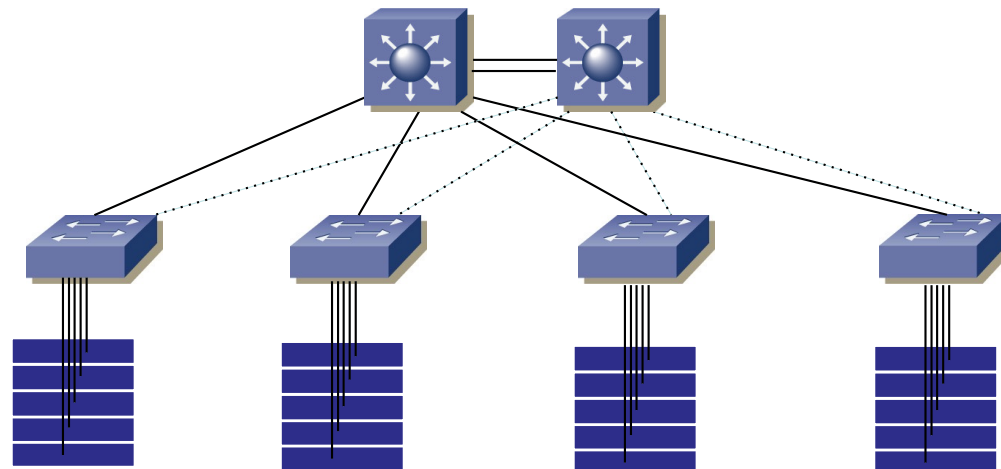


Arquitectura tradicional en el data center: limitaciones

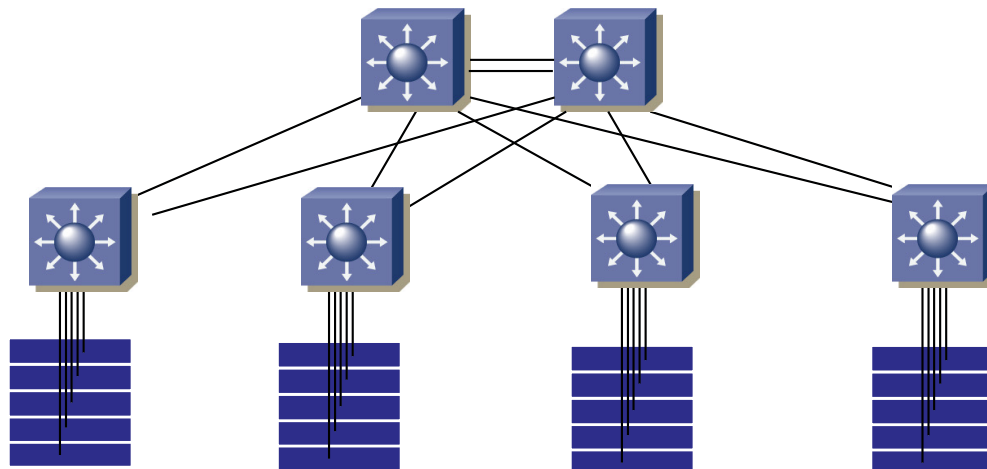
DC tradicional de 2 capas

- 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
- (...)



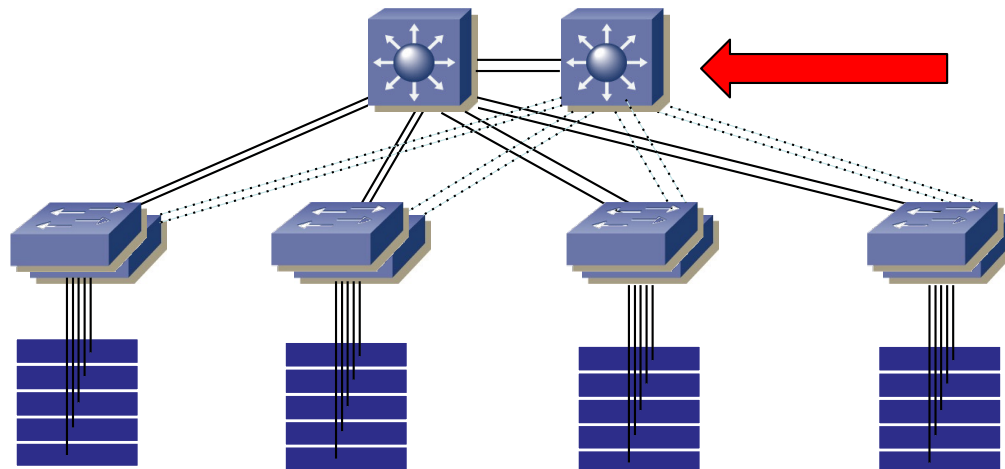
DC tradicional de 2 capas

- 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



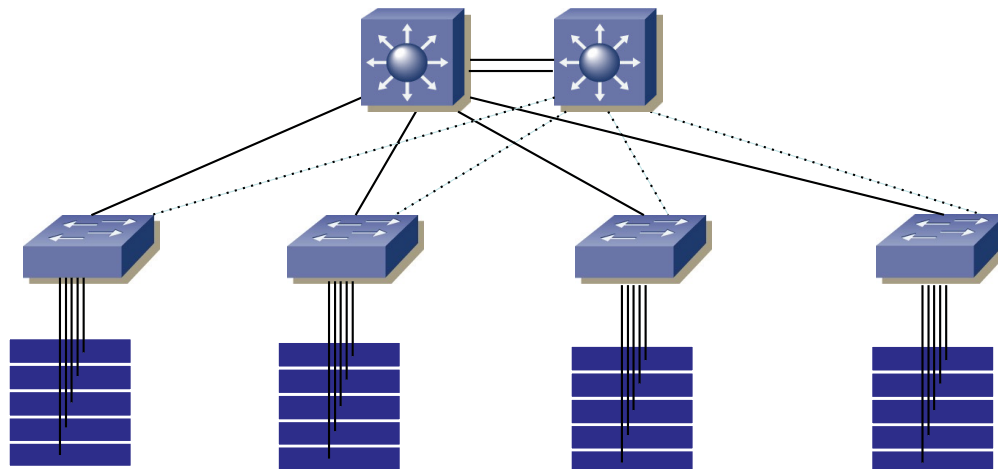
Límites con 2 capas

- 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 (...)



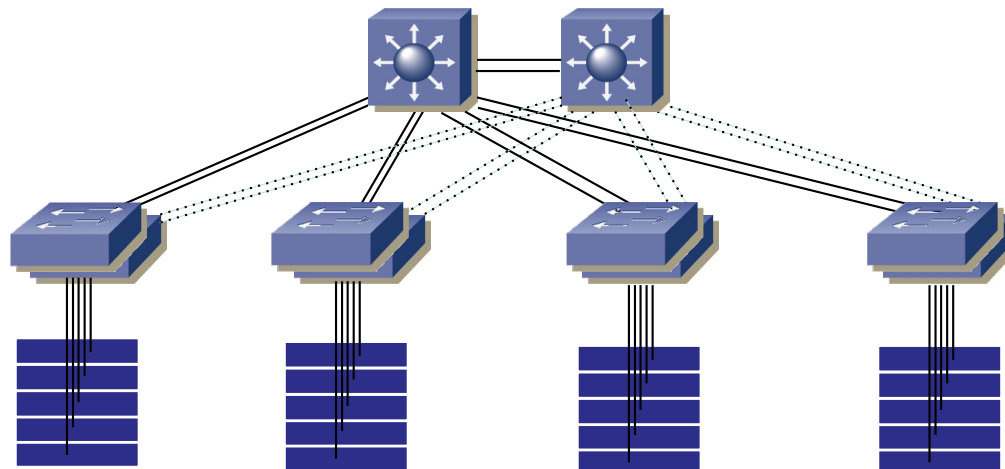
Límites con 2 capas

- 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
 - Máximo de $48 \times 64 = 3072$ servidores
 - (...)



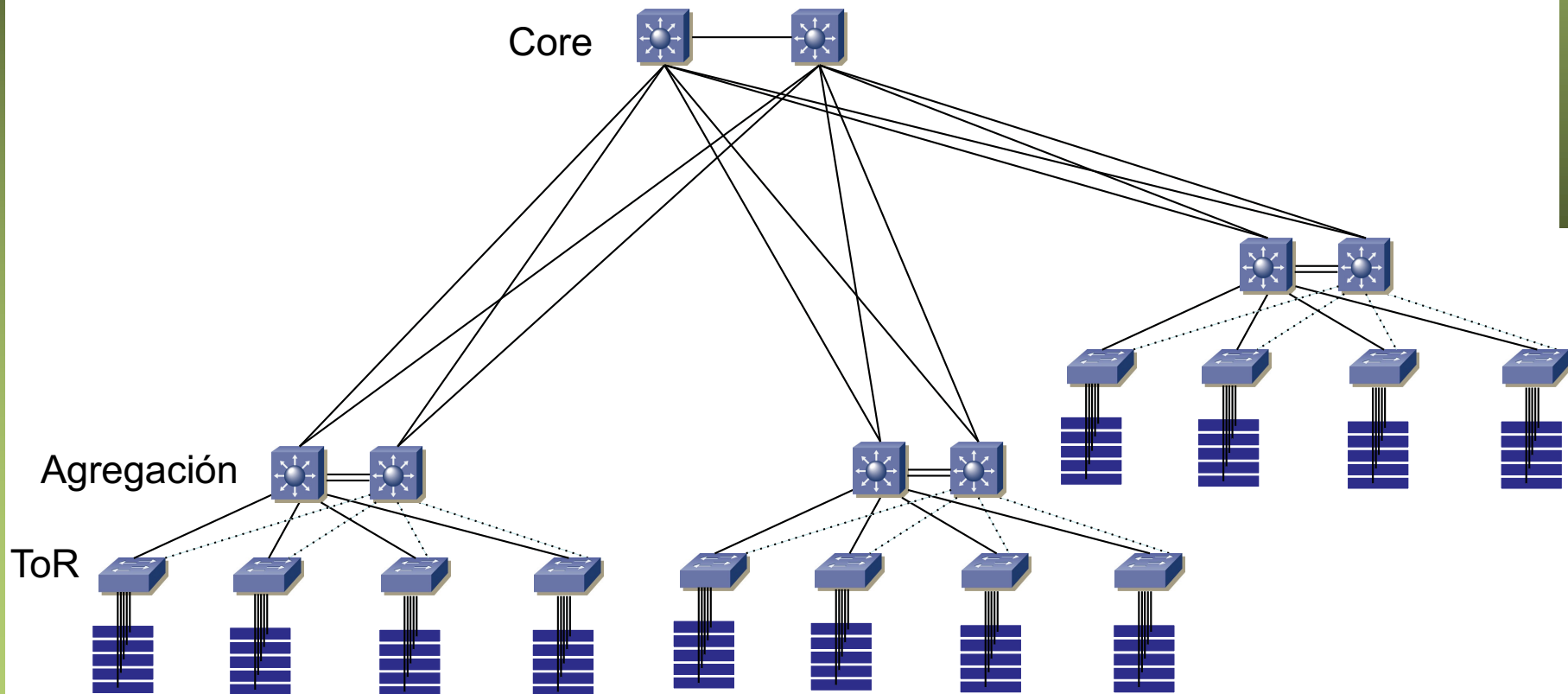
Límites con 2 capas

- 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 n° de servidores por armario pero cada sw. de agregación consume 2 puertos/rack
 - Máximo $48 \times 64 / 2 = 1536$ servidores
- ¿Más?



