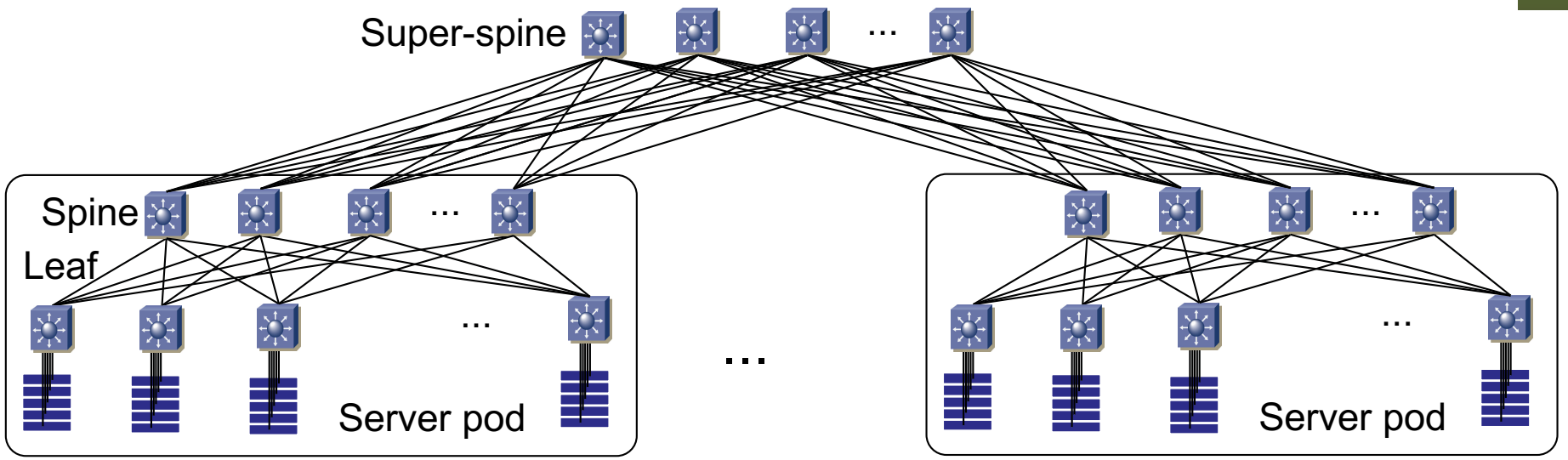
upna
Universidad Pública de Navarra
Nafarroako Unibertsitate Publikoa
Redes de Nueva Generación
Área de Ingeniería Telemática

