

АРХИТЕКТУРА И ДИЗАЙН ЦОД EVPN-VXLAN

Иван Лысогор

ilysogor@juniper.net

октябрь 2017

СТРАТЕГИЯ ПОСТРОЕНИЯ ЦОД

ОТКРЫТАЯ

- Несколько архитектур
- Открытые протоколы
- Установка собственного ПО

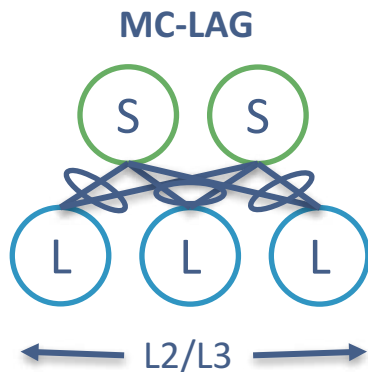
ПРОСТАЯ

- Plug & play
- Автоматическое развертывание
- Аналитика

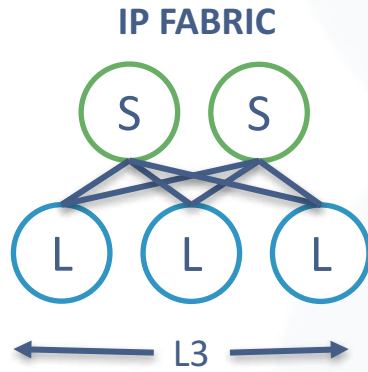
ПРОГРАММИРУЕМАЯ

- Полный набор API
- Основана на Linux
- Структурированные данные

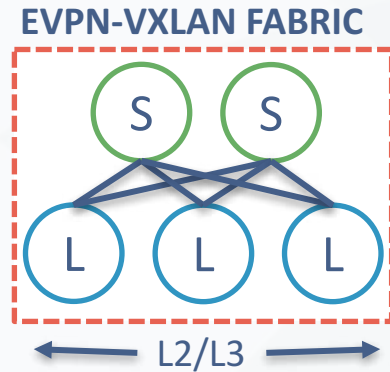
АРХИТЕКТУРА ФАБРИКИ КОММУТАЦИИ ЦОД



Traditional IT

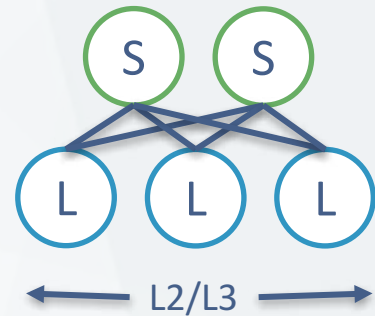


SaaS / Telco



Cloud / Hosting

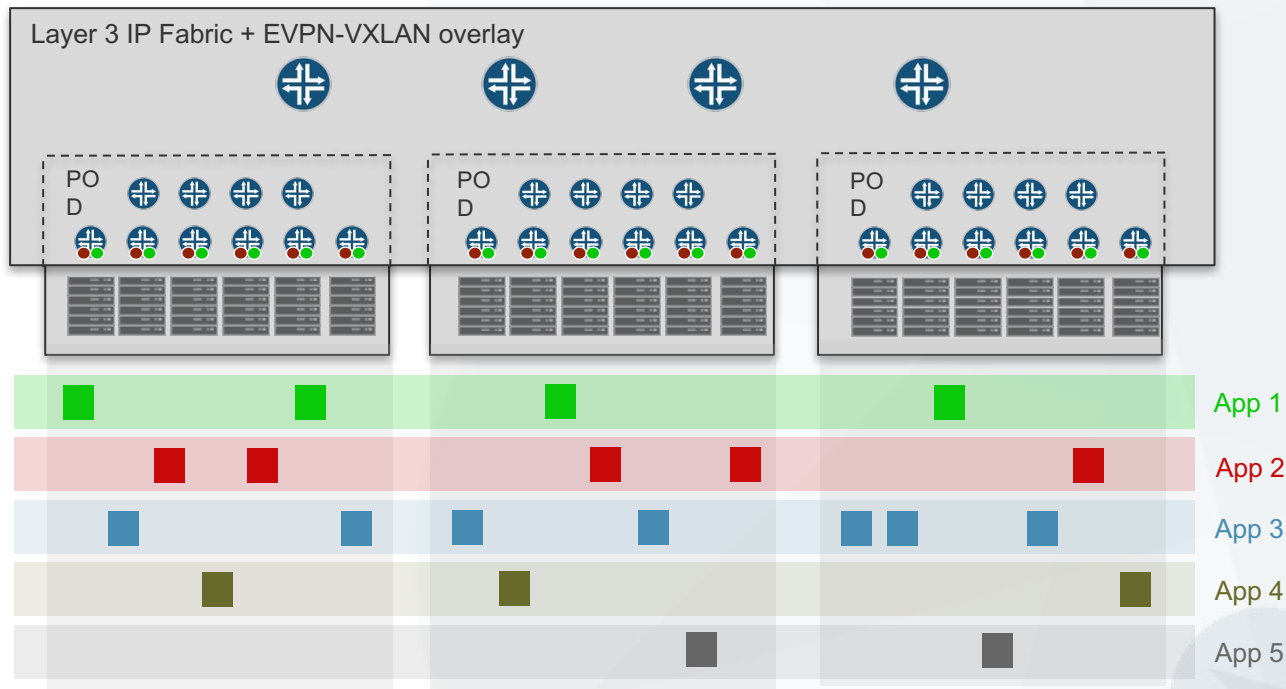
ETHERNET FABRIC
Junos Fusion или VCF



Enterprise ITaaS

ОВЕРЛЕЙ: ПРЕИМУЩЕСТВО ПРИМЕНЕНИЯ

Логическая топология не связана с физической (underlay)

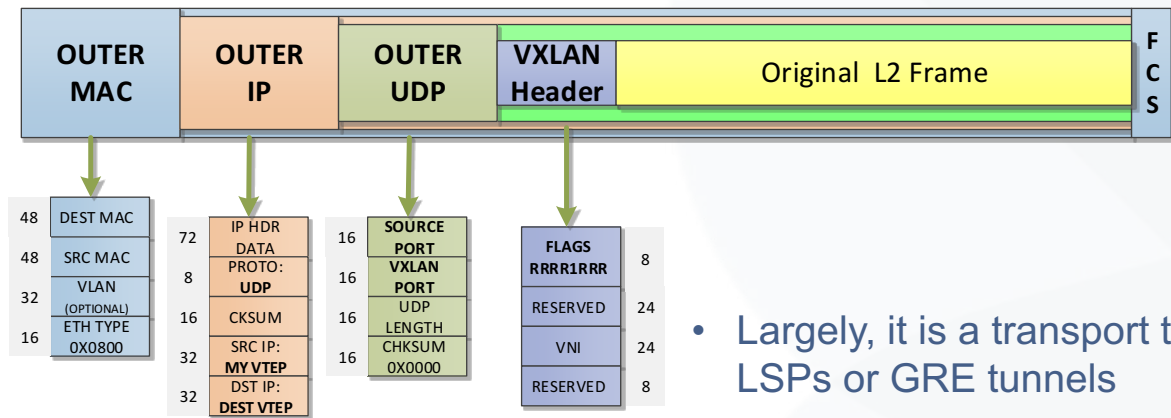




VXLAN

транспорт

VXLAN транспорт



- Largely, it is a transport tunnel alternative for MPLS LSPs or GRE tunnels
- Can be used as a generic tunnel underlay for various applications (L2VPN/L3VPN/other MPLS applications)
- Different from 'regular' VXLAN where the goal is L2 Virtualization and BUM traffic handling.
- VNID is equivalent to MPLS label in transport.
- VXLAN Tunnel Endpoints (VTEP) on Enterprises' WAN PEs

VXLAN транспорт



VXLAN in the Data Center

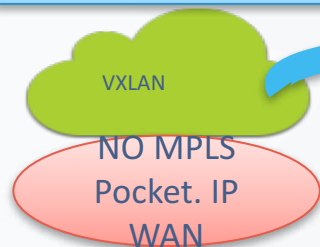
- TOR/Access/Spine layers of DCs have very limited capabilities
- MPLS-phobia inside DCs: Cost and (perceived) complexity
- Flat IP fabric based DCs
- Originated in DCs but gaining momentum in SD WAN

VXLAN in the WAN

- Over The Top (OTT) L2VPN and L3VPN connectivity using IP transport
- VXLAN extends L2/L3 reach over IP WAN
- Cost optimization
- Better than GRE; Entropy friendly