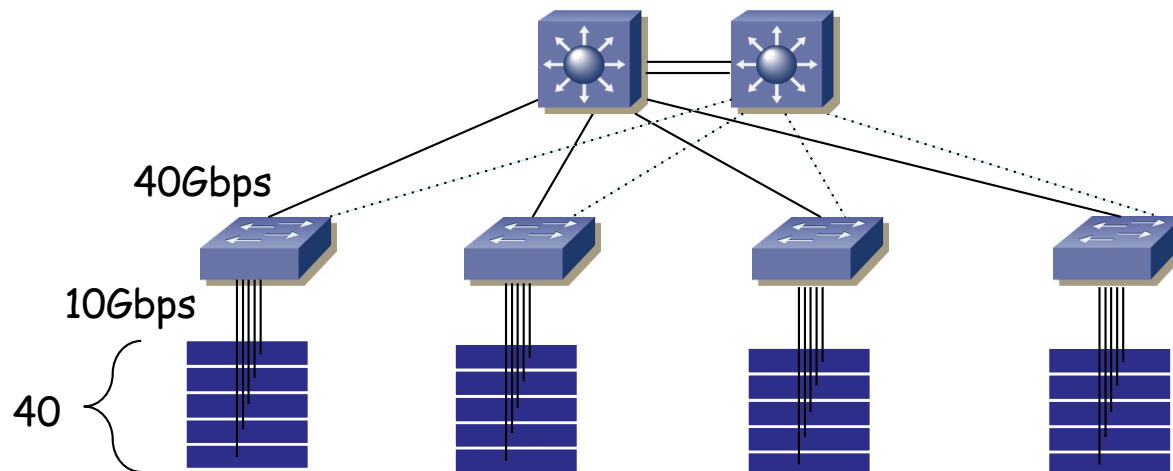
3 capas

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



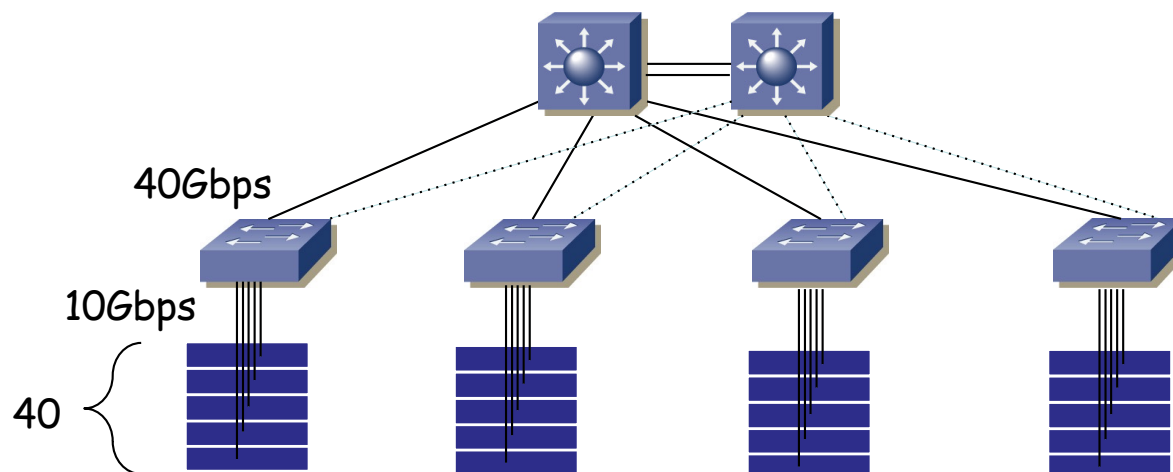
Over-subscription

- 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
 - (...)



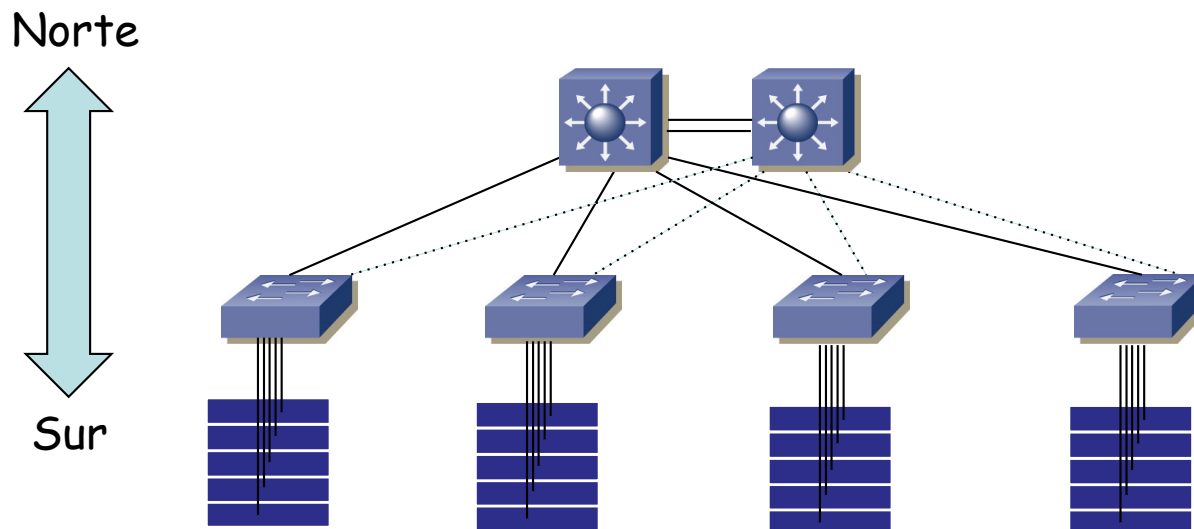
Over-subscription

- 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



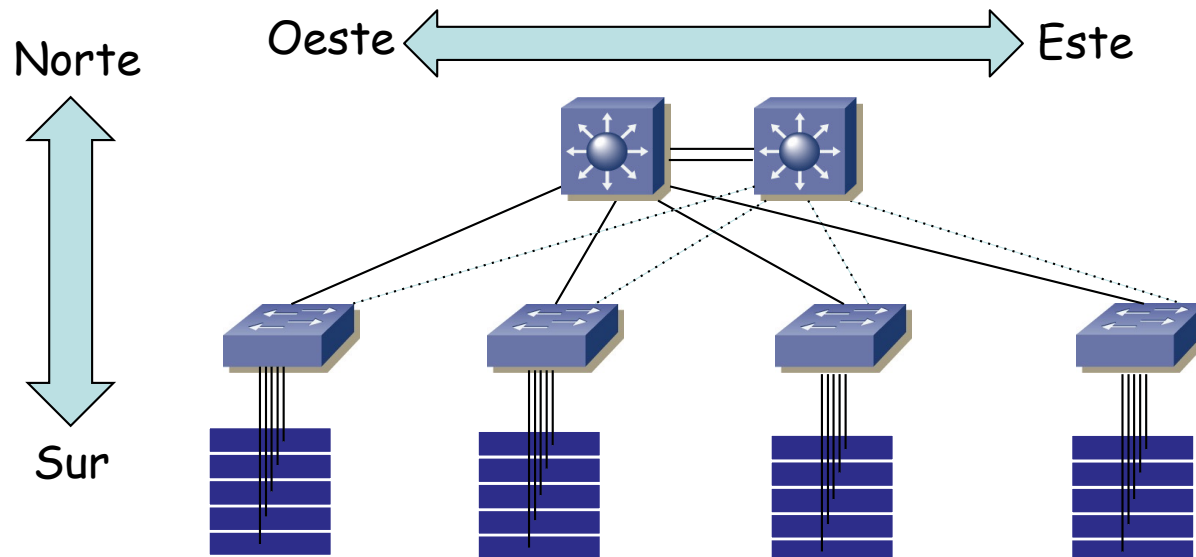
Over-subscription

- 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
- (...)



Over-subscription

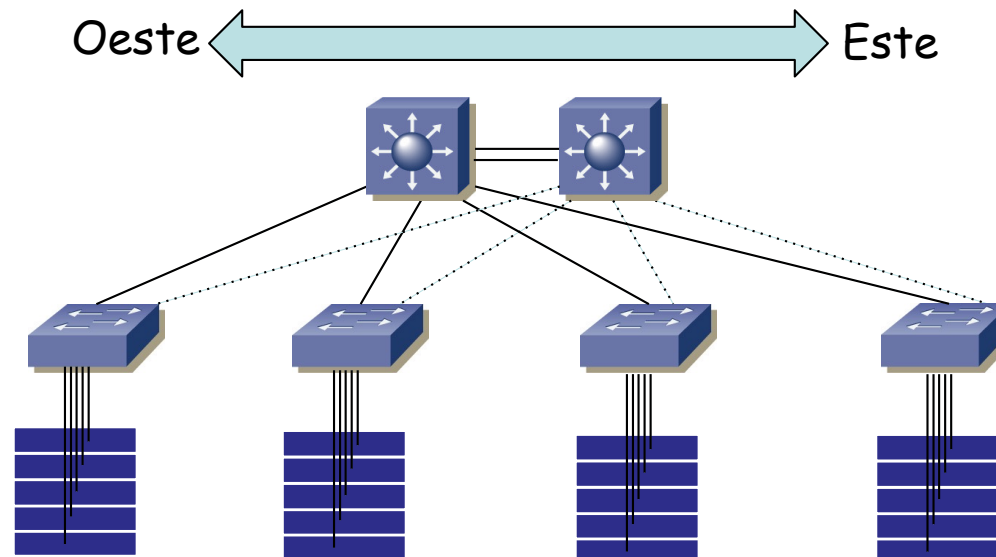
- Altos ratios de sobre-subscripción son razonables cuando tenemos mucho tráfico norte-sur
- Por ejemplo hacia una salida a Internet que sea en realidad el cuello de botella
- No son tan razonables cuando hay mucho tráfico este-oeste
 - Tráfico entre los servidores en distinto rack
 - Aplicaciones distribuidas,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



