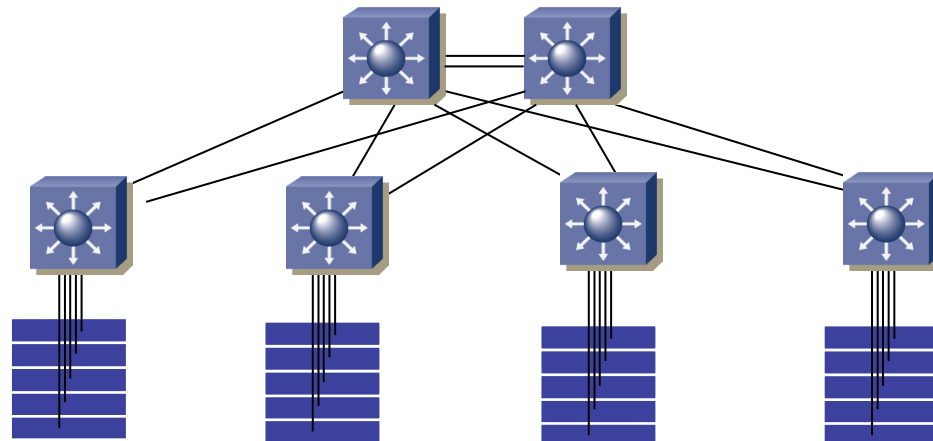


Arquitectura tradicional en el data center: limitaciones

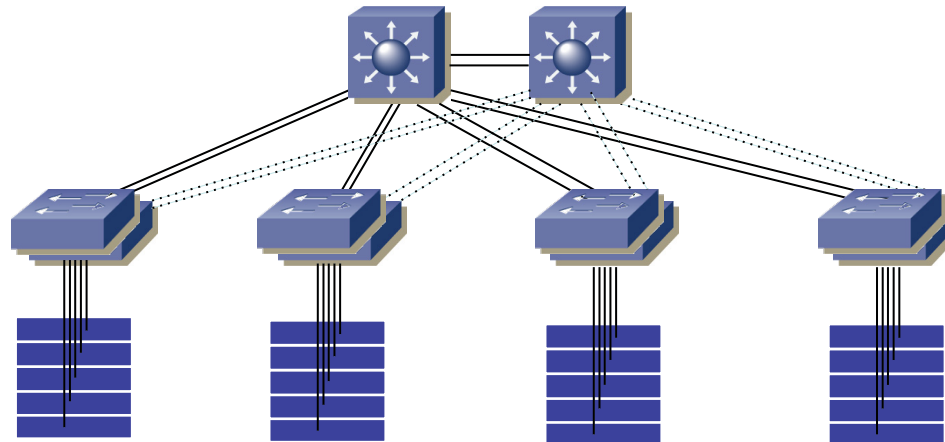
DC tradicional de 2 capas

- Racks de servidores
- Conmutación en dos capas: ToR y agregación
- Conmutación capa 2 en agregación si se necesita que los distintos armarios estén en la misma VLAN
- Si los armarios son servicios independientes estarán en distintas VLANs
- En un mismo armario podría haber diferentes componentes de un servicio separados por VLANs
- Podría conmutarse capa 3 en agregación o en el ToR
- Hoy en día conmutación *line-rate* capa 3



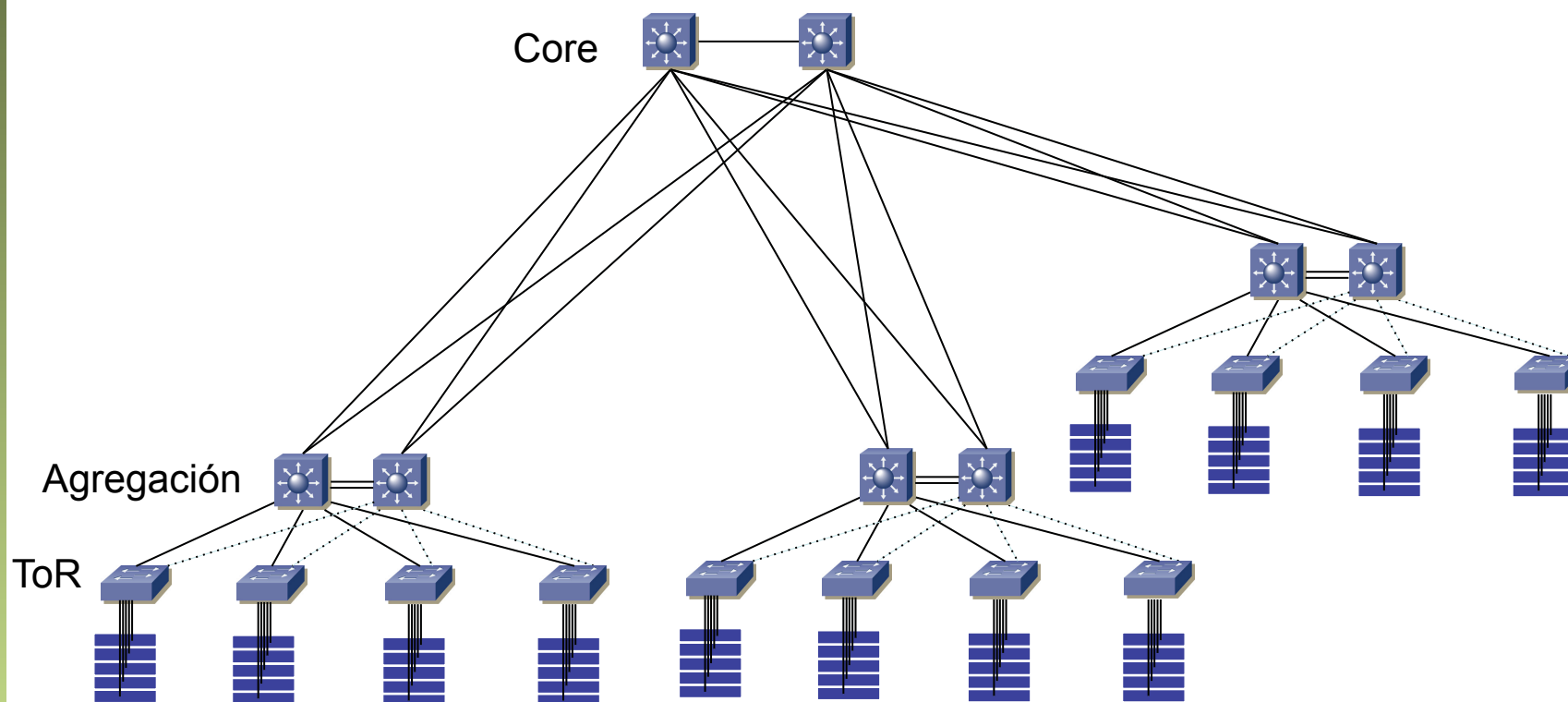
Límites con 2 capas

- Este esquema puede ser con uno o dos ToR por armario (redundancia de conexión de servidor a switch)
- En cualquier caso el límite de escala está en el número de puertos en los conmutadores de agregación
- Ejemplo:
 - Conmutador de acceso de 48 puertos (48 servs/rack) + 2 uplinks
 - Conmutador de agregación con 64 puertos 10Gbps
 - Si hay redundancia en el rack hay mismo nº de servidores por armario pero cada sw. de agregación consume 2 puertos/rack
 - Máximo $48 \times 64 / 2 = 1536$ servidores
- ¿Más?



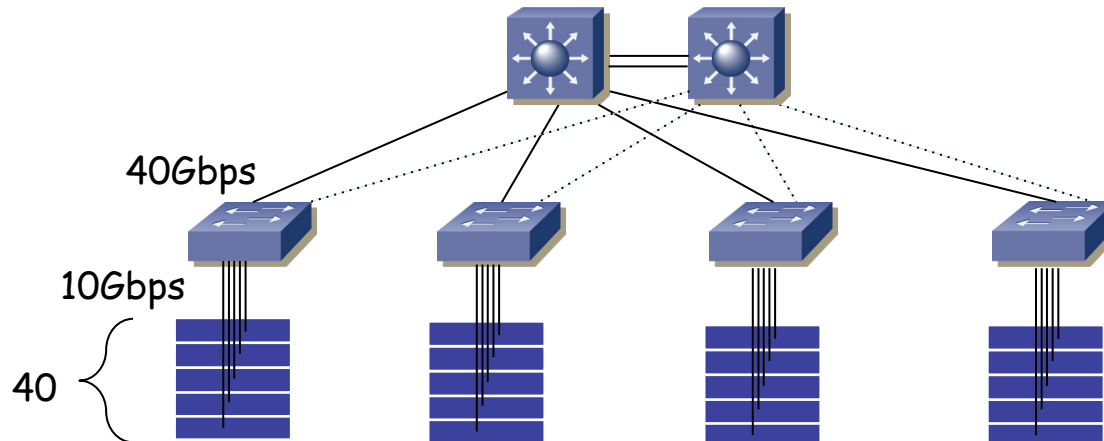
3 capas

- La solución tradicional es añadir una tercera capa
- ¿De qué capacidad son esos enlaces?



