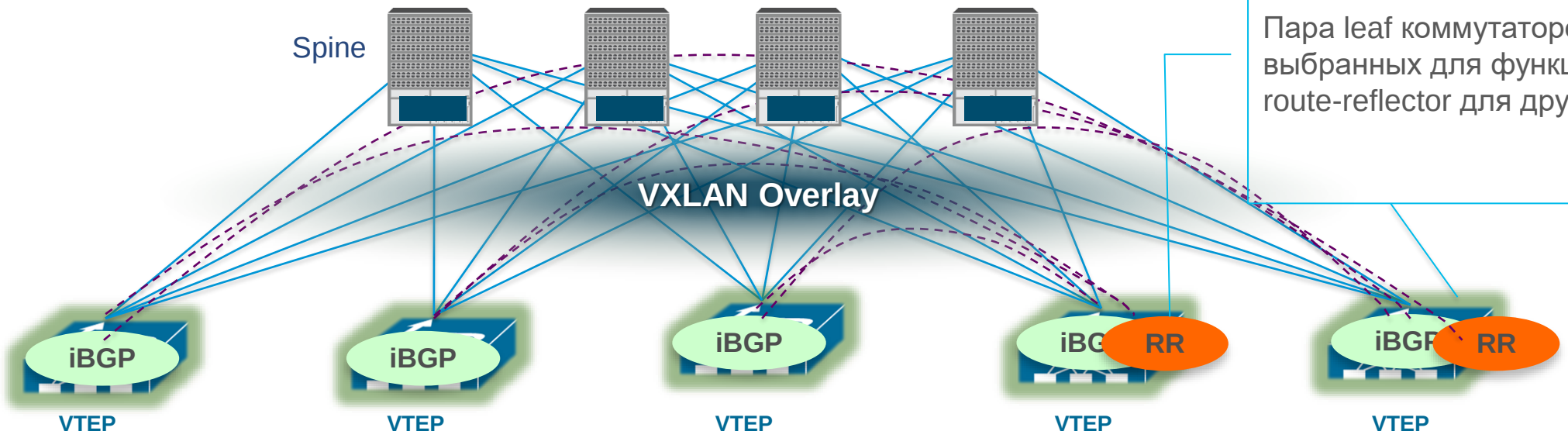
Типичный дизайн VXLAN фабрики с MP-iBGP EVPN



- VTEP только на уровне leaf
- Spine коммутаторы являются iBGP RR
- Spine коммутаторы не требуют поддержки VTEP функций
- iBGP сессии между VTEP и RR
- IGP протокол для IP достижимости между loopback адресами

Варианты дизайна VXLAN фабрики с MP-iBGP EVPN

- Spine коммутаторы не требуют поддержки VTEP функций
- Не требуют и поддержки EVPN
- Чистый IP транспорт

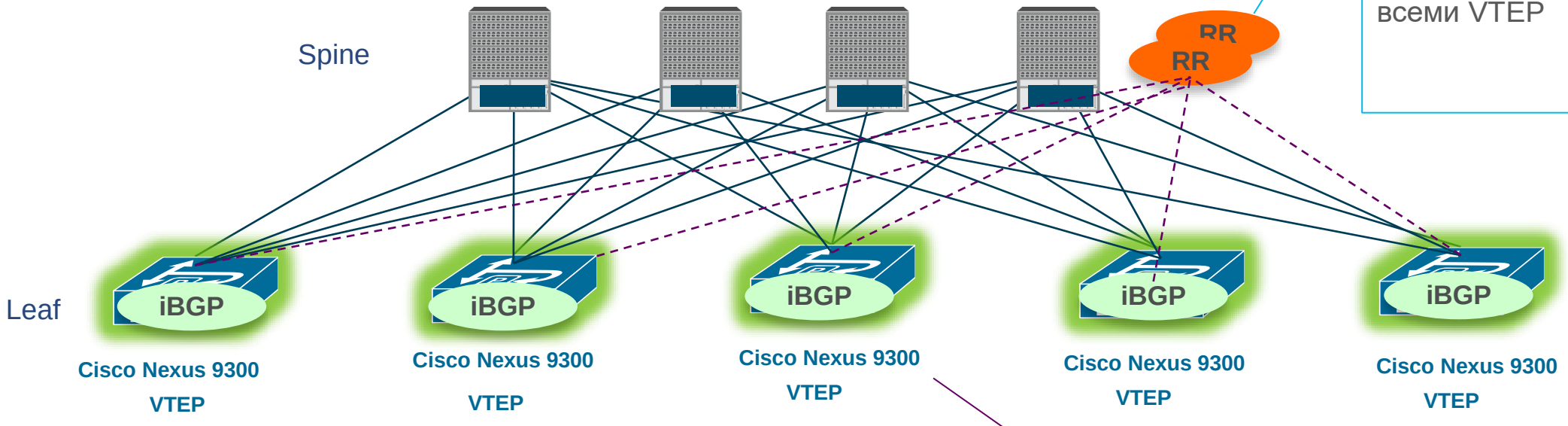


Пара leaf коммутаторов выбранных для функций iBGP route-reflector для других VTEP

