ARCHITETTURE DATACENTERS in tecnologia cisco

Massimiliano Sbaraglia
Evoluzione architetture
Data Centers Concept

- High Availability
- Resiliency and Fault-Tolerant
- Virtualization
- QoS

- Security Policy
- Failover
- Virtual Firewall (vdom)
- Zone-based

- Load Balancing Servers
- Servers Farm

- L2 extension
- DCI
- Disaster Recovery
- Business Continuity

- Vmware Vsphere
- Cloud and SDN
- Vmotion and Fault-Tolerant

- FCoE, FC, FCIP, SCSI, iSCSI
- CIFS NFS
- DAS, NAS, SAN (vSAN)
- Zoning and Fabric

Data Center

- Network
- Firewall
- Balancer
- Storage

- Geocluster
- Computing
Architettura di base tre livelli non-datacenters

- **Core Level**
  - MPLS private
  - Internet public

- **Distribution Level**
  - EBGP
  - static route
  - dynamic route
  - Switch L2/L3 (Catalyst 6500 - 7600)
  - trunk
  - Spanning Tree Domain

- **Access Level**
  - Switch L2 (family 2K, 3K, 4K)
  - 80% data traffic transitante attraverso il distribution level
  - client
  - server

Access Level

Distribution Level

Core Level
Architettura di base fisica di un datacenter

Core Level

Aggregation Level

Access Level

MPLS private

Internet public

EBGP

FW1

link FW failover

L3 FW gateway

FW2

L3 FW gateway

L2 domain

L3 BAL gateway

STP Domain

BAL1

link BAL failover

server

client
Architettura di base logica di una datacenters

**MPLS**

**INTERNET**

- **Public ISP 1**
- **Public ISP 2**

**SLB**

**FW**

**NAT**

- **static route RFC 1918**
- **DMZ inside subnet firewalling**
- **DMZ outside subnet with single NAT pool to Internet**
- **inside subnet balanced**

**INSIDE**

- **HSRP**
- **EBGP**

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

HSRP gestisce HA per traffico outbound (to external)

**INSIDE**

- **VIP gateway**

NAT traduce il piano di indirizzamento privato con un blocco (pool) pubblico annunciato ad entrambi gli ISP Internet
Architettura di base logica di una datacenters con tecniche di virtualizzazione

Ogni livello AC, FL, AL sono separati da security device e livelli di sicurezza

EC + FL + AL = sono livelli virtualizzati all’interno di contesti VDC su Nexus Cisco

EC layer: task è configurare peering eBGP tra il VRF AC (VDC-Egress) verso Edge Router

FL layer: posizione dove sono i servers in VRF FL, visibile da Internet

AL layer: posizione dove sono i servers in VRF AL, invisibile da Internet

pair VSYS instance on Firewall NS-1 - NS-2 between EC and FL layer

pair VSYS instance on Firewall NS-3 - NS-4 between FL and AL layer
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
Architettura con Nexus
vPC, VDC, FEX
Architettura Data Centers con Nexus

- VSS-1
- VSS-2
- VSL
- N7K-1
- N7K-1
- N5K-1
- N5K-2
- N5K-3
- N5K-4
- VSS-1
- VSS-2
- VSL
- N7K-1
- N7K-1
- N5K-1
- N5K-2
- N5K-3
- N5K-4
- Server
- Server
- vPC keepalive
- vPC peer-link
- vPC domain 10
- vPC domain 20
- vPC domain 30
- active link
- active active mode
## vPC Nexus components

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

- eliminates SPT blocked port
- uses all available links and relative bandwidth
- dual-homed servers in active-active mode
- fast-convergence in case of fault link or switch
- split-horizon loop via port-channeling (traffic entering in a port cannot exit the same channel)
- a vPC domain is composed of two peers, each working with its own control-plane
- vPC means a connection in port-channel between two vPC peers and one device in downstream
- vPC domain is built through the configuration of a peer-keepalive (to monitor the condition of the two peers) and a peer-link (for state synchronization of the two peers)
- HA, link-level resiliency

![vPC diagram](image)
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 appartengono al VDC di default; attraverso questo VDC è possibile creare nuovi 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 5K, 6K o 7K (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, the FEX N2K is online for both parent switches N5K.
- In this topology, an event failure of a parent switch does not affect the FEX because both parent switches peers vPC manage their connection simultaneously.
- Requirement ensures that the FEX N2K configuration is the same (including host interfaces) in both switches.