Tráfico Este-Oeste

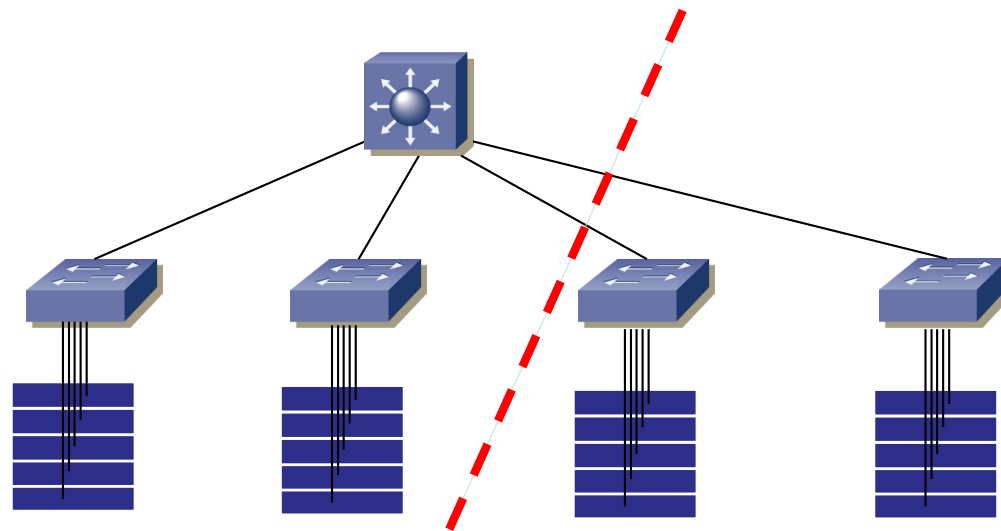
- Ejemplo: Facebook (2013)
- Una petición HTTP se transforma en:
 - 88 búsquedas en caches
 - 35 búsquedas en bases de datos
 - 392 llamadas a procedimientos remotos en el backend
- Una petición con un tamaño de 1KB ha resultado en cerca de 1MB de transferencias en el data center
- ¿Datos almacenados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

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



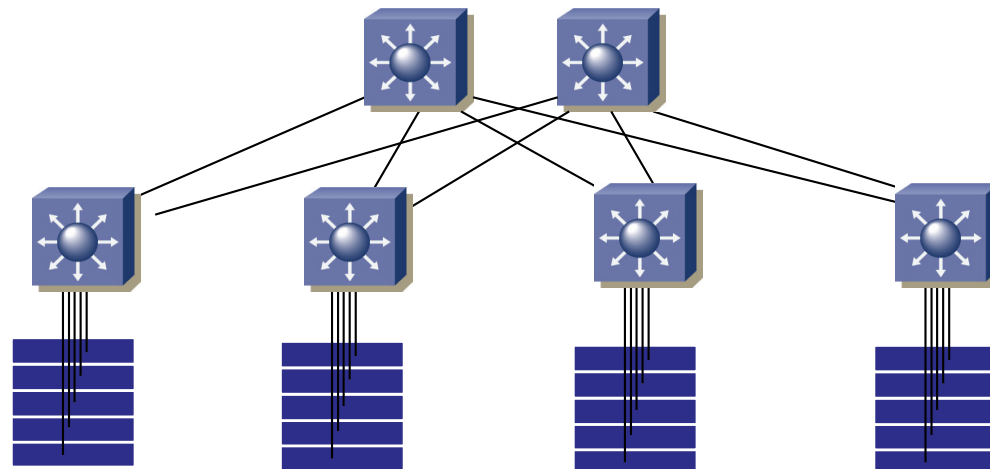
Bisectional bandwidth

- 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



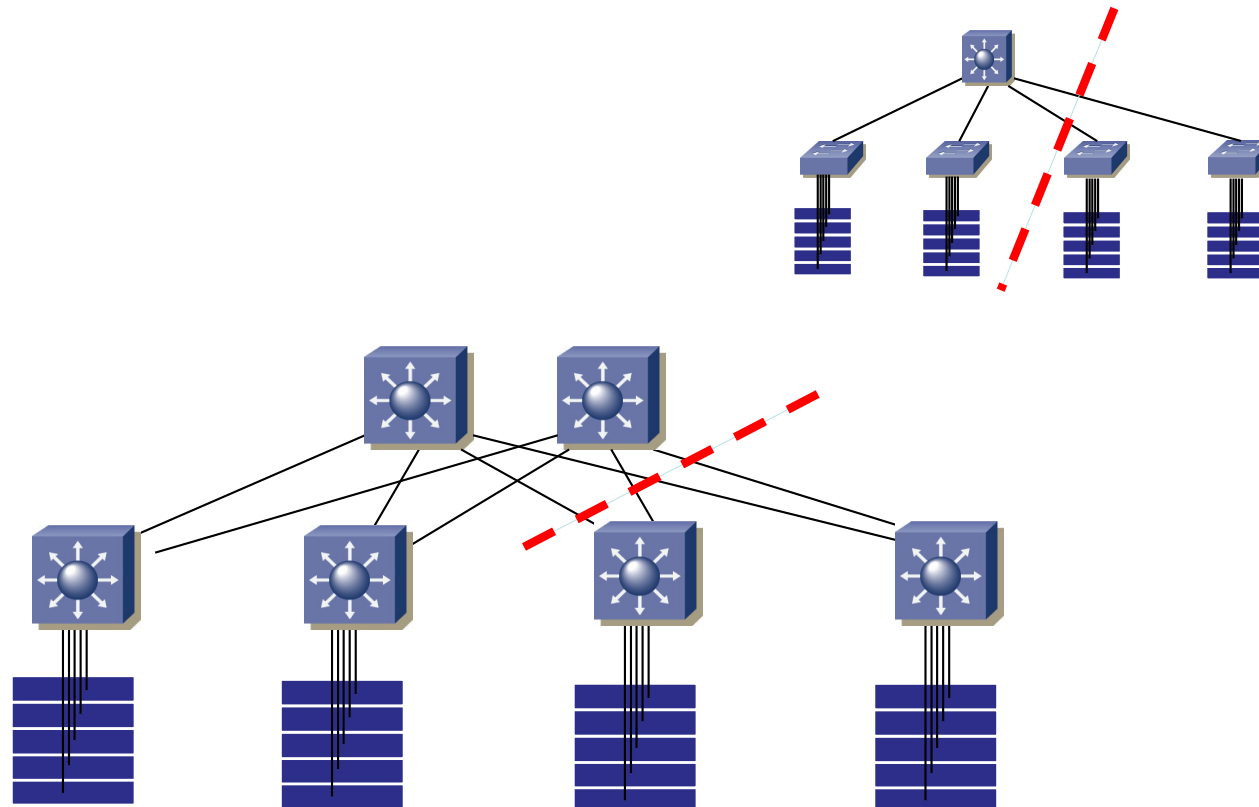
ECMP y biseccional bw

- Aumentamos el ancho de banda de bisección aumentando el número de caminos
- En este caso por ejemplo (...)



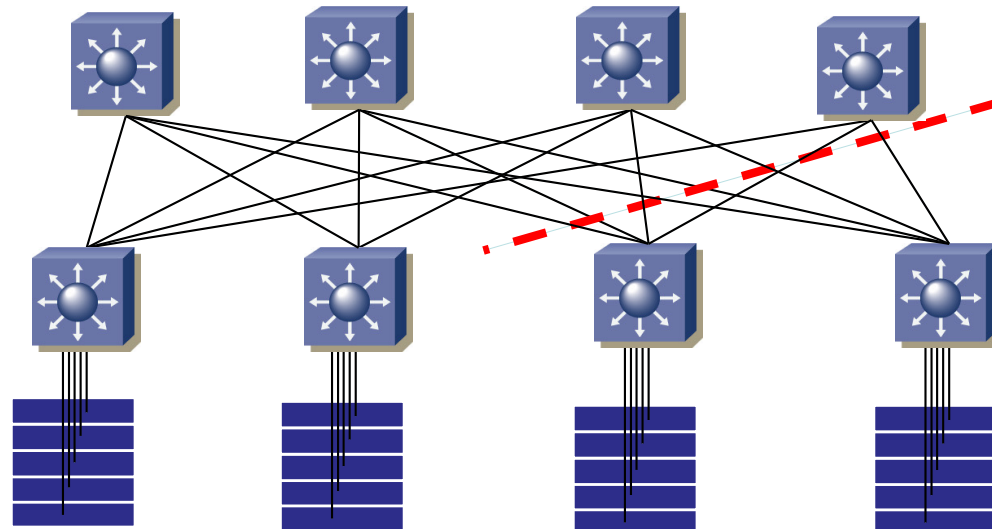
ECMP y biseccional bw

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



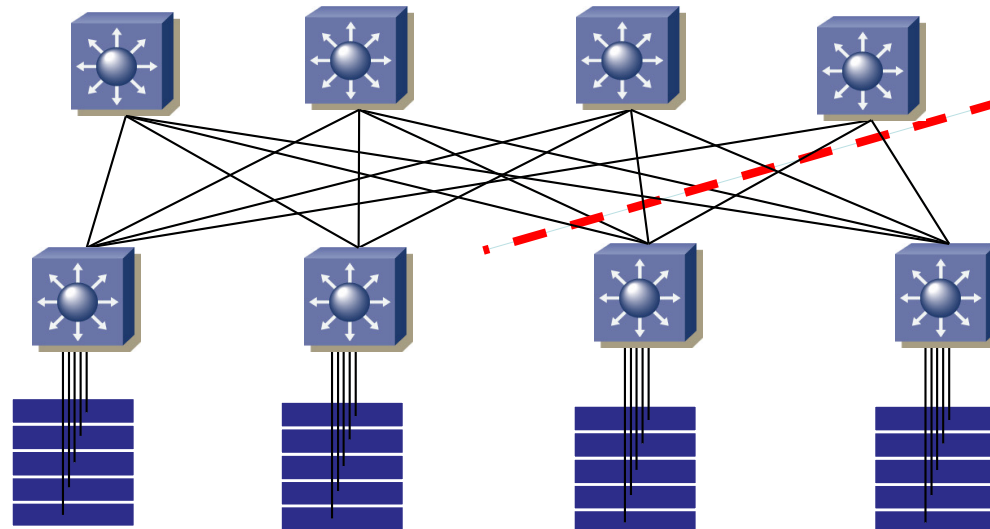
ECMP y biseccional bw

- Aumentamos el ancho de banda de bisección aumentando el número de caminos
- En este caso por ejemplo lo hemos duplicado
- Podríamos seguir aumentándolo
- ¿Inconveniente?



