Наложенная сетевая инфраструктура в центрах обработки данных: использование технологии EVPN и SDN-контроллера Contrail

Ivan Lysogor
ilysogor@juniper.net

Развитие наложенной сетевой инфраструктуры

IETF Network Virtualization Overlay Working Group

Требования:

- IP-based underlay
- Logically centralized authority for network virtualization
- Multi-tenancy
- Endpoint workload mobility
- Cloud speed and agility

Решаемые задачи:

- Network virtualization with massive scale, performance, HA
- Operational simplicity through orchestrated programmability
- Multi-vendor flexibility avoiding lock-in
- Solid networking for IaaS/PaaS/virtual compute systems like:
  NFV-I, OpenStack, Docker, Kubernetes, CloudFoundry...

3 решения от Juniper

Ethernet VPN

Аппаратная поддержка EVPN

VMware NSX

Партнерство с VMware

Contrail Networking

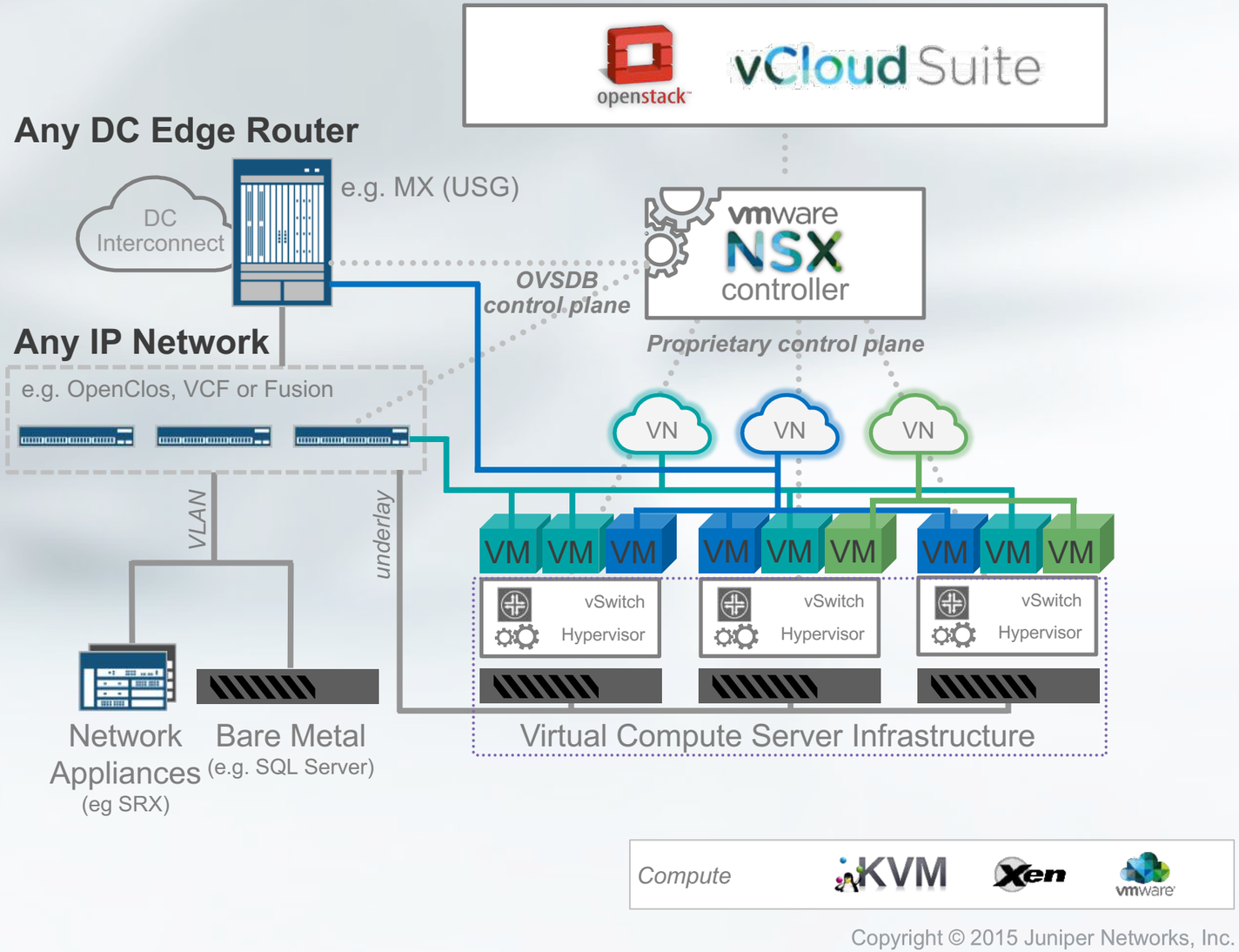
Сетевая виртуализация

Обзор VMware NSX

Платформа сетевой виртуализации NSX

Решение SDN для платформ виртуализации OpenStack and vCloud Suite

- Сетевая виртуализация на базе VXLAN
- Отказоустойчивый кластер контроллеров масштабируется до 5000 виртуальных серверов
- Поддержка OVSDB для интеграции с аппаратными сетевым устройствами
- *Технологический альянс Juniper VMWare:*
  - VXLAN маршрутизация
  - Аппаратный шлюз VXLAN-VLAN
  - Единая платформа управления

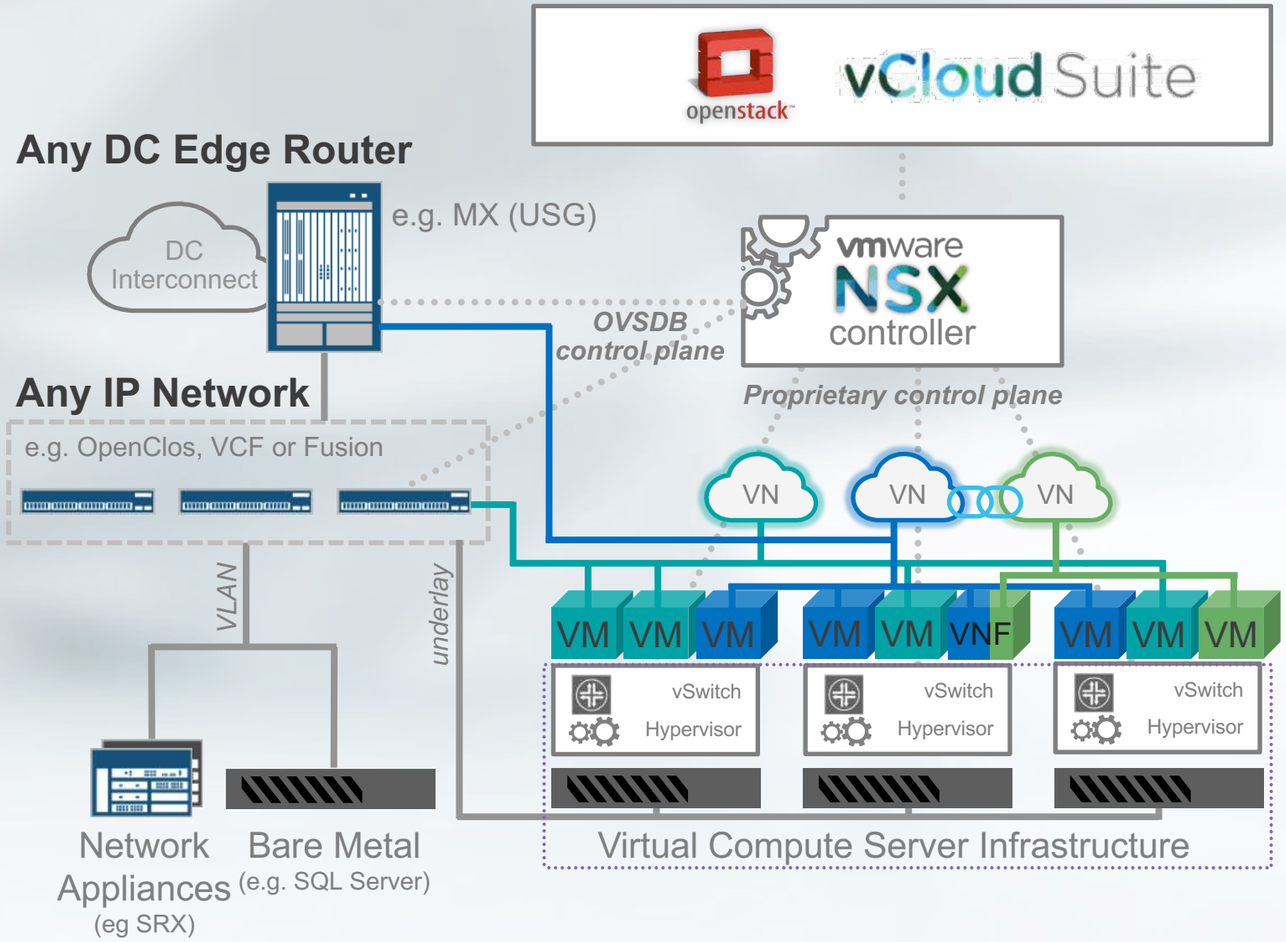


Обзор VMware NSX

ПРЕИМУЩЕСТВА

РЕШЕНИЯ

- **В качестве транспортной инфраструктуры может использоваться любая IP сеть**
- **Естественный путь развития виртуальной платформы, построенной на базе VMware**
- **Большая экосистема партнеров**

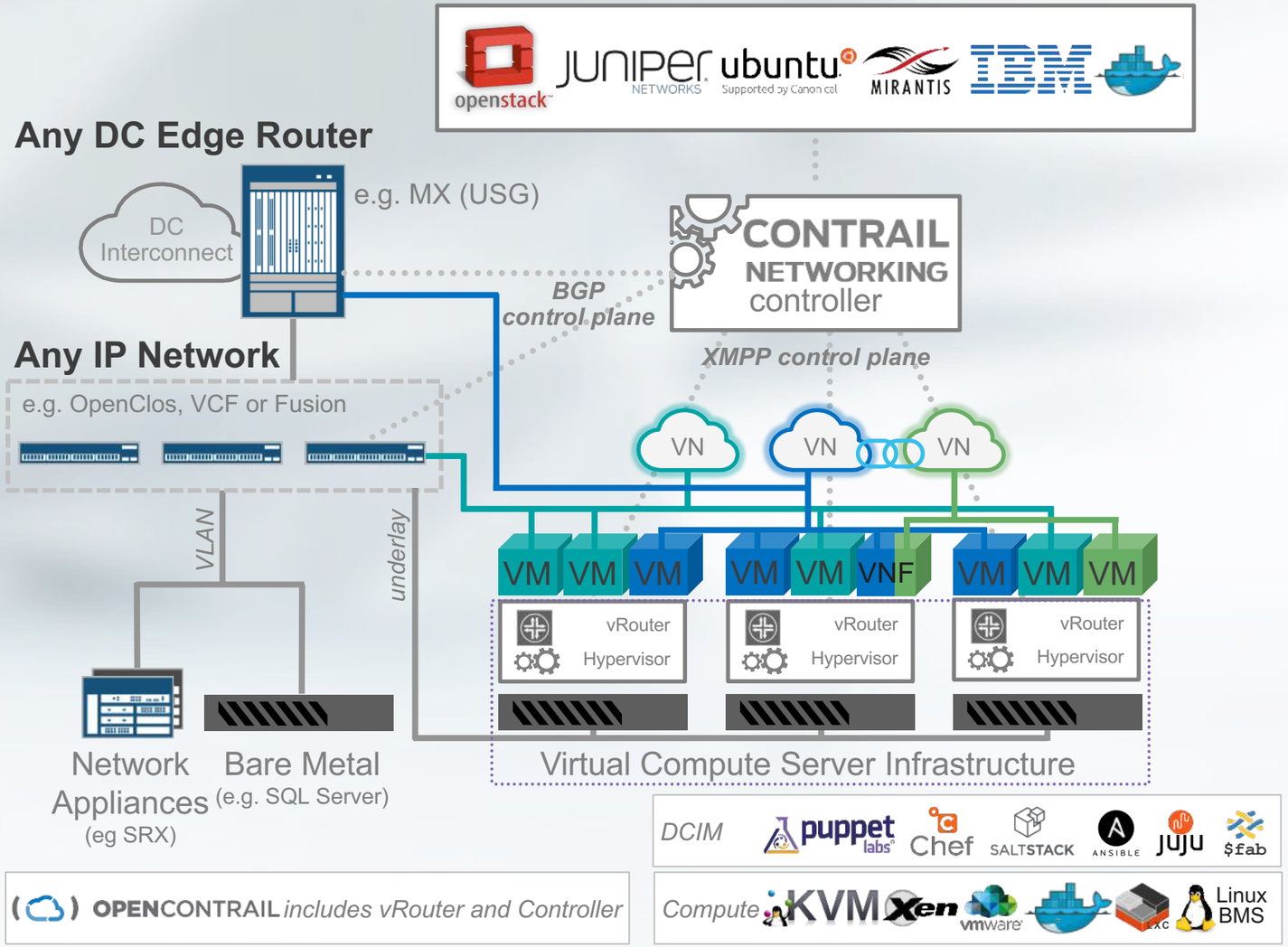


Обзор Contrail

Сетевая виртуализация Contrail

Сетевая виртуализация для OpenStack

- Масштабируемая отказоустойчивая платформа для построения облачной среды с виртуализацией сетевых сервисов
- Решение с открытым исходным кодом
- Поддерживает полный набор сетевых сервисов
- Интеграция с сетевыми сервисами сторонних производителей
- Аналитика и визуализация наложенной и транспортной сетевой инфраструктуры

