Felhő alapú hálózatok (VITMMA02)
Data Center Bridging, Virtuális hálózati technológiák

Dr. Maliosz Markosz

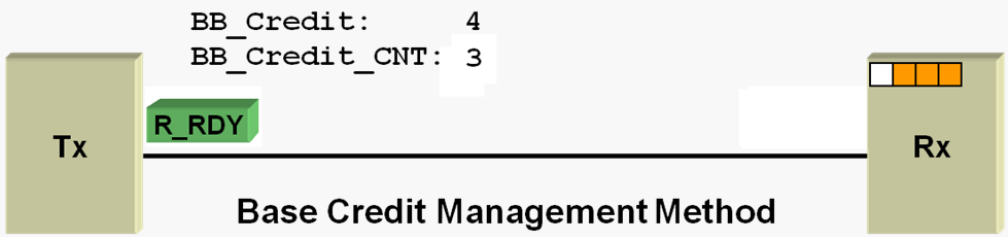
Budapesti Műszaki és Gazdaságtudományi Egyetem
Villamosmérnöki és Informatikai Kar
Távközlési és Médiainformatikai Tanszék

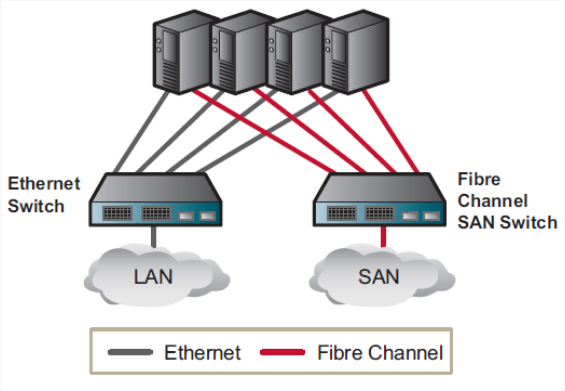
2015. tavasz

DATA CENTER BRIDGING

Háttértár forgalom az adatközpontban

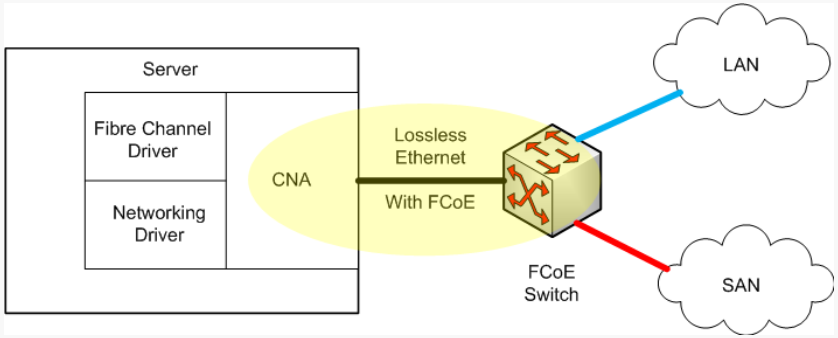
- Adatközpontban korábban
  - Ethernet adatforgalomra
  - Fibre Channel a háttértár forgalomra (SAN – Storage Area Network)
    - külön dedikált hálózat
    - optikai vagy elektronikus interfész
    - 2, 4, 8, 16 Gbps
    - nincs csomageldobás torlódás esetén
      - puffer kredit alapú folyamvezérlés
      - buffer to buffer credit

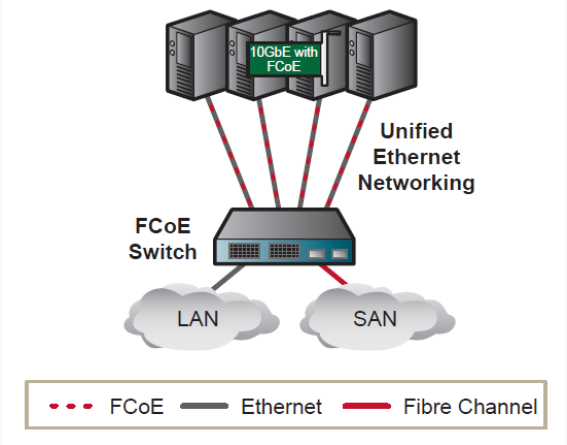




Fibre Channel over Ethernet (FCoE)

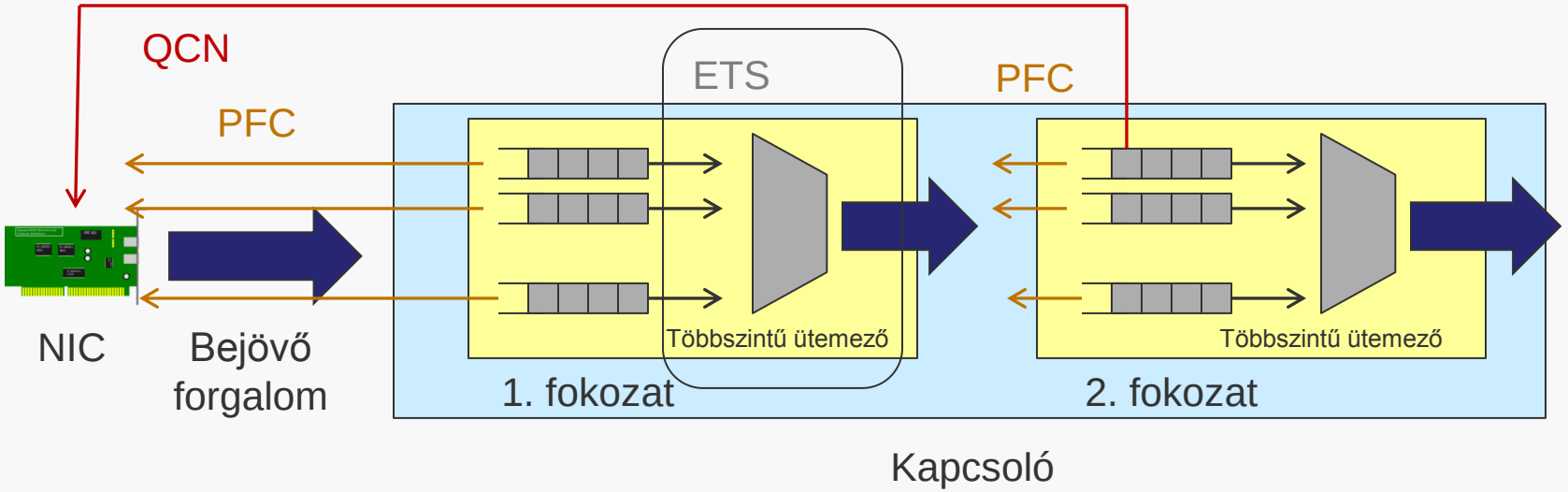
- Ethernet
  - torlódás esetén csomageldobás
  - TCP garantálja a megbízható átvitelt (újraküld)
    - késleltetés ingadozás
    - video és háttértár forgalom számára ez nem ideális
  - kiegészítések szükségesek: DCB





