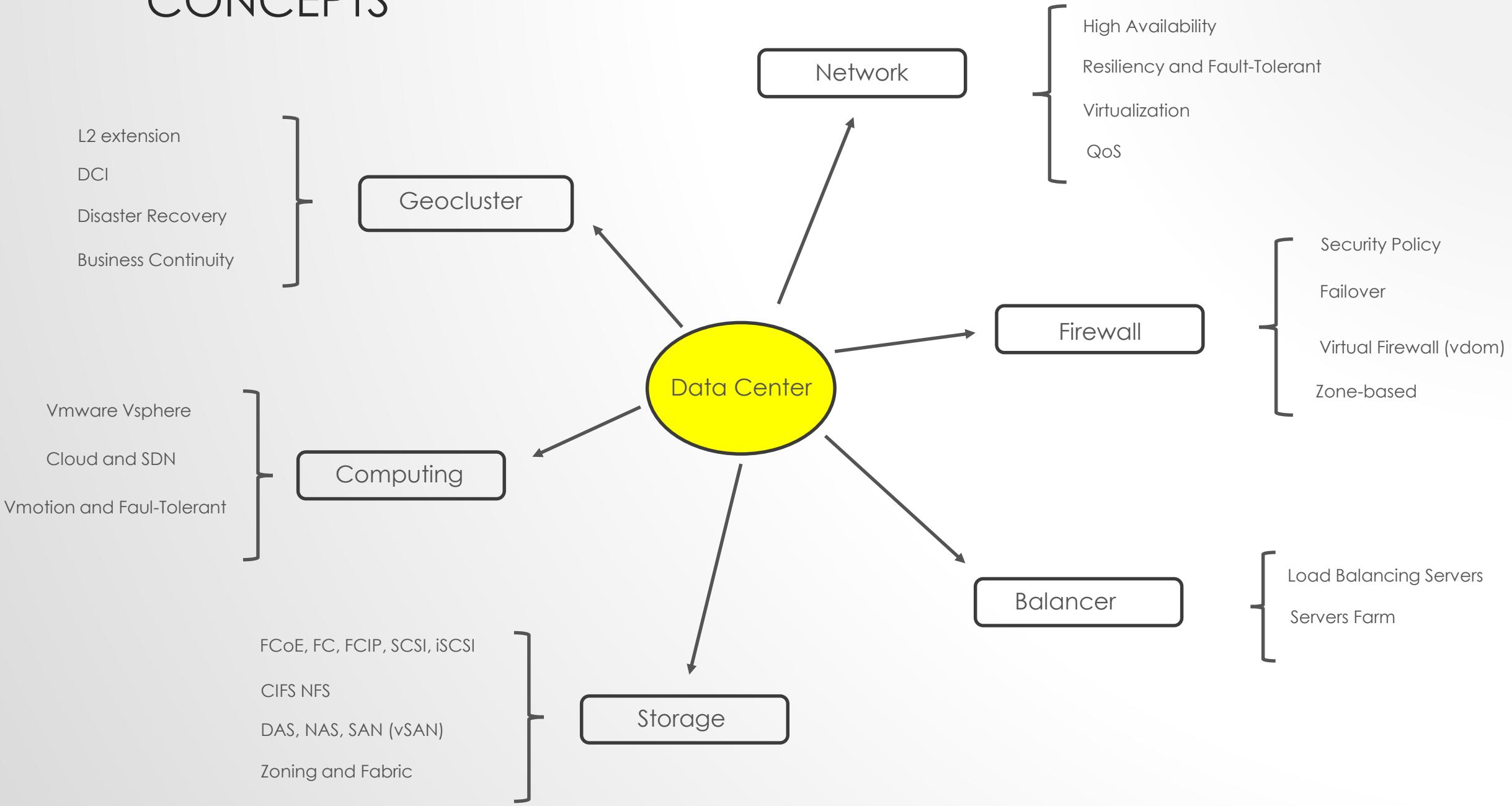
ARCHITETTURE DATACENTERS IN TECNOLOGIA CISCO

Massimiliano Sbaraglia

CONCEPTS

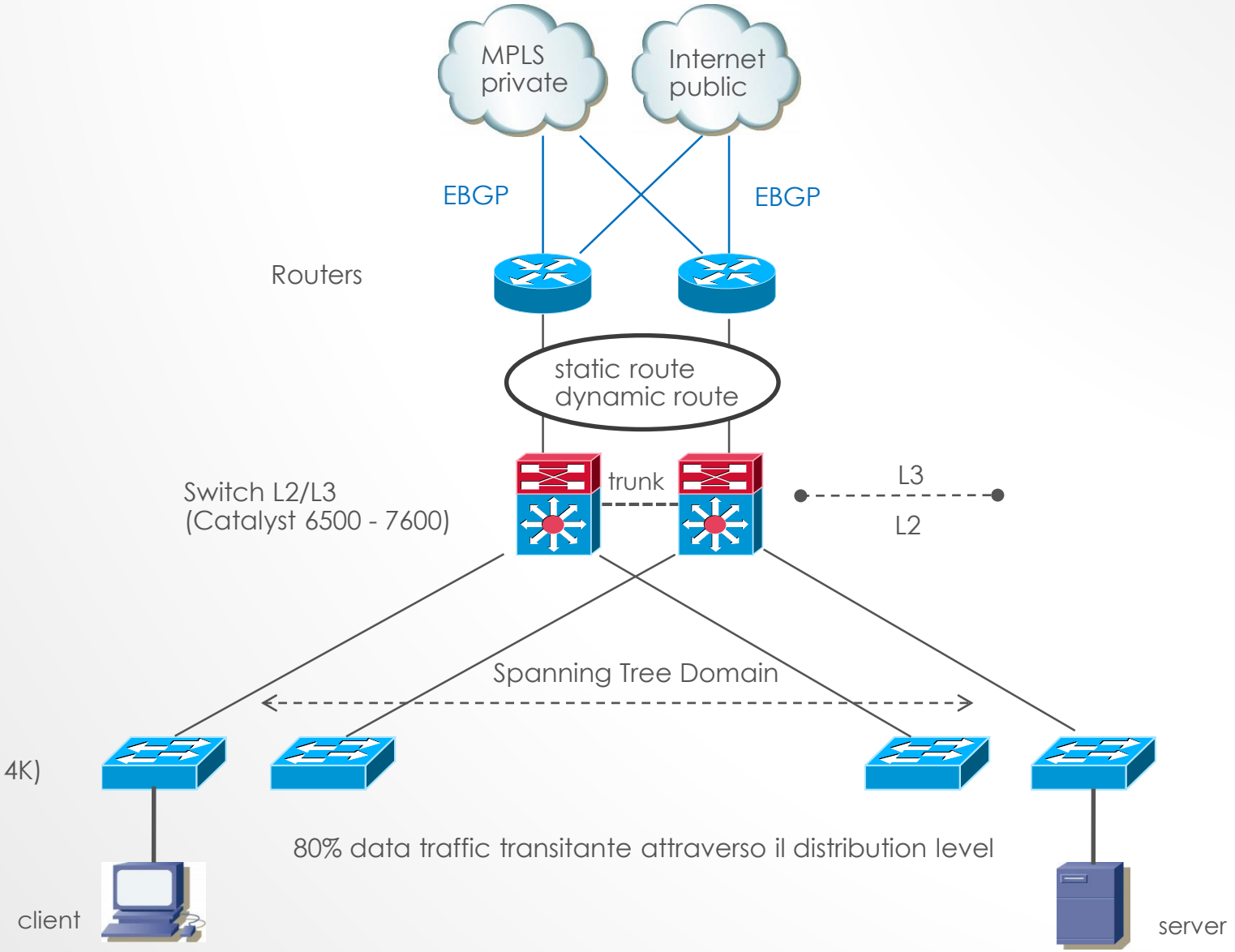

Data Center
Network
High Availability
Resiliency and Fault-Tolerant
Virtualization
QoS
Geocluster
L2 extension
DCI
Disaster Recovery
Business Continuity
Firewall
Security Policy
Failover
Virtual Firewall (vdom)
Zone-based
Computing
Vmware Vsphere
Cloud and SDN
Vmotion and Faul-Tolerant
Balancer
Load Balancing Servers
Servers Farm
Storage
FCoE, FC, FCIP, SCSI, iSCSI
CIFS NFS
DAS, NAS, SAN (vSAN)
Zoning and Fabric

ARCHITETTURA BASE

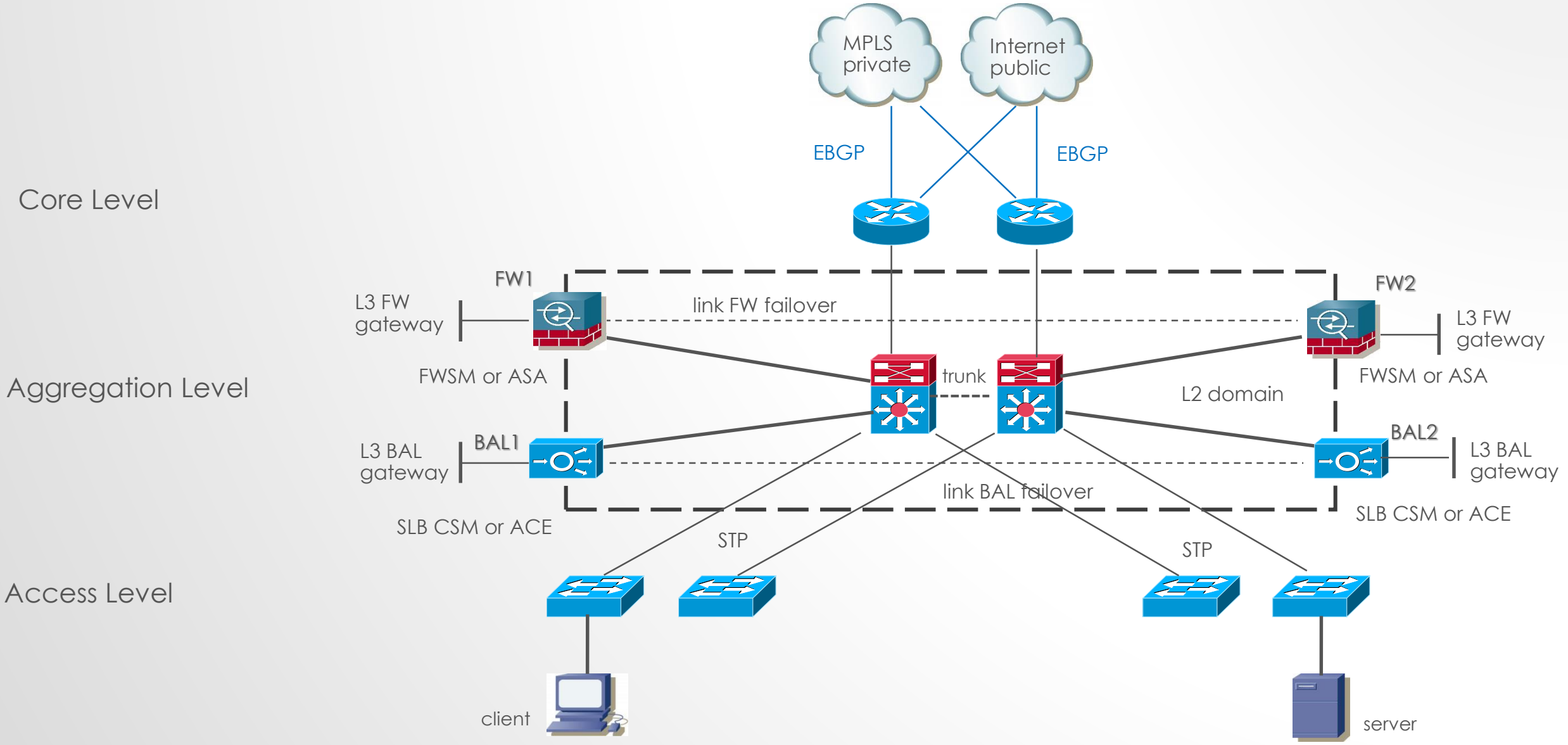
Core Level

Distribution Level

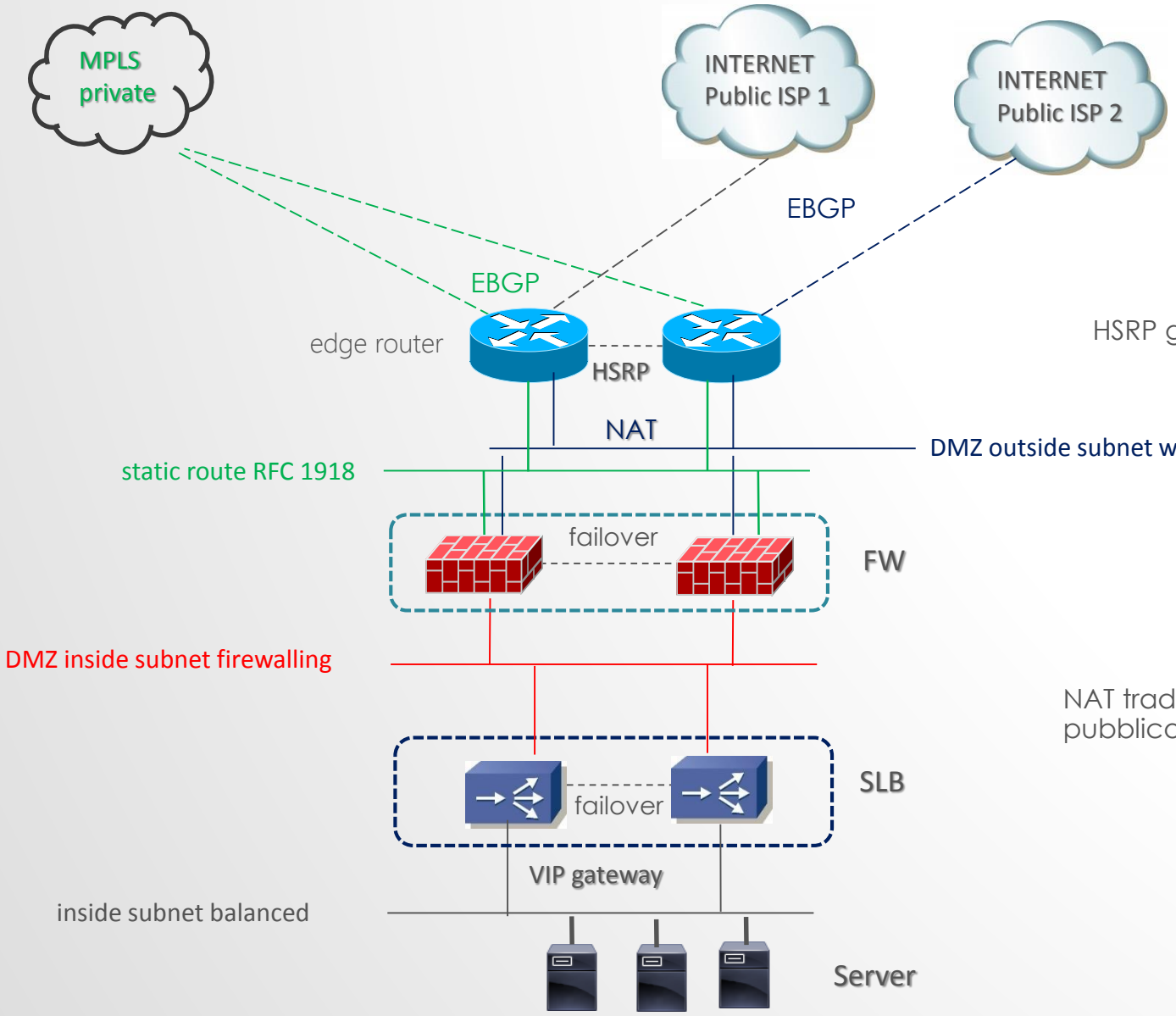
Access Level



ARCHITETTURA DATACENTER BASE



ARCHITETTURA DATACENTER BASE

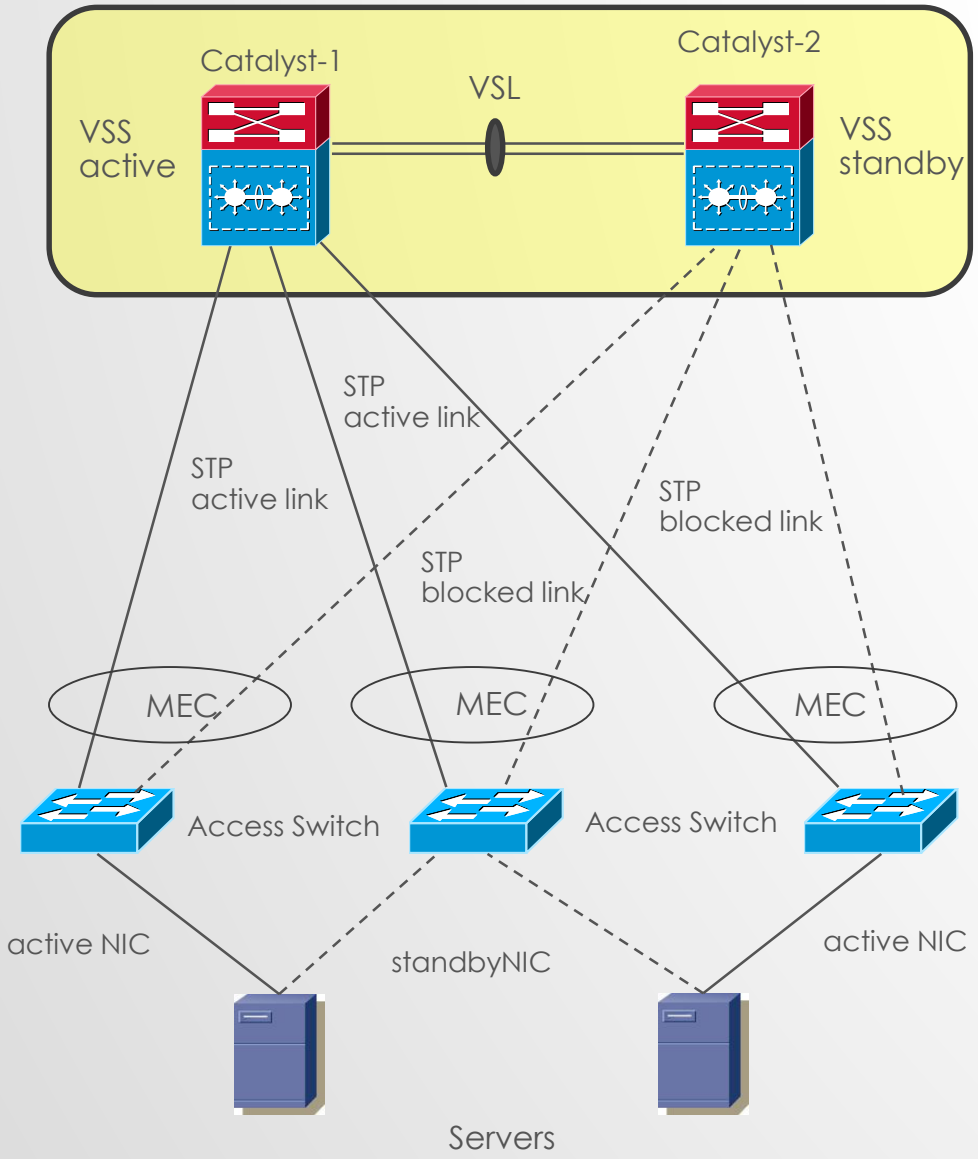


BGP AS-path prepend garantisce failover tra path EBGP per inbound traffic (to internal domain)

HSRP gestisce HA per traffico outbound (to external)

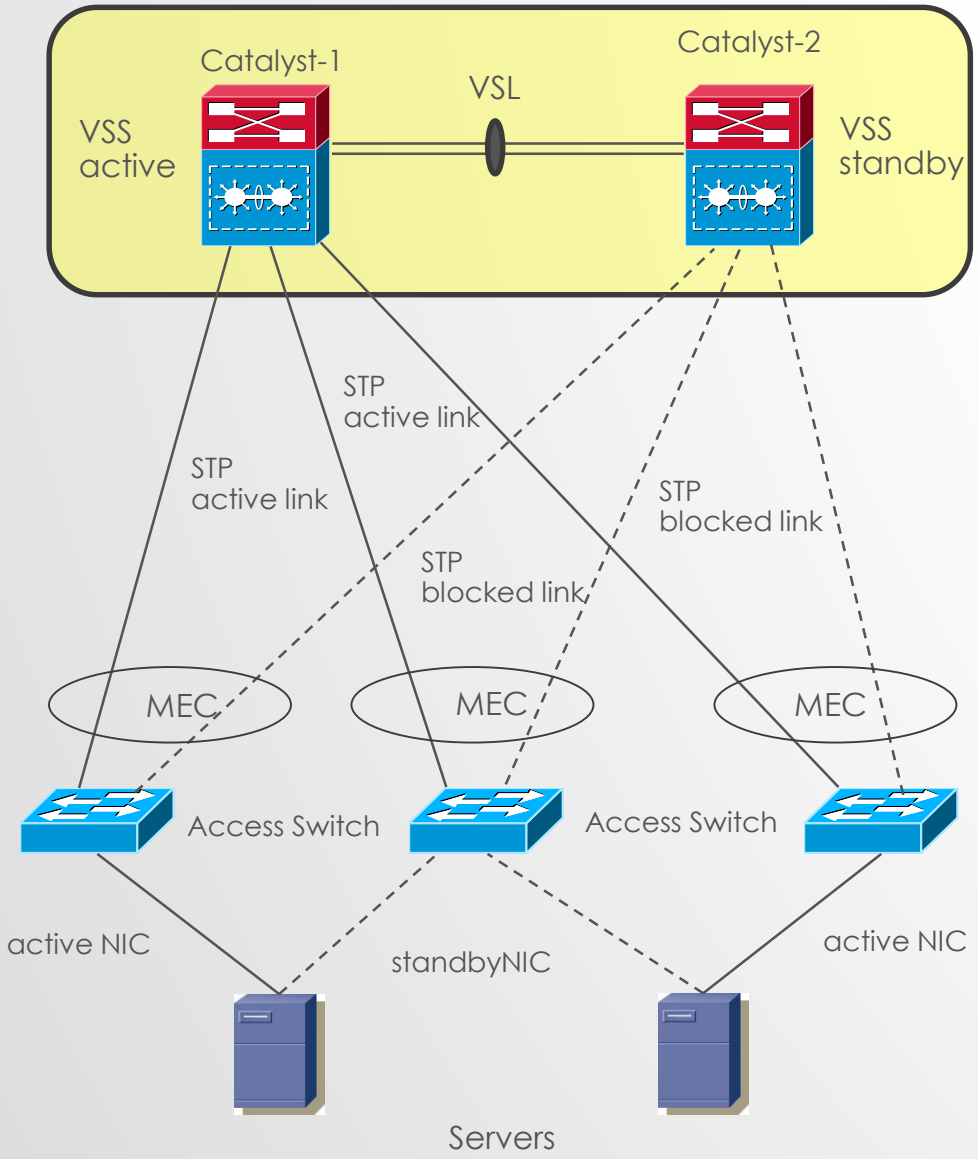
NAT traduce il piano di indirizzamento privato con un blocco (pool) pubblico annunciato ad entrambi gli ISP Internet

VSS (VIRTUAL SWITCHING SYSTEM)



