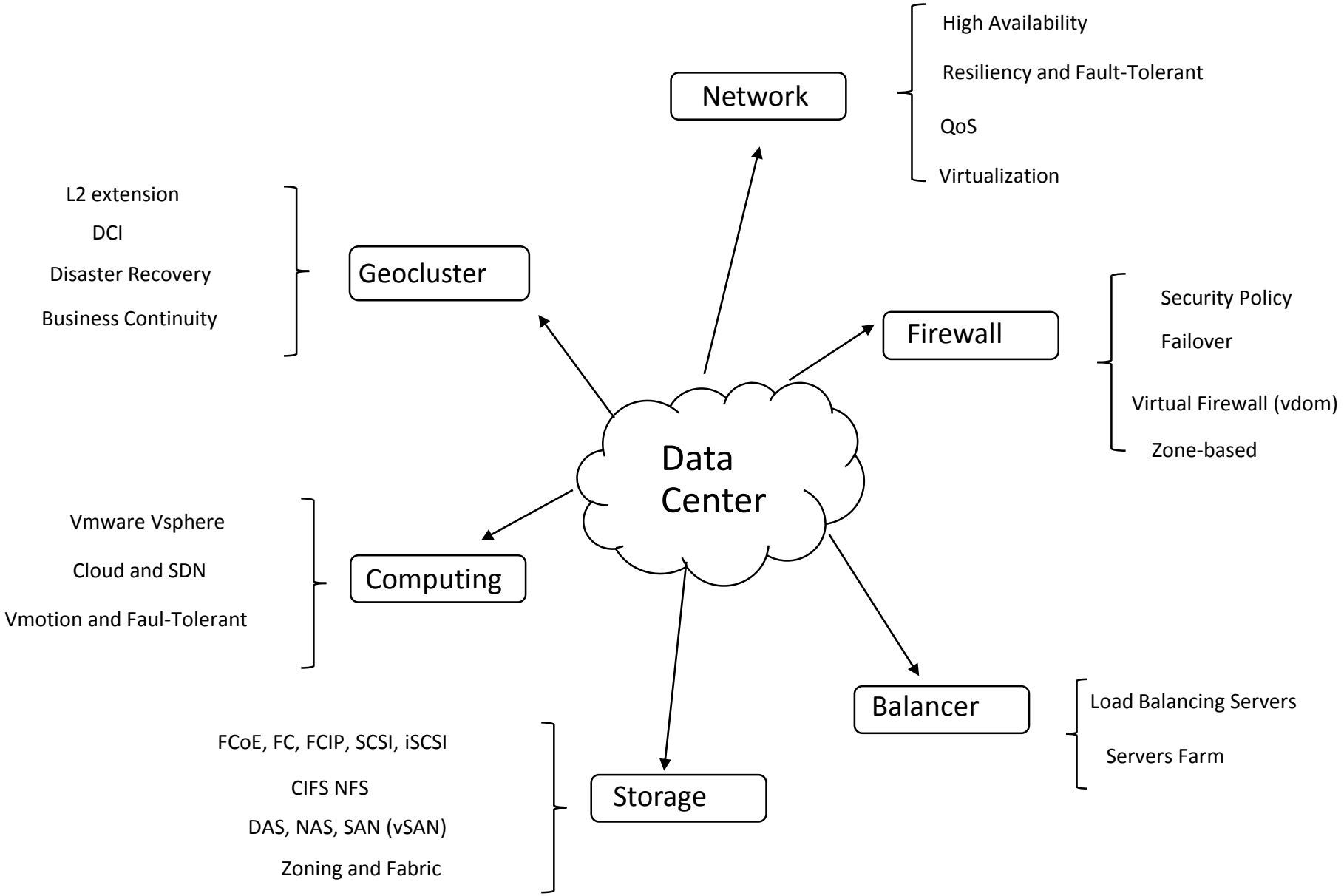
Seminario
ARCHITETTURE DATACENTERS

Massimiliano Sbaraglia

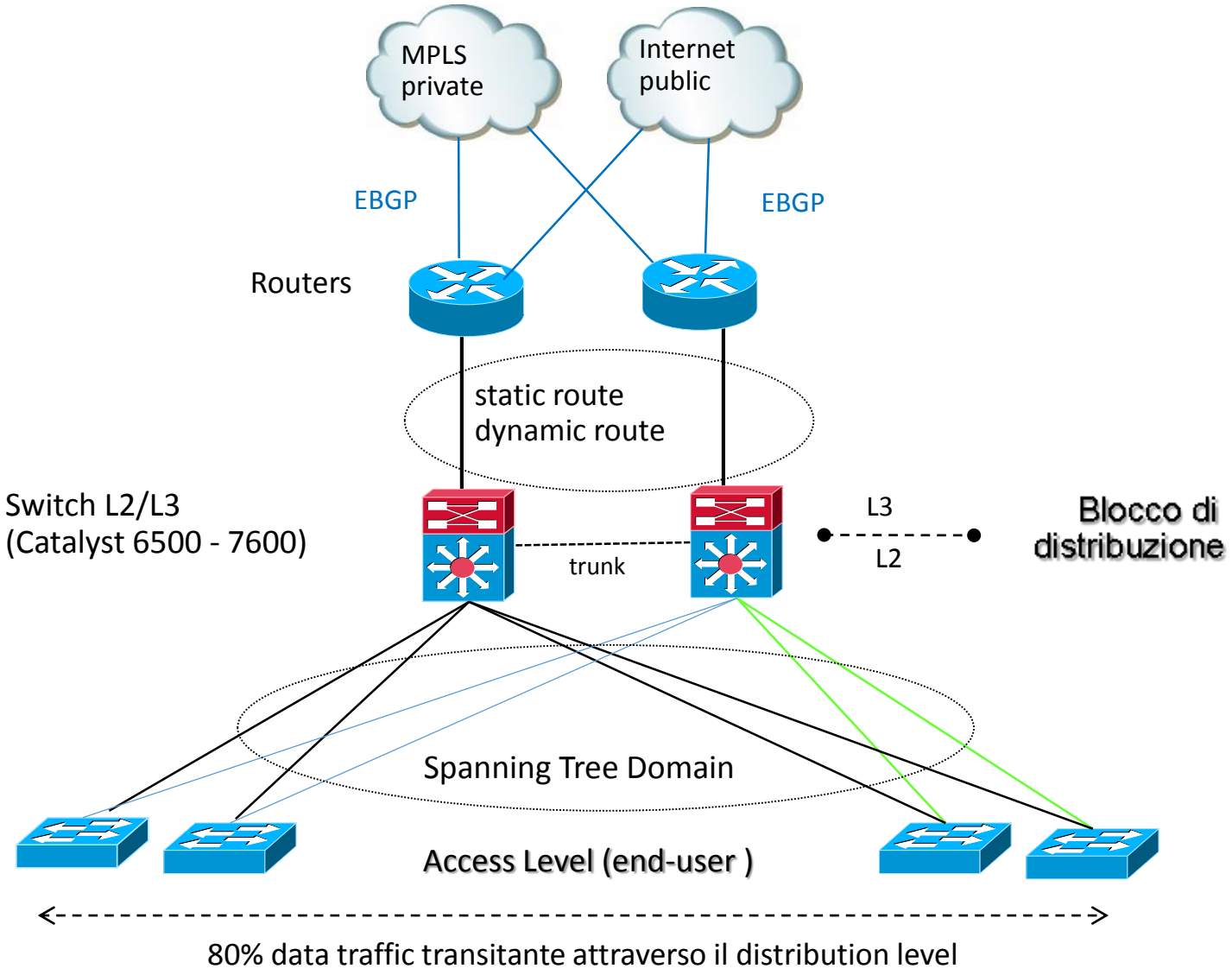
Data Center Concept

- Data Centers Concepts
- Architettura base NON data center
- Architettura base fisica data center
- Architettura base logica active standby data center
- Architettura logica active active data center
- Architettura DCS Nexus con tecniche di virtualizzazione
- Ingegneria principale per un nuovo modello Data Center
- EFP Ethernet Flow Point
- Example Configuration EFP Eth QinQ
- Architettura di un Bridge Domain
- Example Configuration Cisco di un l2vpn bridge domain
- Architettura Bare Metal Data Center Fabric
- Fabric VCS Network Function Virtualization
    - Config example Fabric Pre-Provisioned
- Cluster Firewalls Fortinet
- Architettura AWS Amazon Web Server Cloud data center

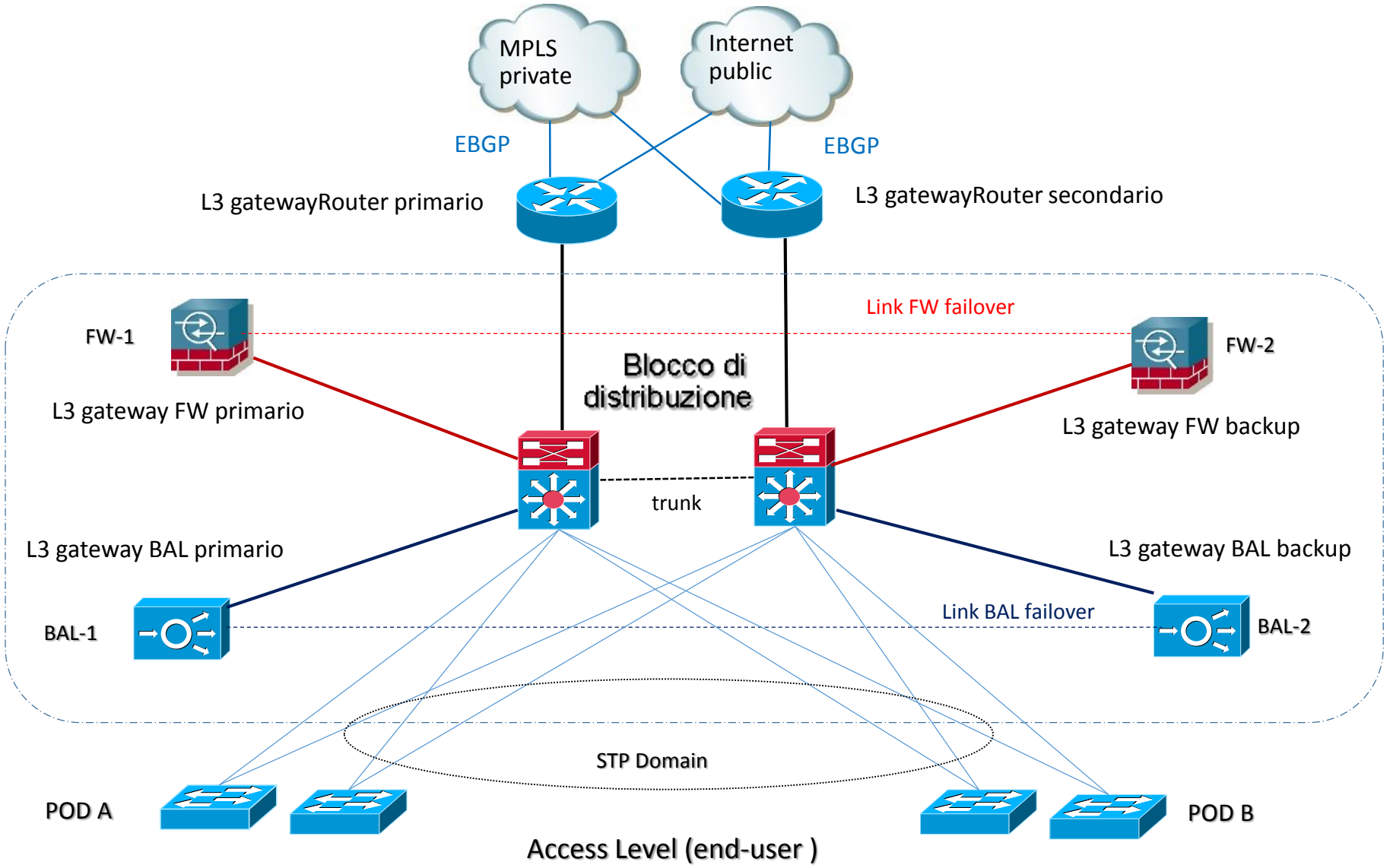
Data Centers Concept



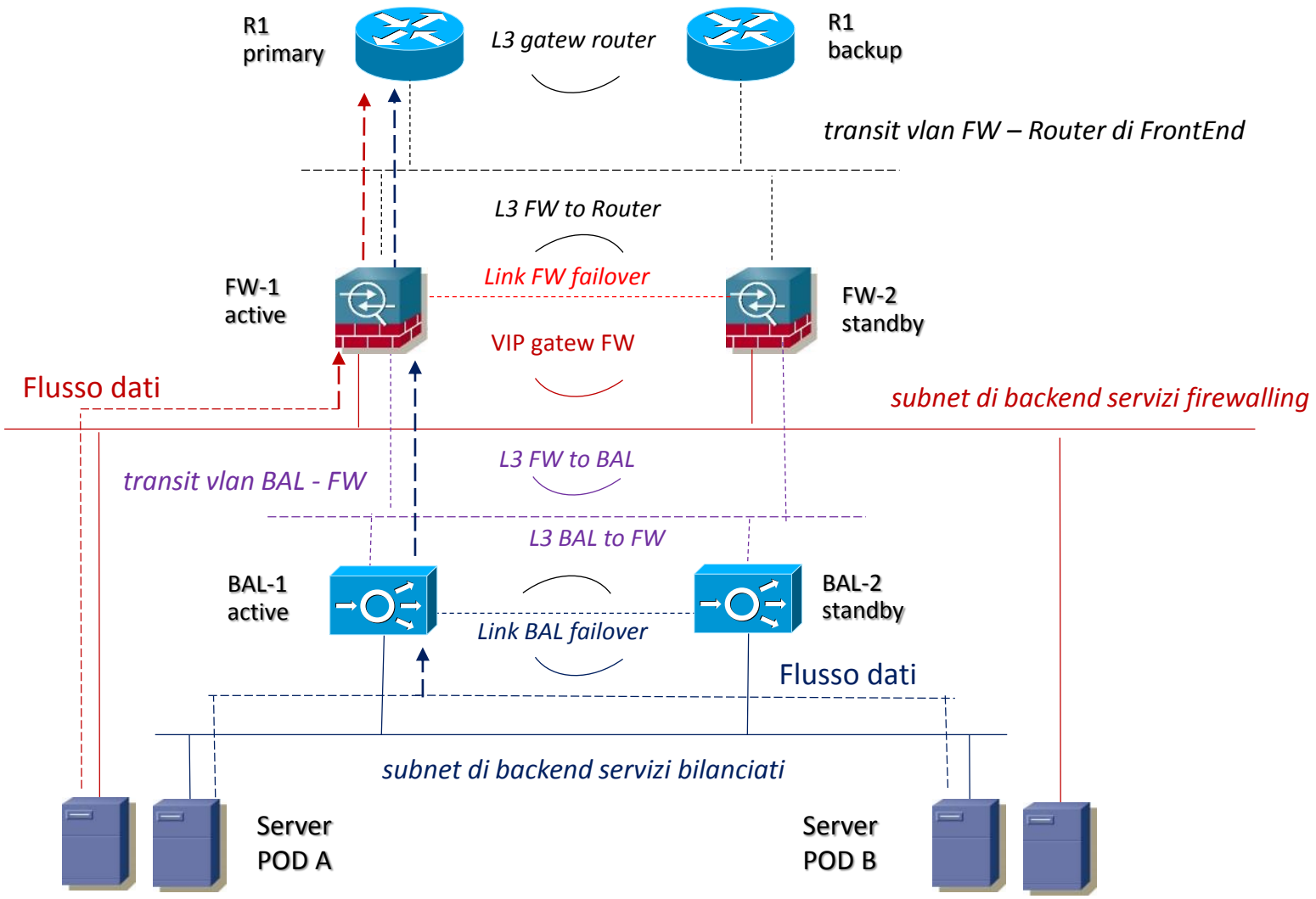
Architettura di base tre livelli non-datacenters



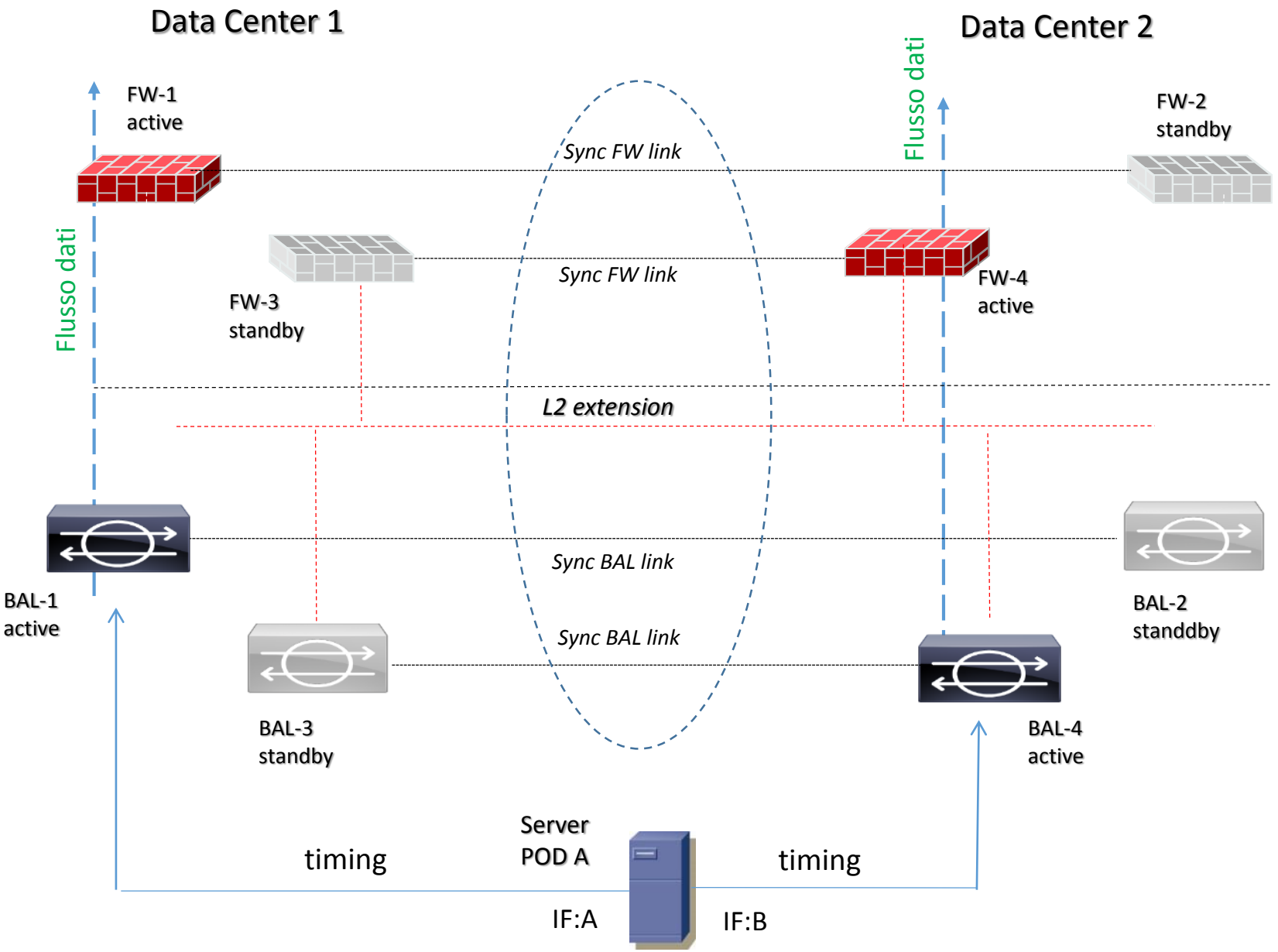
Architettura di base fisica di un datacenters



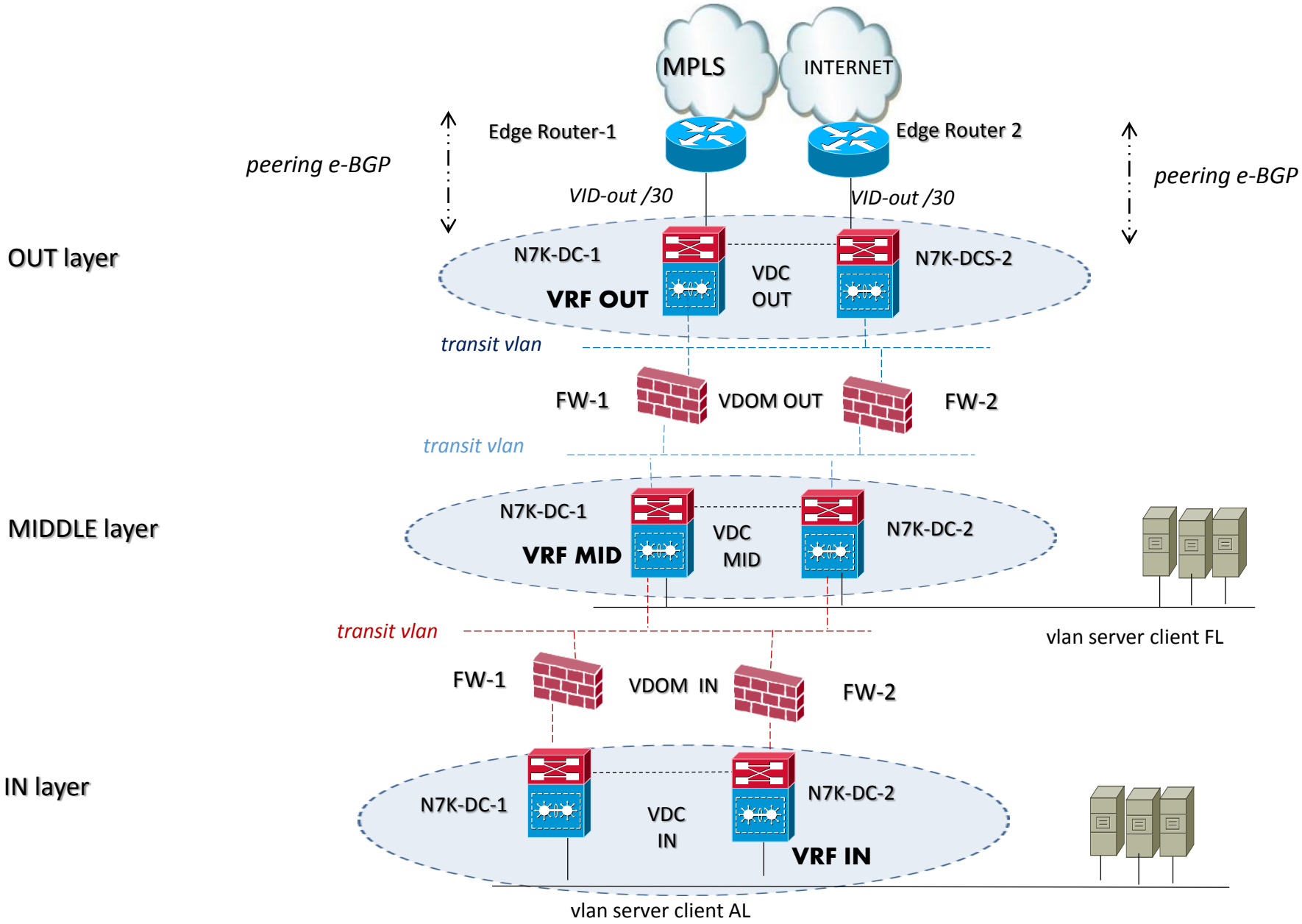
Architettura di base logica di un datacenters (active standby)



Architettura di base logica di un datacenters (active active)



Architettura DCS con Nexus Cisco e tecniche di virtualizzazione



Requisiti principali per un nuovo modello di Data Center

Cloud Computing

- Moving di VM (virtual machine)
- Business Continuity

Consolidamento risorse HW

- Riduzione OpEx (es: riduzione numero apparati)
- Riduzione CapEx (es: risparmio energetico)

Eliminazione tecnologie “legacy”

- Spanning-tree

visibilità network accesso virtuale

- Controllo componente accesso delle VM

SDN

- Definizione di modelli di erogazione standard e API per provisioning automatizzato

Ingegneria principale per un nuovo modello di Data Center

VM Moving

- VPLS
- Implementazione dot1q tunnel (QinQ)

Implementazione accessi virtuali

- Nexus 1000v (VSM, VEM)

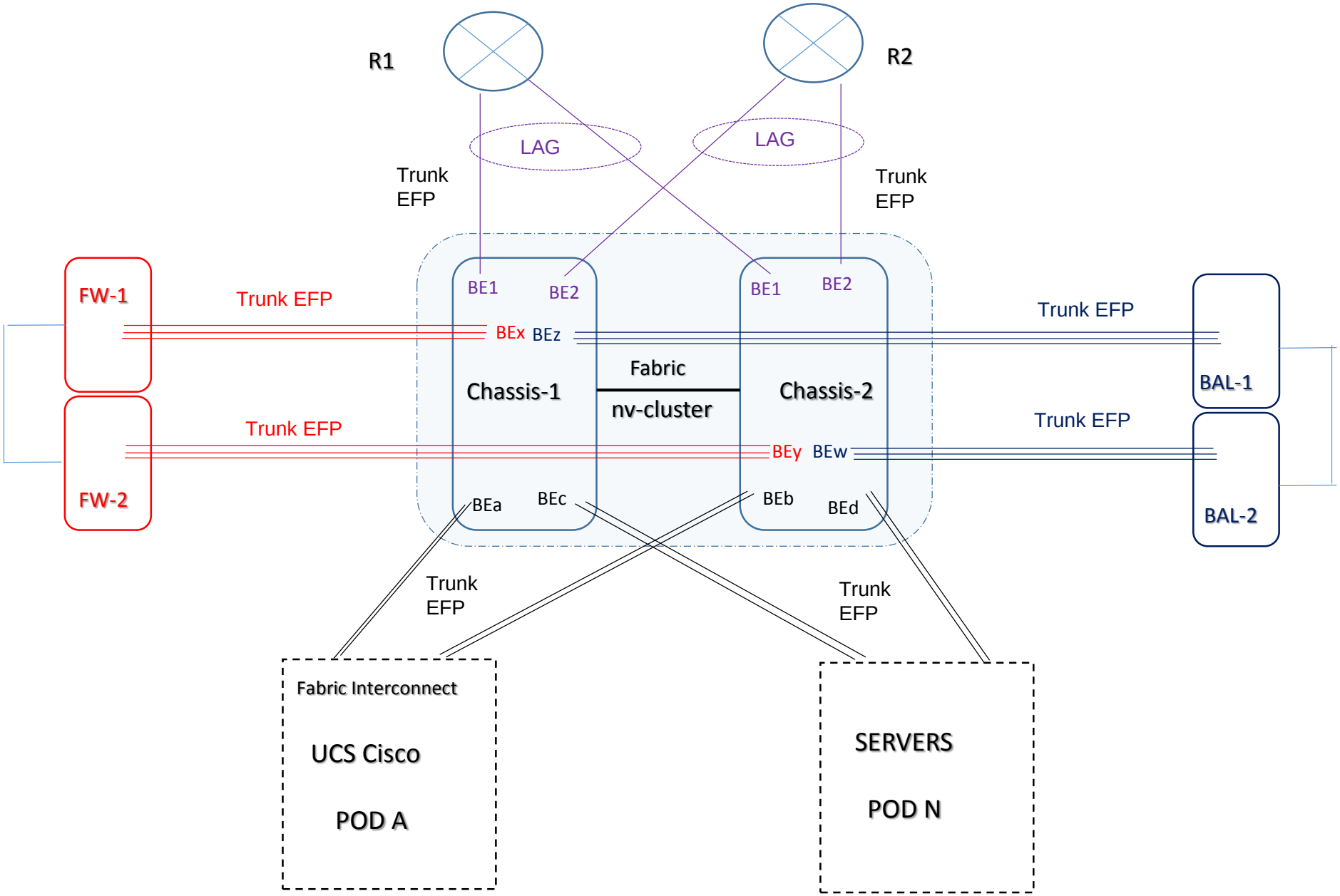
Servizi Applicativi condivisi (SAC)

- NAS
- Repository Linux

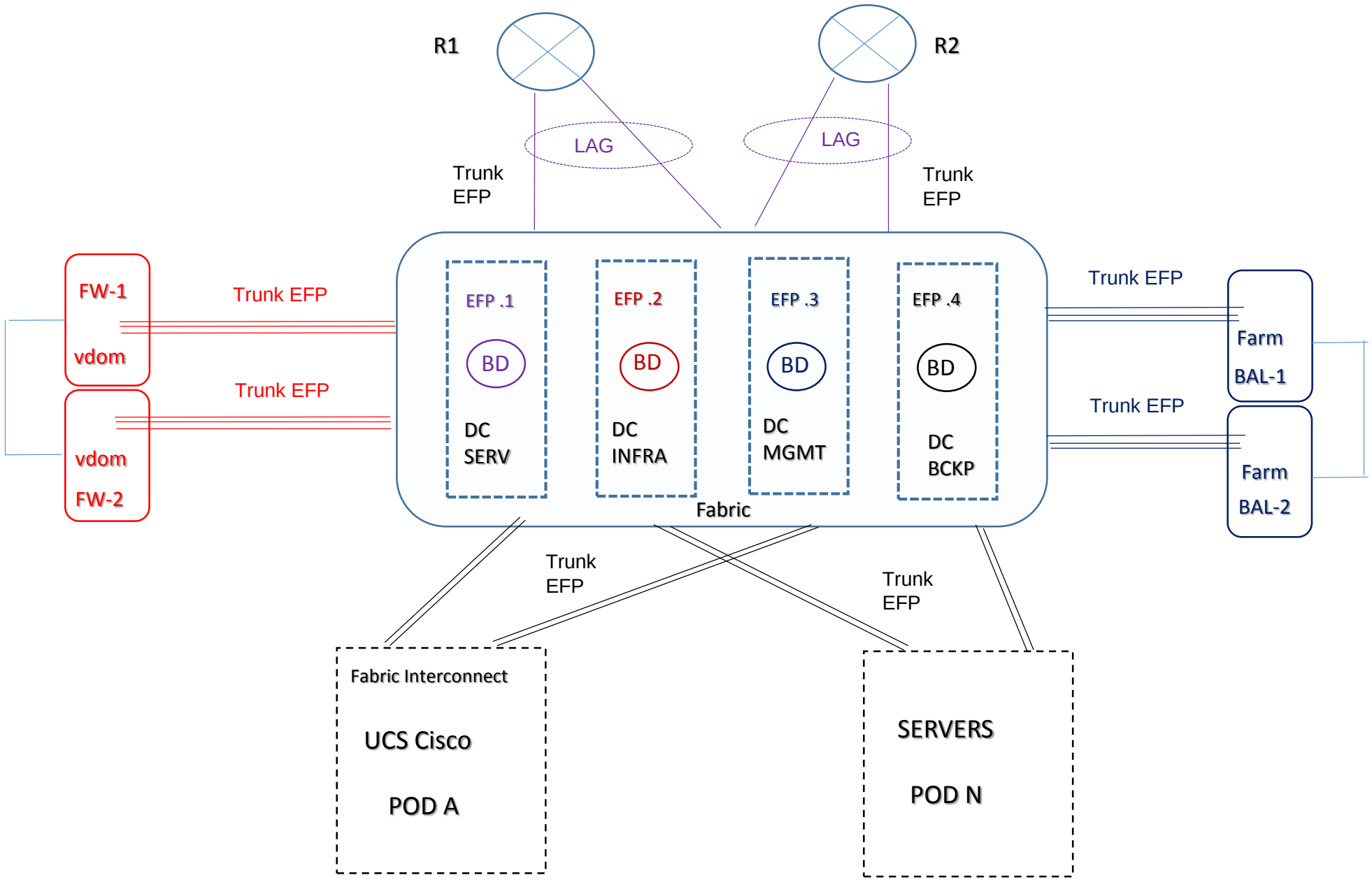
Isolamento domini L2

- Bridge Domain

Ingegneria principale per un nuovo modello di Data Center



Ingegneria principale per un nuovo modello di Data Center



EFP Ethernet Flow Point

Un EFP è una subinterface associata ad un bridge domain

```
interface TenGigEx/y/z/w
description description "to  FW-DC-Backup"
bundle id 4 mode active
!
 interface Bundle-Ether9.4 l2transport
 description FW-BKP_DC-BACKUP
 encapsulation dot1q < range_vlan >
```

L’EFP da sola non è in grado di prendere decisioni di forwarding L2, deve essere necessariamente associata ad un bridge-domain:

```
l2vpn
  bridge group DC
   bridge-domain DC-BKP
   interface Bundle-Ether9.4
!
```

EFP Ethernet Flow Point

Le vlan sono segmentate utilizzando 1 EFP per ogni LAG (Bundle-Ether) che è stata mappata sul BD di riferimento

```
interface Bundle-Etherx.1 l2transport
 description LB_to_DC-SERVER
 encapsulation dot1q < range-vid >
!
interface Bundle-Ethery.1 l2transport
 description LB_to_DC-SERVER
 encapsulation dot1q < range-vid>
!
L2vpn
bridge group DC
bridge-domain DC-SERVER
  interface Bundle-Etherx.1
  interface Bundle-Ethery.1
```

