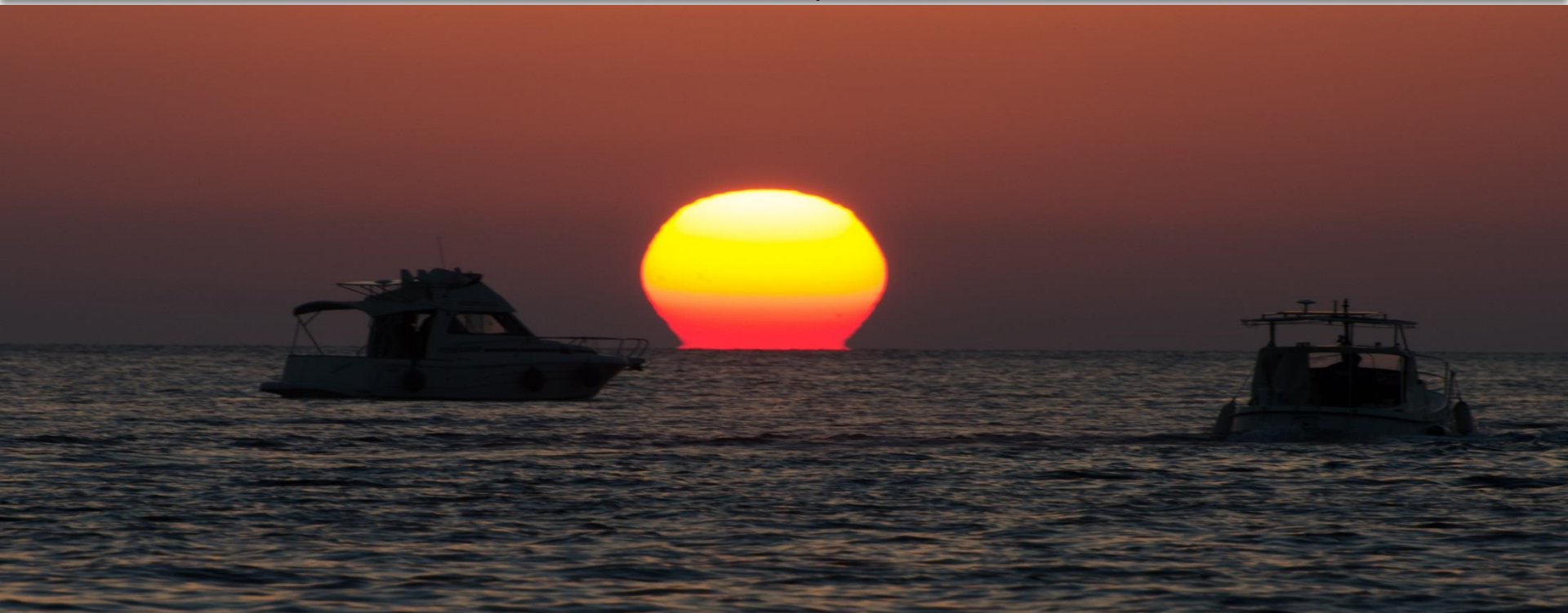
Élet Spanning Tree nélkül

Balla Attila

CCIE #7264

balla.attila@kapsch.net



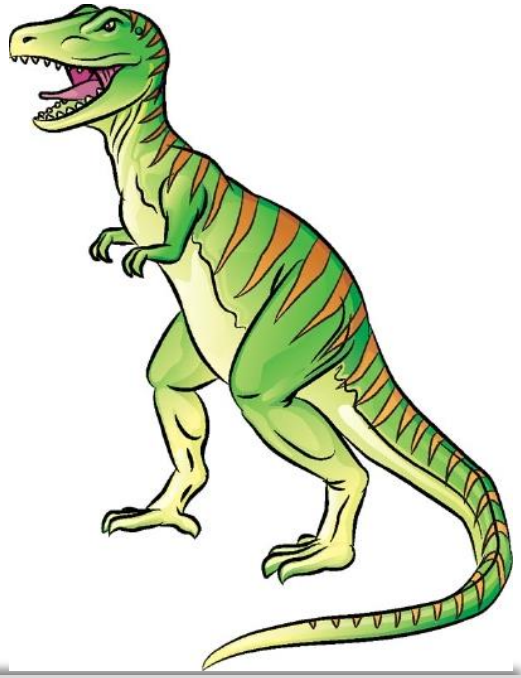
Tartalom

- Emlékeztető
- vPC
- FabricPath
- Melyiket, mikor, hol?

Emlékeztető

Spanning Tree nem a mai technológiákhoz lett fejlesztve

- Sebesség, hibatűrés, számítási kapacitás
- Problémák: kihasználtság, konvergencia, skálázhatóság
- 1985.: Első implementáció, DEC
- 1990.: IEEE 802.1D
- 2001.: IEEE 802.1w RSTP
- 2005.: IEEE 802.1s -> 802.1Q: MSTP



Korábbi STP témakörök

2007.: STP finomhangolása

2008.: Hálózati Megoldások Szerverkonszolidációs Környezetben

- Layer2 Multipathing 3 fázisa
  1. MAC Pinning
  2. MCEC
  3. L2 ISIS

2012.: Veszprém, Layer2 hálózatok tervezése STP nélkül

- MultiChassis EtherChannel

Technológiák

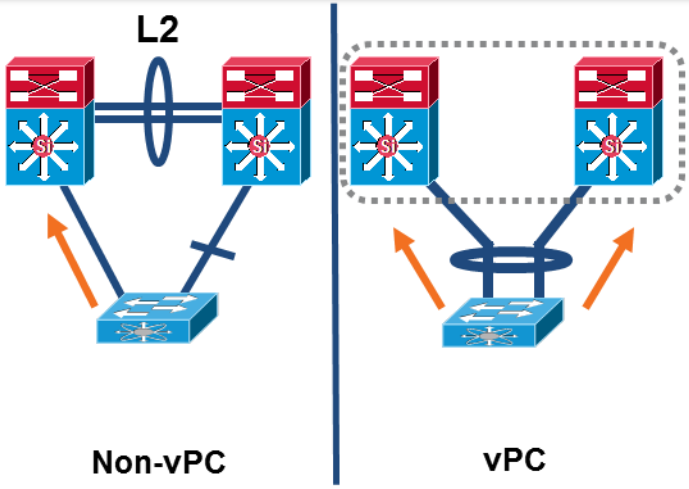
Virtual Port Channel

- MultiChassis EtherChannel
- UNI
- Redundancia növelés
- Kapacitás növelés (nincs STP)
- Gyors konvergencia

FabricPath

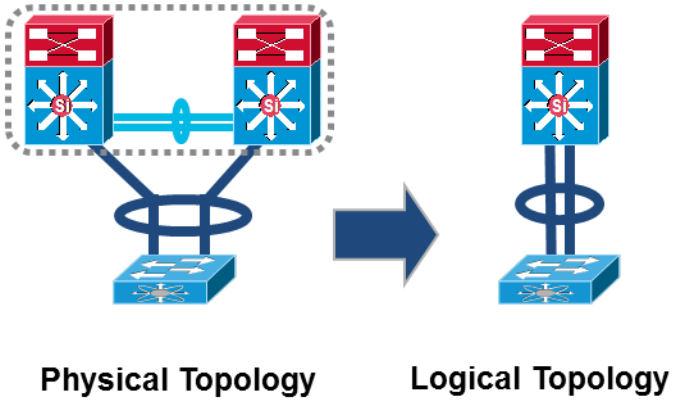
- IS-IS L2 routing
- ECMP routing
- NNI
- Kapacitás növelés (nincs STP)
- Redundancia
- Gyors konvergencia

vPC áttekintés



Bi-sectional BW with vPC

- EtherChannel két különböző switch-hez
- Nincs STP által blokkolt port

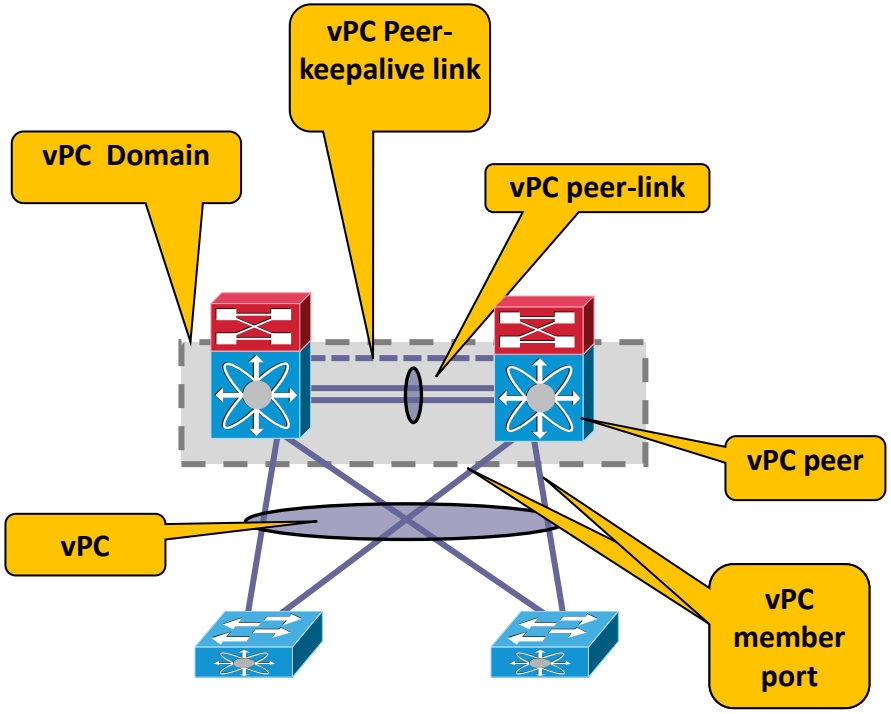


Virtual Port Channel

- Független vezérlési sík
- Egyszerű topológia
- Gyors konvergencia
- vPC Layer2 EtherChannel!