Варианты дизайна VXLAN фабрики с MP-iBGP EVPN

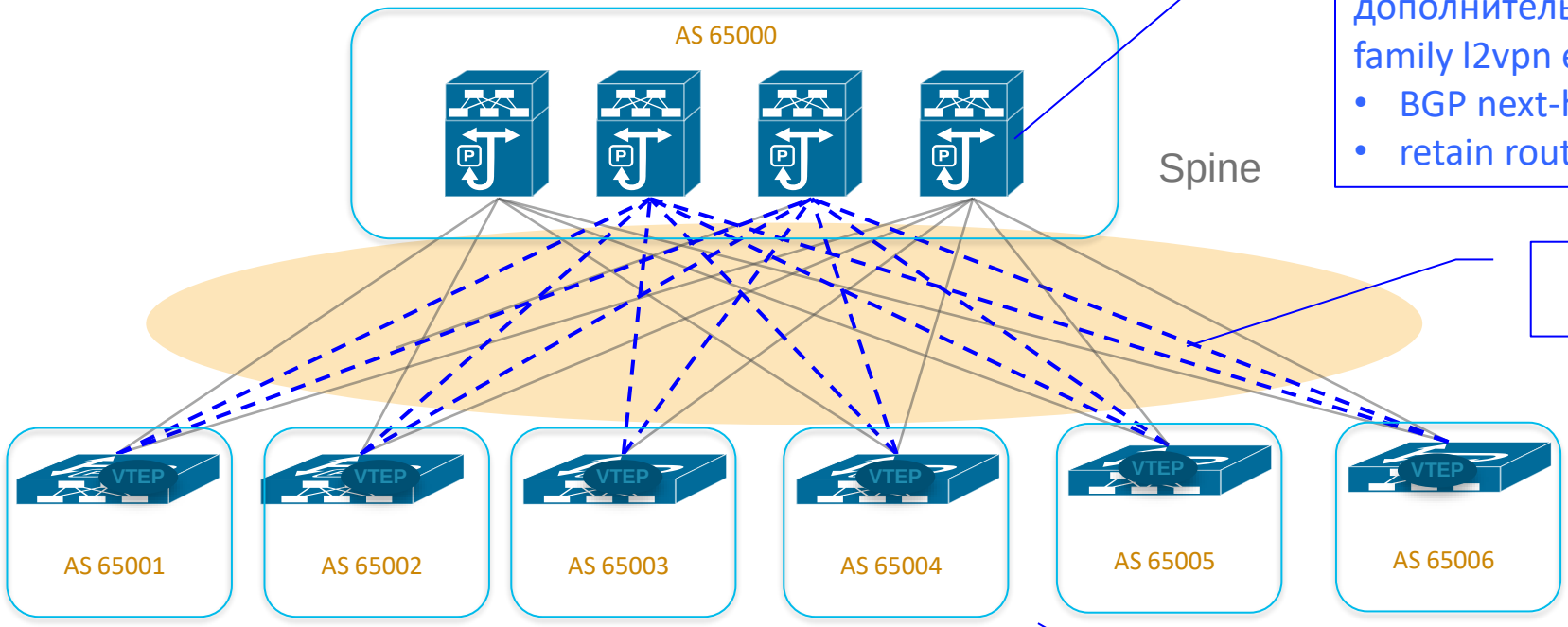
- Spine коммутаторы не требуют поддержки VTEP функций
- Не требуют и поддержки EVPN
- Чистый IP транспорт

Пара выделенных iBGP route-reflector-ов для пиринга со всеми VTEP



All leaf VTEPs run iBGP sessions with the dedicated route reflectors.

Дизайн VXLAN фабрики с MP-eBGP EVPN Multi-AS



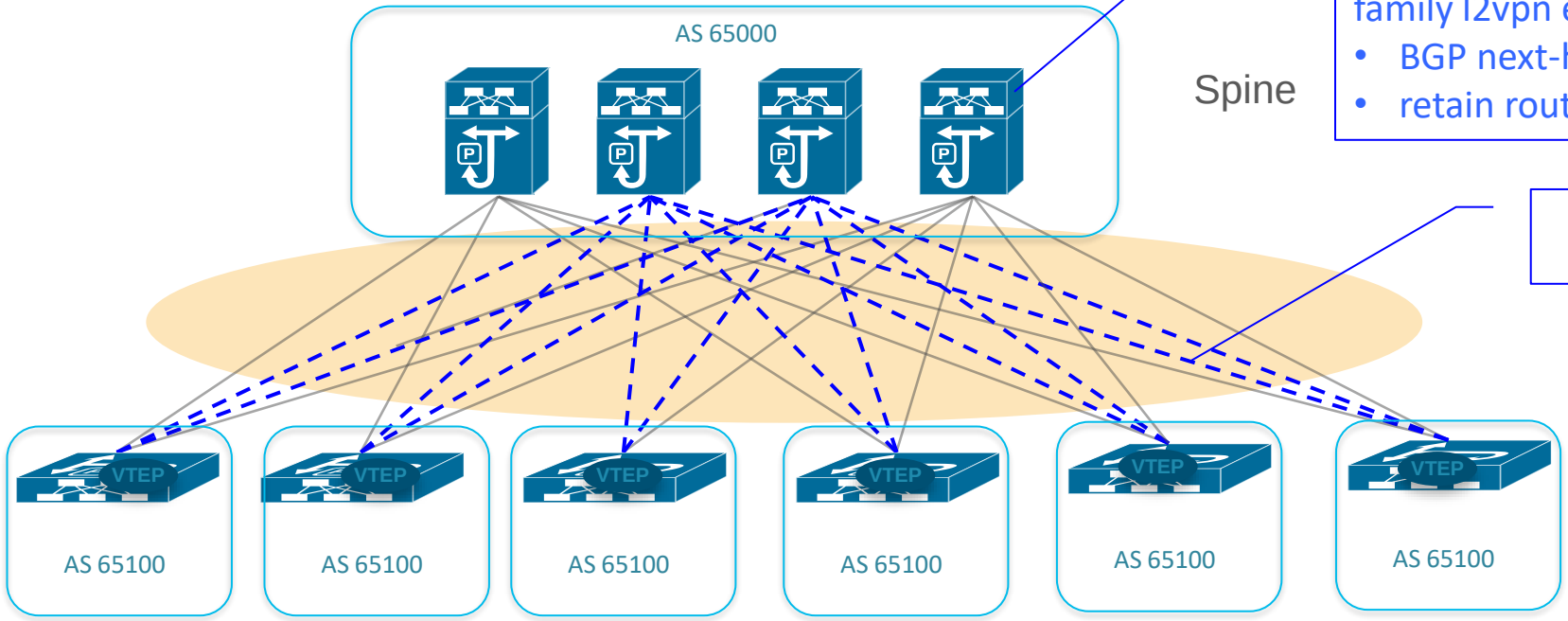
В BGP на Spine необходима дополнительная настройка в address-family l2vpn evpn:

- BGP next-hop unchanged
- retain route-target all

MP-eBGP Sessions

Нужно вручную настроить RT на каждом VTEP-е

Дизайн VXLAN фабрики с MP-eBGP EVPN
Dual-AS



В BGP на Spine необходима дополнительная настройка в address-family l2vpn evpn:
- BGP next-hop unchanged
- retain route-target all

