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

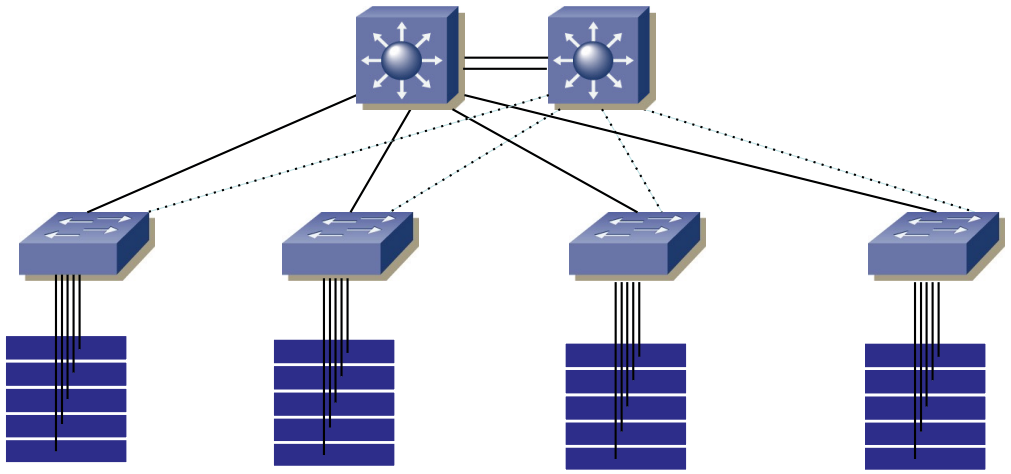
Nuevas arquitecturas para el data center

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

Arquitectura tradicional en el data center: limitaciones

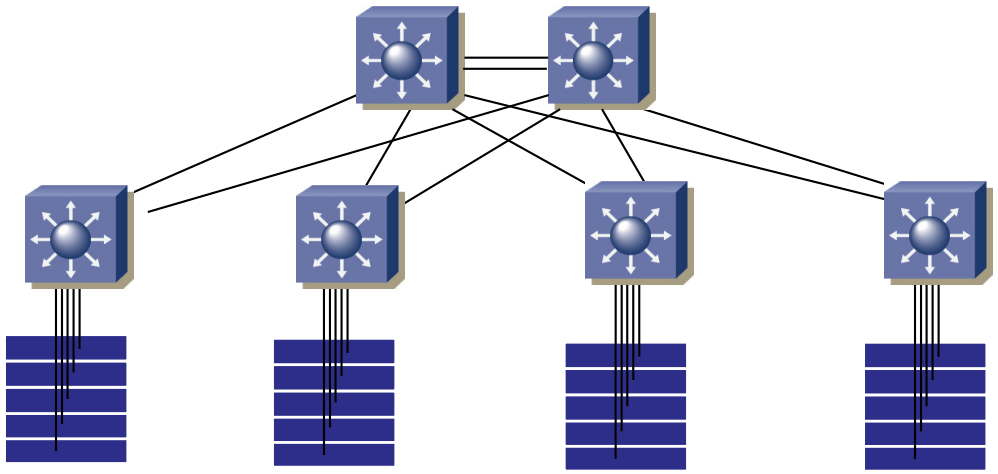
DC tradicional de 2 capas

- 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
- (…)



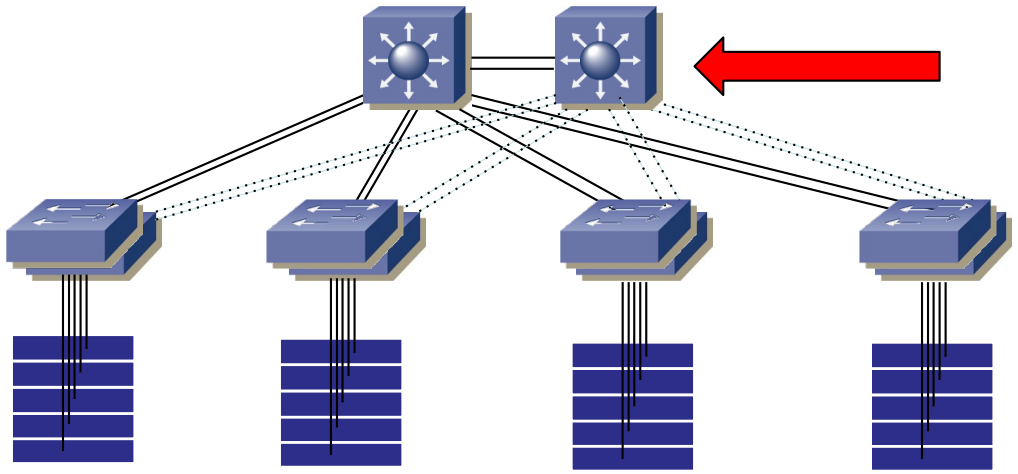
DC tradicional de 2 capas

- 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



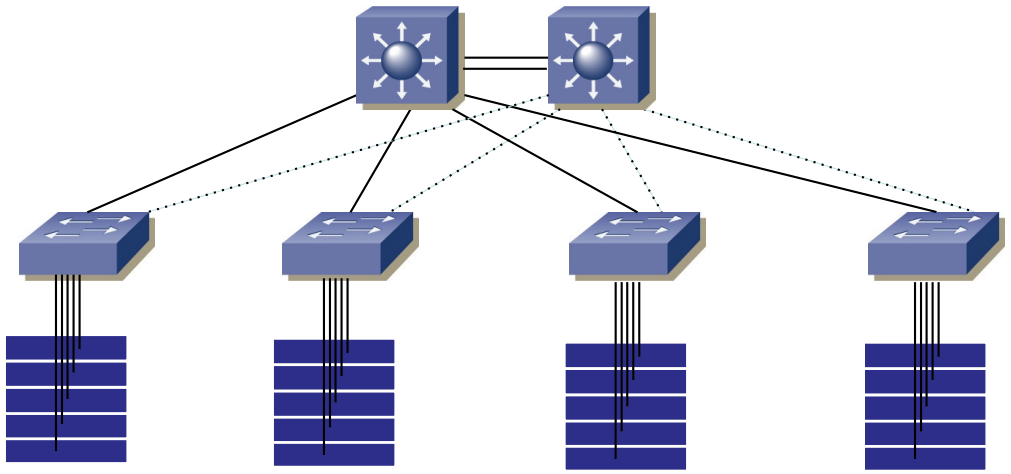
Límites con 2 capas

- Este esquema puede ser con uno o dos ToR por armario
  - Servidor conectado a 2 switches consume un puerto en cada uno (es como tener 2 servidores)
  - Switches agregados (MLAG, vPC, etc)
  - Protección en nivel de aplicación y servidor conectado a un solo switch
- El límite de escala (puertos de acceso) está en el número de puertos en los conmutadores de agregación
- También limita el que todo es un dominio de broadcast
  - Los switches deben aprender todas las direcciones MAC
  - Inundación por todo el data center
  - Debilidad del STP



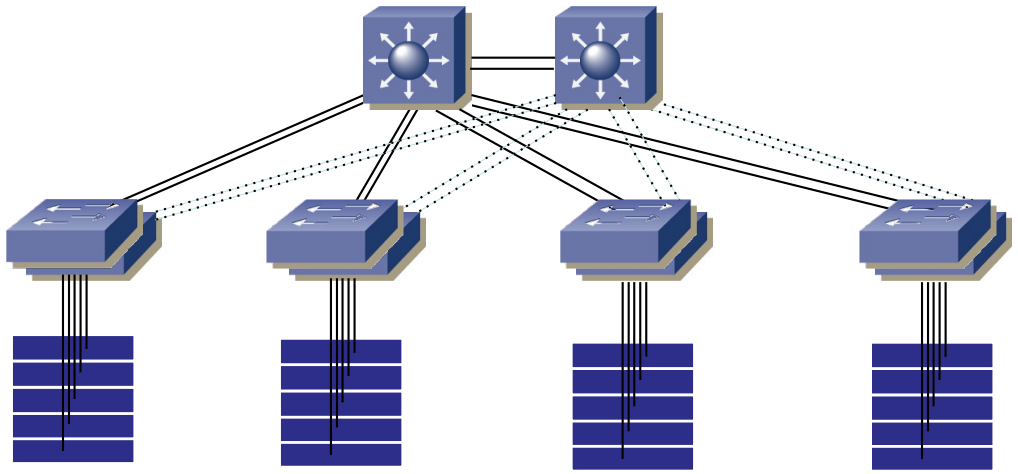
Límites con 2 capas

- Ejemplo:
  - Conmutador de acceso de 48 puertos (48 servs/rack) + 2 uplinks
  - Conmutador de agregación con 64 puertos 10Gbps
  - Máximo de 48x64 = 3.072 servidores
  - (…)