Extending VPNs over vanilla IP

OTT

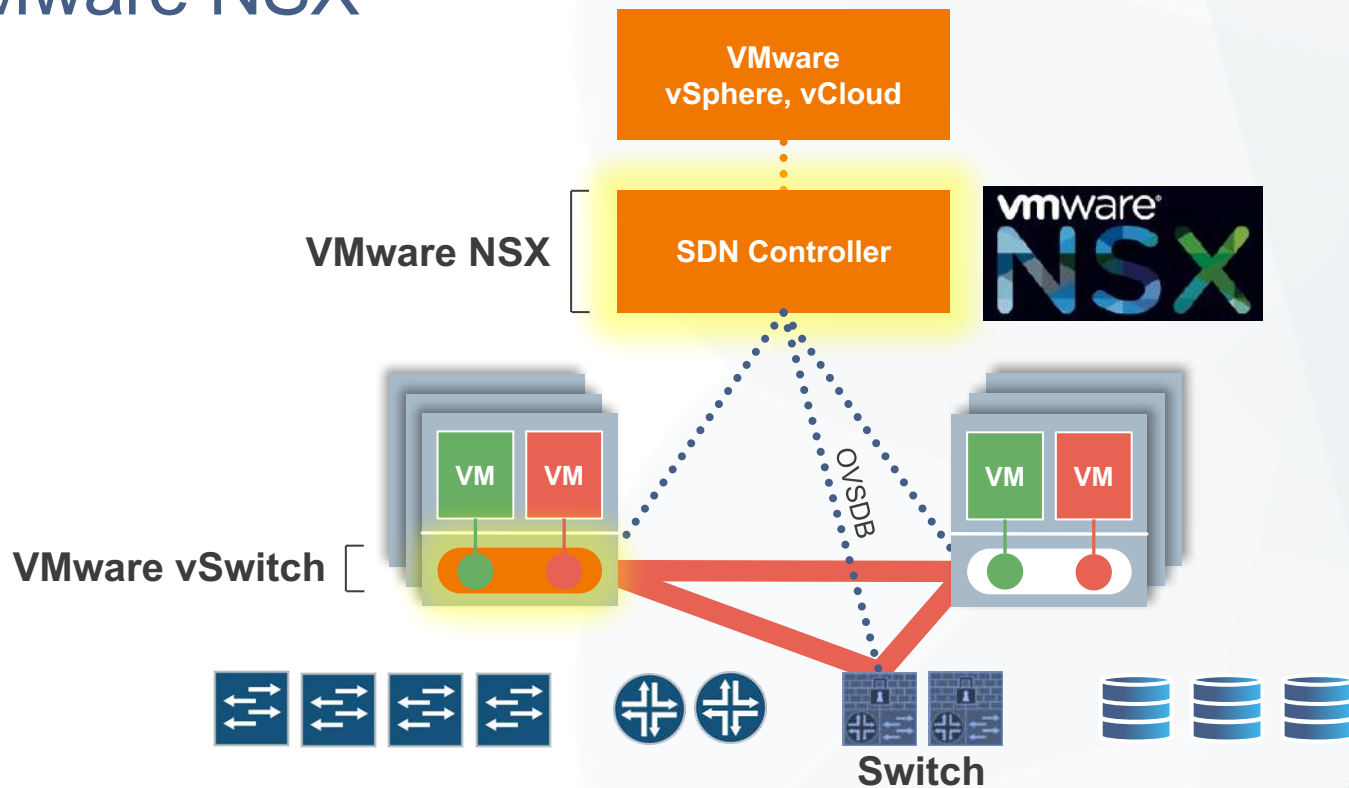


WAN
MPLS

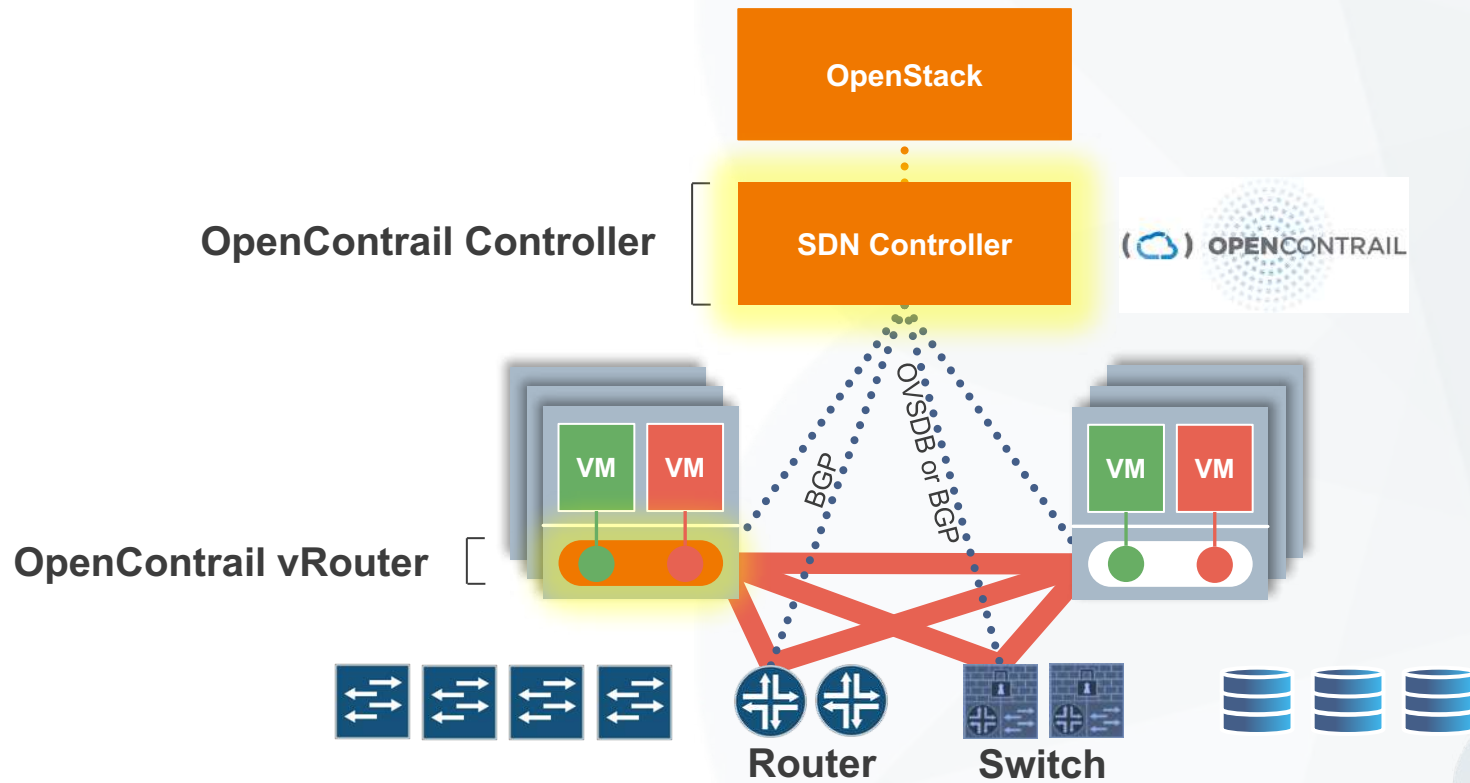
РЕАЛИЗАЦИИ VXLAN

- Multicast VXLAN
Dataplane learning
- Unicast VXLAN
Static Control plane (точки терминации заданы статически)
- OVSDB/VXLAN
Controller: Contrail/VMware NSX
- EVPN-VXLAN
BGP based Control plane

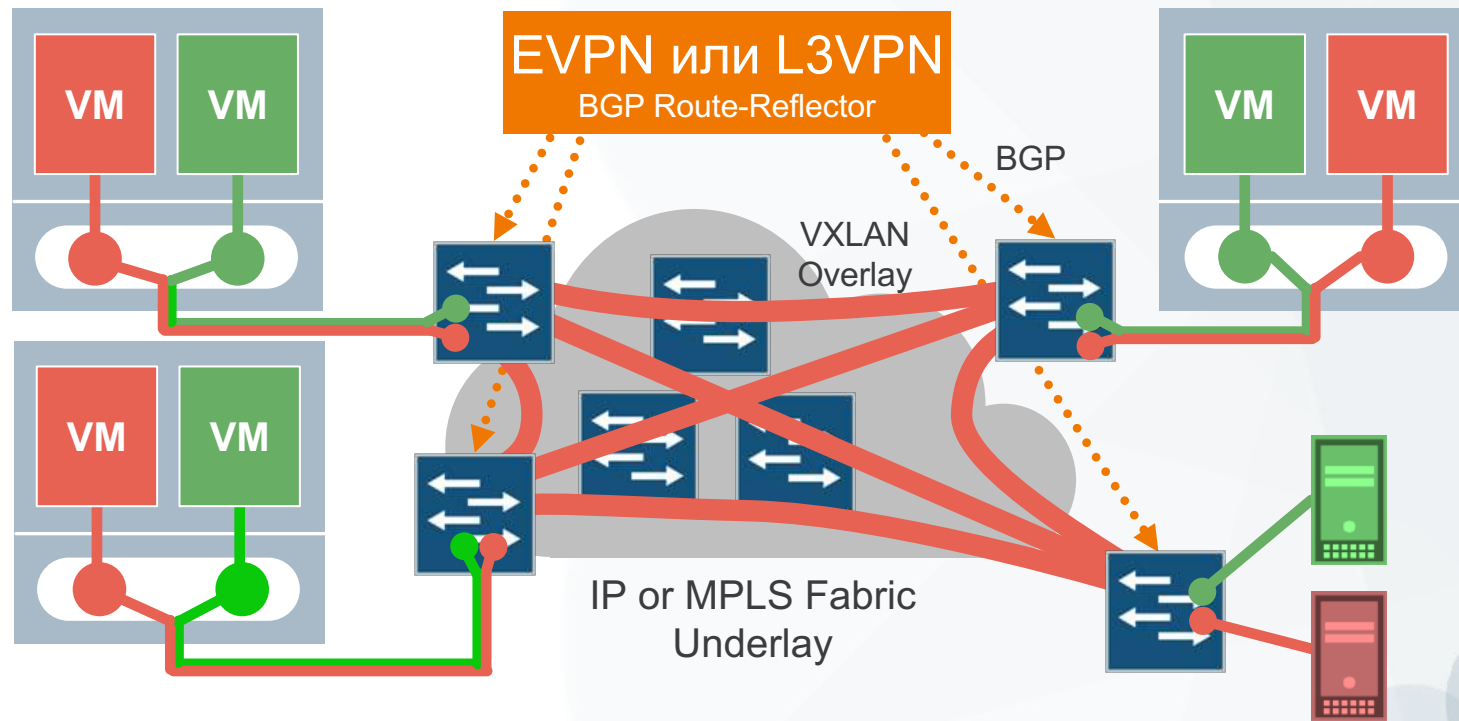
VMware NSX



OpenContrail + Openstack



EVPN: Network only





EVPN

технология

EVPN преимущества

Эффективность

Использование всех каналов для передачи трафика с встроенной защитой от L2 петель