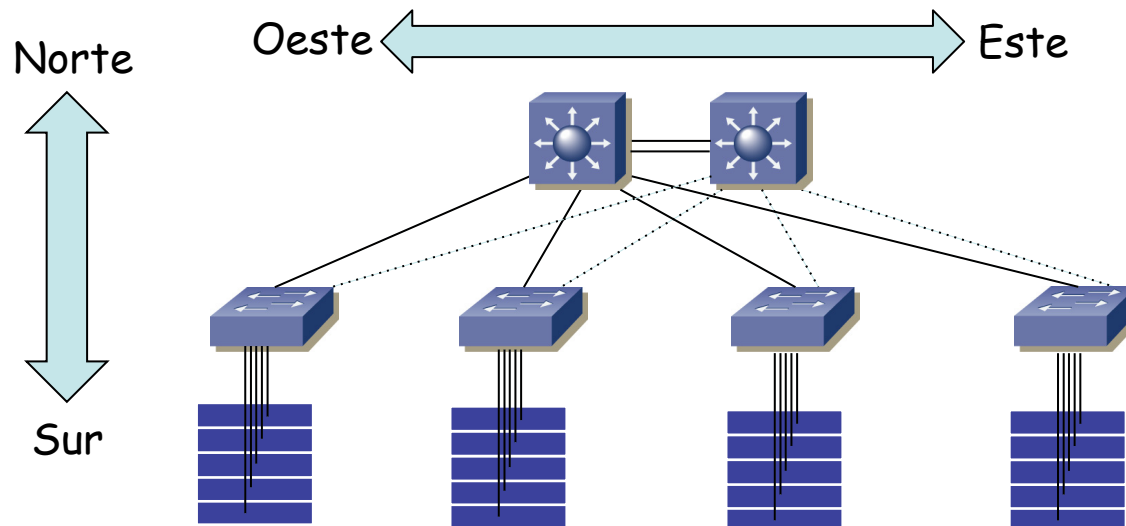
Over-subscription

- Conectamos hosts de tal forma que su tráfico agregado excede el que se puede cursar por los enlaces externos
- Ejemplo:
 - Cada conmutador de acceso: 40 servidores con una NIC a 10Gbps
 - Eso es un máximo de $40 \times 10 = 400$ Gbps al ToR
 - Enlace hacia la capa de distribución es de 40 Gbps
 - Tenemos una sobre-subscripción de 10:1
 - Si el enlace a distribución fuera un LAG de 2×40 Gbps sería un 5:1
 - Un 5:1 para servidores con enlaces 10GE quiere decir en un reparto equitativo 2Gbps por servidor



Over-subscription

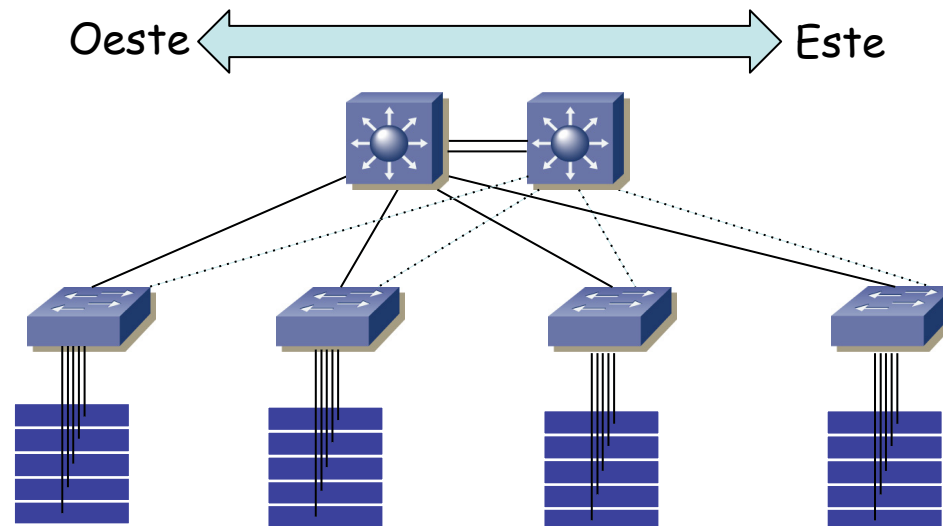
- Altos ratios de sobre-subscripción son razonables cuando tenemos mucho tráfico norte-sur
- Por ejemplo hacia una salida a Internet que sea en realidad el cuello de botella
- No son tan razonables cuando hay mucho tráfico este-oeste
 - Tráfico entre los servidores en distinto rack
 - Aplicaciones distribuidas, tráfico de SAN, movimiento de máquinas virtuales, tráfico entre tiers de aplicación, clustering, etc



Tráfico Este-Oeste

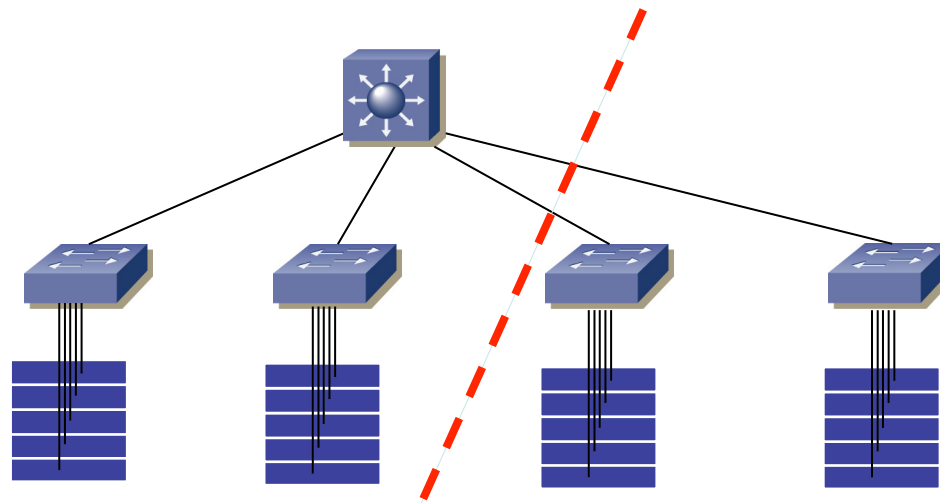
- Ejemplo: Facebook (2013)
- Una petición HTTP se transforma en:
 - 88 búsquedas en caches
 - 35 búsquedas en bases de datos
 - 392 llamadas a procedimientos remotos en el backend
- Una petición con un tamaño de 1KB ha resultado en cerca de 1MB de transferencias en el data center
- ¿Datos almacenados por Facebook? Más de 100PB

N.Farrington and A.Andreyev, "Facebook's Data Center Network Architecture", 2013 IEEE Optical Interconnect Conference



