Adatközponti hálózat - Technológiák a Felhő alapú adatközpontokhoz

Zeisel Tamás
Senior Solution Architect
tamas.zeisel@hpe.com

2017. November 16.

Hewlett Packard Enterprise


200X ??

Miről Beszéltünk 200x-ben Adatközponti Hálózatok kapcsán ?

- Topológia → 3 rétegű model
- Sávszélesség → 10G Ethernet
- Eszközválasztás szempontjai:
  - Operációs rendszer
  - ASIC → Technológiai gazdagság:
    - Tároló/Adat hálózat integráció FCoE
    - DCI Technológiák, Trill ….
    - …..

Hewlett Packard Enterprise

Mi változott 2002-2010 között?

2002 július

2008-2010

2008-2010

Felhő – Cloud Centric adatközpont

Hewlett Packard Enterprise

Mik változik a Felhő alapú adatközpont megjelenésével Funkciók

- Méret – Skálázhatóság
- Egyszerűség – topológia, kialakítás, konfiguráció
- Szabványosság – nem lehet Vendor lock
- Automatizálás szabványosan REST API, Python, Chef, Ansible Puppet
- Operációs rendszer: Linux, Linux, Linux….
- Szolgáltatói – multitenant kialakítás lehetősége
- Speciális funkciók – nem igazán:
    - Tároló hálózati integráció
    - DCI, Trill

Mik változik a Felhő alapú adatközpont megjelenésével Hardware

- Mi a helyzet a Hardware-rel?
- Boltban megvásárolható Merchant Silicon

DUNE NETWORKS

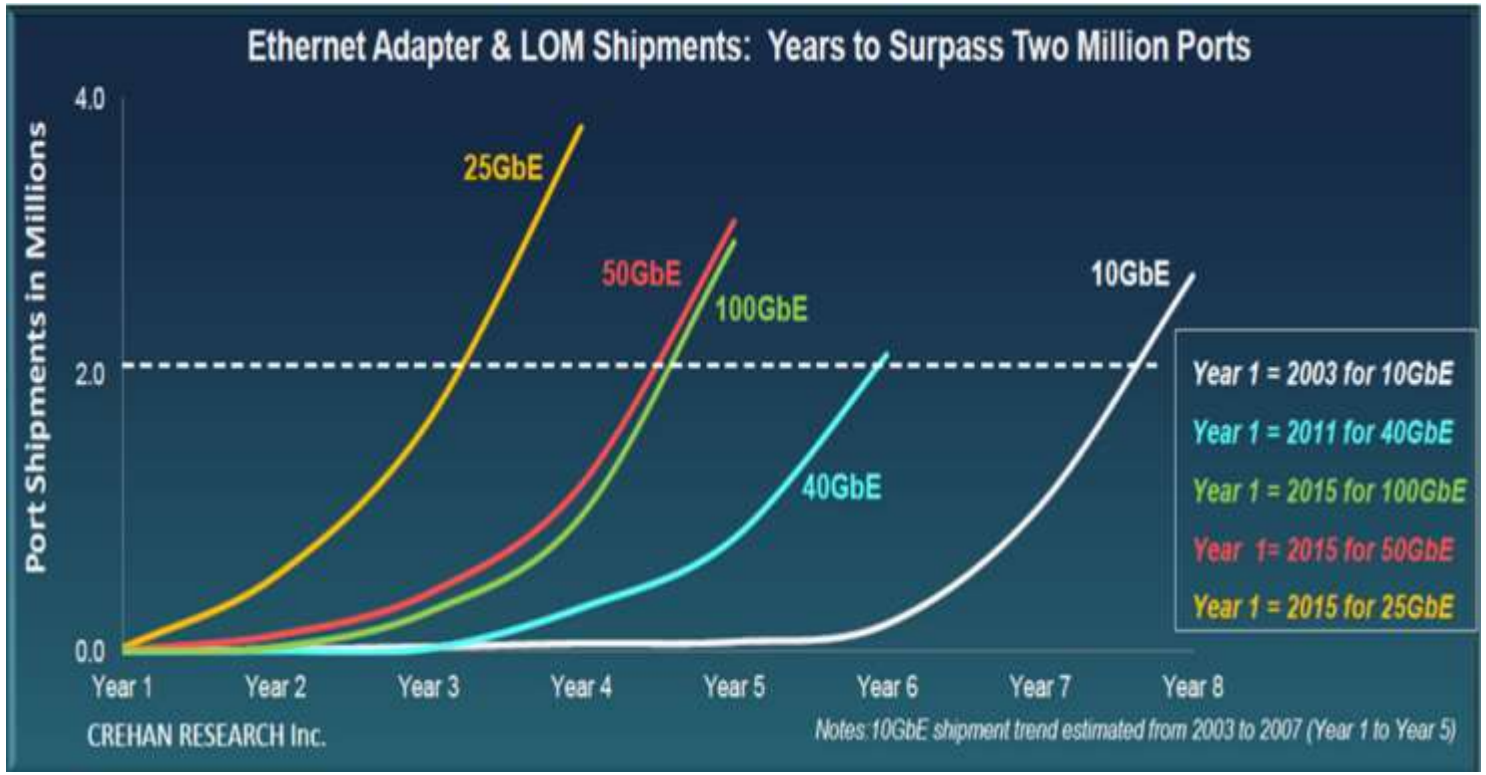
intel

FULCRUM microsystems

Xpliant

BAREFOOT NETWORKS

Sávszélesség 10->40->25/100GbE



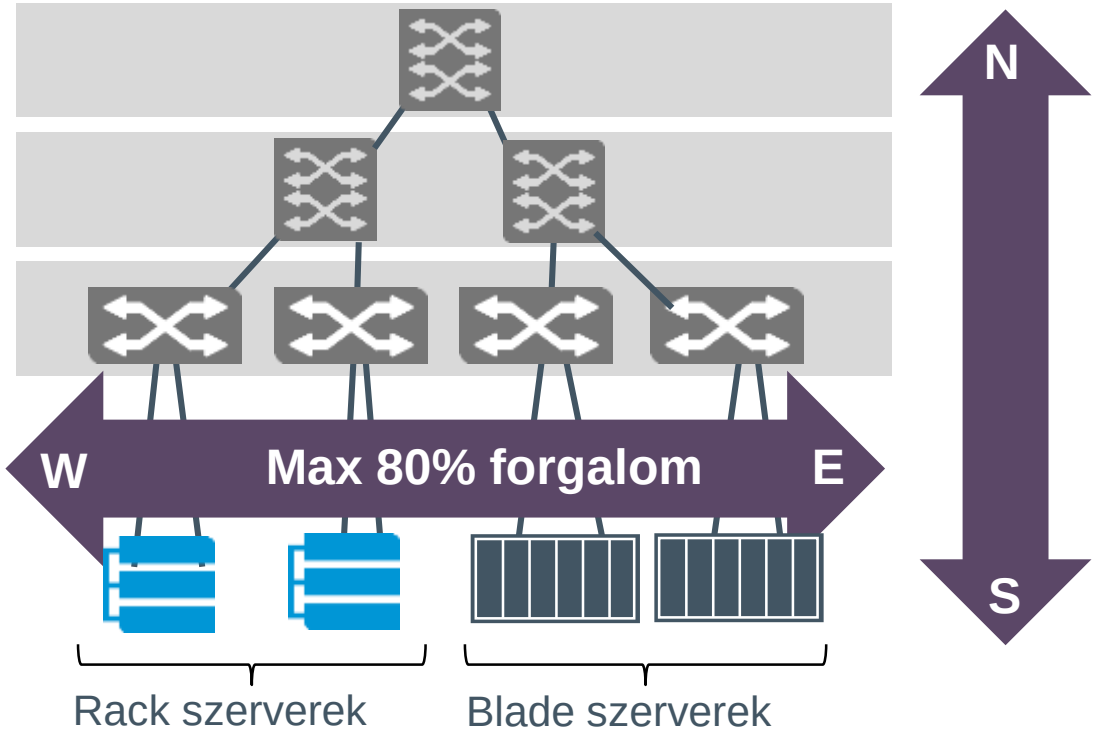
"We predict that 25/50/100GbE will comprise over a third of server access ports, within just three years of shipments"