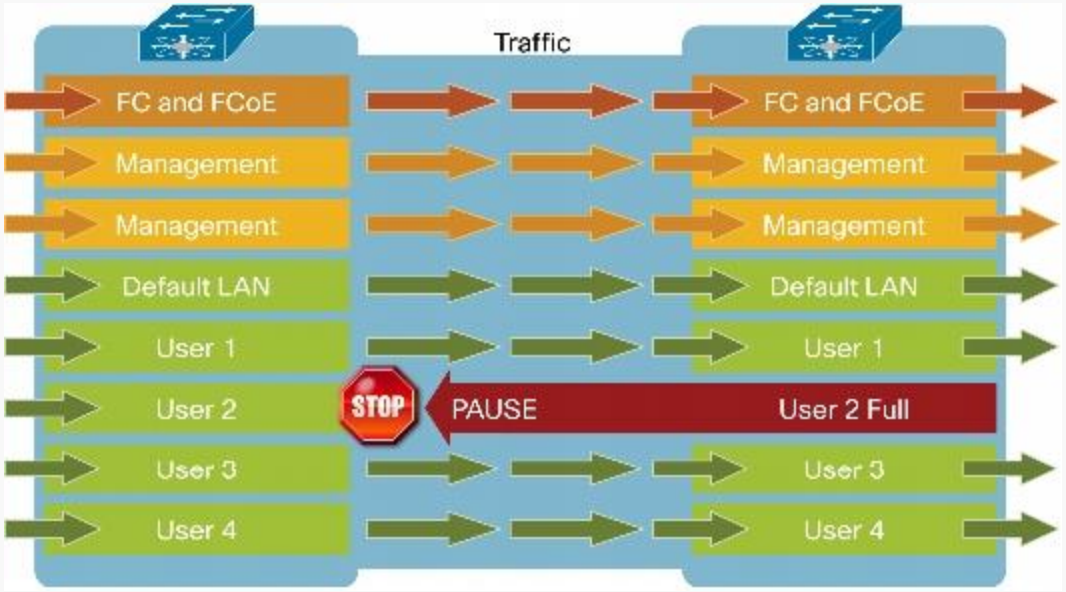
Data Center Bridging

- Ethernet kiegészítések: megbízható(bb) átvitel elérésére a TCP komplexitása nélkül
  - Priority based Flow Control (PFC)
  - Enhanced Transmission Selection (ETS)
  - Quantized Congestion Notification (QCN)
  - Data Center Bridging exchange (DCBx) protocol



Priority based Flow Control

- A torlódásból adódó csomageldobás kiküszöbölésére
- IEEE 802.3Qbb
  - link szintű
  - kapcsolók, vagy kapcsolófokozatok között
- 8 prioritás: virtuális sávok
- kapcsolóban: memória partíciók
  - telítettségi szint
- szünet üzenet: időtartamot tartalmaz



Forrás: Cisco

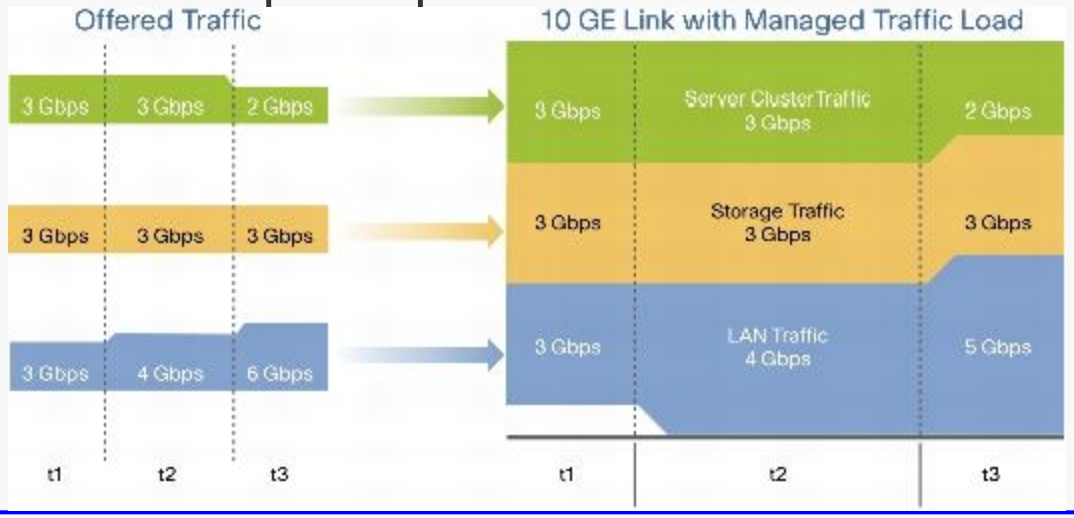
Enhanced Transmission Selection

- IEEE 802.3Qaz
- Forgalmi prioritás osztályok
  - szabályok alapján a fejléc mezők vizsgálatával
    - Access Control List (ACL)
  - VLAN címke 3-bites prioritás mezője
  - osztályok csoportjait kezeli (Traffic Class Group – TCG)
    - min. 3-at kell tudni az ETS-képes kapcsolónak