MP-eBGP Sessions

Все VTEP (leafs) в одной BGP AS

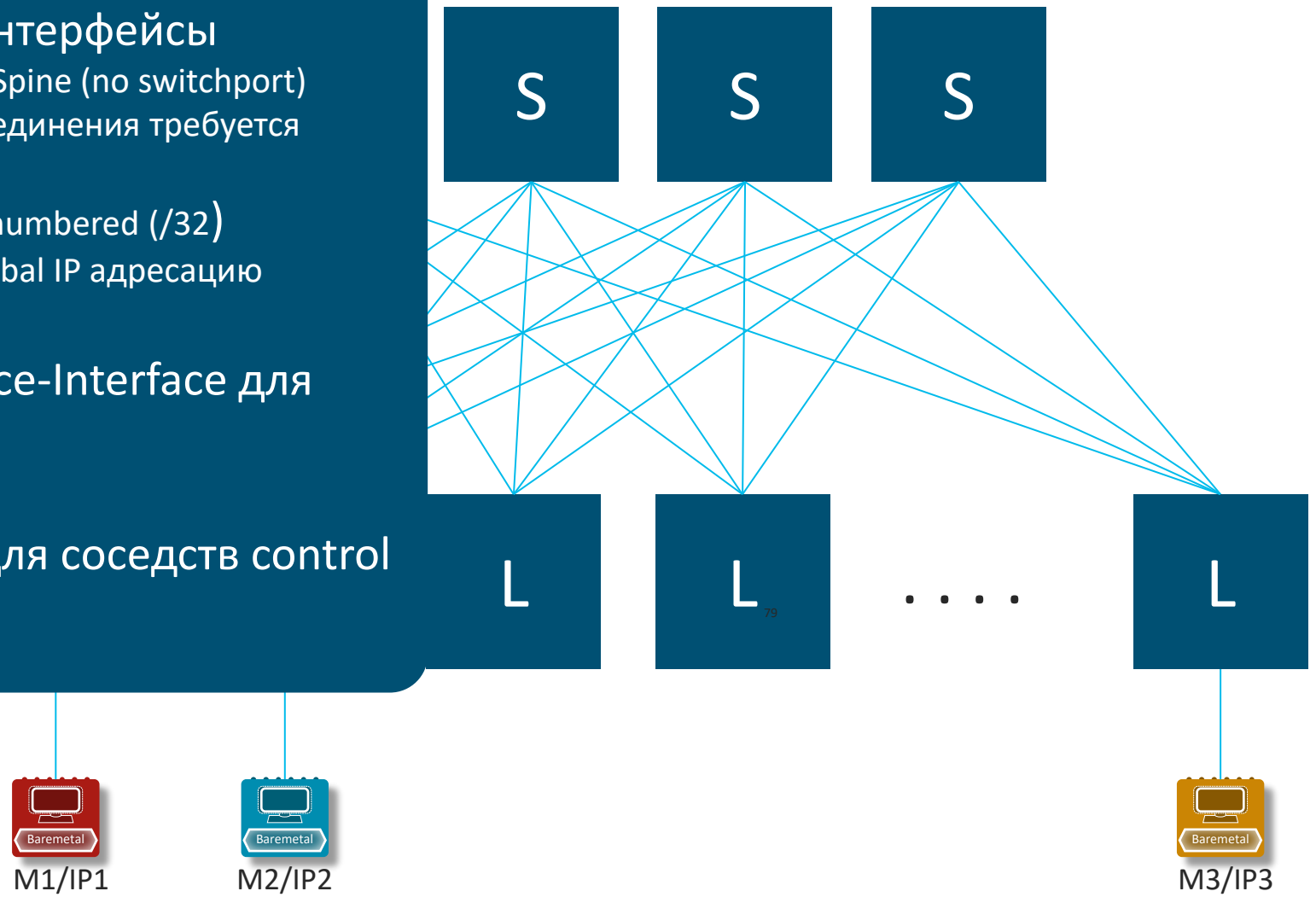
Дизайн Underlay связности

Свойства underlay сети и требования к ней

- Минимальная сложность
  - Минимум маршрутов
  - Только достижимость VTEP и инфраструктурные маршруты
- ECMP
  - Equal Cost Multi-Path для масштабируемости и доступности
- Максимально быстрая сходимость
- Отделение настройки и эксплуатации underlay и overlay
- Не забудьте про MTU с учётом VXLAN заголовка!
- Проблемы в underlay => проблемы в overlay !!!

Интерфейсы и адресация

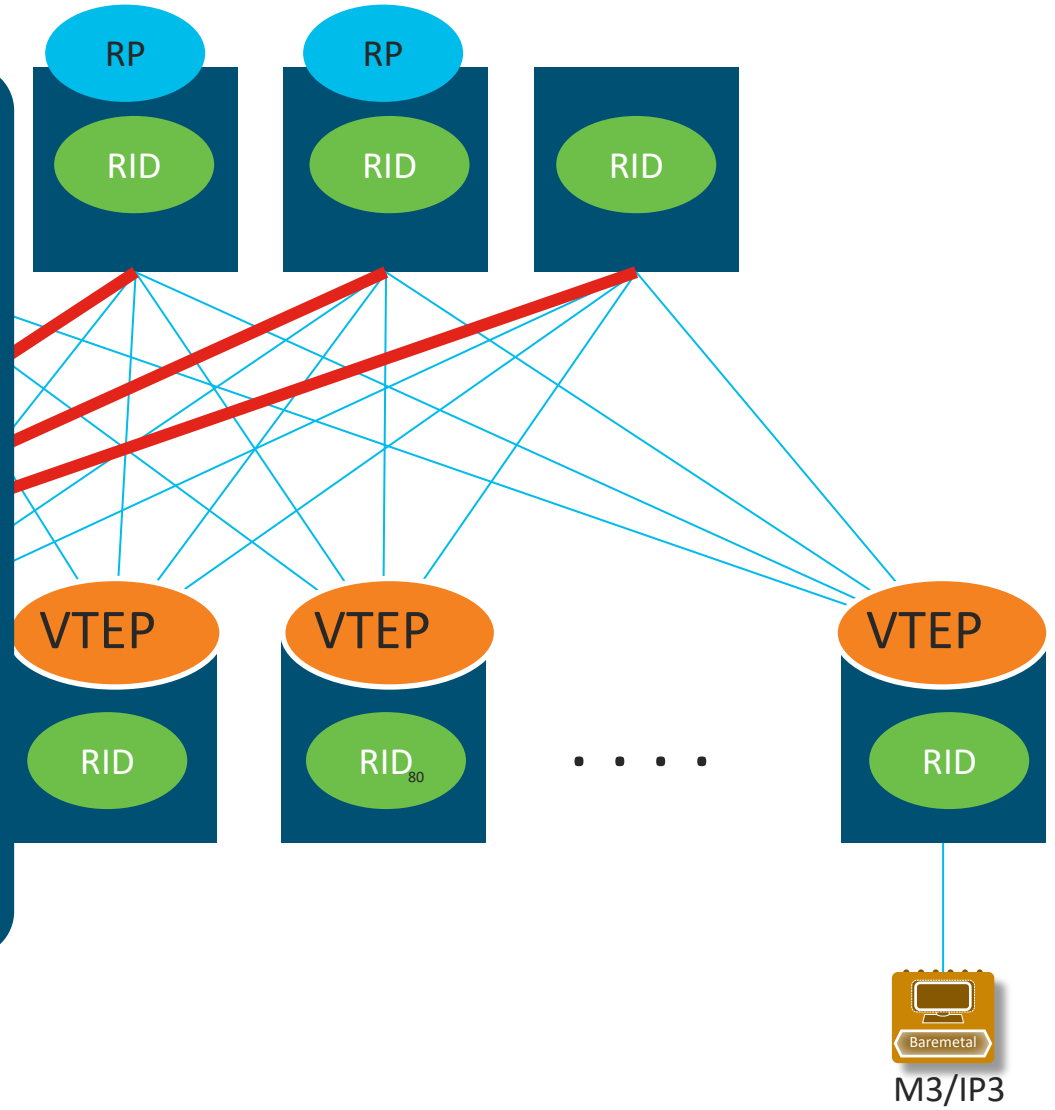
- Маршрутизируемые порты и интерфейсы
  - Интерфейсы Layer-3 между Leaf и Spine (no switchport)
  - Для каждого Point-2-Point (P2P) соединения требуется минимум /31 (IPv4)
  - Альтернатива, использовать IP Unnumbered (/32)
  - IPv6 поддерживает Link-Local и Global IP адресацию
- Используйте Loopback как Source-Interface для VTEP (NVE)
- Используйте *другие* Loopback для соседств control plane