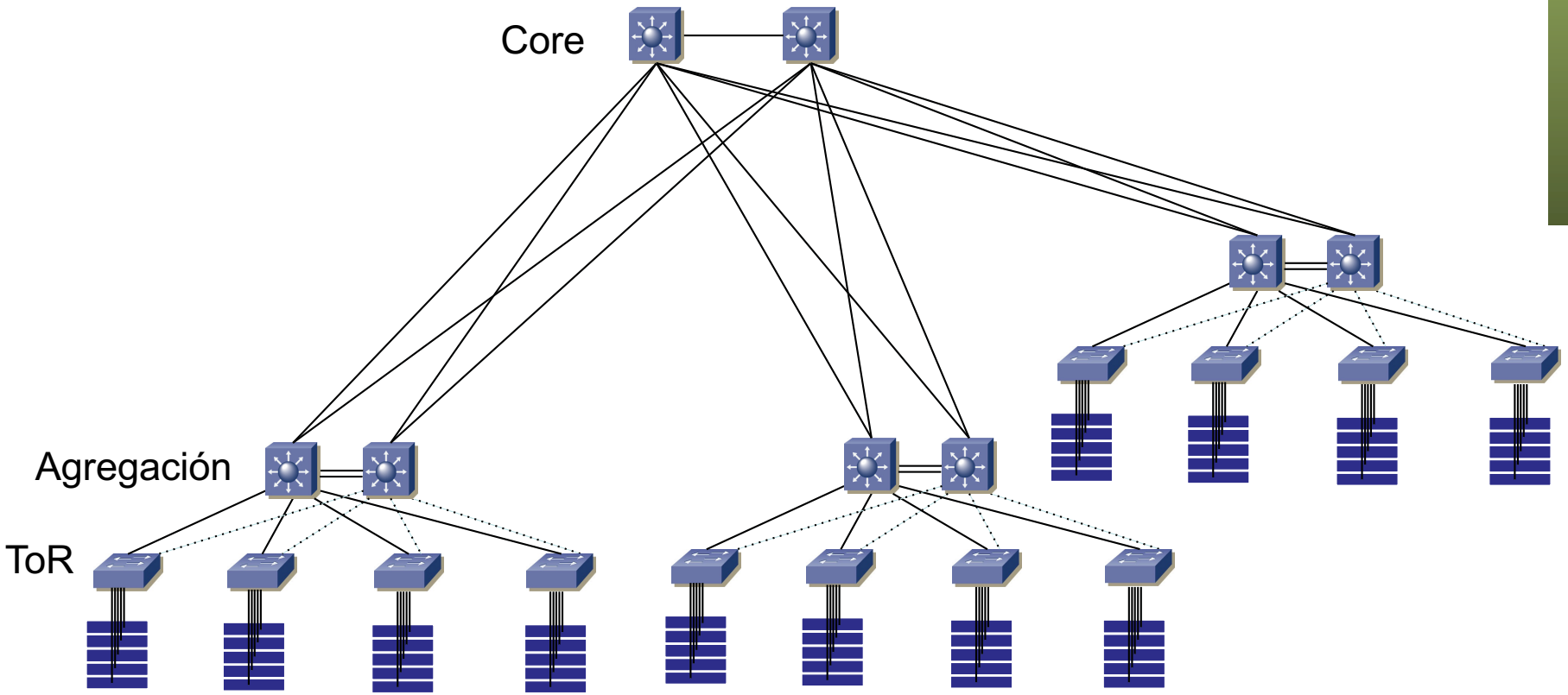
Límites con 2 capas

- 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 48x64/2 = 1.536 servidores
- Número de puertos en switch de agregación x número de puertos en switch de acceso
- ¿Más?



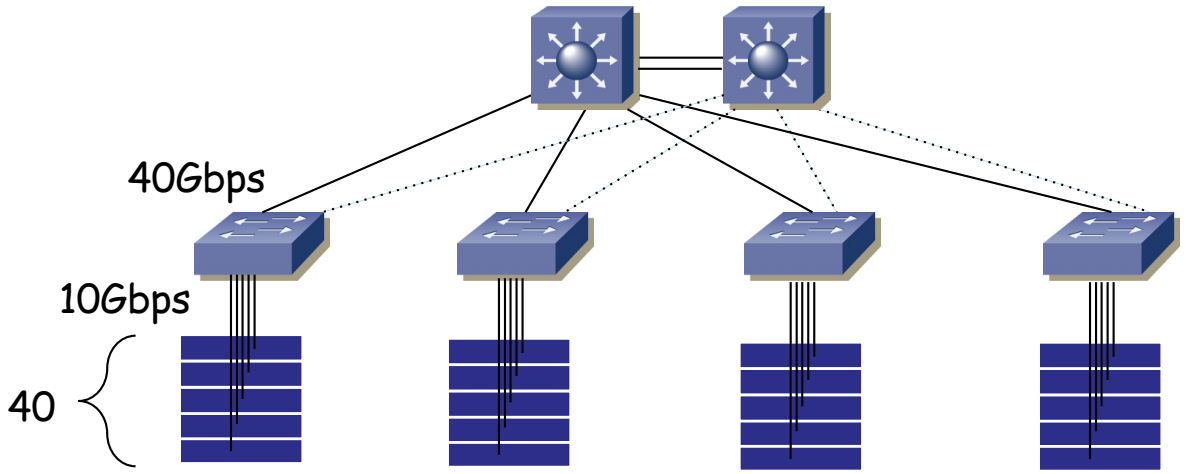
3 capas

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



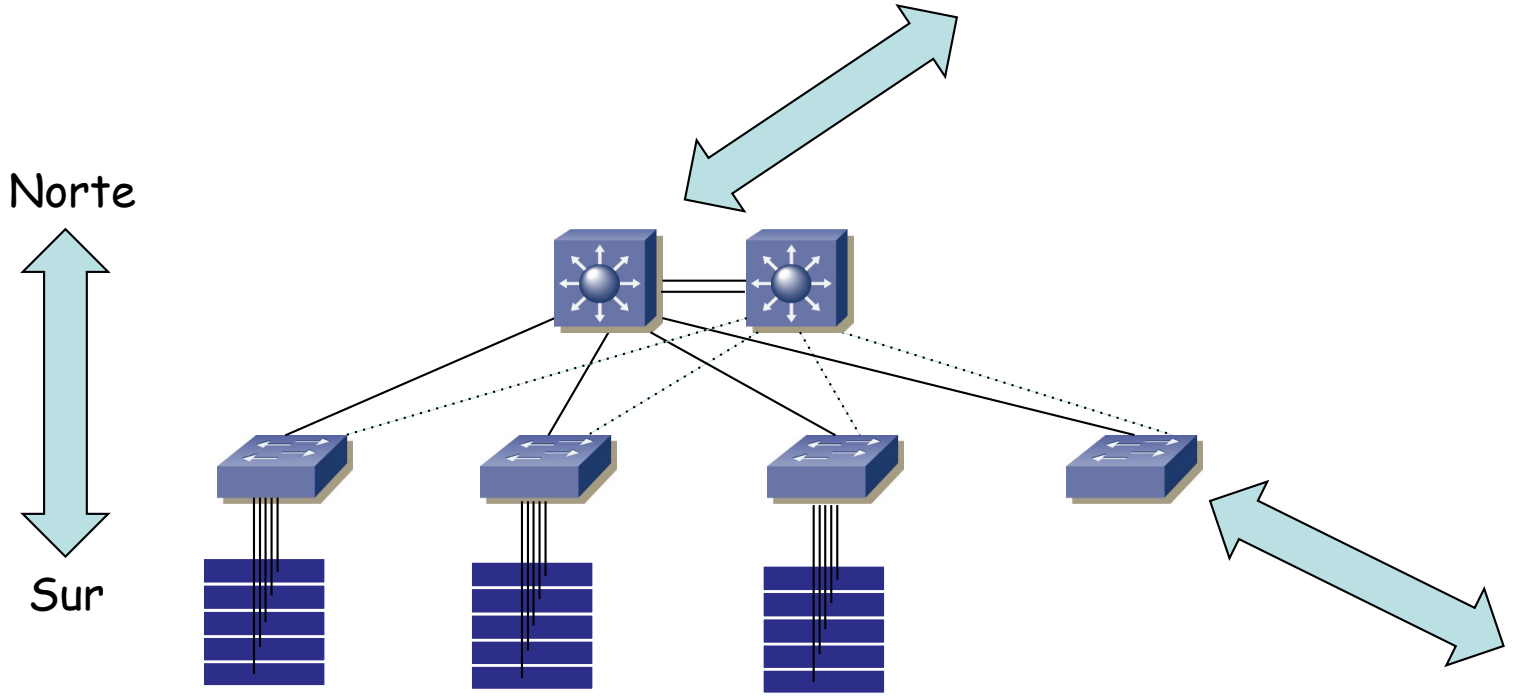
Over-subscription

- Conectamos hosts de tal forma que su tráfico agregado excede al 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 40x10 = 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 2x40 Gbps sería un 5:1
  - Un 5:1 para servidores con enlaces 10GE quiere decir en un reparto equitativo 2Gbps por servidor