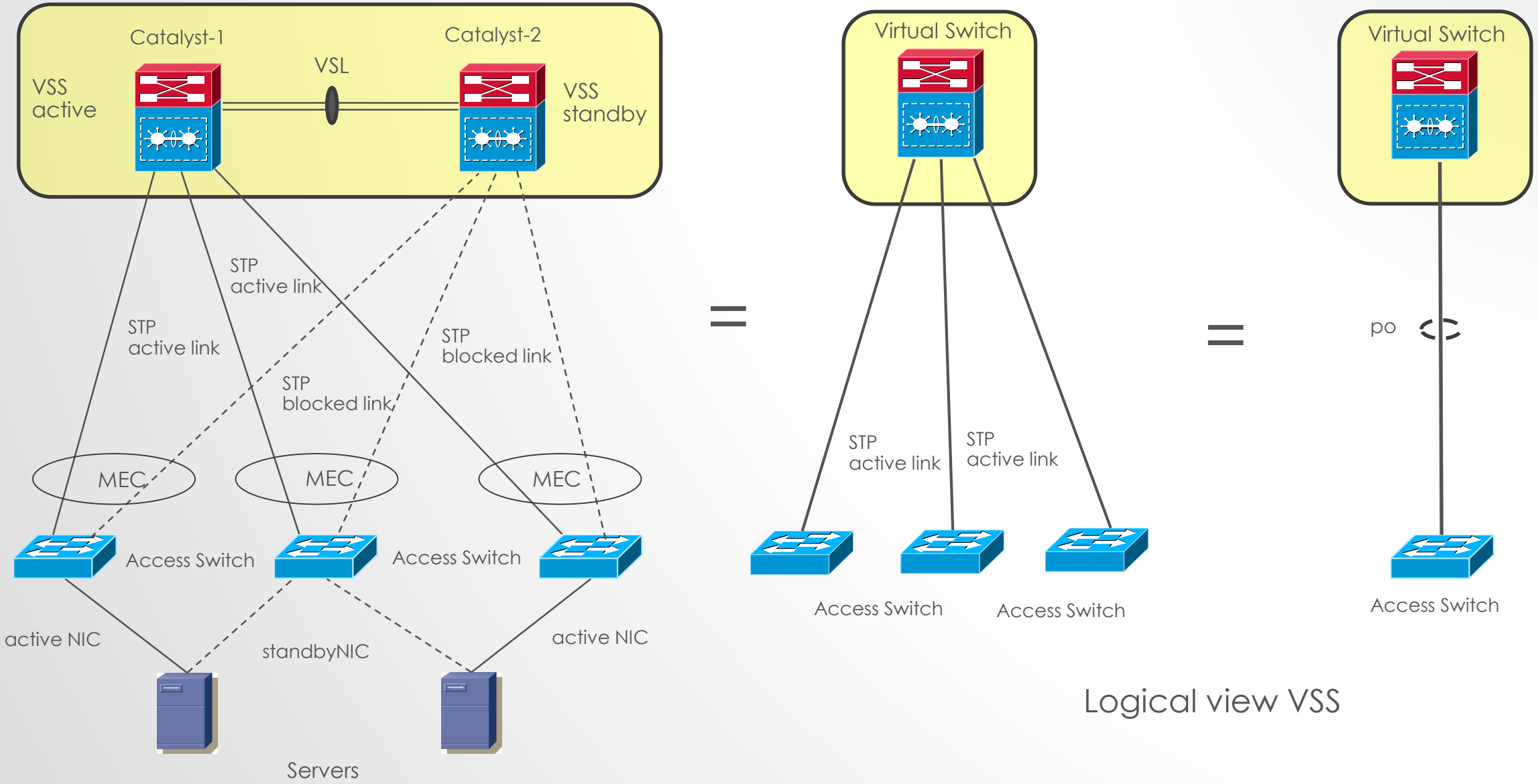
- Un sistema VSS opera via SSO (Stateful Switch Over) tra peers attraverso le Supervisor Engine active e standby ospitate nei rispettivi chassis
- VSS Supervisor Engine active controlla le funzionalità layer 2 (switching) and layer 3 (routing) per entrambi gli chassis
- Il piano di forwarding del traffico è performato da entrambi i peers VSS
- In caso di fault della Supervisor Engine active, quella in stato standby assume il suo ruolo(switchover)
- VSL (Virtual Switch Link) è un collegamento tra i peers VSS per lo scambio di messaggi di controllo processati dalla Supervisor Engine active ma trasmessi e ricevuti su interface presenti nel peer VSS standby
- VSS opera in un contesto di Spanning Tree Protocol; il VSS standby redirige le BPDU STP via VSL verso il peer active.
- Il STP bridge ID è un valore comune ed è calcolato sul MAC address chassis; non cambia a seguito di uno switchover peers.

VSS (VIRTUAL SWITCHING SYSTEM)



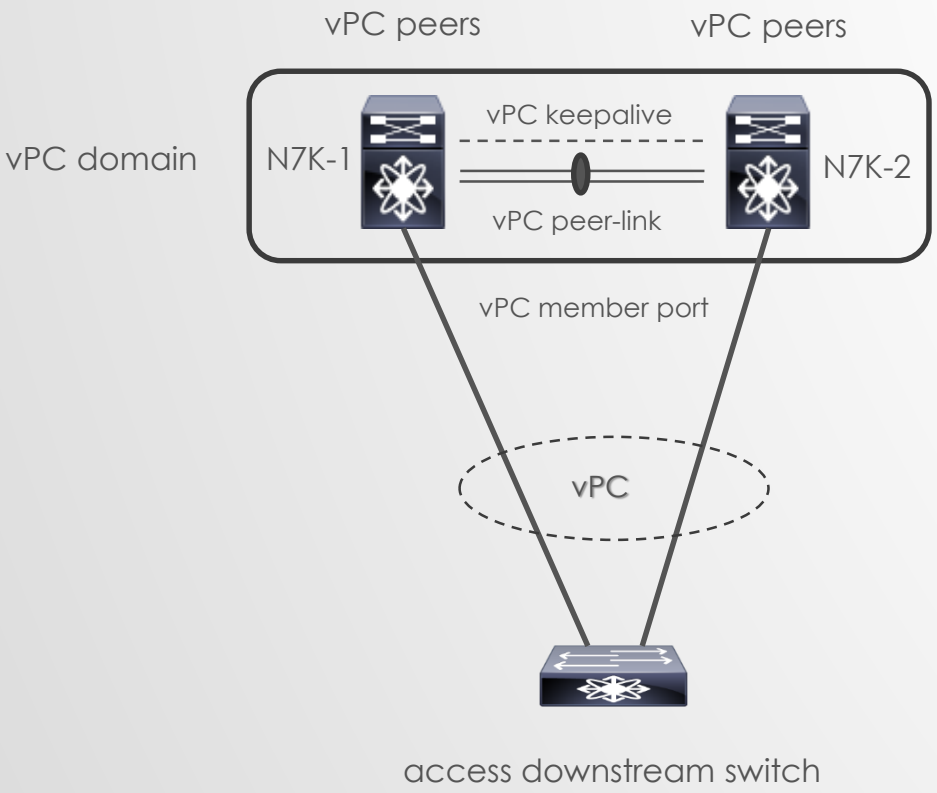
- Un MEC è un collegamento Multihassis Etherchannel tra un nodo di accesso verso entrambi i peers VSS active e standby
- Un VSS MEC può collgare qualsiasi elemento di rete che supporti etherchannel (quale host, serer, switch, router)
- Un VSS MEC supporta protocolli quali LACP (Link Aggregation Protocol) oppure PAgP (Port Aggregation Protocol)
- Il MSFC (Multilayer Switch Feature Card) presente nella Supervisor Engine active lavora a livello 3 (routing) ed entrambi i peers VSS performano il forwarding del traffico sulle rispettive interface sia in ingress che in uscita
- In genere un traffico in ingress è trasmesso (forwarding) da una interfaccia di uscita appartenente allo stesso chassis per ridurre cosi la quantità di dati transitante via VSL.

VSS (VIRTUAL SWITCHING SYSTEM)



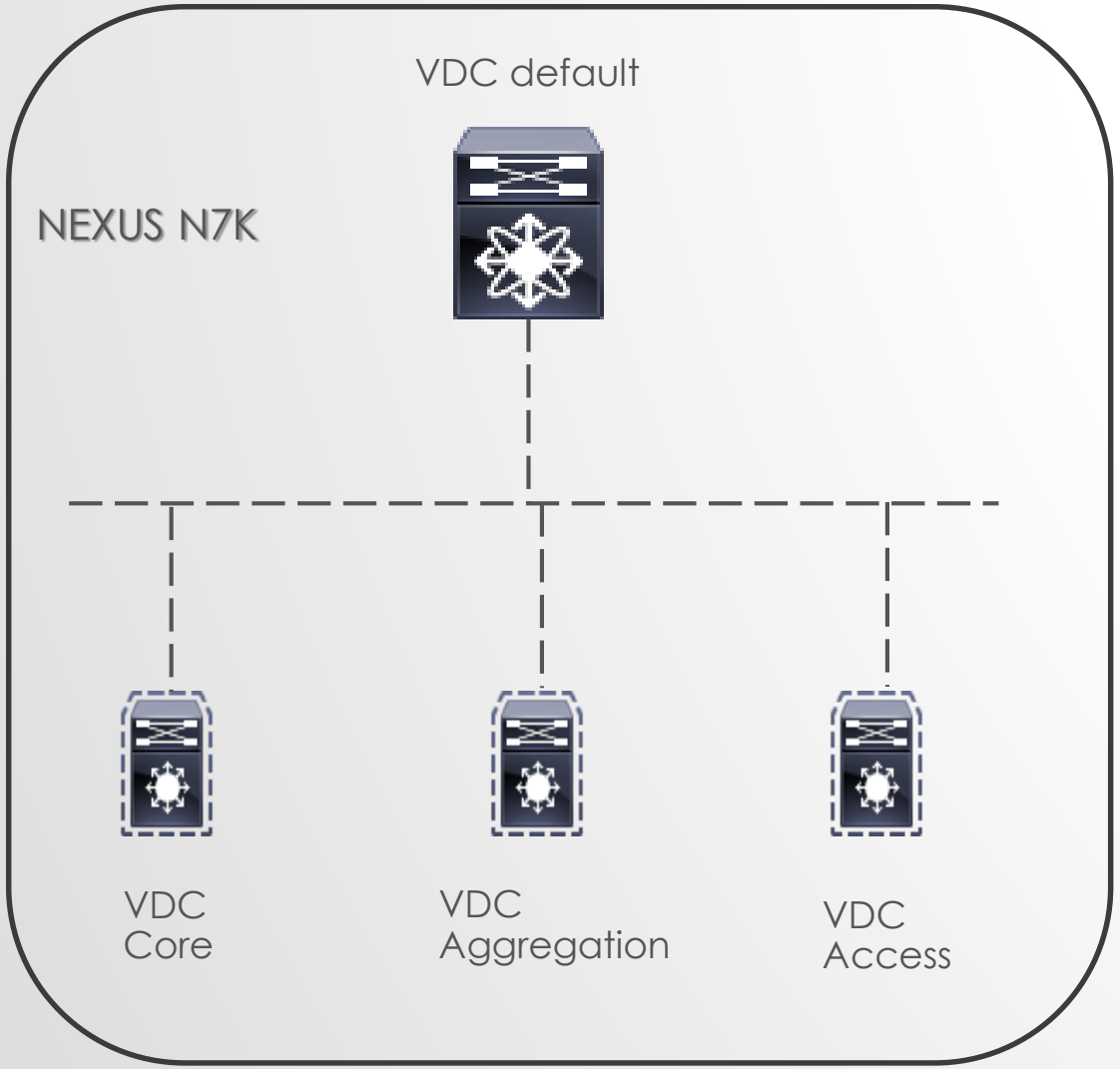
Logical view VSS

VPC CONCEPT



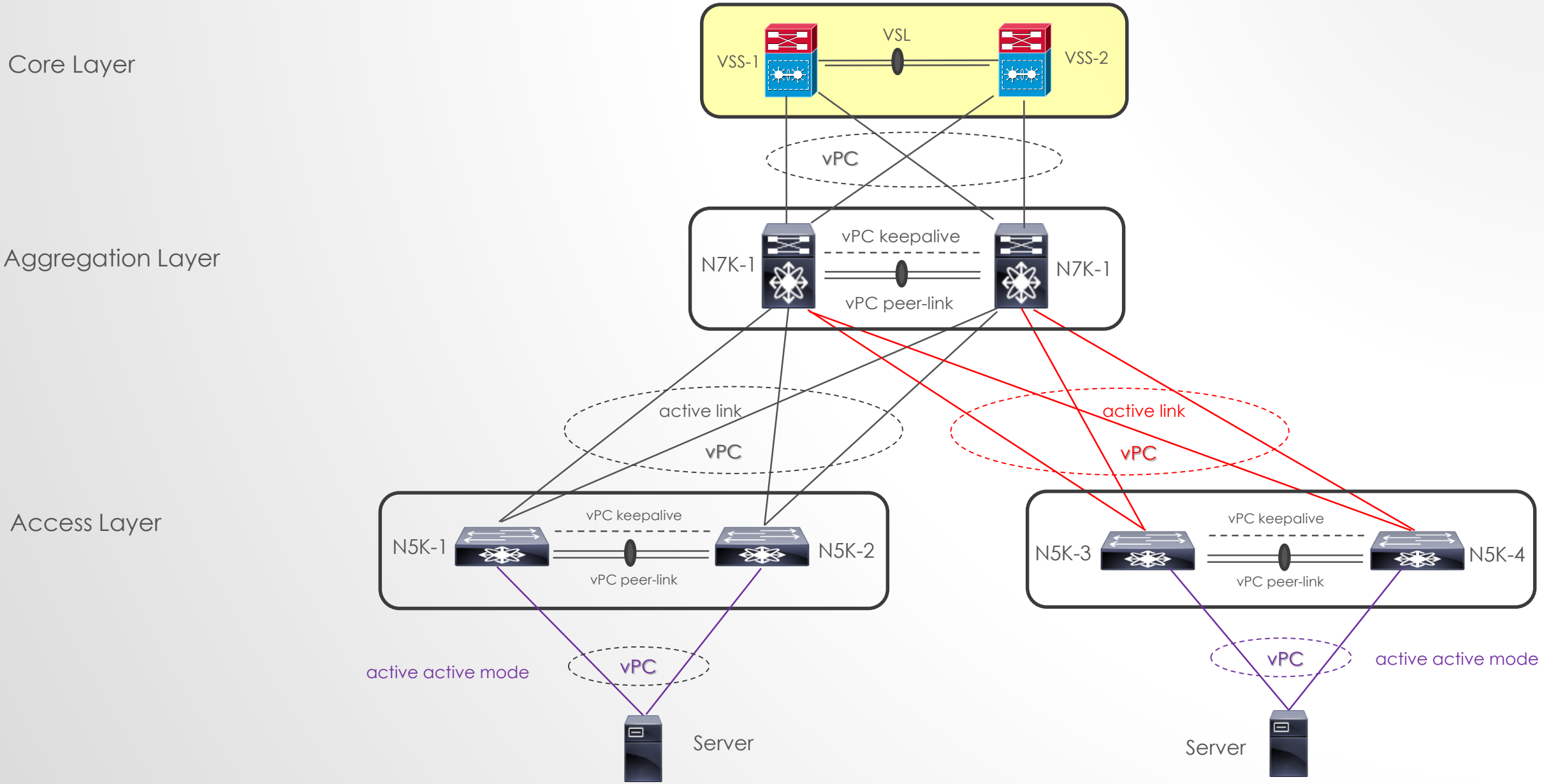
- elimina SPT blocked port
- utilizza tutti I link disponibili e relativa bandwidth
- dual-homed servers in active-active mode
- fast-convergence in caso di fault link or switch
- split-horizon loop via port-channeling (traffico entrante in un po non può uscre dallo stesso port-channel)
- Un vPC domain 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

VDC (VIRTUAL DEVICE CONTEXTS) CONCEPT

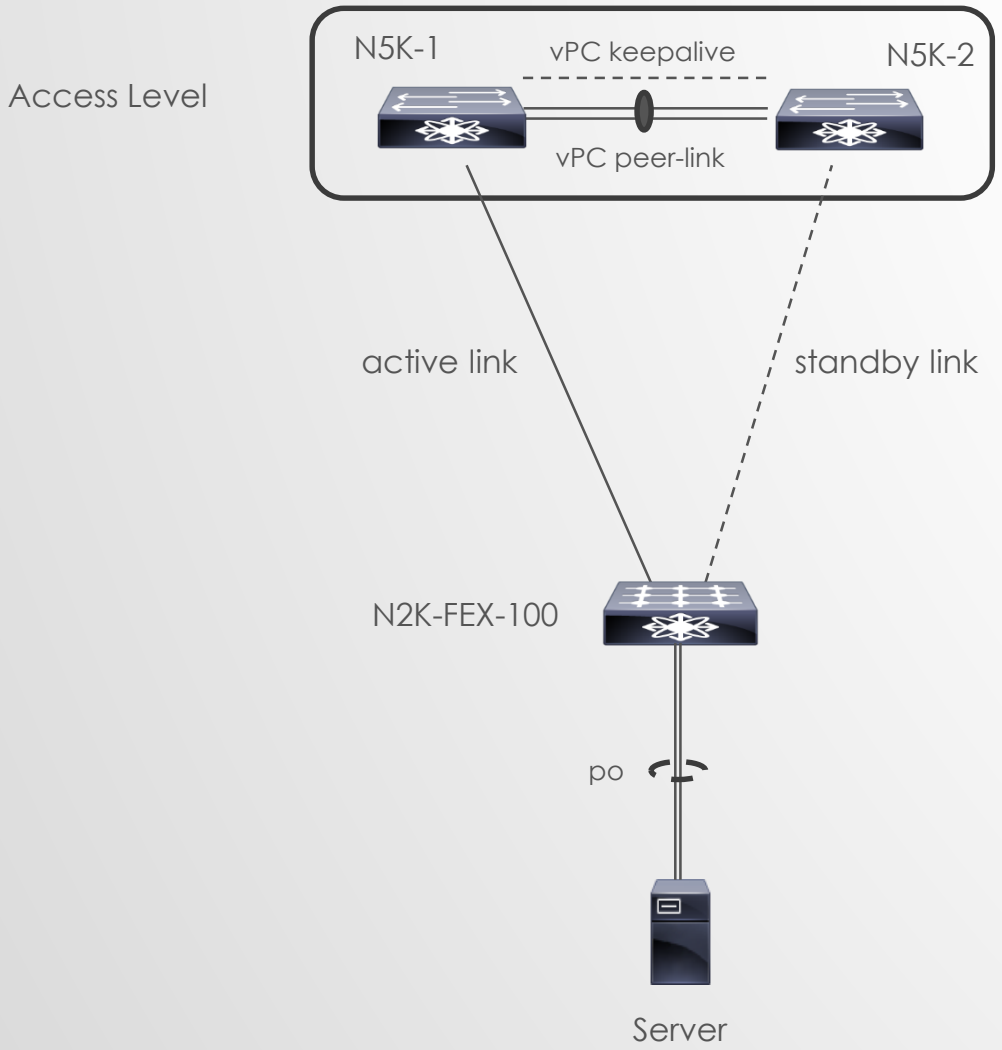


- N7K ha il concetto di VDC
- Il sistema operativo dei Nexus è NX-OS
- inizialmente tutte le risorse hardware (physical ports) e software appartengno al VDC di default; attraverso questo VDC è possibile creare nuivi contesti virtuali ed allocare le risorse di cui sopra ai VDC di competenza consentento 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

VPC ARCHITECTURE WITH NEXUS CISCO



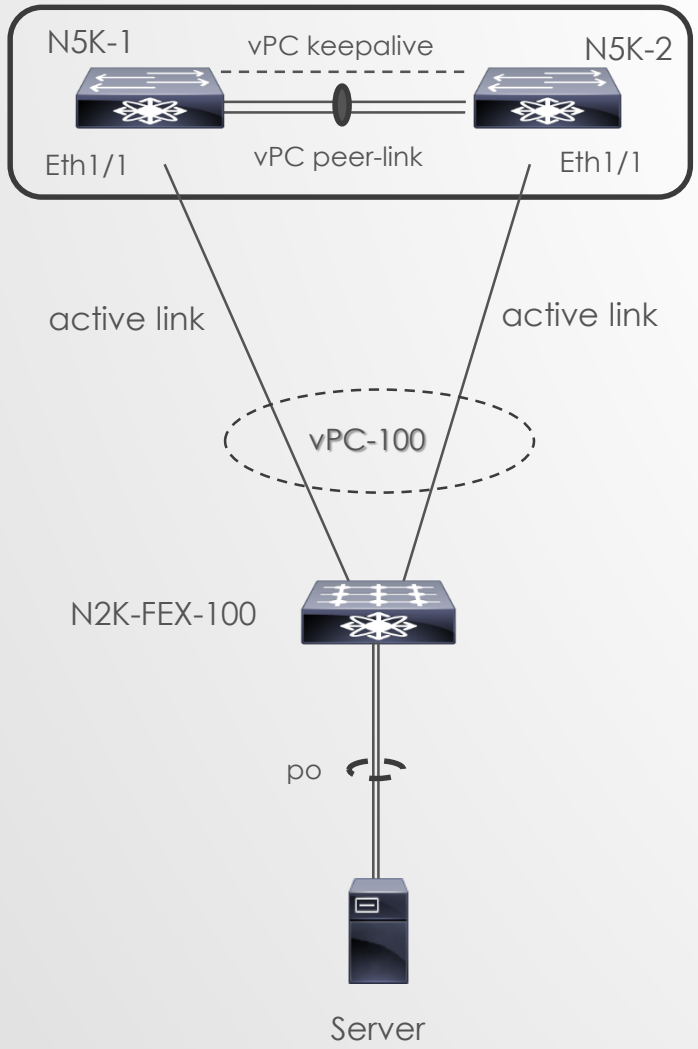
POD NEXUS FEX 1XN2K WITH ACTIVE-STANDBY DUAL-HOMED



- I FEX sono switch cisco gestiti dai loro parent-switch Nexus 5000 (possono essere visti come una estensione modulare dei parent-switch N5K)
- 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-switc 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

