Cisco Fabric Technologiák Fejlődése

Zeisel Tamás
Konzultáns Rendszermérnök

HBONE Tábor 2013

Mi indokolja a Hálózati Fabric Technológiák megjelenését

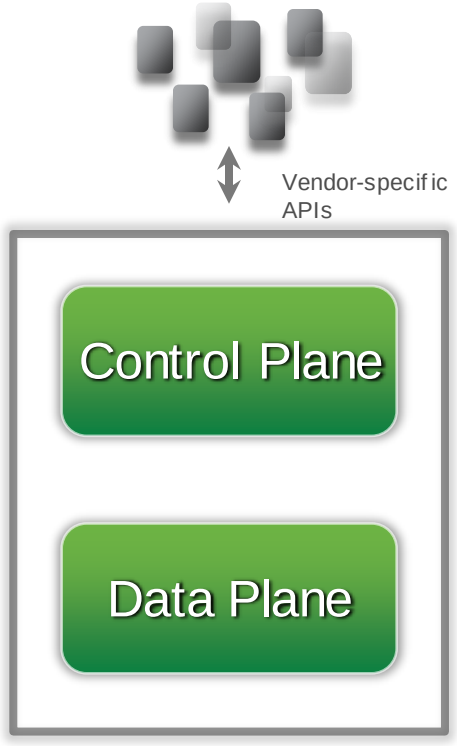
- Méretnövekedés – Cloud
  - Skálázhatóság (VLAN), Bővíthetőség
  - Többfelhasználóós Multi-Tenant környezet
  - Automatizálás igénye
- Virtualizáció előtérbe kerülése
  - Fizikai és Virtuális környezet egységes kezelése
- Egyszerüsítés
  - Kevesebb szintű architektúra
- Controll Plane Data Plane szeparálásának ígénye
  - SDN technológia Megjelenése az iparágban
  - Egységes Dataplane Fabric

Cisco és iparági Fabric megoldások a felmerült ígényekre

- Control Plane Fabric - SDN – onePK Cisco Open Network Environment Platform Kit
- FabricPath technológia továbbfejlődése
  - Dynamic Fabric Automation DFA
- VXLAN Technológia
- Alkalmazás központú Hálózati Fabric
  - Application Centric Infrastructure ACI

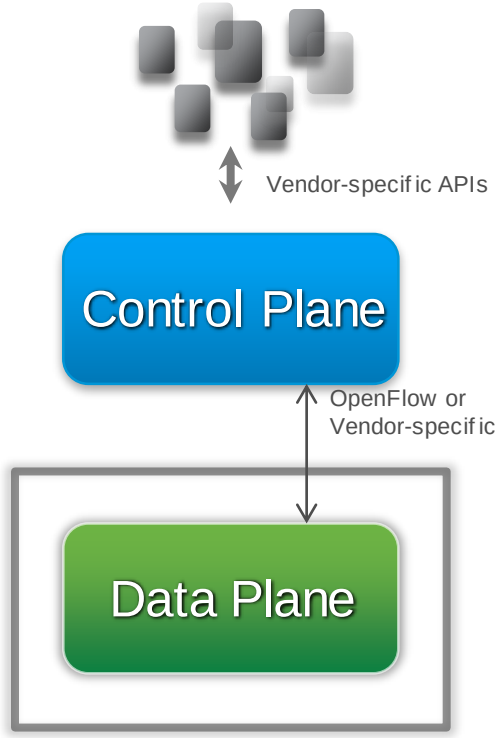
Controll Plane Fabric - SDN

Jelenlegi switch/router



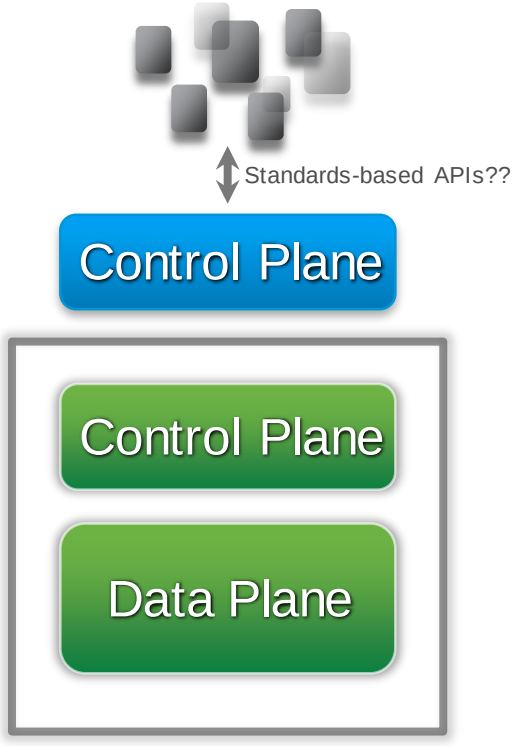
Megbízható, Skálázható

"SDN" Megoldás



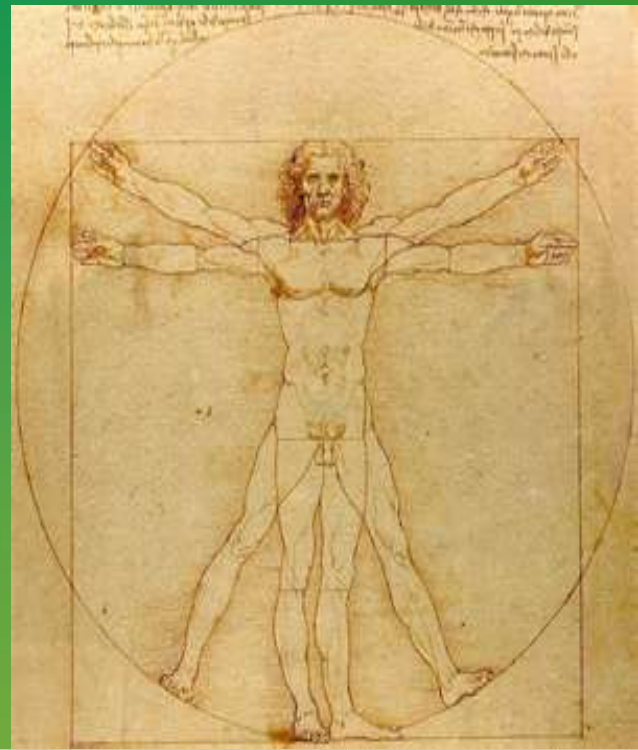
Egyszerű (kevesebb menedzsment pont) Centralizált topológia

Hibrid Model onePK



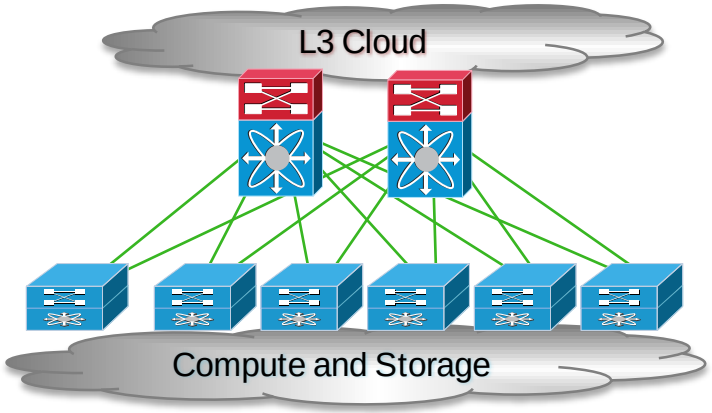
Optimális megoldás

Dynamic Fabric Automation (DFA)

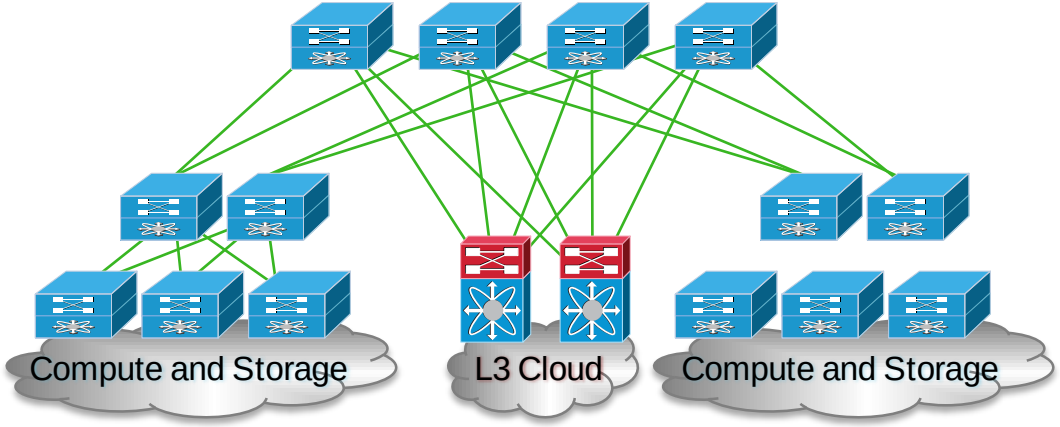


Egyszerüsített optimalizált topológia

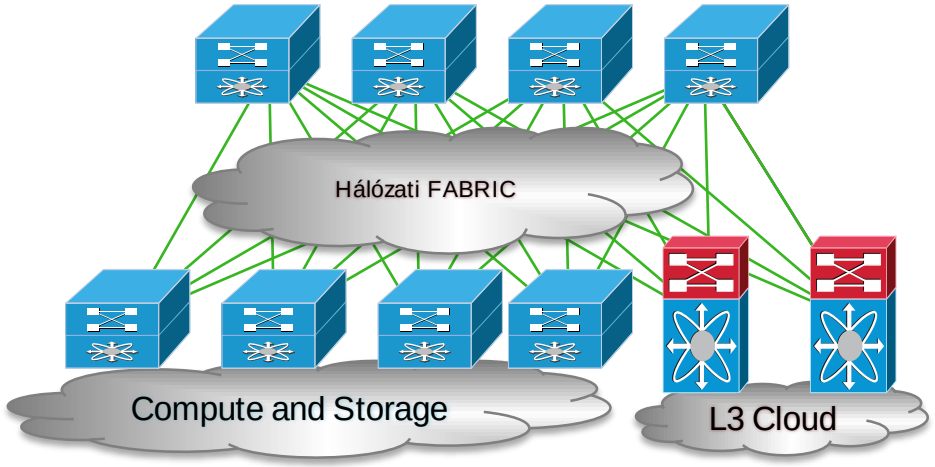
Hagyományosl Access/Aggregation



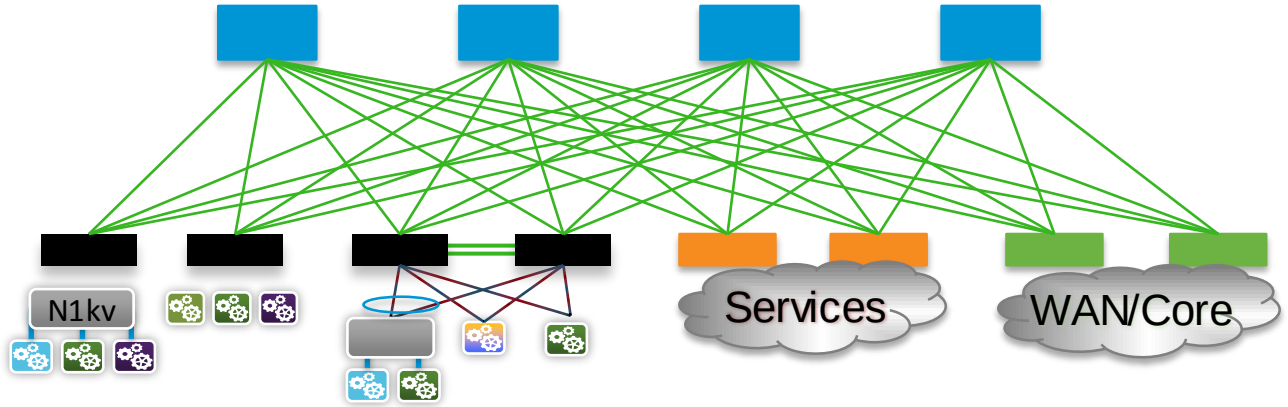
Három rétegű topológia



Két szintű
Spine - Leaf



Dynamic Fabric Automation Architektúra
Csomópontok fajtái



- Spine
- Leaf
- Border Leaf
- Szerviz Leaf
- Virtual Leaf

N1kv

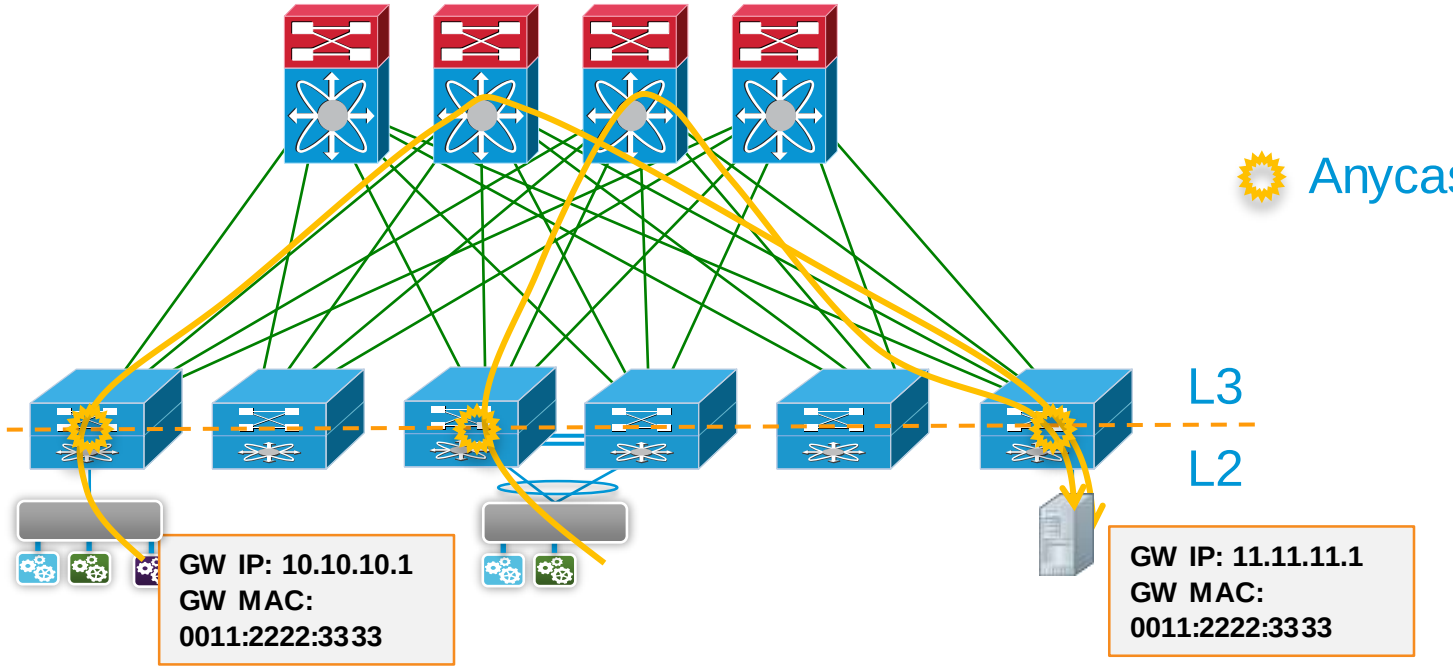
Virtual Machine-ok
Fizikai eszközök
FEX-ek
3rd Party Switchek
UCS
Blade Switchek
Storage-ok

Tűzfalak
Load Balancerek
3rd Party Szerviz eszk.

Routerek
Switcheek
3rd Party eszközök

Különböző leaf szerepek kizárólag logikai megkülönböztetést takarnak

Optimalizált Hálózat
Elosztott Gateway a Leaf node-okban

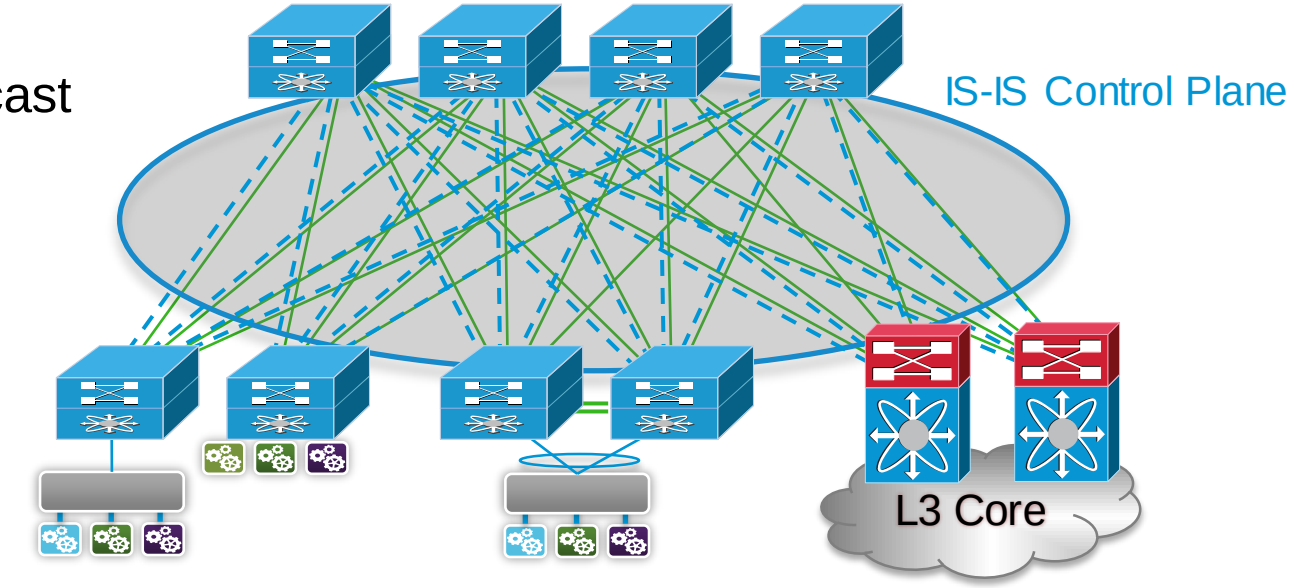


- Bármely subnet bárhol => Bármely leaf bármely subnetnek része lehet
  - Minden leaf elosztott gateway funkcióval rendelkezik bármely subnetben megosztva az IP és MAC címeket (Nincs HSRP)
  - ARP üzenetek a leaf node-okon terminálódnak, No Flooding
- Illeszkedik a VM mozgatásához
- Egységes kommunikáció a fizikai és virtuális gépek között mind L2 mind L3 -ban

Control Plane
1 - IS-IS Fabric Control Plane protokoll

ISIS biztosítja a fabric link state disztribúcióját

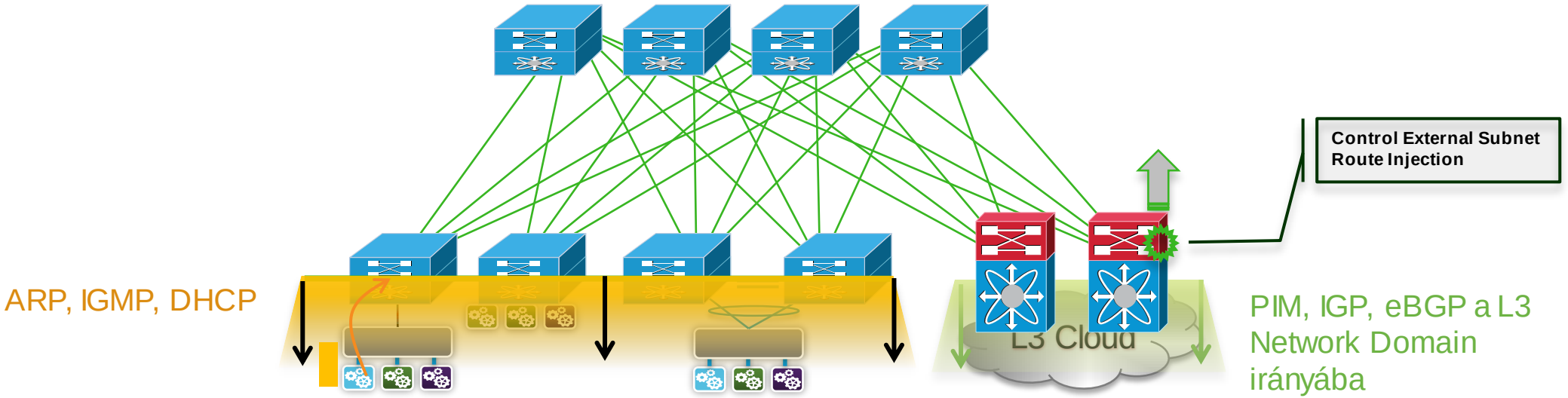
- Fabric node elérhetőség
- Multi-destination fa a multicast és broadcast forgalom kezelésére
- Gyors átállás link/node kiesés esetén



