

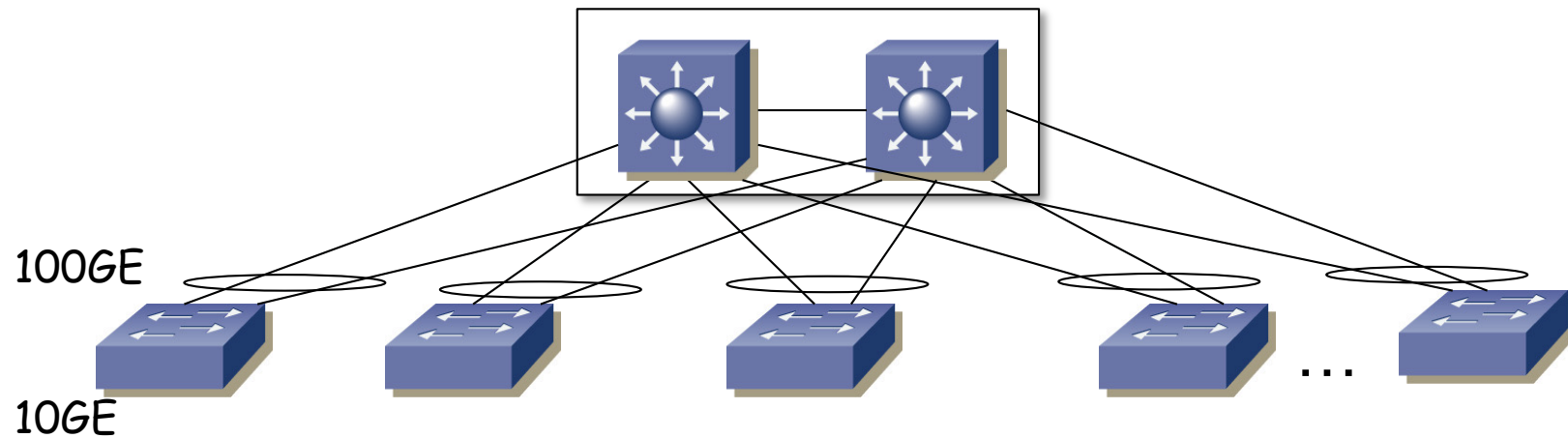
Arquitectura tradicional en el data center: Escalabilidad

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)*

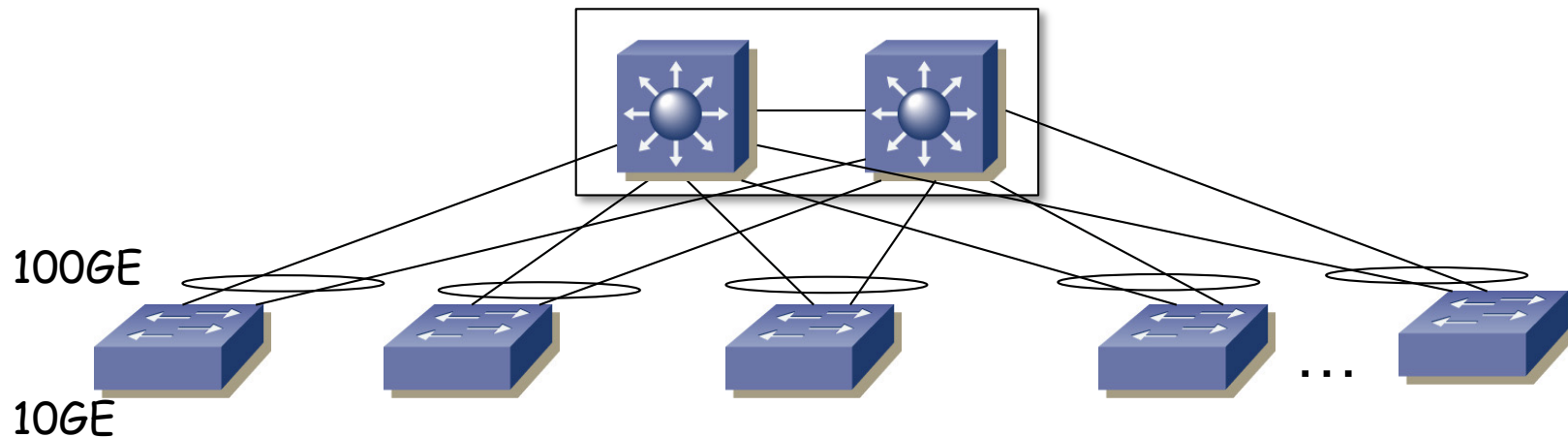
Escalabilidad con MLAG

- (...)