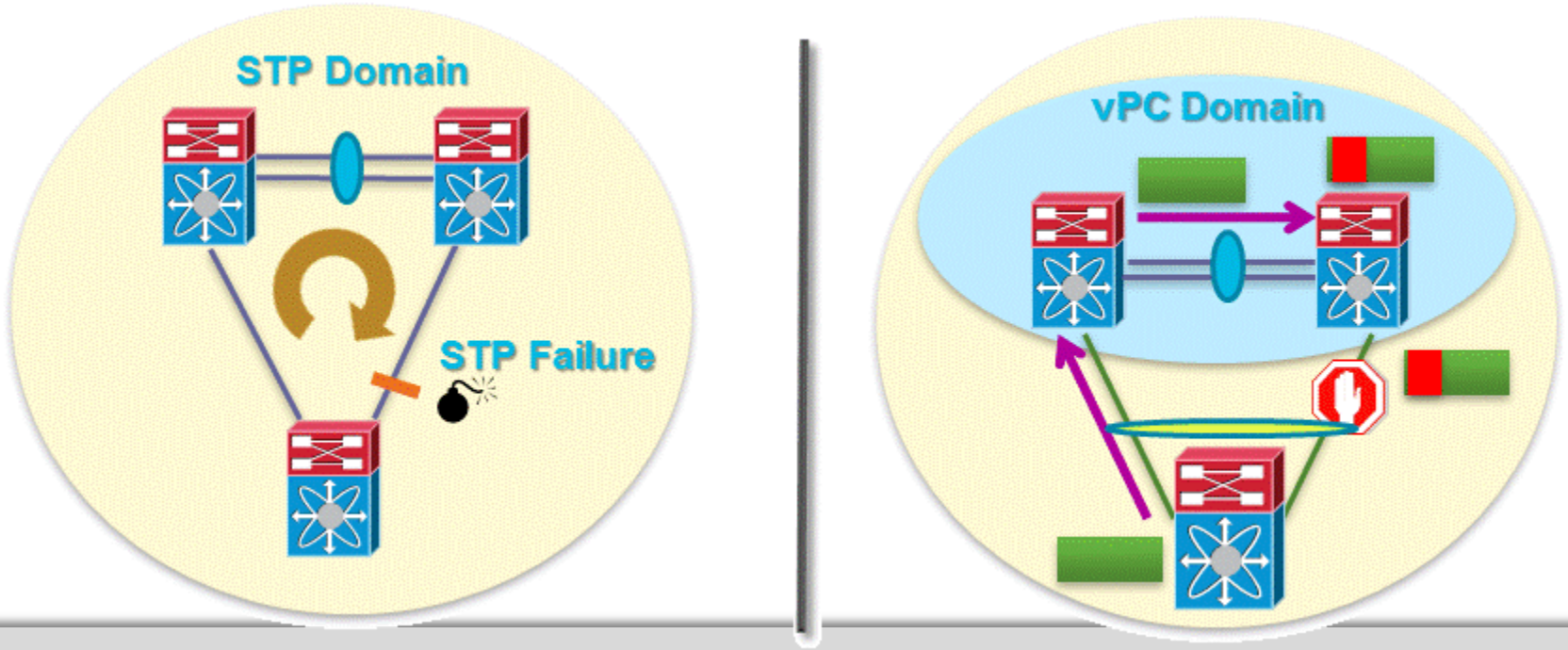
vPC terminológia

- vPC Domain
  - két vPC képes switch
- vPC peer
  - A domain egyik tagja
- vPC member port
  - Az MCEC egy portja
- vPC
  - PortChannel az „access” switch felé
- vPC peer-link
  - vPC peer-ek között, szinkronizáció
- vPC peer-keepalive link
  - Keepalive link a peer-ek között



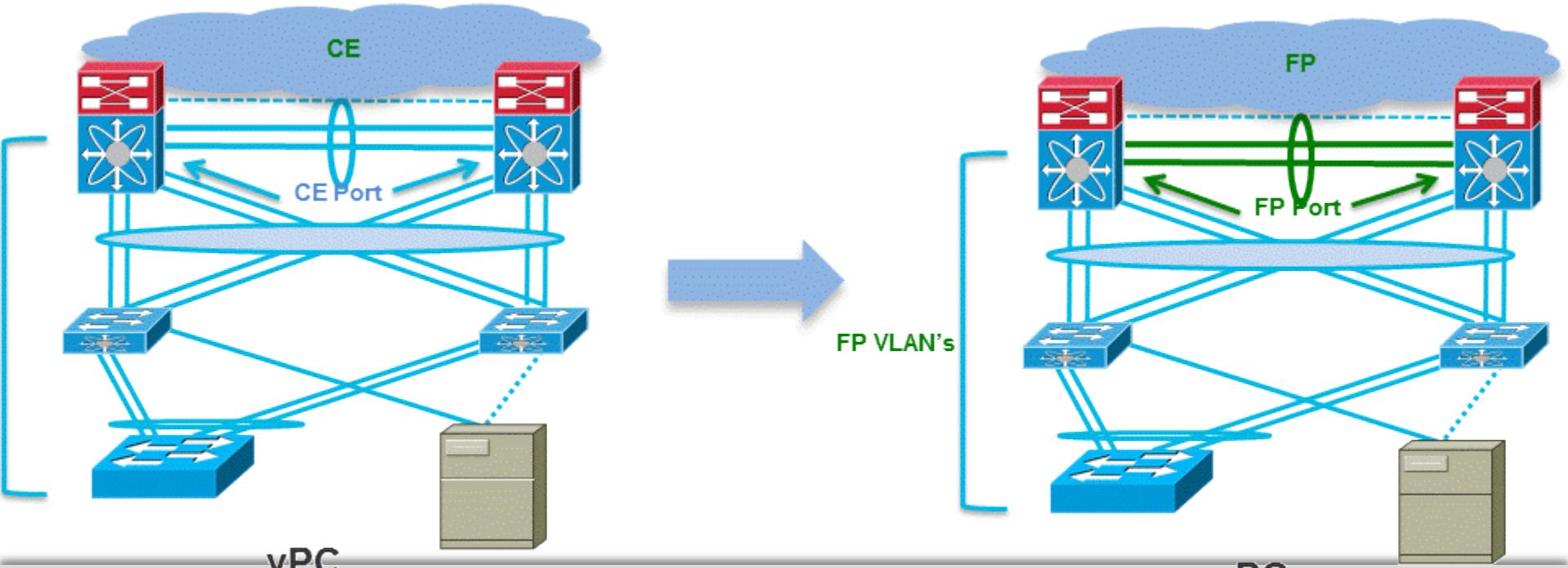
Adatsík hurok elkerülés

- vPC peer-ek lokálisan továbbítják a forgalmat
- Peer-linken tipikusan nincs adatforgalom
- A Peer-linken levő forgalmat megjelöljük



vPC kiterjesztése: vPC+

- Fizikai architektúra ugyanaz
- Funkciók/Koncepció ugyanaz
- Virtual Switch ID
- További előnyök



vPC Hardware támogatás

Platform	vPC peer link	vPC interface
Nexus 7k	✓ kivéve GE interface	✓
Nexus 6k	✓	✓
Nexus 5k	✓	✓

vPC Domain felépítése

Alapvető lépések (sorrend fontos!)