Fabric Control Protocol nem kezeli

- Host Route-okat
- Host által indított vezérlő forgalmat
- Szerver subnet információkat

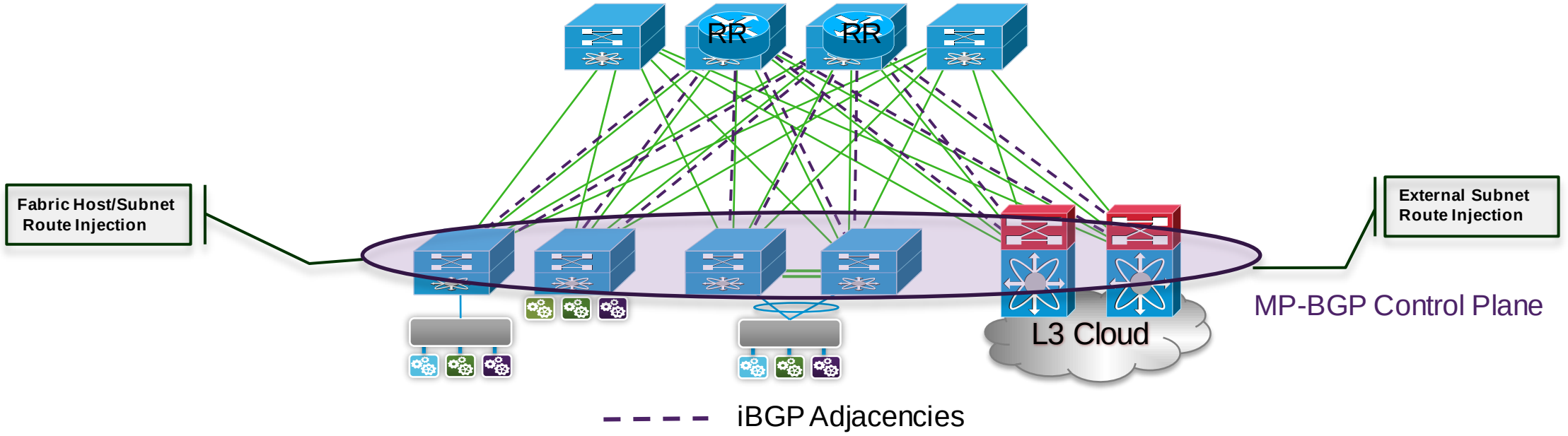
Control Plane
2 – Host Protokollok kezelése



- A szerverek által indított ARP, IGMP forgalom a Leaf node-okon terminálódik
- Külső hálózati PIM, OSPF, eBGP a Border Leaf node-okon terminálódik

Control Plane
3 – Host és Subnet Route disztribúció

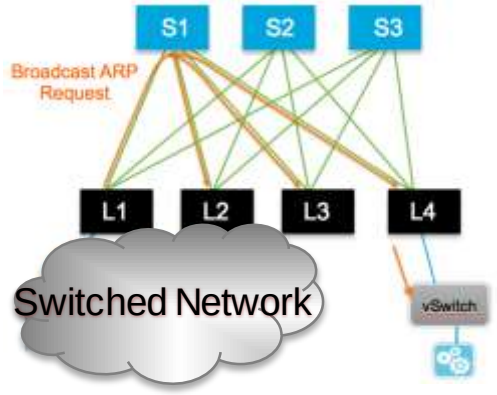
Route-Reflectorok skálázási okokból



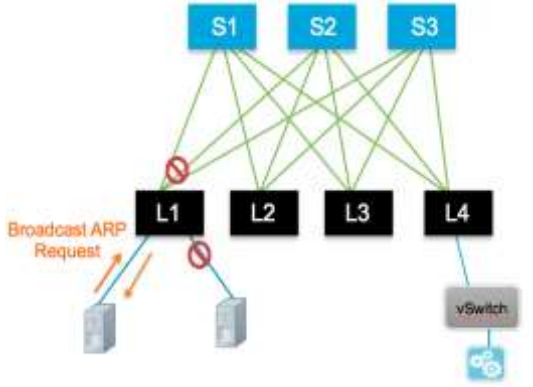
- Host Route információ disztributálása elkülönül a Fabric link state protokoltól
- MP-BGP protokolt használ a leaf node-okon a belső host/subnet route és külső elérhetőség információ disztributálásra
- MP-BGP továbbfejlesztése biztosítja a százeres route méretet és konvergenia idő csökkenést

Data Plane
Csomagtovábbítási üzemmódok

- Anycast-Gateway a javasolt:
  - ✓ Hagyományos switched infrastruktúrához történő kapcsolódás esetén
  - ✓ “silent host” detektálás szükséges



- Proxy-Gateway a javasolt:
  - ✓ Optimalizált továbbítási esetben

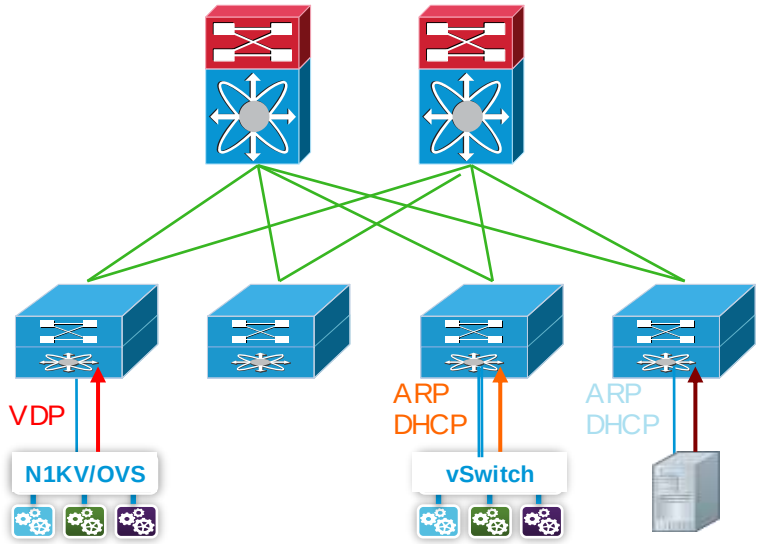


Control Plane – Proxy Gateway Host Felfedezés és törlés

- Proxy-gateway üzemmódban a leaf node-ok fedezik fel a rájuk kapcsolódó fizikai és virtuális végpontokat

- Kapcsolódó Host felderítés

  Control plane aktivitást kíván!!! → ARP/DHCP csomag tartalmazza a host IP címét virtuális host esetén az Edge Virtual Bridging (802.1Qbg) VSI Discovery and Configuration Protocol (VDP) alkalmazása segít a felderítésben



- Távoli host felderítés

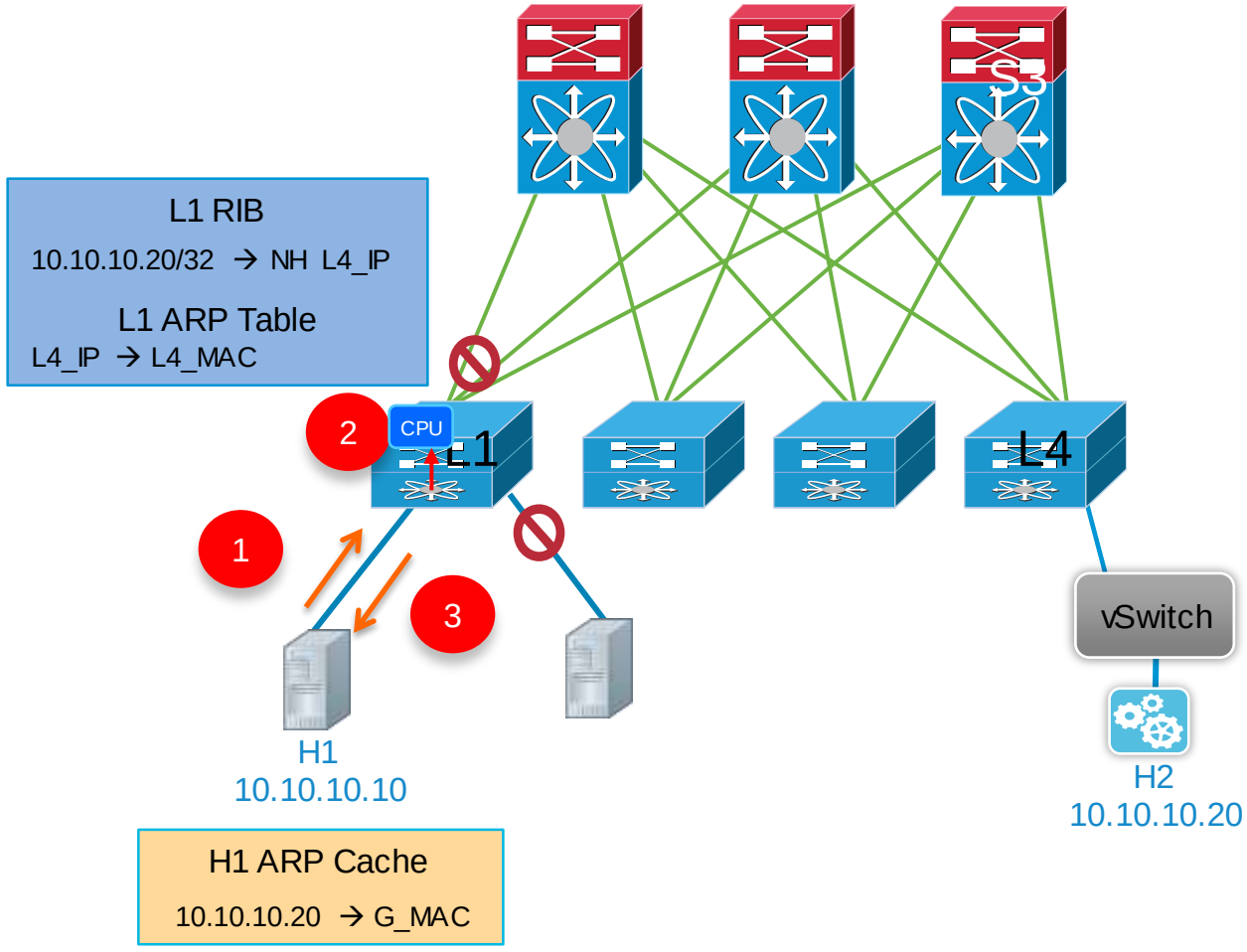
  MP-BGP protokol információ alapján kerül a route az Unicast RIB (URIB) táblába → (FIB HW tábla programozása további üzemmód - L3 conversational learning - függő

- Kapcsolódó és távoli host információ törlés

  Inactivity timer illetve explicit protokol üzenet alapon (VDP értesítés)

IP csomagtovábbítás azonos Subneten belül Proxy-gateway üzemmód

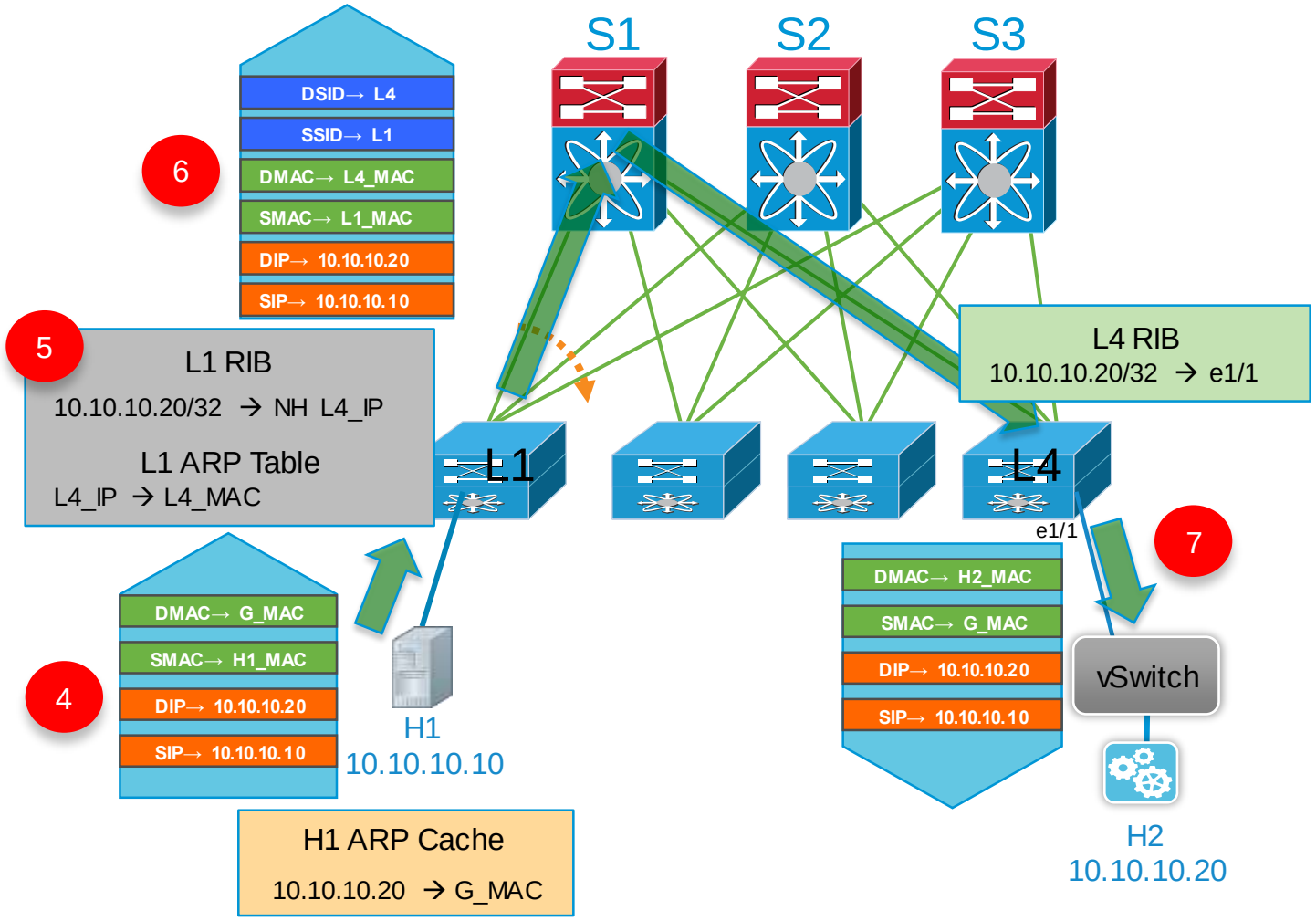
1. H1 ARP kérést küld a H2 –10.10.10.20 – re
2. ARP –ot a Leaf1 lekezeli és átadja a Supervisorának
3. Feltéve, hogy az Leaf1 RIB táblájában szerepel a H2 irányába mutató route a Leaf1 ARP választ küld a saját G_MAC fizikai címmével és a H1 felépíti a saját ARP cache-t



Fontos: Proxy Gateway esetén az ARP kérés nem kerül továbbküldésre sem a Fabricba, sem a L” domainba

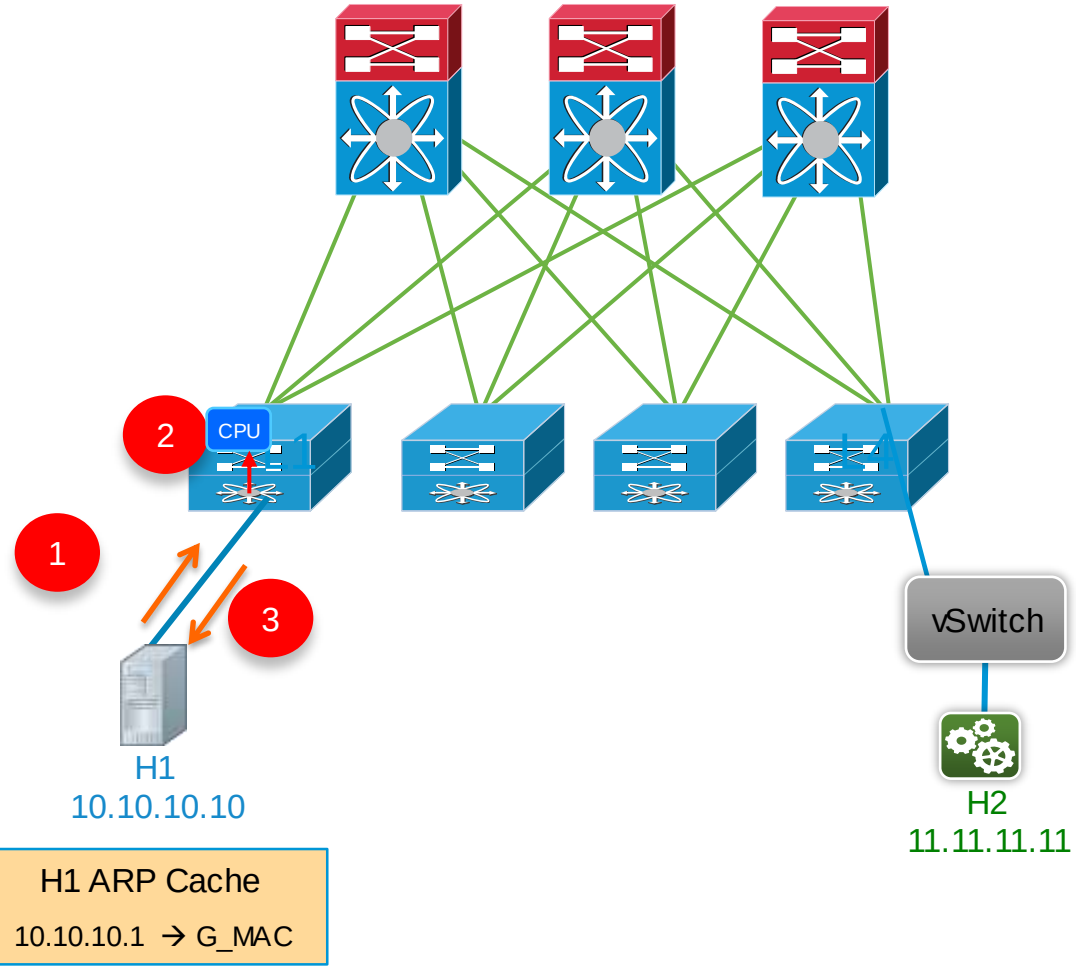
IP csomagtovábbítás azonos Subneten belül Proxy-gateway üzemmód (2)

4. H1 létrehozza a az adatcsomagot a G_MAC cél MAC címmel
5. Leaf1 fogadja a csomagot, leveszi a L2 headert, L3 lookupot végez a cél meghatározásához
6. Leaf1 encapsulálja a csomagot hozzáteszi a Layer 2 és FP header-t, majd továbbküldi a FP csomagot a Fabricban a három azonos súlyú út valamelyikén (S1, S2, S3)
7. Leaf4 veszi a csomagot leveszi a FP és L2 headert L3 lookopot végez és elküldi a csomagot H2 felé