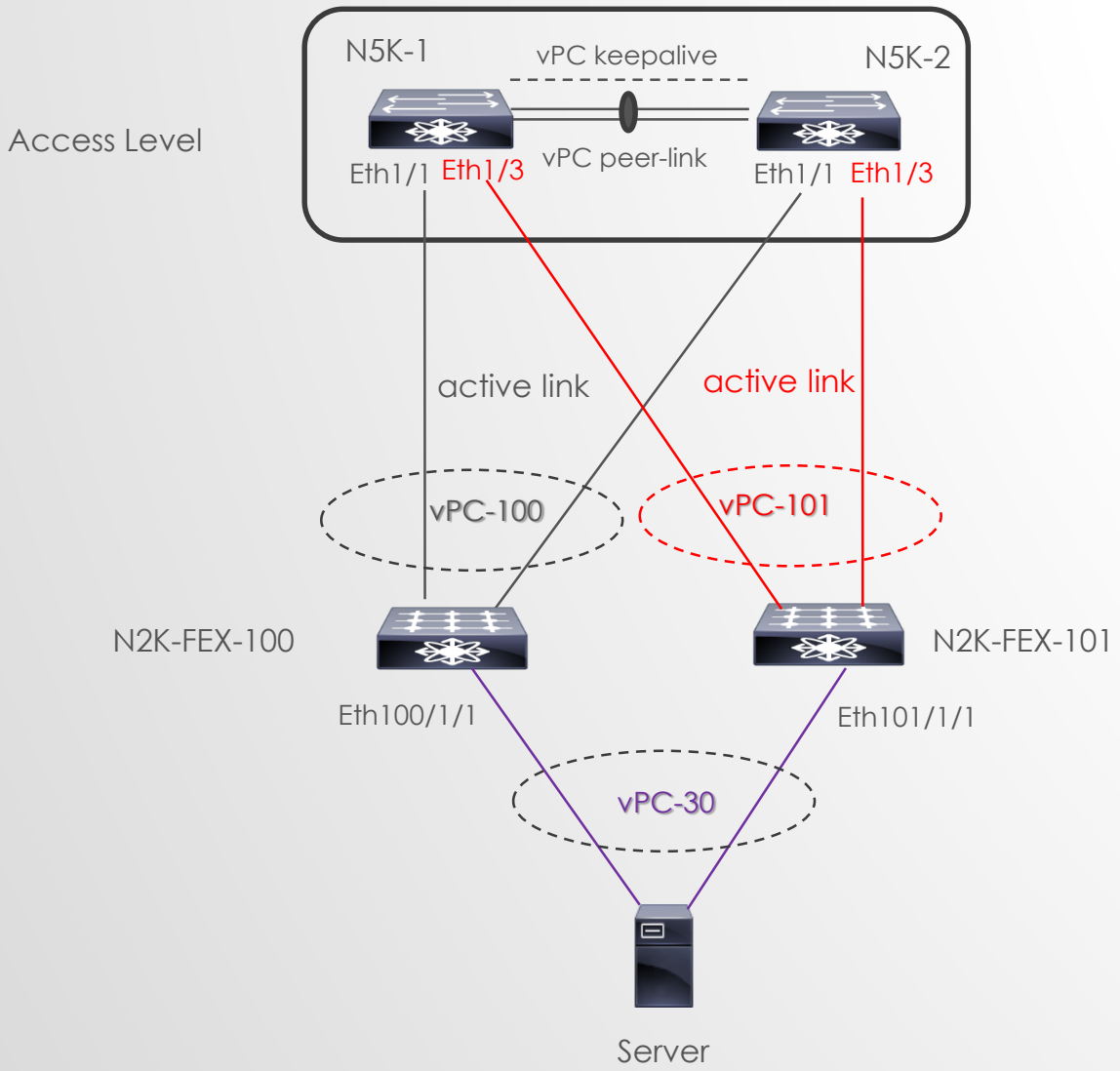
Access Level



- 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:

```
interface eth1/1
switchport mode fex-fabric
fex associate 100
!
interface port-channel 100
switchport mode fex-fabric
fex associate 100
vpc 100
```

