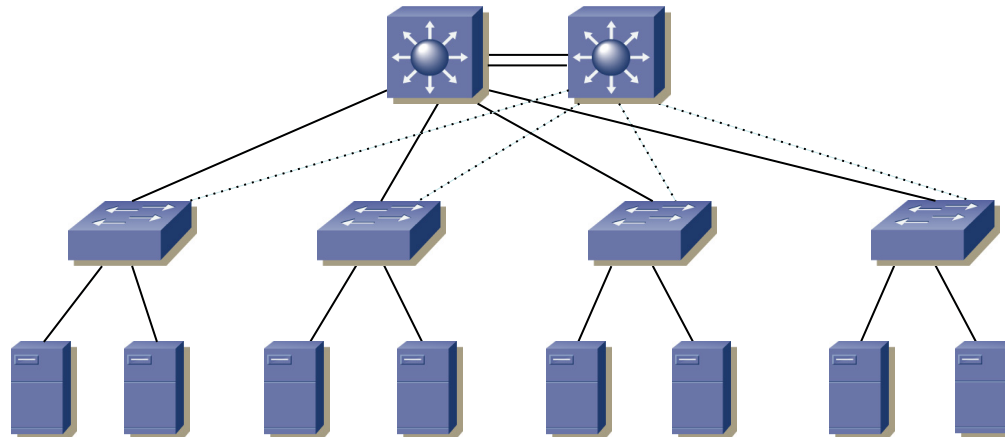


Overlays en el data center

Over-subscription

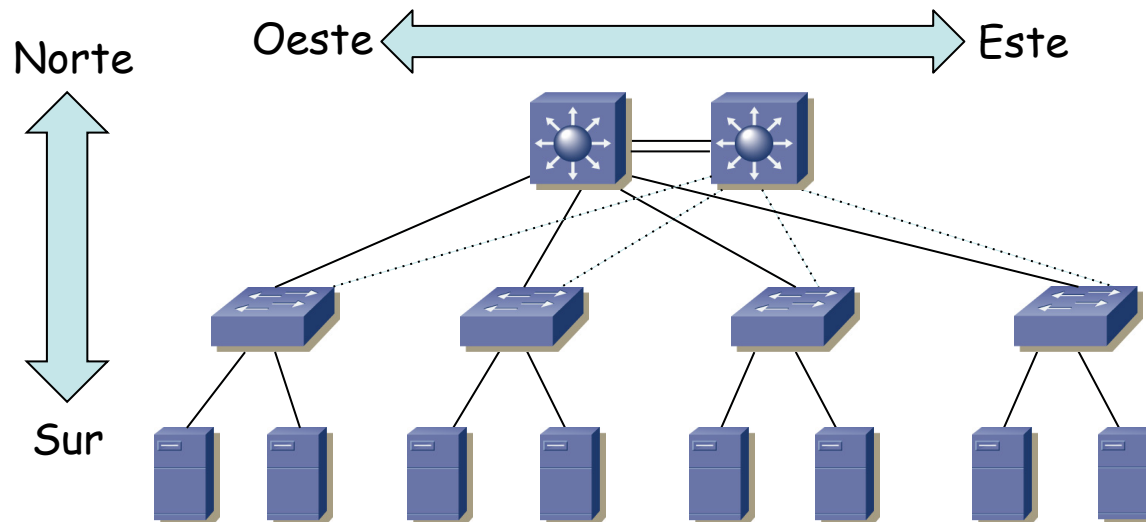
Over-subscription

- Conectamos fuentes 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
 - 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



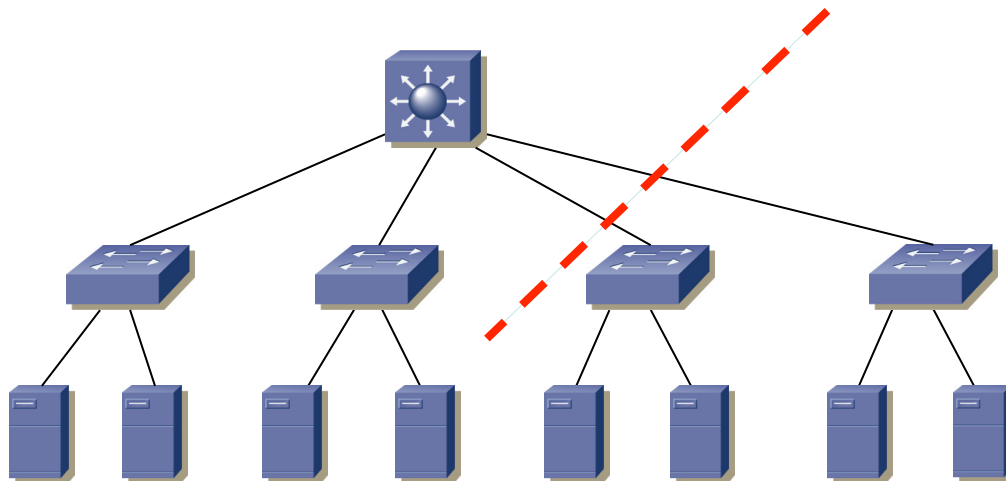
Over-subscription

- 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
- Aplicaciones distribuidas, tráfico de SAN, movimiento de máquinas virtuales, tráfico entre tiers de aplicación, clustering, etc



Bisectional bandwidth

- Una bisección es una partición de la red en dos subconjuntos con igual número de nodos
- 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



Limitaciones de STP

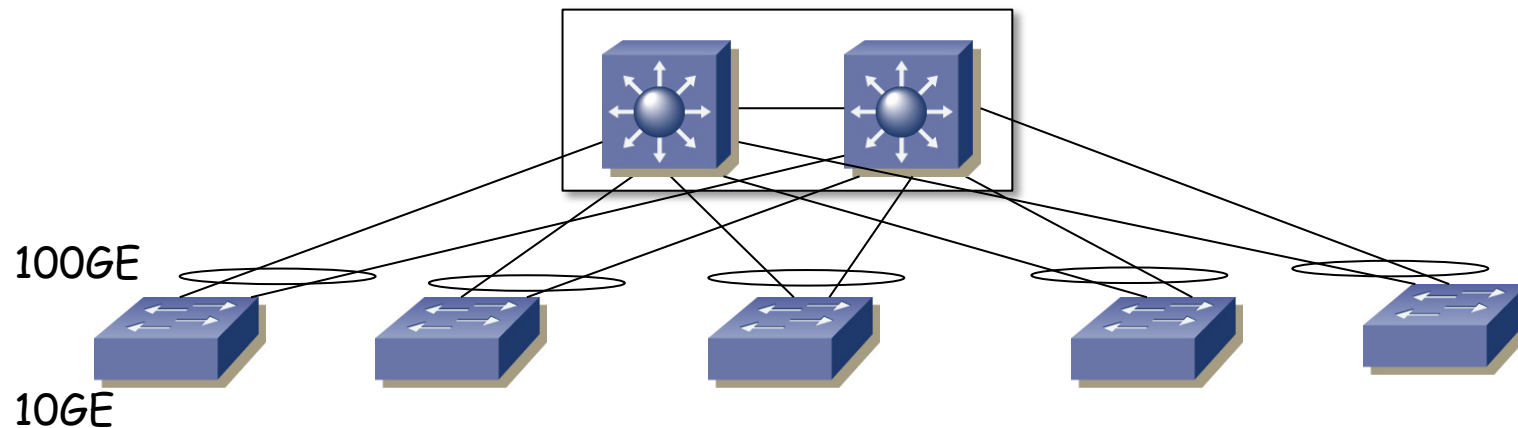
STP

- Limitaciones de STP
 - No soporta multipath para una VLAN
 - Multipath entre diferentes VLANs requiere intervención manual
 - El camino es el más corto solo desde la perspectiva del root
 - Largos tiempos de convergencia
 - Peligro de tormentas de inundación
 - Elección de la raíz no es segura
- Mejoras a STP
 - RSTP, MSTP mejoran los tiempos de convergencia pero siguen en el rango de los segundos
 - Otras mejoras a la convergencia, muchas veces sin estandarizar (Loopguard, BPDU guard, Rootguard, Storm control)
 - No cambian que desactiva puertos para formar un árbol
- Alternativas clásicas a STP
 - Multichassis LAG
 - Routing capa 3

Escalabilidad con MLAG





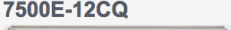

Escalabilidad con MLAG

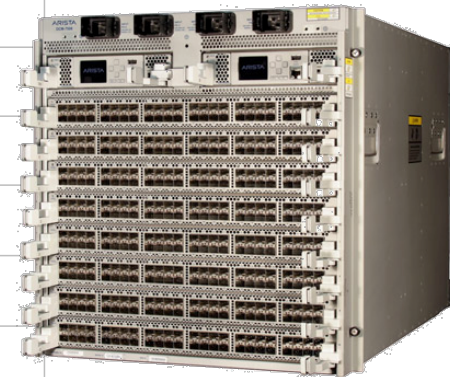
- Tenemos conmutadores con más de 1000 puertos 10GE (¡!)
- Hacia la segunda capa puertos 40GE o 100GE
- (...)



