upna
Universidad Pública de Navarra
Nafarroako Unibertsitate Publikoa
Redes de Nueva Generación
Área de Ingeniería Telemática

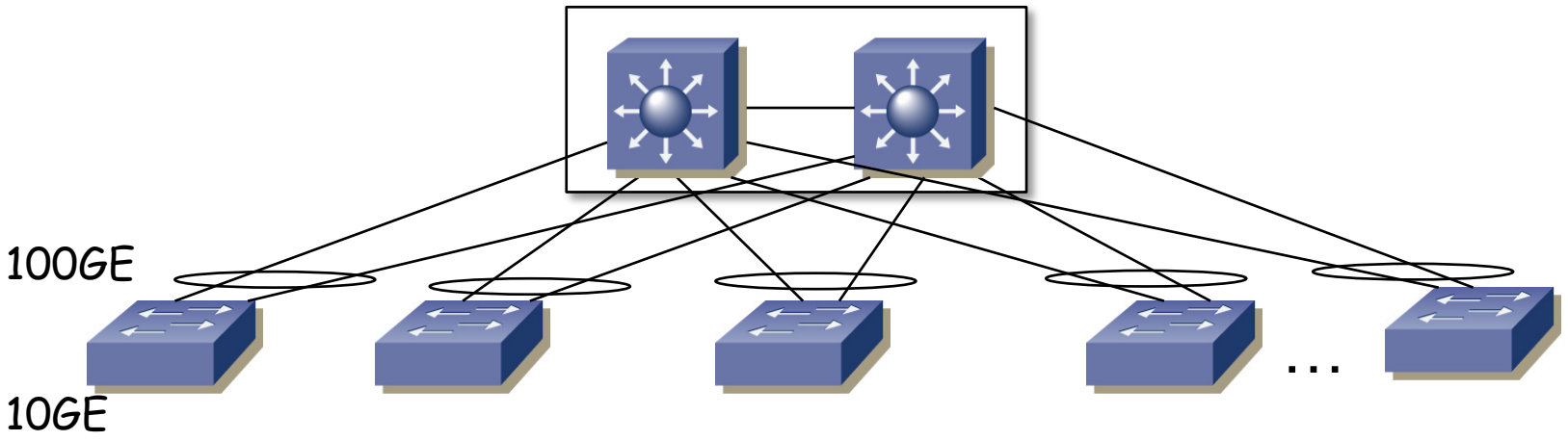
Arquitectura tradicional en el data center: Escalabilidad

Escalabilidad

- Estamos condicionados por el número de puertos en los conmutadores
- La densidad ha ido aumentando con los años
- Veamos un ejemplo simplemente con 2 capas
- En primer lugar con una arquitectura MLAG, es decir, con 2 switches en la capa de agregación (…)
- *(Intentaré poner ejemplos de equipos reales pero por simplicidad me tomaré algunas libertades respecto a lo que pueden hacer)*

Escalabilidad con MLAG

- Acceso: Tenemos conmutadores con más de 1000 puertos 10GE (¡!)
- Hacia la segunda capa puertos 40GE o 100GE
- (…)



Ejemplo: Arista 7512R

- Capacidad de conmutación de 115 Tbps o 51 Bpps
- (Ojo, “Billones” anglosajones = “miles de millones”)
- Hasta 288 GB de buffer, consumo de 12KW)
- 12 slots
- Linecards:
  - La 7500R-36CQ tiene 144 puertos 10GE
  - 144x11 = 1584 puertos 10GE
  - En el slot restante pongamos la misma con puertos 100GE
  - Eso son 36 puertos 100GE