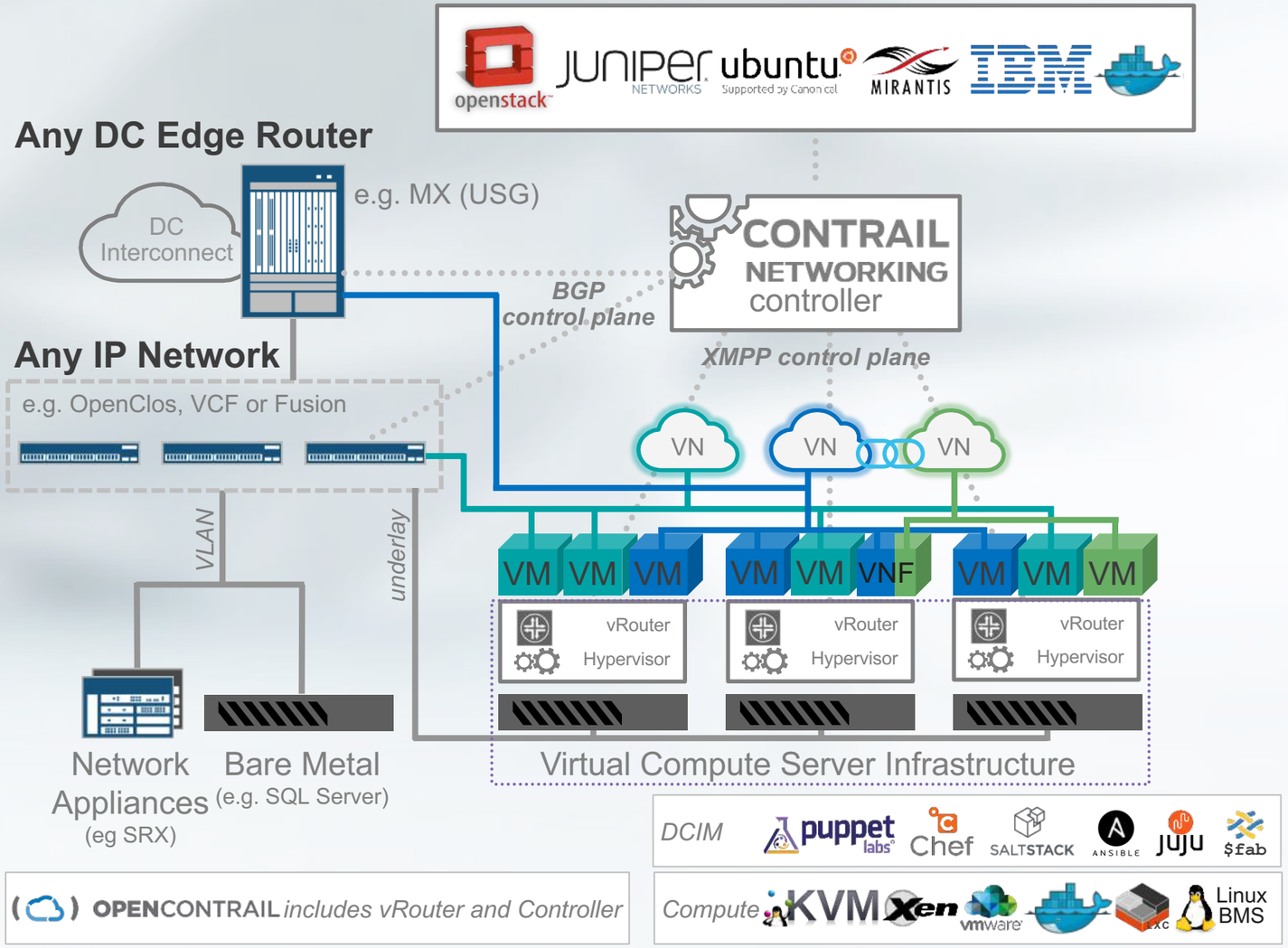


Обзор Contrail

ПРЕИМУЩЕСТВА

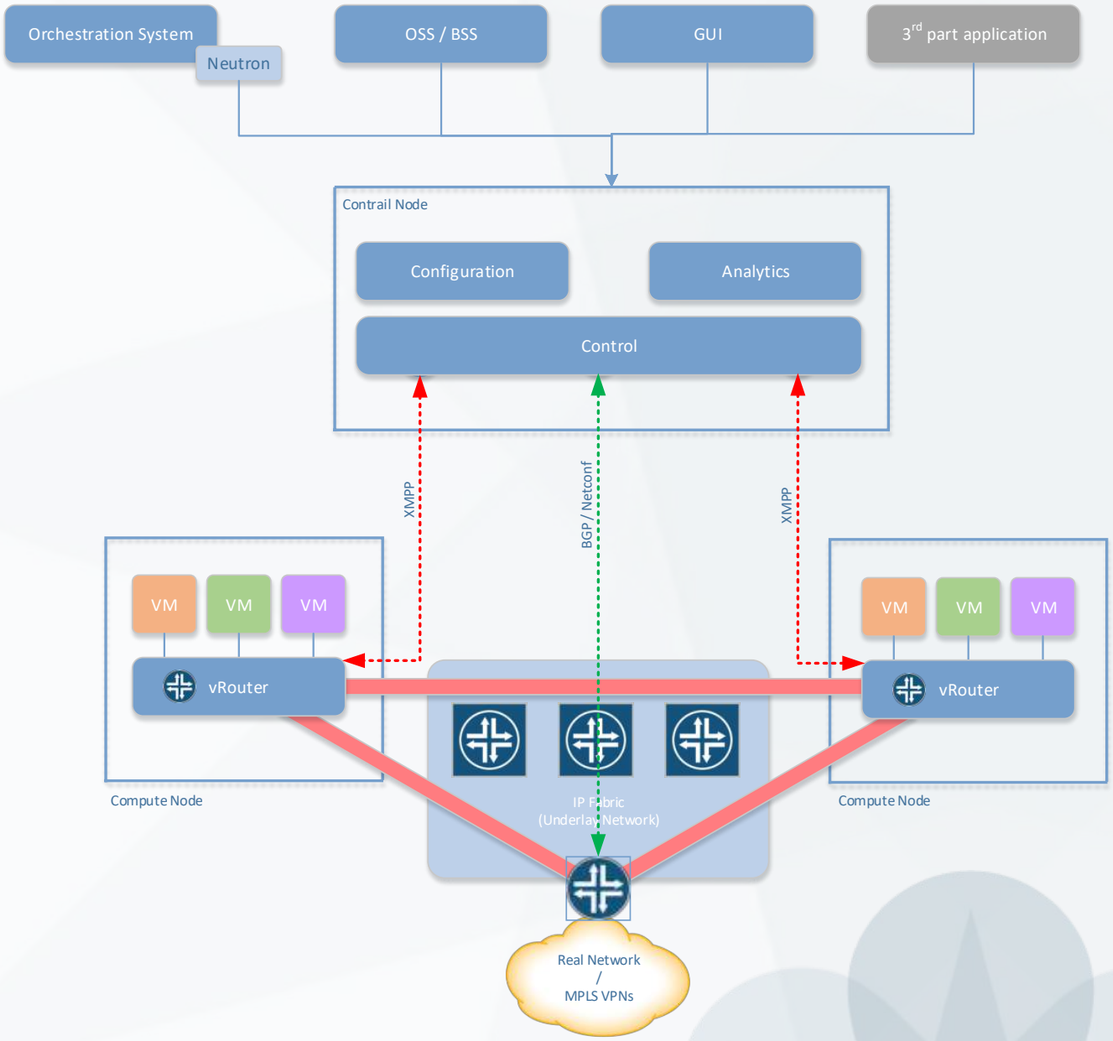
РЕШЕНИЯ

- **В качестве транспортной инфраструктуры может использоваться любая IP сеть**
- **Широкие возможности по автоматизации**
- **Поддержка открытых стандартов для построения наложенной сети:**
  - BGP, EVPN, OVSDB control plane
  - MPLS over GRE/UDP, VXLAN data plane overlays



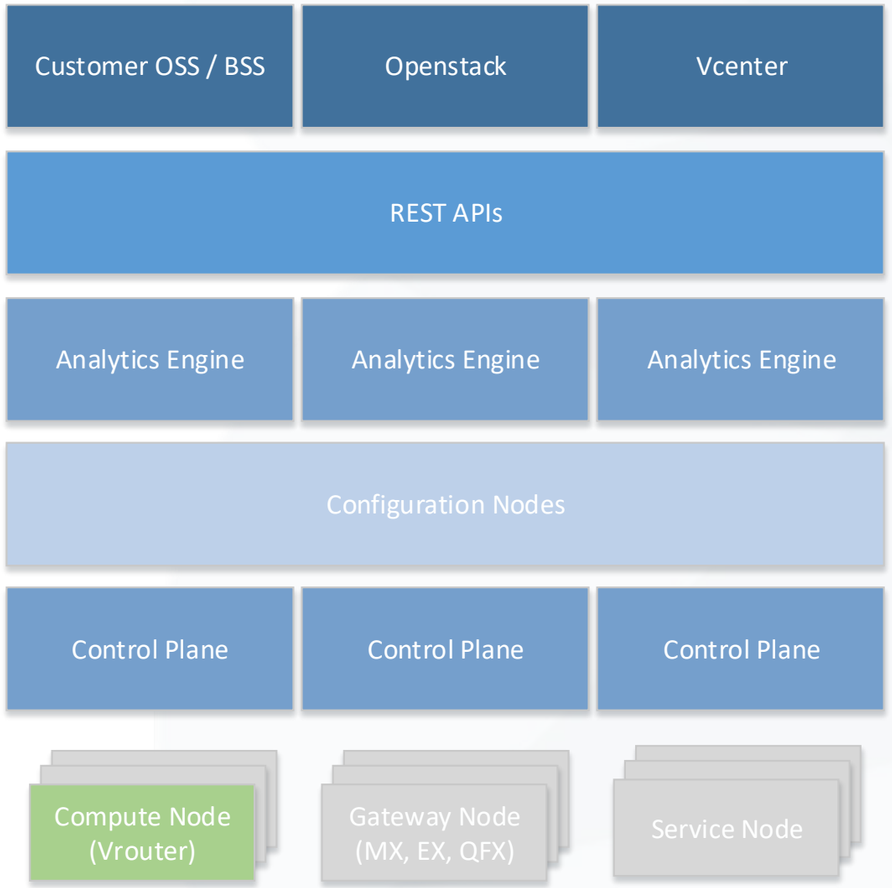
Contrail Overview

- The **Contrail System** consists of two main components: the *OpenContrail Controller* and the *OpenContrail vRouter*.
- Contrail Controller is a logically centralized but physically distributed Software Defined Networking (SDN) controller that is responsible for providing the management, control, and analytics functions of the virtualized network.
- The Contrail vRouter is a forwarding plane (of a distributed router) that runs in the hypervisor of a virtualized server. It extends the network from the physical routers and switches in a data center into a virtual overlay network hosted in the virtualized servers
- The Contrail Controller provides the logically centralized control plane and management plane of the system and orchestrates the vRouters.



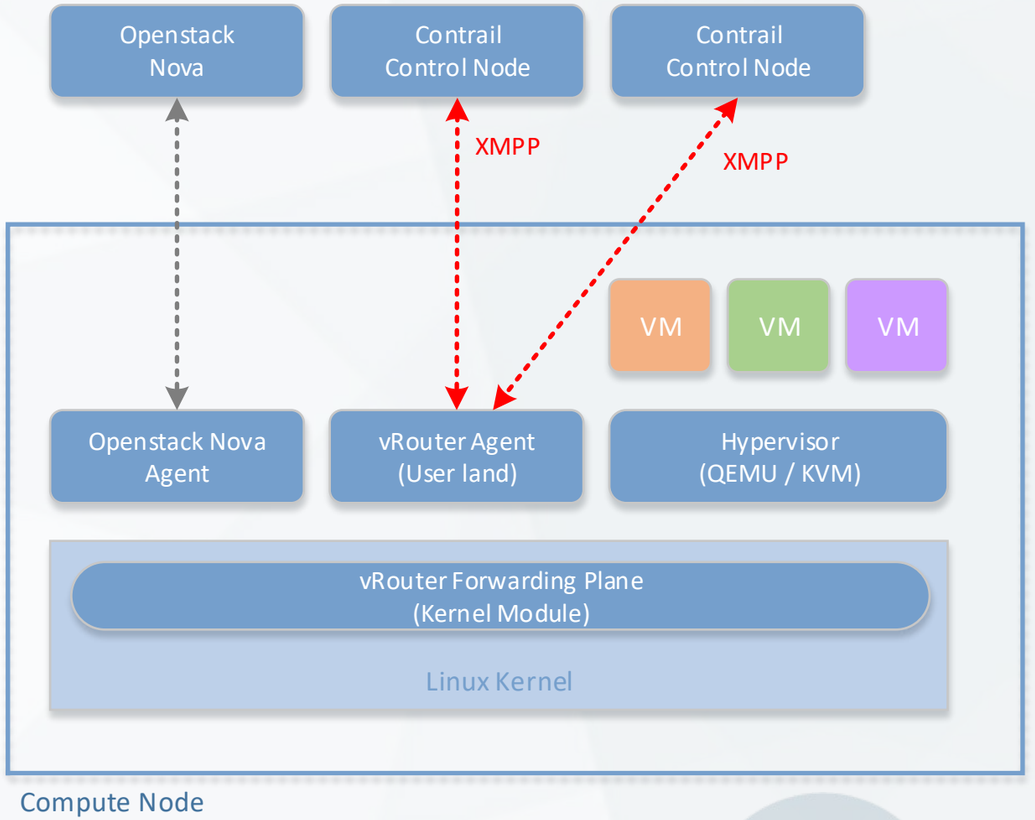
Contrail Overview

- Contrail Stack: Overview of compute node



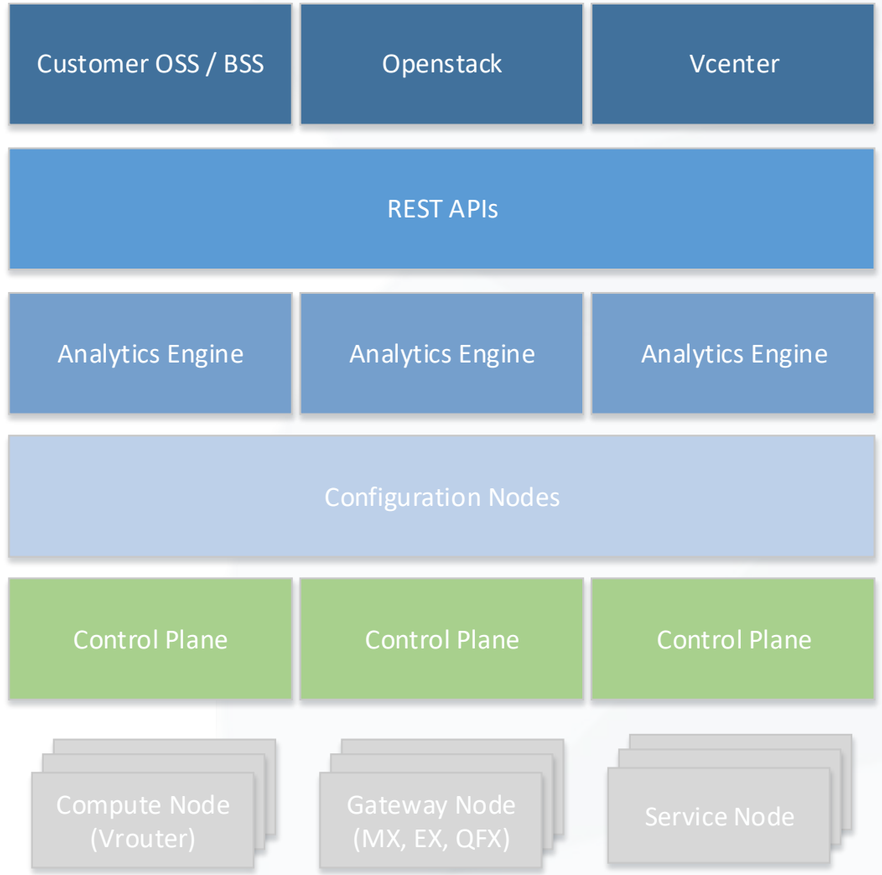
Contrail Overview: Compute Node

- **vRouter** is replaces the Linux Bridge or OVS module in Hypervisor Kernel
- **vRouter** performs bridging (E-VPN) and routing (L3VPN)
- **vRouter** performs networking services like Security Policies, NAT, Multicast, Mirroring, and Load Balancing
- No need for Service Nodes or L2/L3 Gateways for Routing, Broadcast/Multicast, NAT
- Routes are automatically leaked into the VRF based on Policies
- Support for Multiple Interfaces on the Virtual Machines
- Support for Multiple Interfaces from Compute Node to the Switching Fabric