Конвергенция

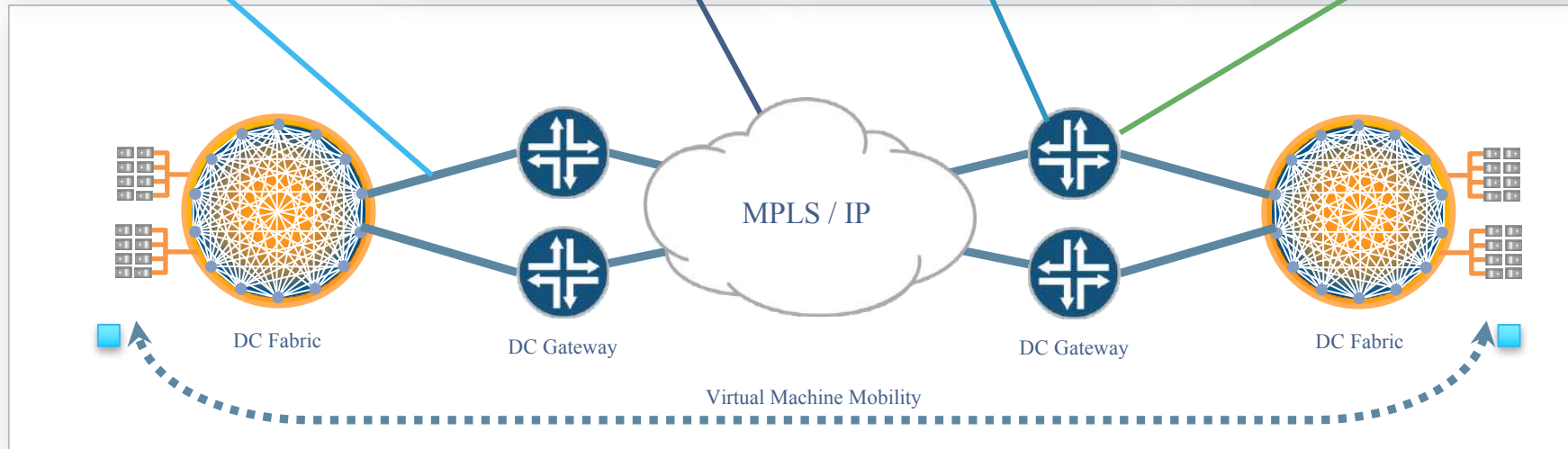
Высокий уровень доступности, быстрая сходимость и перемаршрутизация

L3 и L2

Интеграция L2 и L3 уровней в протокол управления

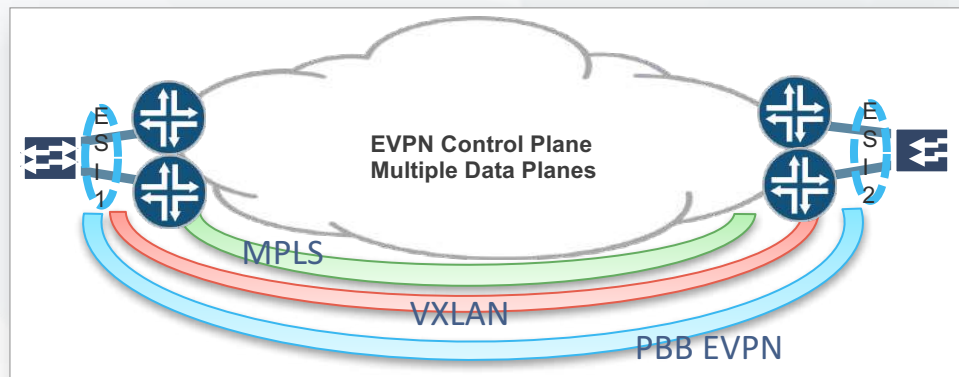
Оптимизация

Оптимизация входящего и исходящего трафика при миграции виртуальных машин



EVPN : Варианты использования

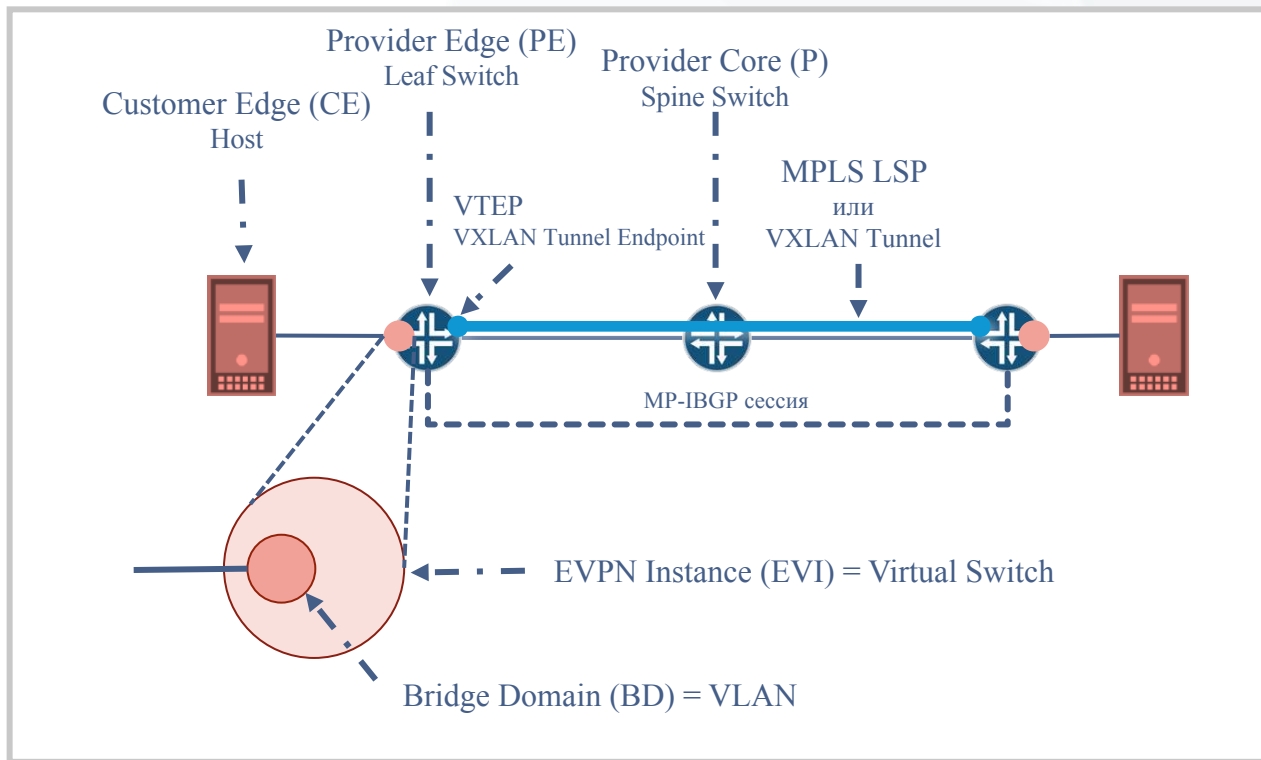
MPLS Transport (WAN)	L2 Service	L3 Service
	MPLS service label advertised for MACs	MPLS service label advertised for subnets
	ELINE, ELAN, ETREE. MEF 2.0 compliant BE/Metro Services	L2/L3 Service Integration



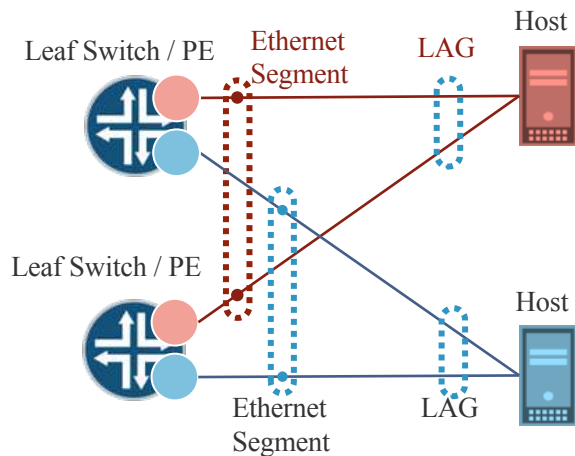
IP Transport (DC, OTT)	L2 Service	L3 Service
	VNIDs advertised for MACs	VNID advertised for subnets
	L2 stretch between VMs over IP OTT L2VPN	Inter-subnet traffic Integration in DC, OTT L3VPN with EVPN

PBB –EVPN / MPLS Transport (WAN)	L2 Service	L3 Service
	MPLS service label advertised for B-MACs C-MACs learned in data plane ELAN, ETREE	No support

Терминология EVPN



Ethernet Segment



Ethernet Segment (ES)

- Набор каналов к одному сайту (хосту)
- Каналы могут быть как Single-Active или All-Active

Ethernet Segment Identifier (ESI)

- 10 байт (например: 00:11:22:33:44:55:66:77:88:99)
- Нулевое значение означает только одно подключение
- Может быть сконфигурирован вручную
- Может быть назначен автоматически