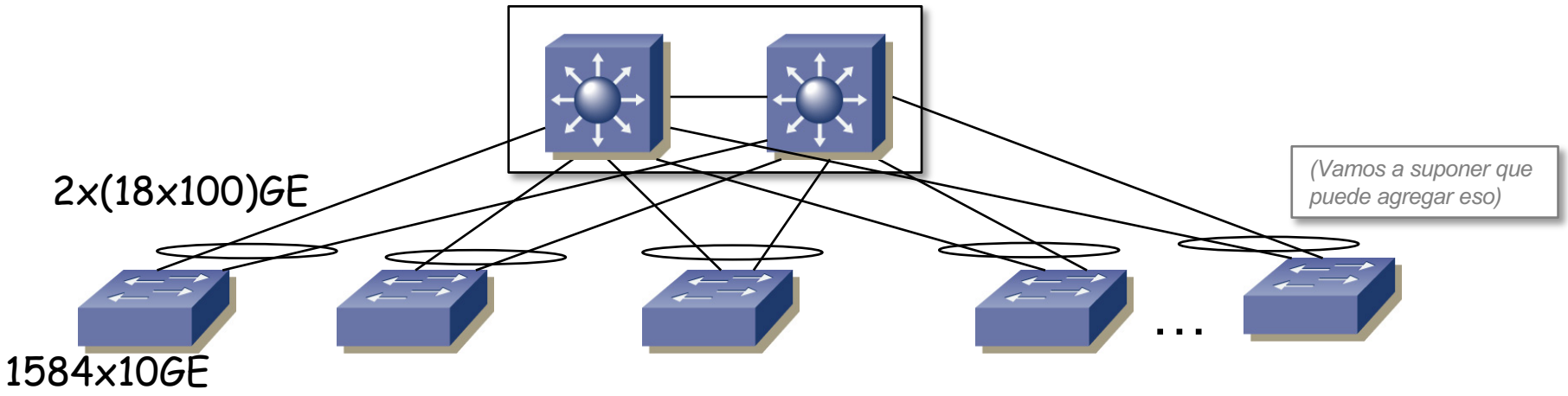
	7500R-36CQ 	7500R-36Q 	7500R-48S2CQ 
Ports	36 x 100G QSFP	30 x 40G / 6 x 100G	48x 10G and 2x 100G
Max 100GbE	36	6	2
Max 50GbE	72	12	4
Max 40GbE	36	36	2
Max 25GbE	144	24	8
Max 10GbE	144	96	56 (48+8)
Port Buffer	24GB	8GB	4GB
Switching Capacity	3.6 Tbps	1.8 Tbps	680 Gbps
Forwarding Rate	4.32 Bpps	1.4 Bpps	720 Mpps

QSFP+ = Quad Small Form-factor Pluggable

http://www.arista.com/en/products/7500r-series

Escalabilidad con MLAG

- Acceso: Tenemos conmutadores con más de 1000 puertos 10GE
- Hacia la segunda capa puertos 40GE o 100GE
- Por ejemplo 1584x10 GE sobre 36 enlaces, 18 a cada switch de agregación, dando una sobre-subscripción 4.4:1 (15840/3600 = 4.4)
- (Suponiendo que su MLAG nos permita ese reparto de enlaces)
- Agregación: Hay conmutadores con más de 100 puertos 100GE
- (…)



Ejemplo: Arista 7512R

- Sin irnos más lejos
- Linecards:
  - La 7500R-36CQ tiene 36 puertos 100GE
  - Con 12 de ellas tendríamos 12x36 = 432 puertos 100GE

