Сеть ЦОД на основе VXLAN/EVPN за 5 дней

День 4: VXLAN/EVPN фабрика – интеграция L4-L7 сервисов

Виктор Пустошилов
Системный Архитектор
08.04.2021

Обсуждение – в Telegram канале:

Программа спринта

День	Тема
5 апреля, понедельник	VXLAN/EVPN фабрика – основы
6 апреля, вторник	VXLAN/EVPN фабрика – клиенты и внешние сегменты, расширенные функции NX-OS
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

Содержание

- Часть 1: Варианты подключения сервисов. Подключение сервисов без Policy Based Routing.
- Часть 2: Подключение сервисов при помощи Policy Based Routing.
  - Demo
- Часть 3: Подключение виртуализированных сервисных функций.
- Часть 4: Подключение сервисов при помощи Enhanced PBR.

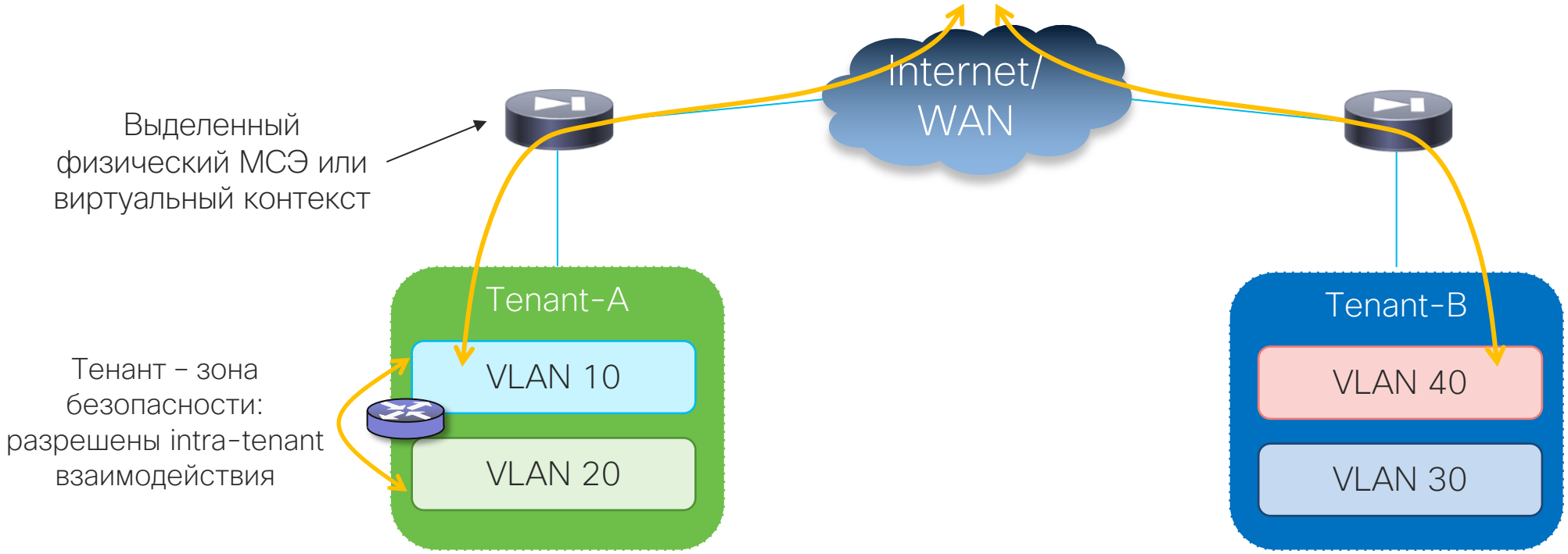
Часть 1: Варианты подключения сервисов

Требования сетевых сервисов

- В ЦОД сервисные устройства как правило работают в одном из двух режимов:
  - Transparent – Layer 2 (или GO THROUGH)
  - Routed – Layer 3 (или GO TO)
    - Шлюз по умолчанию для подсети настраивается на сервисном устройстве (наиболее используемая опция)
    - Шлюз по умолчанию для подсети настраивается на сети и сервисный узел является next hop-ом.
- Выбор режима напрямую влияет на используемые настройки сети
- Необходимо выяснить требования к сетевому сервису (intra-tenant, inter-tenant)

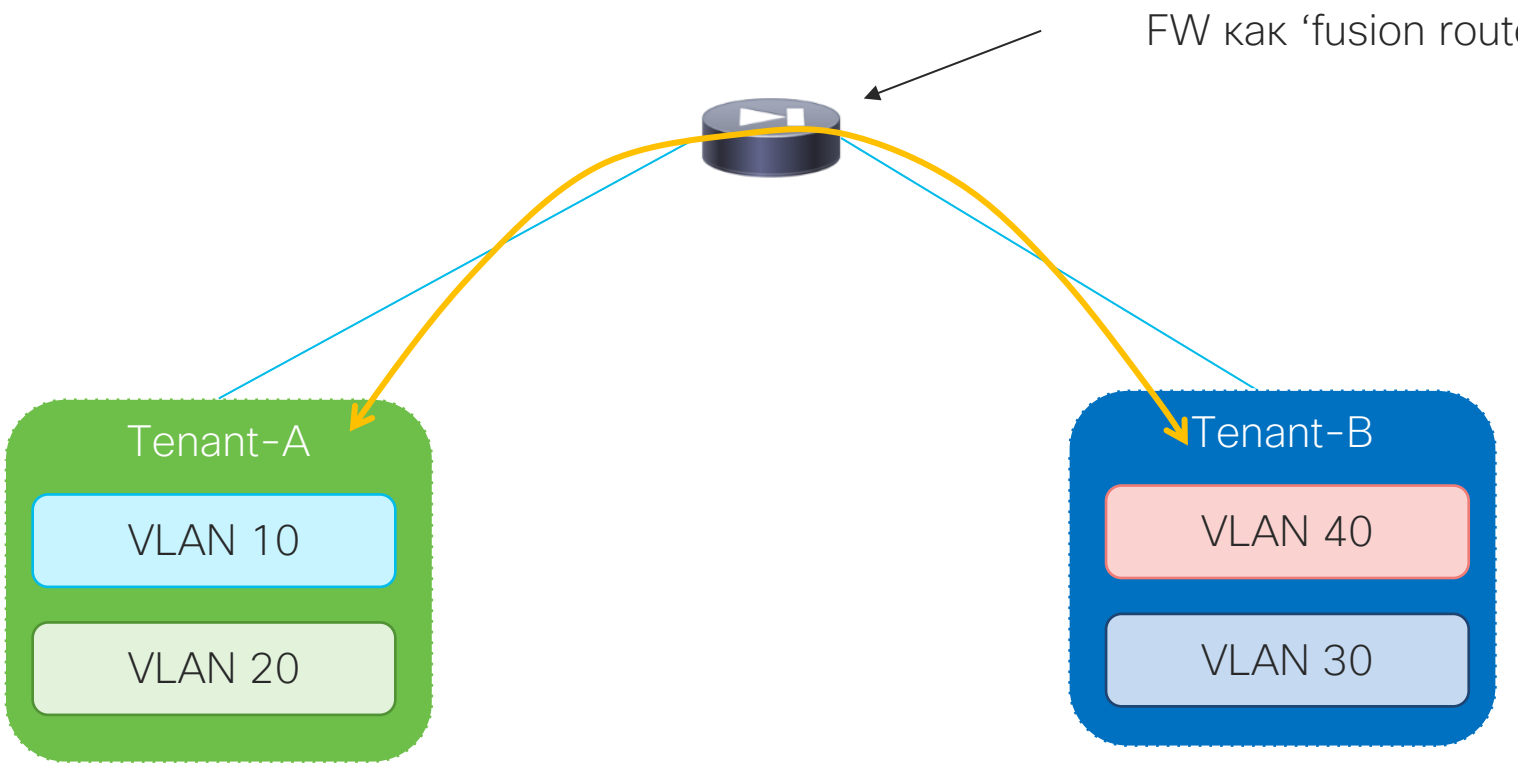
Сервис на периметре сети – Tenant Edge Service

- Фильтрация и применение политик между сетью ЦОД и внешним миром – **между зонами безопасности**



Сервис между сетями тенантов – Inter Tenant Services

- Фильтрация и применение политик между тенантами ЦОД (Inter-VRF) – между зонами безопасности



Сервис внутри сегментов тенанта – Intra Tenant Services

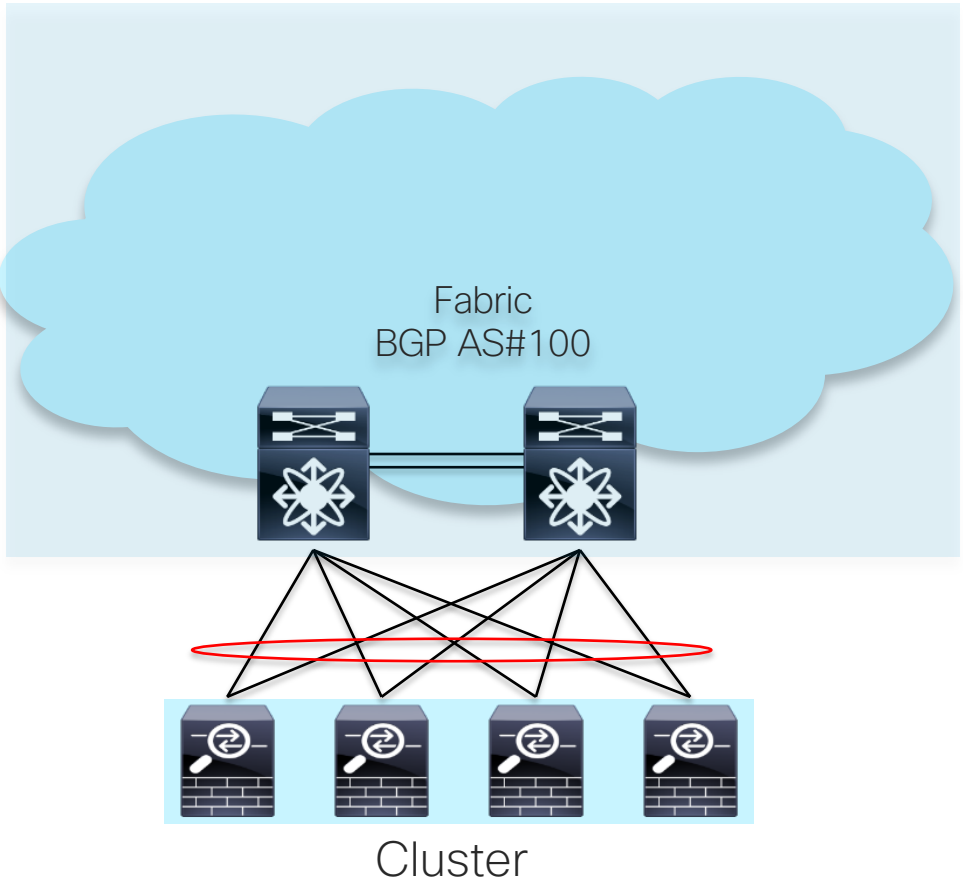
- Фильтрация и применение политик между или внутри сегментов одного тенанта
  - Intra-VRF, inter-subnets

Опция 1 : FW – шлюз по умолчанию
Опция 2 : PBR с FW как L3 hop
Опция 3 : FW в транспарент режиме (редко используется)

Проще говоря

- Tenant Edge Services → Между VRF
- Inter Tenant Services → Между VRF
- Intra Tenant Services → Внутри VRF/Между VLAN

Как физически подключаются сервисные устройства



Cluster

Для кластерных систем подойдет vPC

(Узлы кластера должны подключаться к общей vPC паре коммутаторов)

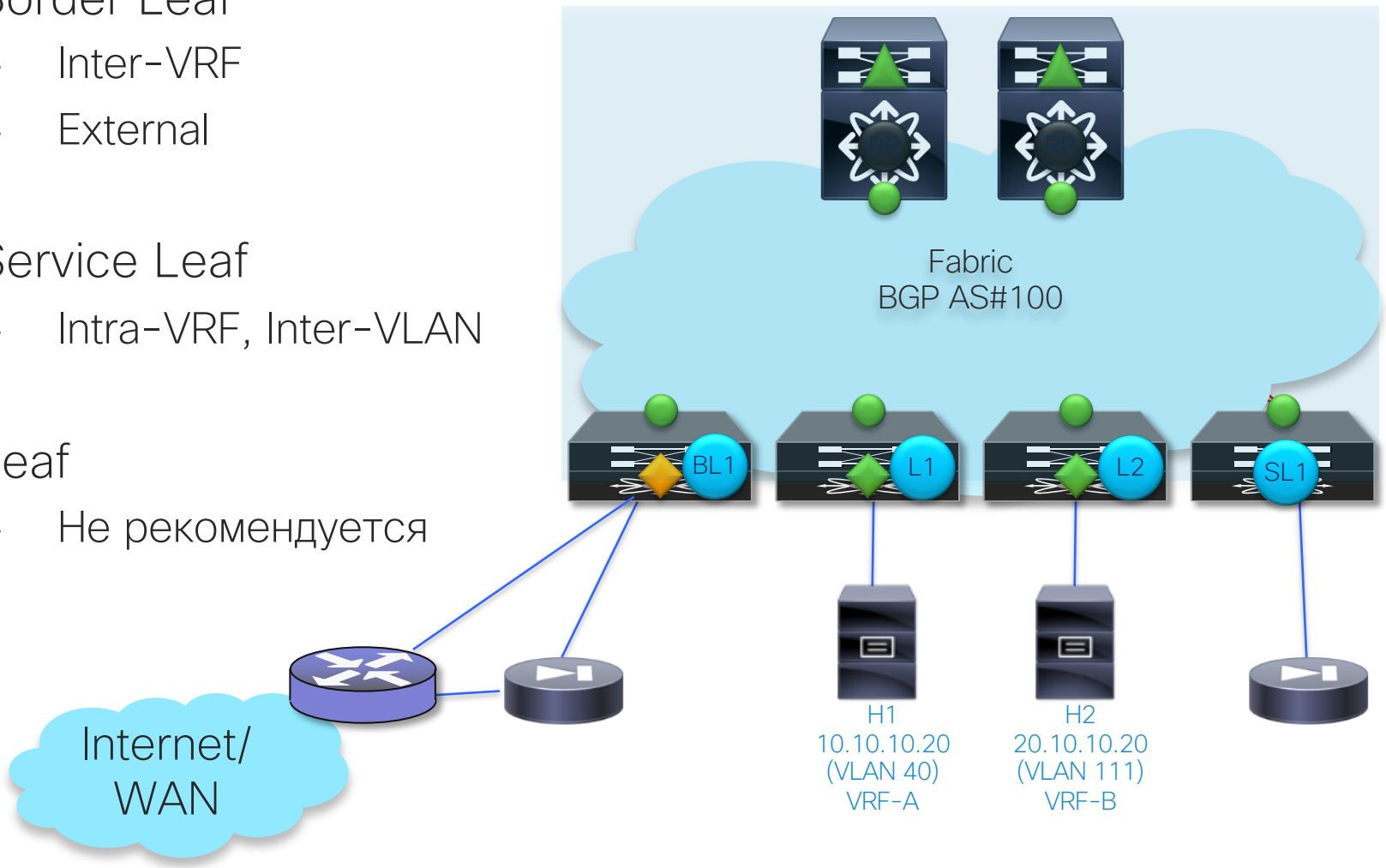
Active/Standby

Для Active/Standby систем vPC не лучший выбор.

(L3 peering через vPC, Multicast routing через vPC)

Где и как подключать разные роли сервисов

- Border Leaf
  - Inter-VRF
  - External
- Service Leaf
  - Intra-VRF, Inter-VLAN
- Leaf
  - Не рекомендуется



SL Service Leaf

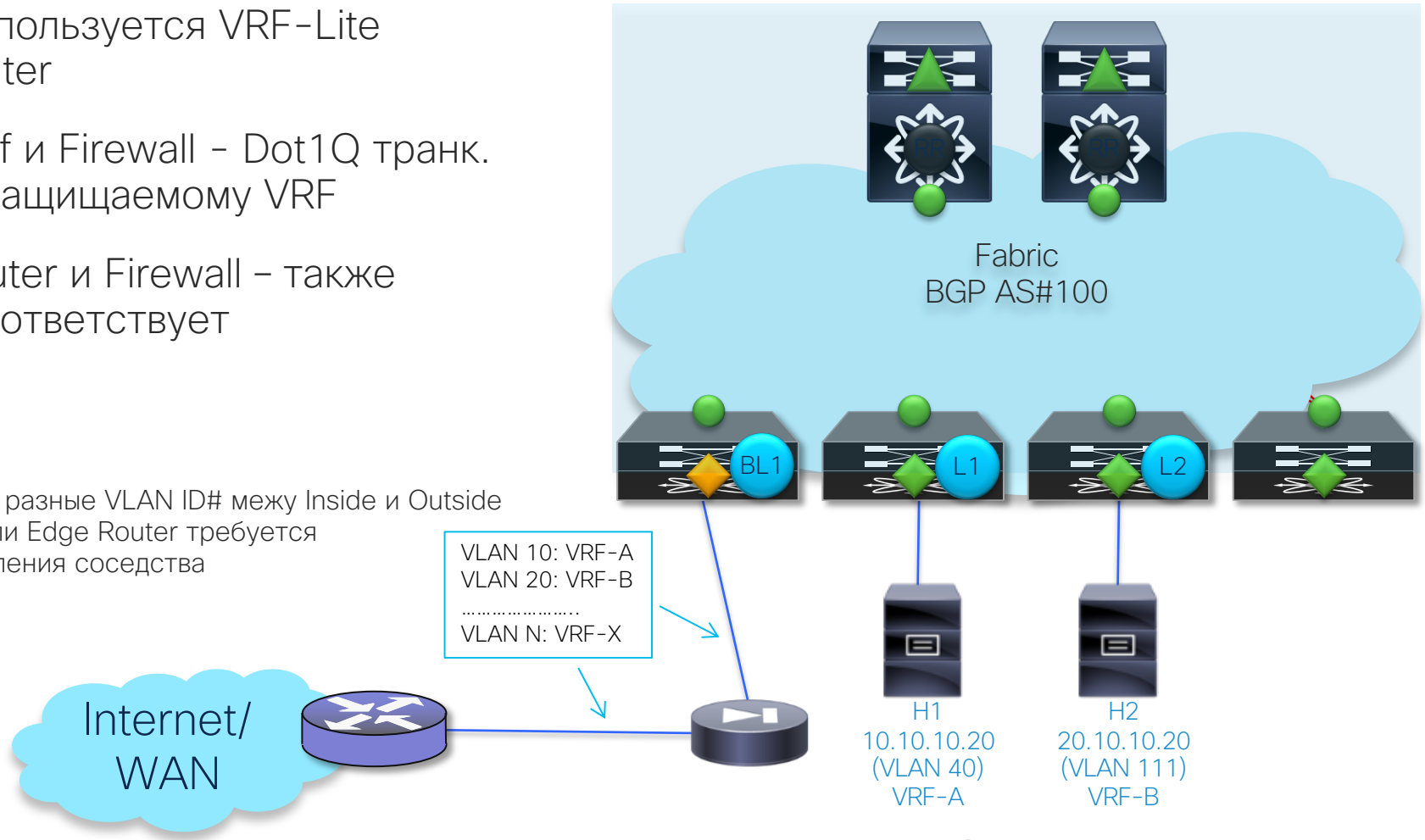
L Leaf

BL Border Leaf

Подключение сервисов без Policy Based Routing.

Inter-VRF Firewall: Transparent Mode

- Transparent Firewall устанавливается в разрыв между Border Leaf и Edge-Router. Используется VRF-Lite между Border Leaf и Edge-Router
- Inside канал между Border Leaf и Firewall - Dot1Q транк. Каждый VLAN соответствует защищаемому VRF
- Outside канал между Edge Router и Firewall – также Dot1Q транк. Каждый VLAN соответствует защищаемому VRF.
- Что нужно помнить:
  - Некоторые Firewall-ы могут использовать разные VLAN ID# межу Inside и Outside интерфейсами, поэтому на Border Leaf или Edge Router требуется соответствующая настройка для установления соседства

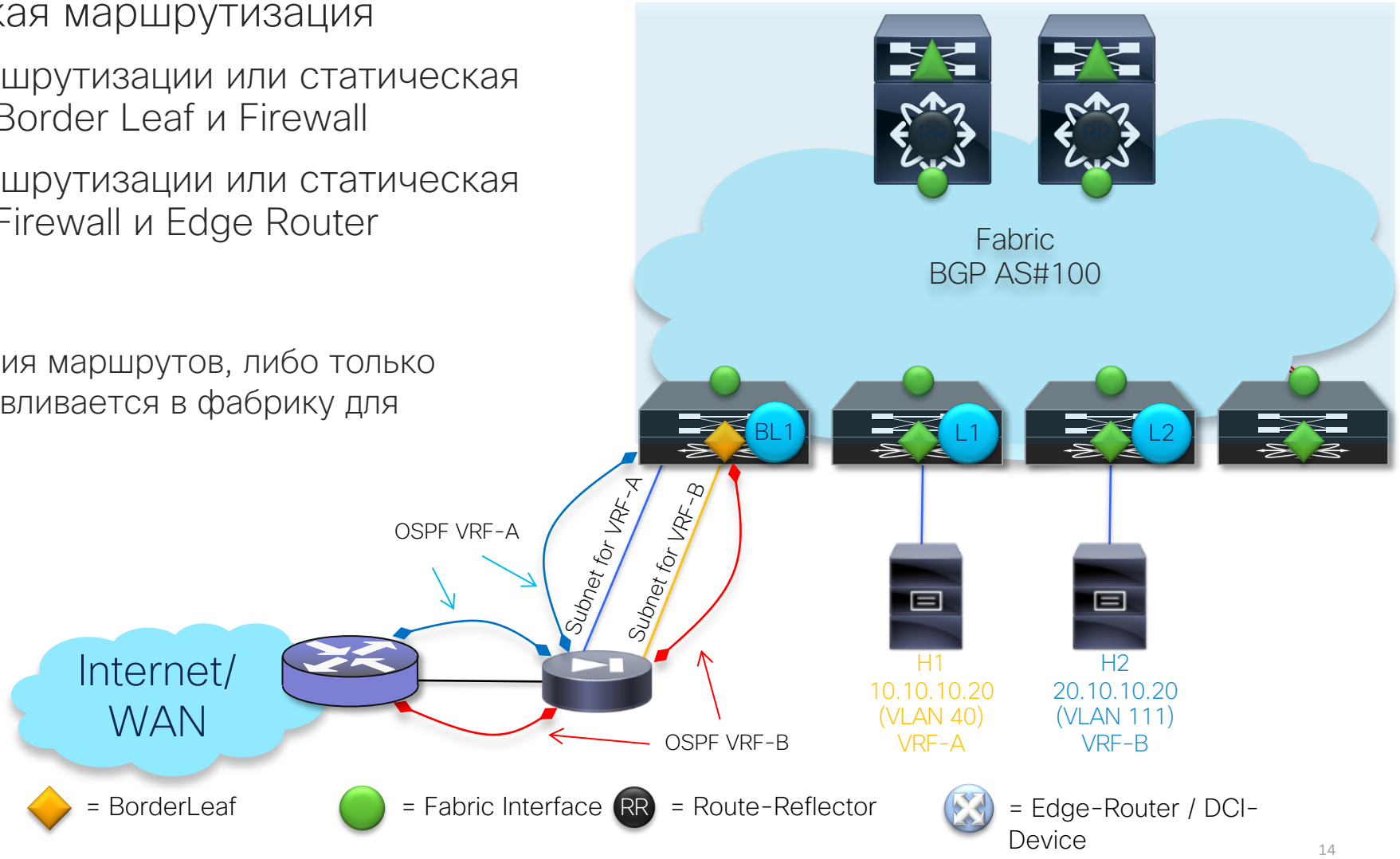


= Spine = Leaf = BorderLeaf = Fabric Interface RR = Route-Reflector = Edge-Router / DCI-Device

Inter-VRF Firewall: Routed Mode (Layer 3)

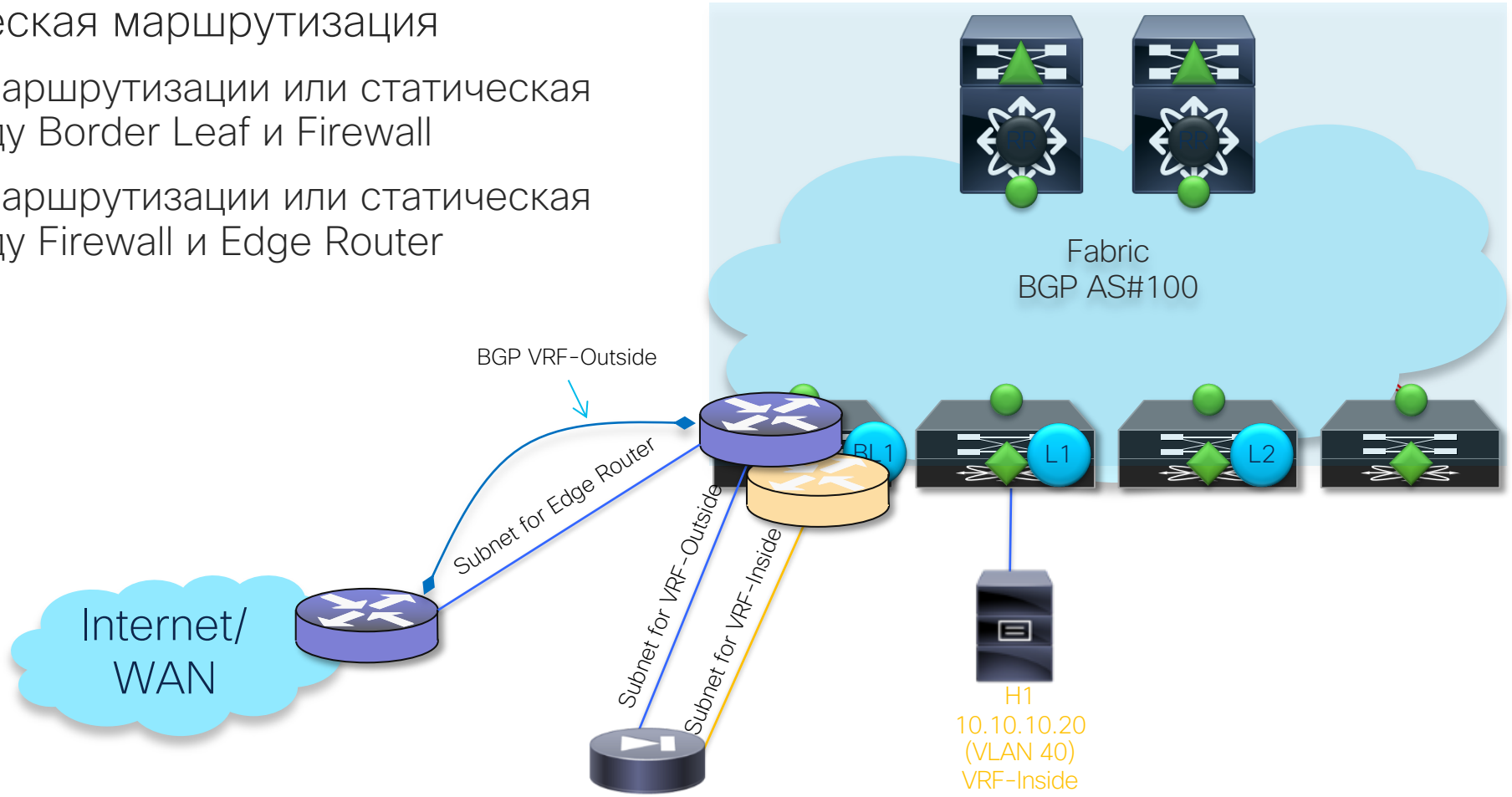
- Динамическая/статическая маршрутизация
  - Per-VRF соседство маршрутизации или статическая маршрутизация между Border Leaf и Firewall
  - Per-VRF соседство маршрутизации или статическая маршрутизация между Firewall и Edge Router