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

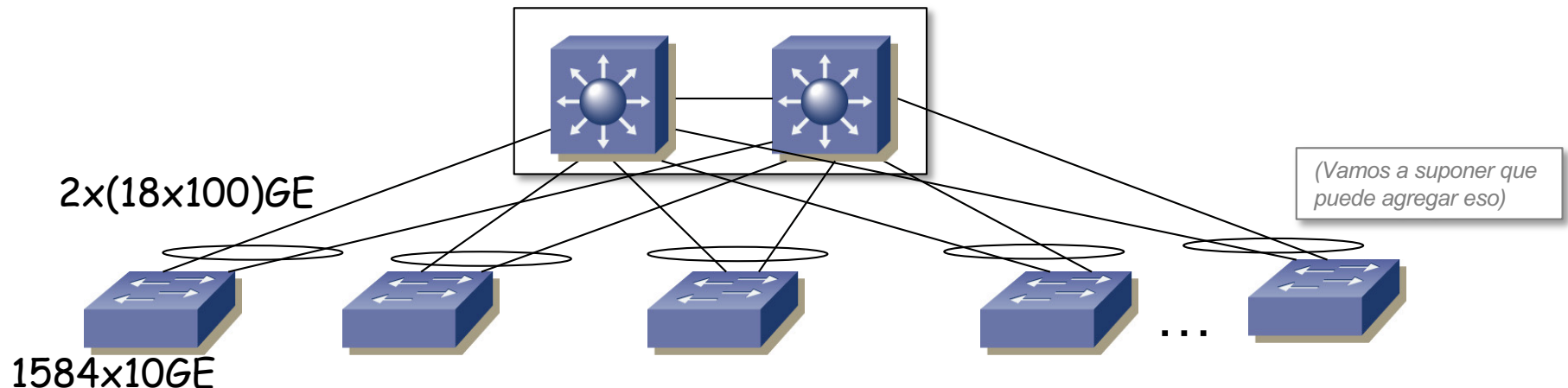


QSFP+ = Quad Small Form-factor Pluggable

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

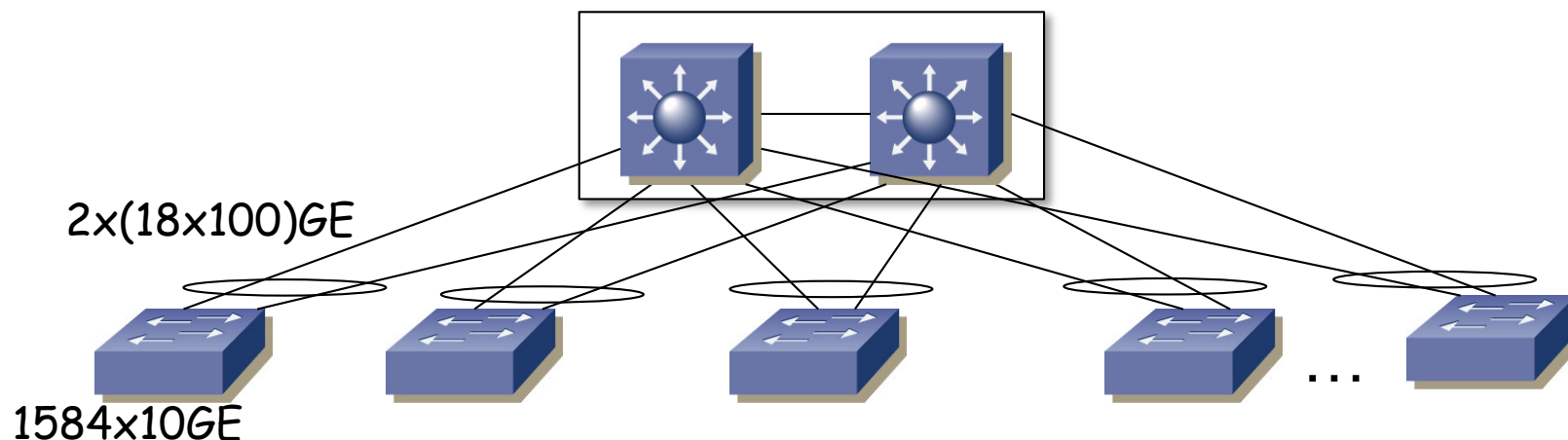


QSFP+ = Quad Small Form-factor Pluggable