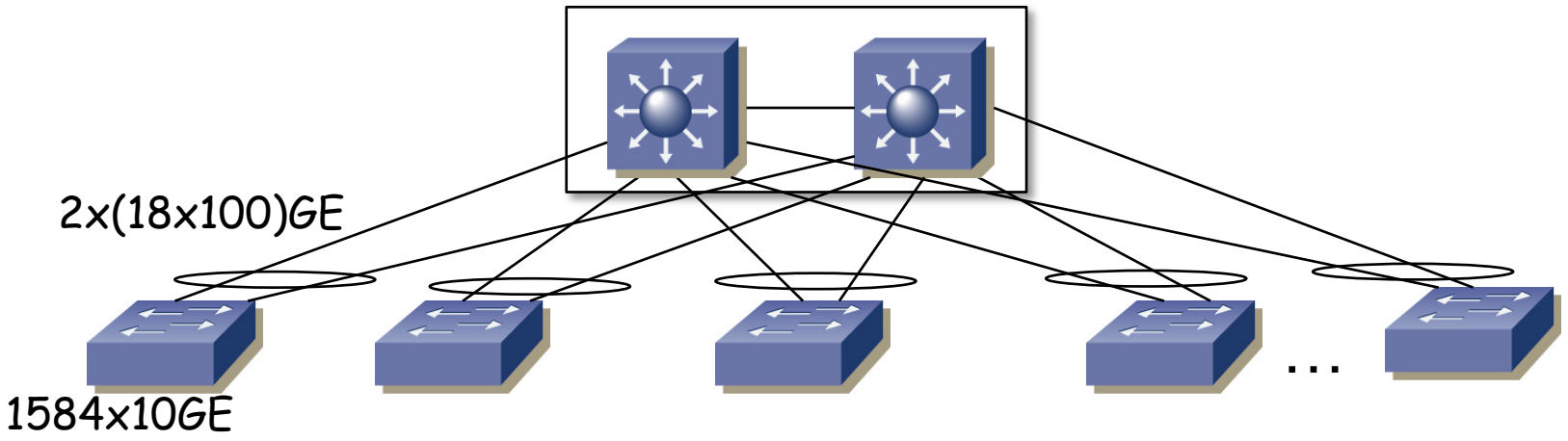
	7500R-36CQ 	7500R-36Q 	7500R-48S2CQ 
Ports	36 x 100G QSFP	30 x 40G / 6 x 100G	48x 10G and 2x 100G
Max 100GbE	36	6	2
Max 50GbE	72	12	4
Max 40GbE	36	36	2
Max 25GbE	144	24	8
Max 10GbE	144	96	56 (48+8)
Port Buffer	24GB	8GB	4GB
Switching Capacity	3.6 Tbps	1.8 Tbps	680 Gbps
Forwarding Rate	4.32 Bpps	1.4 Bpps	720 Mpps