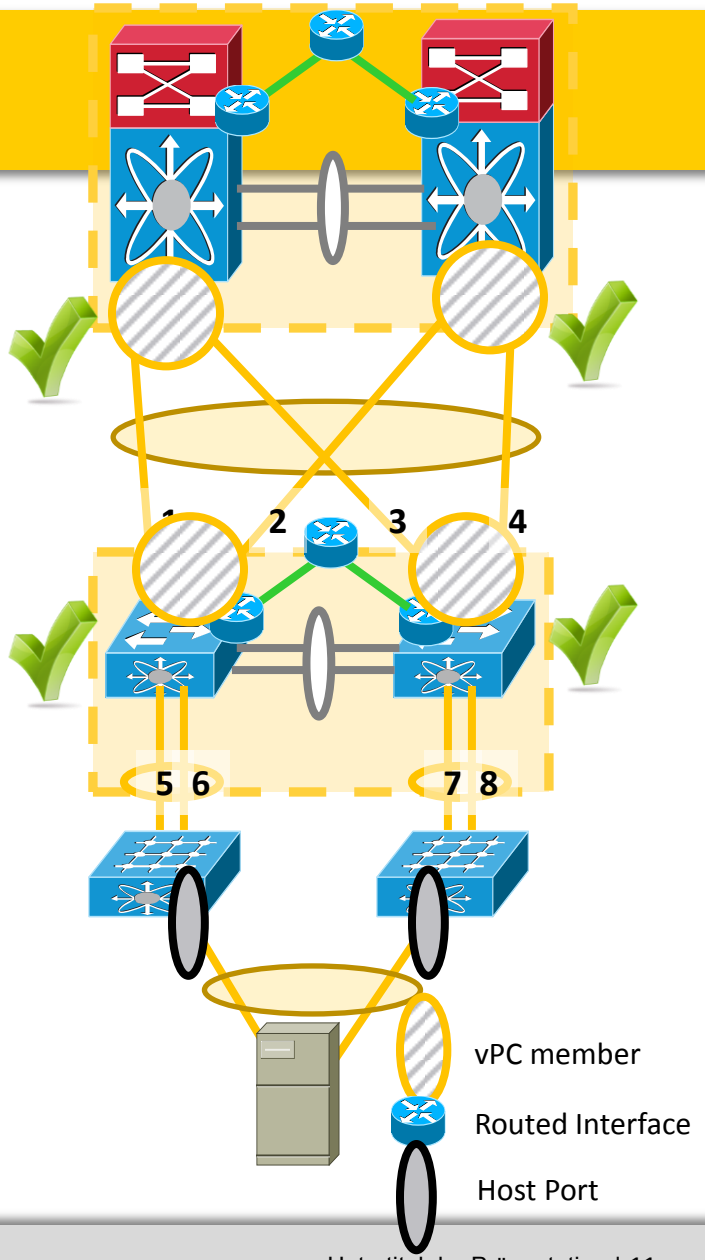
Domain létrehozása

vPC Peer Keepalive összeköttetés

vPC Peer link összeköttetés

Port-Channel-ek és vPC-k felhasználása

Konzisztens konfiguráció

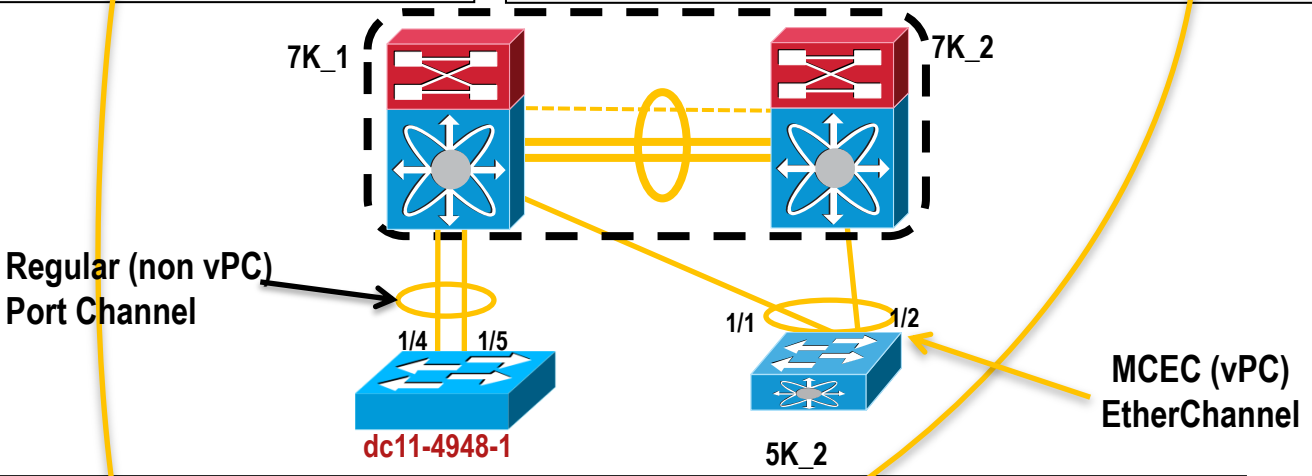


vPC Domain – vPC csatlakozás

LACP szomszéd a virtuális System ID-t látja vPC esetén

```
7K_1# sh vpc role
<snip>
vPC system-mac                  : 00:23:04:ee:be:14
vPC system-priority             : 1024
vPC local system-mac            : 00:0d:ec:a4:53:3c
vPC local role-priority         : 1024
```

```
7K_2 # sh vpc role
<snip>
vPC system-mac                  : 00:23:04:ee:be:14
vPC system-priority             : 1024
vPC local system-mac            : 00:0d:ec:a4:5f:7c
vPC local role-priority         : 32667
```

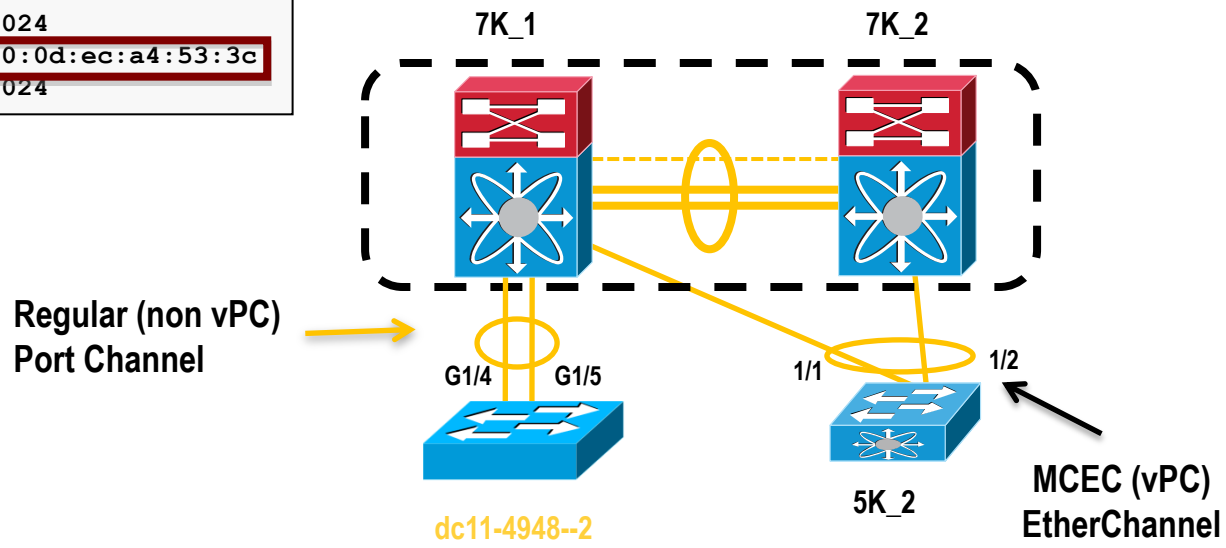



```
5K_2#sh lacp neighbor
<snip>
                  LACP port                          Admin  Oper   Port     Port
Port      Flags   Priority  Dev ID           Age     key    Key    Number   State
E1/1      SA      32768     0023.04ee.be14   6s      0x0    0x801E 0x4104   0x3D
E1/2      SA      32768     0023.04ee.be14   21s     0x0    0x801E 0x104    0x3D
```

vPC Domain – nem vPC csatlakozás

„Local system-mac”-t használja minden nem vPC alapú PDU-hoz (LACP, STP, …)

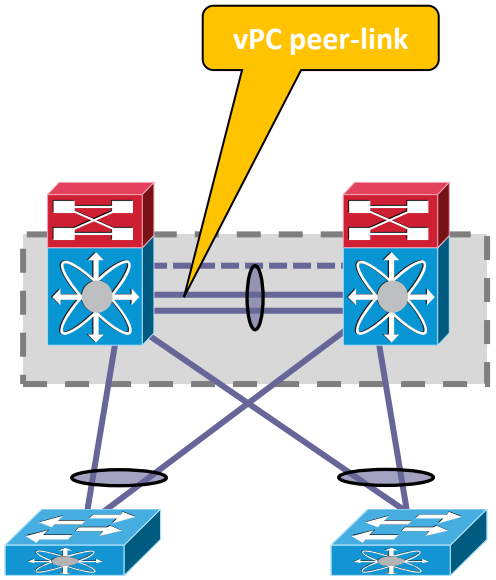
```
7k_1 # sh vpc role
<snip>
vPC system-mac                  : 00:23:04:ee:be:14
vPC system-priority             : 1024
vPC local system-mac            : 00:0d:ec:a4:53:3c
vPC local role-priority         : 1024
```




```
dc11-4948-2#sh lacp neighbor
<snip>
                  LACP port                          Admin   Oper    Port    Port
Port      Flags   Priority  Dev ID           Age     key     Key     Number  State
Gi1/4     SA      32768     000d.eca4.533c    8s     0x0     0x1D    0x108   0x3D
Gi1/5     SA      32768     000d.eca4.533c    8s     0x0     0x1D    0x108   0x3D
```

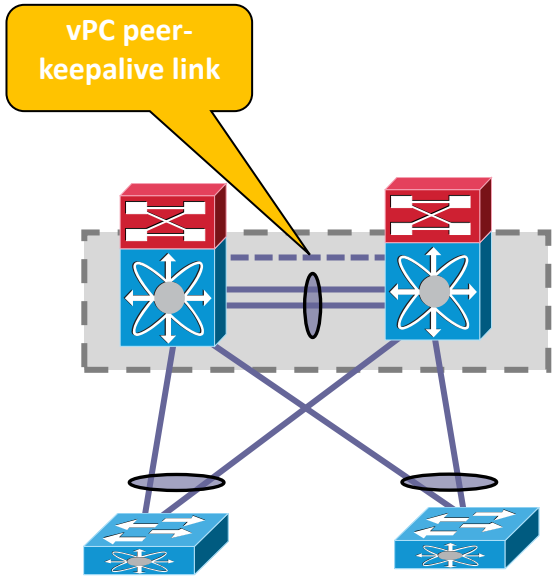
vPC Peer Link

- 802.1Q trunk, Cisco Fabric Services
- Alapvetően csak flood-olt forgalmak
  - STP BPDU, HSRP hello, IGMP update
- Árva portok forgalma
- Erősen ajánlott a 2x10GE
  - Ugyanolyan típusú modulokon



vPC Peer Keepalive Link

- Heart beat üzenetek
- Aktív/aktív detektálás
- IP csomag
  - UDP:3200, 96 byte
  - Timer: 1s
- Összeköttetés típusa
  - Dedikált összeköttetés (GE – EC)
  - Mgmt0 interface-k
  - L3 infratruktúrán keresztül



vPC hibakezelés

Peer-link megy Down-ba

- Keepalive ellenőrzés
- Elsődleges vPC peer nem csinál semmit
- Másodlagos vPC peer blokkolja a vPC portokat
- A másodlagos vPC peer árva eszközei izolálódnak