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

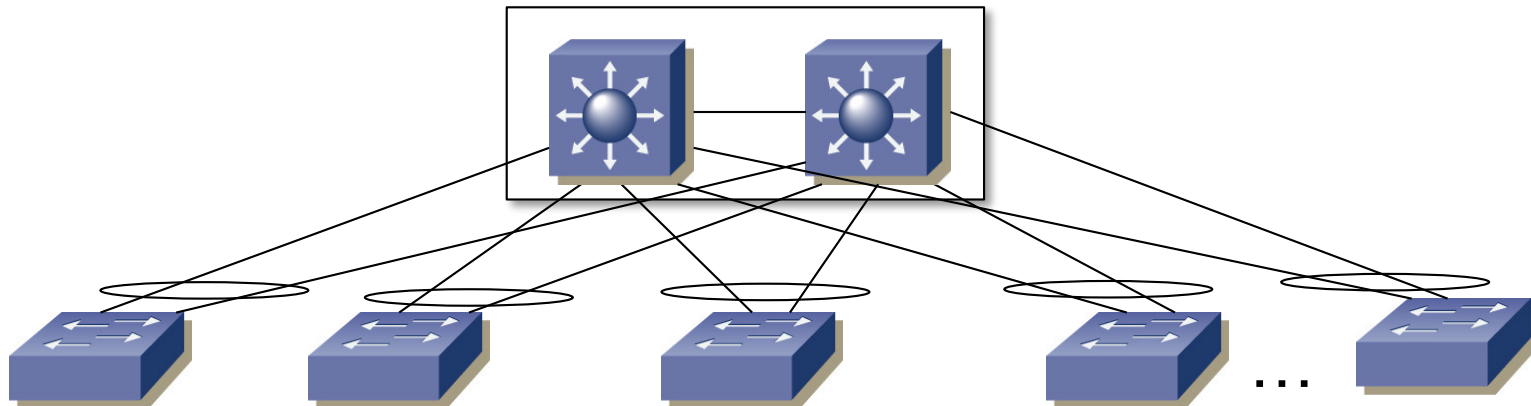
Escalabilidad con MLAG

- 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
- (...)



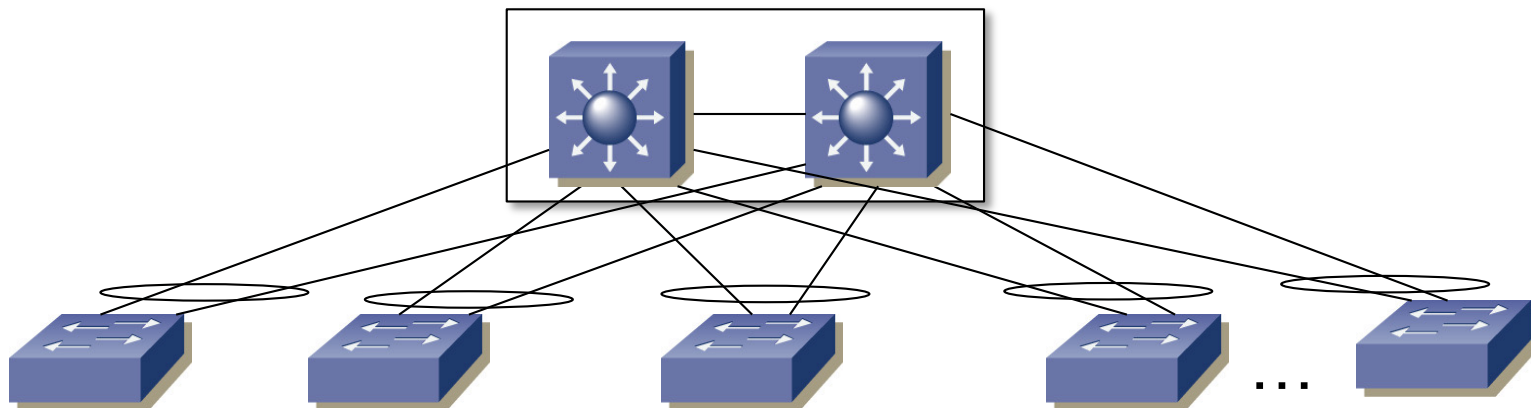
MLAG no bloqueante

- ¿Podríamos hacerlo no-bloqueante?
- (...)