- Выполняется либо суммаризация маршрутов, либо только маршрут по умолчанию устанавливается в фабрику для каждого VRF



Inter-VRF Firewall: VRF “sandwitch” (Layer 3)

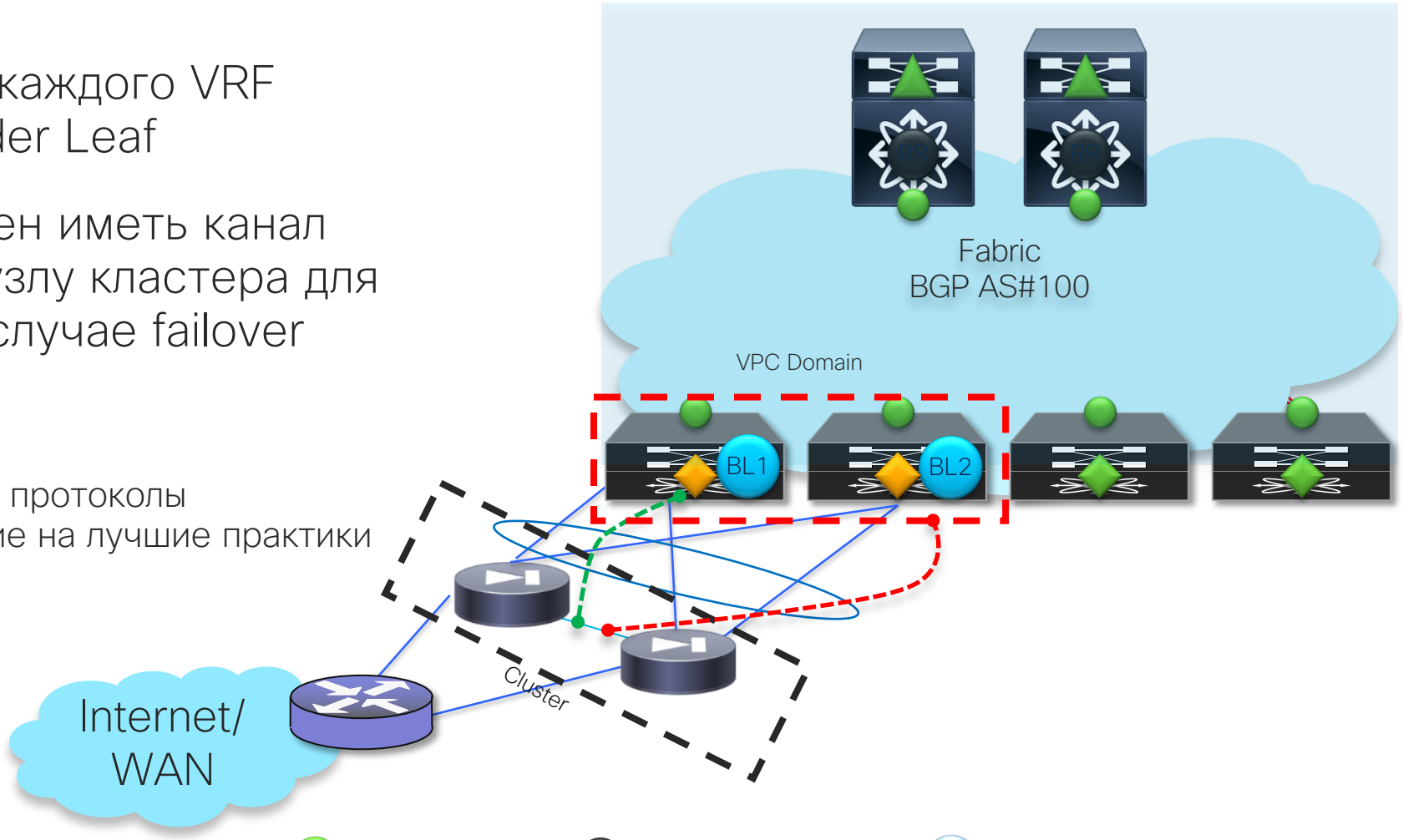
- Динамическая/статическая маршрутизация
  - Per-VRF соседство маршрутизации или статическая маршрутизация между Border Leaf и Firewall
  - Per-VRF соседство маршрутизации или статическая маршрутизация между Firewall и Edge Router



= Spine = Leaf = BorderLeaf = Fabric Interface RR = Route-Reflector = Edge-Router / DCI-Device

Inter-VRF Firewall: Routed Mode (Layer 3)
Отказоустойчивость (Cluster)

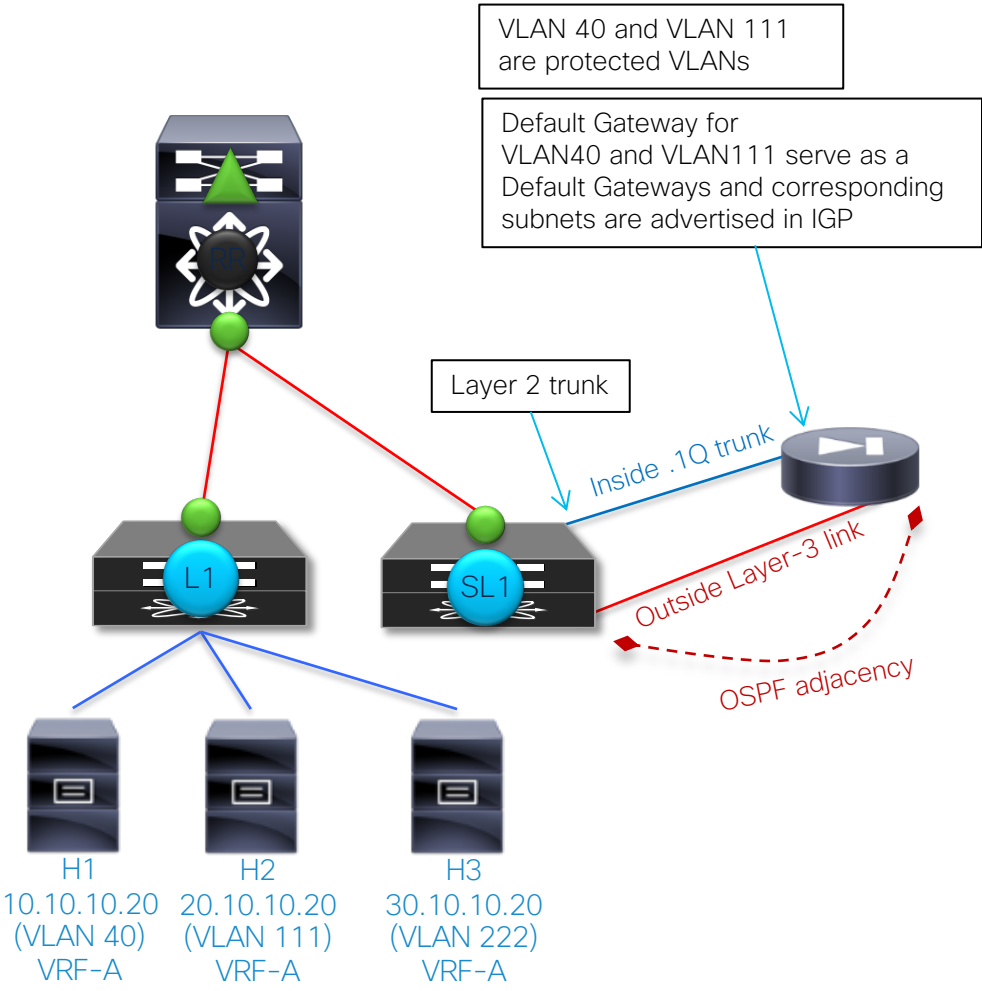
- Peering выполняется для каждого VRF через SVI на каждом Border Leaf
- Каждый Border Leaf должен иметь канал подключения к каждому узлу кластера для сохранения соседства в случае failover
- Если используются динамические протоколы маршрутизации обратите внимание на лучшие практики для настройки L3 поверх vPC



= Spine = Leaf = BorderLeaf = Fabric Interface RR = Route-Reflector = Edge-Router / DCI-Device

Intra-VRF, Inter-VLAN Firewall: Routed Mode

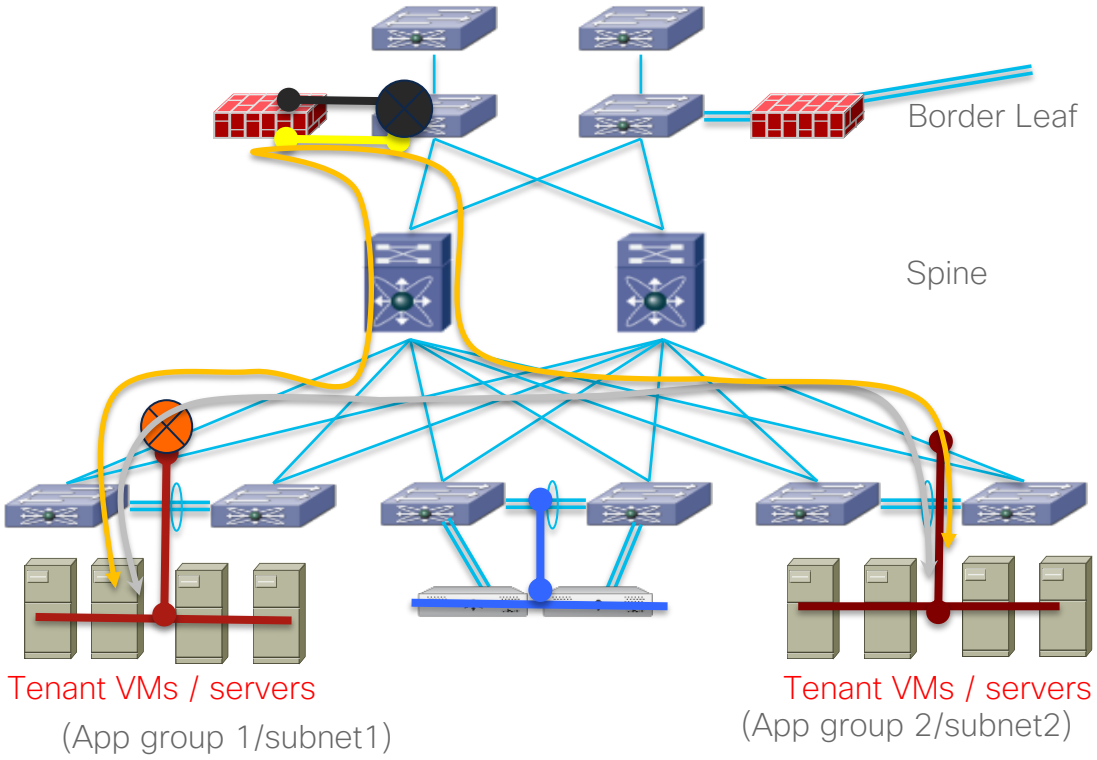
- Firewall можно подключить к Border Leaf или Service Leaf
- Inside канал между Service Leaf и Firewall - Dot1Q trunk. Каждый VLAN соответствует защищаемой подсети.
- Защищаемые VLAN-ы настраиваются в режиме L2 на Service Leaf и на всей фабрике
- Outside канал - Layer 3 point-to-point
- Firewall по одному из протоколов IGP (OSPF, EIGRP или eBGP) устанавливает соседство с Service Leaf через Outside канал
- На Firewall защищаемые VLAN-ы терминируются в BVI (bridged virtual interface) или эквивалент и анонсируются в IGP. Данные BVI-и являются шлюзом по умолчанию для хостов в защищаемых VLAN-х
- На Service Leaf префиксы, получаемые от Firewall, редистрибьютятся в BGP
- Service Leaf анонсирует маршрут по умолчанию 0.0.0.0/0 или specific маршруты на Firewall через IGP



= Spine = Leaf = BorderLeaf = Fabric Interface RR = Route-Reflector = Edge-Router / DCI-Device

Часть 2: Подключение сервисов при помощи Policy Based Routing

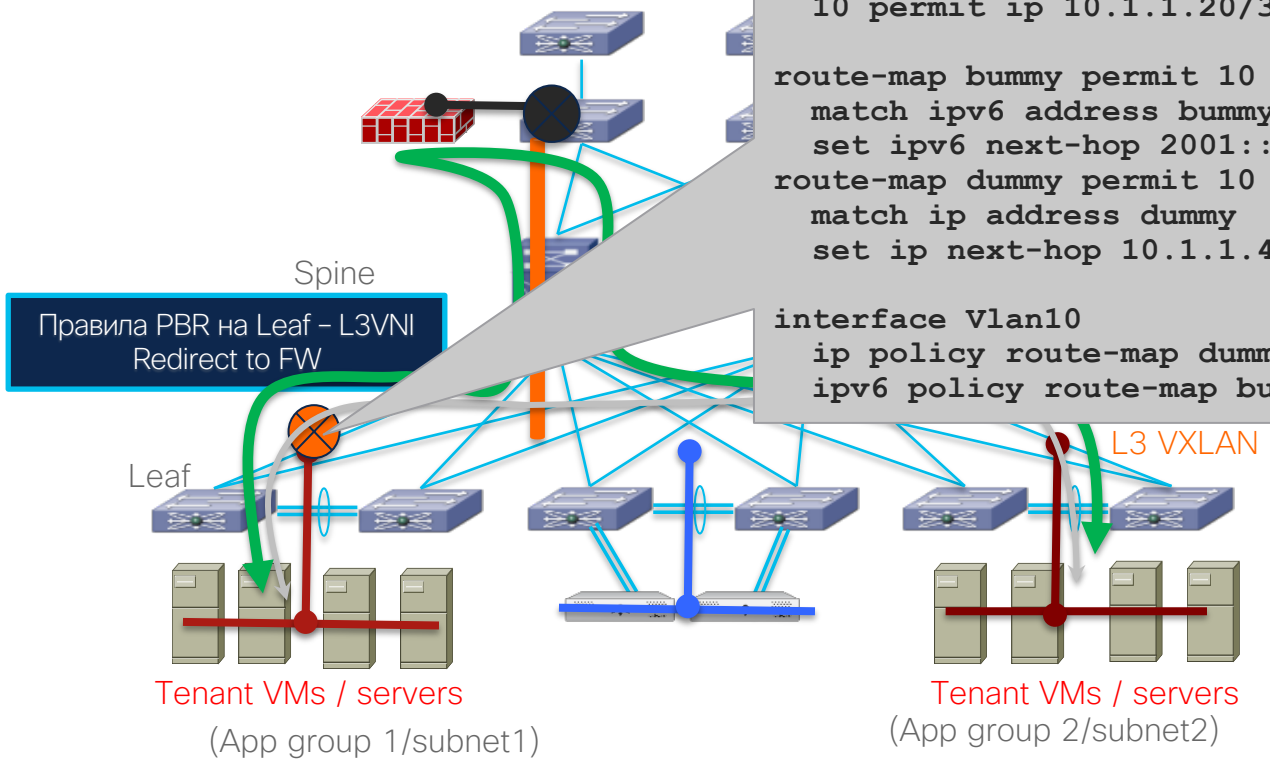
Поддержка PBR на VXLAN BGP EVPN фабрике



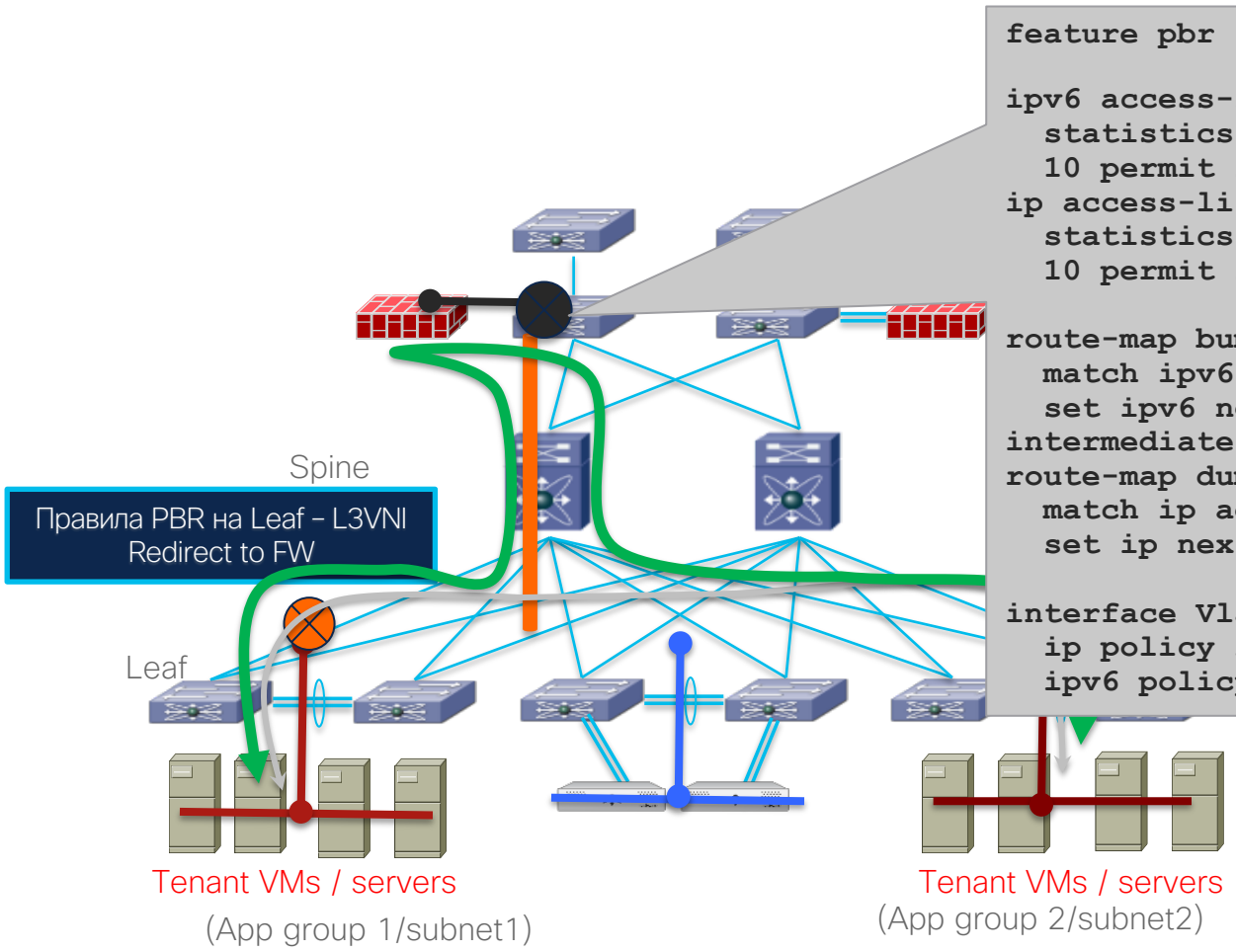
- Перенаправление L3 трафика на основе 5-tuple
- Работает только для маршрутизируемого трафика
- Используется для перенаправления на Load-Balancer или Firewall
- Политика PBR должна быть применена на всех leaf-ах, чтобы гарантировать симметричную передачу

Поддержка PBR на VXLAN BGP EVPN фабрике

```
feature pbr

ipv6 access-list bummy
  statistics per-entry
  10 permit ipv6 2001:10:1:1::20/128 any
ip access-list dummy
  statistics per-entry
  10 permit ip 10.1.1.20/32 any

route-map bummy permit 10
  match ipv6 address bummy
  set ipv6 next-hop 2001::DB8:800:200C:417A <== next-hop host behind some intermediate VTEP
route-map dummy permit 10
  match ip address dummy
  set ip next-hop 10.1.1.40 <== next-hop host behind some intermediate VTEP

interface Vlan10
  ip policy route-map dummy
  ipv6 policy route-map bummy
```



Поддержка PBR на VXLAN BGP EVPN фабрике

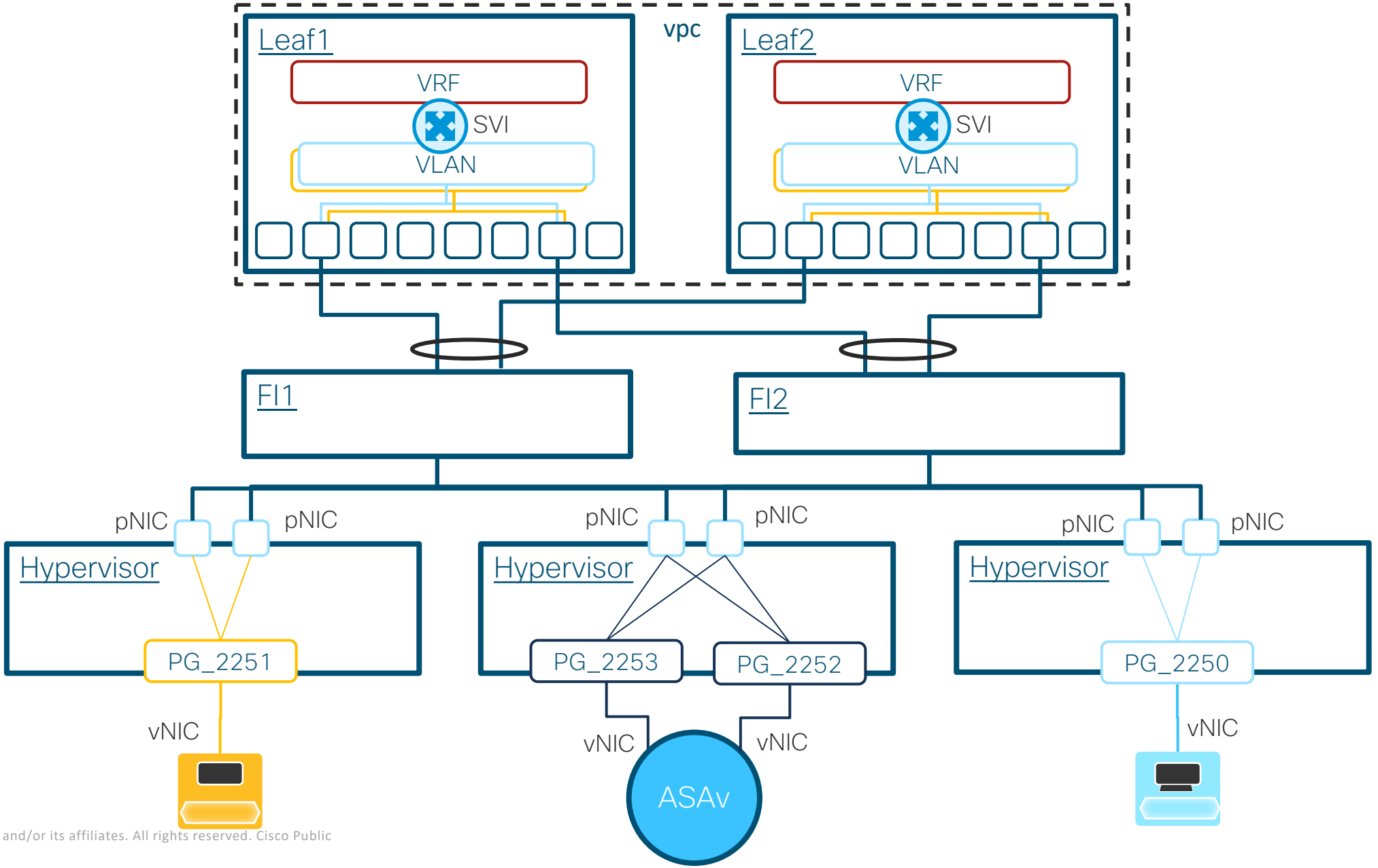



```
feature pbr

ipv6 access-list bummy
  statistics per-entry
  10 permit ipv6 2001:10:1:1::20/128 any
ip access-list dummy
  statistics per-entry
  10 permit ip 10.1.1.20/32 any

route-map bummy permit 10
  match ipv6 address bummy
  set ipv6 next-hop 2001::DB8:800:200C:417A <== next-hop host behind some
intermediate VTEP
route-map dummy permit 10
  match ip address dummy
  set ip next-hop 10.1.1.40 <== next-hop host behind some intermediate VTEP

interface Vlan2500
  ip policy route-map dummy
  ipv6 policy route-map bummy
```