Sameh Boujelbene, Dell'Oro Group

Topológia

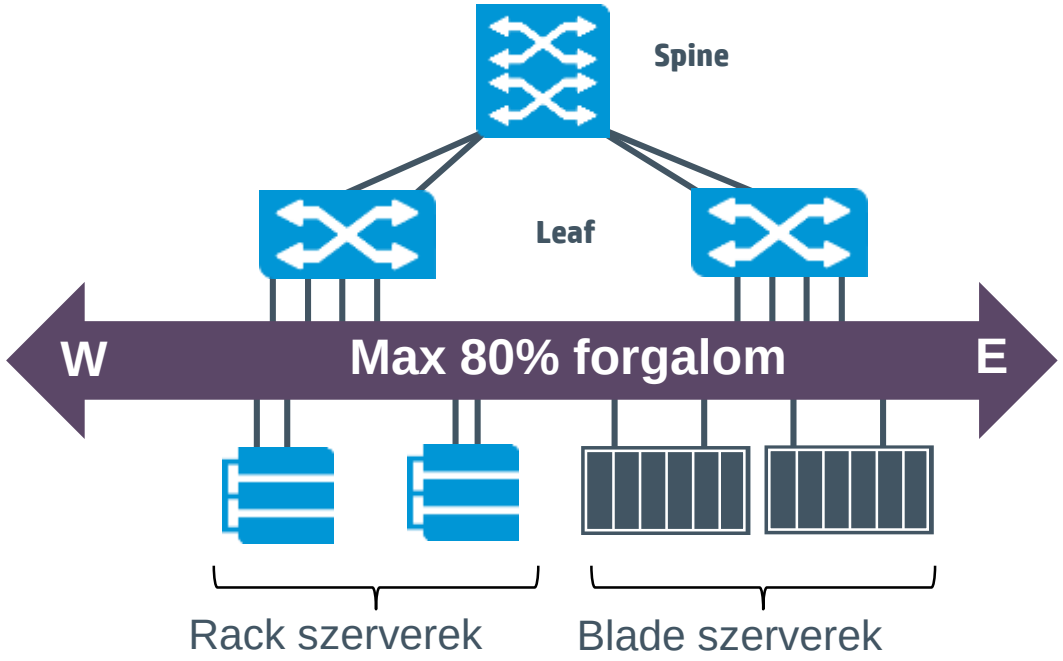
Bonyolult, nagy késleltetésű drága

Hagyományos hierarchikus Model



Egyszerű, hatékony innovatív

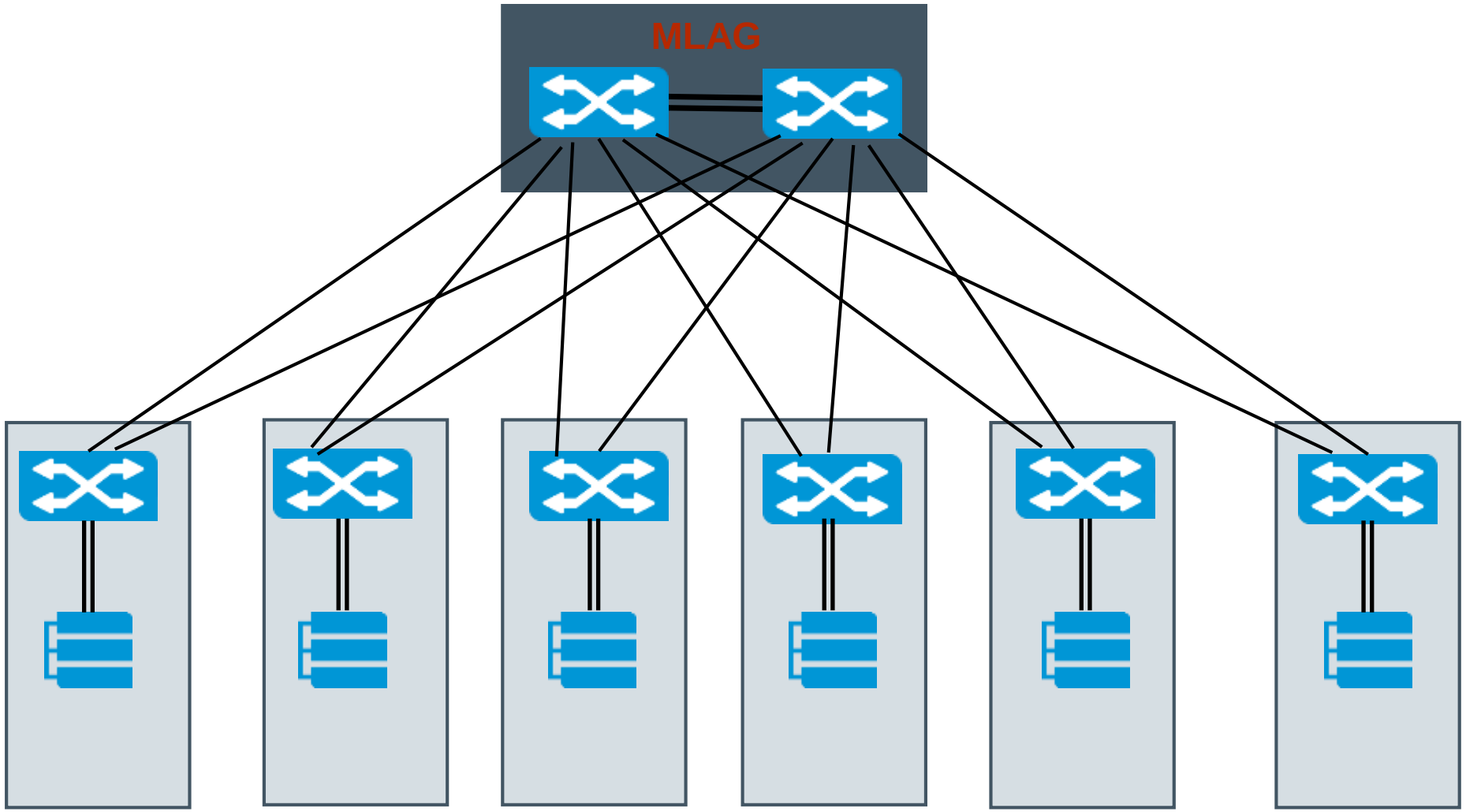
Spine – Leaf 2 szintű



75%-kal csökkentheti a komplexitást

Hewlett Packard Enterprise

Spine – Leaf topológia

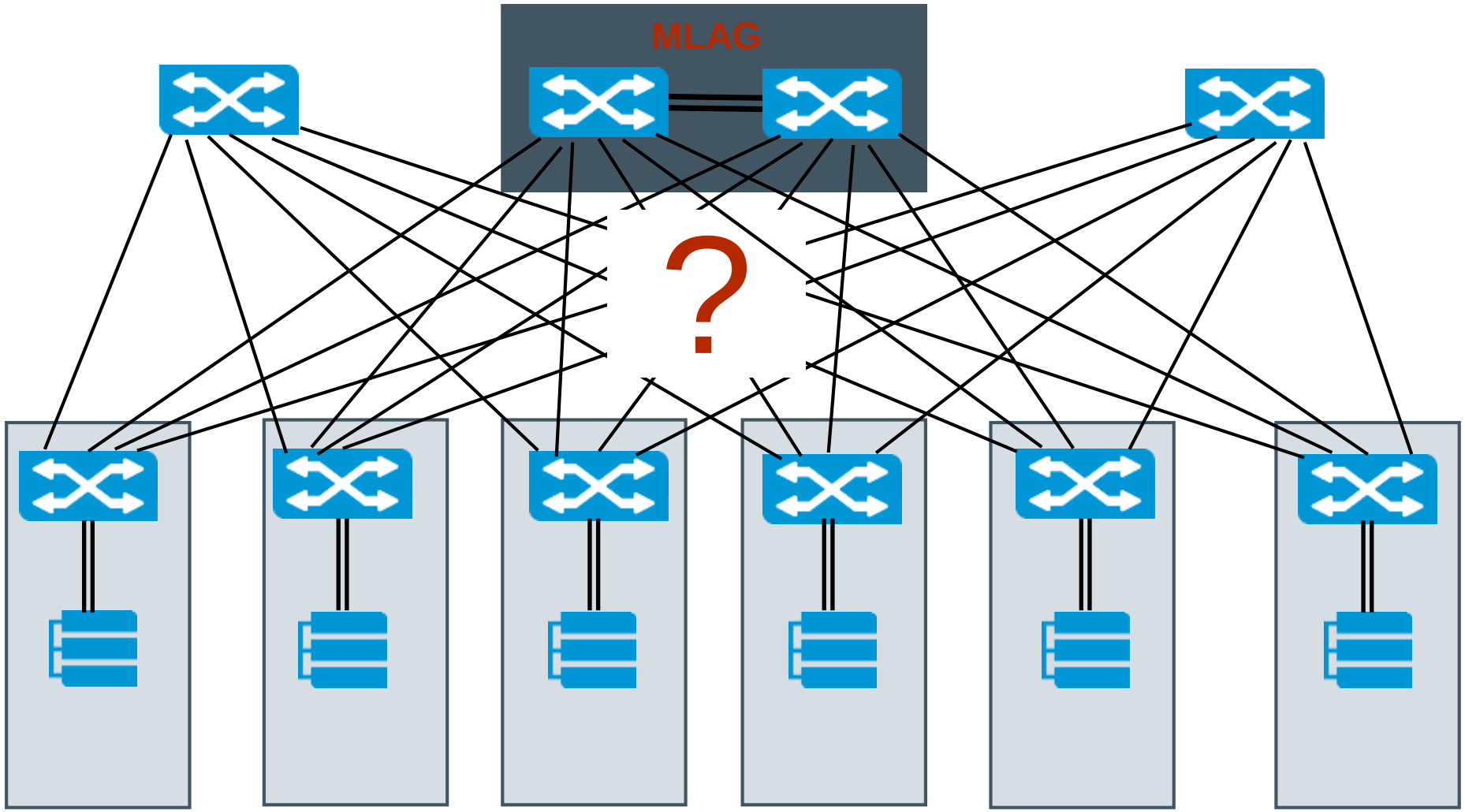


Spine

Leaf

Hewlett Packard Enterprise

Spine – Leaf topológia

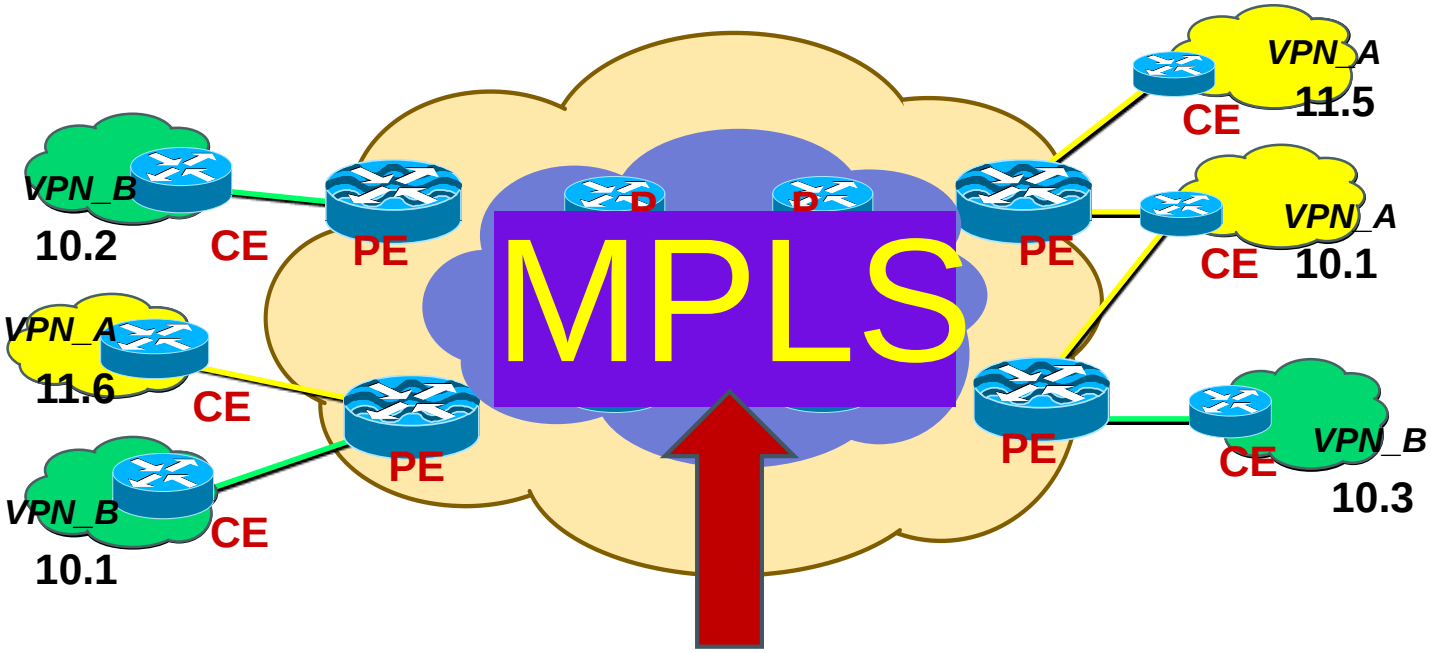


Spine

Leaf

Hewlett Packard Enterprise

Szolgáltatói Multitenant L3 tecnológia
MPLS VPN



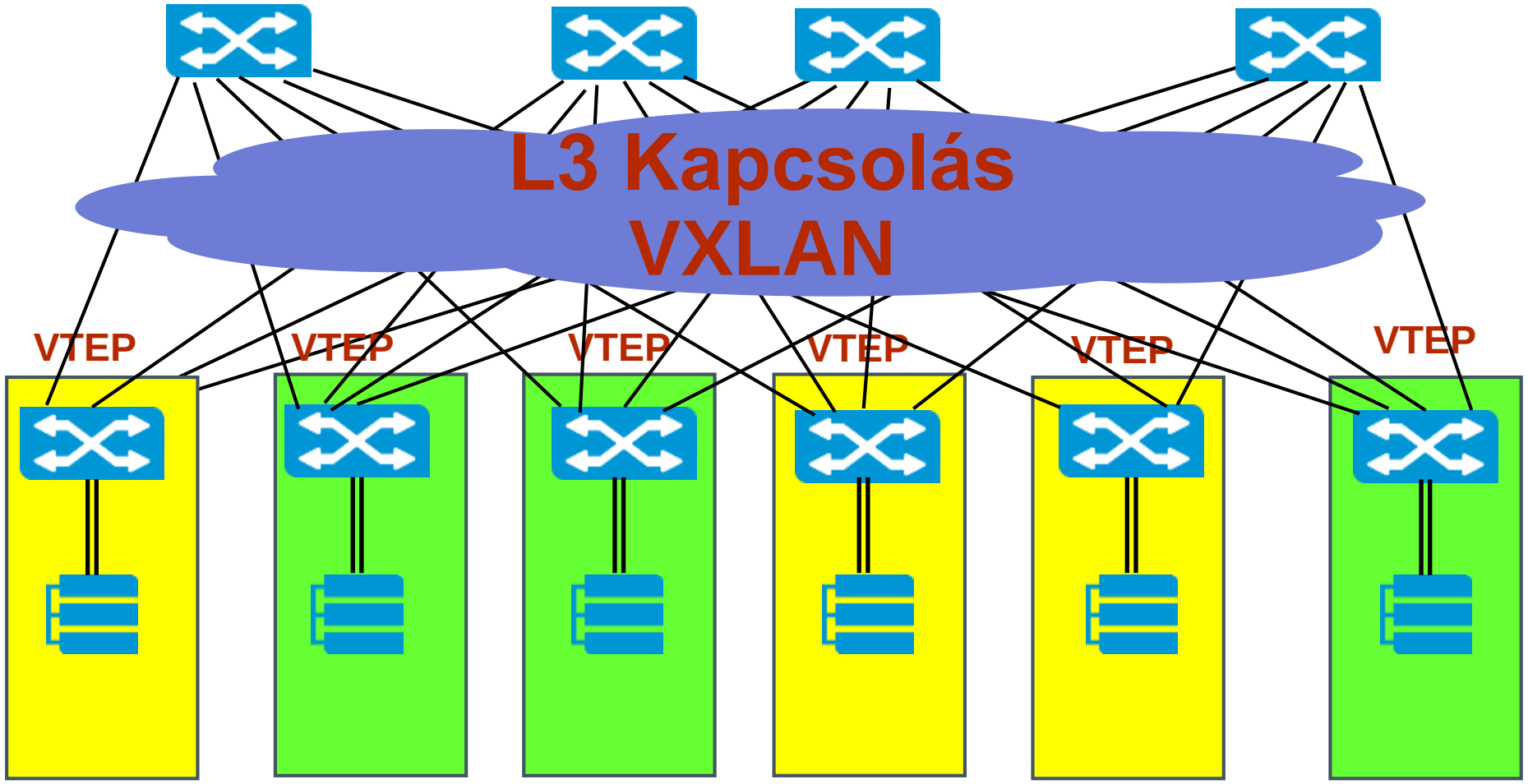
Kapcsolás MPLS Cimke alapon
+ VPN Cimke

Előnyök:

- **Adott Felhasználói csoport (VPN) csak a saját csoportjával beszél közvetlenül**
- **Átlapolt IP Címek használhatók**

Hewlett Packard Enterprise

Mi helyzet L2-őn?