Типы маршрутов NLRI для EVPN

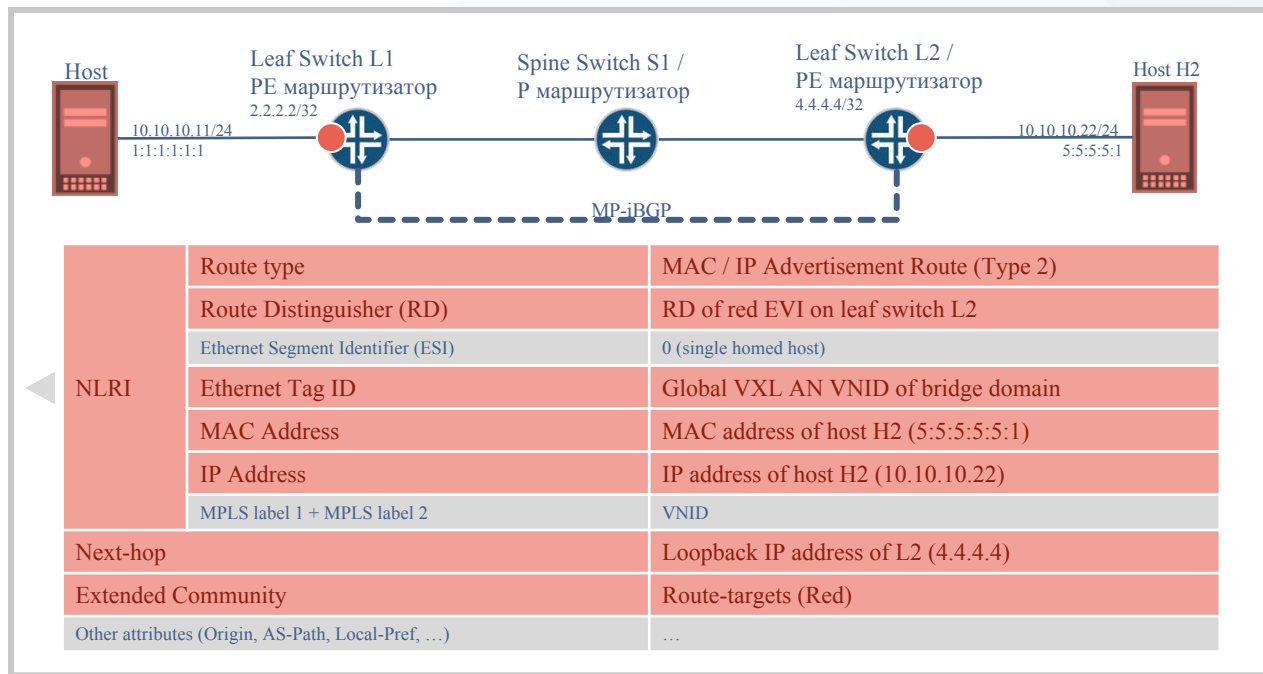
Route Type	Description	Usage	Standard
1	Ethernet Auto-Discovery	PE Discovery and Mass Withdraw	RFC 7432
2	MAC Advertisement	MAC Advertisement	RFC 7432
3	Multicast Route	BUM Flooding	RFC 7432
4	Ethernet Segment Route	ES Discovery and DF Election	RFC 7432
5	IP Prefix Route	IP Route Advertisement	draft-rabadan-l2vpn-evpn-prefix-advertisement

Локальный процесс mac learning



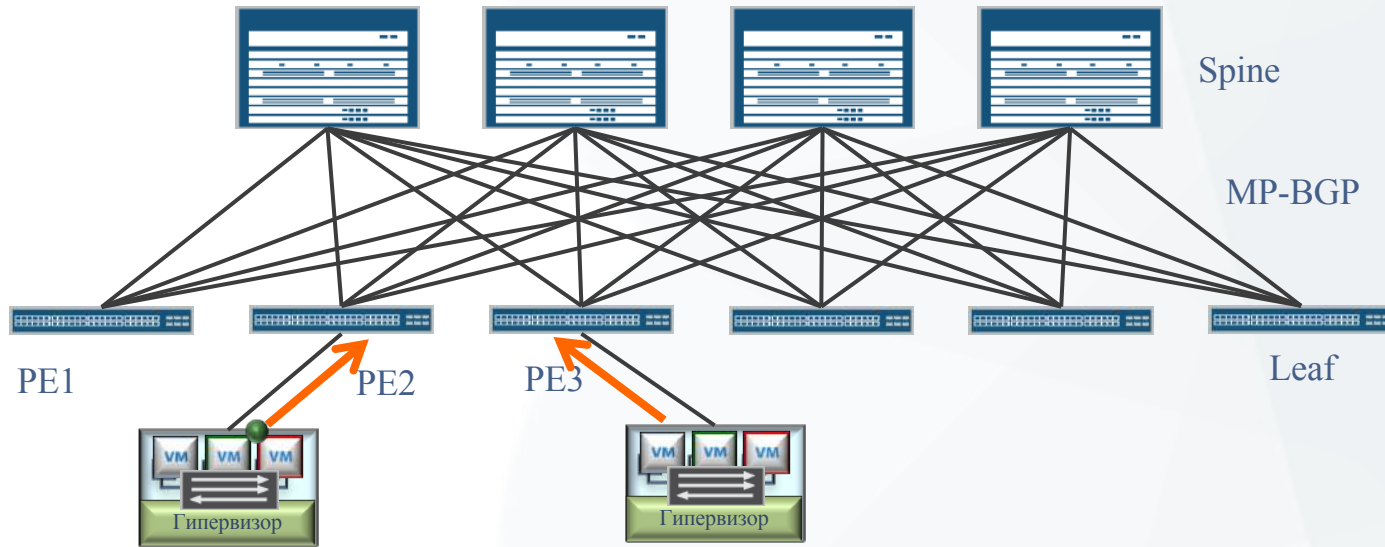
Удаленный процесс mac learning

- MAC / IP advertisement (type 2) routes



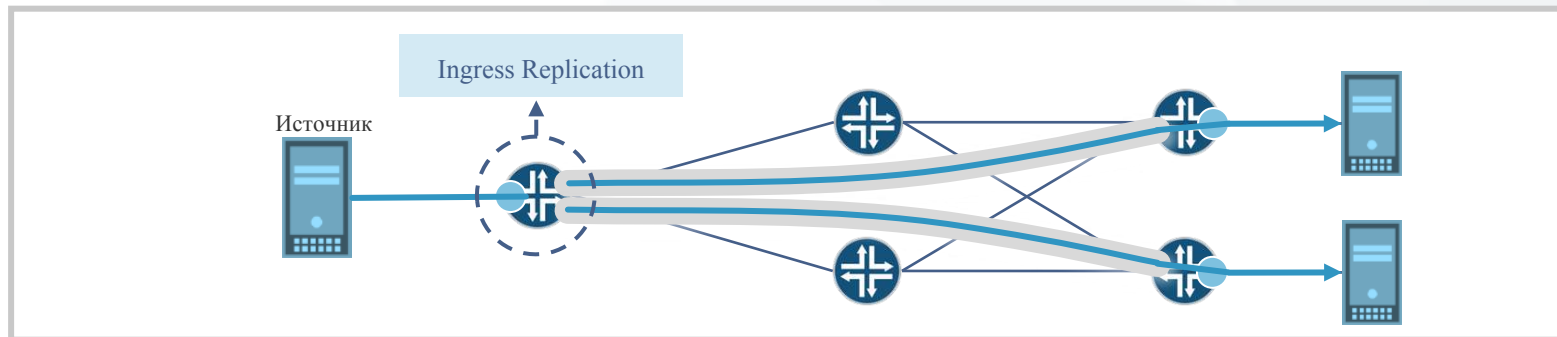
VM Mobility

Миграция mac-адресов виртуальных машин



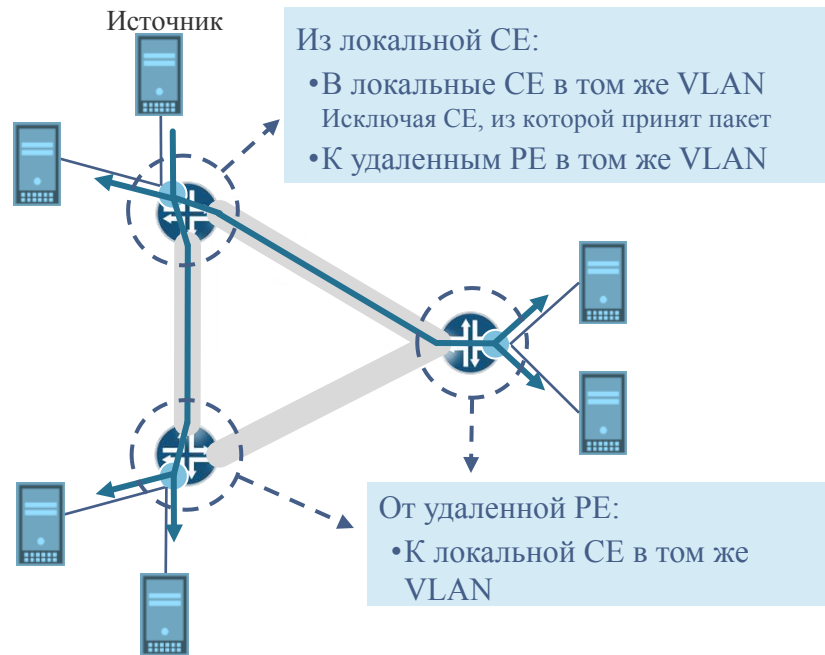
- MAC, анонсируемый PE2, перемещается за PE3.
- PE3 анонсирует этот MAC с новым порядковым номером в ext. Community (seq#)
- Маршрут MAC обновляется на всех PE, при этом PE2 старую запись удаляет

Inclusive multicast (type 3) routes



NLRI	Route type	Inclusive Multicast Ethernet Tag Route (Type 3)	
	Route Distinguisher (RD)	...	
	Ethernet Tag ID	0	
	Originator IP Address	...	
Provider Multicast Service Interface (PMSI) Tunnel		Flags	0 (No leaf information required)
		Tunnel Type	Ingress replication, PIM-SSM, PIM-SM, BIDIR-PIM, ...
		MPLS Label	0 (Not used)
		Tunnel Identifier	Multicast Group IP Address Sender IP Address
Extended Communities		Route-Targets	
Other attributes (Origin, Next-Hop, AS-Path, Local-Pref, ...)		...	