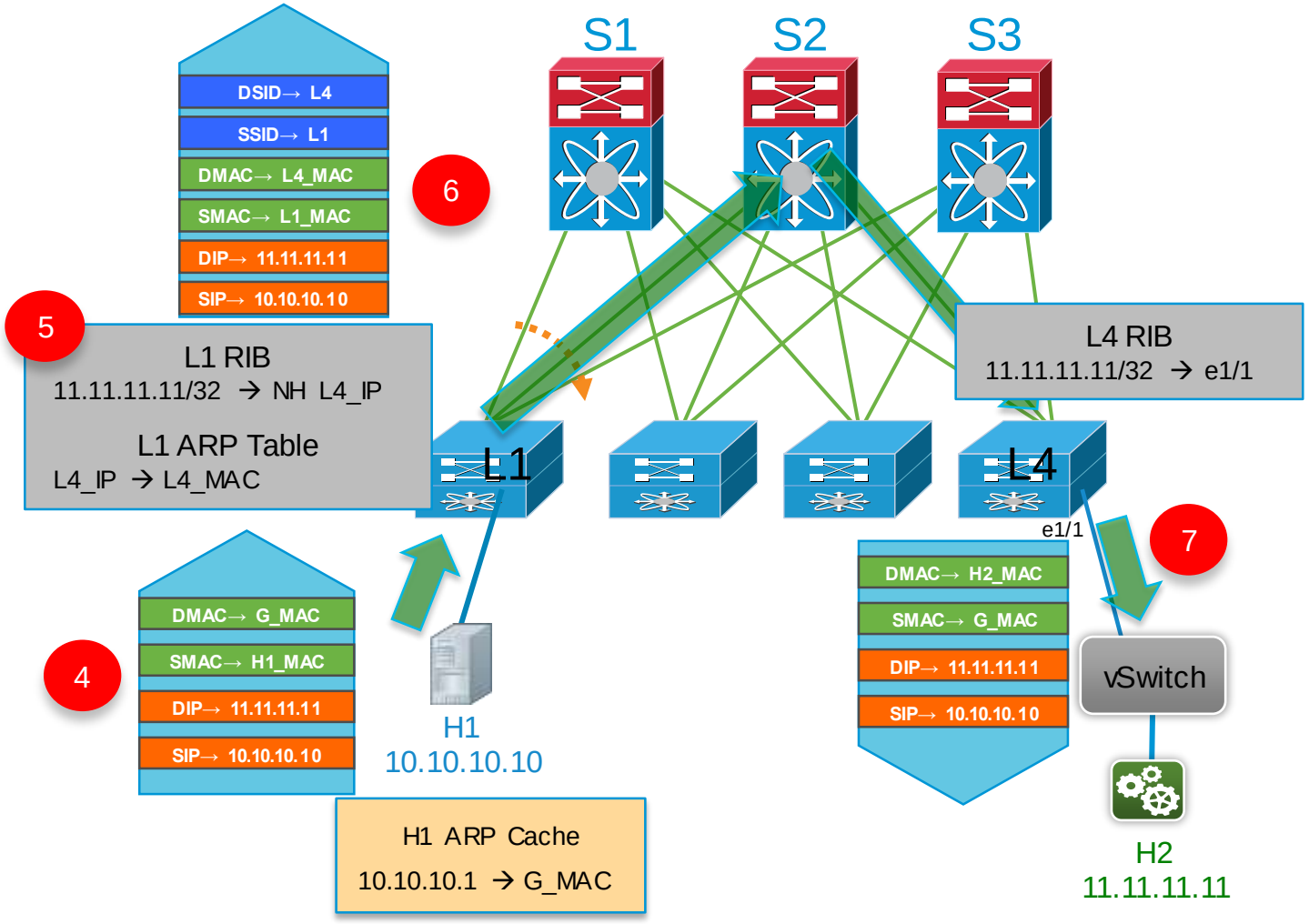
IP csomagtovábbítás különböző Subnetek között Proxy-gateway üzemmód

1. H1 elküldi az ARP kérést a default gateway – 10.10.10.1 felé
2. Az ARP kérést a leaf lekezeli (Supervisora felé küldi)
3. Leaf1 hagyományos default gateway funkciót lát el és elküldi az ARP választ a G_MAC címmel



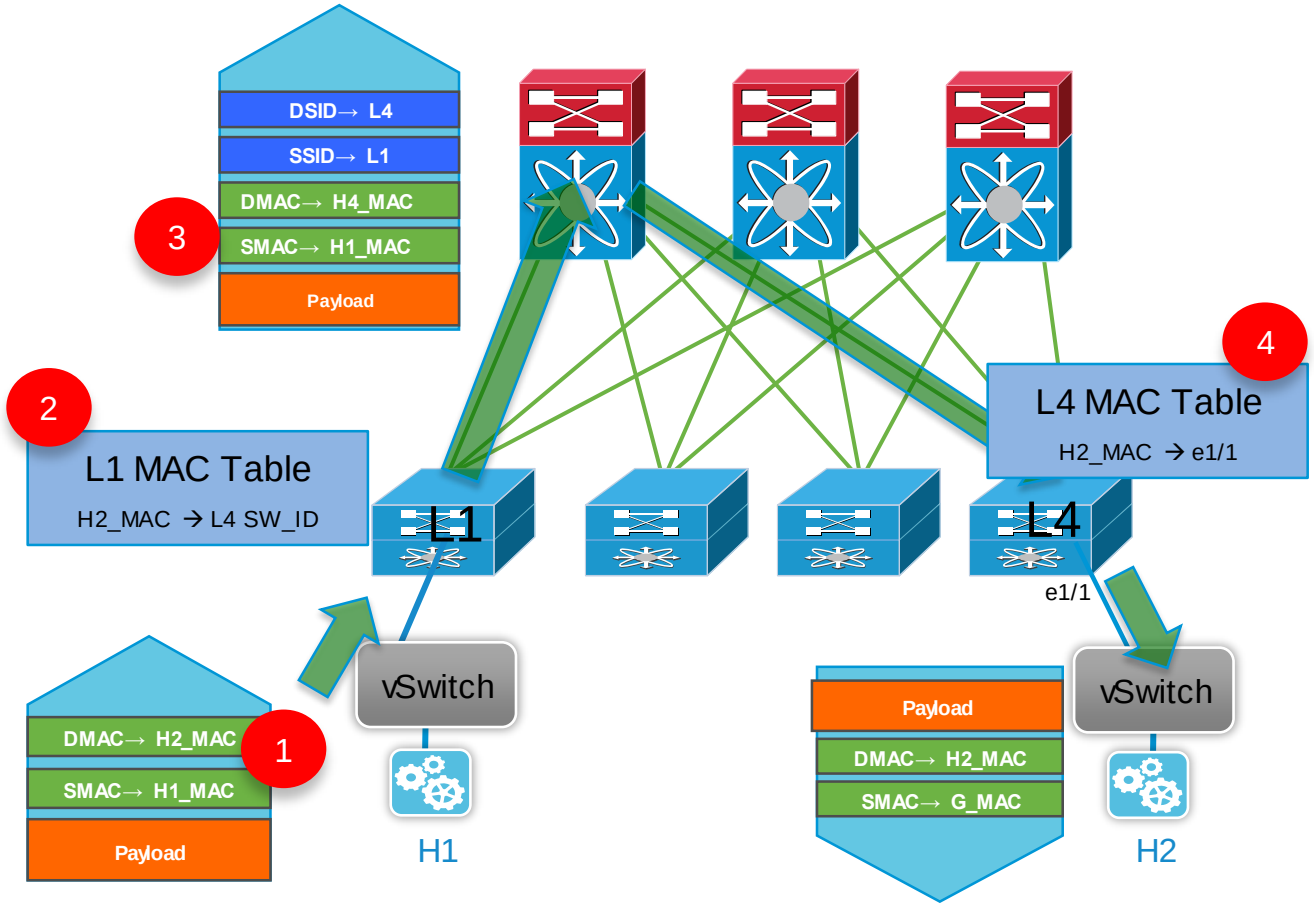
IP csomagtovábbítás különböző Subnetek között Proxy-gateway üzemmód (2)

4. H1 létrehozza a csomagot H2 IP címmel és G_MAC cél MAC címmel
5. Leaf1 fogadja a csomagot, leveszi a L2 headert, L3 lookupot végez a cél meghatározásához
6. Ha H2-re érvényrs route bejegyzés van a routing táblában Leaf1 enkapszulálja a csomagot L2 és FP fejléccel látja el és továbbítja a FP csomagot a Fabricban a három azonos súlyú út valamelyikén (S1, S2, S3)
7. Leaf4 veszi a csomagot leveszi a FP és L2 headert L3 lookopot végez és elküldi a csomagot H2 felé



L2 nem IP tipusú Csomagtovábbítás

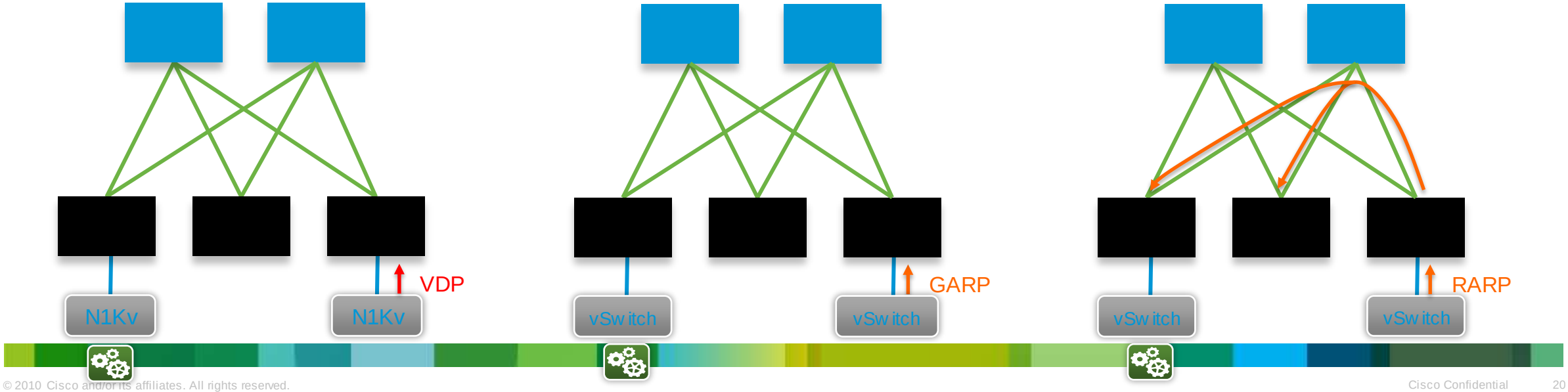
1. H1 elküldi a csomagot a H2 MAC címére
2. A Leaf1 L2 lookupot végez a H2 címre az adott VLANban
3. Leaf1 hozzáteszi a Layer 2 és FP headert mielőtt elküldené a csomagot a fabricba
4. Leaf4 veszi a csomagot leveszi a FP és L2 headert, L3 lookopot végez és elküldi a csomagot H2 felé



Host Mobilitás
VM költözés

Amikor egy VM költözik egy új Leafre, a VM új helyzetét az alábbi módon detektálhajuk:

- VDP: az N1Kv egy explicit control plane üzenetet küld a VM költözésről
- GARP: VM költözéskor, a hypervisor egy Gratuitous ARP (GARP) csomagot küld → a helyi leaf a csomag tartalmában lévő információból kinyeri a VM IP címét
- RARP: VM költözéskor a hypervisor egy Reverse ARP (RARPcsomagot küld → mivel a csomag tartalmában nincs IP információ a fabricnak kell lekezelni a RARP üzenetet, ami triggereli a VM felfedezési folyamatot



Cisco Dynamic Fabric Automation Platform támogatás

Cloud Stack & Orchestration Tools

Compute & Storage

cloudstack
System Center
vmware
openstack
cloupia

Network

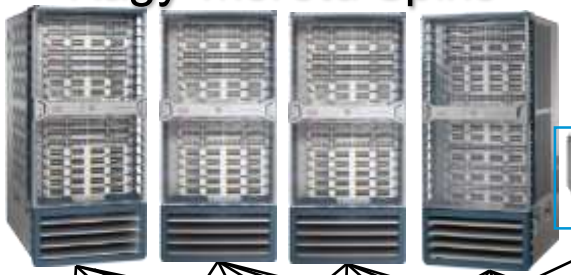
DCNM/CPoM

Network Services

Services Controller

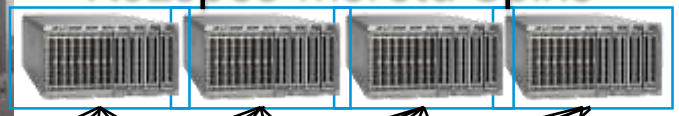
Nexus 7X00 (F2/F2e/F3)

Nagy méretű Spine



Nexus 6004

Közepes méretű Spine

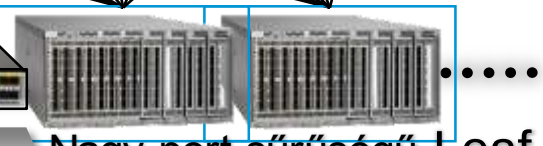


ToR Leaf

Nexus 6001

Fabric Extender

Nexus 2x00



Nagy port sűrűségű Leaf

Nexus 6004

Fabric Extender

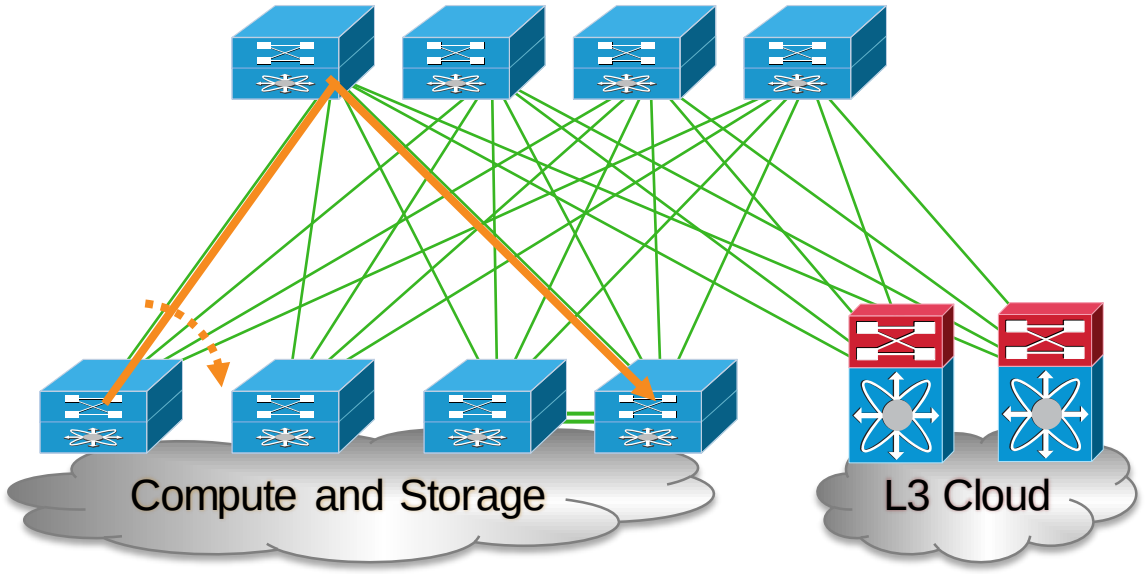
Nexus 2x00

Nexus 7X00 (F3)

Border Leaf

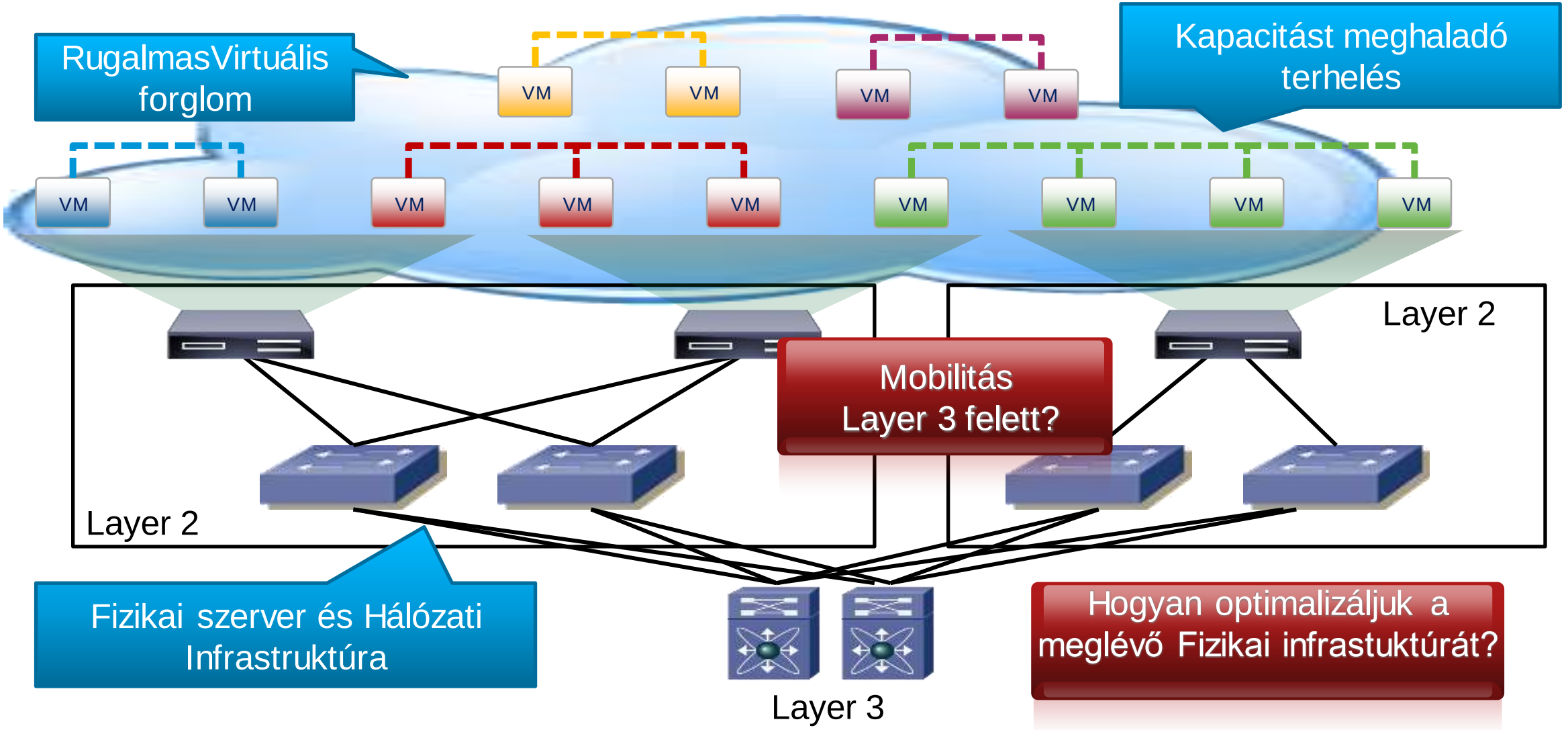
Cisco Dynamic Fabric Automation Összefoglalás

- Nagy sávszélességű két rétegű Architektúra
- L2 és L3 egységes működés
- Optimalizát - ECMP: Unicast vagy Multicast
- Egységes elérhetőség, kiszámítható késleltetés
- Magas redundancia: Node/Link meghibásodás
- Veszteségmentes kis késletetésű valamennyi forgalomra
- Elosztott Gateway üzemmód
- Rugalmasan skálázható, tervezhető, bővíthető