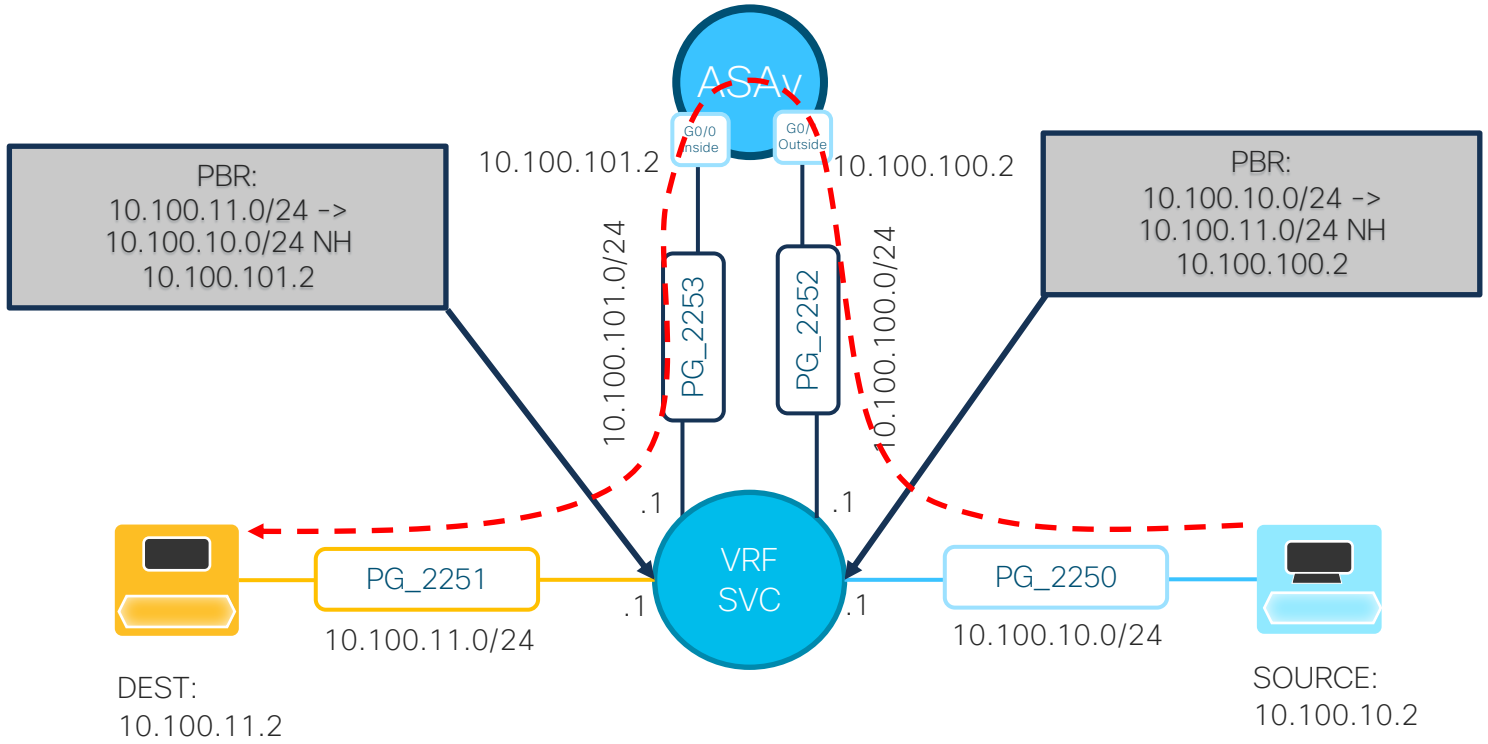
Demo

Сервисное устройство с PBR

DEMO



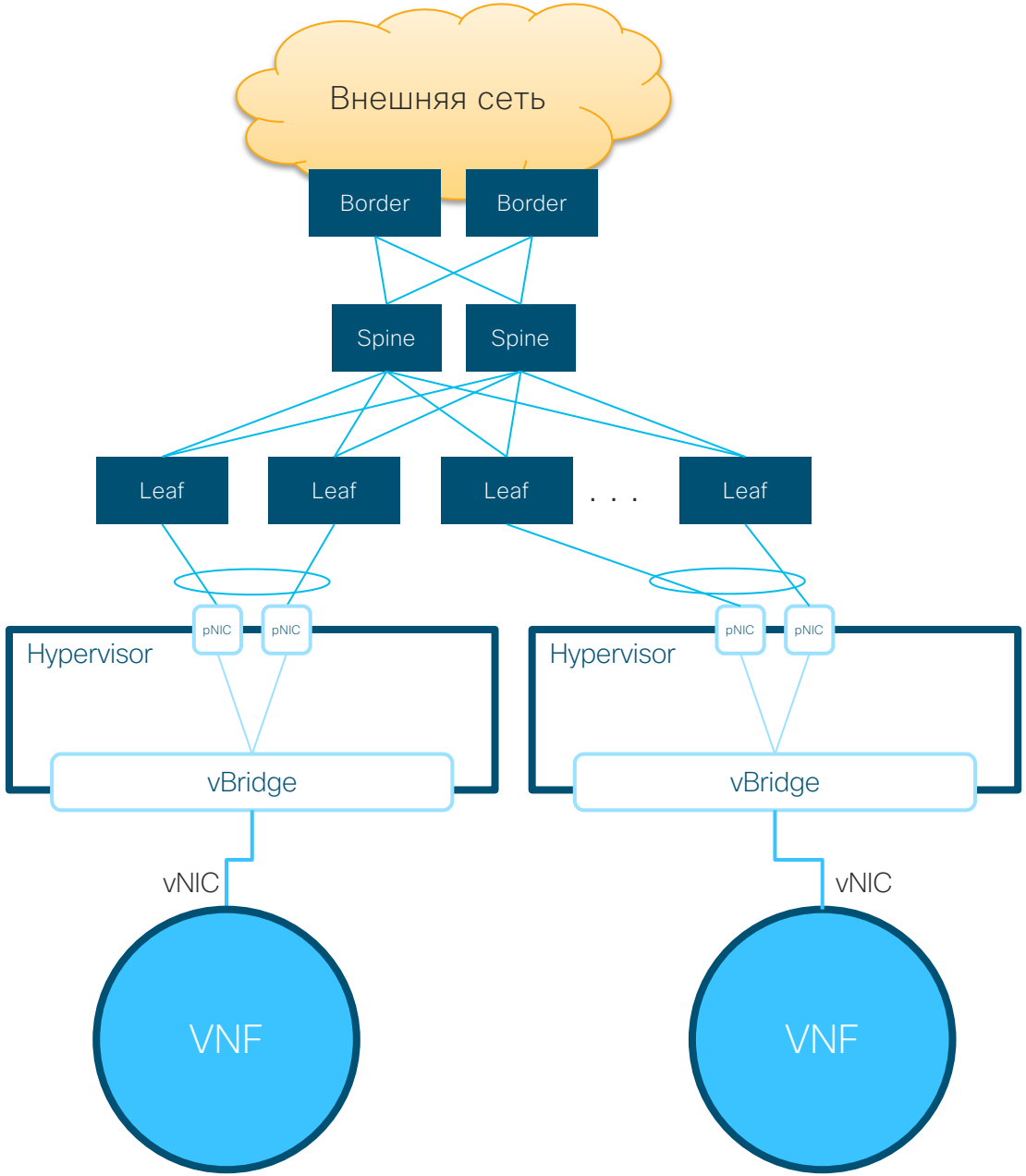
DEMO



Часть 3: Подключение виртуализированных сервисных функций

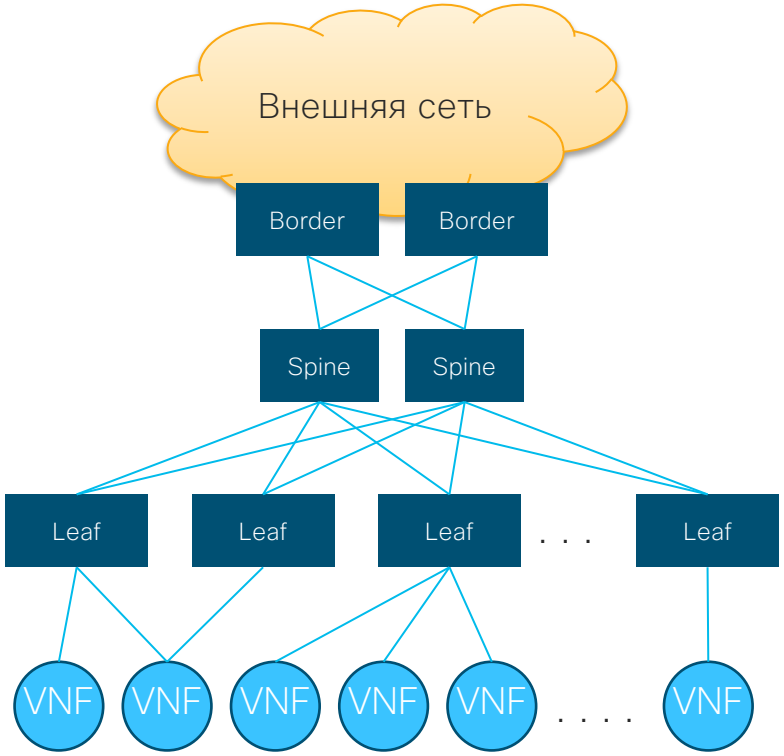
Virtual Network Function (VNF)

- Виртуальная машина (VM) или контейнер (CNF)
  - Например как часть NFVi – Network Function Virtualization Infrastructure
- Часто используемые VNF это виртуальные маршрутизаторы, МСЭ, устройства для балансировки или сервисы вычислительных кластеров (Calico, NSX)
- Разновидность дезагрегации; Железо/Гипервизор (HA/Cluster) и функция (VM)



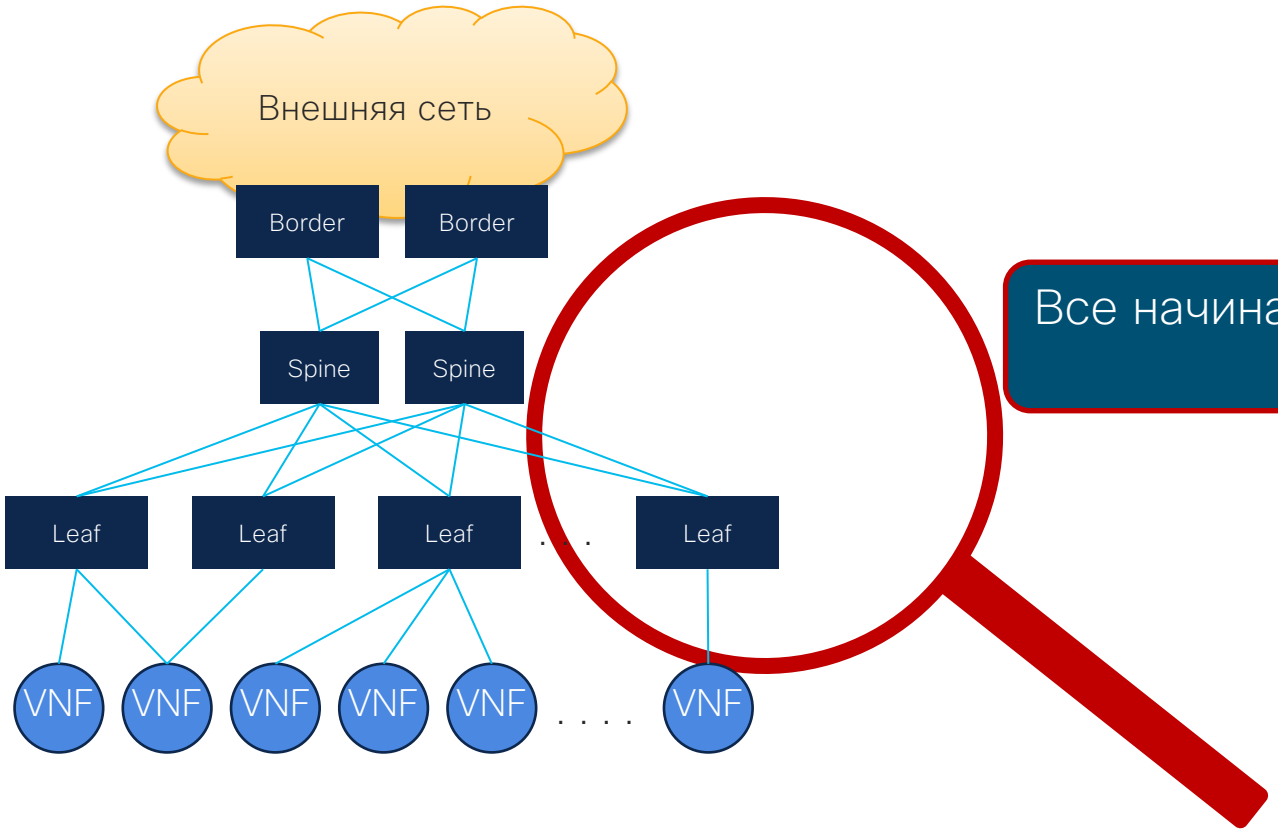
Подключение VNF

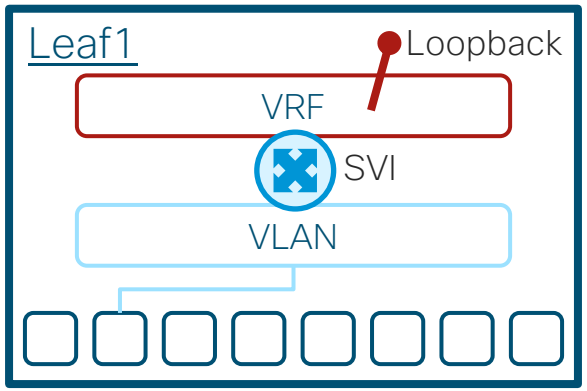
Организация подключения VNF



- На сетевой фабрике на базе VXLAN EVPN настроена подсеть Distributed Anycast Gateway
- Хосты с виртуализированными функциями для подключения используют Access Port, Trunk Ports, Port-Channel или Virtual Port-Channel (vPC)
- На стороне Leaf-коммутатора: выделенный Loopback используется для BGP-сессии (multi-hop routing)
- eBGP (external BGP) между сетевой фабрикой и VNF
  - Контроль за маршрутной информацией
  - Разграничение зон ответственности
- VNF Peering остается локальным для Leaf или пары Leaf

Организация подключения VNF Loopback на фабрике



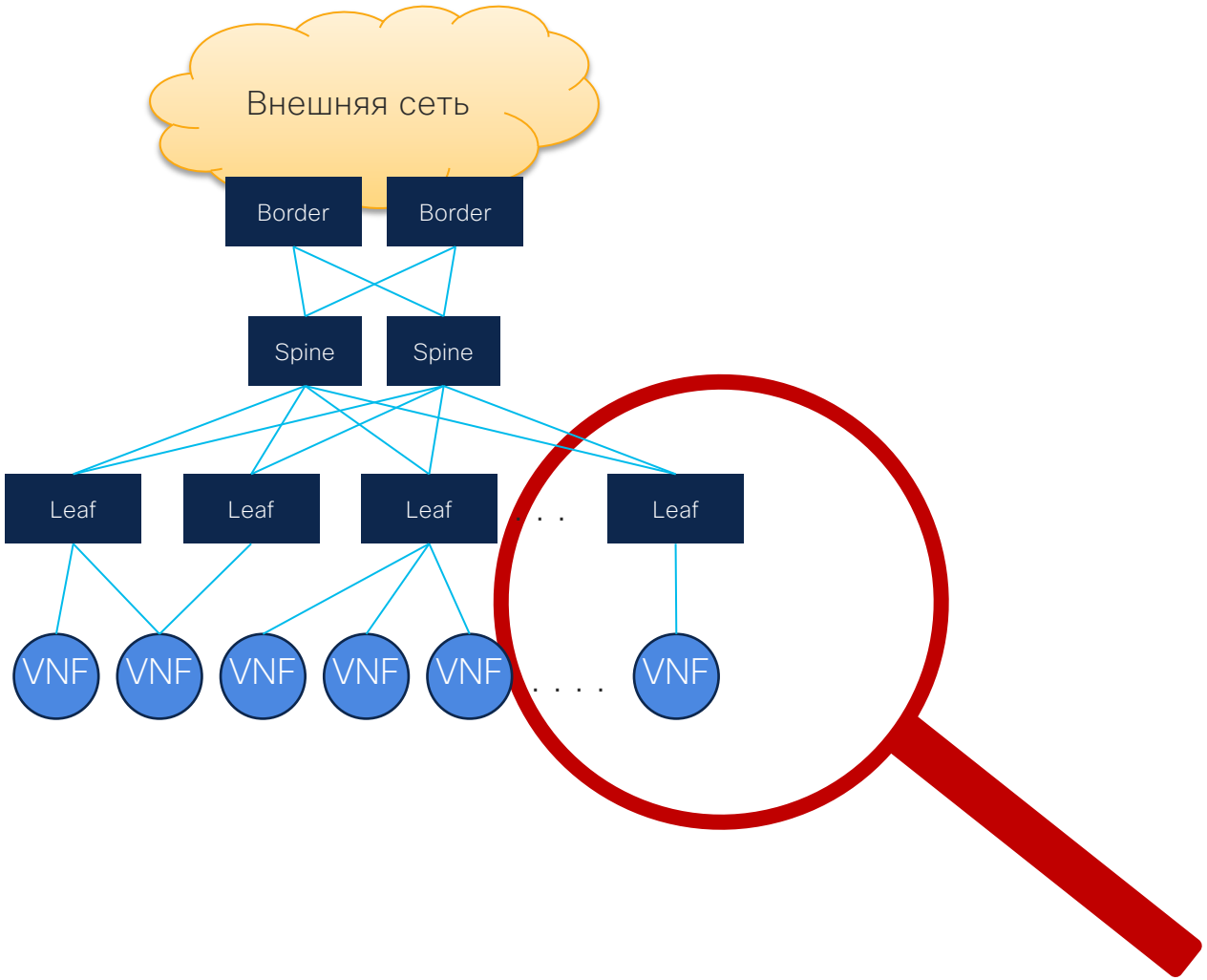


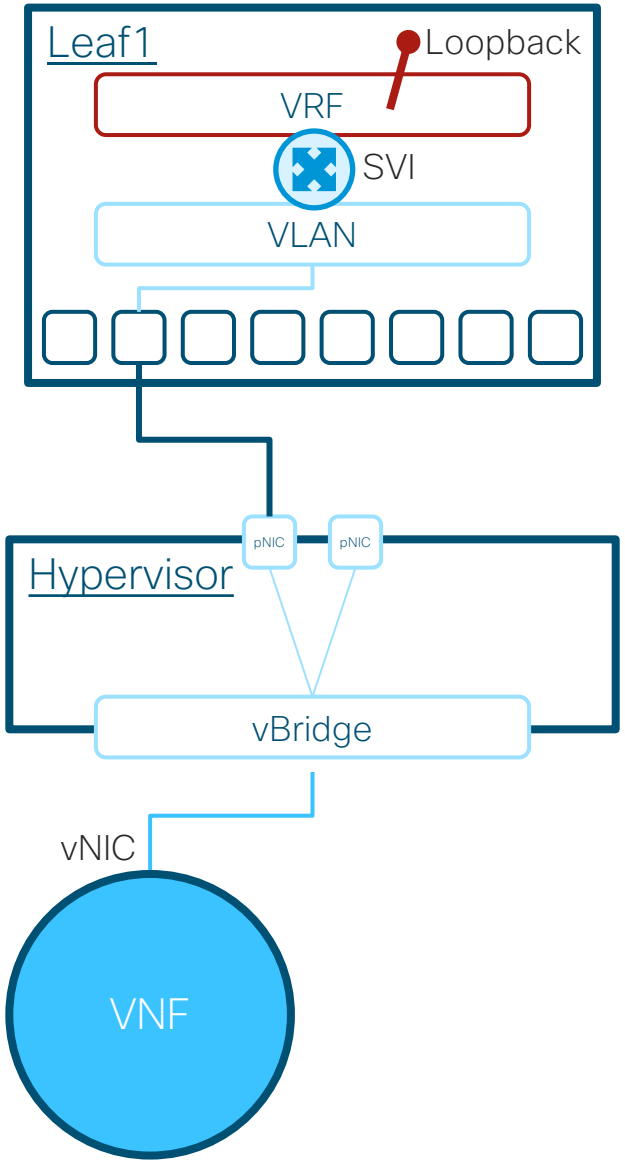
Все начинается с VLAN, SVI и Loopback на стороне Leaf-коммутатора

Всегда использовать Loopback для пиринга со стороны фабрики

Организация подключения VNF

Стандартное подключение внутри гипервизора





Организация подключения VNF

eBGP сессия с vNIC на стороне VNF

```
BGP: 65001
VRF: Service1
Loopback: 10.1.1.5/32
```

```
VLAN: 145
SVI: 192.168.10.1/24 (Anycast Gateway)
```

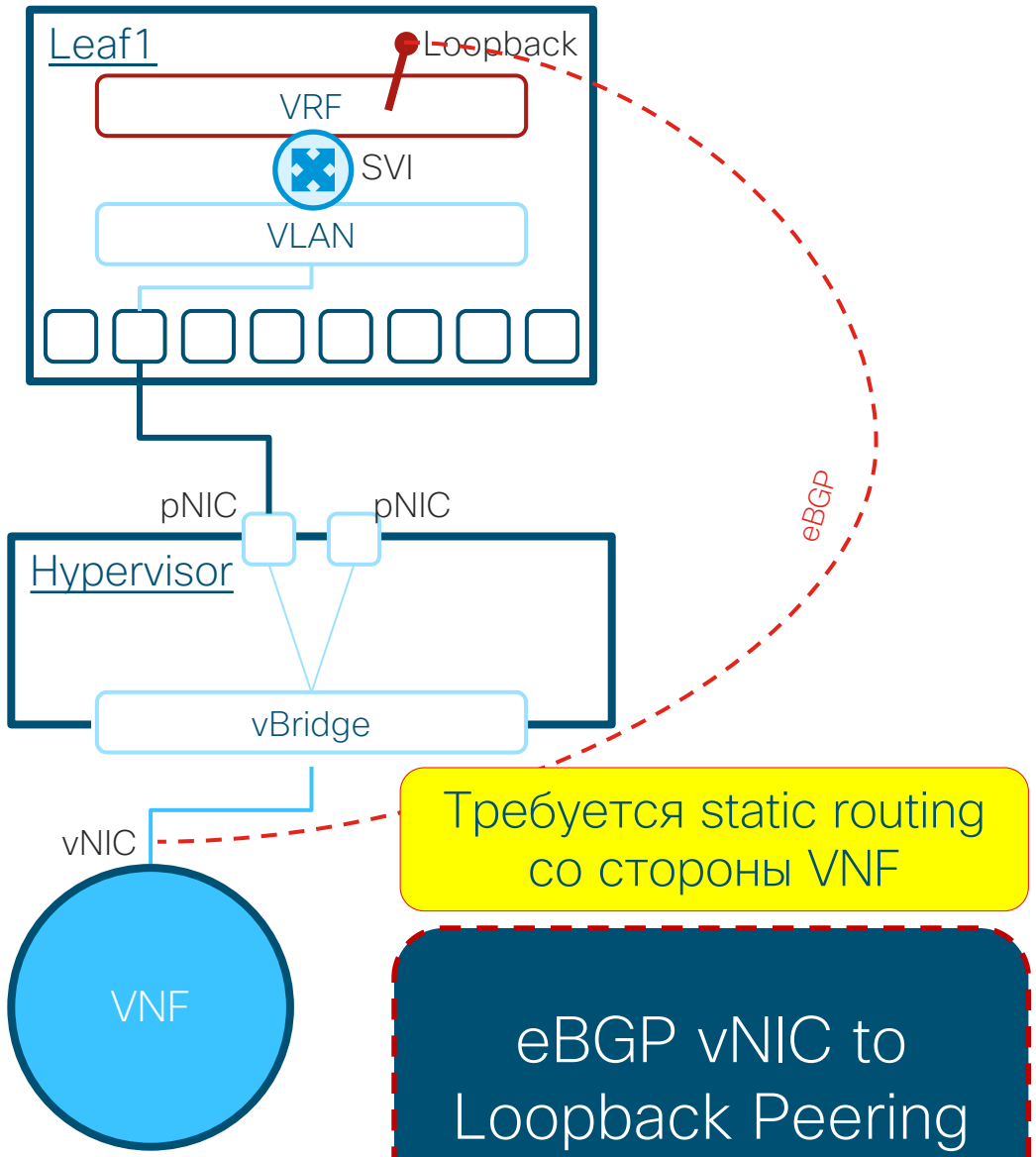
```
Interface: Ethernet1/2
Switchport: Mode Trunk
```

```
pNIC: Tagged Port
vBridge: VLAN 145
```

```
vNIC: 192.168.10.22/24
Default Gateway: 192.168.10.1
```

```
BGP: 65011
Update-Source: vNIC
```

```
Service IP: 172.16.5.0/24
```



Организация подключения VNF

eBGP сессия с Loopback на стороне VNF

```
BGP: 65001
VRF: Service1
Loopback: 10.1.1.5/32
```

```
VLAN: 145
SVI: 192.168.10.1/24 (Anycast Gateway)
```

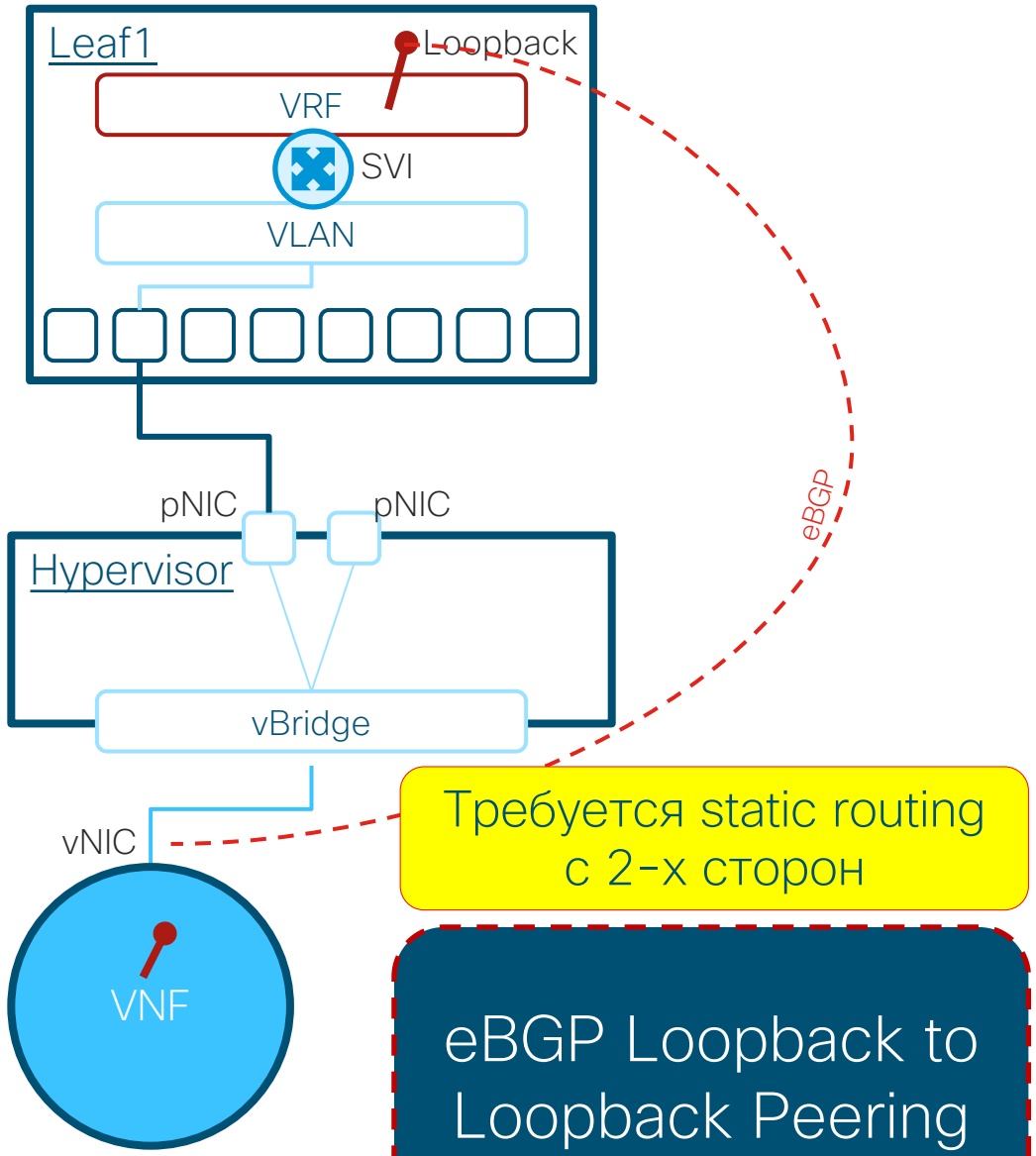
```
Interface: Ethernet1/2
Switchport: Mode Trunk
```

```
pNIC: Tagged Port
vBridge: VLAN 145
```