Split Horizon



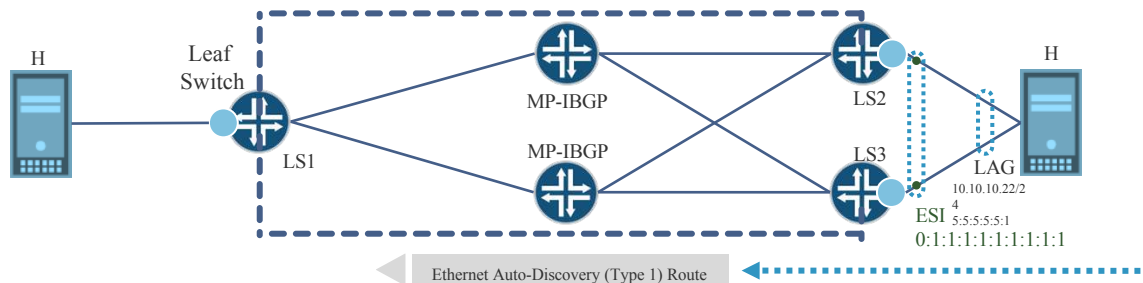
Multi-home All-active



- Устройство доступа подключается стандартной **агрегацией каналов (LAG)**
- Применяется выделенное устройство коммутации для **устранения дублирования BUM** трафика
- Все устройства коммутации используются в **активном режиме без применения MC-LAG**

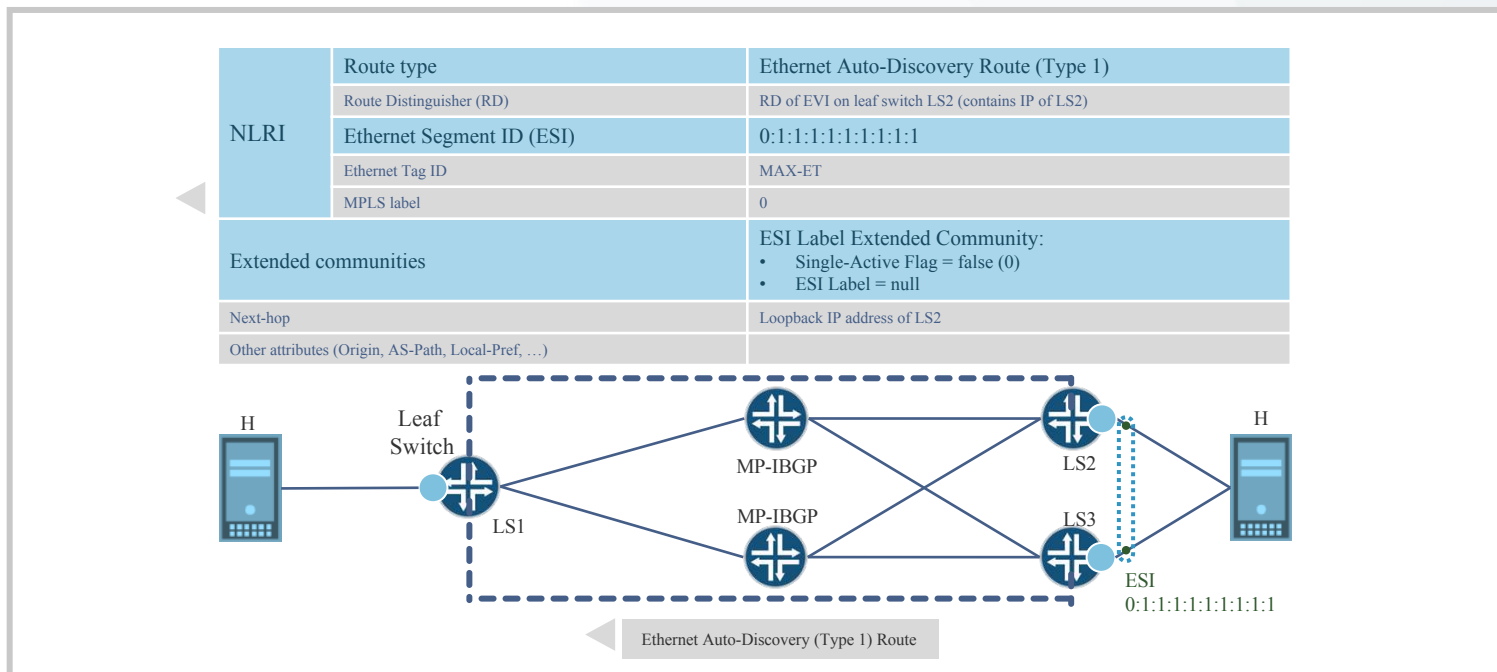
MAC / IP Advertisement (Type 2) Routes

NLRI	Route type	MAC / IP Advertisement Route (Type 2)
	Route Distinguisher (RD)	...
	Ethernet Segment Identifier (ESI)	0:1:1:1:1:1:1:1:1
	Ethernet Tag ID	VNID
	MAC Address	MAC address of host H2 (5:5:5:5:5:1)
	IP Address	IP address of host H2 (10.10.10.22)
	MPLS label 1 + MPLS label 2	VNID
Next-hop		Loopback of LS2
Other attributes (Origin, AS-Path, Local-Pref, ...)		...