**Configurazione:**

```config
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
```

![Diagram: Pod NeXUS FEX 1xN2K with active-active dual-homed](attachment:diagram.png)
Questo 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'interfaccia 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
```

![Diagram](image-url)
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
Architettura Unified Computing System (UCS cisco)
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
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 componenti del server (adapters, BIOS, etc..) sono configurati per accordarsi su specifici parametri (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.

UCS Manager

- Virtual interface
- Uplink selection
- VLAN tag
- VSAN domain
- QoS setting

Fabric Interconnect

UCS Domain
unified connection

UCS Servers

- Server ID
- MAC address
- WWN address
- Firmware package
- BIOS setting
- Boot Order
- Management IP
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 Adapter) e switch Ethernet per il trasporto e l’instradamento di FCoE packets
- Un server connesso ad una rete FCoE rappresenta un iSCSI Initiator (così 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
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 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
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 (adapter 1) per il controllo della comunicazione tra altri VSM e la configurazione di una VEM (virtual Ethernet module)
- Una scheda Ethernet (adapter 2) per il sistema di management (mgmt0)
- Una scheda Ethernet (adapter 3) per la trasmissione di packets inviati da una 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)
Architettura Spine & Leaf
L2 Fabric
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

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

- **SAN Storage**
  - MDS-1
  - FEX-1

- **UCS Servers**
  - MDS-2
  - FEX-2

- **Fabric-Interconnect**
  - LEAF
  - L2 domain

- **L3 domain**
  - Firewall
  - Balancer

- **Egress Routers**
  - SPINE

- **FABRICPATH Cloud SPINE LEAF**
  - STP Free
  - ISIS protocol Path Forwarding based Switch-ID
  - Equal Cost Multi Path (ECMP)
  - vPC+ emulated switch-id feature

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

Fabric-path switch-id 1  
VPC +  
Fabric-path switch-id 2

Nexus 1 Eth 1/46 Nexus 2 Eth 1/47
Eth 1/2

Full-Mesh link-mode fabricpath

Nexus 3  
VPC +  
Nexus 4
Fabric-path switch-id 3  
Fabric-path switch-id 4

Nexus 5 VPC + Nexus 6
Fabric-path switch-id 5  
Fabric-path switch-id 6
NEXUS-1

feature-set fabricpath
feature vpc

interface 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 1 mode active

interface ethernet 1/47
  description B2B
  switchport mode fabricpath
  channel-group 1 mode active

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
description 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.
LISP (locator / identifier separation protocol)

ALT = Alternate Logical Topology
MR = Map Resolver
MS = Map Server
Architettura DCI
Data Center Interconnection
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

CORELAYER

DCILAYER

AGGREGATION LAYER

ACCESS LAYER

COMPUTING

SAN Storage

UCS / Blade Server / VM

Servers

Data Center DR

N7K-1

N7K-2

vPC

vPC

primary

backup

HSRP

INTERNET

L2 extension

Firewalling Level

Applications Security

Balanced Server

vPC domain N5K

vPC domain N7K

vPC
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’infrastrututra 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 adiacenze 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 authoritative 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

<table>
<thead>
<tr>
<th>vlan</th>
<th>MAC</th>
<th>IF</th>
</tr>
</thead>
<tbody>
<tr>
<td>X</td>
<td>xxxx.xxxx</td>
<td>eth 0/0</td>
</tr>
<tr>
<td>X</td>
<td>zzzz.zzzz</td>
<td>IP C</td>
</tr>
</tbody>
</table>

OTV Site VLAN 1000
OTV Site-ID 0x1
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:**

```plaintext
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
```
```plaintext
interface ethernet 4/16
mtu 1600
channel-group 300 mode active
no shut
```
```plaintext
interface ethernet 5/18
mtu 1600
channel-group 300 mode active
no shut
```

**OTV overlay interface:**