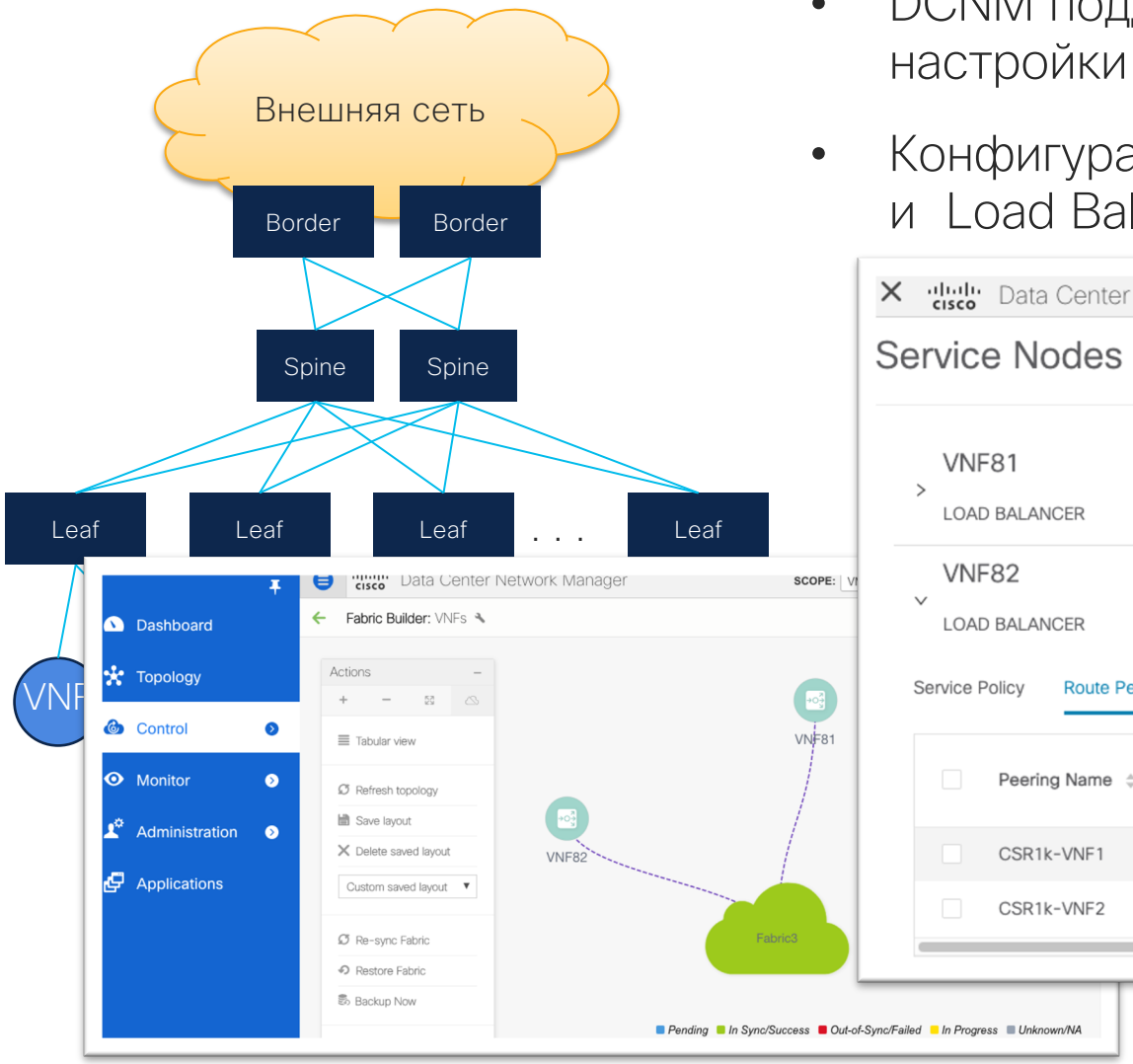
```
vNIC: 192.168.10.22/24
Default Gateway: 192.168.10.1
```

```
BGP: 65011
Loopback: 10.22.22.22/32
```

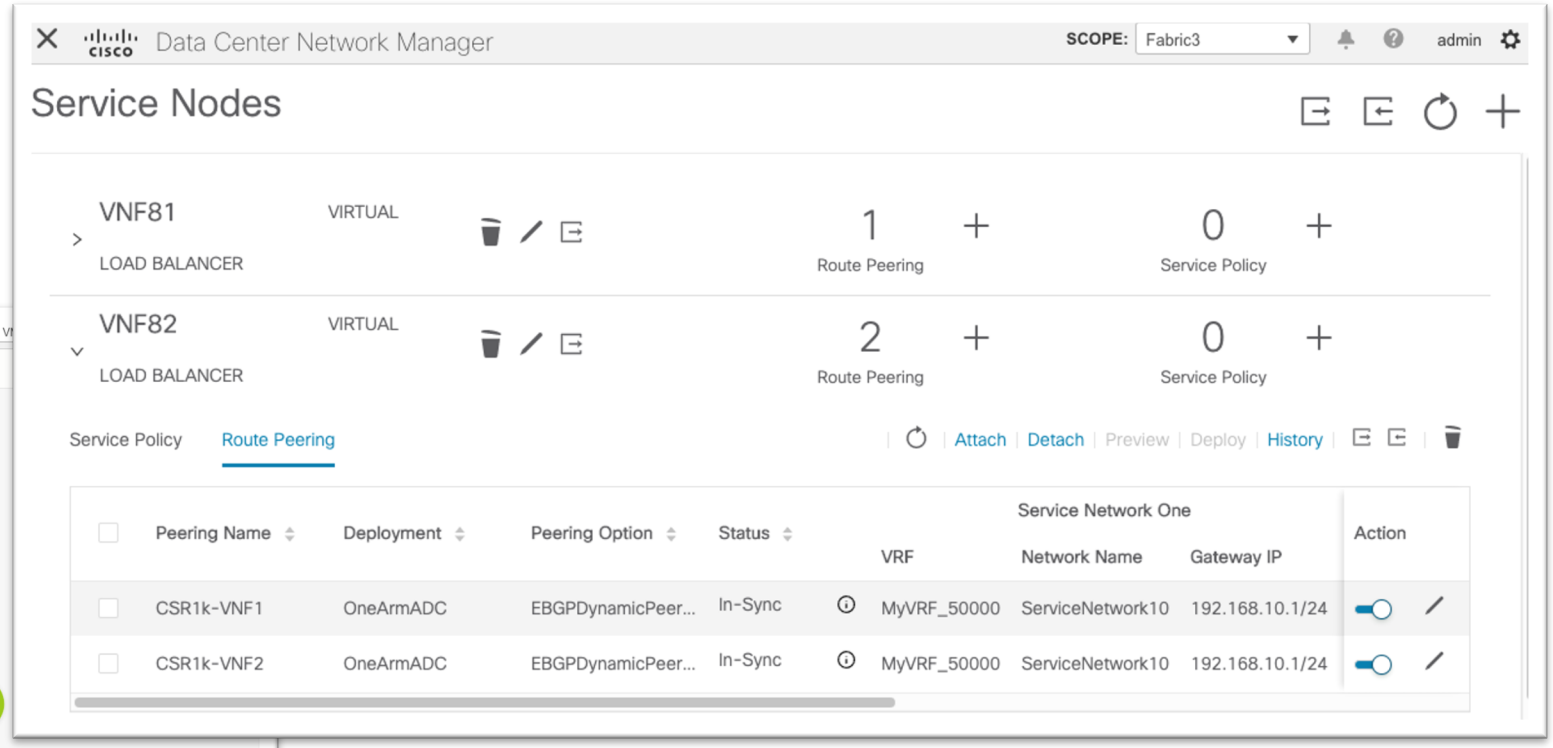
```
Service IP: 172.16.5.0/24
```



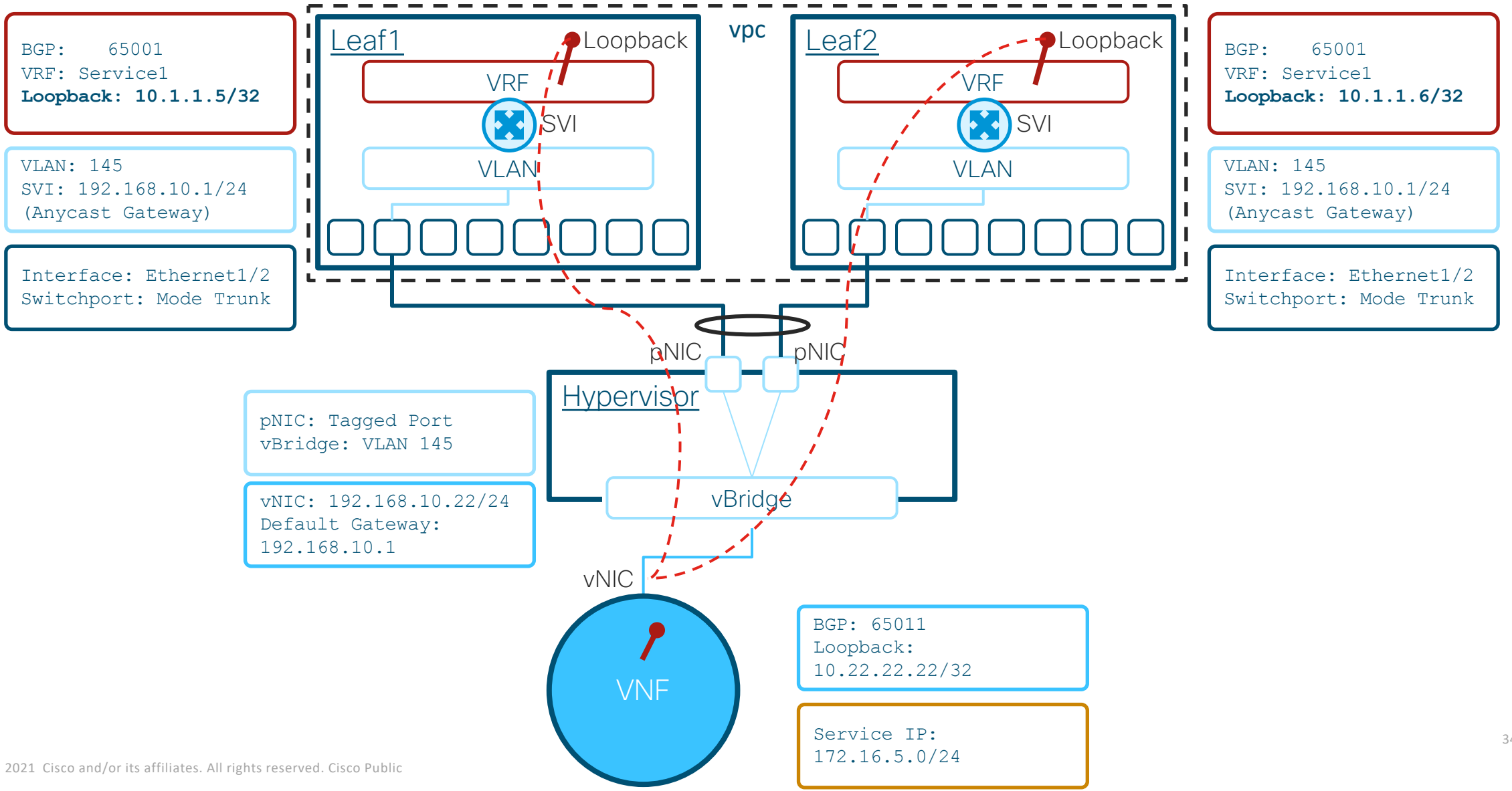
Автоматизация подключения VNF с DCNM



- DCNM поддерживать подключение =VNF через модуль настройки L4-L7 сервисов
- Конфигурационный визард аналогично подключению Firewall и Load Balancer.

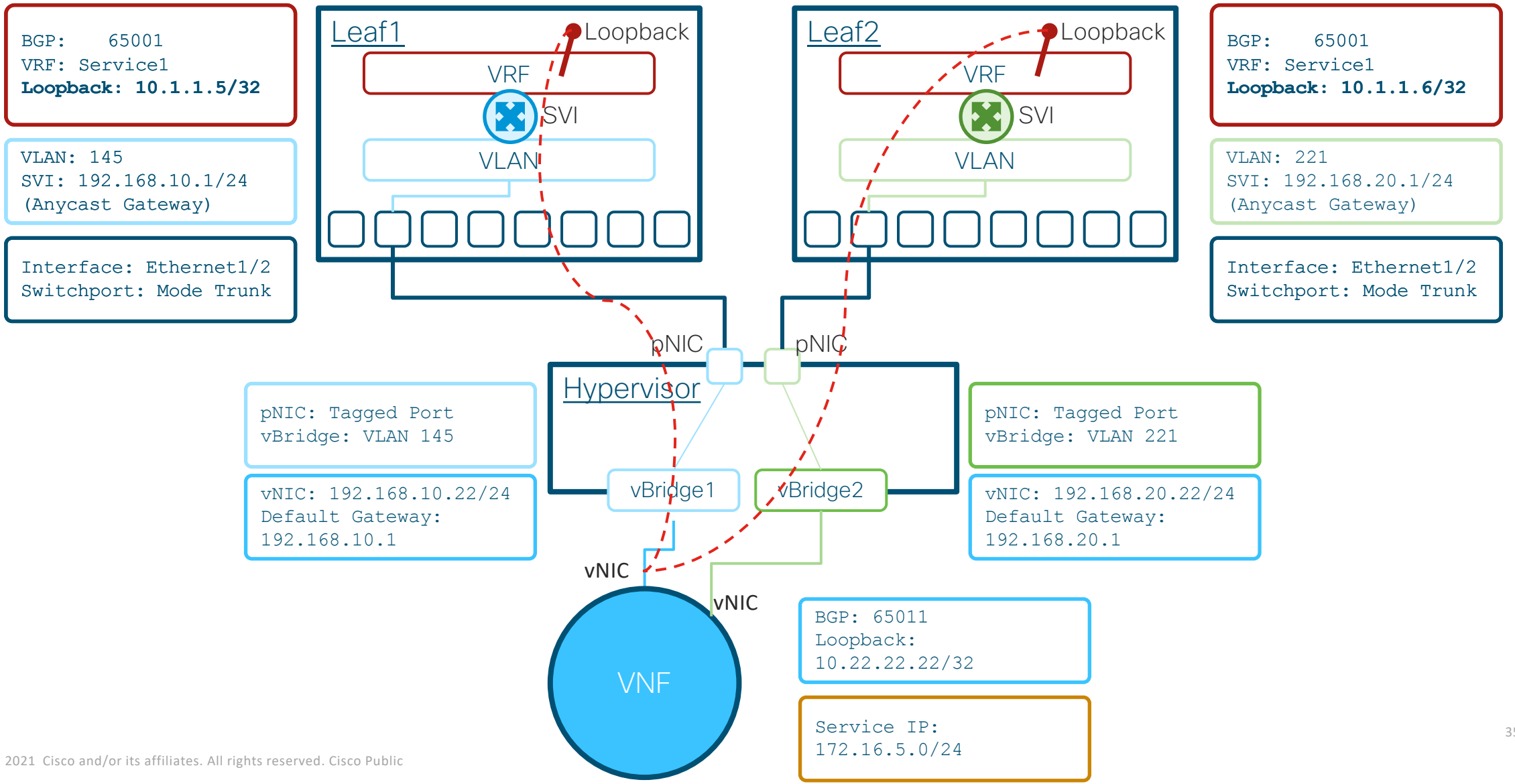


Организация подключения VNF
Отказоустойчивое подключение при помощи vPC

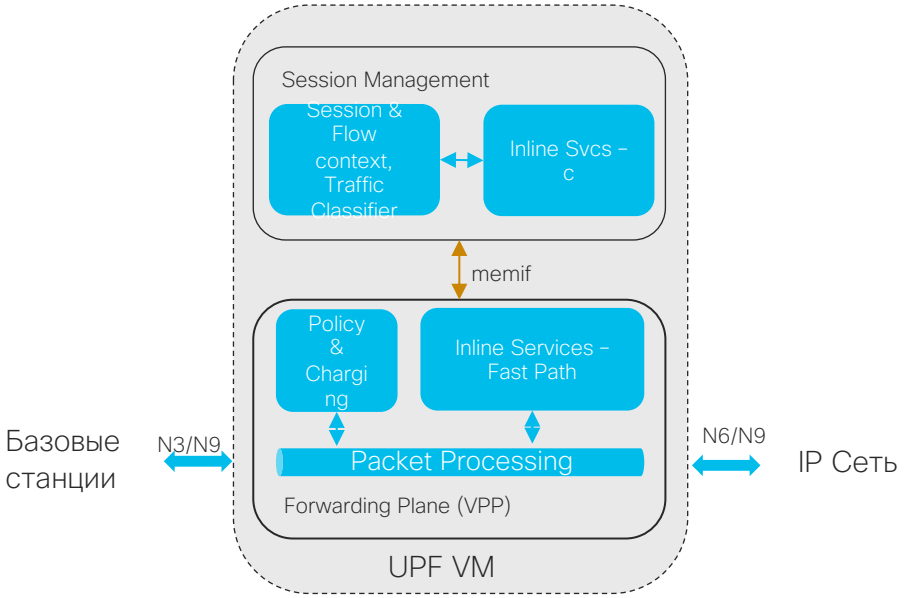


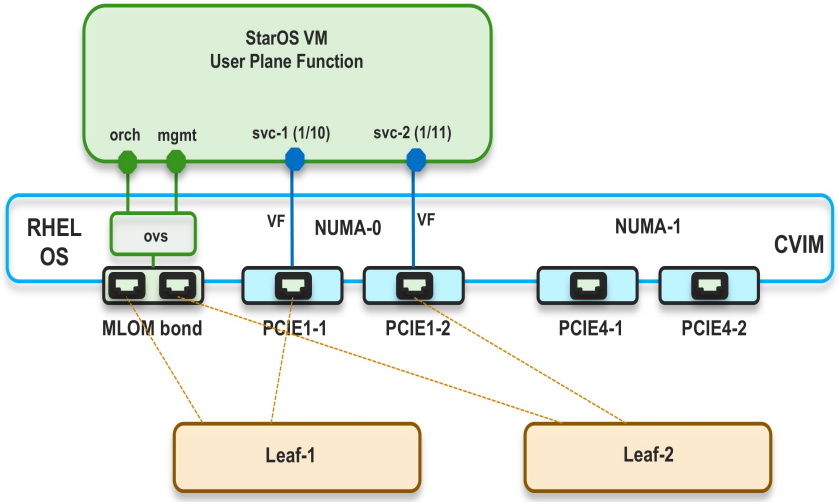
Организация подключения VNF

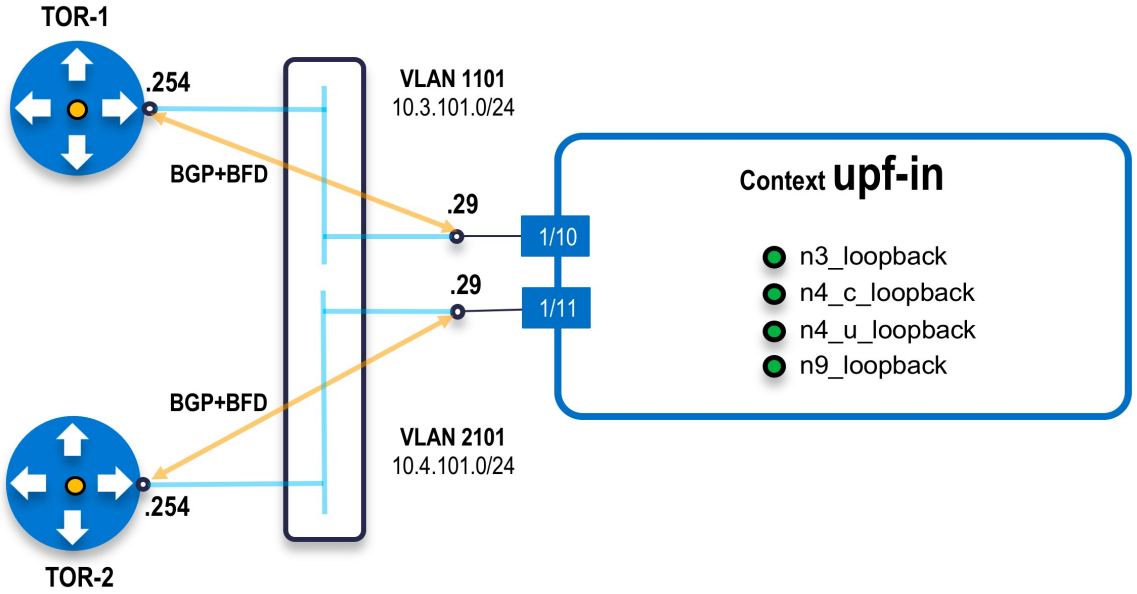
Отказоустойчивое подключение с использованием SR-IOV

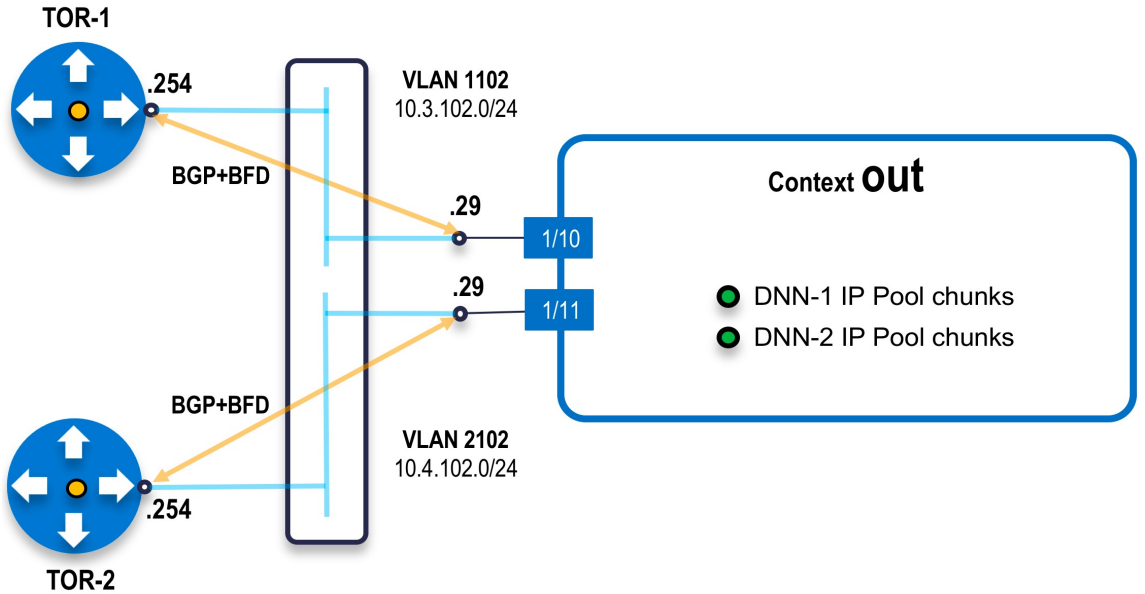


5G User Plane Function (UPF)

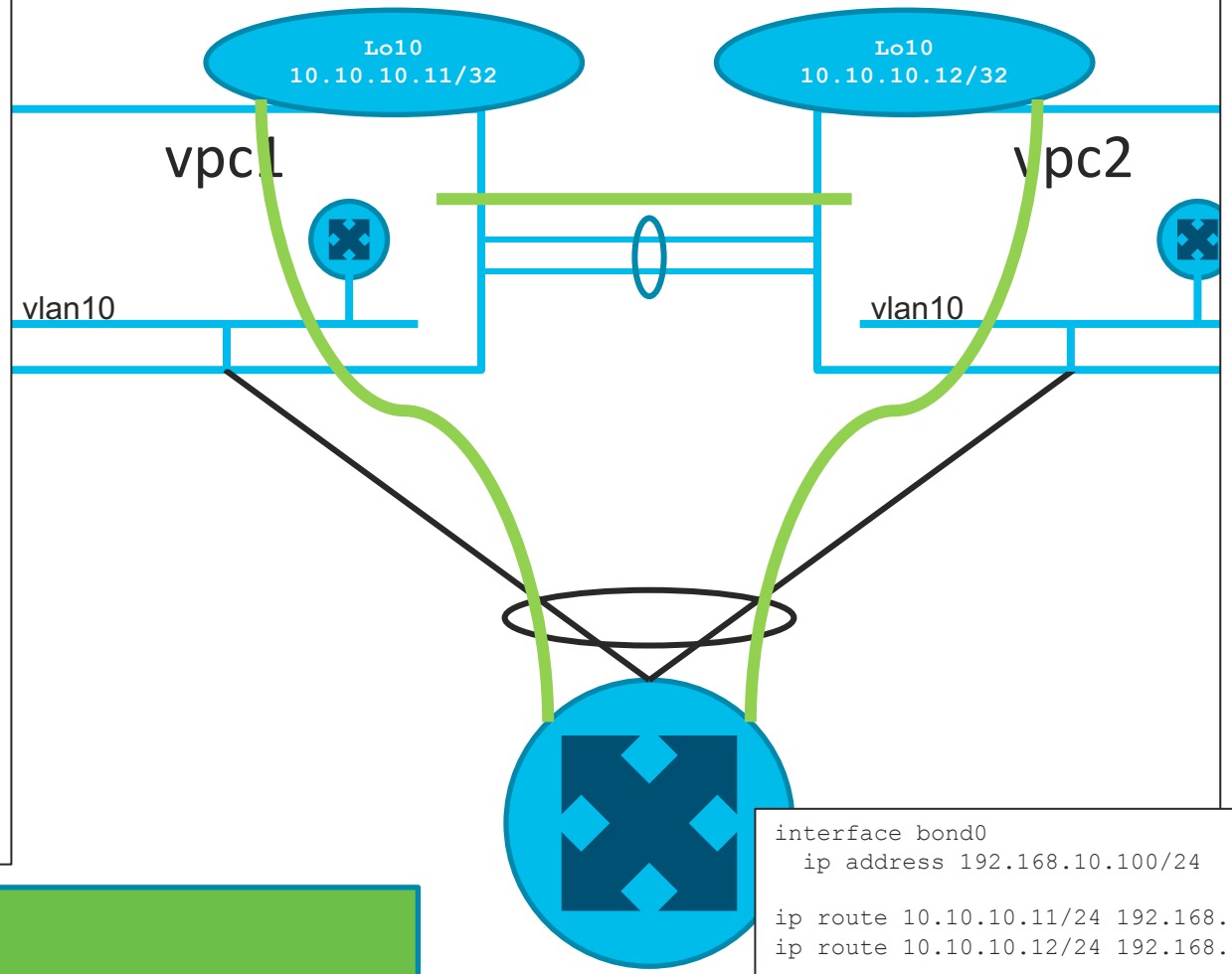








```
vlan 10
  vn-segment 30010
vlan 3967

system nve infra-vlans 3967

interface loopback10
  no shutdown
  vrf member VRF-A
  ip address 10.10.10.11/32 tag 12345

interface Vlan10
  no shutdown
  vrf member VRF-A
  ip address 192.168.10.1/24 tag 12345
  fabric forwarding mode anycast-gateway

interface vlan 3967
  no shutdown
  vrf member VRF-A
  ip address 10.10.0.1/30 tag 12345

router bgp 65501
  vrf VRF-A
    address-family ipv4 unicast
      neighbor 192.168.10.0/24
        remote-as 65502
        ebgp-multihop 5
        update-source loopback 10
        address-family ipv4 unicast
       neighbor 10.10.0.2
         remote-as 65501
         update-source VLAN 3967
         address-family ipv4 unicast
          next-hop-self
```




```
vlan 10
  vn-segment 30010
vlan 3967

system nve infra-vlans 3967

interface loopback10
  no shutdown
  vrf member VRF-A
  ip address 10.10.10.12/32 tag 12345

interface Vlan10
  no shutdown
  vrf member VRF-A
  ip address 192.168.10.1/24 tag 12345
  fabric forwarding mode anycast-gateway

interface vlan 3967
  no shutdown
  vrf member VRF-A
  ip address 10.10.0.2/30 tag 12345

router bgp 65501
  vrf VRF-A
    address-family ipv4 unicast
      neighbor 192.168.10.0/24
        remote-as 65502
        ebgp-multihop 5
        update-source loopback 10
        address-family ipv4 unicast
       neighbor 10.10.0.1
         remote-as 65501
         update-source VLAN 3967
         address-family ipv4 unicast
          next-hop-self
```

```
interface bond0
  ip address 192.168.10.100/24

ip route 10.10.10.11/24 192.168.10.1
ip route 10.10.10.12/24 192.168.10.1

router bgp 65502
  address-family ipv4 unicast
    neighbor 10.10.10.11
      remote-as 65501
  address-family ipv4 unicast
     neighbor 10.10.10.12
       remote-as 65501
```