Peer-keepalivelink és a Peer-link megy Down-ba

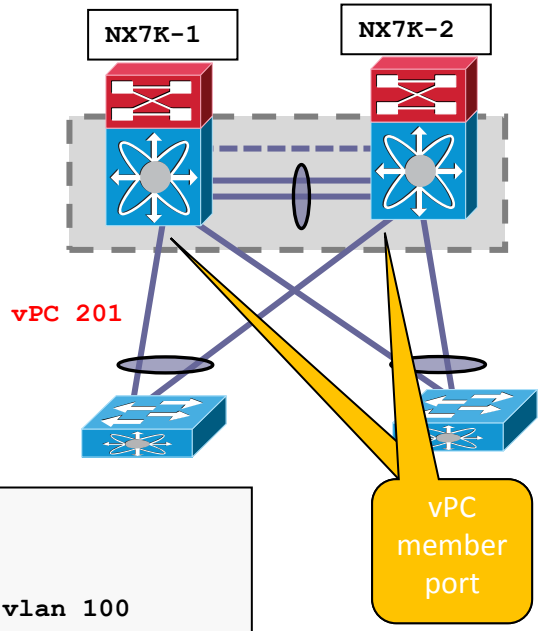
- Elsődleges elsődleges marad
- Másodlagos is elsődleges lesz
- Dual-Active mód
- STP lép életbe
- Forgalom kiesés

vPC konzisztencia ellenőrzés

- Peerlink-en futó CFS protokoll kiterjesztése
- Konfiguráció ellenőrzés – önálló vezérlők miatt
- Két típus
  - Kötelező, Type 1 => suspend állapot (5.2 óta csak a secondary peer-en)
  - Opcionális, Type2 => syslog üzenet
- Type1
  - Port-Channel mode, Link speed, Trunk mode, STP Port type, Loop/Root Guard, MTU
- Type2
  - MAC aging timer, BPDU Filter/Guard, QoS, Port Security, ...

vPC

- vPC member port
- Nexus 7k esetén akár 2x16 aktív port




```
NX7K-1 :
interface port-channel201
  switchport mode trunk
  switchport trunk native vlan 100
  switchport trunk allowed vlan 100-105
  vpc 201
```

```
NX7K-2 :
interface port-channel201
  switchport mode trunk
  switchport trunk native vlan 100
  switchport trunk allowed vlan 100-105
  vpc 201
```

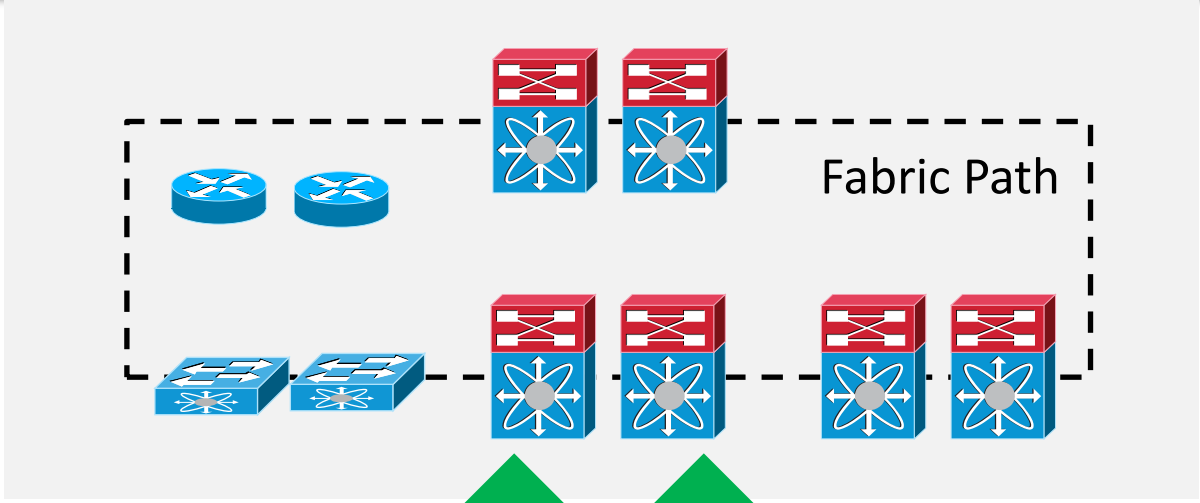
vPC konvergencia idők

- vPC link member failure -> subsecond
- vPC port-channel failover -> subsecond
- vPC Peer-link Failure -> subsecond
- vPC peer-keep-alive Failure -> hitless
- vPC primary/secondary device failure -> subsecond
- vPC Supervisor Failover/Switchover -> hitless
- vPC ISSU device Upgrade/Downgrade -> hitless

N7k, NX-OS: 5.2, 6.0, 6.1

http://www.cisco.com/en/US/docs/switches/datacenter/sw/verified_scalability/b_Cisco_Nexus_7000_Series_NX-OS_Verified_Scalability_Guide.html

FabricPath



Switching

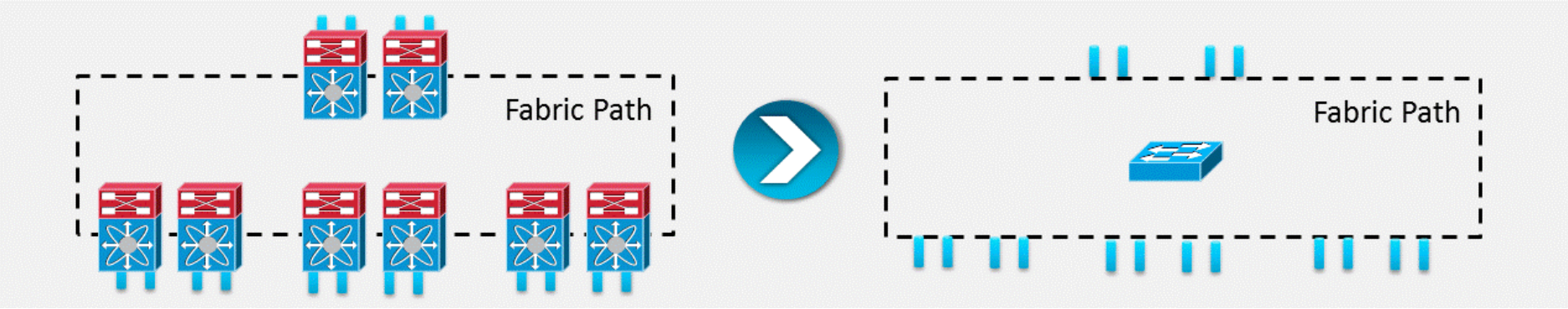
- Könnyű konfigurálás
- Plug & Play
- Rugalmas létesítés

Routing

- Multi-pathing (ECMP)
- Gyors konvergencia
- Skálázható

FabricPath

- Layer2 mindenhol
- STP limitációi nélkül
- Switching & Routing előnyei
  - Könnyű konfigurálás
  - ECMP, konvergencia, skálázhatóság
- Kívülről egy L2 switch
- Belülről egy protokoll köti össze az eszközöket

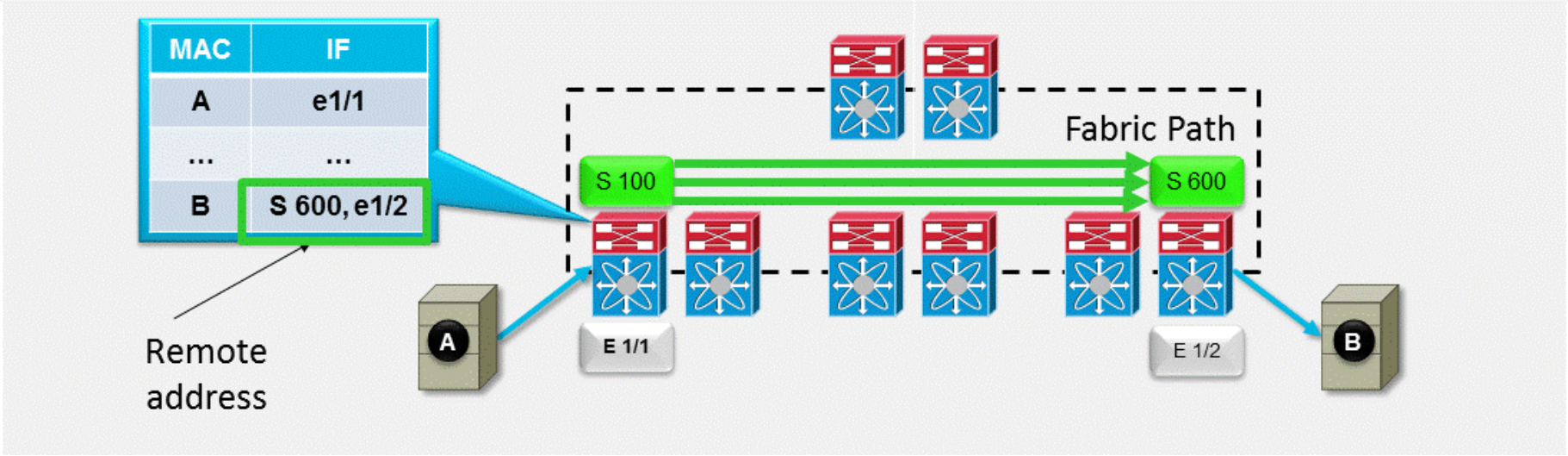


FabricPath

- Bármilyen topológia
- Támogatott eszközök, interface-k
- Nincs STP a fabricpath-on belül

```
N7K(config)# interface ethernet 1/1
N7K(config-if)# switchport mode fabricpath
```