```plaintext
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
```
```plaintext
```
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 (Rendezvous 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).
DCI layer 2 dark-fiber point-to-point

BPDU Filter assicura STP Isolation tra differenti Data Center domain

CORE LAYER

AGGREGATION LAYER

max 100 Km

DWDM Network Line

ACCESS LAYER

Data Center
Primario

Data Center
Secondario
BPDU Filter assicura STP Isolation tra differenti Data Center domain
DCI layer 2 pseudowire Ethernet P2P

BPDU Filter assicura STP Isolation tra differenti Data Center domain
Il flooding del dominio STP è qualcosa di indesiderato via DCI

Soluzione: introduzione del MEC into VPLS
DCI layer 2 with Tunnel GRE

CORELAYER

AGGREGATION LAYER

ACCESS LAYER

Any Transport over MPLS over GRE

Data Center Primario

Data Center Secondario
**VXLAN protocol**

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

Data Centers:

- VMware and Vshere virtualization
- 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 meccanismo 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 encapsulation

<table>
<thead>
<tr>
<th>Field</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Outer MAC Header</td>
<td>18 bytes</td>
</tr>
<tr>
<td>Outer IP Header</td>
<td>20 bytes</td>
</tr>
<tr>
<td>Outer UDP Header</td>
<td>8 bytes</td>
</tr>
<tr>
<td>VXLAN Header</td>
<td>8 bytes</td>
</tr>
<tr>
<td>Original L2 frames</td>
<td></td>
</tr>
<tr>
<td>IP Head Data</td>
<td></td>
</tr>
<tr>
<td>IP Protocol</td>
<td></td>
</tr>
<tr>
<td>IP Head Checksum</td>
<td></td>
</tr>
<tr>
<td>Outer Source IP</td>
<td></td>
</tr>
<tr>
<td>Outer Destination IP</td>
<td></td>
</tr>
<tr>
<td>Outer MAC Header</td>
<td></td>
</tr>
<tr>
<td>Outer Source MAC</td>
<td></td>
</tr>
<tr>
<td>Vlan Type</td>
<td></td>
</tr>
<tr>
<td>Vlan TAG</td>
<td></td>
</tr>
<tr>
<td>Ether Type</td>
<td></td>
</tr>
<tr>
<td>Inner Dest Mac</td>
<td></td>
</tr>
<tr>
<td>Inner Source Mac</td>
<td></td>
</tr>
<tr>
<td>TPID</td>
<td></td>
</tr>
<tr>
<td>TAG</td>
<td></td>
</tr>
<tr>
<td>Payload</td>
<td></td>
</tr>
</tbody>
</table>
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 encapsulation 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 throughput 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 (Directional PIM) provvedono servizi multicast per VXLAN

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

VM = Virtual Machine
VNI = VXLAN Network Identifier
VTEP = VXLAN Tunnel Endpoint
BD = Bridge Domain
Architettura CLOS
L3 Fabric
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

Spine Level
Capacity 40/100 Gbps

Access/Leaf Level
Capacity 1/10 Gbps

Computing Level

ACI Fabric
Nexus 9K

Cluster APIC

WAN MPLS External
External Networks to Leaf level

APIC-M
> 1000 edge ports

APIC-L
< 1000 edge ports

Single-homed
Dual-homed

Physical Server

UCS Fabric Interconnect

UCS B series

VMware vSphere

N1KV VEM

VM

VM

VM
ACI (Application Centric Infrastructure) Cisco Architecture

Spine Level
Capacity 40/100 Gbps

ACI Fabric

Nexus 9K

Cluster APIC

ACI Fabric is IP-based with VXLAN overlay
- NO STP
- VXLAN for encapsulation traffic inside the Fabric MAC-to-IP
- Leaf Switch acts as VTEP
- ISIS underlay routed multipath

VTEP
VTEP
VTEP
VTEP
VTEP
VTEP

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

ACI Fabric

Nexus 9K

Proxied by VTEP1, VTEP2, VTEP3, VTEP6, VTEP7

VM1 (MAC/IP: aaa.bbb.ccc.222, Porta 3)
VM2 (MAC/IP: 10.1.1.1, Porta 12)
VM3 (MAC/IP: zzz.qqq.sss.777, Porta 9)
VM4 (MAC/IP: 10.1.1.3, Porta 6)

Global Station Table (GST)
- Entries: 10.1.1.1, VTEP1
- Entries: 10.1.1.3, VTEP3
- Entries: aaa.bbb.ccc.222, VTEP2
- Entries: zzz.qqq.sss.777, VTEP7

Local Station Table (LST)
- Entries: 10.1.1.1, VTEP1
- Entries: 10.1.1.3, VTEP3
- Entries: aaa.bbb.ccc.222, VTEP2
- Entries: zzz.qqq.sss.777, VTEP7

Proxy Station Table (PST)
- Entries: 10.1.1.1, VTEP1
- Entries: 10.1.1.3, VTEP3
- Entries: aaa.bbb.ccc.222, VTEP2
- Entries: zzz.qqq.sss.777, VTEP7
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 (Attachable 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
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 flooding 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 an 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.