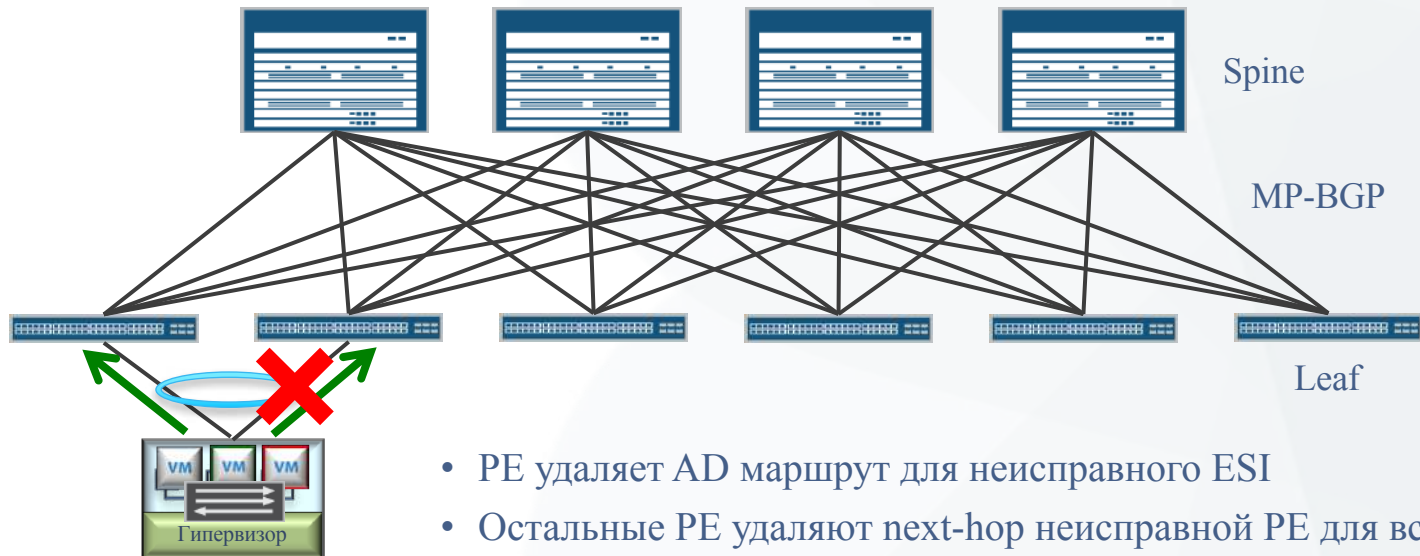
Ethernet Auto-discovery (Type 1) Routes

- для Ethernet segment (ES): многоканальность и быстрая сходимость



Быстрая сходимость

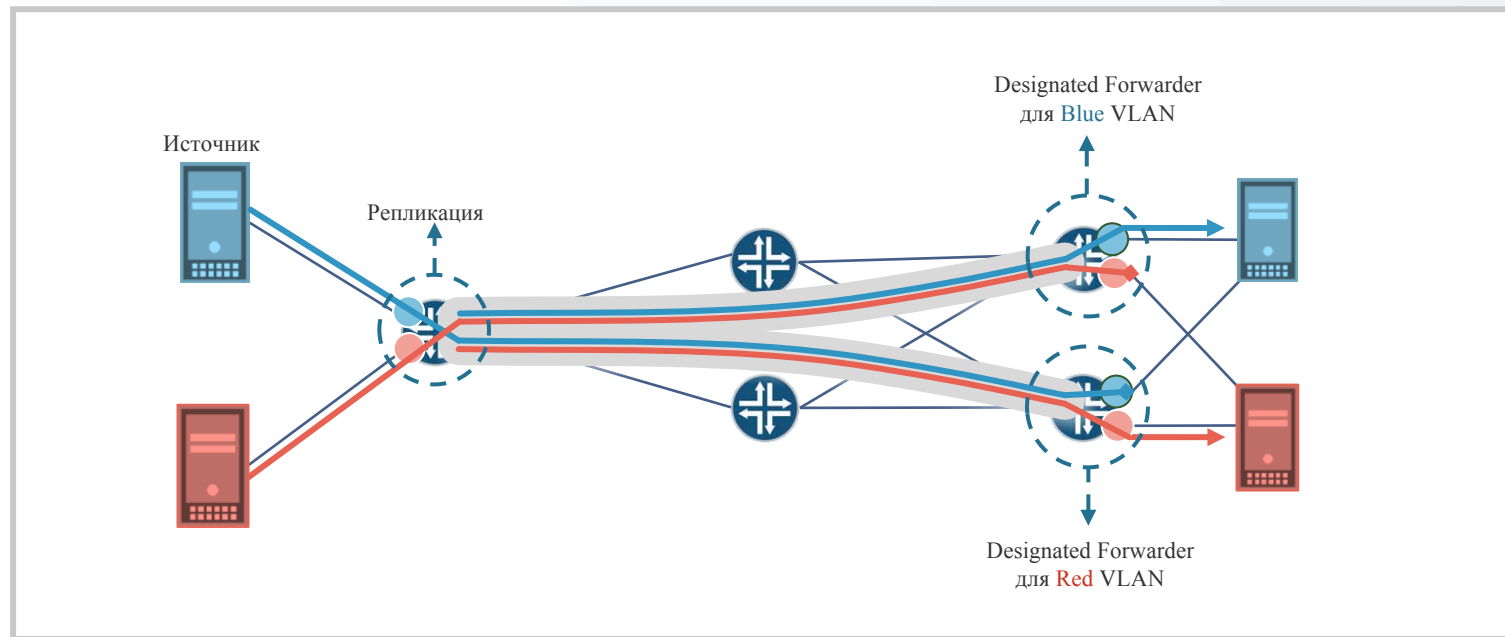
MAC Mass Withdrawal



- PE удаляет AD маршрут для неисправного ESI
- Остальные PE удаляют next-hop неисправной PE для всех MAC/IP адресов, полученных для данного ESI

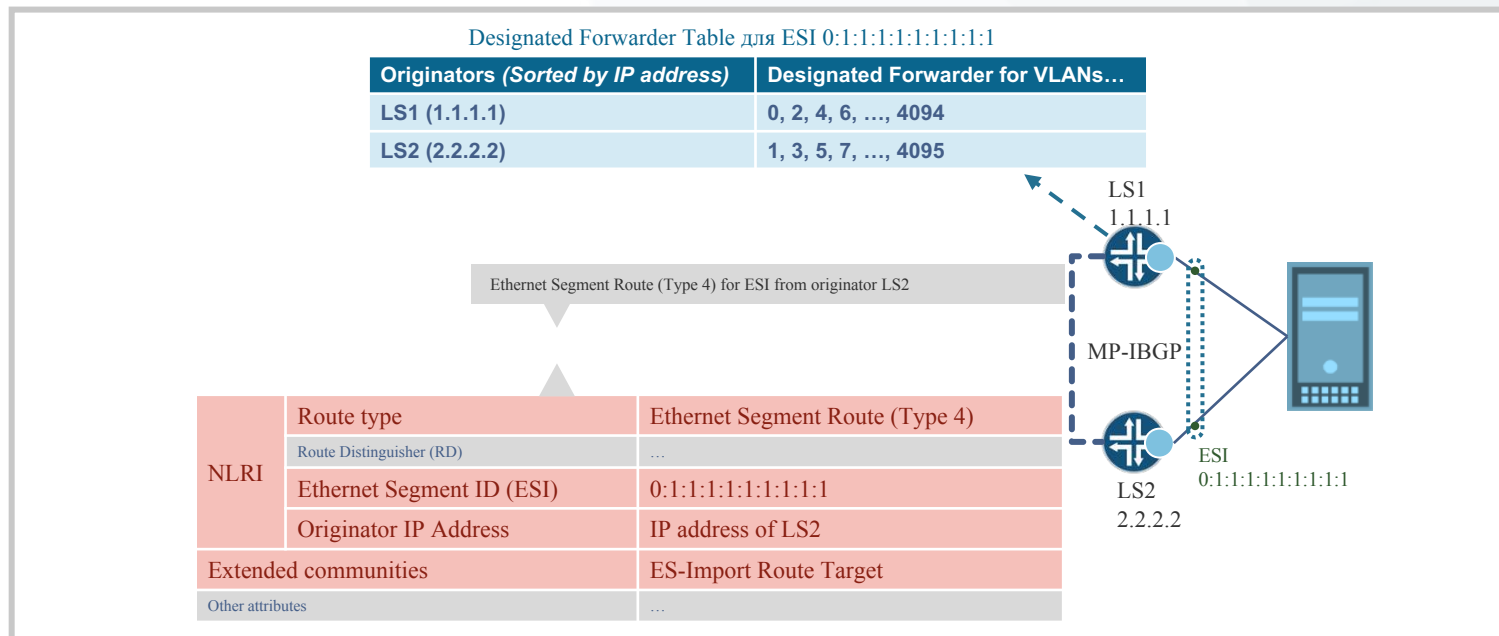
Designated forwarder

- Для передачи broadcast, unknown unicast, multicast (BUM) трафика

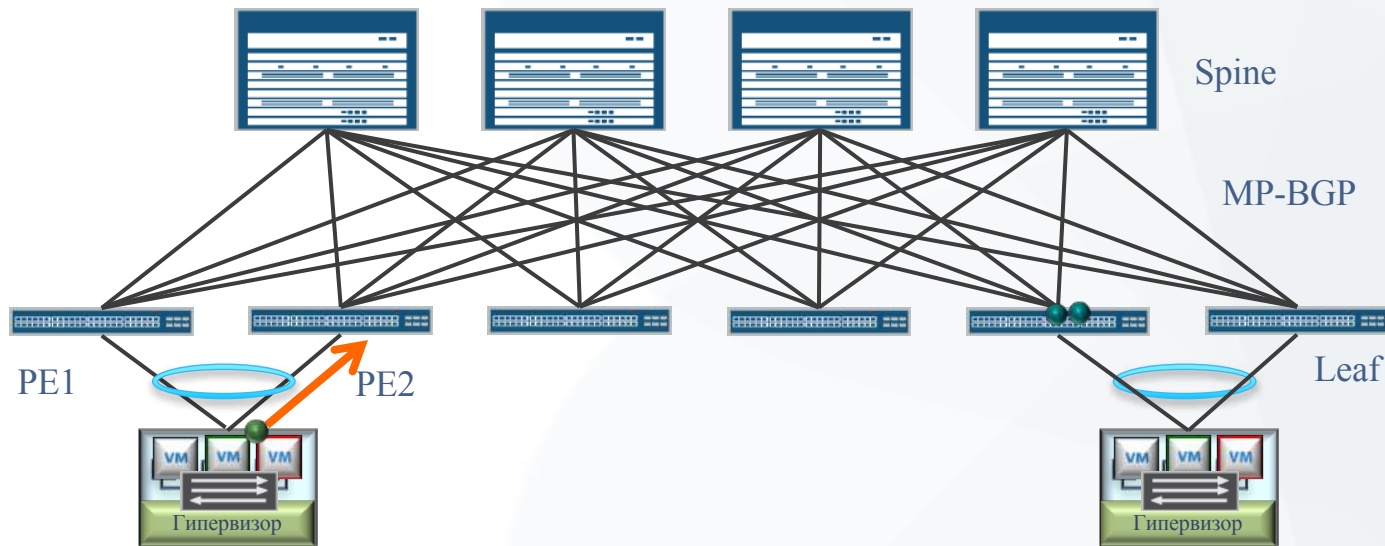


Ethernet Segment (Type 4) Routes

- Необходим для выбора designated forwarder



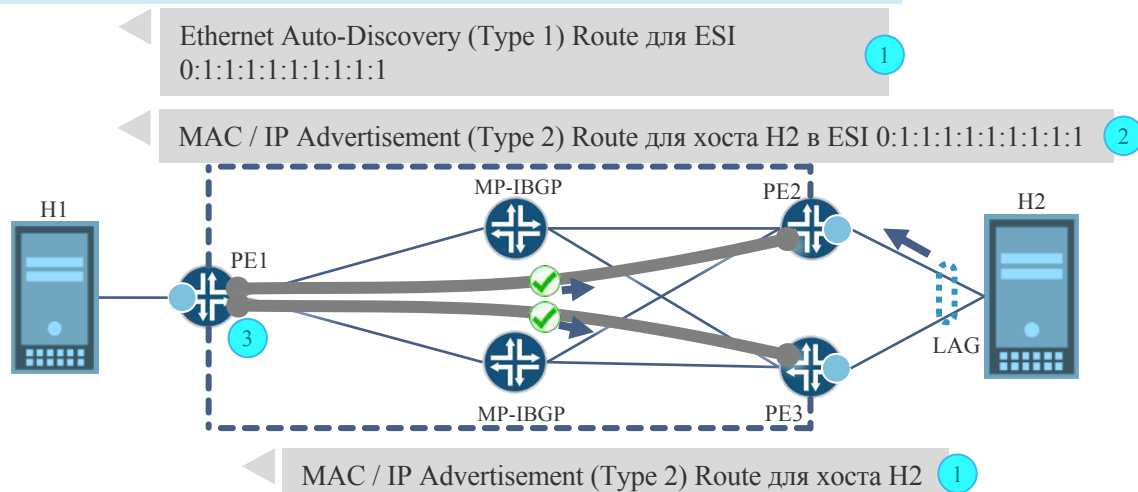
Балансировка нагрузки



- Анонсируемый MAC адрес с PE1 может быть достижим через PE1 и PE2 в одном сегменте ESI
- Удаленные PE устройства могут балансировать трафик между PE, которые анонсируют одинаковый идентификатор ESI