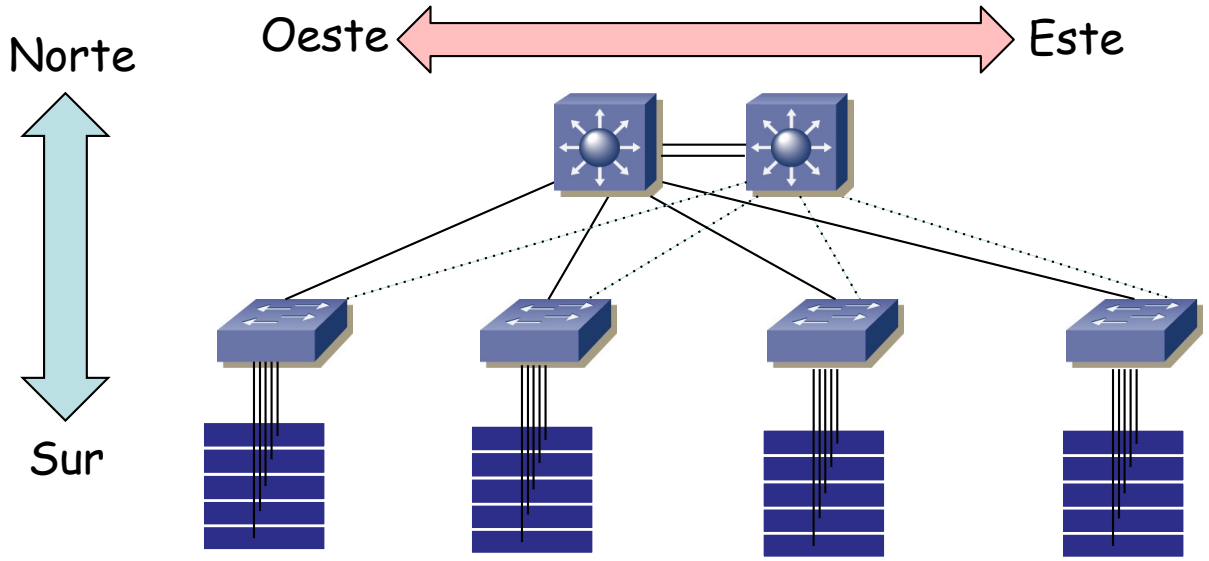
Conectividad externa

- 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
  - A través de los switches de distribución
  - O a través de switches de acceso de borde dedicados



Over-subscription

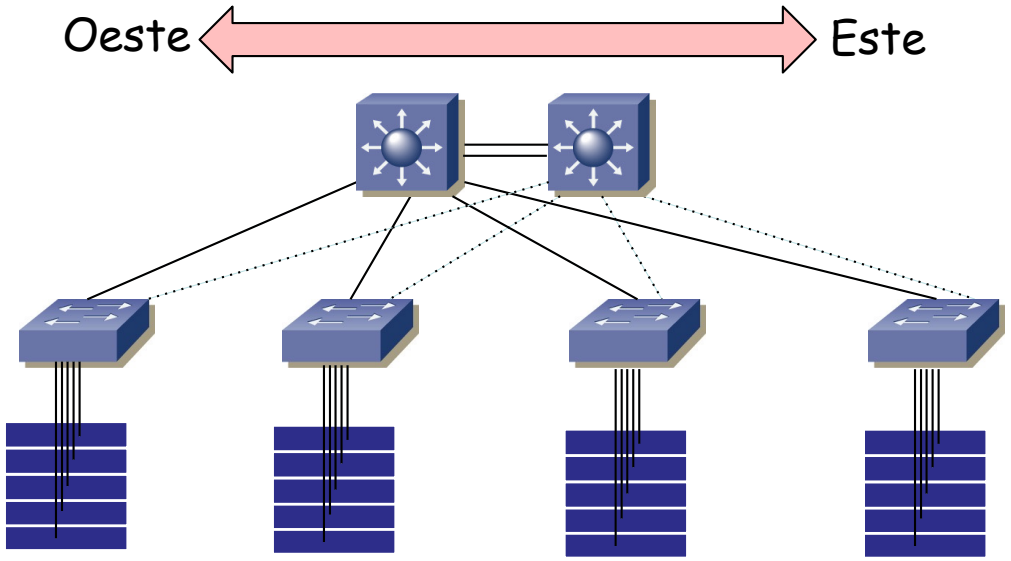
- Altos ratios de sobre-subscripción son razonables cuando tenemos mucho tráfico norte-sur
- Por ejemplo, hacia una salida a Internet que sea en realidad el cuello de botella
- No son tan razonables cuando hay mucho tráfico este-oeste
  - Tráfico entre los servidores en distinto rack
  - Aplicaciones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



Tráfico Este-Oeste

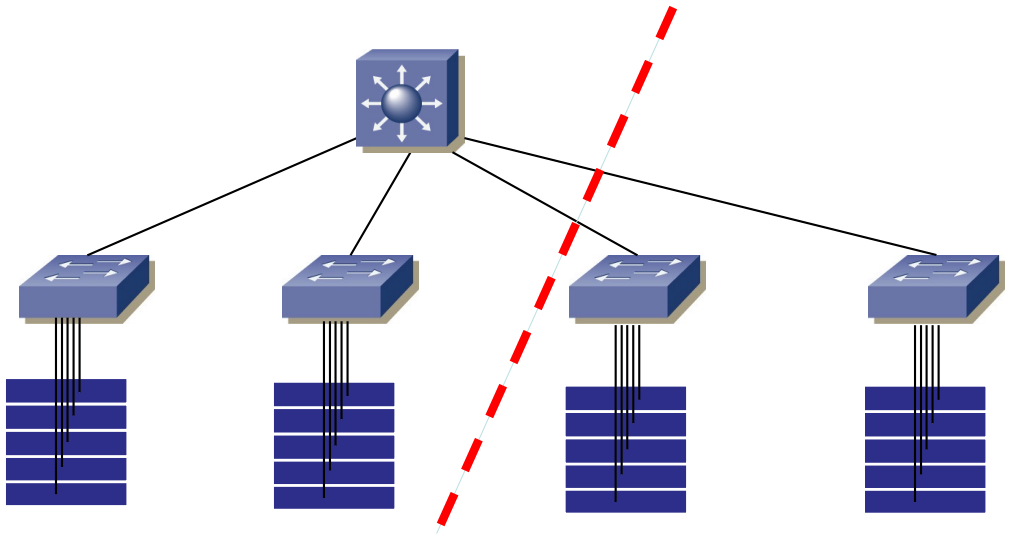
- Ejemplo: Facebook (2013)
- Una petición HTTP se transforma en:
  - 88 búsquedas en caches
  - 35 búsquedas en bases de datos
  - 392 llamadas a procedimientos remotos en el backend
- Una petición con un tamaño de 1KB ha resultado en cerca de 1MB de transferencias en el data center
- ¿Datos almacenados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