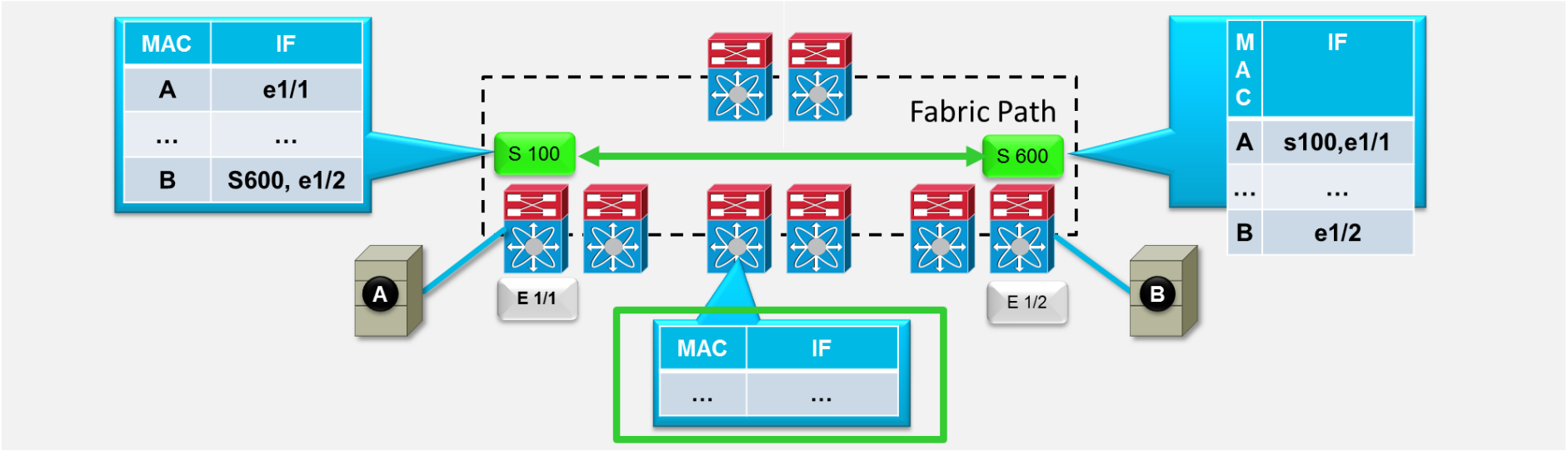
Optimális adattovábbítás



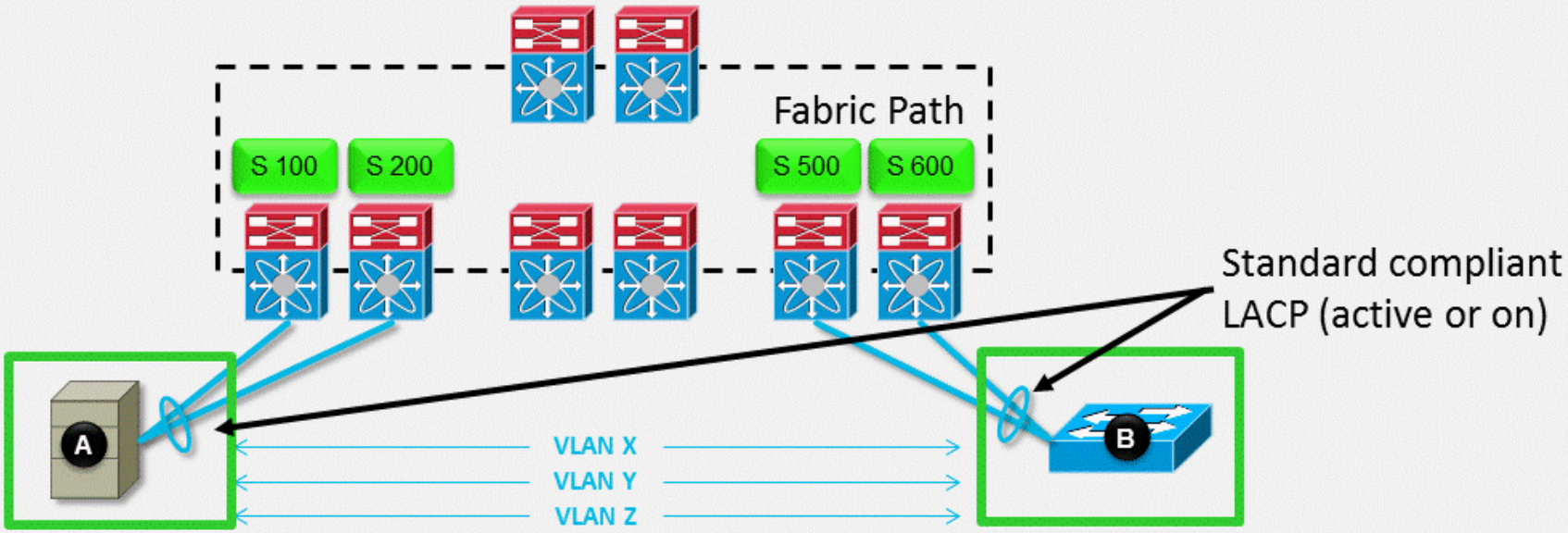
- Egy lookup az ingress oldalon azonosítja a kimenő portot
- Shortest Path, ECMP akár 256! Linken
  - Első szint: FabricPath Core link selection (L3/L4 mezők alapján)
  - Második szint: Port-Channel (src-dst ip)
- Skálázhatóság

Skálázhatóság

- Conversational learning
- MAC address tábla csak az unicast forgalomban résztvevők címét tartalmazza
- Elméletileg végtelen számú host kapcsolódhat



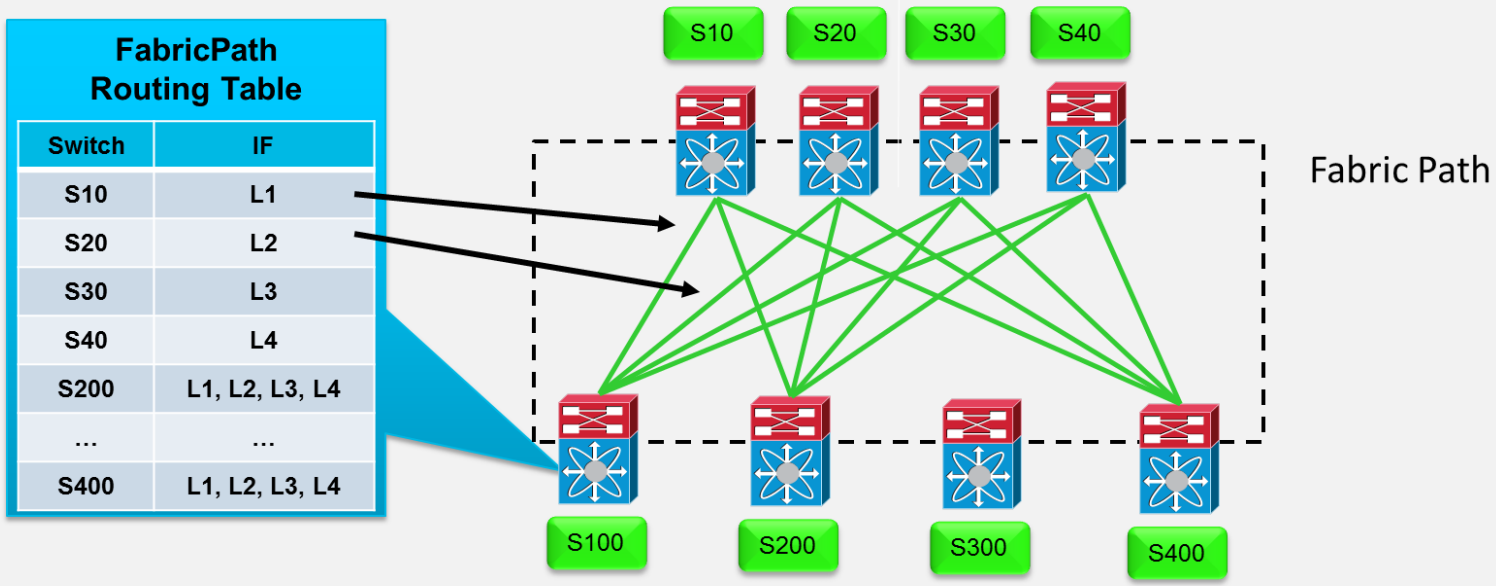
Layer2 integráció – VPC+



- VLAN-ok korlátlan kiterjesztése
- MCEC csatlakozás a Fabrichoz
- Virtuális Fabric Switch-ID a vPC-n belül

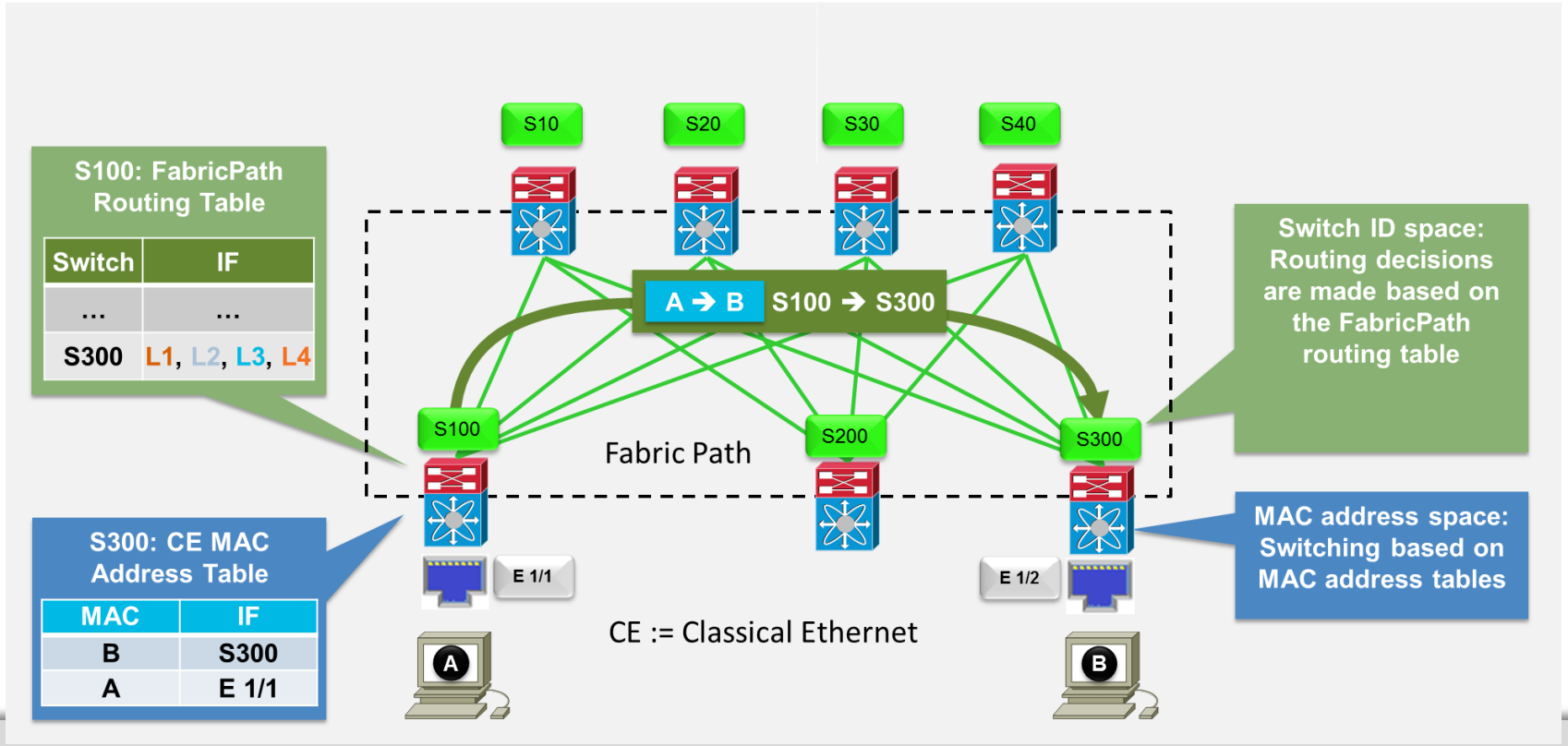
Új Vezérlési Sík

- IS-IS automatikusan rendel címet minden FP switch-hez
- Kiszámolja a legrövidebb utakat
- ECMP támogatás