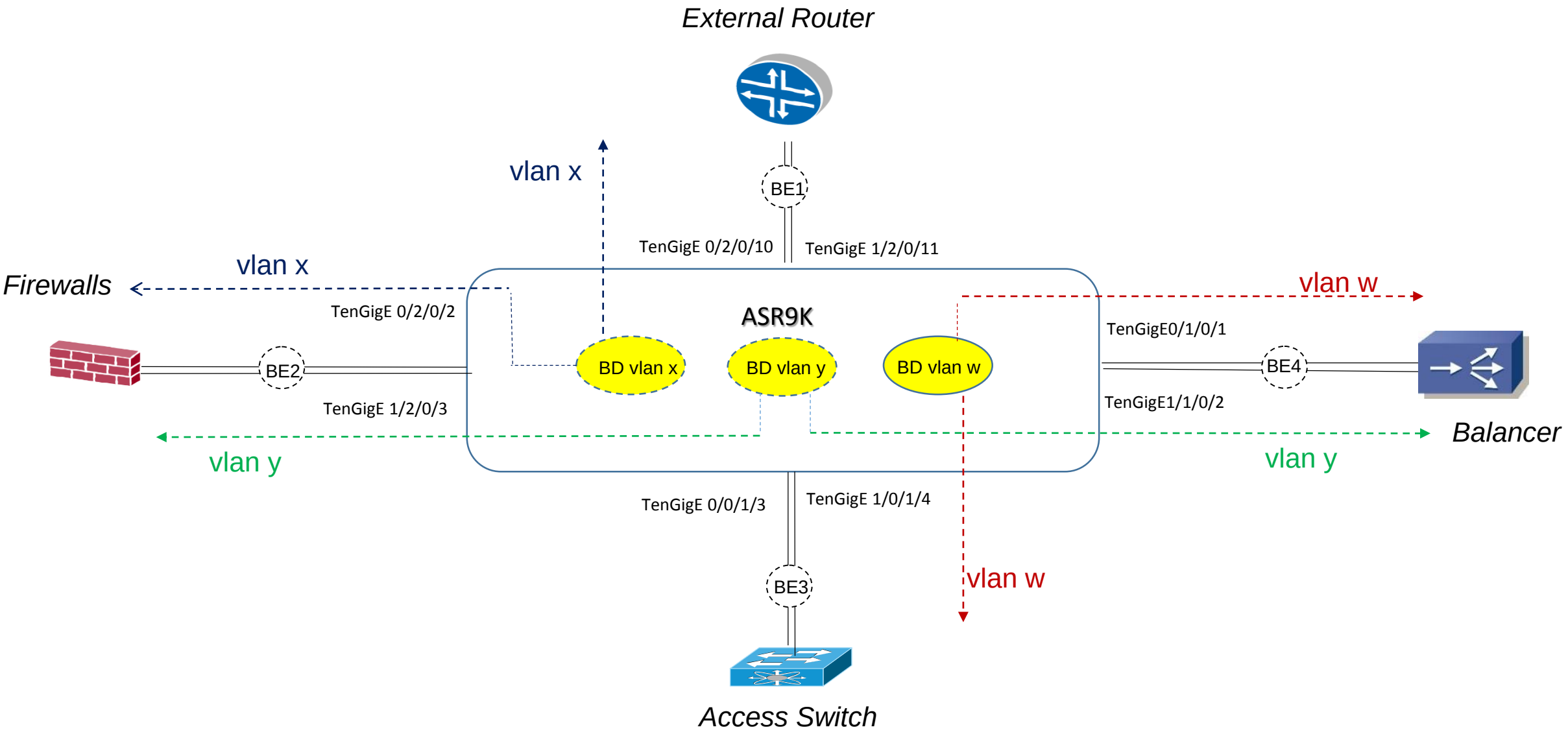
Example configurazione EFP Ethernet QinQ

```
interface Bundle-Etherx.1 l2transport
 description R1_to-DC-SERVER
 encapsulation dot1q <outer-vid> second-dot1q <range-inner-vid>
 rewrite ingress tag pop 1 symmetric
 ethernet-services access-group permit-qinq-vid-based ingress
 ethernet-services access-group permit-qinq-vid-based egress
!
ethernet-services access-list permit-qinq-vid-based
 1 remark “qinq-permit”
 a permit any any vlan x inner-vlan a
 b permit any any vlan x inner-vlan b
 c permit any any vlan x inner-vlan c
 d permit any any vlan x inner-vlan d
 z deny any any
```

rewrite ingress tag pop 1 symmetric: Effettua il pop (eliminazione) dell’outer tag prima dell’ingresso al Bridge Domain associato alla EFP oppure aggiunge l’outer tag configurato nell’encapsulation all’uscita del Bridge Domain associato alla EFP

x: outer tag
a, b, c, d: inner tag

Architettura Bridge Domain



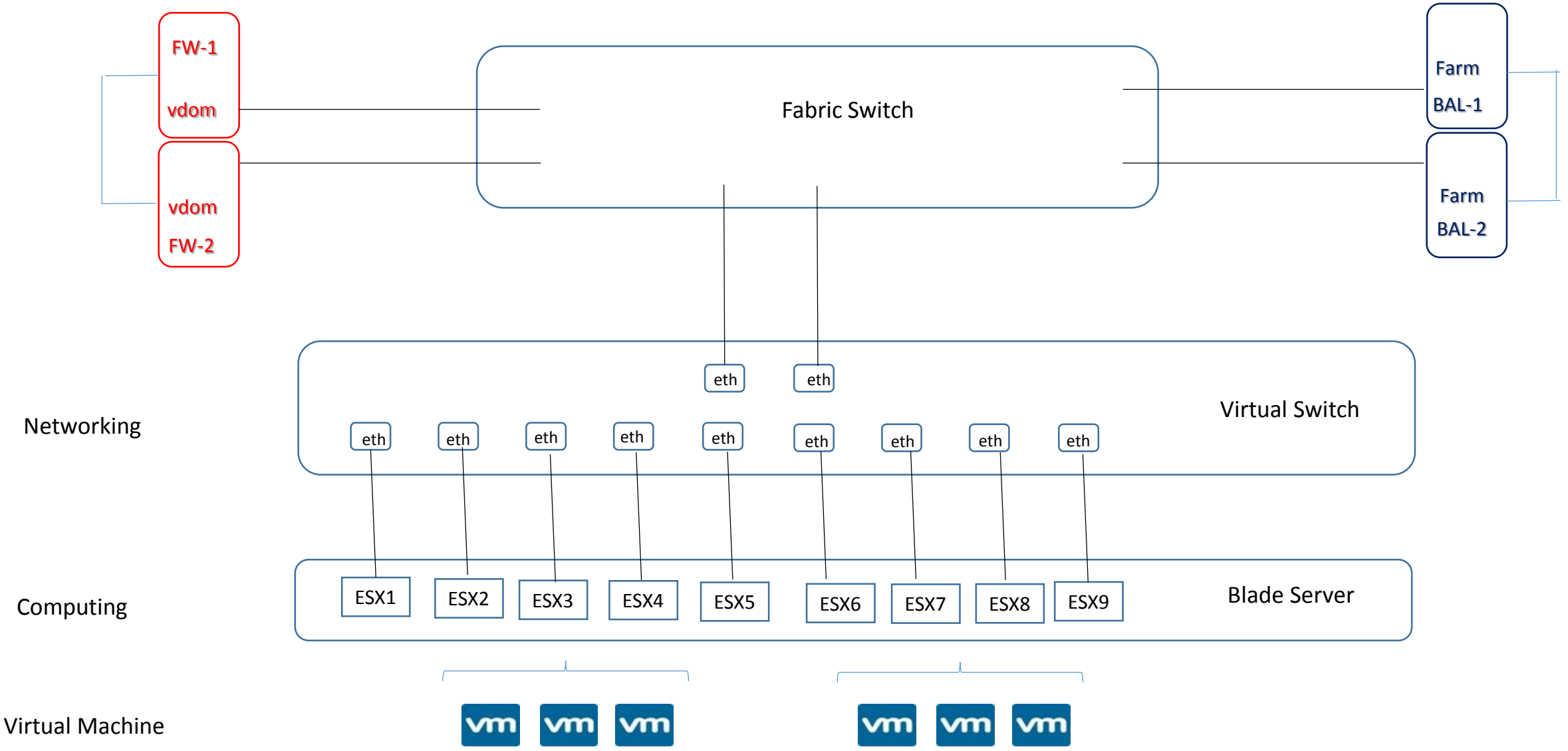
Architettura Bridge Domain configuration cisco example

```
interface TenGigE0/2/0/10
 description description "to MX960 external router"
 bundle id 1 mode active
!
interface TenGigE1/2/0/11
 description description "to MX960 external router"
 bundle id 1 mode active
 !
interface TenGigE0/2/0/2
 description description "to FW security"
 bundle id 2 mode active
 !
interface TenGigE1/2/0/3
 description description "to FW security"
 bundle id 2 mode active
```

```
interface Bundle-Ether1.x l2transport
 description MX960_to_BRIDGE_DOMAIN-VLAN-X
 encapsulation dot1q x
!
interface Bundle-Ether2.x l2transport
 description FW_to_BRIDGE_DOMAIN-VLAN-X
 encapsulation dot1q x
!
interface Bundle-Ether2.y l2transport
 description FW_to_BRIDGE_DOMAIN-VLAN-Y
 encapsulation dot1q y
!
interface Bundle-Ether4.y l2transport
 description BAL_to_BRIDGE_DOMAIN-VLAN-Y
 encapsulation dot1q y
!
```

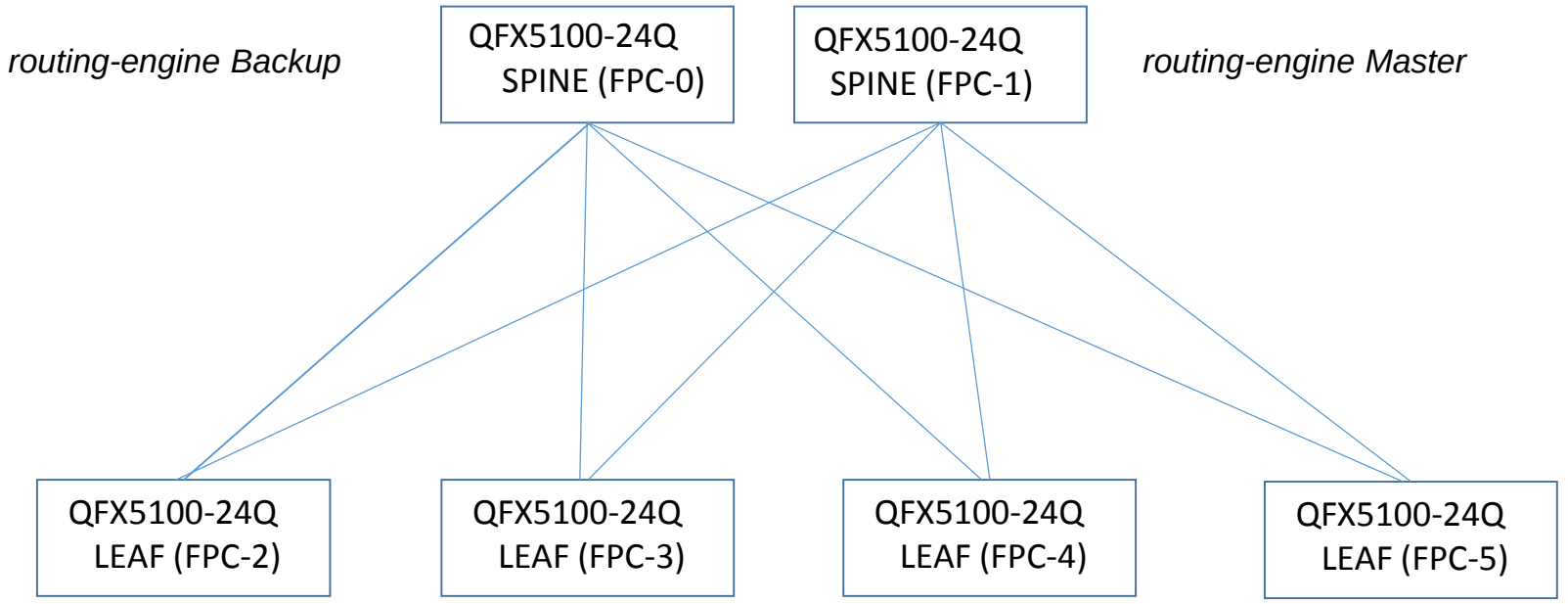
```
l2vpn
 bridge group CLIENTE
!
  bridge-domain BRIDGE-DOMAIN-VLAN-X
   interface Bundle-Ether1.x
   interface Bundle-Ether2.x
  !
  bridge-domain BRIDGE-DOMAIN-VLAN-Y
   interface Bundle-Ether2.y
   interface Bundle-Ether4.y
  !
 bridge-domain BRIDGE-DOMAIN-VLAN-W
  interface Bundle-Ether3.w
  interface Bundle-Ether4.w
!
```

Architettura Bare-Metal Data Centers Fabric



Fabric Network Function Virtualization

La Fabric opera in modalità VCF (Virtual Chassis Fabric) in cui tutti gli switch della Fabric si aggregano a formare logicamente un unico switch L2/L3 nel contesto del quale i due apparati di Spine assolvono il ruolo di routing engine (active/standby) e i nodi Leaf operano concettualmente come line-card.
La Fabric consente l'aggregazione di più porte fisiche, anche di switch differenti, in gruppi LACP. Ciò a scopo di distribuzione del traffico su più interfacce e di alta affidabilità ai guasti.



Fabric Network Function Virtualization

E’ un’architettura CLOS (Spine and Leaf) dove le principali features sono:

- Fabric multi-path: il piano di forwarding di un pacchetto tra i nodi è regolato dal protocollo SPF (Shortest Path First);
- Intelligent Bandwidth Allocaton: il nodo trasmittente considera la quantità di banda disponibile per ogni multi-path tra un nodo e l’altro, allocando le risorse di rete end-to-end;
- Bidirectional MDT (Multicast Distribution Tree): VFC calcola multipli alberi (tree) multicast in modo bidirezionale e performa load-balancing in questi percorsi;
- L2 and L3 capability: in base alla licenza adottata, possiamo avere funzionalità L2 attraverso L3 capability IPv4 e IPv6 (oltre MPLS, BGP, ISIS), inoltre supporta funzionalità quali FCoE, VXLAN, NVGRE, VMware integration;
- Resiliency and High Availability: include redundant routing engine in modalita active-backup, redundant data-plane con modalità active-active uplinks;
- NSSU (No Stop Software Upgrade): disponibile per VFC con doppio RE (Routing Engine) e consente aggiornamenti software senza distruzioni o interruzioni di funzionalità.

Fabric Network Function Virtualization

Sono possibili due configurazioni VCF:

- Preprovisioned: con il controllo di ciascun nodo assegnando un member-ID ed il ruolo a lui assegnato;
- Non-provisioned: è il nodo master che assegna un member-ID a ciascun nodo; il ruolo è determinato dal valore di priority masterschip ed altri fattori che concorrono alla elezione del master;
- Master Routing Engine: il nodo master RE controlla tutta la Fabric VCF
- Backup Routing Engine: il nodo di backup RE resta in standby mode con un kernel (cuore del sistema) e lo stato dei protocolli in uso sincronizzato rispetto al nodo master;
- Line-card: a parte I nodi master e backup, tutti gli altri nodi della VFC hanno ruolo di line-card.

Fabric Network Function Virtualization configuration example

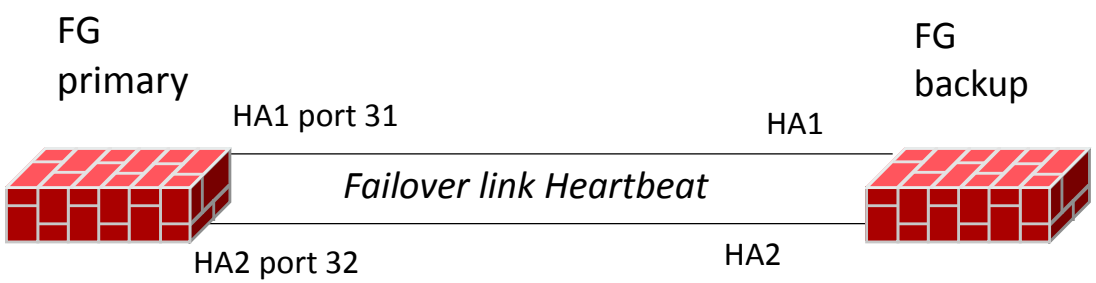
```
virtual-chassis { preprovisioned;
no-split-detection;

member 0 {
 role routing-engine;
 serial-number AAAAAA111111; }

member 1 {
 role routing-engine;
 serial-number BBBBBB222222; }

member 2 {
role line-card;
serial-number CCCCCC333333; }

member 3 {
role line-card;
serial-number DDDDDD444444; }

member 4 {
role line-card;
serial-number EEEEEE555555; }

member 5 {
role line-card;
serial-number FFFFFF666666; }
```

Cluster Firewalls Fortinet

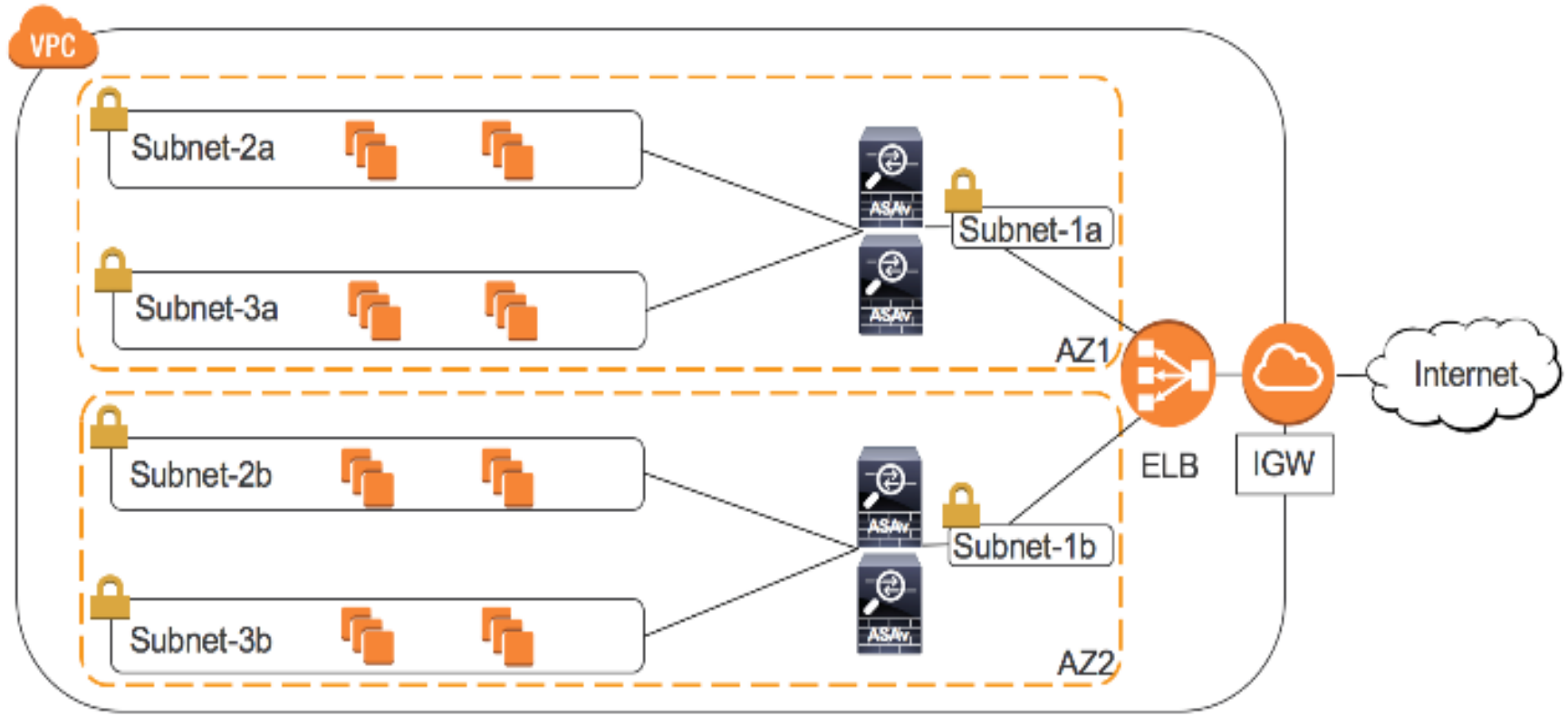
Failover Link Active Passive HA: consiste in una coppia di FW in cluster dove una unità ha ruolo di master e processa tutte le sessioni di comunicazione mentre l'altro FW è in modo subordinato (condizione di standby) non processando nessuna sessione di comunicazione ma mantenendo sincronizzata la configurazione e monitorando lo status del FW primario

Failover Link Active Active HA: consiste in una coppia di FW in cluster dove attraverso un meccanismo di load-balancing (proxy based (cpu intensive)) entrambi i FW processano le sessioni di comunicazione

Virtual clustering per operare con multipli VDOM; possiamo considerare una unit FW primario per alcuni vdom e l'altro unit FW primario per altri vdom



Architettura AWS Amazon Web Server Cloud datacenters



Architettura AWS Amazon Web Server Cloud datacenters

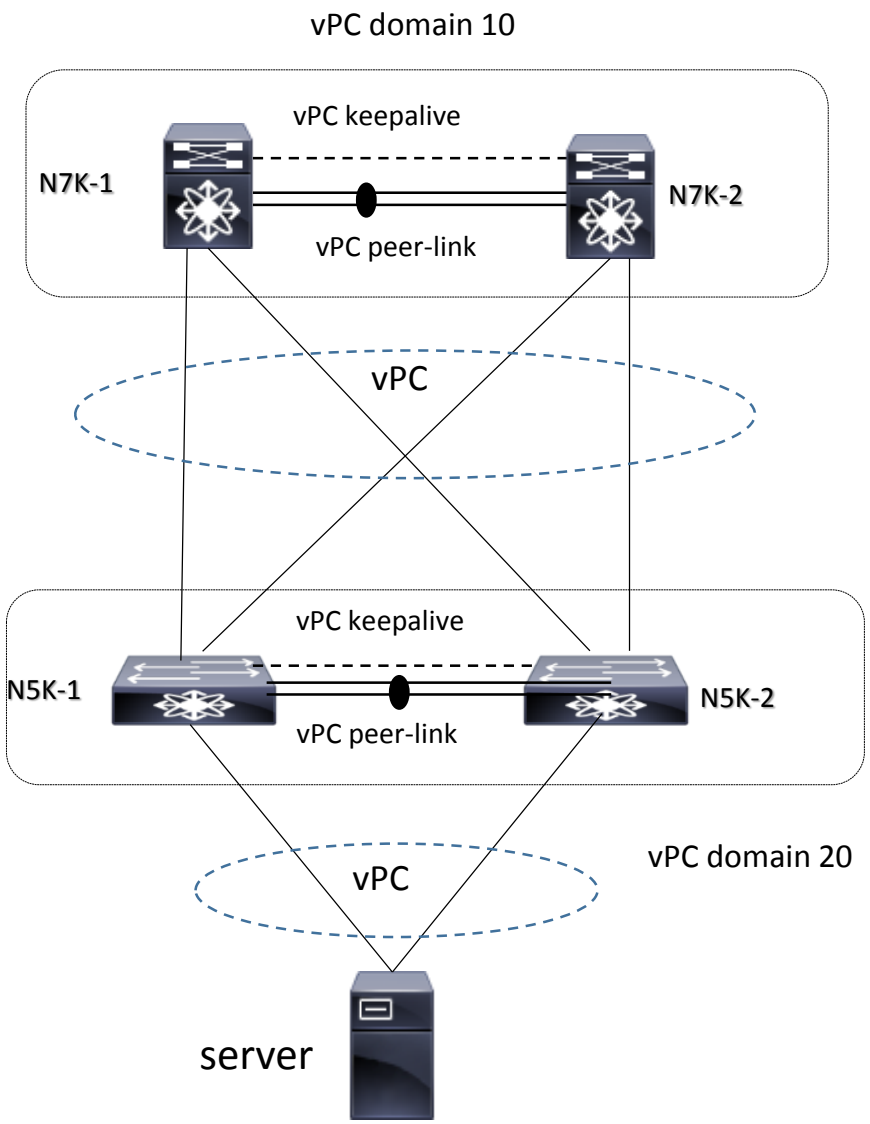
Dal punto di vista Network, bisogna avere familiarità per questa NGVDC (New Generation Virtual Data Center) con queste tematiche:

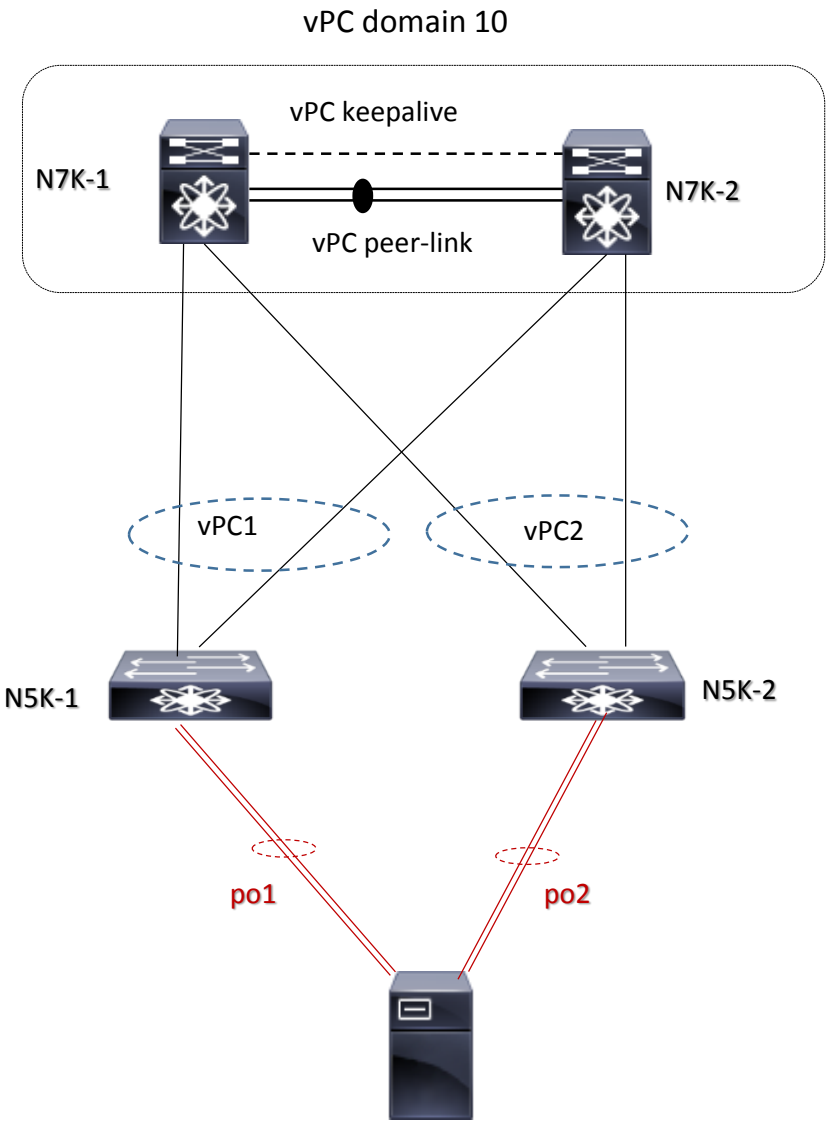
- VPC è un ambiente Private Virtual Cloud interconnesso con altri VPC;
- EC2 Elastic Cloud Computing;
- ELB Elastic Load Balancer;
- Route Table, Subnets, Elastic IP;
- NGFWv elementi virtuali di firewalling;
- AZ per definire zone disponibile oppure no

Architetture Nexus Cisco

- Architettura data centers Nexus Cisco design example
- Architettura data center Nexus Cisco terminologia
- vPC concept
    - vPC peer-keepalive and peer-link config
    - vPC spanning tree vlan-id
    - vPC and STP con peer-switch feature
    - vPC and FHRP con peer-gateway feature
    - vPC broadcast and loop-avoidance feature
    - vPC+ (vpc plus) concepts
- VDC concept
- POD FEX 1xN2K active standby dual-homed
- POD FEX 1xN2K active active dual-homed
- POD FEX 2xN2K con enhanced vPC
- POD FEX 2xN2K straight-through

Architettura Data Centers con Nexus design example





Architettura Data Centers con Nexus terminologia

Termine	Definizione
vPC	è un port-channel tra due vPC peers ed un downstream switch
vPC peer device	è uno dei due vPC peer device (esempio un Nexus 7000)
vPC domain	una coppia di vPC peer rappresenta un dominio vPC
vPC peer-link	è un link utilizzato per sincronizzare gli stati tra i due vPC peers (è buona norma utilizzare un link a 10G)
vPC peer-keepalive	è un link utilizzato (differente rispetto al vpc peer-link) per verificare e monitorare lo stato di vita tra i due peer devices
vPC member port	una o più porte che fanno parte del port-channel a formare un vPC
vPC LAN	sono vlans trasportate via vpc peer-link tra i due peer-devices e verso il downstream switch via vPC
non-vPC LAN	viceversa è una vlan che non transita per vpc peer-link tra i due vpc peers devices e non fa parte di nessun port-channel in vPC
Orphan Port	sono porte collegate a terze parti switch non facenti parte di vPC trunks
CFS (Cisco Fabric Services)	è un protocollo che opera attraverso il vpc peer-link per rendere affidabile la sincronizzazione tra i due vpc peer devices

vPC concepts

- elimina SPT blocked port;
- utilizza tutti I link disponibili e relativa bandwidth;
- dual-homed servers in active-active mode;
- fast-convergence in caso di fault link or switch;
- split-horizon loop via port-channeling (traffico entrante in un port-channel non può uscire dallo stesso port-channel);
- un vPC domain è costituito da due peers, ognuno dei quali lavora con il proprio control-plane;
- vPC significa un collegamento in port-channel tra due vPC peers ed un devices in downstream;
- vPC domain è costruito attraverso la configurazione di un peer-keealive (per monitorare la condizione dei due peer) ed un peer-link (per la sincrozizzazione degli stati dei due peer);
- HA, link-level resiliency

vPC concepts

Con vPC, ogni Nexus mantiene il proprio control-plane ed il proprio management:

- vPC IEEE 802.3ad: è un port-channel tra un devices in downstream e due Nexus Devices con stessa release software
- vPC peer: è uno dei due devices (or VDC) che formano la coppia di Nexus in aggregazione
- vPC member port: è una interfaccia che appartiene ad uno specific vPC port-channel di uno dei due vPC peers
- vPC domain: un unico identificativo per coppia di Nexus (ogni Nexus switch or VDC supporta un solo dominio)
- vPC peer-link: usato per la sincronizzazione dei rispettivi status switches e per il forward del traffic o tra i due peers
- vPC peer-keepalive: usato per la verifica heartbeat tra I due switches vPC peers (in modo esplicito per verificare failure dei peer-link e peer-keepalive)
- CSF (Cisco Fabric Services): è automaticamente abilitato in un vPC e rappresenta la capacità di sincrinizzare lo stato e la configurazione tra i due vPC peers

NOTA CISCO: quando si configura un vPC domain tra i due vPC peers, questi generano un condiviso MAC address che viene usato come un logico switch bridge ID in SpanningTree Protocol.

All'interno di un vPC domain ad ogni peer è assegnato un ruolo: primario e secondario (di default lo switch con il più basso MAC address diventa il primario; è un valore comunque che l'operatore può cambiare)

La comunicazione per il control-plane avviene attraverso il peer-link tra i due peers.

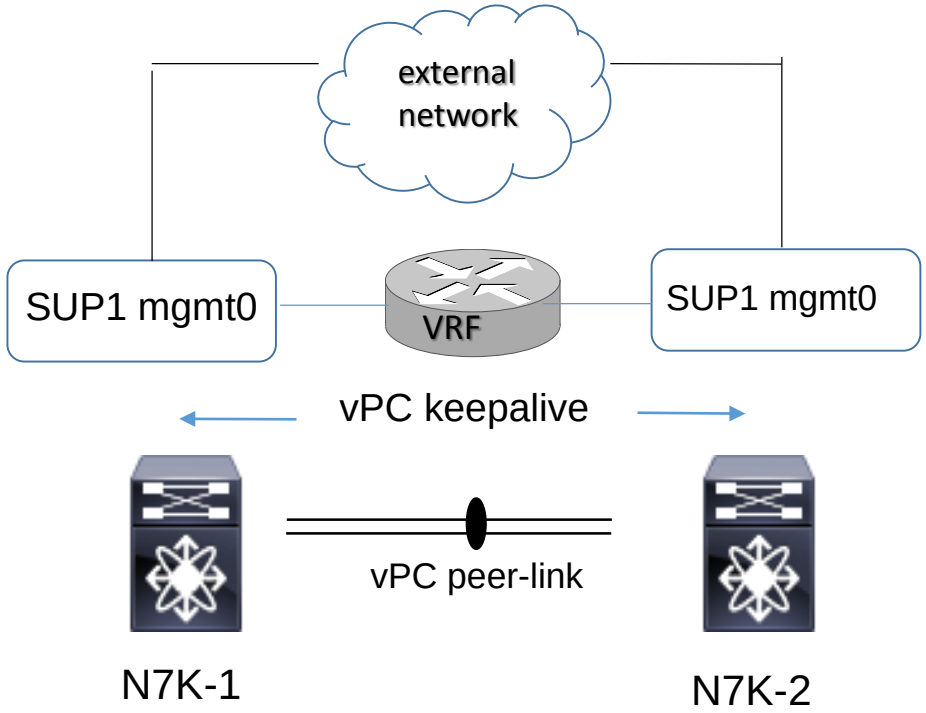
Esiste un meccanismo di loop-avoidance split-horizon loop via port-channeling dove il traffico entrante in un port-channel non può uscire dallo stesso port-channel)

vPC peer-keepalive vrf management

Non usare un cavo diretto tra mgmt0 per evitare perdità di connettività tra i due Nexus SUP; poiché solo la active SUP è l'unica a trasmettere heartbeat, una diretta connessione tra un active ed un standby SUP può risultare come un falso chassis state detection

Usare la mgmt0 interface è la routing-instances di management VRF per la configurazione keepalive (se non è possibile usare una routed interface che appartiene ad una esclusiva VRF)

Usare sempre link differenti tra vpc peer-link e keepalive



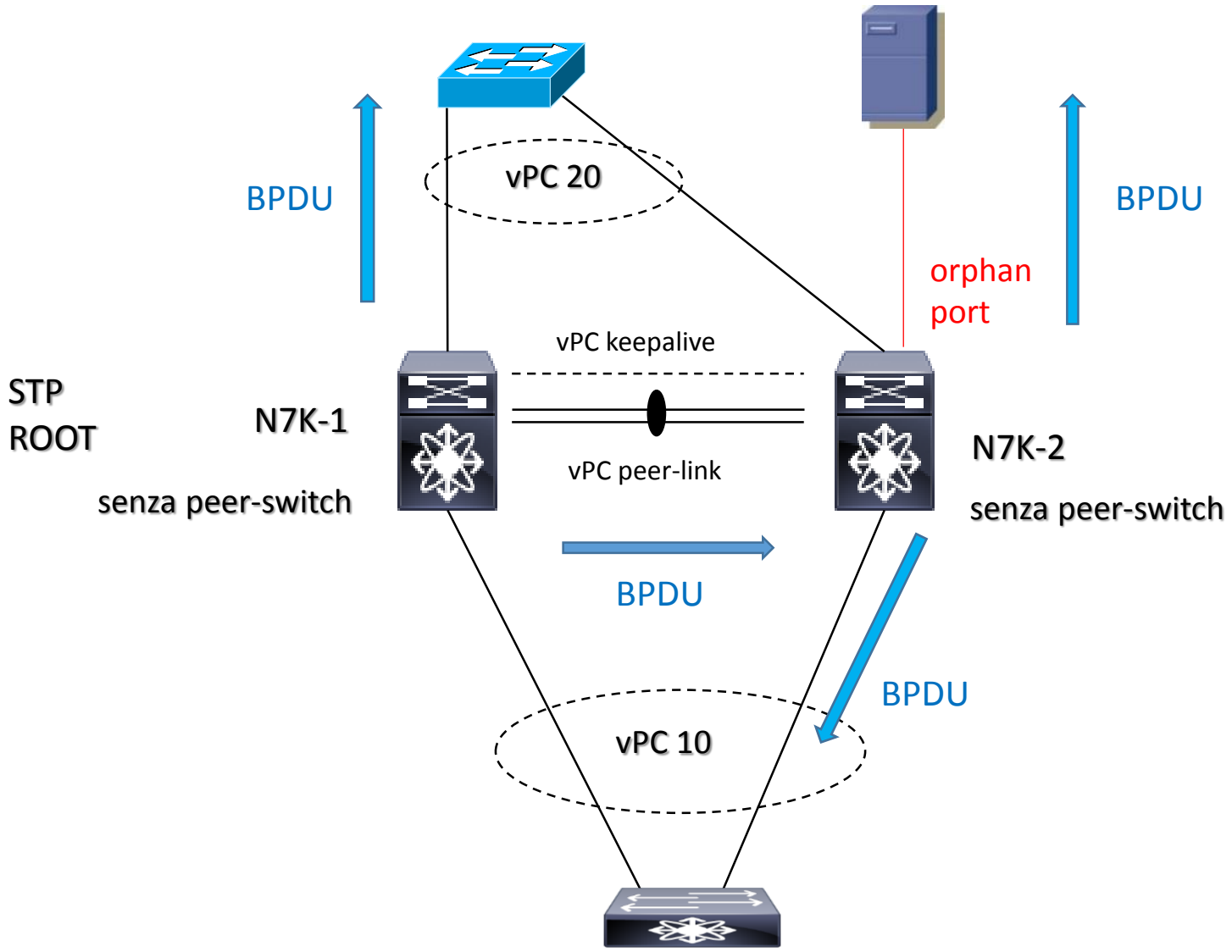
vPC Architectures and STP Spanning Tree Protocol with peer-switch feature