Spine

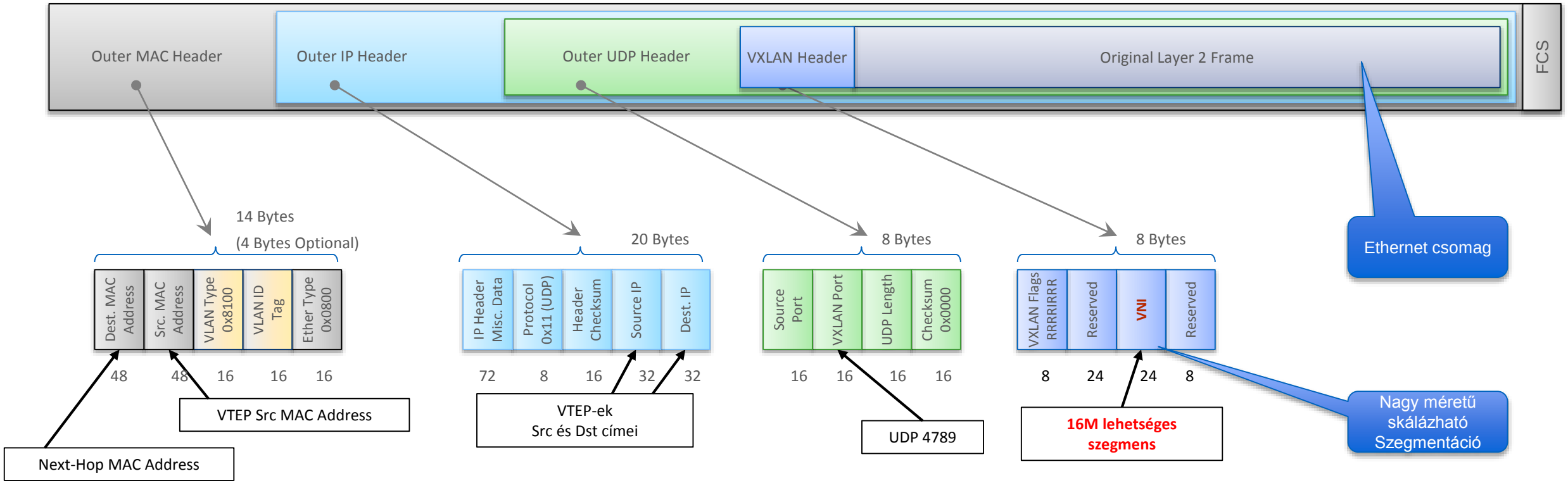
Leaf

VTEP = VXLAN Tunnel Endpoint

Hewlett Packard Enterprise

VXLAN Packet Struktúra

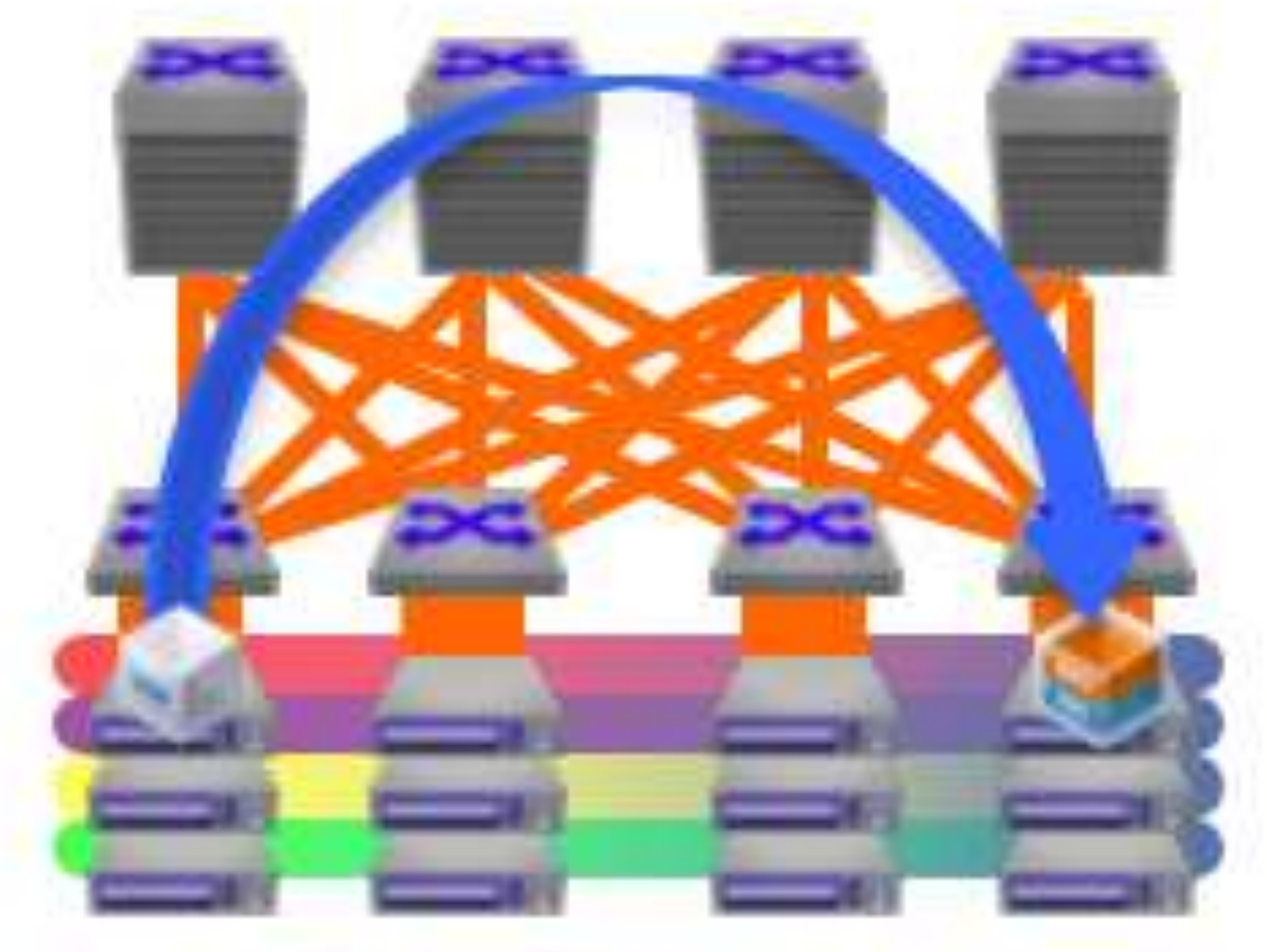
Ethernet in IP



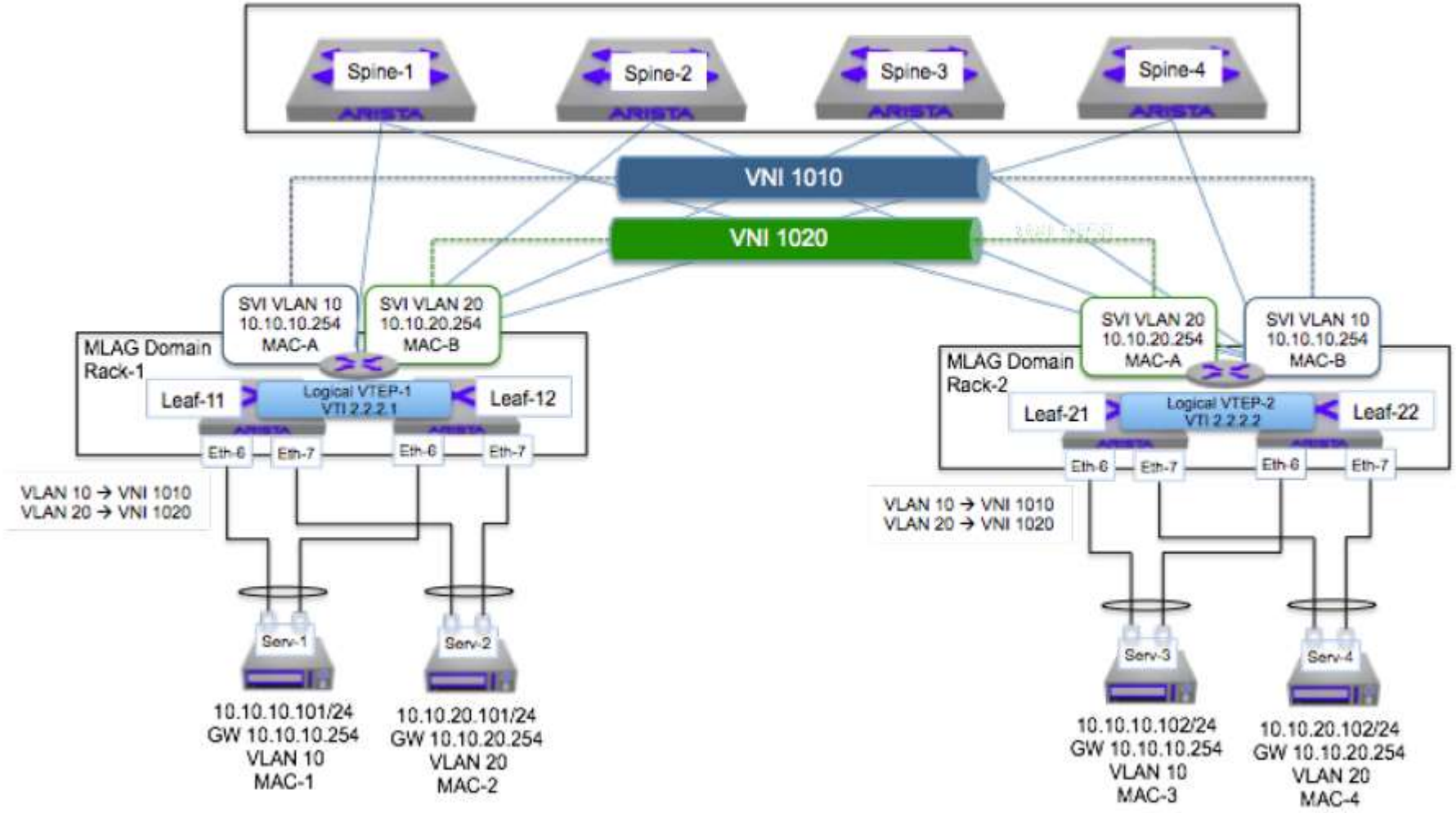
50 (54) Bytes of overhead

Hewlett Packard Enterprise

VXLAN hálózati model



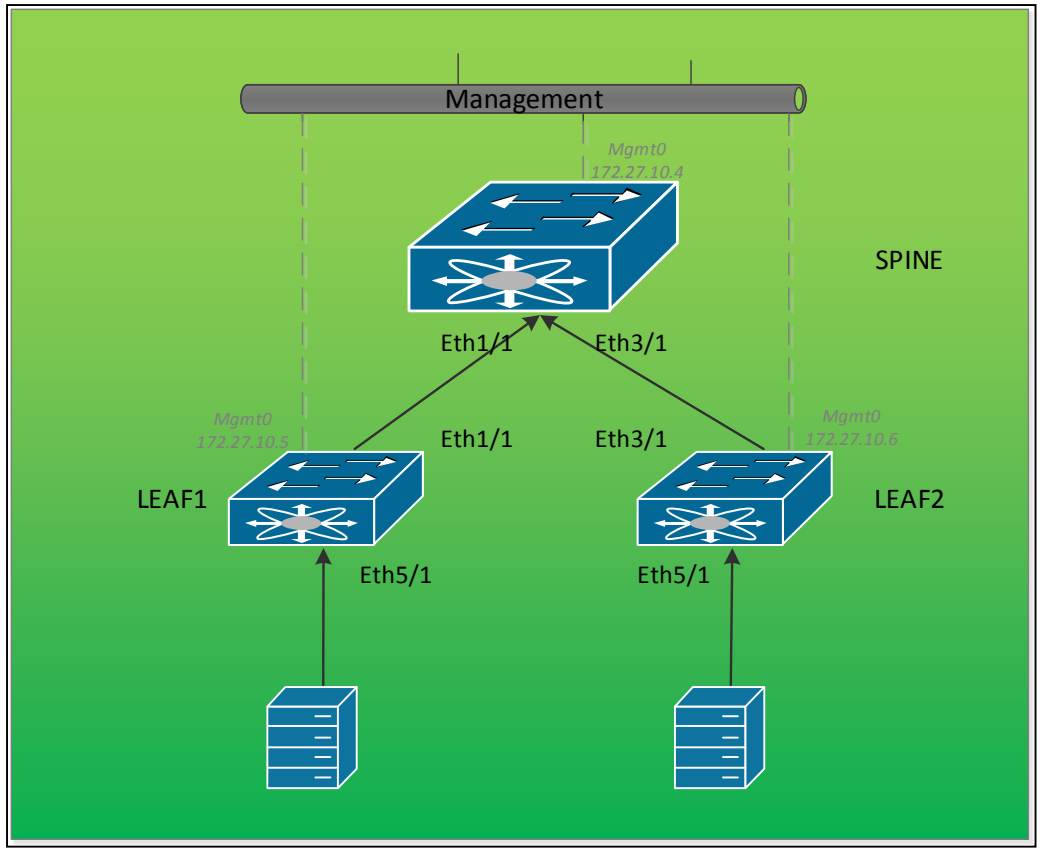
VXLAN – Statikus VXLAN konfiguráció



Hewlett Packard Enterprise

Felhasználói tesztek – Arista eszközökkel – Statikus Control plane

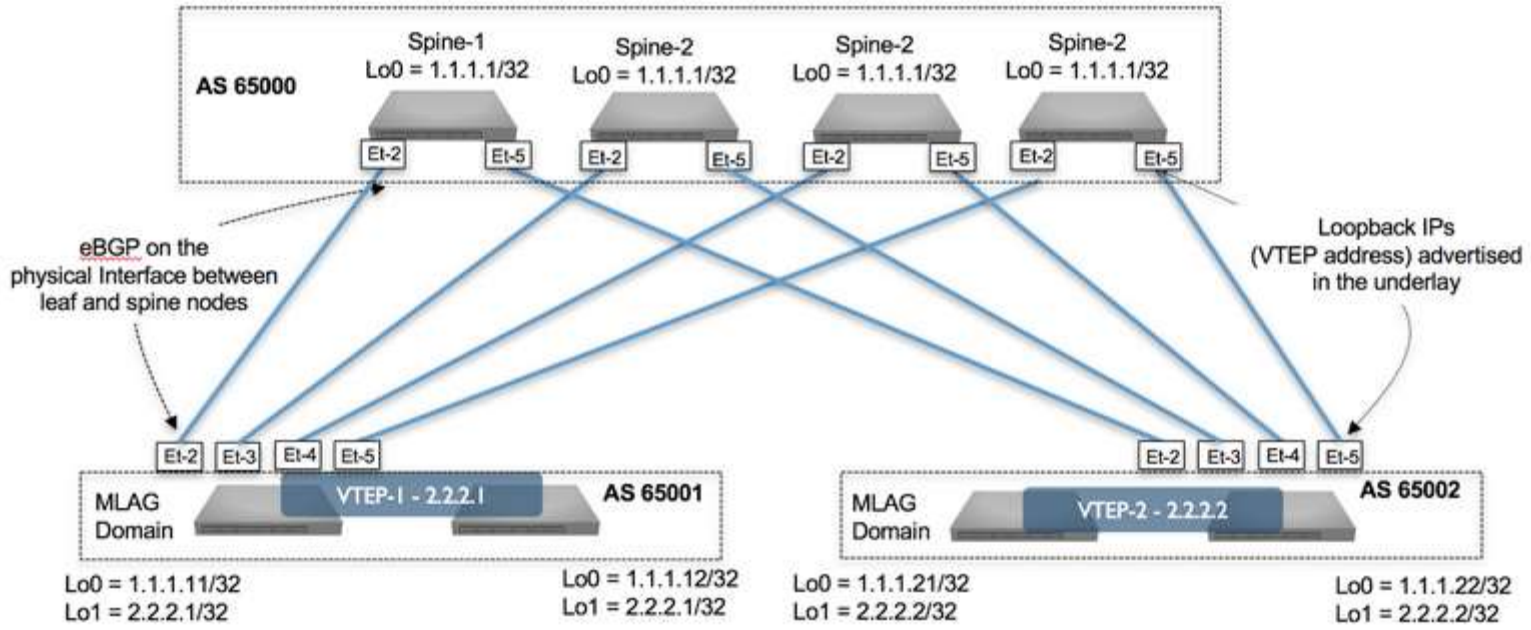
(Kákonyi István, Valkó László, Gelsi Tamás, Zeisel Tamás)




```
interface Ethernet1/1
   mtu 9214
   no switchport
   ip address 10.10.0.1/30
   ip ospf area 0.0.0.0
!
interface Ethernet25/1
   switchport access vlan 100
!
interface Loopback100
   ip address 10.100.1.1/32
!
interface Vlan100
   mtu 9214
   no autostate
   ip address 192.168.32.1/24
!
interface Vxlan1
   vxlan source-interface Loopback100
   vxlan udp-port 4789
   vxlan vlan 100 vni 100