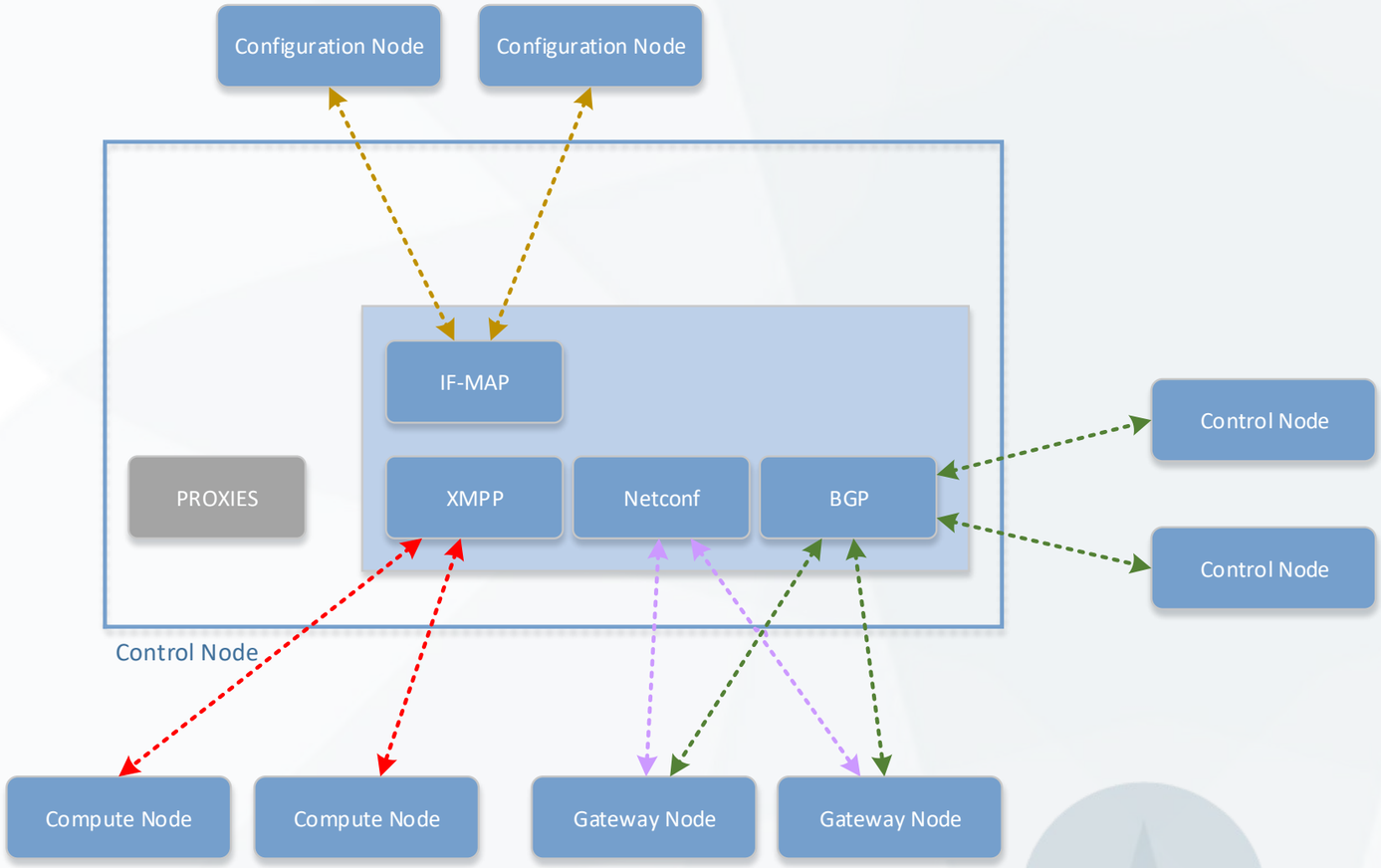
Contrail Overview: Control Plane

- Contrail Stack: Overview of Control Plane



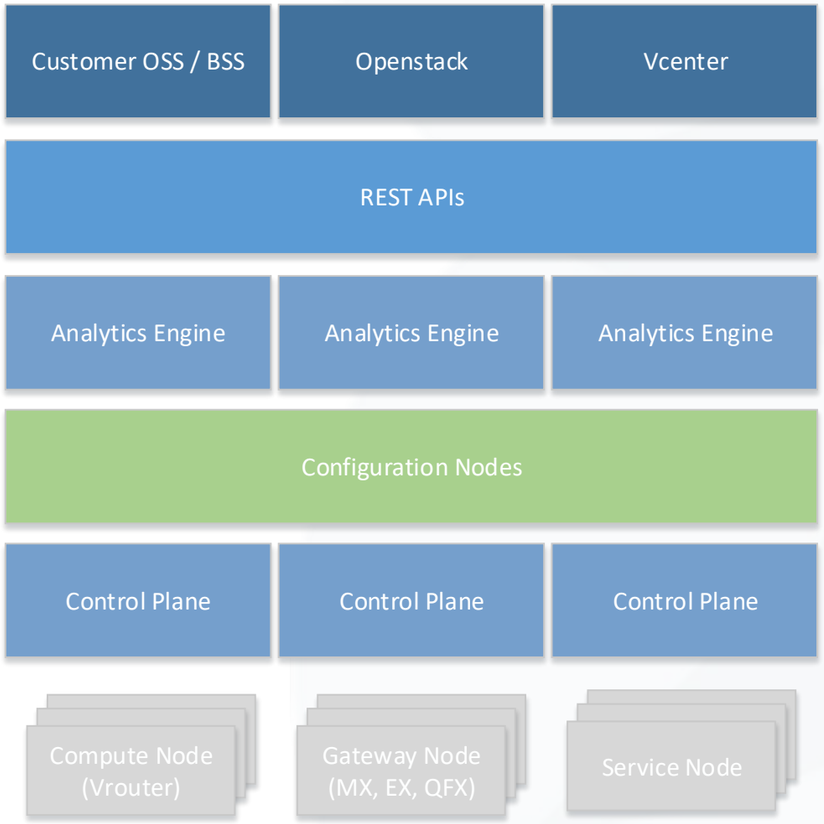
Contrail Overview: Control Plane

- All Control Plane Nodes are active active
- Each vRouter uses XMPP to connect with multiple Control Plane nodes for redundancy
- Each Control Plane Node connects to multiple configuration nodes for redundancy
- BGP is used to connect with Physical Gateway Routers or switches
- Control Plane Nodes federate using BGP



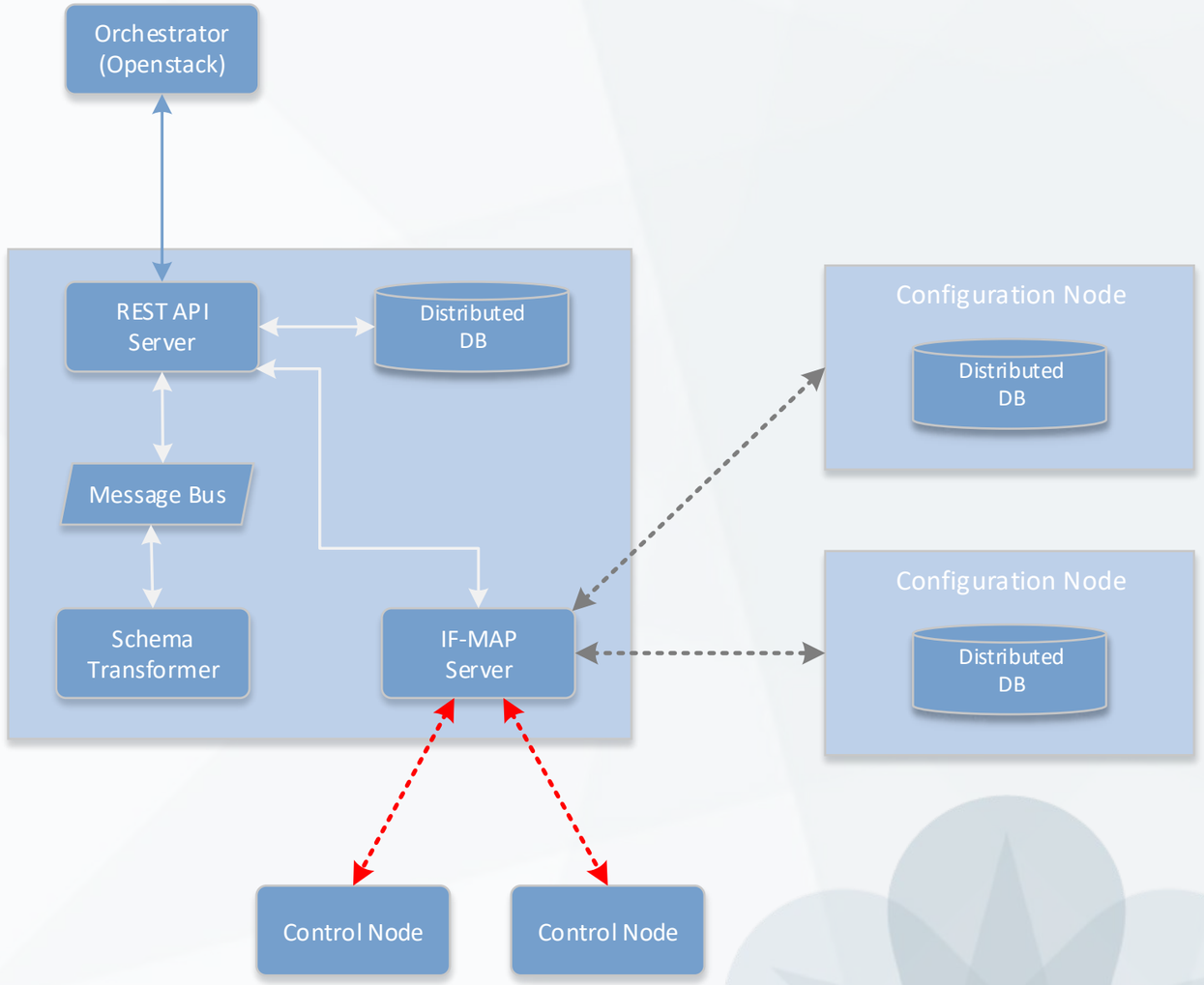
Contrail Overview: Configuration Node

- Contrail Stack: Overview of Configuration Node



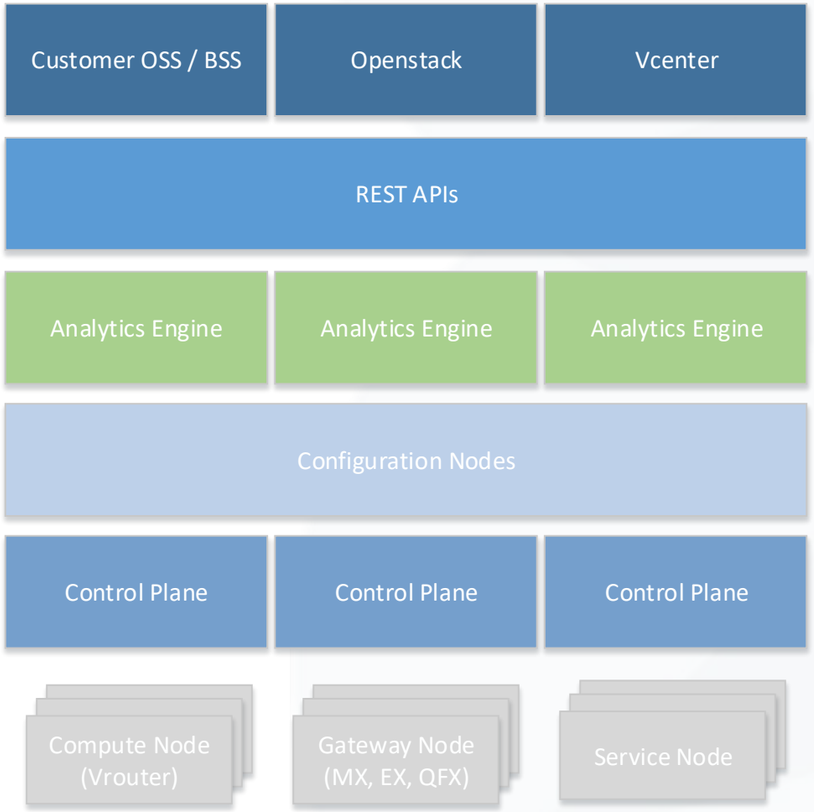
Contrail Overview: Configuration Node

- A **REST API** Server that provides the north-bound interface to an Orchestration System or other application
- A **Rabbitmq** message bus to facilitate communications amongst internal components
- A **Cassandra** database for persistent storage of configuration
- A Schema transformer that learns about changes in the high level data model over the message bus and transforms (or compiles) these changes in the high level data model into corresponding changes in the low level data model
- An **IF-MAP** Server that provides a south bound interface to push the computed low-level configuration down to the Control nodes
- **Zookeeper** (not shown in diagram) is used for allocation unique object identifiers and to implement transactions



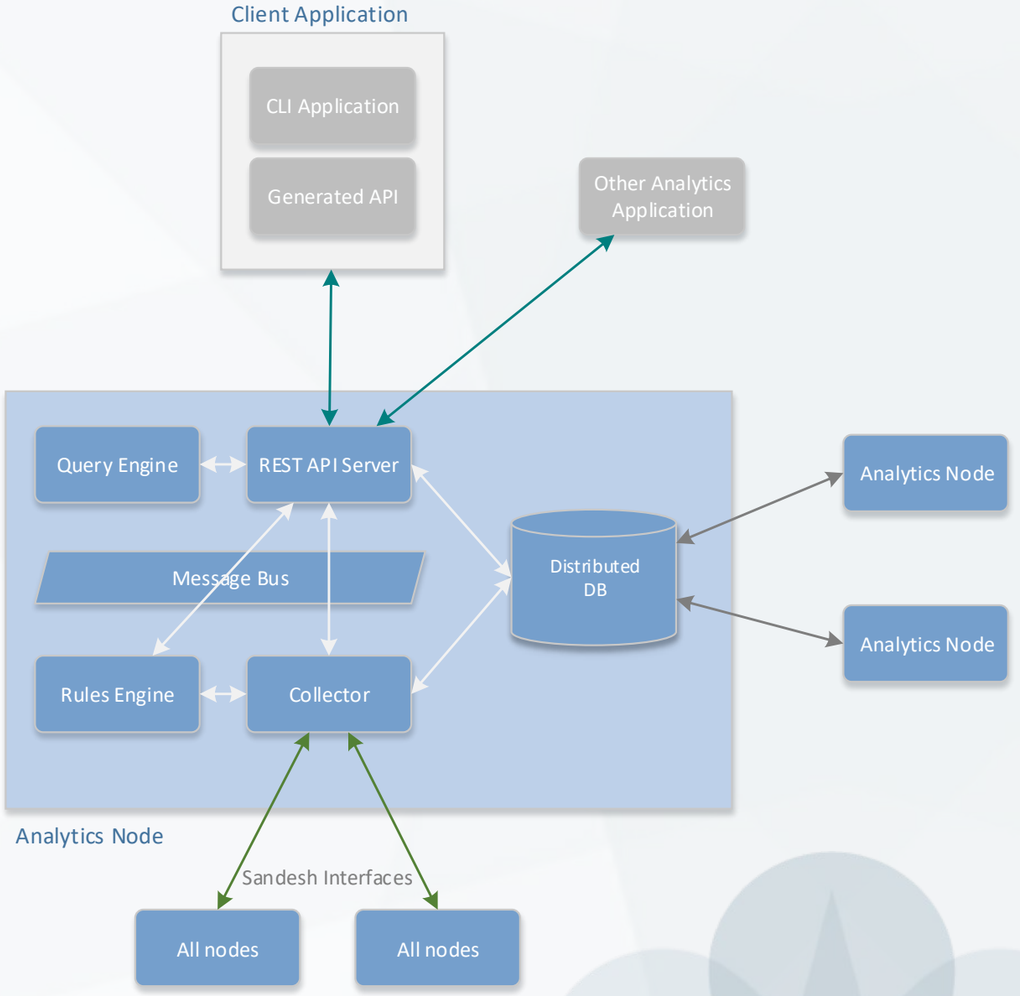
Contrail Overview: Analytics Node

- Contrail Overview: Analytics Node



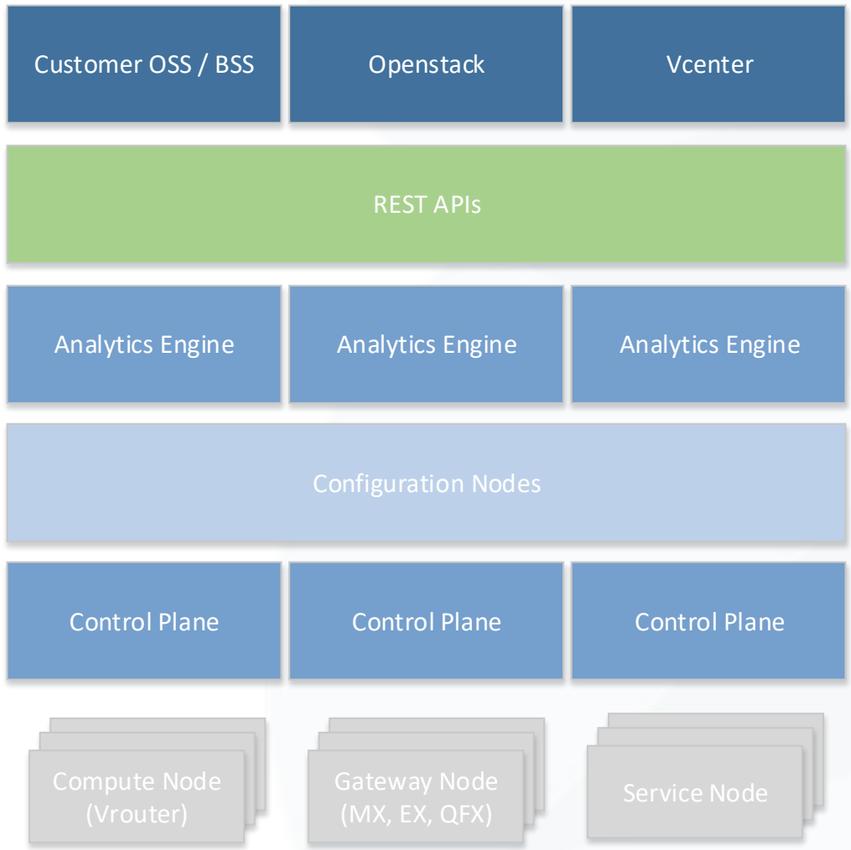
Contrail Overview: Analytics Node

- A Collector that exchanges Sandesh with components in control nodes and configuration nodes to collect analytics information
- A cassandra database for storing this information
- A rules engine to automatically collect operational state when specific events occur
- A REST API server that provides a northbound interface for querying the analytics database and for retrieving operational state.
- A Query engine for executing the queries received over the northbound REST API. This engine provides the capability for flexible access to potentially large amounts of analytics data