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



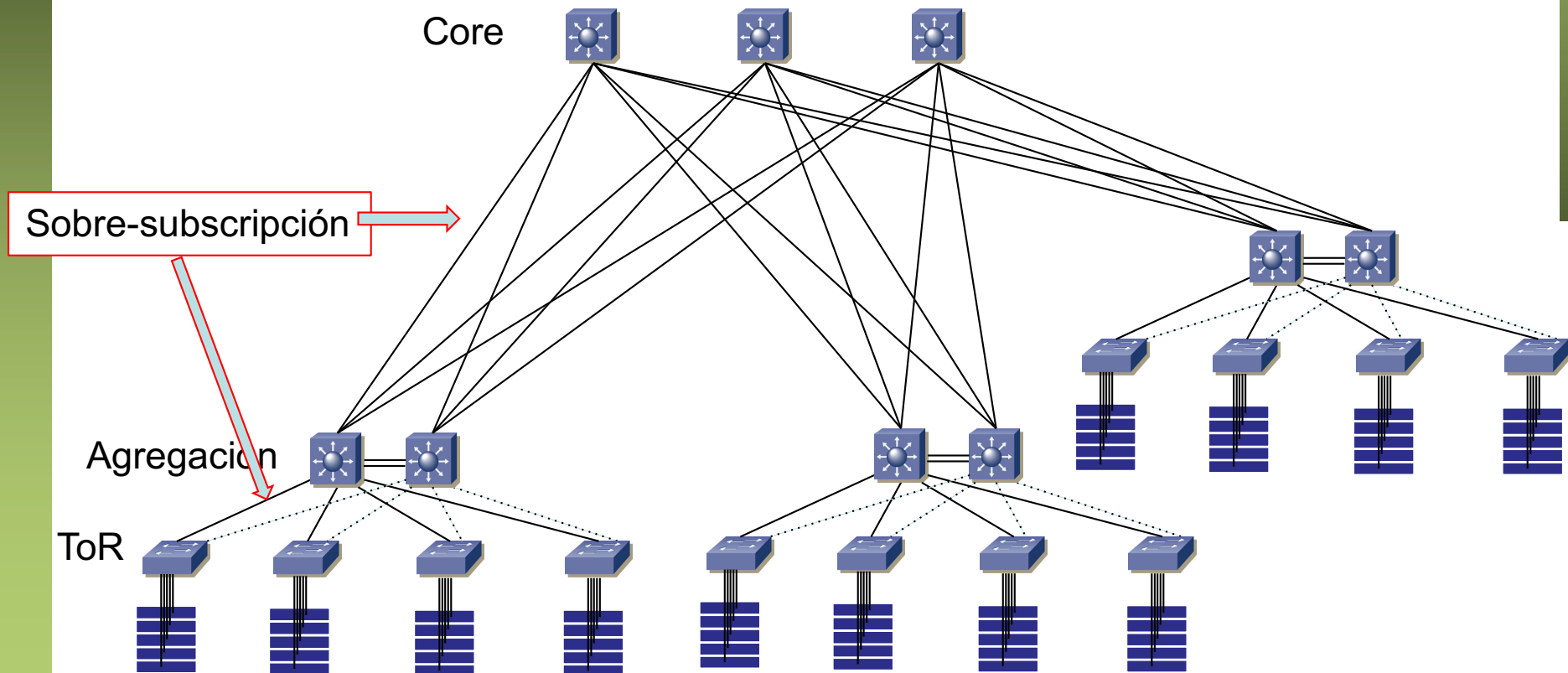
Otro ejemplo

- 200 Access switches con 28 puertos 100 Gb/s c.u.
- 6-way ECMP (6x400 Gb/s)
- Cada access switch $28 \times 100\text{G} = 2.8\text{ Tb/s}$ in, $6 \times 400\text{G} = 2.4\text{ Tb/s}$ out: 1.17:1
- $28 \times 200 = 5600$ puertos 100 Gb/s



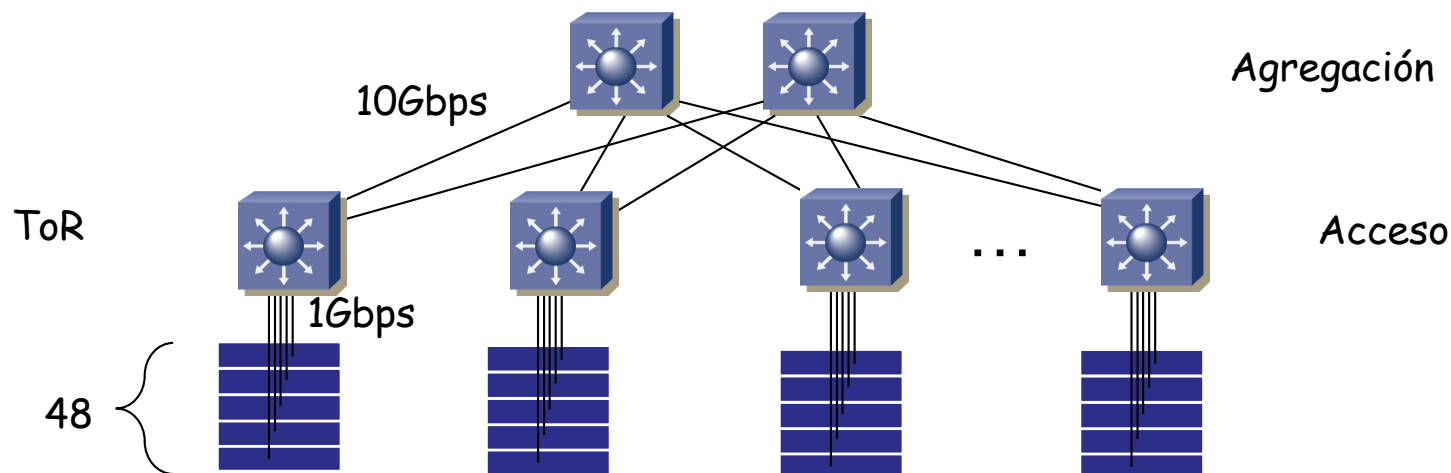
3 capas

- 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 (...)



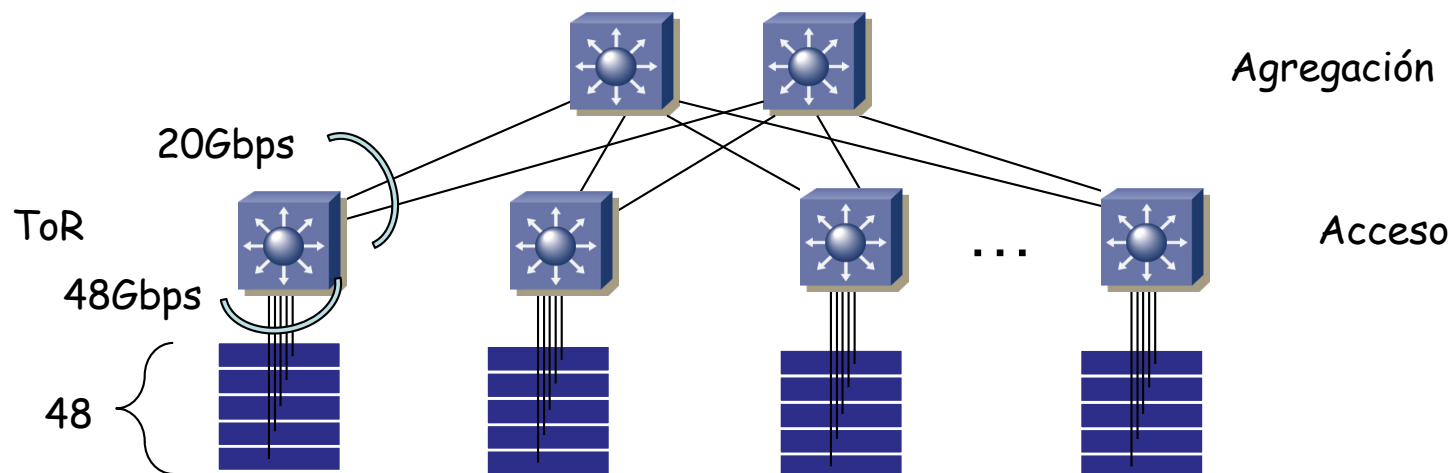
Ejemplo de sobre-subscripción

- ToR de 48x1Gbps + 2x10Gbps
- 48 servidores a 1Gbps por cada ToR
- Enlaces 10Gbps a la capa de agregación (ECMP)
- (...)



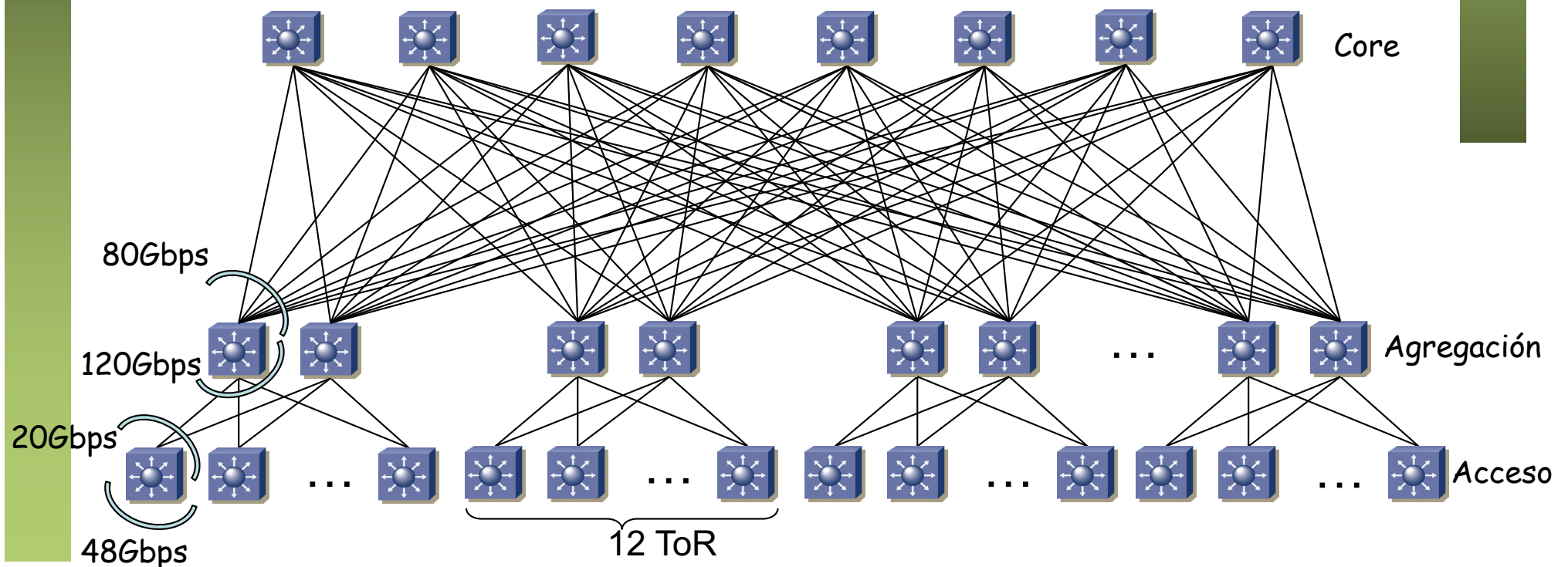
Ejemplo de sobre-subscripción

- 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”