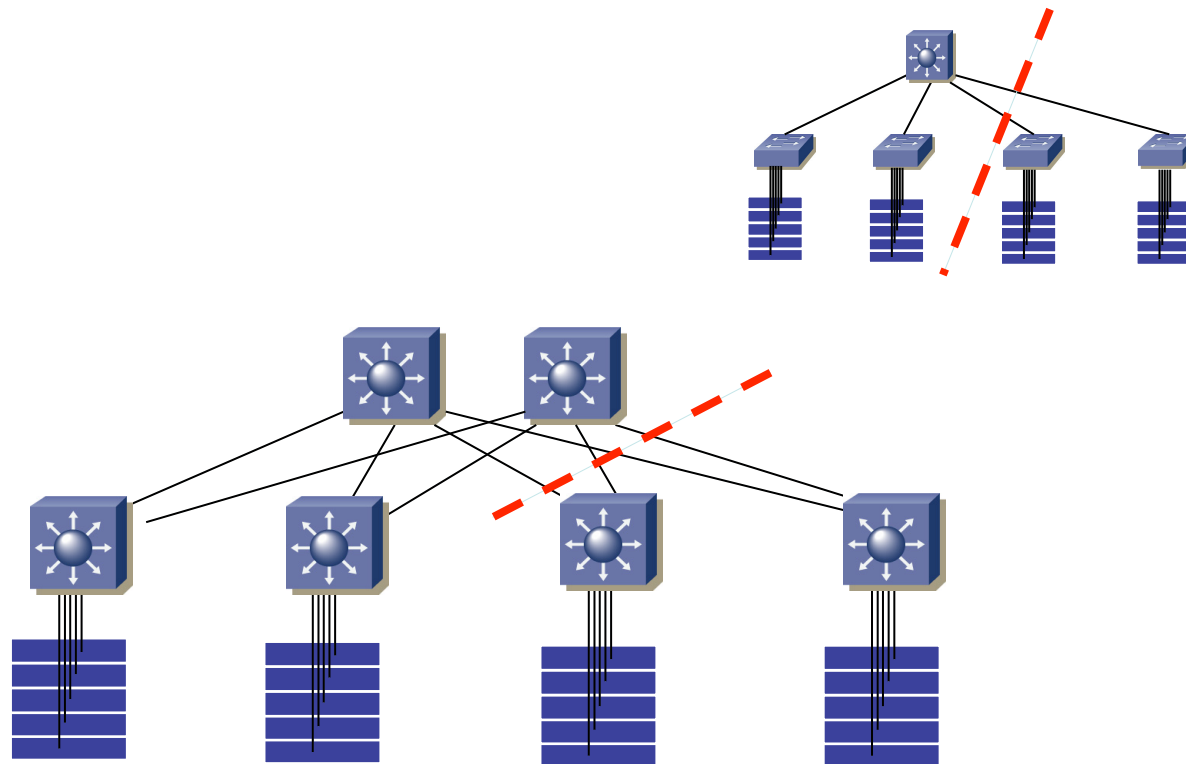
Bisectional bandwidth

- Una **bisección** es una partición de la red en dos subconjuntos con igual número de hosts
- El **ancho de banda de esa bisección** es la suma de las capacidades de los enlaces entre los dos subconjuntos
- En este ejemplo 2x capacidad del enlace de agregación a acceso
- El **ancho de banda de bisección de la red** es el menor ancho de banda de una bisección de la red que se pueda conseguir
- Cuando tenemos mucho tráfico este-oeste queremos un elevado ancho de banda de bisección
- Una topología en 2 capas nos puede dar un alto ancho de banda de bisección si hay muchos enlaces activos entre ellas



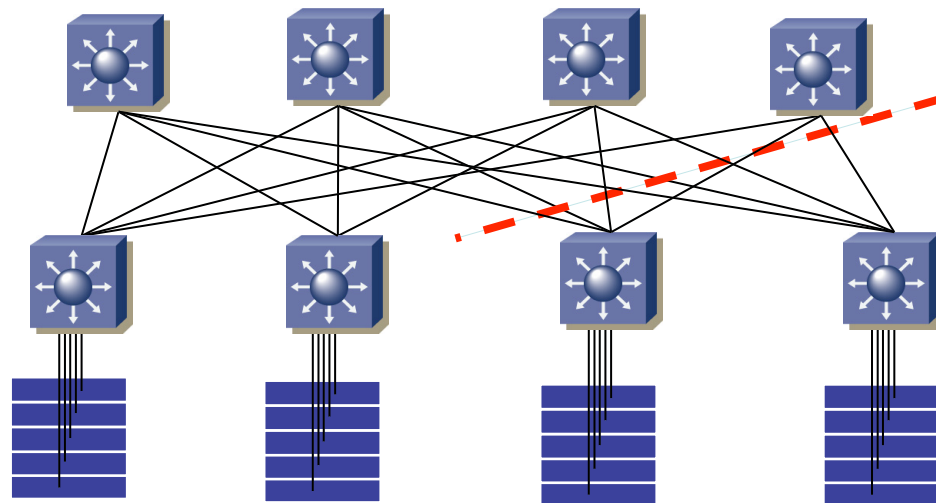
ECMP y biseccional bw

- Aumentamos el ancho de banda de bisección aumentando el número de caminos
- En este caso por ejemplo lo hemos duplicado
- Podríamos seguir aumentándolo (...)



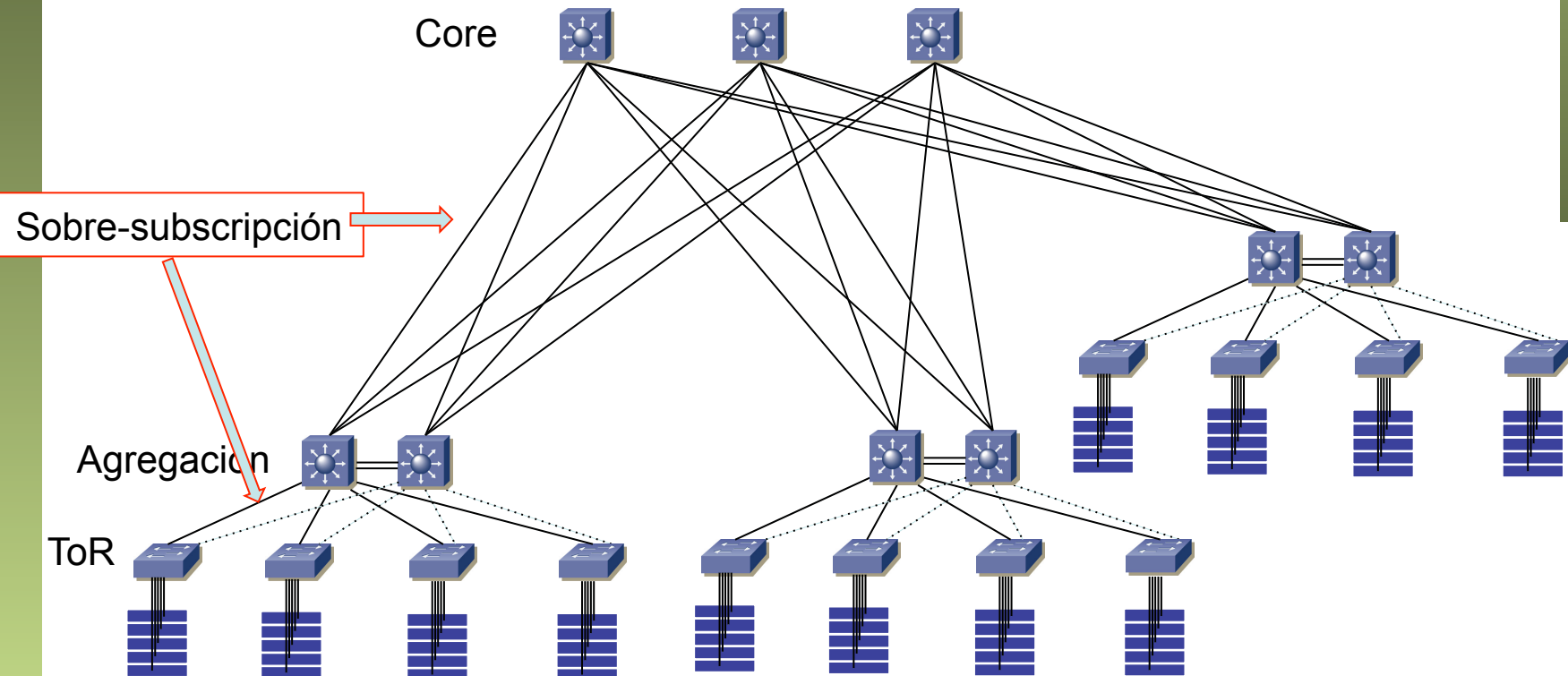
ECMP y biseccional bw

- Aumentamos el ancho de banda de bisección aumentando el número de caminos
- En este caso por ejemplo lo hemos duplicado
- Podríamos seguir aumentándolo
- ¿Inconveniente?
- La conmutación es capa 3 pues en capa 2 STP no nos permite tener caminos alternativos
- Eso quiere decir que no podemos extender VLANs entre los armarios



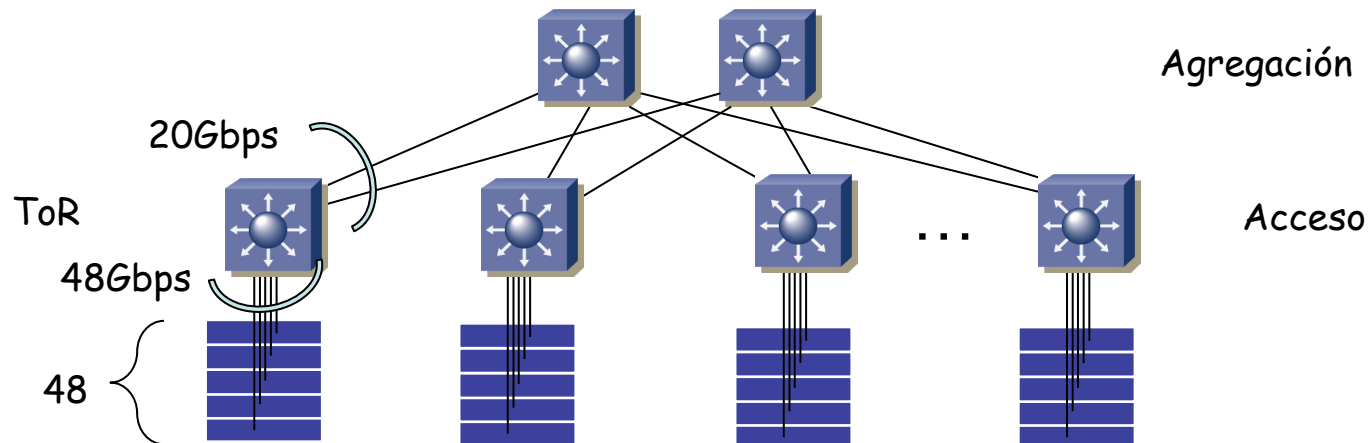
3 capas

- Por supuesto podríamos hacer ECMP entre las capas de agregación y core
- En ese caso sí podemos extender VLANs entre algunos armarios
- Ahora tenemos un segundo punto con sobre-subscripción (...)