   vxlan vlan 100 flood vtep 10.100.1.1 10.100.1.2
!
router ospf 1
   network 10.10.0.1/32 area 0.0.0.0
   network 10.10.1.1/32 area 0.0.0.0
   network 10.100.1.1/32 area 0.0.0.0
   max-lsa 12000
```

Hewlett Packard Enterprise

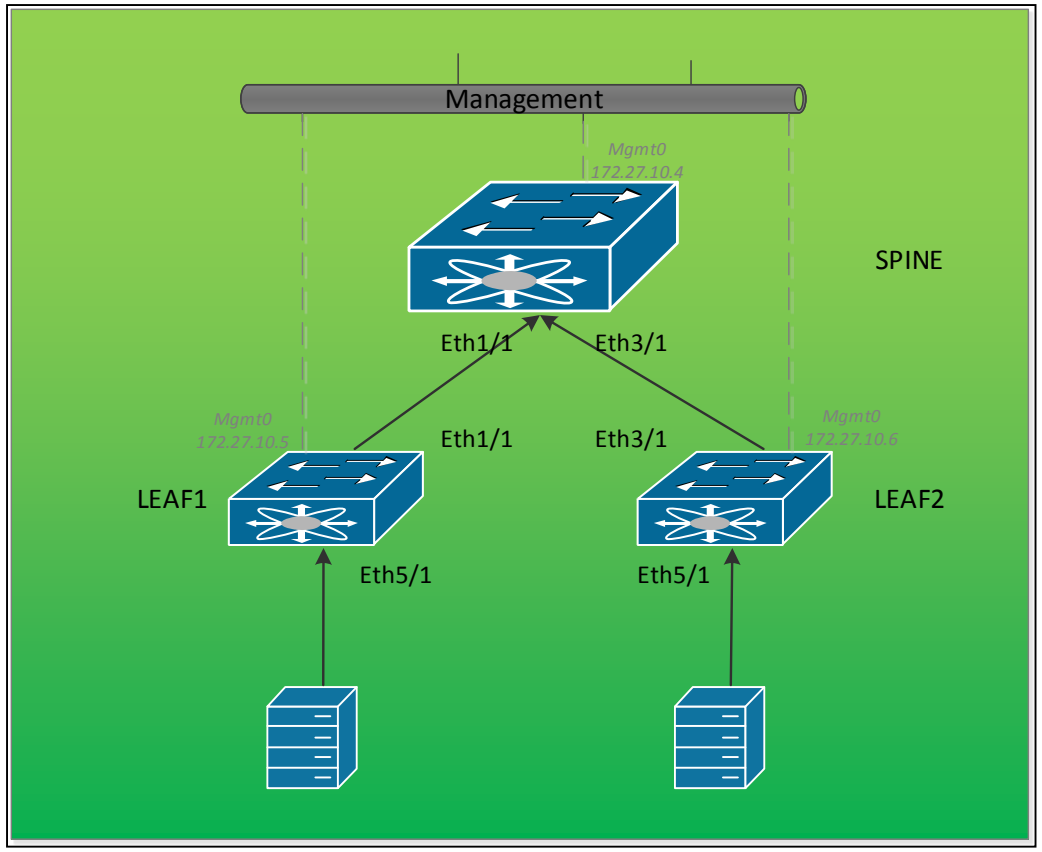
VXLAN – EVPN (eBGP) Control Plane



Hewlett Packard Enterprise

Felhasználói tesztek – Arista eszközökkel – EVPN Control Plane

(Kákonyi István, Valkó László, Gelsi Tamás, Zeisel Tamás)

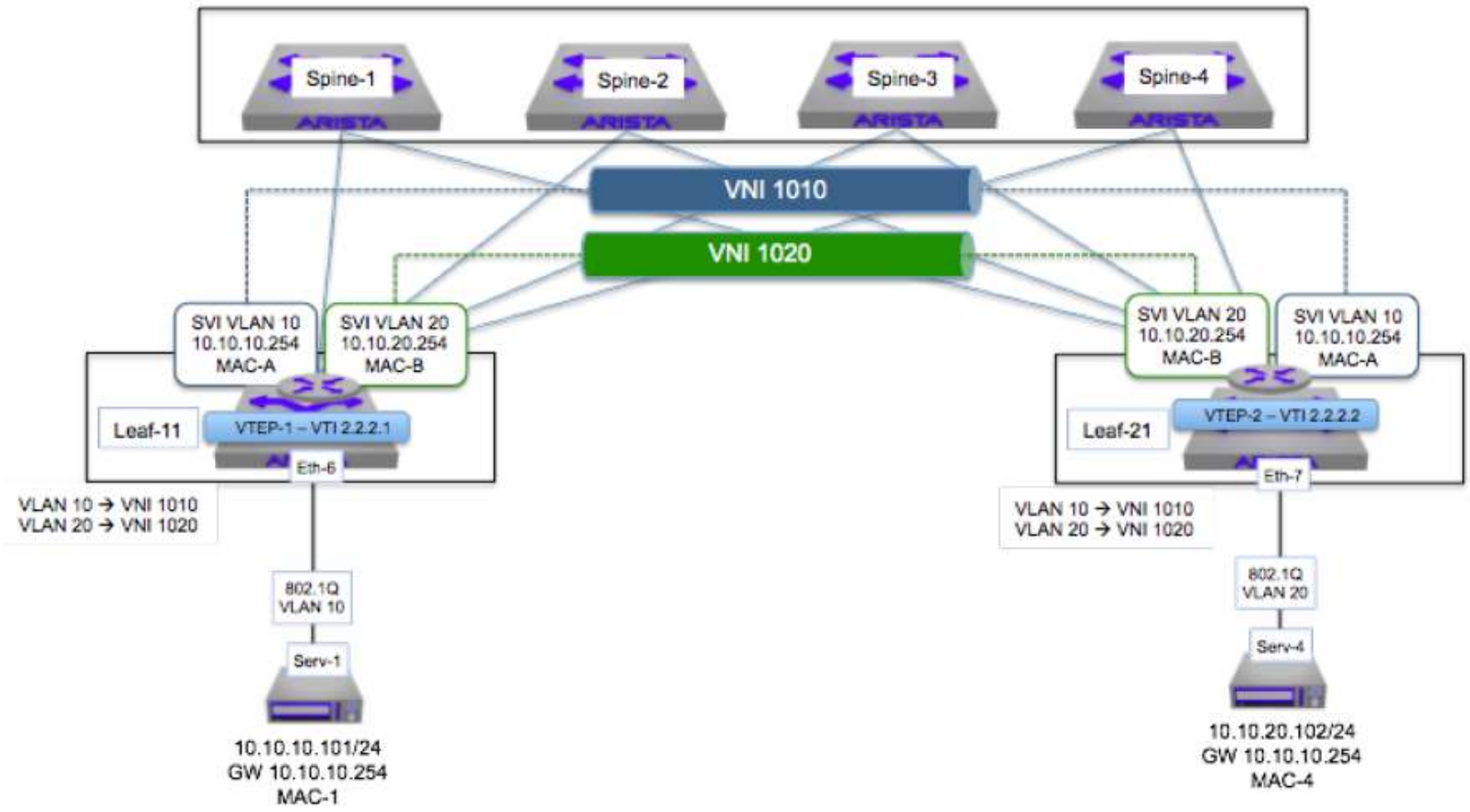


….

```
!
interface Vxlan1
   vxlan source-interface Loopback0
   vxlan udp-port 4789
   vxlan vlan 100 vni 100
!
router bgp 65001
   router-id 10.10.1.1
   neighbor 10.10.0.2 remote-as 65000
   neighbor 10.10.0.2 send-community extended
   neighbor 10.10.0.2 maximum-routes 12000
   !
   vlan 100
      rd 100:100
      route-target both 100:100
      redistribute learned
   !
   address-family evpn
      neighbor 10.10.0.2 activate
   !
   address-family ipv4
      neighbor 10.10.0.2 activate
      network 10.10.1.1/32
```