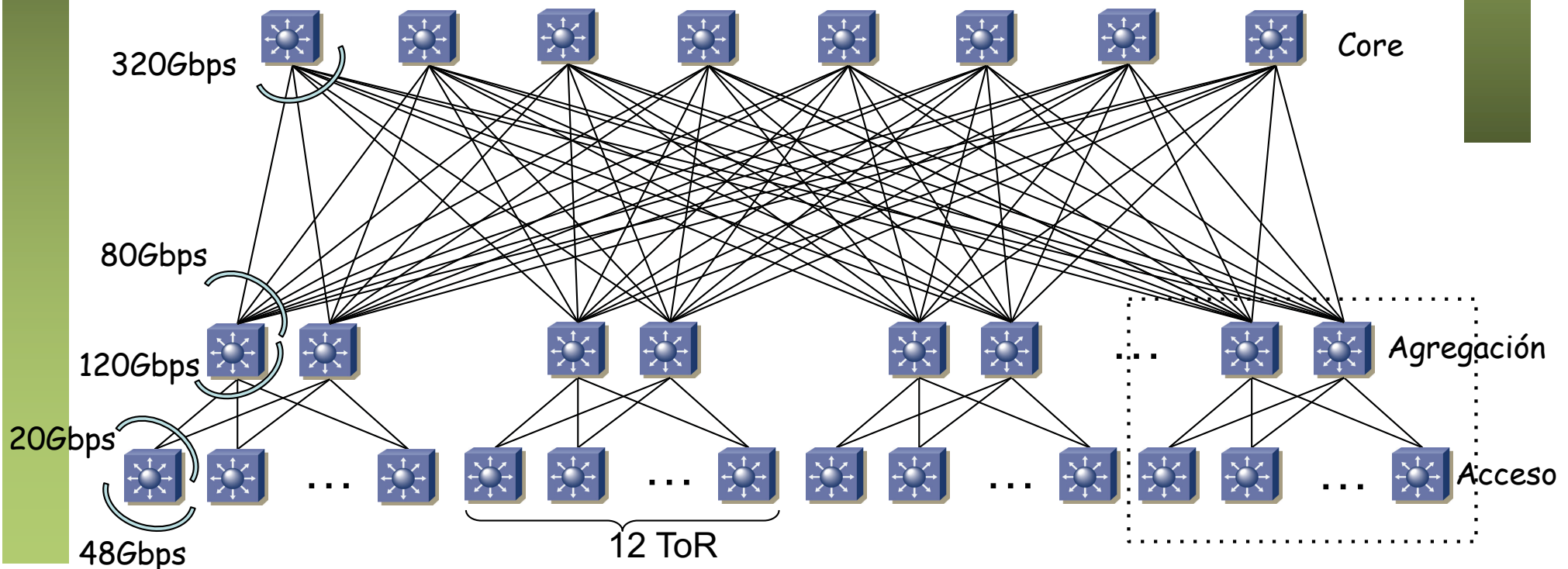
Ejemplo de sobre-subscripción

- 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



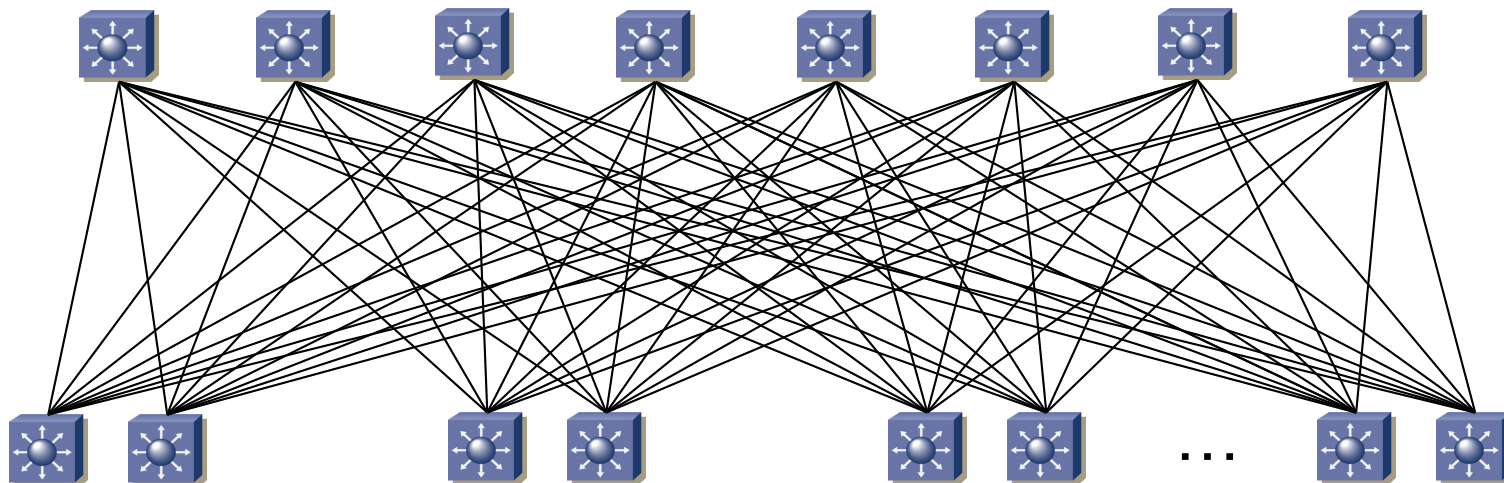
Ejemplo de sobre-subscripción

- 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



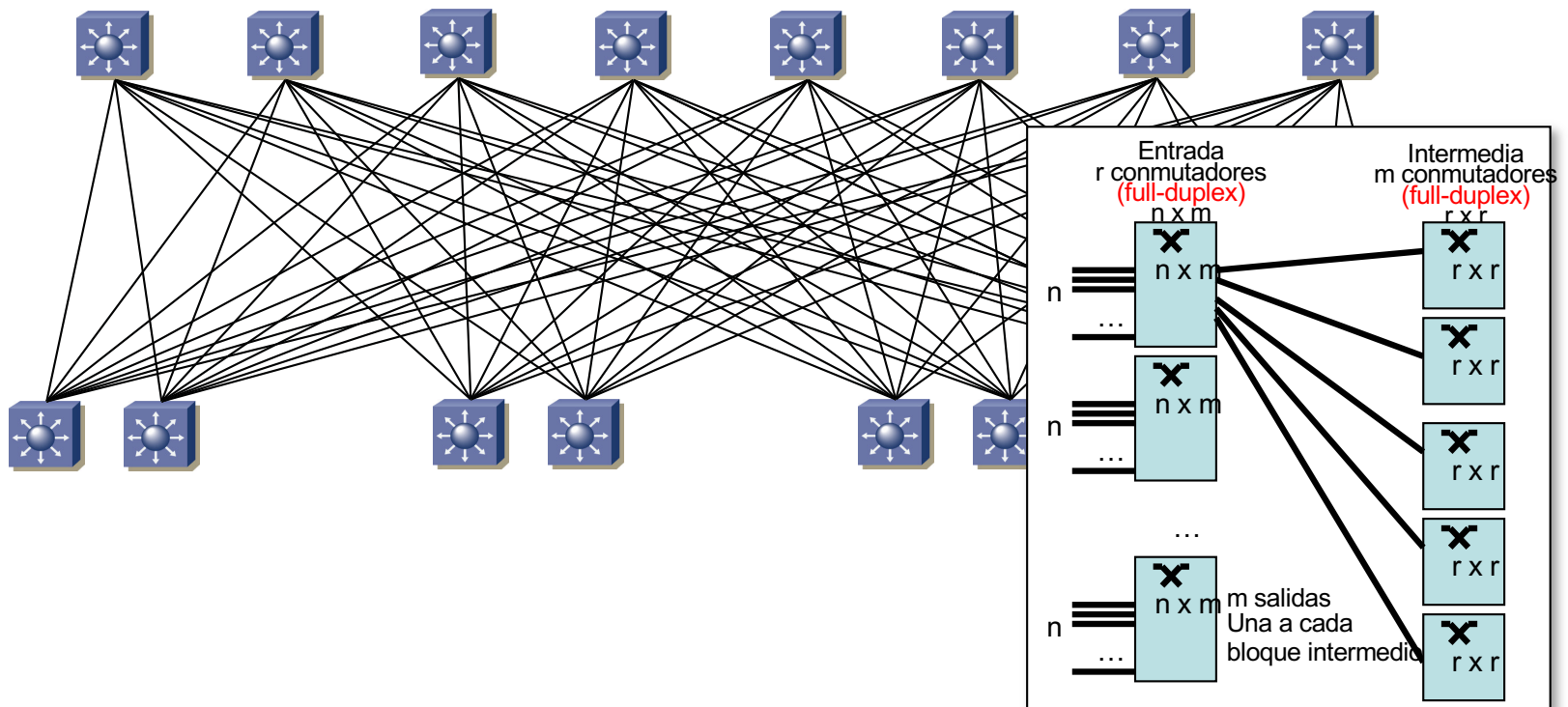
Redes de...