Пример конфигурации

Рекомендации для подключения VNF

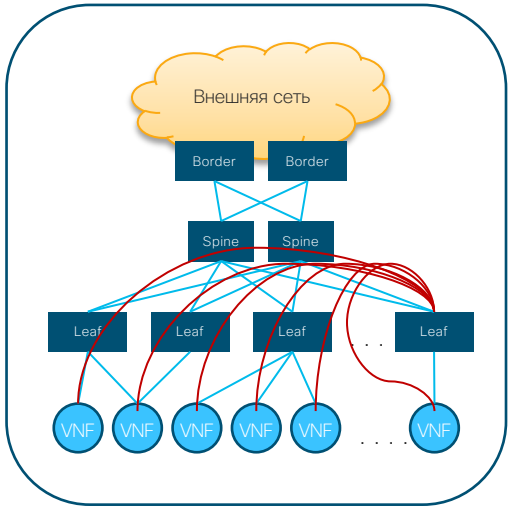
- Использовать лучшие практики
  - Локальный eBGP Route Peering на Leaf коммутатор
  - Централизованный Route Peering требует анализа по масштабируемости и функциональности (см след раздел презентации)
  - VNF Mobility или VNF failover (anycast service) требуется понимать особенности (см след раздел презентации)
- Всегда использовать выделенный Loopback на Leaf для локального eBGP Peering
  - Этот Loopback не нужно анонсировать внутри фабрики
  - Loopback может быть с одним и тем же IP на каждом Leaf для упрощения настройки VNF
    - Требуются уникальные адреса Loopback IP для коммутаторов являющихся частью одного vPC Domain
  - На стороне VNF сессию можно поднимать или с интерфейса или с Loopback
- Для использования BFD требуются основание (в большинстве случаев отсутствует)

Сценарии развертывания – проблемы и решения

Сценарии развертывания
Формулировка проблем

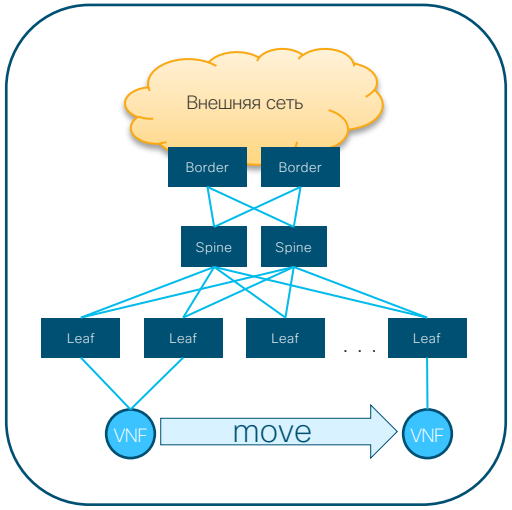
Centralized Routing Peering

Особенности в случае когда для пиринга используется одно выделенное устройство (или пара)



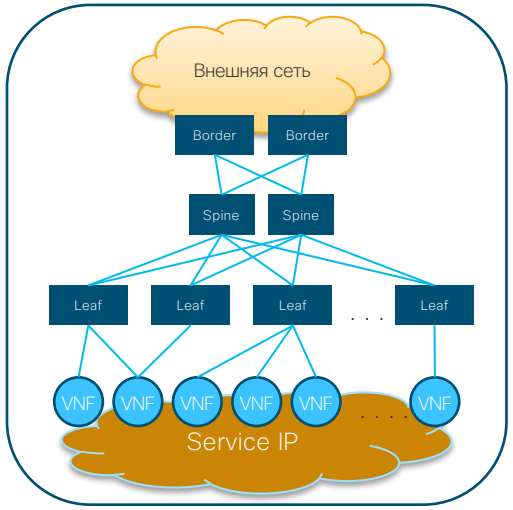
Мобильность VNF

Зависимости и требования



Anycast сервисы

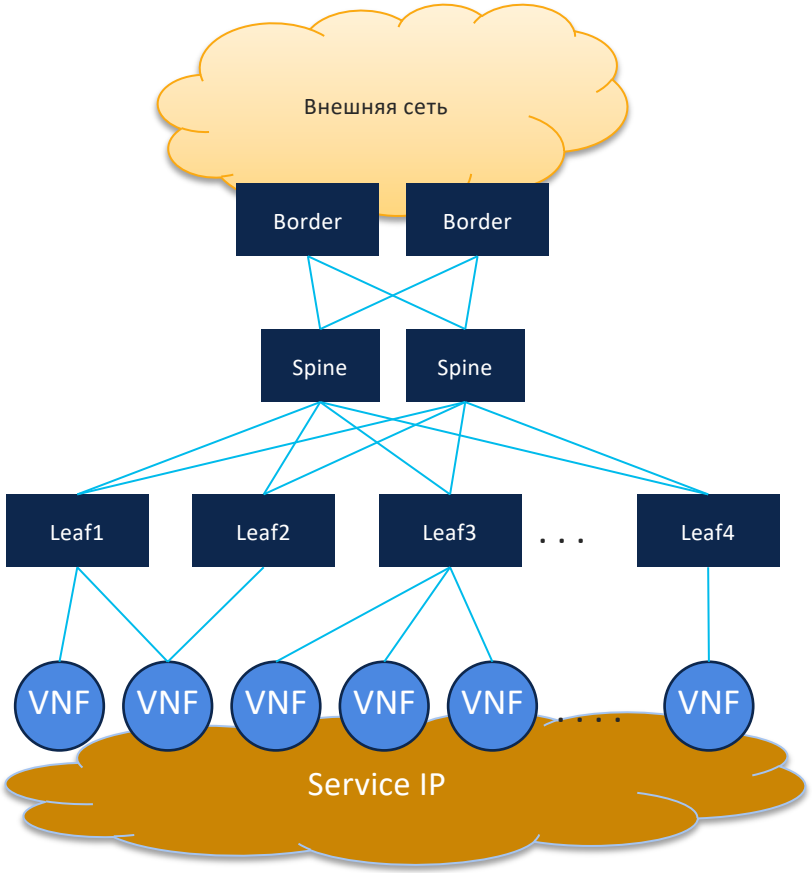
Распределенный сегмент «Service IP или Service Subnet» с требованиями по более гранулярному распределению трафика per-VNF



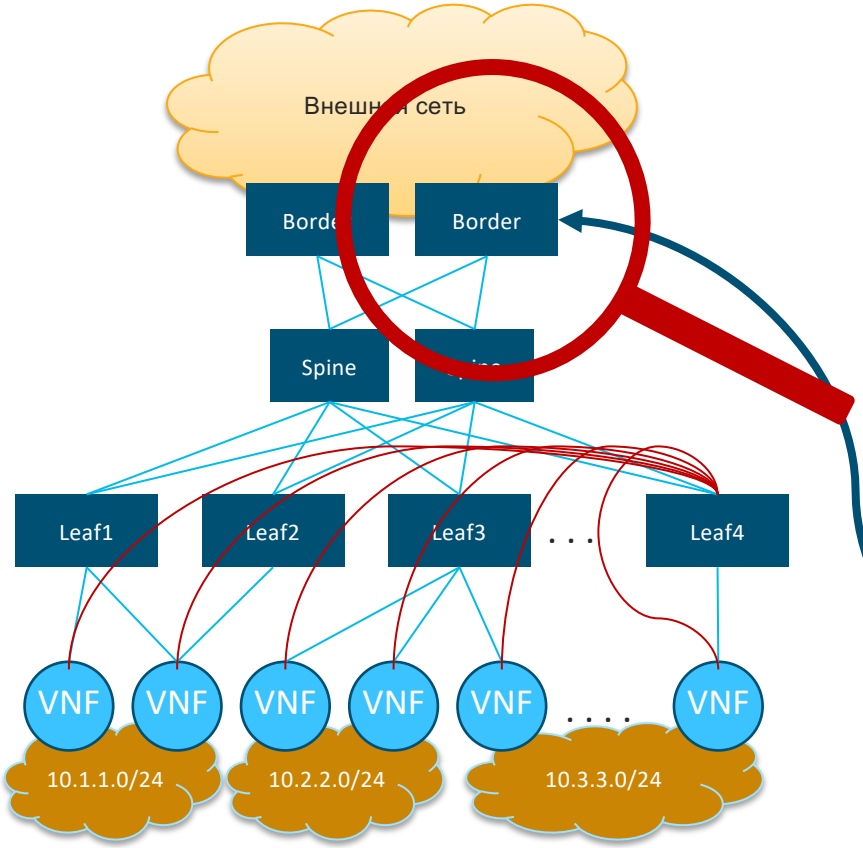
Для решения этих проблем требуется больше, чем просто организовать подключение

Сценарии развертывания
Формулировка проблем

Проблемы	EVPN AF
Avoid Traffic Tromboning (Recursive Routing)	Потенциальный риск
VNF Mobility (Centralized Peering)	Возможно
How Many VNF per Leaf? (Proportional Multipath)	Нет видимости



Centralized VNF Routing Peering
Проблема масштабирования (количество сессий)



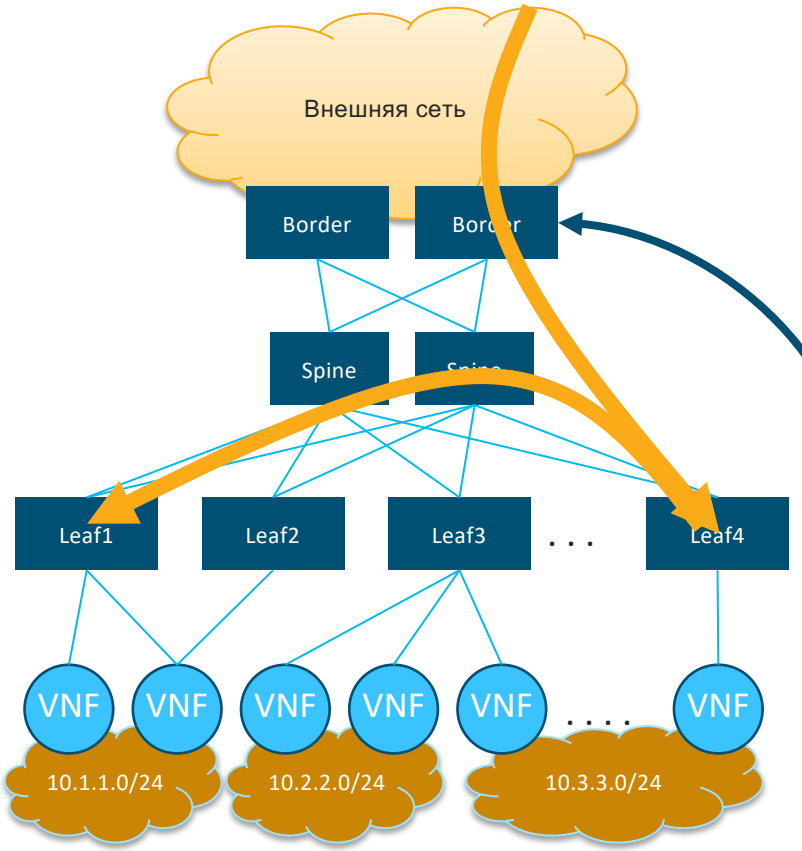
- VNF пирится с централизованными устройствами; может быть Leaf, Border Leaf или похожие
- Routing Peering происходит с VNF на Loopback Interface внутри VXLAN EVPN VRF
- Используется eBGP Multi-Hop через VXLAN транспорт
- Нужно контролировать количество сессий и трафик (Control-Plane and Data-Plane)

Routing Table

Prefix	Next-Hop
10.1.1.0/24	Leaf4
10.2.2.0/24	Leaf4
10.3.3.0/24	Leaf4

Centralized VNF Routing Peering

Проблема “притягивания” трафика одним узлом с которым пирятся VNF



- Routing Prefixes анонсируются одним устройством
- Next-Hop в апдейтах остается неизменным

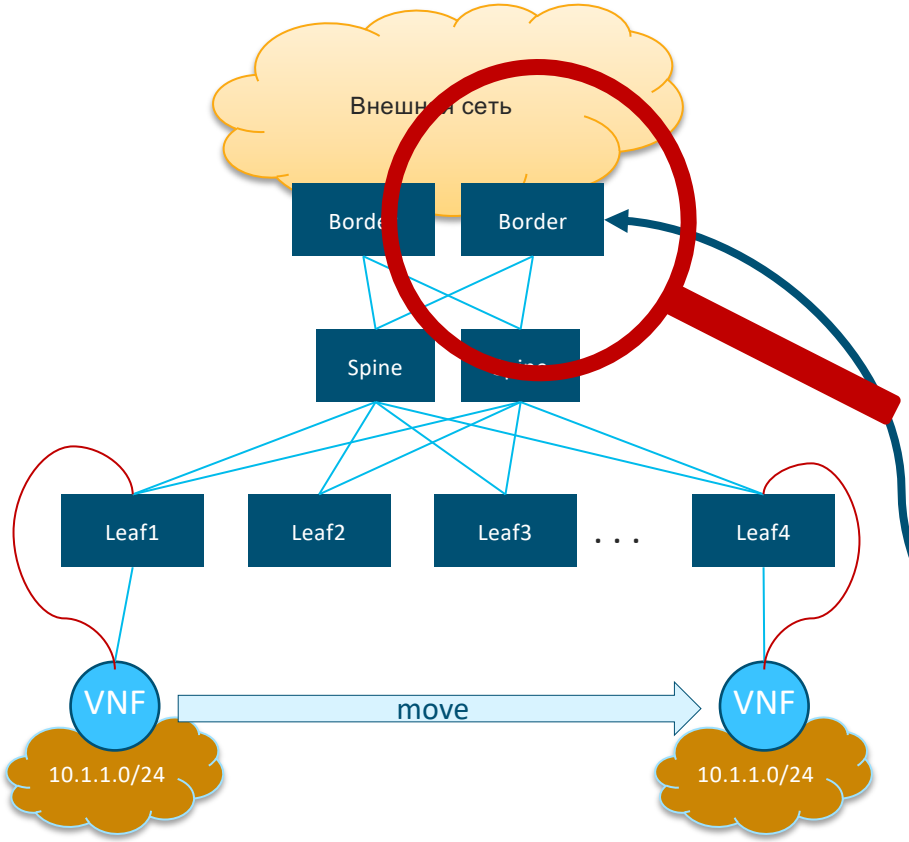
Трафик “притягивается” к одному устройству

Routing Table

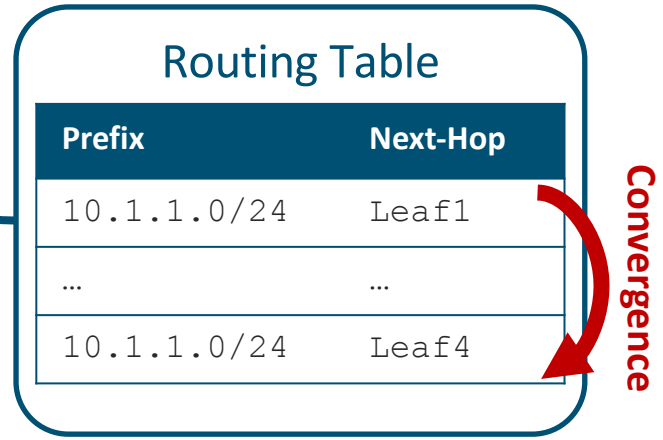
Prefix	Next-Hop
10.1.1.0/24	Leaf4
10.2.2.0/24	Leaf4
10.3.3.0/24	Leaf4

Мобильность VNF

Установление соединения заново после миграции вносит потери и задержку

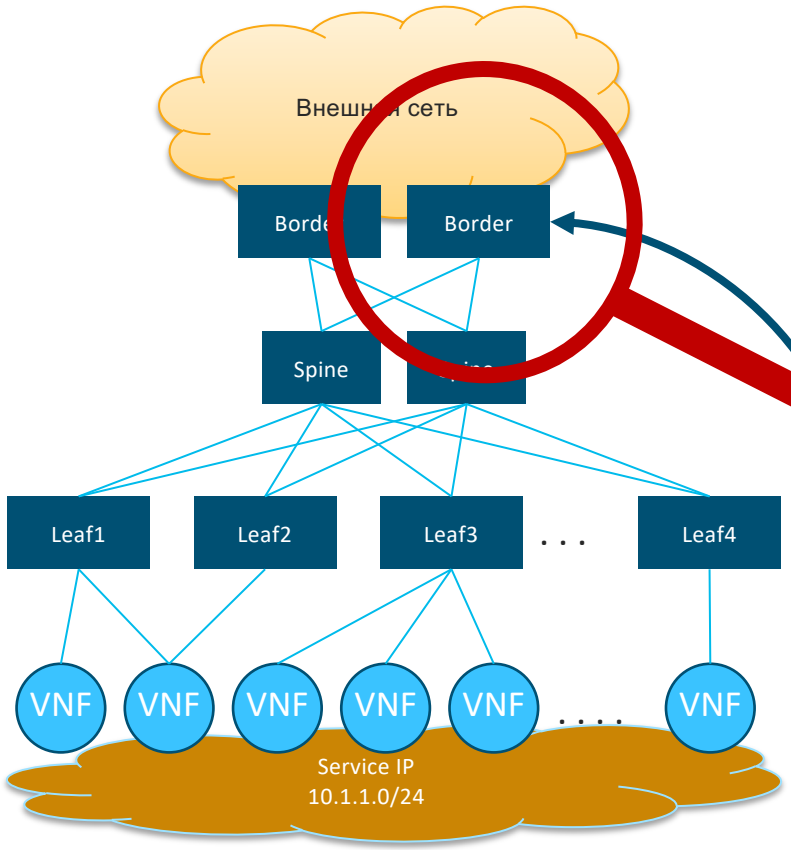


- Physical Network Function (PNF) никогда не перемещается, тогда как VNF имеет такую возможность
- В случает локального eBGP Route Peering соседские отношения должны быть переустановлены после перемещения
- Установление соединения вносит потери и задержку в передаче данных



Anycast сервис

Неэффективное распределение трафика без учета количества VNF за каждым Leaf



- Многие VNF анонсируют один и тот же Service IP или Service Subnet
- Количество VNF за каждым Leaf не прозрачно и не принимается во внимание при балансировке
- Балансировка происходит по принципу «per-Leaf» base; может приводит к неэффективному распределению трафика
  - Например каждому Leaf достается только ¼ трафика, даже если Leaf3 имеет более высокую плотность VNF за собой

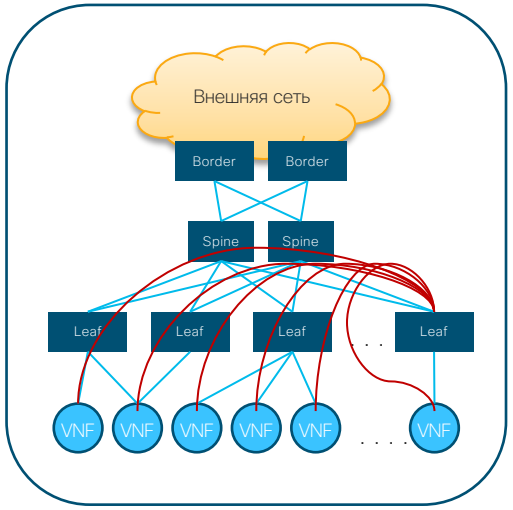
Routing Table

Prefix	Next-Hop
10.1.1.0/24	Leaf1
	Leaf2
	Leaf3
	Leaf4

Сценарии развертывания
Формулировка проблем

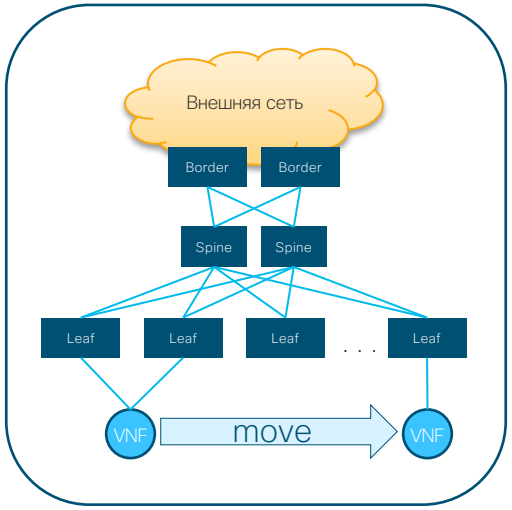
Centralized Routing Peering

Особенности в случае когда для пиринга используется одно выделенное устройство (или пара)



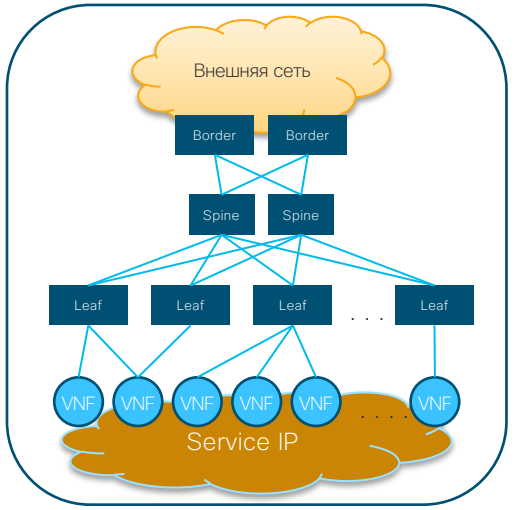
Мобильность VNF

Зависимости и требования



Anycast сервисы

Распределенный сегмент «Service IP или Service Subnet» с требованиями по более гранулярному распределению трафика per-VNF

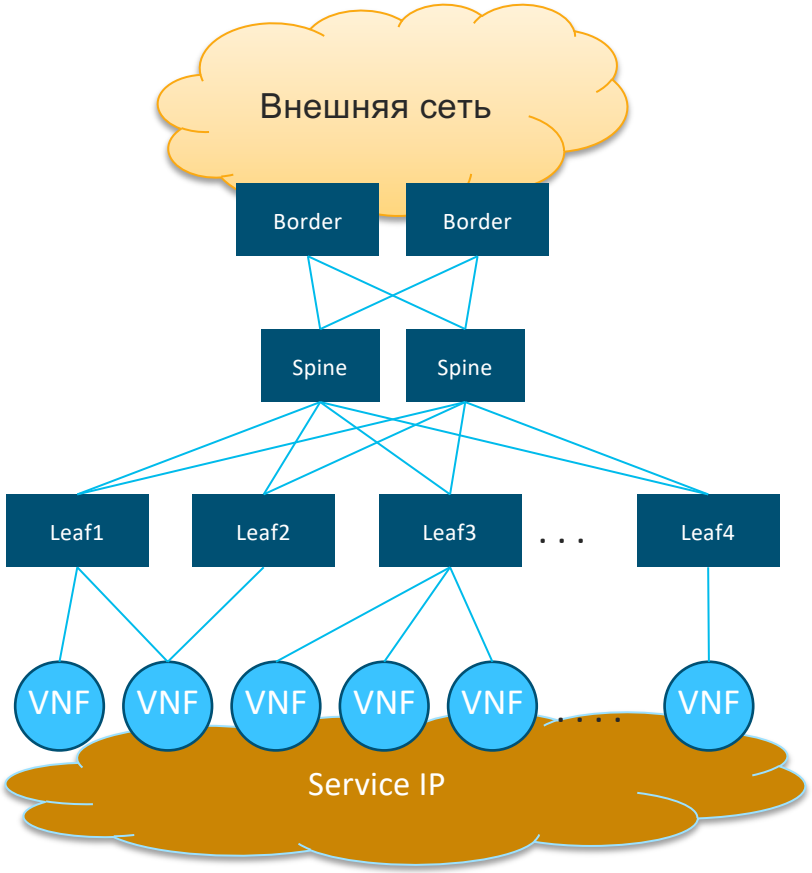


”Магическое” решение

Оптимизация передачи данных для VNF

Проблемы	EVPN AF	EVPN AF with VNF Extensions*
Avoid Traffic Tromboning (Recursive Routing)	Потенциальный риск	Да
VNF Mobility (Centralized Peering)	Возможно	Да
How Many VNF per Leaf? (Proportional Multipath)	Нет видимости	Да**

*draft-ietf-bess-evpn-inter-subnet-forwarding (Section 9.2)
**additional Optimization