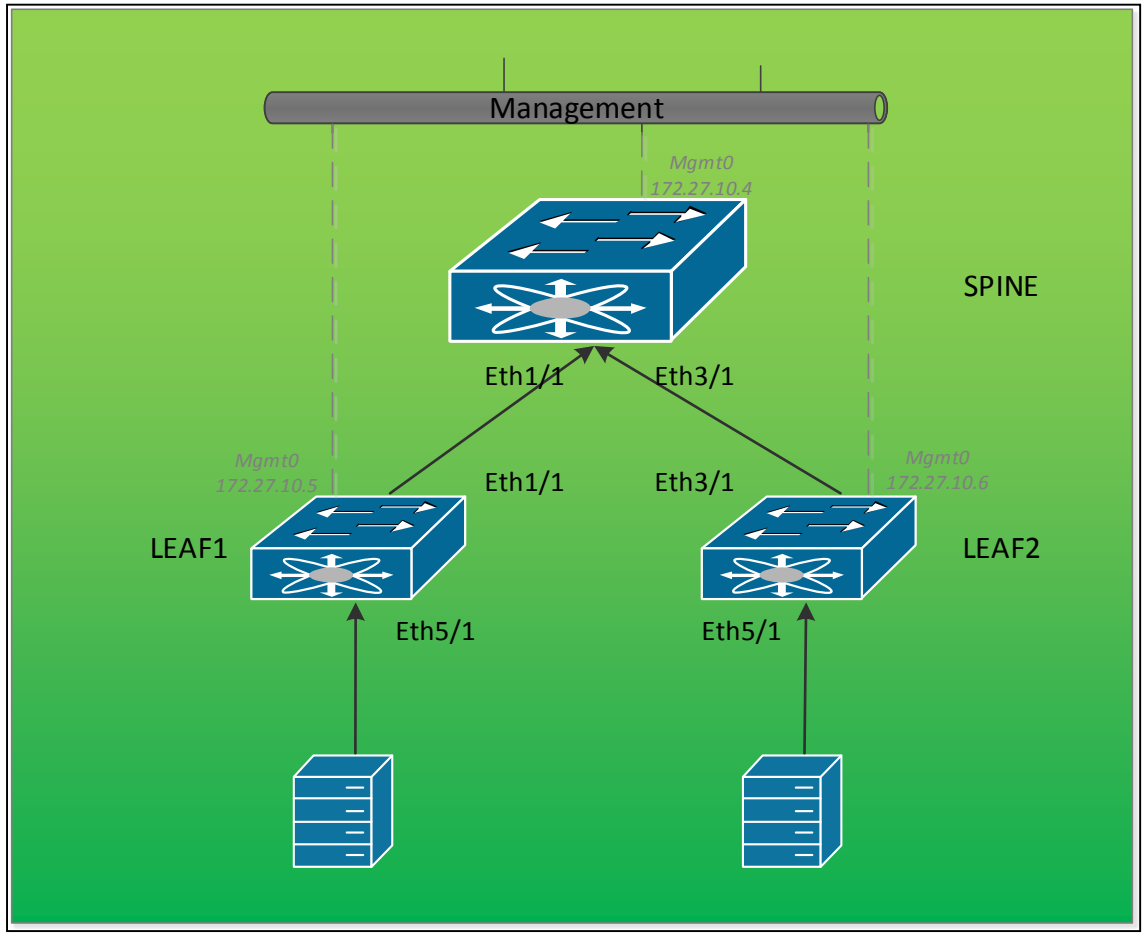
Hewlett Packard Enterprise

További kihívások VXLAN Környezetben
IRB – Integrated Routing and Bridging



Hewlett Packard Enterprise

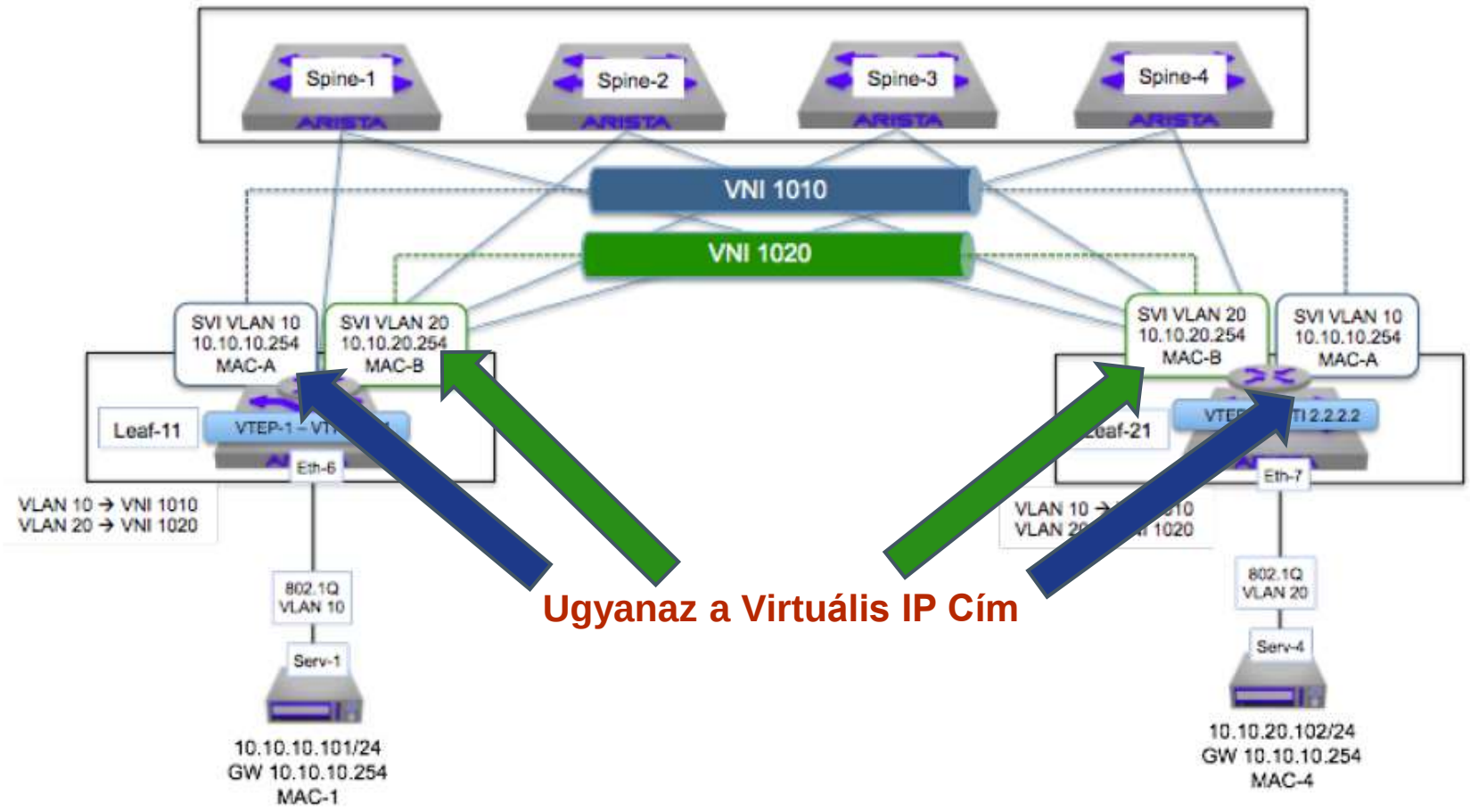
További kihívások VXLAN Környezetben
IRB – Integrated Routing and Bridging




```
interface Vlan100
  mtu 9214
  no autostate
  ip address 192.168.32.1/24
!
interface Vlan200
  mtu 9214
  no autostate
  ip address 192.168.42.1/24
!


interface Vxlan1
  vxlan source-interface Loopback0
  vxlan udp-port 4789
  vxlan vlan 100 vni 100
  vxlan vlan 200 vni 200


!
interface Recirc-Channel100
  no switchport
  switchport recirculation features vxlan

interface Ethernet 4/1
  traffic-loopback source system device mac
  channel-group recirculation 1
```

Hewlett Packard Enterprise

További kihívások VXLAN Környezetben
Routing – Anycast router, DCI



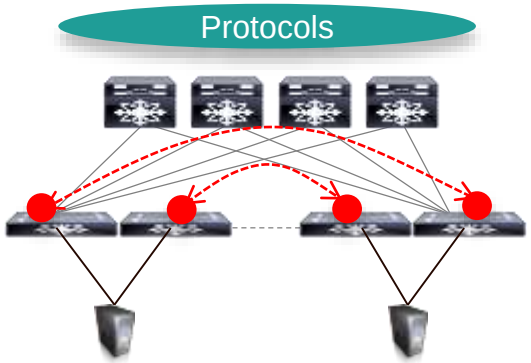
Hewlett Packard Enterprise

Fizikai és Virtuális világ összekapcsolása VXLAN segítségével

Hewlett Packard
Enterprise

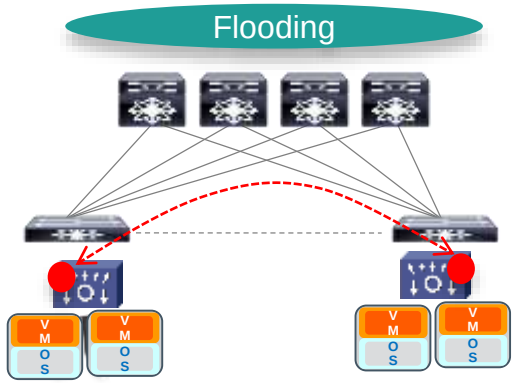
VXLAN Tunnel indítás/végződtetés lehetőségei

Hálózati eszköz



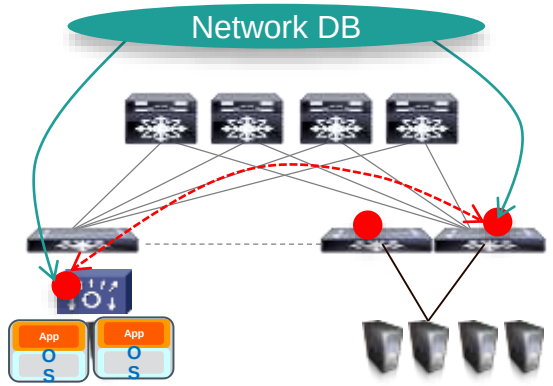
- Router/switch tunnel végpont
- Hagyományos VPN megoldás