QSFP+ = Quad Small Form-factor Pluggable

http://www.arista.com/en/products/7500r-series

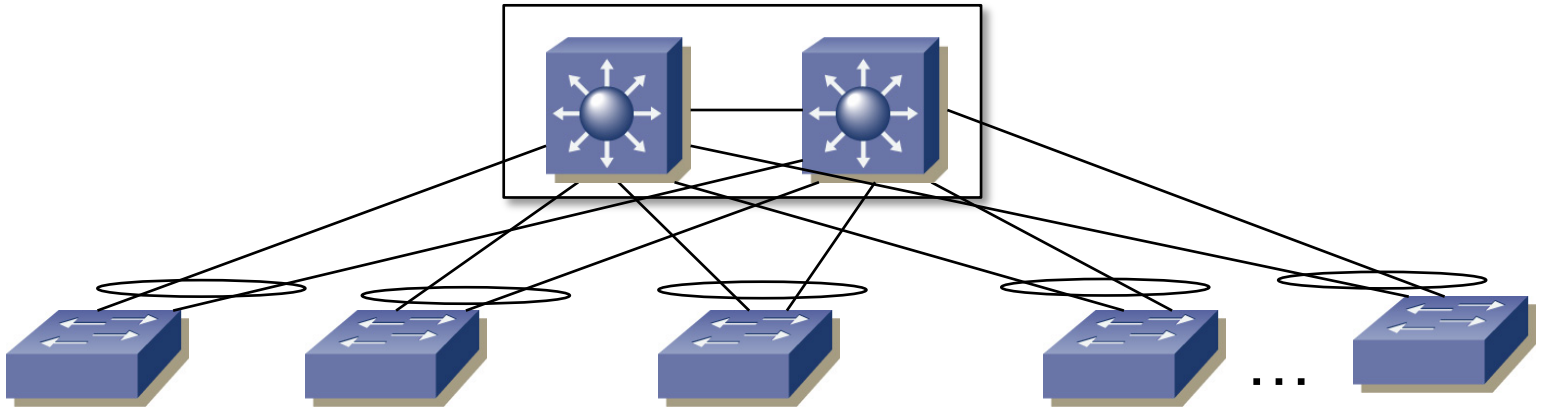
Escalabilidad con MLAG

- Acceso: Tenemos conmutadores con más de 1000 puertos 10GE
- Hacia la segunda capa puertos 40GE o 100GE
- Por ejemplo 1584x10 GE sobre 36 enlaces, 18 a cada switch de agregación, dando una sobre-subscripción 4.4:1 (15840/3600 = 4.4)
- Agregación: Hay conmutadores con más de 100 puertos 100GE
- Con 432x100GE, donde cada conmutador de acceso consume 18 puertos, podríamos tener 24 (24x18=432) conmutadores de acceso
- Eso son 1584x24 = 38.016 hosts con un puerto 10GE cada uno y una sobre-subscripción 4.4 a 1
- ¿1500 hosts por switch? Cableado EoR
- (…)



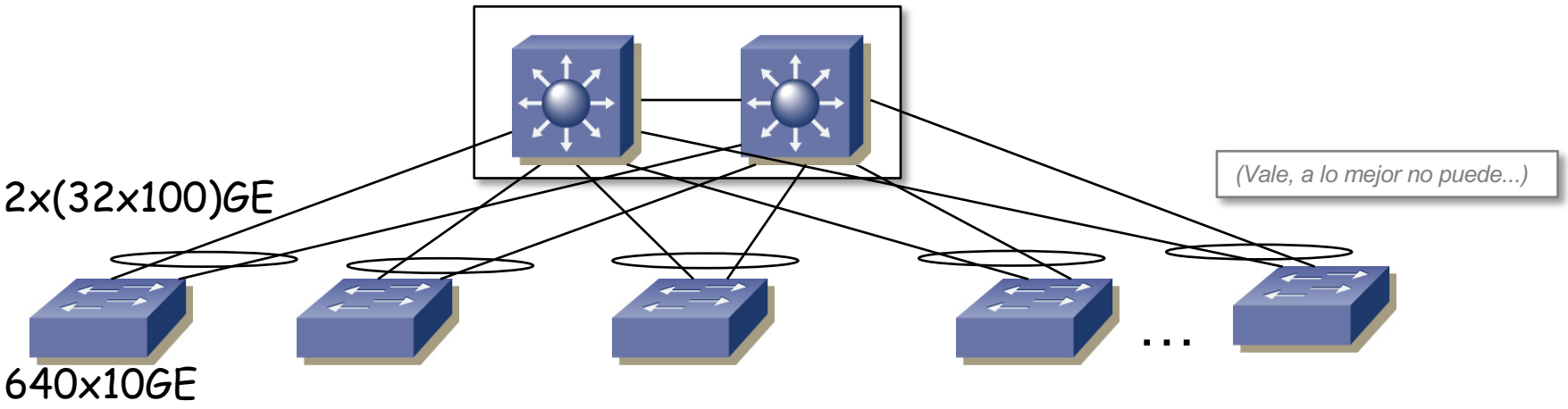
MLAG no bloqueante

- ¿Podríamos hacerlo no-bloqueante?
- Configuración del switch de acceso:
  - 8x144 = 1152 puertos 10GE (11.5 Tbps)
  - 4x36 = 144 puertos 100GE (14.4 Tbps) en uplinks
  - Podríamos ajustarlo pues cada puerto 100GE puede sacar 4x10GE
  - En una de las 4 tarjetas 100GE ponemos 22x100GE y (14x4)x10GE :
  - 8x144+56 = 1208 puertos 10GE (12.1 Tbps)
  - 3x36+22 = 130 puertos 100GE (13 Tbps)
  - En cada switch de acceso 12.9 Tbps desde los hosts para 13 Tbps uplink
  - (…)