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



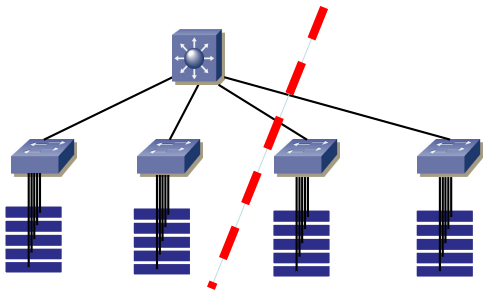
Bisectional bandwidth

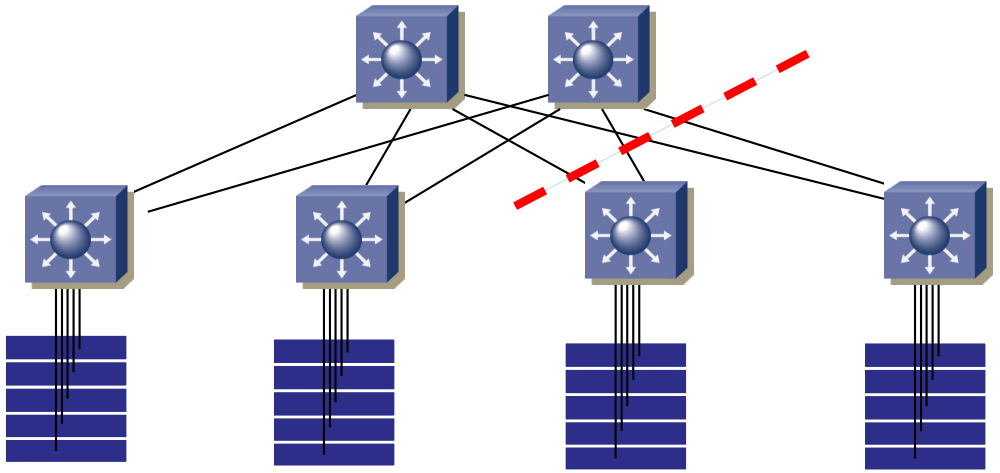
- Una **bisección** es una partición de la red en dos subconjuntos con igual número de hosts
- El **ancho de banda de esa bisección** es la suma de las capacidades de los enlaces entre los dos subconjuntos
- En este ejemplo 2x capacidad del enlace de agregación a acceso
- El **ancho de banda de bisección 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 bisectional bw

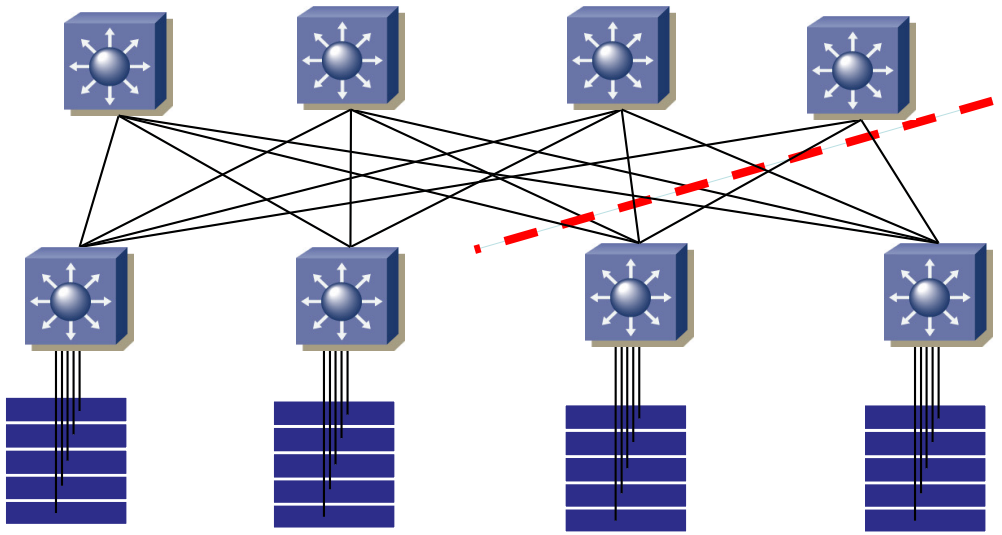
- Aumentamos el ancho de banda de bisección aumentando el número de caminos
- En este caso por ejemplo lo hemos duplicado
- Podríamos seguir aumentándolo (…)





ECMP y bisectional 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
- Pero al menos evitamos STP y el dominio de broadcast extendido

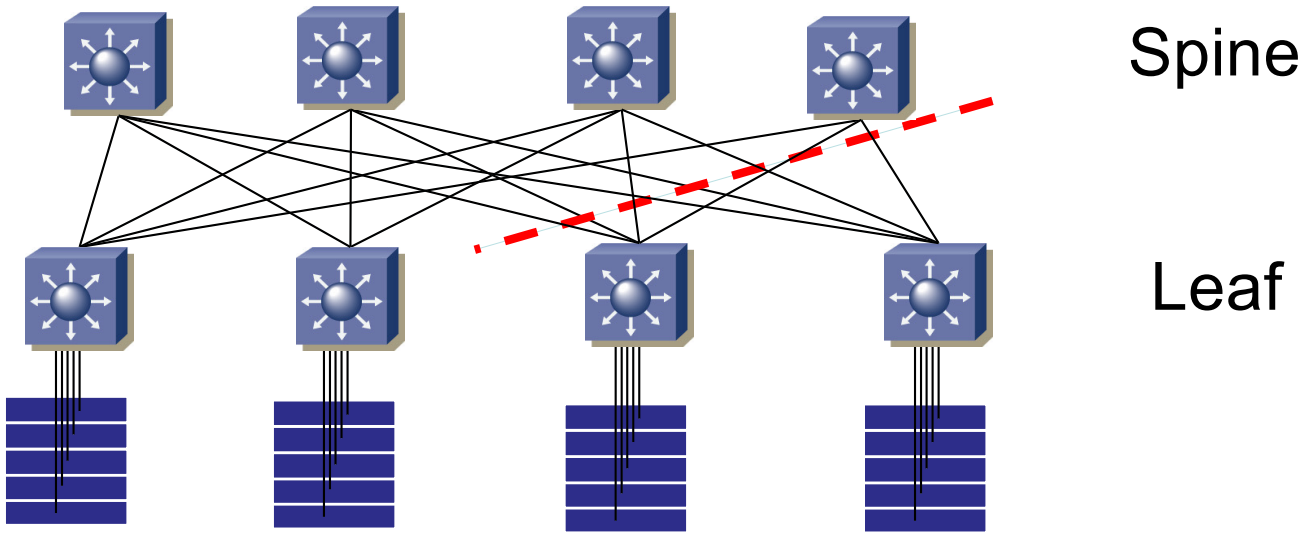


Leaf + Spine

Evolución

- Servidores 1GE, uplinks 10GE
- Servidores 10GE, uplinks 40GE
- 2017: Servidores 25GE, uplinks 100GE
- 2019: Servidores 50GE, uplinks 100GE
- 2024: Servidores hasta 100GE, uplinks 400GE
- Igual que en MLAG: nº puertos Spine x nº puertos Leaf
- Aumentando nº spines aumenta la capacidad

MSDC = Massively Scalable Data Center



Ejemplo: Server pod

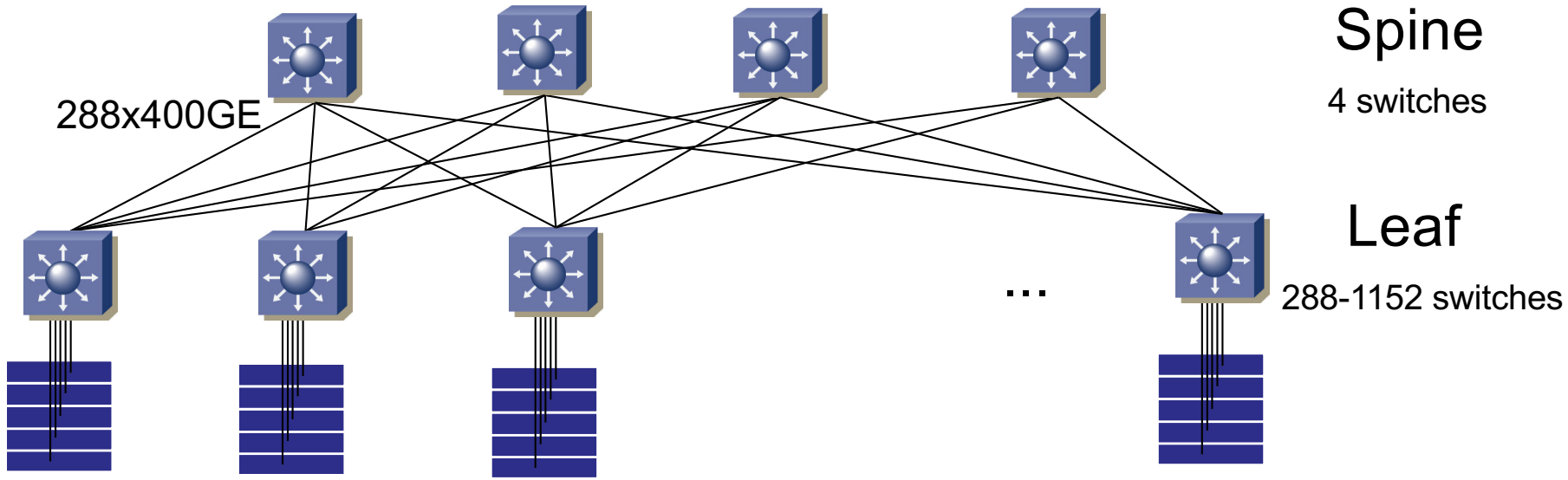
- Spine: Cisco Nexus 9808
  - 8 slots
  - Line card N9K-X9836DM-A
    - 36 @ 400G ó 4x36 @ 100G (breakout cable)