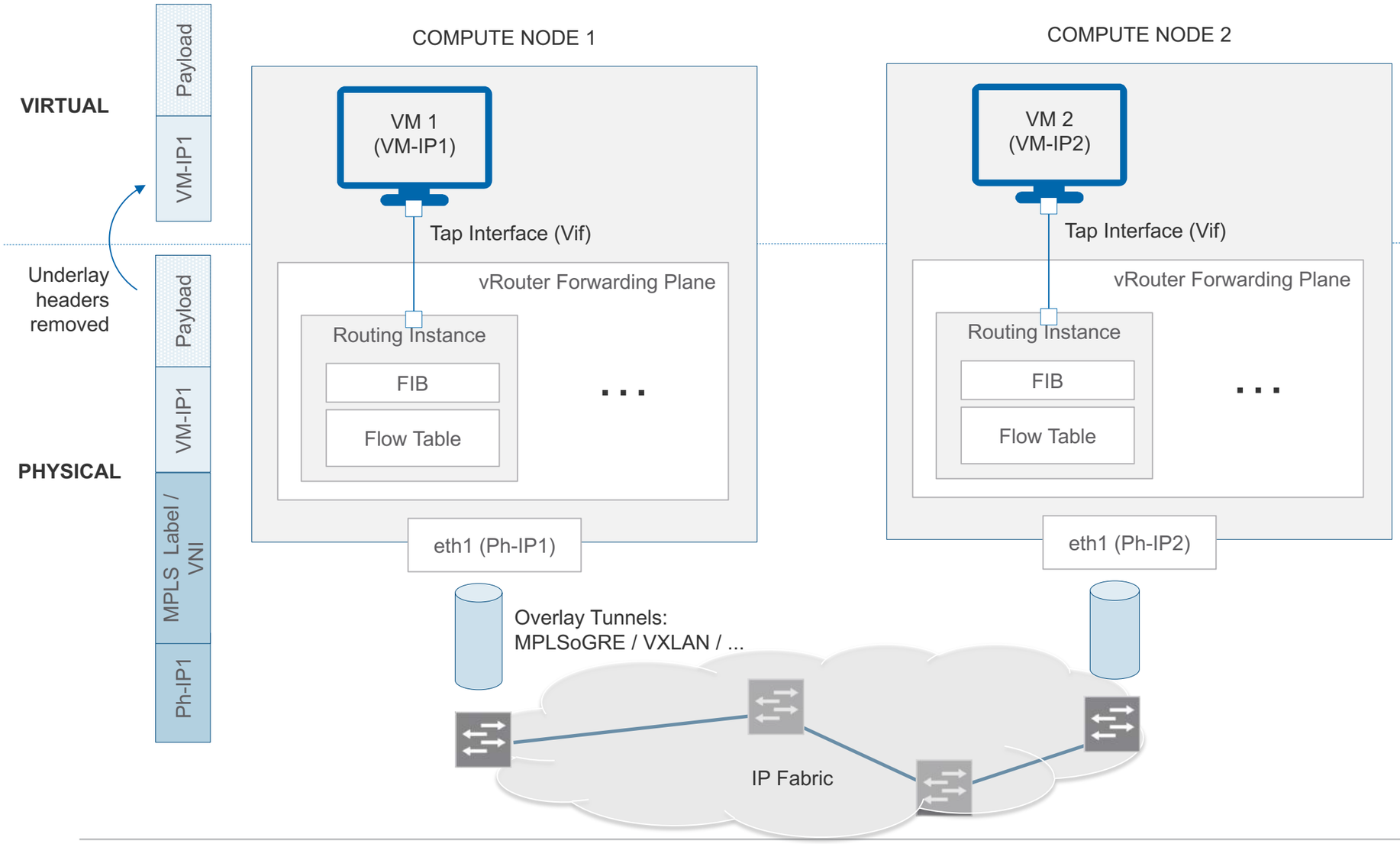


Contrail Overview: Northbound API

- Contrail Overview: Northbound API



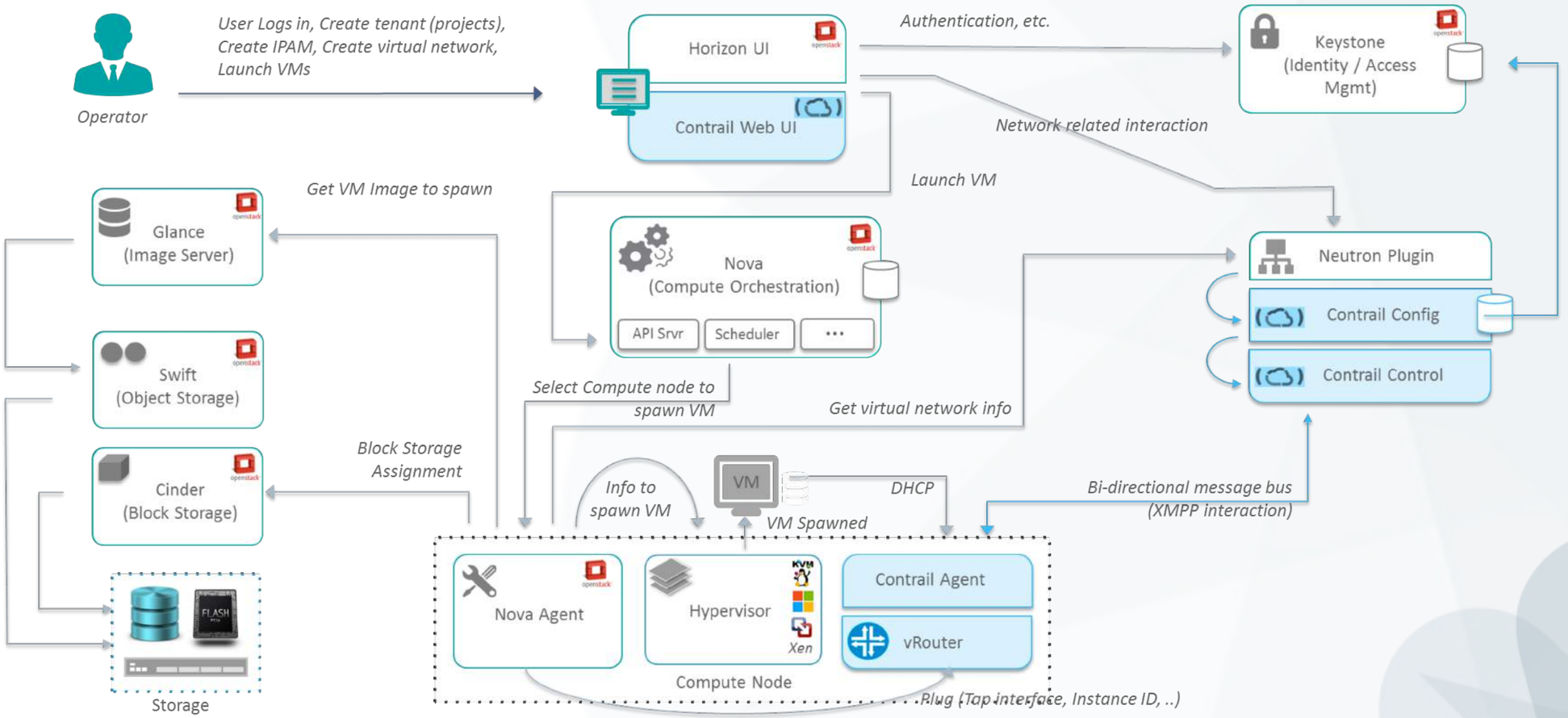
COMPUTE NODE: FORWARDING / TUNNELING



1. Guest OS ARPs for destination within subnet or default GW
2. VRouter receives the ARP and responds back with VRRP MAC
3. Guest OS sends traffic to the VRRP MAC, vRouter encapsulates the packet with appropriate MPLS/VNI tag and GRE header
4. Physical Fabric Routers on Physical IP Address
5. Returning packets get forwarded to appropriate Routing Instance by the MPLS/VNI tag
6. vRouter de-capsulates the packet, and forwards it to the Guest OS

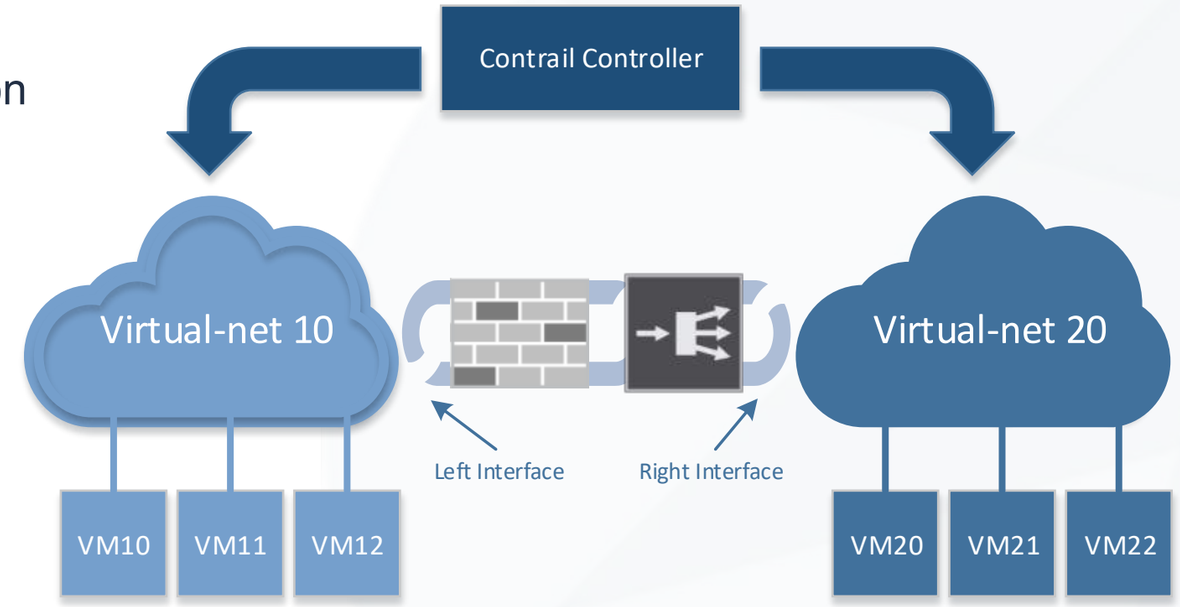
Contrail Overview

- Openstack & Contrail Components



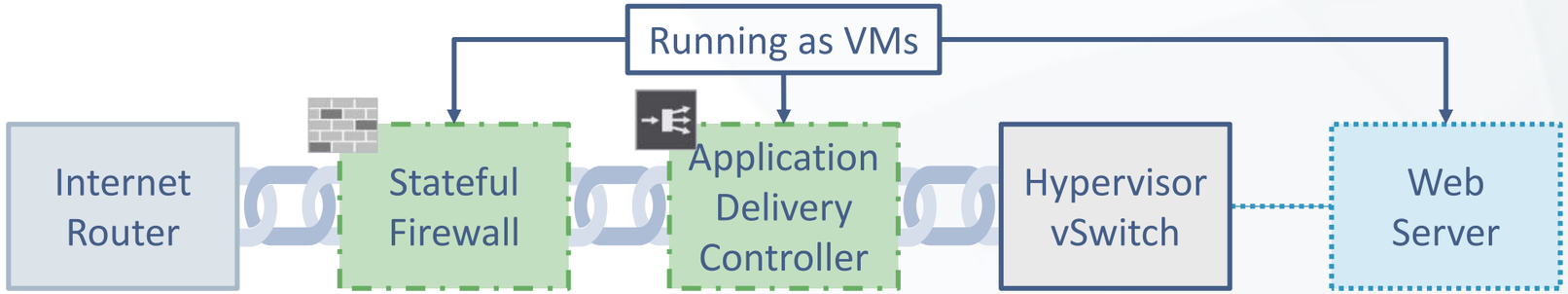
Contrail Overview

- Service Chaining overview
  - By means of services chains you introduce by just a single click services like:
    - Firewall
    - IDP
    - UDP
    - Load Balancing
    - DDoS prevention

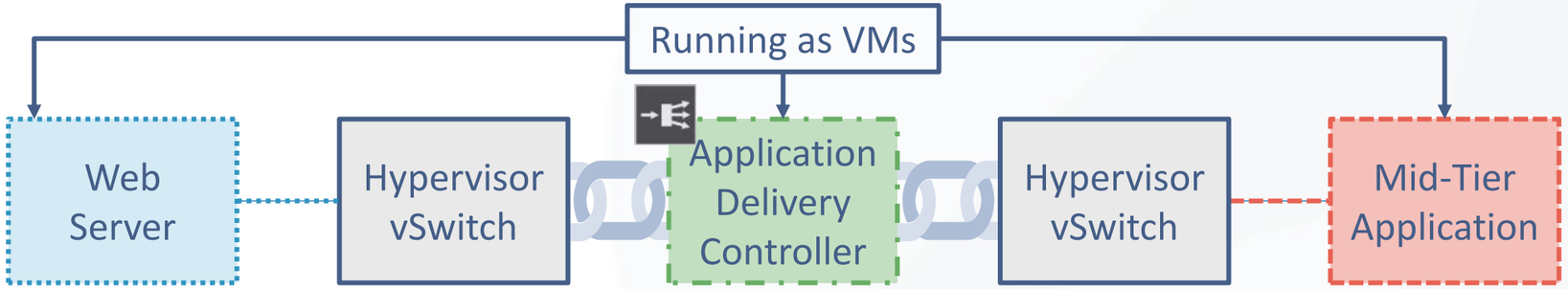


Contrail Overview

- Service Chaining examples
  - Use case 01:



  - Use case 02:

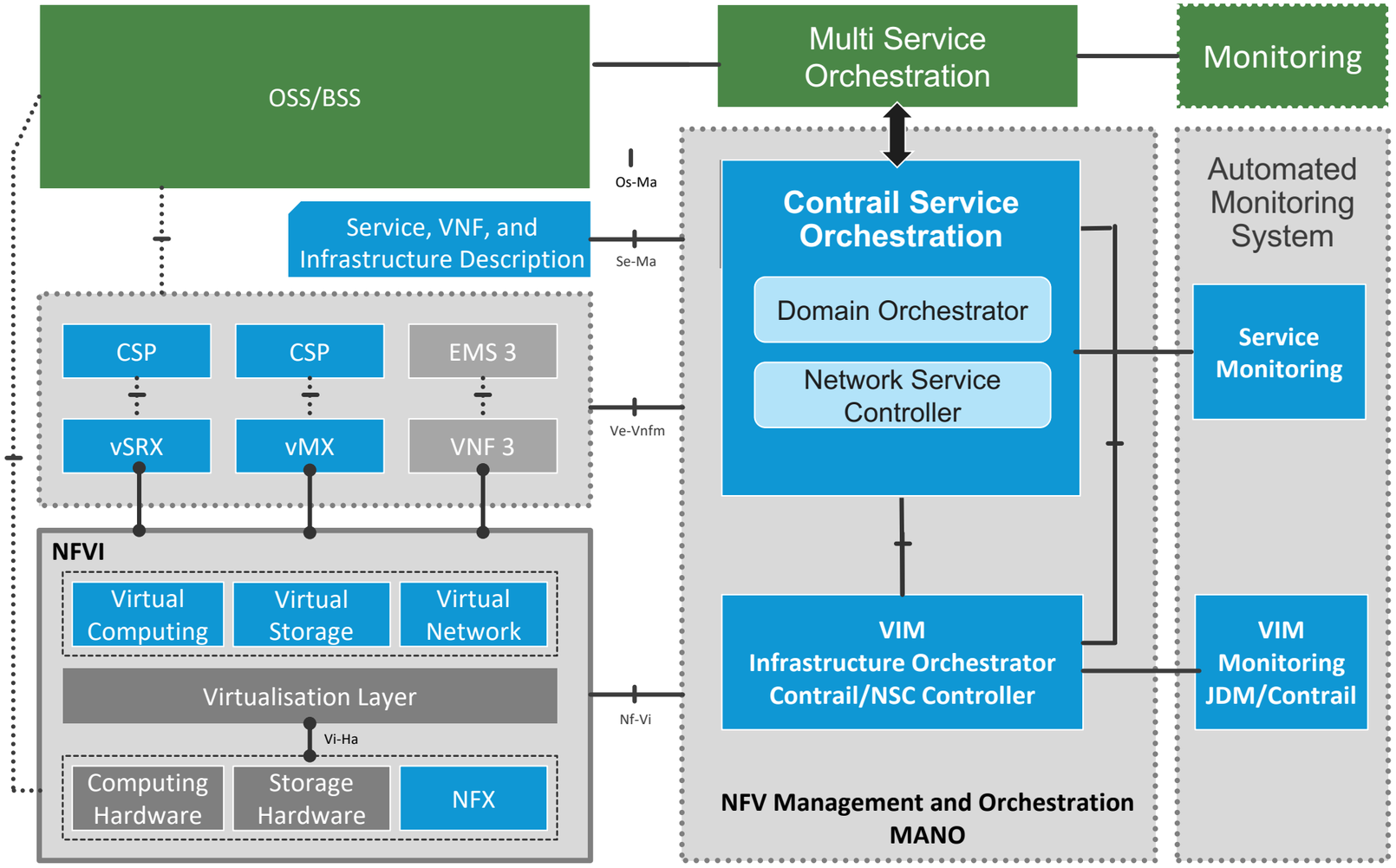


CSO – ETSI NFV Reference Architecture

Solution Components

- OSS (Partner)
- Customer Facing
  - Multi Service Orchestration (Partner)
- Resource Facing
  - CSO Domain Orchestrator
- SD-WAN / NFX Device Controller
  - Network Services Controller
- Infrastructure Orchestration
  - Canonical Openstack
- SDN Controller
  - Contrail SDN (Centralized), NSC (SD-WAN)