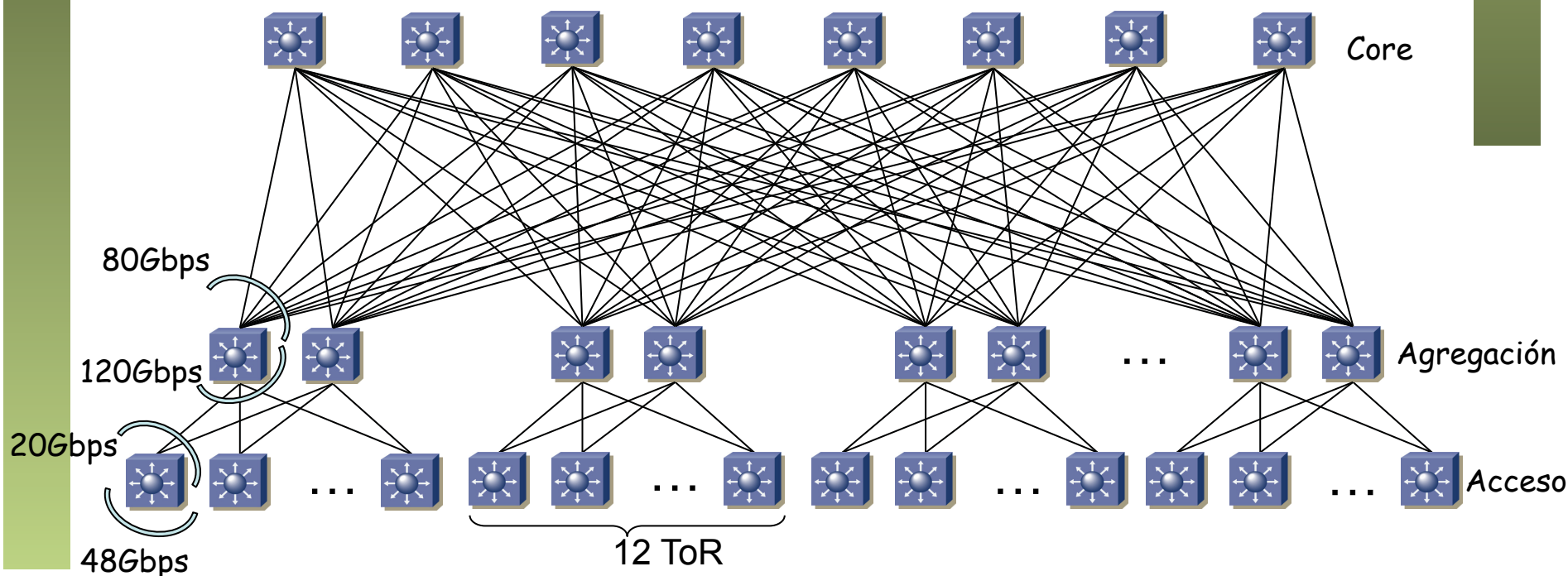
Ejemplo de sobre-subscripción

- ToR de 48x1Gbps + 2x10Gbps
- 48 servidores a 1Gbps por cada ToR
- Enlaces 10Gbps a la capa de agregación (ECMP)
- $20\text{Gbps}/48\text{servidores} = 416\text{Mbps/servidor}$
- $48\text{Gbps}:20\text{Gbps} = 2.4:1$
- Una agregación “2.4 a 1”



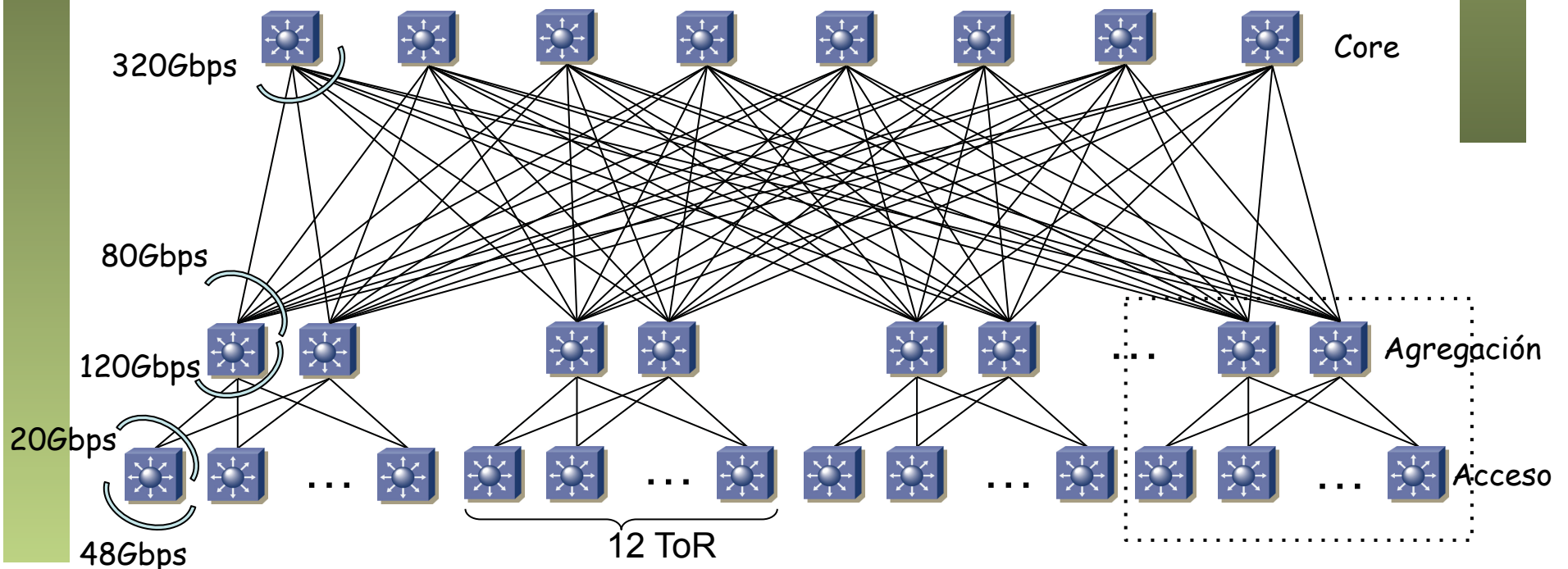
Ejemplo de sobre-subscripción

- Switch de agregación con 20 interfaces a 10Gbps (8+12)
- 8x10Gbps hacia el núcleo → 8-way ECMP 80Gbps hacia el núcleo
- 12x10Gbps hacia acceso → 12 conmutadores de acceso bajo cada uno de agregación
- 12x10Gbps = 120Gbps de la capa de agregación sobre 80Gbps a core
- $120:80 = 1.5:1$, una agregación 1.5 a 1



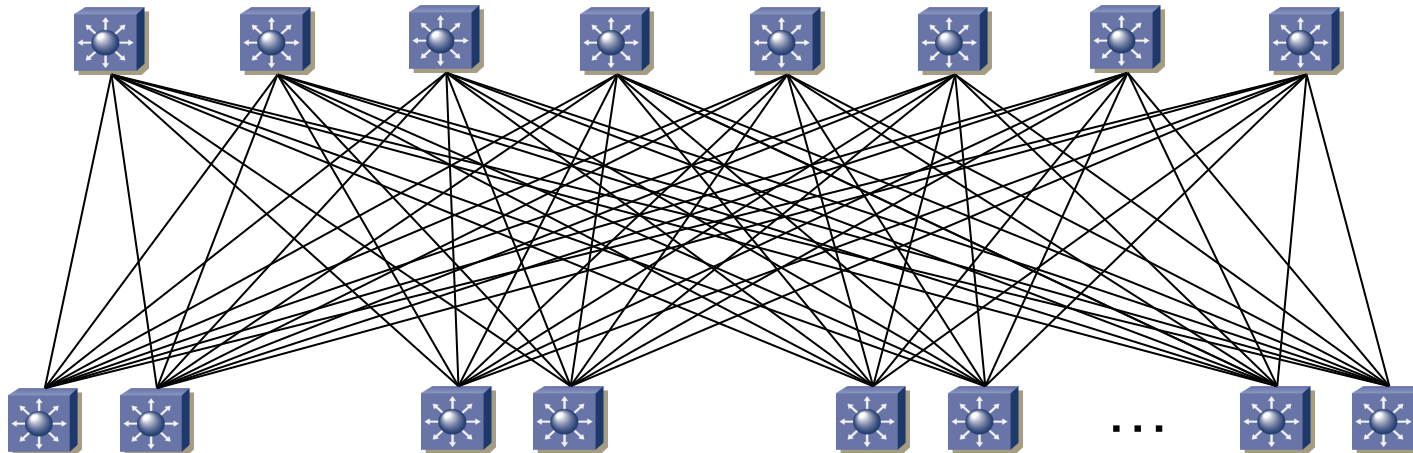
Ejemplo de sobre-subscripción

- Switch de core con 32 interfaces a 10Gbps
- Eso le permite tener por debajo 16 bloques de agregación
- Cada bloque de agregación $48 \times 12 = 576$ servidores
- Una pareja de switches de agregación tiene 160Gbps al núcleo para $48 \times 12 = 576$ servidores, lo cual da unos 277Mbps/servidor
- En total $576 \times 16 = 9216$ servidores



Redes de...