- Leaf: Cisco Nexus 93400LD-H1
  - 48 @ 10/25/50G
  - 4 @ 400G o 16 @ 100G

https://www.cisco.com/c/en/us/products/collateral/switches/nexus-9000-series-switches/nb-06-nexus-93400ld-h1-switch-ds-cte-en.html
https://www.cisco.com/c/en/us/products/collateral/switches/nexus-9000-series-switches/nexus9800-series-switches-ds.html

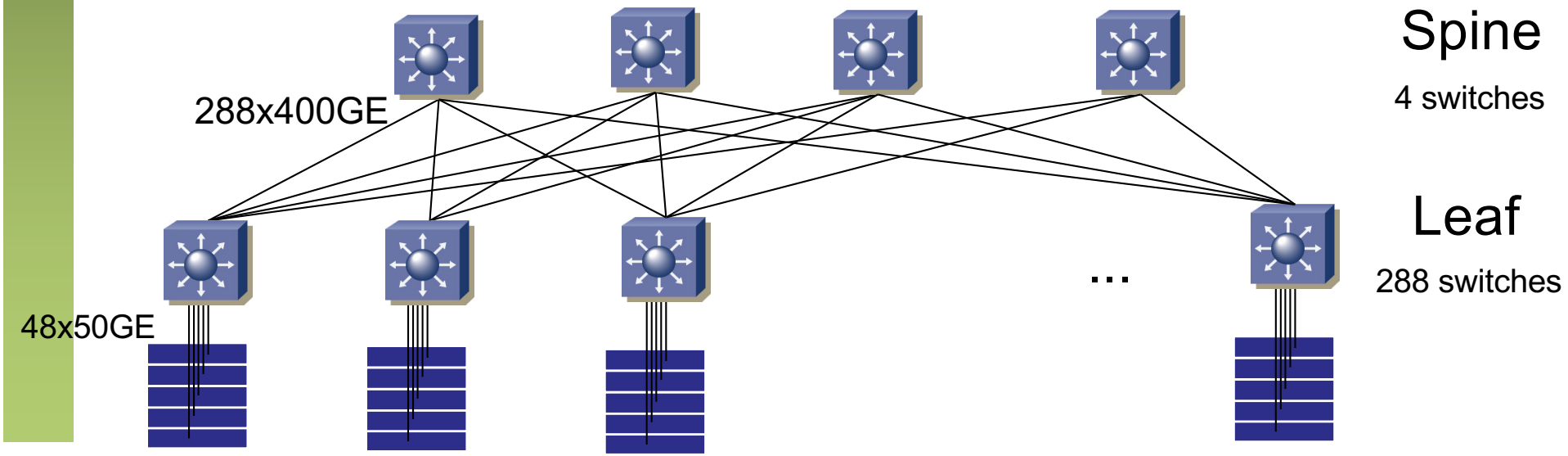
Server pod: Ejemplo 1

- 4 spine switches:
  - Cada uno 8 x line cards N9K-X9836DM-A
  - 8x36=288 puertos/switch @ 400GE o 8x4x36=1152 @ 100GE
  - 288 puertos/switch x 400 Gb/s = 115.2 Tb/s/switch → 288 leaf switches
  - 1152 puertos/switch → 1152 leaf switches
- (...)



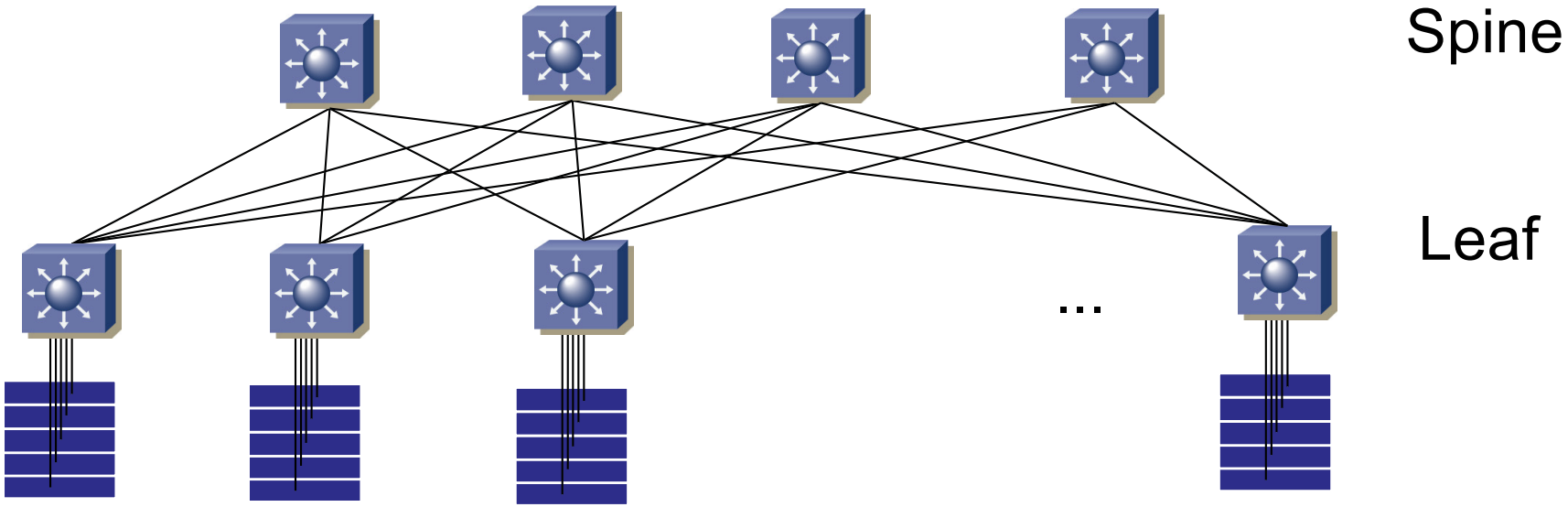
Server pod: Ejemplo 1

- 4 spine switches:
  - Cada uno 8 x line cards N9K-X9836DM-A
  - 8x36=288 puertos/switch @ 400GE o 8x4x36=1152 @ 100GE
  - 288 puertos/switch x 400 Gb/s = 115.2 Tb/s/switch → 288 leaf switches
  - 1152 puertos/switch → 1152 leaf switches
- Usando 4 spines y 288 leaf switches:
  - Cada leaf 4 puertos @ 400 Gb/s links (o 16 @ 100 Gb/s) uplink
  - En total 288x48=13.824 puertos @ 10/25/50 Gb/s
  - Cada Leaf 48x50G=2.4Tb/s
  - Con 4 uplinks 400GE: 4x400G=1.6Tb/s → over-subscription 1.5:1 (2.4:1.6)



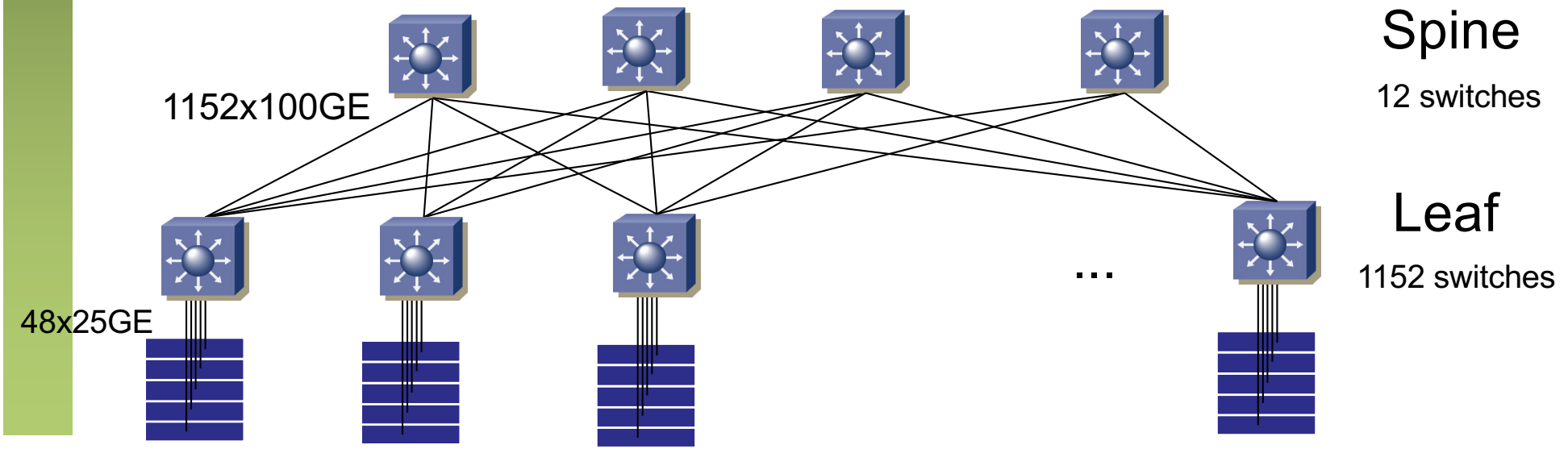
Sin sobre-subscripción

- Leaf a Spine: 1.6Tb/s con hasta 4 puertos @ 400G o hasta 16 puertos @ 100G
- En un caso son 4 Spines, en el otro 16 Spines, todos switches iguales
- Misma capacidad de interconexión, diferente disponibilidad ante fallos
- Diferente cableado de leaf a spine, diferente nivel de ECMP
- Diferente máxima capacidad por flujo (no si 400G=4 lanes x 100G)
- Si 1.6Tb/s / 50G = 32 puertos/switch, entonces sobre-subscripción (1:1)
- Si 25G x 48 = 1200G, vale con 12 puertos @ 100G o 3 puertos @ 400G
- Eso son 12 ó 3 Spines