Con vPC NON si elimina lo Spanning Tree Protocol

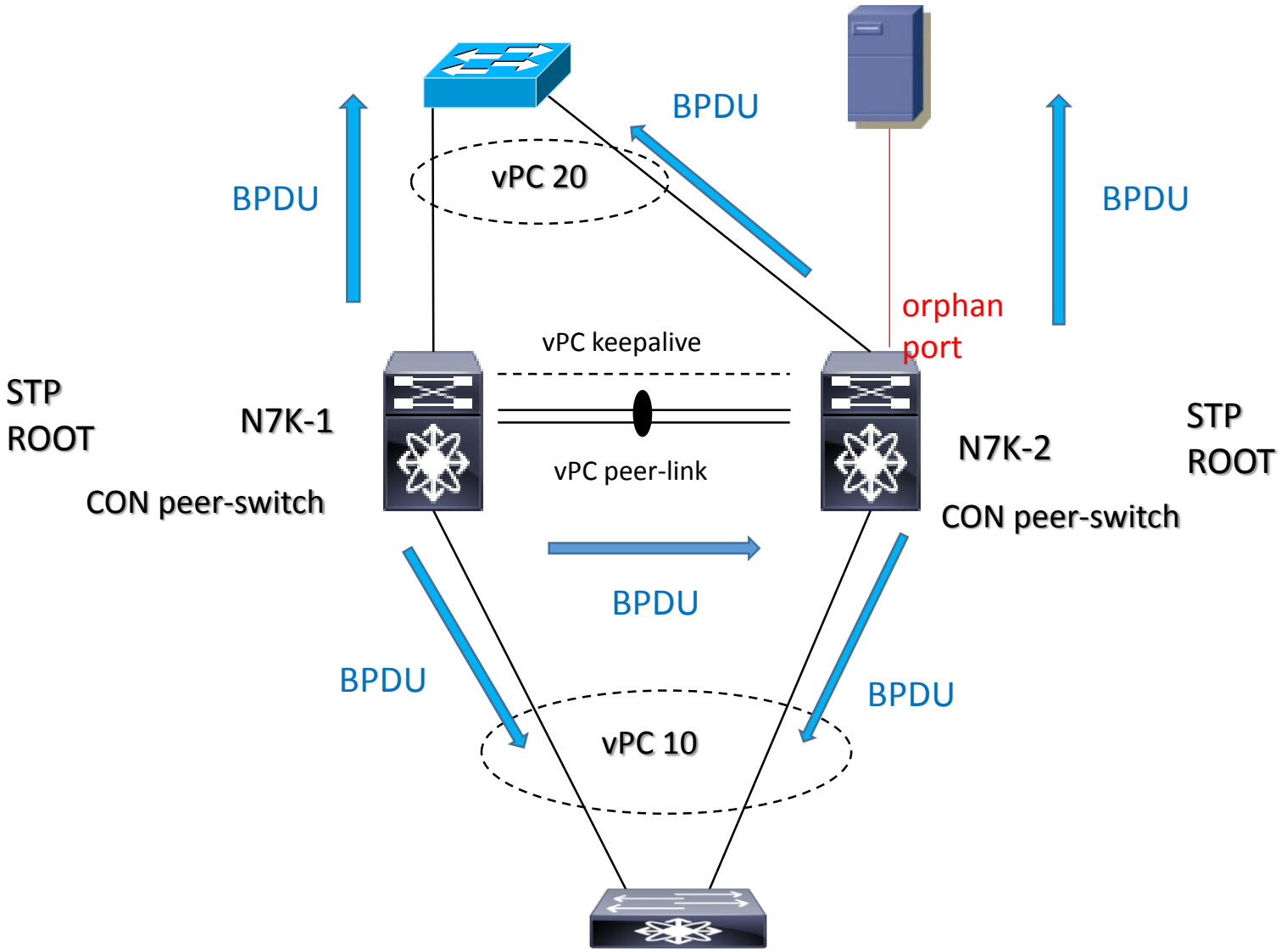
Il peer primario è responsabile per tutte le comunicazioni STP attraverso i vPC configurati; il peer secondario genera quindi BPDU solo per le interfaccie non membri vPC ma solo attraverso le sue non-vPC interface (esempio le orphan port)

- peer-switch: permette ad entrambi i peer vPC switches di emulare un singolo STP bridge; cosi entrambi gli switch trasmettono BPDU down su tutte le loro interface usando lo stesso STP bridge ID (il vPC system MAC address è usato via le BPDU)

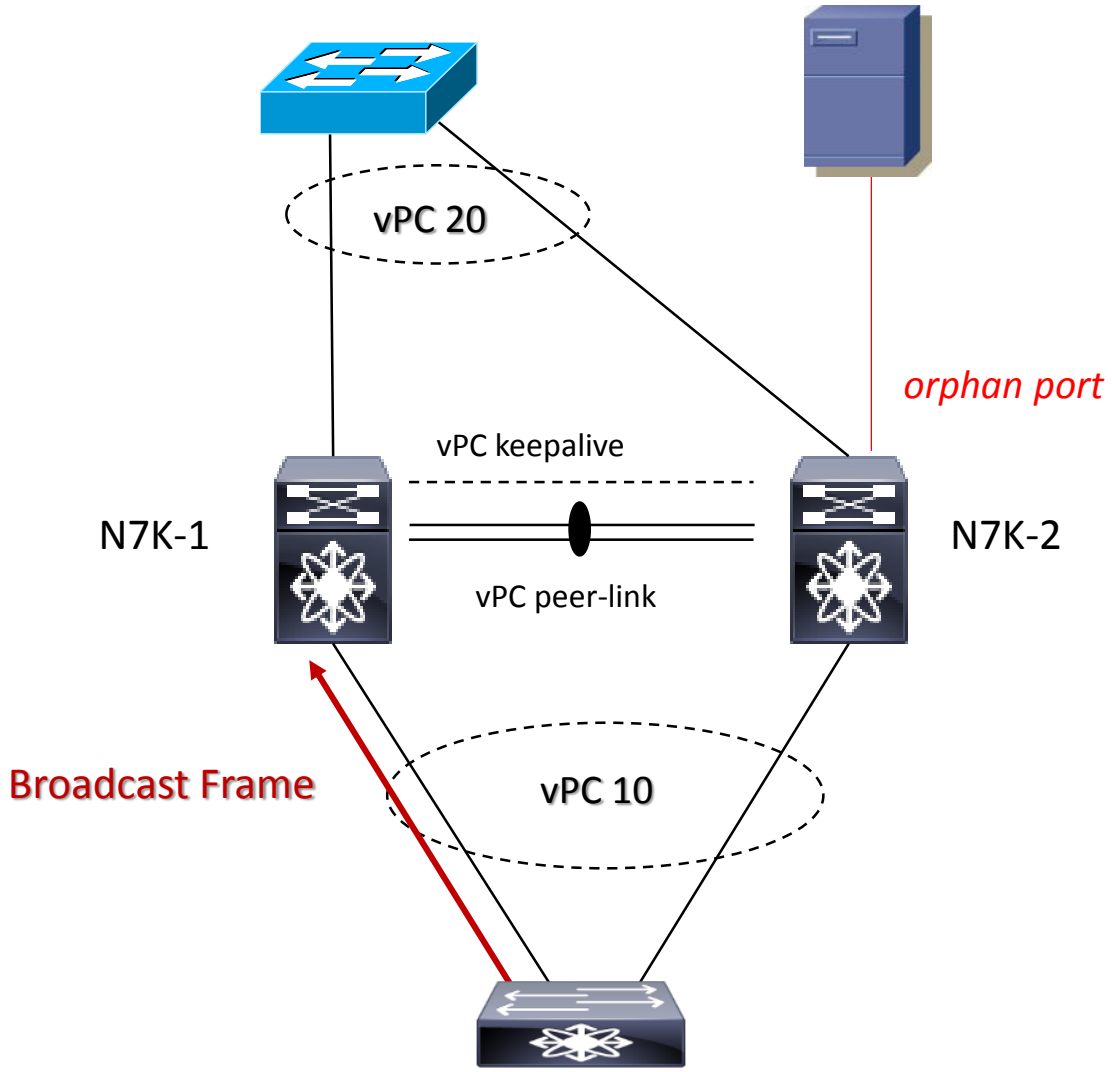
vPC Architectures and STP Spanning Tree Protocol wiithout peer-switch feature (1/1)



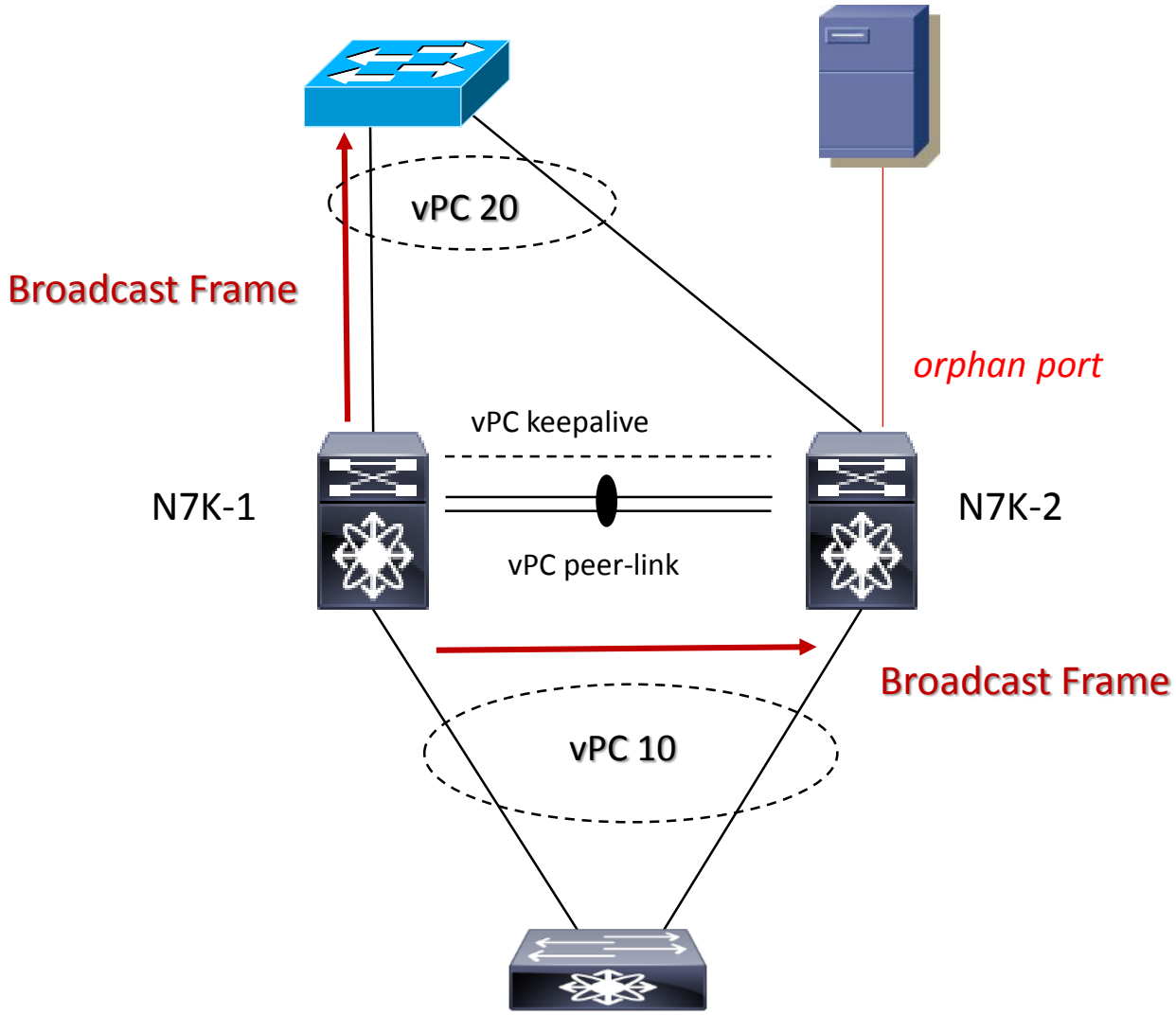
vPC Architectures and STP Spanning Tree Protocol with peer-switch feature (1/2)



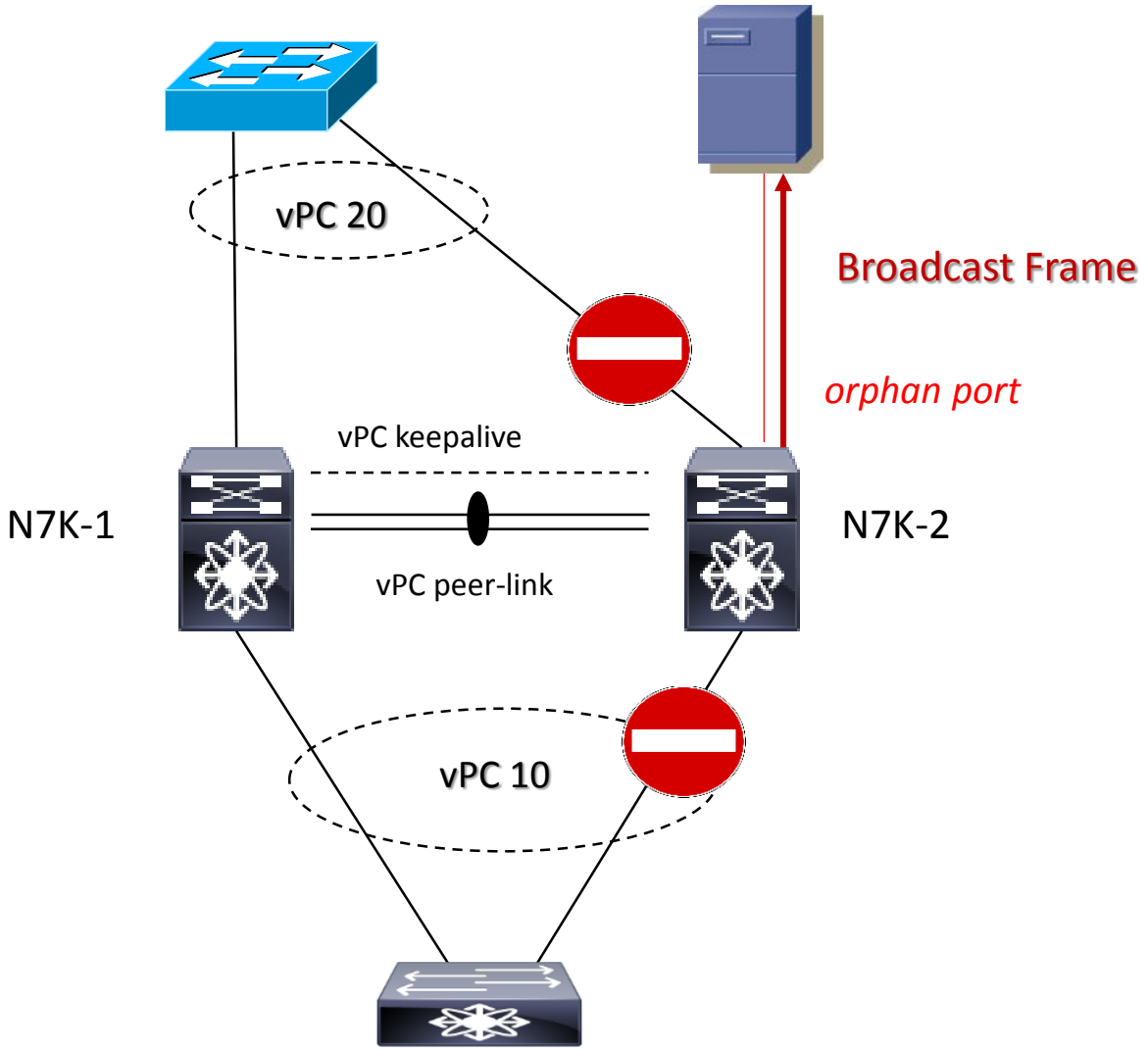
vPC broadcast frame forwarding and loop-avoidance (1/1)



vPC broadcast frame forwarding and loop-avoidance (1/2)



vPC broadcast frame forwarding and loop-avoidance (1/3)



vPC Architectures and First Hop Routing Gateway with peer-gateway feature

vPC peers può essere configurator come default gateway, affinchè entrambi gli switches hanno layer 3 capability.

La trasmissione in upstream (from Server) è bilanciata via un hash algoritmo attraverso il vPC

NOTA:
Non tutti i server, però, necessitano di bilanciamento (ad esempio server NAS Network Storage) o server con particolari NIC teaming perché non trasmettono nessun ARP request per scoprire indirizzi MAC, ma semplicemente usano il MAC address ricevuto nella frame originale di risposta.

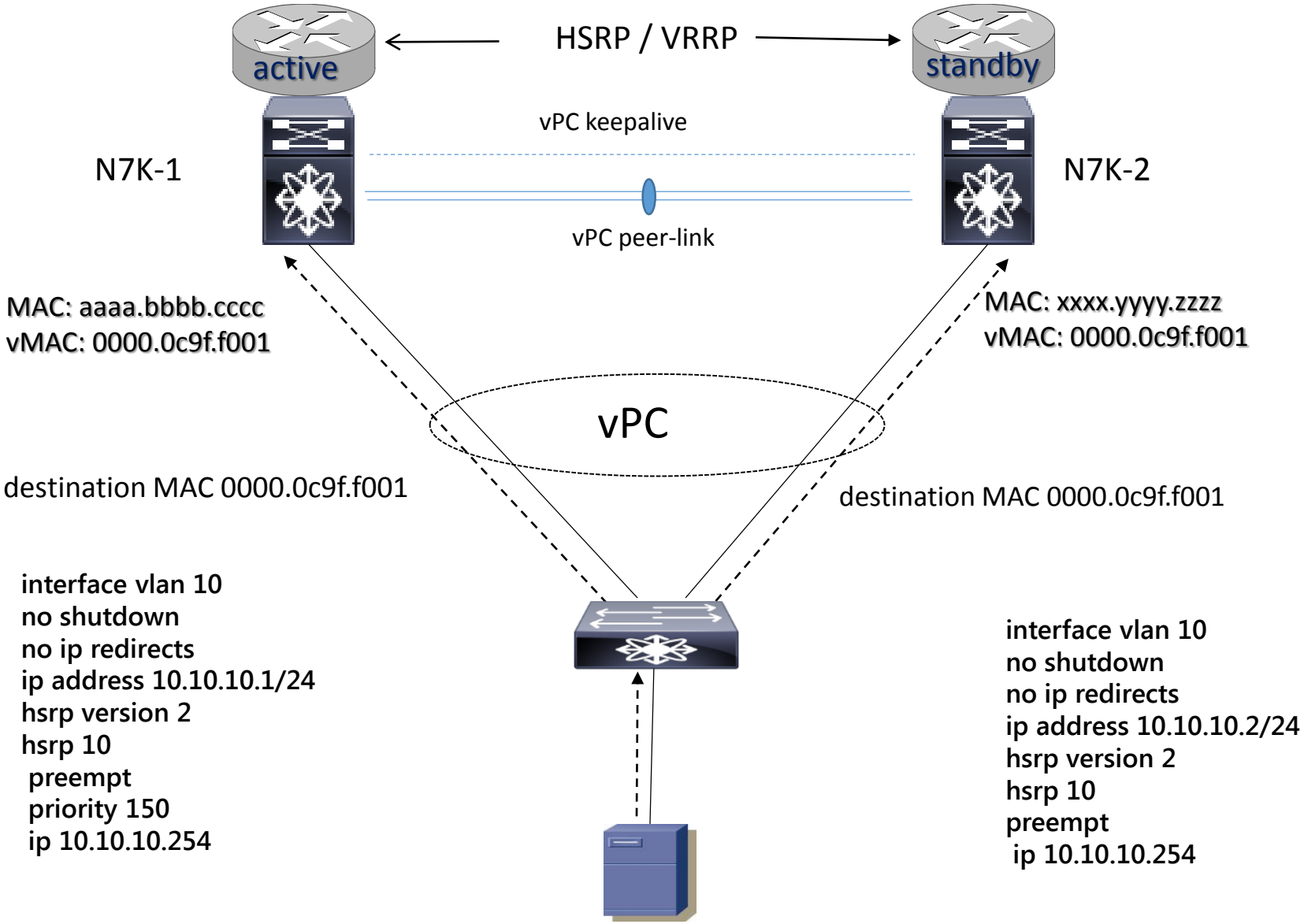
- Il NAS trasmette un pacchetto IP diretto al MAC aaaa.bbbb.cccc (int vlan 10 N7K-1), usando però una interfaccia collegata al N7K-2 perchè cosi ha deciso l'hash algoritmo vPC port-channel;

- N7K-2 switches questa frame verso il N7K-1 usando il proprio MAC address;

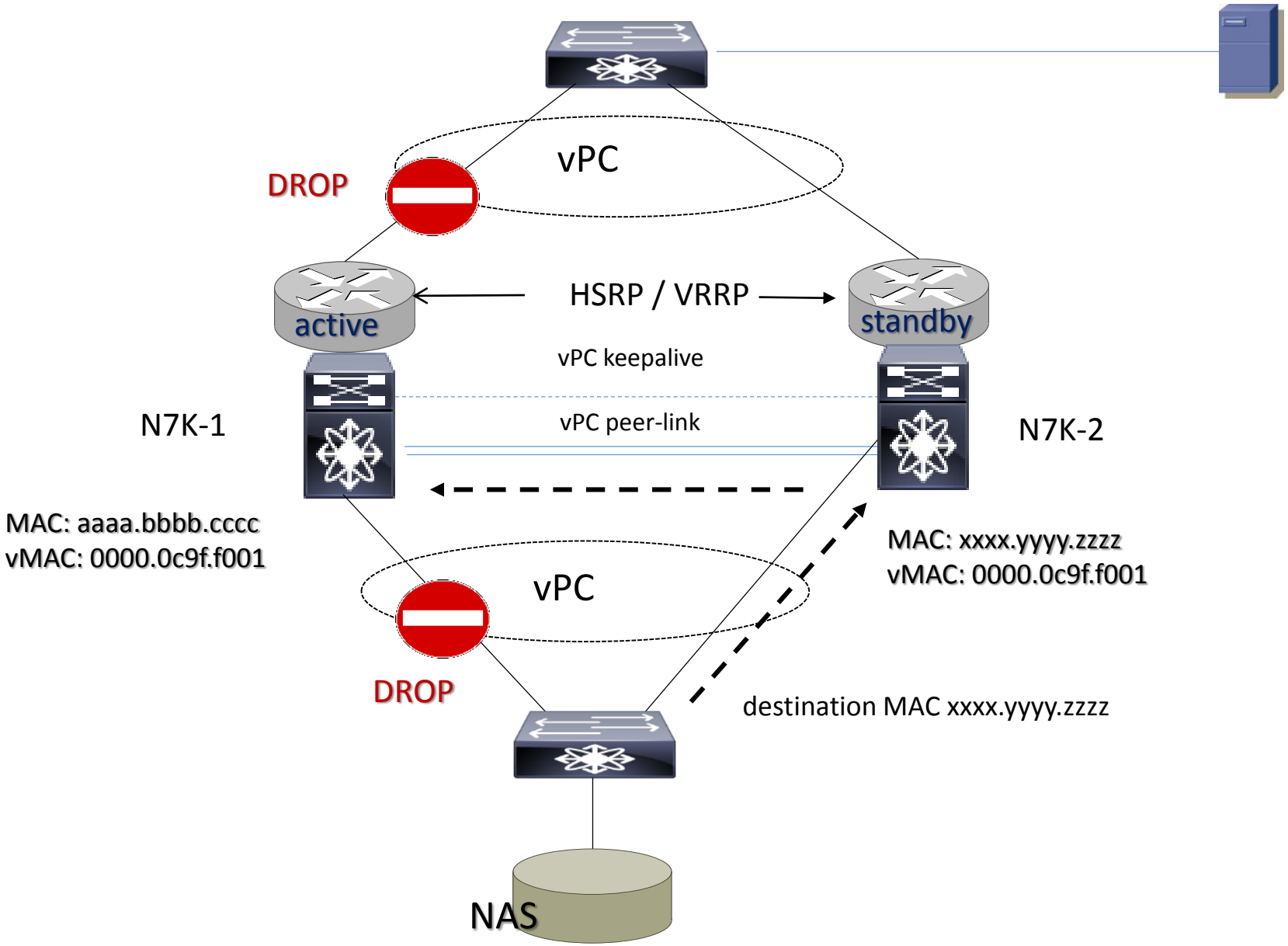
- N7K-1 può bloccare questo pacchetto perchè supposto essere trasmesso fuori da o verso un vPC (split-horizon loop via port-channeling dove il traffico entrante in un port-channel non può uscire dallo stesso port-channel)

peer gateway è una caratteristica che risolve il problema indicato nella slide 1/2; in questo caso entrambi gli switches sono configurati come peer gateway e sono abilitati a ruotare pacchetti che sono diretti ai propri peer MAC address

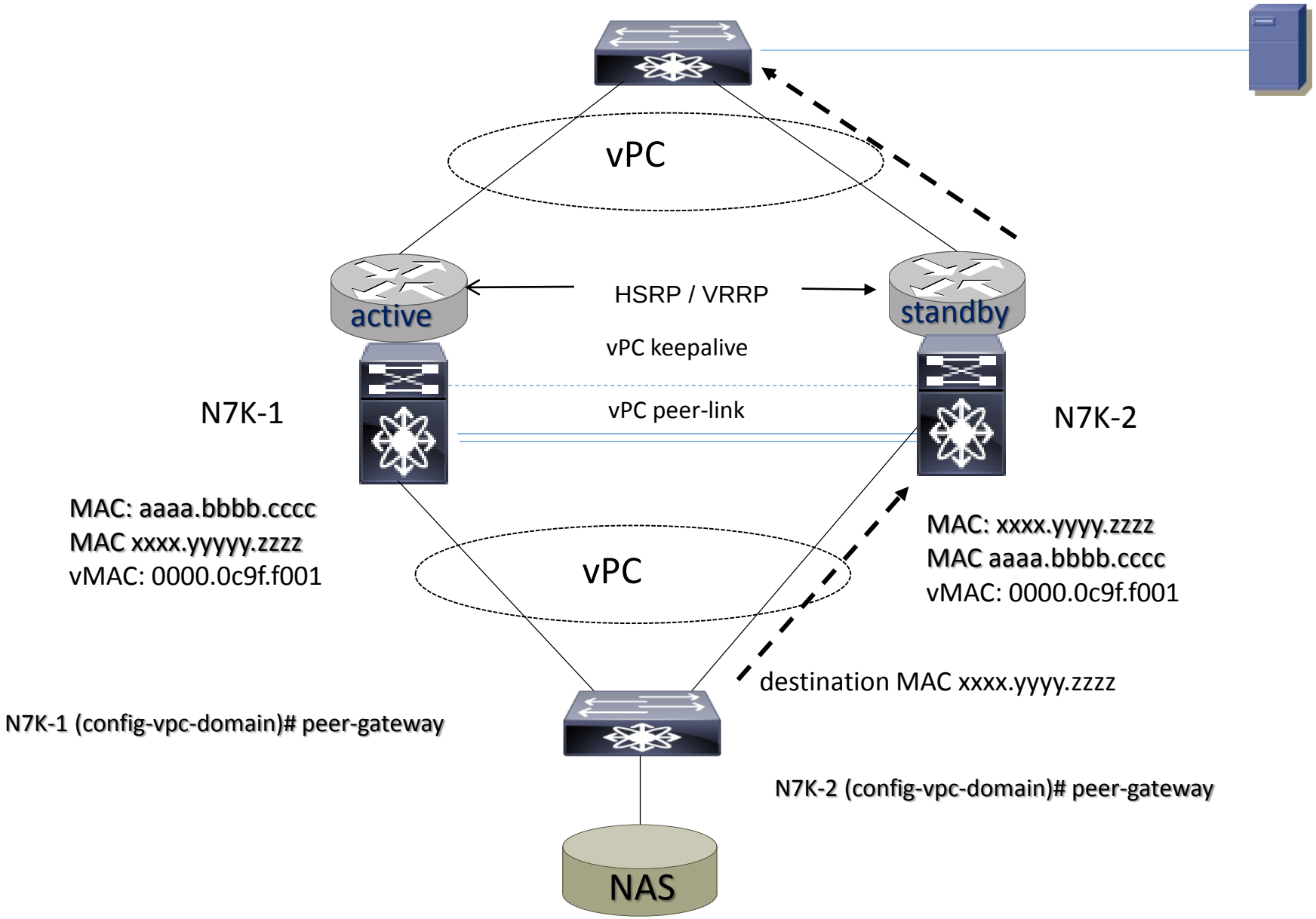
vPC Architectures and First Hop Routing Gateway with peer-gateway feature (1/1)



vPC Architectures and First Hop Routing Gateway with peer-gateway feature (1/2)

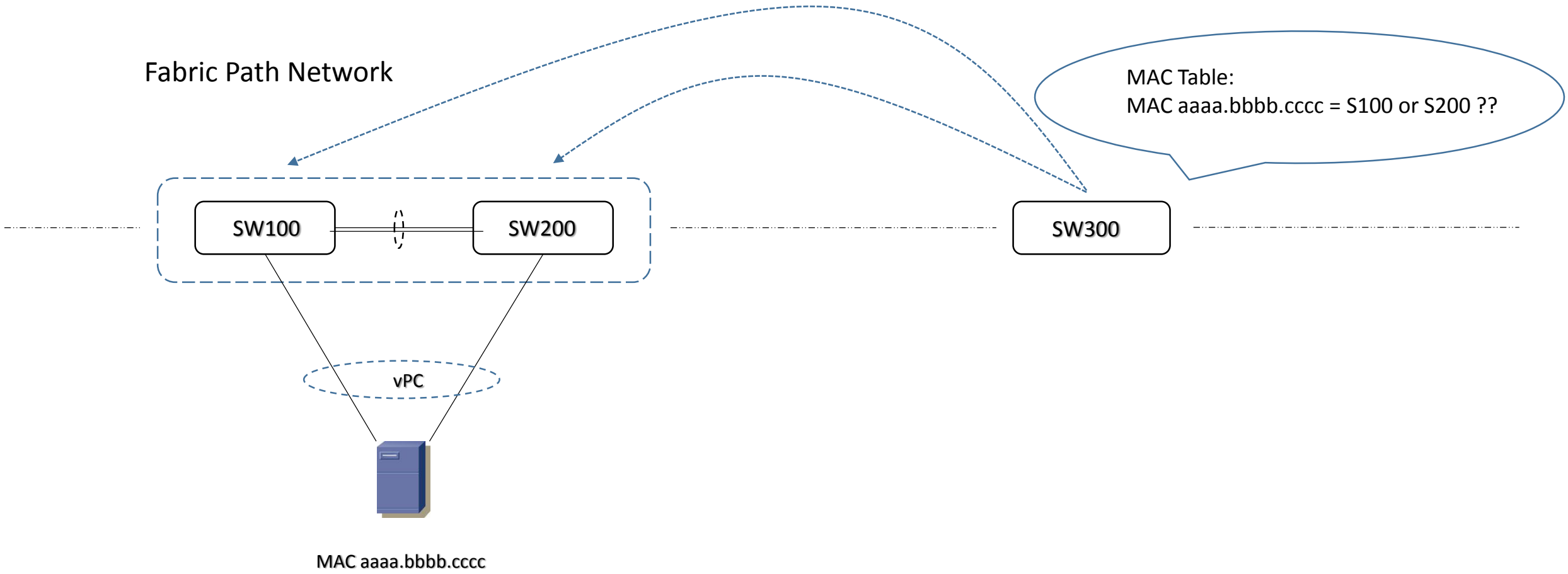


vPC Architectures and First Hop Routing Gateway with peer-gateway feature (1/3)



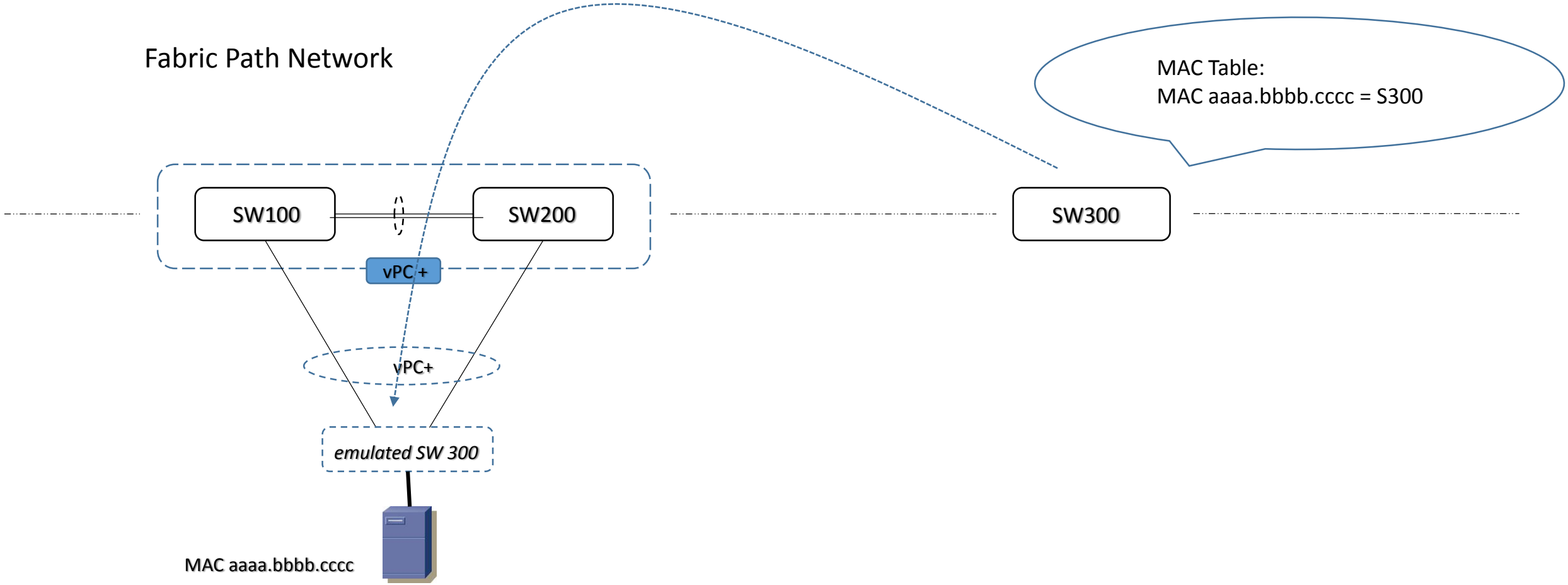
vPC Plus (vPC+) Architectures Design Cisco Models (1/2)

vPC+ è una tecnologia di virtualizzazione Cisco che provvede ad un Layer 2 multipathing quale FP verso connessioni a switch non-FabricPath; in altre parole la differenza tra vPC+ ed vPC è che vPC+ performa la formazione di un " emulated FabricPath switch " la quale garantisce il load balancing di frames dirette verso un virtual port-channel attraverso una rete FabricPath



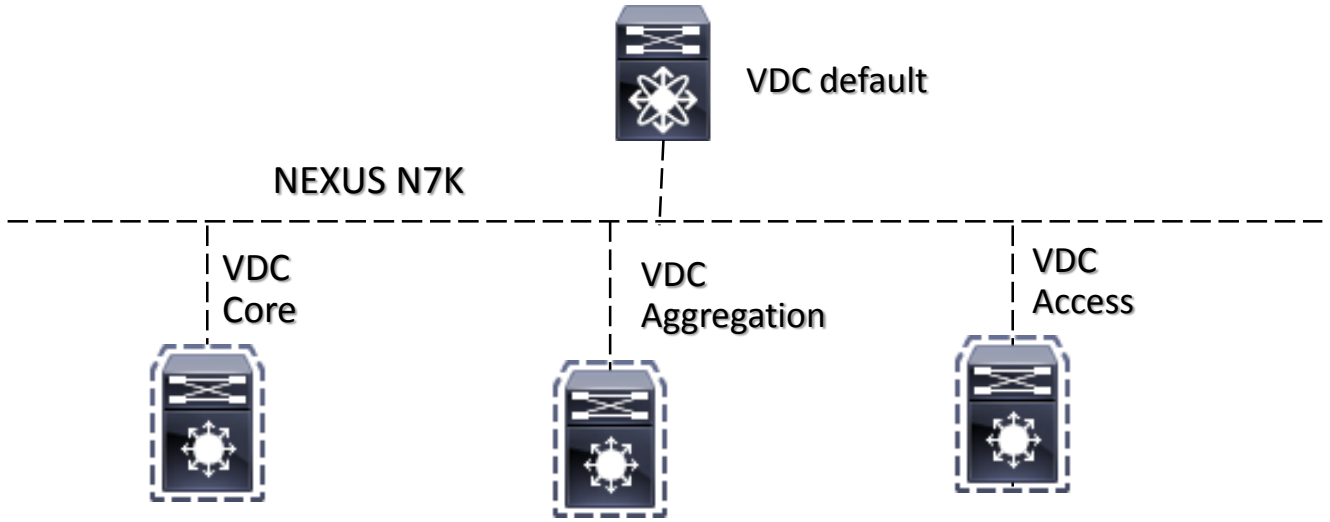
vPC Plus (vPC+) Architectures Design Cisco Models (2/2)

vPC+ è una tecnologia di virtualizzazione Cisco che provvede ad un Layer 2 multipathing quale FP verso connessioni a switch non-FabricPath; in altre parole la differenza tra vPC+ ed vPC è che vPC+ performa la formazione di un " emulated FabricPath switch " la quale garantisce il load balancing di frames dirette verso un virtual port-channel attraverso una rete FabricPath



VDC concepts

- solo il Nexus 7K ha il concetto di VDC
- il sistema operativo dei Nexus è NX-OS
- inizialmente tutte le risorse hardware (physical ports) e software appartengno al VDC di default; attraverso questo VDC è possibile creare nuivi contesti virtuali ed allocare le risorse di cui sopra ai VDC di competenza consentendo una completa separazione dei protocolli di livello 2 e 3.
- a seconda della supervisor engine presente è possibile collegare da 4 ad 8 VDC Virtual Device Context
- l'interfaccia di mngt0 (out-of-band management) permette invece di gestire tutti i VDC creati; comunque ogni VDC ha un suo indirizzo IP di management che permette la trasmissione di informazioni syslog, SNMP, etc.
- se esiste un dominio Storage, è possibile creare un VDC dedicato per il trasporto di traffico FCoE



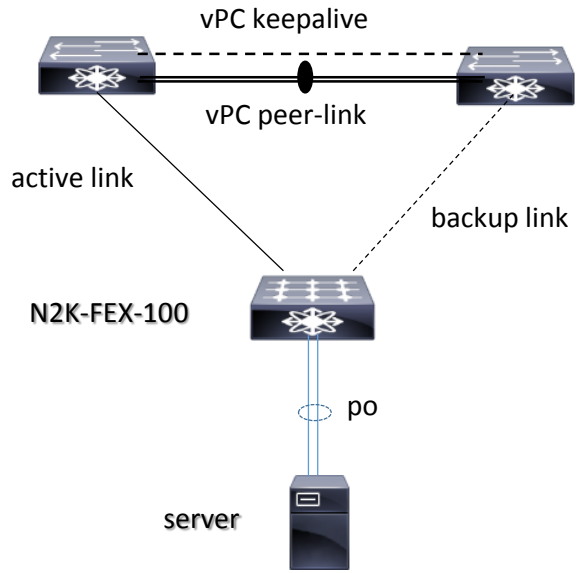
POD NeXUS FEX 1xN2K with active-standby dual-homed

I FEX sono switch cisco gestiti dai loro parent-switch Nexus (possono essere visti come una estensione modulare dei parent-switch)

In questa configurazione il FEX N2K è nello stato Online con il Nexus N5K-1 e rimane nello stato Connected nel N5K-2 perchè è già registrato dal primo

La connessione verso il N5K-2 (standby) non è usato per il trasporto del data traffic; la transazione da un parent-switch ad un'altro ha una attesa di circa 40 secondi prima che il Fabric Extender (FEX) diventa Online.