MLAG no bloqueante

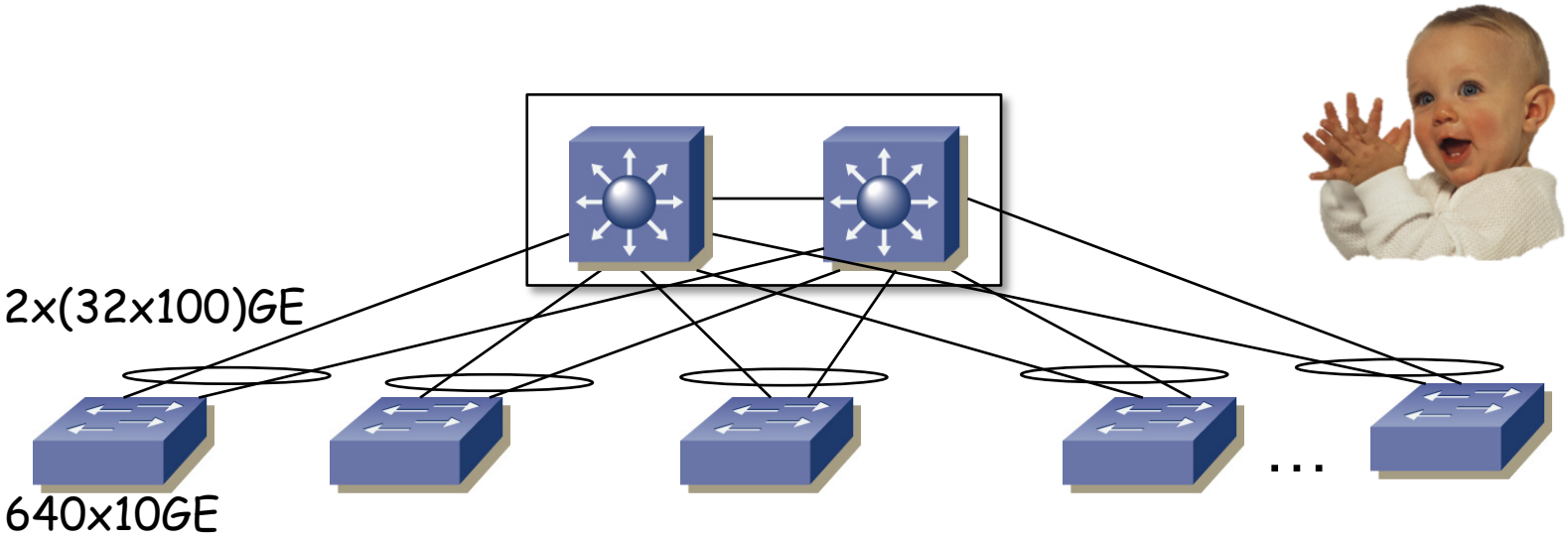
- 8x144+56 = 1208 puertos 10GE (12.1 Tbps)
- 3x36+22 = 130 puertos 100GE (13 Tbps)
- En cada switch de acceso 12.1 Tbps desde hosts para 13 Tbps uplink
- ¿Pero y si este switch soportara un máximo de 64 puertos por MLAG? Así en realidad tendríamos:
  - 64x100 = 6.4 Tbps en uplinks
  - 640x10 = 6.4 Tbps a hosts (quedan slots sin usar)
- Hemos perdido puertos a hosts para mantener la sobre-subscripción 1:1
- (…)



MLAG no bloqueante

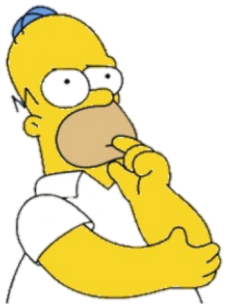
- El switch de agregación podía tener 432 puertos 100GE
- Cada 32 puertos a un switch así que tenemos hasta 13 conmutadores de acceso (14x32 = 448 > 432)
- Así que en total 13x640 = 8.320 hosts con puertos 10GE non-blocking
- Antes teníamos 38.016 hosts con sobre-subscripción 4.4:1

- Sobrevive ante un fallo de switch de agregación pero con una pérdida del 50% del BW de bisección