- Cada conmutador de agregación conectado a cada uno de la capa del núcleo
- (...)





No bloqueante

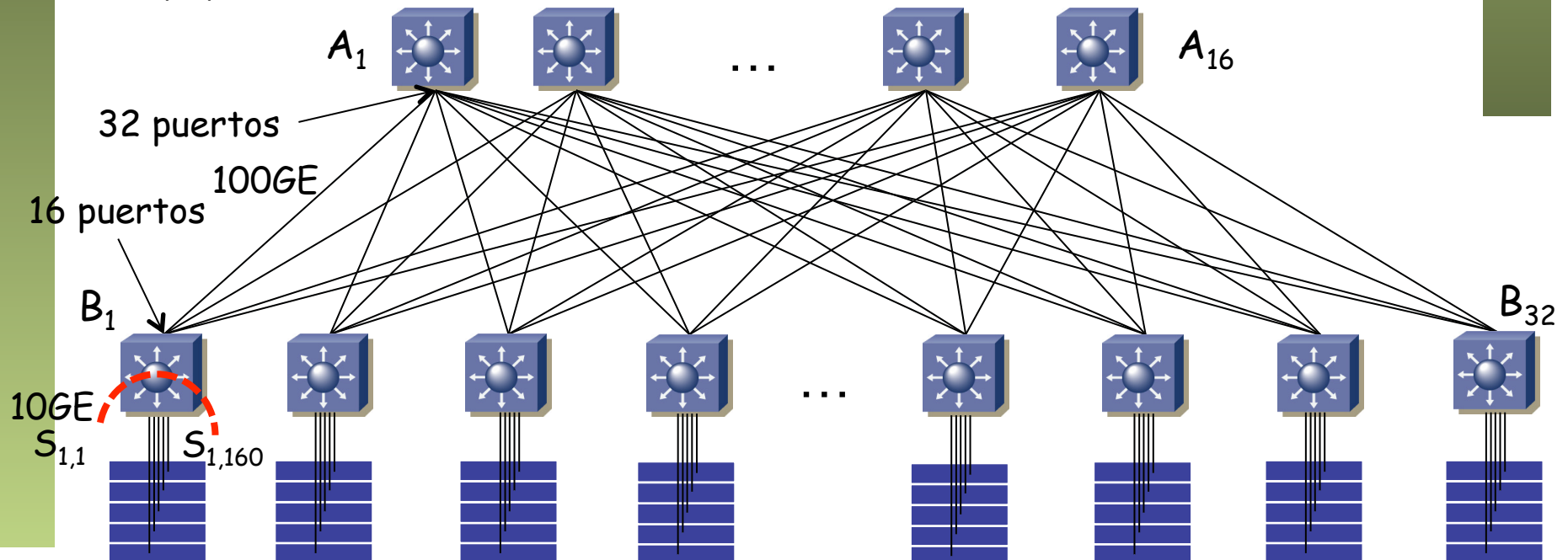


No bloqueante

- ¿Podemos hacer la red sin sobre-subscripción?
- Sería el equivalente a una red de Clos “rearrangeably non-blocking”
- Ejemplo (...)

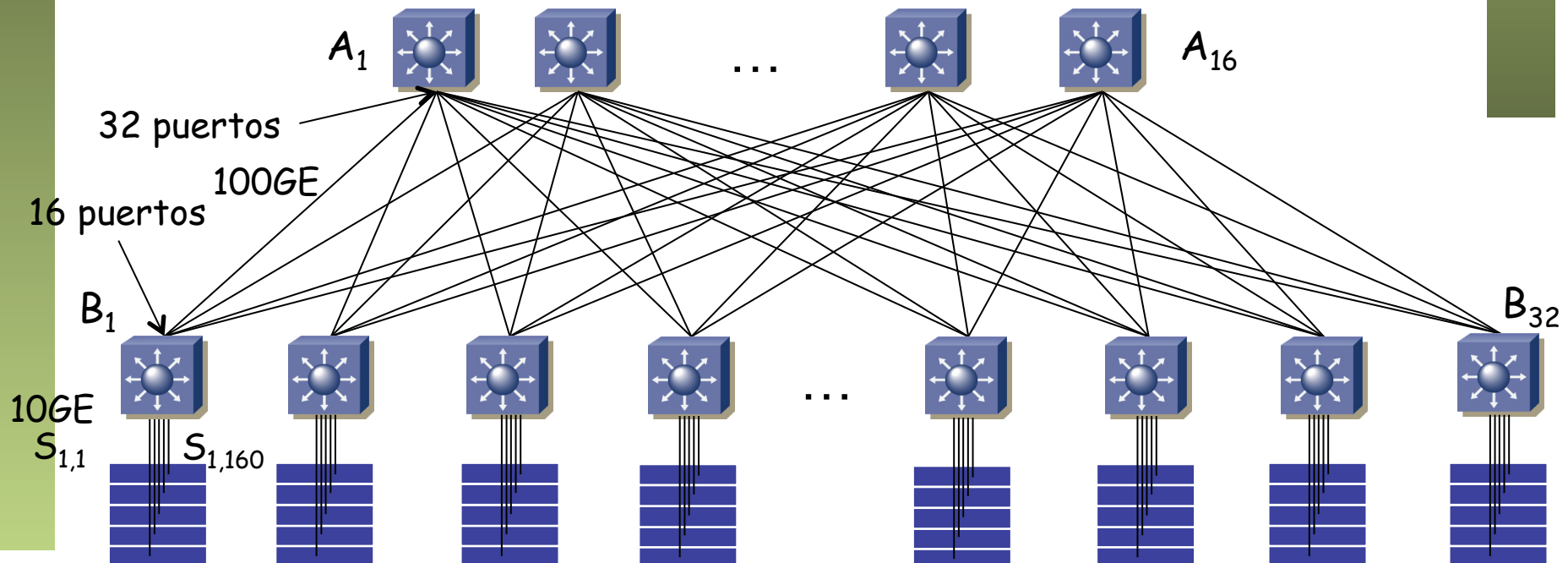
Ejemplo

- Cada conmutador de acceso da conectividad 10GE a 160 interfaces
- Recibe así un máximo de $160 \times 10 = 1.6$ Tbps
- 32 conmutadores en la capa de acceso
- En total $160 \times 32 = 5120$ servidores
- 16 conmutadores en la capa de agregación
- Enlaces de 100GE entre acceso y agregación
- Entonces de cada conmutador de acceso salen $16 \times 100 = 1.6$ Tbps
- Over-subscription 1:1 en la capa de acceso
- (...)



Ejemplo

- De cada conmutador de acceso salen $16 \times 100\text{GE} = 1.6 \text{ Tbs}$
- Cada conmutador de agregación recibe 32 enlaces 100GE (3.2 Tbps)
- En total puede haber fluyendo $32 \times 16 \times 100 = 51.2 \text{ Tbps}$
- Sin bloqueo si no tienen bloqueo interno los conmutadores
- $32 \times 16 = 512$ enlaces entre los conmutadores
- Eso son bastantes cables a tender sin errores
- Y si quieres hacer cambios en la topología o ampliarla es costoso