Per evitare questa situazione possiamo considerare una connessione di tipo active-active con vPC



POD NeXUS FEX 1xN2K with active-active dual-homed

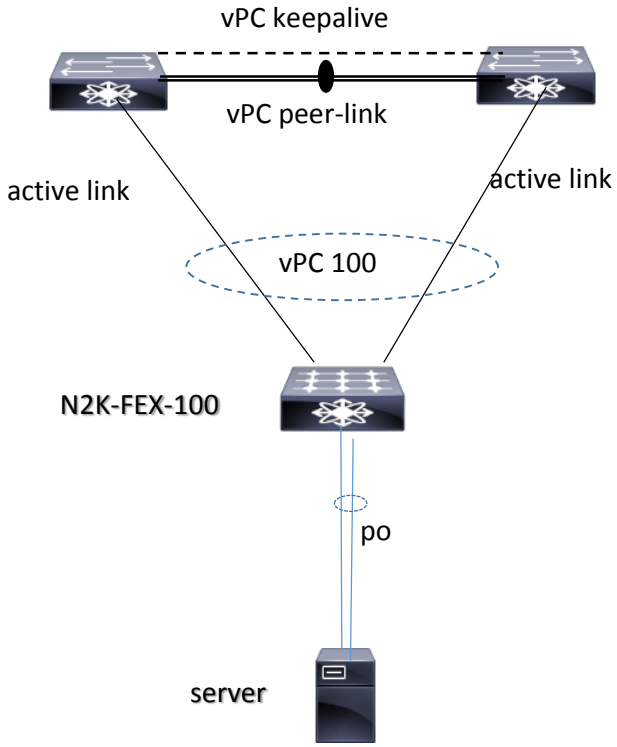
In active-active configuration, il FEX N2K è nello stato Online per entrambi I parent-switch N5K.

In questa topologia un eventuale failure di un parent-switch non ha effetto sul FEX perchè entrambi i parent-switch peers vPC gestiscono la sua connessione simultaneamente.

Requisito prevede che la configurazione FEX N2K sia la stessa (incluso le host interfaces) in entrambi gli switch

Configurazione:

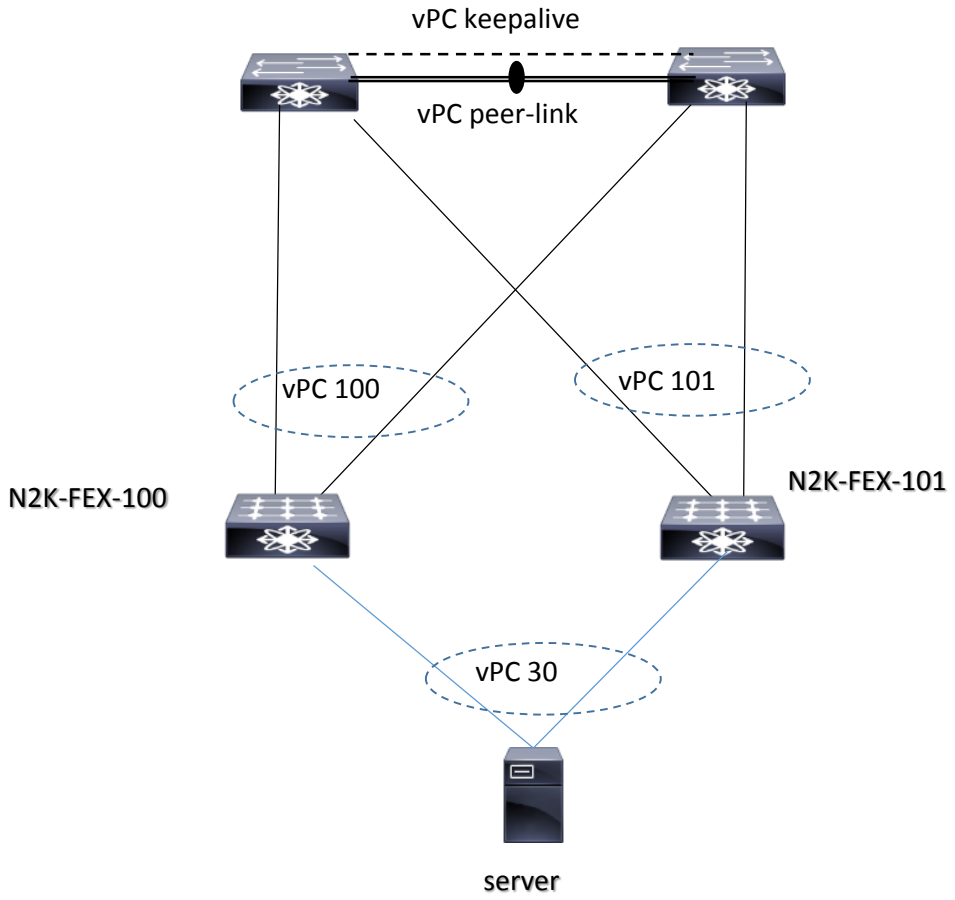
```
feature fex
!
fex 100
pinning max-links 1
description "FEX100"
!
interface eth1/1
switchport mode fex-fabric
channel-group 100
fex associate 100
!
interface port-channel 100
switchport mode fex-fabric
fex associate 100
vpc 100
```



POD NeXUS FEX 2xN2K with enhanced vpc

Questa configurazione con doppio FEX prevede una EvPC capacità, mantenendo la stessa configurazione per entrambi i parent-switch N5K e rilasciando un port-channel per l'nterfaccia di collegamento al server che si cerca di aggregare:

```
interface port-channel 30
description "to Server"
switchport mode trunk
switchport trunk allowed vlan 10-19, 20-29, 30-39
!
interface Ethernet 100/1/1
description "to Server FEX100"
switchport mode trunk
switchport trunk allowed vlan 10-19, 20-29, 30-39
channel-group 30 mode active
!
interface Ethernet 101/1/1
description "to Server FEX101"
switchport mode trunk
switchport trunk allowed vlan 10-19, 20-29, 30-39
channel-group 30 mode active
!
```



POD NeXUS FEX 2xN2K with straight-through

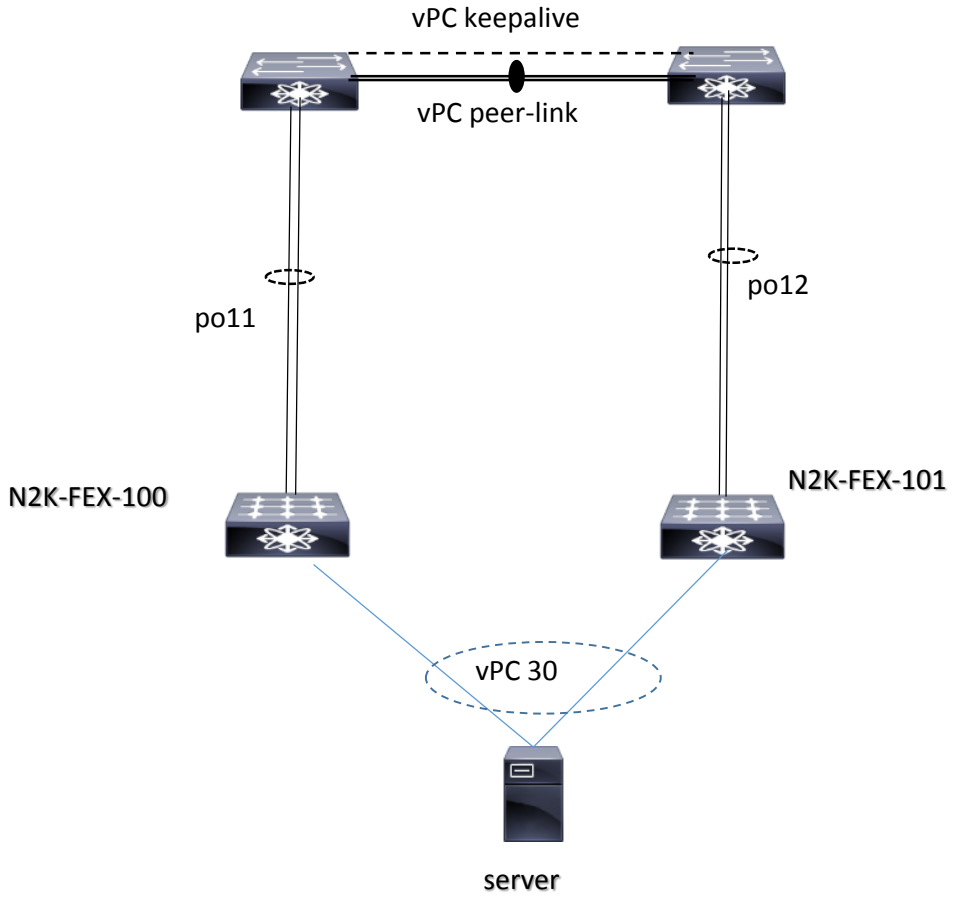
In questa topologia la configurazione vPC lato server mantiene una modalità active-active evitando perdita di connettività in caso di fault di uno dei due parent-switch N5K
Ogni FEX usa due aggregate link Fabric verso i rispettivi parent-switch

N5K-1

```
interface po11
vpc 30
!
interface eth 110/1/1
vpc 30
```

N5K-2

```
interface po12
vpc 30
!
interface eth 120/1/1
vpc 30
```



Architetture Unified UCS Cisco

- Unified Computing System UCS Cisco
- Unified UCS design models
- FCoE concepts
- UCS Services Profile
- Virtual Machine (VM) concepts
- Nexus 1000v cisco VSM + VEM
- Nexus 1000v Vsphere ESXi UCS diagram

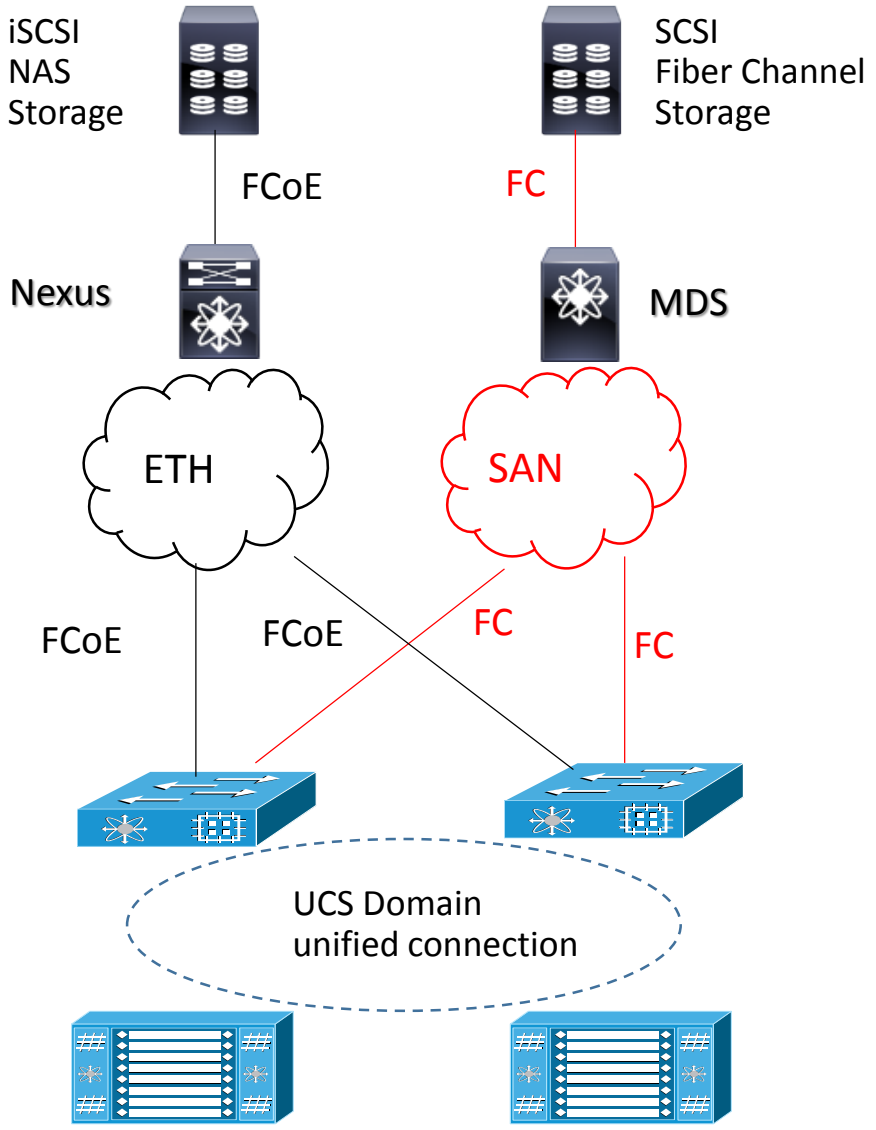
Unified computing system (UCS)

Unified Computing System (UCS) significa un insieme di Servers, Storage e tecnologie di virtualizzazione all'interno di una stessa architettura.

L'interoperabilità tra un sistema UCS Servers e le infrastrutture di rete IP e SAN è gestita da devices chiamati Fabric Interconnect

- Servers blade UCS serie B
- Servers Rack UCS serie C
- Servers di archiviazione UCS serie S
- Software di gestione UCS Manager
- Fabric Interconnect UCS + Fabric Interconnect Extender

Unified computing system (UCS)



FCoE concepts

FCoE mappa le frame FC su una rete IEEE 802.3 Ethernet full-duplex con connessioni a 10G senza modificare tutte le funzionalità proprie del FC (zoning, lun, etc..);

Sono necessarie apposite schede di rete chiamate CNA (Converged Network Adpater) e switch Ethernet per il trasporto e l'instradamento di FCoE packets;

Un server connesso ad una rete FCoE rappresenta un iSCSI Initiator (cosi come un server SCSI nativo collegato in FC), mentre uno Storage Array connesso tramite FCoE rappresenta uno iSCSI target ;

Per operare FCoE ha bisogno di una rete lossless Ethernet che garantisca un trasporto senza perdita di pacchetti indispensabile per uno scambio di dati SCSI incapsulato all'interno di pacchetti Fibre Channel

Per operare, FCoE ha bisogno di una rete lossless ethernet che garantisca un trasporto senza perdita di pacchetti indispensabile al trasporto del traffico SCSI che è incapsulato all'interno dei pacchetti Fibre Channel.

Una SAN è una dedicata area Network che provvede all'accesso di dati su livelli di blocchi Storage Servers

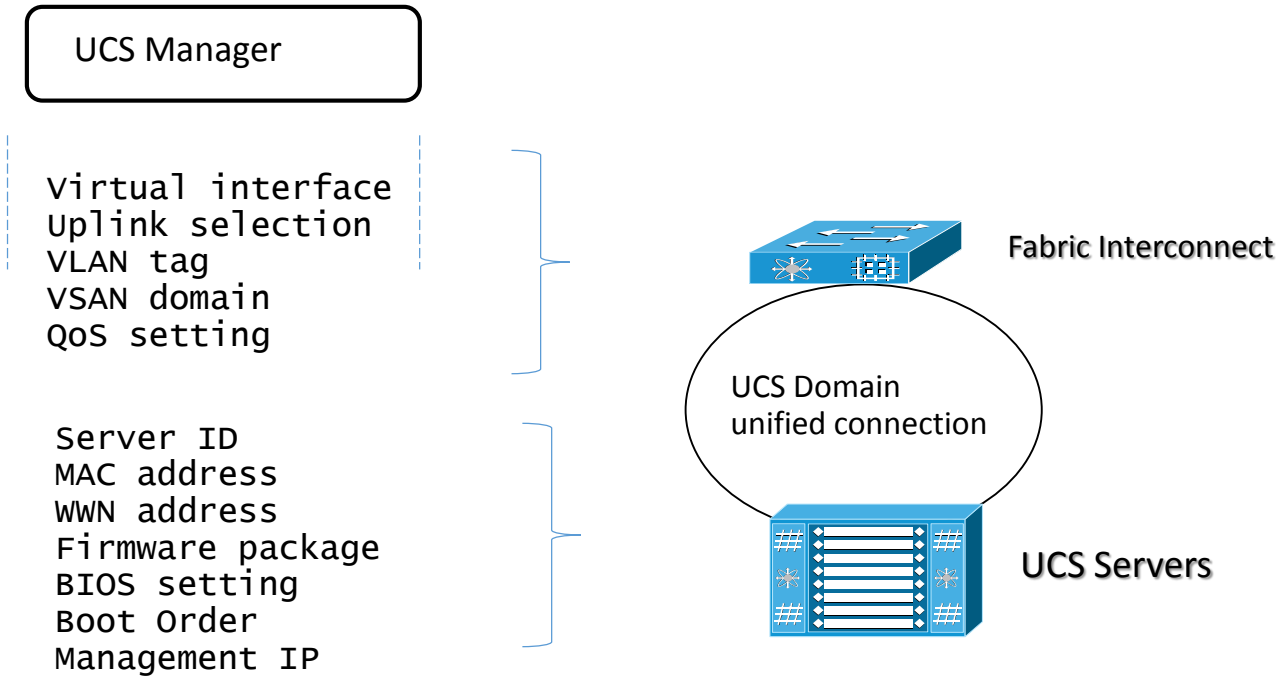
Una SAN tipicamente non è accessibile attraverso standard network ma solo attraverso la sua dedicata piattaforma di rete;

La connessione verso devices SAN è possibile attraverso una particolare estensione definità in una card chiamata HBA (Host Bus Adpater)

UCS services profle

Un blade o Rack UCS server deve essere associate ad un “ service profile ” ed ogni associazione ha una relazione 1:1 con un server.

Quando un service profile è associato ad un server, sia fabric-interconnect che le componente del server (adpters, BIOS, etc..) sono configurati per accordarsi su specifici parmetri (virtual-interface eth o FC, unico VID, LAN connectivity (MAC address), SAN connectivity (wwn), firmware package and version, IP address di management, etc…
Un service profile è una entità virtuale all’interno del sistema di gestione UCS Manager



Virtual Server (VM) concept

Una VM (Virtual Machine) emula un server fisico per sistema operativo, applicazioni, IP address e collegamento verso una rete (vnic);

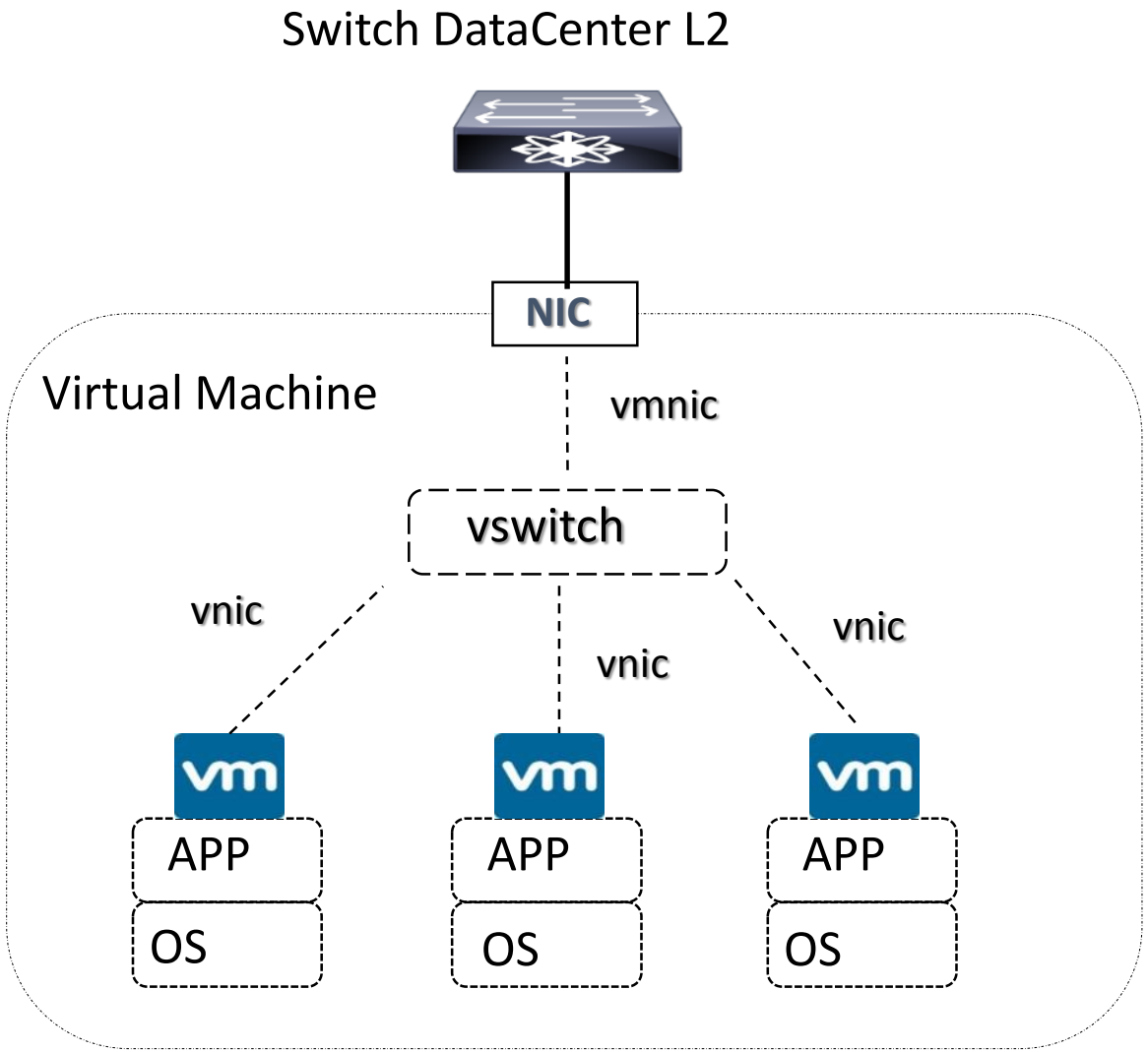
VMware ha introdotto il concetto di vswitch (virtual switch) che altro non è che un Hipervisor che emula tutte le funzionalità di un vero layer 2 switch;

Questo vswitch, quindi, provvede a collegamenti di tipo access ports verso le VM (vnic) e collegamenti uplinks verso physical NIC (collegamento definite vmnic) permettendo 802.1q tagging e MAC address table per trasmettere frame Ethernet basate sul loro valore di destination MAC;

Un vswitch offre configurazioni di tipo port-group; un port-group può contenere vlan-id, security feature, shaping definendo percentuali di banda utilizzabile e NIC teaming (vmnic load-balancing, network failover detection, switch notification, failure behavior);

Cisco ha introdotto Nexus 1000V quale elemento virtuale che emula le funzionalità di un distribute vswitch vmware DVS attraverso proprie API (Application Programmable Interface) rilasciate attraverso NX-OS vCenter operations

Virtual Server (VM) concept



Nexus 1000V cisco VSM + VEM

VSM (Virtual Supervisor Module): è il piano di controllo e management del Nexus 1000V;

VSM monitorizza lo stato di tutti gli switch e le loro interface, la tabella MAC address e comunica con un tool di management virtualizzato quale Vcenter VMware, permettendo la sincronizzazione ed automazione tra la rete ed i servers;

Una scheda Ethernet (adpter 1) per il controllo della comunicazione tra altri VSM e la configurazione di una VEM (virtual Ethernet module);

Una scheda Ethernet (adpter 2) per il sistema di management (mgmt0);

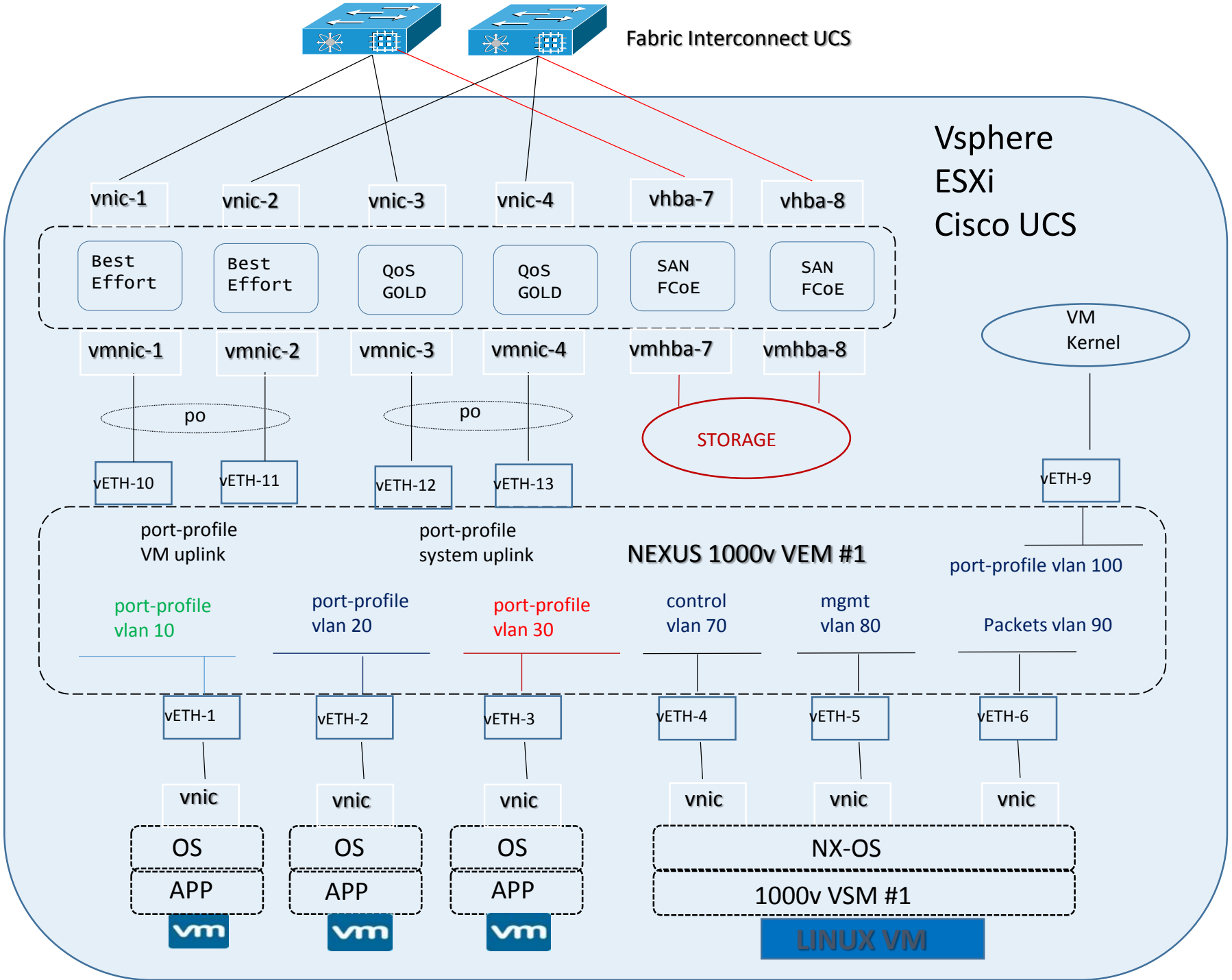
Una scheda Ethernet (adpter 3) per la trasmissione di packets inviati da uan VEM verso il VSM per essere maggiormente analizzati (esempio: CDP, LACP, IGMP snooping, SNMP e Netflow);

Nexus 1000V può essere configurato in modalità active-standby con due differenti VSM per ridondanza

VEM (Virtual Ethernet Module) condivide un dominio di broadcast (vlan) per il controllo layer 2 con il VSM;

Ogni VEM richiede uno specific VM-Kernel interface (vmknic) per comunicare con il VSM (layer 3 control mode);

Port Profile è una collezione di interface-level configuration per creare delle network policy


Fabric Interconnect UCS
Vsphere
ESXi
Cisco UCS
vnic-1
vnic-2
vnic-3
vnic-4
vhba-7
vhba-8
Best Effort
Best Effort
QoS GOLD
QoS GOLD
SAN FCoE
SAN FCoE
VM Kernel
vmnic-1
vmnic-2
vmnic-3
vmnic-4
vmhba-7
vmhba-8
po
po
STORAGE
vETH-10
vETH-11
vETH-12
vETH-13
vETH-9
port-profile VM uplink
port-profile system uplink
NEXUS 1000v VEM #1
port-profile vlan 100
port-profile vlan 10
port-profile vlan 20
port-profile vlan 30
control vlan 70
mgmt vlan 80
Packets vlan 90
vETH-1
vETH-2
vETH-3
vETH-4
vETH-5
vETH-6
vnic
vnic
vnic
vnic
vnic
vnic
OS
OS
OS
NX-OS
APP
APP
APP
1000v VSM #1
LINUX VM

Architetture Spine Leaf

- Vantaggi di una rete Spine Leaf
- Fabric Path concept
- Fabric Path design
- Fabric Path data center
- Fabric Path example configuration
- Trill
- LISP

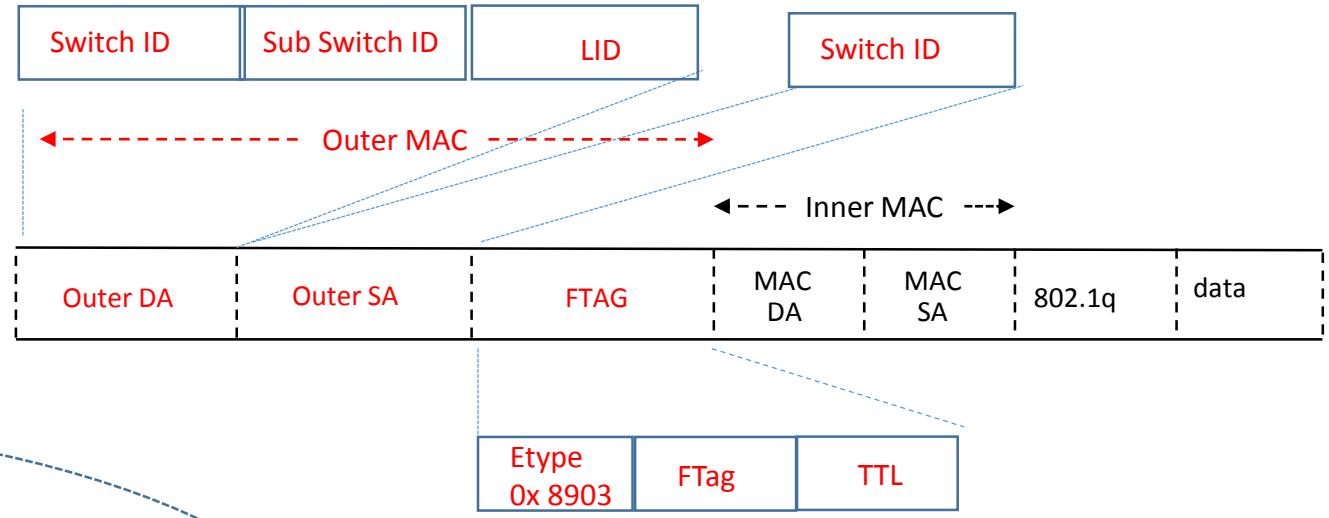
Vantaggi di una architettura Spine Leaf

- Architettura a due livelli a costruire una Fabric Switch (unico dominio);
- Alta scalabilità (possibilità di inserimento nuovi elementi) ed una grande capacità in numero di porte;
- Riduzione OpEx (es: riduzione numero apparati rispetto ad una tradizionale rete a tre livelli);
- Riduzione CapEx (es: risparmio energetico);
- Spanning Tree Free;
- L3 Ethernet equal-cost multipath (ECMP Load Balancing);
- avere funzionalità L2 (switching) attraverso L3 capability IPv4 e IPv6 (oltre MPLS, BGP, ISIS), inoltre supporta funzionalità quali FCoE, VXLAN, NVGRE, VMware integration

Fabric-Path Cisco Spine Leaf

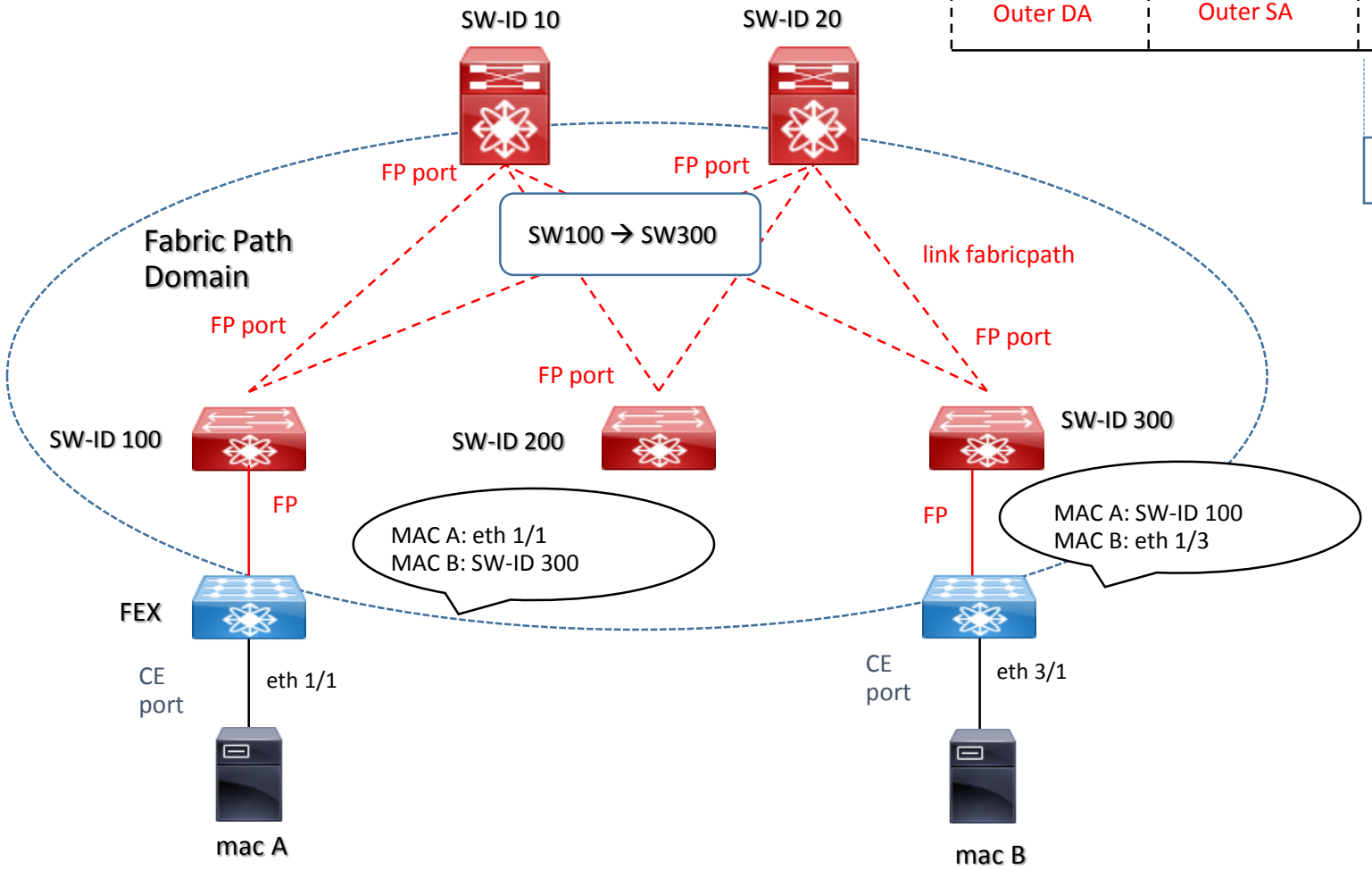
- FabricPath è una tecnologia Cisco con Nexus devices a livello di accesso, distribuito all'interno di un solo datacenters;
- Le frame FP è usata per incapsulare standard frame ethernet per attraversare un dominio fabricpath, basato su un nuovo header chiamato Switch-ID;
- ISIS routing protocol è utilizzato per lo scambio di informazioni riguardo la raggiungibilità degli switch-ID;
- Usando SPF (Shortest Path First), ISIS permette l'uso di multipli equal-cost path tra due end-points FP;
- La prevenzione e la riduzione dei loop è disponibile nel piano dati; i frame Cisco FabricPath includono un campo time-to-live (TTL) simile a quello usato in IP e viene applicato anche un controllo Reverse Path Forwarding (RPF)
- FP utilizza multi-destination tree per trasmettere pacchetti broadcast, multicast e unknown unicast frame;
- Da un punto di vista di un edge-switch (è uno switch che permette connessioni FP e STP) tutto il dominio FabricPath è visto come un solo Virtual STP bridge;
- FTAG descrive e segmenta un multipath mappando una frame ethernet con vlan-id ad una specifica topologia FP a livello edge-switch
- Cisco FabricPath supporta ECMP a 16 vie; pertanto, possono essere attivi fino a 16 percorsi tra due dispositivi nella rete. Poiché ciascuno di questi 16 percorsi può essere esso stesso un PortChannel a 16 porte, la soluzione può effettivamente fornire 2,56 Tbps di larghezza di banda