Интерфейсы и адресация

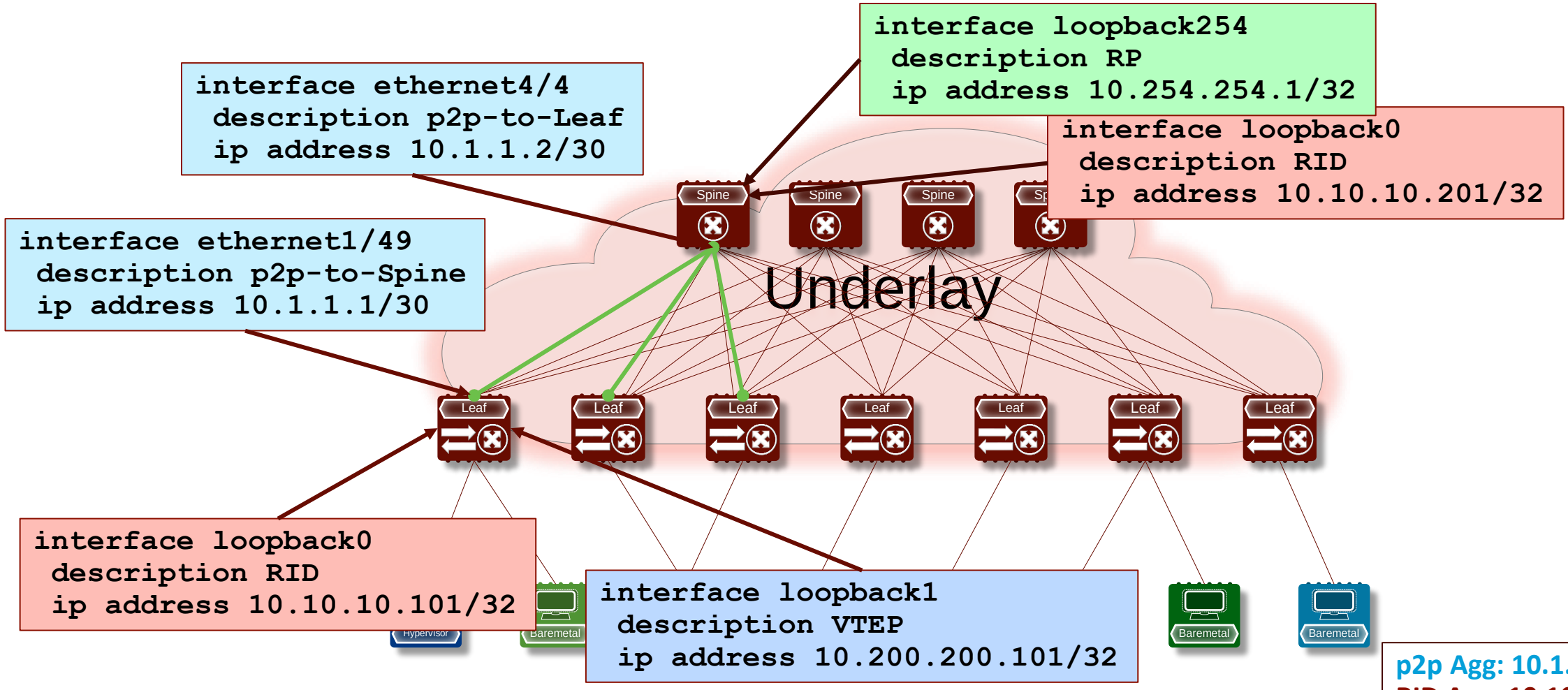
p2p Agg: 10.1.1.0/24

- Подготовьте план IP адресации
  - Начиная с NX-OS 9.3(1) поддерживается и IPv6
- Разделите IP адреса интерфейсов по функциям в разные диапазоны (агрегаты)
  - Unicast Routing – Routing Protocol Peering (p2p)
  - Unicast Routing – Routing Identifier (RID)
  - VTEP и VPC
  - Multicast Rendezvous-Point (RP)



10.254.254.1/32

Пример IP адресации



p2p Agg: 10.1.1.0/24
RID Agg: 10.10.10.0/24
VTEP Agg: 10.200.200.0/24
RP Agg: 10.254.254.0/24

Маршрутизация underlay – OSPF или IS-IS

- OSPF: обратите внимание на Network Type
- Network Type Point-2-Point
  - Предпочтительно (только LSA type-1)
  - Нет выборов DR/BDR
  - Хорошо подходит для маршрутизируемых интерфейсов/портов (оптимально с точки зрения LSA DB)
  - Полный пересчет SPF в случае изменений на линках

- IS-IS: что за CLNS такой?....
  - Не зависит от IP (CLNS)
  - Хорошо подходит для маршрутизируемых интерфейсов/портов
  - Нет пересчета SPF в случае изменений на линках
  - Быстрая сходимость
  - Аналог единственной зоны – использование Level-2
  - Не все хорошо с ним знакомы…

Для информации:

Работы по оптимизации IGP: Flood Optimization и Leaf/Spine Networks

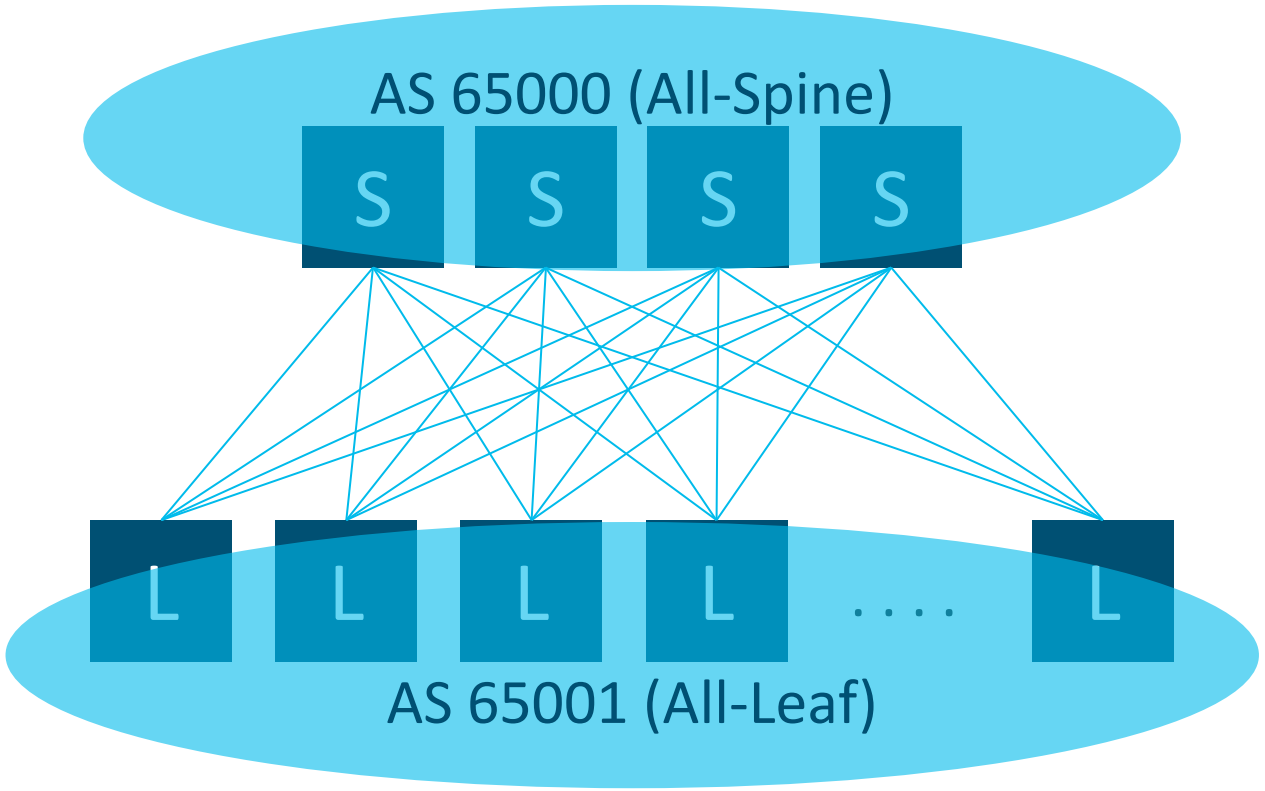
`draft-ietf-lsr-isis-spine-leaf-ext` & `draft-li-lsr-dynamic-flooding`

Маршрутизация underlay – BGP???

- eBGP Underlay маршрутизация «в духе SP/MSDC»
  - Придумана для не-оверлейных L3 фабрик (RFC7938)
    - Значительный объем «подстройки» для VXLAN фабрики
  - Две модели внедрения
    - Two-AS
    - Multi-AS
- BGP - Distance Vector (Path Vector) протокол
  - Знает только про AS_Path, не про скорости или стоимость
  - Не забудьте включить ECMP
- Сохраняется ли «независимость» overlay/underlay??
- «Разный» пиринг
  - На физических интерфейсах для underlay
  - На Loopback-ах для EVPN (BFD не нужен)

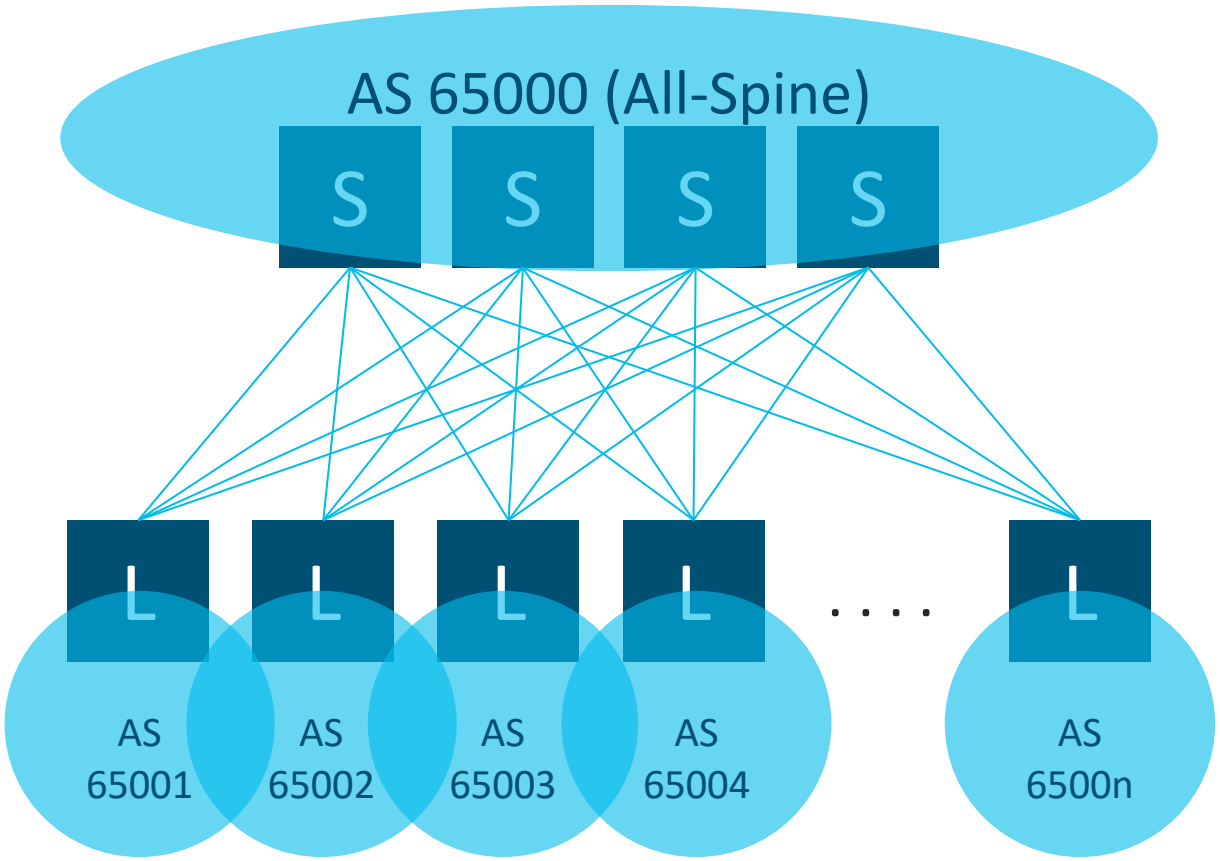
*Подробнее о том, почему eBGP для underlay и уровня управления overlay не очень хорошая идея:
https://learningnetwork.cisco.com/s/blogs/a0D3i000002eeaAEAQ/the-magic-of-superspines-and-rfc7938-with-overlays-guest-post