Ejemplo: Arista 7508

- Capacidad de conmutación de 30Tbps o 14.4Bpps
- (Ojo, “Billones” anglosajones = “miles de millones”)
- 8 slots
- Linecards:
 - La 7500E-36Q tiene 144 puertos 10GE
 - 144x7 = 1152 puertos 10GE
 - En el slot restante pongamos una 7500E-12CM
 - Eso son 12 puertos 100GE

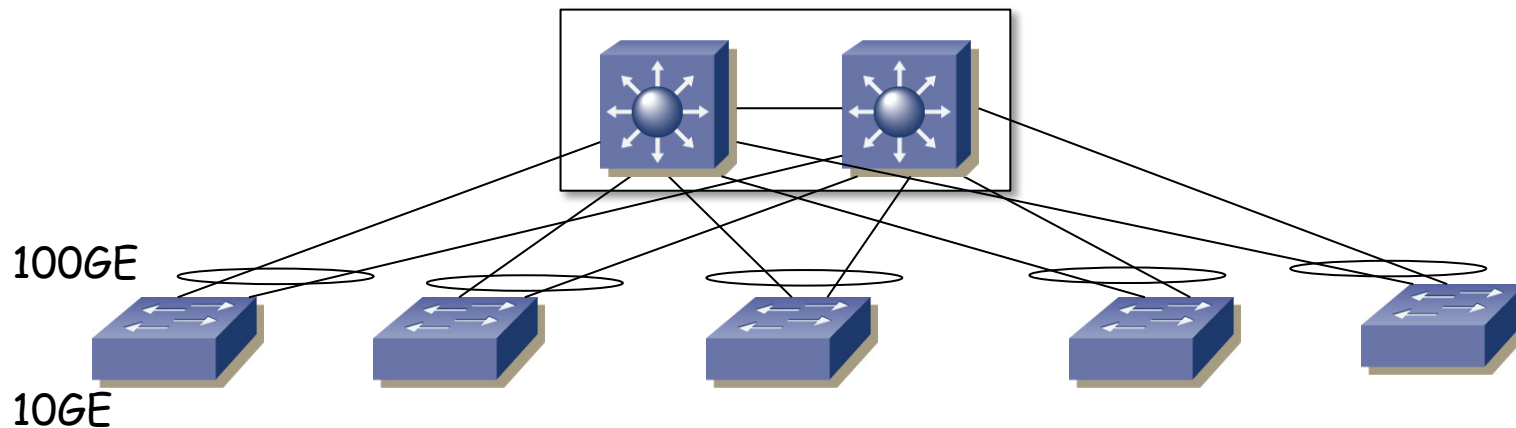
	Ports (type)	Interfaces			Port Buffer	Forwarding Rate	Switching Capacity
		10GbE	40GbE	100GbE			
7500E-48S 	48 SFP+	48	-	-	9GB	720Mpps	960Gbps
7500E-72S 	48 SFP+, 2 MXP	72	-	2	9GB	900Mpps	1.44Tbps
7500E-36Q 	36 QSFP+	144	36	-	18GB	1.8Bpps	2.88Tbps
7500E-12CM 	12 MXP	144	36	12	18GB	1.8Bpps	2.88Tbps
7500E-12CQ 	12 QSFP 100	48	12	12	18GB	1.8Bpps	2.4Tbps
7500E-6C2 	6 CFP2	60*	12*	6	9GB	900Mpps	1.2Tbps



QSFP+ = Quad Small Form-factor Pluggable Plus (un puerto 40G ó 4 puerto 10G)




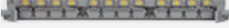


Escalabilidad con MLAG

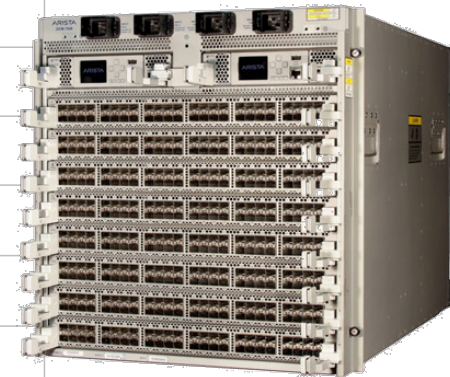
- Tenemos conmutadores con más de 1000 puertos 10GE
- Hacia la segunda capa puertos 40GE o 100GE
- Con un MLAG tendríamos 1000x10GE sobre 2x100GE o una sobresuscripción de 50:1 (¡!)
- Hay conmutadores que permiten cerca de 100 puertos 100GE
- (...)



Ejemplo: Arista 7508

- Sin irnos más lejos
- Linecards:
 - La 7500E-12CM tiene 12 puertos 100GE
 - Con 8 de ellas tendríamos $8 \times 12 = 96$ puertos 100GE

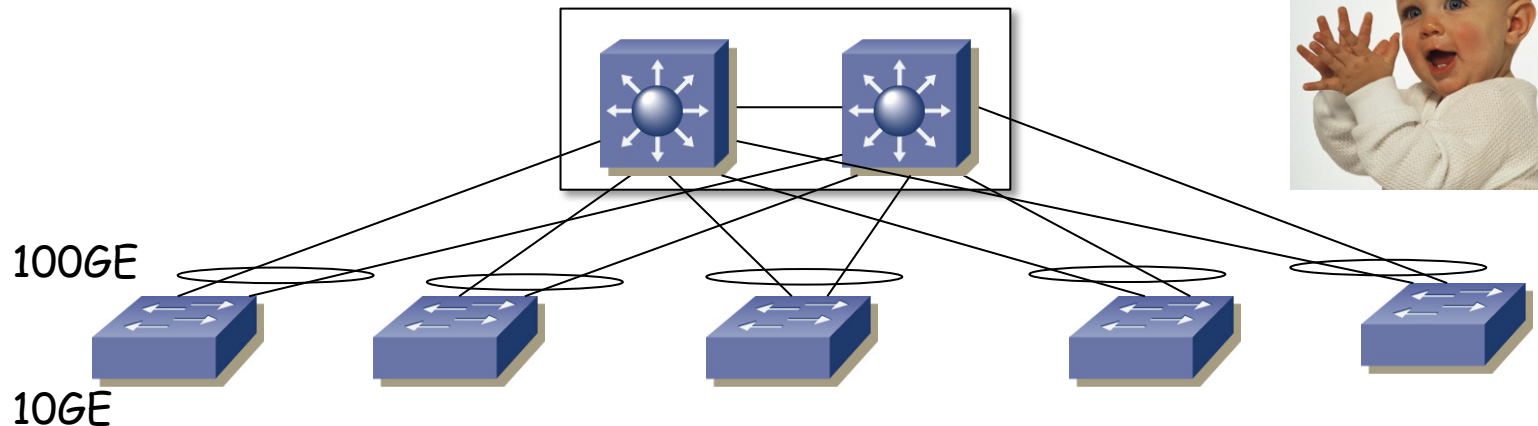
	Ports (type)	Interfaces			Port Buffer	Forwarding Rate	Switching Capacity
		10GbE	40GbE	100GbE			
7500E-48S 	48 SFP+	48	-	-	9GB	720Mpps	960Gbps
7500E-72S 	48 SFP+, 2 MXP	72	-	2	9GB	900Mpps	1.44Tbps
7500E-36Q 	36 QSFP+	144	36	-	18GB	1.8Bpps	2.88Tbps
7500E-12CM 	12 MXP	144	36	12	18GB	1.8Bpps	2.88Tbps
7500E-12CQ 	12 QSFP 100	48	12	12	18GB	1.8Bpps	2.4Tbps
7500E-6C2 	6 CFP2	60*	12*	6	9GB	900Mpps	1.2Tbps



QSFP+ = Quad Small Form-factor Pluggable Plus (un puerto 40G ó 4 puerto 10G)

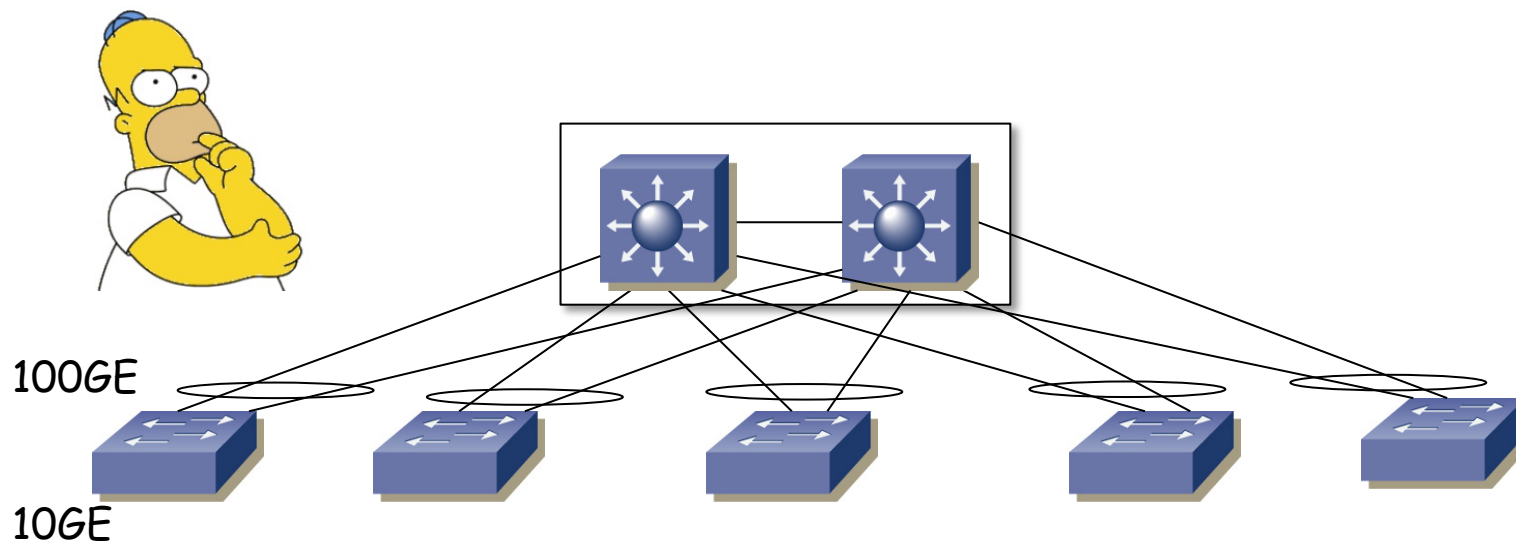
Escalabilidad con MLAG

- Tenemos conmutadores con más de 1000 puertos 10GE
- Hacia la segunda capa puertos 40GE o 100GE
- Con un MLAG tendríamos 1000x10GE sobre 2x100GE o una sobresuscripción de 50:1 (¡!)
- Hay conmutadores que permiten cerca de 100 puertos 100GE
- Entonces podríamos tener 100 conmutadores en el acceso
- Eso son $1000 \times 100 = 100.000$ hosts con un puerto 10GE cada uno
- (...)