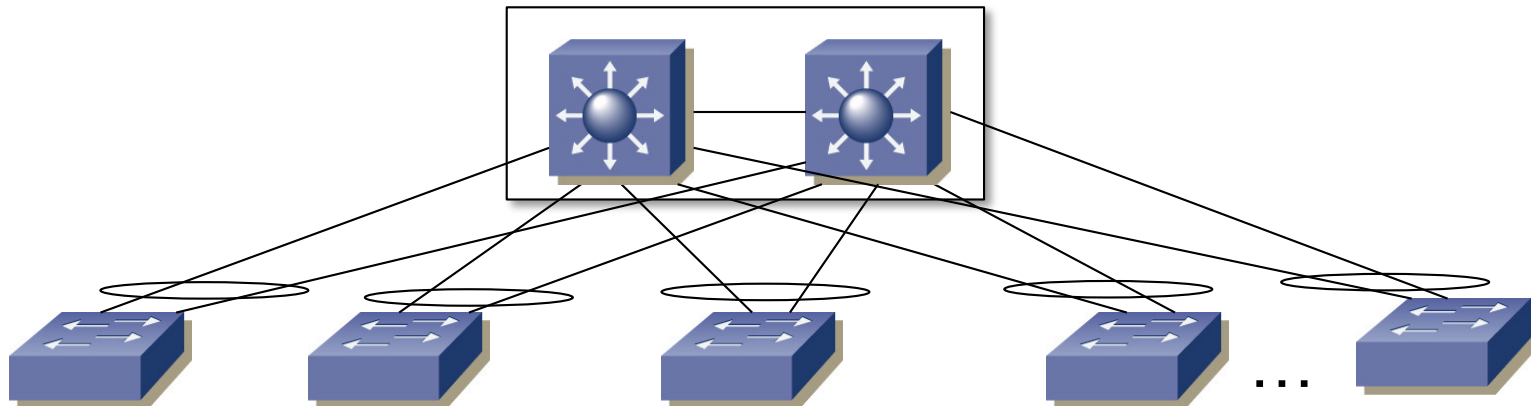
MLAG no bloqueante

- ¿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 4x10GE
 - (...)