Fabric-Path Cisco Spine Leaf



SPINE

LEAF



Switch ID = numero unico identifica ogni switch Fabric Path

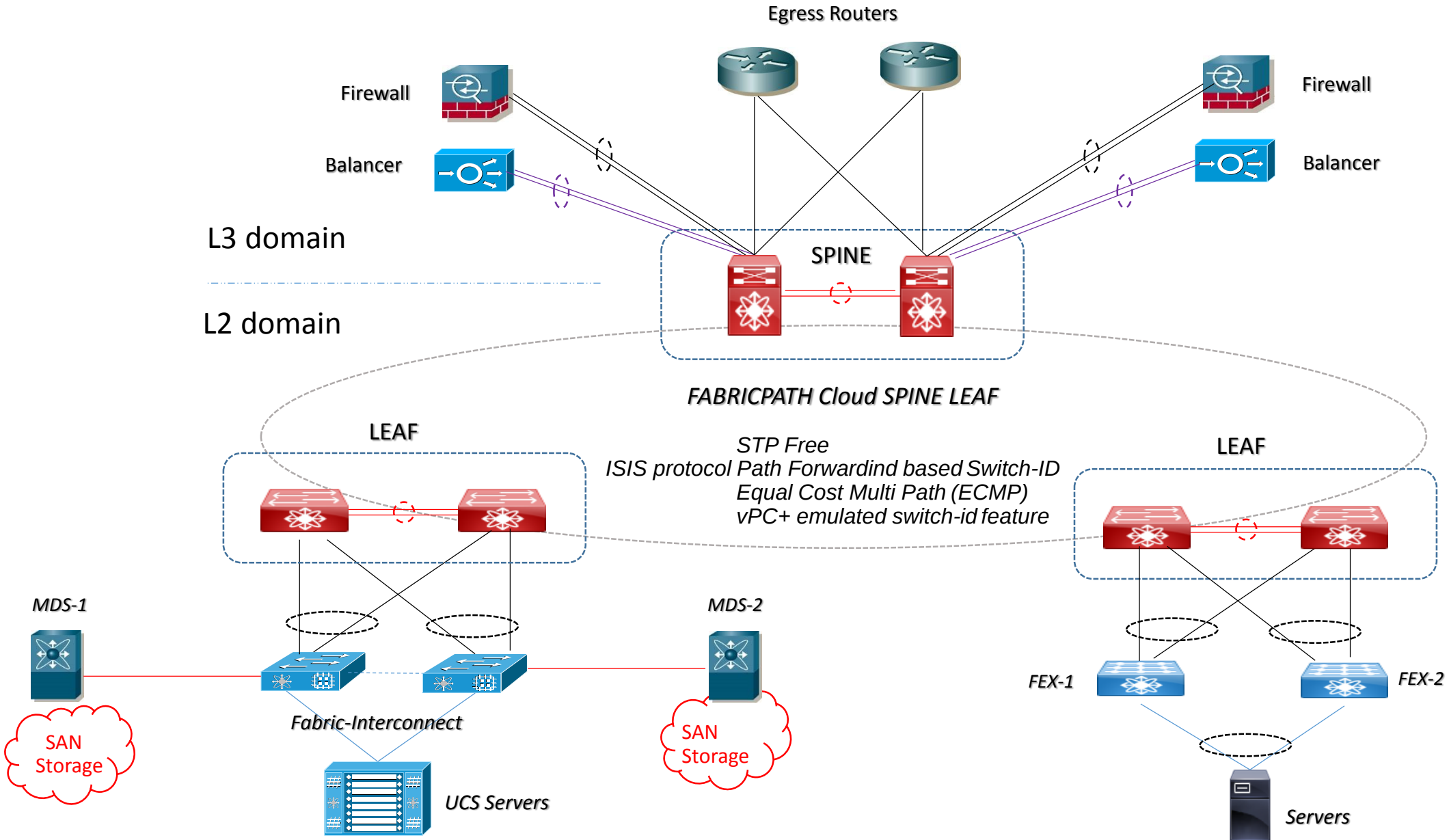
Sub Switch ID = identifica devices /host connessi via vPC+

LID = Local ID, identifica la porta destinazione o sorgente

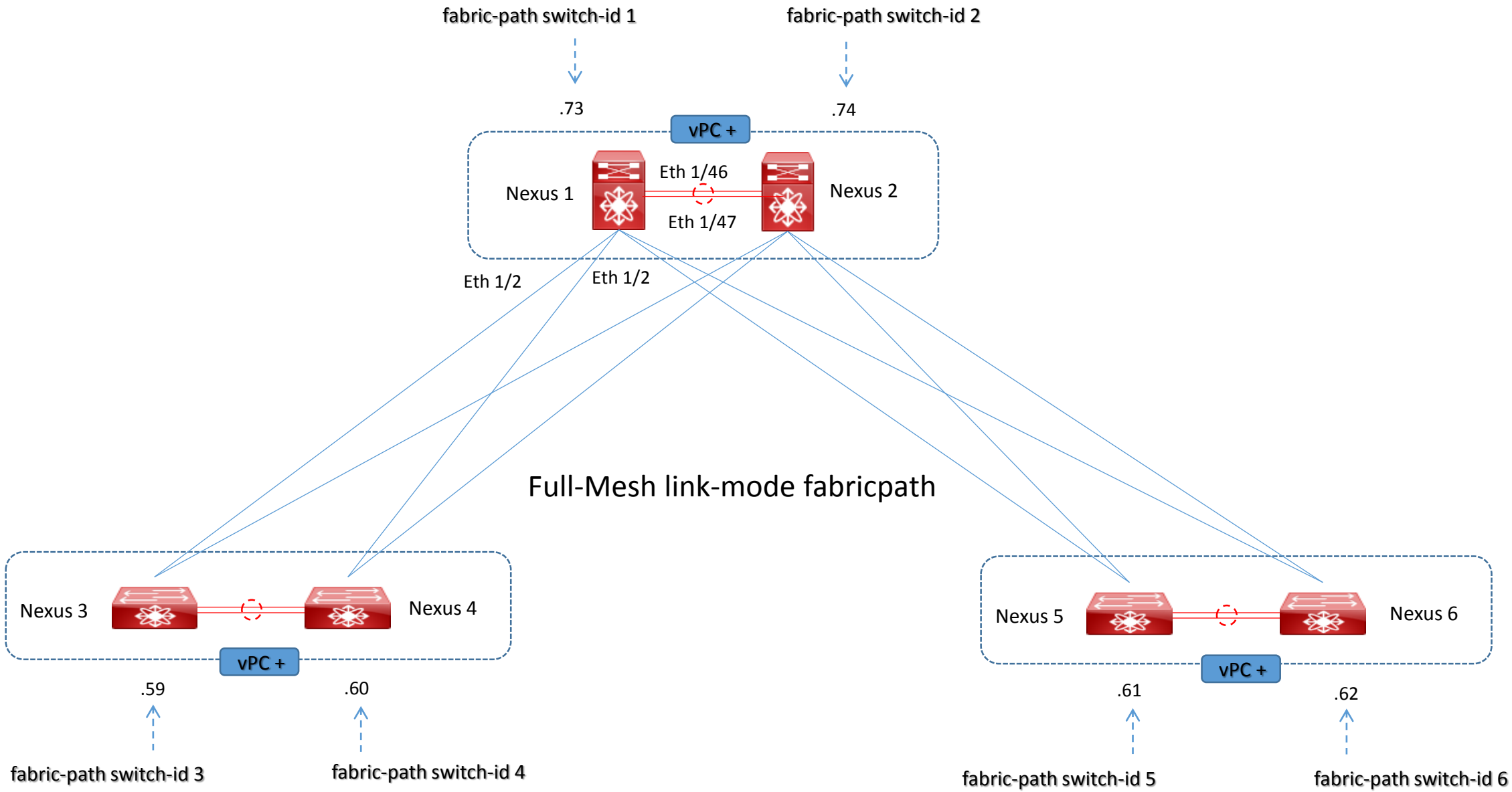
FTAG = Forwarding Table e identifica la topologia o l'albero di distribuzione

TTL = viene decrementato a seguire ogni hop del dominio in modo da prevenire un loop infinito della frame

Fabric-Path Cisco Spine Leaf Data Centers



Fabric-Path Cisco Spine Leaf Data Centers Configuration (1/1)



Fabric-Path Cisco Spine Leaf Data Centers Configuration (1/2)

NEXUS-1

```
feature-set fabricpath
feature vpc
!
inteface mgmt0
 ip address 192.168.100.73/24
!
vrf context management
 ip route 0.0.0.0/0 192.168.100.1
!
vpc domain 1
 role priority 1
 peer-keepalive destination 192.168.100.74 source 192.168.100.73
 fabricpath switch-id 1
!
interface port-channel 1
 description peer-link
 switchport mode fabricpath
 vpc peer-link
!
interface ethernet 1/46
 description B2B
 switchport mode fabricpath
 channel-group 1mode active
!
interface ethernet 1/47
 description B2B
 switchport mode fabricpath
 channel-group 1mode active
!
```

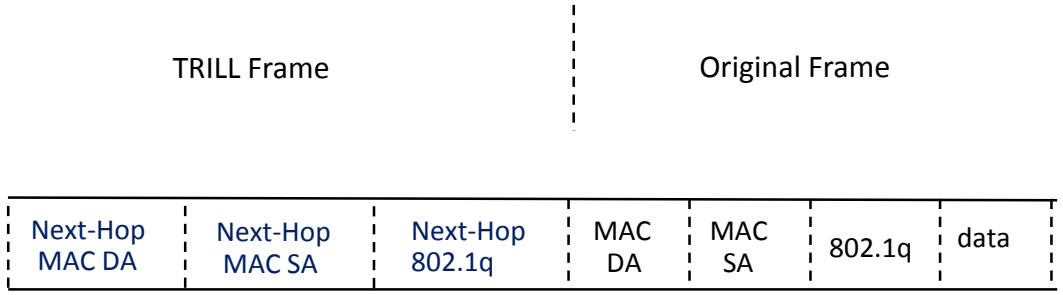
Fabric-Path Cisco Spine Leaf Data Centers Configuration (1/3)

NEXUS-1

```
vlan 100
mode fabricpath
vlan 200
mode  fabricpath
vlan 300
mode fabricpath
!
spanning-tree vlan 100,200,300 priority 8192
!
interface ethernet 1/2-3
 descri ption To-Nexus-34
 switchport mode fabricpath
 channel-group 34 mode active
 no shutdown
!
interface port-channel 34
 description link-FP To-Nexus-34
switchport
switchport mode fabricpath
no shutdown
```

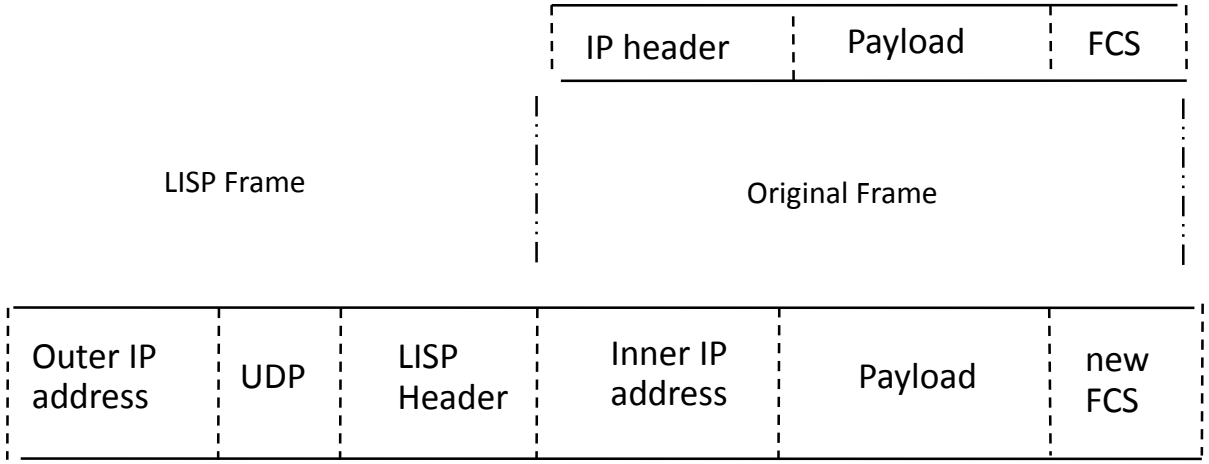
TRILL (transparent interconnection of lots of links)

- TRILL è una tecnologia L2 multipath a livello di accesso (come FabricPath);
- E' implementato da devices conosciuti come RBridge (routing bridges) che aggiunge un nuovo encapsulation in modo incrementale, ripetendo l'originale IEEE 802.3 ethernet frame che può passare attraverso intermediate Router Bridge;
- TRILL utilizza ISIS per lo scambio di informazioni di controllo e raggiungibilità tra end-points RB, calcolando il miglior percorso per pacchetti unicast e calcolare un albero di distribuzione (distribution tree) per destinazioni multiple di frame;
- Le informazioni di un End-Host possono essere apprese attraverso il protocollo ESADI (End-Station Address Distribution Information) le cui frame sono regolarmente encapsulate in TRILL frame;
- TRILL può usare un massimo di 4000 segmenti di rete (vlans)



LISP (locator / identifier separation protocol)

- LISP è progettato per ambienti datacenter dove è previsto un moving di un end-point ed i suoi parametri di rete (addressing) non cambiano ma semplicemente la sua locazione;
- RLOC (Routing Locators): descrive la topologia e locazione di un end-point e quindi è usato questo parametro per trasmettere traffico;
- EID (End-Point ID): è utilizzato per indirizzare end-points separati dalla topologia della rete;
- ITR (Ingress Tunnel Router) and ETR (Egress Tunnel Router): sono i devices che operano encapsulation (ingress) ed de-encapsulation (egress) di pacchetti IP-based EID attraverso una IP Fabric;
- LISP è conosciuto come una tecnologia Layer 3 che comprende IPv4 e IPv6 per overlay e underlay;
- LISP assicura virtual segmenti di rete (vlans) aggiungendo un header di 24 bit instance-id che permette di estendere sino a più di 16 milioni di virtual segment; questo meccanismo è settato dal ITR.

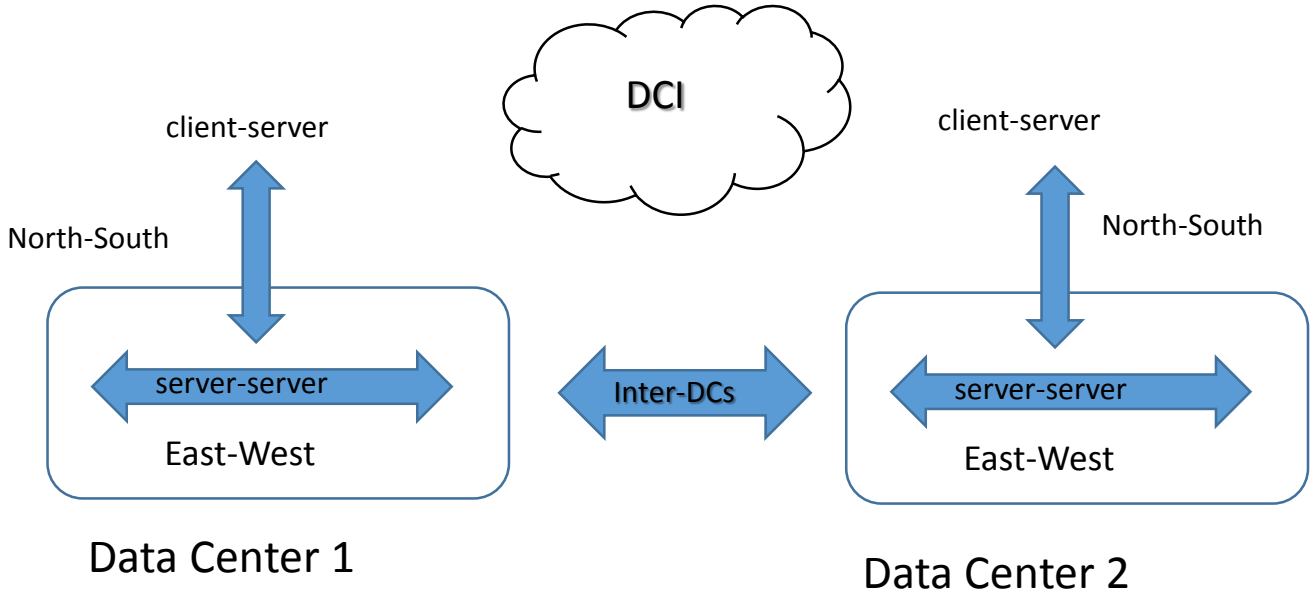


DCI Interconnection

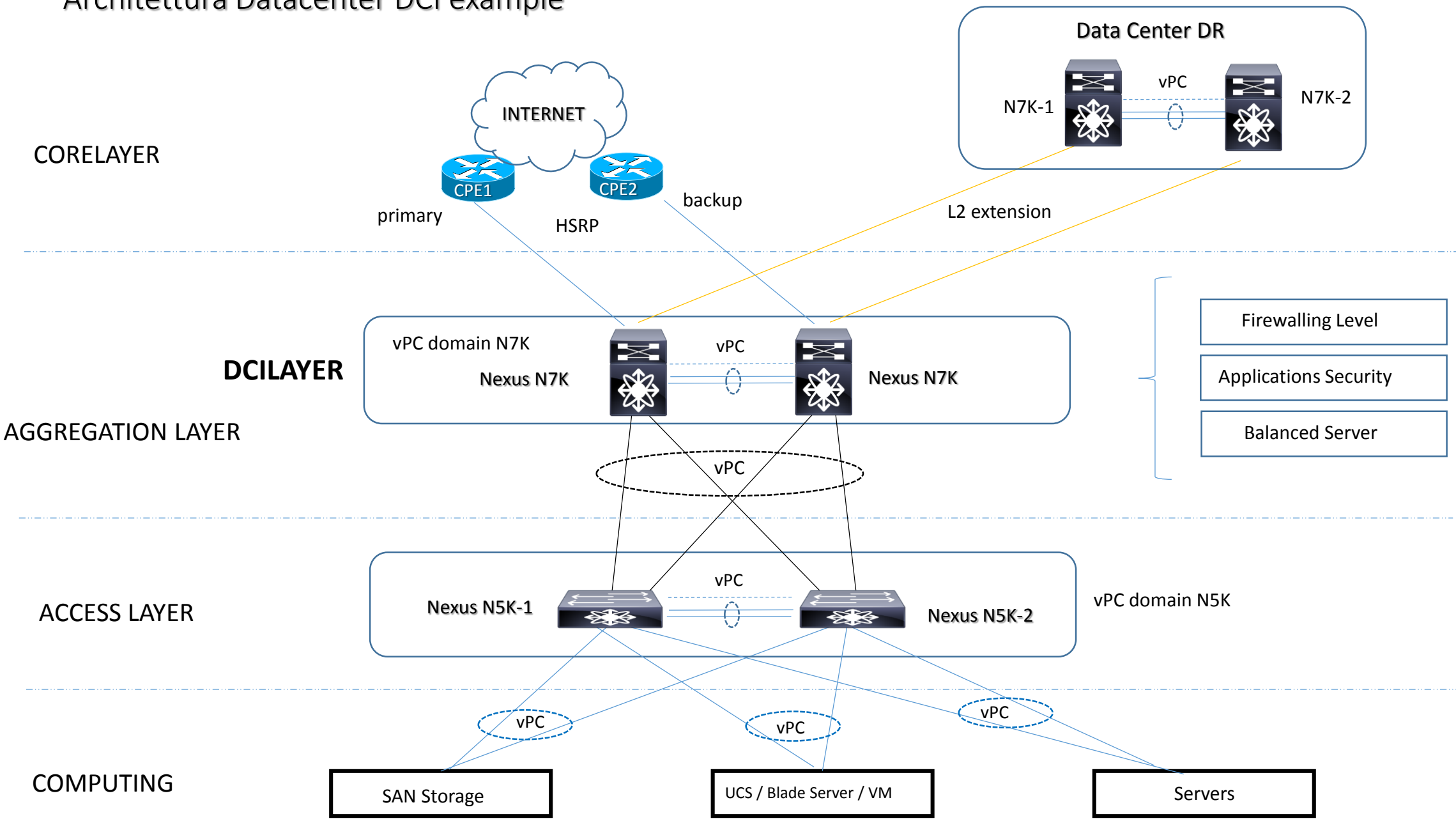
- DCI Layer 2 and Layer 3 concept
- DCI architettura example
- DCI OTV cisco
- DCI OTV cisco example configuration
- DCI OTV multicast enable
- DCI OTV unicast enable
- DCI Layer 2 Dark-Fiber P2P design
- DCI Layer2 Dark-Fiber Ring design
- DCI Layer 2 pseudowire ethernet P2P
- DCI Layer 2 VPLS standard design
- DCI Layer 2 with tunnel GRE design

DCI layer 2 and layer 3 concept

- DCI Layer 2 è inerente a tecniche di mobilità di VM e IP address
- DCI Layer 3 riguarda soprattutto ad operazioni di transazione e replicazione di database in cluster, e la sincronizzazioni di applicazioni in cluster
- Replicazioni Sincrone di dati Storage (generalmente utilizzato all'interno di un solo datacenter) e dipende da fattori quali RPO ed RTO (Recovery Point Object e Recovery Time Object)
- Replicazioni Asincrone di dati Storage (utilizzato tra inter-datacenters via DCI) e dipende sempre da fattori quali RPO ed RTO
- RPO indica la quantità di dati persi che possono essere considerati accettabili dal momento che un fault avviene
- RTO indica la quantità di tempo di ripristino dal momento che un fault avviene



Architettura Datacenter DCI example



DCI OTV CISCO (overlay transport virtualization)

- OTV è una infrastruttura inter-datacenters e provvede a L2 extensions preservando fault-isolation, resilienza e load-balancing;
- Il requisito è che deve esserci connettività IP tra i due datacenters;
- OTV introduce il concetto di Layer 2 MAC routing (MAC in IP) che abilita il piano di controllo (control-plane) di annunciare la raggiungibilità MAC addressess; con il piano di controllo MAC address learning, OTV non trasmette (flood) unknown unicast traffic e il traffico ARP è trasmesso solo in modo controllato;
- OTV non propaga BPDU STP attraverso l'infrastruttura di trasporto overlay;
- OTV utilizza Nexus Cisco con VDC (Virtual Context Domain) ed è mandatorio avere vlans extended con layer 3 SVI (switched virtual interface) per una data vlan;
- La funzionalità site-vlan è utilizzata per la scoperta di edge devices remoti in una topologia multi-homed: in aggiunta al site-vlan, l'edge devices mantiene una seconda OTV adiacenza con gli altri edge devices appartenenti allo stesso datacenter

DCI OTV CISCO (overlay transport virtualization)

- OTV Edge Device: performa le funzionalità e le operazioni OTV; riceve le frame ethernet traffic per tutte le vlans soggette ad L2-extensions tra data centers OTV peers e dinamicamente le incapsula dentro IP packets che sono trasmessi via overlay transport infrastructure;

- OTV internal interface: sono le interfacce di un edge device che connette il datacenter locale con una configurazione generalmente in trunk trasportando multiple vlans. Non prevedono nessuna configurazione OTV compliant;

- OTV join interface: sono le interfacce uplink di un edge device che si affacciano alla rete core overlay IP; questo tipo di interfacce sono point-to-point layer 3 routed, subinterface, port-channel oppure port-channel subinterface (No loopback) ed hanno lo scopo di essere le sorgenti di traffico OTV incapsulato e trasmesso verso l'infrastrutura overlay;

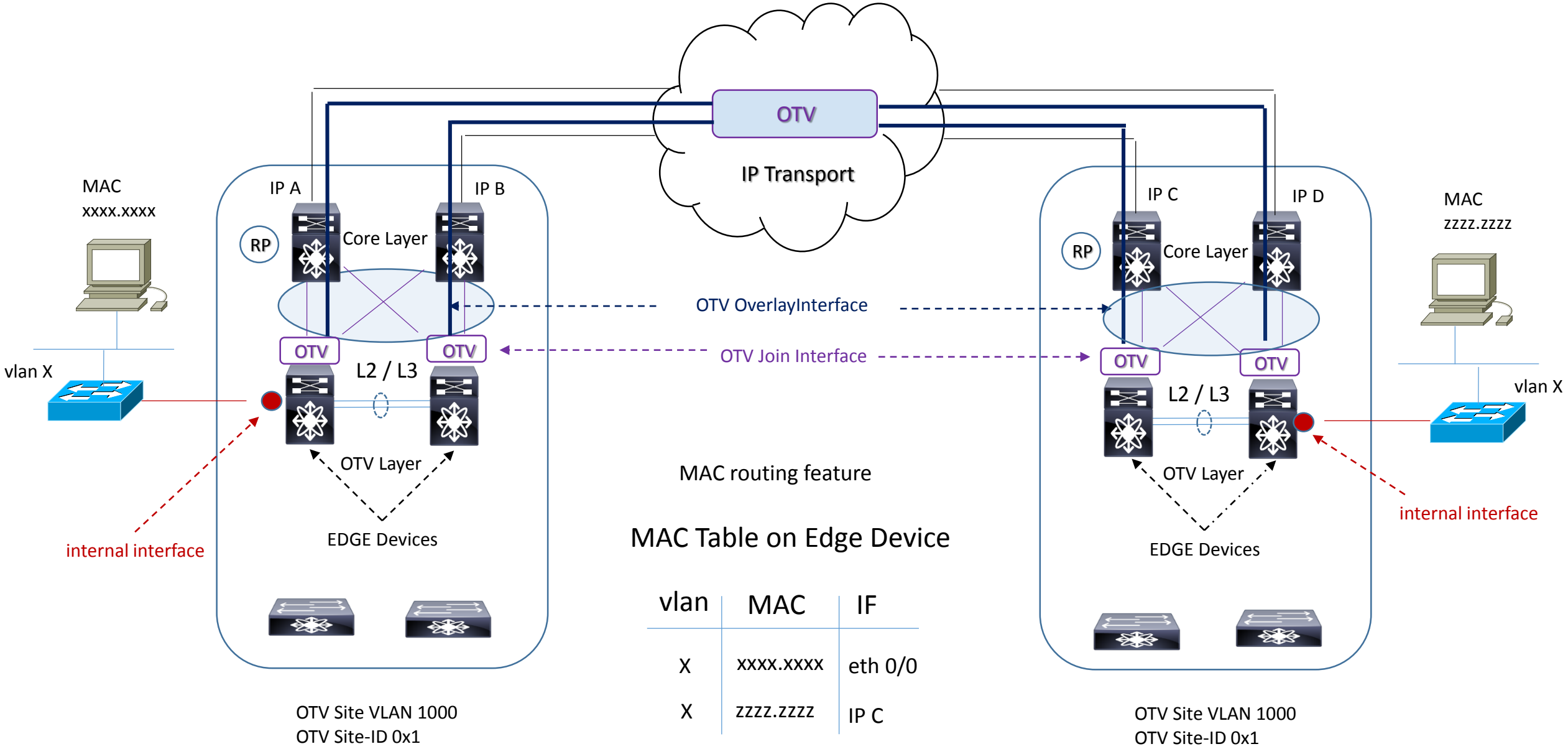
- OTV overlay interface: sono interfacce logiche virtuali dove risiede tutta la configurazione OTV; incapsula le frame layer 2 in IP unicast o multicast packets che sono trasmesse verso altri datacenters. Questo pemette agli edge device di performare un dinamico encapsulations;

- OTV site vlan: è una funzionalità utilizzata per scoprire altri Edge Devices in una topologia multi-homed;

- OTV site ID: sappiamo che le adiancenze OTV sono costruite via le join interface attraverso la rete IP overlay; ogni edge device all'interno dello stesso site hanno lo stesso site-id configurato; dalla release NX-OS 5.2.1 una seconda OTV adiancenza è mantenuta con lo scopo di protezione in caso di partizionamento di site-vlan tra edge devices all'interno dello stesso site;

- AED authoritatve edge device: è responsabile della trasmissione di layer 2 traffic incluso unicast, multicast e broadcast; è responsabile di annunciare la raggiungibilità dei mac-addresses verso i datacenters remoti

DCI OTV CISCO (overlay transport virtualization)



MAC routing feature

MAC Table on Edge Device

vlan	MAC	IF
X	xxxx.xxxx	eth 0/0
X	zzzz.zzzz	IP C

DCI OTV CISCO configurazione internal interface

OTV internal interface:

```
interface port-channel 200
switchport
switchport mode trunk
switchport trunk native vlan 100
switchport trunk allowed vlan 10,12,14,20-30,40-50,70-99,1000
spanning-tree port type normal
mac packet-classify
!
interface  ethernet 3/23
switchport
switchport mode trunk
switchport trunk native vlan 100
switchport trunk allowed vlan 10,12,14,20-30,40-50,70-99,1000
spanning-tree port type normal
channel-group 200 mode active
no shut
!
interface  ethernet 7/23
switchport
switchport mode trunk
switchport trunk native vlan 100
switchport trunk allowed vlan 10,12,14,20-30,40-50,70-99,1000
spanning-tree port type normal
channel-group 200 mode active
no shut
!
```

DCI OTV CISCO configurazione join and overlay interface

OTV join interface:

```
interface port-channel 300
mtu 1600
ip address 172.16.1.1/30
ip ospf network point-to-point
ip router ospf 10 area 0.0.0.0
ip igmp version 3
no shut
!
interface  ethernet 4/16
mtu 1600
channel-group 300 mode active
no shut
!
interface  ethernet 5/18
mtu 1600
channel-group 300 mode active
no shut
!
```

OTV overlay interface:

```
interface overlay 1
otv join-interface port-channel 300
otv control-group 239.1.1.1
otv data-group 232.0.0.0/24
otv extend-vlan 10,12,14,20-30,40-50,70-99
no shut
!
```

OTV MULTICAST enabled transport overlay

OTV Edge Devices sono configurati per unirsi ad uno specifico ASM (Any Source Multicast) group; in questo modo ogni OTV edge devices diventa receiver e source multicast traffic;

Le interfacce in upstream layer 3 debbono essere configurate in PIM sparse-mode ed ogni device deve specificare il SSM group da usare;

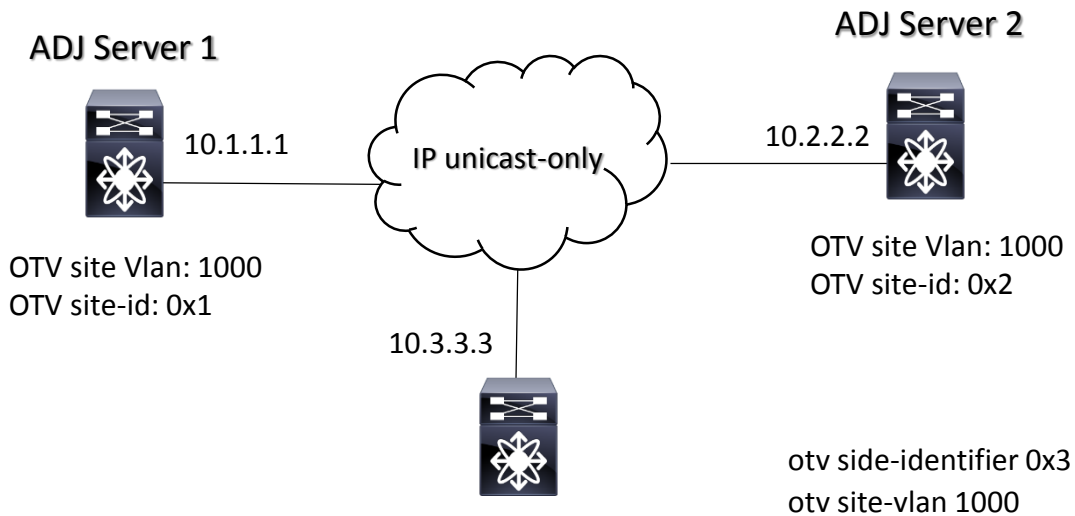
Un RP (Rendezvou Point) router deve essere definito (due RP per ridondanza, dove quest'ultima può essere ottenuta usando Anycast RP);

OTV unicast enabled transport overlay

Nella situazione dove non è possibile avere un Multicast Overlay Transport, è possibile utilizzare un trasporto di tipo unicast-only; la differenza sta che ogni Edge Device deve creare multiple copie di ogni control-plane packet relativo ad ogni edge devices remoto facente parte dello stesso logical overlay interface.

