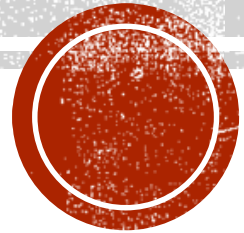
VXLAN OVERVIEW

Massimiliano Sbaraglia



VXLAN ENVIRONMENT

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

- Data Centers
  - VMware and Vshere virtualizzazation
  - 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:

- Permette di collegare VMs server in differenti IP network come se fossero all'interno di una stesso dominio Layer 2

VXLAN IMPLEMENTATION

VXLAN viene supportato da una infrastruttura:

- Multicast
  - IGMP
  - PIM

- IP routing protocols:
  - OSPF
  - ISIS
  - BGP

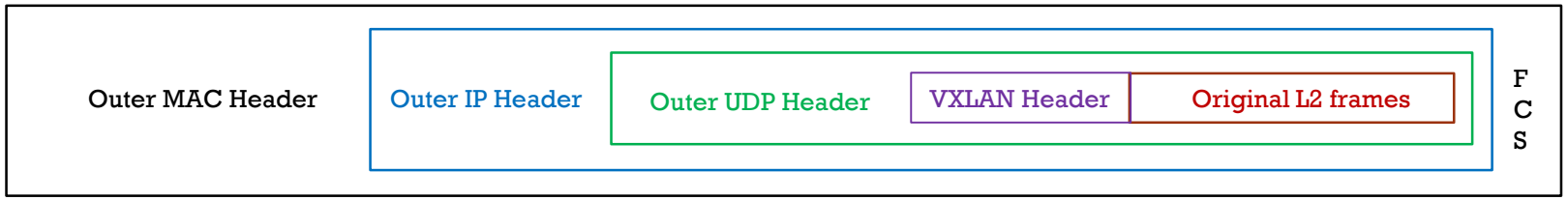
- IP Gateway:
  - VTEP (Vlan Tunnel End Point) provvede ad incapsulare e decapsulare servizi layer 2 to VXLAN.
  - VTEP possono essere:
    - Virtual Bridges Hipervisor
    - VXLAN aware VM application
    - Router/Switch hardware