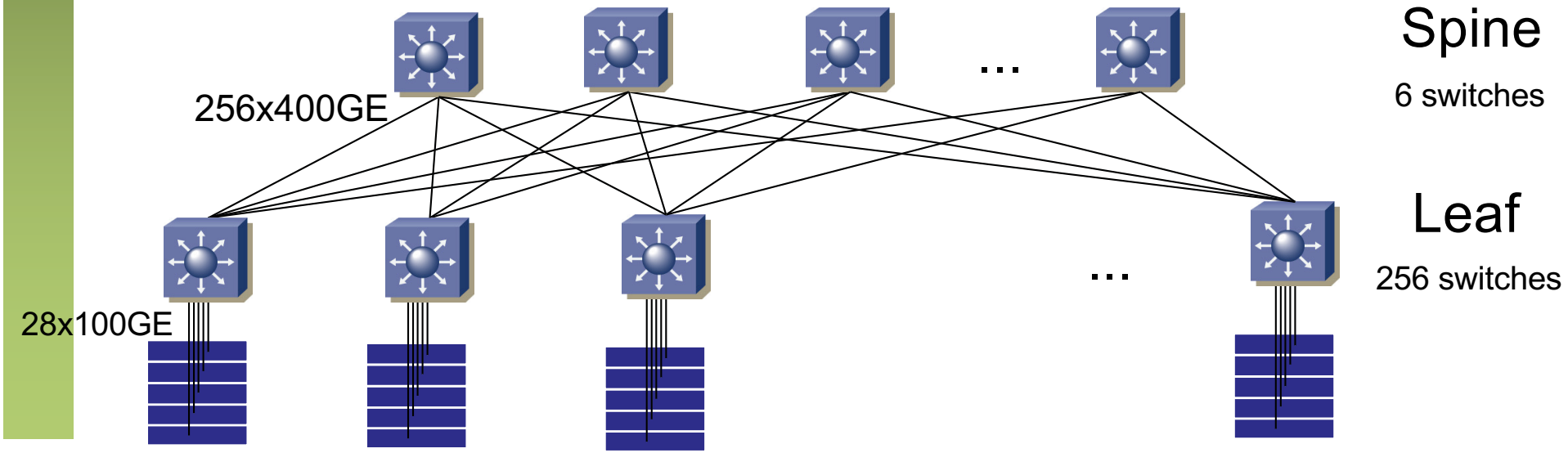
Server pod: Ejemplo 2

- 12 spine switches:
  - 1152 puertos/switch @ 100G → 1152 leaf switches
  - 1152 x 12 = 13.824 cables entre armarios
- 1152 leaf switches:
  - Cada uno 12 puertos @ 100 Gb/s uplink
  - En total 1152x48=55.296 puertos @ 10/25/50 Gb/s
  - Si cada Leaf 48x25G=1.2Tb/s, sin sobre-subscripción
  - Si cada Leaf 48x50G=2.4Tb/s, sobre-subscripción 2:1
- Miles de armarios en los grandes centros de datos



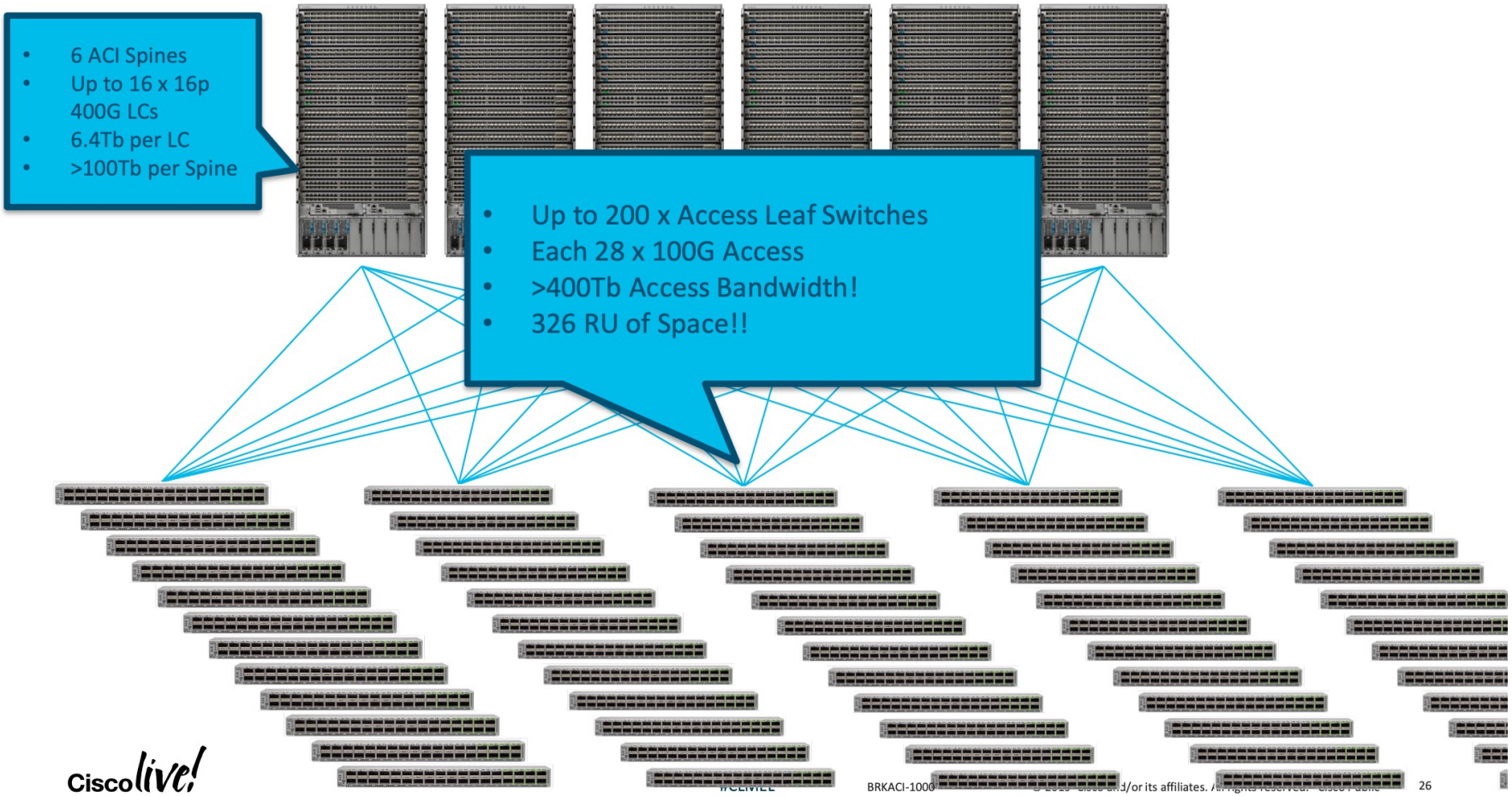
Server pod: Ejemplo 3

- Cambiando el hardware
- 6 spine switches (modular):
  - 16 slots, line cards de 16 puertos @ 400 Gb/s, 6.4 Tb/s/card
  - 256 puertos/switch x 400 Gb/s = 102.4 Tb/s/switch
- 256 leaf switches (1U ToR):
  - 28 puertos x 100 Gb/s (divisibles) = 2.8 Tb/s hacia spines
  - 6 x 400 Gb/s links = 2.4 Tb/s (sobresubscripción 1.17:1)
  - 256 x 28 = 7.168 puertos @ 100G