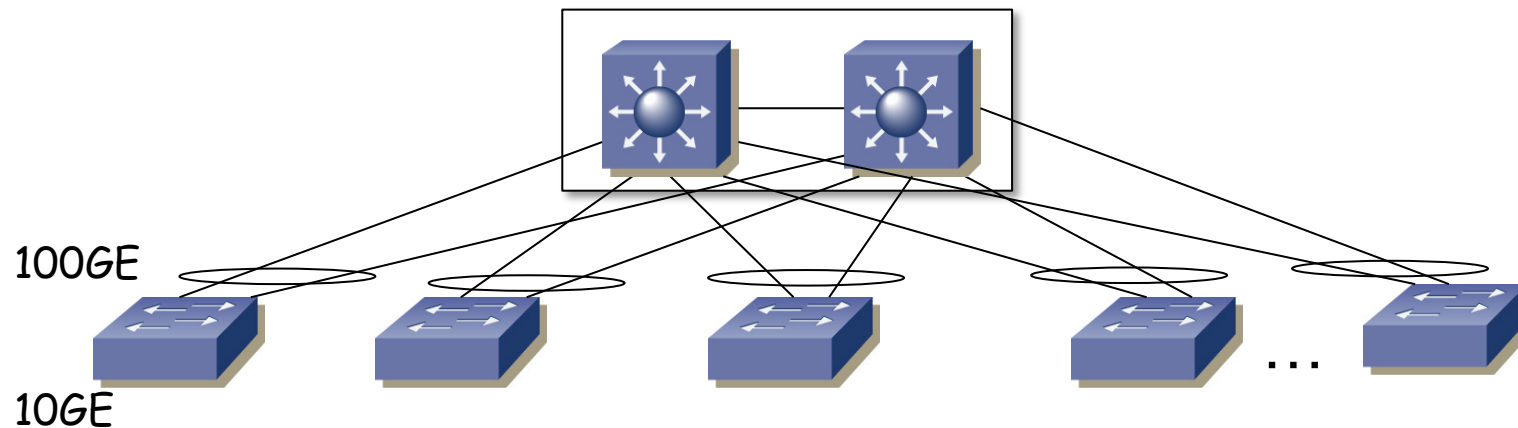
Escalabilidad

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
 - $144 \times 11 = 1584$ 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

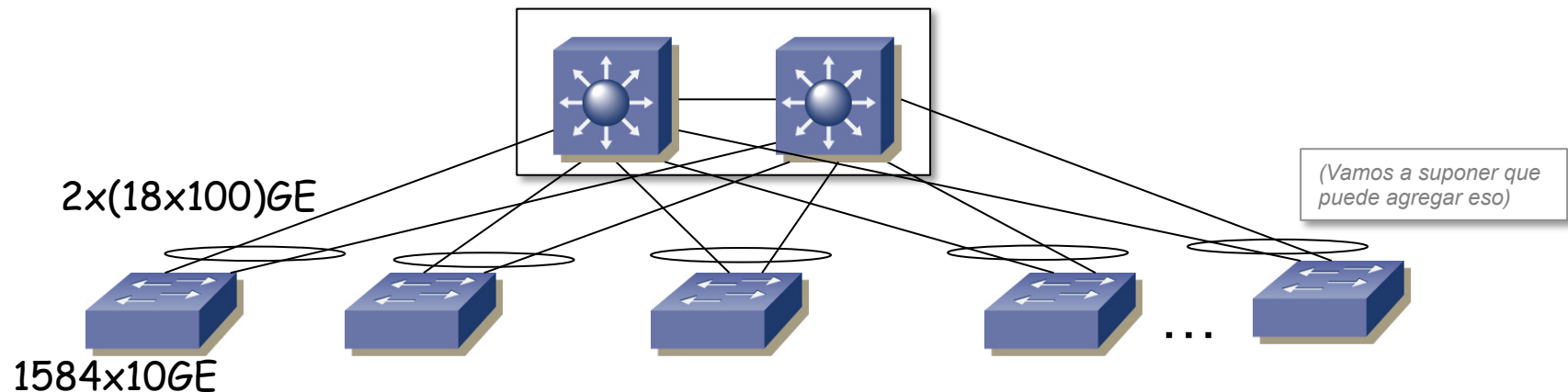


QSFP+ = Quad Small Form-factor Pluggable

<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 ($15840/3600 = 4.4$)
- 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 $12 \times 36 = 432$ 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

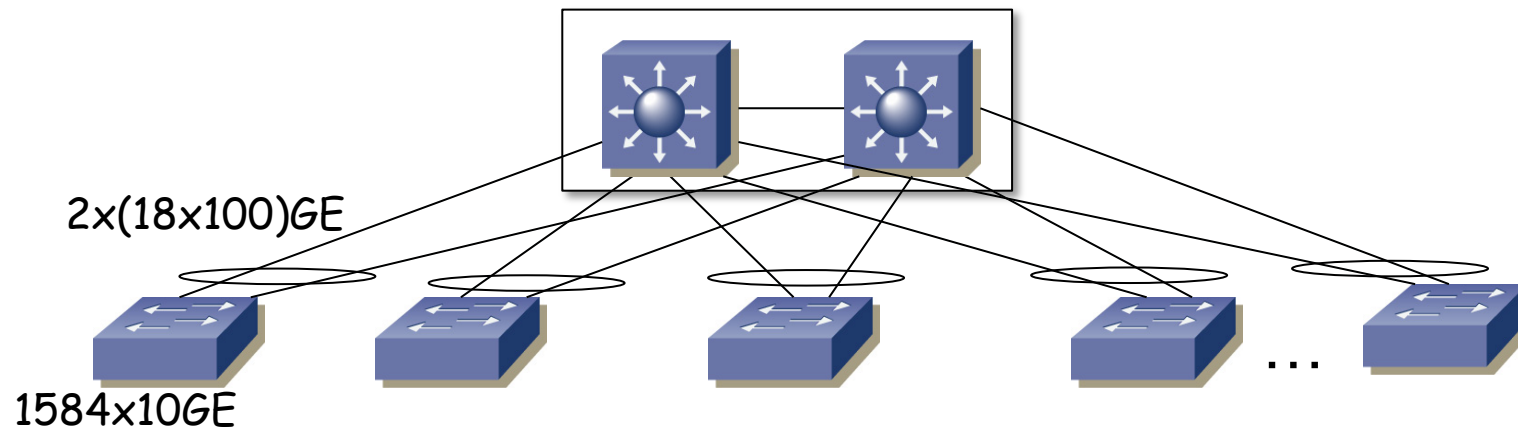


QSFP+ = Quad Small Form-factor Pluggable

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

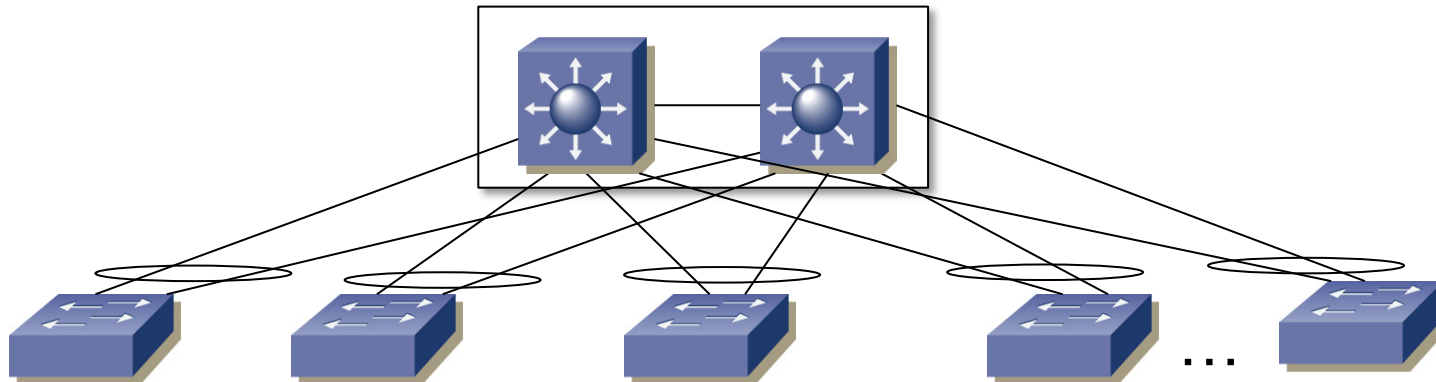
Escalabilidad con MLAG

- 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 ($24 \times 18 = 432$) conmutadores de acceso
- Eso son $1584 \times 24 = 38.016$ hosts con un puerto 10GE cada uno y una sobre-subscripción 4.4 a 1
- ¿1500 hosts por switch? Cableado EoR
- (...)