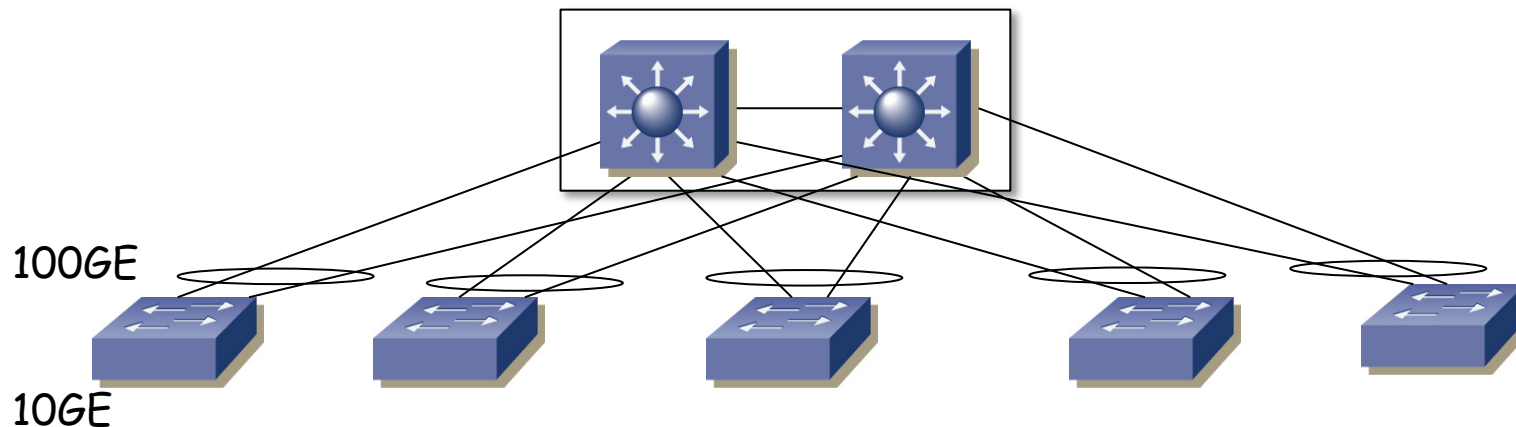
Escalabilidad con MLAG

- Tenemos conmutadores con más de 1000 puertos 10GE
- Hacia la segunda capa puertos 40GE o 100GE
- Con un MLAG tendríamos 1000x10GE sobre 2x100GE o una sobresuscripción de 50:1 (¡!)
- Hay conmutadores que permiten cerca de 100 puertos 100GE
- Entonces podríamos tener 100 conmutadores en el acceso
- Eso son 1000x100 = 100.000 hosts con un puerto 10GE cada uno
- ¿ Algún problema ?
- (...)



Escalabilidad con MLAG

- Tenemos conmutadores con más de 1000 puertos 10GE
- Hacia la segunda capa puertos 40GE o 100GE
- Con un MLAG tendríamos 1000x10GE sobre 2x100GE o una sobresuscripción de 50:1 (¡!)
- Hay conmutadores que permiten cerca de 100 puertos 100GE
- Entonces podríamos tener 100 conmutadores en el acceso
- Eso son $1000 \times 100 = 100.000$ hosts con un puerto 10GE cada uno
- Si en cada servidor tenemos 20 VMs entonces $20 \times 100K = 2M$ direcciones MAC, solo con las MACs de los servidores virtuales
- (...)



Ejemplo: Arista 7508

- 256K direcciones MAC puede almacenar la CAM de una de las tarjetas

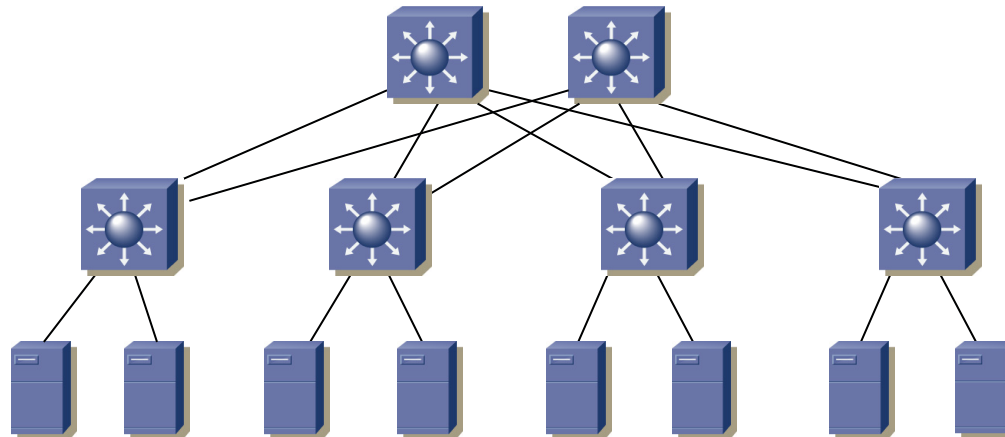
	7500E Linecards
Latency	4usec
MAC Table Size	256K
Maximum v4/v6 Host Routes	128K
Maximum IPv4 Route Prefixes	64K
Maximum IPv6 Route Prefixes	12K
Maximum Multicast Groups	64K
Maximum Egress Routes	30K per port group
Maximum LAG Groups	1K
Maximum LAG Members	64 ports
Maximum ECMP Fanout	64-way
Maximum ACL Entries	12K per port group
Buffer per 10GbE Port	Up to 125MB



Oversubscription con ECMP

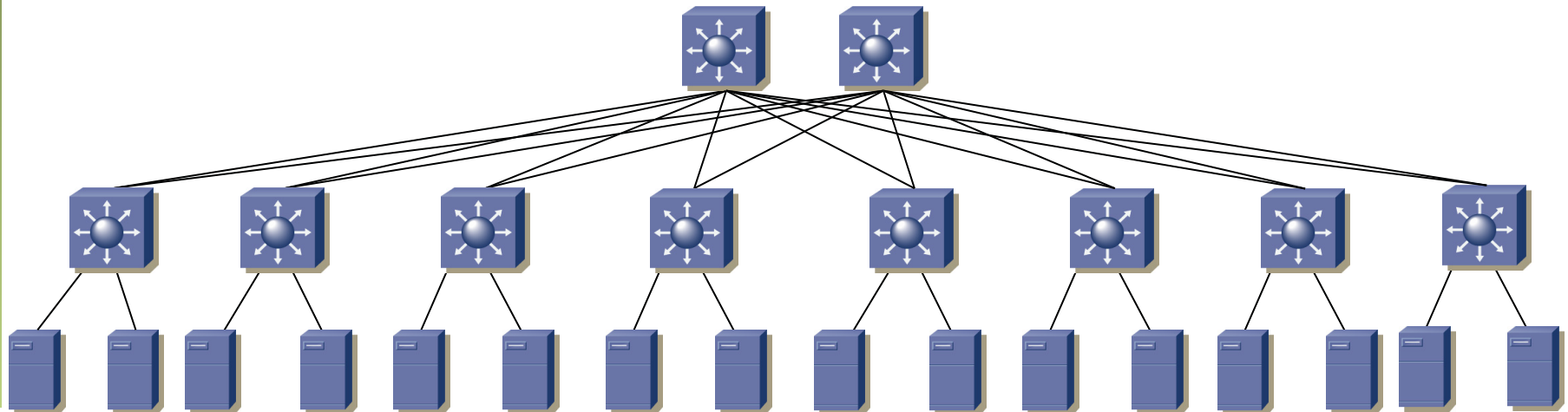
ECMP

- Empleando routing capa 3 podemos aprovechar los múltiples caminos
- No se desactivan enlaces ni tenemos inundación de broadcast
- Puede crecer la topología en la medida en que los equipos puedan crecer en número de puertos
- (...)



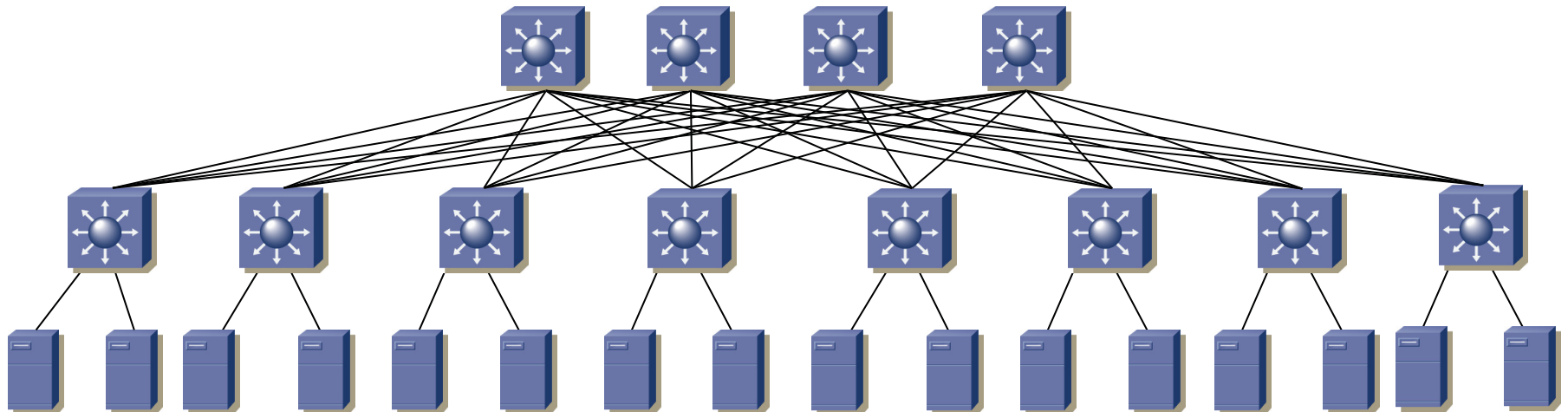
ECMP

- Aumentar el número de conmutadores en la capa de acceso está limitado por el número de puertos de los equipos de agregación
- Es más sensible a la caída de un equipo de agregación
- No mejora la sobre-subscripción
- Por ejemplo, con conmutadores de acceso de 48 puertos 10GE y 2 puertos 40GE hacia agregación sería de $48 \times 10 : 2 \times 40 = 6:1$
- Podemos mejorar la sobre-subscripción con LAGs a la capa de agregación
- Por ejemplo en este caso con LAGs de 2 enlaces sería de 3:1 (4 puertos de 40GE en cada conmutador de acceso)
- (...)