Ejemplo

- Muy parecido a esto:



Ejemplo

- Muy parecido a esto:

Cisco Nexus 9516:
16 line card slots

N9K-X9716D-GX line card:
16-port 400GE QSFP-DD

Cisco Nexus 93600CD:
28 x 100/40-Gb/s (QSFP28) + 8 x 400/100-Gb/s (QSFP-DD)
System buffer: 80MB
Switching capacity: 12 Tb/s
Forwarding capacity: 4.000 Mpps

Spine 800GE

- Switches modulares con puertos 800G
- Ejemplo:
  - Arista 7816LR
  - 16 slots x 36 puertos/slot @ 800G = 576 puertos

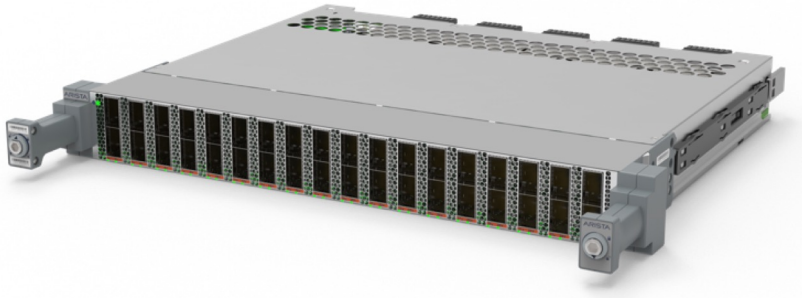
Arista 7800R Series Modular Data Center Switches

System Scale and Performance

- 460 Tbps (920 Tbps FDX) fabric capacity
- Up to 173 Billion packets per second
- Up to 28.8 Tbps per slot
- Up to 576 wire-speed 800G ports
- 100G, 200G, 400G and 800G modes
- Deep packet buffer up to 32 GB /line card

Optimized for AI/ML and HPC Clusters

- Fully scheduled lossless interconnect
- 100% efficient cell based load balancing
- Distributed hardware scheduler
- Per port Virtual Output Queuing to eliminate head of line blocking
- Hardware accelerated link health management
- Under 3 microsecond latency (64 bytes)
- Ultra Ethernet ready

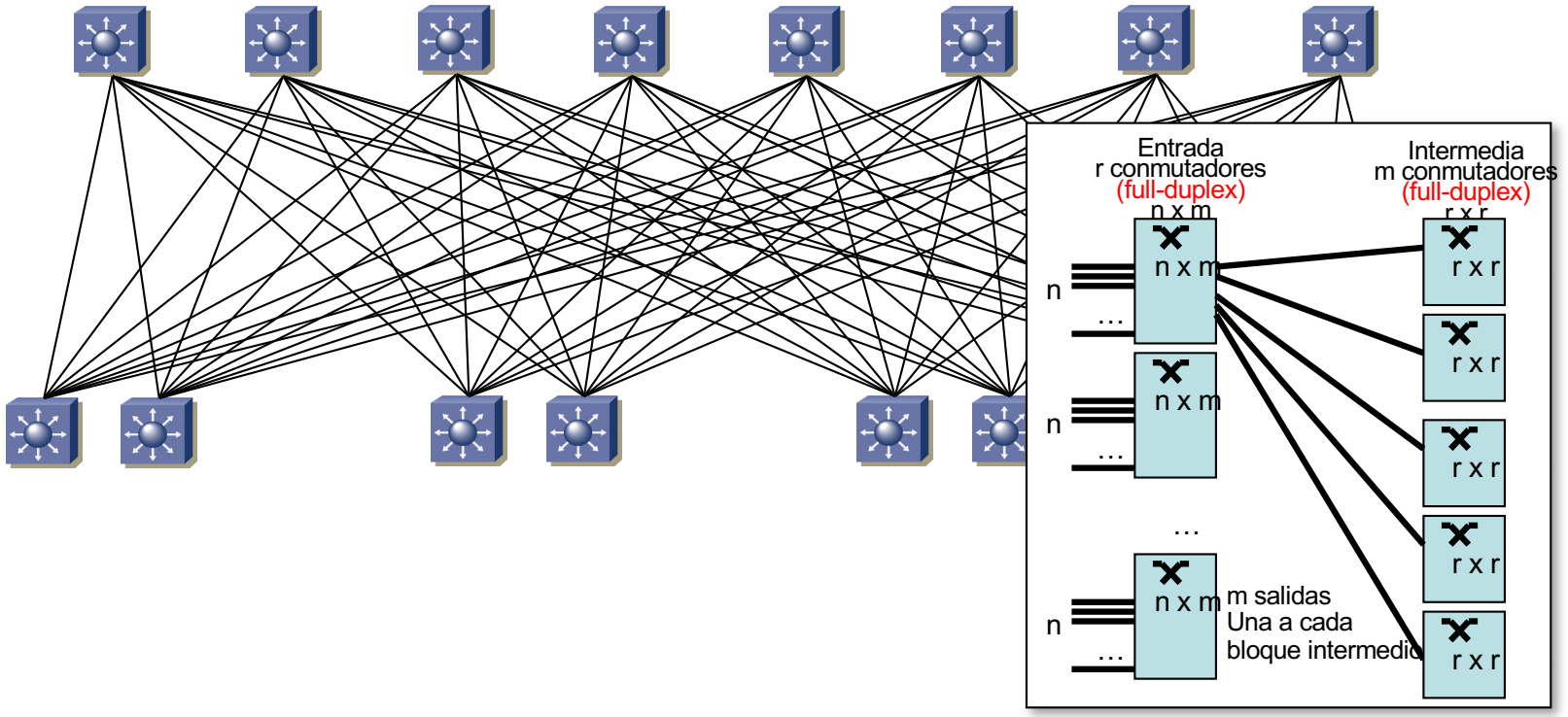


7800R4 36PE series: 36 port 800G OSFP line card

- Up to 36 800G ports per line card or 72 400G ports
- 28.8 Tbps of forwarding and 10.8 Bpps with 32GB of buffer
- Available in three specifications

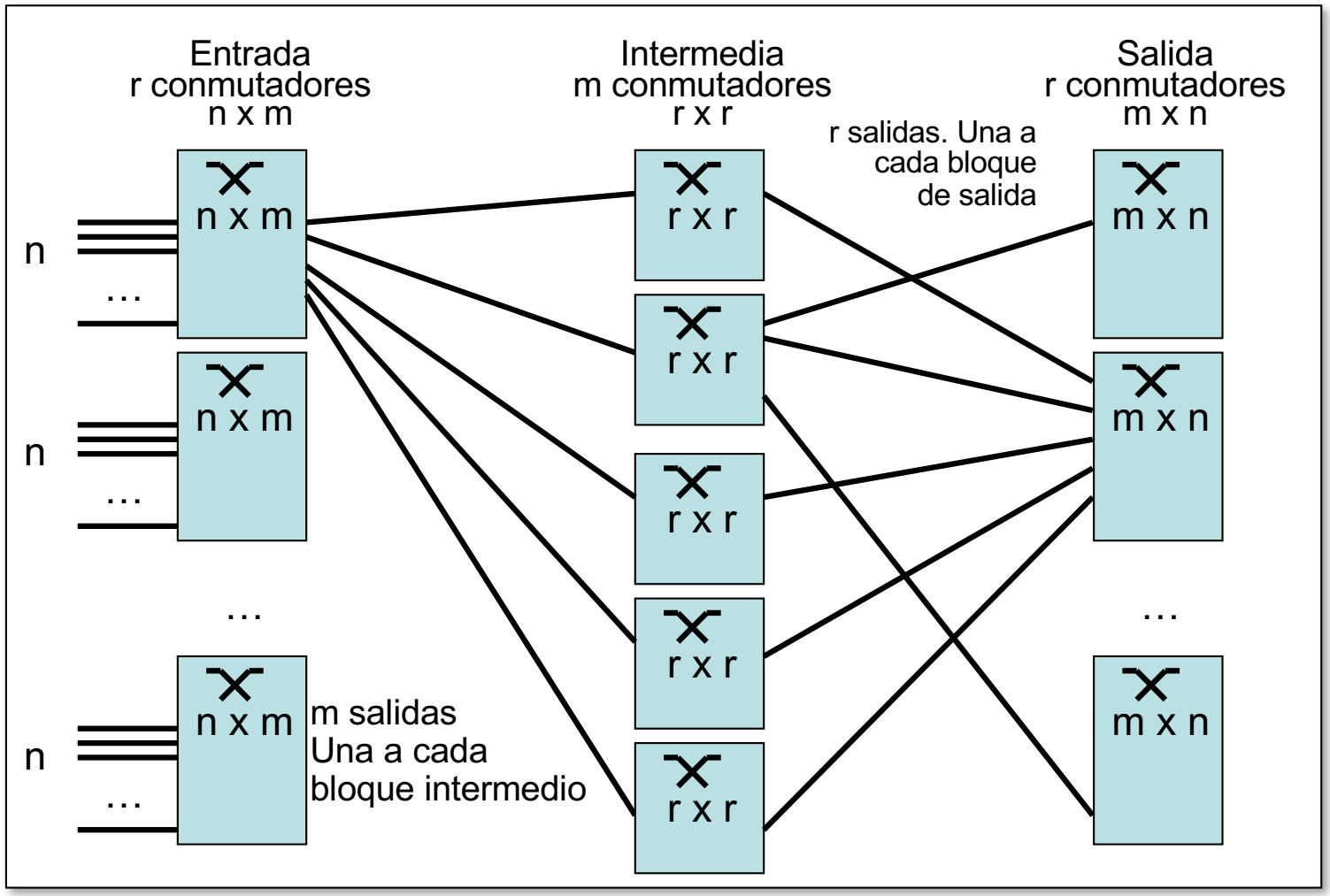
Redes de…

- Cada leaf switch conectado a un spine switch
- ¿Clos?