Juniper Partner

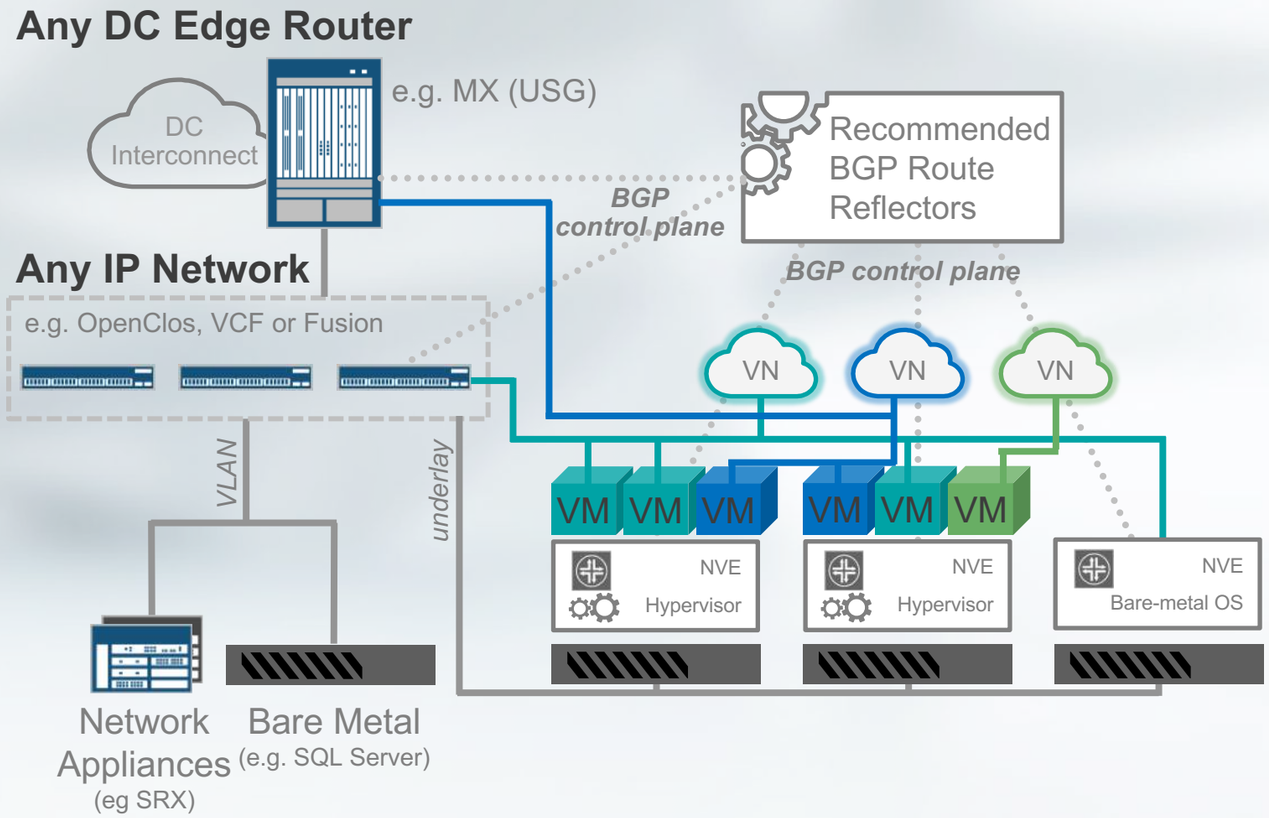


Обзор EVPN-NVO

Ethernet VPN внутри дата-центра

Масштабируемость L2 сетей при помощи технологии VXLAN

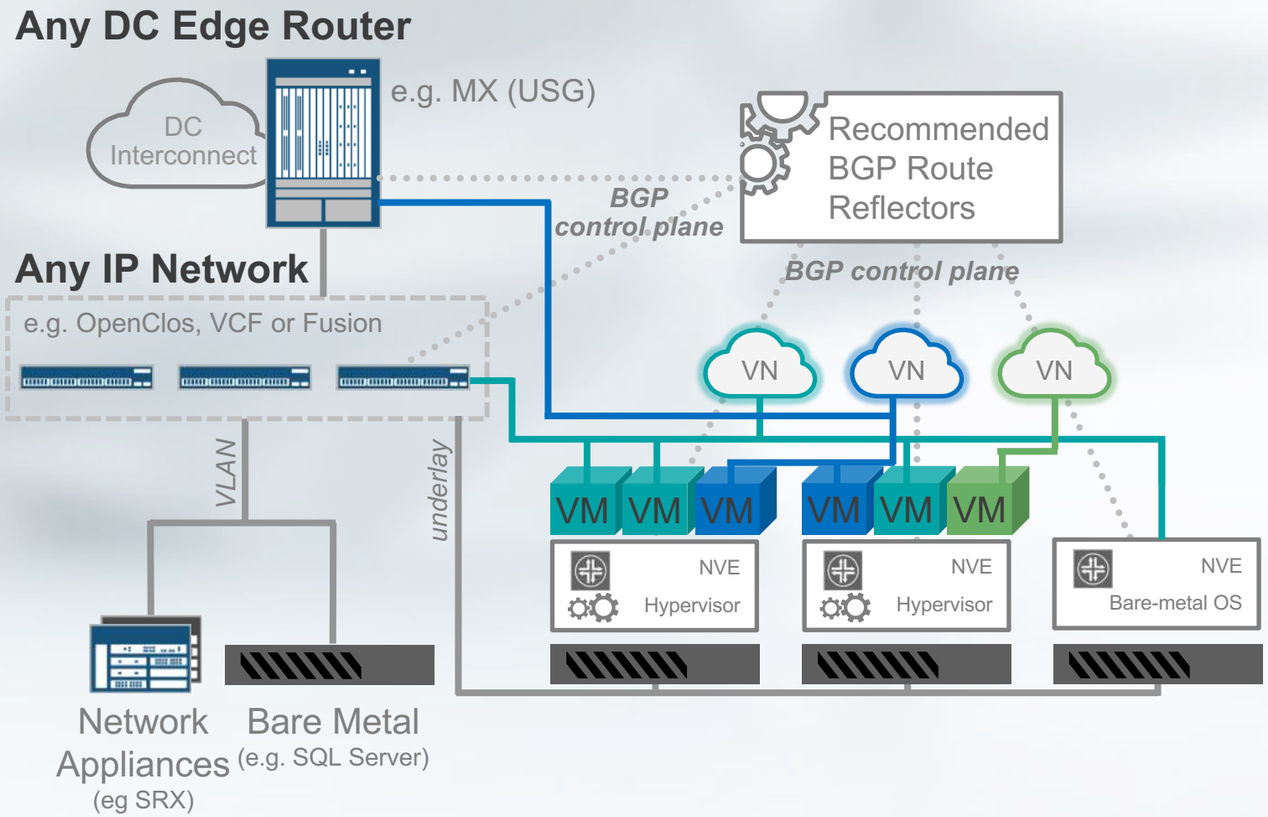
- Отсутствие широковещательного трафика. Использование BGP для передачи информации об IP и MAC адресах
- Использование стандартизованных IETF протоколов
- Терминация туннелей на платформе виртуализации или коммутаторе доступа
- Балансировка нагрузки по всем линкам
- В качестве data plane может использоваться VXLAN, MPLSoGRE, MPLS или NVGRE



Обзор EVPN-NVO

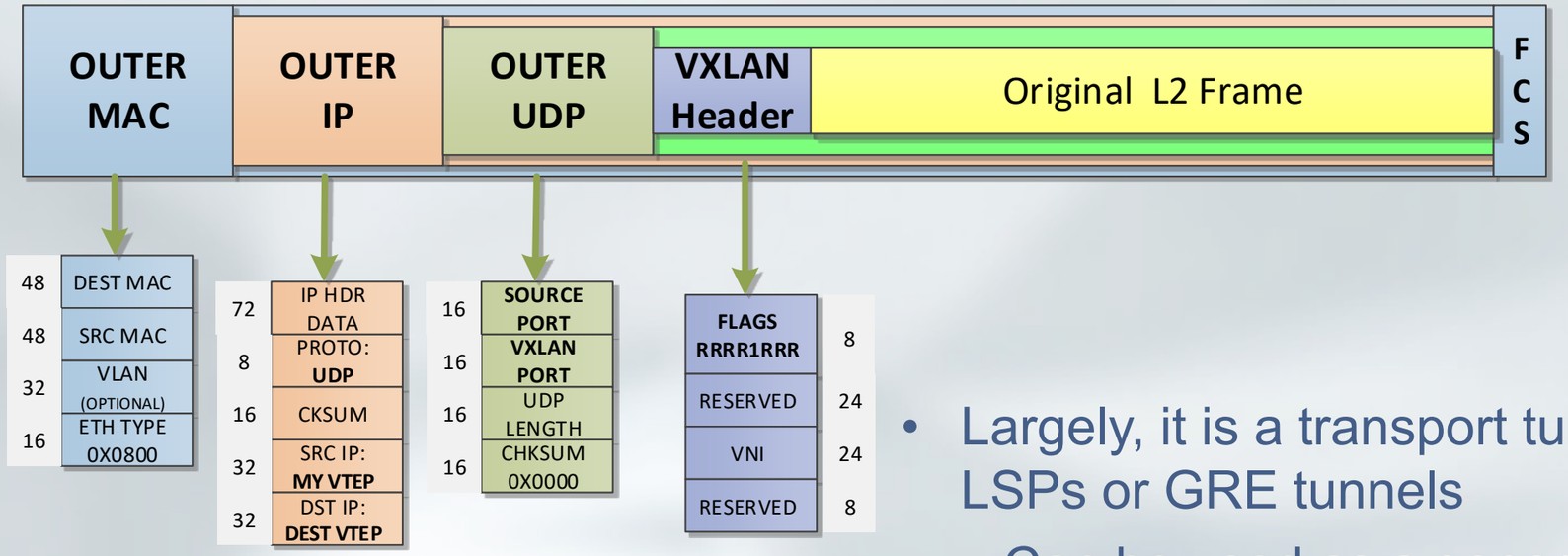
ПРЕИМУЩЕСТВА РЕШЕНИЯ

- В качестве транспортной инфраструктуры может использоваться любая IP сеть
- Единая технология как внутри дата-центра так и для обеспечения связности между дата-центрами
- Открытый стандарт, который поддерживается Juniper, Cisco, ALU



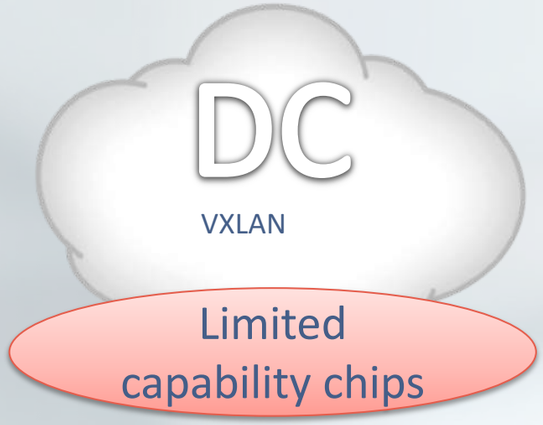
VXLAN
транспорт

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

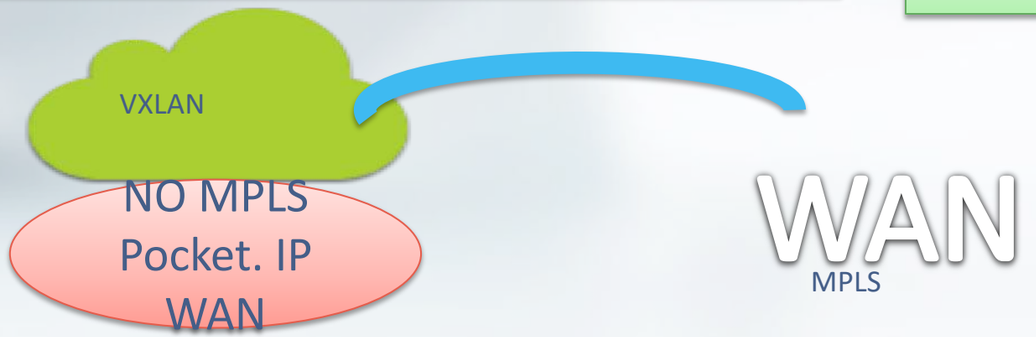
- TOR/Access/Spine layers of DCs have very limited capabilities
- MPLS-phobia inside DCs: Cost and (perceived) complexity
- Flat IP fabric based DCs
- Originated in DCs but gaining momentum in SD WAN

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



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

EVPN

технология

EVPN преимущества

Эффективность

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

Конвергенция

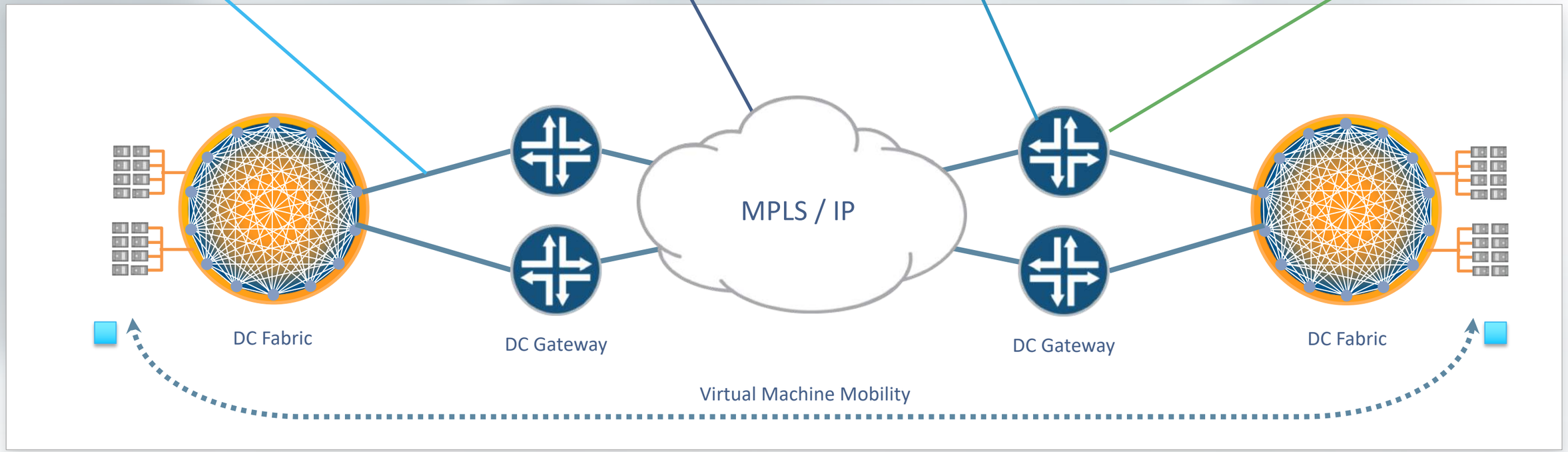
Высокий уровень доступности, быстрая сходимость и перемаршутизация