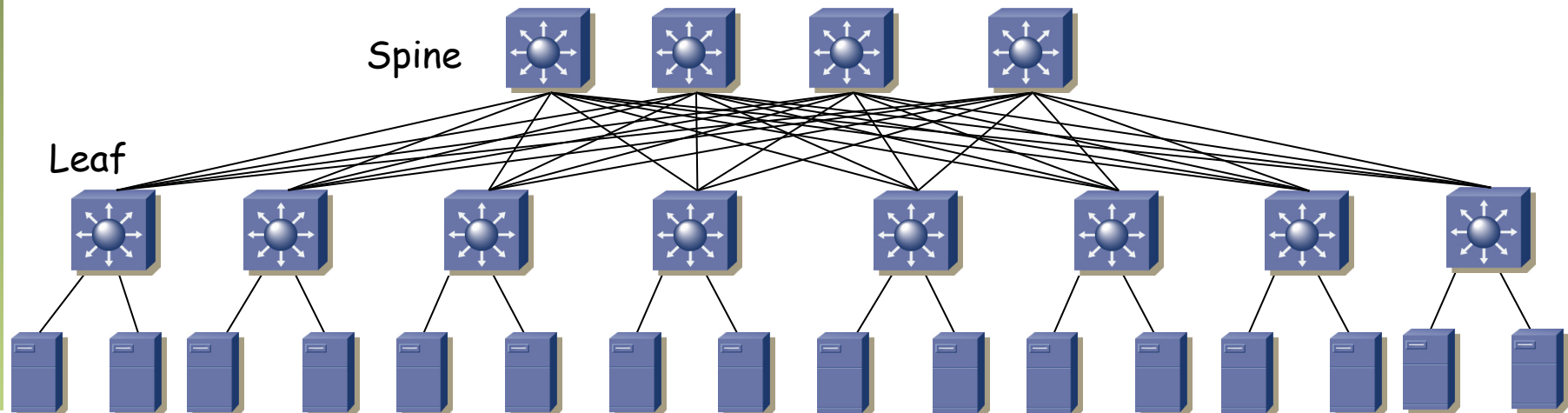
ECMP

- También podemos aumentar el número de conmutadores de agregación
- Pasamos por ejemplo de 2 caminos a 4 caminos de igual coste
- Es menos sensible a la caída de un equipo
- Con los mismos 4 enlaces de los 2 LAGs de 2 enlaces tenemos el mismo nivel de sobresuscripción (3:1)



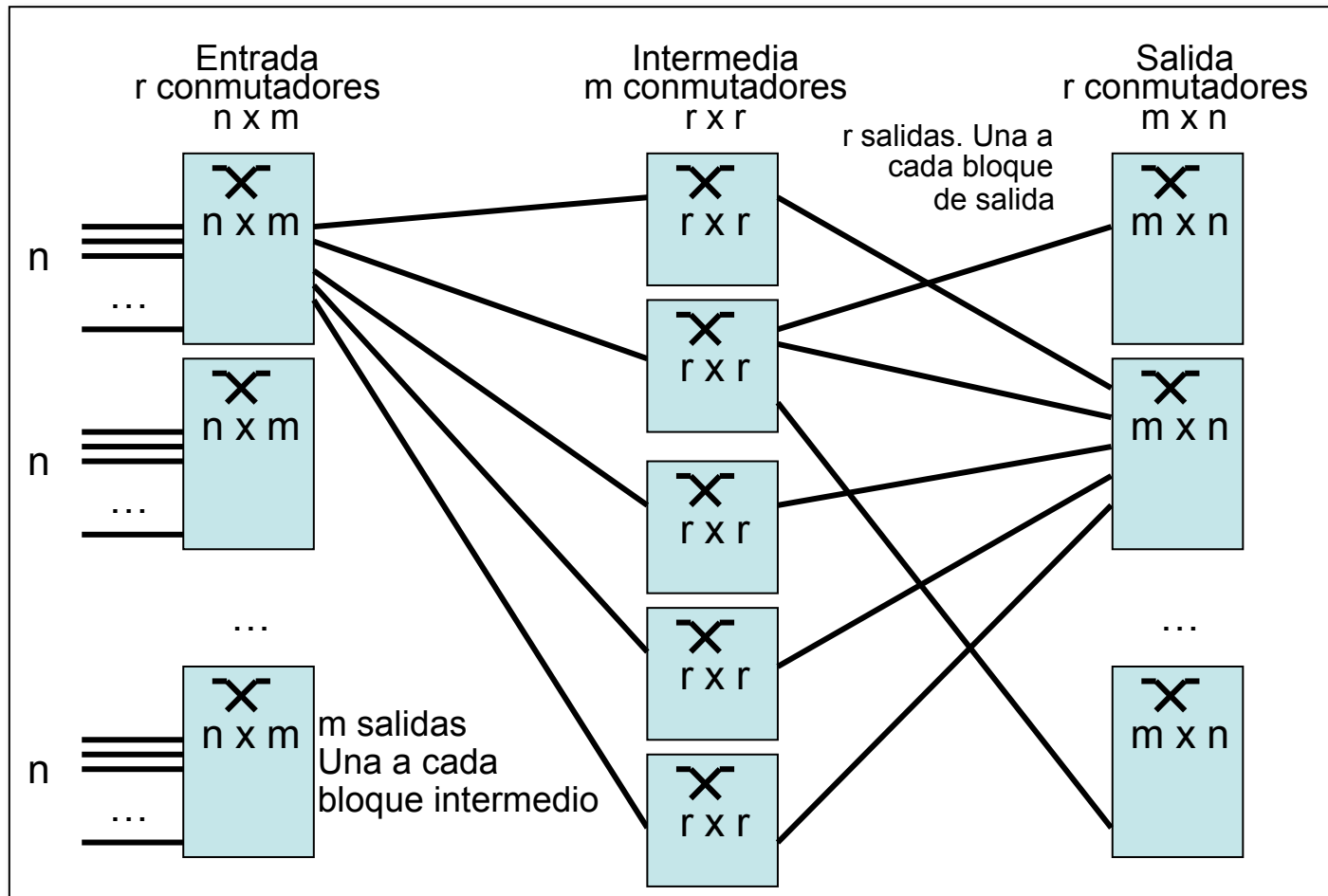
Redes de Clos

- Cada conmutador de la capa inferior está conectado con cada uno de la capa superior
- Si no hay *oversubscription* tenemos una arquitectura sin bloqueo
- Si tenemos una tecnología que saque provecho al multipath
- Estamos creando un gran conmutador o lo que se viene a llamar un “*fabric*”
- Oversubscriptions hasta 3:1 son habituales



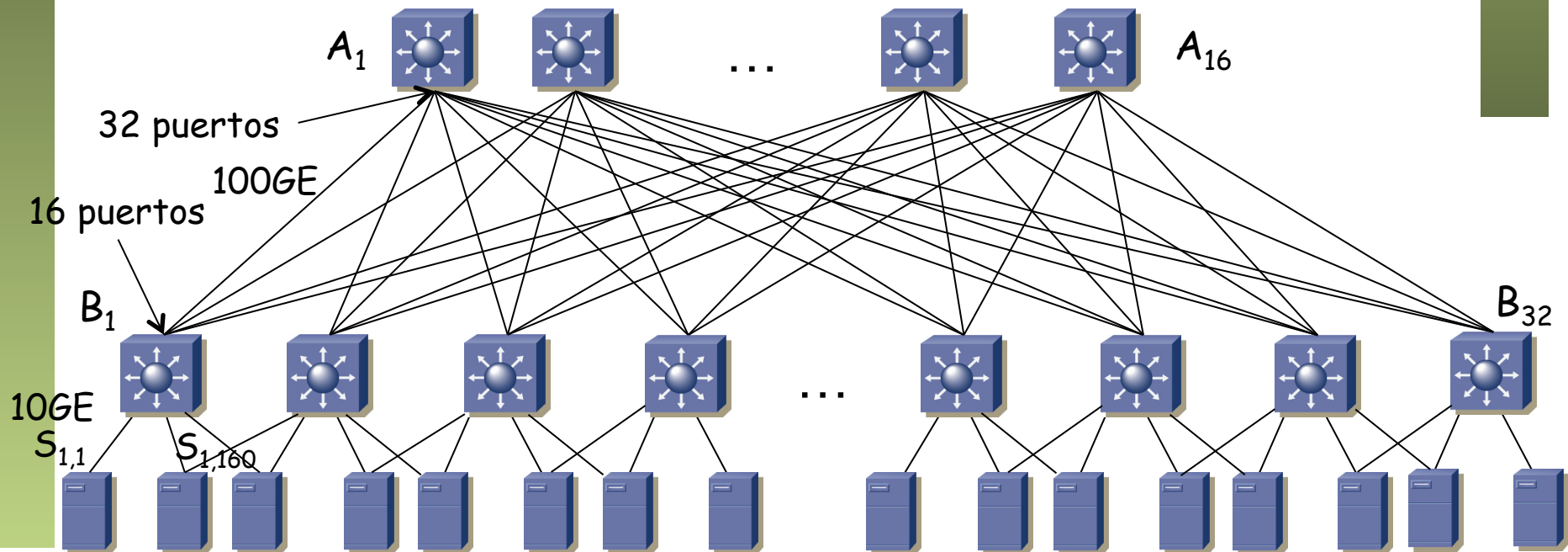
Redes de Clos

- Ya lo vimos al hablar de la 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