Aliasing

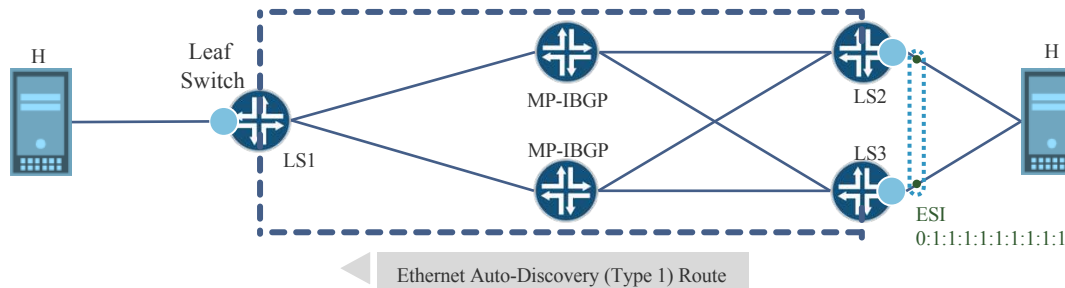
1. PE1 получает Ethernet Auto-Discovery (type 1) маршрут от PE2 и PE3
2. PE1 получает MAC/IP Advertisement (type 2) маршрут только от PE2
 - PE1 проинформирован, в каком ESI находится хост H2
 - PE1 проинформирован, что этот ESI доступен через PE2 и PE3
3. PE1 может балансировать трафик (ECMP) к хосту H2 через два LSP к PE2 и PE3



Ethernet Auto-discovery (Type 1) Routes

- для EVPN Instance (EVI): балансировка нагрузки – aliasing

NLRI	Route type	Ethernet Auto-Discovery Route (Type 1)
	Route Distinguisher (RD)	–
	Ethernet Segment ID (ESI)	0:1:1:1:1:1:1:1
	Ethernet Tag ID	VNID
	MPLS label	0
Next-hop		Loopback IP address of LS2
Other attributes (Origin, AS-Path, Local-Pref, ...)		





Layer 3

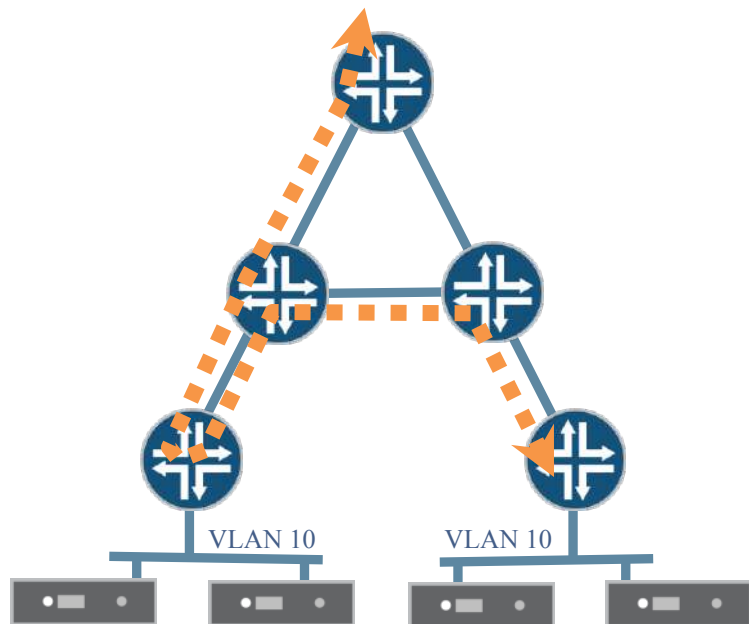
маршрутизация

EVPN: функциональность L3 шлюза

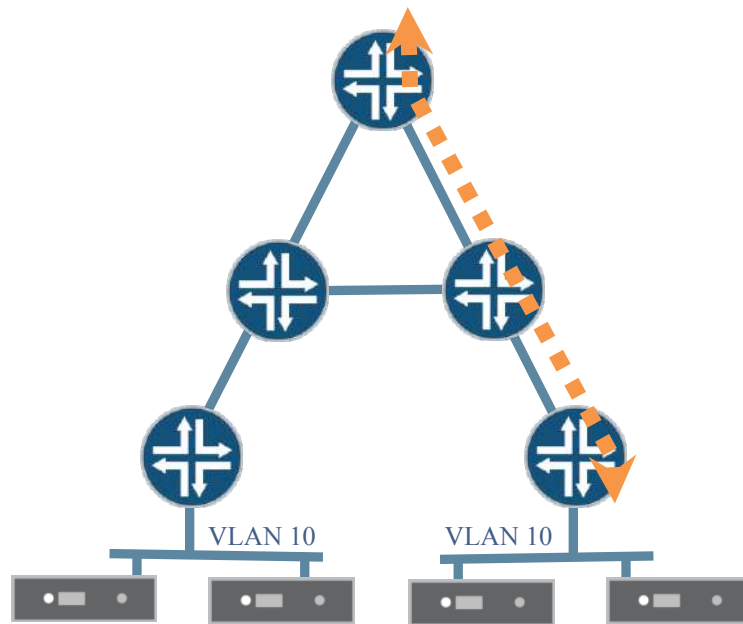
- Установка IRB интерфейса для каждого bridge domain необходима для:
 - обеспечения L3 связности
 - оптимальной передачи трафика между VLAN внутри и между дата-центрами
- Синхронизация MAC/IP хоста
 - маршрутизатор изучает ARP сообщения для составления таблицы соответствия MAC и IP хоста
 - анонсирует IP адрес внутри MAC маршрута
- Синхронизация MAC/IP шлюза по-умолчанию
 - IRB IP адрес конфигурируется как «default gateway» для EVI
 - EVPN PE анонсируют все локальные MAC и IP для IRB интерфейсов, промаркированные default gw community (настраиваемый режим)
 - Для одного VLAN может быть несколько шлюзов по-умолчанию с одинаковым MAC и IP адресом anycast
 - PE маршрутизируют пакеты, которые были направлены на MAC шлюза
 - PE обеспечивают Proxy ARP для IP шлюза и направляют ответ с полученным MAC адресом