VXLAN PROTOCOL

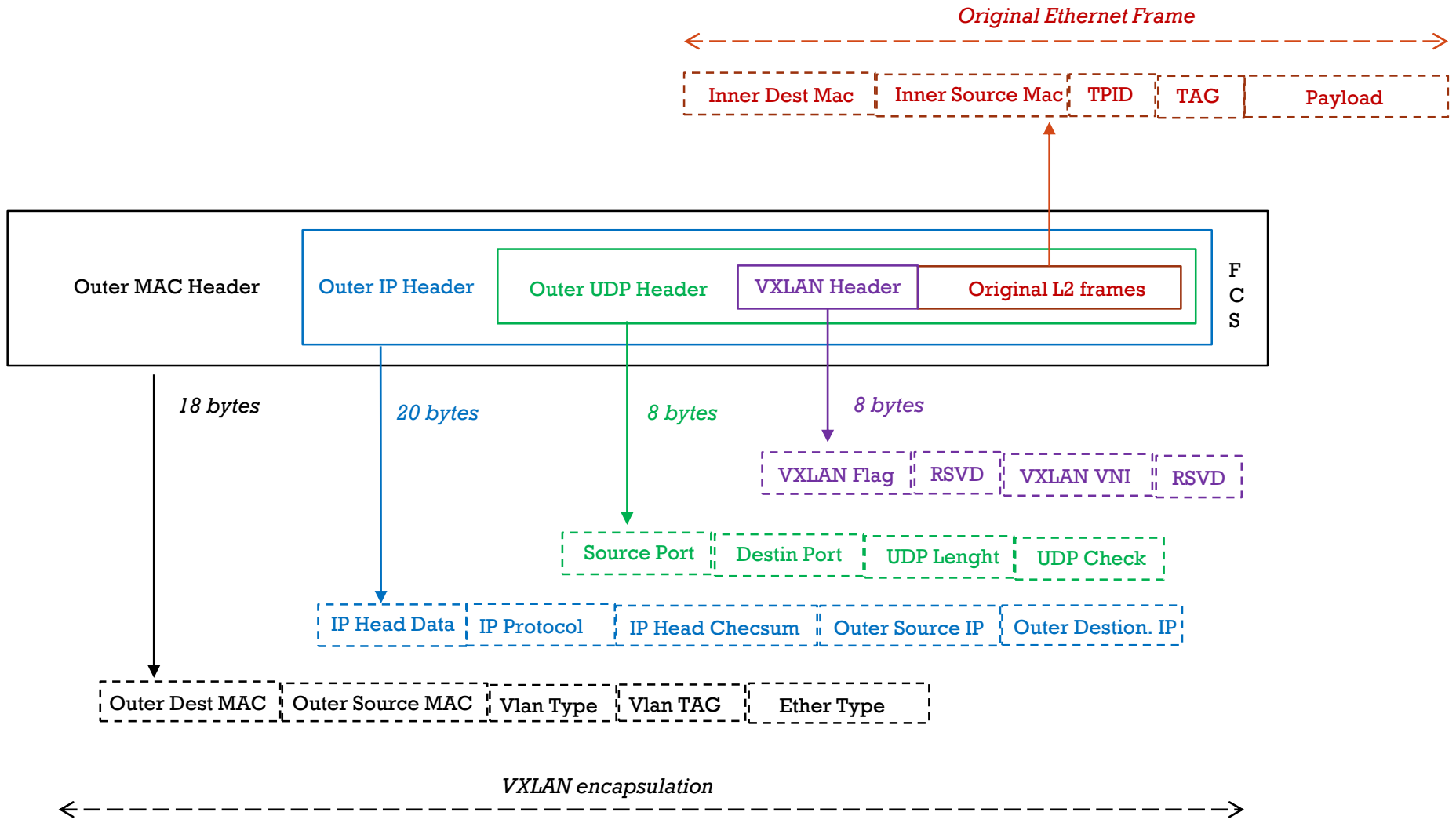
- Ogni VXLAN segment è associato con un unico 24 bit VXLAN Network Identifier differente chiamato VNI.
- Questo 24 bit VNI permette di scalare da il classico 4096 vlans con 802.1q a più di 16 milioni di possibili virtual networks
- Le VMs servers all'interno di un dominio layer 2 utilizzano la stessa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