L3 и L2

Интеграция L2 и L3 уровней в протокол управления

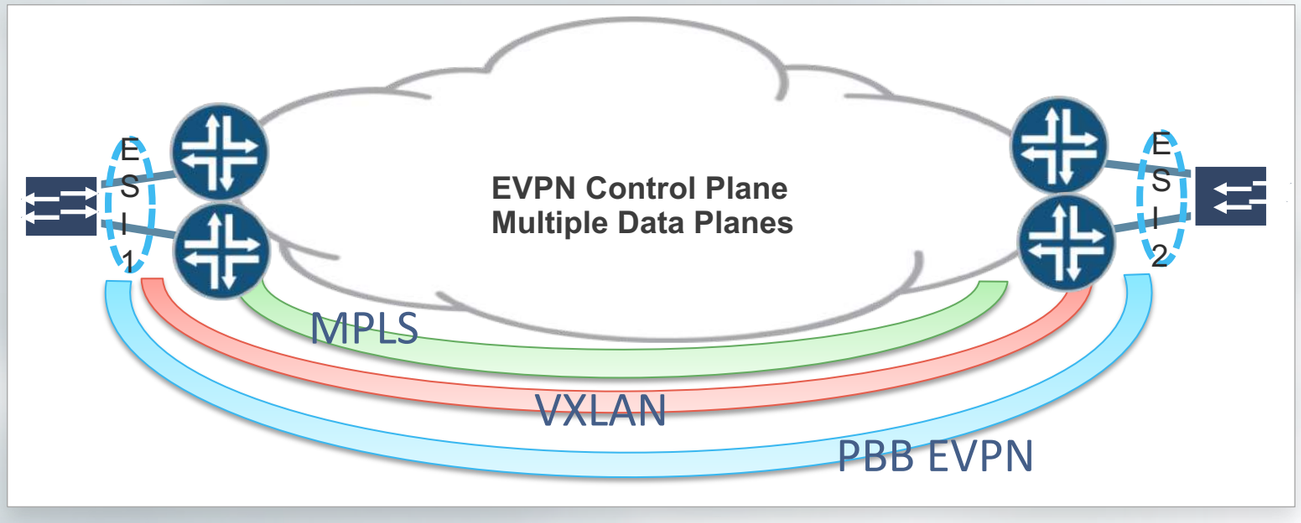
Оптимизация

Оптимизация входящего и исходящего трафика при миграции виртуальных машин



EVPN : Варианты использования

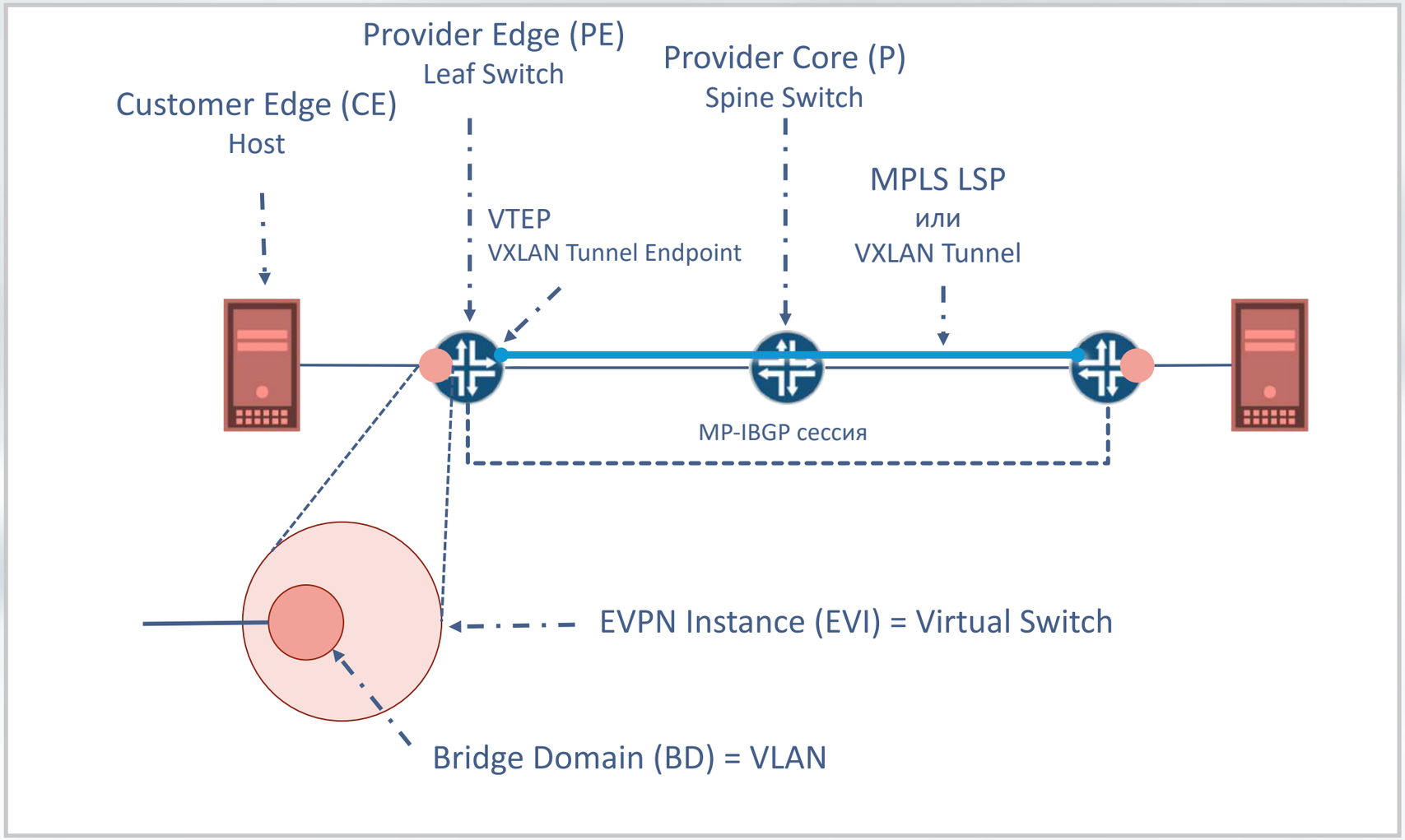
MPLS Transport (WAN)	L2 Service	L3 Service
	MPLS service label advertised for MACs	MPLS service label advertised for subnets
	ELINE, ELAN, ETREE. M 2.0 compliant BE/Met Services	L2/L3 Service Integration



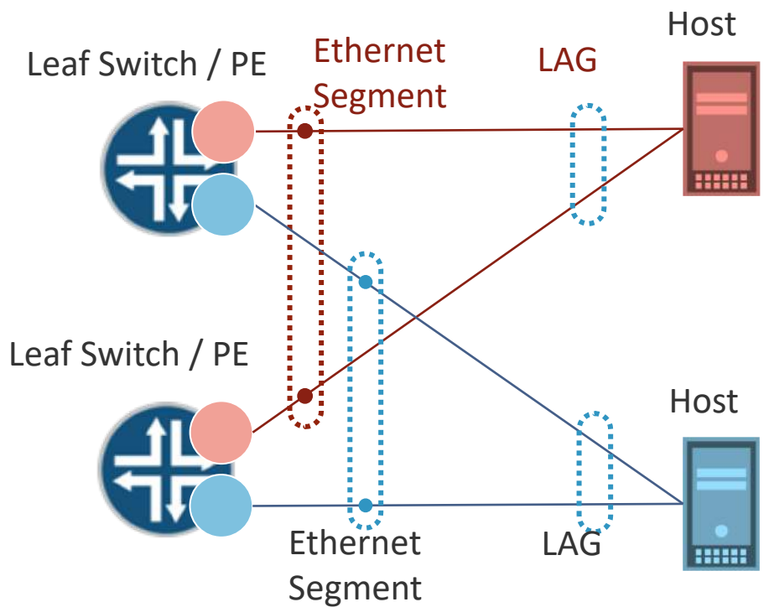
IP Transport (DC, OTT)	L2 Service	L3 Service
	VNIDs advertised for MACs	VNID advertised for subnets
	L2 stretch between VMs over IP OTT L2VPN	Inter-subnet traffic Integration in DC, OTT L3VPN with EVPN

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

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

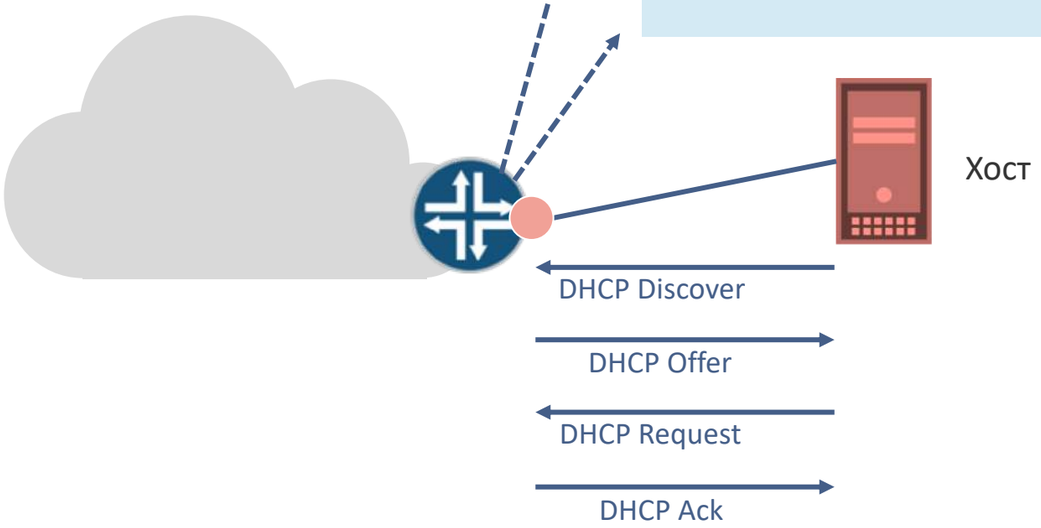
Типы маршрутов NLRI для EVPN

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

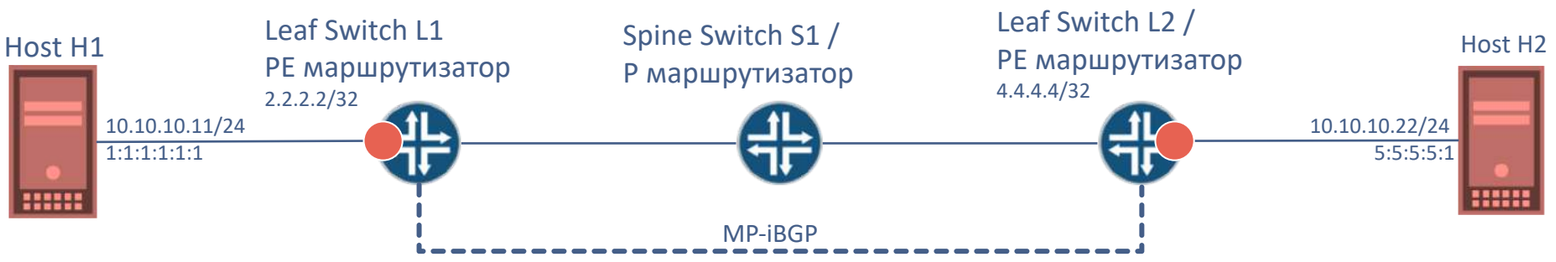
Обязательно: коммутатор обучается MAC адресу хоста отслеживанием адреса источника (source address) каждого ethernet фрейма

Опционально: коммутатор изучает IP адрес хоста отслеживанием DHCP или ARP сообщений



Удаленный процесс mac learning

- MAC / IP advertisement (type 2) routes

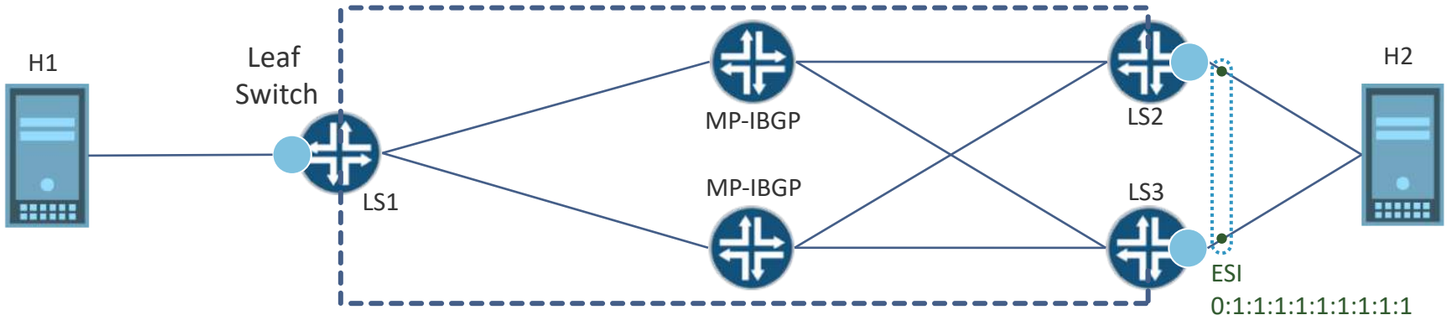


NLRI	Route type	MAC / IP Advertisement Route (Type 2)
	Route Distinguisher (RD)	RD of red EVI on leaf switch L2
	Ethernet Segment Identifier (ESI)	0 (single homed host)
	Ethernet Tag ID	Global VXL AN VNID of bridge domain
	MAC Address	MAC address of host H2 (5:5:5:5:5:1)
	IP Address	IP address of host H2 (10.10.10.22)
	MPLS label 1 + MPLS label 2	VNID
Next-hop		Loopback IP address of L2 (4.4.4.4)
Extended Community		Route-targets (Red)
Other attributes (Origin, AS-Path, Local-Pref, …)		…

Ethernet Auto-discovery (Type 1) Routes

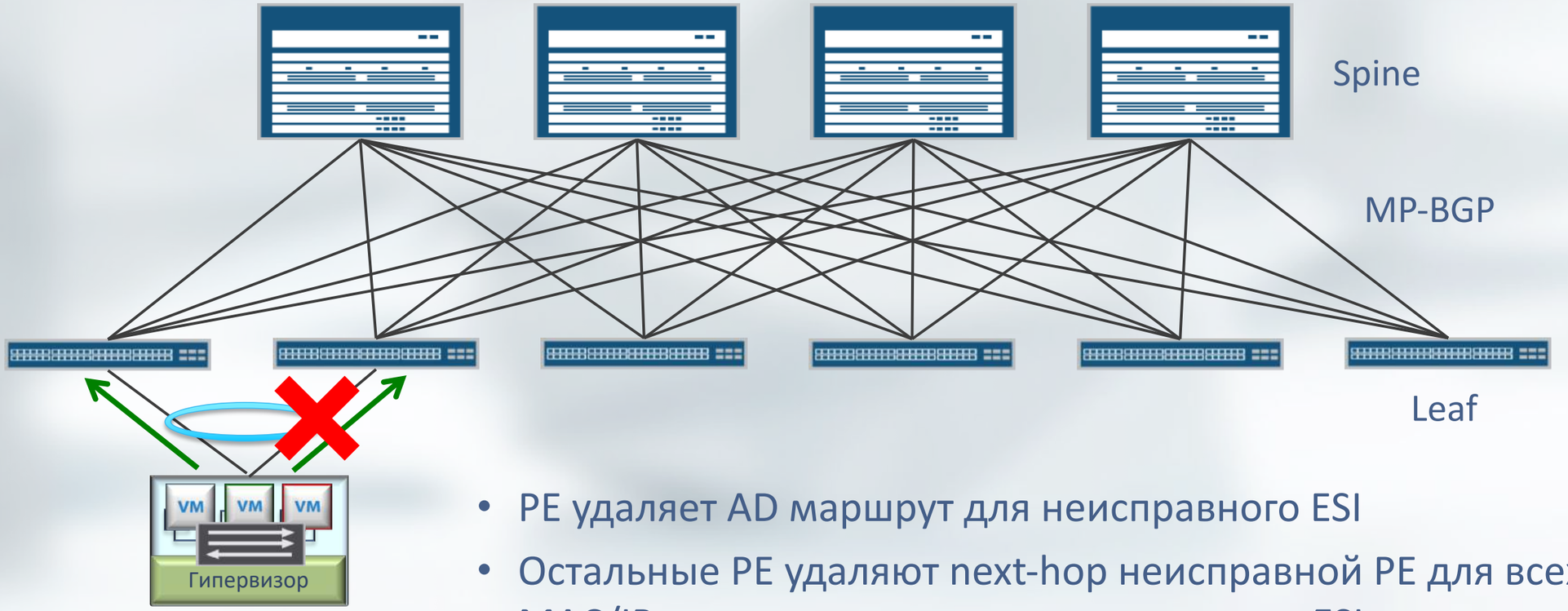
- для Ethernet segment (ES): многоканальность и быстрая сходимость