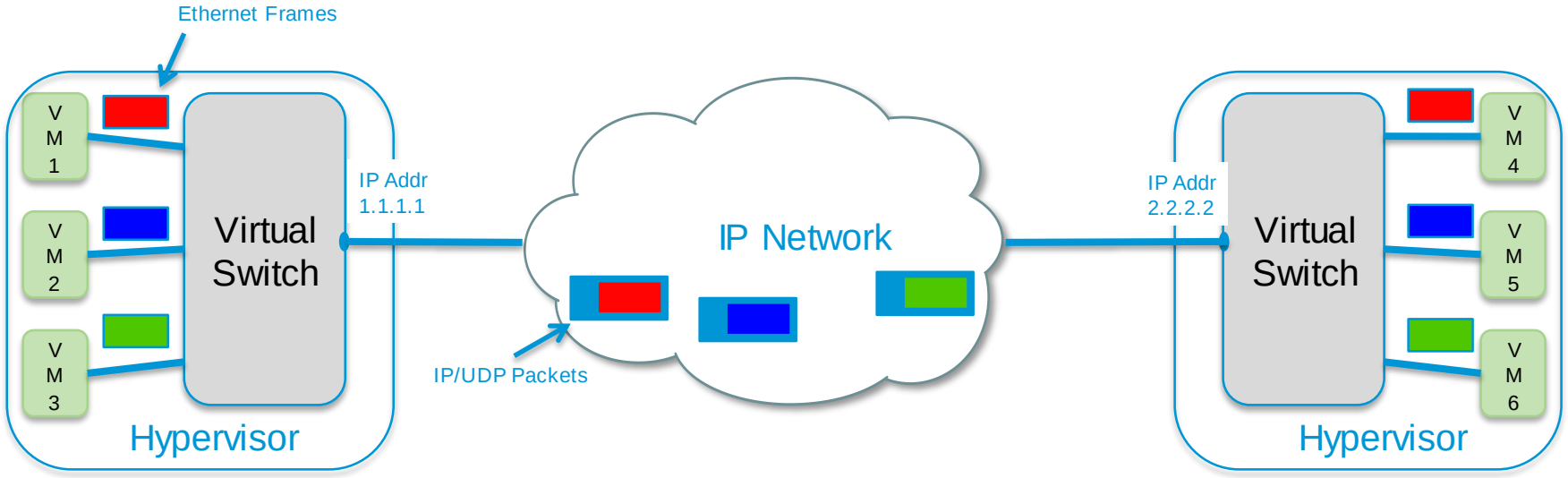


Virtuális gépek Hálozati Fabric Megoldása

Cloud technológia megjelenése a Fizikai hálózati környezetben



Overlay Network

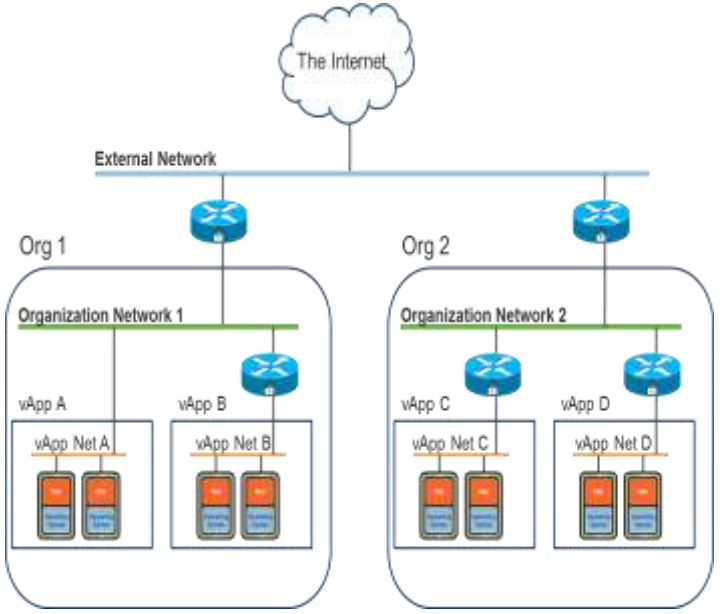


Sok Felhasználós (Multi-Tenant) L2 szegmens ígénye

- Mind a MAC mind az IP cím tartományok átlapolódhatnak (akár egy felhasználónál is különböző virtuális alkalmazásoknál)

  Mindem átlapoló címtaromány elkülönített szegmenst ígényel

- VLAN: 12 bit ID= 4K
- VXLAN: 24 bit IDs = 16M



Mi a VXLAN?

- VLAN plusz egy X a közepén ☺
- A VXLAN ugyanazt nyújtja a virtuális gépeknek mint a hagyományos VLAN fizikai környezetben
- Az X = eXtensible
  - Skálázhatóság
  - Több L2 szegmens mint VLANokkal
  - VLANok kiterjesztése
- VXLAN egy Overlay Network technologia
  - MAC Over IP/UDP
- A Cisco, VMware és más hypervisor network technológiában érdekelt gyártó benyújtotta a VXLAN szabvány tervezetet az IETF-hez (draft-mahalingam-dutt-dcops-vxlan-01.txt)

VXLAN Versus NVGRE

Network Virtualization Generic Routing Encapsulation

Hasonlóságok

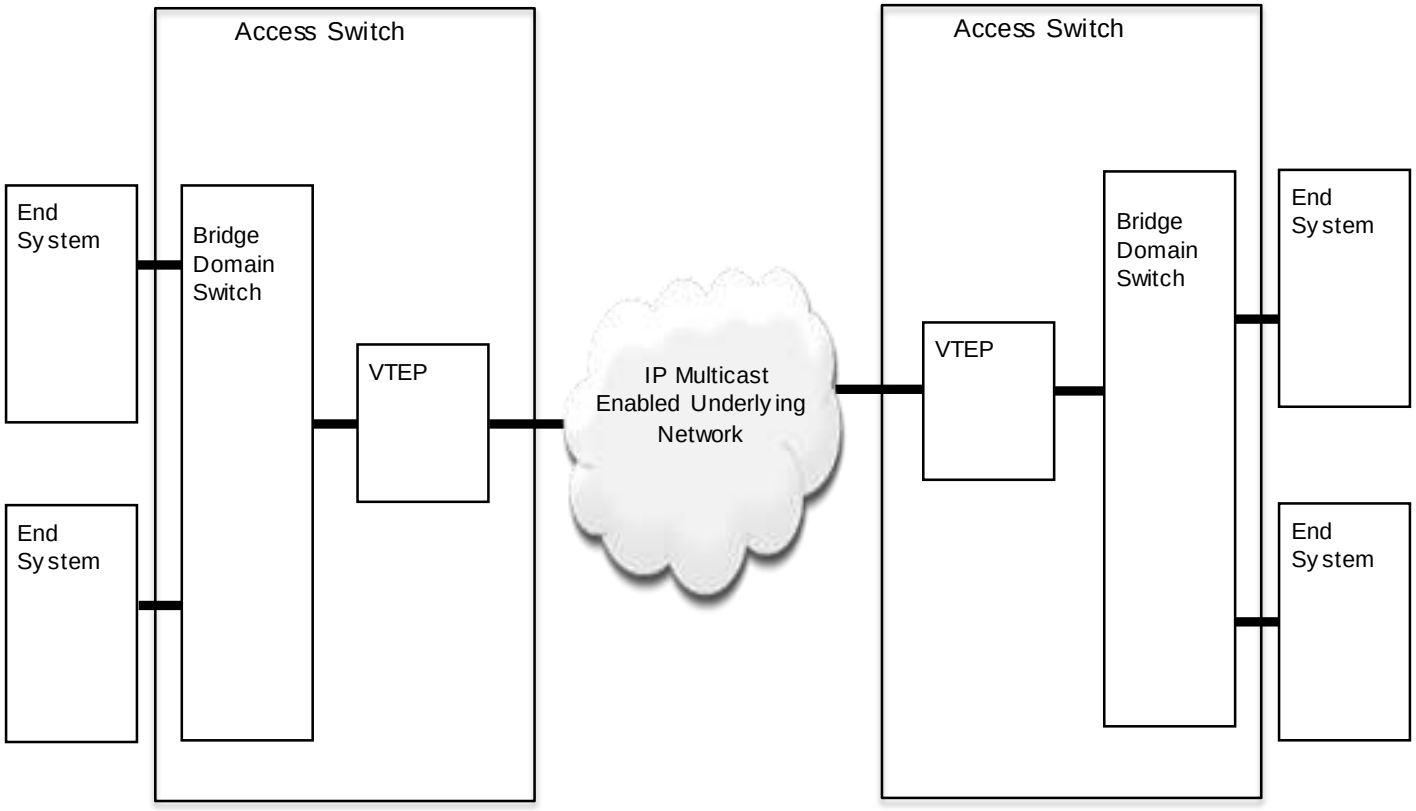
- IP Transport
  - Tuunel technológia az access switchek között
- IP Multicast
  - Broadcast és multicast csomagokkal
- 24 Bit Segment ID

Nexus 1000V a Hyper-V-n támogatni fogja az NVGRE-t

Különbségek

- IETF Draft
  - VXLAN: Cisco, VMware, Citrix, Red Hat, Broadcom, Arista
  - NVGRE: Microsoft, Intel, Dell, HP, Broadcom, Arista, Emulex
- Enkapszuláció
  - VXLAN: UDP 50 bytes
  - NVGRE: GRE 42 bytes
- Firewall ACL VXLAN UDP port alapon
  - Nehéz a GRE protocol field alapján
- Forwarding Logic
  - VXLAN: Flooding/Learning
  - NVGRE: Not specified

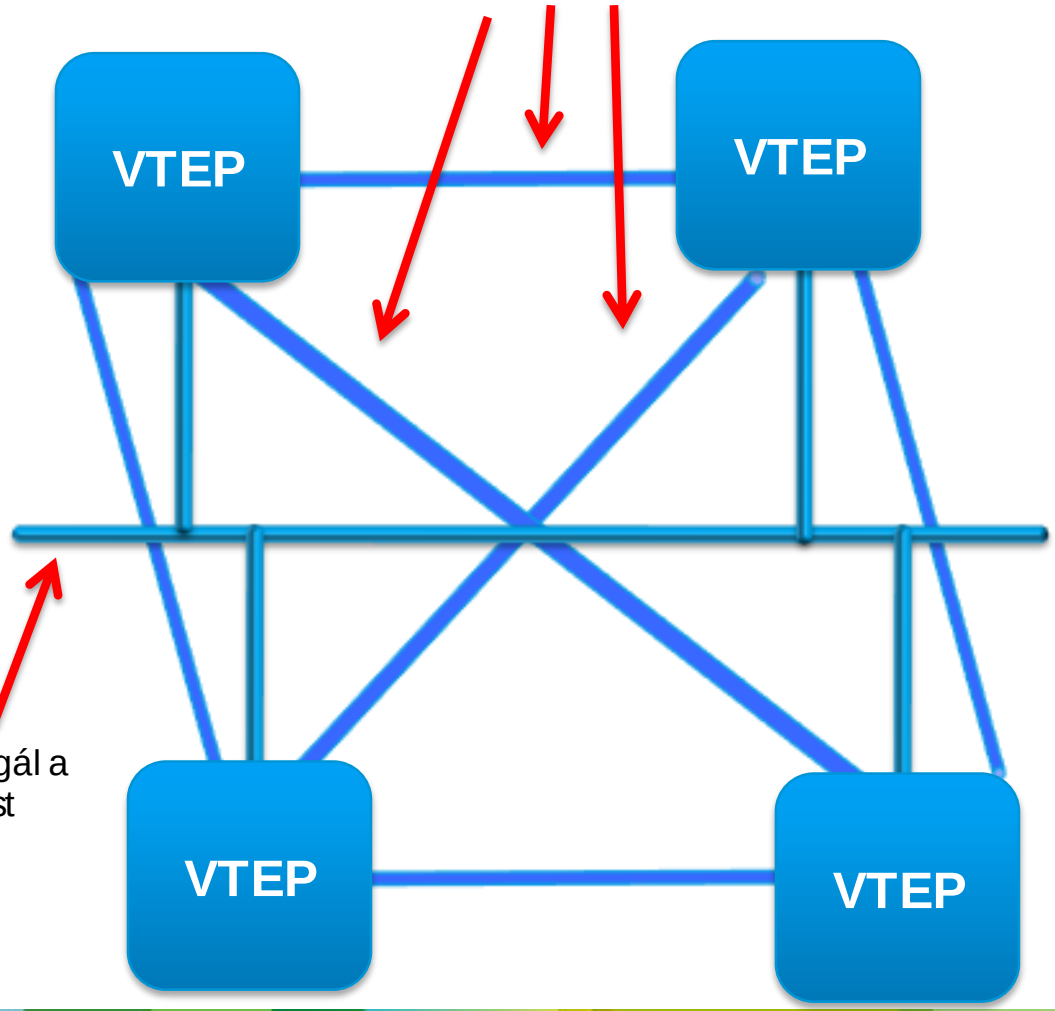
VXLAN Network Model



VTEP = VXLAN Tunnel End Point

VXLAN Data Plane Model

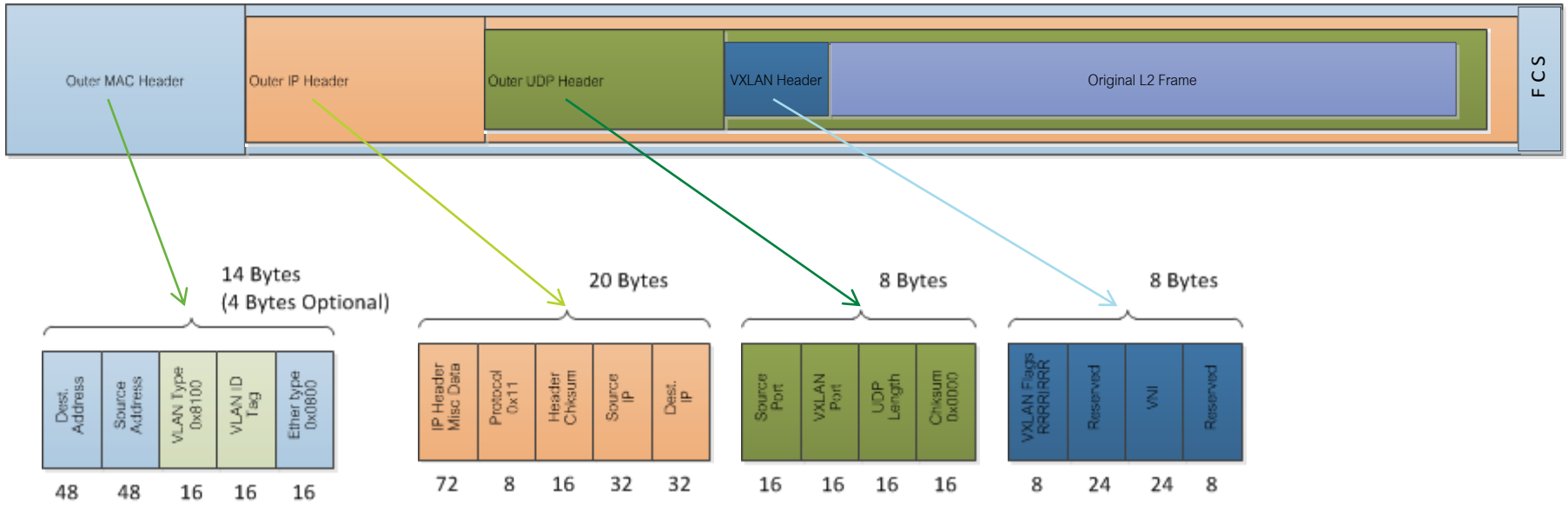
Közvetlen Unicast tunnelek az egyes VTEPek között (Known Unicast csomag továbbítás)



VXLAN IP Any Source Multicast Group szolgál a releváns VTEP unknown/broadcast/multicast forgalom átvitelére

VXLAN csomag struktúra

- Eredeti L2 csomag VXLAN fejléccel



UDP fejlécben egy rögzitett UDP destination port tartozik a VXLAN forgalomhoz

UDP source port hash algoritmussal keletkezik a belső IP Ethernet fejlécből

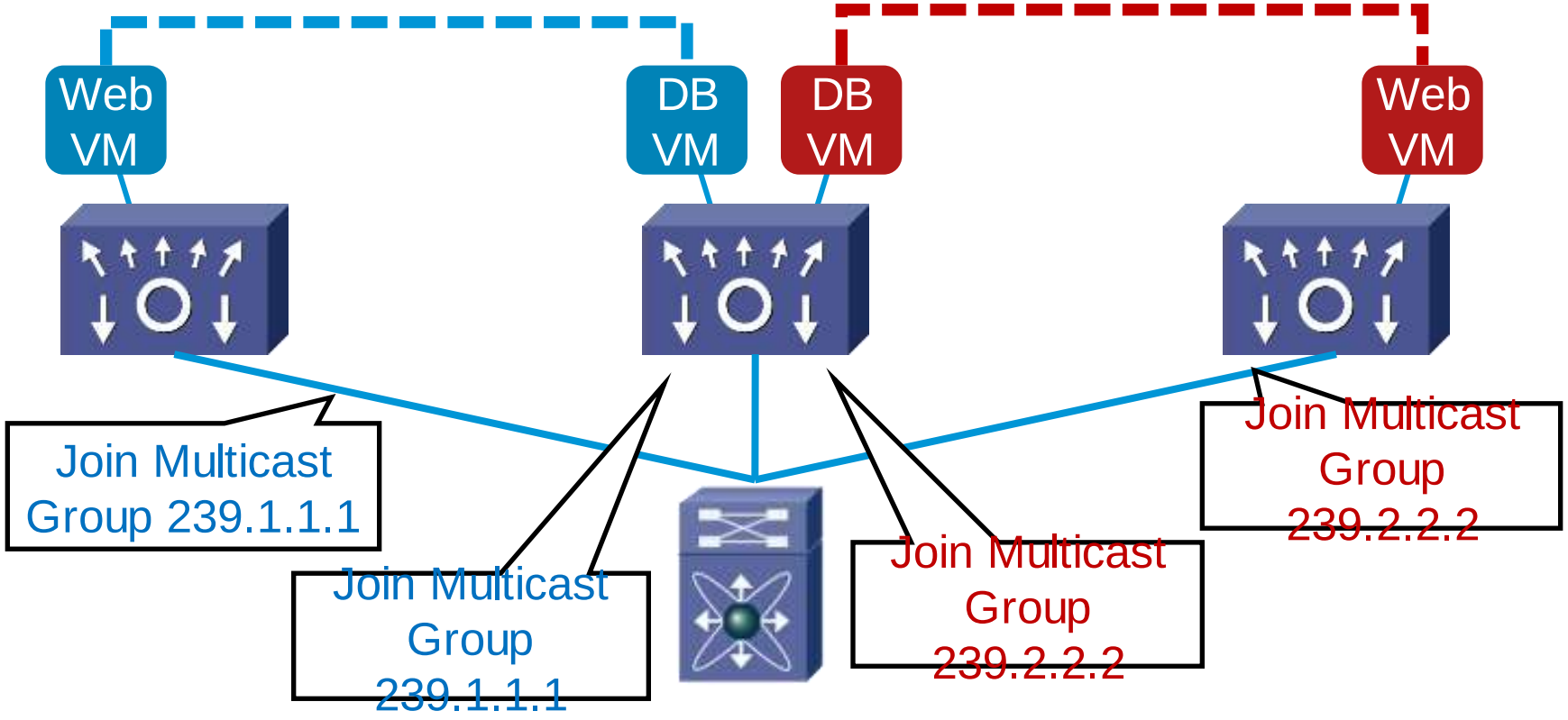
IP csomag destination és source címei a megfelelő VTEP IP címek

Külső MAC fejléc source címe a VTEP MAC és a destination a next hop MAC

Külső MAC csomag opcionális VLAN tag-et tartalmazhat (amennyiben Trunk-ön megy át)