Prio1: 2 Gbps

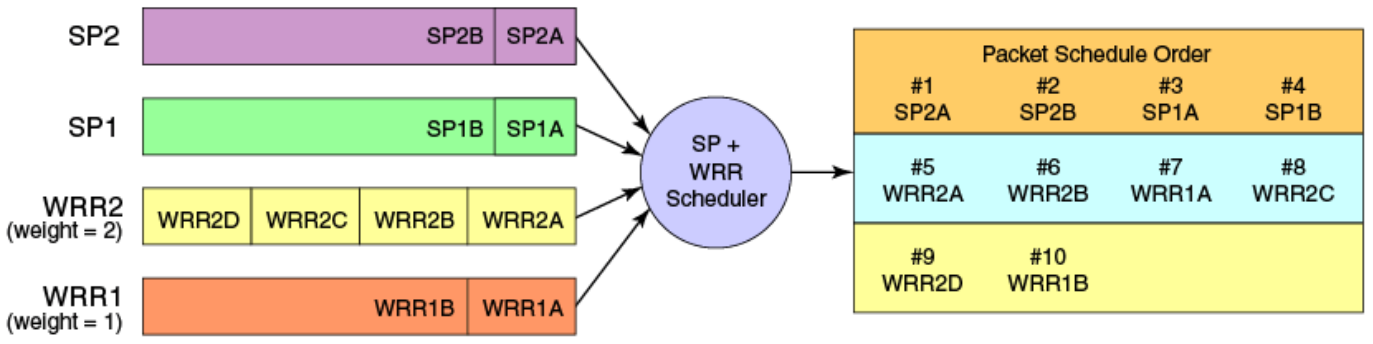
Prio2: 3 Gbps

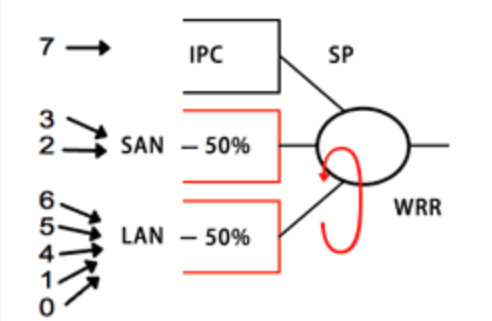
Prio3: 5 Gbs



Enhanced Transmission Selection

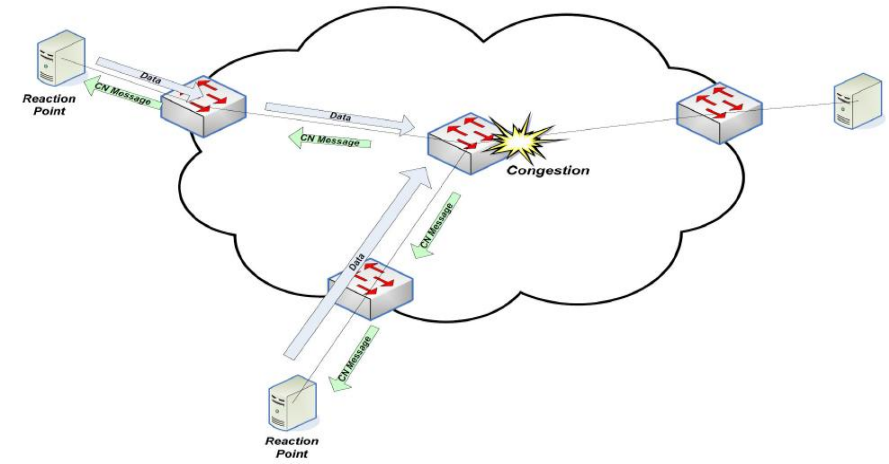
- Sávszélesség allokáció
  - osztály csoportonként (max. 8) minimum garantált bitsebesség
    - 1%-os egységekben, ±10% pontosság
  - a kihasználatlan sávszélességet a többi osztály ki tudja használni
- Megvalósítása: ütemezés (scheduling) és forgalom szabályozás (rate limiting, shaping)





Quantized Congestion Notification

- PFC + ETS
  - veszteségmentes átvitel és sávszélesség garancia
  - gyors reakcióidő
  - azonban: többfokozatú kapcsolókban és adatközpontokban sok kapcsolón áthalad a forgalom
- QCN: tranziens torlódások kiküszöbölésére
  - a forgalom forrását értesíti (ent-to-end)
  - nagyobb időskála
  - torlódási pont
    - sorhosszak vizsgálata, mintavételezés (függ a telítettségtől)
    - a telítettségtől függően visszajelzési érték kiszámítása (6 bitre kvantálás)
    - forrás MAC címre küldés
      - 1-10% közti vg.-gel
    - mintavételezési idő utánállítás
  - beavatkozási pont
    - kimenő forgalom korlátozása a visszajelzett érték függvényében
    - idővel újra növeli a sebességet



Quantized Congestion Notification

- Ritkán implementálják…
  - a szabályozási kör sok tényezőtől függ
    - torlódási pont reakcióideje, visszajelzés megérkezési ideje a forráshoz, a beavatkozási pont reakcióideje
  - finom beállítást igényel
  - hosszú ideig tartó forgalomra ideális
  - a véletlen mintavételezés bizonytalansága
  - a forrásnál minden torlódási ponthoz külön sort kellene fenntartani
  - L2 alhálózatban működik
    - egy útválasztón áthaladva másik QCN tartományba kerül
  - hardveres implementáció lenne a megfelelő a sebességhez
    - összes kapcsoló és hálózati kártya cseréje

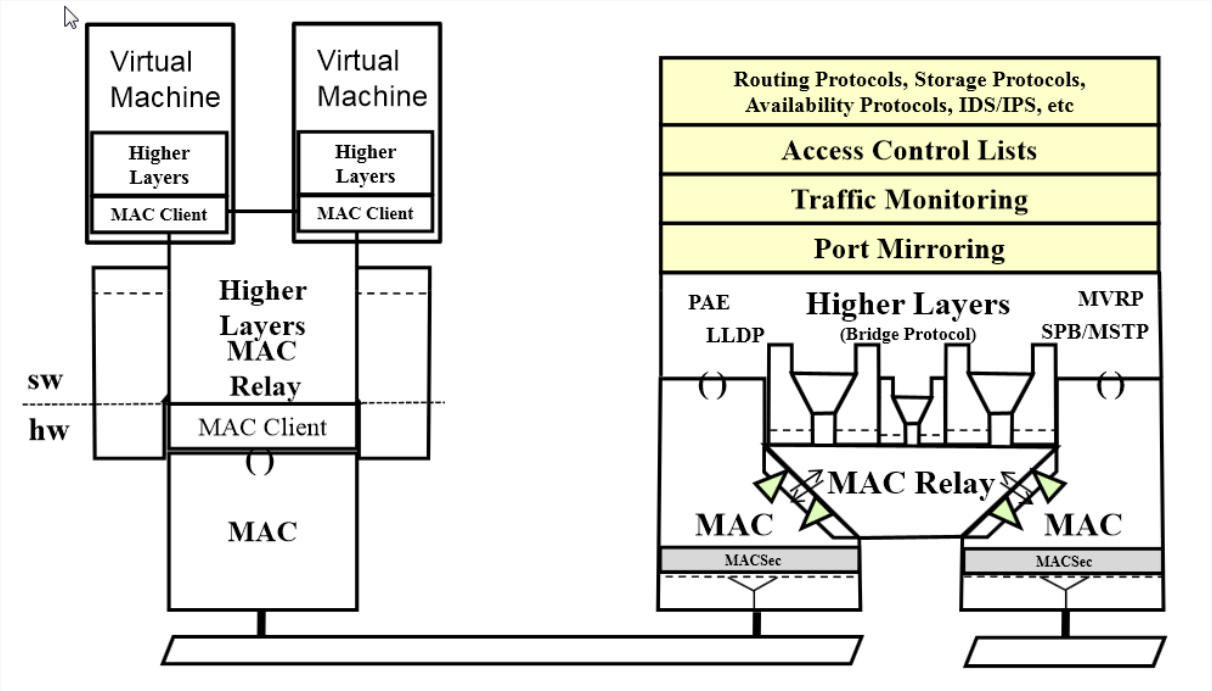
Data Center Bridging exchange (DCBx)