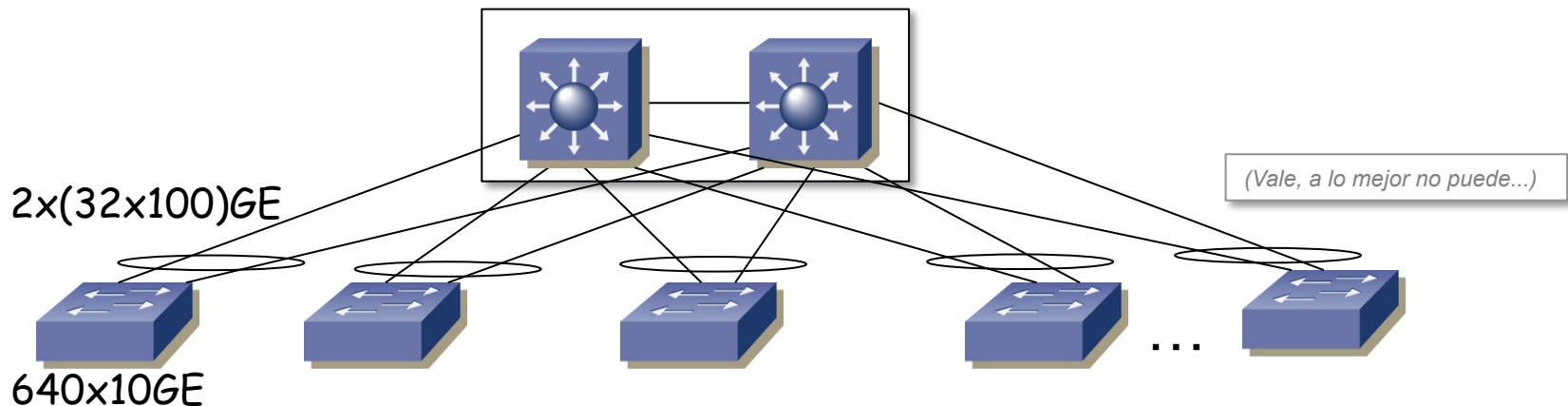
MLAG no bloqueante

- ¿Podríamos hacerlo no-bloqueante?
- Configuración del switch de acceso:
 - $8 \times 144 = 1152$ puertos 10GE (11.5 Tbps)
 - $4 \times 36 = 144$ puertos 100GE (14.4 Tbps) en uplinks
 - Podríamos ajustarlo pues cada puerto 100GE puede sacar 4×10 GE
 - En una de las 4 tarjetas 100GE ponemos 22×100 GE y $(14 \times 4) \times 10$ GE :
 - $8 \times 144 + 56 = 1208$ puertos 10GE (12.1 Tbps)
 - $3 \times 36 + 22 = 130$ puertos 100GE (13 Tbps)
 - En cada switch de acceso 12.9 Tbps desde los hosts para 13 Tbps uplink
 - (...)



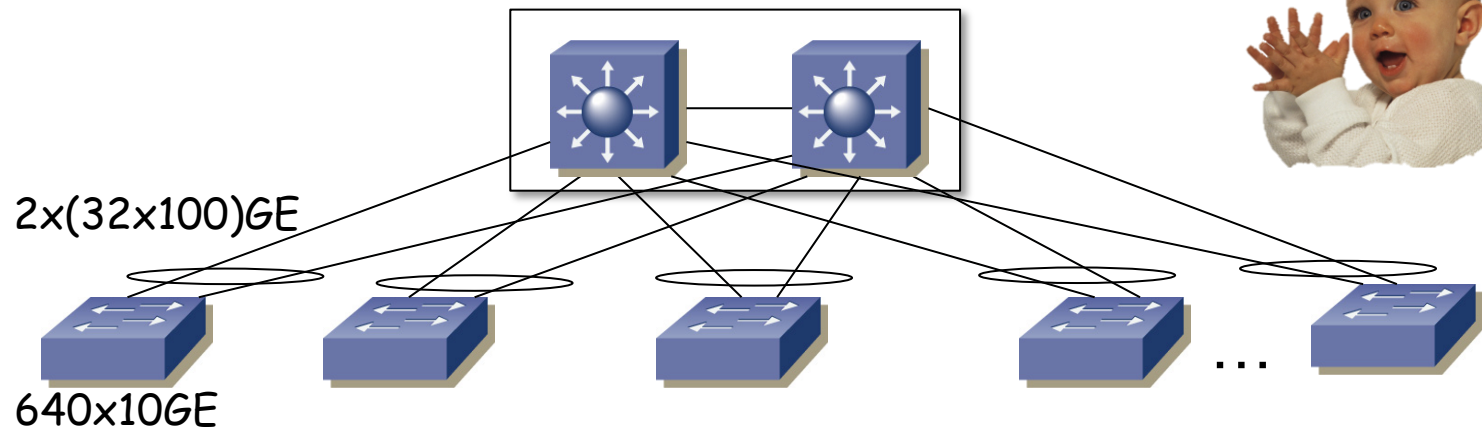
MLAG no bloqueante

- $8 \times 144 + 56 = 1208$ puertos 10GE (12.1 Tbps)
- $3 \times 36 + 22 = 130$ 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:
 - $64 \times 100 = 6.4$ Tbps en uplinks
 - $640 \times 10 = 6.4$ Tbps a hosts (quedan slots sin usar)
- Hemos perdido puertos a hosts para mantener la sobre-subscripción 1:1
- (...)



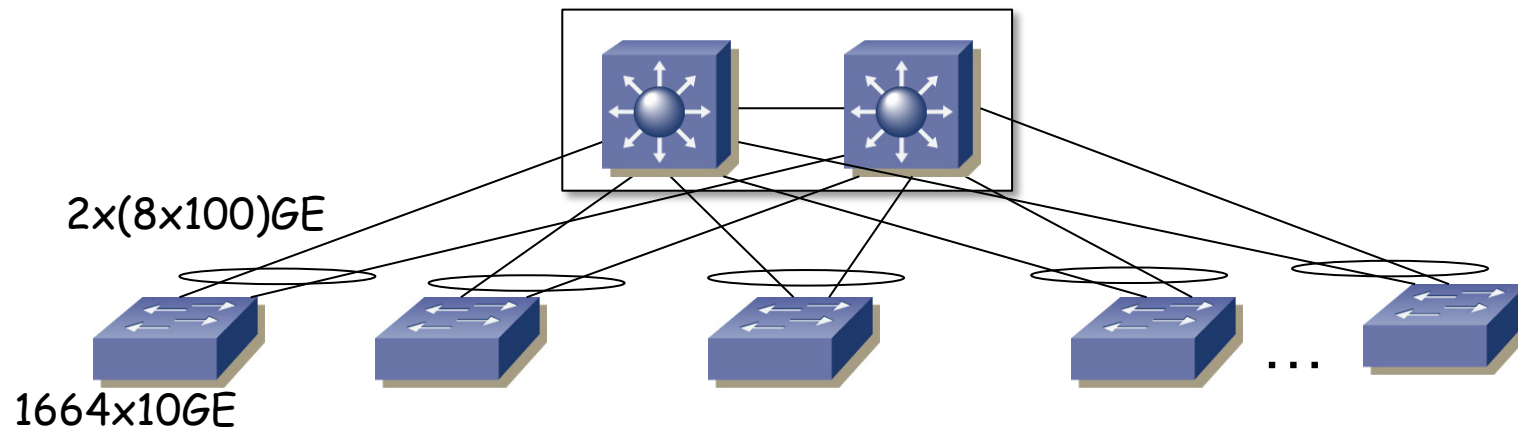
MLAG no bloqueante

- El switch de agregación podía tener 432 puertos 100GE
- Cada 32 puertos a un switch así que tenemos hasta 13 conmutadores de acceso ($14 \times 32 = 448 > 432$)
- Así que en total $13 \times 640 = 8.320$ hosts con puertos 10GE non-blocking
- Antes teníamos 38.016 hosts con sobre-subscripción 4.4:1



Escalabilidad con MLAG

- ¿Y si nos permitimos una sobre-subscripción mayor?
- Por ejemplo:
 - $12 \times 144 = 1728$ puertos 10GE
 - Por cada puerto 100GE quitamos 4 puertos 10GE
 - Ponemos 16 puertos 100GE de uplink nos quedan 1664 puertos 10GE
 - Tenemos 16640:1600 o una sobre-subscripción 10:1
 - Ahora un switch de agregación de 432 puertos 100GE agrega 54 conmutadores de acceso ($432/8=54$)
 - Así que un total de $54 \times 1664 = 89.856$ hosts con un 10:1 aprox.

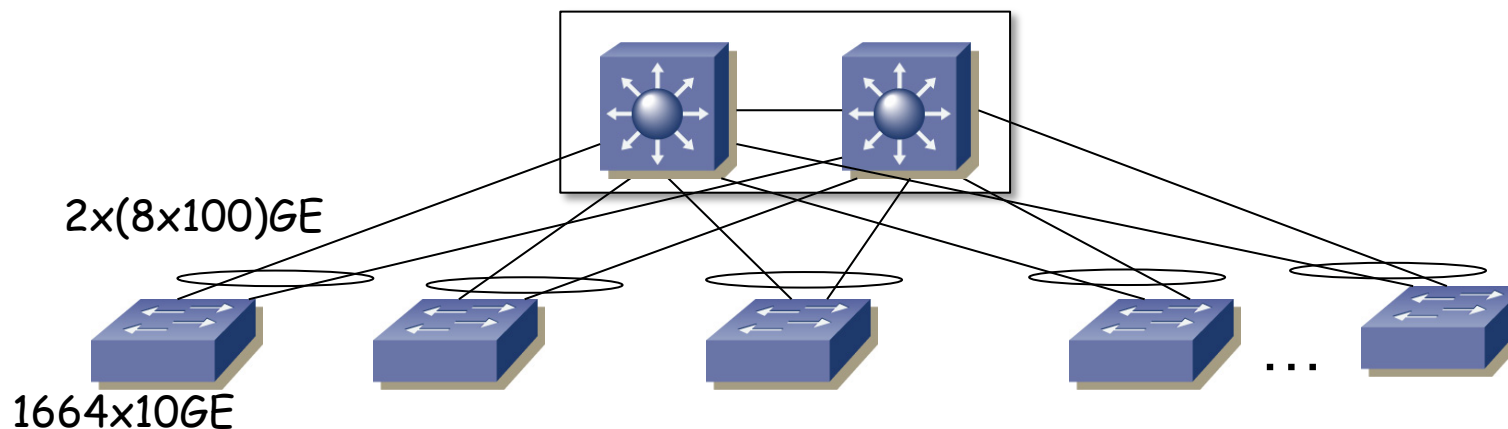


Escalabilidad con MLAG

- 89.856 puertos 10GE
- Supongamos que en cada host tenemos una máquina con 2 CPUs de 8 cores
- Y por ejemplo decidimos correr 20 VMs
- Eso son 20×89.856 o casi 1.8 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