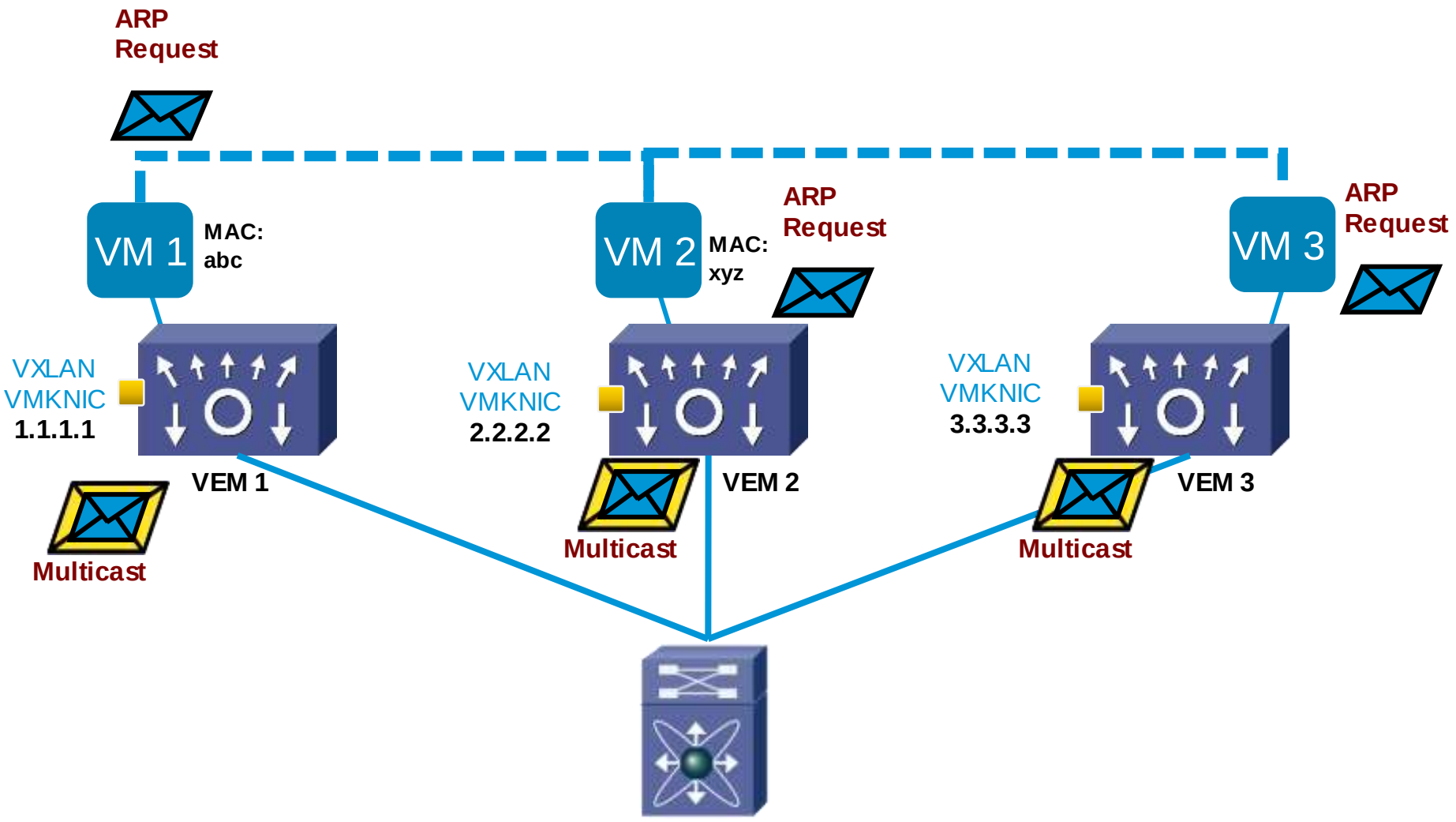
VTEP IGMP alapú kommunikációja

- IGMP-t használ a VXLANhoz rendelt Multicast Group-hoz történő csatlakozáshoz



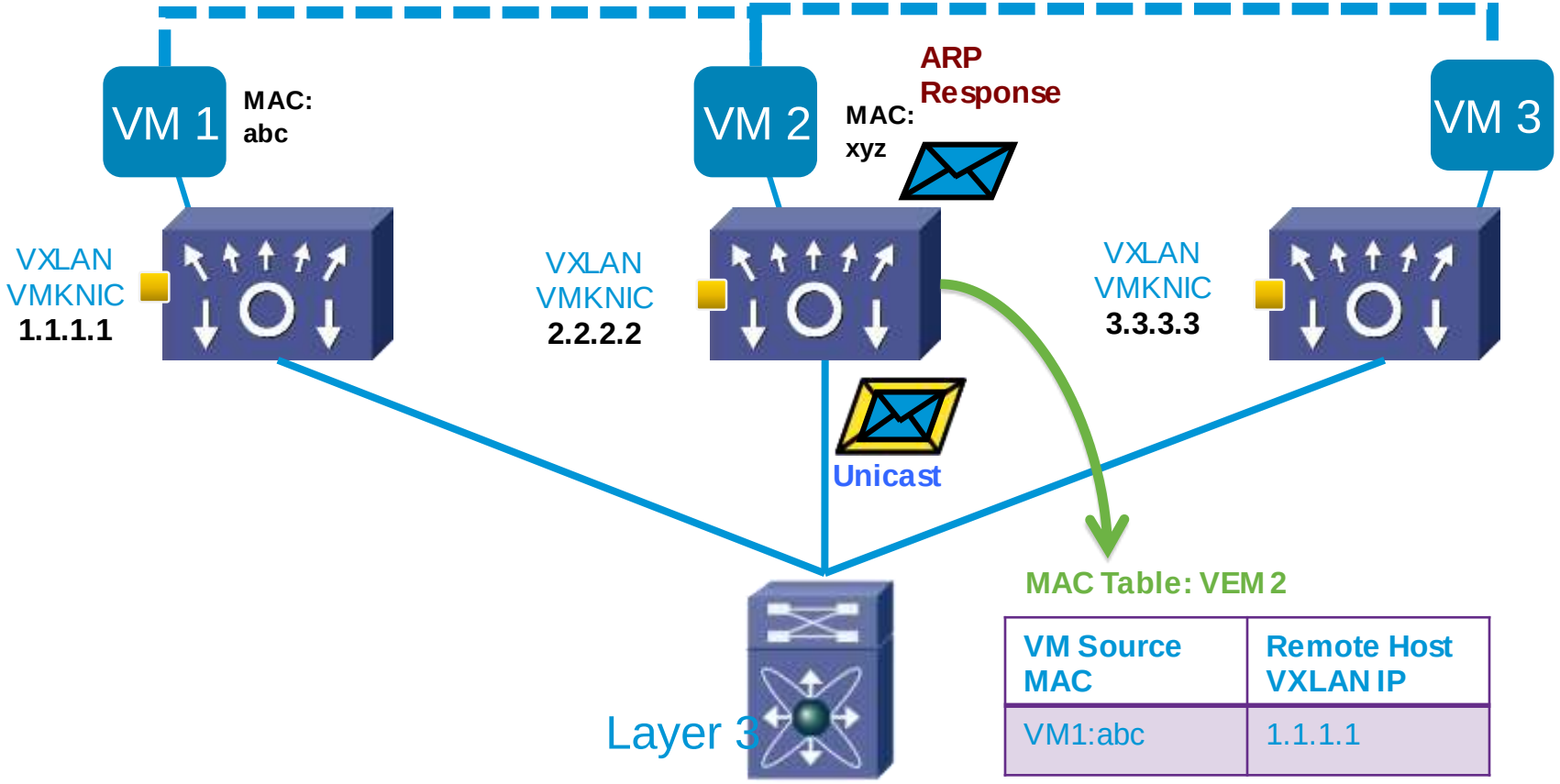
VXLAN Adatátvitel

- VM1 VXLAN-on történő kommunikál VM2-vel



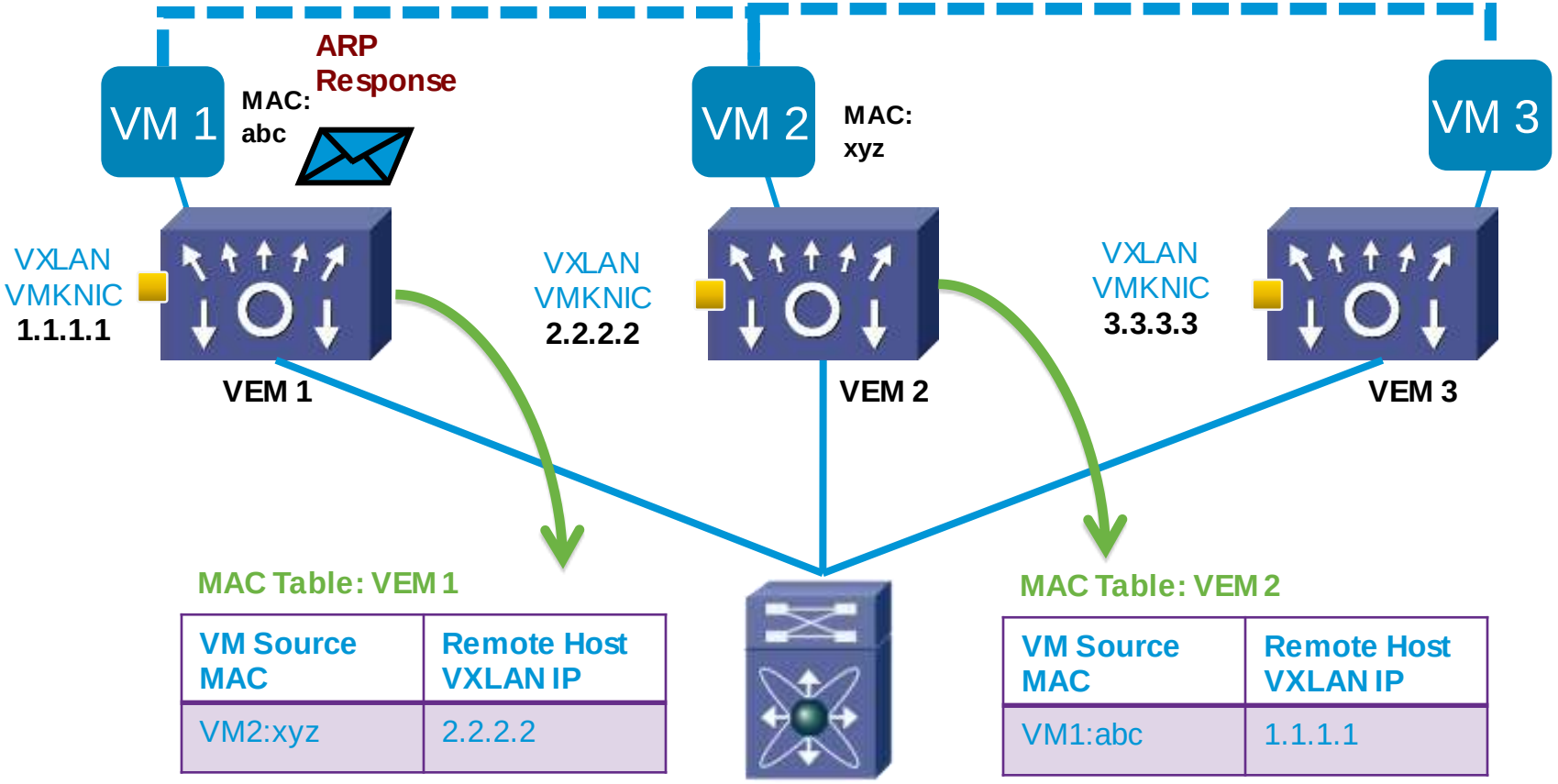
VXLAN Adatátvitel

- VM1 VXLAN-on történő kommunikál VM2-vel



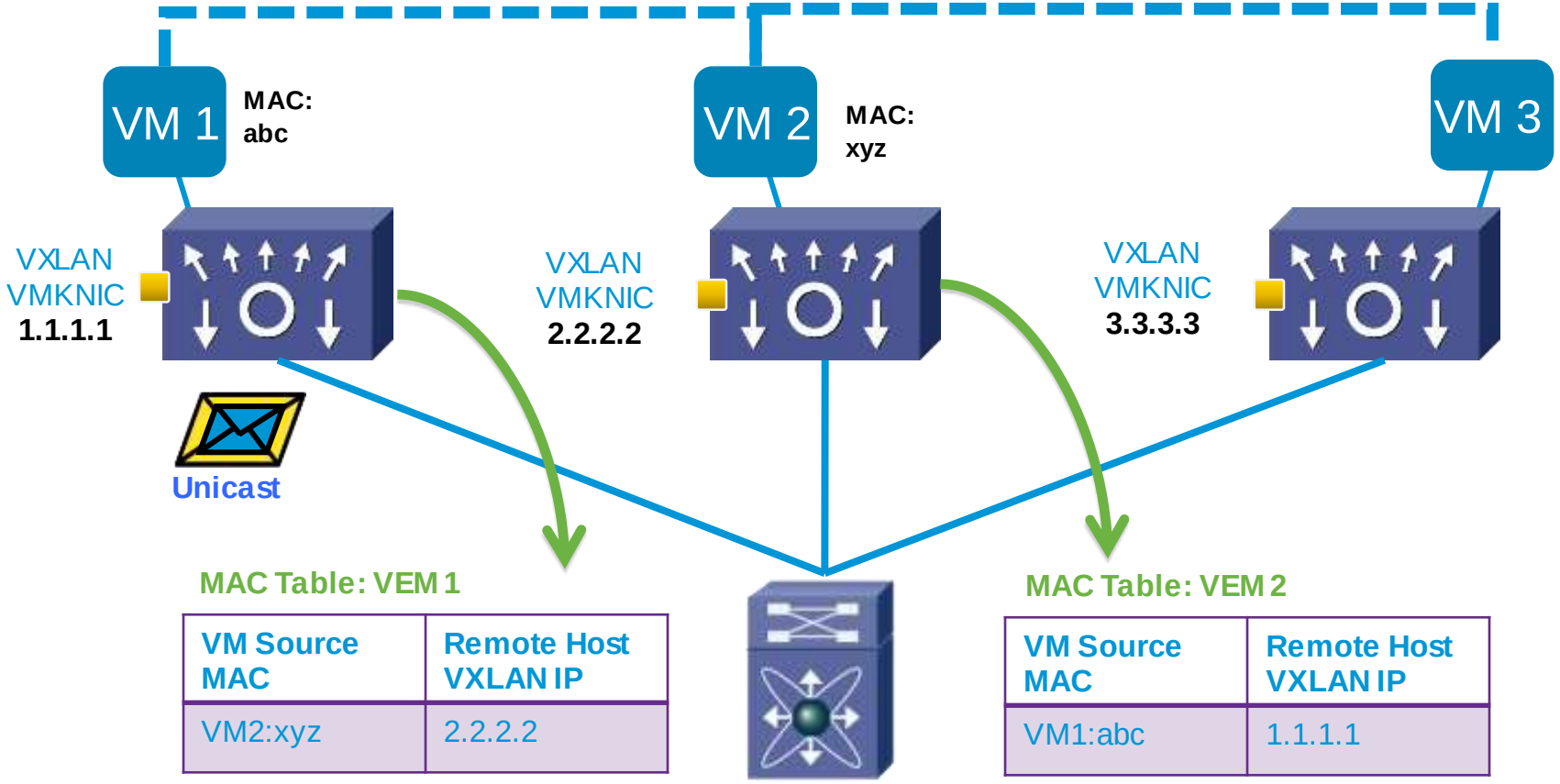
VXLAN Adatátvitel

- VM1 VXLAN-on történő kommunikál VM2-vel



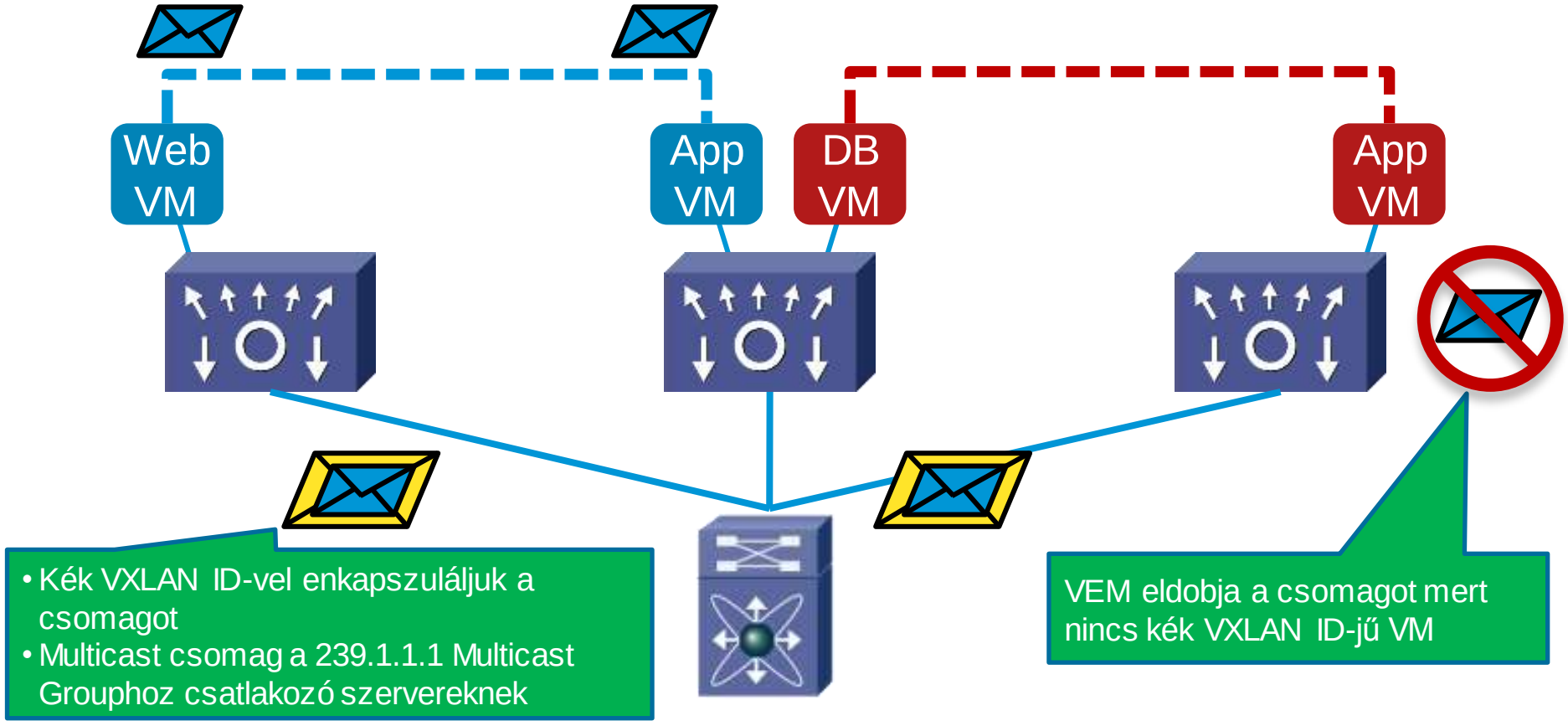
VXLAN Adatátvitel

- VM1 VXLAN-on történő kommunikál VM2-vel



Több VXLAN közös Multicast Groupban

- Kék & piros VXLAN osztozik a 239.1.1.1 Multicast Groupon



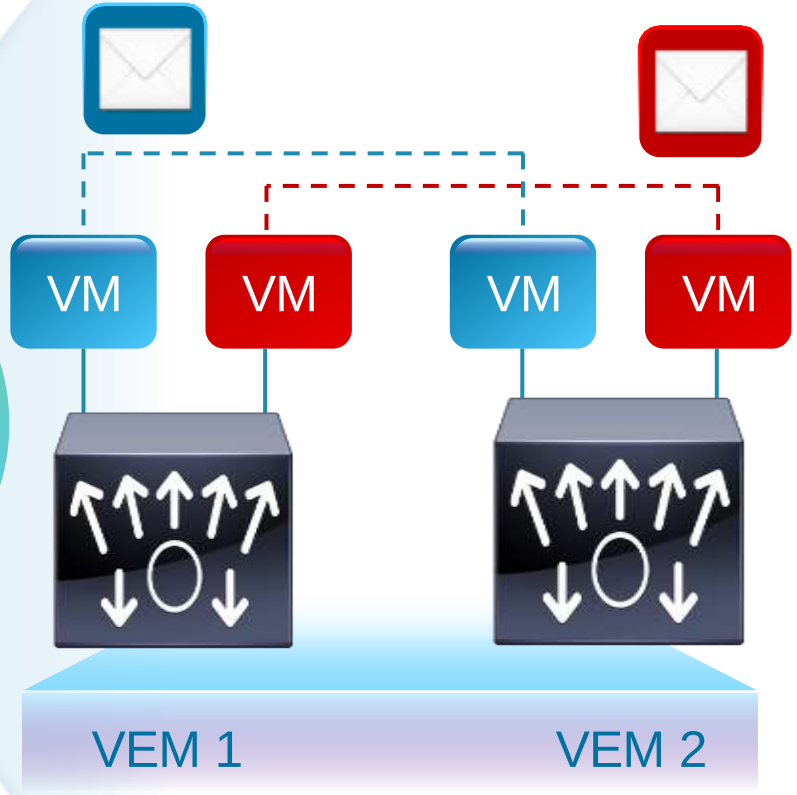
VM Broadcast csomagok több szerverhez is eljutnak
Broadcast Domain korlátozódik a VXLAN Szegmensre