Unicast Routing – eBGP Two-AS Model



- eBGP Two-AS – да, так можно!
- eBGP пиринг для Underlay
  - Retain Route-Targets
  - Disable BGP AS-Path check
- Underlay отвечает за достижимость!
  - Анонсируйте Loopback-и
- Специальные требования для Overlay Control-Plane
  - Next-Hop Unchanged
  - Disable BGP AS-Path check

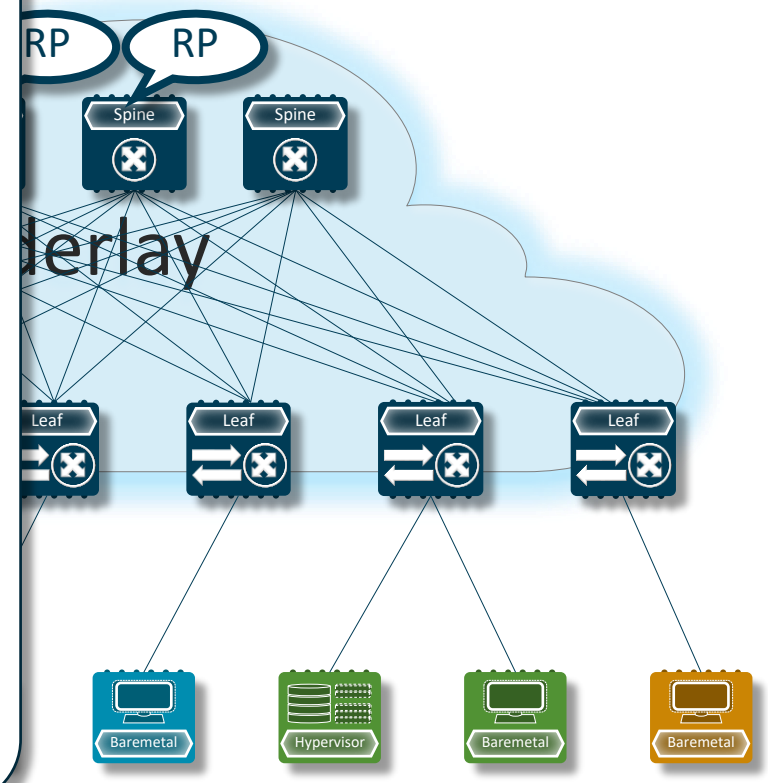
Unicast Routing – eBGP Multi-AS Model



- eBGP Multi-AS, воспроизводим Internet
- eBGP пиринг для Underlay
  - Retain Route-Targets
  - Disable BGP AS-Path check
  - Next-Hop Unchanged
- Underlay отвечает за достижимость!
  - Анонсируйте Loopback-и
- Специальные требования для Overlay Control-Plane
  - Next-Hop Unchanged

BUM репликация в Underlay с помощью multicast
PIM ASM

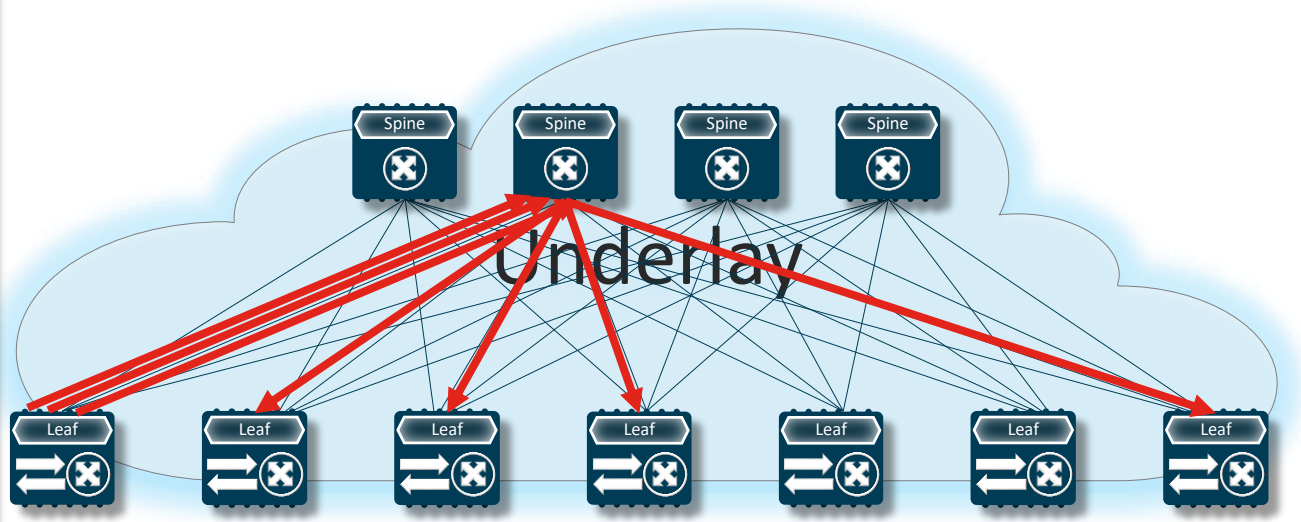
- PIM Any-Source-Multicast (ASM)
- Поддержка оборудованием
  - Nexus 9000 / Nexus 7000 (F3/M3)
  - ASR 1000 / ASR 9000
- Резервирование RP
  - PIM Anycast-RP или MSDP
- Source-Trees (однонаправленные)
  - 1 Source Tree для каждого VTEP для каждой Multicast группы



Настройка метода распространения BUM трафика выполняется для VNI
Разные методы multicast и ingress replication могут комбинироваться на одном устройстве

BUM репликация на VTEP: Ingress Replication

- «Умножение» пакетов
- Поддержка платформами
  - Nexus 9000
- Ingress Replication
  - Хост отправляет 1 пакет на VTEP
  - VTEP на входе инкапсулирует пакет и размножает его по числу VTEP в данном L2VNI
- Не требуется отдельного control plane



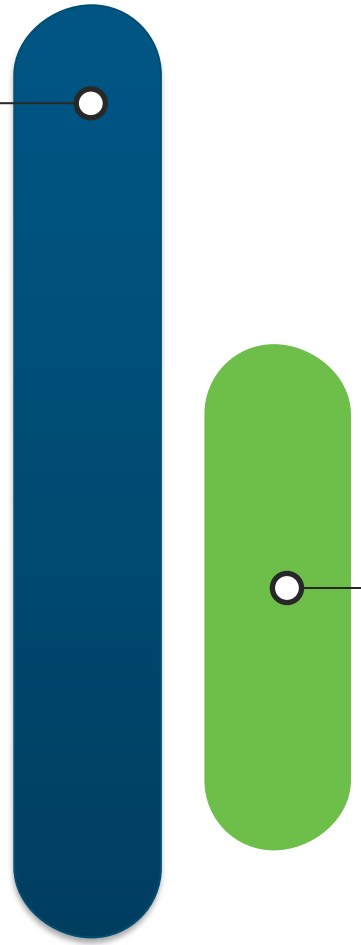
Настройка метода распространения BUM трафика выполняется для VNI
Разные методы multicast и ingress replication могут комбинироваться на одном устройстве