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



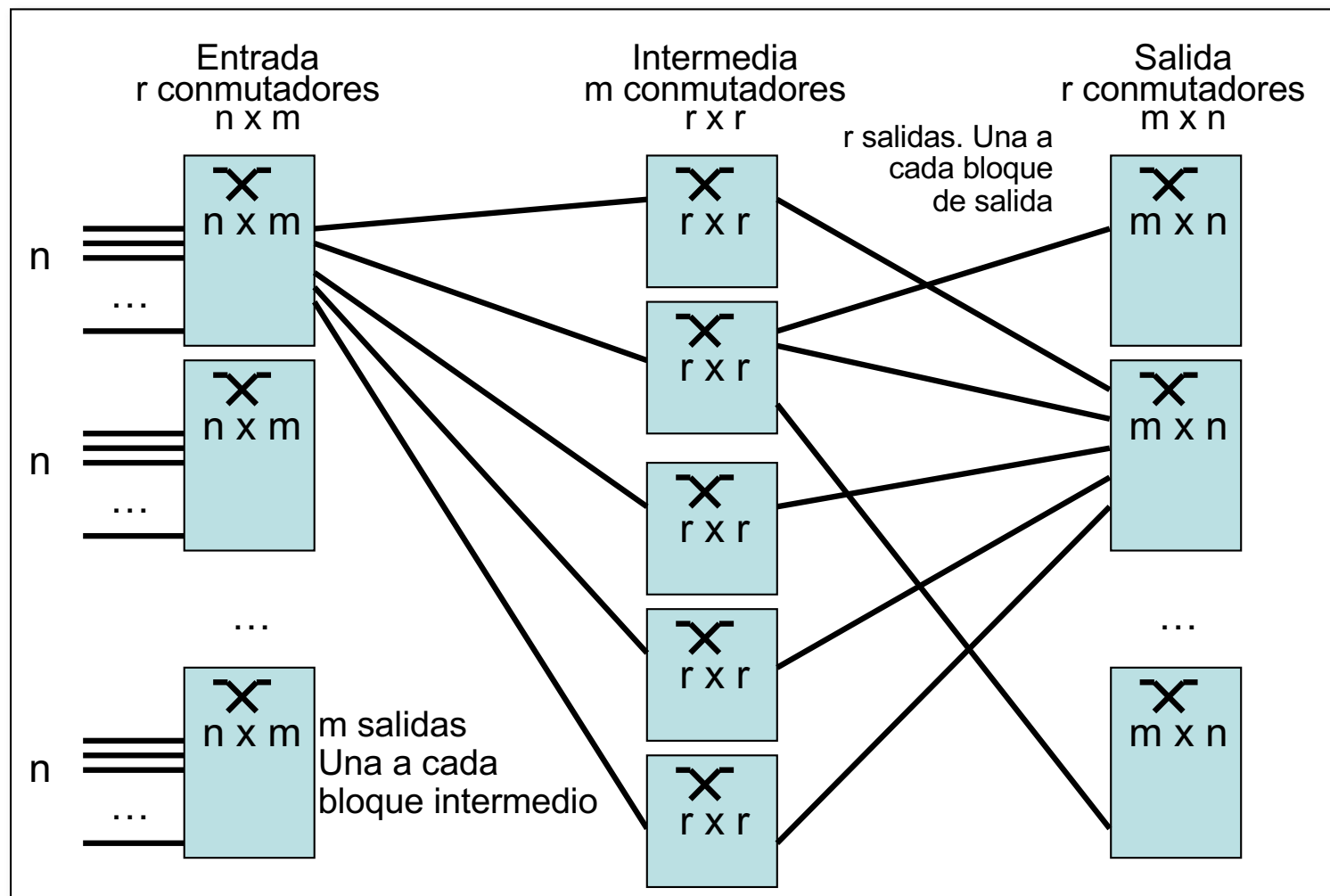
Redes de...

- Cada conmutador de agregación conectado a cada uno de la capa del núcleo
- ¿Clos?



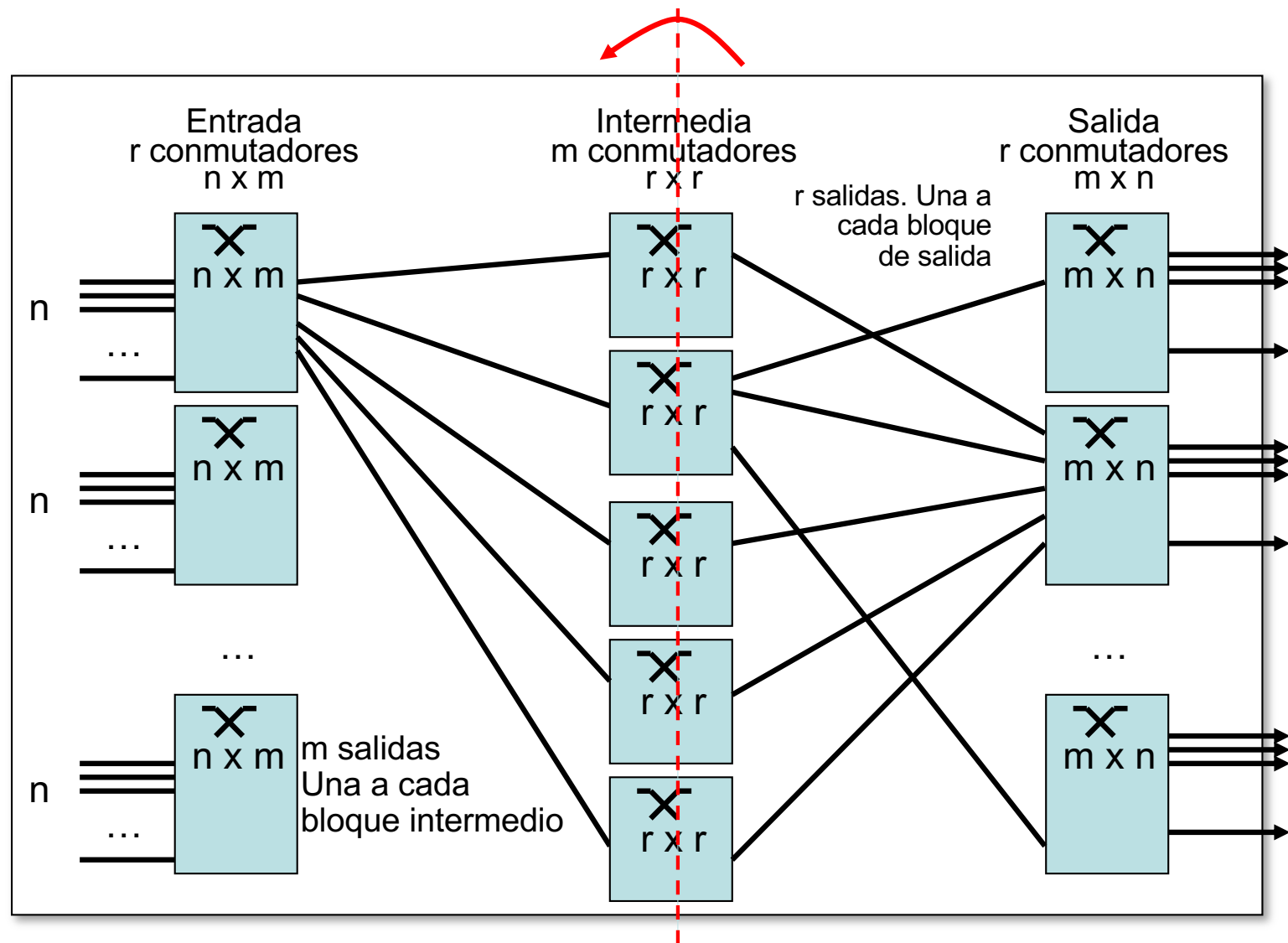
Redes de (Charles) Clos

- Básico en arquitectura de conmutadores de circuitos
- Hay múltiples etapas y los elementos de la etapa x se conectan solo con los de $x-1$ y los de $x+1$ pero no con los de x



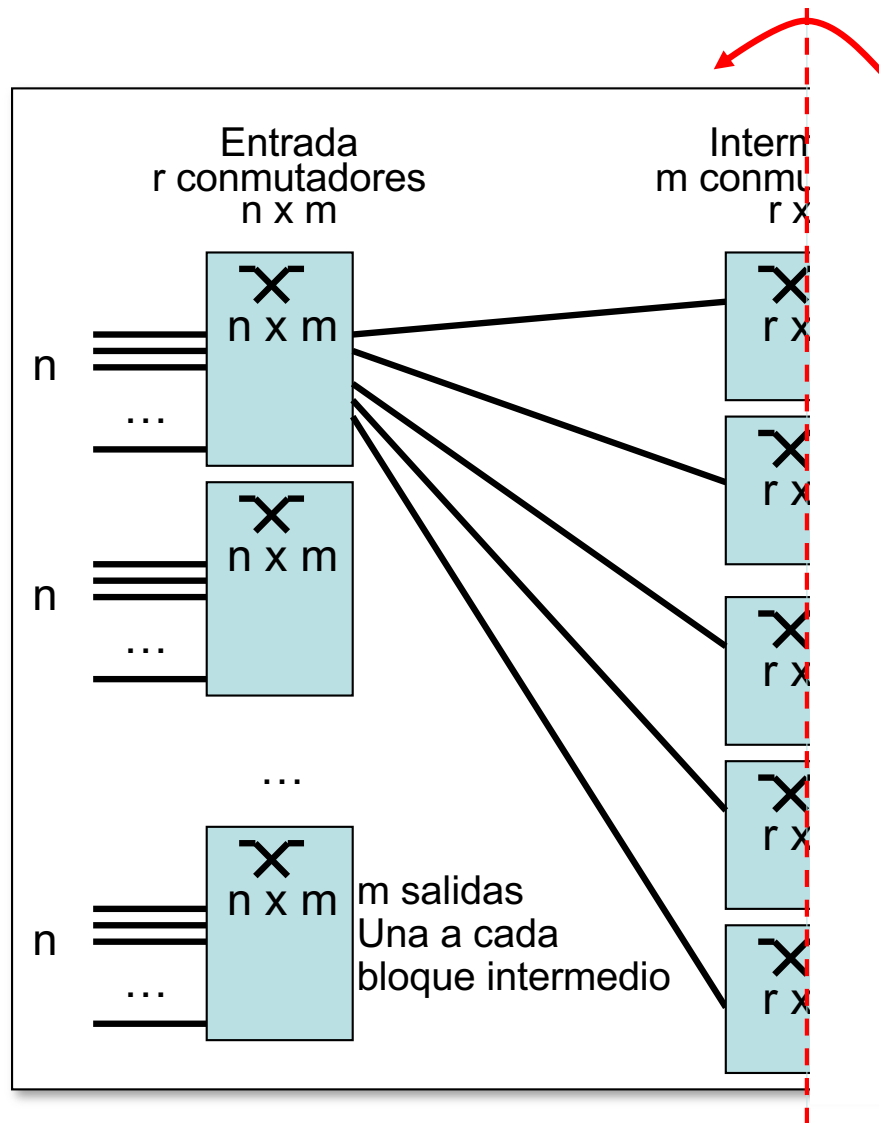
Folded Clos / Fat Tree

- En realidad este diseño es para enlaces unidireccionales
- Con enlaces bidireccionales se “dobla” esta topología (...)