VXLAN csomagtovábbítás

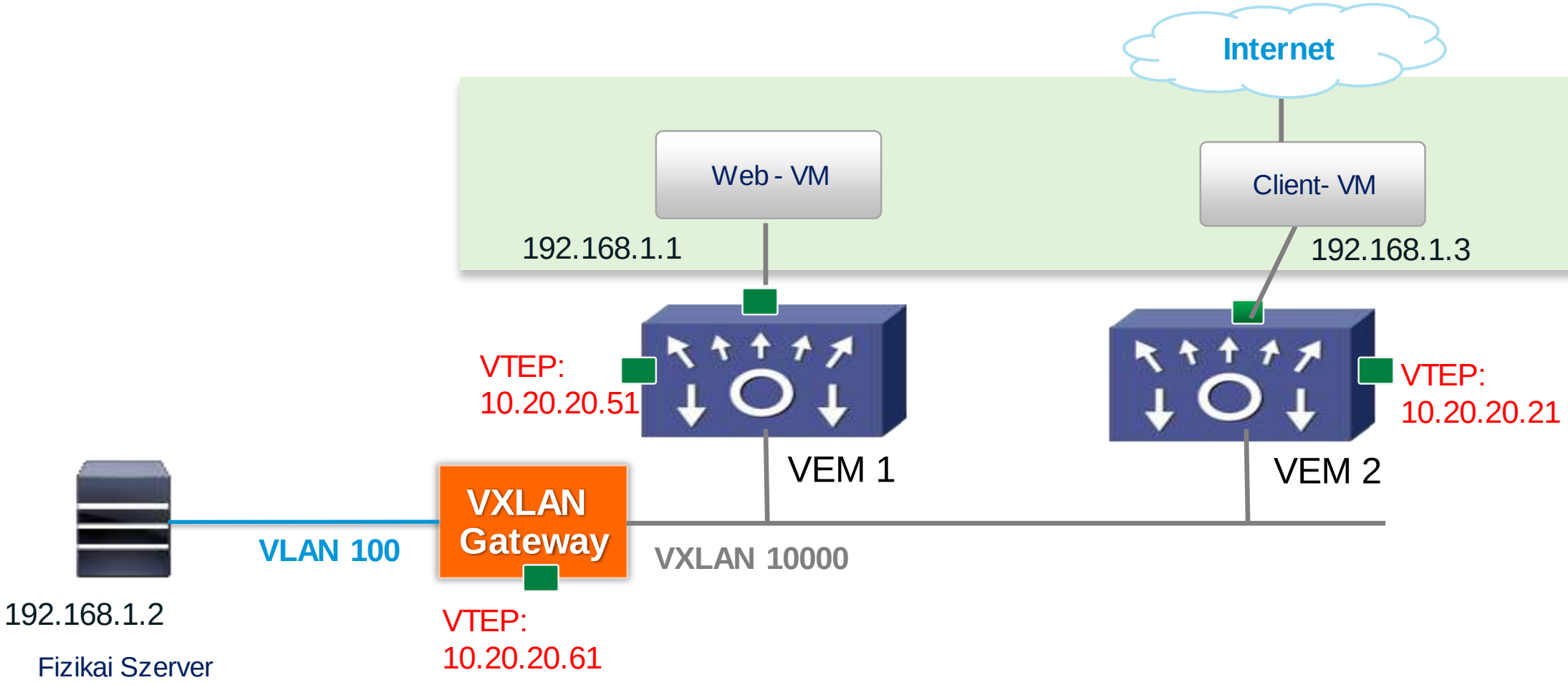
- Forwarding mechanizmus hasonló Layer 2 bridge-hez: Flood + learn
  - VEM megtanulja a VM source (MAC, host VXLAN IP) paramétereit

- Broadcast, multicast, és unknown unicast traffic
  - Multicast módon történik

- Unicast
  - Közvetlenül (nem mcast) enkapszuációval történik a destination VEM,-VXLAN IP címére.



VXLAN-VLAN Gateway topológia



VXLAN Előnyök

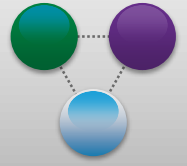
- Virtuális L2 szegmens, ami átnyúlik a fizikai L2 határokon - L3 domének fölött
- Nagy mértékben skálázható L2 szegmens több felhasználós (Multi-tenant környezetben)
- Hálózati szegmens ígény szerinti létrehozása a hálózat átkonfigurálása nélkül
- Rugalmas VM üzemeltetés az adatközpont bármely fizikai helyén
- VXLANs transzparensen működik valamennyi Adatközponti switchen, routeren

Alkalmazásközpotú hálózati Fabric

Application Centric Infrastructure

Alkalmazásközpotú Infrastruktúra
Application Centric Infrastructure - ACI

HAGYOMÁNYOS HÁLÓZATI MODEL

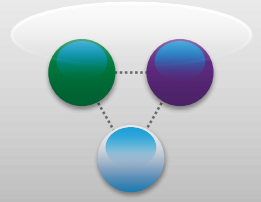


Önálló dobozok hálózatba kapcsolása

Kihívások 1G -> 1/10G

Kihívások 10G -> 40G

JELENLEGI SDN/FABRIC MODEL



Software-alapú Hálózavirtualizáció

Menedzsment rendszer alapú fabric kezelés

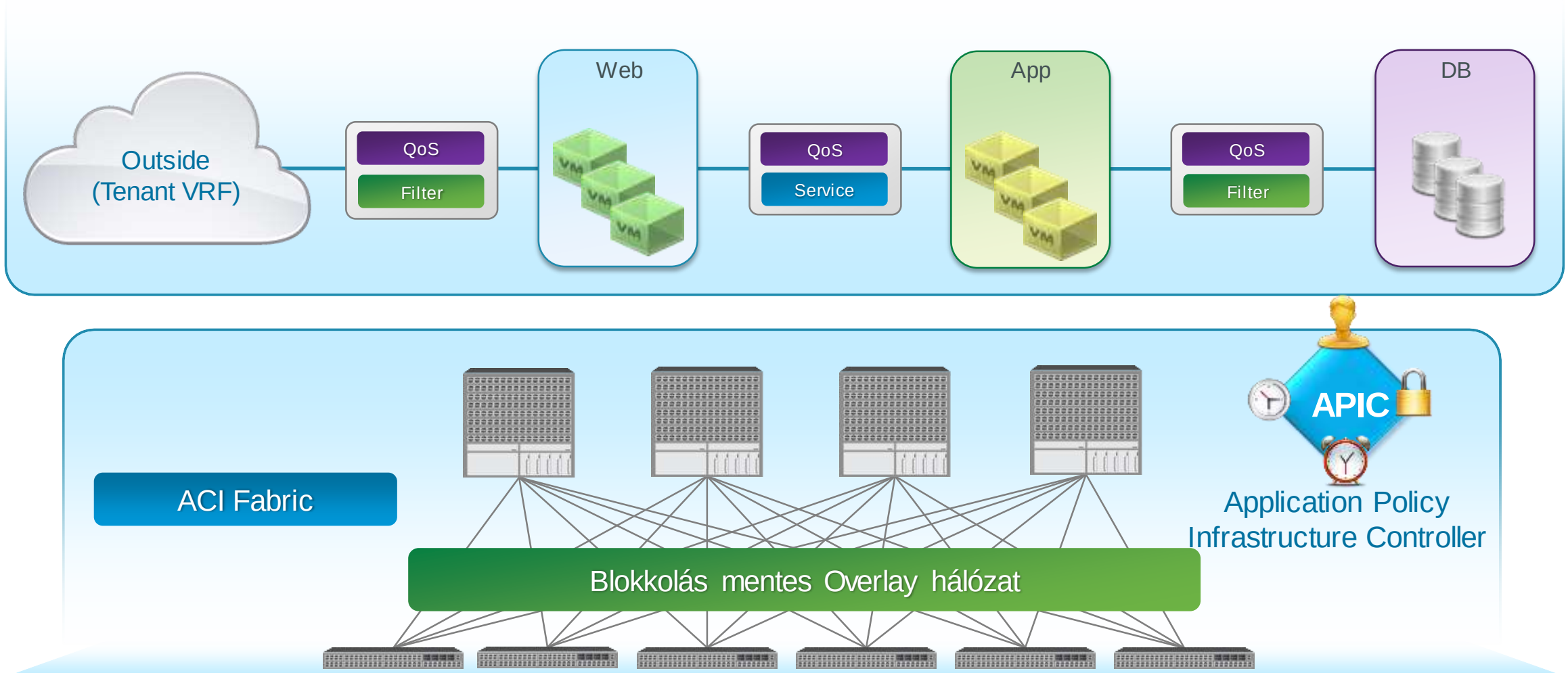
VXLAN, DFA

JÖVŐ MODEL

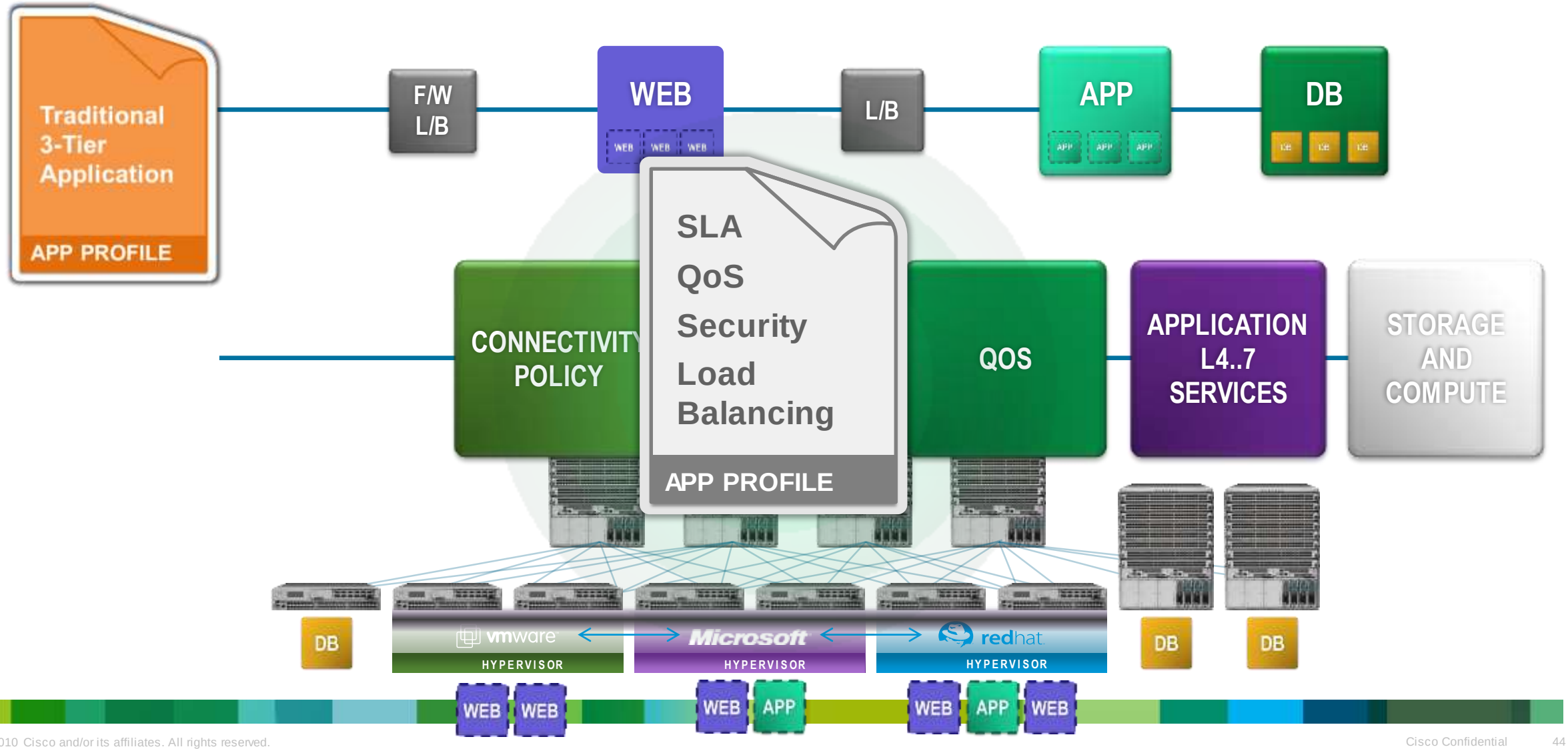
Központosított Automatizálás, Security, és Alkalmazás Profilok

ACI Architektúra

Alkalmazásközpontú Infrastruktúra
ACI - Stateless Hardware filozófia

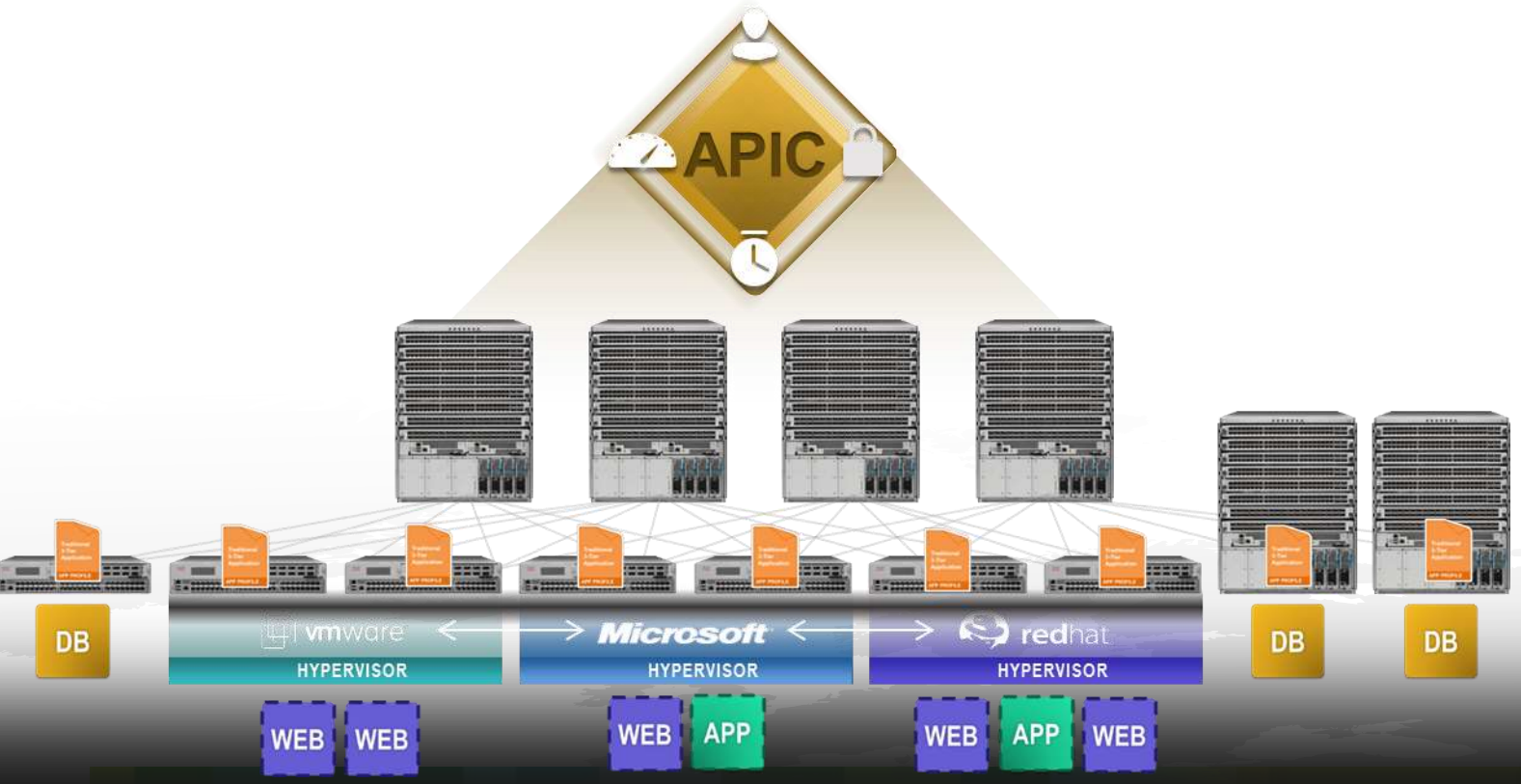


Application Network Profile:
Bármely Fizikai és Virtuális alkalmazás, bárhol lehet