Un nuovo concetto di adiancenza è introdotto: **OTV adjacency server**; ogni OTV device cerca di unirsi ad una specifica logical overlay interface avendo il bisogno di registro verso il server inviando hello message; questi messaggi servono al server per costruire una lista di tutti gli OTV devices che dovranno far parte dello stesso dominio overlay (unicast-replication-list).

```
otv side-identifier 0x1
otv site-vlan 1000
interface overlay 1
 otv join-interface port-channel 300
 otv adjacency-server unicast-only
 otv extend-vlan 10,12,14,20-30,40-50,70-99
```

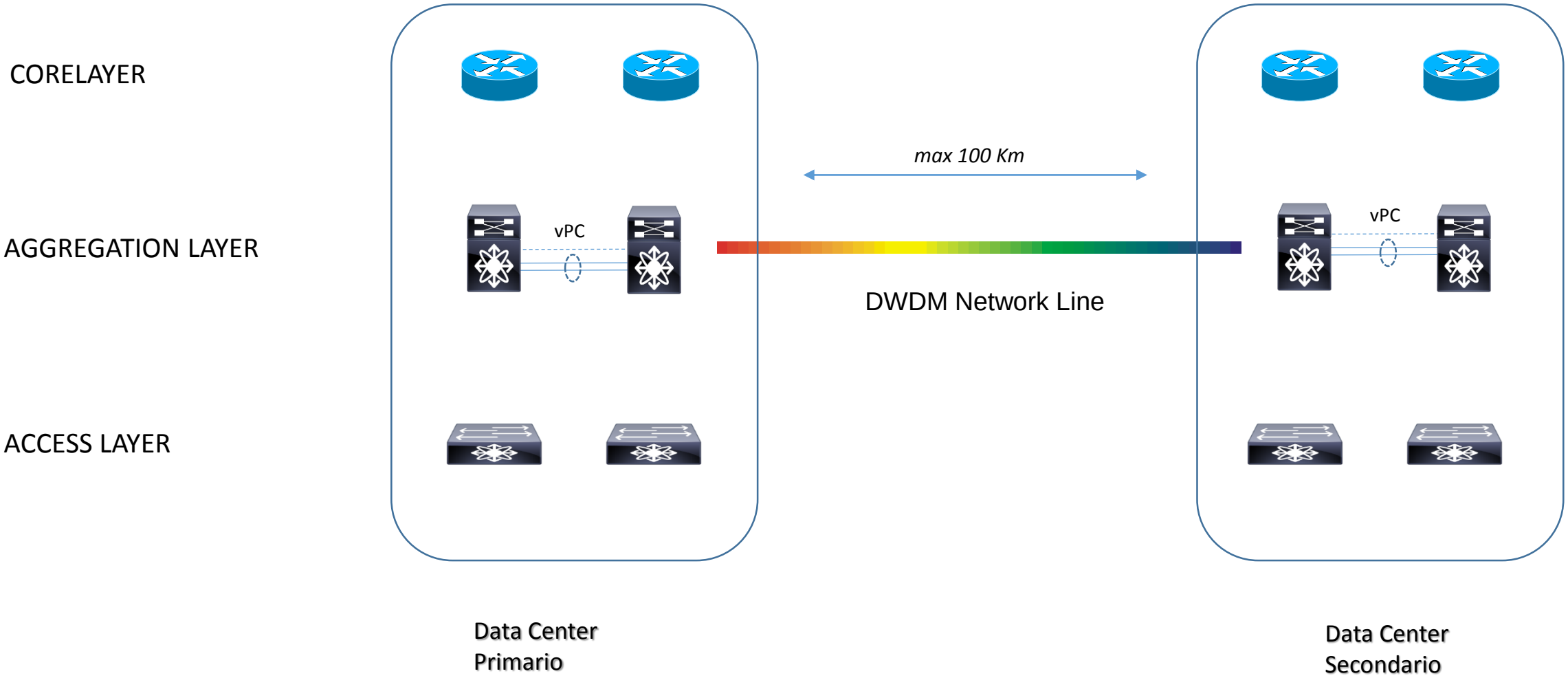



```
otv side-identifier 0x2
otv site-vlan 1000
interface overlay 1
 otv join-interface port-channel 300
 otv adjacency-server unicast-only
 otv use-adjacency-server 10.1.1.1 unicast-only
 otv extend-vlan 10,12,14,20-30,40-50,70-99
```

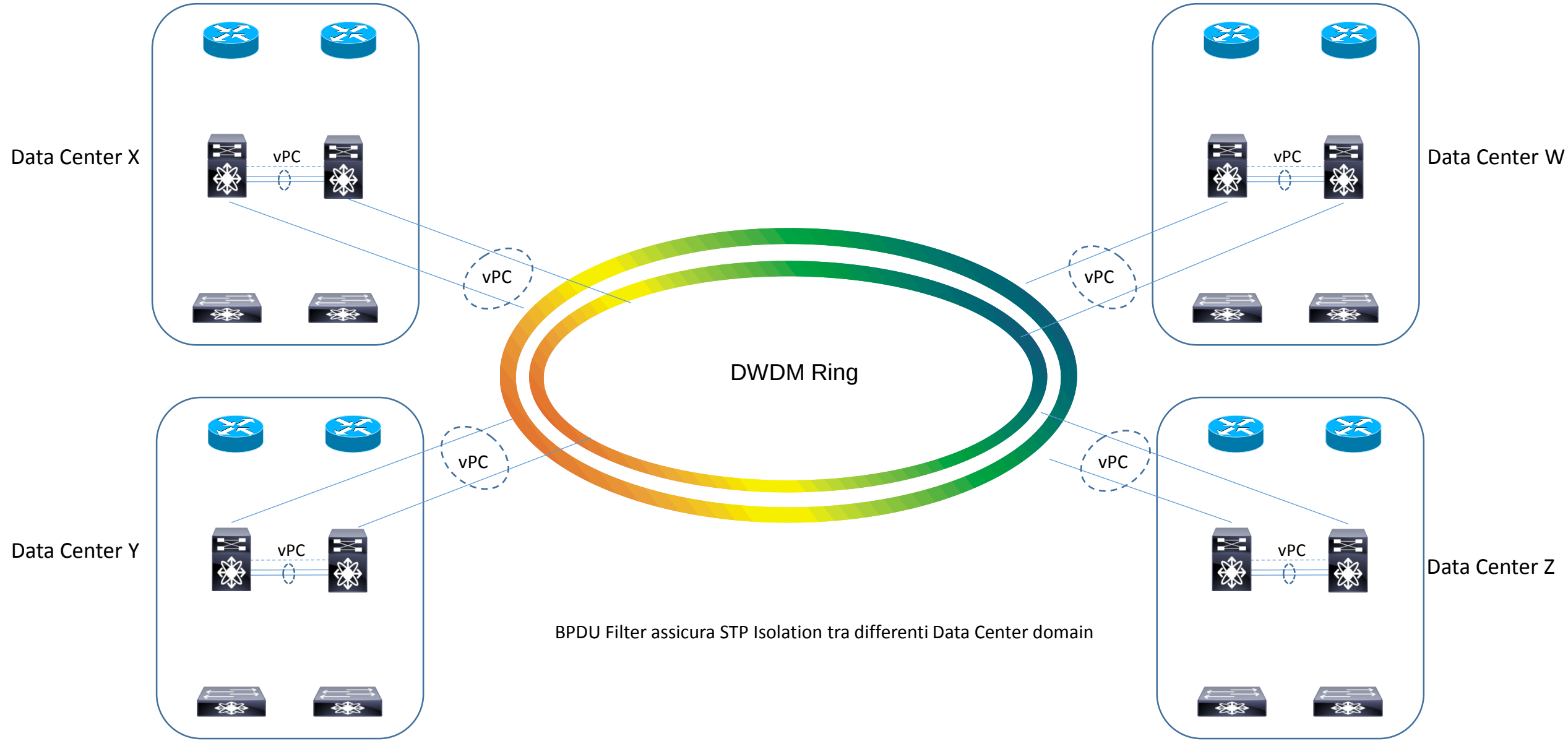
```
otv side-identifier 0x3
otv site-vlan 1000
interface overlay 1
 otv join-interface port-channel 300
 otv use-adjacency-server 10.1.1.1  20.2.2.2 unicast-only
 otv extend-vlan 10,12,14,20-30,40-50,70-99
!
```

DCI layer 2 dark-fiber point-to-point

BPDU Filter assicura STP Isolation tra differenti Data Center domain

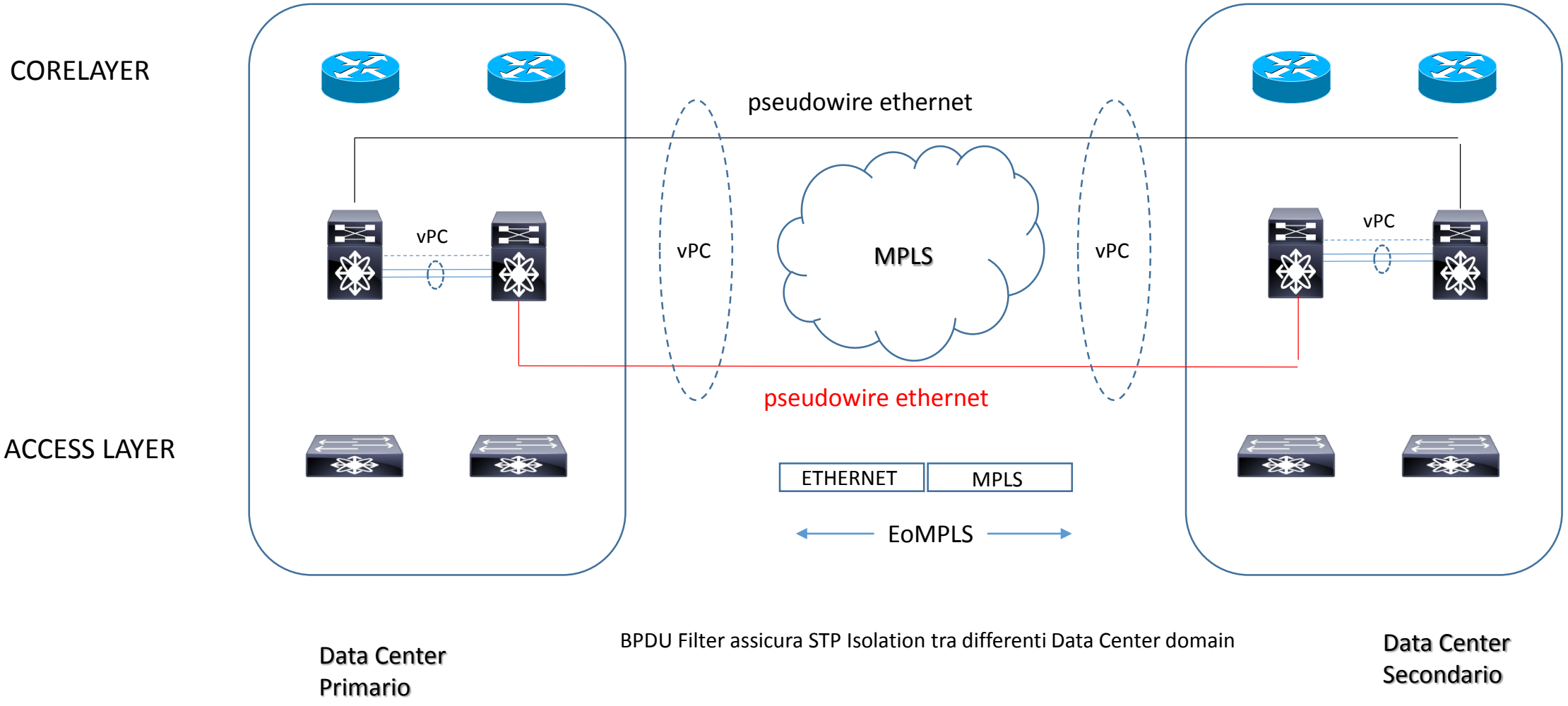


DCI layer 2 dark-fiber ring

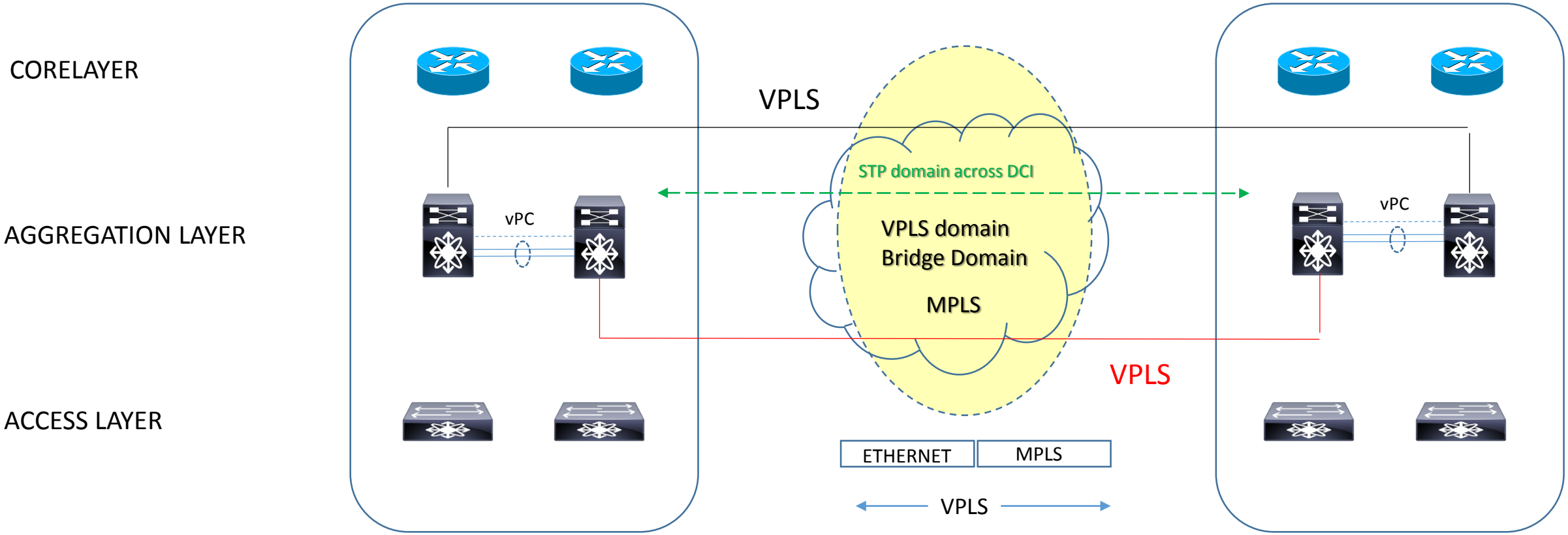


BPDU Filter assicura STP Isolation tra differenti Data Center domain

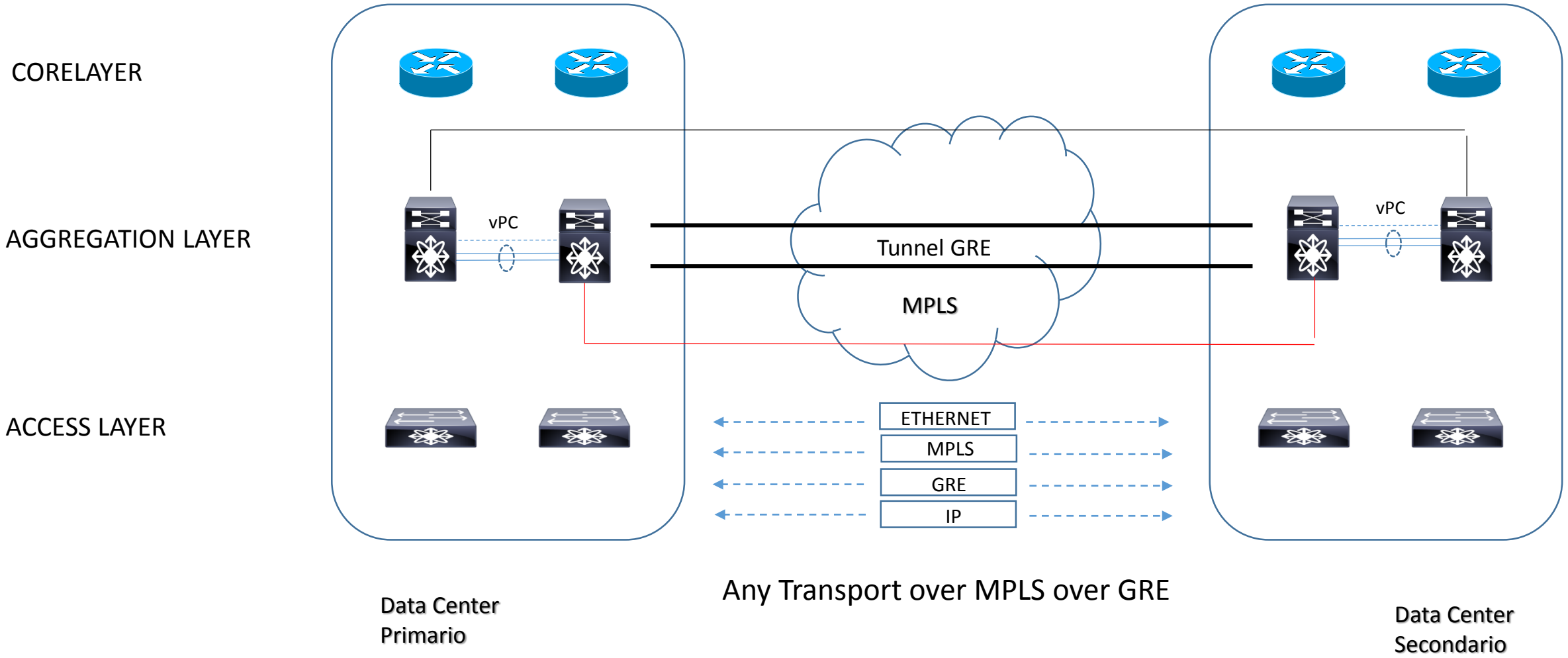
DCI layer 2 pseudowire Ethernet P2P



DCI layer 2 VPLS Ethernet standard



DCI layer 2 with Tunnel GRE



VXLAN

- VXLAN protocol
- VXLAN Header
- VXLAN Considerazioni tecniche
- VXLAN design example

VXLAN protocol

VXLAN (Vlan Extensible LAN) viene utilizzato per i seguenti ambienti:

Data Centers:

- VMware and Vsherevirtualizzazation
- Vmotion
- Multi-Tenant offrendo capacità di scalare la limitazione classica del 802.1q Vlans

VXLAN è un meccanismo che permette di aggregare e tunnellizzare (VTEP) multipli layer 2 subnetwork attraverso una infrastruttura layer 3 IP network:
VXLAN viene supportato da una infrastruttura:

Multicast
IGMP
PIM

IP routing protocols:

OSPF
ISIS
BGP

IP Gateway:

VTEP (Vlan Tunnel End Point) provvede ad incapsulare e decapsulare servizi layer 2 to VXLAN.

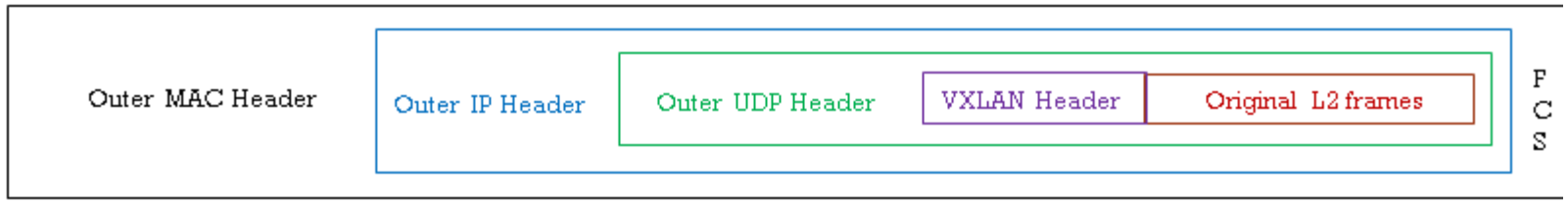
VTEP possono essere:
Virtual Bridges Hipervisor
VXLAN aware VM application
Router/Switch hardware

VXLAN protocol

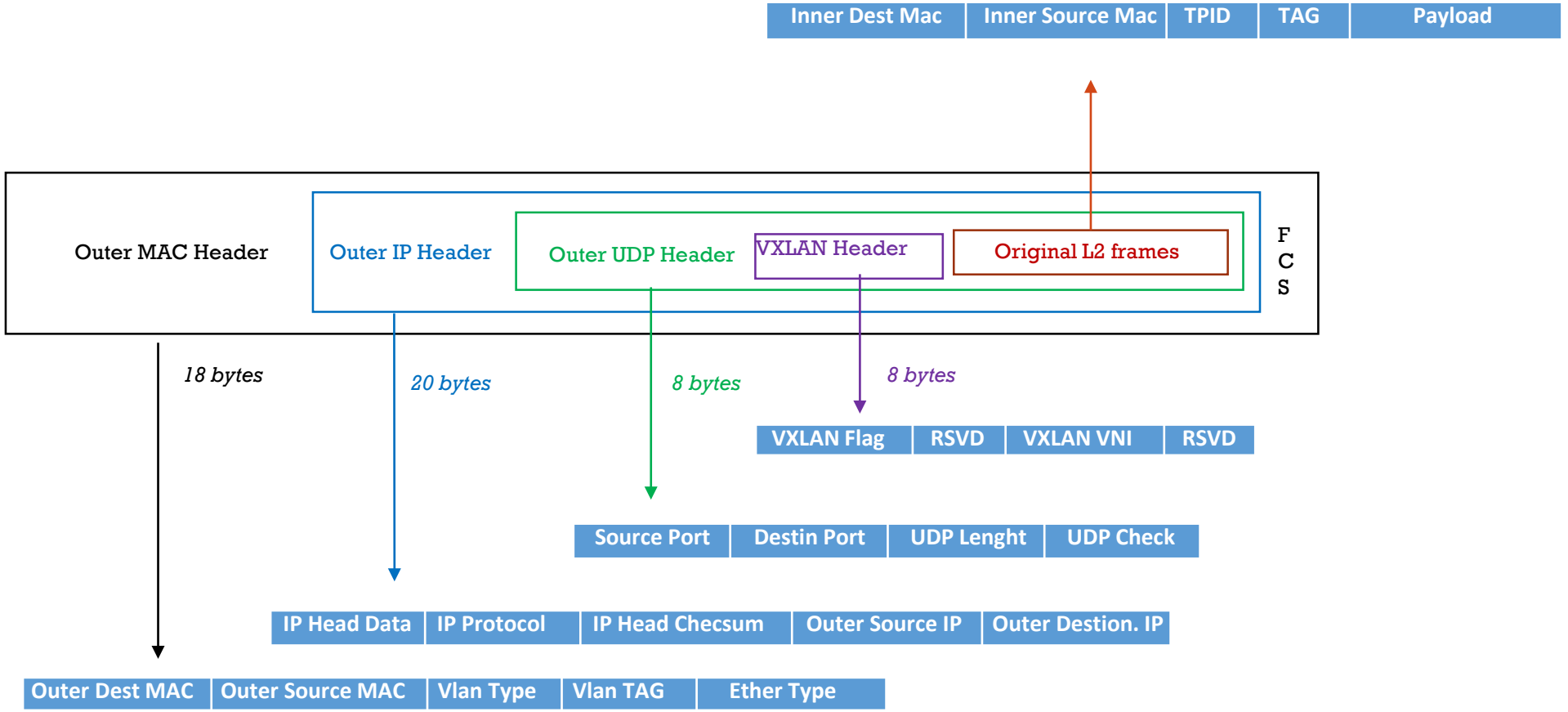
- Ogni VXLAN segment è associato con un unico 24 bit VXLAN Network Identifier differente chiamato VNI;
- Questo 24 bit VNI permette di scalare da il classico 4096 vlans con 802.1q a più di 16 milioni di possibili virtual networks;
- Le VMs servers all'interno di un dominio layer 2 utilizzano la stessa subnet IP e sono mappati con lo stesso valore VNI;
- VXLAN mantiene l'indentità di ciascuna VMs mappando il valore di MAC address della VM con il valore VNI (possiamo avere duplicate MAC address all'interno di un datacenters domain ma con il limite che non possono essere mappati con lo stesso VNI);
- VMs appartenenti ad uno specifico VNI non richiedono speciali configurazioni a supporto perché il meccanisco di encapsulation/de- encapsulation subnets ed il mapping VNI viene gestito dal gateway VTEP;
- Il gateway VTEP deve essere configurato associando il dominio L2 or L3 al VNI network value e quest'ultimo ad un gruppo IP multicast; quest'ultima configurazione permette ai VTEP la costruzione di una forwarding table attraverso l'infrastruttura di rete;
- La sincronizzazione della configurazione VTEP può essere automatizzata grazie a strumenti di gestione quali VMware Orchestrator, Open, Vswitch, Rancid e/o altri.

VXLAN protocol

- Nel caso il MAC sorgente ed il MAC destinazione si trovino nella stesso host, il traffico viene performato all'interno del Vswitch e nessuna azione VXLAN (encapsulation/decapsulation) viene intrapresa;
- Se, invece, il MAC destinazione si trova su altro ESX host, le frames vengono encapsulate in una VXLAN header dal VTEP sorgente e trasmesse al VTEP destinazione, sulla base delle loro informazioni contenute nella forwarding table;
- Per traffico di tipo unknow unicast oppure broadcast/multicast, il VTEP sorgente encapsula il frames in un VXLAN header ed associa esso ad una VNI multicast address (questo include all ARPs request, Boot-p/DHCP request, etc..); i VTEP destinazione (residenti in altri ESX host) ricevono questo multicast frames e lo processano come se fosse un frames unicast.



VXLAN header



VXLAN header format

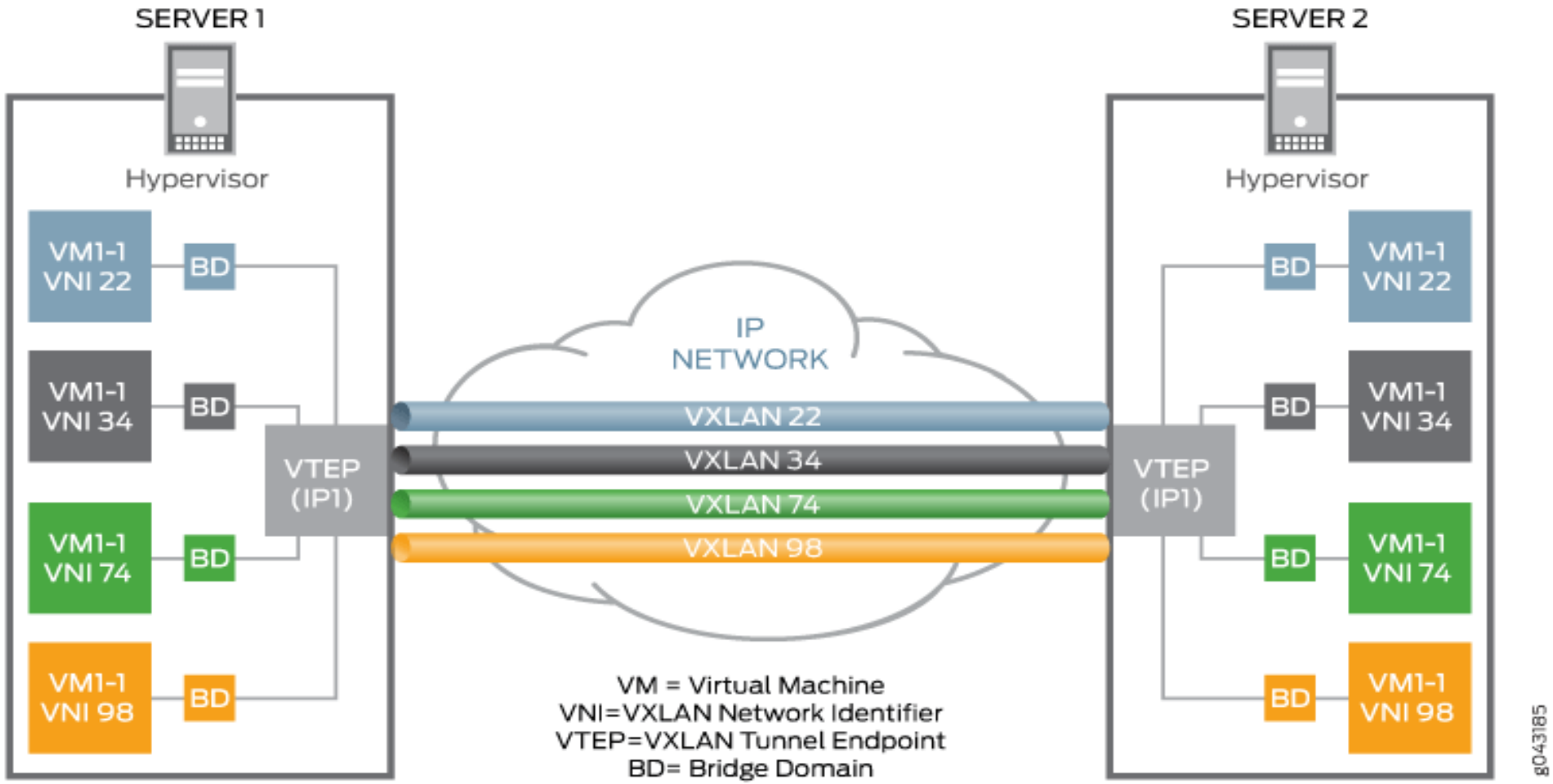
- VXLAN Header:
  - Flag: composto da 8 bits dove il 5° bit (flag) indica un valido valore VNI (i restanti sette bits sono riservato e settati a zero)
  - VNI: valore di 24 bits, provvede a rilasciare un unico identifier per segmento VXLAN; possiamo avere più di 16 milioni di VXLAN segments all'interno di un singolo dominio L2
- UDP Header:
  - Outer UDP: si riferisce alla porta sorgente all'interno dell' outer UDP Header ed è dinamicamente assegnata dal VTEP sorgente; la porta di destinazione è tipicamente la well-know UDP port 4789 (può comunque variare su base implementazione)
  - UDP Checksum: dovrebbe essere settato a zero (0x0000) dal VTP sorgente; nel caso il VTEP destinazione riceve un checksum non uguale a zero, la frame dovrebbe essere scartata
- IP Header:
  - Protocol: settato al valore 0x11 ed indica un UDP packets
  - IP sorgente: è l'indirizzo IP del VTEP sorgente associato con la inner frame source
  - IP destinazione: è l'indirizzo IP del VTEP destinazione corrispondente alla inner frame destination
- Ethernet Header:
  - Outer Ethernet: rappresenta l'indirizzo MAC del VTP sorgente associato con la inner frame source mentre il destination MAC address è l'indirizzo MAC del routing next- hop per raggiungere il VTEP destinazione (l'outer Ethernet header può essere taggato con un IEEE 802.1q per il trasporto in rete)
  - VLAN: default 802.1q tagged protocol identifier
  - Ethertype: settato a 0x0800 per identificare un pacchetto IPv4

VXLAN considerazioni

- VXLAN encapsulaton header aggiunge 50 byte ad un frame Ethernet; pertanto è richiesto l'uso di jumbo frame settato;
- VXLAN richiede una buona quantità di banda per supportare il traffico; è preferibile progettare una rete VXLAN con un troughput di almeno 10Gb;
- L'uso di IP standard aiuta VXLAN ad offrire opzioni di Vmotion VM su lunga distanza e alta affidabilità;
- Assicurare sempre che VXLAN Vmotion /HA heartbit round trip delay non superi la soglia di 10 msec (ad esempio nei casi di disaster recovery oppure mirrored data centers application);
- IP multicast services è usato per pacchetti di tipo unknow unicast, broadcast/multicast all'interno di un dominio VXLAN;
- È da settare sempre un gruppo multicast per ogni VNI segment;
- PIM sparse, Dense sparse e BIDIR (Birectional PIM) provvedono servizi multicast per VXLAN

Feature capability	802.1q VLAN	VXLAN
Number of virtual network	4K: limited by spanning tree	16+ million: limited by number of multicast groups supported by multicast network
Network diameter	As far as 802.1q permitted	As far as PIM multicast groups permitted
Network packet size	1.5K or 9K	Add 50 bytes to VXLAN header
Multicast requirement	NO	PIM, SM, DM, BIDIR (number of group defines number of virtual network
Routing support	Any 802.1q capable router/switch	Any router or switch working with Vmware Vshield, vEdge, and VTEP gateway routers
ARP cache	Limits the VM supported per vlan	Cache on Vmware or VTEP limit VMs supported per VNI
MAC table	VM MAC address count against switch MAC table limits	VTEP MAC address count against switch MAC table limits

VXLAN design example



Architetture CLOS Fabric

- ACI cisco architettura
- ACI cisco control-plane
- ACI cisco policy-based
- ACI cisco access-policy
- ACI cisco L2 step di configurazione
- ACI cisco L2 option extended to external domain
- ACI cisco L3 step di configurazione
- ACI cisco L3 extended to external domain
- EVPN MP-BGP
- EVPN MP-BGP control-plane
- EVPN MP-BGP route-type
- EVPN MP-BGP and ASN underlay design
- EVPN distributed anycast protocol gateway
- EVPN learning process end-point information
- EVPN intra-subnet and inter-subnet communication
- EVPN I-BGP config example VTEP VXLAN
- EVPN I-BGP config example control-plane overlay
- EVPN I-BGP config example VLAN to VXLAN
- EVPN I-BGP config example routng resource on VXLAN
- EVPN I-BGP config example distributed IP anycast
- EVPN I-BGP config example routing on VXLAN
- EVPN I-BGP config example IGP (OSPF)

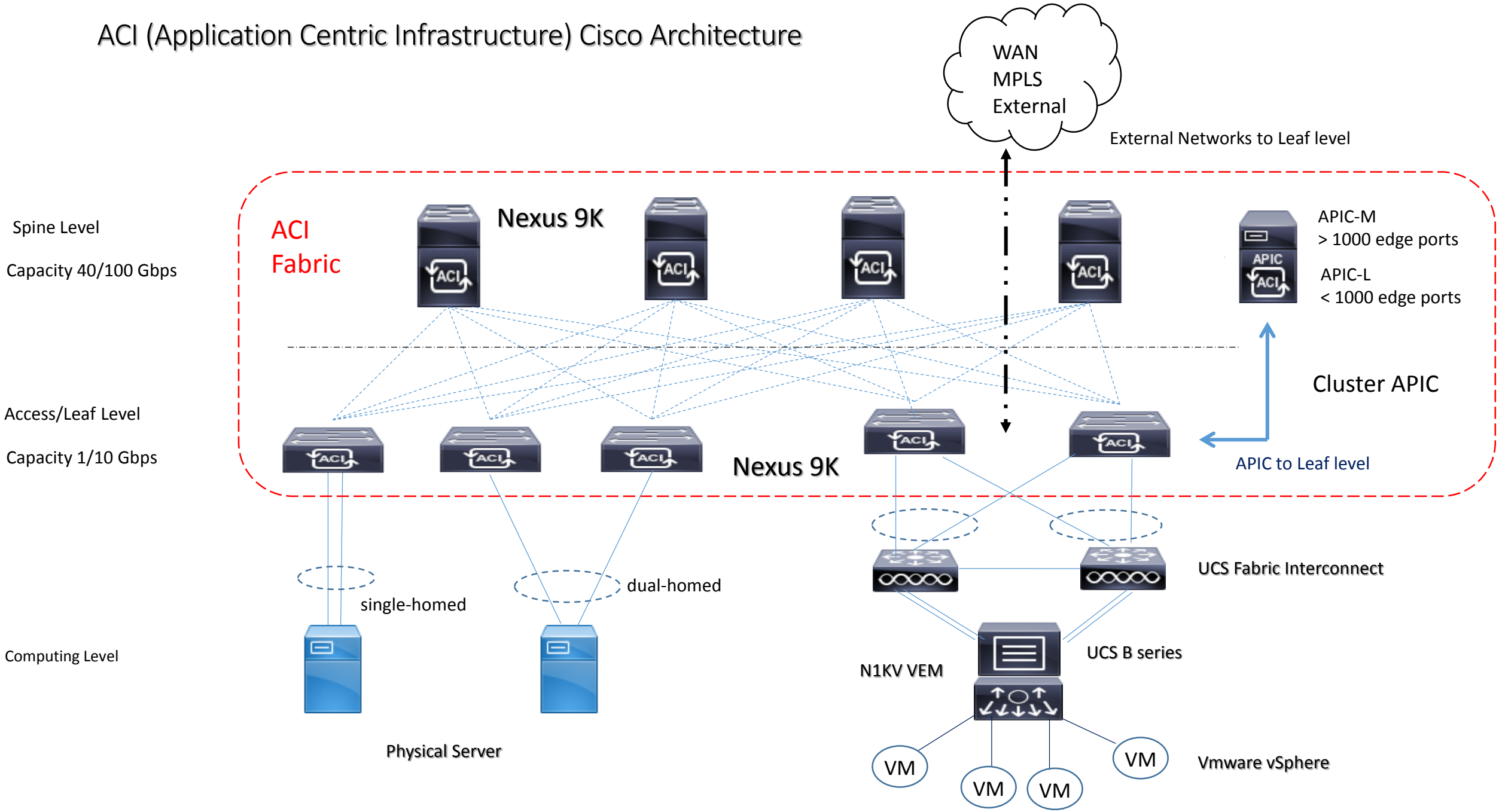
ACI (Application Centric Infrastructure) Cisco

- Cisco ACI (Application Centric Infrastructure) è basato sul concetto di group-based policy SDN;
- End-User ACI può definire una serie di regole senza la conoscenza e/o informazioni che derivano dalla struttura networking;
- Cisco APIC (Application Policy Infrastructure Controller) è responsabile della gestione centralizzata delle policies configurate e distribuirle a tutti i nodi facenti parte della ACI Fabric;
- Cisco ACI è disegnato per scalare in modo trasparente nei confronti di cambiamenti di connettività, bandwidth, tenants e policies; la sua architettura è di tipo spine-leaf che si presta efficientemente a introdurre e/o cambiare requisiti di rete;
- Cisco ACI include servizi layer 4 to layer 7, APIs (Application Programming Interface), virtual networking, computing, storage resources, wan routers, orchestration services.