Управление VXLAN/EVPN фабрикой

Data Center Network Manager (DCNM)

Единый продукт, три режима эксплуатации

- VXLAN EVPN фабрика (с поддержкой традиционных технологий)
- IP Media Network Controller (PMN)
- SAN (MDS & Nexus)



Инструмент эксплуатации

- Real-Time Topology
- Integrated Compute Visibility
- Performance Monitoring
- Fault Management
- Configuration Compliance
- Image Management, Upgrades and RMA

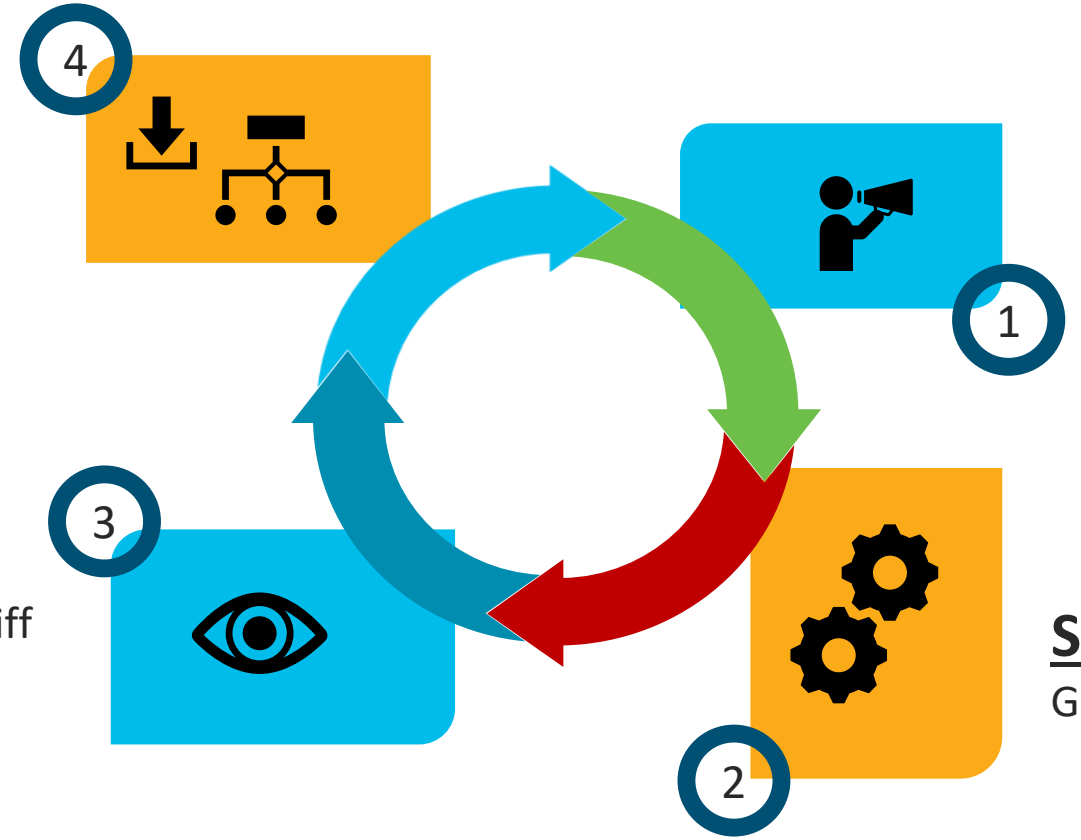
Сквозная настройка сетевых ресурсов

- GUI/API-based provisioning
- Multi-Fabric & Multi-Site
- Network Configuration Backup & Restore

Задачи «нулевого», «первого» и «второго» дня эксплуатации!

Режим LAN Fabric в DCNM 11

Deploy
Centralized config push



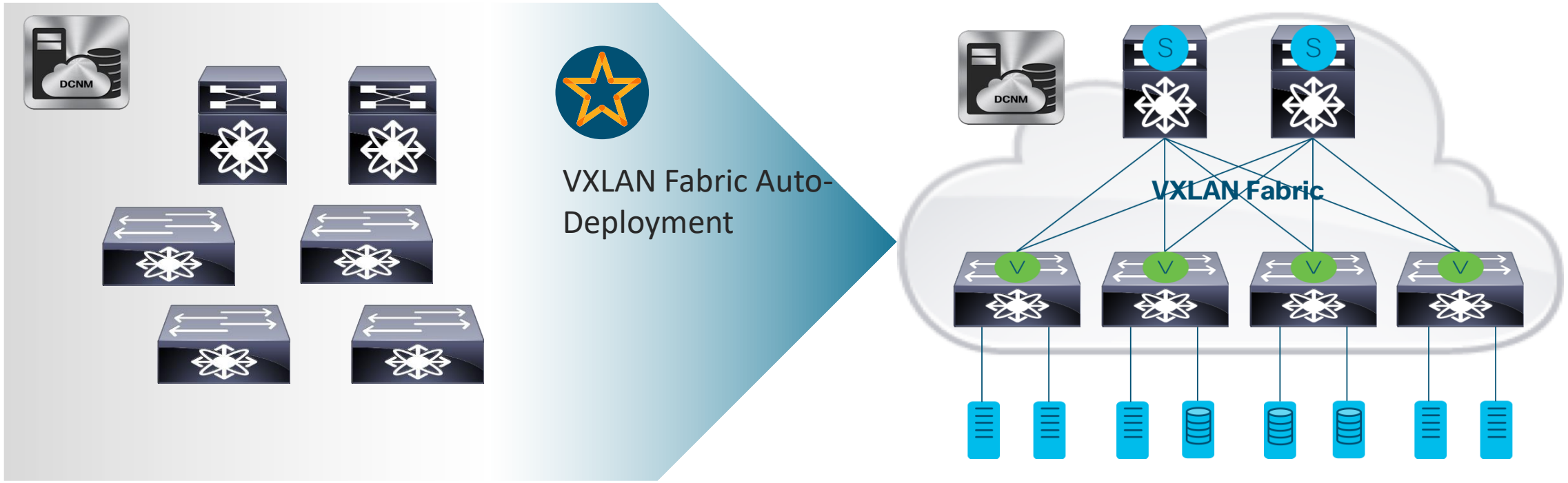
Define
Define Intent based on best practices
- Underlay
- Interfaces
- Overlay

Preview
Side-by-side diff

Save
Generates configuration based on intent

Фабрика готова за несколько минут!