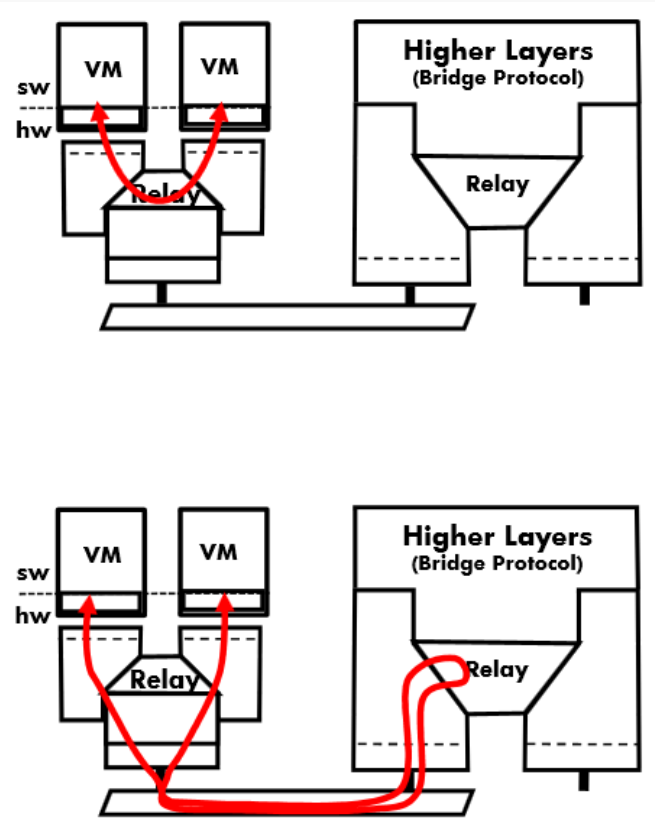
- Koordináció a képességekről szomszédos kapcsolók között
  - PFC
    - prioritások, forgalmi sávok száma
  - ETS
    - sávszélesség egységek
- Link Level Discovery Protocol (LLDP) üzenetekben
  - Type-Length-Value struktúrában
- Működés
  - küldő oldal
    - *javaslat* a másik oldali eszköznek a paraméterek beállítására
    - periodikus küldés
  - vevő oldal
    - paraméterek beállítása a másik oldali eszköztől kapott konfiguráció figyelembe vételével
    - adatbázis frissítés a vett adatok alapján
  - nincsen nyugtázás
    - nem foglalkozik azzal, hogy a másik oldal mit állít be

VIRTUÁLIS HÁLÓZATI MEGOLDÁSOK

Edge Virtual Bridging

- ToR fizikai kapcsoló ⇔ virtuális kapcsoló (Virtual Ethernet Bridge – VEB) képességek
  - szűrés, biztonság, monitorozás, stb.

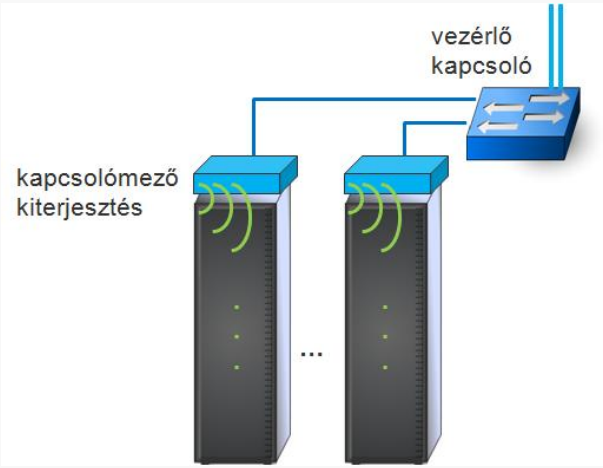




Forrás: Pat Thaler et al., IEEE 802 Tutorial: Edge Virtual Bridging, 2009.

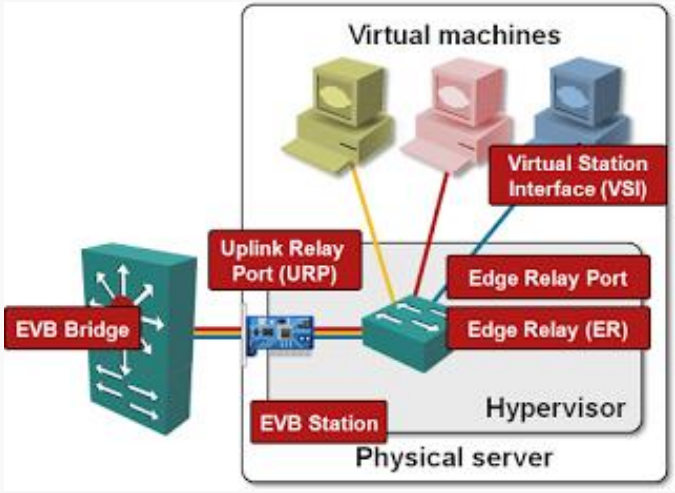
Edge Virtual Bridging

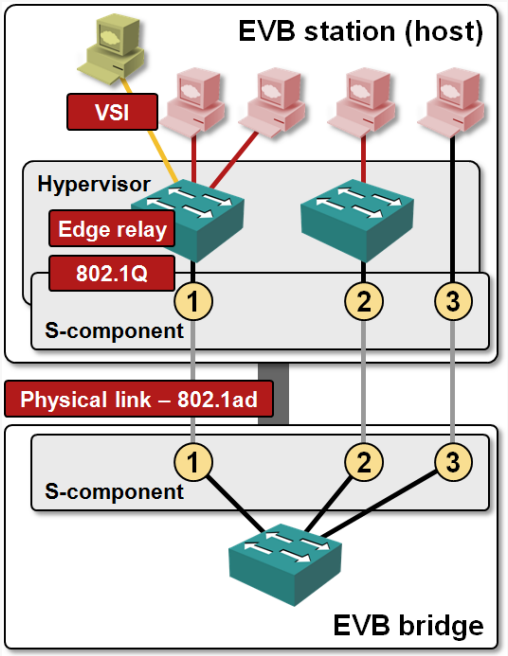
- EVB: IEEE standard
  - a virtuális és fizikai kapcsoló közötti interakció
    - cél: minden forgalom egységes kezelése
  - Virtual Ethernet Port Aggregation (VEPA) 802.1Qbg
    - minden forgalom átmegy a fizikai kapcsolón
  - Virtual Network Tag (VN-Tag), Bridge Port Extension 802.1Qbh, 802.1BR (E-Tag)
    - vezérlő kapcsoló által konfigurált portok
      - a kiterjesztett kapcsolómezőn (S-Tag)
      - szerver fizikai hálózati kártyáján (VN-Tag)
        - minden vNIC-hez külön VN-Tag
        - extra fejléc, benne Virtual Interface (VIF)