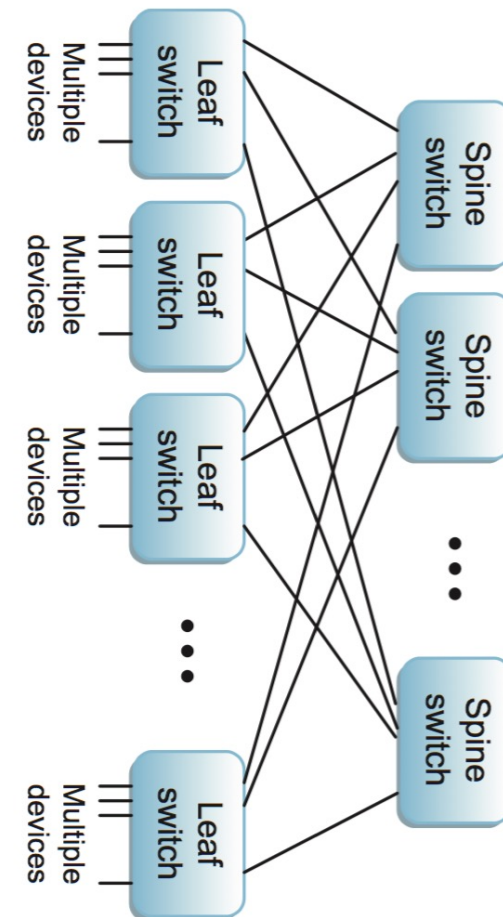
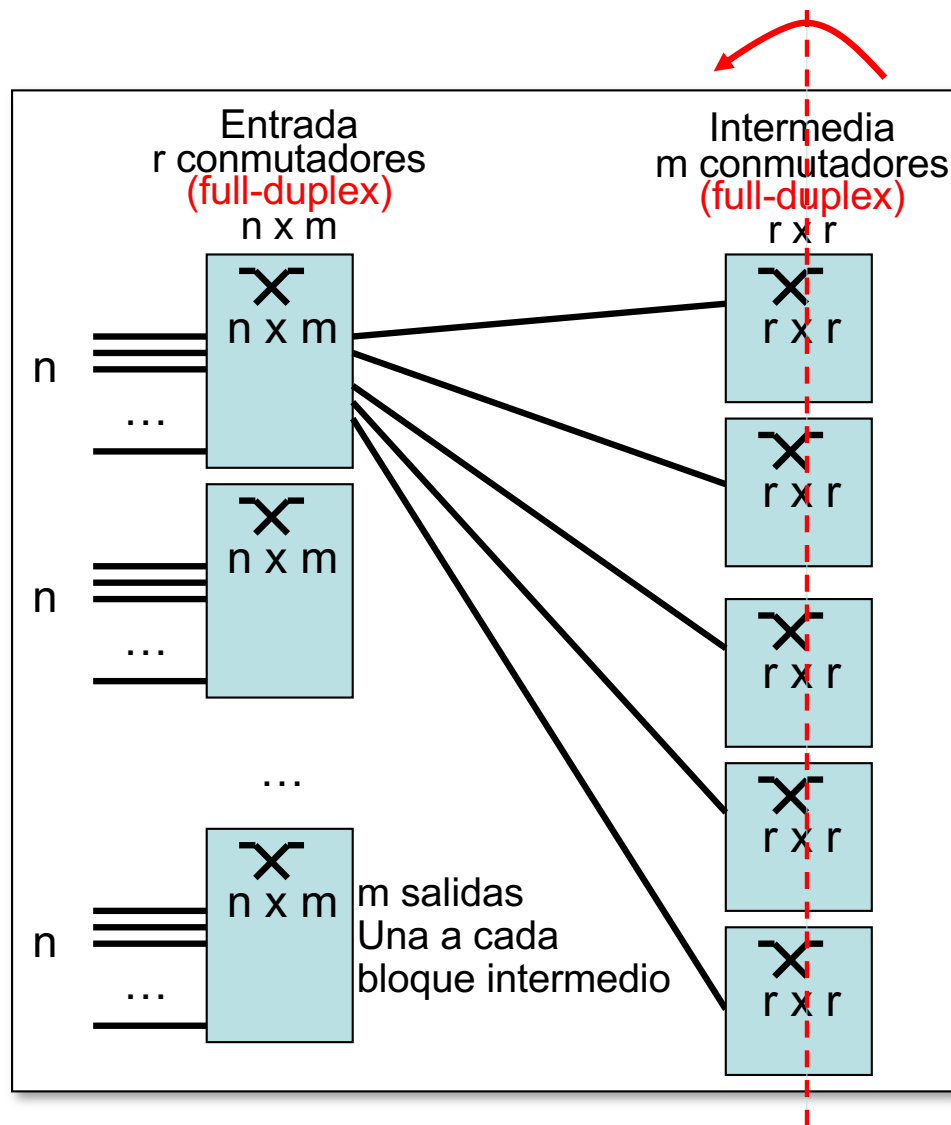
Folded Clos / Fat Tree

- En realidad este diseño es para enlaces unidireccionales
- Con enlaces bidireccionales se “dobla” esta topología (...)



Folded Clos / Fat Tree

- Aquí hemos hablado de la arquitectura interna del conmutador pero volveremos a esta topología al hablar de la interconexión de ellos





No bloqueante

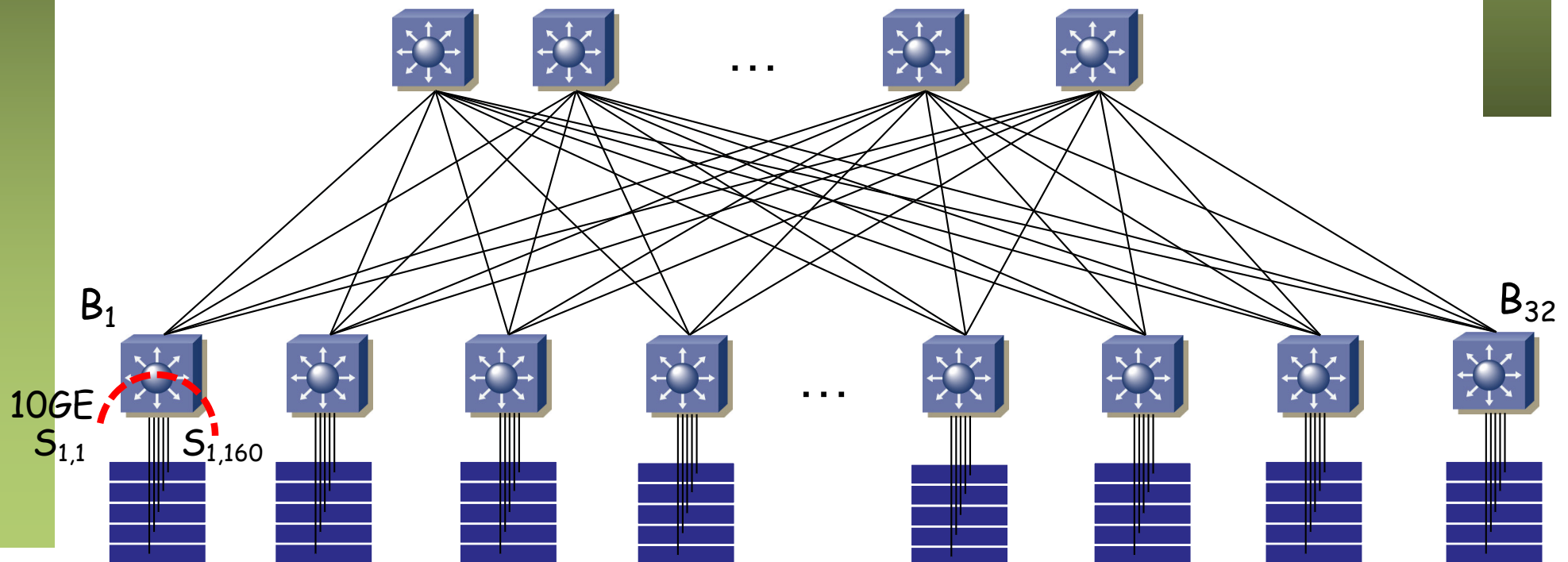


No bloqueante

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

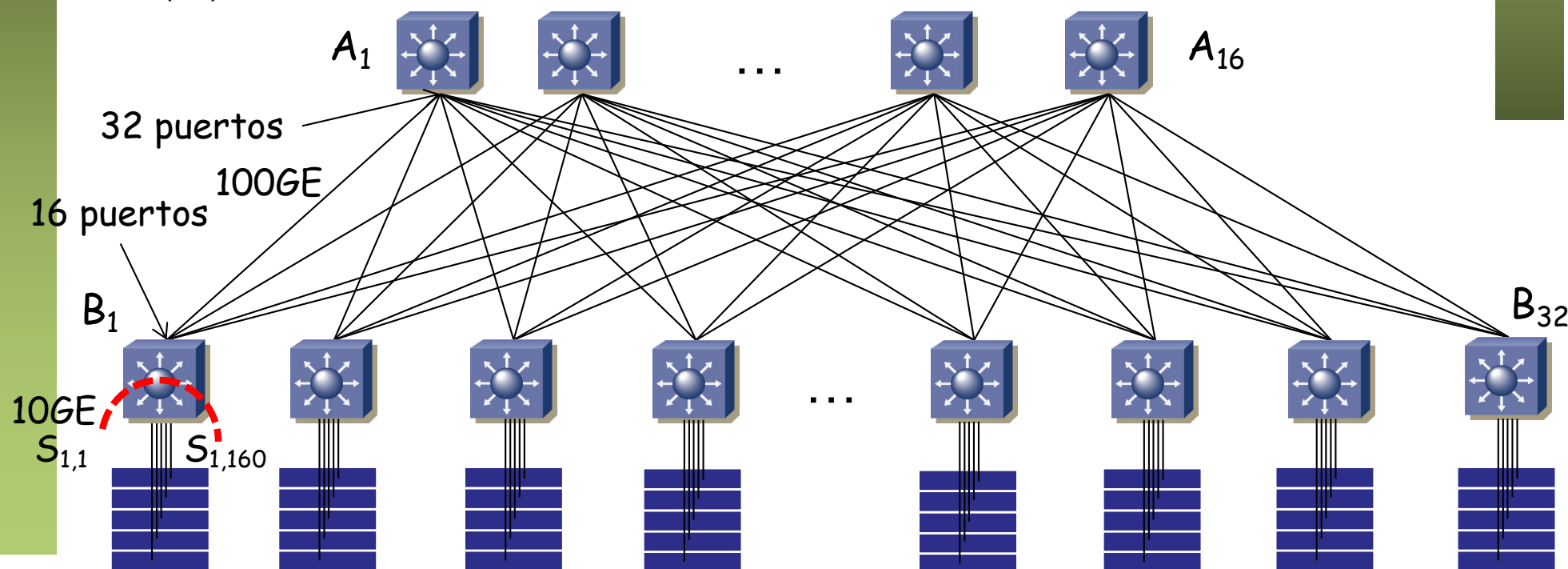
Ejemplo

- 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
- (...)