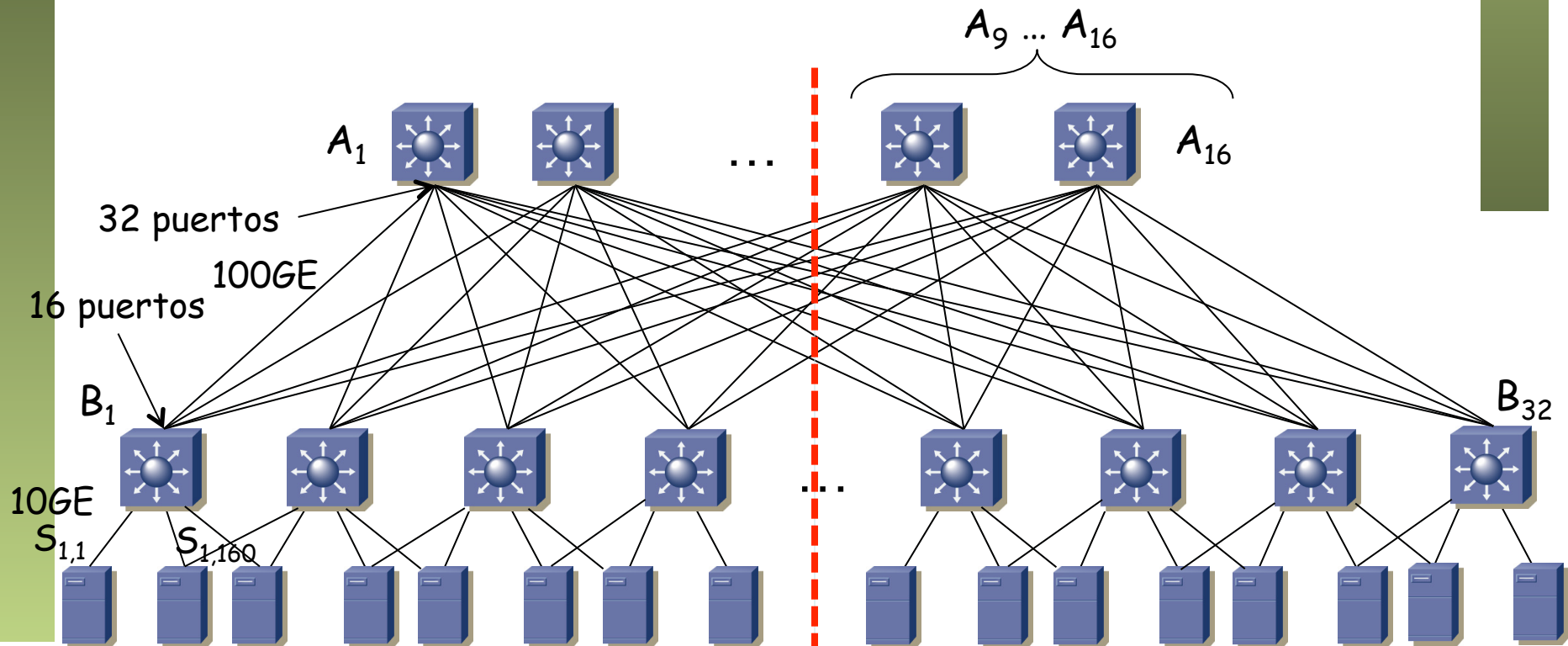
Ejemplo

- 32 conmutadores en la capa de acceso, servidores dual-homed 10GE
- Cada conmutador de acceso da conectividad 10GE a 160 interfaces
- $160 \times 32 / 2 = 2560$ servidores (dual-homed)
- 16 conmutadores en la capa de agregación
- Enlace de 100GE entre acceso y agregación
- (...)



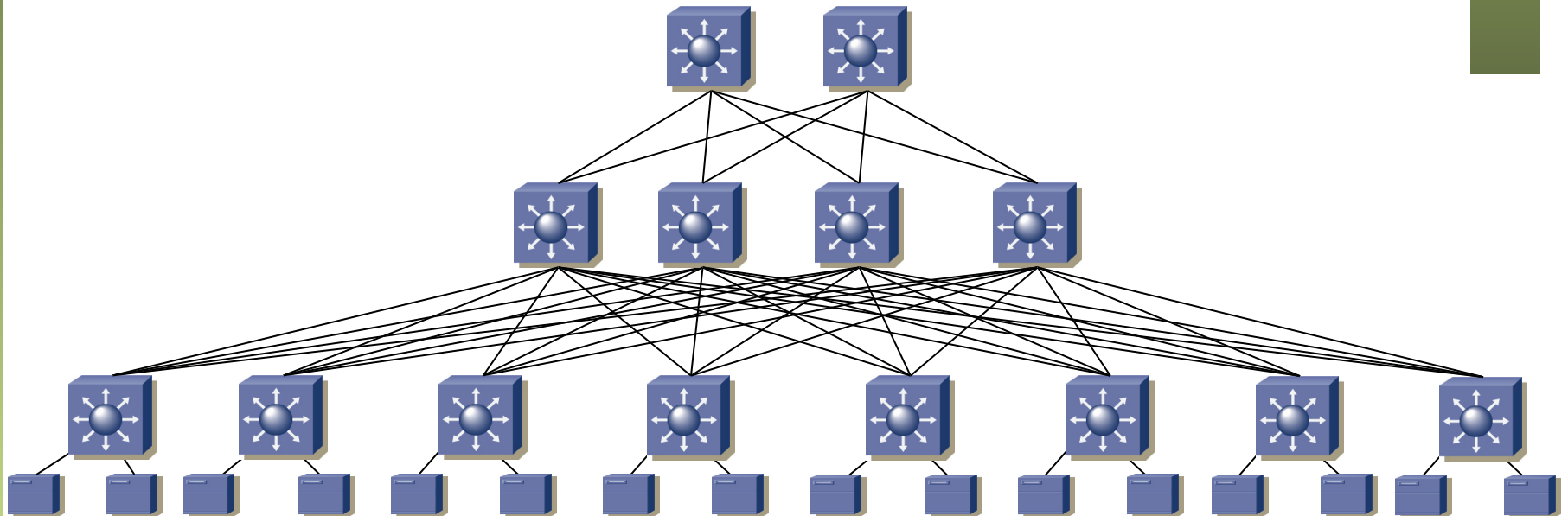
Ejemplo

- De cada conmutador de acceso salen $16 \times 100\text{GE} = 1.6 \text{ Tbs}$
- Cada conmutador de agregación recibe 32 enlaces 100GE
- En total puede haber fluyendo $32 \times 16 \times 100 = 51.2 \text{ Tbps}$
- Sin bloqueo si no tienen bloqueo interno los conmutadores
- BW de bisección $8 \times 16 \times 2 \times 100\text{G} = 25.6 \text{ Tbps}$



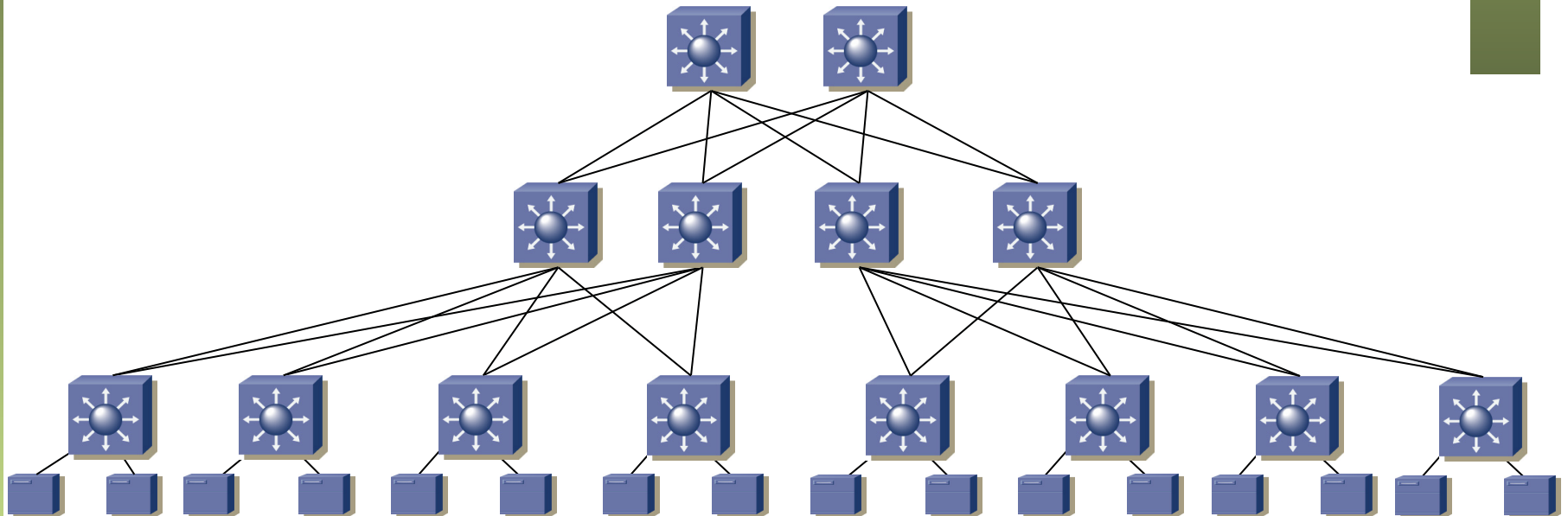
ECMP

- ¿ Tendría sentido un tercer tier ?
- Hay un mejor camino por la capa de agregación, así que no se emplearía
- (...)



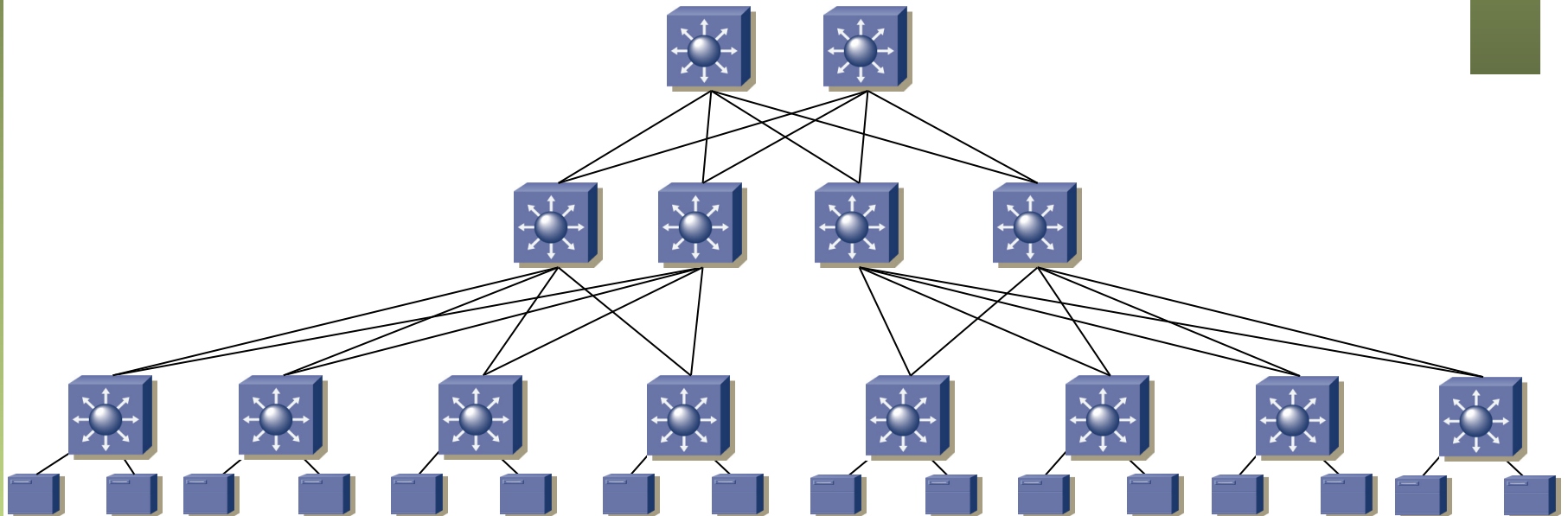
ECMP

- ¿ Tendría sentido un tercer tier ?
- Hay un mejor camino por la capa de agregación, así que no se emplearía
- En esta disposición hay módulos en la capa de agregación que agregan un conjunto de conmutadores de la capa de acceso
- Permite mayor escalabilidad porque los conmutadores de acceso no se enlazan a todos los de agregación
- La conectividad entre los módulos la da el core