Оптимизация трафика Inter VLAN

Передача трафика без оптимизации

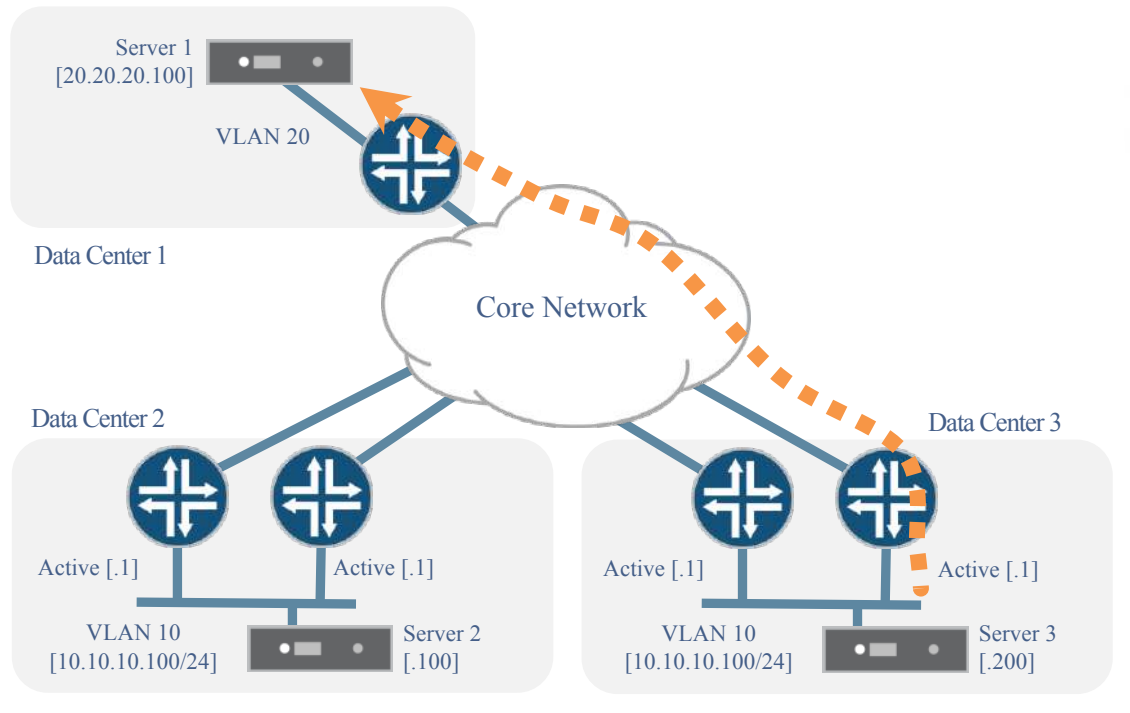


Эффективная маршрутизация



Оптимизация исходящего трафика

Задача: обеспечить оптимальный путь от сервера 3 к серверу 1



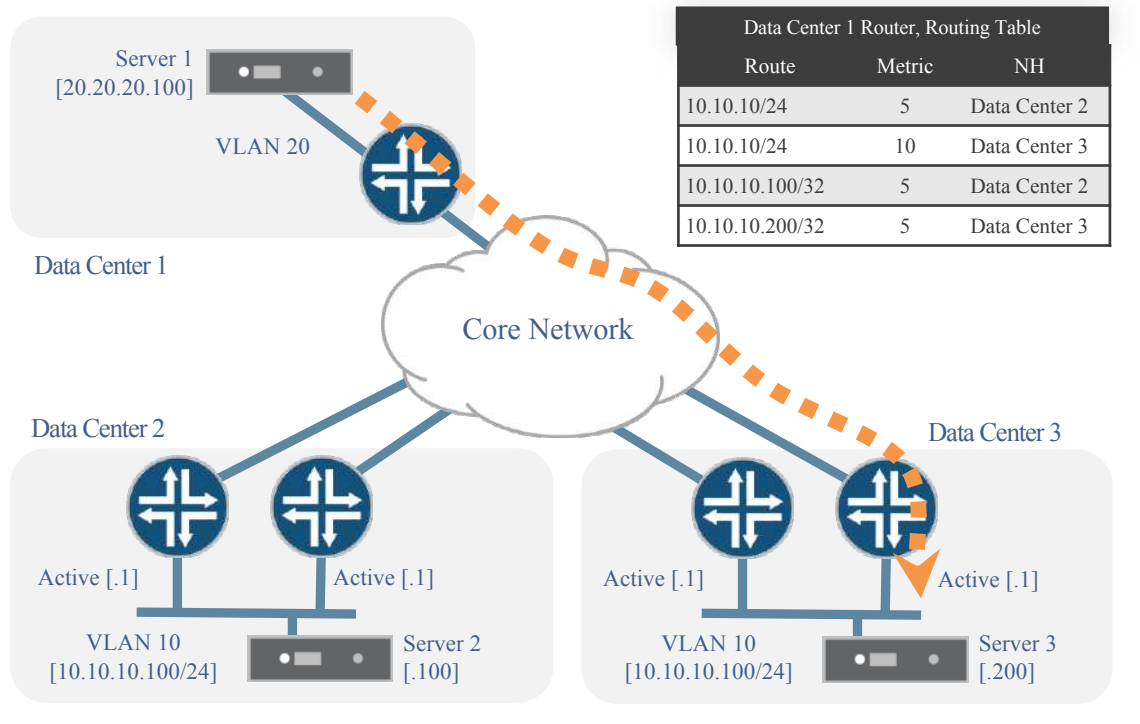
Решение

Виртуализация и распространение Default gateway на все маршрутизаторы, обслуживающие VLAN трафик

Исходящий трафик может быть направлен для любой маршрутизатор для VLAN 10. При этом маршрутизация всегда будет в локальном дата-центре.

Оптимизация входящего трафика

Задача: обеспечить оптимальный путь от сервера 1 к серверу 3

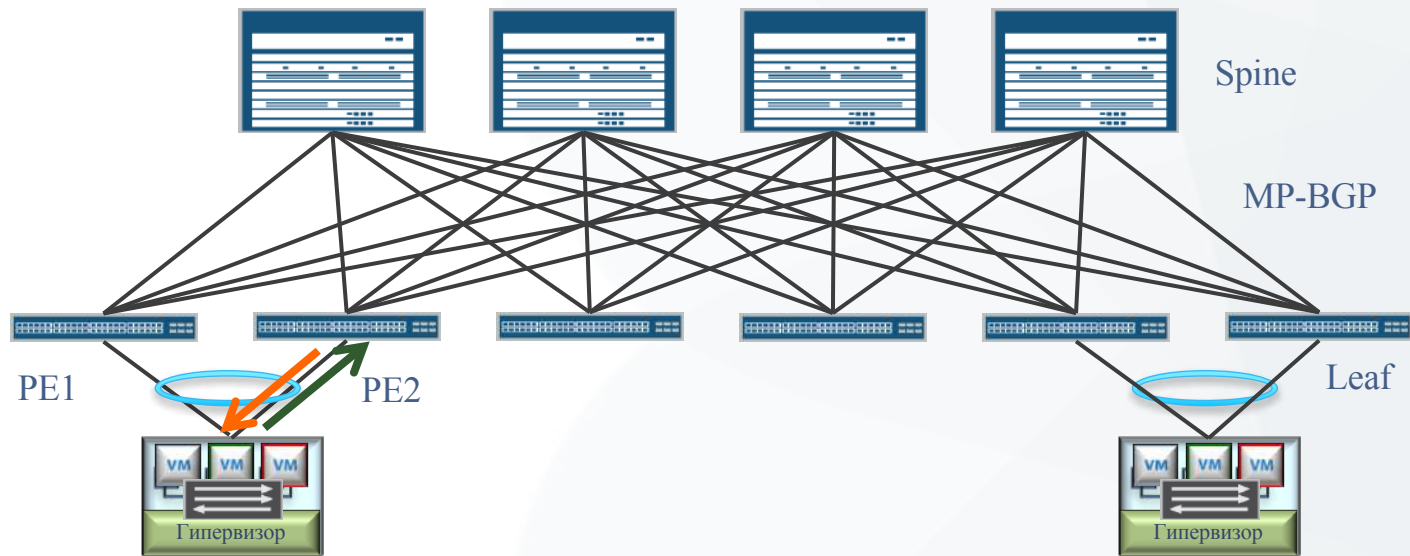


Решение

В дополнении к маршруту сети 10.10.10.0/24 маршрутизатор отправителя обладает специфичным маршрутом /32 до конечного хоста.

Входящий трафик к серверу 3 будет направлен напрямую через WAN от DC 1 к DC3.

Proxy ARP



- PE2 обладает ARP записями MAC/IP.
- PE2 детектирует ARP запрос для шлюза по-умолчанию и напрямую формирует ARP ответ для своих пиров.



EVPN Route Type 5

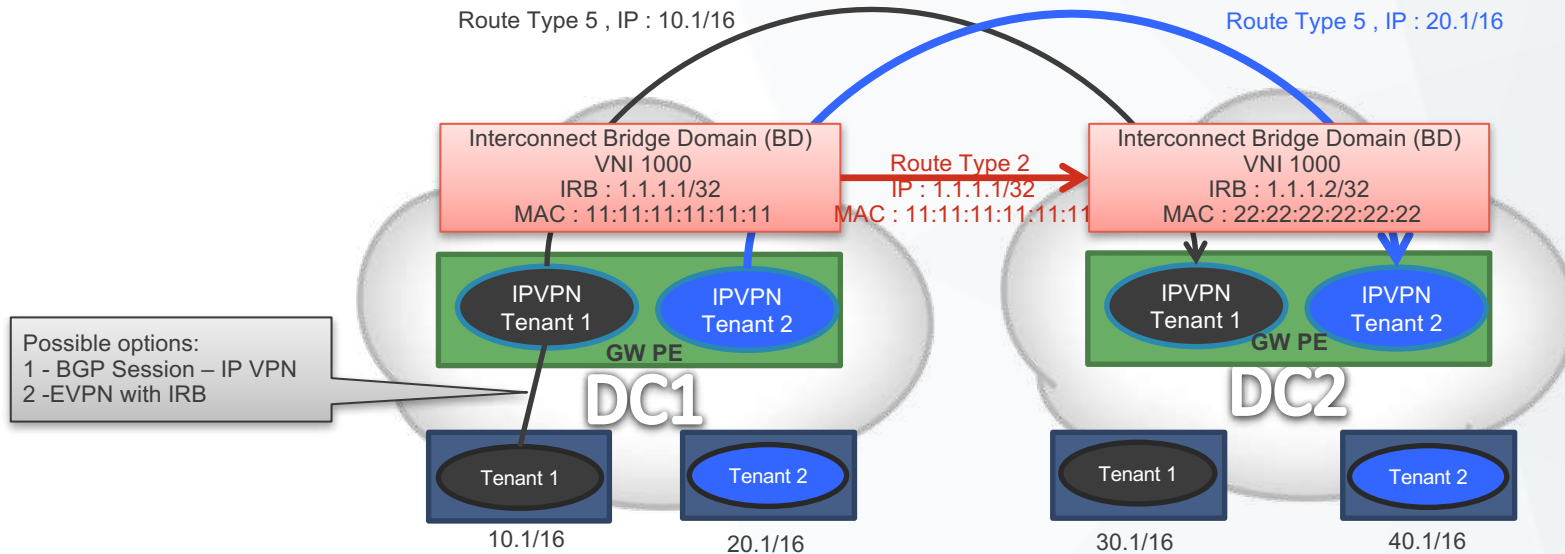
DCI with Route Type 5

- **When to use Route Type 5 for DCI**
 - No L2 stretch required between DCs
 - MACs belonging to a DC customer can be summarized by an IP prefix
 - End to end unified EVPN Solution. Use EVPN for the DC as well as DCI

– IEFT Information

- **draft-rabadan-l2vpn-evpn-prefix-advertisement**
- We implement Section 5.4 of the above draft and are fully compliant with it.
- RLI 25985 (Planned for 16.2R1)

Route Type 5 : Gateway Address Model



Route Type 5 : Supportability Matrix

Platform	Chip Set	VXLAN Routing Support	General Route Type 5 support
QFX5100	Trident 2	Hardware limitation	No
QFX5110	Trident 2+	Yes	17.3R1
QFX5200	Tomahawk	Hardware limitation	No
QFX10002	Q5	Yes	15.1X53-D30
QFX10008/16	Q5	Yes	16.2
MX Series	Trio	Yes	17.1R1



DESIGN GUIDES

- Software as a Service

https://www.juniper.net/techpubs/en_US/release-independent/solutions/information-products/pathway-pages/sg-001-software-as-a-service.pdf

- Infrastructure as a Service: Contrail and OVSDDB

http://www.juniper.net/techpubs/en_US/release-independent/solutions/information-products/pathway-pages/sg-002-infrastructure-as-a-service.pdf

- Infrastructure as a Service: EVPN and VXLAN

http://www.juniper.net/techpubs/en_US/release-independent/solutions/information-products/pathway-pages/sg-003-evpn-vxlan.pdf

- Improve Data Center Interconnect, L2 Services with Juniper's EVPN

<https://www.juniper.net/assets/us/en/local/pdf/whitepapers/2000596-en.pdf>

- Juniper Networks EVPN Implementation for Next-Generation Data Center Architectures

<https://www.juniper.net/assets/us/en/local/pdf/whitepapers/2000606-en.pdf>

Спасибо!