L2 konfiguráció automatizálása

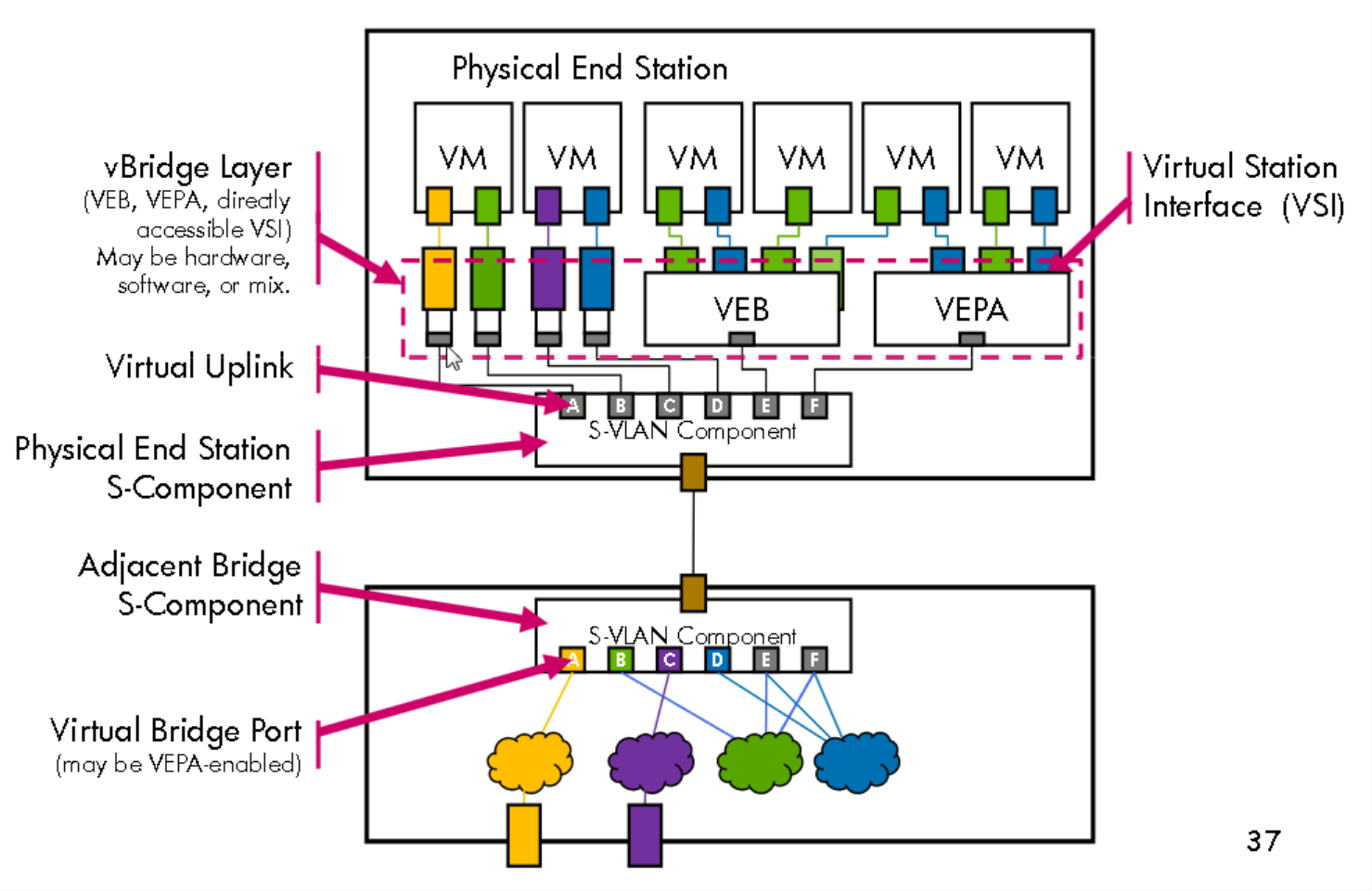
- Edge Virtual Bridging
  - Virtual Station Interface (VSI): VM hálózati interfésze
  - VSI Discovery and Configuration Protocol (VDP)
    - az EVB bridge információt kap a VM indítása előtt a hypervisortól
  - VN-Tag: plusz fejléc a virtuális interfészek azonosítására (Cisco)
    - lokálisan a vezérlő és a kiterjesztett kapcsolómező között
  - S-component
    - logikai 802.1Q összeköttetések multiplexlása egy fizikai szakaszon (Q-in-Q)





Edge Virtual Bridging

» lehet kombinálni a megoldásokat



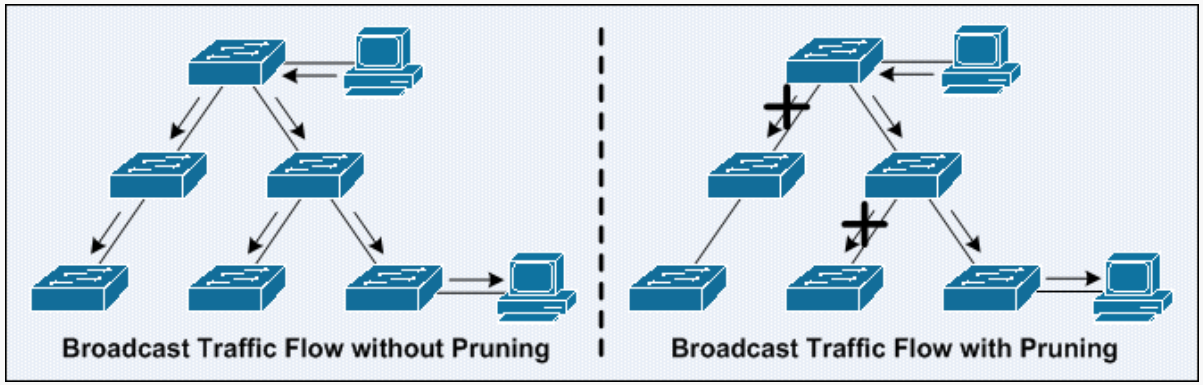
Forrás: Pat Thaler et al., IEEE 802 Tutorial: Edge Virtual Bridging, 2009.

Összehasonlítás

- Virtuális kapcsoló
  - továbbítás MAC + VID alapján
  - nem szükséges
    - MAC cím tanulás, mert VM címek be lehetnek konfigurálva
    - STP, mert a hálózat szélén helyezkedik el
  - forgalom a szerveren belül marad
    - kívülről nem látható, analizálható, szűrhető
    - jobb teljesítmény azonos szerveren lévő VM-ek között
  - nincs egységes menedzselés a fizikai kapcsolókkal
  - CPU és memória használat
- EVB
  - minden forgalom fizikai kapcsolón keresztül (fejlettebb képességek)
  - kevesebb hálózat konfigurációs feladat
  - nagyobb forgalom és késleltetés
  - VEPA
    - továbbítás MAC + VID alapján
    - virtuális kapcsoló funkció megmarad
    - változatlan Ethernet keretek
    - forgalom visszaküldése a beérkező porton
  - VN-Tag
    - továbbítás címke alapján
    - új keretformátum
- Alkalmazás szempontjából
  - VEPA: hypervisor támogatás szükséges
  - VN-Tag: speciális hálózati kártya szükséges
  - irányok
    - fizikai kapcsoló funkciók integrálása a virtuálisba
    - más hálózat virtualizáció és alagút megoldások (VXLAN, NVGRE, stb.)