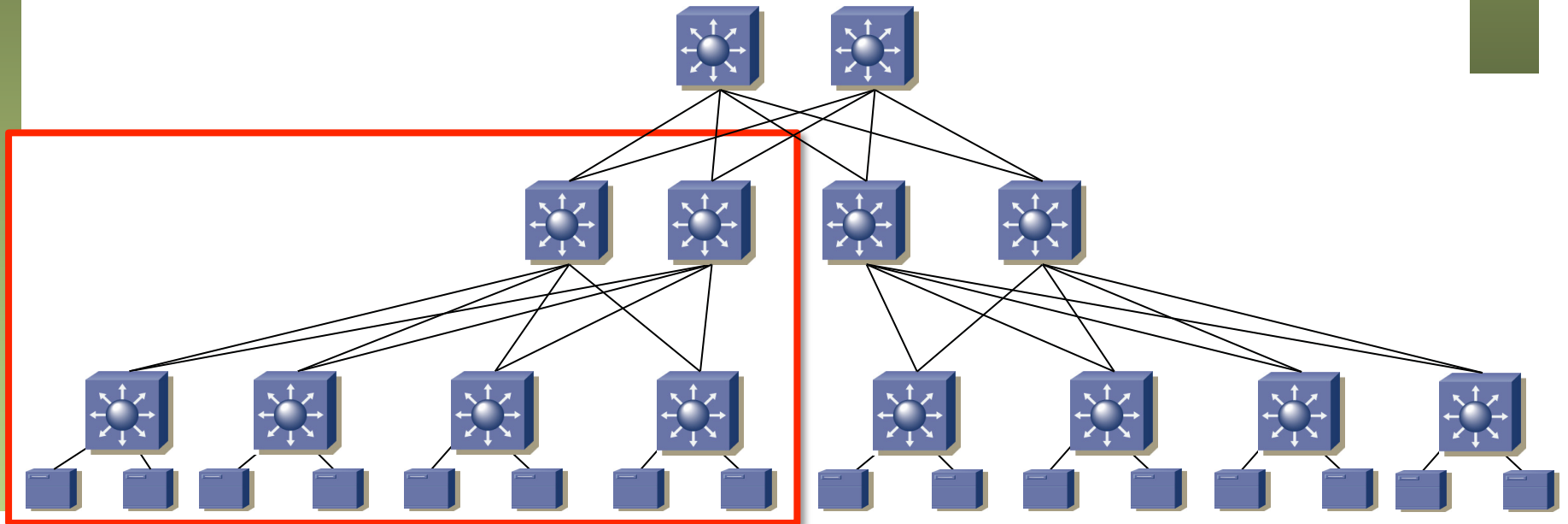
ECMP

- Ejemplo de sobre-subscripción
 - Servidores 10GE, 48 por conmutador de acceso
 - Enlaces de agregación y al núcleo 40GE
 - En la capa de acceso es un $48 \times 10 : 2 \times 40 = 6:1$ (1.6Gbps por servidor)
 - En la capa de distribución es un $4 \times 40 : 2 \times 40 = 2:1$
 - Dado que teníamos 1.6Gbps que llegaban por servidor, entonces con un 2:1 serían solo unos 800Mbps por servidor
 - Es lo mismo que decir que de agregación a core tenemos 4×40 Gbps a repartir entre 4×48 servidores



“Pods”

- Uno o más racks de servidores
- Incluye los equipos que les proveen de conectividad y almacenamiento
- Un cliente de un data center puede comenzar con un “pod” y crecer posteriormente
- Esto no es particular del caso de conmutación en capa 3
- De hecho sería normal la conmutación en capa 2 para poder extender las VLANs a otro “pod”



Oversubscription con ECMP

Alternativas a STP

- Hemos comentado
 - Emplear conmutación capa 3
 - Usar agregación multichassis
- Conmutación capa 3 no permite comunicación en capa 2
- Eso es un problema para ciertas aplicaciones, en especial con funcionalidades de clustering
- Por ejemplo no permite la movilidad de las VMs
- Agregación multichassis está limitada a dominios en el orden de los miles de hosts

Alternativas a STP

Overlays en el data center

Overlays

- Permiten que las tablas de direcciones MAC de los conmutadores no crezcan con el número de hosts
- Para ello intentan evitar que los conmutadores del núcleo aprendan las direcciones MAC de los hosts
- Esto lo van a hacer encapsulando las tramas Ethernet de los hosts extremo
- Para entornos con mucho tráfico este-oeste en vez de norte-sur
- Alternativas existentes:
 - TRILL (Transparent Interconnection of Lots of Links)
 - SPB (Shortest Path Bridging)
 - FabricPath (similar a TRILL)
 - VXLAN (Virtual Extensible LAN)
 - OTV (Overlay Transport Virtualization)
 - NVGRE
 - etc



TRILL

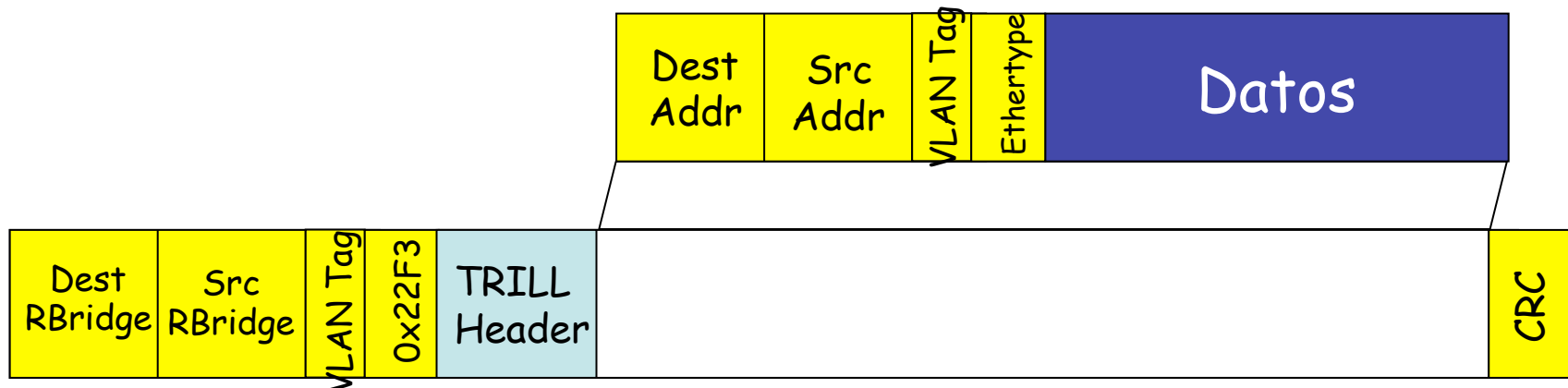


TRILL

- *Transparent Interconnection of Lots of Links*
- IETF (RFCs 6325 y otras)
- Pretende sustituir a STP
- El conmutador que lo implementa se conoce como un RBridge (Routing Bridge)
- Lo básico
 - Los RBriges y enlaces o LANs puenteadas que los interconectan forman un “campus”
 - Transportan las tramas Ethernet por ese campus encapsulándolas en otras tramas Ethernet (MAC in MAC)
 - Esa cabecera adicional incluye una cuenta de saltos
 - El camino por el campus lo calcula IS-IS (permite ECMP)
 - Se desencapsula en el RBridge de salida hacia el destino
- Está especificado su transporte sobre Ethernet y sobre PPP

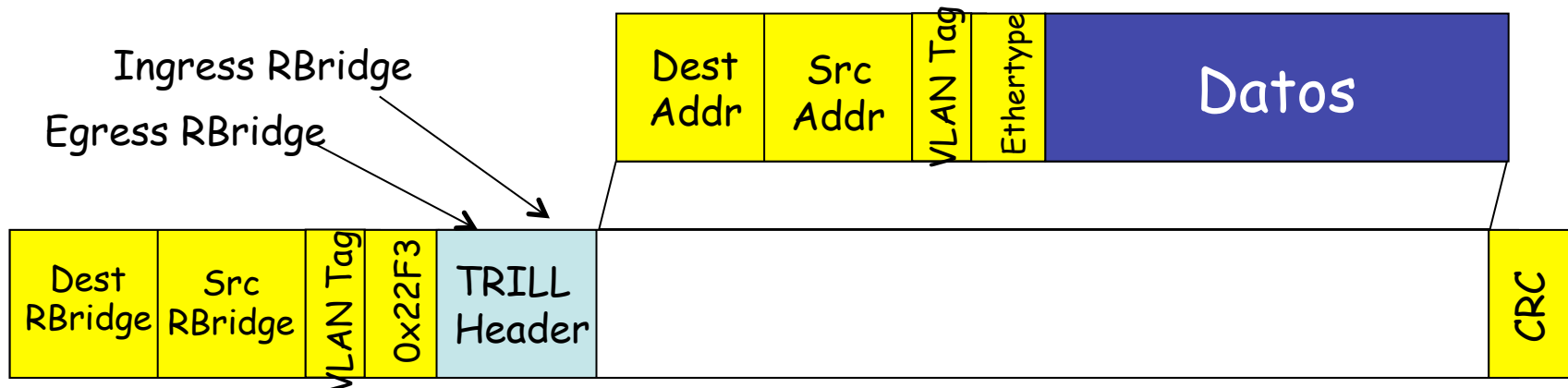
TRILL sobre Ethernet

- MAC origen y destino son de los RBridges
- Puede llevar etiqueta de VLAN si los conmutadores del campus TRILL la necesitan
- Ethertype 0x22F3
- TRILL añade su propia cabecera (la vemos más adelante)
- A continuación la trama que ha llegado al RBridge frontera
- Si la trama original no llevaba etiqueta de VLAN se le añade
- Los conmutadores del Campus TRILL (sean RBridges o no) van a reenviar en base a las direcciones de la cabecera exterior