Cisco ACI consiste in:

- Un insieme di software e hardware devices che costituiscono una Fabric
- APIC per la gestione delle policies centralizzata
- AVS (Application Virtual Switch) per virtual network edge level
- Integrazione di fisiche e virtuali infrastrutture
- Un aperto ecosistema di network, storage, management e orchestration vendor

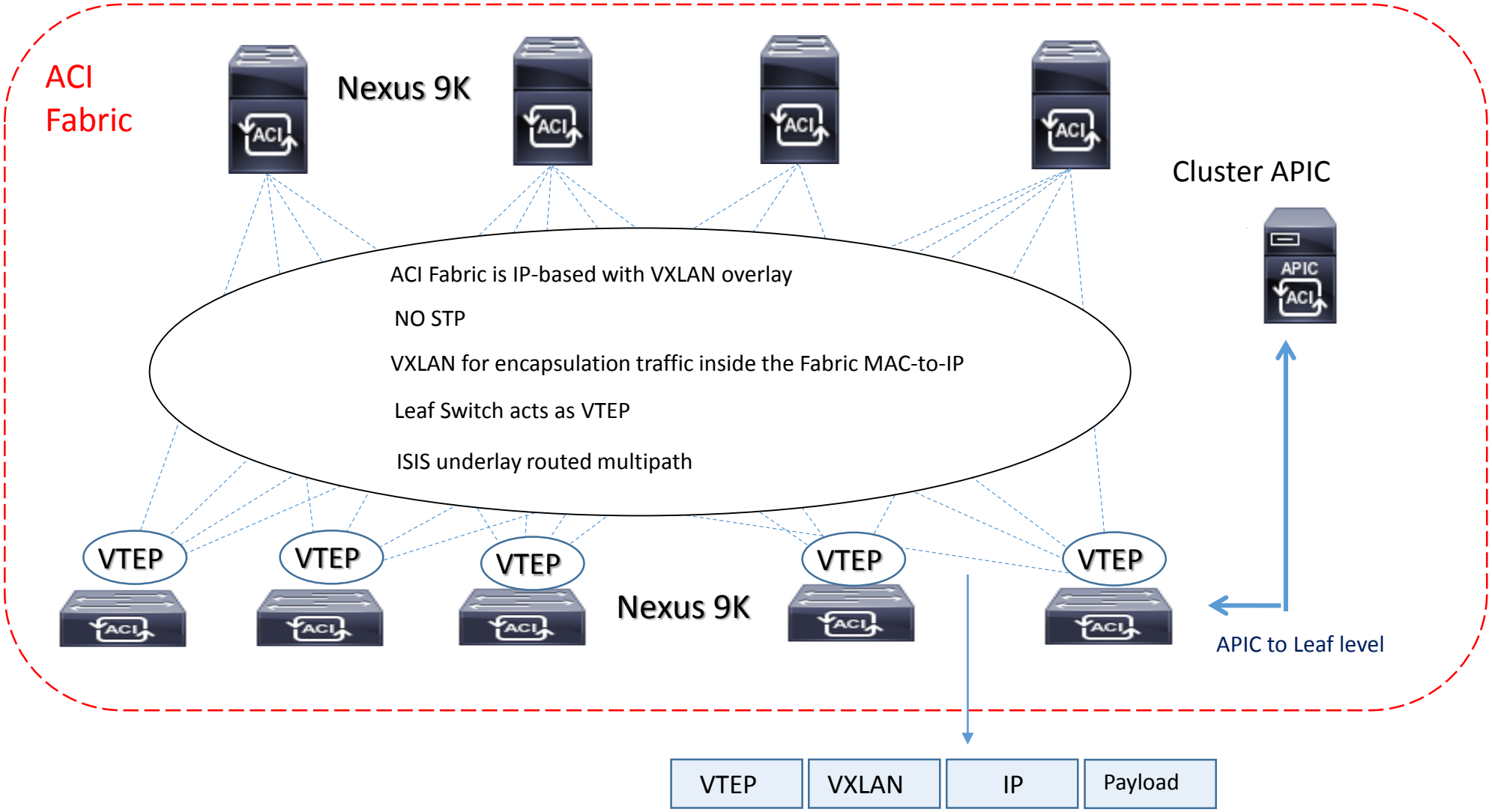
ACI (Application Centric Infrastructure) Cisco Architecture



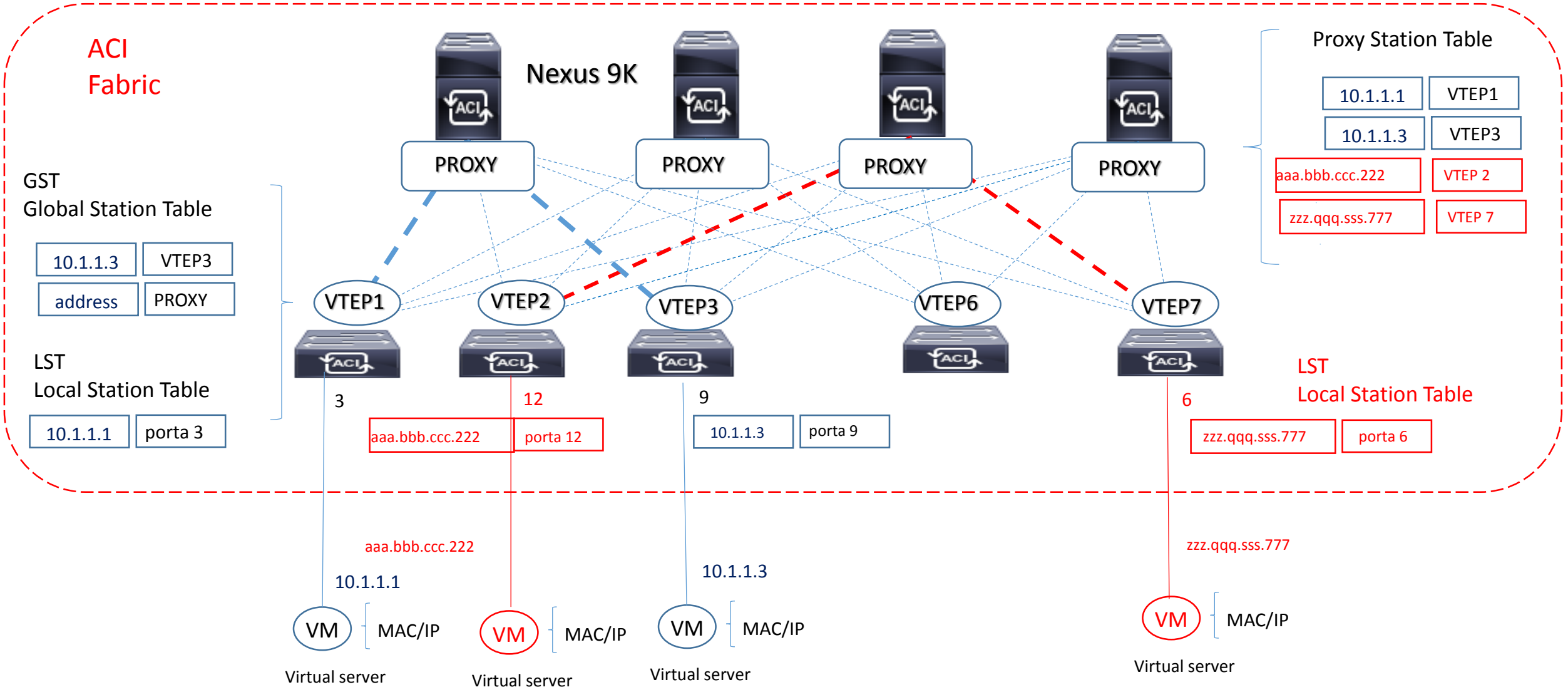
ACI (Application Centric Infrastructure) Cisco Architecture

Spine Level

Capacity 40/100 Gbps



ACI (Application Centric Infrastructure) Cisco Control-Plane with mapping database



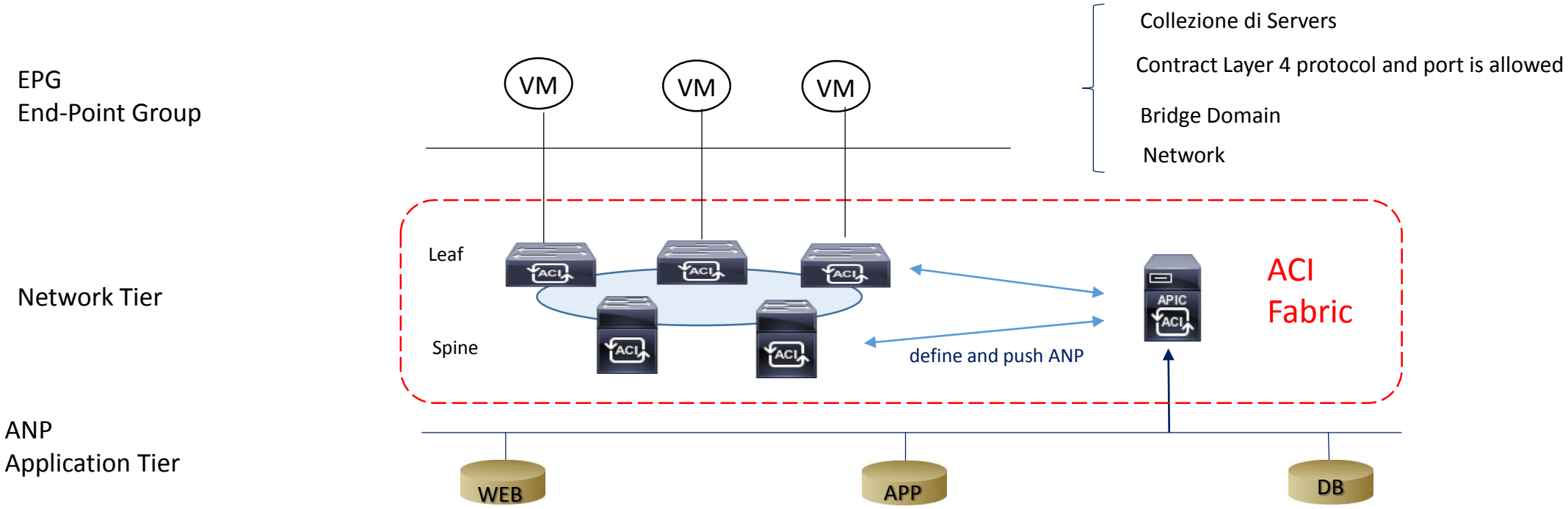
ACI (Application Centric Infrastructure) Cisco Policy Based

Cisco APIC (Application Policy Infrastructure Controller): è responsabile della gestione centralizzata delle policies configurate e distribuirle a tutti i nodi facenti parte della ACI Fabric;

ANP (Application Network Profile): contiene le policies dei sistemi applicativi;

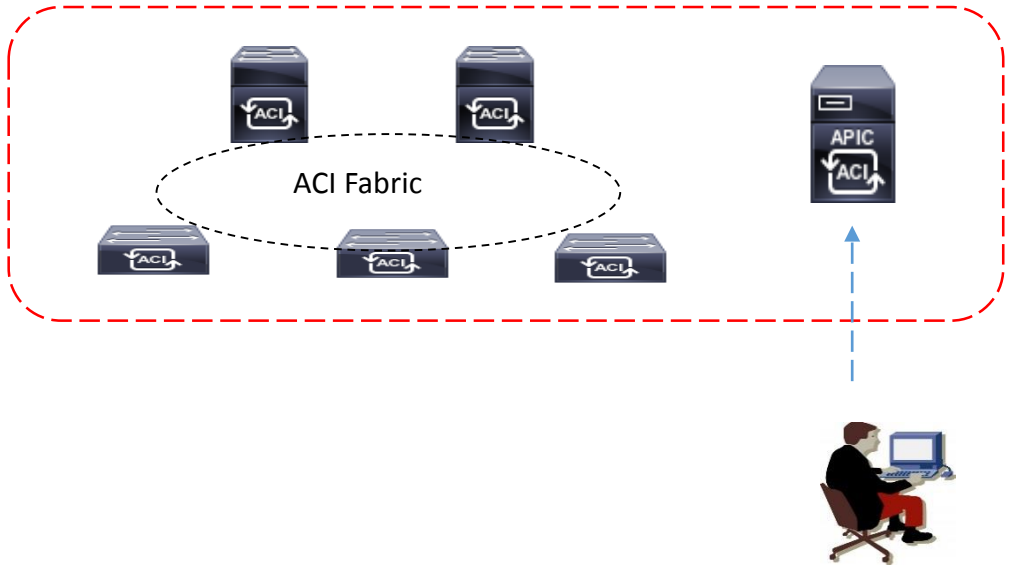
EPG (End Point Group): consiste di un numero di end-point groups rappresentati da uno o più servers all'interno di uno stesso segmento di rete (vlans);

Contract: consiste di policies che definiscono il modo con cui comunicano tra loro gli EPG.



ACI (Application Centric Infrastructure) Cisco Access Policy

- vlan pool: definisce un singolo segmento di rete (vlan) oppure un pool di vlans;
- Physical Domain: definisce un dominio (scopo) dove è creato il vlans pool;
- AAEP (Attachabe Access Entity Profile): definisce un modo di raggruppare multipli domini applicabili ad un profilo su base interfaccia;
- Interface Policy and Profile: questa policy definisce i parametri richiesti come può essere un LLDP, LACP, etc; contiene la interface policy e specifica a quale port number deve essere applicata usando la port-selector;
- Switch Profile: applica il profilo su base interfaccia con la policy associata ad uno o più multiple access Leaf Nodes



Switch Profile

Interface Policy and Profile

Attachable Access Entity Profile (AAEP)

Physical Domain

Vlan Pool

VM MAC/IP

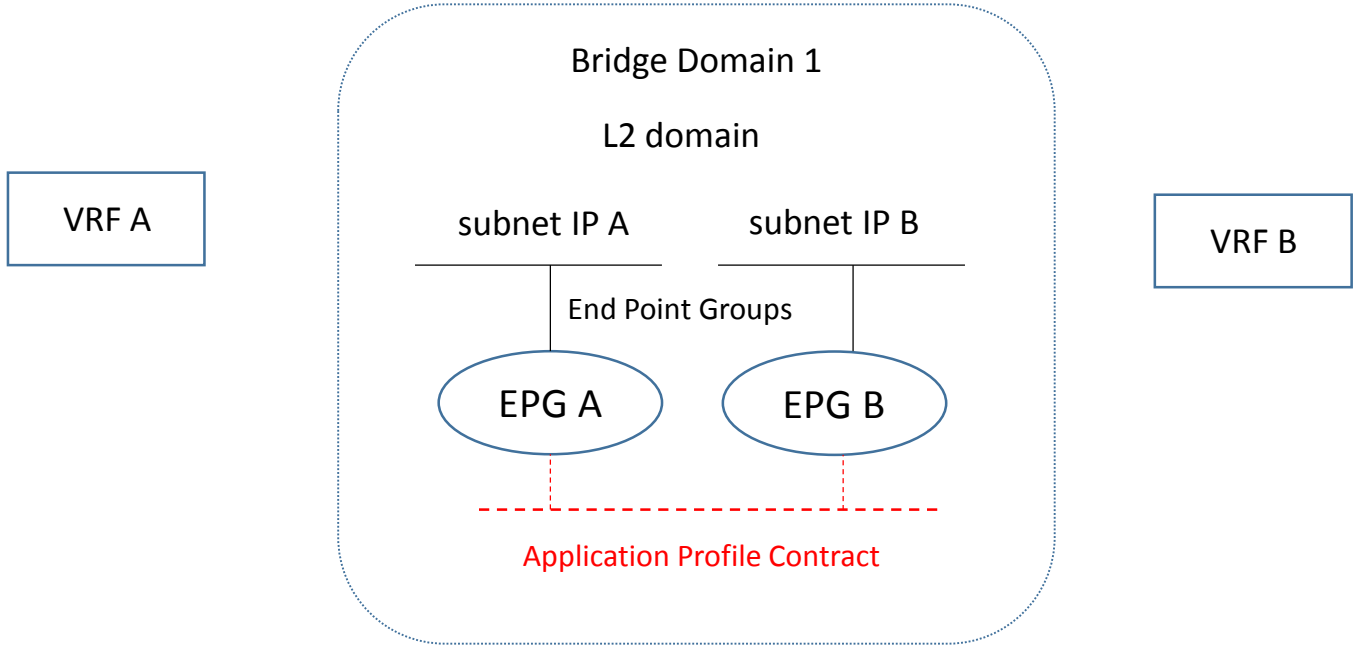
Virtual server

VM MAC/IP

Virtual server

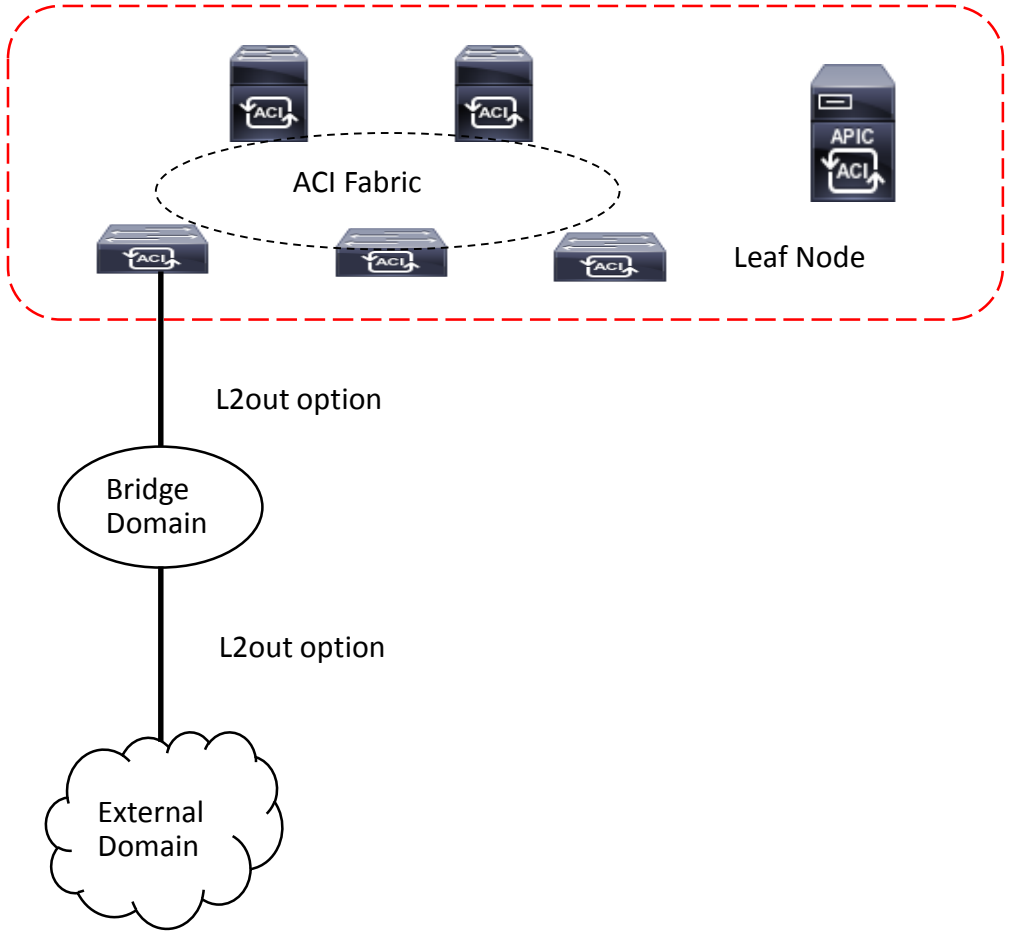
ACI (Application Centric Infrastructure) Cisco layer 2 steps di configurazione

- VRF instances
- BD (Bridge Domain) associato alla VRF instance (senza abilitare nessun layer 3 IP SVIs subnet)
- Configurazione del Bridge Domain per ottimizzare la funzionalità di switching (hardware-proxy-mode) usando il mapping database oppure il tradizionale flood-and-learn
- EPG (End Point Group) relazionandoli ai bridge domain di riferimento; possiamo avere multipli EPG associati allo stesso bridge domain
- Creare policy Contracts tra EPG come necessario; possiamo anche considerare una comunicazione tra diversi EPG senza ausilio di filtri, settando la VRF instance in modalità < unenforced >
- Creare access policies switch e port profiles assegnando i parametri richiesti, associate al nodo Leaf di pertinenza



ACI (Application Centric Infrastructure) Cisco layer 2 option extending to external domain

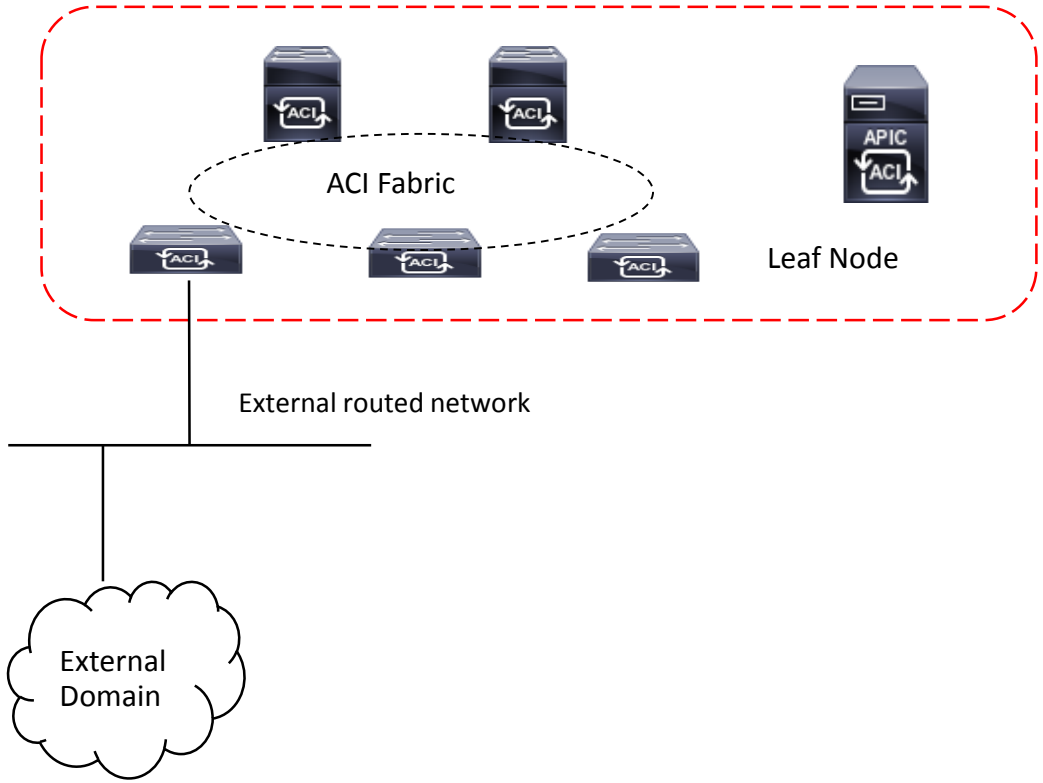
- Enable fllooding of layer 2 unknown unicast
- Enable ARP flooding
- Disable unicast routing (può essere abilitato successivamente ad una fase di migrazione ad esempio se gli end-point usano come IP gateway il sistema ACI Fabric
- L2Out option provvede ad una L2 extension da ACI Fabric ad un External domain bridged network



ACI (Application Centric Infrastructure) Cisco layer 3 steps di configurazione

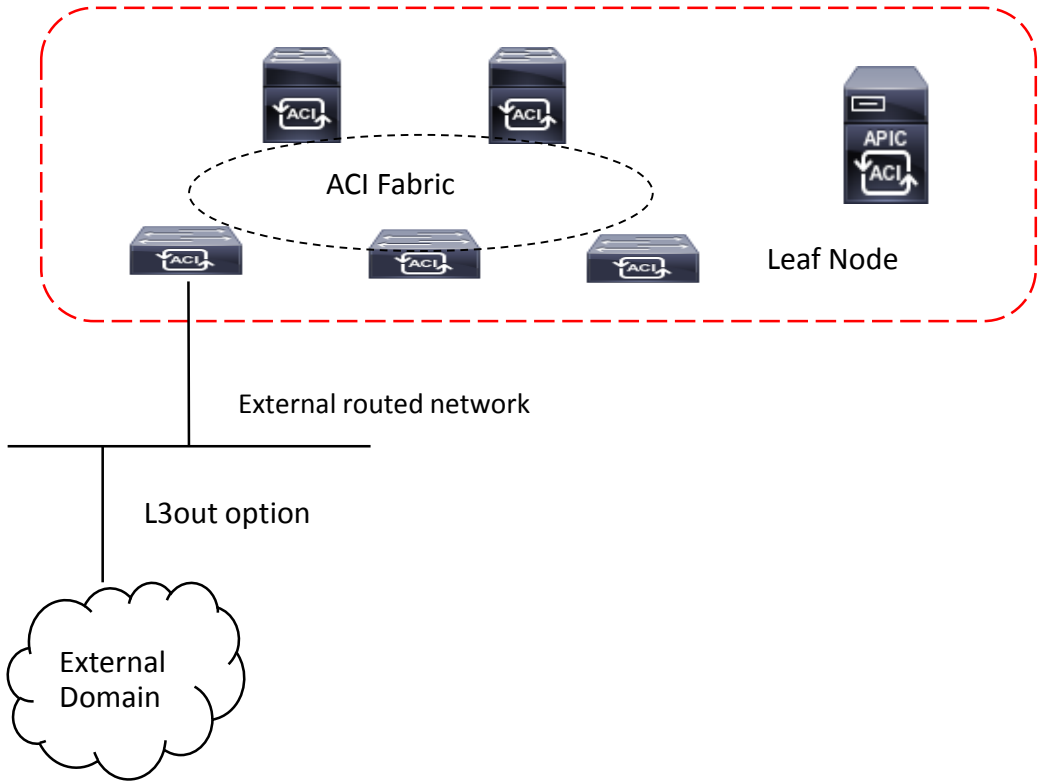
- **Layer 3 interface routed:** usata quando si connette un determinato external devices per tenant /VRF
- **Subinterface with 802.1q tagging:** usata quando vi è una connessione condivisa ad un determinato external devices attraverso tenants/ VRF-lite
- **Switched Virtual Interface (SVI):** usata quando entrambi i layer L2 ed L3 di connessione sono richiesti sulla stessa interfaccia

La propagazione di external network all'interno di un dominio ACI Fabric utilizza il MP-BGP (Multi Protocol BGP) tra Spine e Leaf (si può avere anche la funzionalità di Route Reflector abilitato a livello Spine) all'interno di un unico AS



ACI (Application Centric Infrastructure) Cisco layer 3 option extending to external domain

- Create un external routed network
- Set a layer 3 border leaf node for the L3 outside connection
- Set a layer 3 interface profile for the L3 outside connection
- Repeat step 2 and 3 if you need to add additional leaf nodes/interface
- Configure an external EPG (ACI Fabric maps the external L3 router to the external EPG by using the IP prefix and mask
- Configure a contract policies between the external and internal EPG (without this all connectivity to the outside will be blocked)



EVPN MP-BGP

EVPN (Ethernet Virtual Private Network) collega un gruppo di users sites usando un virtual bridge layer 2;

Tratta indirizzi MAC come address ruotabili e distribuisce queste informazioni via MP-BGP;

Utilizzato in ambienti Data Centers multi-tenancy con end-point virtualizzati; supporta encapsulamento VXLAN e lo scambio di indirizzi IP host e IP-Prefix.

EVPN MP-BGP control plane

- informazioni layer 2 (MAC address) e layer 3 (host IP address) imparate localmente da ogni VTEP sono propagate ad altri VTEP permettendo funzionalità di switching e routing all'interno della stessa fabbrica;
- le routes sono annunciate tra VTEP attraverso route-target policy;
- utilizzo di VRF e route-distinguisher per routes/subnet;
- Le informazioni layer 2 sono distribuite tra VTEP con la funzionalità di ARP cache per minimizzare il flooding;
- le sessioni L2VPN EVPN tra VTEP possono essere autenticate via MD5 per mitigare problematiche di sicurezza (Rogue VTEP)

In genere un data centers IaaS costruito su una architettura Spine-Leaf utilizza per migliorare le sue performance di raggiungibilità layer 2 e 3 un processo ECMP (Equal Cost Multi Path) via IGP.

In caso di crescita della Fabric con la separazione multi-tenant, si può pensare a meccanismi di scalabilità come il protocollo BGP e scegliere se utilizzare Internal-BGP oppure external-BGP in considerazione anche di meccanismi ECMP molto utili in ambienti datacenters

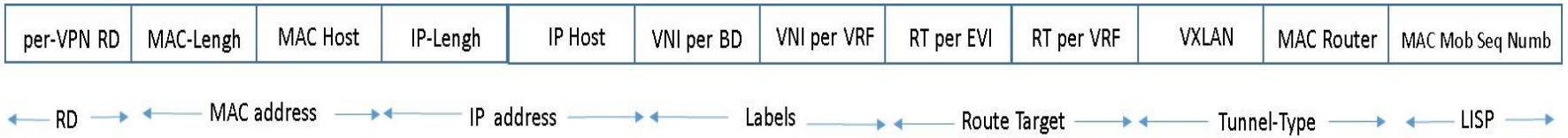
IBGP richiede sessioni tra tutti i PE VTEP e l'impiego di Router Reflector aiuta molto in termini di scalabilità delle sessioni configurati a livello Spine; questo tipo standard di soluzione, in ogni caso, riflette solo il best-single-prefix verso i loro client ed nella soluzione di utilizzare ECMP bisogna configurare un BGP add-path feature per aggiungere ECMP all'interno degli annuncia da parte dei RRs

EBGP, invece, supporta ECMP senza add-path ed è semplice nella sua tradizionale configurazione; con EBGP ogni devices della Fabric utilizza un proprio AS (Autonomous System)

EVPN MP-BGP route-type

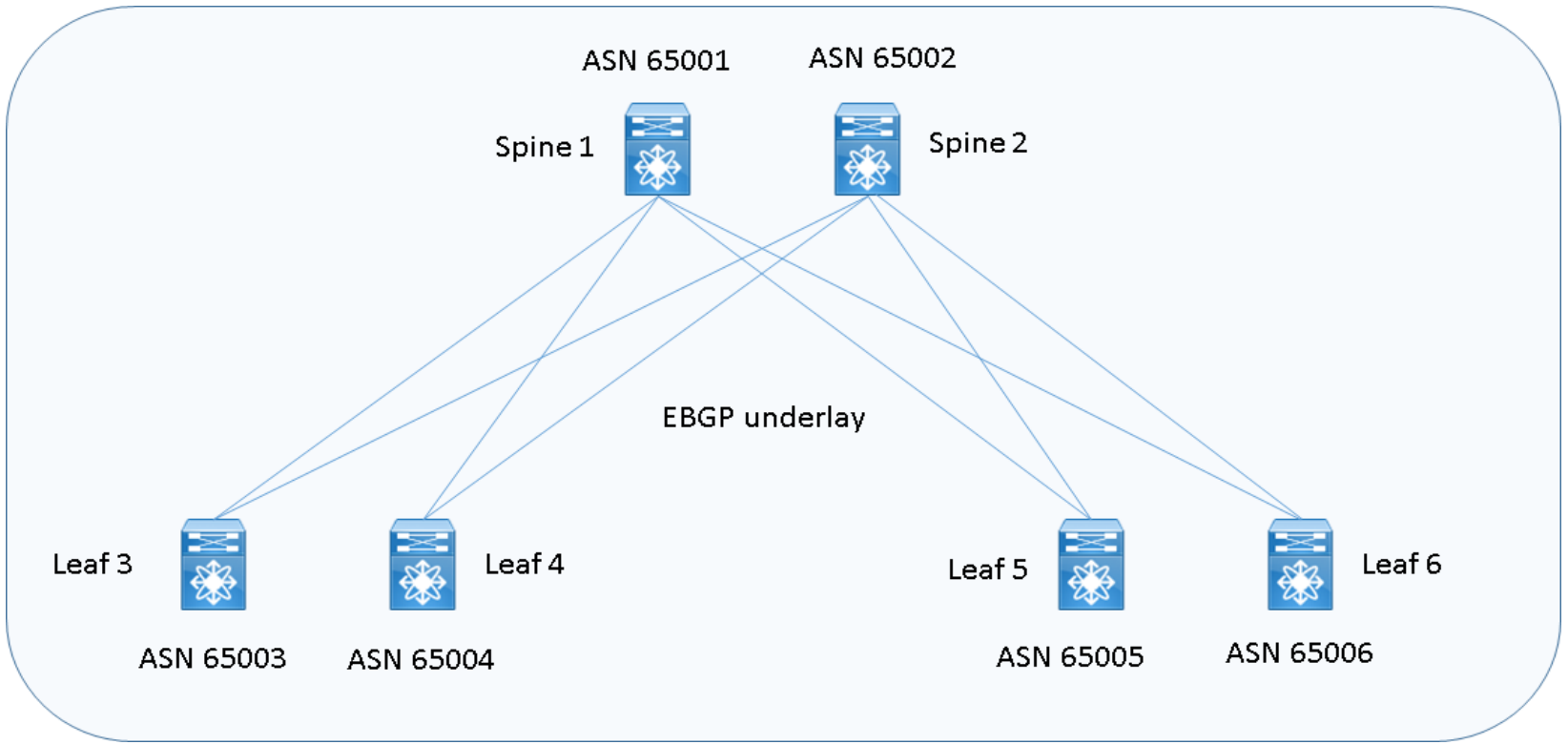
MP-BGP EVPN utilizza due routing advertisement:

- ✓ **Route type 2:** usato per annunciare host MAC ed IP address information per gli endpoint direttamente collegati alla VXLAN EVPN Fabric, ed anche trasportare extended community attribute, come route-target, router MAC address e sequence number

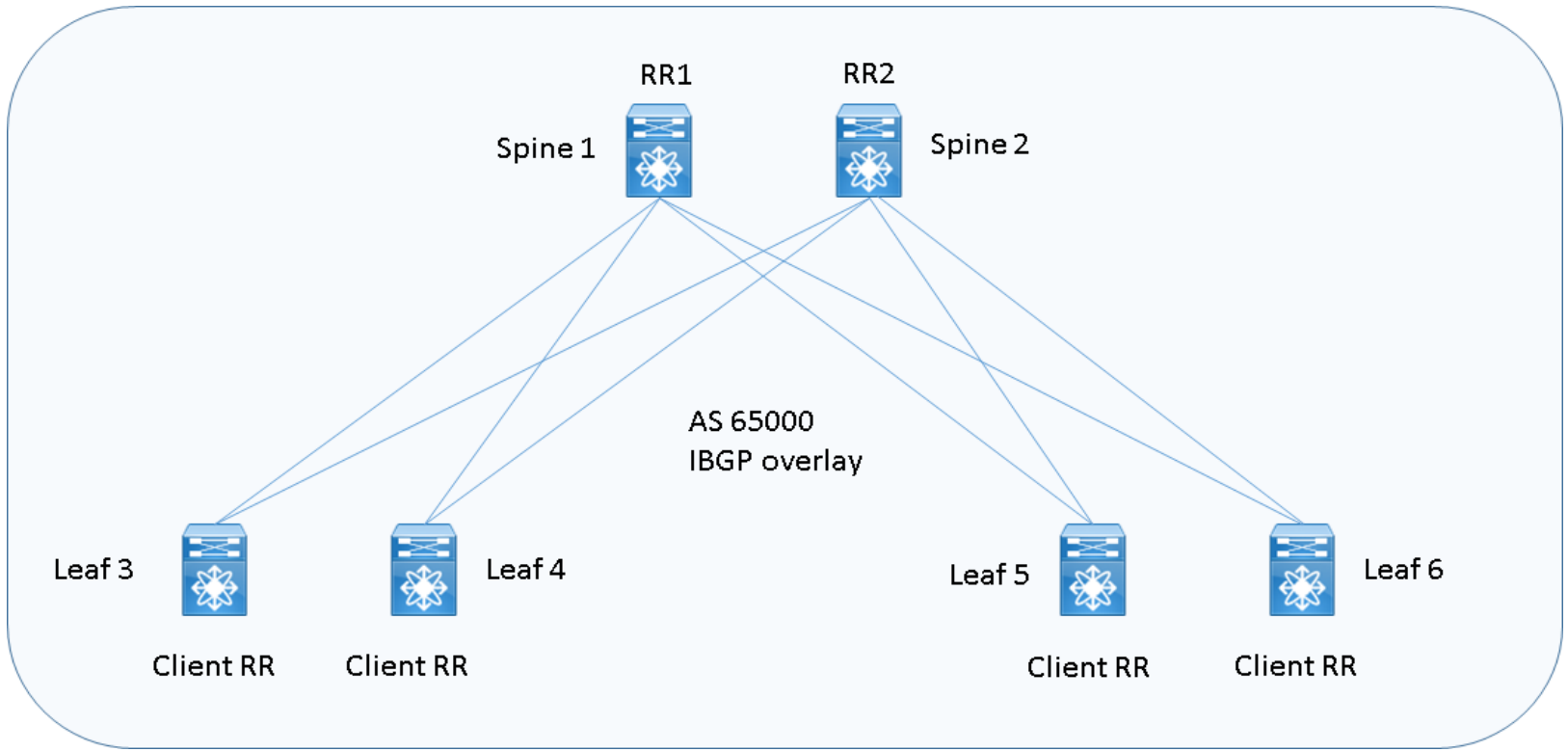


- ✓ **Route type 5:** annuncio di IP Prefix oppure host routes (loopback interface) ed anche trasporto di extended community attribute, come route-target, router MAC address e sequence number

EVPN E-BGP and ASN underlay design



EVPN I-BGP and ASN underlay design



Distributed Anycast Protocol Gateway

Protocolli FHRP quali HSRP, VRRP e GLBP hanno funzionalità di alta affidabilità layer 3 attraverso meccanismi active-standby routers e VIP address gateway condiviso.
Distributed Anycast Protocol, supera la limitazione di avere solo due routers peers HSRP/VRRP in ambienti Data Centers, costruendo una VXLAN EVPN VTEP Fabric con una architettura di tipo Spine-Leaf.