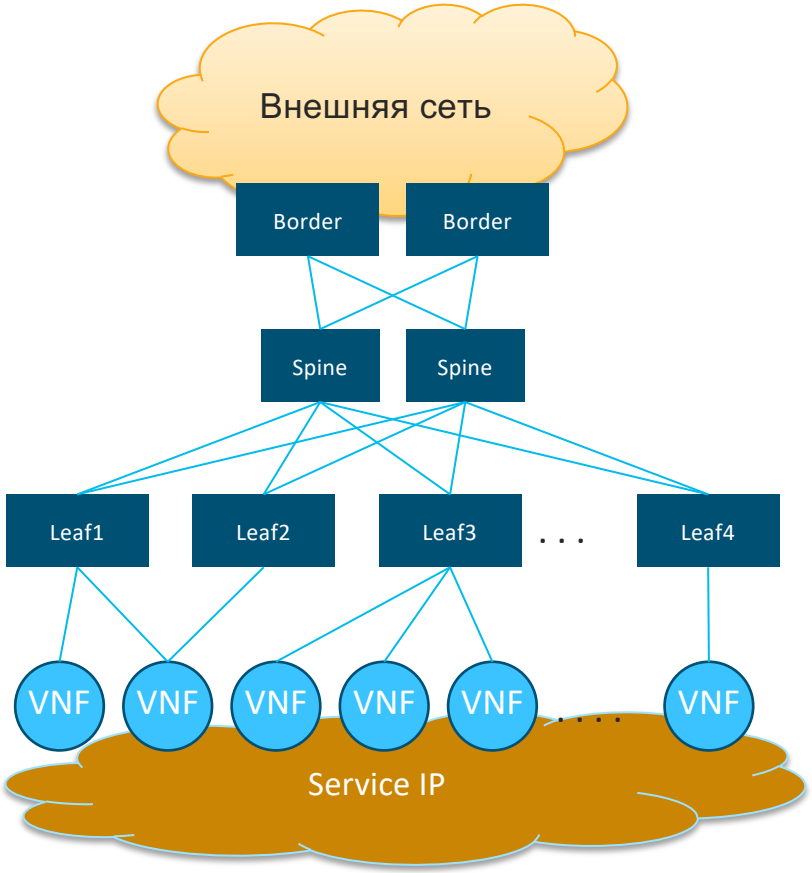


Оптимизация передачи данных для VNF

Проблемы	EVPN AF	EVPN AF with VNF Extensions*
Avoid Traffic Tromboning (Recursive Routing)	Потенциальный риск	Да
VNF Mobility (Centralized Peering)	Возможно	Да
How Many VNF per Leaf? (Proportional Multipath)	Нет видимости	Да**

*draft-ietf-bess-evpn-inter-subnet-forwarding (Section 9.2)
**additional Optimization

Обогащение EVPN маршрутов дополнительной информацией о “Service IP”



Анонсирование GW-IP
IETF-драфт, уже доступно в NX-OS

tools.ietf.org/html/draft-ietf-bess-evpn-prefix-advertisement-11

3.1 IP Prefix Route Encoding

An IP Prefix Route Type for IPv4 has the Length field set to 34 and consists of the following fields:

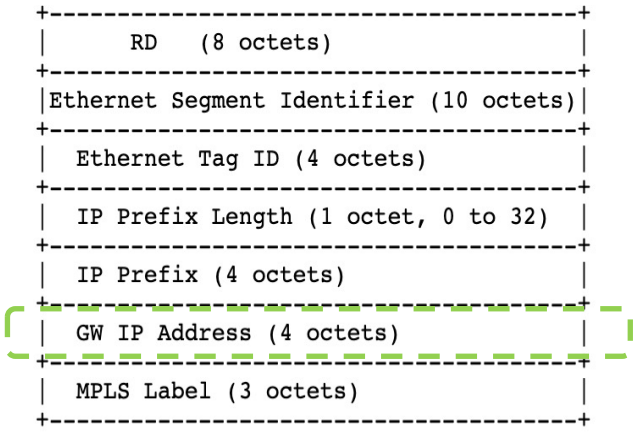


Figure 3 EVPN IP Prefix route NLRI for IPv4

An IP Prefix Route Type for IPv6 has the Length field set to 58 and consists of the following fields:

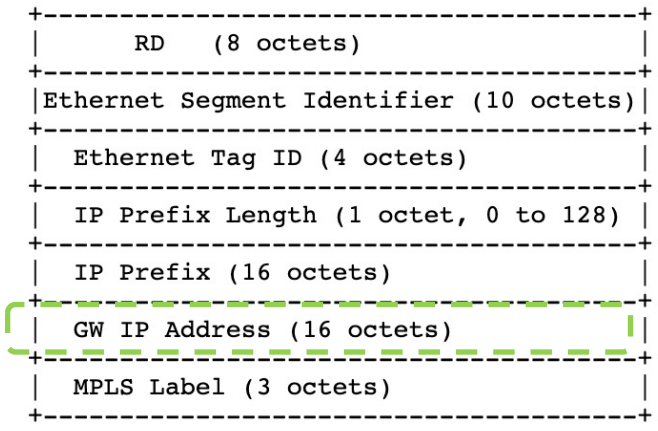


Figure 4 EVPN IP Prefix route NLRI for IPv6

Export Gateway-ip

- Для re-originated EVPN type-5 маршрутов имеется две опции обогащения маршрутной информации при помощи gateway-ip
  1. **“export-gateway-ip”**: полученный «next-hop» будет использоваться как атрибут gw-ip для EVPN type-5 маршрутов. Это происходит для всех маршрутов для AF и VRF где применяется настройка
  2. **“set evpn gateway next-hop”**: есть возможность для гранулярного выбора маршрутов, которые будут обогащены при помощи gateway-ip

Export Gateway-ip

```
router bgp 100
  ...
  vrf cust_1
    address-family ipv4 unicast
      export-gateway-ip
      ...
    address-family ipv6 unicast
      export-gateway-ip
      ...
```

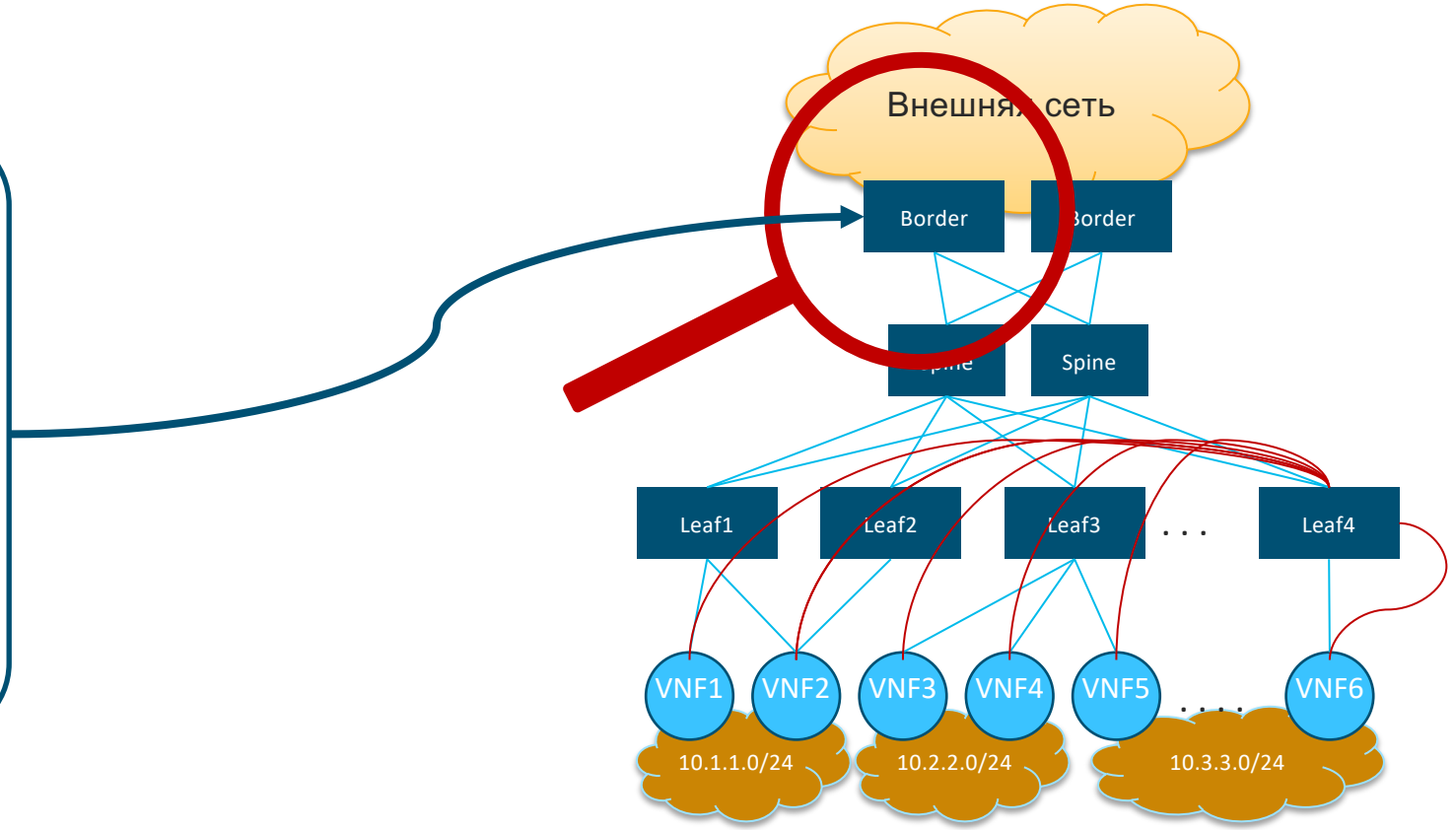
```
ip prefix-list vm-pfx-list seq 10 permit
200.0.0.0 eq 8

route-map export-l2evpn-rtmap permit 10
  match ip prefix-list vm-pfx-list
  set evpn gateway next-hop

route-map passall permit 10

Vrf context cust_1
  address-family ipv4 unicast
    export map export-l2evpn-rtmap
```

Recursive Routing for Centralized Routing Peering
VNF Forwarding Optimization (1)

С атрибутом Gateway-IP

Routing Table (EVPN Type5)

Prefix	Next-Hop	GW-IP
10.1.1.0/24	Leaf4	VNF1
10.1.1.0/24	Leaf4	VNF2
10.2.2.0/24	Leaf4	VNF3
10.2.2.0/24	Leaf4	VNF4
10.3.3.0/24	Leaf4	VNF5
10.3.3.0/24	Leaf4	VNF6

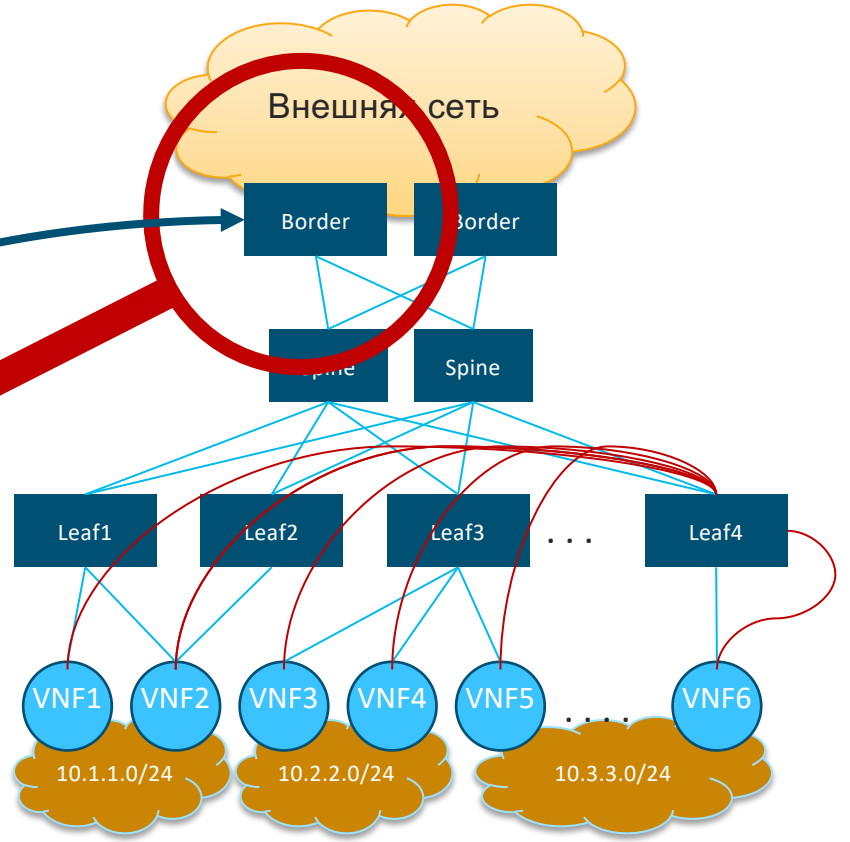


Recursive Routing for Centralized Routing Peering
VNF Forwarding Optimization (2)

С атрибутом Gateway-IP

Routing Table (EVPN Type5)

Prefix	Next-Hop	GW-IP
10.1.1.0/24	Leaf4	VNF1
10.1.1.0/24	Leaf4	VNF2
10.2.2.0/24	Leaf4	VNF3
10.2.2.0/24	Leaf4	VNF4
10.3.3.0/24	Leaf4	VNF5
10.3.3.0/24	Leaf4	VNF6



Как выучивается информация о местонахождении VNF?

Recursive Routing for Centralized Routing Peering
VNF Forwarding Optimization (3)

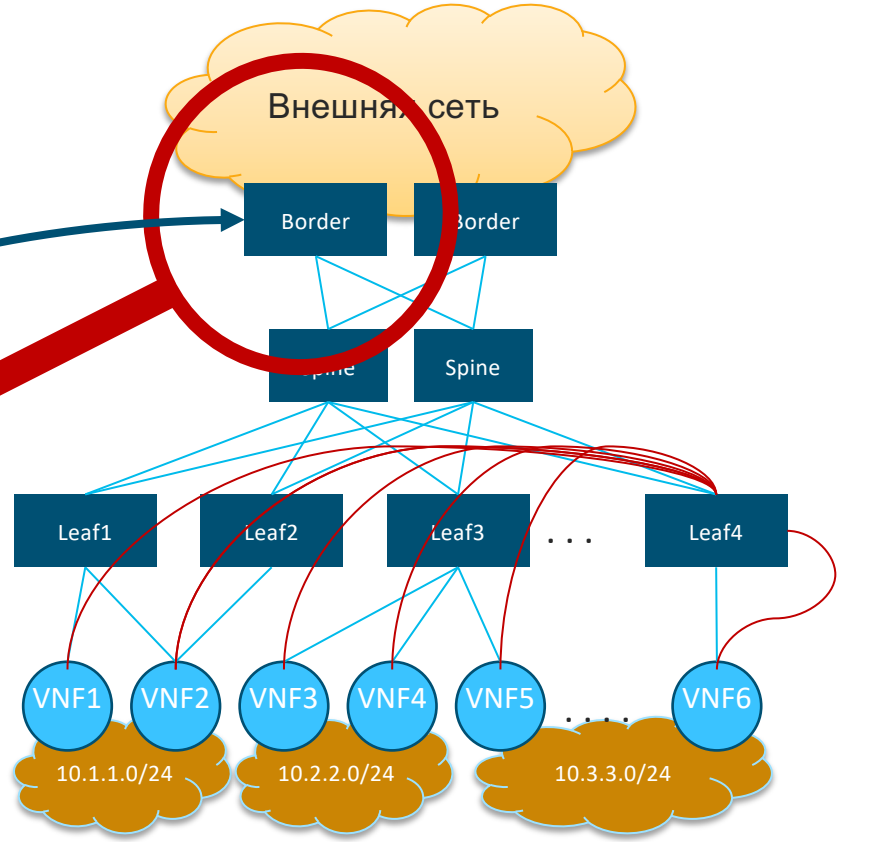
С атрибутом Gateway-IP

Routing Table (EVPN Type5)

Prefix	Next-Hop	GW-IP
10.1.1.0/24	Leaf4	VNF1
10.1.1.0/24	Leaf4	VNF2
10.2.2.0/24	Leaf4	VNF3
10.2.2.0/24	Leaf4	
10.3.3.0/24	Leaf4	
10.3.3.0/24	Leaf4	

Host Table (EVPN Type2)

Prefix	Next-Hop
VNF1	Leaf1
VNF2	Leaf1, Leaf2
VNF3	Leaf3
VNF4	Leaf3
VNF5	Leaf3
VNF6	Leaf4



VNF это фактически хосты подключенные к фабрике

Recursive Routing for Centralized Routing Peering
VNF Forwarding Optimization (4)

С атрибутом Gateway-IP

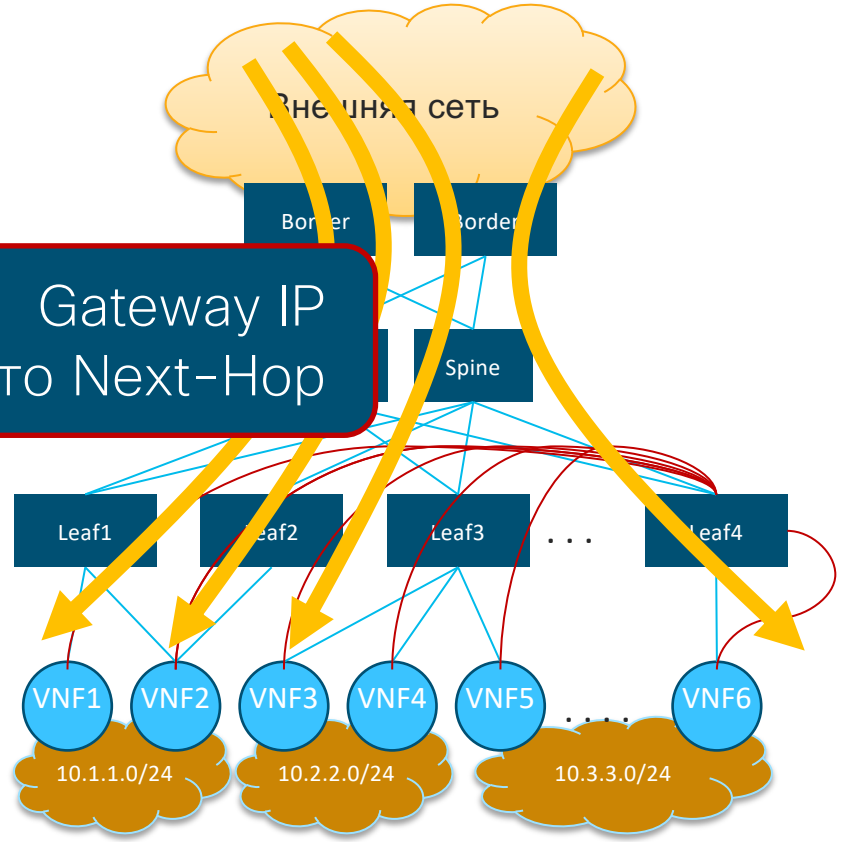
Routing Table (EVPN Type5)

Prefix	Next-Hop	GW-IP
10.1.1.0/24	Leaf4	VNF1
10.1.1.0/24	Leaf4	VNF2
10.2.2.0/24	Leaf4	VNF3
10.2.2.0/24	Leaf4	
10.3.3.0/24	Leaf4	
10.3.3.0/24	Leaf4	

Host Table (EVPN Type2)

Prefix	Next-Hop
VNF1	Leaf1
VNF2	Leaf1, Leaf2
VNF3	Leaf3
VNF4	Leaf3
VNF5	Leaf3
VNF6	Leaf4

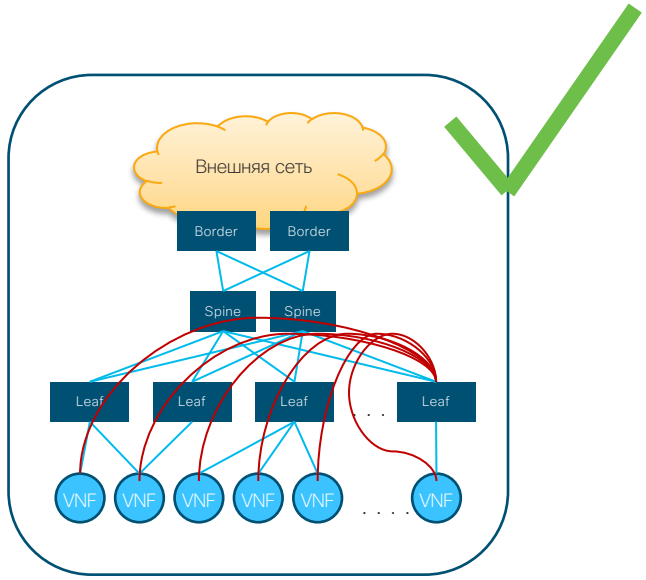
Используется Gateway IP (GW-IP) вместо Next-Hop



Сценарии развертывания
Проблема № 1 – решена

Centralized Routing Peering

Особенности в случае когда для пиринга используется одно выделенное устройство (или пара)



Recursive Routing для мобильности VNF
VNF Forwarding Optimization (1)

С атрибутом Gateway-IP

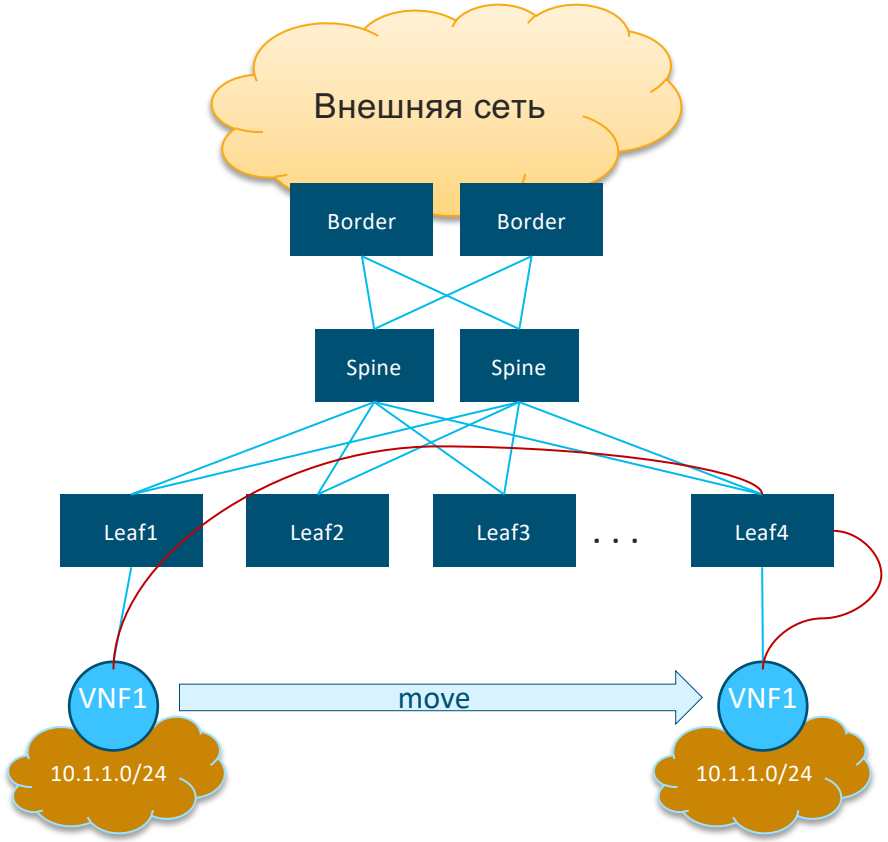
Routing Table (EVPN Type5)

Prefix	Next-Hop	GW-IP
10.1.1.0/24	Leaf4	VNF1
10.1.1.0/24	Leaf4	VNF1

Host Table (EVPN Type2)

Prefix	Next-Hop
VNF1	Leaf1 (Seq1)
…	…
VNF1	Leaf4 (Seq2)

Convergence



Recursive Routing для мобильности VNF
VNF Forwarding Optimization (2)

С атрибутом Gateway-IP

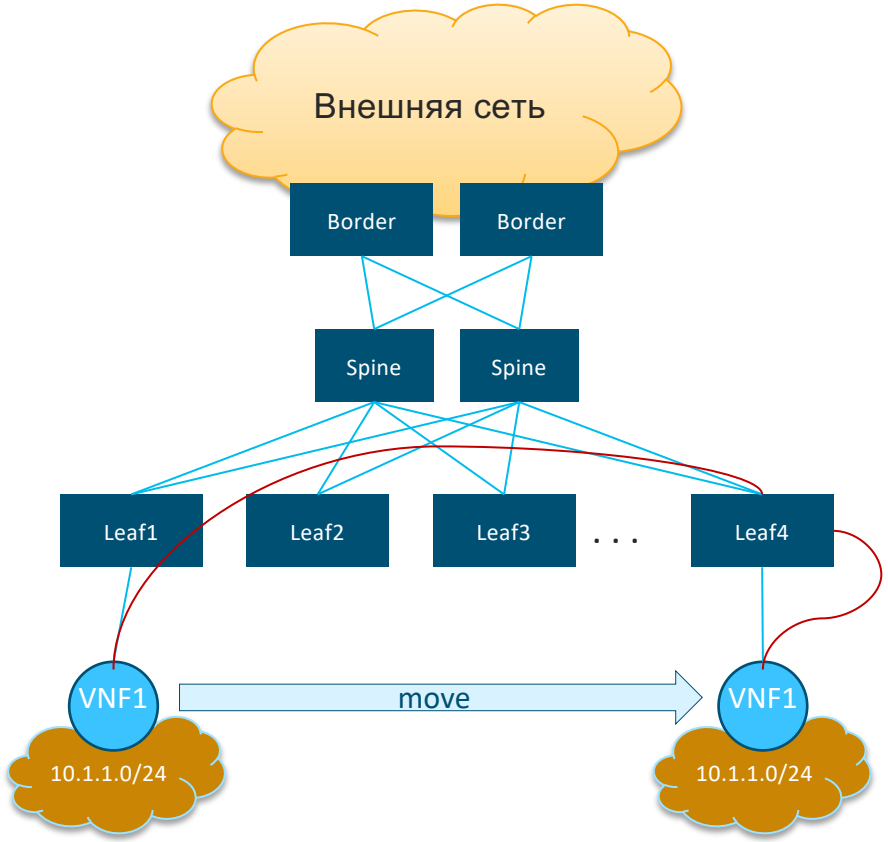
Routing Table (EVPN Type5)

Prefix	Next-Hop	GW-IP
10.1.1.0/24	Leaf4	VNF1
10.1.1.0/24	Leaf4	VNF1

Host Table (EVPN Type2)

Prefix	Next-Hop
VNF1	Leaf1 (Seq1)
…	…
VNF1	Leaf4 (Seq2)

Convergence



На изменение расположения VNF реагирует EVPN Host-Mobility; No Routing Change!