VXLAN FRAME ENCAPSULATION AND FORWARDING

- 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 FORMAT



VXLAN HEADER FORMAT

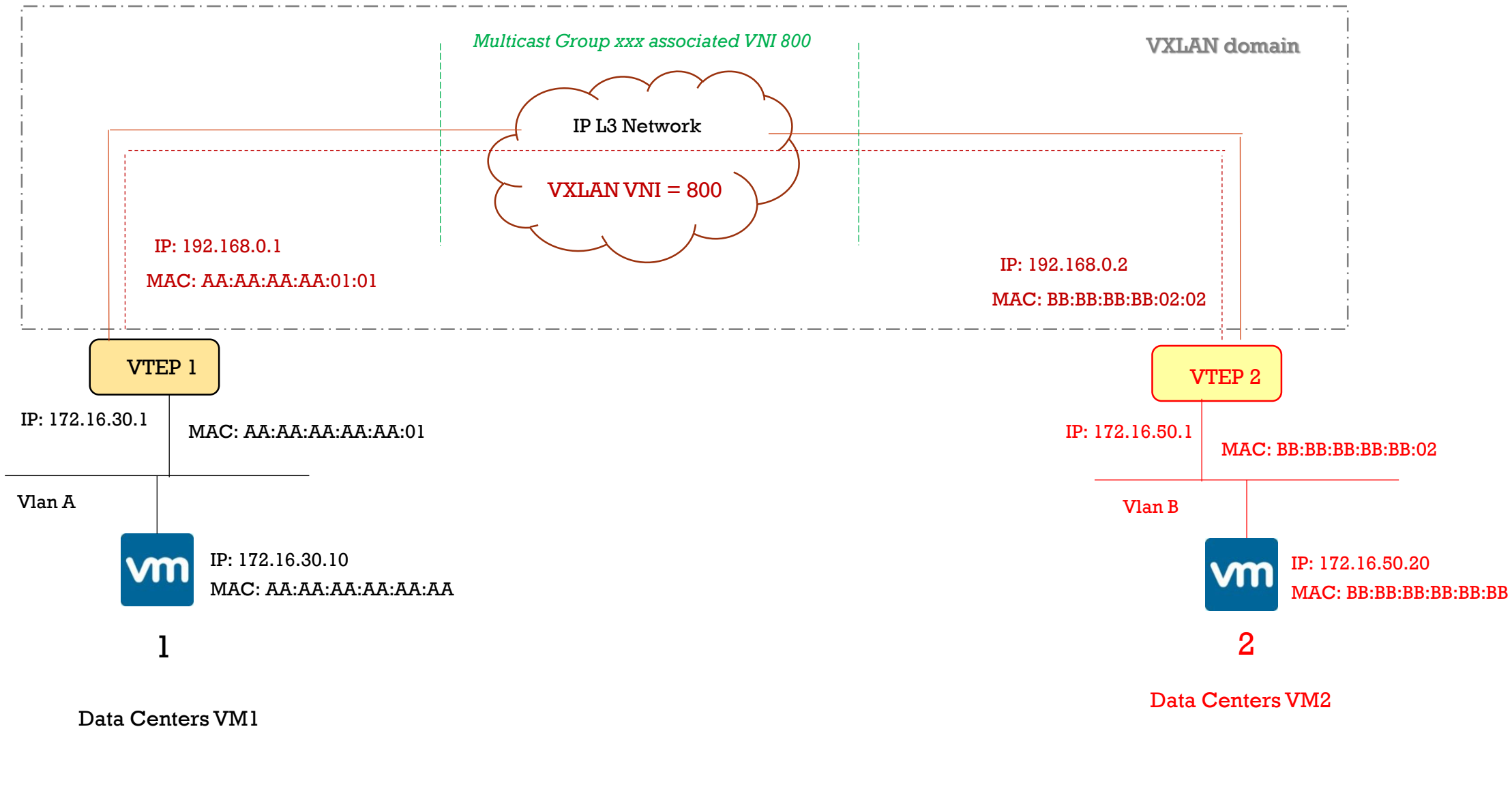
- VXLAN Header:
  - Flag: composto da 8 bits dove il 5° bit (flag) indica un valido valore VNI (i restanti sette bits sono riservato e settati a zero)
  - VNI: valore di 24 bits, provvede a rilasciare un unico identifier per segmento VXLAN; possiamo avere più di 16 milioni di VXLAN segments all'interno 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 EXAMPLE



VXLAN EXAMPLE

1. VM1 invia una richiesta ARP associata all'indirizzo IP 192.168.0.2
2. La richiesta ARP viene gestita dal VTEP1 ed incapsulata in un pacchetto multicast associato al gruppo multicast mappato con il VNI 800
3. Tutti i VTEP associati con il VNI 800 ricevono il pacchetto e aggiungono alla loro tabella il mapping VTEP1/VM1 MAC address
4. VTEP2 riceve il pacchetto multicast, decapsula il pacchetto e trasmette il frames su tutte le porte associate con il VNI 800
5. VM2 riceve la richiesta ARP e risponde alla VM1 con il suo MAC address
6. VM2 incapsula la risposta in un pacchetto IP unicast e lo trasmette verso VM1 (il pacchetto diventa unicast dal momento in cui VM2 ha imparato l'associazione VTEP1/VM1 MAC dal pacchetto sorgente che ha originato la richiesta ARP)
7. VTEP1 riceve la risposta, decapsula il pacchetto e lo invia alla VM1

A questo punto la connessione tra i due servers VM1 e VM2 è stabilita; il traffico unicast con sorgente il VTEP1 quindi seguirà il seguente path:

- IP Source: 192.168.0.1
- IP Destination: 192.168.0.2 e protocol ID settato su UDP oppure 0x0011
- VXLAN VNI: 800
- Standard UDP header con checksum settato a 0x0000 e la VXLAN destination port settato con la corretta IANA port based on vendor
- Standard MAC header con l'indirizzo next-hop MAC address; in questo caso il next-hop è il router con l'indirizzo MAC address AA:AA:AA:AA:01:01

- VTEP2 riceverà il pacchetto dal VTEP1, il processo di decapsulamento è fatto dal valore UDP header; quindi VTEP2 passerà il frame al virtual switch ed alle porte associate con il valore VNI 800

VXLAN CONSIDERATION

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

VXLAN SUMMARY

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