Virtuális hálózati megoldások

- STP problémák: útvonalválasztás (pl. IS-IS) MAC címekre
  - Shortest Path Bridging MAC (SPBM)
- VLAN-ok száma korlátos: még egy VLAN címke hozzáadása
  - Q-in-Q, provider bridging, (IEEE 802.1ad)
- MAC cím korlát: még egy MAC cím fejléc hozzáadása
  - Provider Backbone Bridges (PBB), 802.1ah
  - Transparent Interconnection of Lots of Links (TRILL)
    - bridging + routing
- Hypervisor elárasztás ellen: VM-eket figyelembe venni
  - VLAN pruning (nyesés, metszés): felesleges forgalom eliminálása
- Maghálózati elárasztás ellen: VLAN pruning a maghálózatban

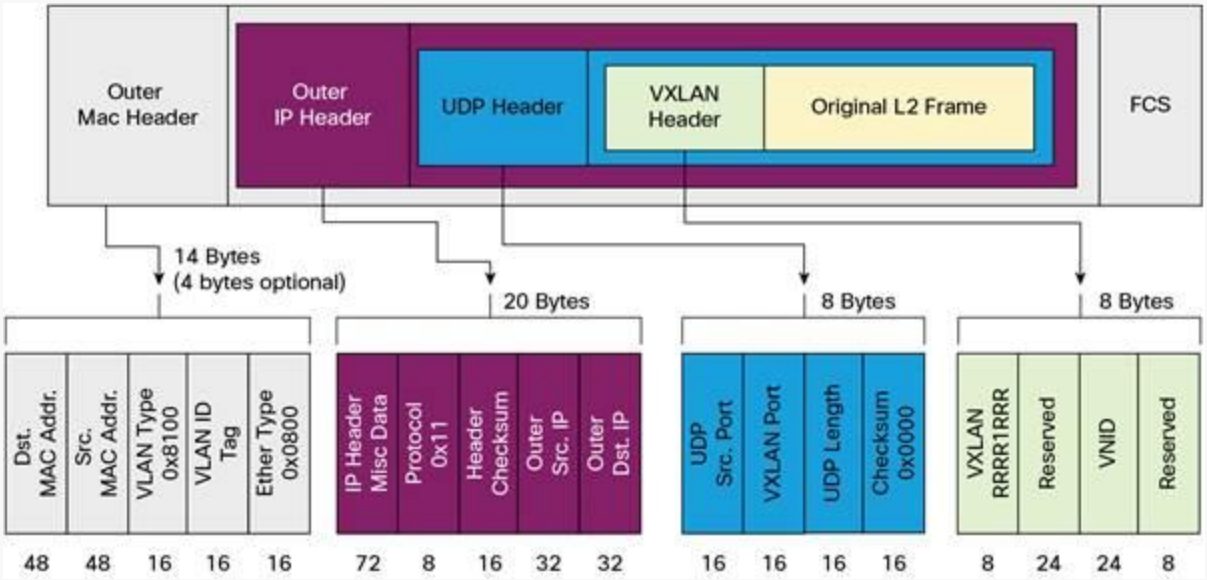


Hálózat virtualizáció

- VN-Tag csak a VM-et azonosítja, az ügyfelet nem
- Ügyfél szintű szeparáció támogatás
  - Virtual Extensible LAN (VXLAN) – RFC 7348
    - Cisco, VMware
    - virtuális L2 hálózati forgalom átvitele L3 fizikai hálózaton
  - Network Virtualization using Generic Routing Encapsulation (NVGRE)
    - Microsoft, Intel, HP, Dell
  - Generic Network Virtualization Encapsulation (GENEVE)
    - a fenti kettő fúziója
  - Stateless Transport Tunneling (STT)
    - Nicira ⇨ VMware

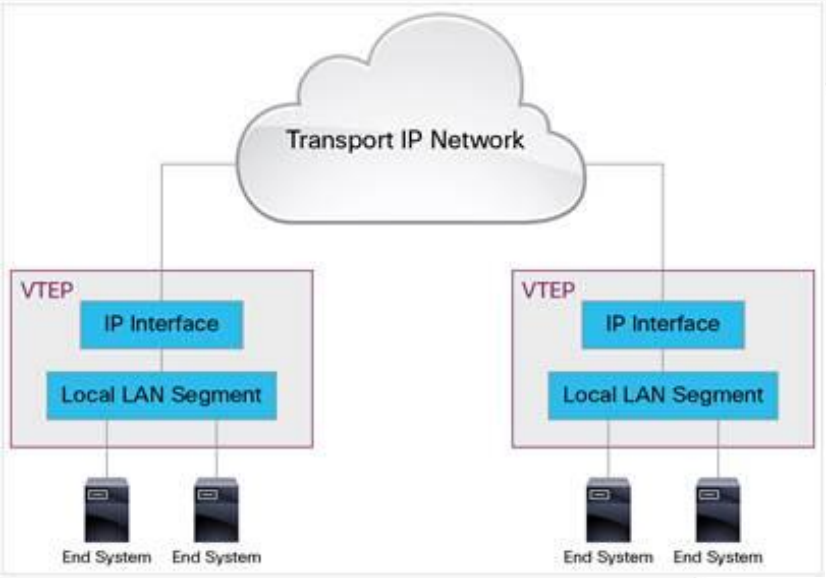
VXLAN

- ügyfél eredeti L2 kerete
  - eredeti MAC címmel és VLAN címkével
- MAC-in-UDP
- VXLAN és UDP fejléc
  - VXLAN network ID (VNID) – ez azonosítja az ügyfelet
    - 24 bit ⇨ 16 millió ügyfél
- fizikai hálózat: IP útvonalválasztás (Layer3)