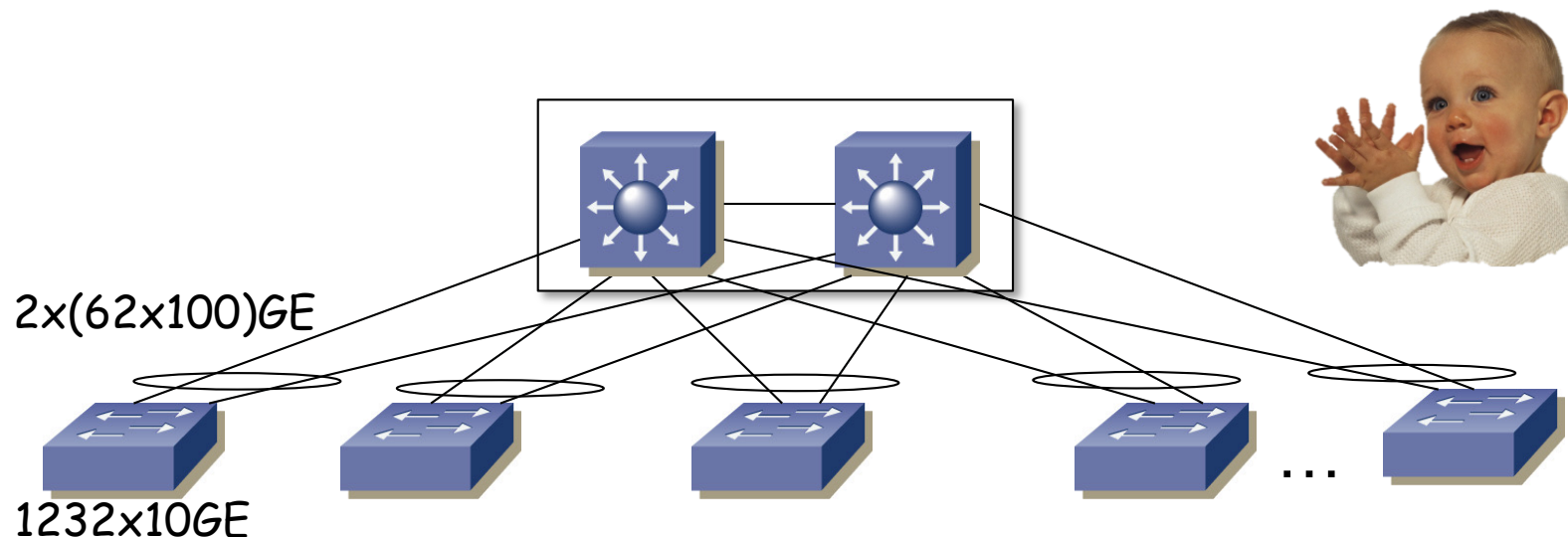
MLAG no bloqueante

- ¿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 16×100 GE y $(20 \times 4) \times 10$ GE :
 - $8 \times 144 + 80 = 1232$ puertos 10GE (12.3 Tbps)
 - $3 \times 36 + 16 = 124$ puertos 100GE (12.4 Tbps)
 - En cada switch de acceso 12.3 Tbps desde los hosts para 12.4 Tbps uplink
 - (...)