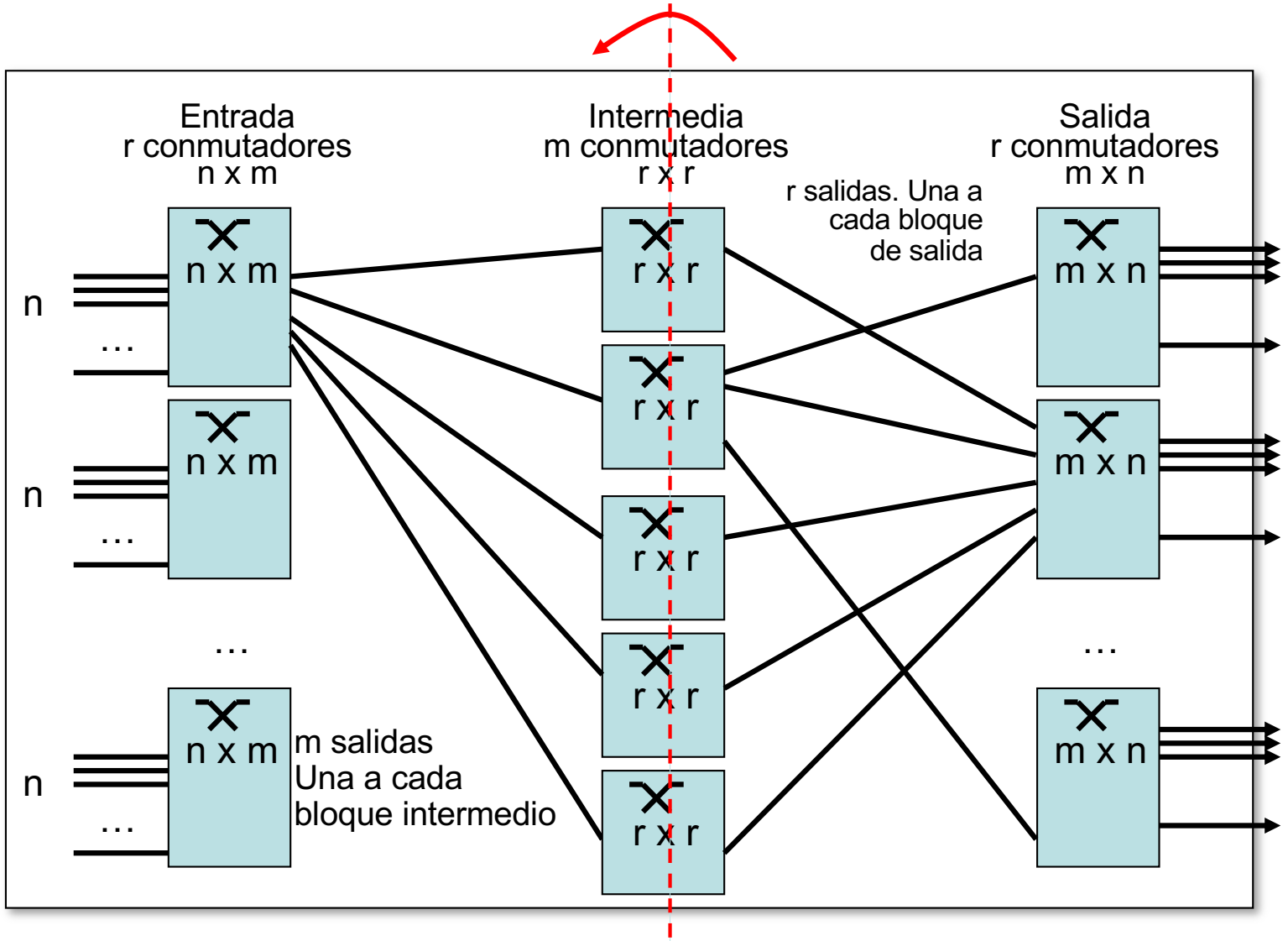
Redes de (Charles) Clos

- 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



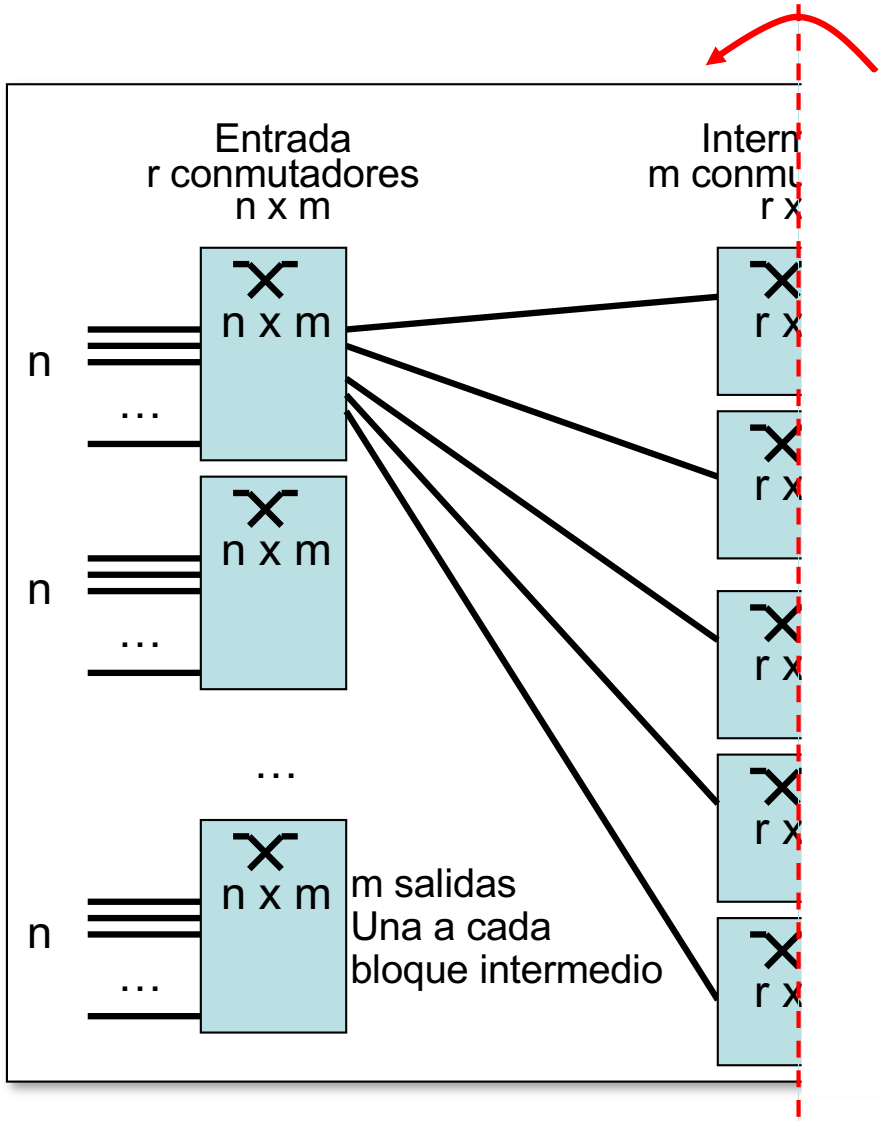
Folded Clos / Fat Tree

- 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 interconexión de conmutadores, pero en telefonía hablábamos de arquitectura de un gran conmutador en base a conmutadores pequeños (¿similar?)