NLRI	Route type	Ethernet Auto-Discovery Route (Type 1)
	Route Distinguisher (RD)	RD of EVI on leaf switch LS2 (contains IP of LS2)
	Ethernet Segment ID (ESI)	0:1:1:1:1:1:1:1:1:1
	Ethernet Tag ID	MAX-ET
	MPLS label	0
Extended communities		ESI Label Extended Community: • Single-Active Flag = false (0) • ESI Label = null
Next-hop		Loopback IP address of LS2
Other attributes (Origin, AS-Path, Local-Pref, …)		



Ethernet Auto-Discovery (Type 1) Route

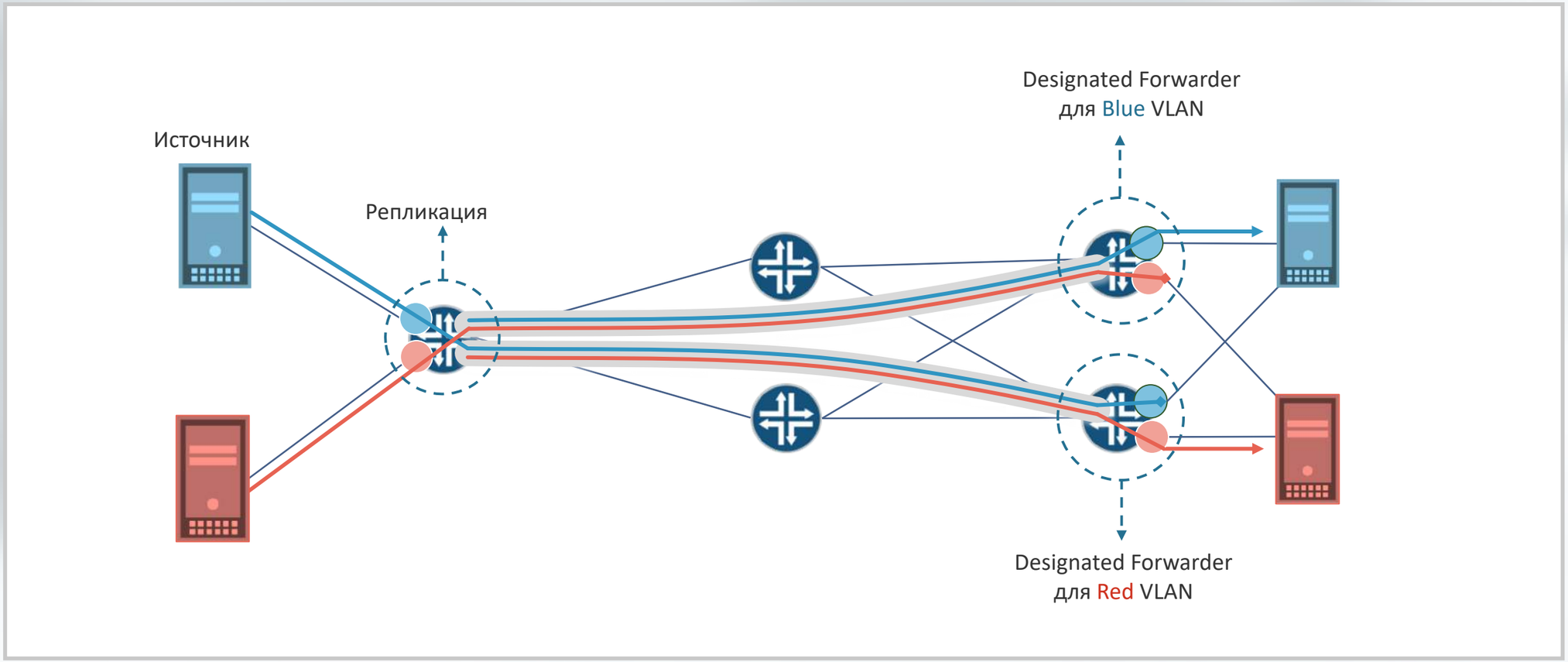
Быстрая сходимость
MAC Mass Withdrawal



- 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

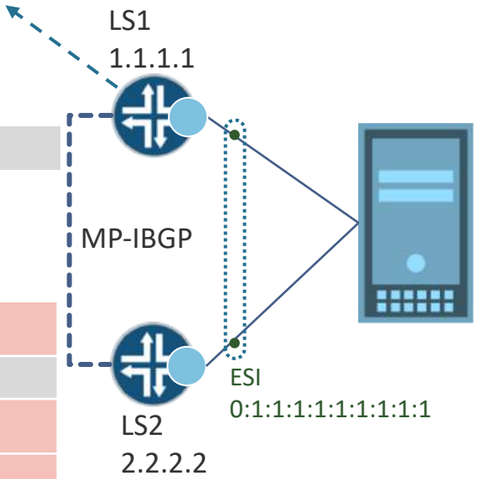


Ethernet Segment (Type 4) Routes

- Необходим для выбора designated forwarder

Designated Forwarder Table для ESI 0:1:1:1:1:1:1:1:1:1

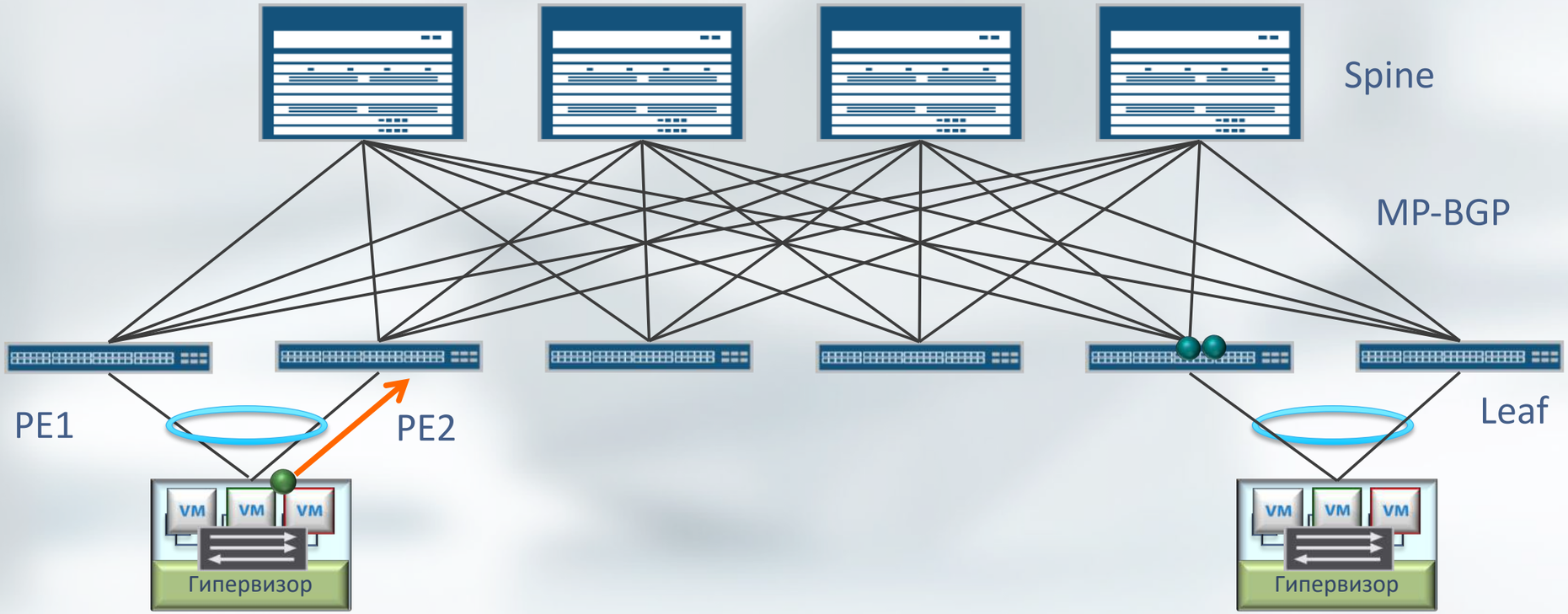
Originators *(Sorted by IP address)*	Designated Forwarder for VLANs…
LS1 (1.1.1.1)	0, 2, 4, 6, …, 4094
LS2 (2.2.2.2)	1, 3, 5, 7, …, 4095



Ethernet Segment Route (Type 4) for ESI from originator LS2

NLRI	Route type	Ethernet Segment Route (Type 4)
	Route Distinguisher (RD)	…
	Ethernet Segment ID (ESI)	0:1:1:1:1:1:1:1:1:1
	Originator IP Address	IP address of LS2
Extended communities		ES-Import Route Target
Other attributes		…

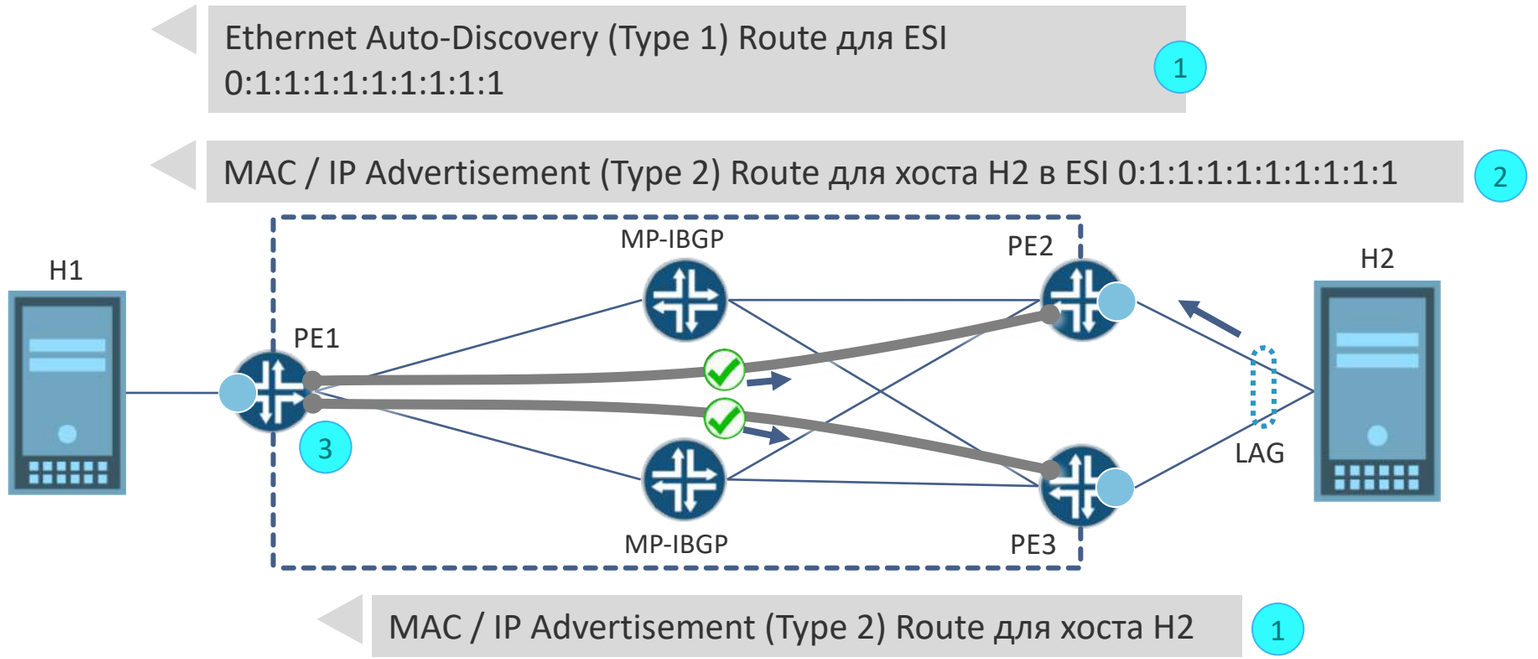
Балансировка нагрузки



- 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



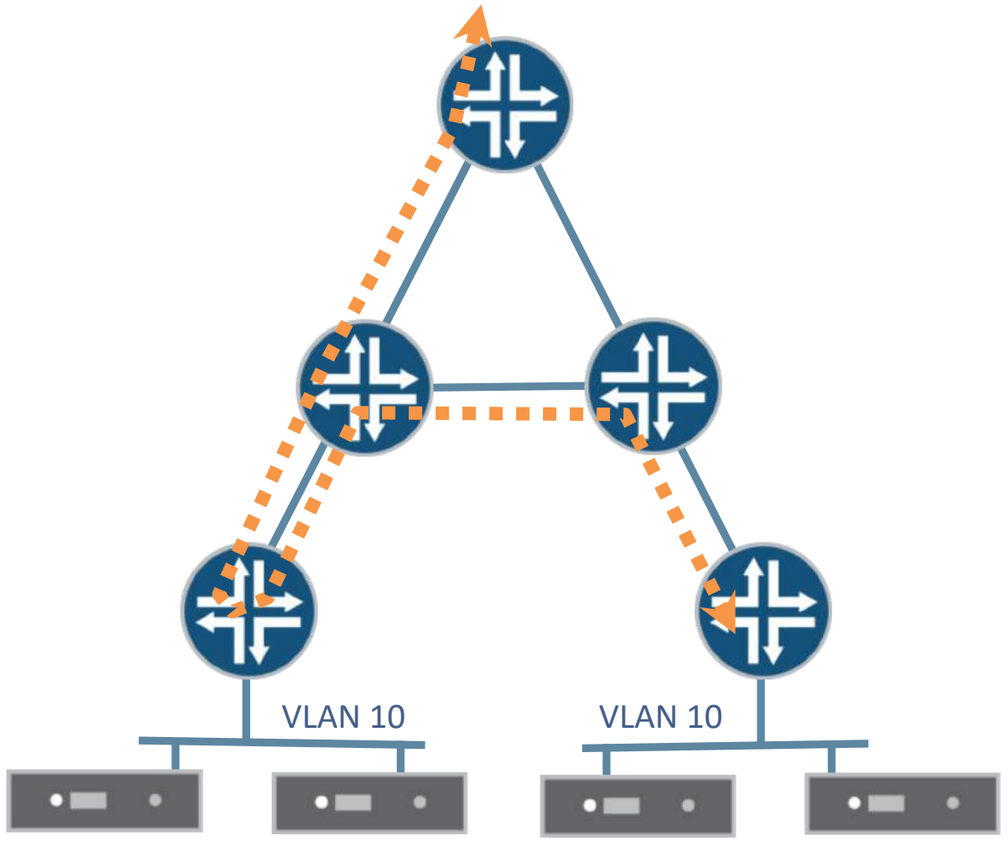
Layer 3
маршрутизация

EVPN: функциональность L3 шлюза

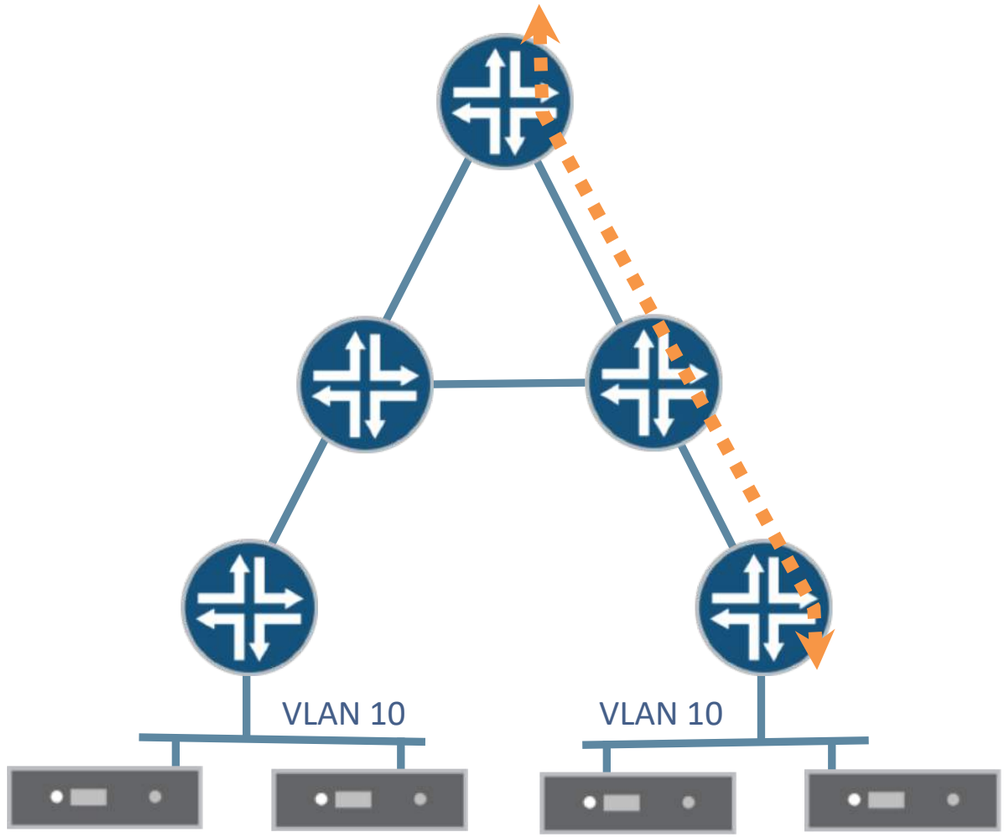
- Установка IRB интерфейса для каждого bridge domain необходима для:
  - обеспечения L3 связности
  - оптимальной передачи трафика между VLAN в внутри и между дата-центрами
- Синхронизация MAC/IP хоста
  - маршрутизатор изучает ARP сообщения для составления таблицы соответствия MAC и IP хоста
  - анонсирует 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