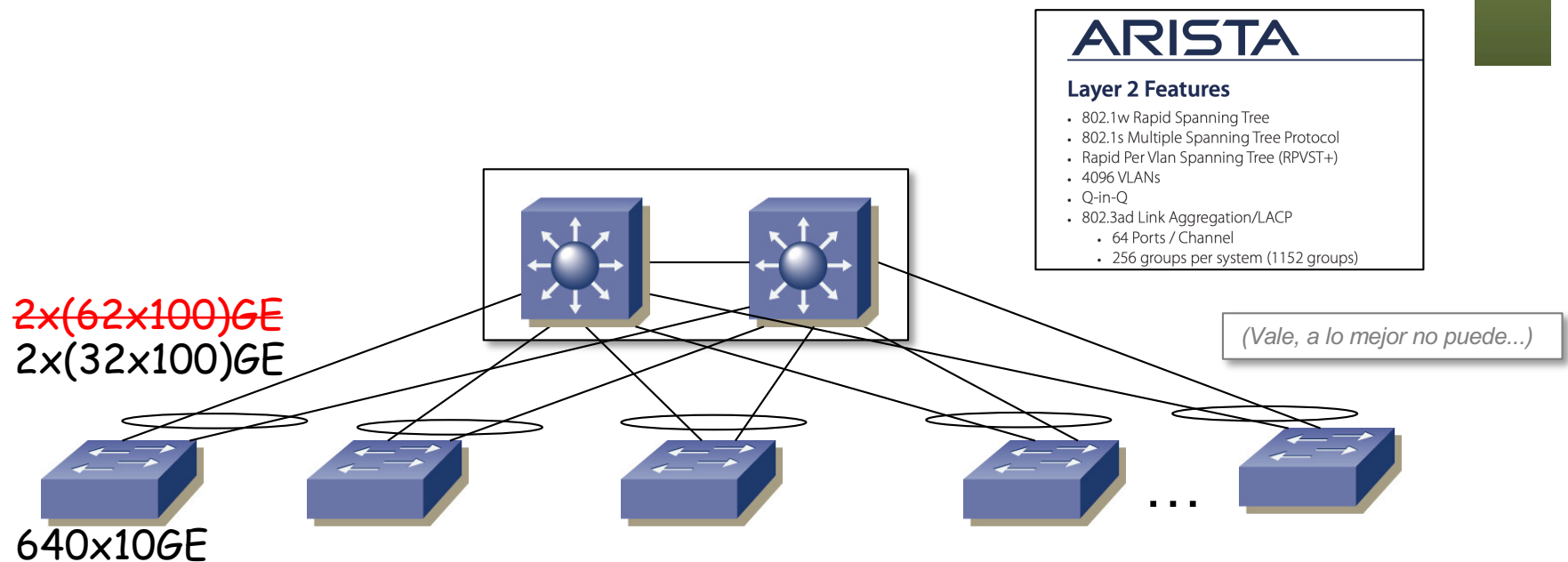
MLAG no bloqueante

- $8 \times 144 + 80 = 1232$ puertos 10GE (12.3 Tbps)
- $3 \times 36 + 16 = 124$ puertos 100GE (12.4 Tbps)
- En cada switch de acceso 12.3 Tbps desde hosts para 12.4 Tbps uplink (124 puertos en el LAG de cada switch, 62 a c.u. del MLAG)
- ¿Cuántos puertos 10GE?
- El switch de agregación podía tener 432 puertos 100GE
- Cada 62 puertos a un switch así que tenemos hasta 6 conmutadores de acceso ($7 \times 62 = 434 > 432$)
- Así que en total $6 \times 1.232 = 7.392$ hosts con puertos 10GE non-blocking
- Antes teníamos 38.016 hosts con sobre-subscripción 4.4:1



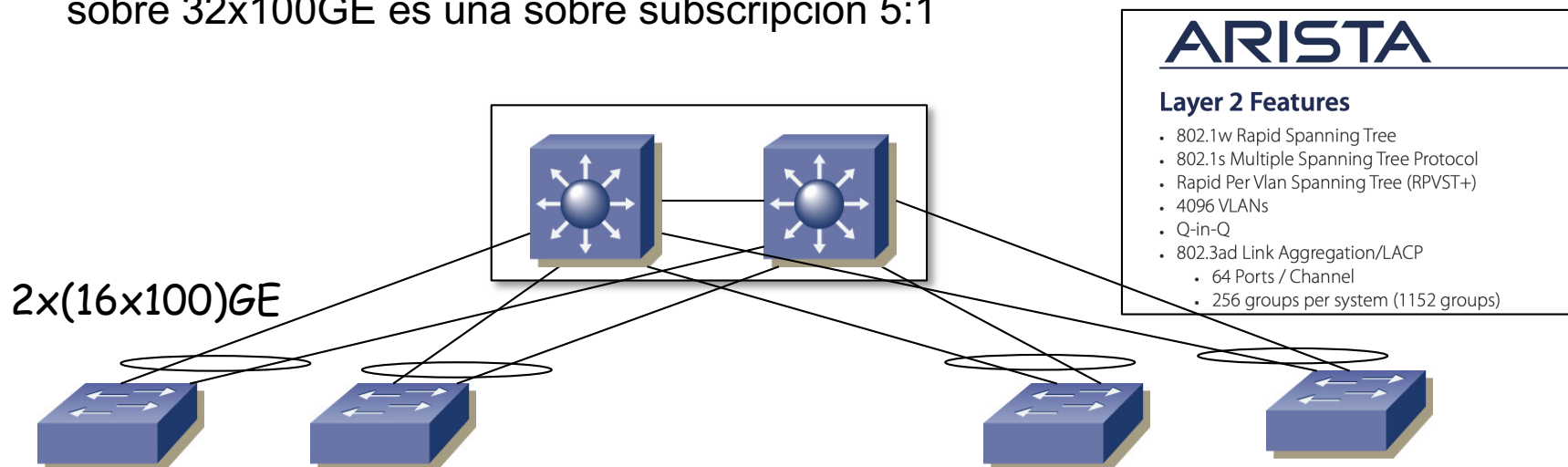
MLAG (limitaciones)

- ¿Pero y si el switch soportara un máximo de 64 puertos por LAG? 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 (limitaciones)

- Hay limitación de puertos en el LAG
- ¿Limitación de puertos en el MLAG?
- Si la limitación fuera la misma:
 - ¿Cada switch de agregación 64 puertos en el MLAG?
 - Por ejemplo 4 switches de acceso, 16 puertos a cada switch de agregación
 - $32 \times 100\text{Gb/s} = 3.2\text{Tb/s}$ desde cada switch de acceso hacia el MLAG
 - Si queremos mantener el 1:1 podríamos tener 320 puertos 10GE en cada switch (4 switches x 320 = 1.280 puertos 10GE)
 - Entonces en acceso no merece la pena ese mismo modelo de switch (él solo ya daba más de 1.500 puertos 10GE)
 - Con ese switch prácticamente lleno: 1.584x10GE por switch de acceso sobre 32x100GE es una sobre subscripción 5:1