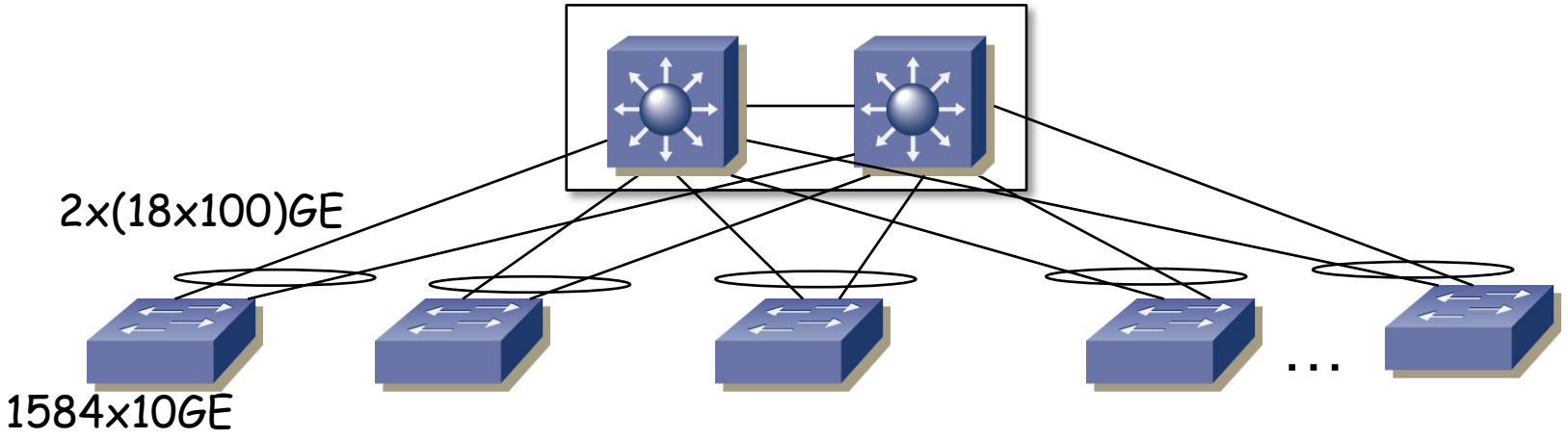


Escalabilidad con MLAG

- 38.016 puertos 10GE con sobre-subscripción 4.4:1
- Supongamos que en cada host tenemos 2 CPUs de 8 cores
- Y por ejemplo decidimos correr no 20VMs sino 100 contenedores
- Eso son 100x38.016 o casi 4 millones de direcciones MAC (…)



	7500R Linecards
Latency	Under 3.5usec
MAC Table Size	768K
Maximum IPv4 Host Routes	768K
Maximum IPv6 Host Routes	768K
Maximum ACL Entries	24K
Maximum IPv4 Route Prefixes	Over 1M
Maximum IPv6 Route Prefixes	768K
Maximum Multicast Routes	768K
Maximum ECMP	128-way