Recursive Routing для мобильности VNF
VNF Forwarding Optimization (3)

С атрибутом Gateway-IP

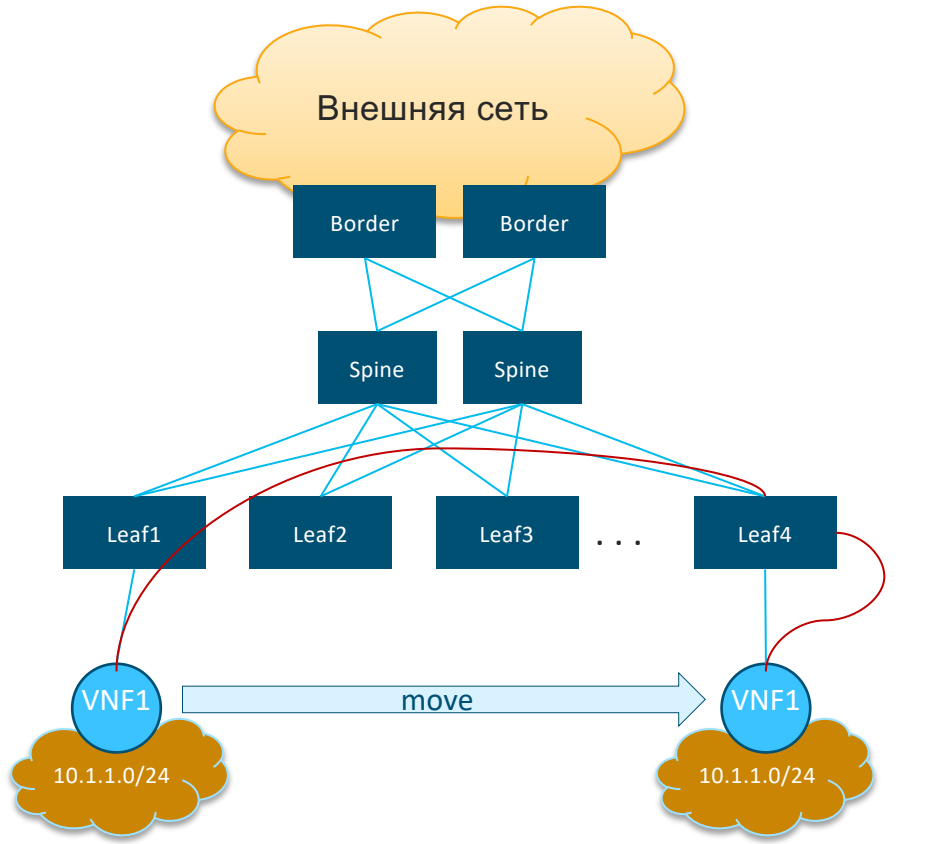
Routing Table (EVPN Type5)

Prefix	Next-Hop	GW-IP
10.1.1.0/24	Leaf4	VNF1
10.1.1.0/24	Leaf4	VNF1

Host Table (EVPN Type2)

Prefix	Next-Hop
VNF1	Leaf1 (Seq1)
…	…
VNF1	Leaf4 (Seq2)

Convergence

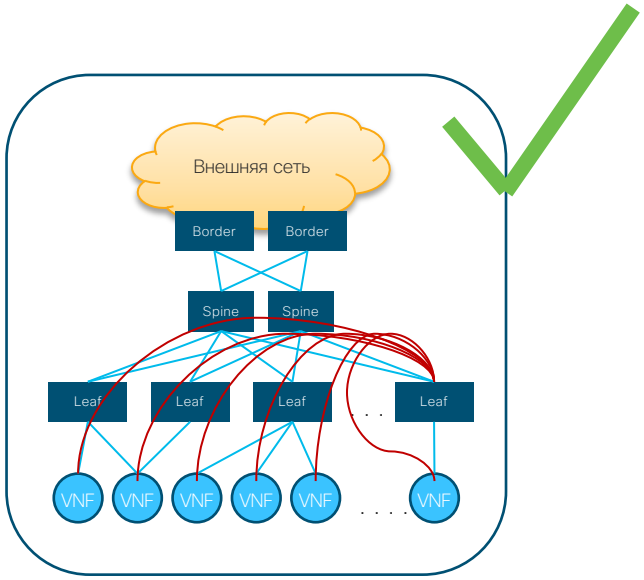


VNF это фактически хосты подключенные к фабрике

Сценарии развертывания
Проблема № 2 – решена

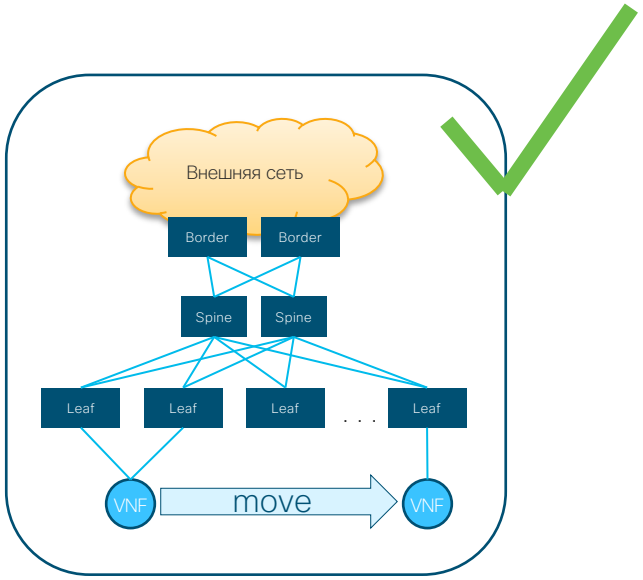
Centralized Routing Peering

Особенности в случае когда для пиринга используется одно выделенное устройство (или пара)

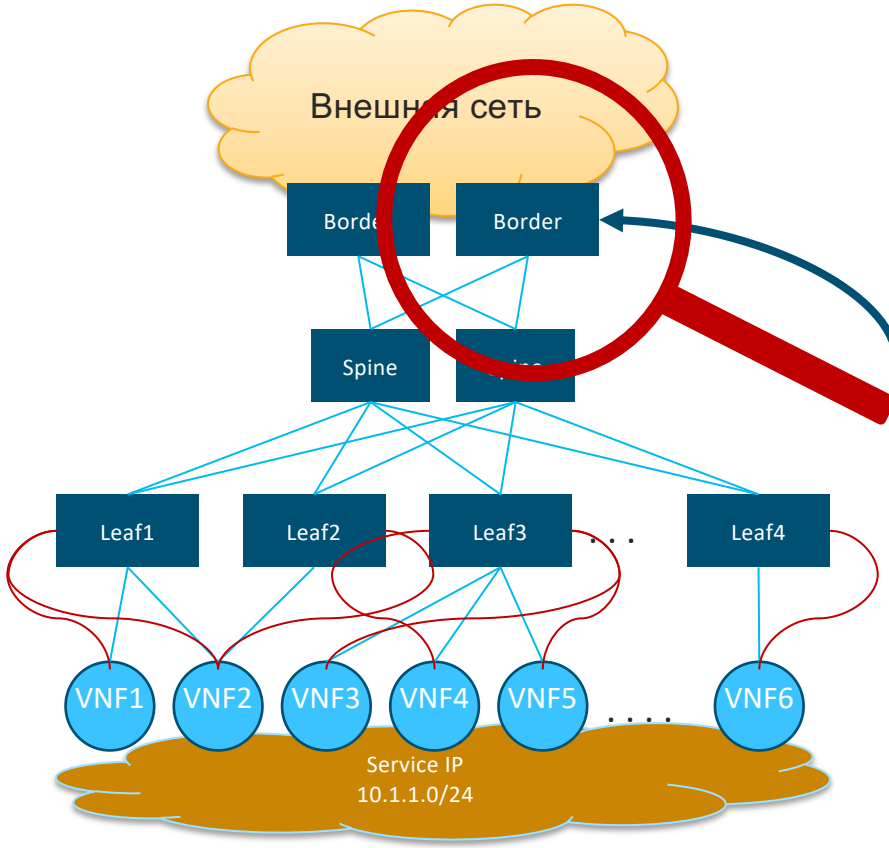


Мобильность VNF

Зависимости и требования



Proportional Multipath для Anycast сервисов
VNF Forwarding Optimization (1)

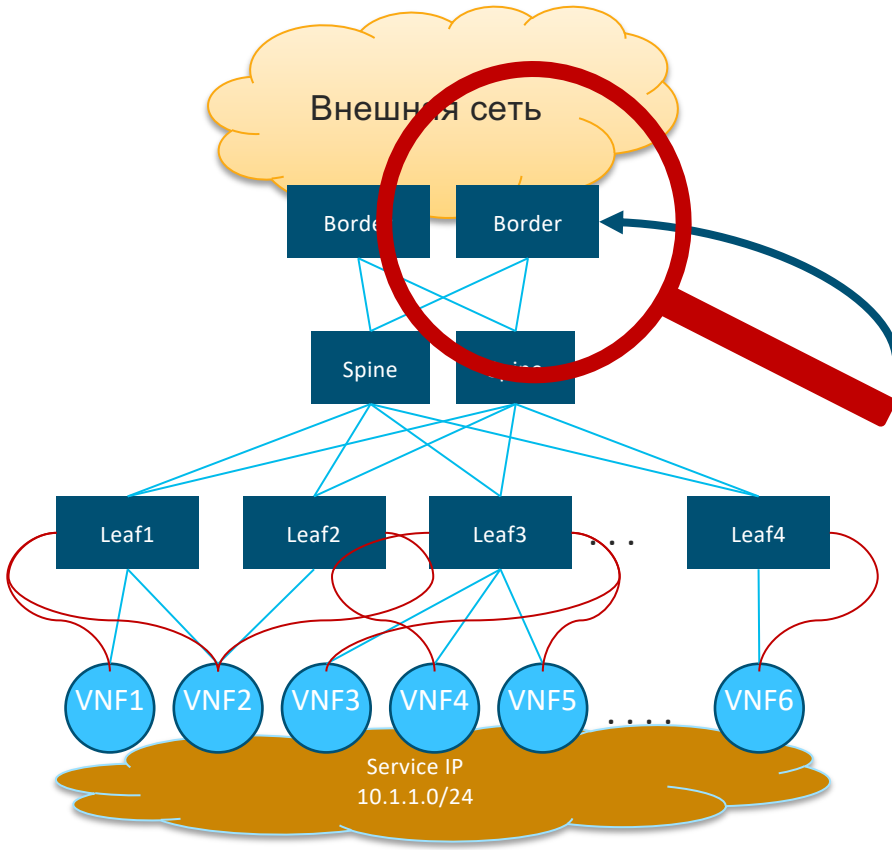


С атрибутом Gateway-IP

Routing Table (EVPN Type5)

Prefix	Next-Hop	GW-IP
10.1.1.0/24	Leaf1	VNF1
10.1.1.0/24	Leaf1	VNF2
10.1.1.0/24	Leaf2	VNF2
10.1.1.0/24	Leaf3	VNF3
10.1.1.0/24	Leaf3	VNF4
10.1.1.0/24	Leaf3	VNF5
10.1.1.0/24	Leaf4	VNF6

Proportional Multipath для Anycast сервисов
VNF Forwarding Optimization (2)



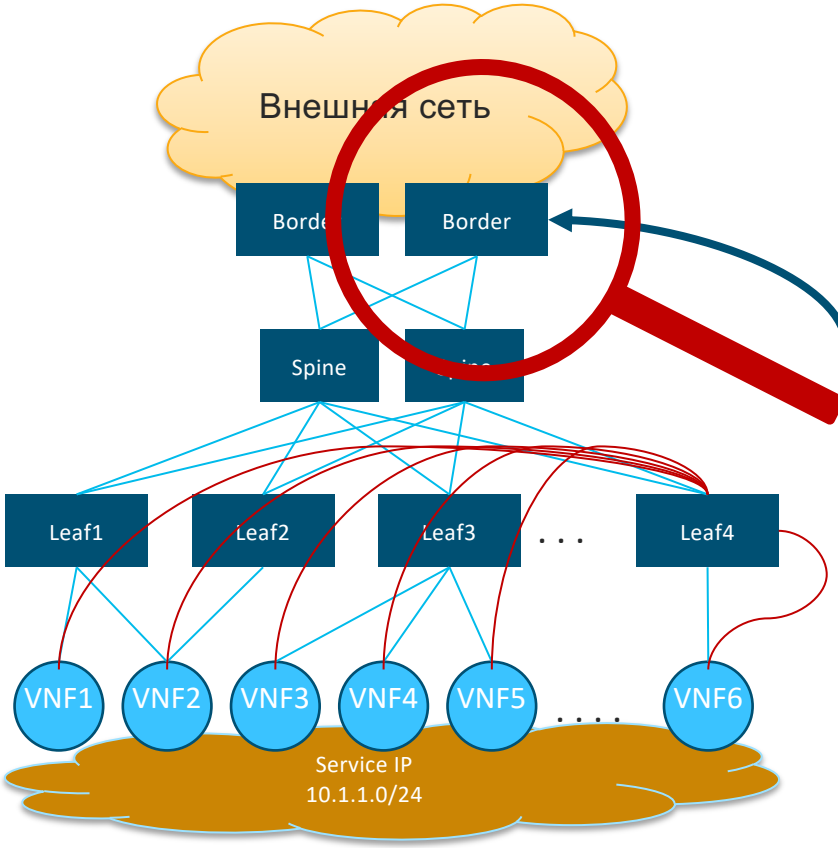
С атрибутом Gateway-IP

Routing Table (EVPN Type5)

Prefix	Next-Hop	GW-IP
10.1.1.0/24	Leaf1	VNF1
10.1.1.0/24	Leaf1	VNF2
10.1.1.0/24	Leaf2	VNF2
10.1.1.0/24	Leaf3	VNF3
10.1.1.0/24	Leaf3	VNF4
10.1.1.0/24	Leaf3	VNF5
10.1.1.0/24	Leaf4	VNF6

Каждый маршрут, который анонсирует VNF выучивается индивидуально; Leaf не агрегирует эту информацию

Proportional Multipath для Anycast сервисов
VNF Forwarding Optimization (3)



С атрибутом Gateway-IP

Routing Table (EVPN Type5)

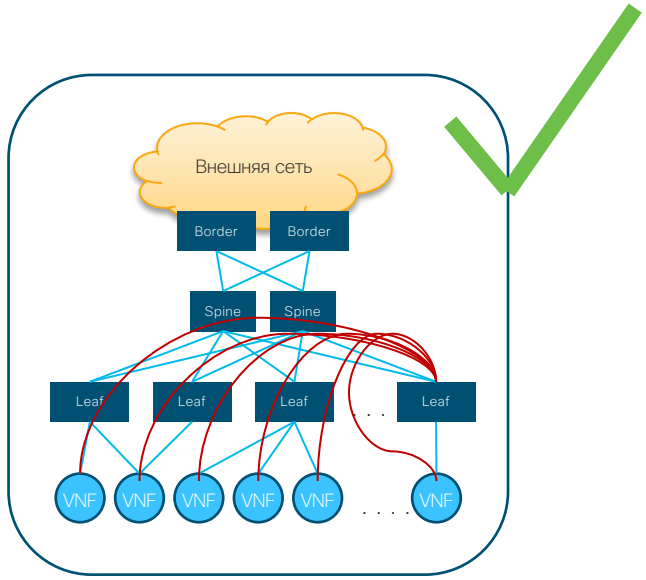
Prefix	Next-Hop	GW-IP
10.1.1.0/24	Leaf4	VNF1
10.1.1.0/24	Leaf4	VNF2
10.1.1.0/24	Leaf4	VNF3
10.1.1.0/24	Leaf4	VNF4
10.1.1.0/24	Leaf4	VNF5
10.1.1.0/24	Leaf4	VNF6

Может быть использовано совместно Centralized Routing Peering

Сценарии развертывания
Проблема № 3 – решена

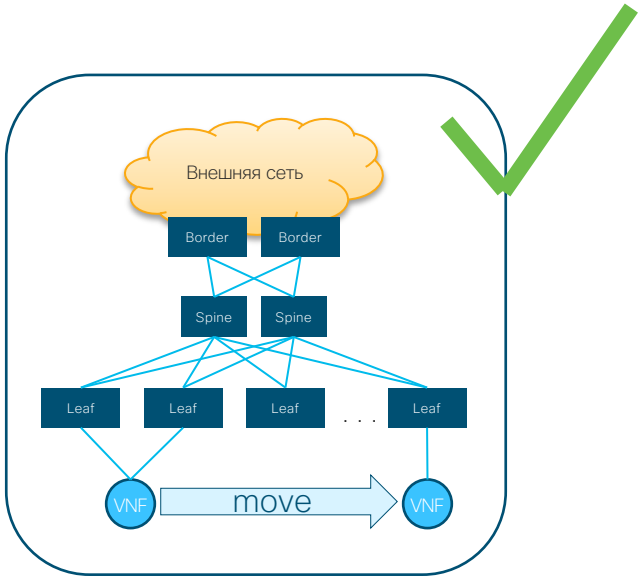
Centralized Routing Peering

Особенности в случае когда для пиринга используется одно выделенное устройство (или пара)



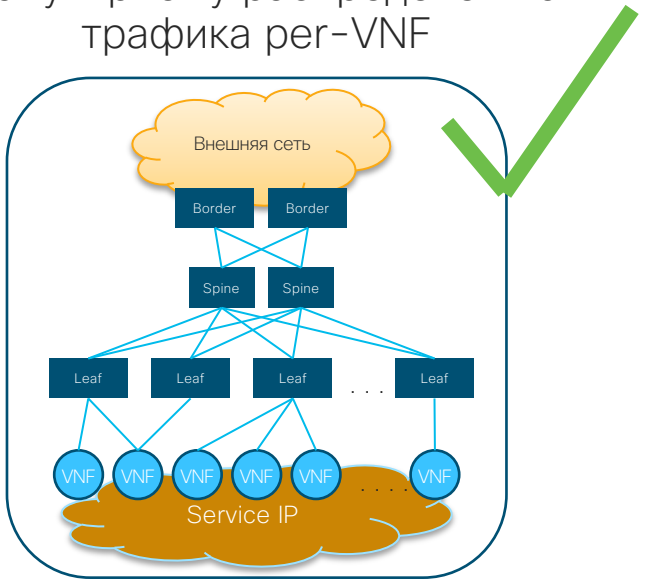
Мобильность VNF

Зависимости и требования



Anycast сервисы

Распределенный сегмент «Service IP или Service Subnet» с требованиями по более гранулярному распределению трафика per-VNF



Proportional Multipath for VNF

Nexus 9000 VXLAN Configuration Guide (NX-OS 9.3)

https://www.cisco.com/c/en/us/td/docs/switches/datacenter/nexus9000/sw/93x/vxlan/configuration/guide/b-cisco-nexus-9000-series-nx-os-vxlan-configuration-guide-93x/b-cisco-nexus-9000-series-nx-os-vxlan-configuration-guide-93x_appendix_011010.html

Рекомендации по использованию “export-gateway-ip”

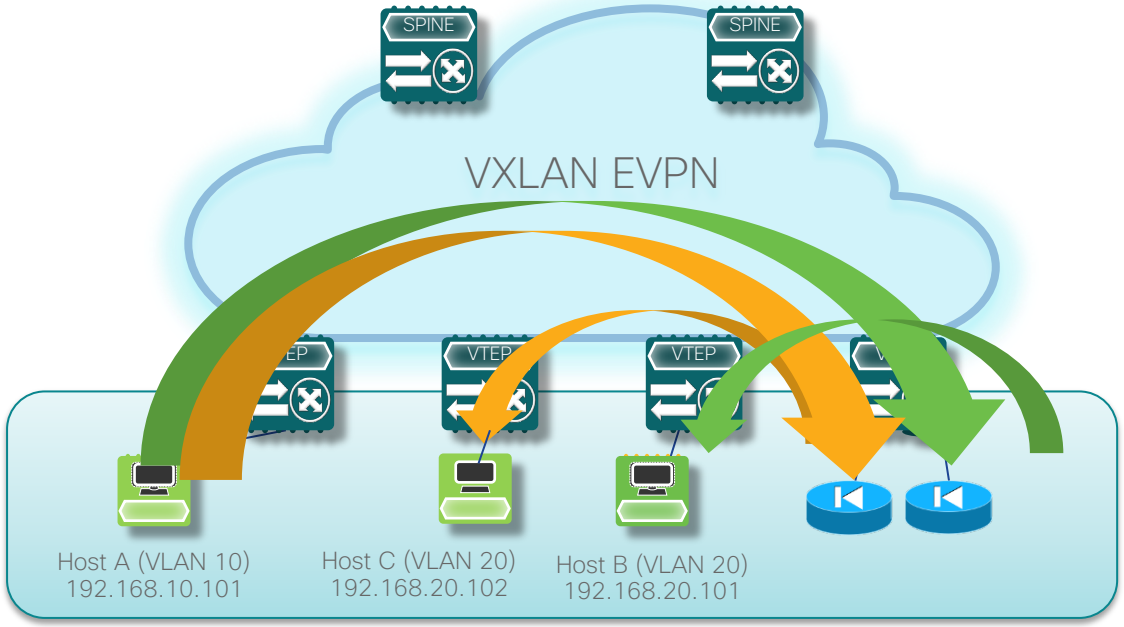
- Использование опции “export-gateway-ip” только для рассматриваемых сценариев, не используйте как настройку по умолчанию для всего
  - Используйте “export-gateway-ip” где требуется
  - При необходимости задействуйте Route-Map
- Рекомендуемый релиз – начиная с NX-OS 9.3(5)
  - Known route evaluation based practice based Administrative Distance (AD)
  - Option available for Mixed-Path
  - Рекомендован к использованию
- Поддерживаются следующие модели
  - Cisco Nexus 9364C, 9300-EX и 9300-FX/FX2/FX3 платформы; Cisco Nexus 9500 платформа с модулями фабрик N9K-C9508-FM-E2 и -EX или -FX линейными картами
- Перемещения VNF между сайтами не поддерживаются

Часть 4: Подключение сервисных устройств при помощи EPBR

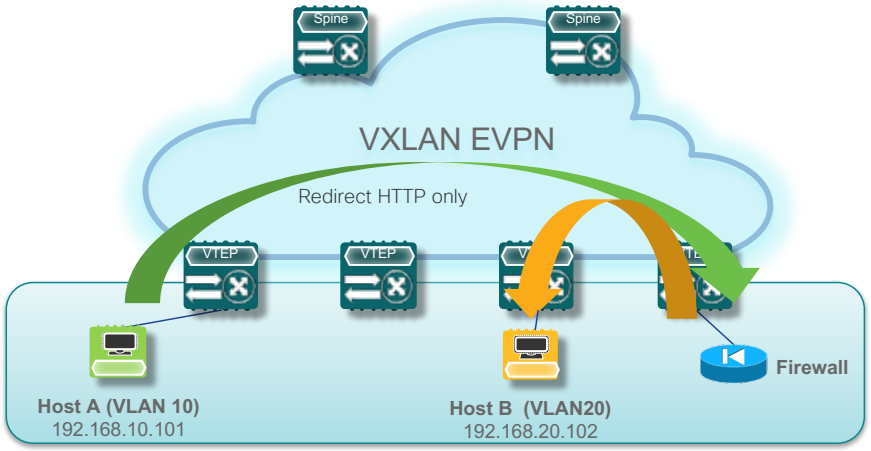
Новая функция: Enhanced Policy Based Redirection

Что обеспечивает EPBR:

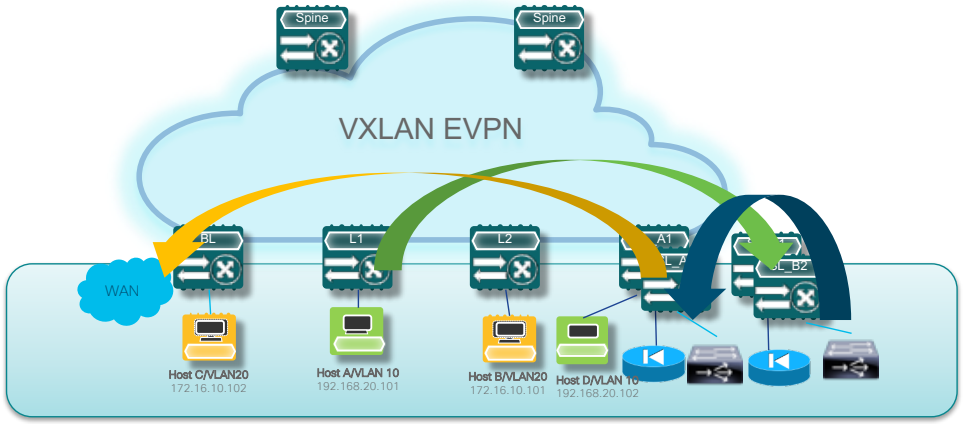
- Возможность описать сервисные устройства, подключенные к фабрике
- Возможность выбора конкретного трафика для перенаправления
  - Оптимизация используемых ресурсов на сервисном устройстве
- Возможность горизонтального масштабирования
  - При помощи symmetric load-balancing
- Упрощение конфигурации PBR политик
- Мониторинг состояния здоровья
  - Гибкие «пробники» (probes)
- Работает на аппаратном уровне!! (не влияет на производительность)



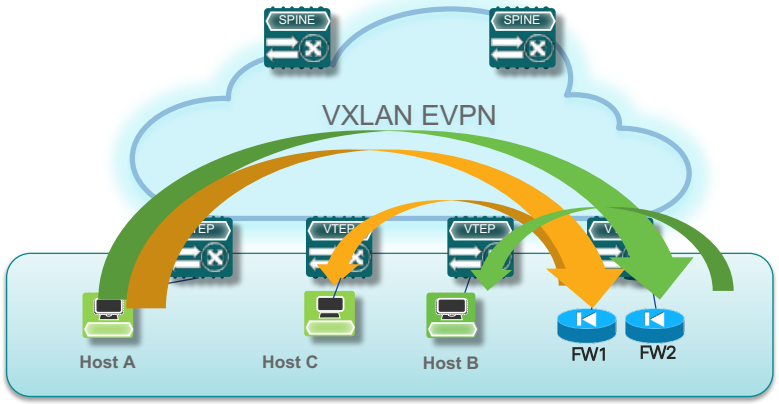
Сценарии для EPBR



Выборочное перенаправление трафика



Выборочное перенаправление на разные сервисные цепочки



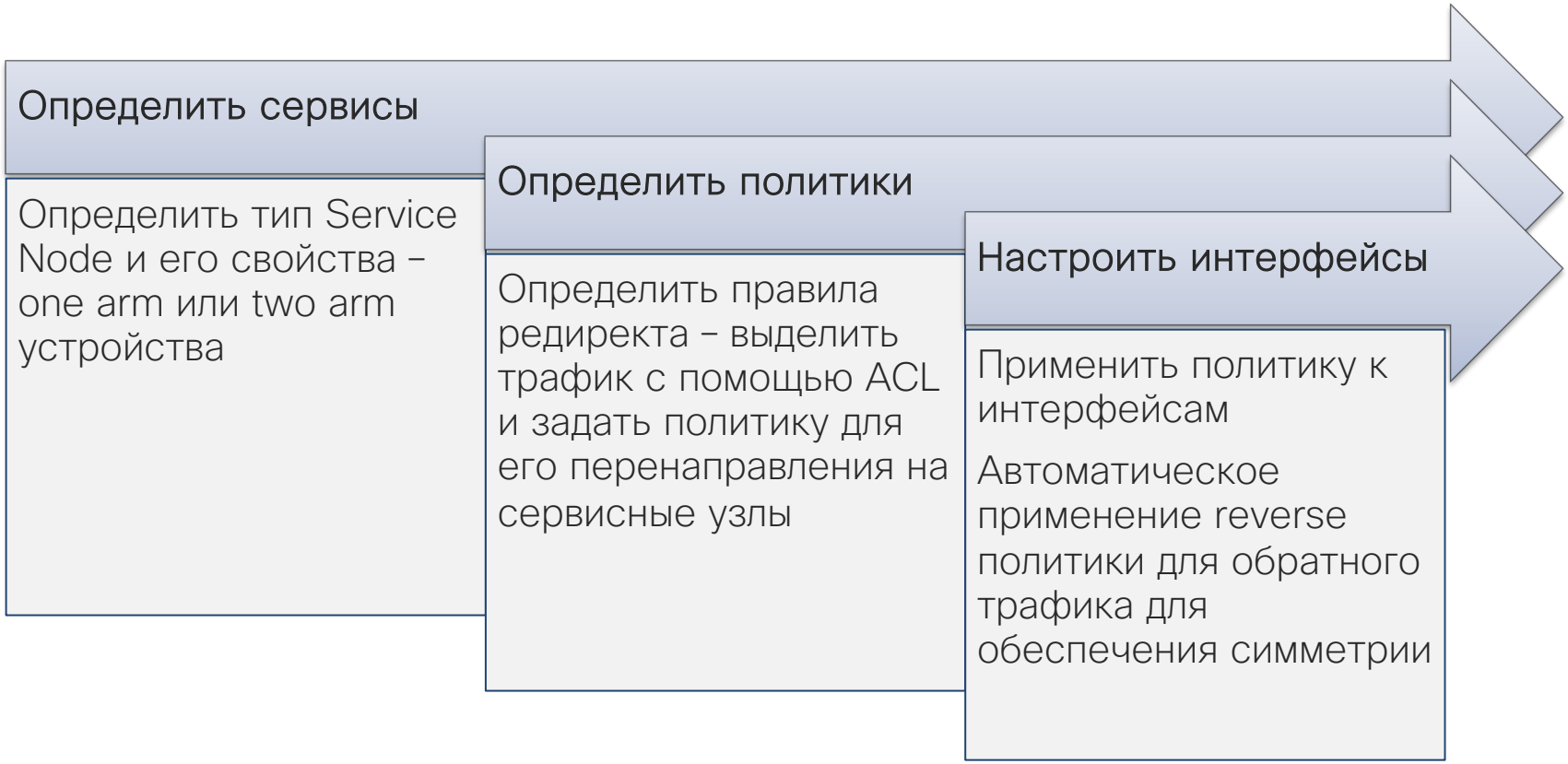
Перенаправление + балансировка

И другие ..