Escalabilidad con MLAG

- ¿ 38.016 puertos 10GE con sobre-subscripción 4.4:1 ? Supongamos que sí
- Supongamos que en cada host tenemos 2 CPUs de 12 cores y 24 threads c.u.
- Y por ejemplo decidimos correr 24 VMs... o 200 contenedores
- Con una vNIC por contenedor, puenteada a la LAN, son 200x38.016 o más de 7.6 millones de direcciones MAC



Nombre del producto	Estado	Fecha de lanzamiento	Núcleos	Frecuencia de turbo máxima	Frecuencia base del procesador	Caché	TDP
☐ Procesador Intel® Xeon® Platinum 8490H (caché de 112,5 MB, 1,90 GHz)	Launched	Q1'23	60	3.50 GHz	1.90 GHz	112.5 MB	350 W



	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 MLAG

- ¿ 38.016 puertos 10GE con sobre-subscripción 4.4:1 ? Supongamos que sí
- Supongamos que en cada host tenemos 2 CPUs de 12 cores y 24 threads c.u.
- Y por ejemplo decidimos correr 24 VMs... o 200 contenedores
- Con una vNIC por contenedor, puenteada a la LAN, son 200x38.016 o más de 7.6 millones de direcciones MAC



Nombre del producto	Estado	Fecha de lanzamiento	Núcleos	Frecuencia de turbo máxima	Frecuencia base del procesador	Caché	TDP
☐ Procesador Intel® Xeon® Platinum 8490H (caché de 112,5 MB, 1,90 GHz)	Launched	Q1'23	60	3.50 GHz	1.90 GHz	112.5 MB	350 W



	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

No puede ser todo un dominio capa 2



Escalabilidad con ECMP

- Conmutadores por ejemplo con soporte para 128-way ECMP
- Podemos cambiar esta topología MLAG por una ECMP (...)

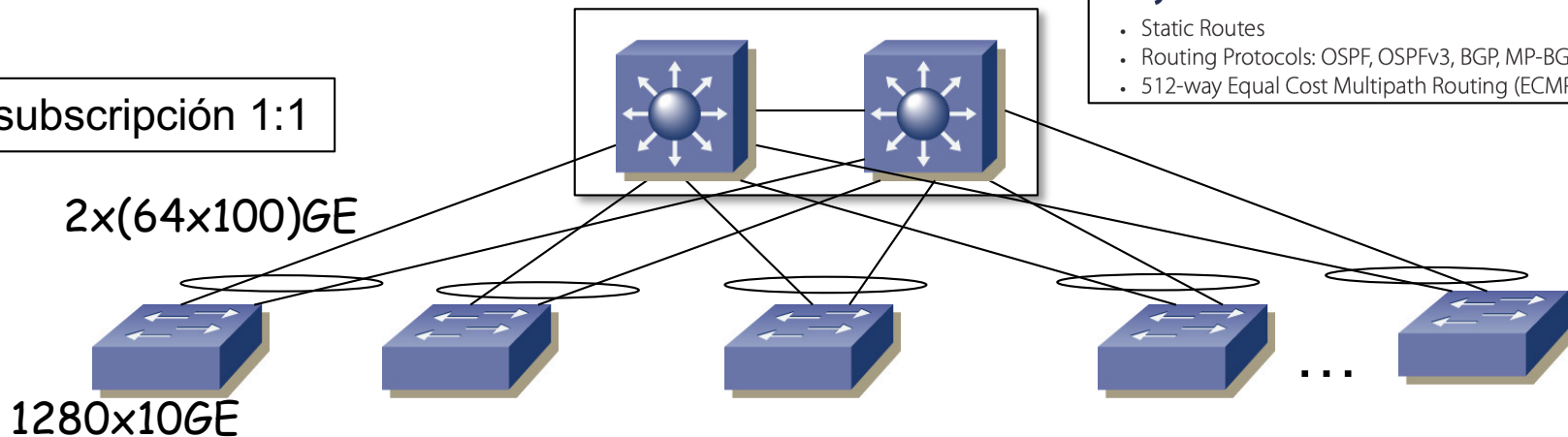
Layer 3 Features

- Static Routes
- Routing Protocols: OSPF, OSPFv3, BGP, MP-BGP, IS-IS, and RIPv2
- 512-way Equal Cost Multipath Routing (ECMP)