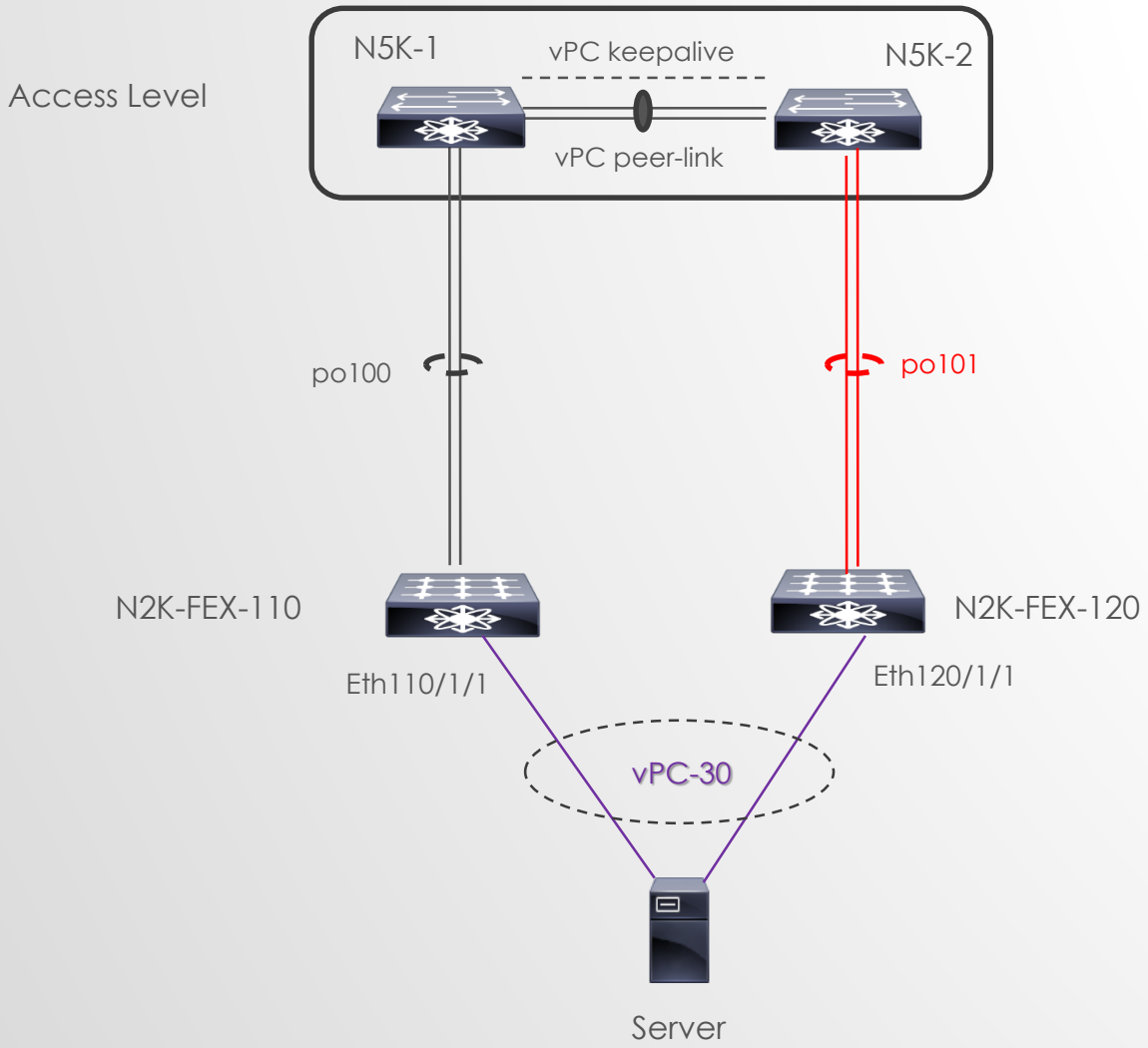
POD NEXUS FEX 2XN2K WITH ENHANCED VPC



- 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 po30
switchport access vlan 30
!
interface Ethernet 100/1/1
switchport access vlan 30
speed 1000
channel-group 30
!
interface Ethernet 101/1/1
switchport access vlan 30
speed 1000
channel-group 30
!
```

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

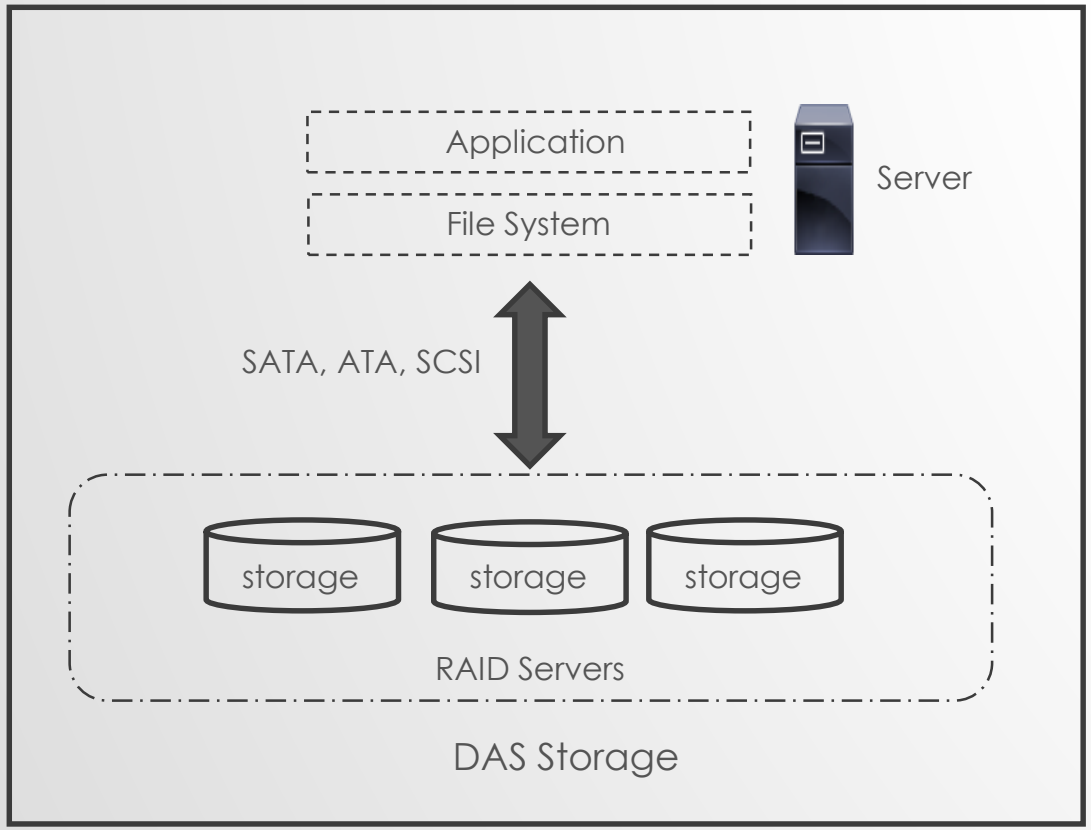
N5K-1
```
interface po11
vpc 30
!
interface eth 110/1/1
vpc 30
```

N5K-2
```
interface po12
vpc 30
!
interface eth 120/1/1
vpc 30
```

ARCHITETTURE STORAGE SAN VSAN

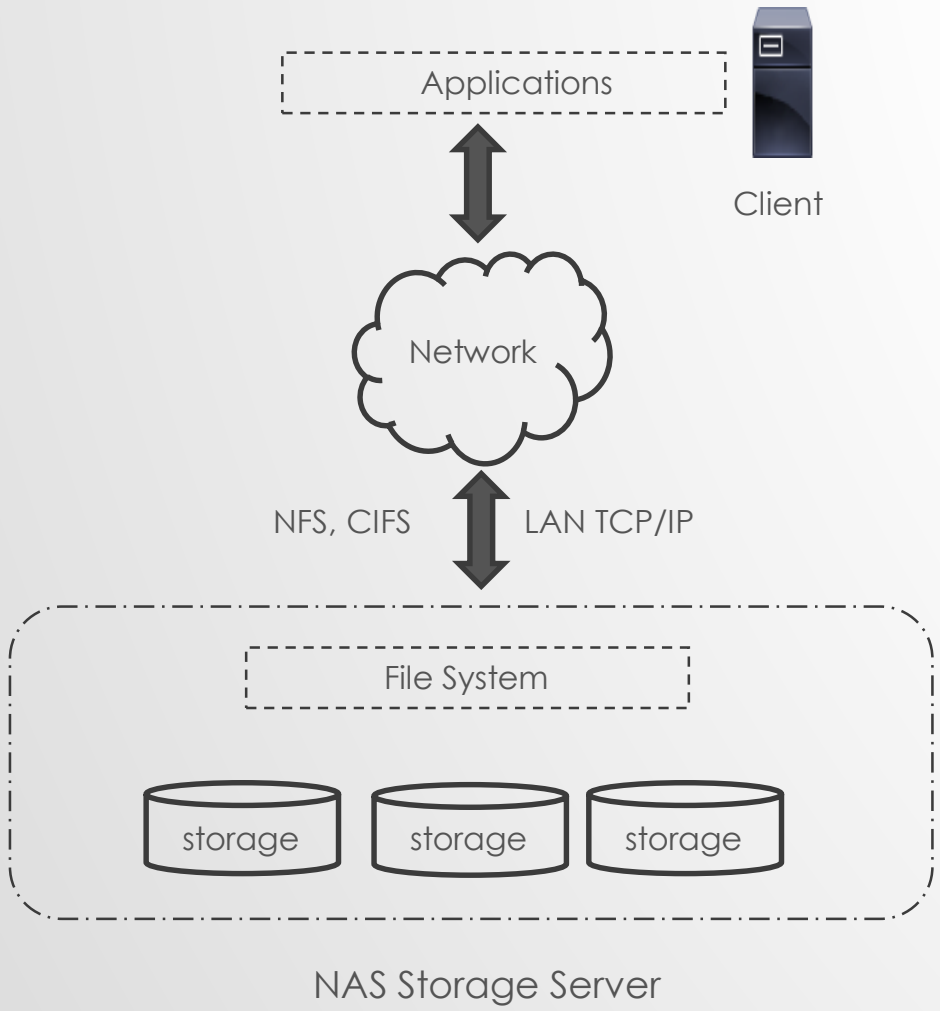
Massimiliano Sbaraglia

DAS (DIRECTLY ATTACHED STORAGE)



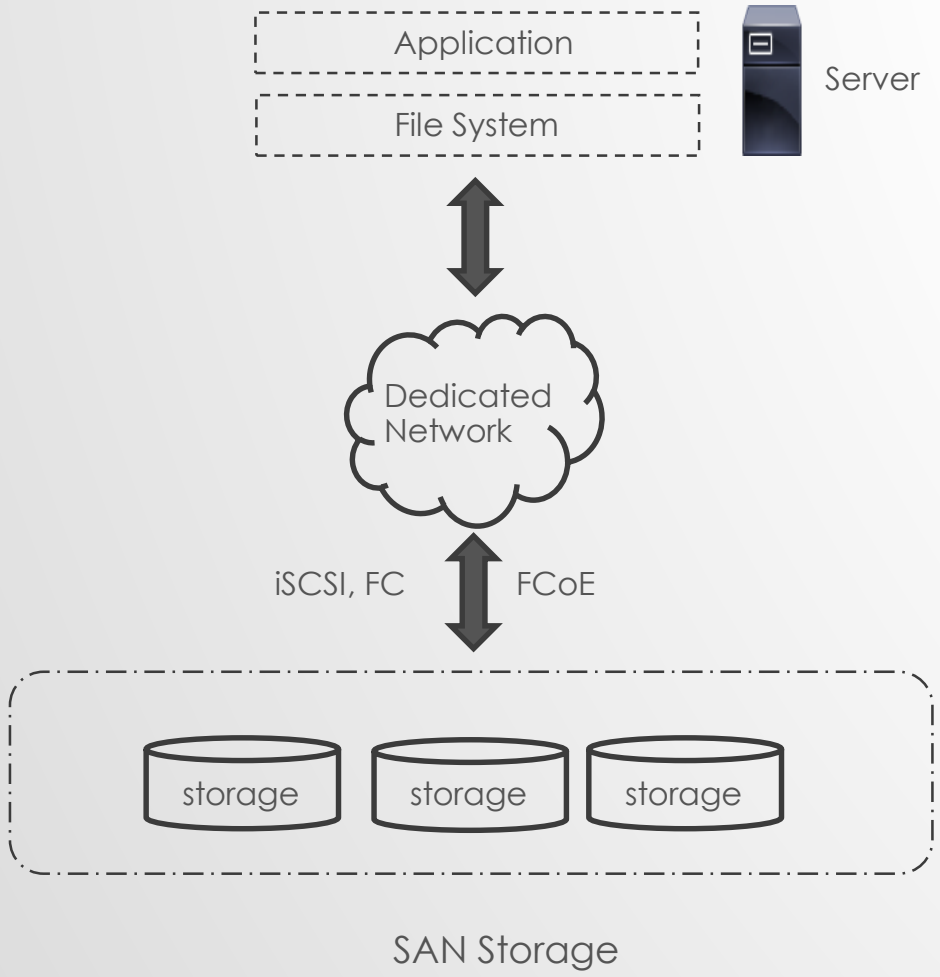
- Un DAS è uno storage sub-system direttamente collegato via cavo al server
- Hard Disk, Solid Drive, Optical Hard Drive, External Drives
- Può essere compost da un RAID (Redundant Array Indipendent Disk) combinando multipli hard drive all'interno di una unità logica di dischi
- Non esiste nessuna Network tra il collegamento di un sistema DAS ed un Server (nessuno switch, router, bridge, etc..)
- Protocolli quali ATA, SATA, SCSI, SAS, USB e FC permettono il collegamento di sistemi DAS

NAS (NETWORK ATTACHED STORAGE)



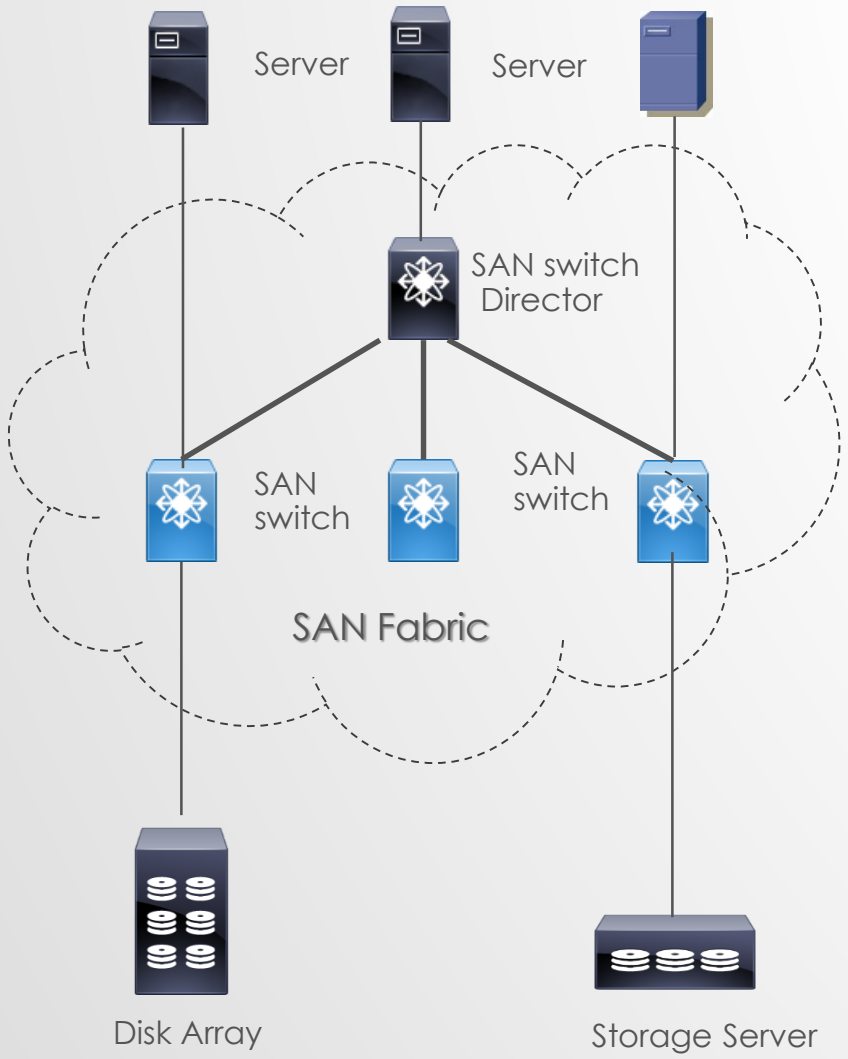
- Un NAS è un file-level storage server collegato attraverso una rete (netrwork) e provvede all'accesso dei dati in modo etereogeneo
- L'accesso ai dati, consolidato per sistemi UNIX è il protocollo NFS (network File System) e per i sistemi Windows (Microsoft) è il protocollo CIFS (Common Internet File Sustem)
- I NAS sono gestibili via browser digitando un indirizzp IP della rete
- Un NAS ha una interfaccia di rete per il collegamento alla rete aziendale
- I NAS sono indicati per lo scambio di piccolo quantità di informazioni ed il protocollo TCP/IP si presta bene a questa funzionalità.

SAN (STORAGE AREA NETWORK)



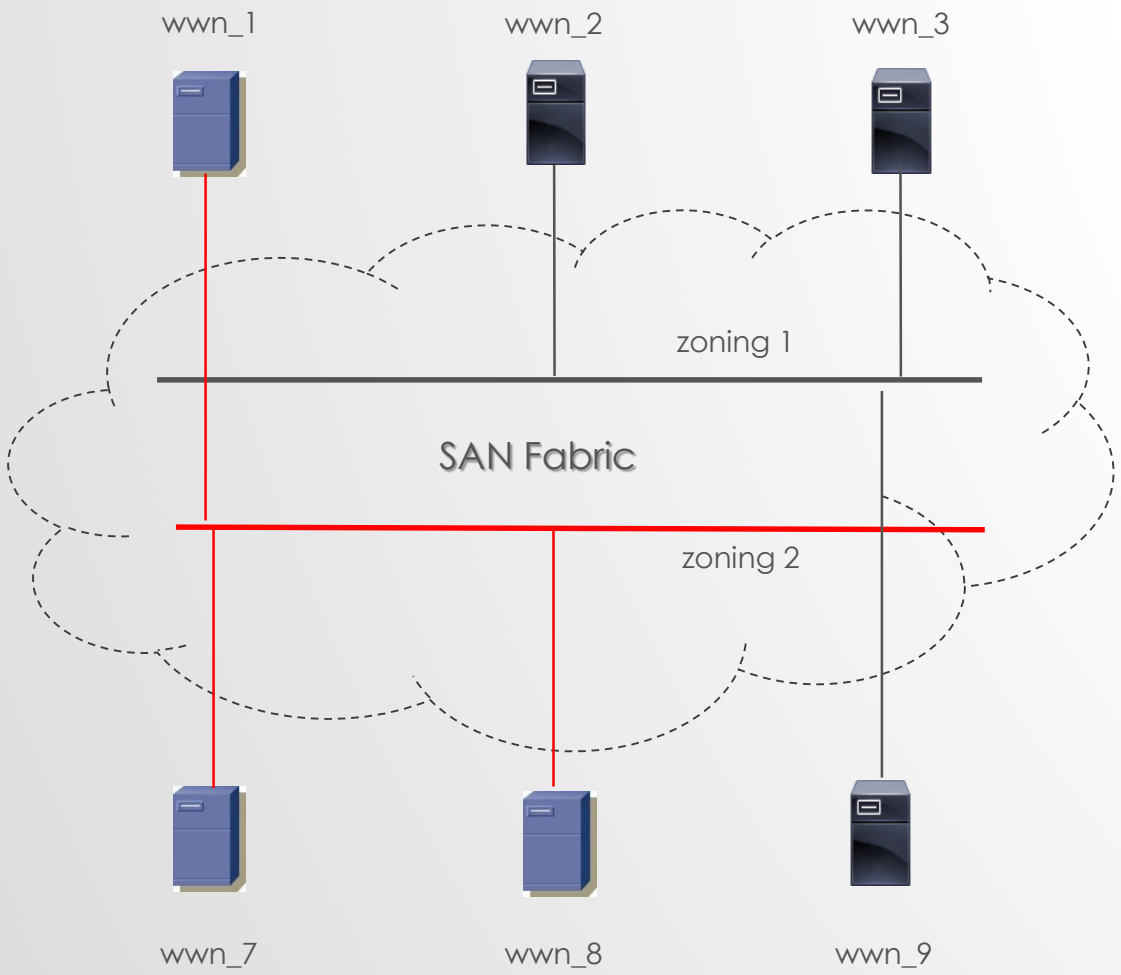
- Un SAN prevede una rete di switch dedicata per l'accesso dei dati tra blocchi di Servers
- La connessione verso devices di rete SAN è possibile grazie ad una estensione definita in una HBA (Host Bus Adpater)
- I protocolli impiegati sono iSCSI, FC, FCoE
- Una SAN si adatta bene ad ambienti dove lo scambio di dati è di grandi dimensioni (Terabyte di dati)
- Le SAN sono dotate di sistemi di protezione dei dati e ridandanza di hardware per evitare interruzioni di servizio
- L'assenza di interruzioni di servizio è garantita dalla presenza di una rete di tipo lossless Ethernet, indispensabile per il trasporto di pacchetti SCSI incapsulato dentro pacchetti Fiber Channel over Ethernet
- In una SAN vi sono I concetti di VSAN, Zoning (WWN World Wide Name) e LUN (Logical Unit Number to Server)

SAN CONCEPT



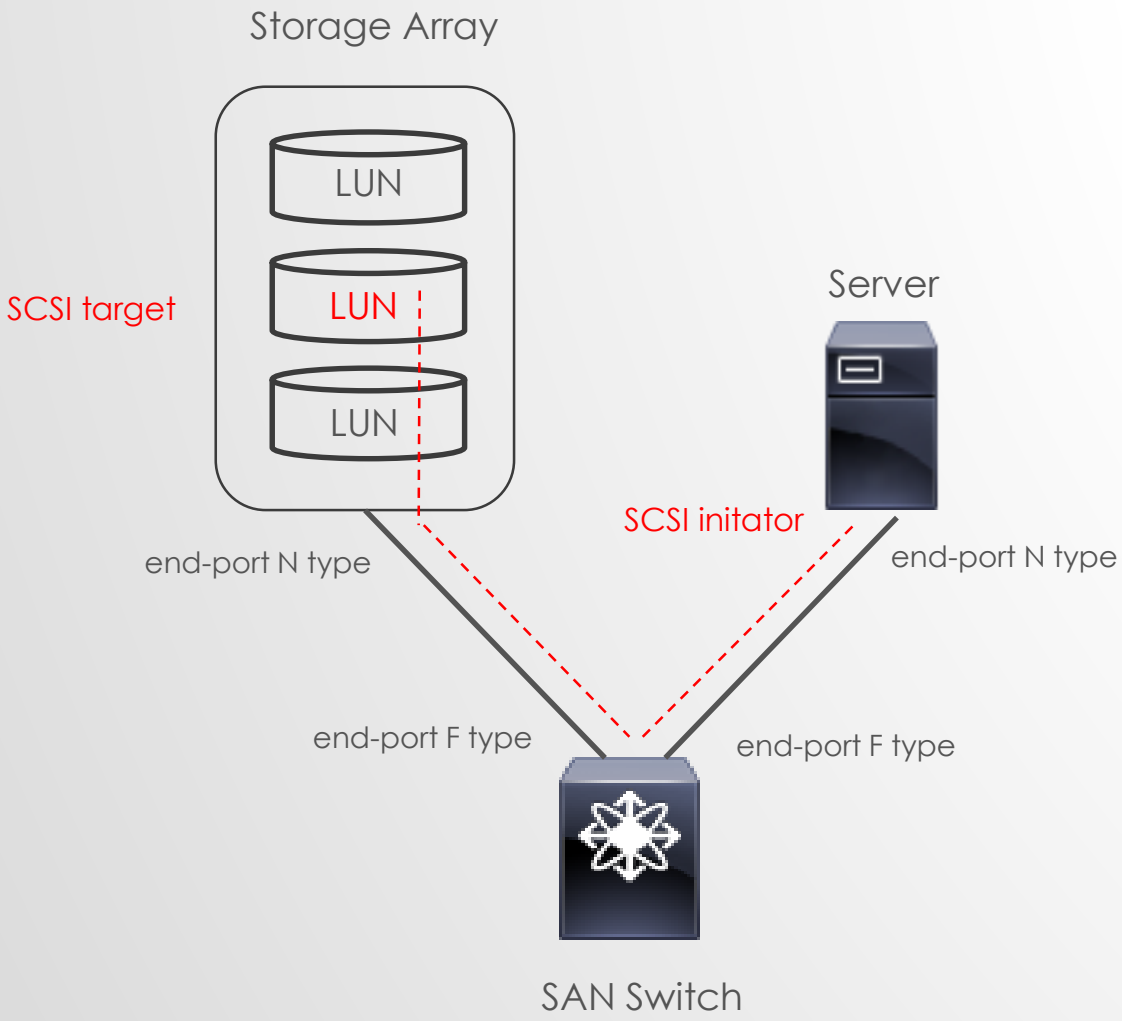
- Una SAN è un insieme di switch connessi tra loro a formare una Fabric, di cui solo uno ha funzioni di director ed è suo compito assegnare un domain-ID agli switch secondari al momento di attivazione di una SAN
- Una SAN deve essere caratterizzata da resilienza, ossia resistere a malfunzionamenti della rete (attraverso la duplicazione di link)
- Una SAN deve essere caratterizzata da ridondanza con la duplicazione dei suoi component sino ad arrvare alla duplicazione dell'intera Fabric (in questo caso abbiamo una Dual-Fabric SAN)