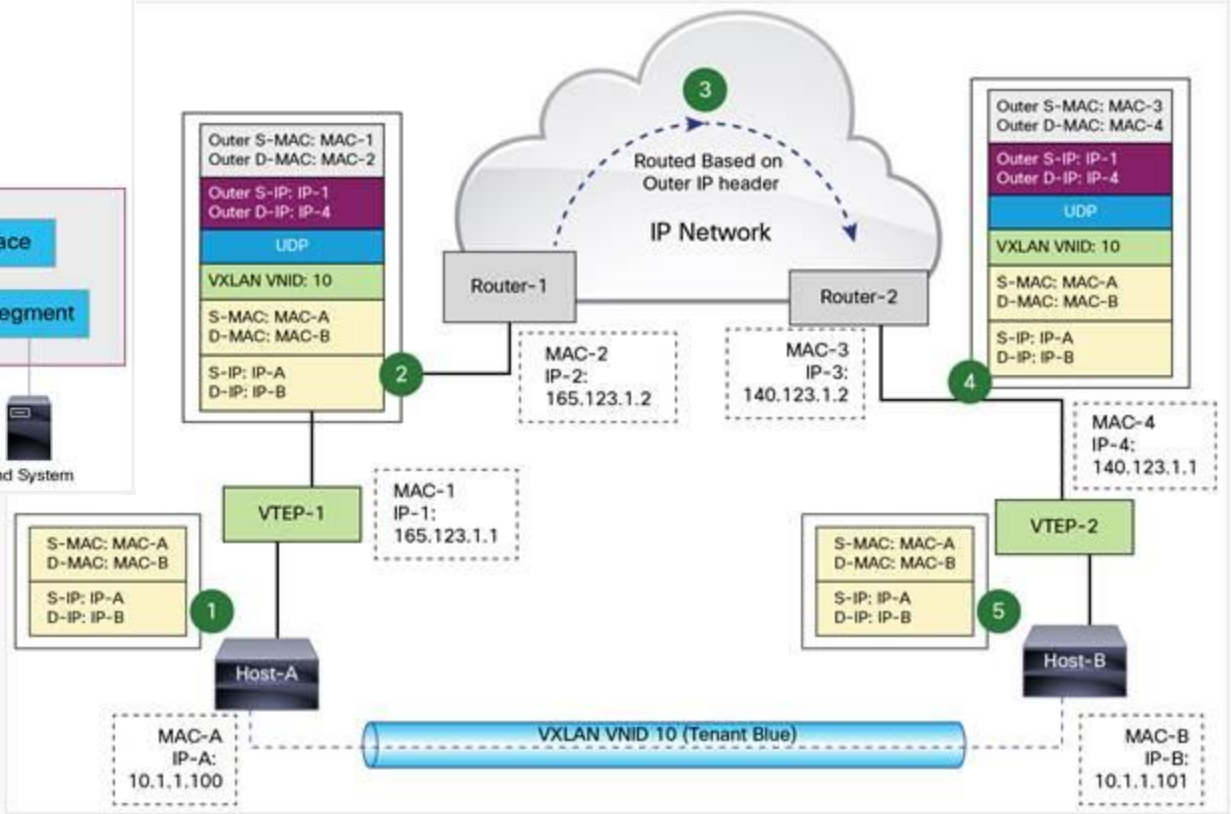


VXLAN

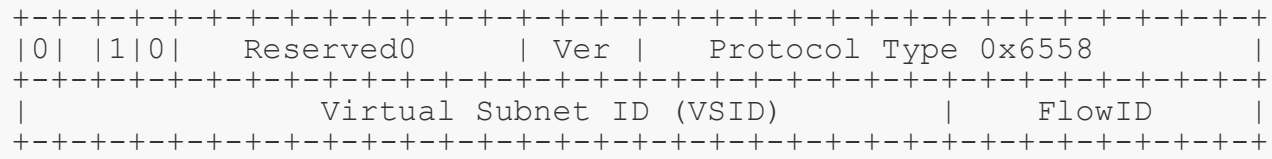
- VXLAN Tunnel End Point (VTEP)
- MAC-to-VTEP táblák tanulás útján (IP multicast)
  - egy VNI összes VTEP-je egy multicast csoportban



ECMP



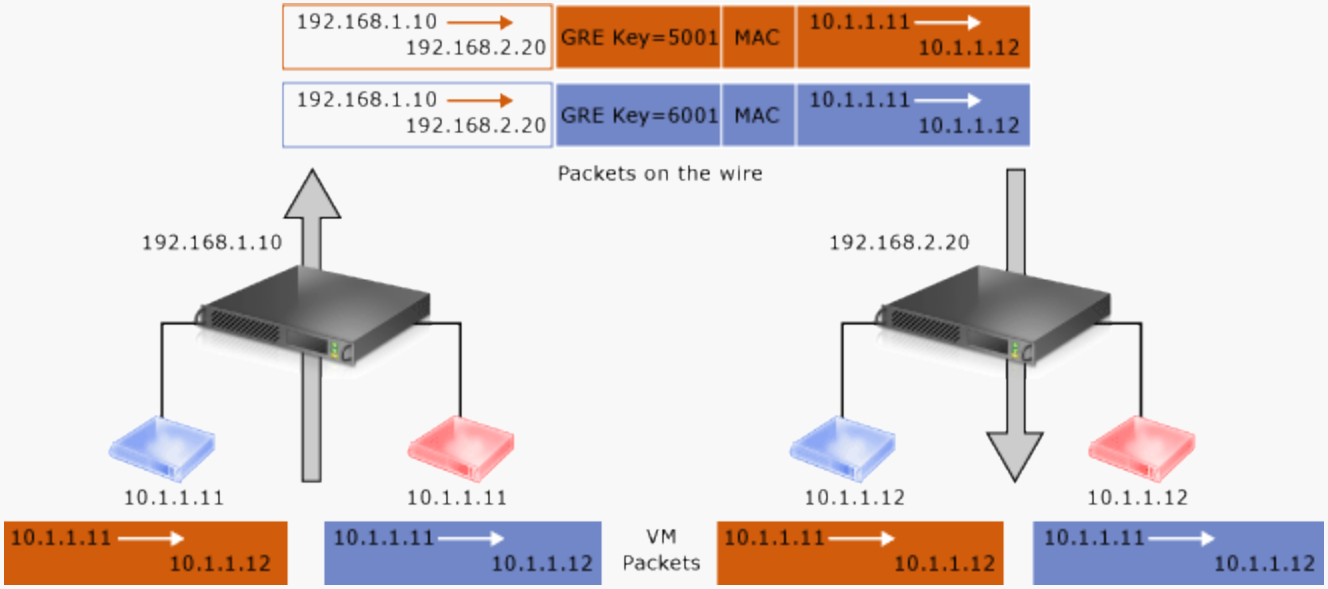
NVGRE

- hasonló a VXLAN-hoz
- alapja: Generic Routing Encapsulation (GRE)
  - általános fejléc
  - sok különböző protokollra
  - pont-pont kapcsolat
- NVGRE
  - GRE fejléc
  - 

    - Virtual Subnet Identifer (VSID) 24 bit ⇨ 16 millió ügyfél
    - FlowID: opcionális, egyedi folyamazonosító
      - ECMP hash számításhoz
  - belül nincs VLAN címke (vagy levételre kerül)
    - VSID-be kódolják