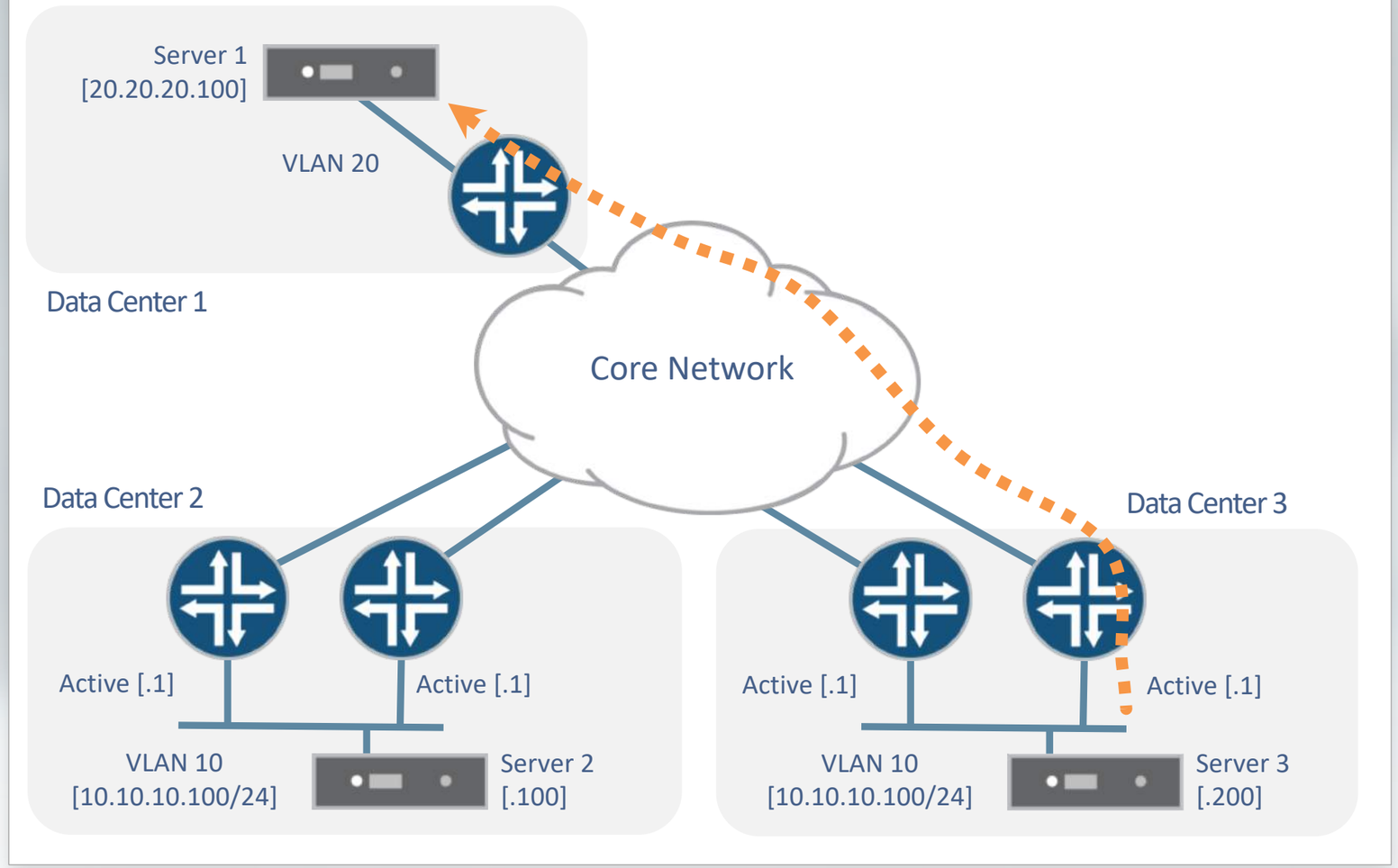


Эффективная маршрутизация



Оптимизация исходящего трафика

Задача: обеспечить оптимальный путь от сервера 3 к серверу 1



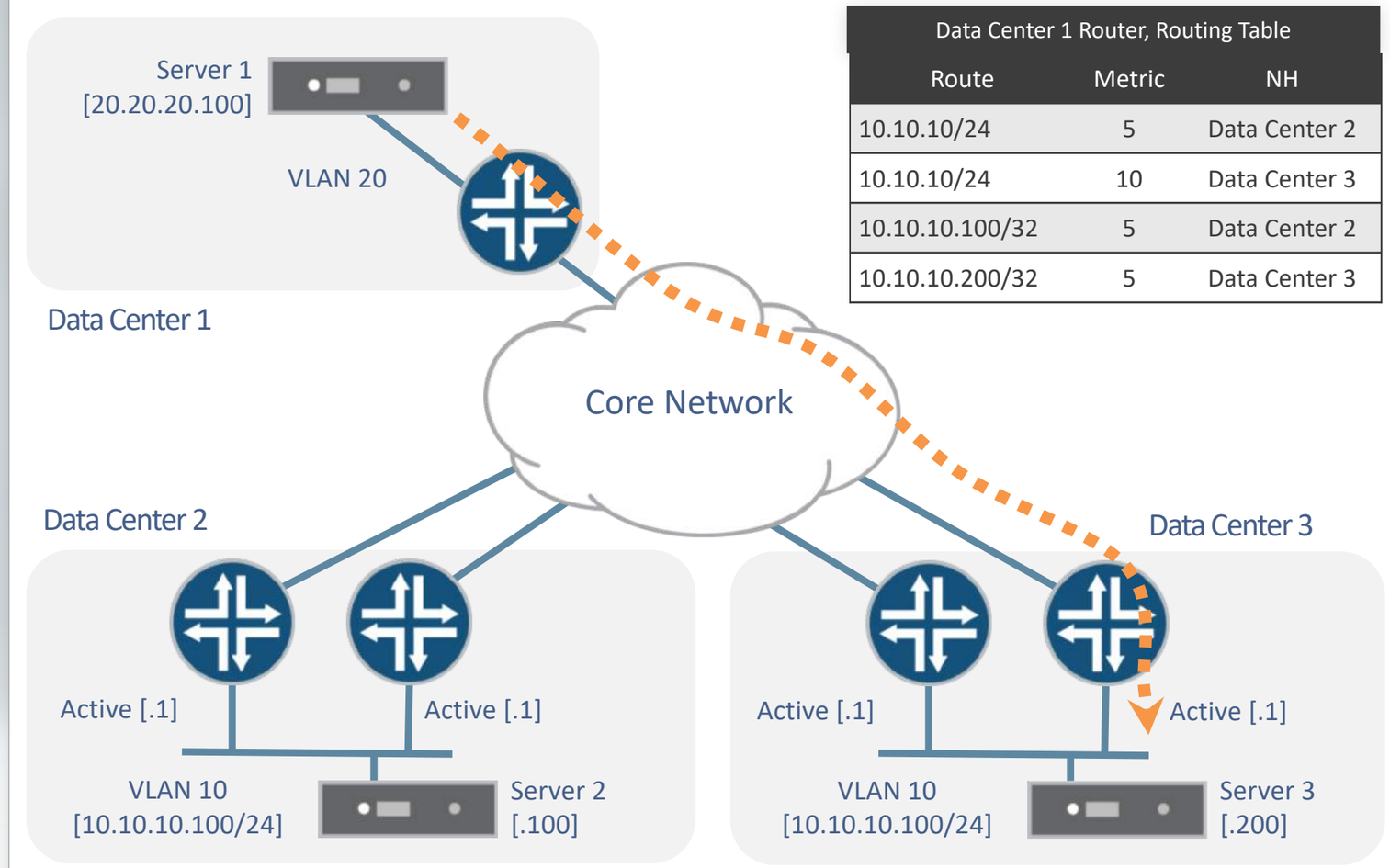
Решение

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



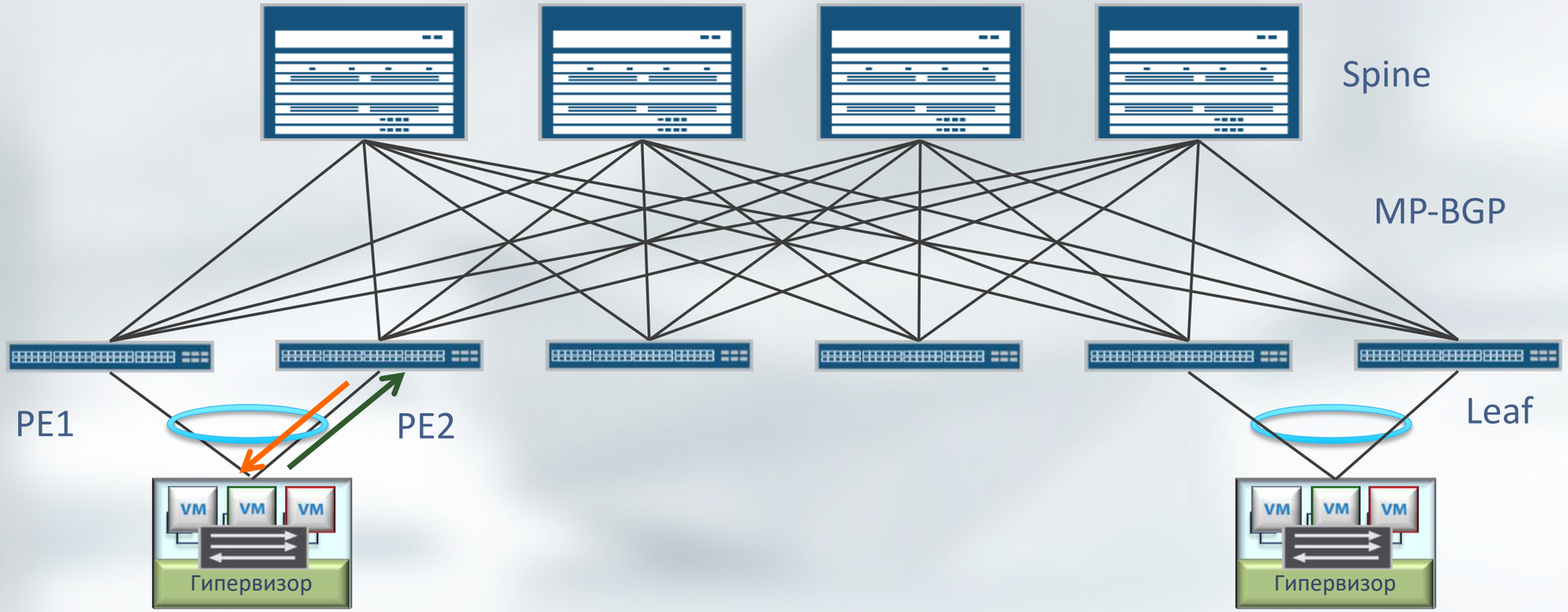
Data Center 1 Router, Routing Table		
Route	Metric	NH
10.10.10/24	5	Data Center 2
10.10.10/24	10	Data Center 3
10.10.10.100/32	5	Data Center 2
10.10.10.200/32	5	Data Center 3

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

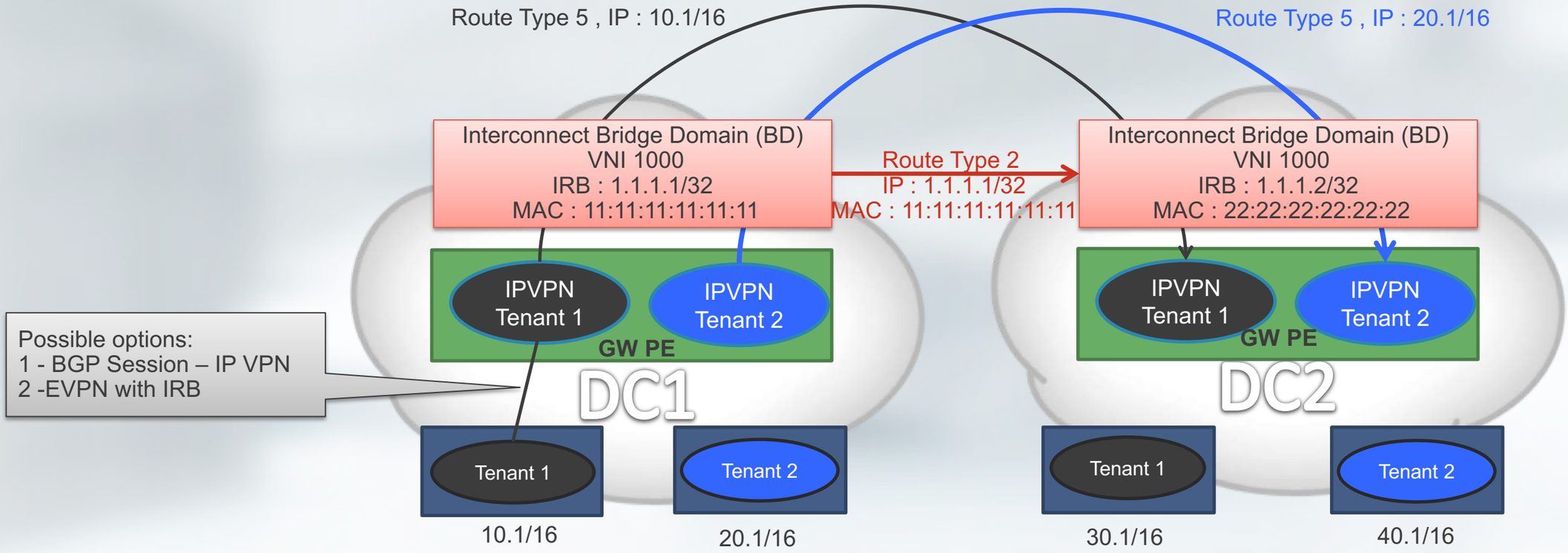
DCI with Route Type 5

- **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	17.4R1
QFX5200	Tomahawk	Hardware limitation	No
QFX10002	Q5	Yes	15.1X53-D30
QFX10008/16	Q5	Yes	16.2
MX Series	Trio	Yes	17.1R1

Спасибо!