TRILL sobre Ethernet

- En cada salto entre RBridges las direcciones MAC más exteriores son de los RBridges que envían y reciben esa trama
- Es decir, “Dest RBridge” es la dirección del siguiente salto
- “Src RBridge” es la dirección del salto anterior
- Parecido al caso en que los RBridges fueran routers
- Los RBridges frontera (entrada a la campus y salida) están indicados en la cabecera de TRILL



TRILL Header

- Nicknames
 - Cada RBridge posee un *nickname* con el que se le hace referencia en las PDUs de TRILL y sirve para identificarlo de cada a IS-IS
 - Los nicknames son números de 2 bytes
 - Los nicknames se eligen mediante un proceso automático con información añadida a los mensajes de IS-IS

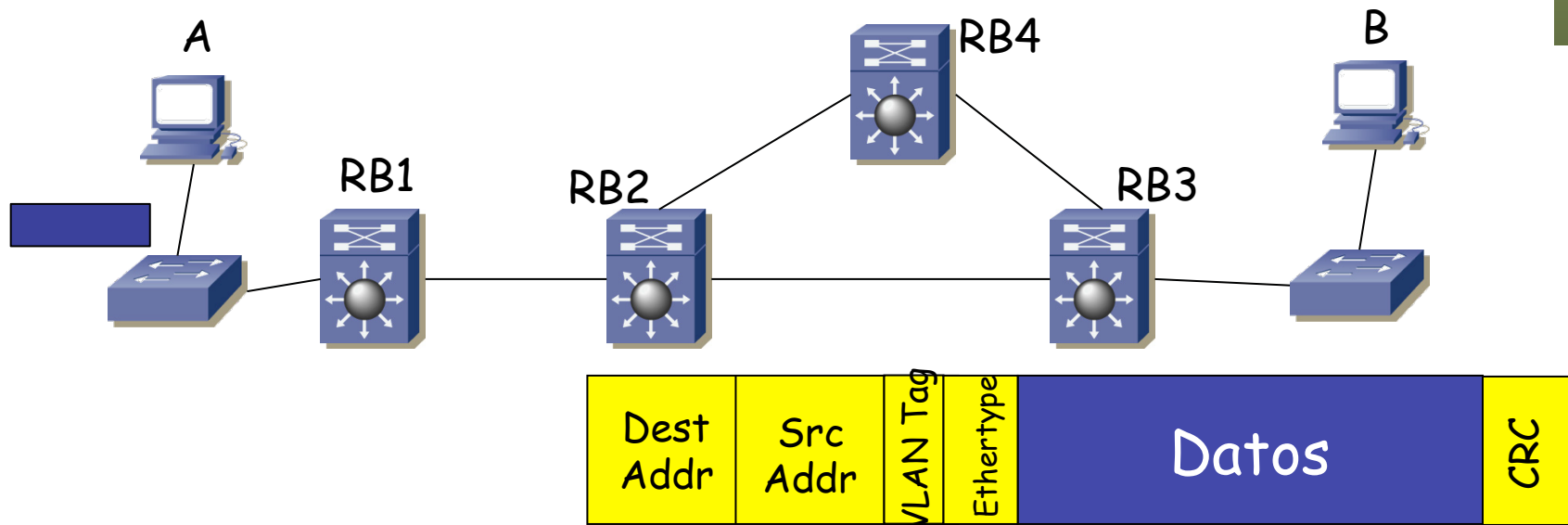
- Campos de la cabecera
 - V = Version (2 bits)
 - R = Reserved (2 bits)
 - M = Multi-Destination (1 bit)
 - ExtLng = Length of TRILL Header Extensions
 - Hop = Hop Limit (6 bits)
 - Egress RBridge Nickname = nickname del RBridge de salida del campus hacia el host destino
 - Ingress RBridge Nickname = nickname del RBridge de entrada al campus de la trama desde el host origen

TRILL Ethertype	V	R	M	ExtLng	Hop
Egress RBridge Nickname	Ingress RBridge Nickname				

TRILL data y control paths

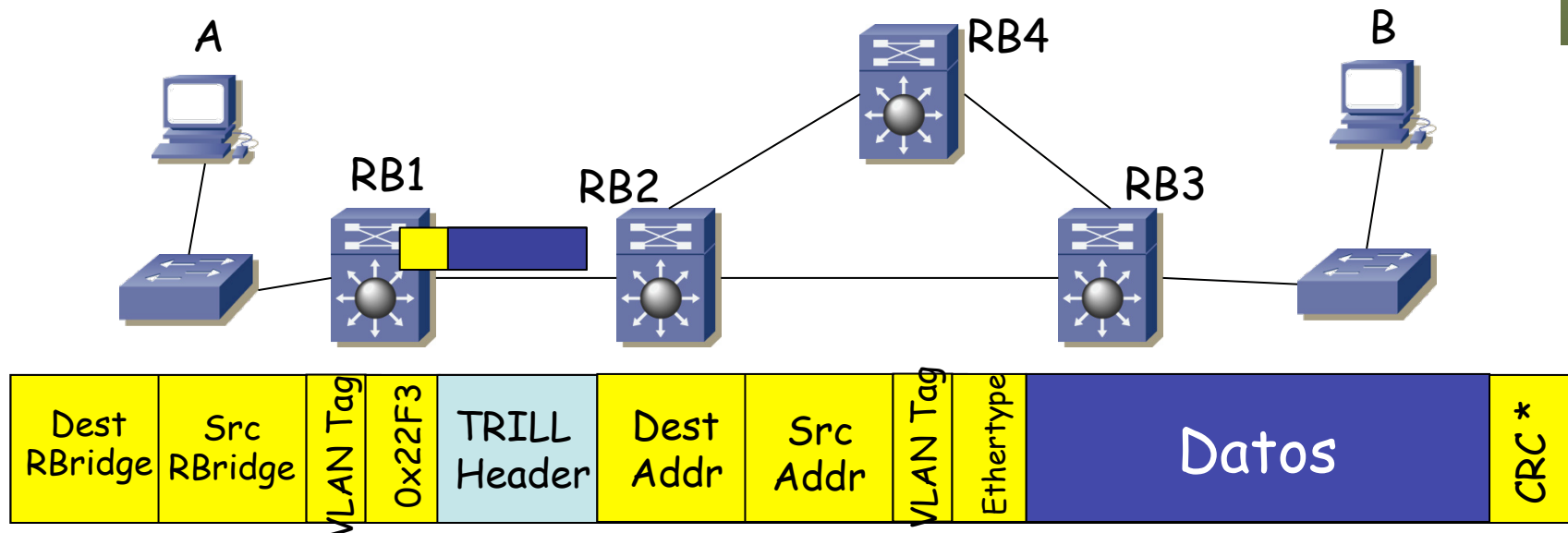
TRILL data path

- Trama Ethernet original
 - Dirección MAC origen de A
 - Dirección MAC destino de B



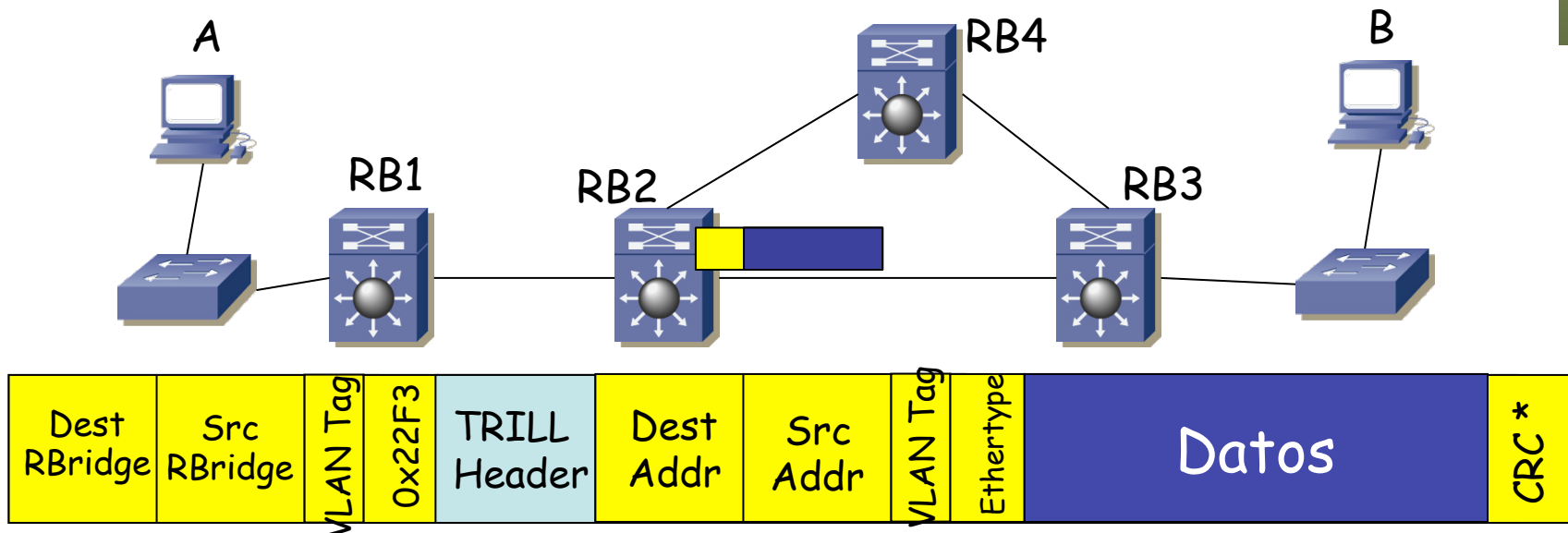
TRILL data path

- La trama llega a RB1
- Calcula cuál es el siguiente salto en el campus TRILL hacia B
- Encapsula esa trama:
 - Dest RBridge = MAC de RB2
 - Src RBridge = MAC de RB1
 - Egress RBridge Nickname = Nickname de RB3
 - Ingress RBridge Nickname = Nickname de RB1
 - TTL = n