- Una configurazione Dual-Fabric SAN consente di far fronte non solo a guasti di tipo hardware, ma anche a errori operative; consente anche, se realizzata con dispositivi posizionati in ambienti differenti (non all'interno di uno stesso rack ad esempio), di effettuare manutenzione o sostituzione di apparati senza impatto per la produzione

ZONING CONCEPT



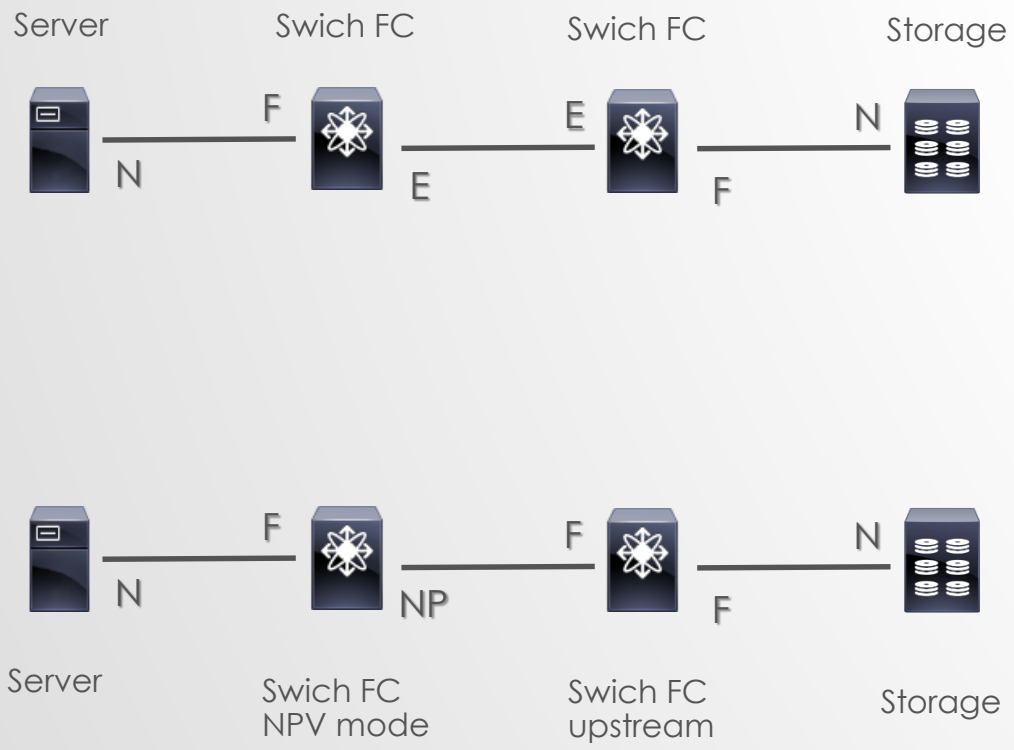
- Lo zoning è una tecnica che permette di raggruppare ed isolare un gruppo di dispositivi all'interno di una zona appartenente ad una SAN
- E' utile per separare ambienti con diversi sistemi operative quali Unix e Windows
- E' utile per scopi di maggiore sicurezza dedicati a diversi ambienti di sviluppo, test e collaudo sulla stessa SAN
- Lo zoning può essere effettuato in modalità o hardware o software, quest'ultimo è implementato da parametric di configurazione a livello switch
- I membri di una zona sono identifcati tramite WWN (World Wide Name), associato a livello di nodo oppure a livello di porta
- Per indiviaduare i dispositivi storage disponibile quando un server di collega ad una SAN vi è un servizio a livello di switch chiamato SNS (Simple Name Server)

LUN CONCEPT (LOGICAL UNIT NUMBER TO SERVER)



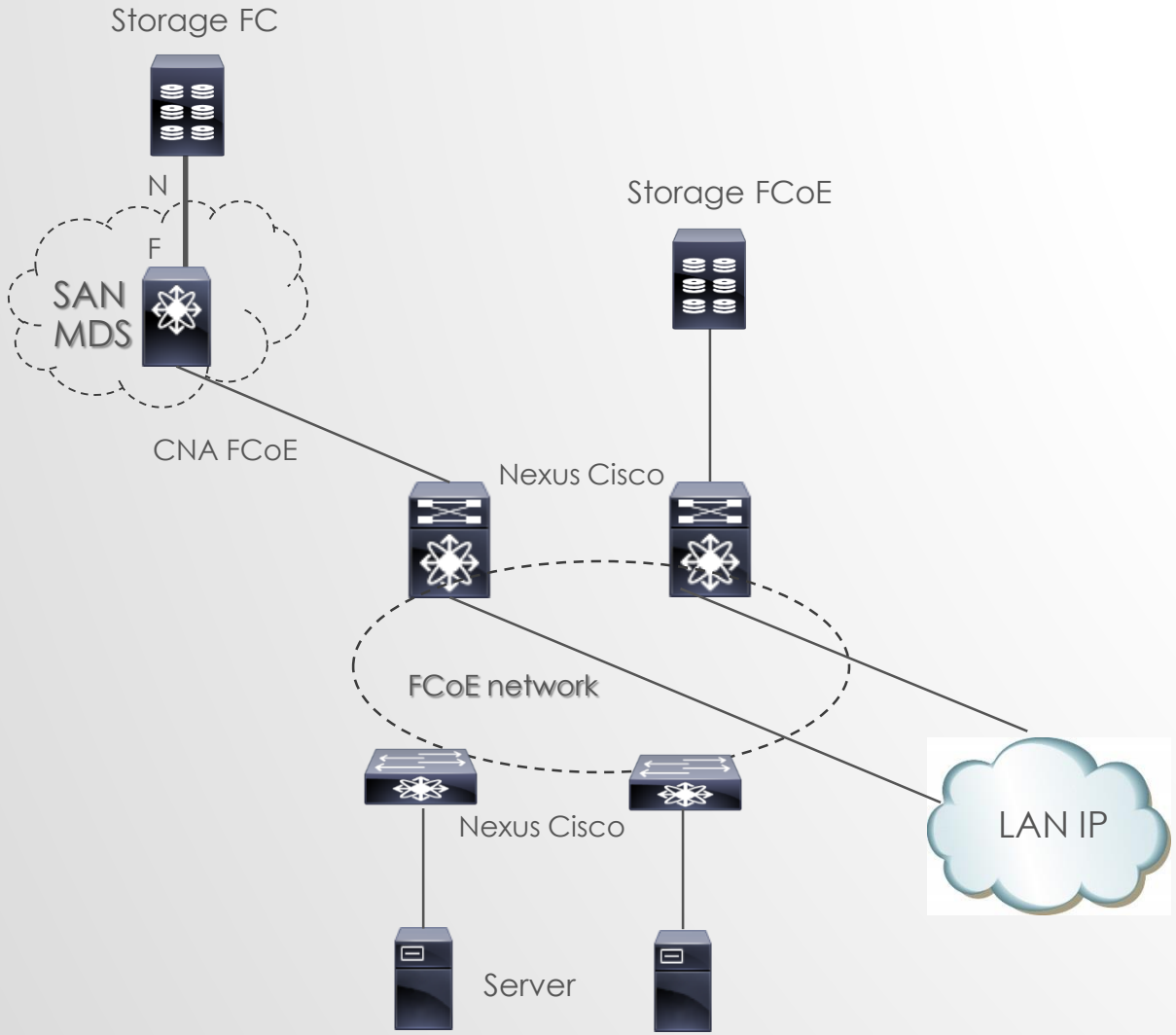
- In uno Storage Array (un insieme di dischi storage)deve essere configurator il parametro LUN , quando questi sono collegati ad una SAN Fabric
- LUN zoning è una configurazione a livello switch e provvede ad isolare flussi di dati I/O attraverso una Fabric tra end-ports
- LUN masking è una configurazione a livello storage controller (può essere fatta anche a livello switch) ed è ristretto all'accesso di un server per designare SCSI/iSCSI target ed i relativi LUN a lui assegnati

FIBRE CHANNEL CONCEPT



- FC point to point: permette una connessione diretta tra Servers senza impiego di network devices
- FC Arbitrated Loop: permette una connessione sino a 127 devices in una connessione looped mode
- FC Fabric: permette di scambiare flussi di dati attraverso switch Fiber Channel
- N port type (Node): modalità di connessione di un Server point to point oppure via Fabric
- NL port type (Node Loop): come sopra ma in una configurazione looped mode
- F port type (Fabric): tipo di porta appartente ad uno switch FC in connessione con una porta N type
- FL port type (Fabric Loop): come una F port ma in una configurazione looped mode
- E port type (Expansion): appartente ad uno switch FC che si collega ad altri switch FC con porte di tipo E (a creare una connessione di tipo trunk ISL)
- NPV (N port virtualization): emula una N port connessa ad un upstream switch con F port (senza connessione ISL)

FIBRE CHANNEL OVER ETHERNET CONCEPT

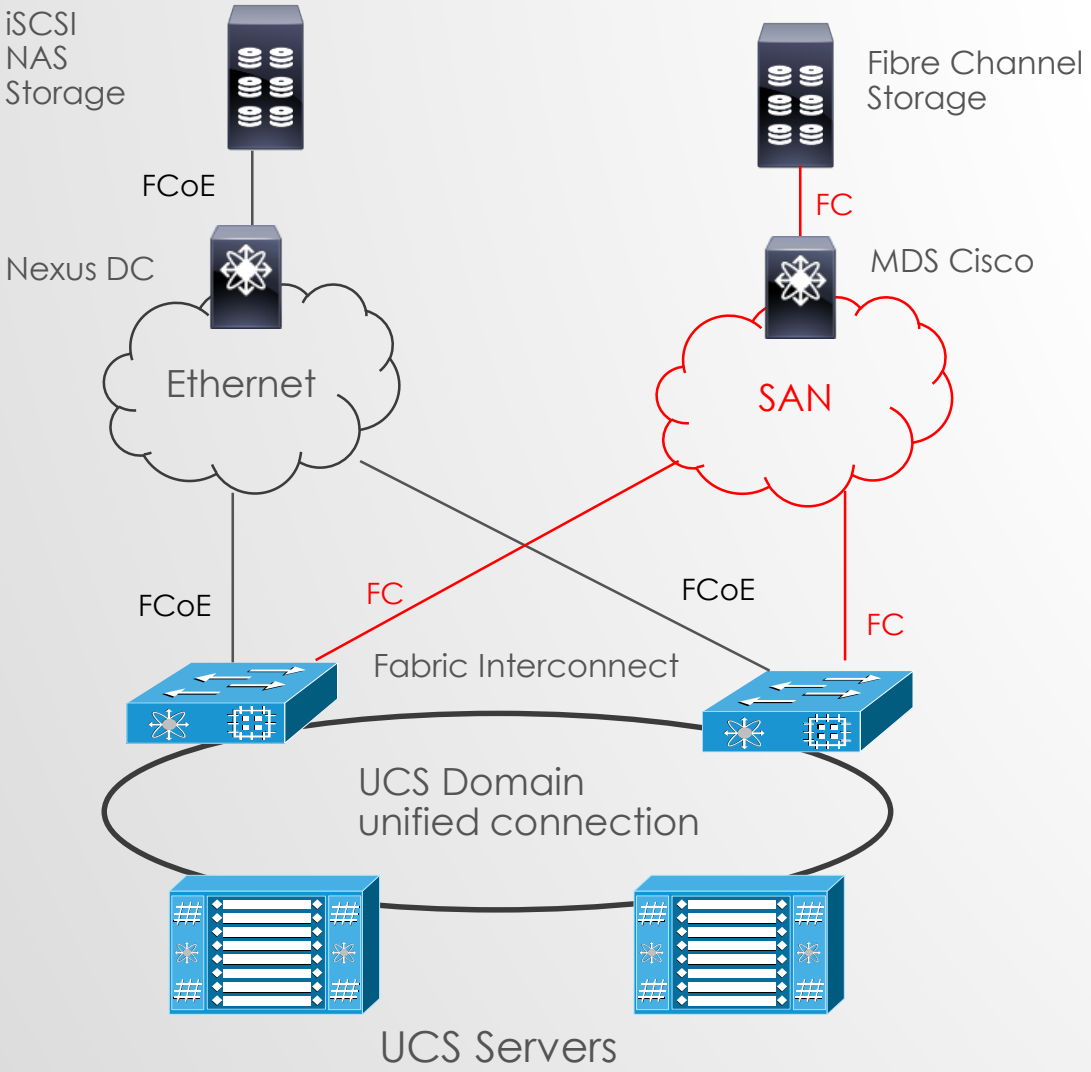


- FCoE mappa le frame FC su una rete IEEE 802.3 Ethernet full-duplex con connessioni a 10G senza modificare tutte le funzionalità proprie del FC (zoning, lun, etc..)
- Sono necessarie apposite schede di rete chiamate CNA (Converged Network Adpater) e switch Ethernet per il trasporto e l'instradamento 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

ARCHITETTURE UNIFIED COMPUTING AND VMWARE

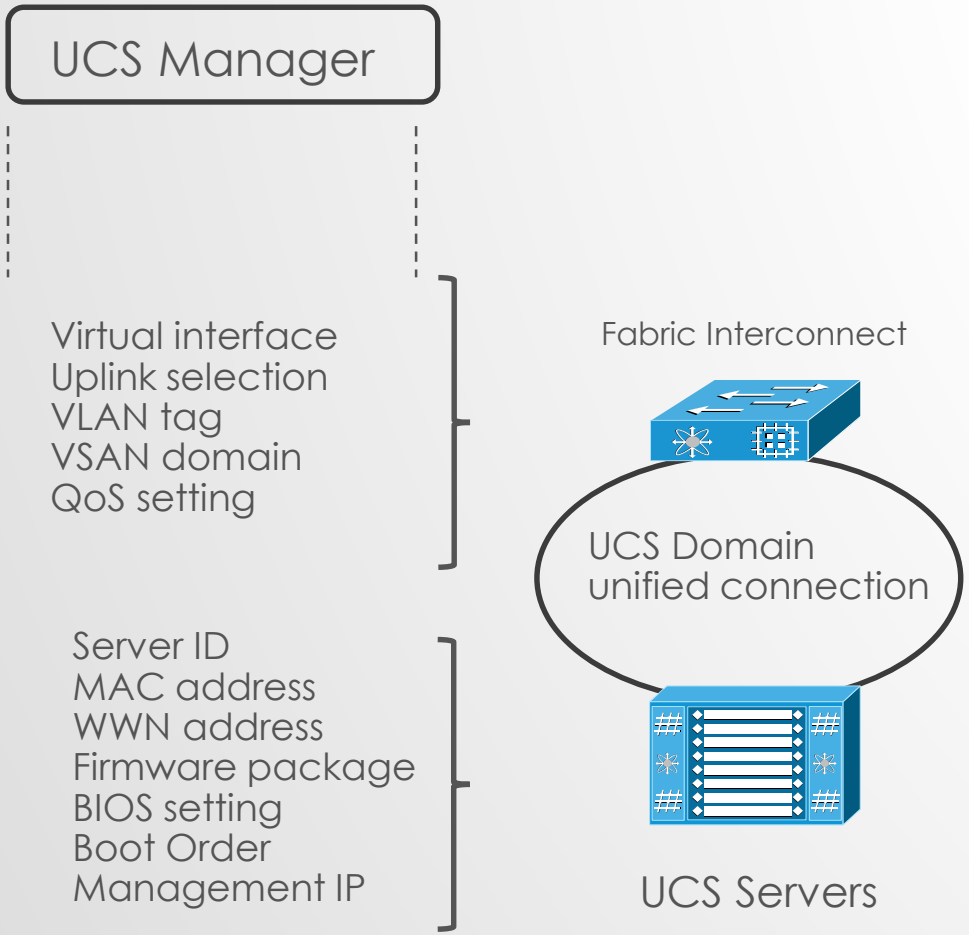
Massimiliano Sbaraglia

UNIFIED COMPUTING SYSTEM (UCS)



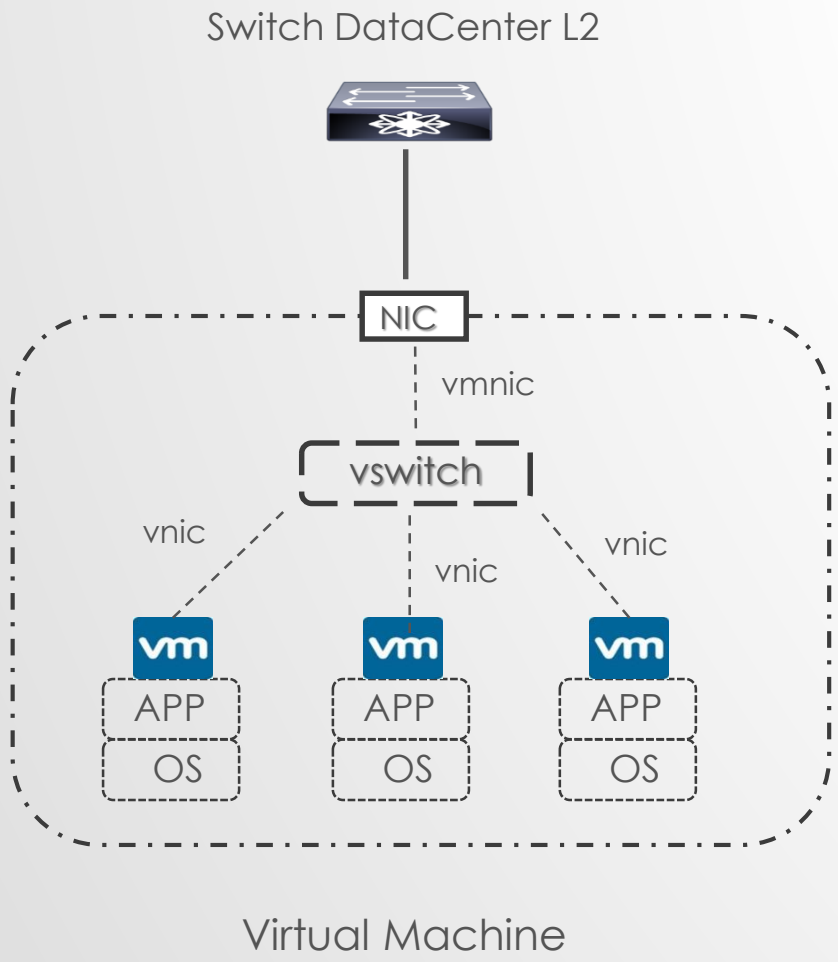
- Unified Computing System (UCS) significa un insieme di Servers, Storage e tecnologie di virtualizzazione all'interno di una stessa architettura.
- L'interoperabilità tra un sistema UCS Servers e le infrastrutture di rete IP e SAN è gestita da devices chiamati Fabric Interconnct