  ```plaintext
  +----------------+----------------+----------------+----------------+----------------+----------------+----------------+----------------+----------------+----------------+----------------+----------------+----------------+
  | per-VPN RD     | MAC-Length      | MAC Host       | IP-Length       | IP Host         | VNI per BD      | VNI per VRF     | RT per EVI       | RT per VRF       | VXLAN           | MAC Router      | MAC Mob Seq Numb |
  +----------------+----------------+----------------+----------------+----------------+----------------+----------------+----------------+----------------+----------------+----------------+----------------+----------------+
  | RD             | MAC address     | IP address     | Labels          | Route Target    | Tunnel-Type     | USP            |                |                |                |                |                |                |
  +----------------+----------------+----------------+----------------+----------------+----------------+----------------+----------------+----------------+----------------+----------------+----------------+----------------+
  ```

- **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.

  ```plaintext
  +----------------+----------------+----------------+----------------+----------------+----------------+----------------+----------------+----------------+----------------+----------------+----------------+----------------+----------------+
  | VRF RD         | ETH Segment     | ETH Tag         | IP-Length       | IP Prefix       | IP Gateway      | L3 VNI          | RT per VRF      | VXLAN           | MAC Router      |                |                |                |
  +----------------+----------------+----------------+----------------+----------------+----------------+----------------+----------------+----------------+----------------+----------------+----------------+----------------+----------------+
  ```
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 Protocol 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

<table>
<thead>
<tr>
<th>NH</th>
<th>HOST SOUR</th>
<th>HOST DEST</th>
<th>VLAN</th>
<th>VXLAN</th>
<th>TYPE</th>
</tr>
</thead>
<tbody>
<tr>
<td>VTEP-1</td>
<td>S1 MAC/IP</td>
<td>S2 MAC/IP</td>
<td>10</td>
<td>1010</td>
<td>2</td>
</tr>
<tr>
<td>VTEP-2</td>
<td>S2 MAC-IP</td>
<td>S1 MAC/IP</td>
<td>10</td>
<td>1010</td>
<td>2</td>
</tr>
<tr>
<td>VTEP-N</td>
<td>S1 MAC/IP</td>
<td>S12 MAC/IP</td>
<td>10</td>
<td>1010</td>
<td>L2 extension</td>
</tr>
</tbody>
</table>

VTEP-1 \(\text{BD VXLAN 1010}\)  
VTEP-2  
VTEP-N  

L2 broadcast ARP to S12 MAC/IP

<table>
<thead>
<tr>
<th>VTEP</th>
<th>NEXT HOP</th>
<th>HOST</th>
<th>TYPE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>S1 MAC/IP</td>
<td>LOCAL</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>S2 MAC/IP</td>
<td>VTEP-2</td>
<td>2</td>
</tr>
<tr>
<td>1</td>
<td>S12 MAC/IP</td>
<td>VTEP-N</td>
<td>2</td>
</tr>
</tbody>
</table>

VTEP-12 \(\text{L2 DCI L2 extension VNI = 1010} \)  