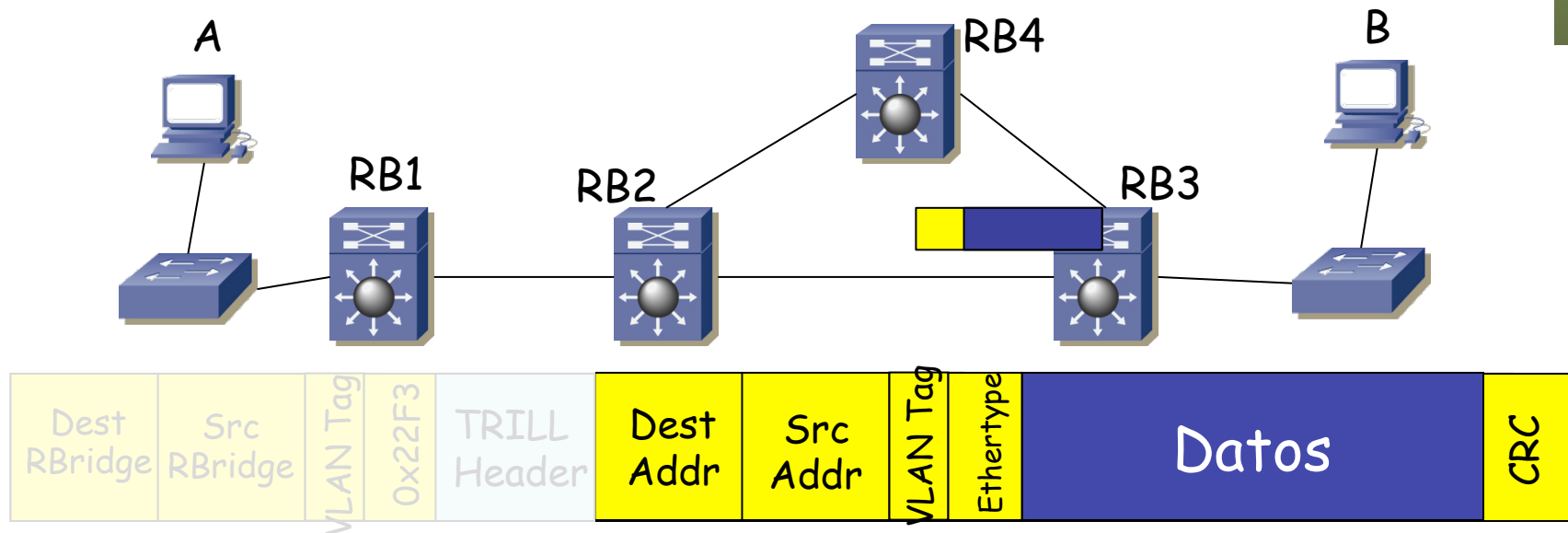
TRILL data path

- La trama llega a RB2
- Calcula cuál es el siguiente salto en el campus TRILL hacia B
- Modifica esa trama:
 - Dest RBridge = MAC de RB3
 - Src RBridge = MAC de RB2
 - Egress RBridge Nickname = Nickname de RB3 (no cambia)
 - Ingress RBridge Nickname = Nickname de RB1 (no cambia)
 - TTL = TTL – 1 (se tira la trama si al recibirla tiene TTL=0; en IP es si vale 0 cuando se va a enviar)



TRILL data path

- La trama llega a RB3
- Calcula cuál es el siguiente salto en el campus TRILL hacia B
- Desencapsula esa trama
- Los RBridges, en cierto modo se han comportado como routers



TRILL control path

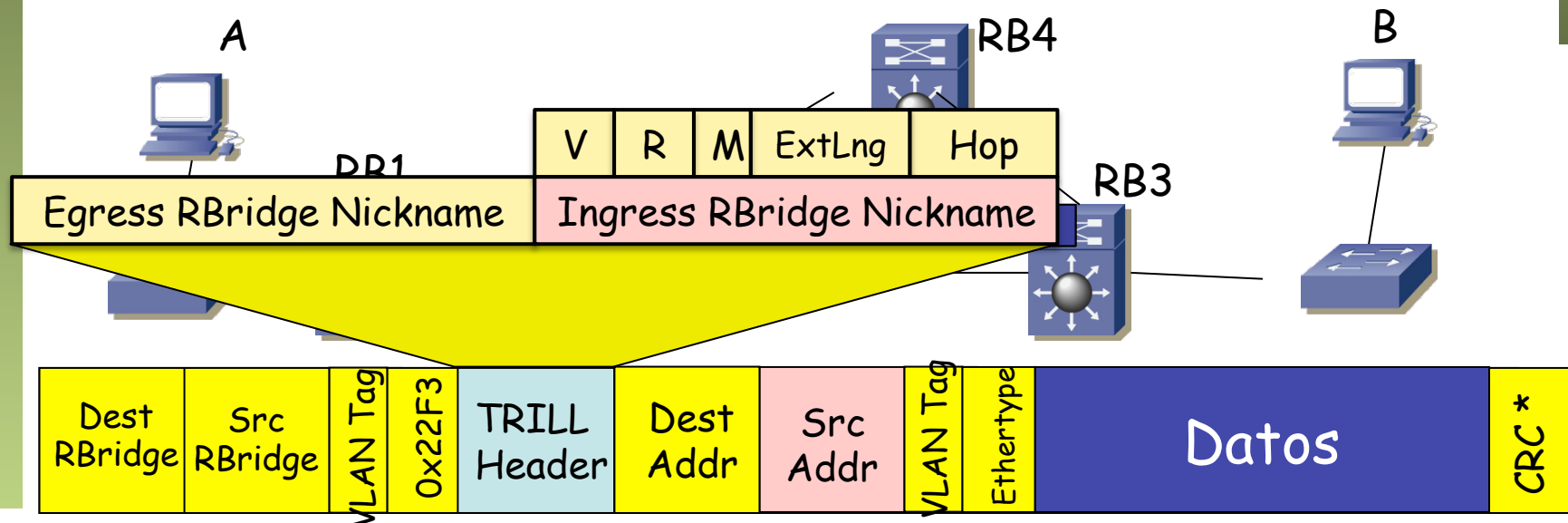
- IS-IS directamente sobre el nivel de enlace
- Ethertype 0x22F4
- Todos los mecanismos típicos de un protocolo link-state
- Más añadidos específicos para TRILL (por ejemplo en el tema de routers designados)



Más sobre TRILL

Aprendizaje

- RBridge frontera aprende direcciones MAC de hosts remotos junto con:
 - RBridge por el que acceden al campus
 - RBridge siguiente salto hacia ese egress RBridge
- Lo hace principalmente en base a los paquetes de TRILL que recibe
- Solo los RBridges frontera necesitan aprender direcciones MAC de los hosts
- Pueden aprender también mediante ESADI (opcional)
 - *End-Station Address Distribution Information*
 - Un RBridge puede anunciar MACs de hosts a otros RBridges
 - Se transporta en tramas TRILL

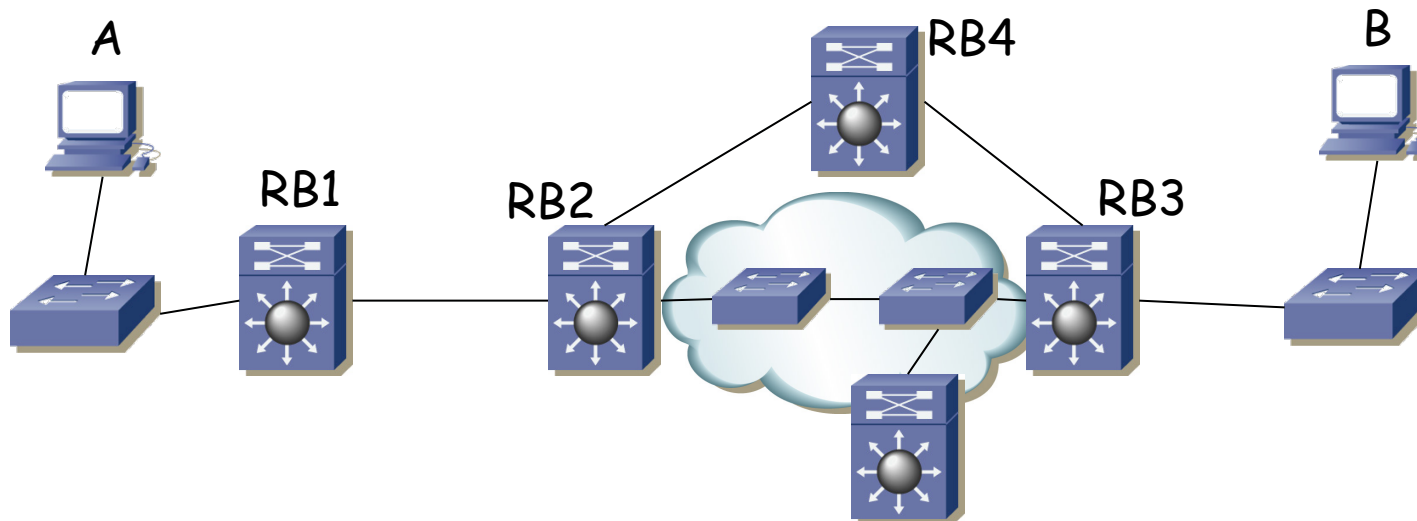


Multidestination

- Casos
 - Tramas unicast para las que no se conoce dónde está el destino
 - Tramas multicast
 - Tramas broadcast
- Los RBridges construyen árboles de distribución para las tramas multicast
- Lo hace con el mismo IS-IS (no hace falta otro protocolo)
- Cada árbol incluye todos los RBridges del campus y las VLANs
- Puede hacer *pruning*
- Se marcan las tramas con un bit en la cabecera de TRILL
- El Nickname del egress RBridge especifica el árbol
- Los árboles son bidireccionales
- Sería suficiente con un árbol pero calcula múltiples, lo cual le permite multipath también para el multicast
- Los nodos hacen una comprobación de RPF

TRILL y puentes

- Entre dos RBridges puede haber un enlace directo o una LAN con puentes
- Puede haber varios RBridges en una LAN con puentes



FabricPath y TRILL

- FabricPath es propietario de Cisco
- El plano de control es como en TRILL, es decir, IS-IS sobre L2
- El plano de datos es similar por emplear encapsulación MAC in MAC
- Las direcciones MAC son asignadas localmente, jerárquicamente
 - SwitchID es el identificador único del switch (manual o automático)
 - SubSwitchID para vPC+
 - PortID puede usarse para indicar el puerto en que está el host
 - EndnodeID se puede emplear para distinguir al host origen/destino
 - OOO/DL indica si se puede emplear balanceo por paquete
 - FTag (*Forwarding Tag*) indica una topología lógica que debe emplear
 - Ethertype 0x9003
- Emplea el SwitchID y el FTAG para las decisiones de reenvío