Cisco DCNM: автоматическое развёртывание VXLAN фабрики
С использованием шаблонов и POAP



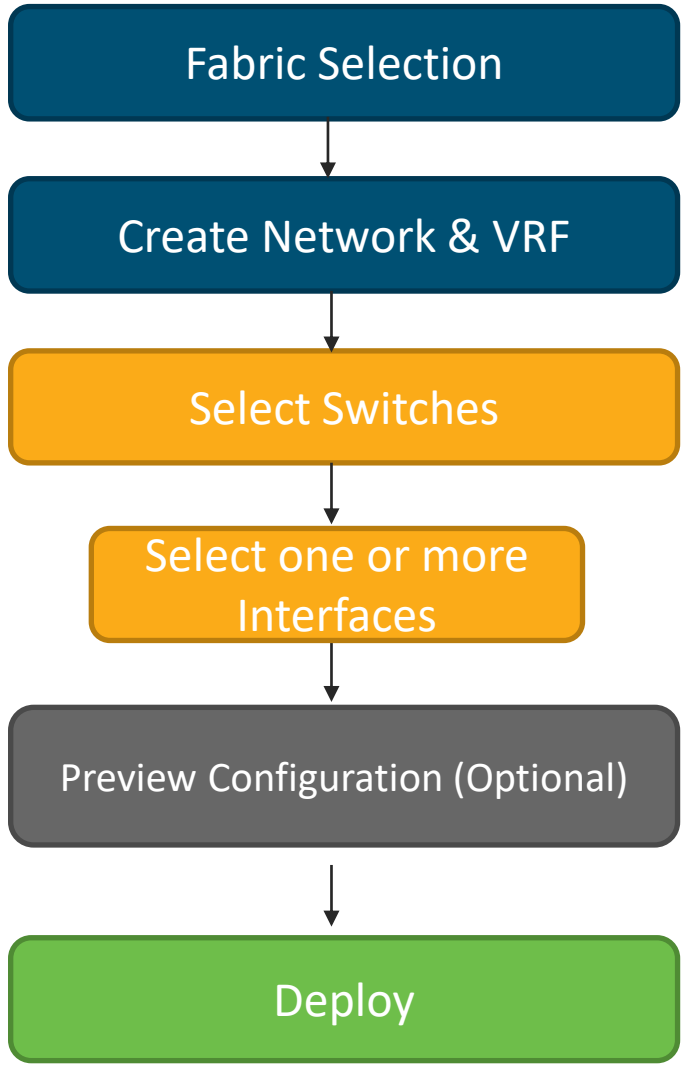
Turn-Key VTEP Deployment

Simplified Deployment

Managed-Fabric Operations

Управление оверлеями «сверху-вниз»

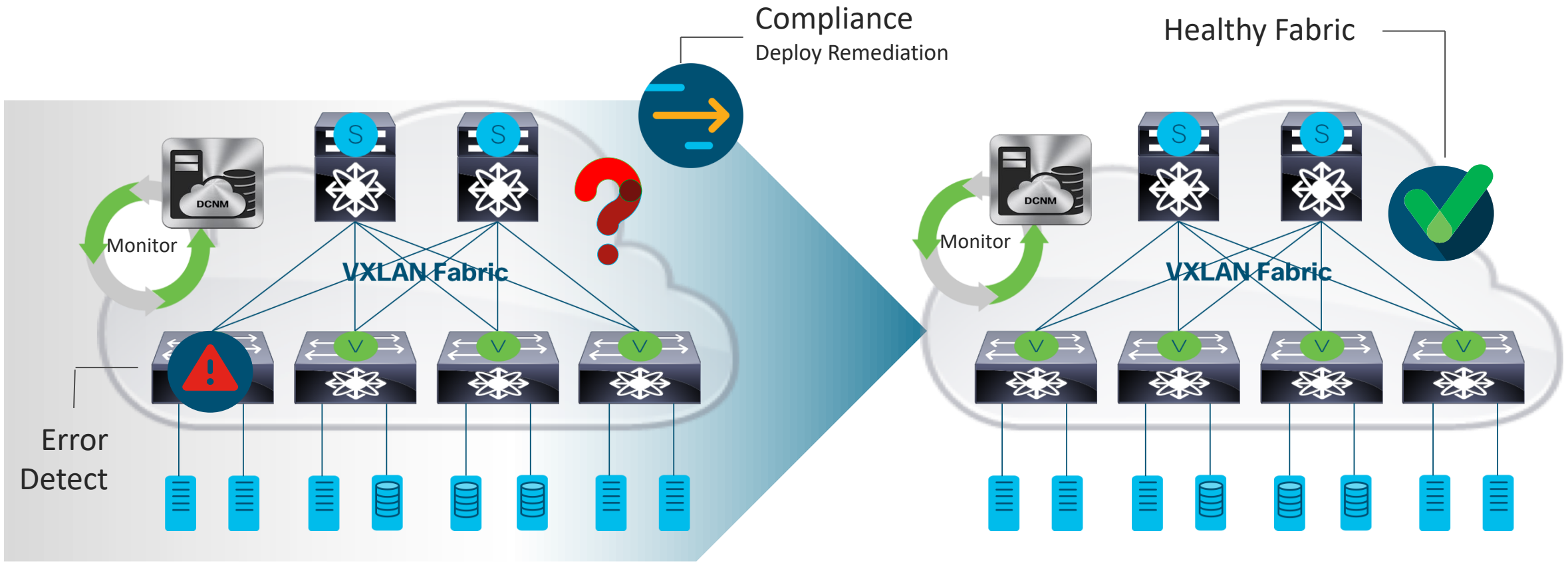
- Настройка Top-Down через GUI или REST APIs
- Создание сетей /VRF Creation на основе политик
- Настройка на коммутаторах и/или интерфейсах
- История настроек для коммутатора/сети
- Отслеживание ресурсов для VNIs, VLANs и т.д.
- Централизованное управлением всеми настройками, включая оверлеи



Настройка внешней связности для внедрения на основе VXLAN-EVPN

- Настройка внешней связности от Border к WAN
  - VRF Lite с использованием субинтерфейсов
- Настройка Layer-2 и Layer-3 DCI связности с помощью EVPN Multi-Site
  - Настройка внешнего пиринга для Multi-Site Underlay и Overlay с использованием Multi-Site Domain (MSD)
  - MSD – «фабрика фабрик»
    - Определение сетей и VRF один раз
    - Поддержка TRM

Fabric Compliance – Underlay / Overlay / Access



Compliance

Fabric Reliability & Visibility

Operations Confidence

Эксплуатационные функции в DCNM

- Мониторинг
  - Отказы, предупреждения
  - Температура
  - Статистика
  - Интеграция с VMM и контейнерными доменами
  - Мониторинг endpoints в MP-BGP EVPN таблицах
  - …
- Запуск дополнительных приложений для эксплуатации
  - Nexus Insights


CISCO