- Servers blade UCS serie B
- Servers Rack UCS serie C
- Servers di archiviazione UCS serie S
- Software di gestione UCS Manager
- Fabric Interconnect UCS + Fabric Interconnect Extender

UCS SERVICE PROFILE



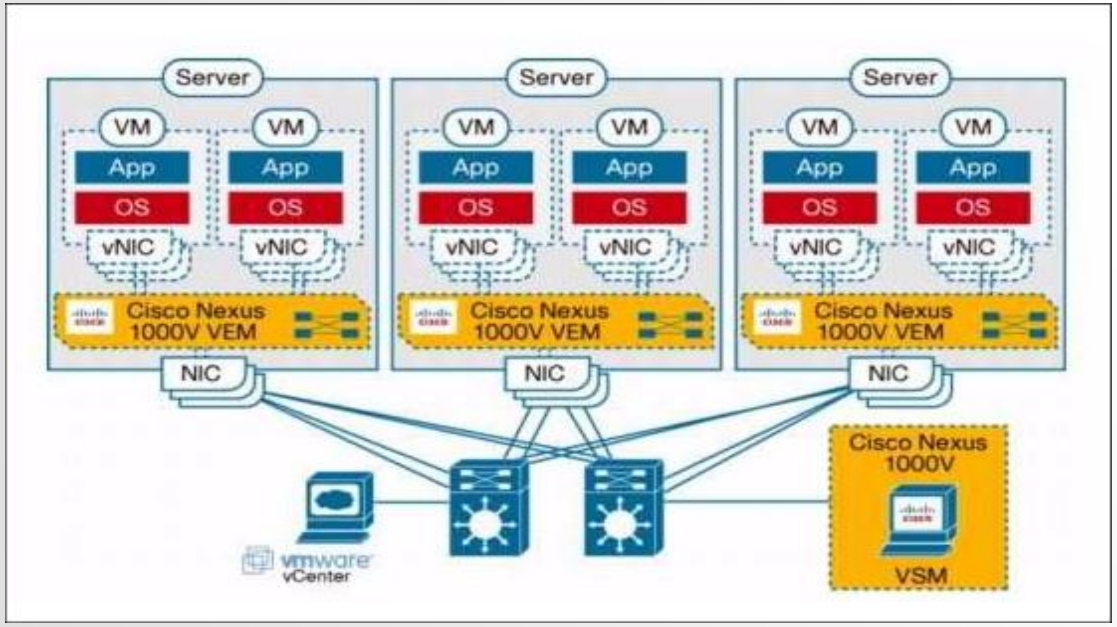
- Un blade o Rack UCS server deve essere associate ad un " service profile " ed ogni associazione ha una relazione 1:1 con un server.
- Quando un service profile è associato ad un server, sia fabric-interconnect che le component del server (adpters, BIOS, etc..) sono configurati per accordarsi su specifici parmetri (virtual-interface eth o FC, unico ID, LAN connectivity (MAC address), SAN connectivity (wwn), firmware package and version, IP address di management, etc…
- Un service pofile è una entità virtuale all'interno del sistema di gestione UCS Manager

VIRTUAL SERVER CONCEPT



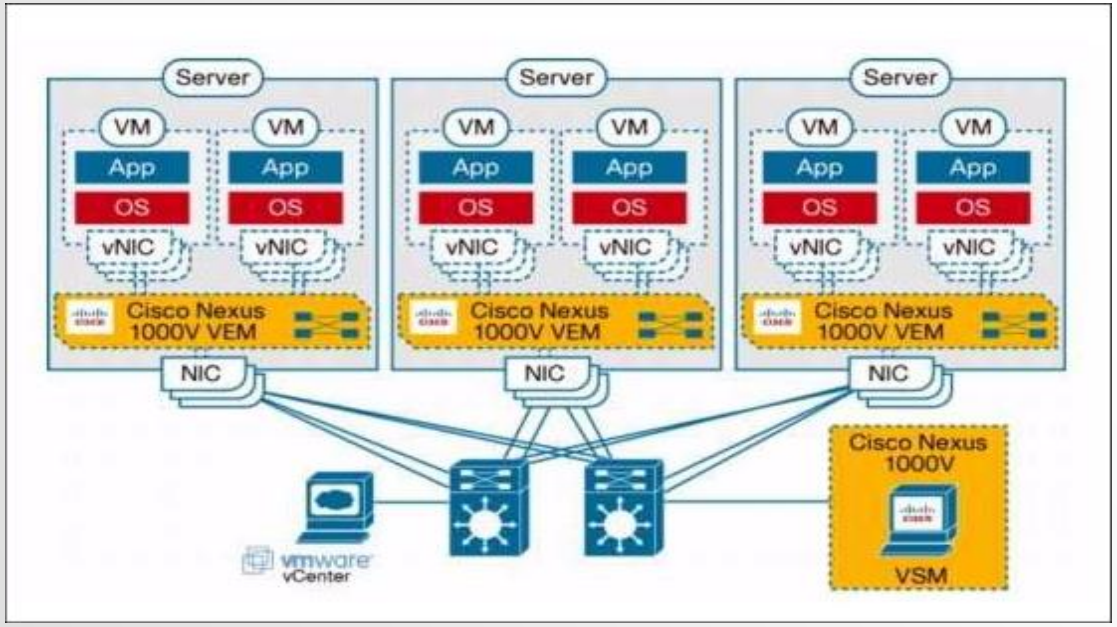
- Una VM (Virtual Machine) emula un server fisico per sistema operativo, applicazioni, IP address e collegamento verso una rete (vnic)
- Vmware ha introdotto il concetto di vswitch (virtual switch) che altro non è che un Hipervisor che emula tutte le funzionalità di un vero layer 2 switch
- Questo vswitch, 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 notofocation, failure behavior)
- Cisco ha introdotto Nexus 1000V quale element virtuale che emula le funzionalità di un distribute vswitch vmware DVS attraverso proprie API (Application Programmable Interface) rilasciate attraverso NX-OS vCenter operations

NEXUS 1000V CISCO VSM (VIRTUAL SUPERVISOR MODULE)



- 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

NEXUS 1000V CISCO VEM (VIRTUAL ETHERNET MODULE)

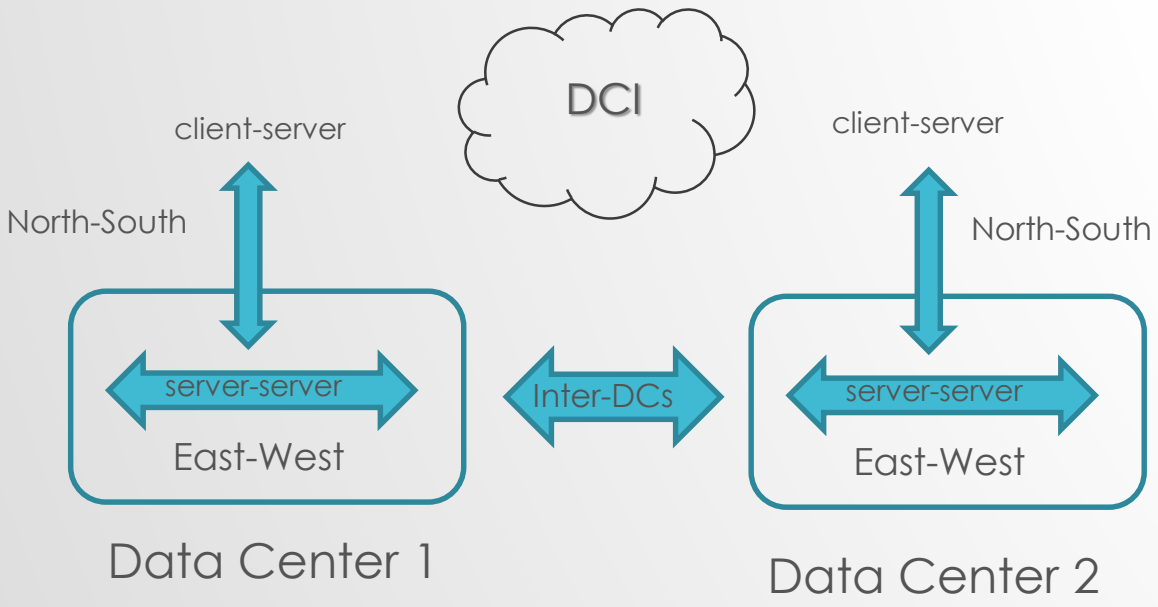


- VEM (Virtual Ethernet Module) condivide un dominio di broadcast (vlan) per il controllo layer 2 con il VSM
- Ogni VEM richiede uno specific VM-Kernel interface (vmknic) per comunicare con il VSM (layer 3 control mode)
- Port Profile è una collezione di interface-level configuration per creare delle network policy (il port profile non solo è per il Nexus 1000 ma può essere presente anche in altri NX-OS device)

DATACENTERS INFRASTRUCTURE DCI AND L2-EXTENSION

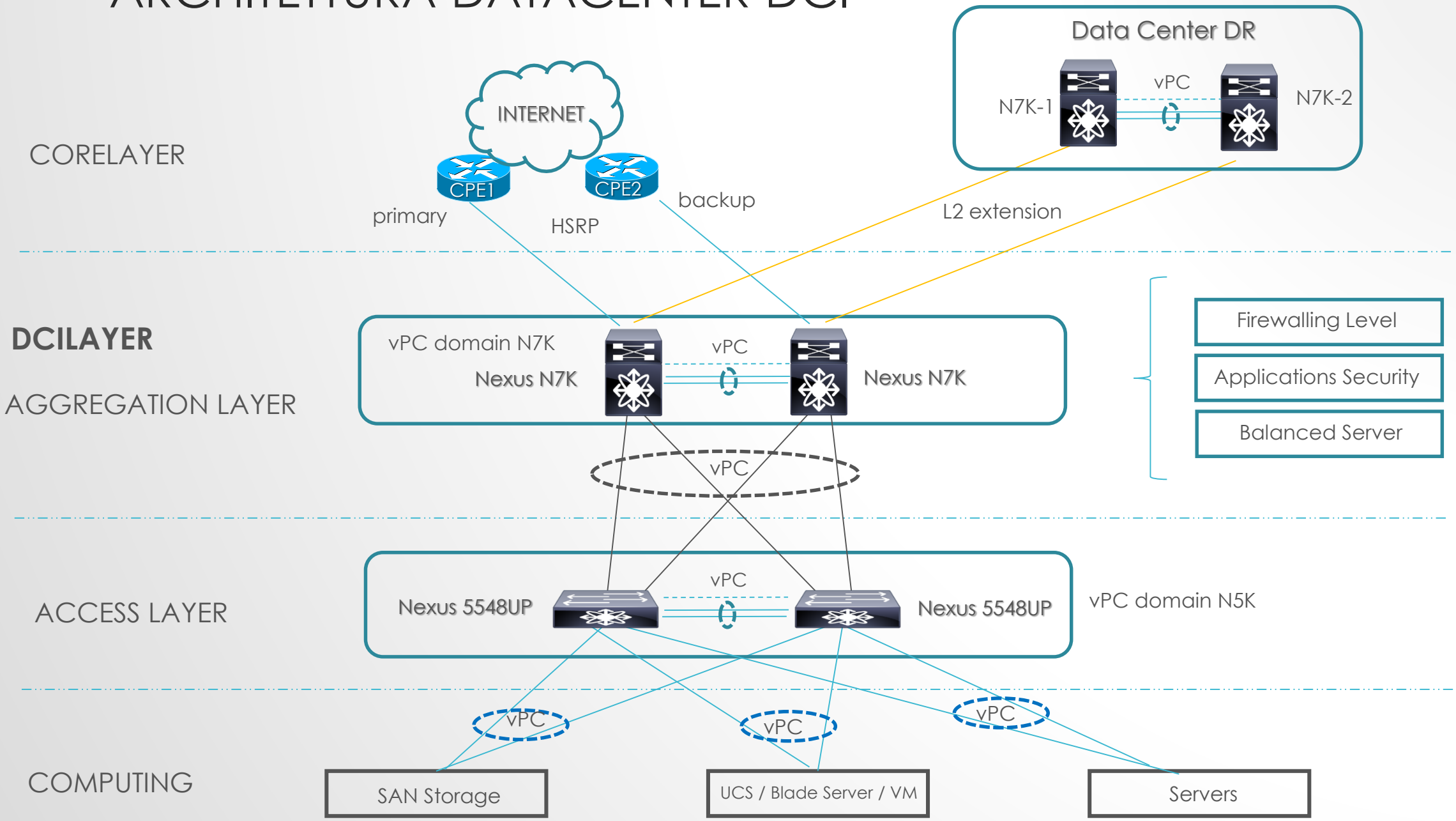
Massimiliano Sbaraglia

DCI LAYER 2 AND LAYER 3 CONCEPT

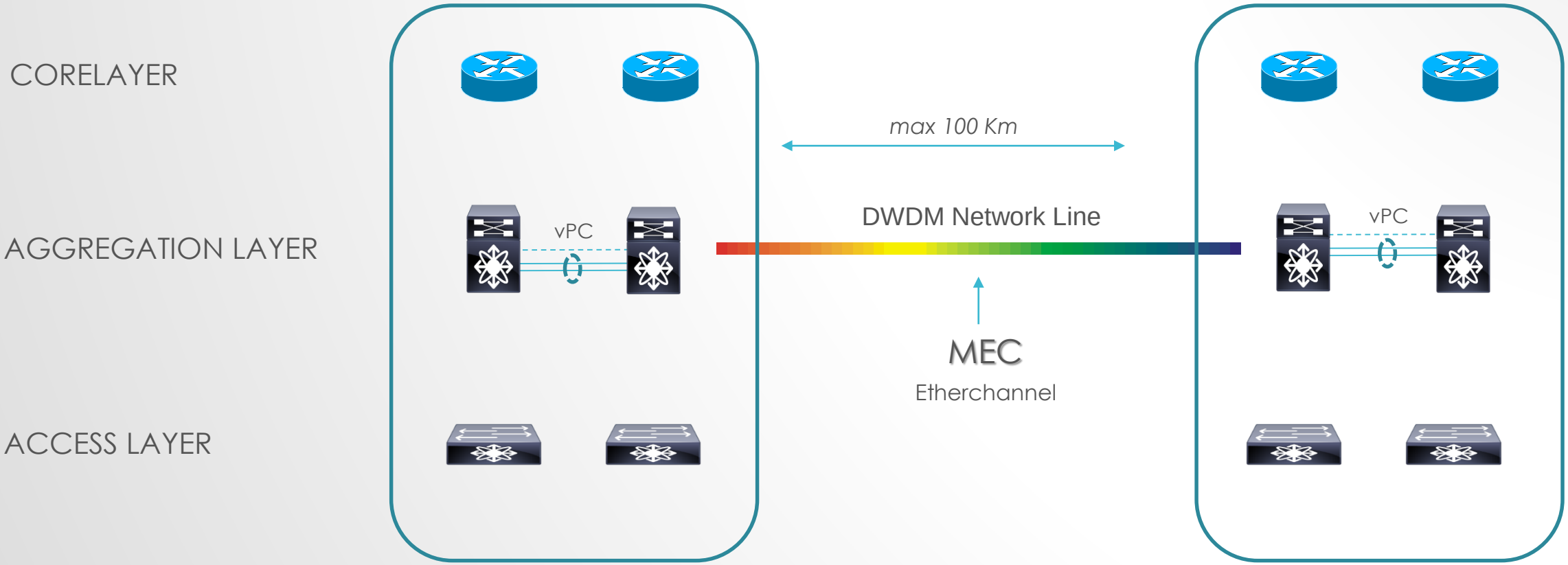


- DCI Layer 2 è inerente a tecniche di mobilità di VM e IP address
- DCI Layer 3 riguarda soprattutto ad operazioni di transazione e replicazione di database in cluster, e la sincronizzazioni di applicazioni in cluster
- Replicazioni Sincrone di dati Storage (generalmente utilizzato all'interno di un 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



DATACENTER DCI LAYER 2 DARK-FIBER POINT-TO-POINT

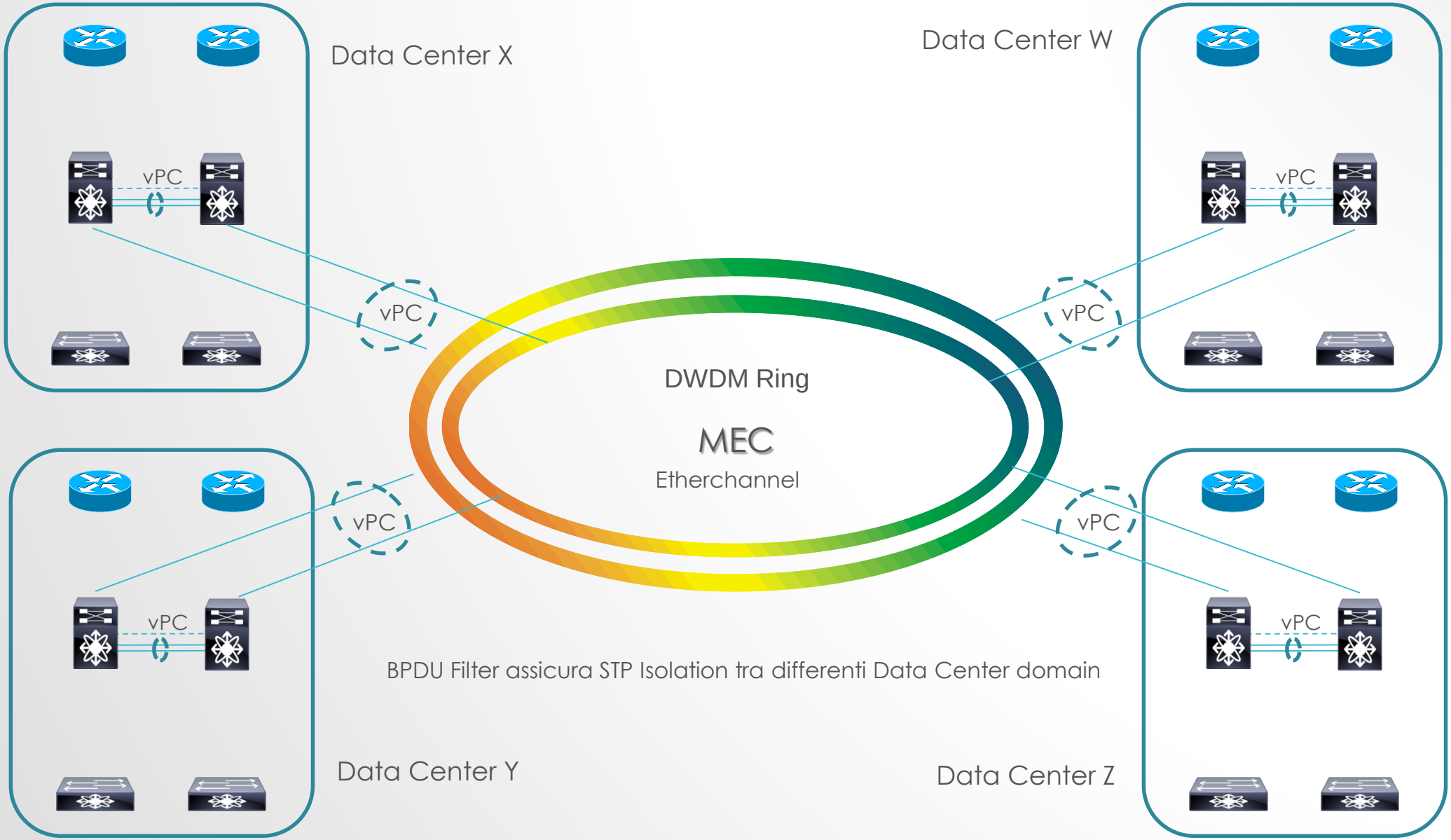


BPDU Filter assicura STP Isolation tra differenti Data Center domain

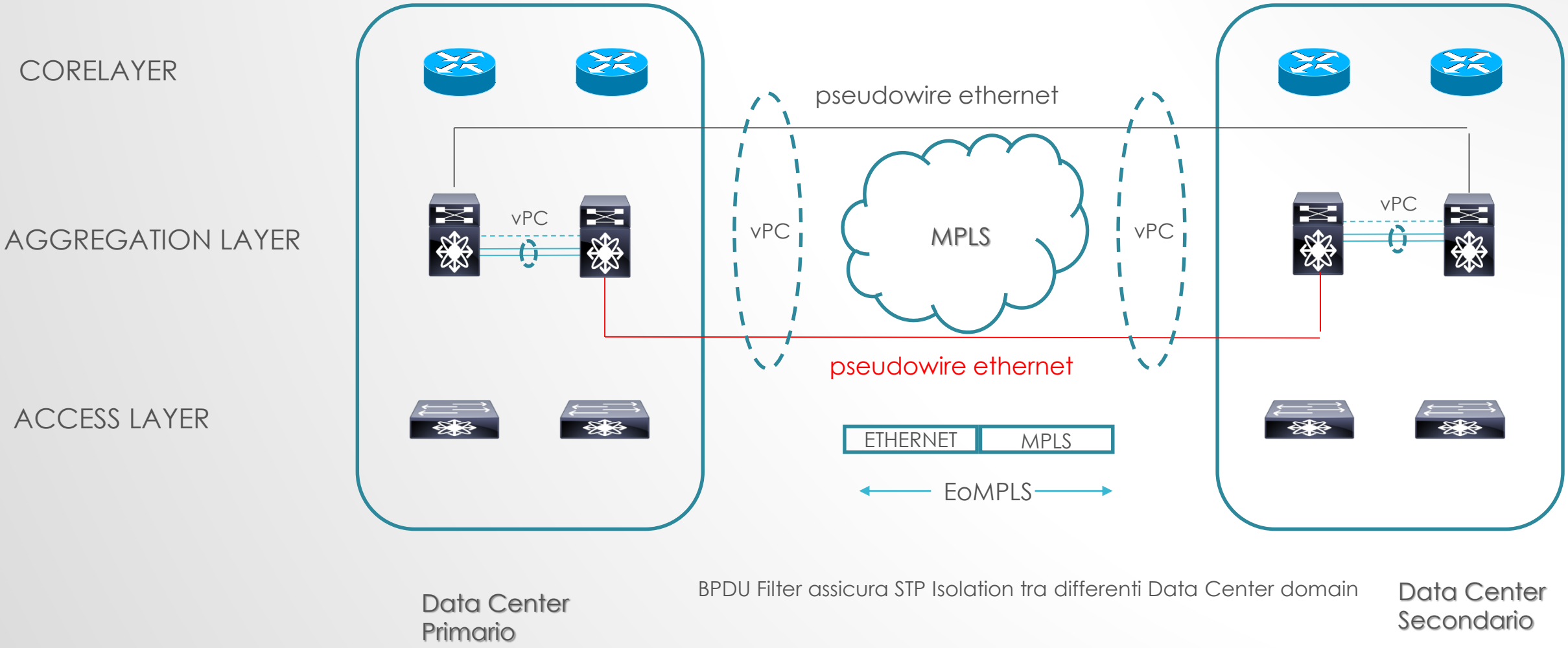
Data Center
Primario

Data Center
Secondario

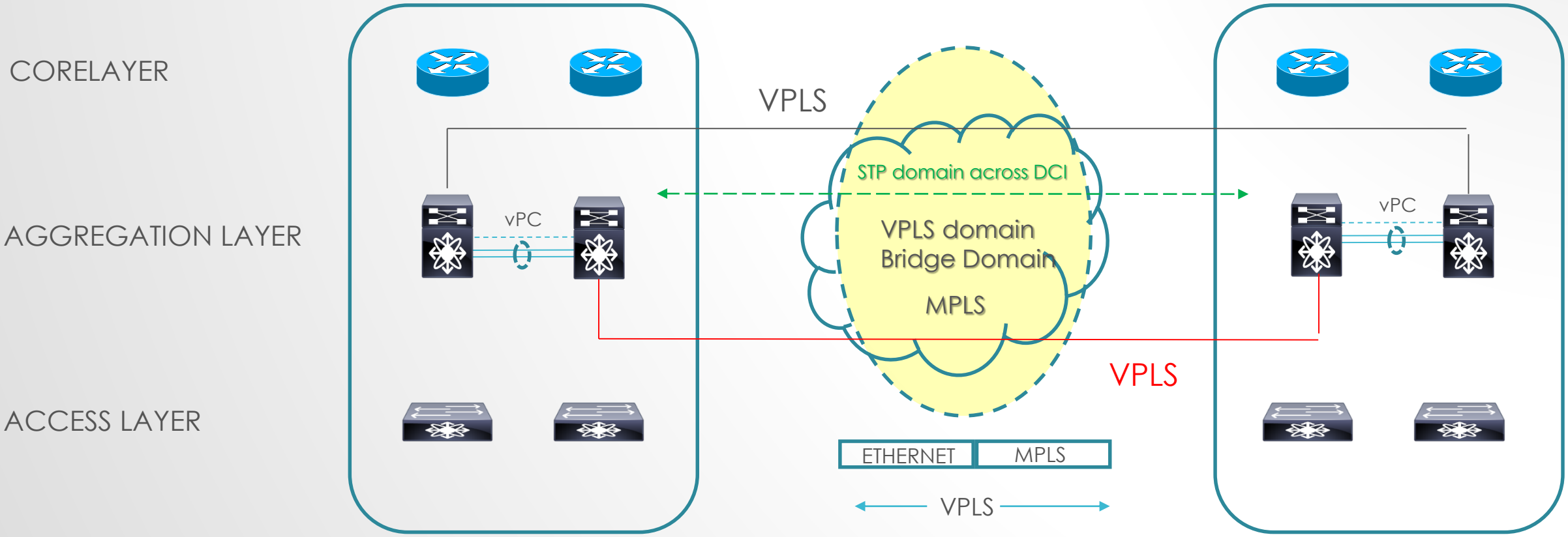
DATACENTER DCI LAYER 2 DARK-FIBER RING



DATACENTER DCI LAYER 2 PSEUDOWIRE ETHERNET P2P



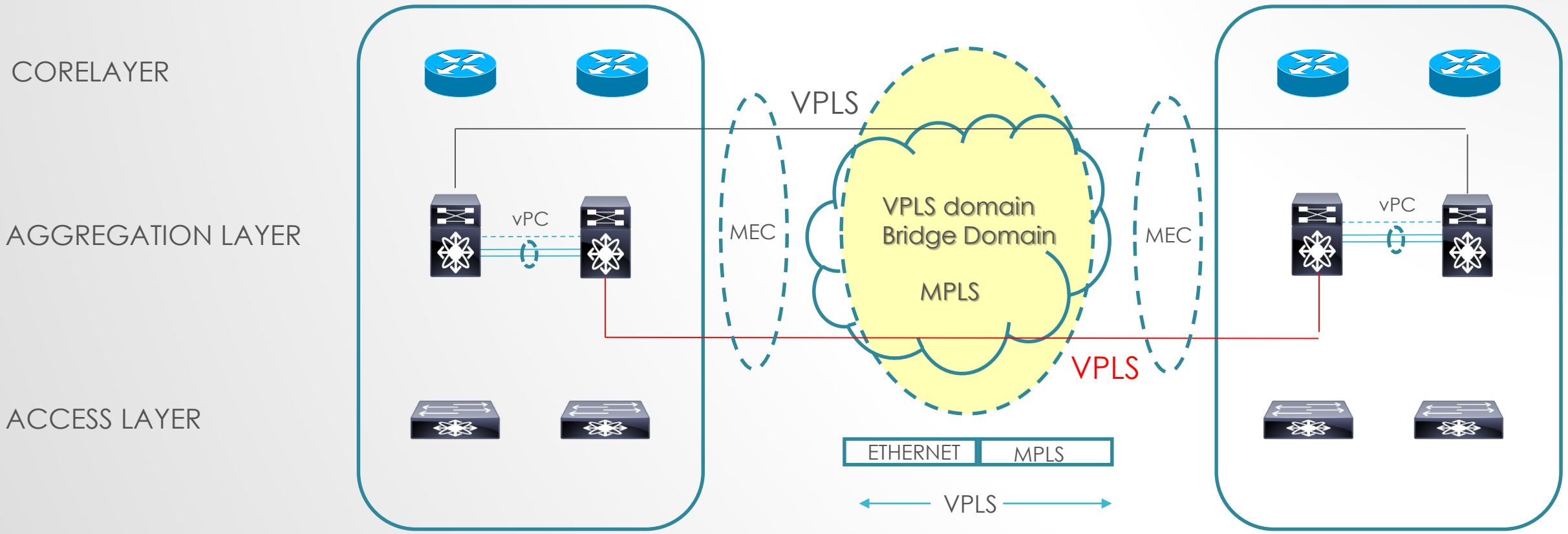
DATACENTER DCI LAYER 2 VPLS STANDARD



Il flooding del dominio STP è qualcosa di indesiderato via DCI

Soluzione: introduzione del MEC into VPLS

DATACENTER DCI LAYER 2 A-VPLS (ADVANCED)



Data Center Primario

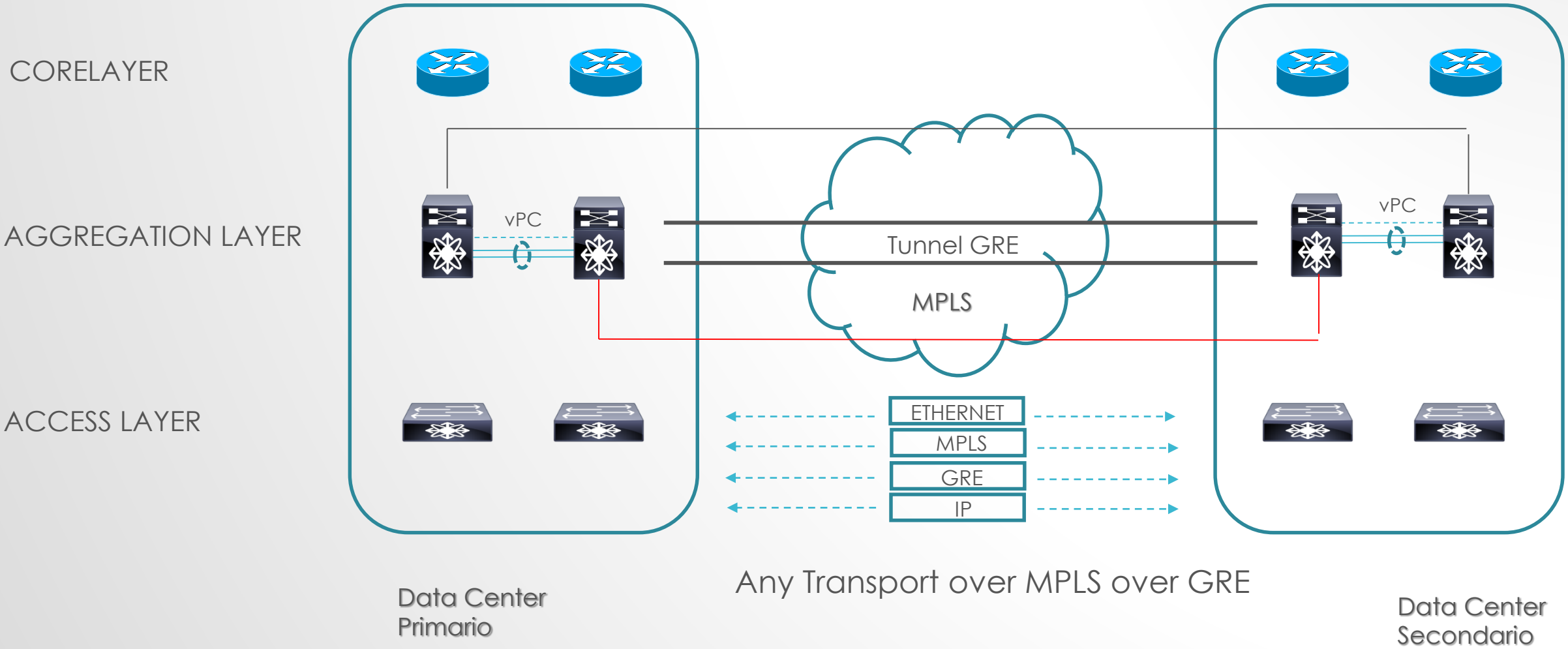
Load-balance traffic across multiple core interface using ECMP