Host alapú



- Virtuális végpontok
- Virtualizációs platform

Hybrid megoldás

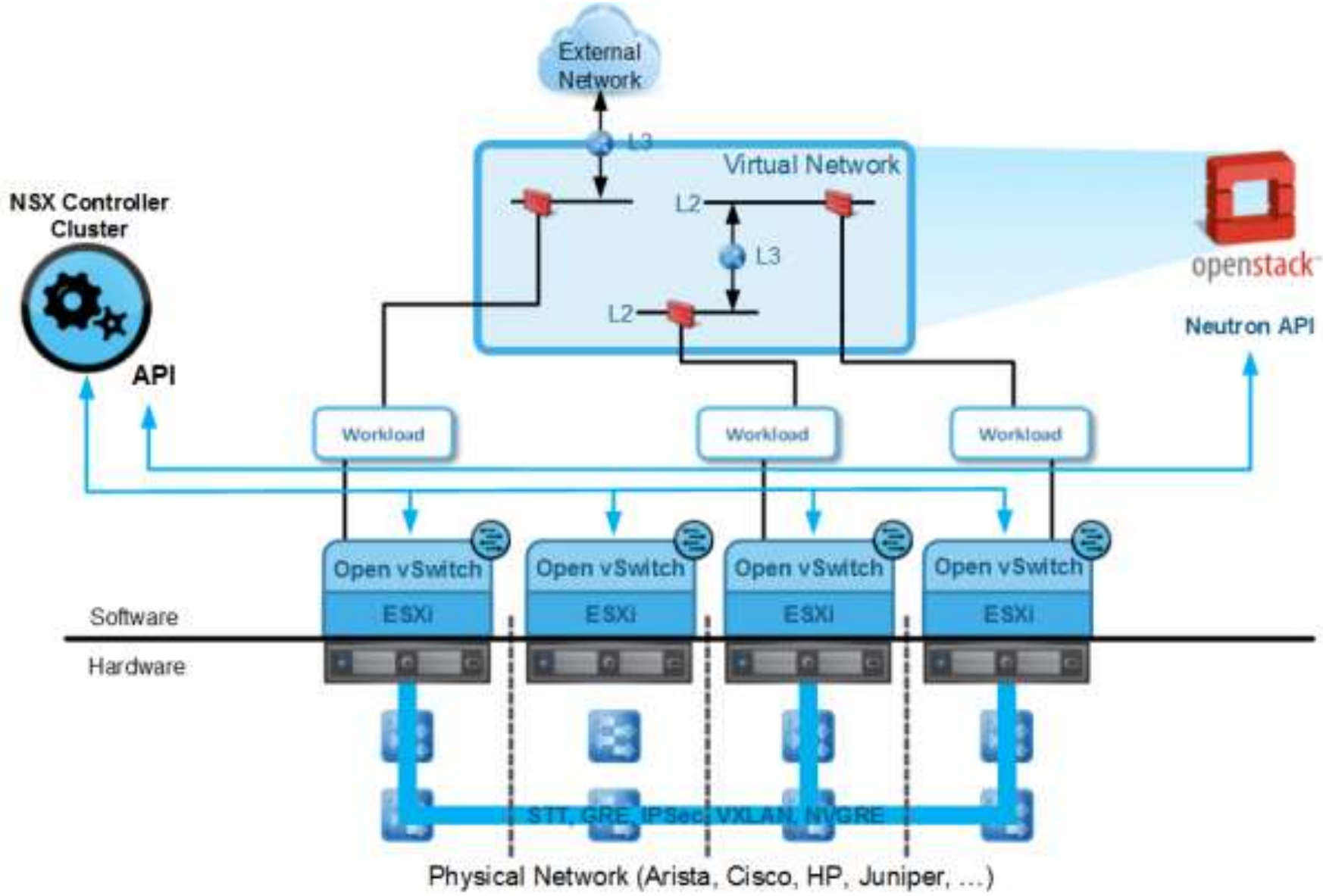


- Fizikai és virtuális végpontok
- Open Standard megoldás

VXLAN Tunnel End-points

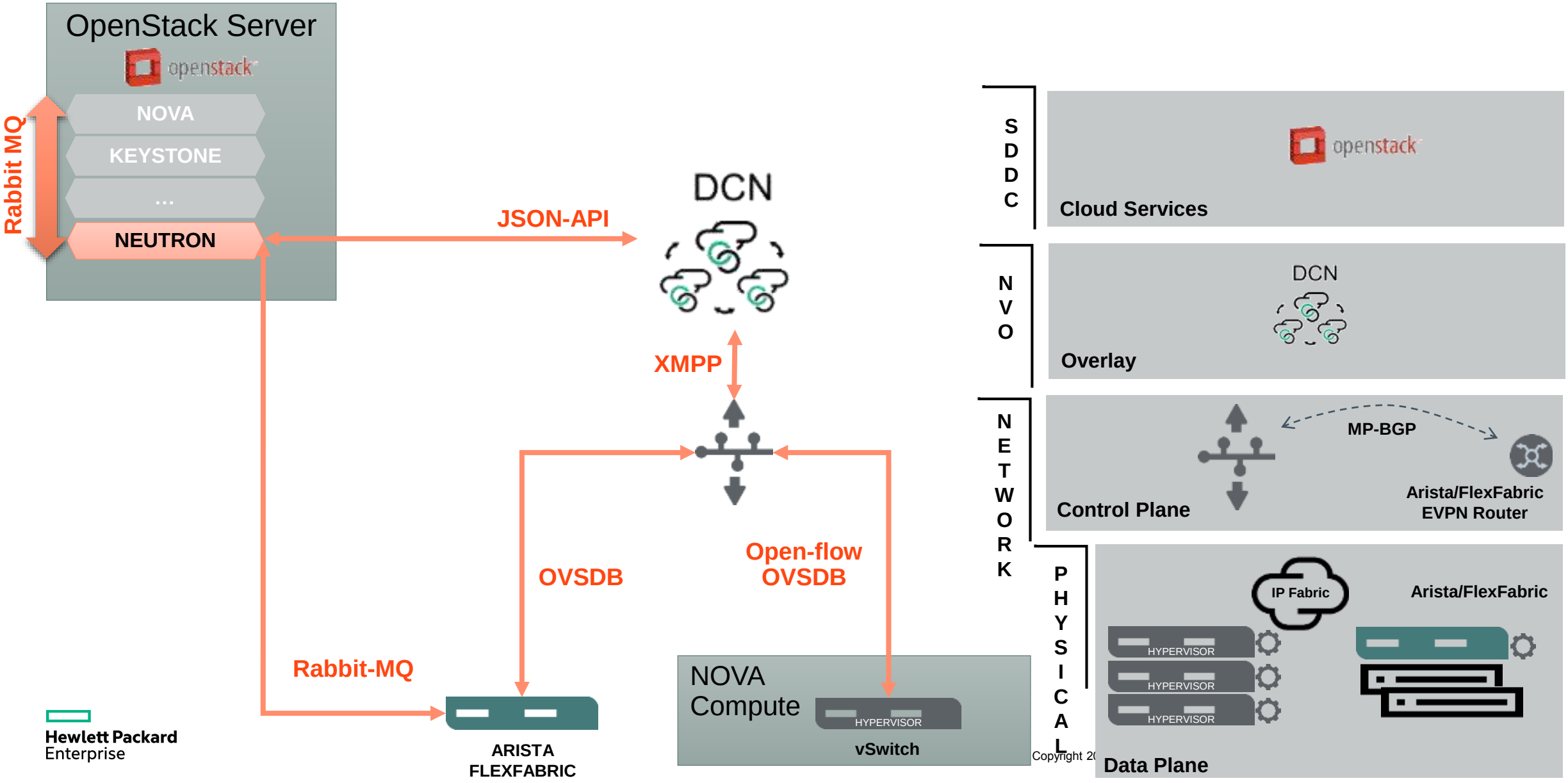
Hewlett Packard Enterprise

VMWare NSX/OpenStack architektúra



Hewlett Packard Enterprise

Hibrid megoldás - OpenStack integráció



Hewlett Packard Enterprise

Technológiát látjuk

Mi legyen az eszköz kiválasztás szempontja?

Hewlett Packard
Enterprise

Eszköz kiválasztás szempontjai

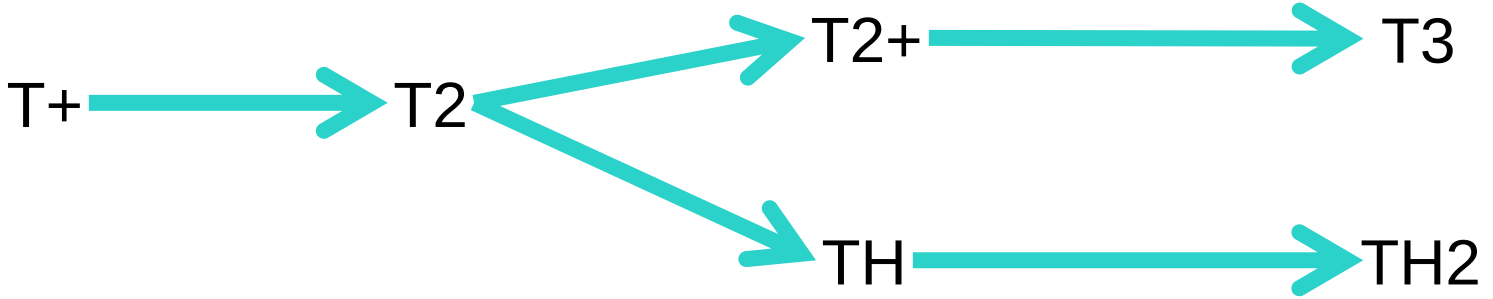
- **Spine: port száma meghatározza a Rackek számát!!**
  - **Tipikusan 32 portos, de van gyártó akinek van 64 portos eszköze**
  - **Mi legyen a gerinc sávszélessége 40G vagy 100G**

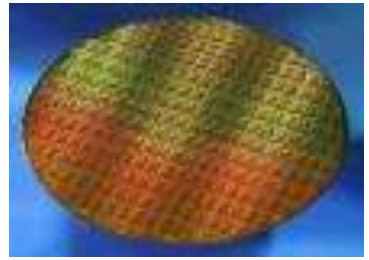
- **Leaf: portszáma meghatározza a rackben a szerverszámot**
  - **Tipikusan 48 port + 4-6 40G vagy 100G uplink**
  - **Mi legyen a szerver csatlakozás sávszélessége**
    100G, 40G, 25G, 10G

Hewlett Packard Enterprise

Eszköz kiválasztás szempontjai

- ASIC továbbra is fontos de sokszor csak a generáció





64x10G 32x40G 32x100G 64x100G

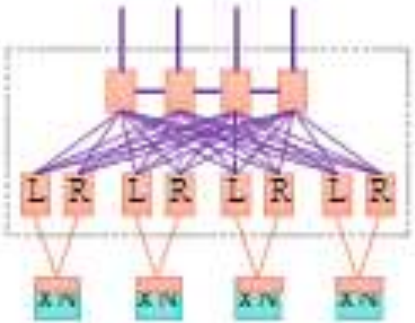
- Deep Buffer – különösen Software Defined storage és Video stream alkalmazás esetén:

8x10G 6x40G 6x100G 18x100G

High Level ASIC Comparison – Leaf/Spine (Pizza Box)

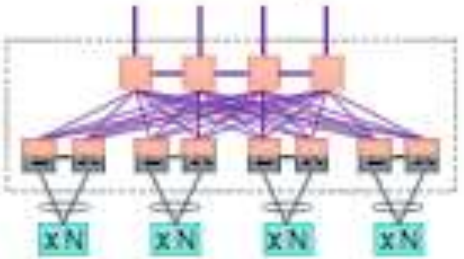
	Trident3	Tomahawk2	Mellanox Spectrum/Spectrum2	Cavium Xpliant
Bandwidth	Up to 3.2 Tb/s	Up to 6.4 Tb/s	Up to 3.2 Tb/s / Up to 6.4 Tb/s	Up to 3.2 Tb/s
Serdes	128 * 25G	256 * 25G	128 * 25G / 256 * 25G	128 * 35G
Typical form factor	48 * 25G + 8 * 100G 32 * 100G	64 * 100G	32 * 100G	32 * 100G 128 * 25G
L2 MAC	288K	264K	256K / 512K	320K
MPLS	Yes	Yes	Yes	Yes
L3 Hosts (IPv4/Ipv6)	168K	146K/78K	256K/128K / 512K / 256K	160K / 80K
Next Hop	64K	48K	TBD	16K
LPM IPv4/IPv6 (MAX)	350K	320K/168K	256K/128K / 512K/256K	164K / 16K L3 mode up to 1M IPv4 prefixes
ECMP Paths	TBD	32K	4K / 4K	
NAT/NATPT	TBD	2K entries/1K sessions	NA	Programmable
Tunnels	16K	32K MPLS / 8K VNIDs	16K / 16K	16K
VXLAN	L2/L3 IPv4	Packet recirculation	L2/L3 Ipv4	L2/L3 Ipv4
VXLAN IPv6	L2/L3 IPv6	Packet recirculation	TBD	Programmable
ACLs	4*6K Ingress	4*6K Ingress/4*1K egress	36K / 512K	Variable
Buffering	32MB	42MB	16MB (fully shared)	24MB (fully shared)
Queues	TBD	10 UC + 10MC	TBD	4K queues/24 queues per port