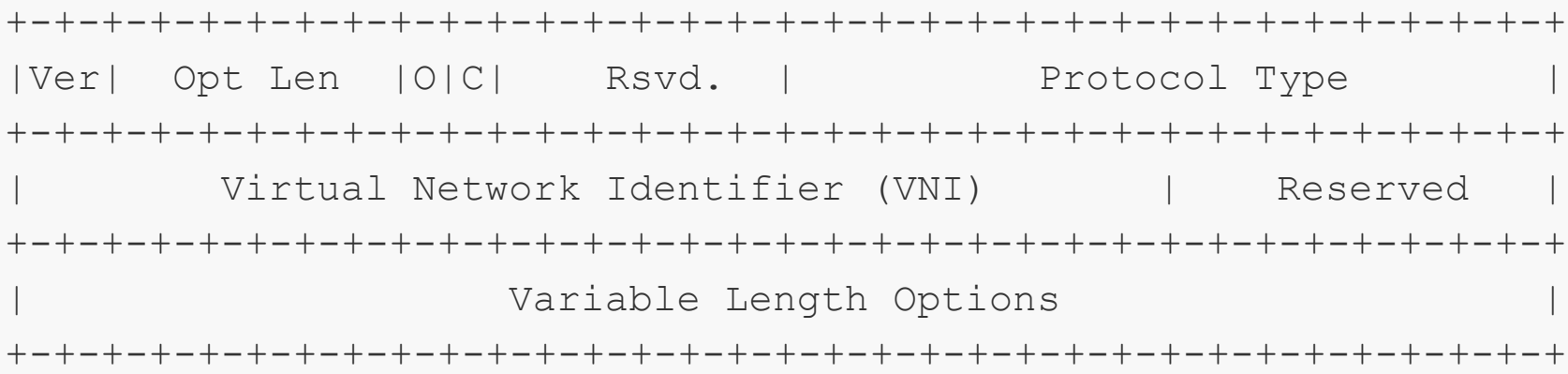
NVGRE

- Network Virtual Endpoint (NVE)
  - VSID és DMAC alapján a címzetthez kapcsolódó NVE IP címére küldés
- az Internet draft nem specifikálja
  - a cím információk terjesztését
  - VLAN információ helyreállítását



Generic Network Virtualization Encapsulation

- MAC-in-UDP over IPv4/IPv6
- univerzális, kiterjeszthető megoldási javaslat
- csak a beágyazási formátumot definiálja
- opcionális mezők
  - nem fix mezőhosszak, rugalmasság
- Geneve fejléc:

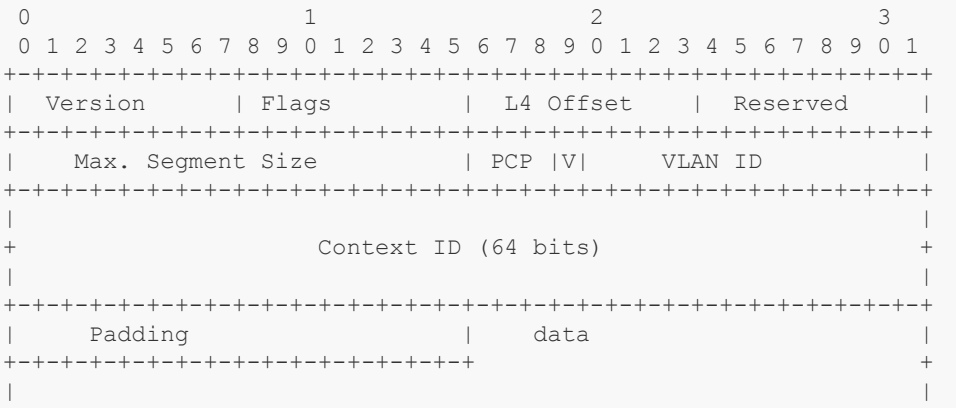


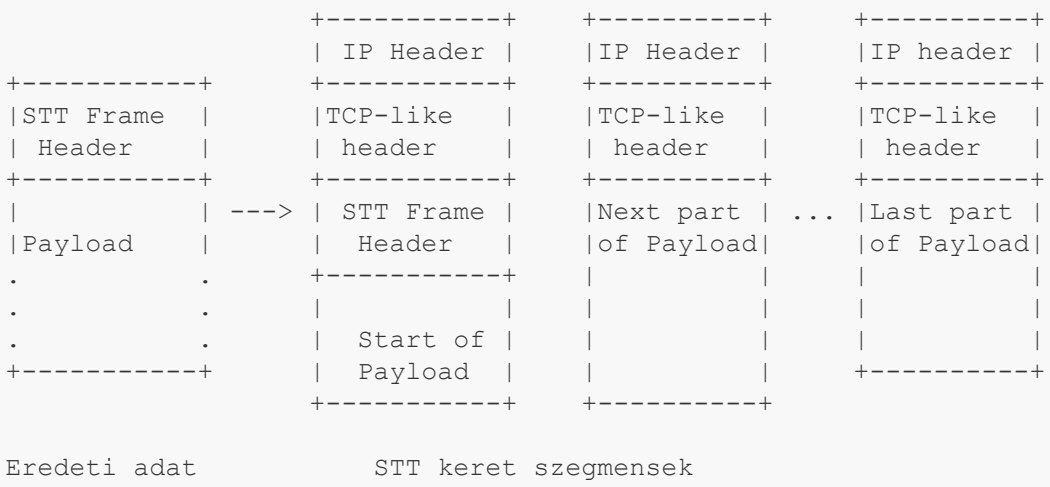
Alagút végződés helye

- hypervisor/vSwitch-ben
  - legközelebb a VM-ekhez
  - CPU erőforrás
  - TCP segmentation offload (TSO), checksum offload támogatás
- fizikai hálózati kártyán
  - offload támogatás a tunnel protokoll fejlécre is
- fizikai kapcsolón
  - forrás VM nem ismert
  - VNID/VSID meghatározásához kell a belső MAC cím

Stateless Transport Tunneling (STT)

- elsődlegesen vSwitch-ek közötti kommunikációra
- komplexebb, mint az előzőek
- max. 64 kbyte-os Ethernet keretet kezel
  - maximum transmission unit (MTU)
  - TCP segmentation offload kihasználása a hálózati kártyán
- STT fejléc
  - 64 bites Context ID mező
- a feldarabolt adatok elé TCP-szerű/IP/Eth fejléc
  - ez alapján állítja össze a nagyméretű keretet





Összehasonlítás

	VXLAN	NVGRE	STT
Plusz bájtok	50 (VLAN: +4)	42 (VLAN: +4)	Első szegmens: 76 Továbbiak: 58 (VLAN: +4)
Protokoll	UDP	GRE	TCP
Ügyfél megkülünböztetés	24 bit VNID	24 bit VSID	64 bit Context ID
ECMP-hez megkülönböztetés (belső⇨külső folyam)	Forrás UPD port	VSID + FlowID (8bit)	Forrás TCP port

Források

- Pat Thaler et al., IEEE 802 Tutorial: Edge Virtual Bridging, 2009.
- Overlay Virtual Networking Explained, Ivan Pepelnjak, NIL Data Communications, 2011.