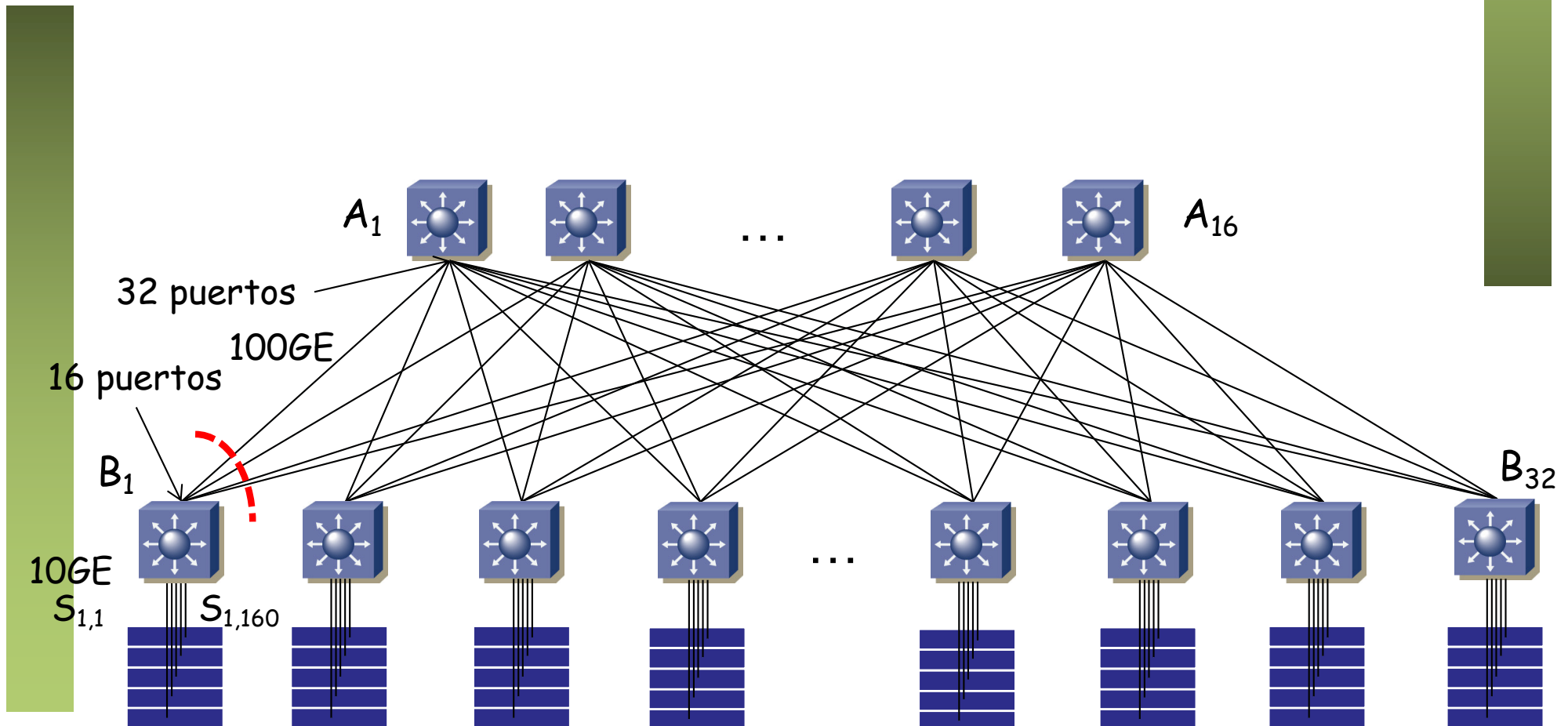
Ejemplo

- 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
- (...)



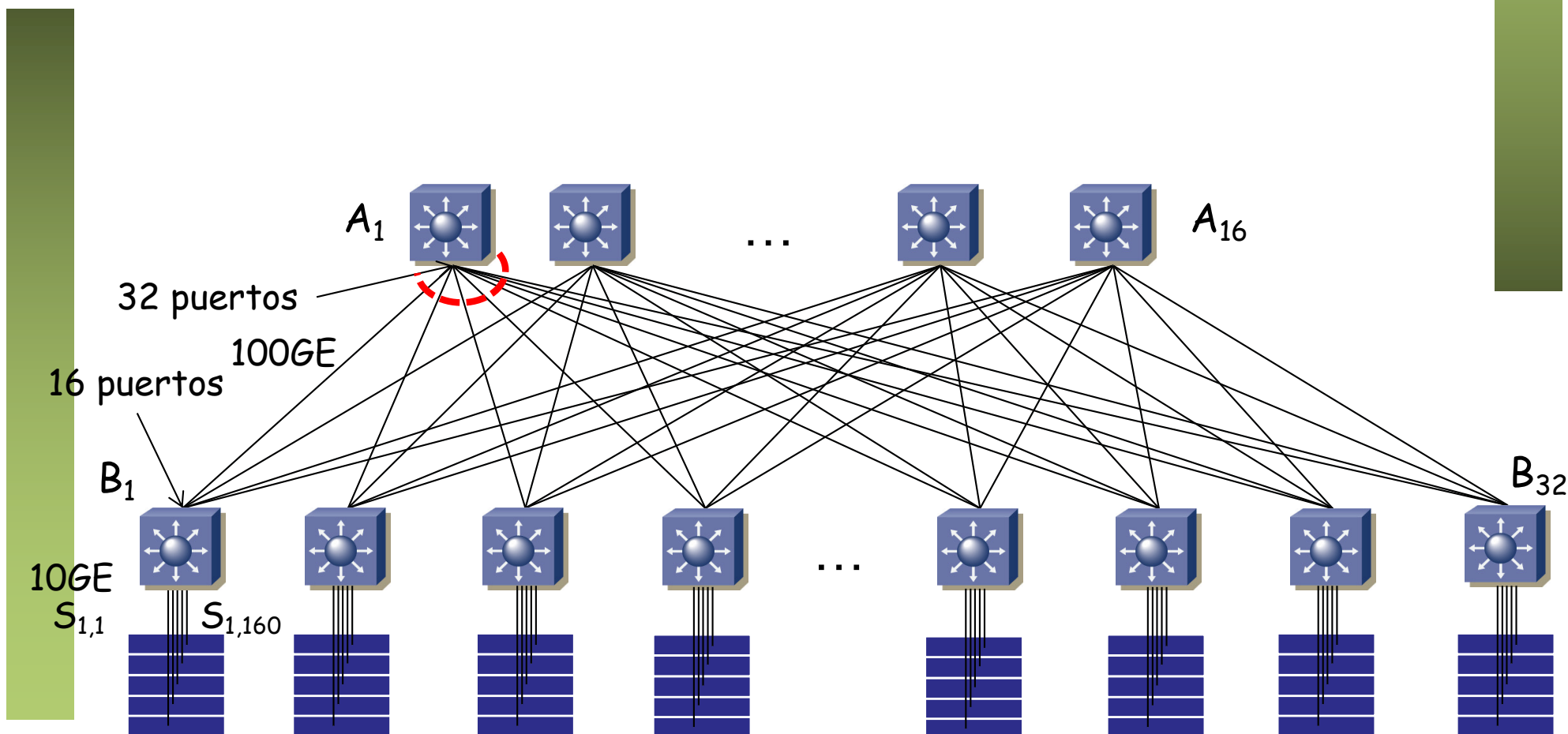
Ejemplo

- De cada conmutador de acceso salen $16 \times 100\text{GE} = 1.6\text{ Tbs}$
- (...)



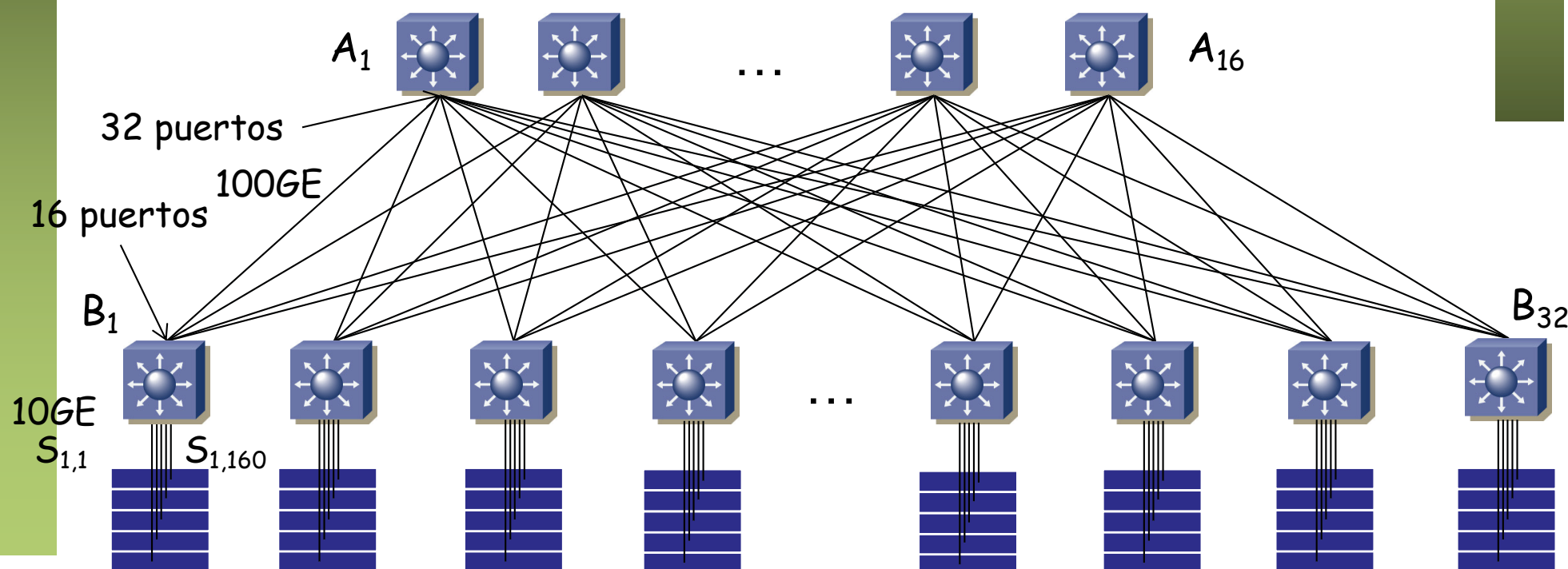
Ejemplo

- 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)
- (...)



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



Arquitectura tradicional en el data center: limitaciones