Alkalmazás szintű monitorozás

Egyetlen közös felület elosztott környezetben



Traditional 3-Tier App

HEALTH SCORE

96%

LATENCY

5 Microsecond(s)

DROP COUNT

25 Packets Dropped

VISIBILITY

7 VMs

3 Physical

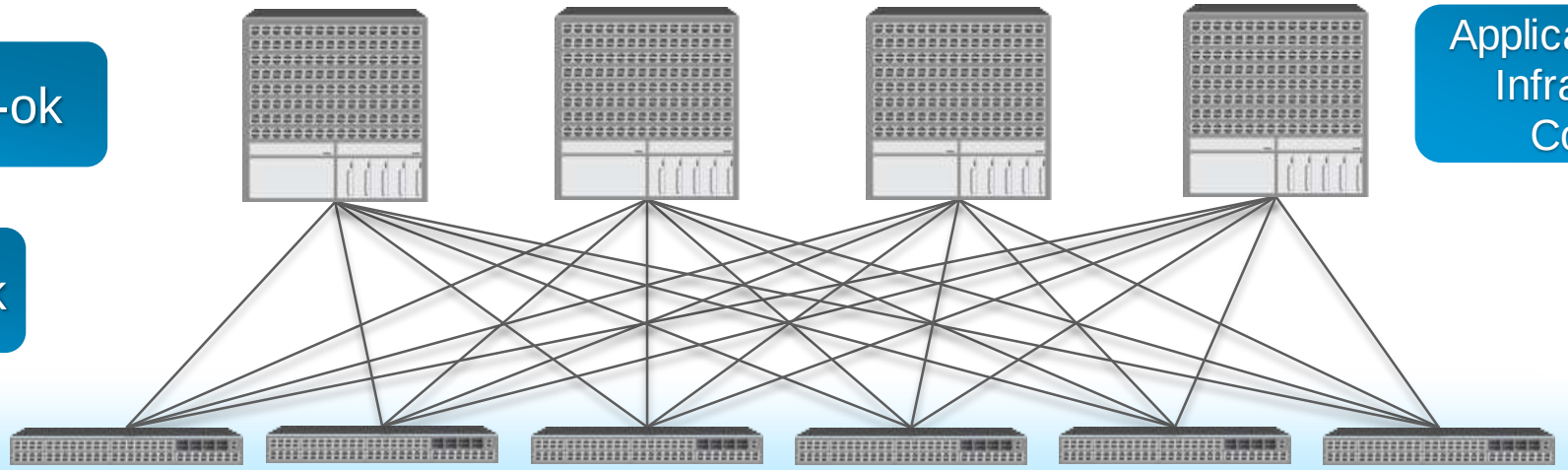
Application Delivery Controller

Firewall

ACI Fabric jellemzők

ACI Spine Node-ok

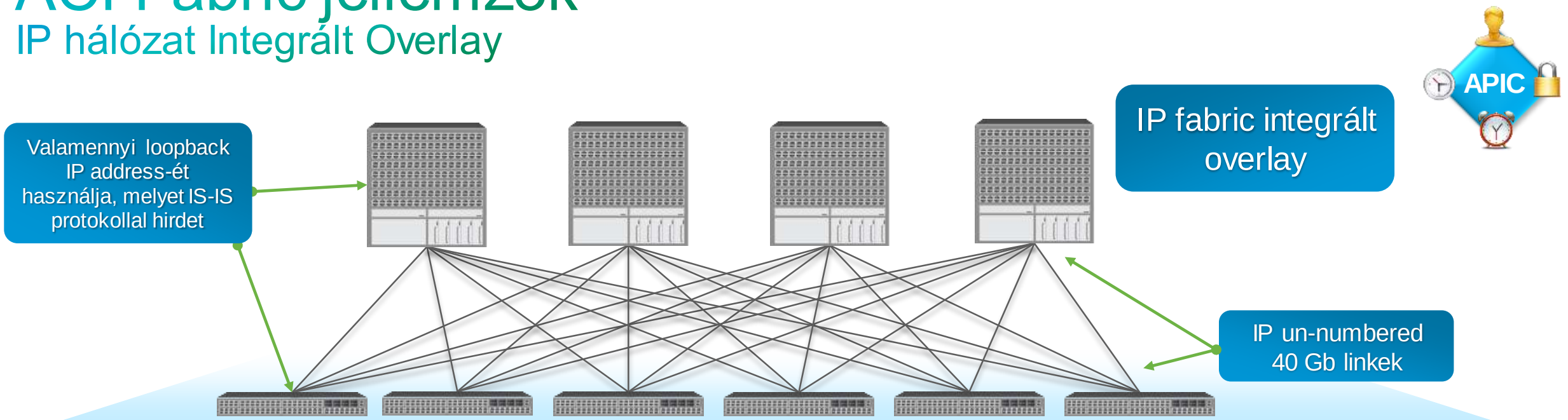
ACI Leaf Node-ok



Application Plolicy Infrastructure Controller

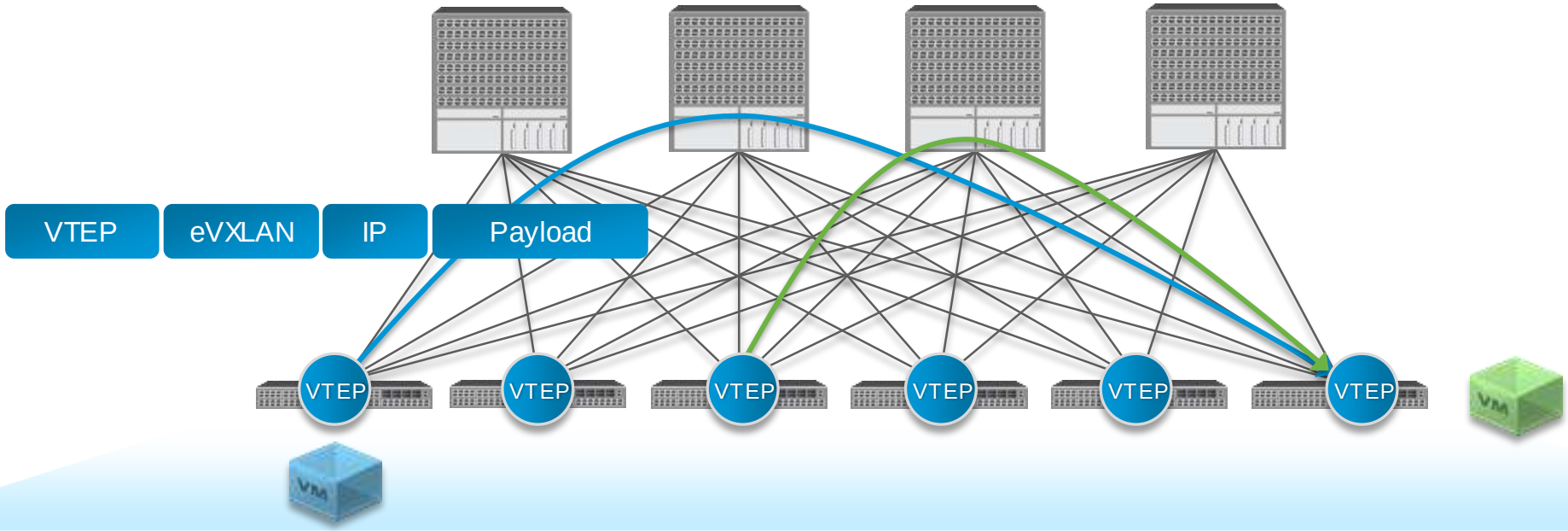
- ACI Fabric biztosítja:
  - A végpontok különválasztását a hely, policy, és az alatta lévő topológiától független paraméterektől
  - Normalizálja a különböző bejövő enkapszulációs mechanizmusokat: 802.1Q VLAN, IETF VXLAN, IETF NVGRE
  - Elosztott Layer 3 gateway optimalizálja a a Layers 3 és Layer 2 csomag továbbítást
  - Támogatja a szabványos bridging és routing technológiákat hely kötöttségek nélkül (Bármely IP cím bárhol lehet)
  - Szerviz szolgáltatás beillesztése, forgalom átirányítás
  - IP control plane (ARP, GARP) vezérlő üzenet forgalom kiszűrése

ACI Fabric jellemzők
IP hálózat Integrált Overlay



- ACI Fabric a fabric szélein IP routing technológián alapul amin host routing alapú overlay hálózatot valósítunk meg
- Valamennyi végberendezés forgalma ezen overlay technológián keresztül kerül átvitelre.
- Miért használunk integrált overlayt?
  - Mobilitás, skálázhatóság, több felhasználói csoport teljes elkülönítése, integrálás multi hypervisor környezethez
  - Adat forgalom a különböző policykhoz tartozó explicit meta data információt hordoz (flow-szintű vezérlő információ)

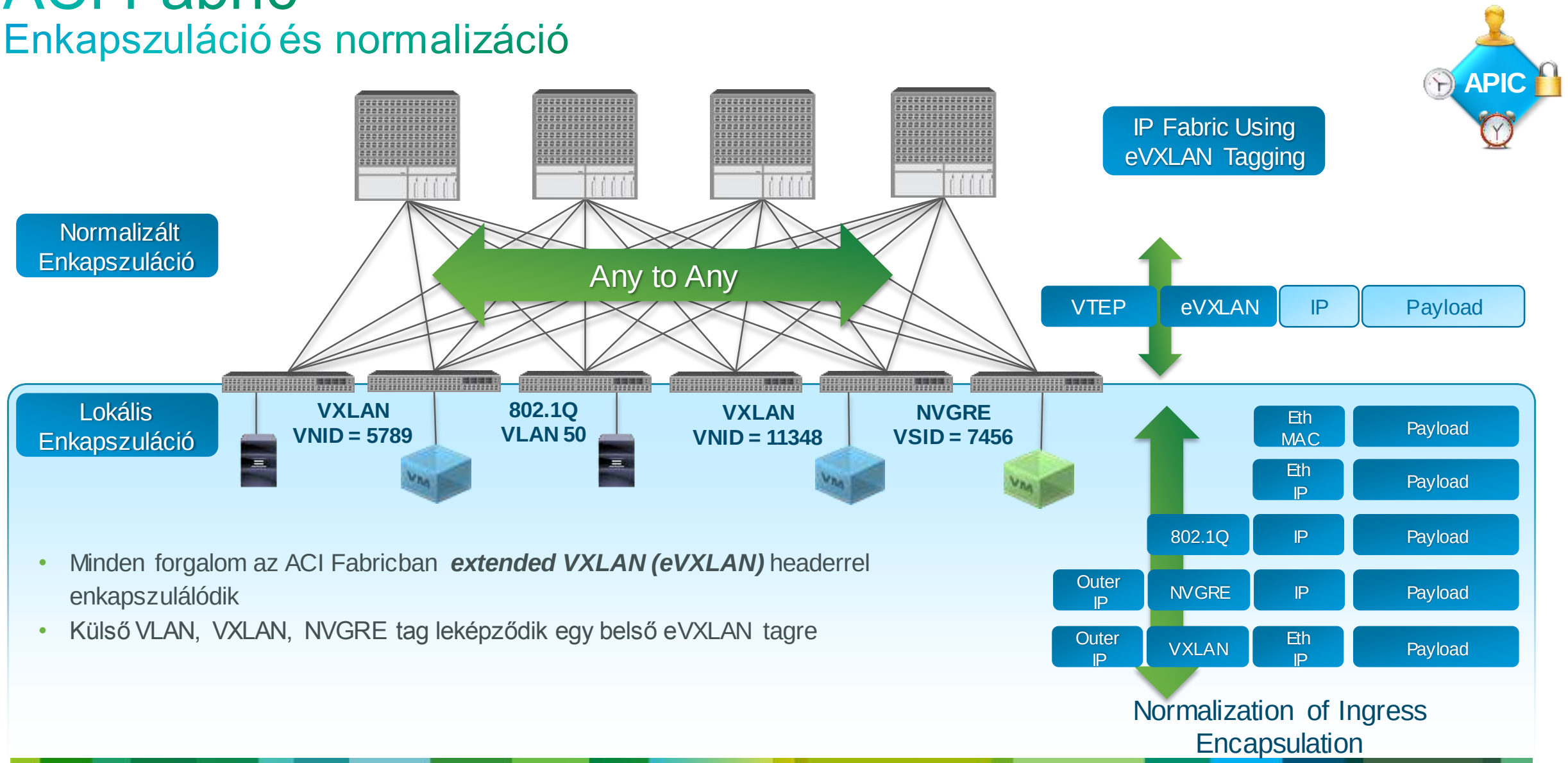
ACI Fabric
Hely és policy azonosítok leválasztása



- ACI Fabric leválasztja a végpont azonosítóját (címét) a hely locator azonosítójáról (VTEP címéről)
- Csomagtovábbítás a Fabricban VTEPek (eVXLAN tunnel endpointok) között történik felhasználva a eVXLAN headerben megtalálható policy azonosítót
- A belső MAC vagy IP cím leképzést a hely azonosítóra a VTEP elosztott adatbázis alapján végzi

ACI Fabric

Enkapszuláció és normalizáció

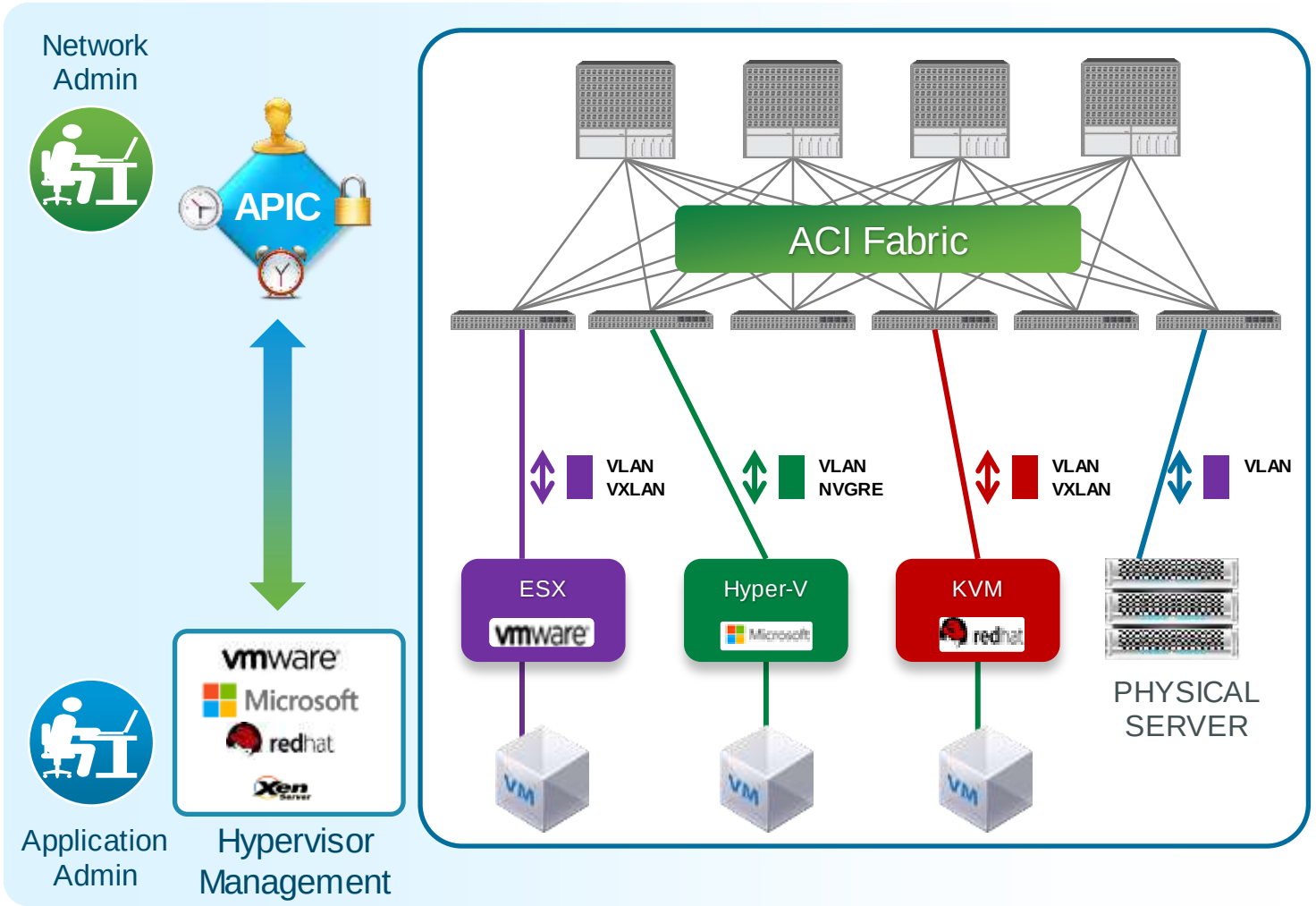


- Minden forgalom az ACI Fabricban ***extended VXLAN (eVXLAN)*** headerrel enkapszulálódik
- Külső VLAN, VXLAN, NVGRE tag leképződik egy belső eVXLAN tagre

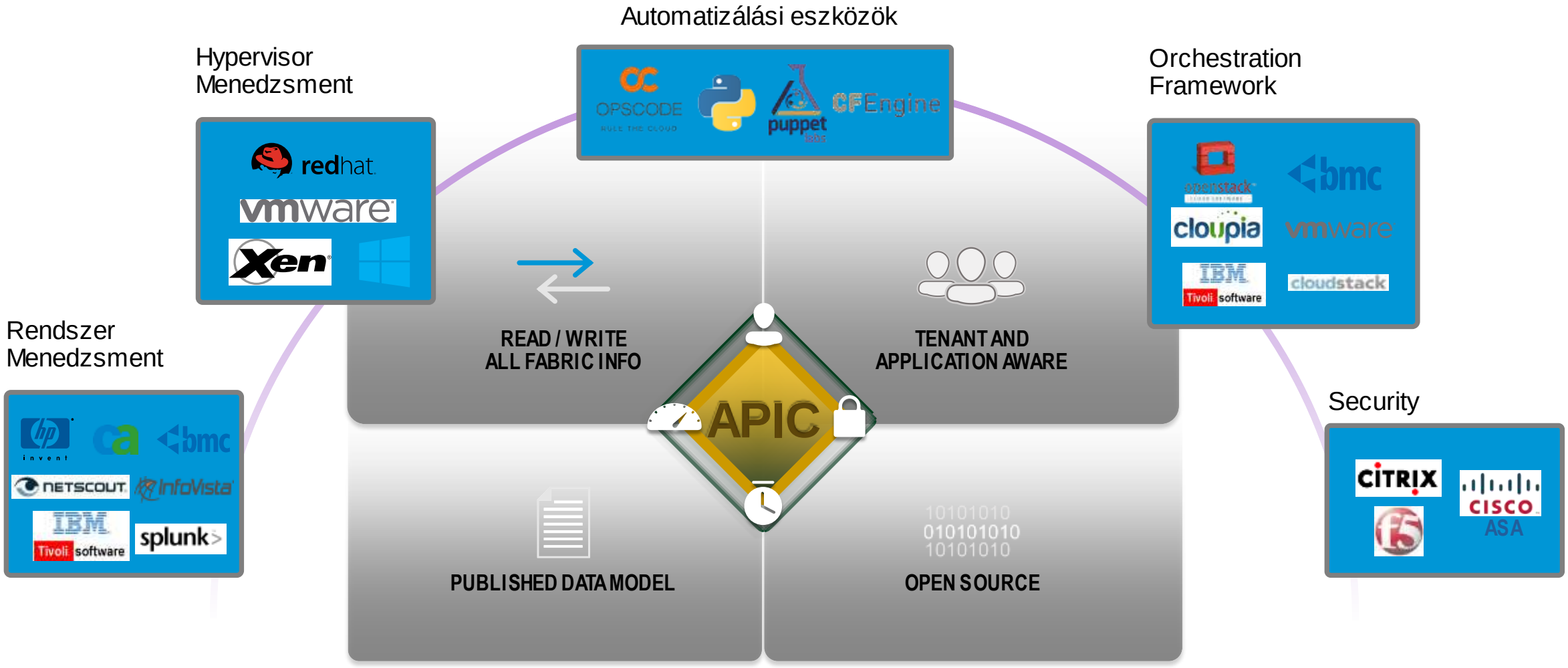
Multi-Hypervisor képes Fabric

Virtuális – Fizikai végpont Integráció

- Multi-hypervisor Fabric
  Nincs Hypervisor kötöttség
- Fizikai és virtuális VLAN, VxLAN, NVGRE technológiákat használó
- végpontok közötti Integrált gateway
- Normalizált NVGRE, VXLAN, és VLAN hálózatok közötti kommunikáció

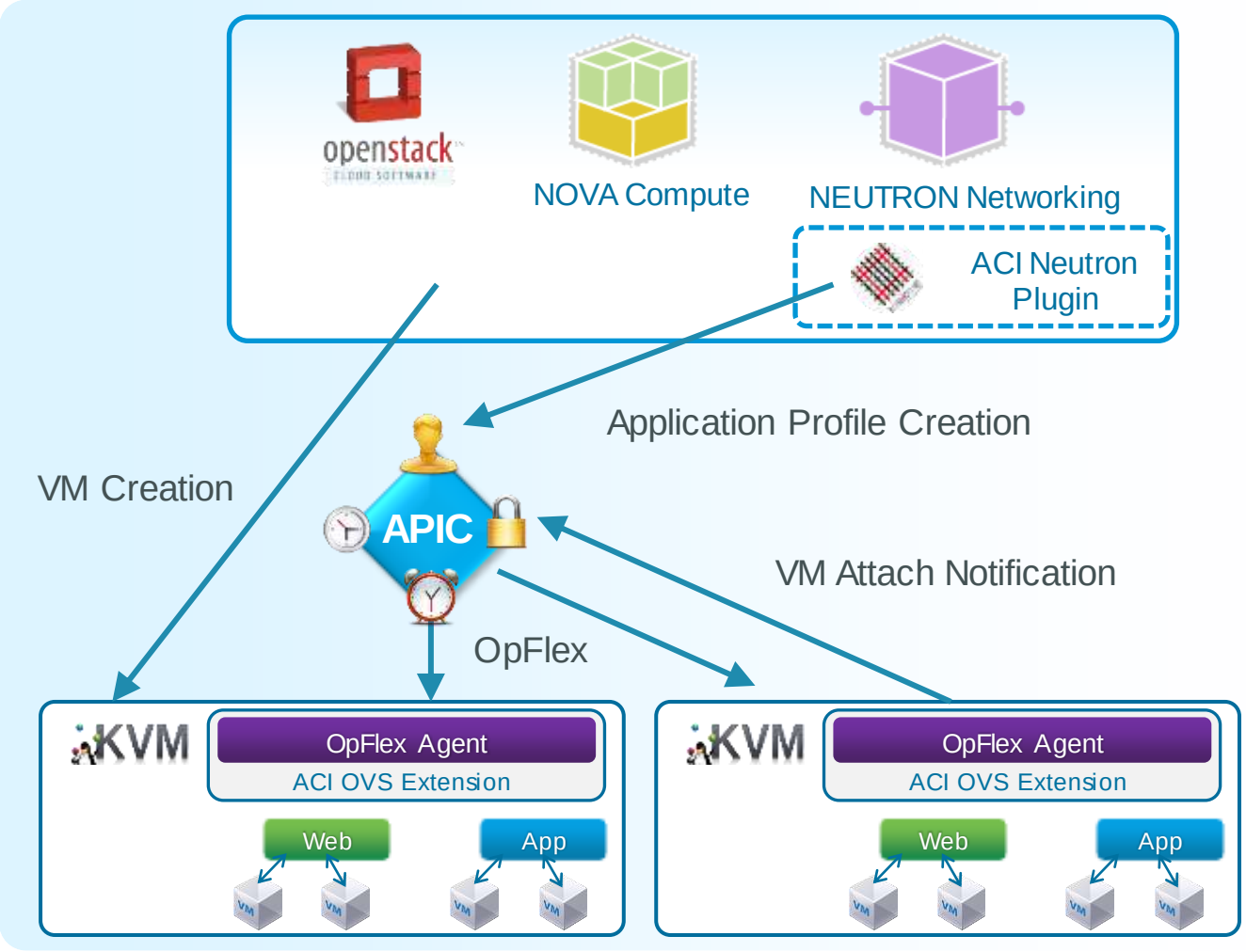


Open Ecosystem, Open API



OpenStack Integráció

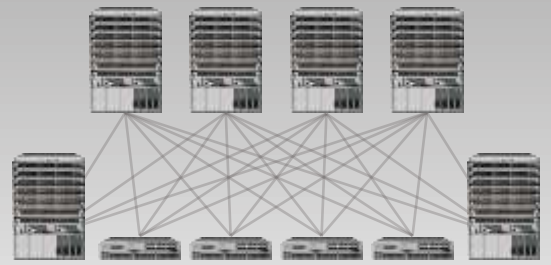
- OpenVSWitch Plugin - ACI OpFlex Agent
- APIC vezérli az OpenVSWitch Plugint
- ACI neutron plugin