L3 Fabric

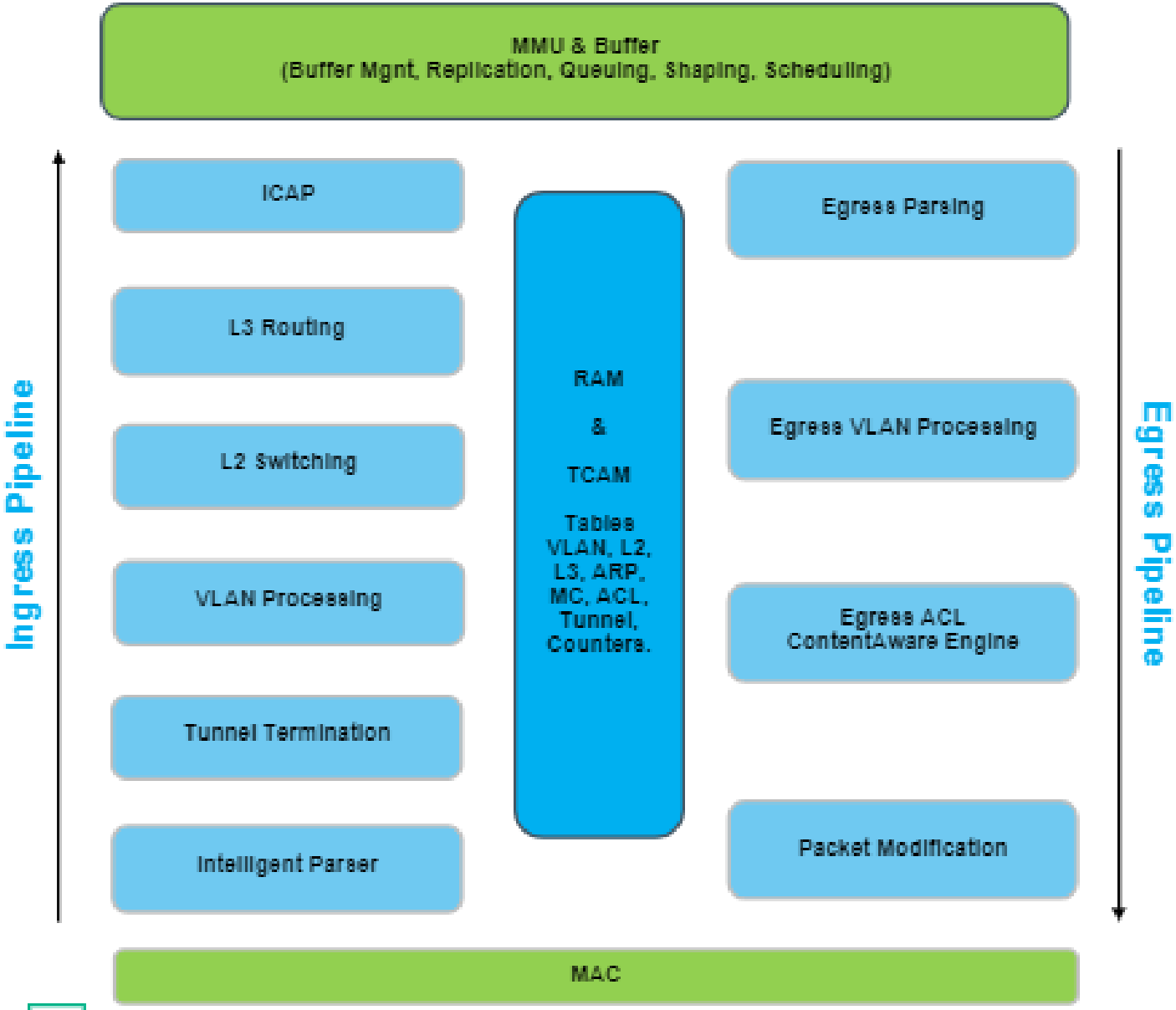


#1 BGPv4/ECMP/BFD
#2 OSPFv2/ECMP/BFD

Overlay



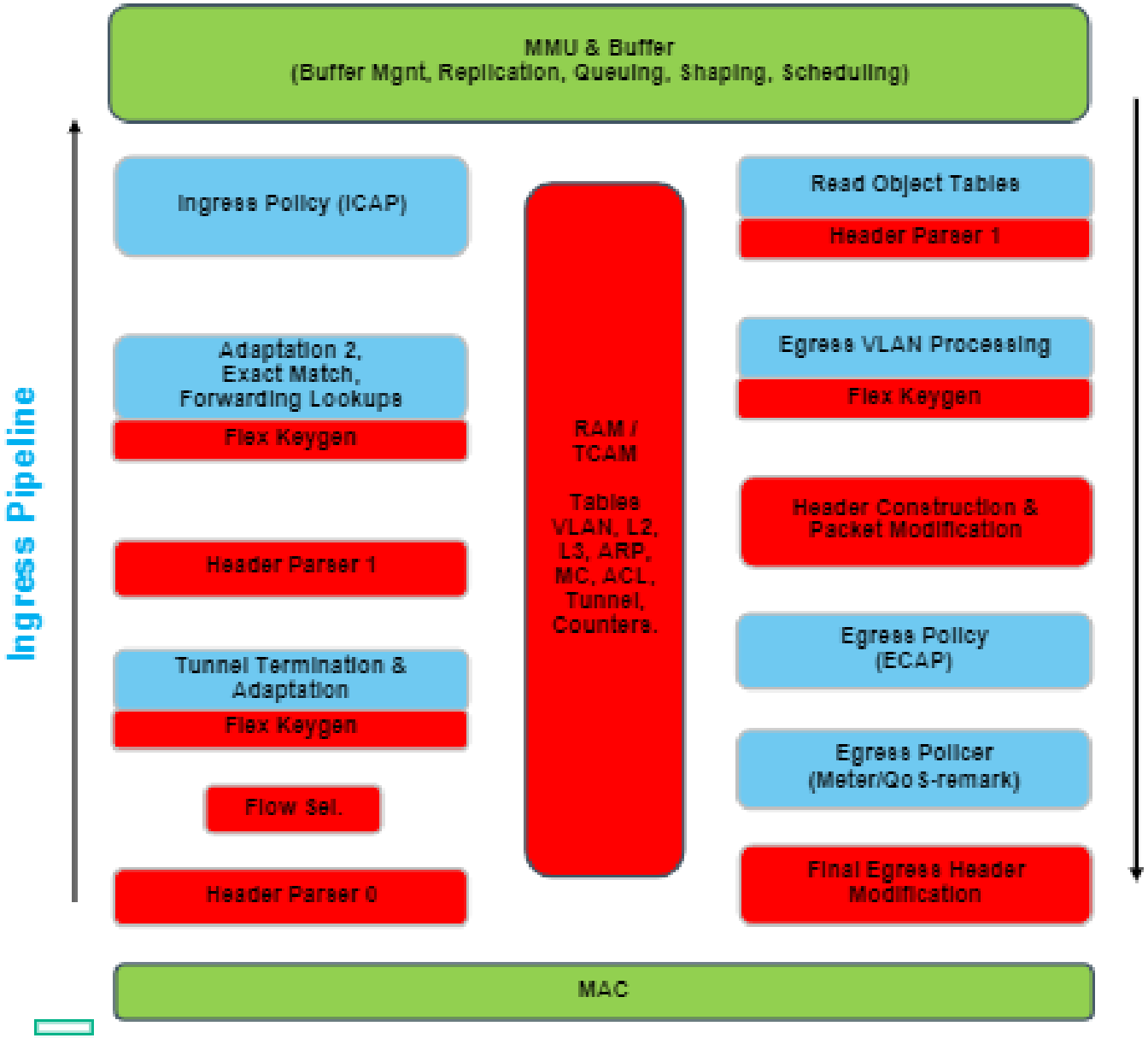
#1 BGP EVPN



BRCM TOMAHAWK ("TH")

- ✓ 16nm
- ✓ 6.4T, 256*25G/64 * 100G and 3.6 as 144 * 25G
- ✓ Throughput: 4.2 BPPS (Line rate @ 250 Bytes)
- ✓ 4 Pipelines, each pipeline supports 16 cores
- ✓ Each Core has 4 x 10G~25G SERDES
- ✓ 42 MB buffer (vs TH 6MB)
- ✓ Each egress port consists of 10 UC queues & 10 MC queues. Four-level hierarchical scheduling.
- ✓ **ECMP Dynamic Flow Distribution - New**
- ✓ UFT: 256K/ECMP (Groups/Members): 32K/16K/Next Hop: 48K
- ✓ Broadview v2 (network congestion, 1588 time stamping, etc..)
- ✓ **TH2 BW capacity: 2 * TH**
- ✓ **TH2 scalability: 2 * TH**
- ✓ **Foundation for next step – TH3 (12.8 Tbps), 32 * 400Gb/s with native 50G PAM4, target 40% less power than TH2 (350W for 32 * 400G)**
- ✓ **Improvements in scale, DLB/DGM (ECMP)**
- ✓ **Programmability: no**

Primary target: Hyperscalers in (desperate !!!) need of BW



BRCM TRIDENT3 ("TD3")

- TD3 BW capacity 2xTD2+ (3.2T/2.0T/1.08T versions)
- Up to 128*25G/64*50G/32*100G
- TD3 scalability~ 2x TD2+.
- 3.2T – 32x100G / 128x25G; 2.03 BPPS. 2 Pipelines.
- Bring native 25G to the Trident family
- 32MB Fully-Shared on-chip buffer. (TD2+ 16MB)
- Enhanced table sharing & table sizes: UFT for L2 & L3; UAT (Unified Adaptation Table) for MPLS. Significantly Enhanced Tables Scaling
- Enhanced MPLS Segment Routing
- Programmability
- DLB (Dynamic Load Balancing) on HG/LAG/ECMP: Link selection based on real-time link loading.
- BroadView v2 Instrumentation / Telemetry
- Enhanced Control Plane (on-chip 4-core 64-bit ARM CPU) with PCIe Gen3 and 2*10G for Mgmt
- 160W@3.2T

"Programmability"

- Flexible Packet Parsers
- Flexible Packet Editor
- Flexible Lookup Key Generation for major tables

Programmable engines

Primary target: Large Enterprise

Eszköz kiválasztás szempontjai

- Skálázhatóság:

  vni number – talán az egyik legfontosabb paraméter
  Szokásos paraméterek: MAC, Routing tábla méret, BGP neighbor szám stb.

- Késleltetés!

  400-500nsec vagy 2-4μsec – Cut through, Store and Forward kapcsolás

  Packet Buffer méret 16-20MByte vagy 200-400MByte

- Menedzsment:

  Operációs rendszer nyitottsága – mennyire Linux

  CLI, Automatizálási lehetőségek (python, REST API, Chef, Puppet Ansible)

  Gyártó SDN rendszere – szabványosság

  OpenStack, NSX integráció

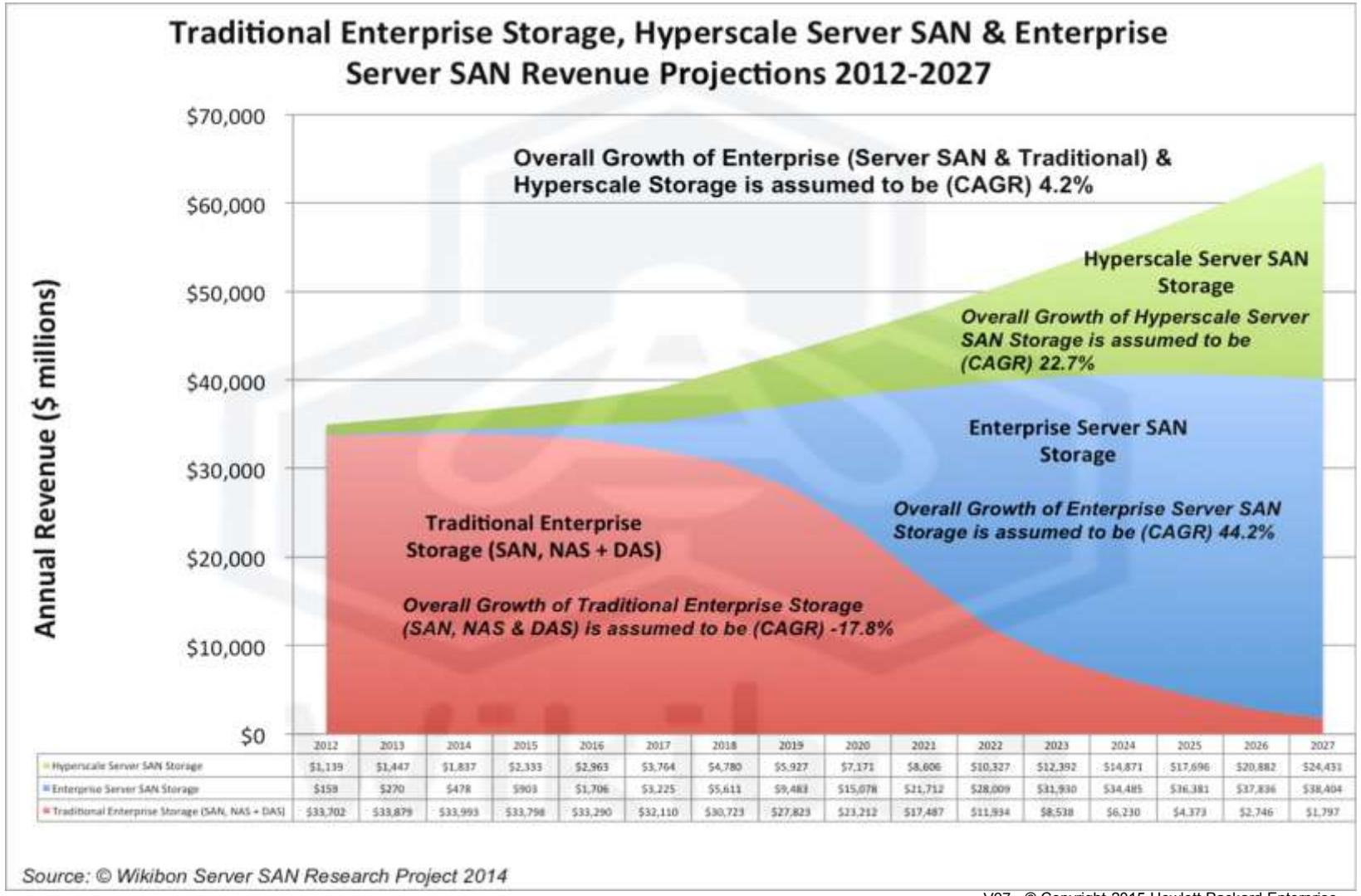
  Telemetria!