EPBR это:

- Сервисные цепочки – последовательность сервисов
- L3 балансировка средствами фабрики
- Собственно PBR
- Проверка “жизни”

EPBR – простота реализации



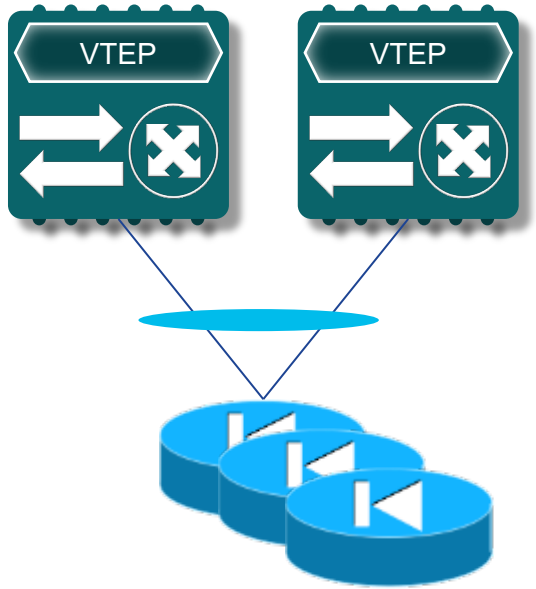
Требуется лицензия Advantage

https://www.cisco.com/c/en/us/td/docs/switches/datacenter/sw/nx-os/licensing/guide/b_Cisco_NX-OS_Licensing_Guide/b_Cisco_NX-OS_Licensing_Guide_chapter_01.html

EPBR – определение сервиса и мониторинг

- Определить свойства сервисного устройства
  - IP адрес
  - Принадлежность VRF
  - Интерфейс (L3/SVI)
  - Probe
  - Reverse IP (если это two arm устройство)
  - Если кластер, то всех его участников определить под одним «service node»

- Доступные следующие методы мониторинга сервисного устройства – probes:
  - ICMP
  - TCP
  - UDP
  - DNS
  - HTTP



EPBR – синтаксис определения сервиса

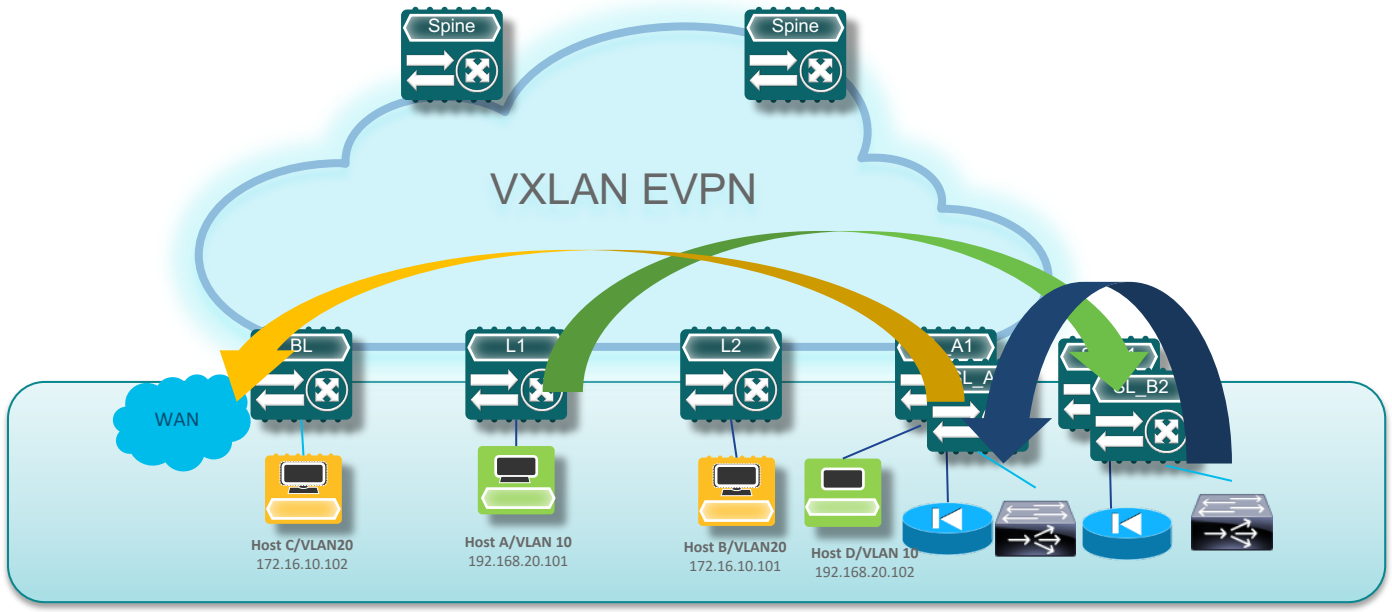
```
epbr service <service_name>
   [probe <probe_type> [probe_parameters]]
   [service-interface <interface_name>]
   [vrf <vrf_name>]
   service-endpoint {ip <ipv4_address> | ipv6 <ipv6_address>} [interface <interface_name>]
                    [probe <probe_type> [probe_parameters]]
                   [reverse {ip <ipv4_address> | ipv6 <ipv6_address>} [interface <interface_name>]]
   service-endpoint {ip <ipv4_address> | ipv6 <ipv6_address>} [interface <interface_name>]
                    [probe <probe_type> [probe_parameters]]
                   [reverse {ip <ipv4_address> | ipv6 <ipv6_address>} [interface <interface_name>]]
```

Пример описания МСЭ кластера из 2-х узлов в режиме Active-Active

epbr service FIREWALL_CLUSTER_A
  probe icmp
  vrf TENANT_A
  service-endpoint ip 172.16.1.200 interface VLAN100 *!!! Interface facing the firewall*
        reverse ip 172.16.2.200 interface VLAN101 *!!! applies for two arm device - this IP maps to the interface IP of the service node 2nd ARM Reverse ip*
  service-endpoint ip 172.16.1.201 interface VLAN100
        reverse ip 172.16.2.201 interface VLAN101

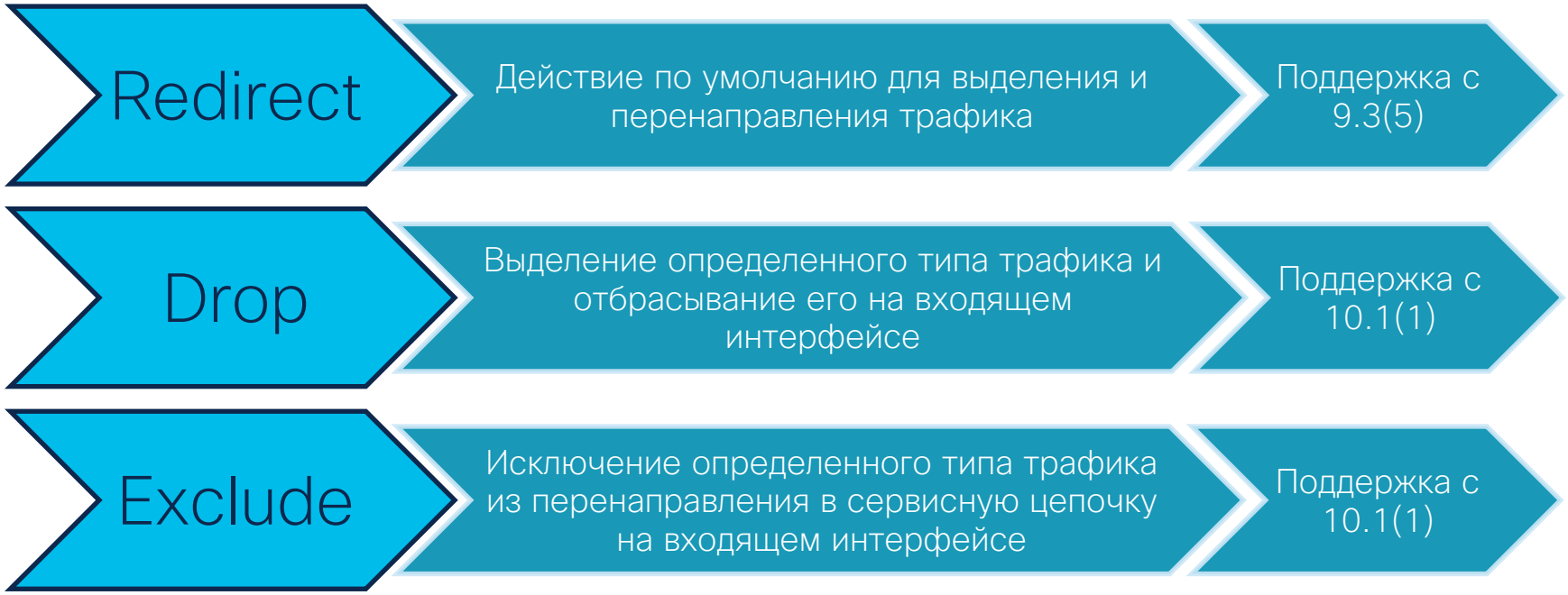
EPBR – определение политики

- Определить правила перенаправления трафика
  - Выбор трафика для редиректа при помощи ACL (только правила Permit)
  - Set Service – определить сервисное устройство для перенаправления
  - Определение метода балансировки
  - Определение fail-action – поведение в случае выхода из строя сервисного устройства



EPBR – типы действий для ACL в политике

- Для match записи в политике ePBR поддерживается три типа действий



- В политике EBPR можно определять несколько match записей с разными ACL
- В политике EBPR может быть только одна match запись с действием drop и/или exclude ACL

EPBR – синтаксис определения политики

```
epbr policy <policy_name>
   match {[ip address <ipv4_acl_name>] | [ipv6 address<ipv6_acl_name>]} [redirect | drop |exclude]
   <seq_no> set service <service_name> [fail-action <bypass|drop|forward>] [load-balance method <src-ip|dst_ip>]
   <seq_no> set service <service_name> [fail-action <bypass|drop|forward>] [load-balance method <src-ip|dst_ip>]
   match {[ip address <ipv4_acl_name>] | [ipv6 address<ipv6_acl_name>]} [redirect | drop |exclude]
   <seq_no> set service <service_name> [fail-action <bypass|drop|forward>] [load-balance method <src-ip|dst_ip>]
```

epbr policy Tenant_A-Redirect

match ip address exclude_acl exclude
match ip address drop_acl drop
match ip address WEB !!! ACL matches web traffic
10 **set service FIREWALL fail-action drop load-balance method src-ip** *!! Firewall cluster is first element in chain. If Firewall is down, drop the trafic*
20 **set service TCP_Optimizer fail-action bypass drop load-balance method src-ip** *!! TCP Optimizer is 2nd element in the chain. If it is down, simply bypass it*
match ip address APP
10 **set service FIREWALL fail-action drop load-balance method src-ip**

Если МСЭ недоступен, трафик отбрасывается, если TCP optimizer недоступен, то трафик передается дальше

Node Fail-Action – поведение в случае выхода из строя сервисного устройства

EPBR поддерживает следующие действия

Bypass - трафик передается на следующее устройство в случае если текущее устройство вышло из строя

Drop - трафик отбрасывается

Forward - трафик передается по правилам таблицы маршрутизации на устройстве

EPBR – применение политики

- Применение политики
  - Применение на интерфейсах
    - Tenant SVI
    - L3VNI SVI
    - L3 Interfaces
- Направление редиректа
  - Forward (по умолчанию)
  - Reverse
    - Автоматическое создание правил редиректа для обратного трафика

EPBR – применение политики (синтаксис)

```
interface <interface_name>
   epbr <ip|ipv6> policy <policy_name> [reverse]
```

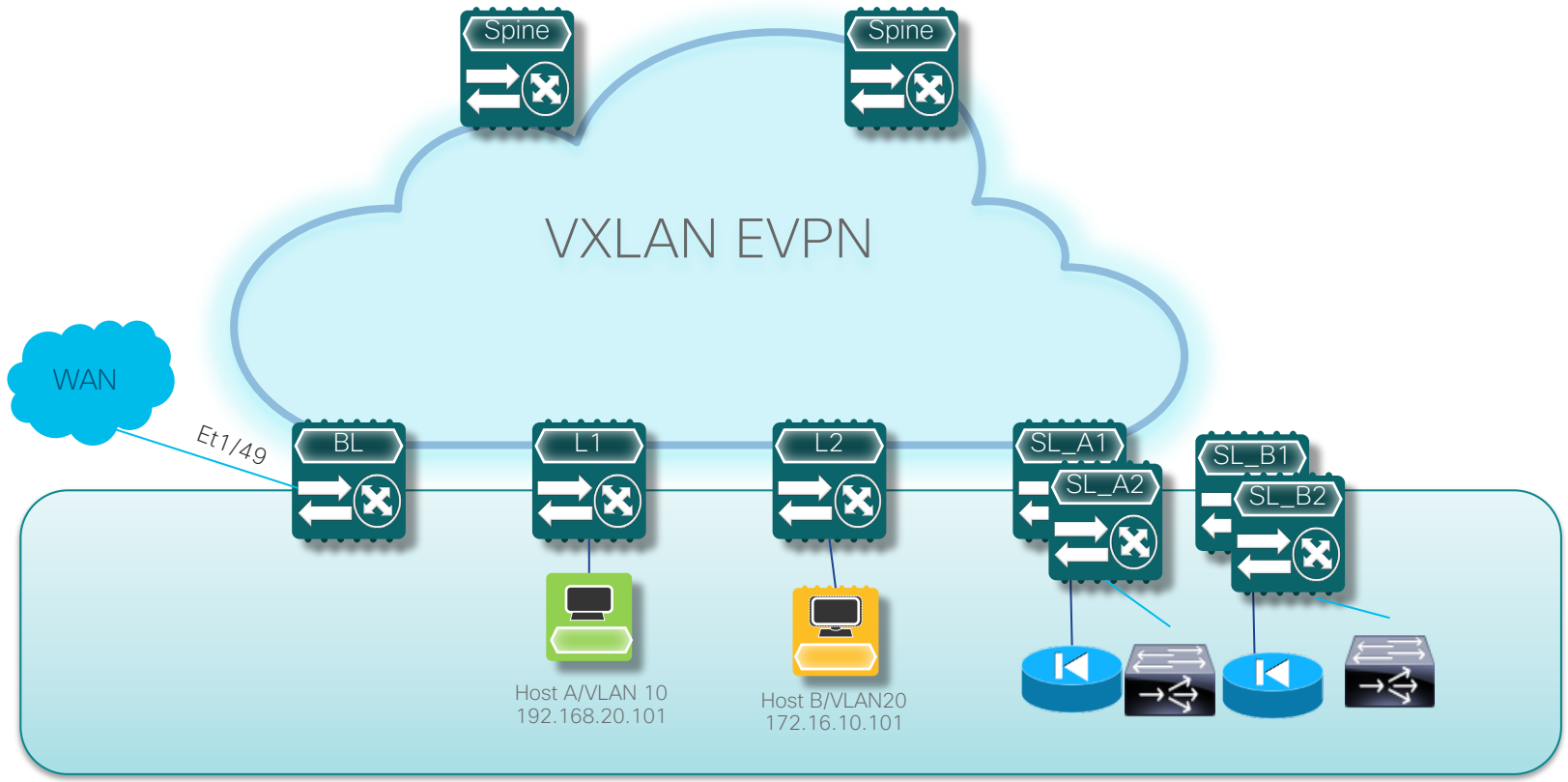
interface Vlan10
  !Tenant SVI
  epbr ip policy **Tenant_A-Redirect**

interface vlan 2010
  !L3 VNI SVI
  epbr ip policy Tenant_A-Redirect
  epbr ip policy Tenant_A-Redirect reverse *!! User can optionally define reverse chain sequence for symmetric traffic. The matc ACL src/dst is reversed. The elements in chain is reversed. If load-balancing was src-ip then in reverse the load-balancing becomes dst-ip*

Для обратного трафика все правила «переворачиваются»
И также для обратного трафика тратятся ресурсы TCAM

- Примеры использования ePBR – Подключение МСЭ и LB в режиме Active/Standby

Сценарий использования ePBR вместе с VxLAN фабрикой



Service Nodes в режиме Active/Standby

FW_inside в VLAN 101

FW_outside в VLAN 102

LB в VLAN 200

FW и LB в одном VRF как хосты

Требования

- Приложение 1 (собственное):
  - Весь трафик идущий с порта 7800 из VLAN 10 и VLAN 20 должен попасть сначала на LB потом на FW_IN, FW_OUT и потом передаваться по правилам маршрутизации
- Web приложение:
  - Весь трафик на порт 80 из VLAN 10 и 20 должен попасть сначала на FW_IN, FW_OUT и потом передаваться по правилам маршрутизации
- Весь остальной трафик передается по правилам маршрутизации
- Правила Fail-action
  - Если LB недоступен, bypass.
  - Если FW недоступен, drop.
- Для обратного трафика должна поддерживаться симметрия

Настройка на серверных Leaf-коммутаторах

```
epbr service FIREWALL
   vrf TENANT_A
   service-endpoint ip 172.16.1.200 interface VLAN100
      probe icmp frequency 4 retry-down-count 1 retry-up-count 1
           timeout 2 source-interface loopback9
     reverse ip 172.16.2.200 interface VLAN101
      probe icmp frequency 4 retry-down-count 1 retry-up-count 1
           timeout 2 source-interface loopback10
epbr service LB
   vrf TENANT_A
   service-endpoint ip 172.17.1.200 interface VLAN200
      probe tcp 7800 enable source-interface loopback9

ip access-list custom_app permit tcp 172.16.10.0/24 eq 7800 any
ip access-list custom_app permit tcp 192.168.20.0/24 eq 7800 any

ip access-list web permit tcp 172.16.10.0/24 any eq 80
ip access-list web permit tcp 192.168.20.0/24 any eq 80

ebpr policy P1
 statistics
  match ip address custom_app
    load-balance buckets 16 method src-ip
    10 set service LB fail-action bypass
    20 set service FIREWALL fail-action drop
  match ip address web
    load-balance buckets 16 method src-ip
    10 set service FIREWALL fail-action drop
```



```
interface vlan 10
 vrf member tenant_a
 ip address 172.16.10.1/24
 fabric forwarding mode anycast-gateway
 epbr policy P1

interface vlan 20
 vrf member tenant_a
 ip address 172.16.20.1/24
 fabric forwarding mode anycast-gateway
 epbr policy P1
```

- Политика применяется на anycast gateway SVI
- Трафик от серверов передается на первое сервисное устройство

Настройка на сервисных Leaf-коммутаторах

```
epbr service FIREWALL
   vrf TENANT_A
   service-endpoint ip 172.16.1.200 interface VLAN100
      probe icmp frequency 4 retry-down-count 1 retry-up-count 1
           timeout 2 source-interface loopback9
     reverse ip 172.16.2.200 interface VLAN101
      probe icmp frequency 4 retry-down-count 1 retry-up-count 1
           timeout 2 source-interface loopback10
epbr service LB
   vrf TENANT_A
   service-endpoint ip 172.17.1.200 interface VLAN200
      probe tcp 7800 enable source-interface loopback9

ip access-list custom_app permit tcp 172.16.10.0/24 eq 7800 any
ip access-list custom_app permit tcp 192.168.20.0/24 eq 7800 any

ip access-list web permit tcp 172.16.10.0/24 any eq 80
ip access-list web permit tcp 192.168.20.0/24 any eq 80

ebpr policy P1
 statistics
  match ip address custom_app
    load-balance buckets 16 method src-ip
    10 set service LB fail-action bypass
    20 set service FIREWALL fail-action drop
  match ip address web
    load-balance buckets 16 method src-ip
    10 set service FIREWALL fail-action drop
```



```
Interface vlan 999
 ! L3 VNI SVI
 vrf member tenant_a
 ip forward
 epbr policy P1
 epbr policy P1 reverse
```

- Политика применяется на L3VNI SVI
- Политика применяется к прямому и обратному трафику

Настройка на Border Leaf-коммутаторах

```
epbr service FIREWALL
   vrf TENANT_A
   service-endpoint ip 172.16.1.200 interface VLAN100
      probe icmp frequency 4 retry-down-count 1 retry-up-count 1
           timeout 2 source-interface loopback9
     reverse ip 172.16.2.200 interface VLAN101
      probe icmp frequency 4 retry-down-count 1 retry-up-count 1
           timeout 2 source-interface loopback10
epbr service LB
   vrf TENANT_A
   service-endpoint ip 172.17.1.200 interface VLAN200
      probe tcp 7800 enable source-interface loopback9

ip access-list custom_app permit tcp 172.16.10.0/24 eq 7800 any
ip access-list custom_app permit tcp 192.168.20.0/24 eq 7800 any

ip access-list web permit tcp 172.16.10.0/24 any eq 80
ip access-list web permit tcp 192.168.20.0/24 any eq 80

ebpr policy P1
 statistics
  match ip address custom_app
    load-balance buckets 16 method src-ip
    10 set service LB fail-action bypass
    20 set service FIREWALL fail-action drop
  match ip address web
    load-balance buckets 16 method src-ip
    10 set service FIREWALL fail-action drop
```



```
interface Eth1/49
 epbr policy P1 reverse
```

- Политика применяется на внешнем интерфейсе border leaf
- Трафик, который попадает на border со внешнего интерфейса это возвращающийся трафик ‘return traffic’
- Политика применяется только в обратном направлении (reverse direction)

EPBR – Ограничения и масштабирование

Требования к аппаратуре и ПО

- EPBR доступен с релиза Cisco NX-OS 9.3(5) нас Cisco Cloud Scale ASICs
- Nexus 9500 Series с линейными картами EX, FX
- Nexus 9300 EX, FX и FX2

EPBR рекомендации и ограничения для VxLAN фабрики

- In VXLAN fabric, service chaining cannot be done to devices within same VLAN. All devices must be present in separate VLANs.
- Currently EPBR is supported within a single site in a VXLAN multisite fabric.
- Active/Standby chain is supported with two service nodes with no restrictions.
- Active/Standby chain with three or more service nodes in chain requires no two nodes of different type behind same service leaf
- Loopback interfaces must be configured as source interface for probes and placed in same VRF as the service nodes
- Each ACL used within a EPBR policy match statement can contain up to 256 entries (ACEs)

EPBR – правила расчета требуемых TCAM ресурсов

- EPBR создает PBR правила и использует ING-RACL TCAM
- Каждая запись ACL использует одну или несколько TCAM entry
- Если одна и та же политика применяется на нескольких интерфейсах в одном VRF, то TCAM entry переиспользуются.
- При использовании балансировки количество программируемых ACE-ов зависит от количества buckets настроенных в ePBR политике. 8 buckets означает TCAM usage =8x ACE-ов в match ACL.
- Forward и reverse (return) – две разные политики, значит используют 2x TCAM ресурсов при применении reverse политики.

```
show hardware access-list tcam region
                                        NAT ACL[nat] size =    0
                              Ingress PACL [ing-ifacl] size =    0
                                        VACL [vacl] size =    0
                             Ingress RACL [ing-racl] size = 1792
                           Ingress RBACL [ing-rbacl] size =    0
                         Ingress L2 QOS [ing-l2-qos] size =  256
                 Ingress L3/VLAN QOS [ing-l3-vlan-qos] size =  512
                              Ingress SUP [ing-sup] size =  512
        Ingress L2 SPAN filter [ing-l2-span-filter] size =  256
        Ingress L3 SPAN filter [ing-l3-span-filter] size =  256
                           Ingress FSTAT [ing-fstat] size =    0
                                        span [span] size =  512
                          Egress RACL [egr-racl] size = 1792
                            Egress SUP [egr-sup] size =  256
                  Ingress Redirect [ing-redirect] size =    0
          Ingress Netflow/Analytics [ing-netflow] size =    0
                           Ingress NBM [ing-nbm] size =    0
                          TCP NAT ACL[tcp-nat] size =    0
               Egress sup control plane[egr-copp] size =    0
        Ingress Flow Redirect [ing-flow-redirect] size =    0
                         MCAST NAT ACL[mcast-nat] size =    0
```

Параметры масштабирования

ePBR Verified Scalability Limits (Unidimensional)

Feature	Supported Platforms	Verified Limits
Maximum services per switch	Nexus 9300 and 9500 switches	1503
Endpoints per service	Nexus 9300 and 9500 switches	32
ePBR policies per switch	Nexus 9300 and 9500 switches	150
Policies per VRF	Nexus 9300 and 9500 switches	16
Services per chain	Nexus 9300 and 9500 switches	6
Match per policy	Nexus 9300 and 9500 switches	16
Aces per match	Nexus 9300 and 9500 switches	256

Заключение

- DCNM предоставляет удобные средства автоматизации настроек подключения сервисных устройств
- Подключение виртуализированных сервисных устройств к EVPN-фабрике
  - Обогащение EVPN type-5 апдейтов информацией о подключении VNF
  - Просто и изящно
  - Покрывает множество сценариев
- EPBR
  - Простой и мощный механизм создания сервисных цепочек в рамках EVPN сайта

Вопросы?


CISCO