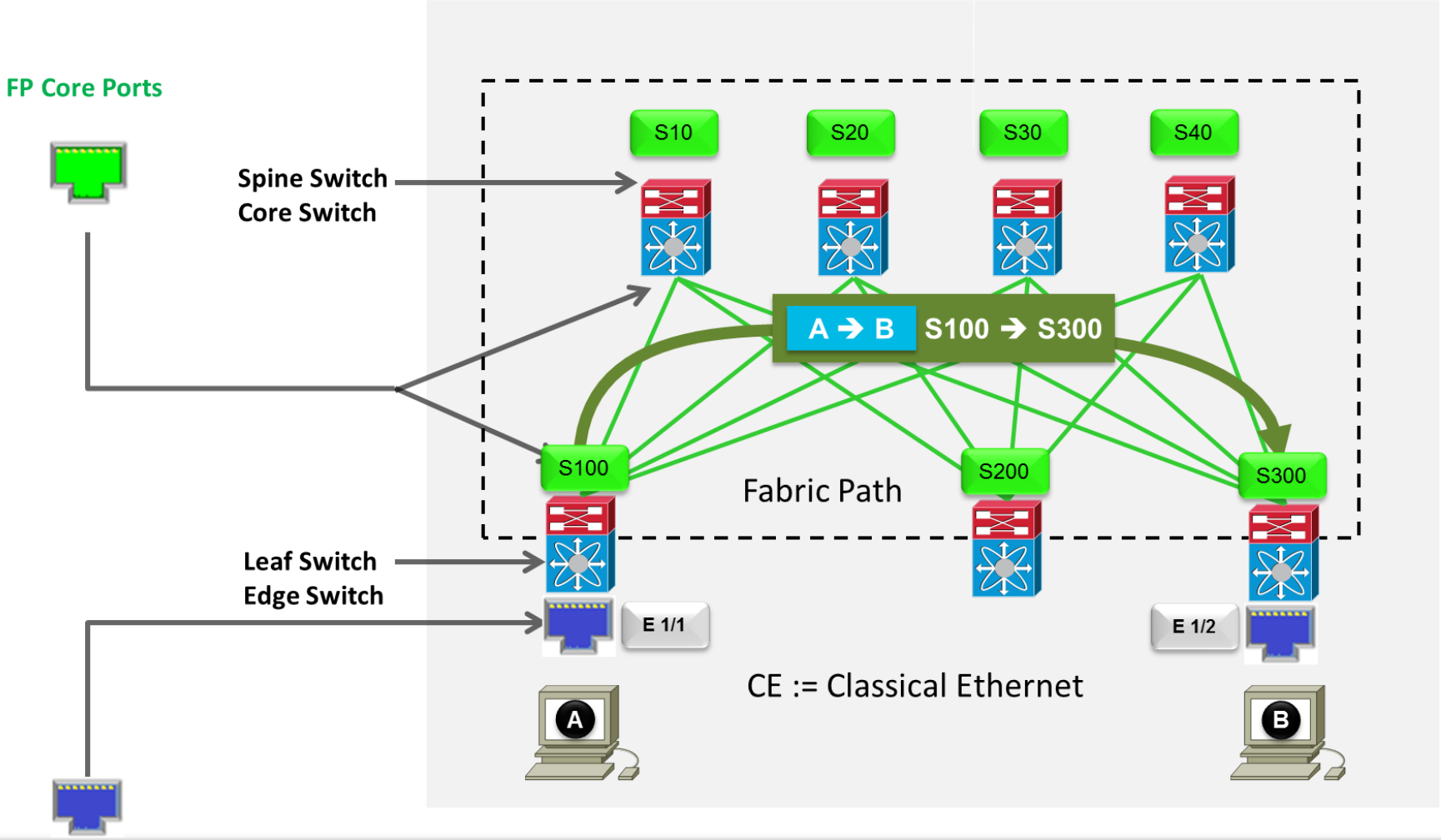


Továbbítási táblák

- MAC cím/Switch ID karbantartása a hálózat szélén
- Csomagok enkapszulálása Fabric-ba



FabricPath Terminológia



FabricPath Vezérlési Sík

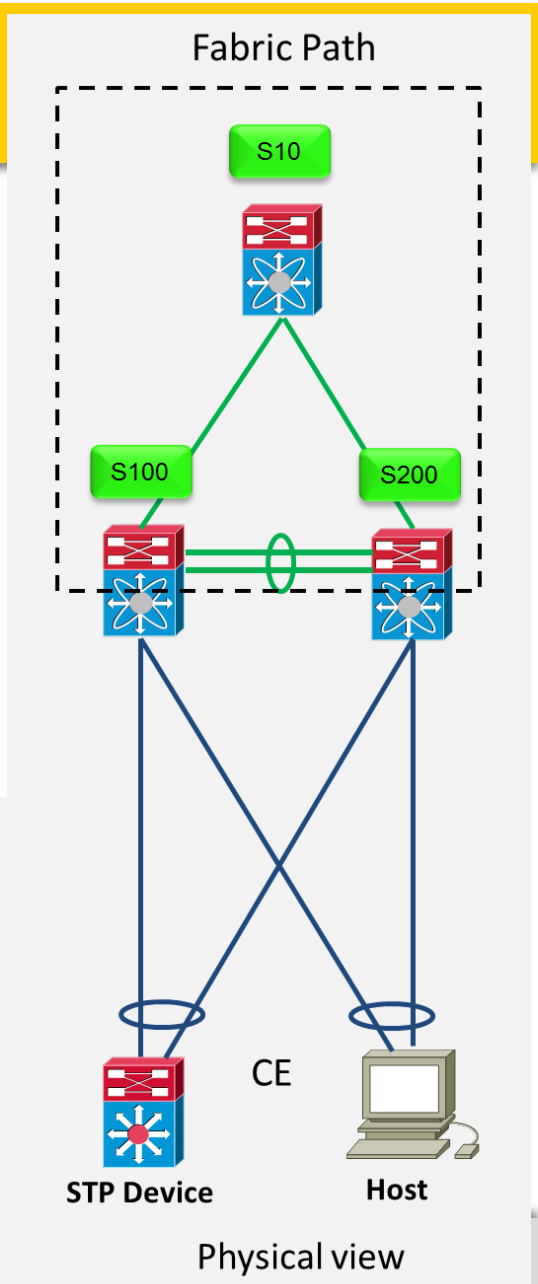
- Plug & Play
- IS-IS automatikusan hozzárendeli a címet minden egyes elemhez
- Shortest Path számítás, ECMP támogatás
  - Link cost, 400Gbps a referencia bw
- Megbízható összeköttetés szükséges
  - DWDM esetén automatikus lézer kikapcsolás
  - Időzítők hangolása nem javasolt
- Egy FP switch-ben két tábla:
  - FP routing table, MAC address table

Hardware támogatás

- ASIC támogatás szükséges
- Nexus 7000, F kártyák
  - Csak az F kártyák, vagy F kártyához csatlakozott FEX
  - FP Core és FP Edge is csak F kártyán
  - vPC+ is csak F-s kártyán működik
- Nexus 6000
- Nexus 5000
  - Nem kell L3 modul

VLAN & vPC+

- vPC domain alatt fabricpath switch-id konfigurálás
- VLAN átvitel engedélyezés
- Root priority

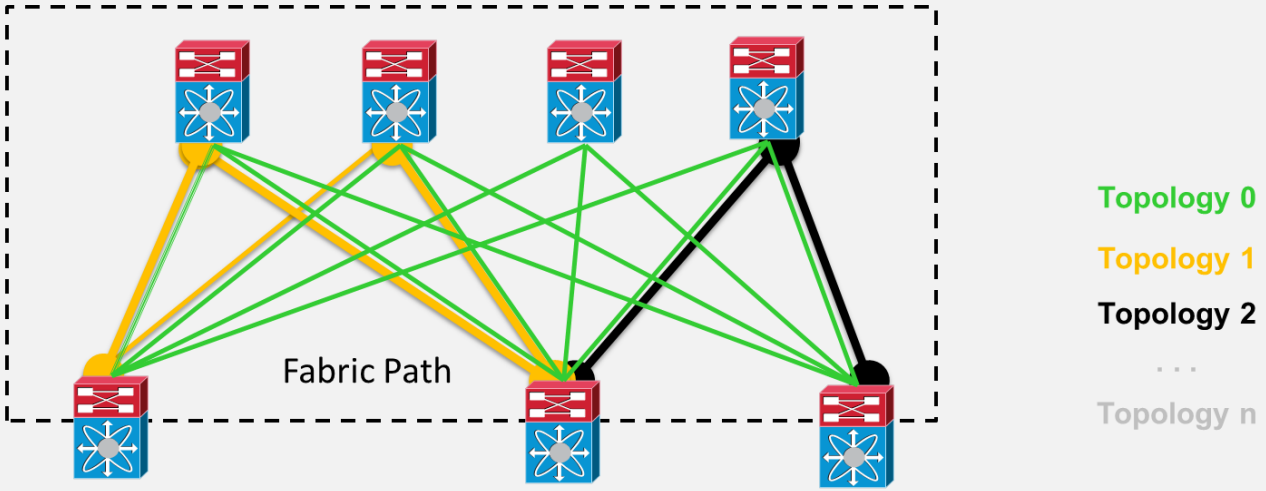



```
S100(config)# vlan 42
S100(config-vlan)# mode ?
  ce          Classical Ethernet VLAN mode
  fabricpath  Fabricpath VLAN mode

S100(config-vlan)# mode fabricpath
S100(config)# spanning-tree pseudo-information
  vlan 42 root priority
```