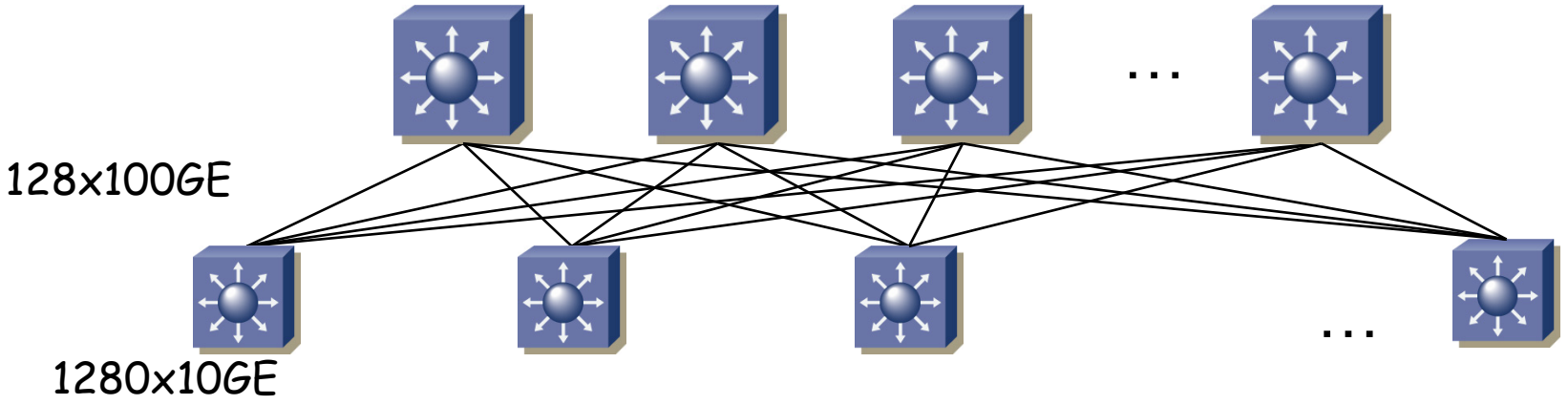


Escalabilidad con ECMP

- Conmutadores por ejemplo con soporte para 128-way ECMP
- Podemos cambiar esta topología MLAG por una ECMP
- Un enlace 100GE de cada switch de acceso a cada uno de agregación
- Siguen siendo 12.8Tbps en uplinks y sobre-subscripción 1:1
- Pero ahora cada switch de agregación puede recibir enlaces de 432 switch de acceso
- Eso son hasta 432x1280 = 552.960 hosts
- ¿Cables? 128x432 = 55.296
- Se conocen data centers con 360.000 hosts

https://youtu.be/4e97g7_qSxA

Sobre-subscripción 1:1



Problemas

- El precio de estos equipos de alto nivel de agregación es elevado
  - CAMs/TCAMs/SRAMs muy grandes
  - Alto consumo eléctrico, componentes costosos
  - Hay que pagar el coste de diseño de este hardware/software (¡no venden muchos!)
- Internamente:
  - Estos conmutadores son redes de conmutación, como hemos visto antes
  - Es decir, están construidos a partir de chips de conmutación interconectados
  - Pueden tener sobre-subscripción interna (bloqueo interno)

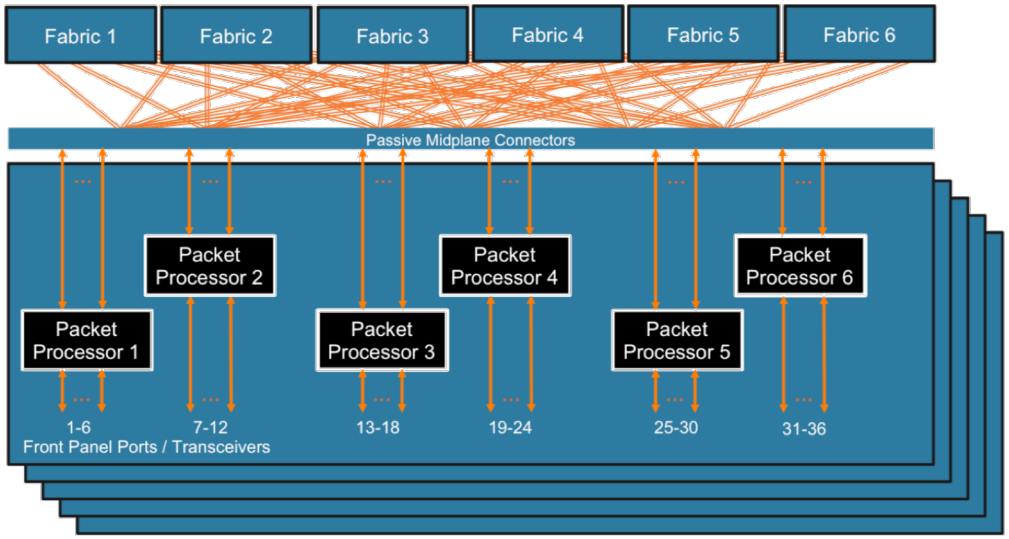


Figure 3: Distributed Forwarding within an Arista 7500R Series

http://www.arista.com/assets/data/pdf/Whitepapers/Arista7500RSwitchArchitectureWP.pdf

El mismo problema

- Creamos una red de interconexión de conmutadores
- Los de agregación son de alto coste porque internamente
- Son una red de interconexión de chips de conmutación
- (…)