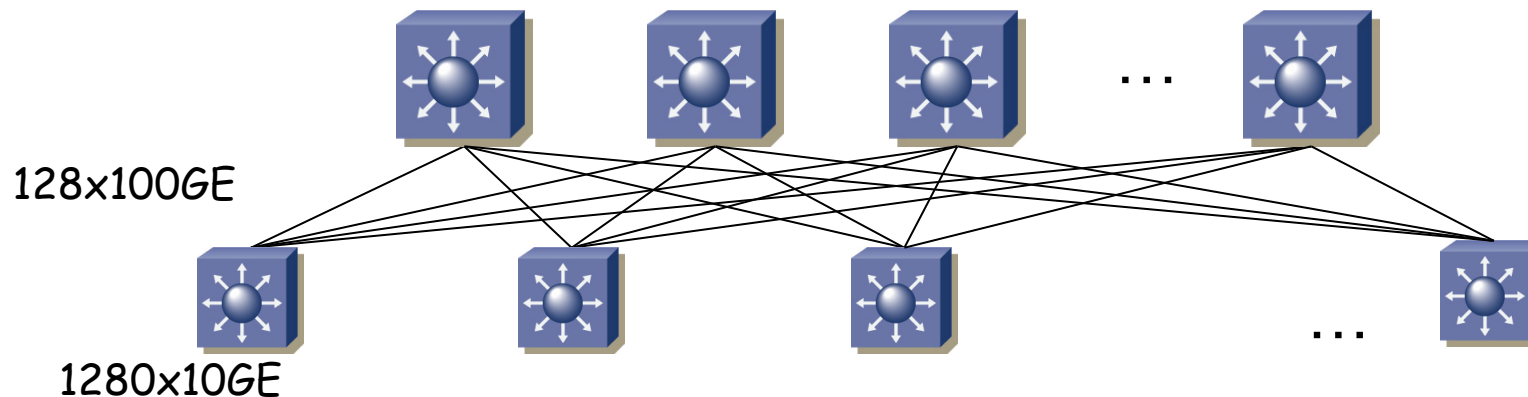


Escalabilidad con ECMP

- 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 $432 \times 1280 = 552.960$ hosts
- ¿Cables? $128 \times 432 = 55.296$
- 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 contruidos a partir de chips de conmutación interconectados
 - Pueden tener sobre-subscripción interna (bloqueo interno)

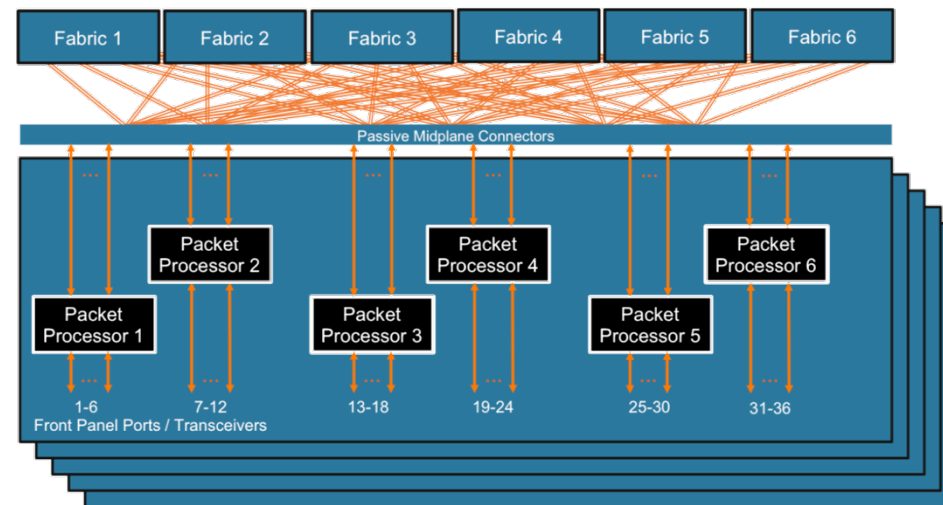
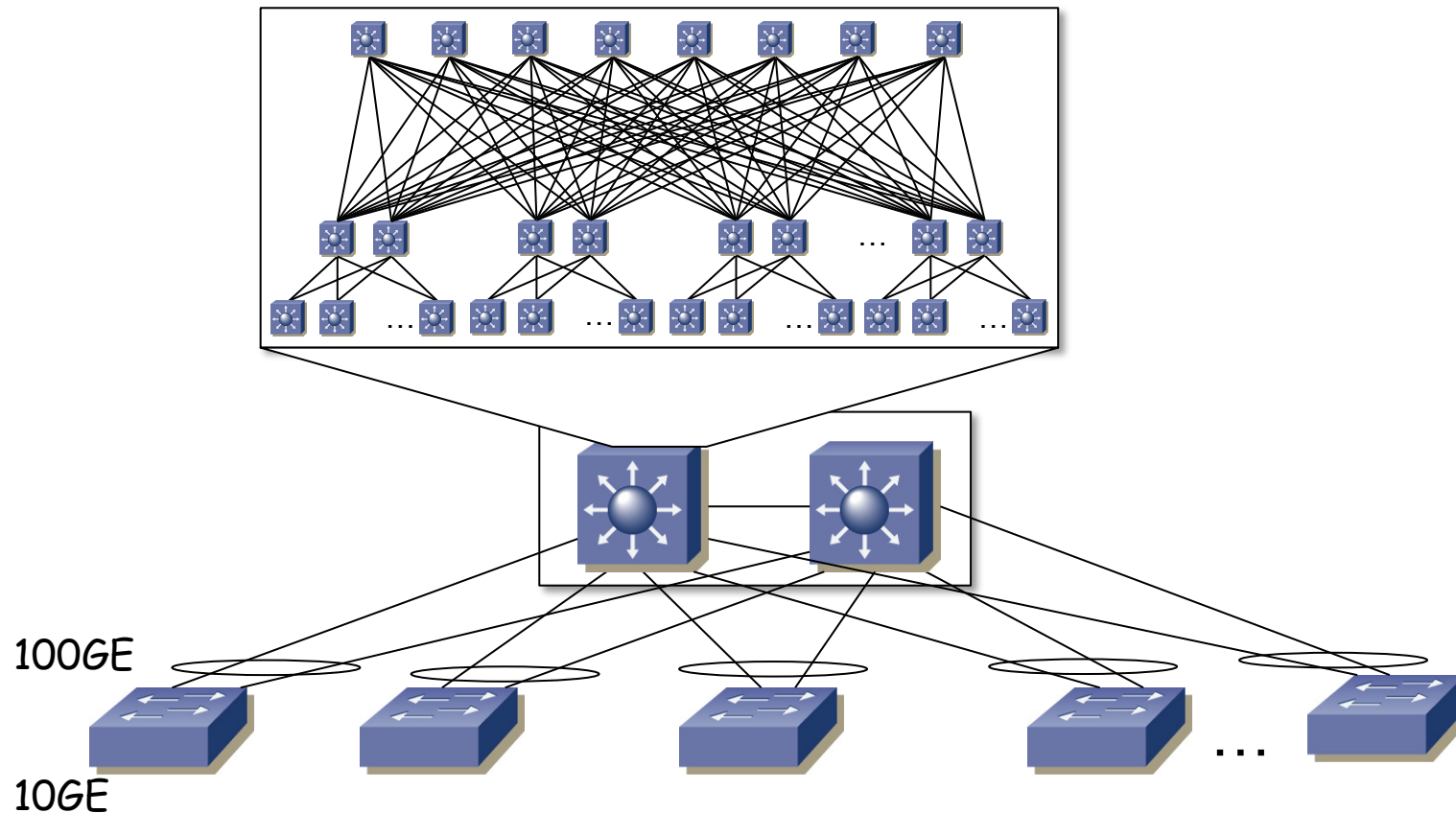


Figure 3: Distributed Forwarding within an Arista 7500R Series

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
- (...)



¿Otra solución?

- 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
- Otra alternativa es crear el data center con conmutadores sencillos
- Elementos