ACI Fabric megvalósítás Standalone & Fabric

Közös Hardware Platform, Két működési mód

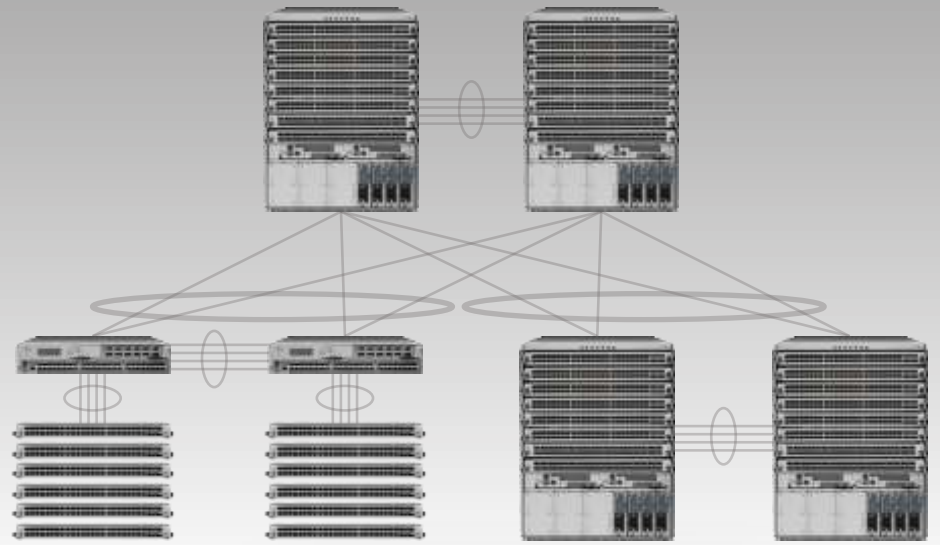
APPLICATION CENTRIC INFRASTRUCTURE



Agility and Visibility | Simplicity | Automation | Scale and Performance | Security | Open

Q2 2014

NX-OS



PROGRAMOZHATÓSÁG
40 GigE – ÁR/ÉRTÉK
Hagyományos hálózati model

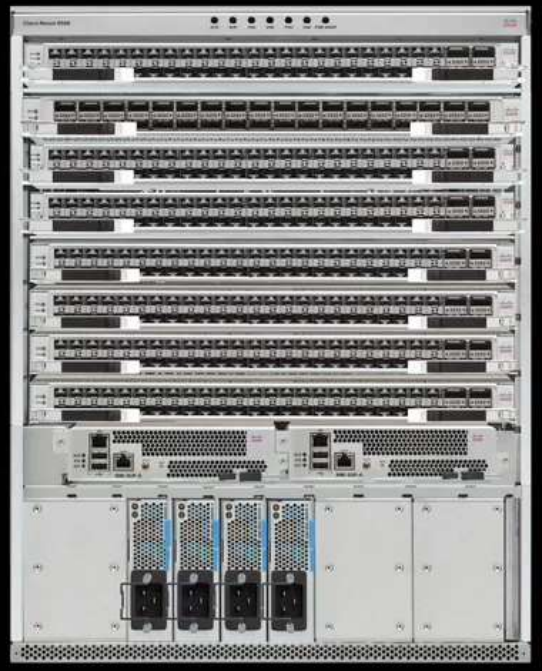
Q4 2013

Hardware, Software, ASIC és rendszer innováció

TELJESÍTMÉNY
VEZETŐ ÁR/MODUL SÁVSZÉLESSÉG
1.92 Tbps per slot
100G képes

KÖLTSÉGEK
KEDVEZŐ ÁR STRUKTÚRA
1G -> 1/10GT
10G -> 40G migráció

MERCHANT+ CUSTOM ASIC
Cisco ASIC Innováció



PROGRAMOZHATÓSÁG
JSON/XML API
Linux Container
Felhasználói app.

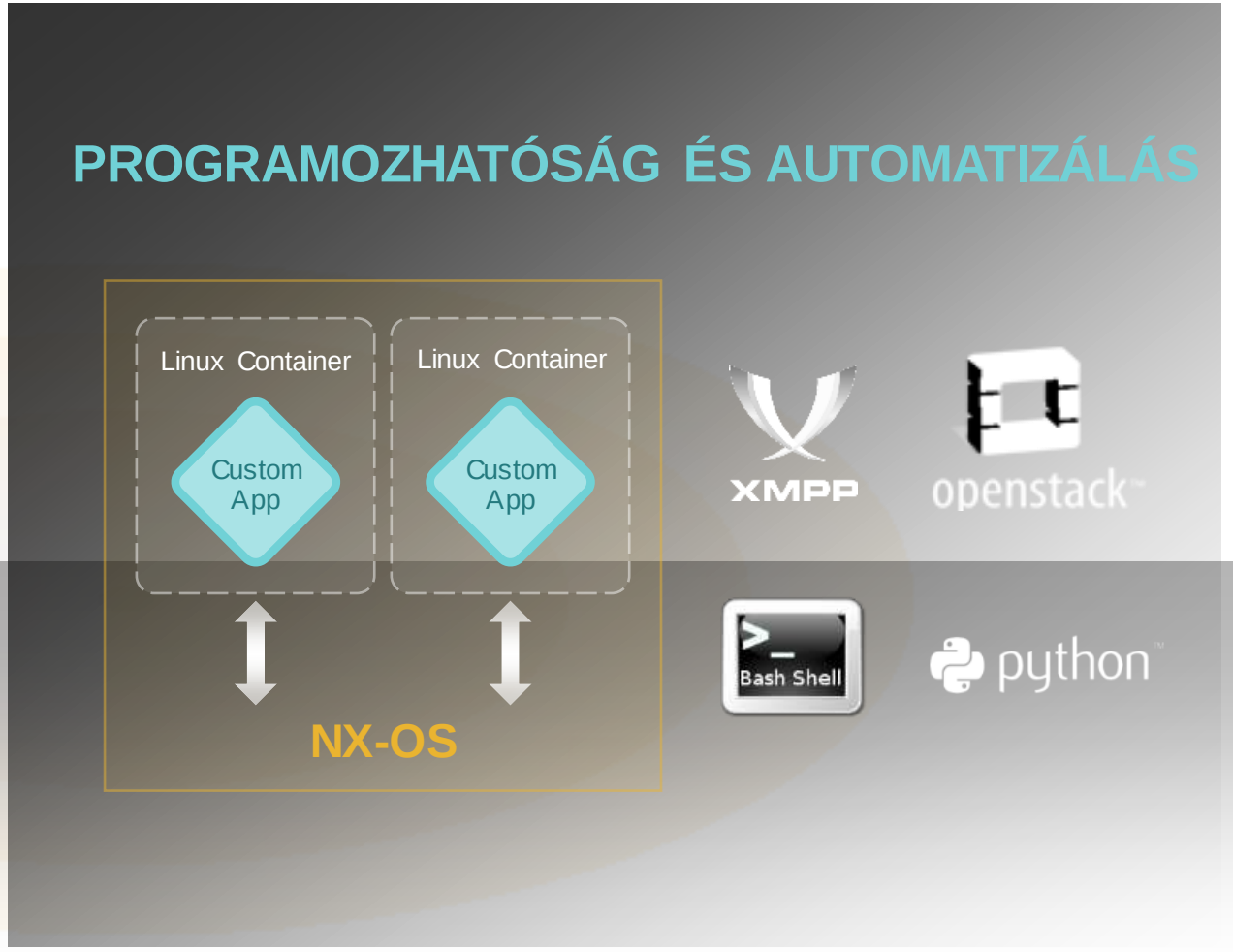
ENERGIA HATÉKONYSÁG
BACKPLANE MENTES DESIGN
15% kevesebb fogyasztás és hűtés ígény

NEXUS 9000

KÖLTSÉGEK TELJESÍTMÉNY PORT SŰRŰSÉG PROGRAMOZHATÓSÁG ENERGIA HATÉKONYSÁG

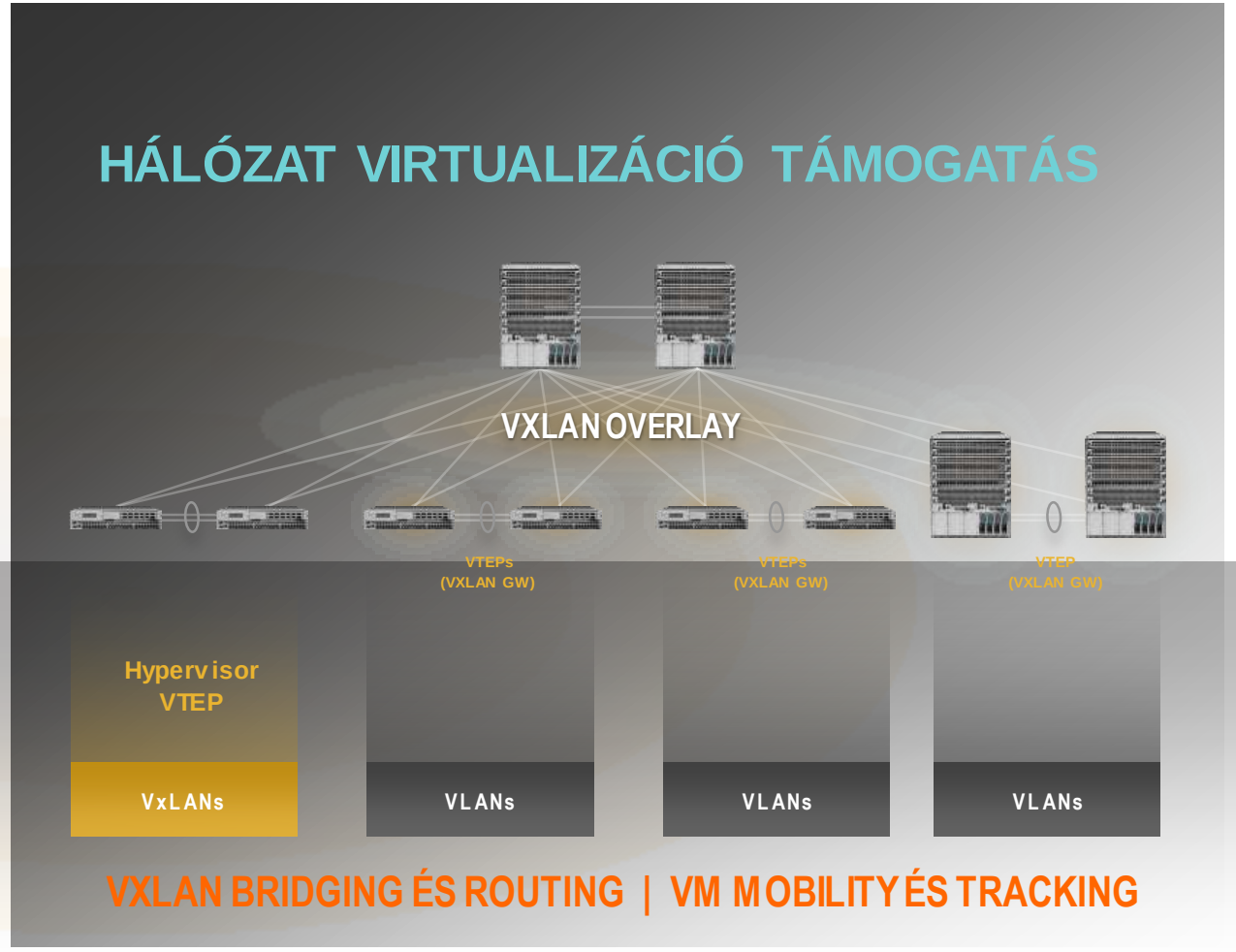
NXOS Továbbfejlesztés

Nexus 9000



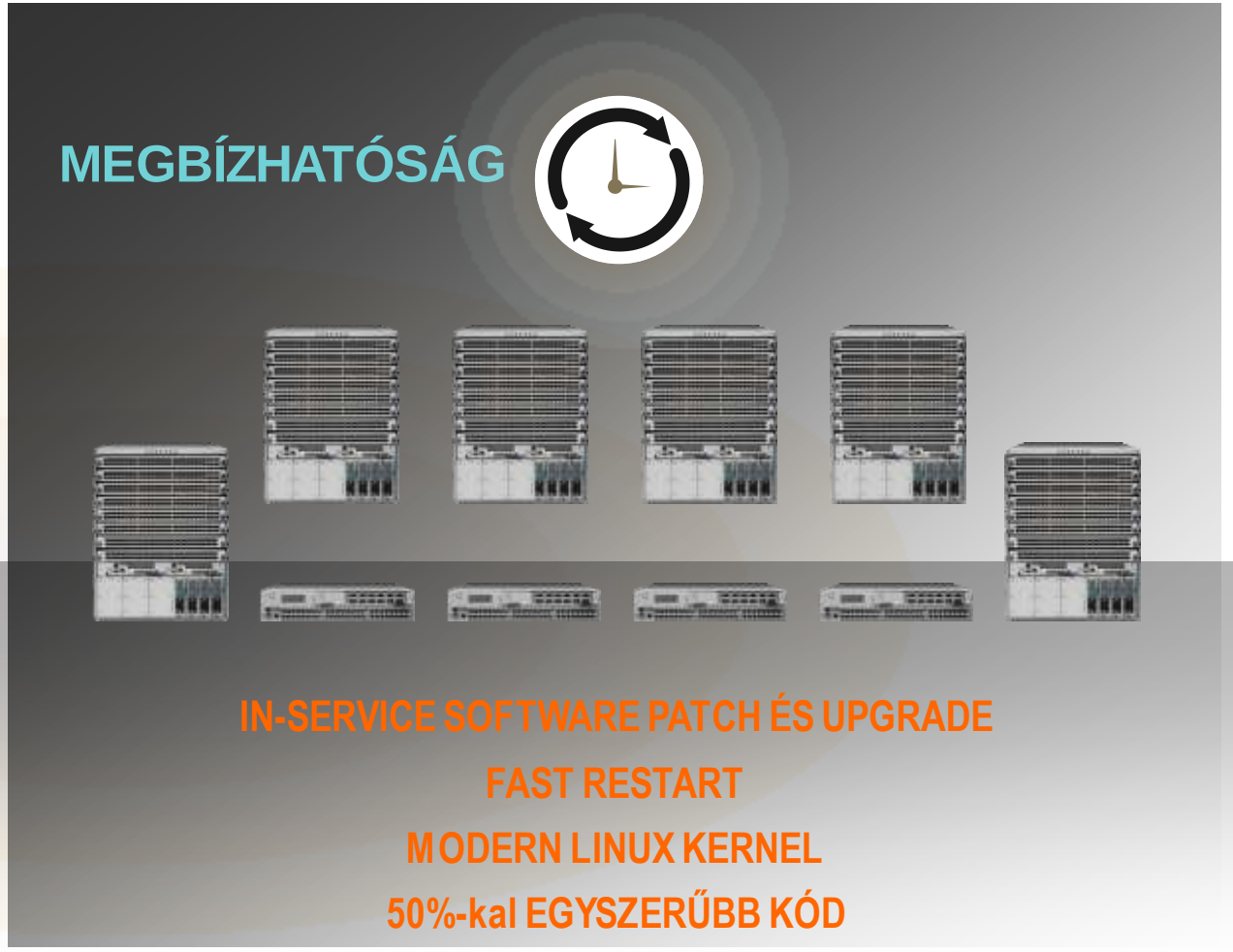
NXOS Továbbfejlesztés

Nexus 9000



NXOS Továbbfejlesztés

Nexus 9000



NXOS Továbbfejlesztés

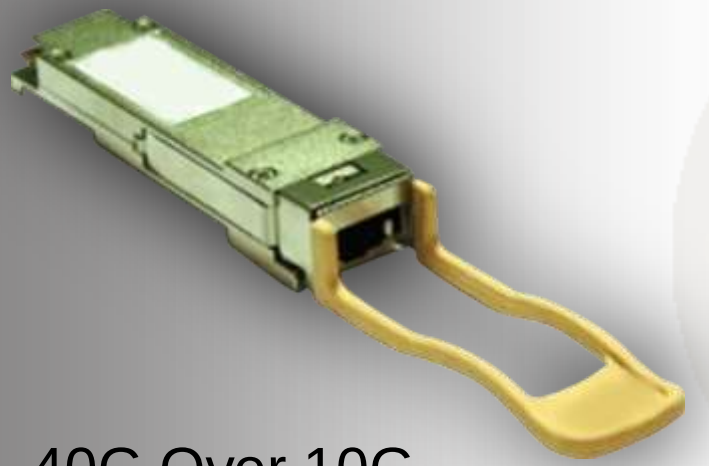
Nexus 9000

Nexus 9500 and 9300

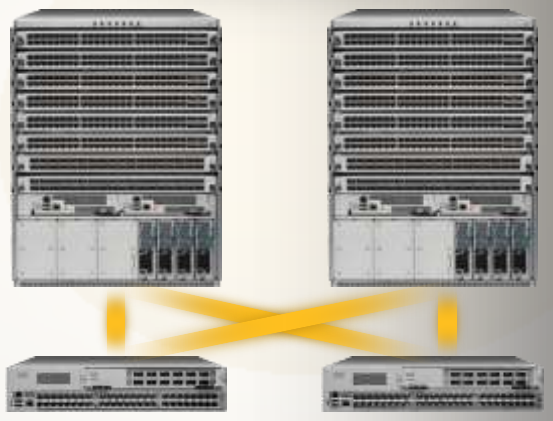
UPGRADELHETŐ
APPLICATION-CENTRIC INFRASTRUCTURE-RE

Optikai Fejlesztések

40G BiDi Optics



40G Over 10G
Multimode Fiber



JELENTŐS KÖLTSÉG MEGTAKARÍTÁS A KÁBELEZÉSI INFRASTRUKTÚRA VÁLTOZATLAN MARAD

Roadmap

Configuration Management

Orchestration Frameworks

DevOps Toolkit

40G Aggregation
NX-OS
36 QSFP+

OCTOBER

End-of-Row Access
and 10G Aggregation
48 10GT + 4 QSFP+
48 10GF + 4 QSFP+

Top–of-Rack
96 10GT + 8 QSFP+
48 10GF + 12 QSFP+

Q1CY14

40G Fixed and
Modular Spine Fabric

Top–of-Rack
Fabric

Q2CY14

Thank you.