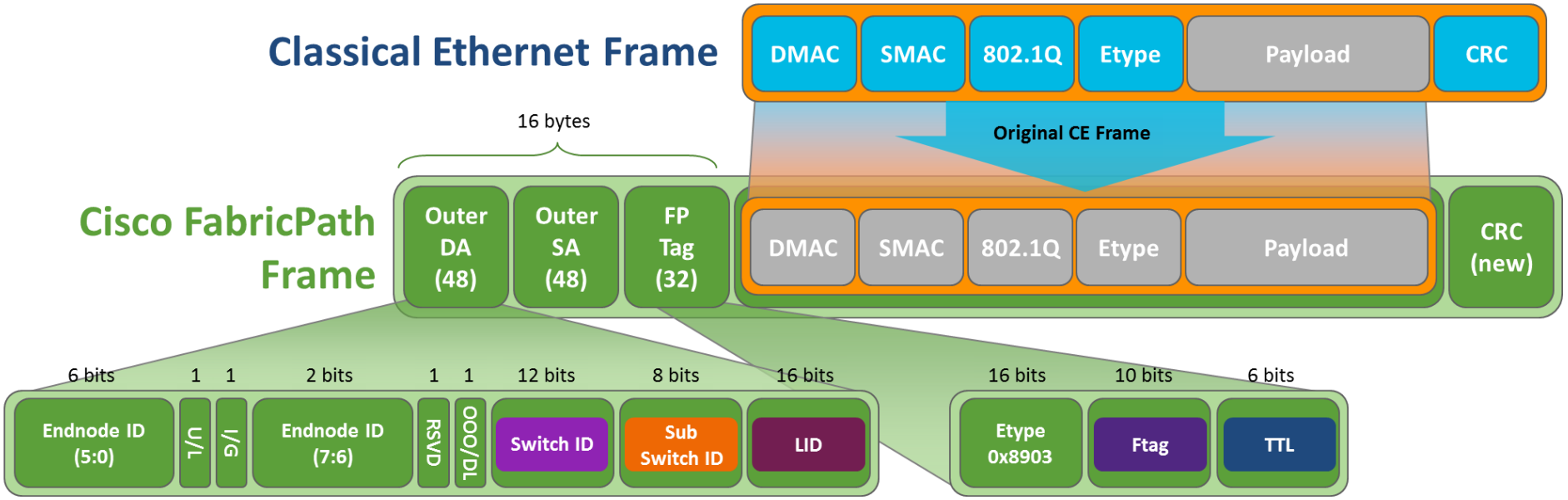
Multi-Topology támogatás

- Nexus 5k-n 2 topológia
- Nexus 7k-n 8 topológia

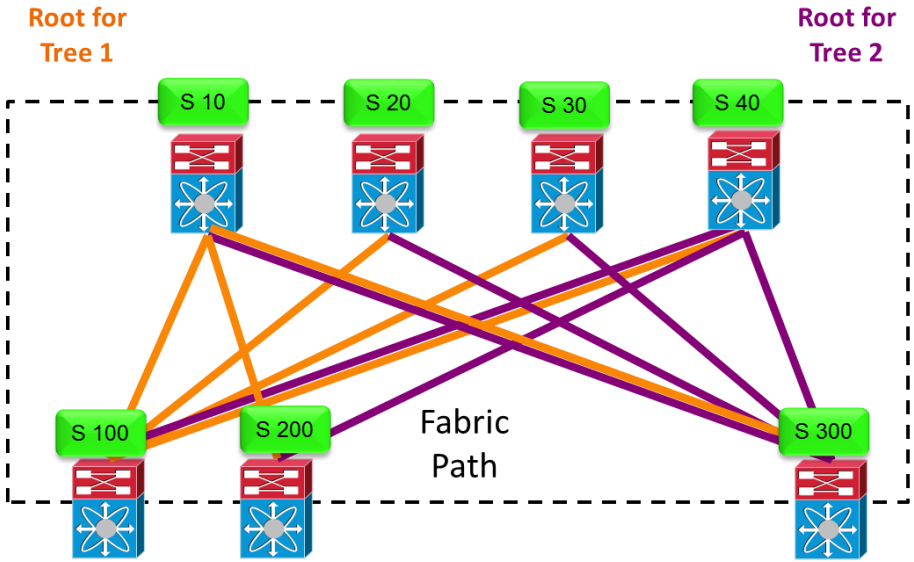


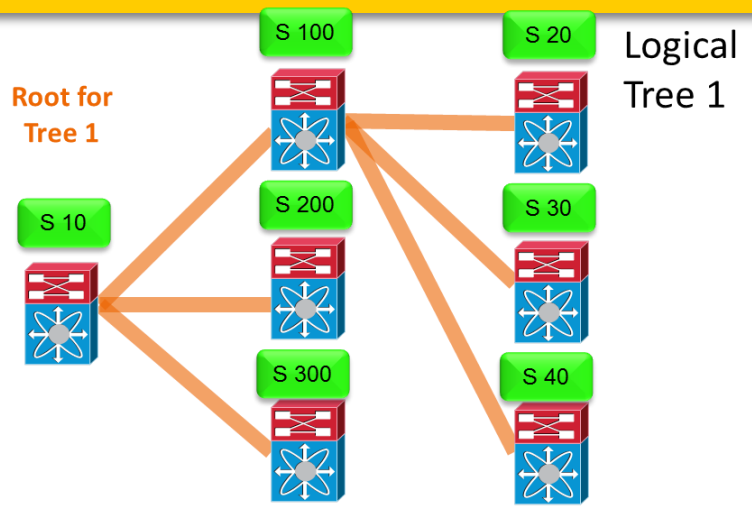
FabricPath keret

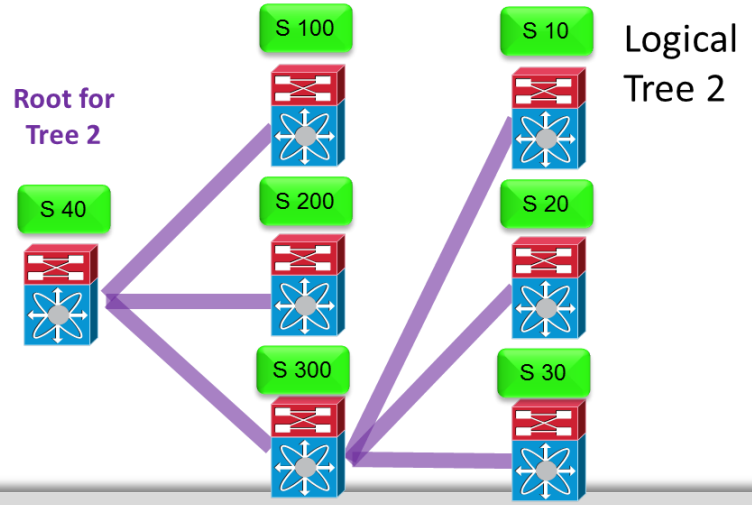
- Switch ID – Egyedi azonosító
- Sub-Switch ID – Egyedi azonosító VPC+ domain-enként
- LID – Local ID, port azonosító
- Ftag (Forwarding tag) – Unique number identifying topology and/or distribution tree



FabricPath Multi Destination Tree

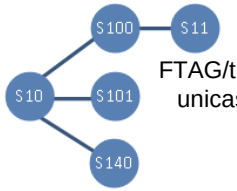




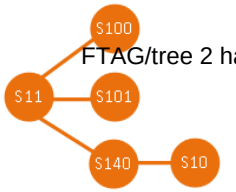


- Primary Tree Root a legmagasabb prioritású Switch (System ID, Switch ID)
- Globális FTAG hozzárendelés mindkét kifeszítő fához
- FTAG1: unknown unicast, broadcast, multicast
- FTAG2: multicast only

ARP Request – Broadcast



FTAG/tree 1 handles unknown unicast, broadcast and some multicast



FTAG/tree 2 handles multicast only

Multidestination Trees on S10

Tree	IF
1	L1,L3,L5
2	L5

ftag

encap

DSID→FF
Ftag→1
SSID→100
DMAC→FF
SMAC→A
Payload

Root for Tree 1

Root for Tree 2

S10

S11

decap

DSID→FF
Ftag→1
SSID→100
DMAC→FF
SMAC→A
Payload

L1 L2 L3 L4 L5 L6

S100

S101

S140

e1/1

e2/2

Host A

DMAC→FF
SMAC→A
Payload

Payload
SMAC→A
DMAC→FF

Host B

Multidestination Trees on S100

Tree	IF
1	L1,L2
2	L2

Broadcast

ftag

Multidestination Trees on S140

Tree	IF
1	L5
2	L5,L6

FabricPath MAC Table on S100

Switch	IF
A	e1/1 (local)

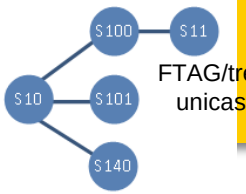
FabricPath MAC Table on S140

Switch	IF

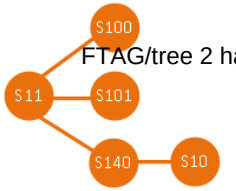
Step 1 :: Host A communicates to Host B for the first time – Sends ARP request to B
Step 2 :: S100 adds A into MAC table as the result of new source learning on CE port
Step 3 :: Since destination MAC is all 'F'; S100 floods this frame out all CE ports
[Learn MACs of directly-connected devices unconditionally]
Step 4 :: Meanwhile, S100 selects 'Tree 1', marks this in the FabricPath header and floods this frame out all FabricPath ports (L1, L2) that are part of Tree 1
Step 5 :: S10 floods this frame further, out (L3, L5) based on local info about Tree 1
Step 6 :: S101 and S140 remove the FabricPath header and flood the frame out all local CE ports.