Sobre-subscripción 1:1

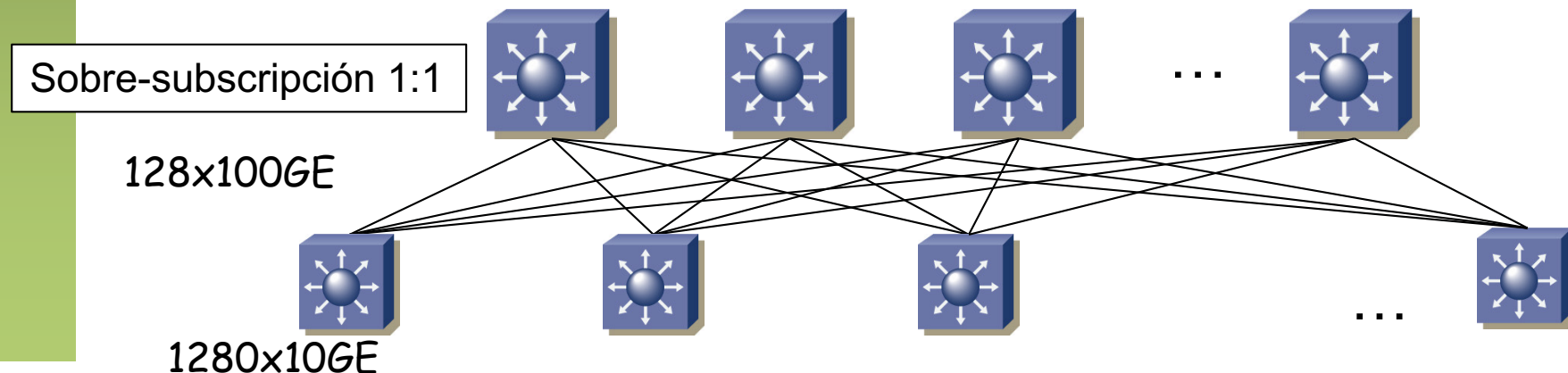
2x(64x100)GE

1280x10GE



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
- (...)

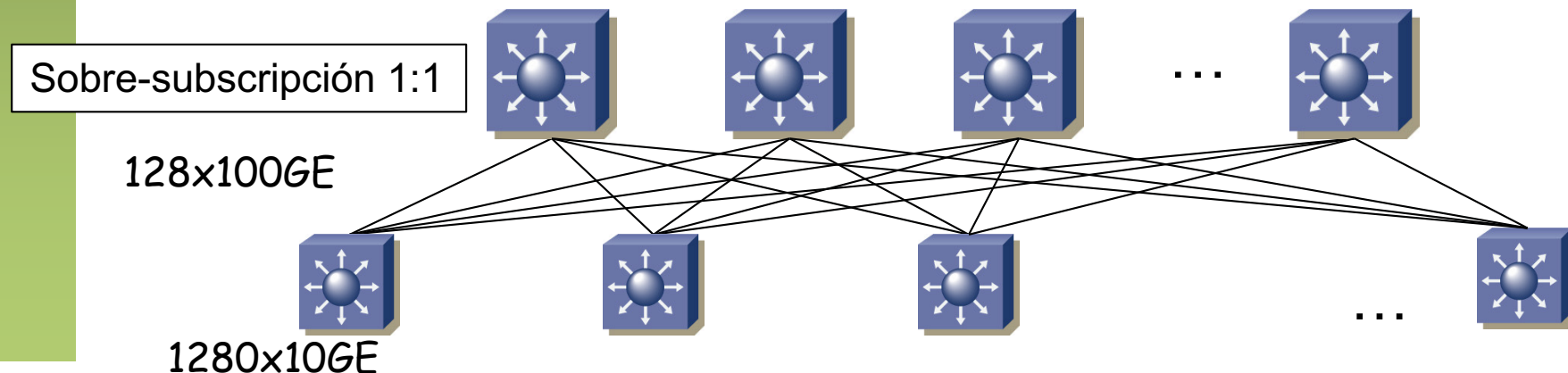


Escalabilidad con ECMP

- 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

<https://www.racksolutions.com/news/blog/how-many-servers-does-a-data-center-have/>



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!)
- (...)

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...

Escalabilidad