CLI enhancements to facilitate configuration of L2VPN A-VPLS

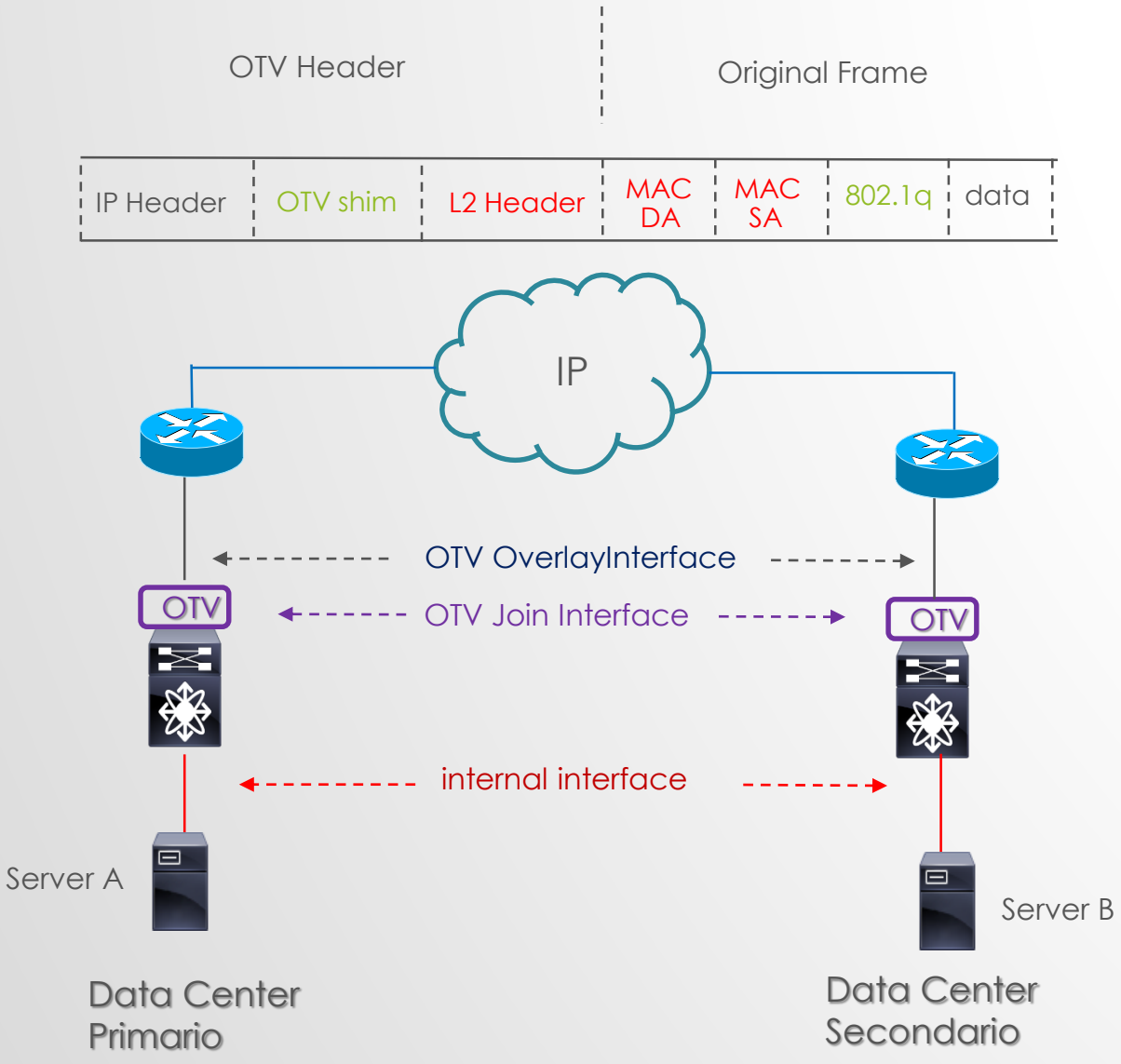
Redundant DCI with edge switches roles

Data Center Secondario

DATACENTER DCI LAYER 2 WITH GRE TUNNEL

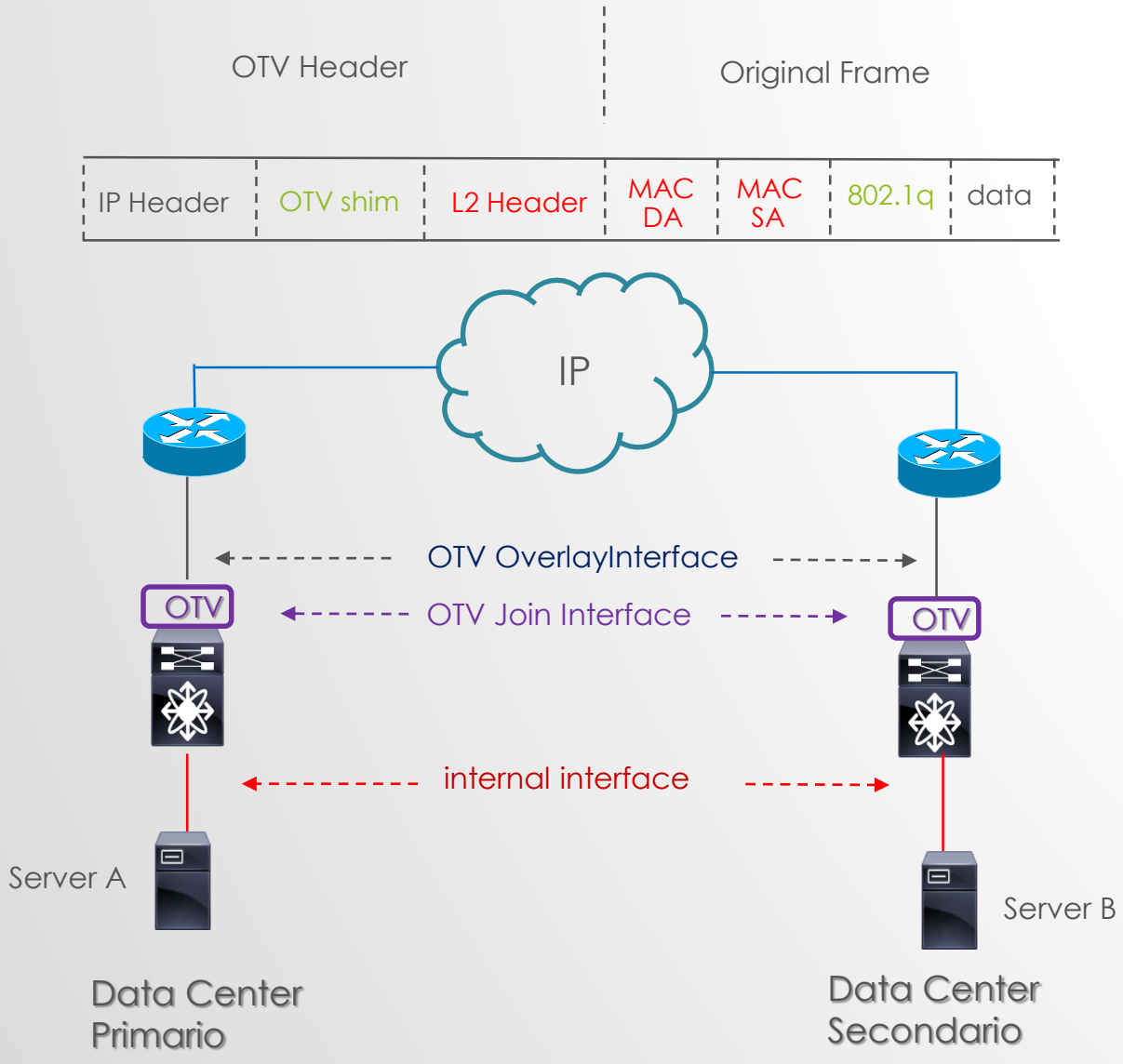


DCI OTV CISCO (OVERLAY TRANSPORT VIRTUALIZATION)



- OTV è una infrastruttura inter-datacenters e provvede a L2 extensions presenrvando 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)



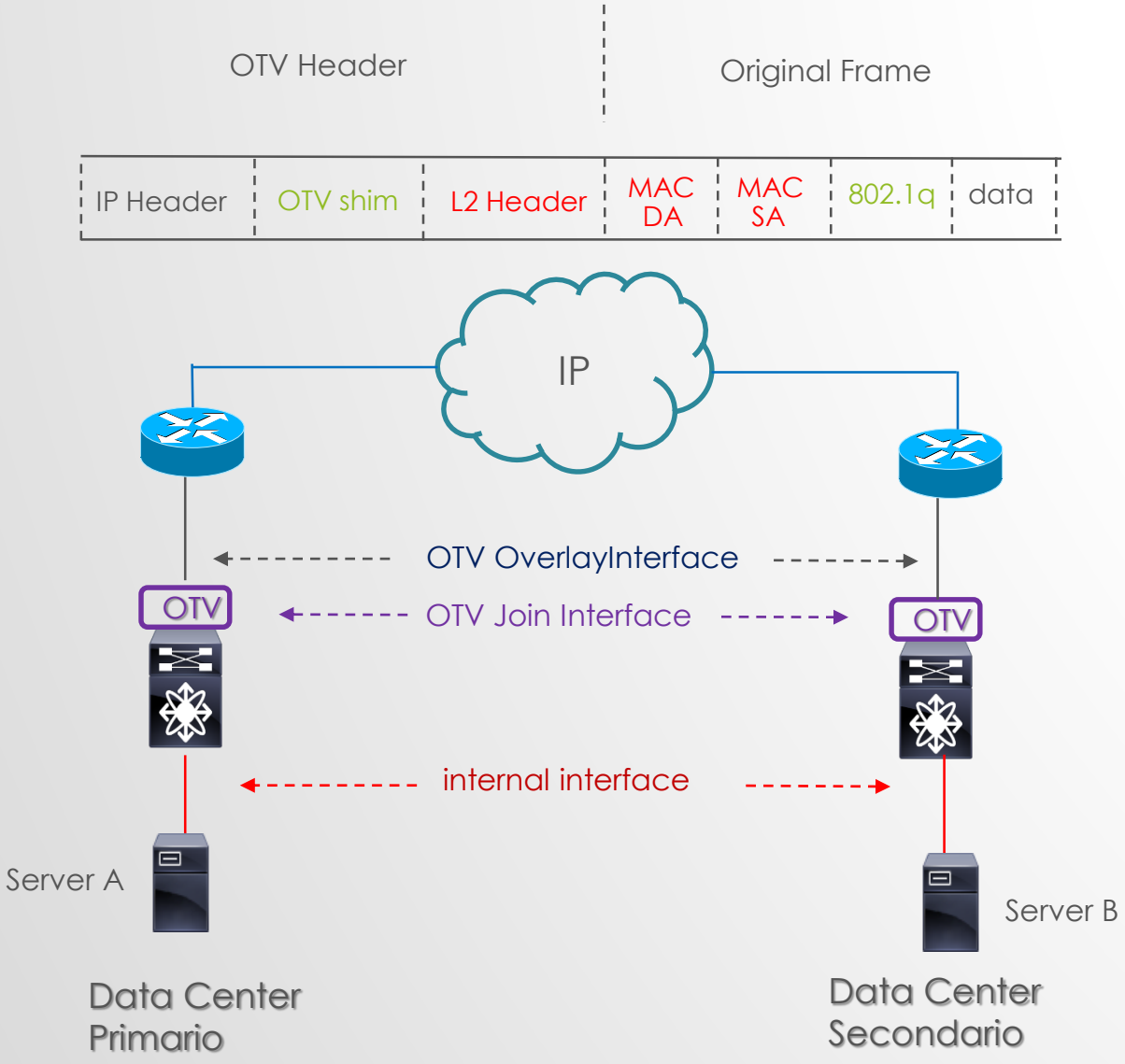
- **OTV Edge Device:** performa le funzionalità e le operazioni OTV; riceve le frame ethernet traffic per tutte le vlans soggette ad L2-extensions tra data centers OTV peers e dinamicamente le incapsula dentro IP packets che sono trasmessi via overlay transport infrastructure

- **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

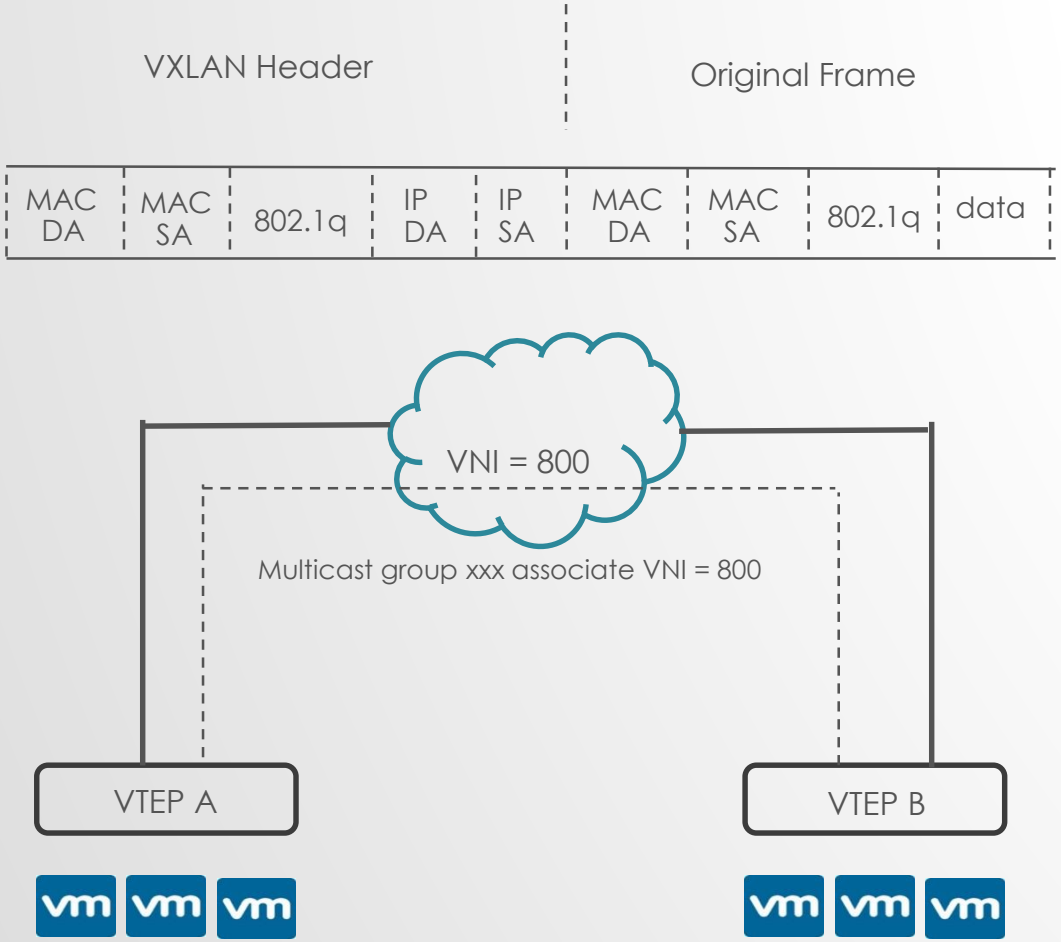
- **OTV overlay interface:** sono interfacce logiche virtuali dove risiede tutta la configurazione OTV; incapsula le frame layer 2 in IP unicast o multicast packets che sono trasmesse verso altri datacenters. Questo pemette agli edge device di performare un dinamico encapsulations.

DCI OTV CISCO (OVERLAY TRANSPORT VIRTUALIZATION)



- **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 VXLAN (VIRTUAL EXTENSIBLE LAN)



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

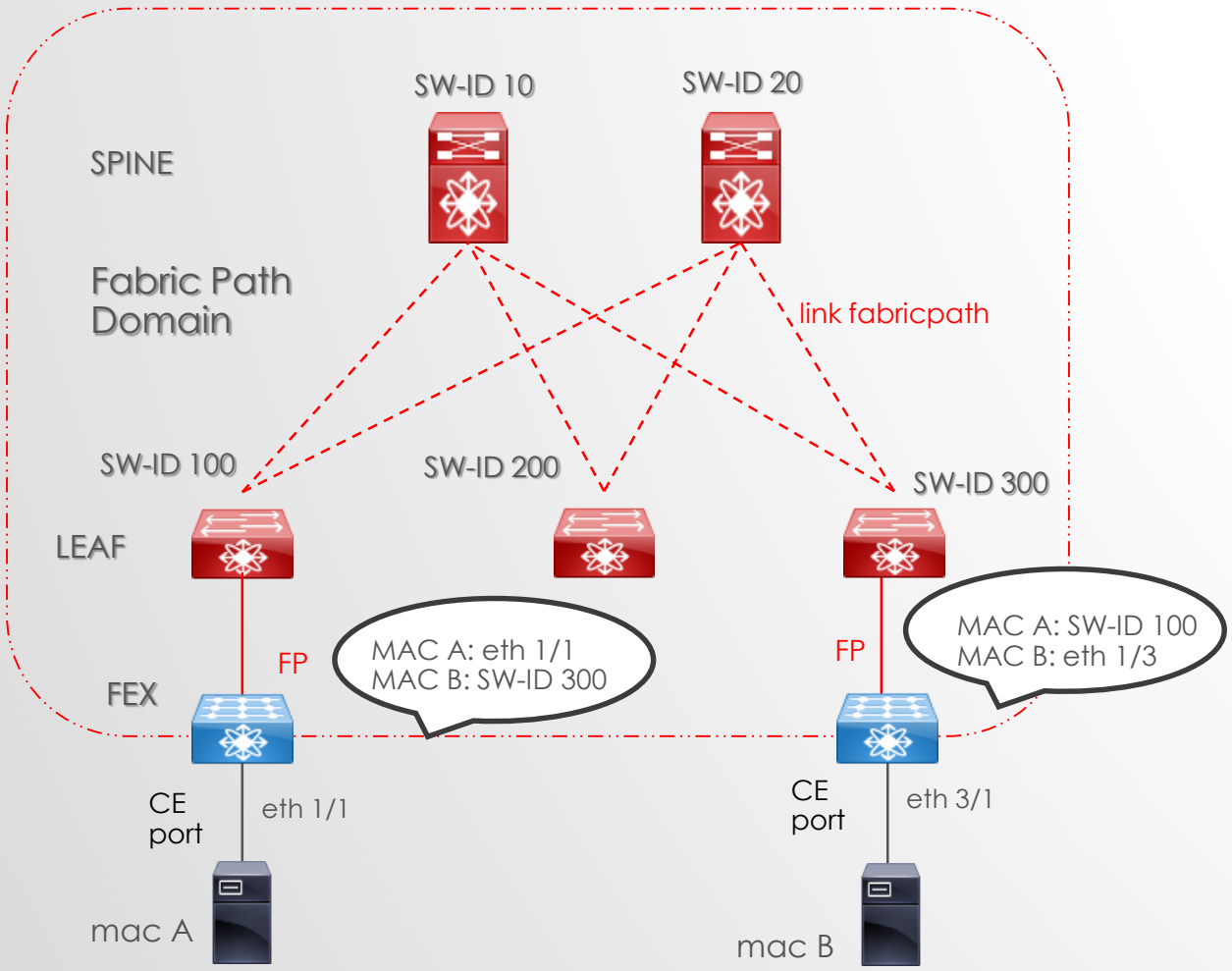
ARCHITETTURE CLOS DATACENTERS SPINE AND LEAF

Massimiliano Sbaraglia

VANTAGGI DI UNA ARCHITETTURA SPINE LEAF

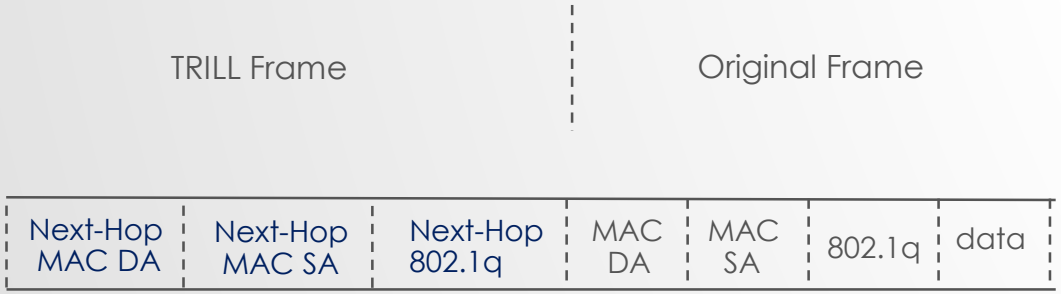
- Architettura a due livelli a costruire una Fabric Switch (unico dominio)
- Alta scalabilità (possibilità di inserimento nuovi elementi) ed una grande capacità in numero di porte
- Riduzione OpEx (es: riduzione numero apparati rispetto ad una tradizionale rete a tre livelli)
- Riduzione CapEx (es: risparmio 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

CISCO FABRICPATH



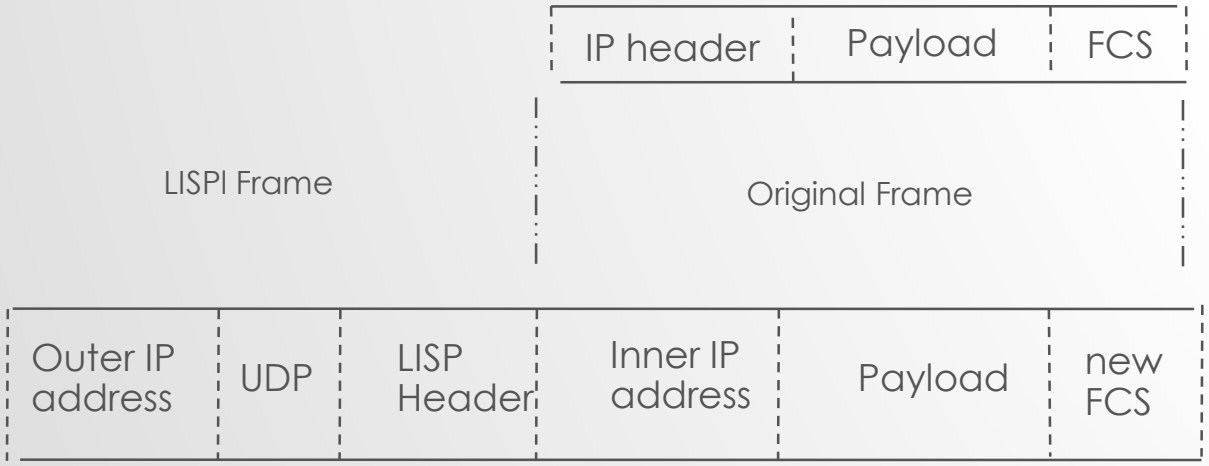
- FabricPath è una tecnologia Cisco con Nexus devices a livello di accesso, distribuito all'interno di un solo datacenters
- Le frame FP è usata per incapsulare standard frame ethernet per attraversare un dominio fabricpath, basato su un nuovo header chiamato Switch-ID
- ISIS routing protocol è utilizzato 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
- Utilizzo della tecnologia vPC enhanced tra peers FP spine and leaf (IEEE802.3ad)
- FP utilizza multidestination 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

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

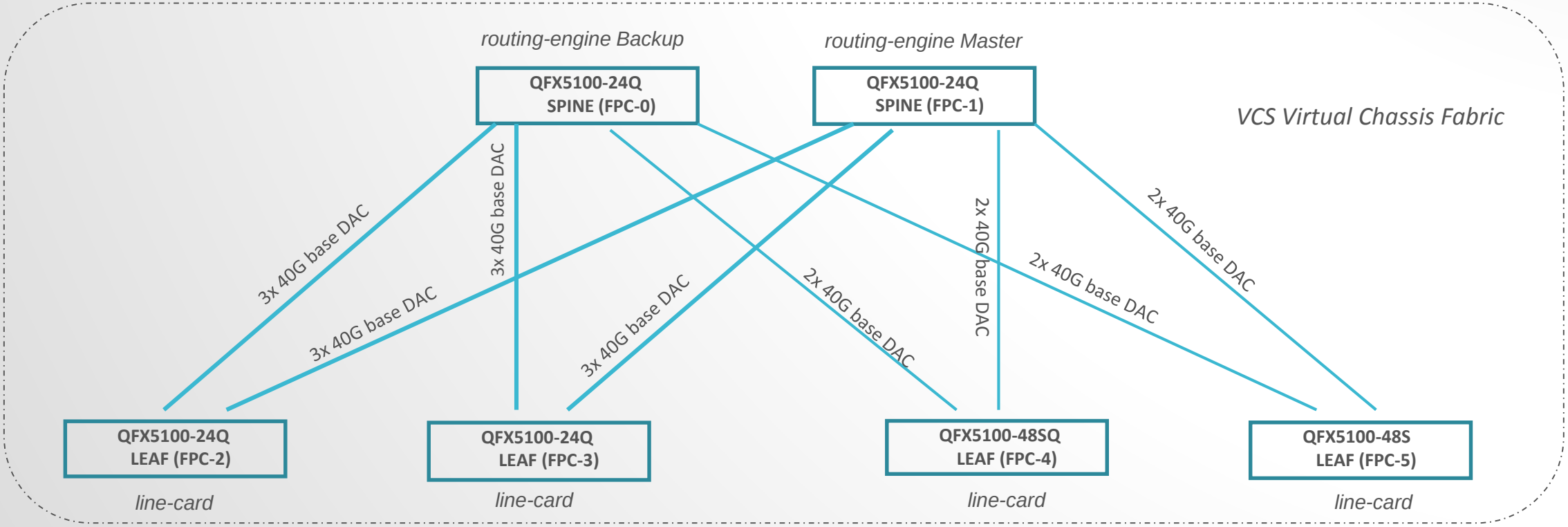


- LISP è progettato per ambienti datacenter dove è previsto un moving di un end-point ed i suoi parametri di rete (addressing) non cambiano ma semplicemente la sua locazione
- RLOC (Routing Locators): descrive la topologia e locazione di un end-point e quindi è usato questo parametro per trasmettere 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.

ARCHITETTURA SPINE-LEAF WITH QFX5100 JUNIPER

- 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 linecard.

- 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.



ARCHITETTURA SPINE-LEAF WITH QFX5100 JUNIPER

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 in tutte le porte VFC), 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.

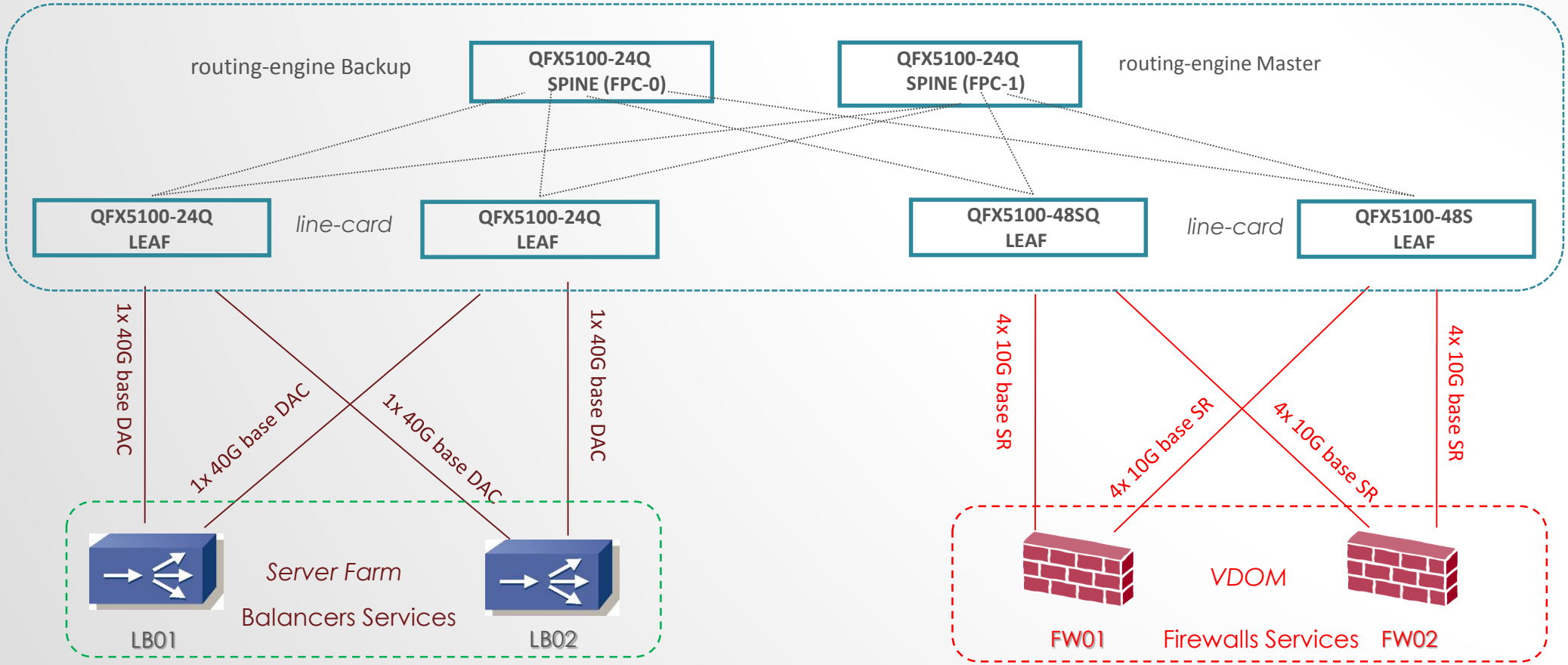
ARCHITETTURA SPINE-LEAF WITH QFX5100 JUNIPER

Sono possibili due configurazioni VCF:

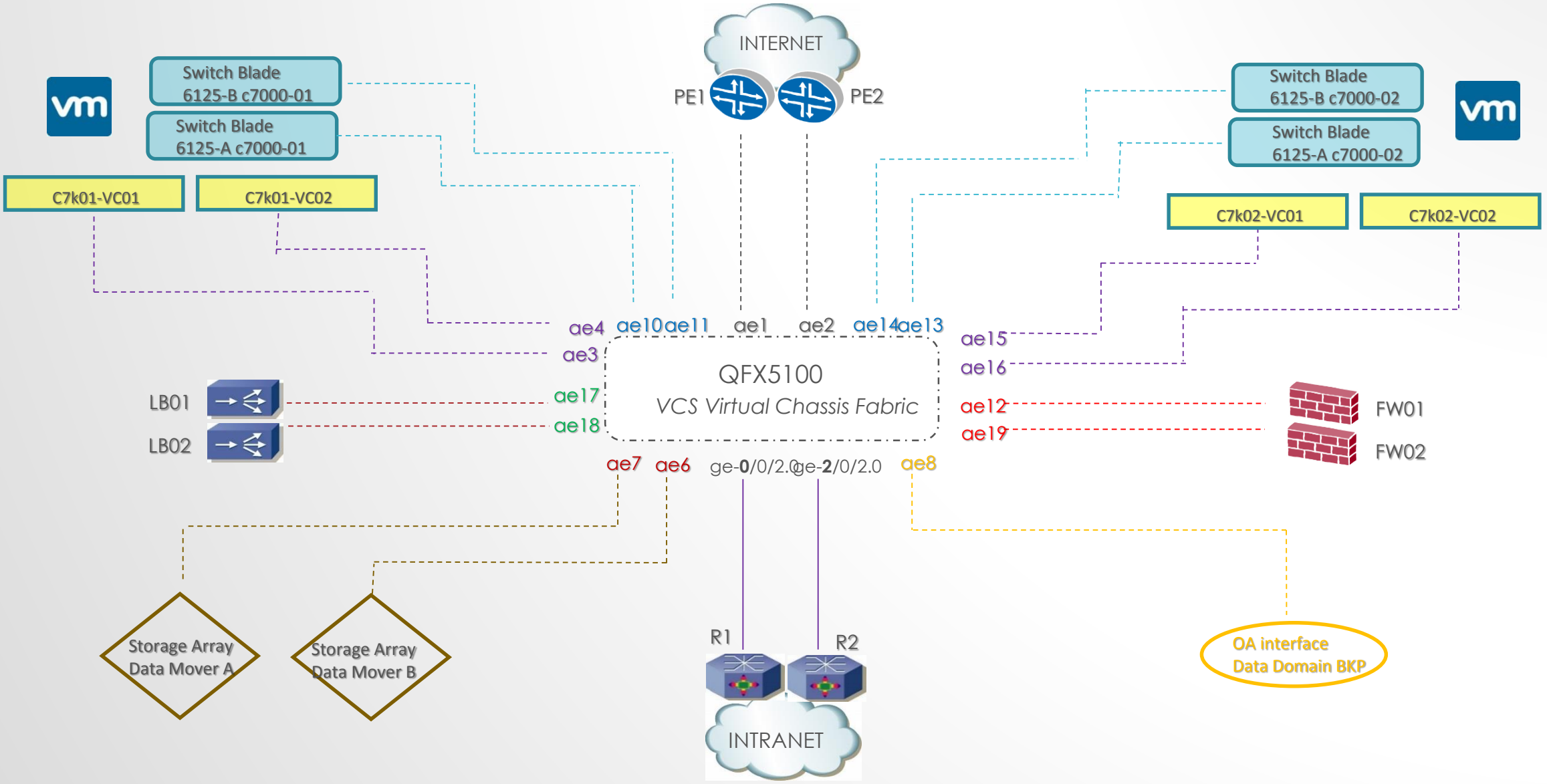
- **Preprovisioned:** con il controllo di ciascun nodo assegnando un member-ID ed il ruolo a lui assegnato;
- **Non-provisioned:** è il nodo master che assegna un member-ID a ciascun nodo; il ruolo è determinato dal valore di priority masterschip ed altri fattori che concorrono alla elezione del master;
- **Master Routing Engine:** il nodo master RE controlla tutta la Fabric VCF
- **Backup Rouing 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.

ARCHITETTURA SPINE-LEAF WITH QFX5100 JUNIPER

Il collegamento di sistemi hardware e software come bilanciatori e firewalls per ambienti datacenter debbono essere collegati su QFX5100 aventi ruolo di line-card:



VCF IN AMBIENTE COMPUTING NFV



CISCO ACI APPLICATION CENTRIC INFRASTRUCTURE

Massimiliano Sbaraglia

ACI CONCEPTS

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;

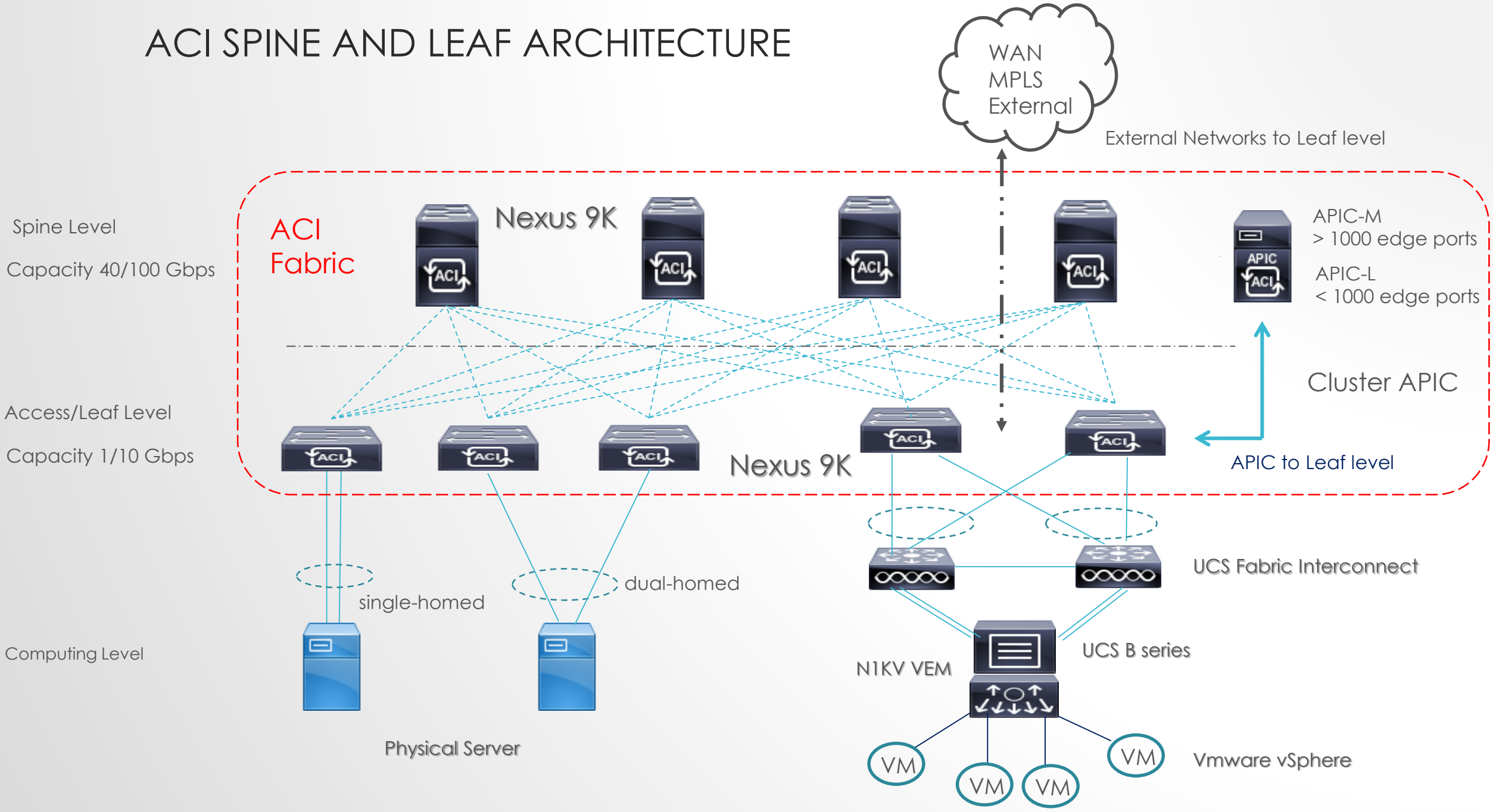
Cisco ACI è disegnato per scalare in modo trasparente nei confronti di cambiamenti di connettività, bandwidth, tenants e policies; la sua architettura è di tipo spine-leaf che si presta efficientemente a introdurre e/o cambiare requisiti di rete;

Cisco ACI include servizi layer 4 to layer 7, APIs (Application Programming Interface), virtual networking, computing, storage resources, wan routers, orchestration services.

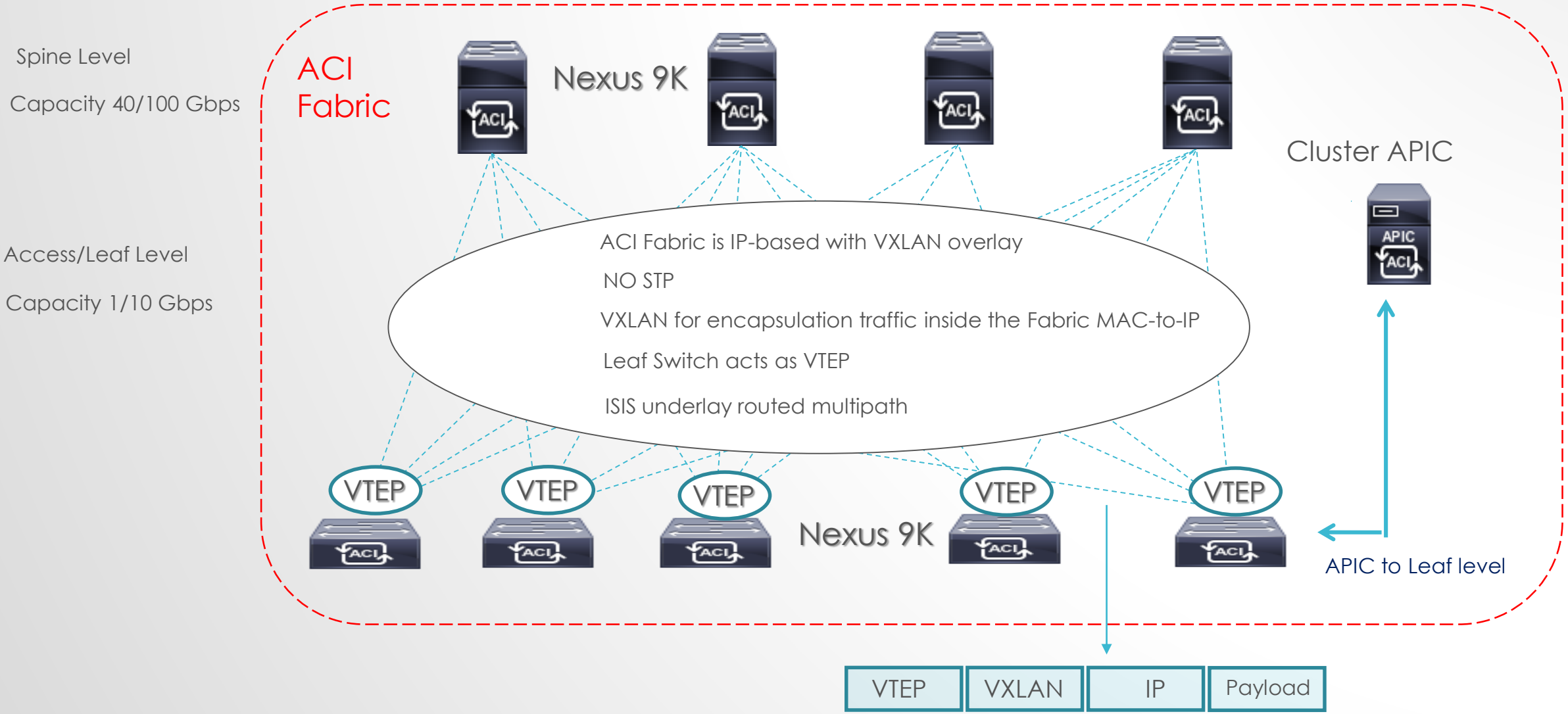
Cisco ACI consiste in:

- 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

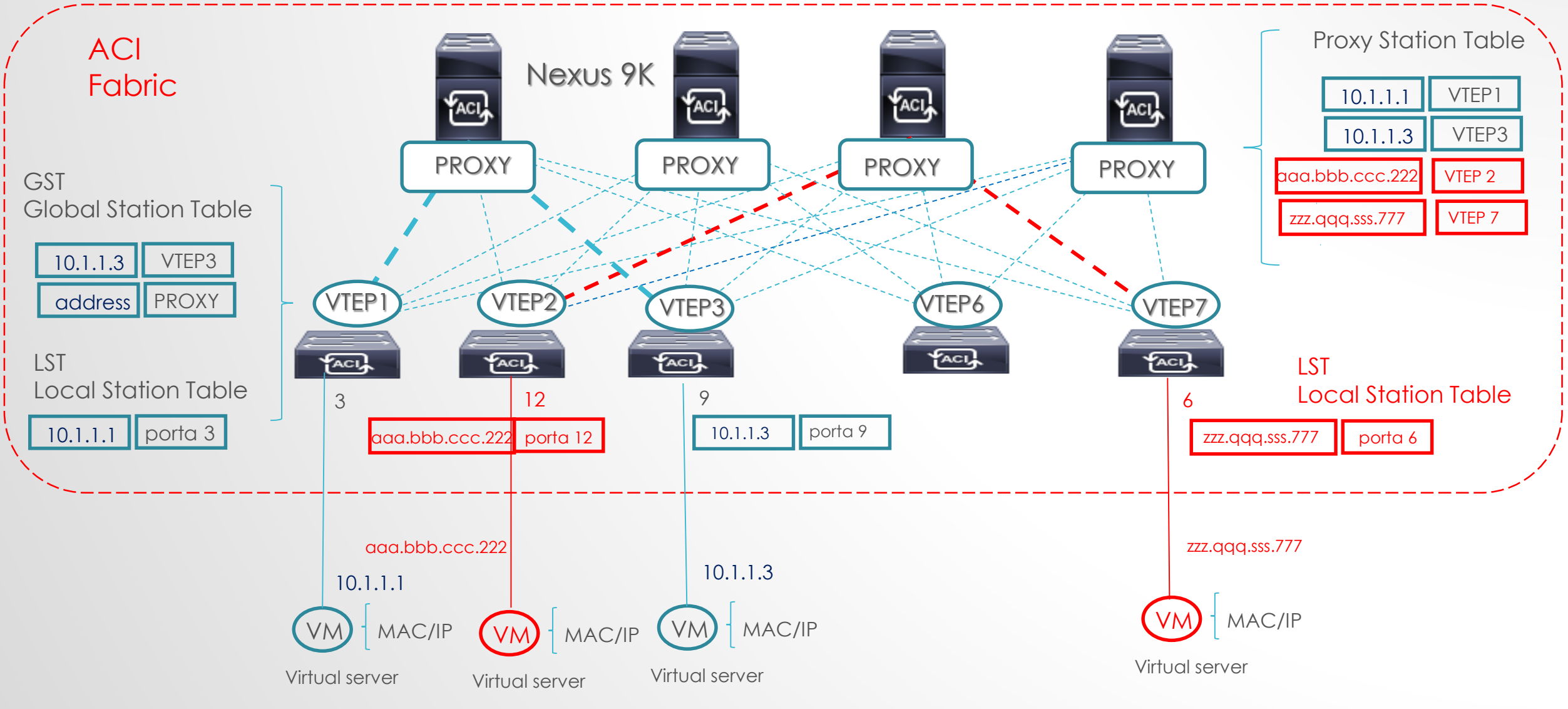
ACI SPINE AND LEAF ARCHITECTURE



ACI FABRIC ARCHITECTURE



ACI CONTROL-PLANE WITH MAPPING DATABASE



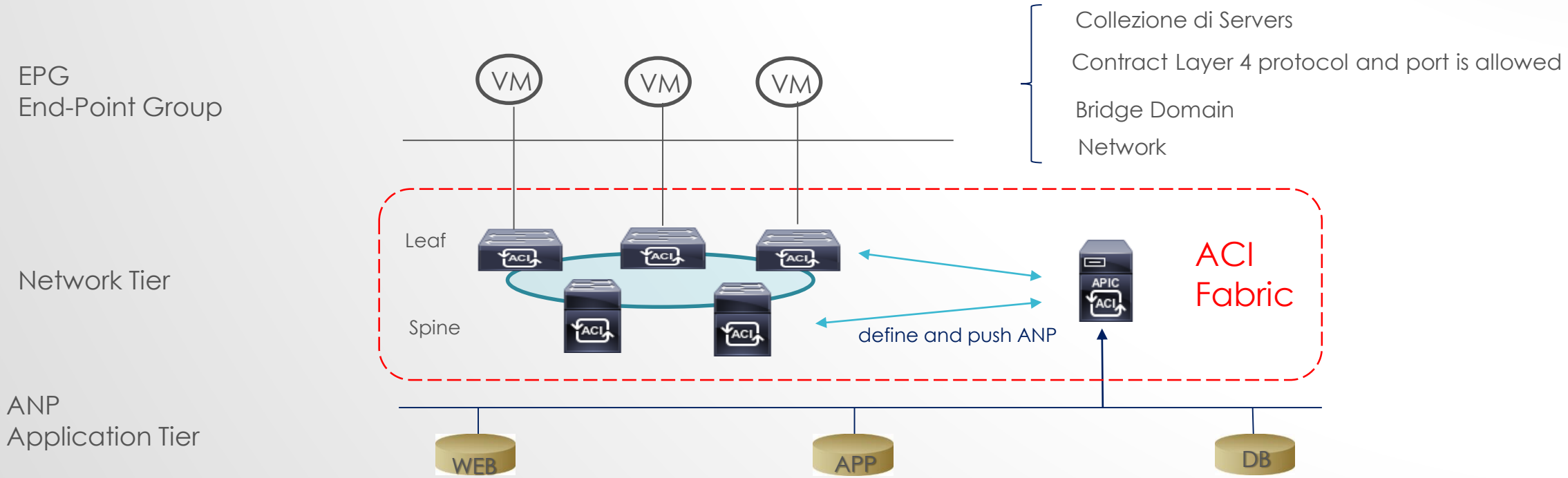
ACI POLICY-BASED APPROACH

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 FABRIC ACCESS POLICY

Switch Profile

Interface Policy and Profile

Attachable Access Entity Profile (AAEP)

Physical Domain

Vlan Pool

VM MAC/IP

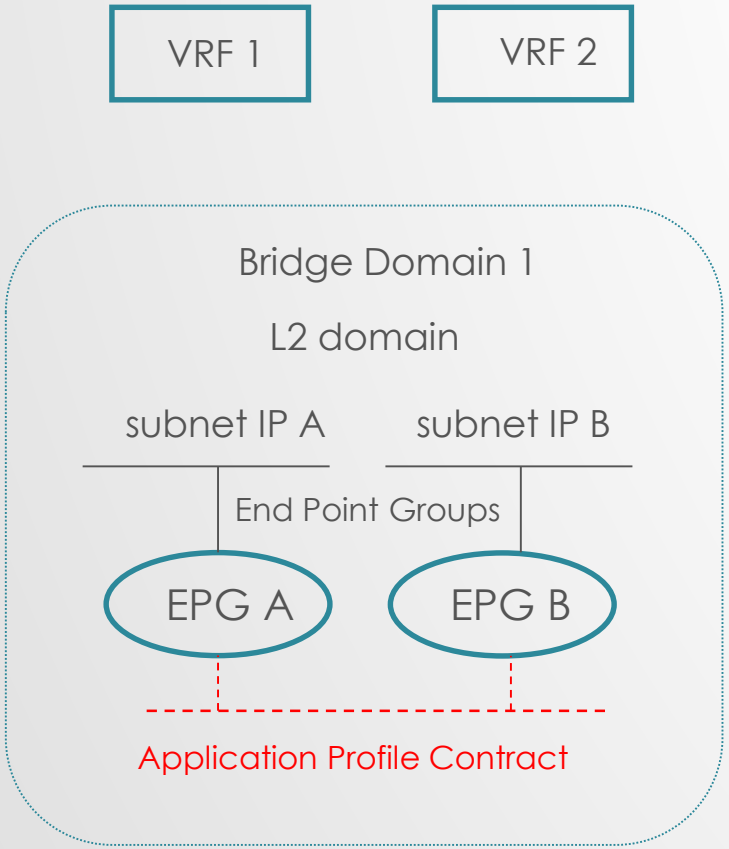
Virtual server

VM MAC/IP

Virtual server

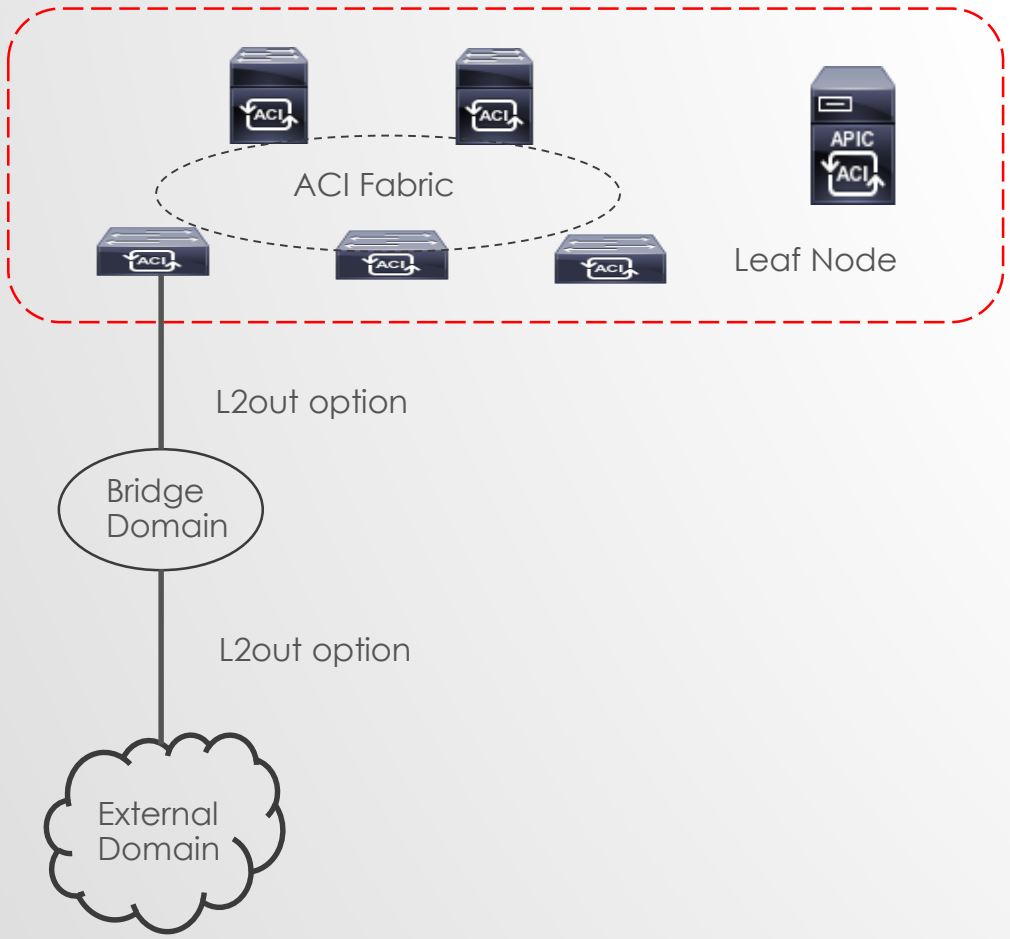
- **vlan pool:** definisce un singolo segmento di rete (vlan) oppure un pool di vlans
- **Physical Domain:** definisce un dominio (scopo) dove è creato il vlans pool
- **AAEP (Attachabe Access Entity Profile):** definisce un modo di raggruppare multipli domini applicabili ad un profilo su base interfaccia
- **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

ACI FABRIC LAYER 2 STEPS DI CONFIGURAZIONE



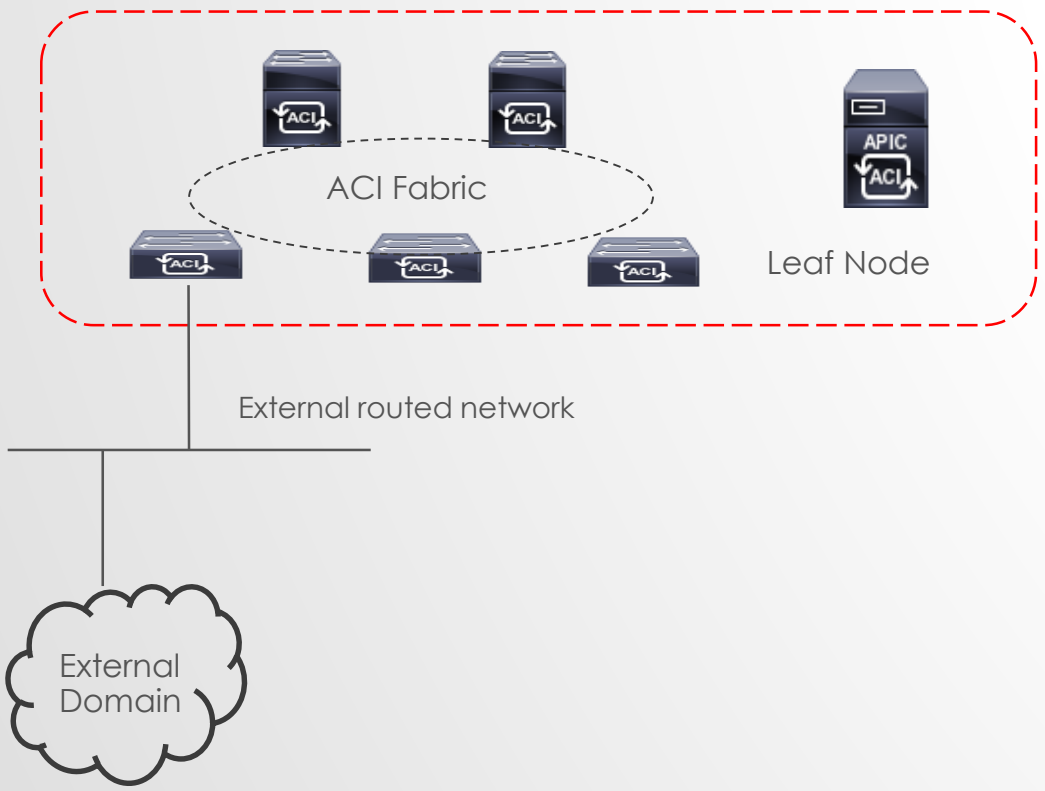
- VRF instances
- BD (Bridge Domain) associato alla VRF instance (senza abilitare nessun layer 3 IP SVIs subnet)
- Configurazione del Bridge Domain per ottimizzare la funzionalità di switching (hardware-proxy-mode) usando il mapping database oppure il tradizionale flood-and-learn
- EPG (End Point Group) 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 FABRIC LAYER 2 EXTENDING TO EXTERNAL DOMAIN



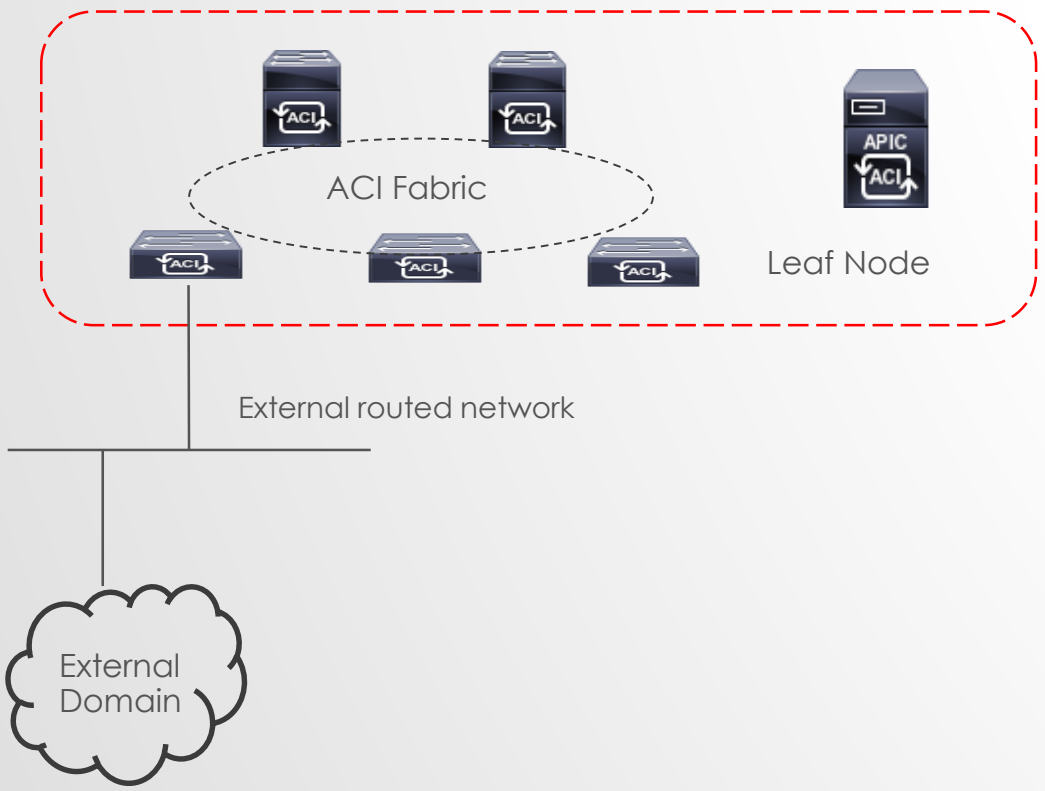
- Enable fllooding of layer 2 unknown unicast
- Enable ARP flooding
- Disable unicast routing (può essere abilitato successivamente ad una fase di migrazione ad esempio se gli end-point usano come IP gateway il sistema ACI Fabric
- L2Out option provvede ad una L2 extension da ACI Fabric ad un External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

ACI FABRIC EXTERNAL NETWORK PARAMETERS



- **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 FABRIC EXTERNAL NETWORK L3-OUT OPTION



- 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)