Don't Learn Remote MAC since DMAC is unknown / is a Flooded Frame

ARP Reply – Unknown Unicast



FTAG/tree 1 handles unknown unicast, broadcast and some multicast



FTAG/tree 2 handles multicast only

Multidestination Trees on S10

Tree	IF
1	L1,L3,L5
2	L5

ftag

Multidestination Trees on S100

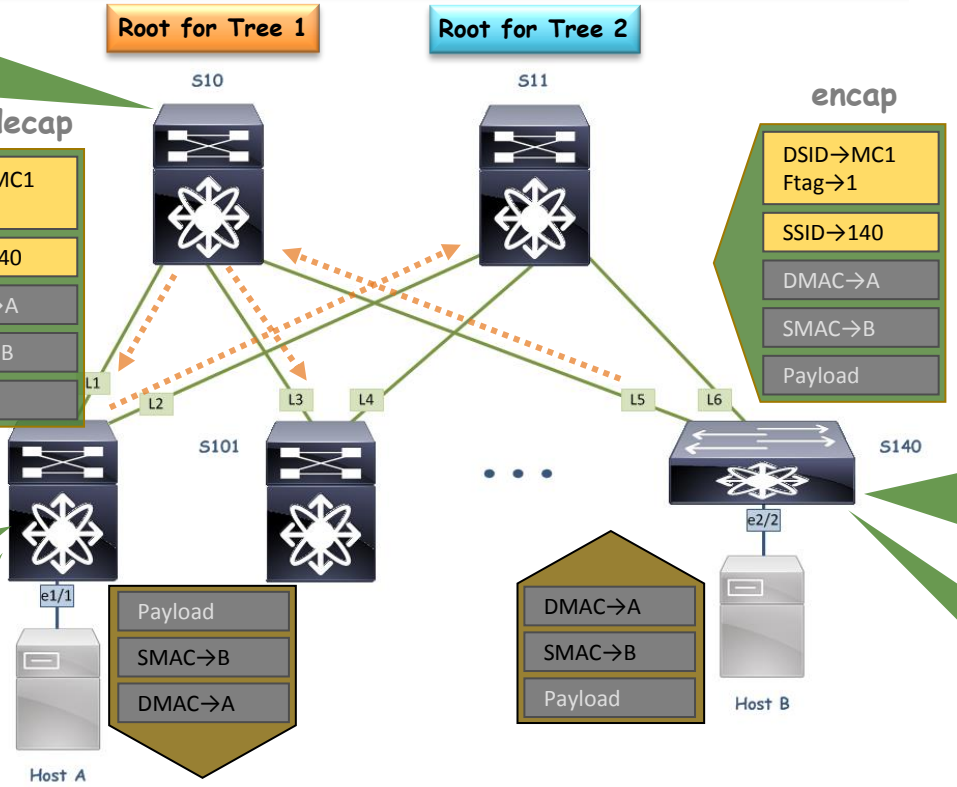
Tree	IF
1	L1,L2
2	L2

ftag

FabricPath MAC Table on S100

Switch	IF
A	e1/1 (local)
B	S140 (remote)

If DMAC is Known then Learn Remote MAC



Unknown

Multidestination Trees on S140

Tree	IF
1	L5
2	L5,L6

FabricPath MAC Table on S140

Switch	IF
A	
B	e2/2 (local)

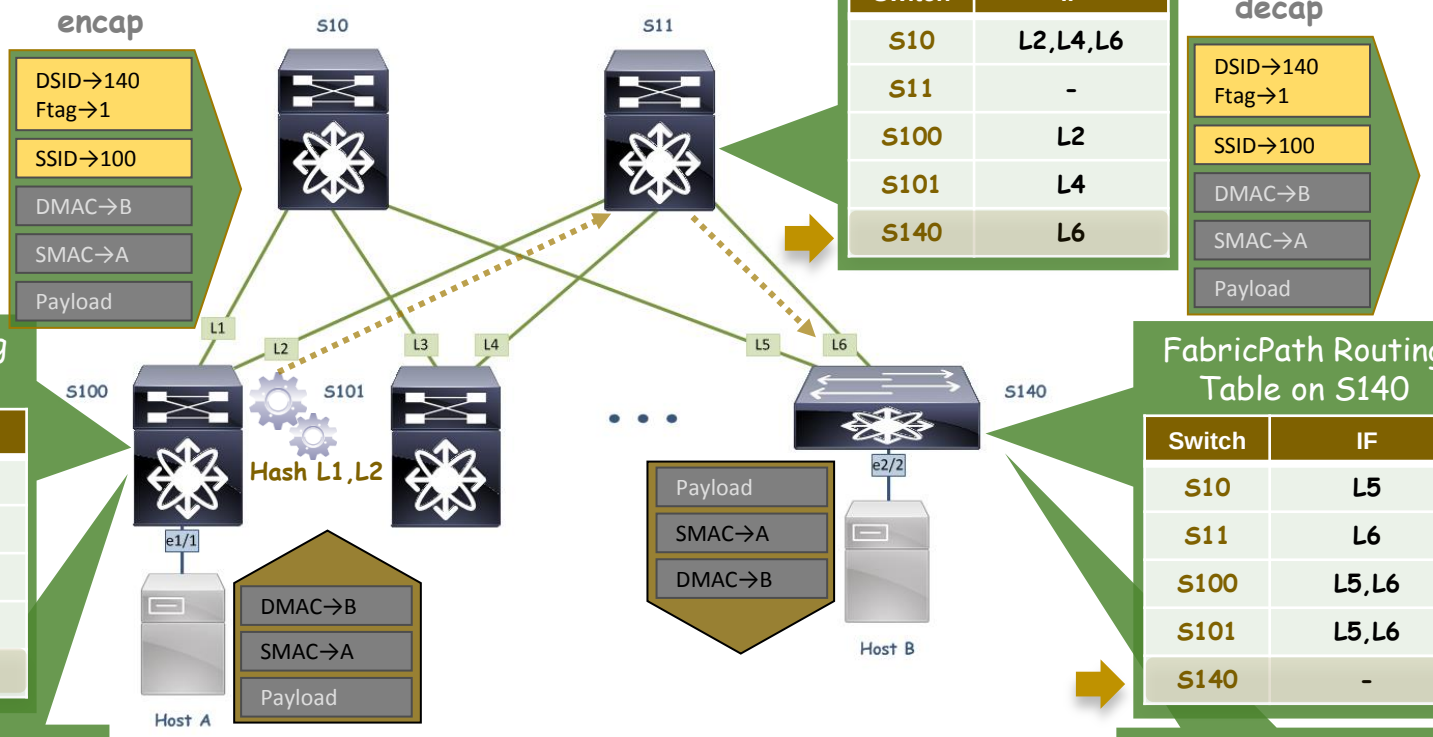
MAC A is Unknown

Step 1 :: Host B sends ARP Reply back to Host A
Step 2 :: S140 adds B into the MAC Table from source learning on CE port
Step 3 :: Since A is unknown, S140 floods the frame out all CE ports
Step 4 :: Meanwhile, S140 selects Tree 1, marks this in the FabricPath header and floods this frame out all FabricPath ports (L5) that are part of Tree 1
Step 5 :: S10 floods this frame further (L1, L3) along Tree 1
Step 6 :: S100 floods this frame further (L2) along Tree 1. Also, upon removing the FabricPath header, S100 finds host A was learned locally. Therefore adds B to the MAC Table as remote, associated with S140

A →

Known Unicast

Destination Switch ID is used to make routing decisions through the FabricPath core & no MAC learning or lookups required inside the FP core



FabricPath Routing Table on S11

Switch	IF
S10	L2,L4,L6
S11	-
S100	L2
S101	L4
S140	L6

FabricPath Routing Table on S100

Switch	IF
S10	L1
S11	L2
S100	-
S101	L1,L2
S140	L1,L2

FabricPath Routing Table on S140

Switch	IF
S10	L5
S11	L6
S100	L5,L6
S101	L5,L6
S140	-

FabricPath MAC Table on S100

Switch	IF
A	e1/1 (local)
B	S140 (remote)

FabricPath MAC Table on S140

Switch	IF
A	S100 (remote)
B	e2/2 (local)

Step 1 :: Host A starts sending traffic to Host B after ARP resolution
Step 2 :: S100 finds B was learned as remote; associated with S140, encap all subsequent frames to B with S140 as destination in FP header
Step 3 :: S100 Routing Table indicates multiple paths to S140; runs ECMP hash and this time S100 selects L2 as next-hop
Step 4 :: Routing Table lookup at S11 indicates L6 as next hop for S140
Step 5 :: S140 finds itself as destination in FabricPath header and B is also known locally; decaps FP header, adds A as remote & associates with S100

FabricPath Összefoglaló

- Rugalmas
  - Routing
- Egyszerű
  - Nincs STP, rögzített címzés
- Hatékony
  - ECMP
- Skálázható
  - MAC címek ott tárolódnak, ahol szükség is van rájuk

Alkalmazások

- vPC+ – tipikusan UNI port
  - Nincs SPOF
- FabricPath – tipikusan NNI port
  - Topológia független
  - Nagyobb rugalmasság
- Eredmény:
  - Nagy rendelkezésre állás
  - Gyors konvergencia
  - Skálázhatóság
  - Jobb kapacitás kihasználtság

Élet Spanning Tree nélkül

Q & A

Balla Attila

attila.balla@kapsch.net