Distributed Anycast Proocol offre i seguenti vantaggi:

- ✓ stesso IP address gateway per tutti gli Edge Switch; ogni endpoint ha come gateway il proprio local VTEP il quale ruota poi il traffico esternamente ad altri VTEP attraverso una rete IP core (questo vale sia per VXLAN EVPN costruito come Fabric locale che geograficamente distribuito);

- ✓ la funzionalità di ARP suppression permette di ridurre il flooding all'interno del proprio dominio di switching (Leaf to Edge Switch);

- ✓ permette il moving di host/server continuando a mantenere lo stesso IP address gateway configurato nel local VTEP, all'interno di ciascuna VXLAN EVPN Fabric locale o geograficamente distribuita;

- ✓ No FHRP Filtering tra VXLAN EVPN Fabrics

Learning Process End-Point information

Il processo di learning Endpoint avviene a livello Edge Switch Leaf Node di una VXLAN EVPN Fabric, dove l'endpoint è direttamente connesso; le informazioni MAC address a livello locale sono calcolate attraverso la tabella di forwarding locale (data-plane table) mentre l'IP address è imparato attraverso meccanismi di ARP, GARP (Gratitous ARP) oppure IPv6 neighbor discovery message.

Una volta avvenuto il processo di apprendimento MAC + IP a livello locale, queste informazioni vengono annunciate dai rispettivi VTEP attraverso il MP-BGP EVPN control-plane utilizzando le EVPN route-type 2 advertisement trasmette a tutti i VTEP Edge devices che appartengono alla stessa VXLAN EVPN Fabric.

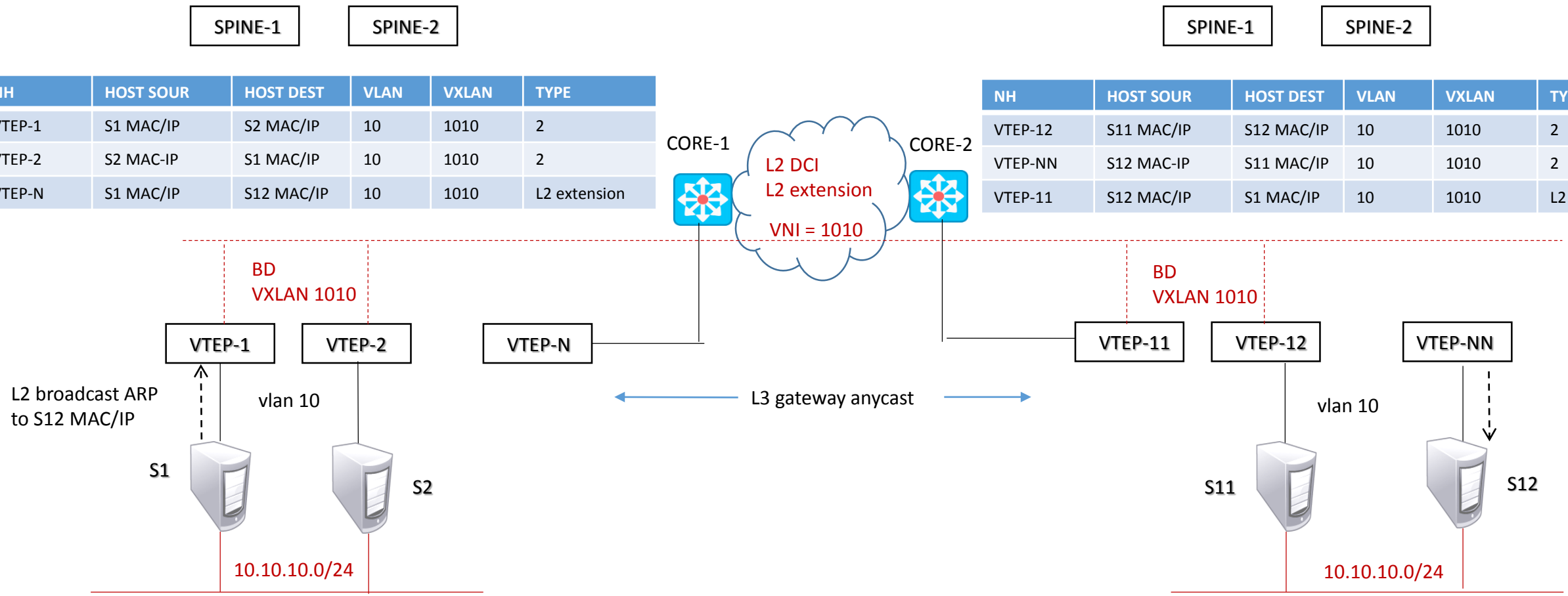
Di conseguenza, tutti gli edge devices imparano le informazioni end-point che appartengono ai rispettivi VNI (VXLAN segment Network Identifier) ed essere importate all'interno della propria forwarding table.

Intra-Subnet and Inter-Subnet communication via EVPN Fabric

La comunicazione tra due end-point intra-subnet (stessa subnet IP) ubicati su EVPN Fabric differenti è stabilito attraverso la combinazione di creare un bridge domain L2 VXLAN (all'interno di ogni Fabric) e un L2 extension segment di rete IP address tra Fabrics;

La comunicazione tra due end-point inter-subnet (differente subnet IP) avviene sempre tra due endpoint EVPN ubicati in differenti Fabrics, ma con due differenti subnets IP default gateway.

Intra-Subnet design communication via EVPN Fabric



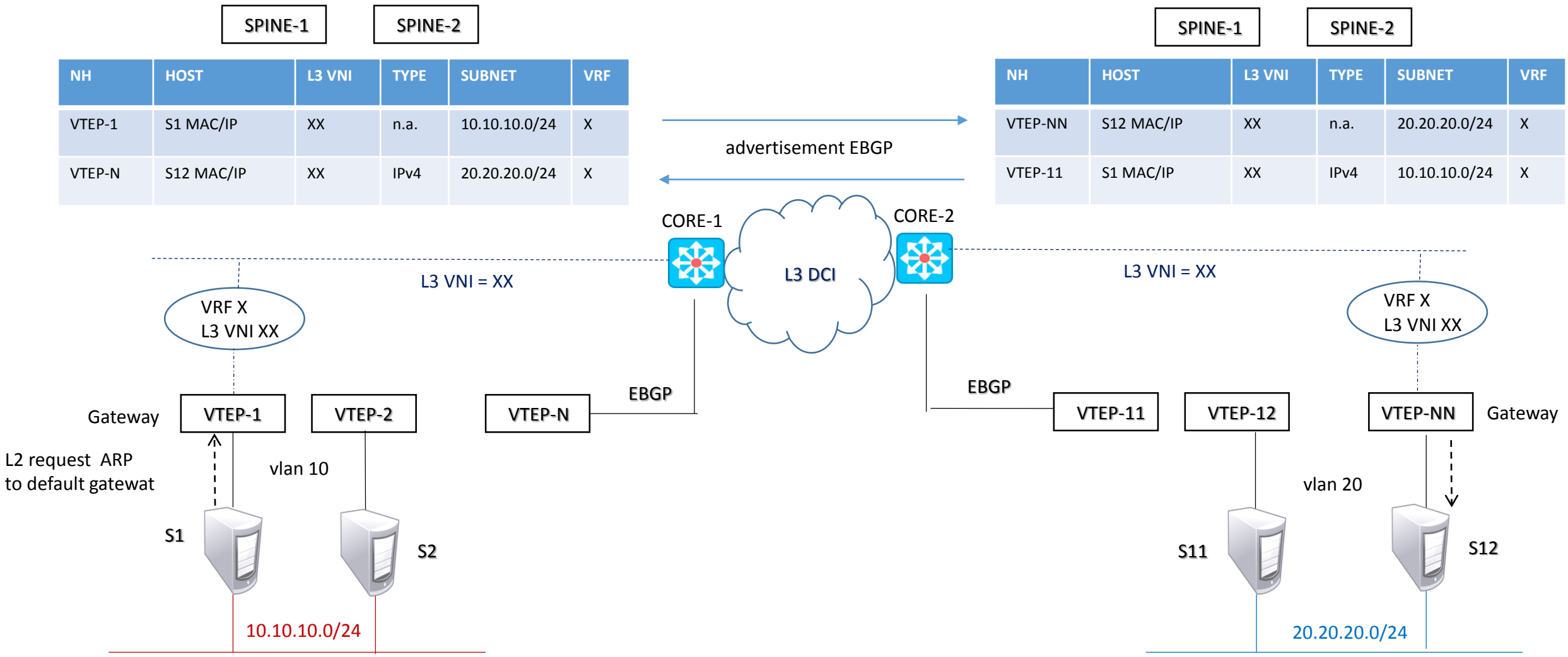
NH	HOST SOUR	HOST DEST	VLAN	VXLAN	TYPE
VTEP-1	S1 MAC/IP	S2 MAC/IP	10	1010	2
VTEP-2	S2 MAC-IP	S1 MAC/IP	10	1010	2
VTEP-N	S1 MAC/IP	S12 MAC/IP	10	1010	L2 extension

NH	HOST SOUR	HOST DEST	VLAN	VXLAN	TYPE
VTEP-12	S11 MAC/IP	S12 MAC/IP	10	1010	2
VTEP-NN	S12 MAC-IP	S11 MAC/IP	10	1010	2
VTEP-11	S12 MAC/IP	S1 MAC/IP	10	1010	L2 extension

VTEP	NEXT HOP	HOST	TYPE
1	S1 MAC/IP	LOCAL	
1	S2 MAC/IP	VTEP-2	2
1	S12 MAC/IP	VTEP-N	2

VTEP	NEXT HOP	HOST	TYPE
VTEP-NN	S12 MAC/IP	LOCAL	
VTEP-NN	S11 MAC/IP	VTEP-12	2
VTEP-NN	S1 MAC/IP	VTEP-11	2

Inter-Subnet design communication via EVPN Fabric



NH	HOST	L3 VNI	TYPE	SUBNET	VRF
VTEP-1	S1 MAC/IP	XX	n.a.	10.10.10.0/24	X
VTEP-N	S12 MAC/IP	XX	IPv4	20.20.20.0/24	X

NH	HOST	L3 VNI	TYPE	SUBNET	VRF
VTEP-NN	S12 MAC/IP	XX	n.a.	20.20.20.0/24	X
VTEP-11	S1 MAC/IP	XX	IPv4	10.10.10.0/24	X

VTEP	NEXT HOP	HOST	TYPE
1	S1 MAC/IP	LOCAL	
1	S2 MAC/IP	VTEP-2	2
1	S12 MAC/IP	Request ARP	5

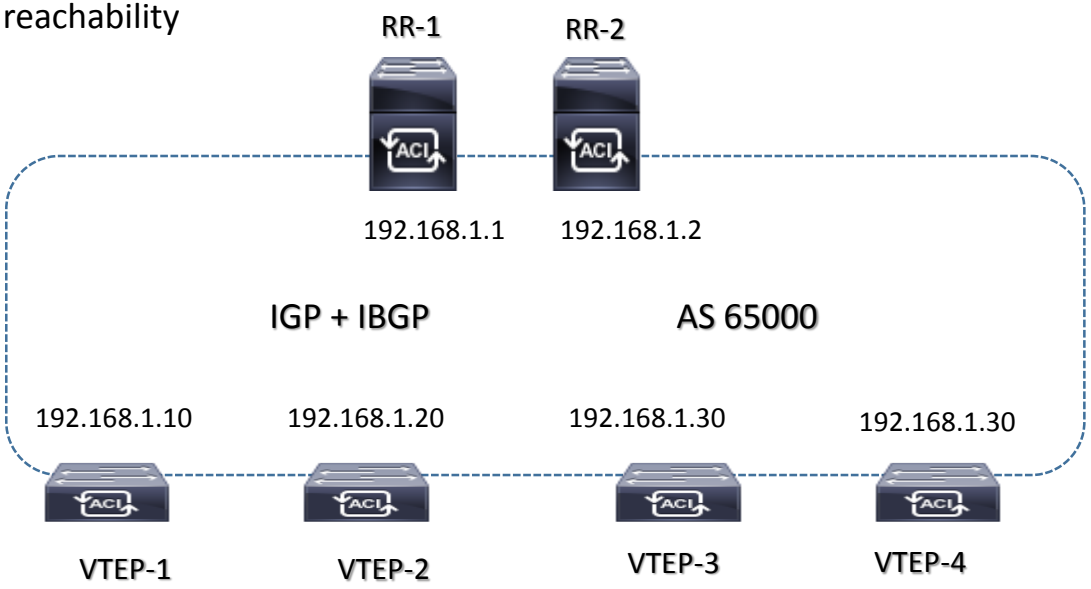
VTEP	NEXT HOP	HOST	TYPE
VTEP-NN	S12 MAC/IP	LOCAL	
VTEP-NN	S11 MAC/IP	VTEP-12	2
VTEP-NN	S1 MAC/IP	Request ARP	5

EVPN I-BGP Configurations VTEP (VXLAN Tunnel End-Point)

feature bgp
feature nv overlay → enable VTEP (required on Leaf or Border)
fature nv overlay evpn → enable EVPN control-plane in BGP

@ only on LEAF

interface nve1 → enable interface VTEP
source-interface loopback0 → enable source interface with loopback
host-reachability protocol bgp → enable BGP for host reachability



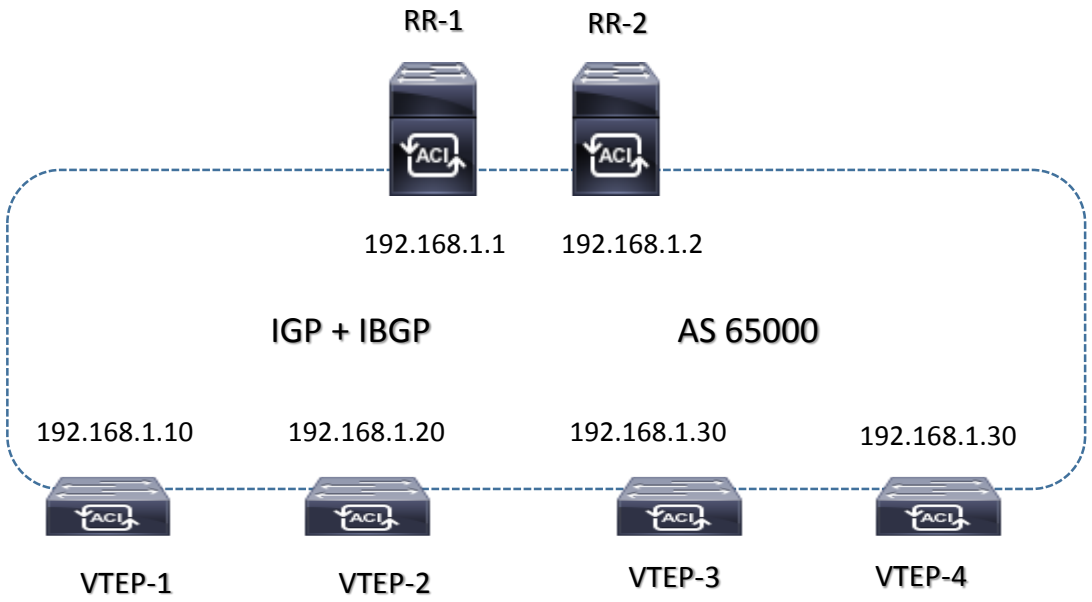
EVPN I-BGP Configurations Overlay Control Plane

 # SPINE RR1

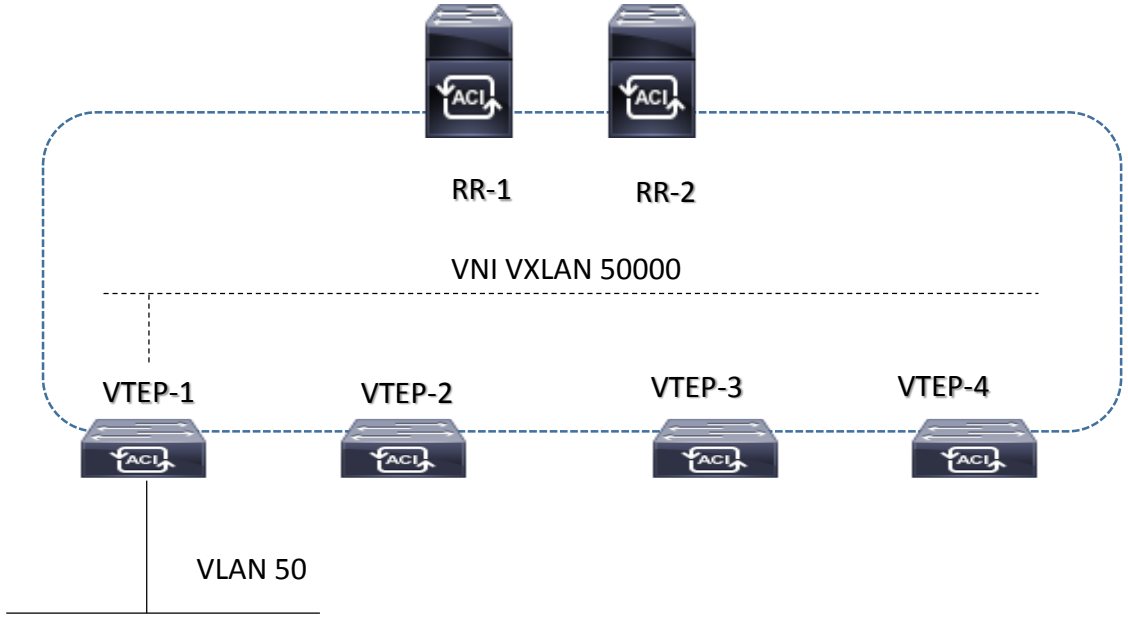
```
router bgp 65000
router-id 192.168.1.1
address-family ipv4 unicast
neighbor 192.168.1.10  remote-as 65000
  update-source loopback0
  address-family l2vpn evpn
  send-community both
  route-reflector client
```

 # LEAF VTEP-1

```
router bgp 65000
router-id 192.168.1.10
address-family ipv4 unicast
neighbor 192.168.1.1 remote-as 65000
  update-source loopback0
  address-family l2vpn evpn
  send-community both
```



EVPN I-BGP Configurations VLAN to VXLAN



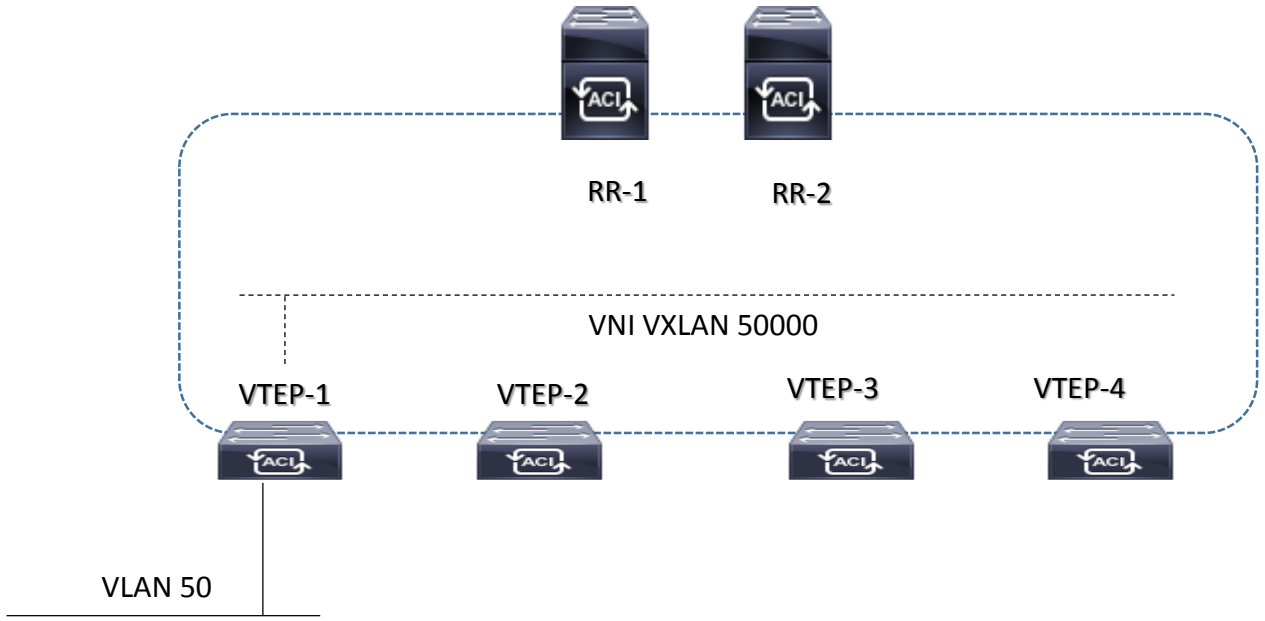
 # Mapping IEEE 802.1q vlan-id TO VXLAN VNI

```
feature  vn-segment-vlan-based
!
vlan 50
  vn-segment 50000
!
evpn
  vni 50000 l2
  rd auto                          → # RD is default calculated as VNI:BGP Router ID
  route-target import auto         → # RT is default calculated as BGP AS:VNI
  route-target export auto
!
interface  nve1
  source-interface loopback0
  host-reachability protocol bgp
  member vni 50000
    mcast-group 239.239.239.10
    suppress-arp
```

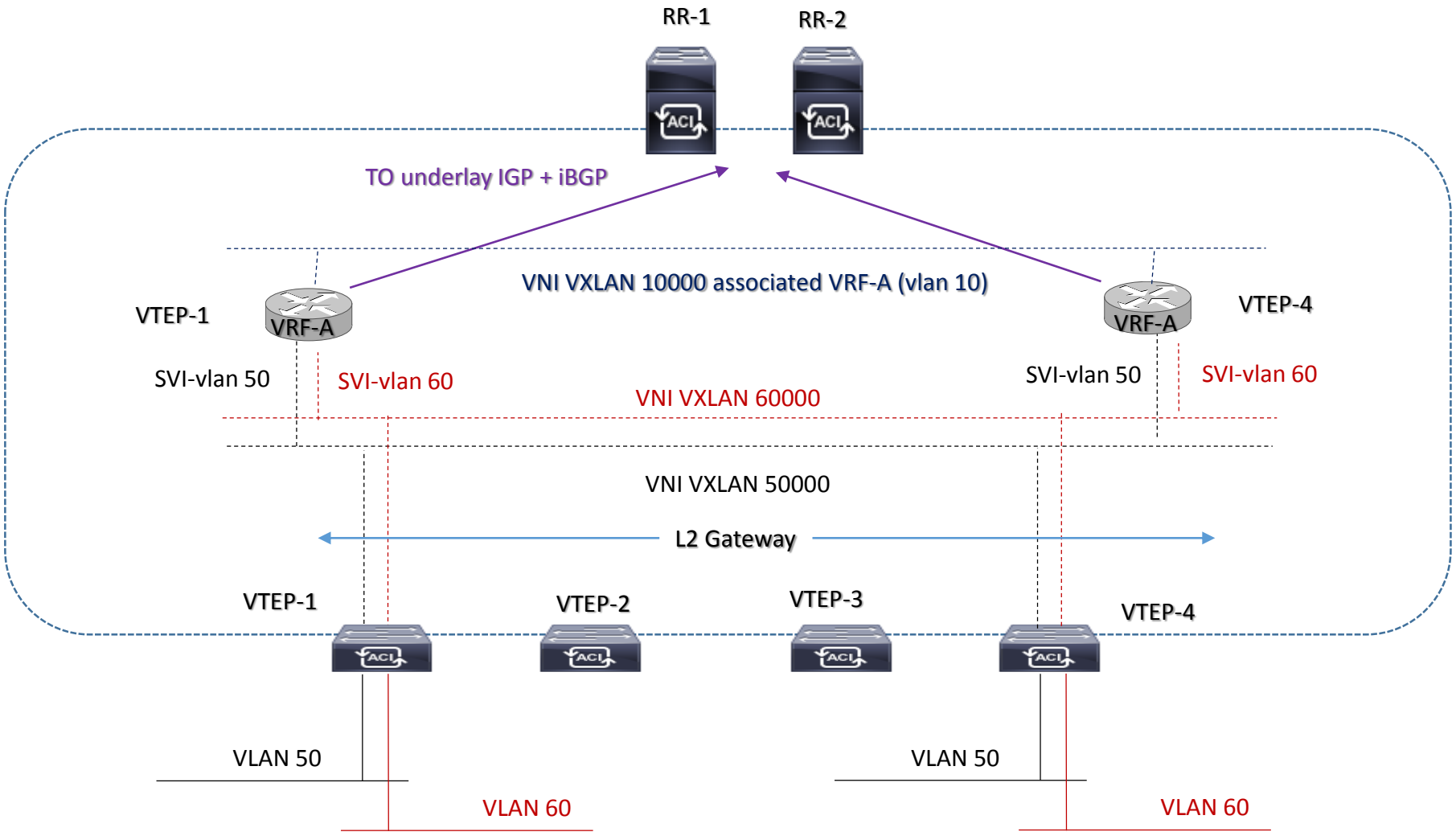
EVPN I-BGP Configurations Routing Resource on VXLAN

 # Define VLAN for VRF routing instances

```
vlan 50
  vn-segment 50000
!
interface  vlan 50
 no shutdown
 mtu 9216
 vrf member VRF-A
 ip forward
!
vrf context VRF-A
 vni 50000
 rd auto
   address-family ipv4unicast
   route-target both auto
   route-target both auto evpn
```



EVPN I-BGP Design Distributed IP Anycast Gateway



Vlan-ID ha significato solo locale al VTEP

EVPN I-BGP Configurations Distributed IP Anycast Gateway

Define VLAN 50 and 60

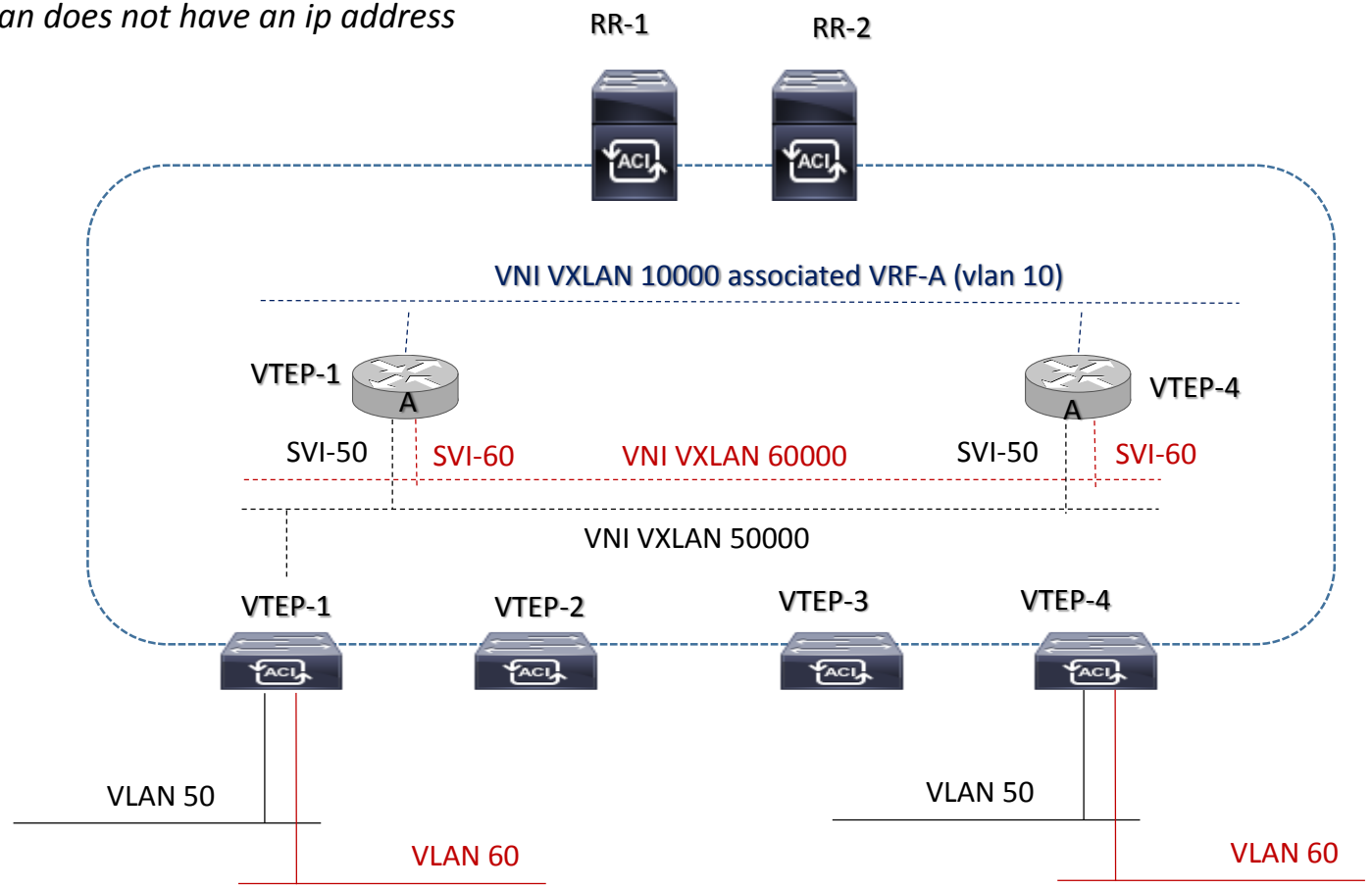
```
features interface-vlan
fabric-forwarding anycast-gateway-mac  < mac-address >
!
vlan 50
  vn-segment 50000
!
vlan 60
  vn-segment 60000
!
interface  vlan 50
 no shutdown
 mtu 9216
 vrf member VRF-A
 ip address 50.50.50.1/24 tag 123
fabric forwarding mode anycast-gateway
!
interface  vlan 60
 no shutdown
 mtu 9216
 vrf member VRF-A
 ip address 60.60.60.1/24 tag 123
fabric forwarding mode anycast-gateway
```

→ *un MAC address per VTEP; tutti i VTEP dovrebbero avere lo stesso MAC Address*

EVPN I-BGP Configurations Routing on VXLAN (1/1)

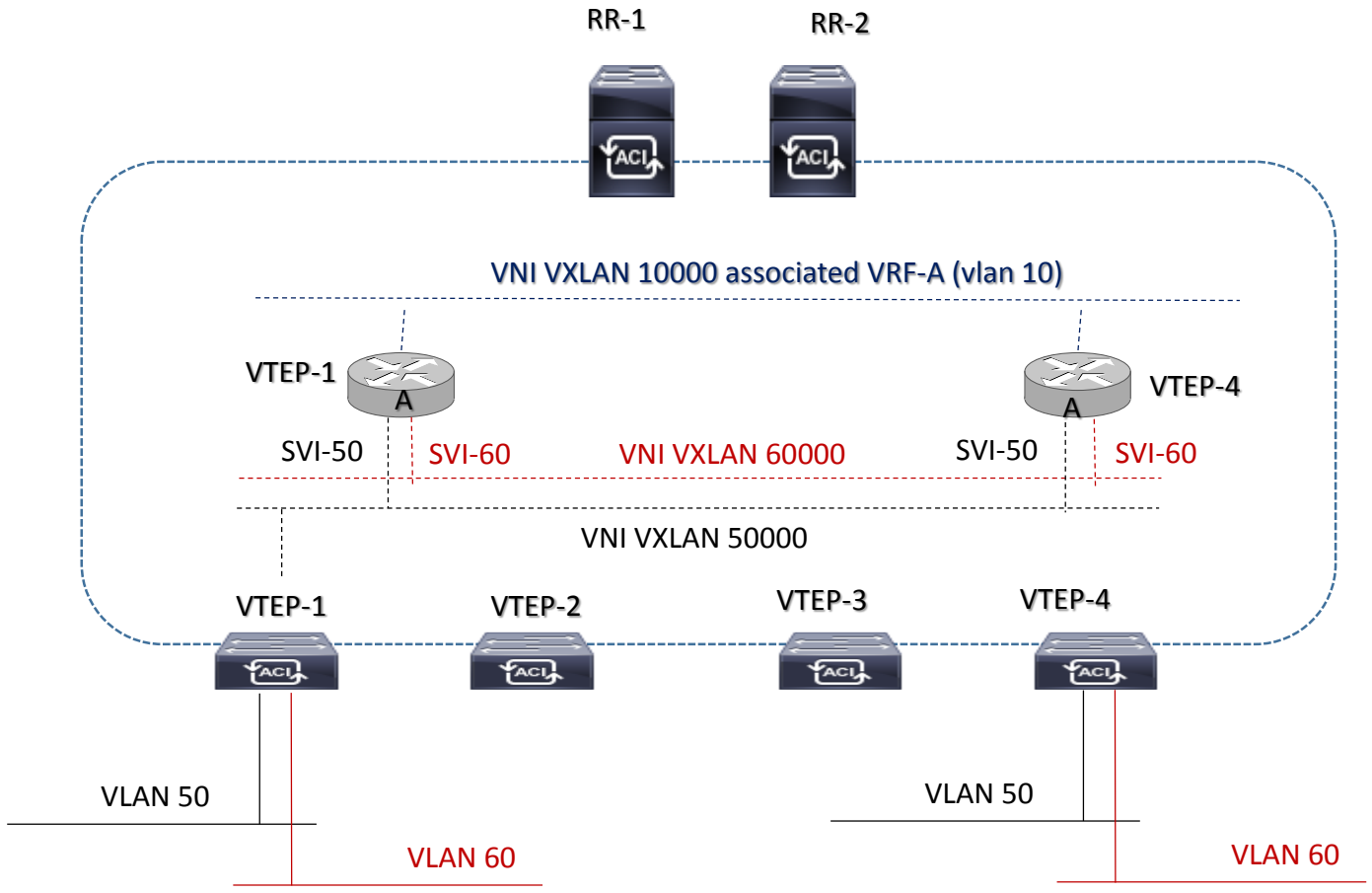
```
vlan 10          → # vlan 10 is used as Layer 3 VNI to route inter-vlan routing
vn-segment 10000
!
interface vlan 10 → # Layer 3 VNI associated interface vlan does not have an ip address
vrf member VRF-A
no shutdown
!
interface  nve1
  source-interface loopback0
  host-reachability protocol bgp
  member vni 50000
    mcast-group 239.239.239.10
    suppress-arp
member vni 10000 associate-vrf
!
member vni 60000
    mcast-group 239.239.239.11
    suppress-arp
member vni 10000 associate-vrf
!
```

segue ./.



EVPN I-BGP Configurations Routing on VXLAN (1/2)

```
route-map RED-SUBNET permit 10
match 123
!
router bgp 65000
 vrf VRF-A
 advertise l2vpn evpn
 redistribuite direct route-map RED-SUBNET
 maximum-path ibgp 2
```



EVPN I-BGP Configurations IGP with OSPF

VTEP1:

```
feature ospf
feature pim
!
ip pim rp-address 192.168.1.1 group-list 224.0.0.0/4
ip pim ssm range 232.0.0.0/8
!
interface ethernet ½
desscription to-SPINE
no switchport
ip address 10.1.1.2/30
ip route ospf UNDERLAY area  0.0.0.0
ip pim sparse-mode
no shutdown
!
interface loopback 0
description «loopback for BGP»
ip address 192.168.1.10/32
ip route ospf UNDERLAY area  0.0.0.0
ip pim sparse-mode
no shutdown
!
router ospf UNDERLAY
```