VTEP-11  
VTEP-12  
VTEP-NN

<table>
<thead>
<tr>
<th>VTEP</th>
<th>NEXT HOP</th>
<th>HOST</th>
<th>TYPE</th>
</tr>
</thead>
<tbody>
<tr>
<td>VTEP-NN</td>
<td>S12 MAC/IP</td>
<td>LOCAL</td>
<td></td>
</tr>
<tr>
<td>VTEP-NN</td>
<td>S11 MAC/IP</td>
<td>VTEP-12</td>
<td>2</td>
</tr>
<tr>
<td>VTEP-NN</td>
<td>S1 MAC/IP</td>
<td>VTEP-11</td>
<td>2</td>
</tr>
</tbody>
</table>
Inter-Subnet design communication via EVPN Fabric

## Network Design

### Core 1
- **VRF X**
- **L3 VNI XX**

### Core 2
- **VRF X**
- **L3 VNI XX**

### L3 DCI

### EBGP Advertisement
- **Gateway**

### VTEP-1
- **Next Hop**
- **10.10.10.0/24**

### VTEP-2
- **Next Hop**
- **20.20.20.0/24**

### VTEP-N
- **Next Hop**
- **Request ARP**

### Table:

<table>
<thead>
<tr>
<th>VTEP</th>
<th>NEXT HOP</th>
<th>HOST</th>
<th>TYPE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>S1 MAC/IP</td>
<td>LOCAL</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>S2 MAC/IP</td>
<td>VTEP-2</td>
<td>2</td>
</tr>
<tr>
<td>1</td>
<td>S12 MAC/IP</td>
<td>Request ARP</td>
<td>5</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>VTEP</th>
<th>NEXT HOP</th>
<th>HOST</th>
<th>TYPE</th>
</tr>
</thead>
<tbody>
<tr>
<td>VTEP-NN</td>
<td>S12 MAC/IP</td>
<td>LOCAL</td>
<td></td>
</tr>
<tr>
<td>VTEP-NN</td>
<td>S11 MAC/IP</td>
<td>VTEP-12</td>
<td>2</td>
</tr>
<tr>
<td>VTEP-NN</td>
<td>S1 MAC/IP</td>
<td>Request ARP</td>
<td>5</td>
</tr>
</tbody>
</table>
EVPN I-BGP Configurations VTEP (VXLAN Tunnel End-Point)

feature bgp
feature nv overlay
feature nv overlay evpn

→ enable VTEP (required on Leaf or Border)
→ enable EVVPN control-plane in BGP

@ only on LEAF

interface nve1
source-interface loopback0
host-reachability protocol bgp

→ enable interface VTEP
→ enable source interface with loopback
→ 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
   route-target import auto
   route-target export auto
!
interface nve1
   source-interface loopback0
   host-reachability protocol bgp
   member vni 50000
      mcast-group 239.239.239.10
      suppress-arp
```

- # RD is default calculated as VNI:BGP Router ID
- # RT is default calculated as BGP AS:VNI

VLAN 50

VNI VXLAN 50000

VTEP-1 VTEP-2 VTEP-3 VTEP-4

RR-1 RR-2

→ # RD is default calculated as VNI:BGP Router ID
→ # RT is default calculated as BGP AS:VNI
EVPN I-BGP Configurations Routing Resource on VXLAN

# Define VLAN for VRF routing instances

```
vlan 50	n v-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

EN

VNI VXLAN 10000 associated VRF-A (vlan 10)

VNI VXLAN 60000

VNI VXLAN 50000

VTEP-1

SVI-vlan 50

SVI-vlan 60

VTEP-2

VTEP-3

VTEP-4

SVI-vlan 50

SVI-vlan 60

TO underlay IGP + iBGP

SVI

VLAN 50

VLAN 60

L2 Gateway

VRF-A

VRF-A

VRF-A

VRF-A

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 ./.

segue ./.
route-map RED-SUBNET permit 10
match 123
!
router bgp 65000
vrf VRF-A
advertise l2vpn evpn
redistribue 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
description 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
```