Mi a helyzet a régen annyit emlegetett Nagyvállalati technológiákkal?

Hewlett Packard
Enterprise

Hálózati Storage technológiák – Tároló/Adat hálózat integráció

- Mi a helyzet a tároló technológiával?



http://wikibon.org/wiki/v/The_Rise_of_Server_SAN

Hálózati Storage technológiák – Tároló/Adat hálózat integráció

- Mi a helyzet az FCoE-vel?

RHAT Linux 8 előzetes információk:

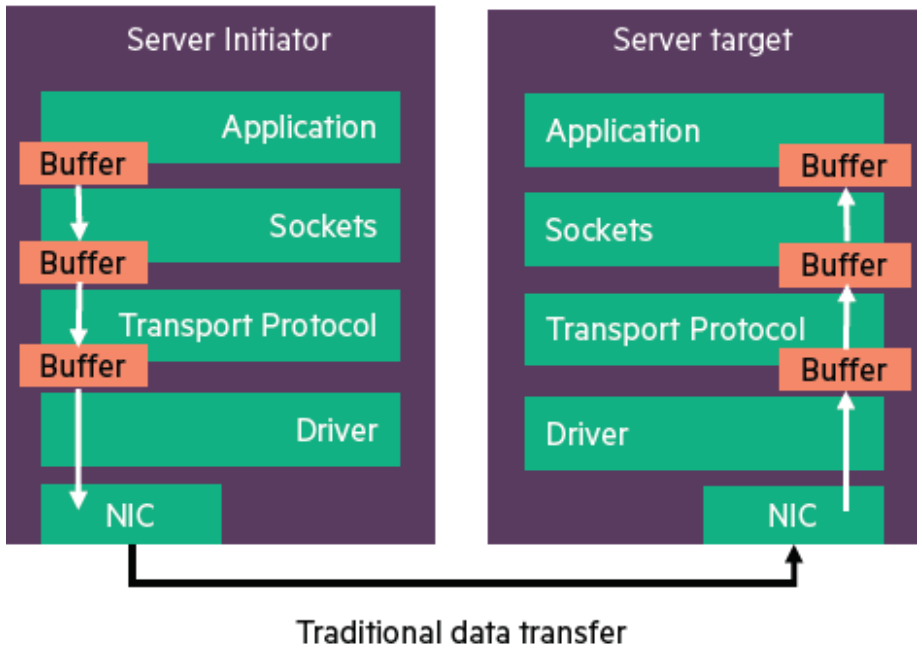
„FCoE storage technologies have been deprecated

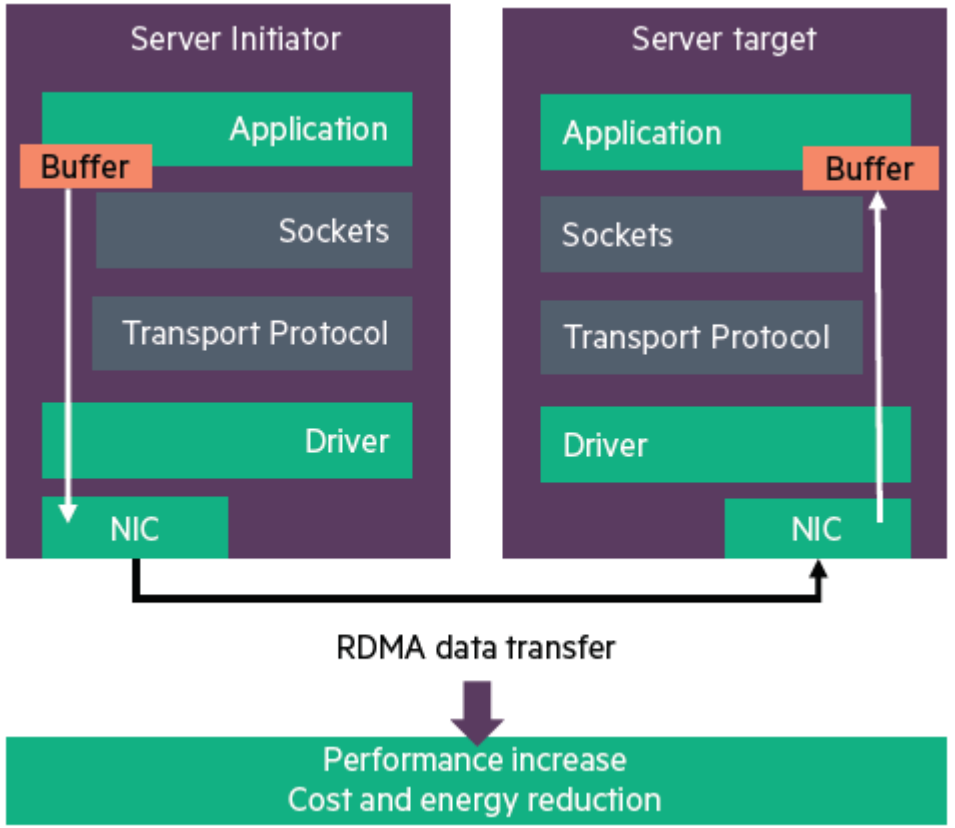
The Fibre Channel over Ethernet (FCoE) storage technologies have been deprecated due to limited customer adoption after several years of availability. FCoE storage technology will remain supported for the life of Red Hat Enterprise Linux 7. The deprecation notice indicates intent to remove FCoE support in a future major release of Red Hat Enterprise Linux.”

https://access.redhat.com/documentation/en-us/red_hat_enterprise_linux/7/html/7.4_release_notes/chap-red_hat_enterprise_linux-7.4_release_notes-deprecated_functionality#idm140026759655120

Hewlett Packard Enterprise

Hálózati Storage technológiák – Technológiák Software Defined Storage környezetben RDMA





Hewlett Packard Enterprise

Hálózati Storage technológiák – Technológiák Software Defined Storage környezetben RDMA

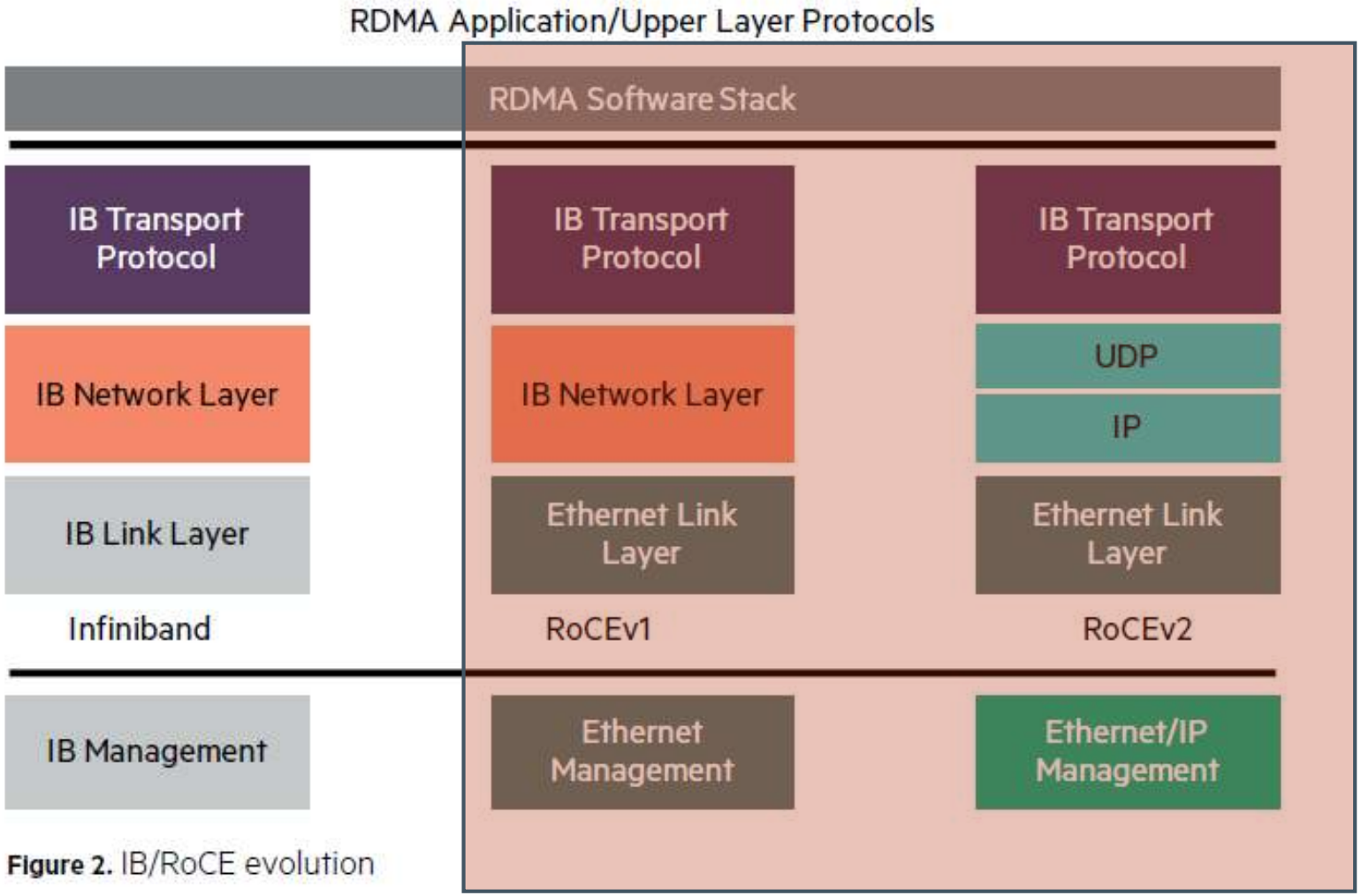


Figure 2. IB/RoCE evolution

Hálózati Storage technológiák – Technológiák Software Defined Storage környezetben RDMA-RoCE előnyök

- Hatékonyabb átvitel sokkal alacsonyabb késleltetés:

  Pl. VMWare:

  - Teljes vMotion forgalom 36%-kal gyorsabb.
  - 30% nagyobb másolási képesség.
  - 84% - 92%-kal alacsonyabb CPU terhelés.

Hálózati Storage technológiák – Technológiák Software Defined Storage környezetben RoCE megvalósítás

- RoCE alapja a veszteségmentes Ethernet DCB

Feature	RoCEv1	RoCEv2
Priority-based Flow Control	Yes	Yes
Enhanced Transmission Selection (ETS)	Yes	Yes
Data Center Bridging Exchange (DCBX)	Yes	Yes
Quantized Congestion Notification (QCN)	Yes	No
IP Explicit Congestion Notification (ECN)	No	Yes

Hewlett Packard
Enterprise

Köszönjük a figyelmet!