Nuevas arquitecturas para el data center

3 capas: super-spines

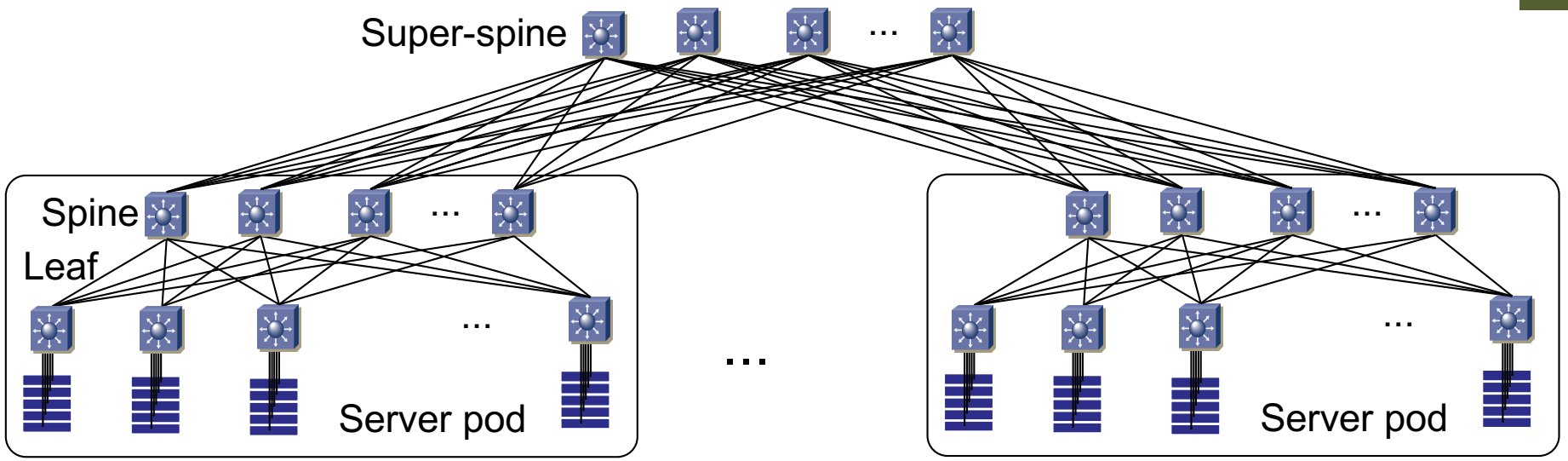
3 capas

- Leaf, Spine, Super-spine switches
- Full mesh entre parejas
- Aumenta número de servidores aumentando server pods
- Límite depende de puertos de super-spines
- Aumentamos BW aumentando número de super-spines



3 capas: Ejemplo

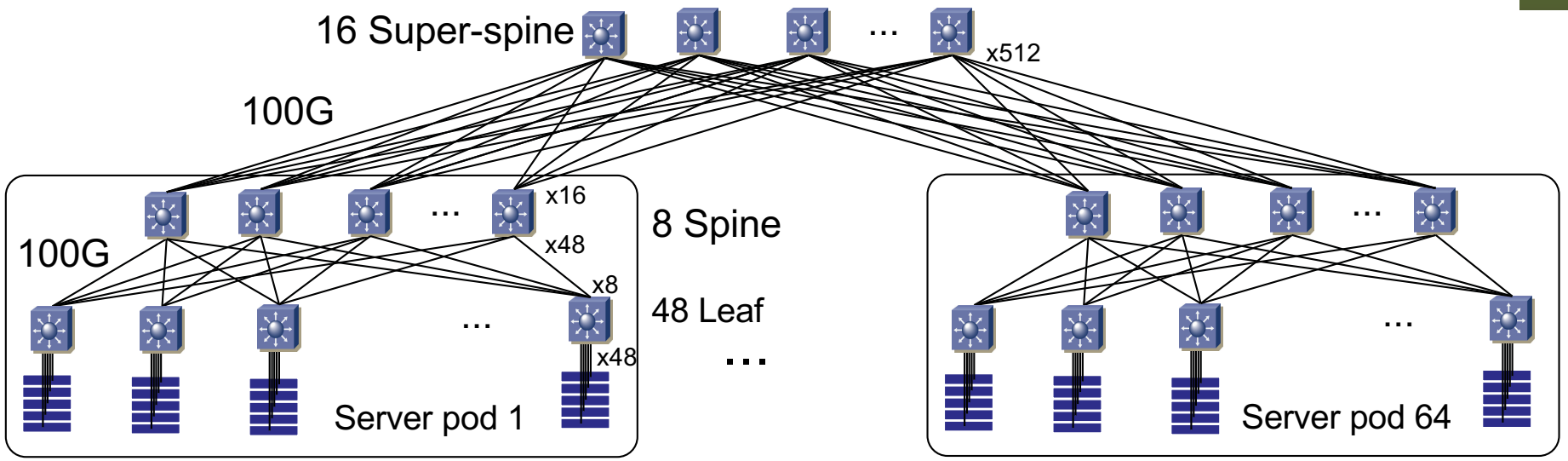
- Leaf: 48 puertos @ 10/25/50G + 16 @ 100G (ó 4 @ 400G)
- Spine: 256 @ 100G (ó 64 @ 400G)
- Super-spine: 1152 @ 100G (ó 288 @ 400G)



3 capas: Ejemplo

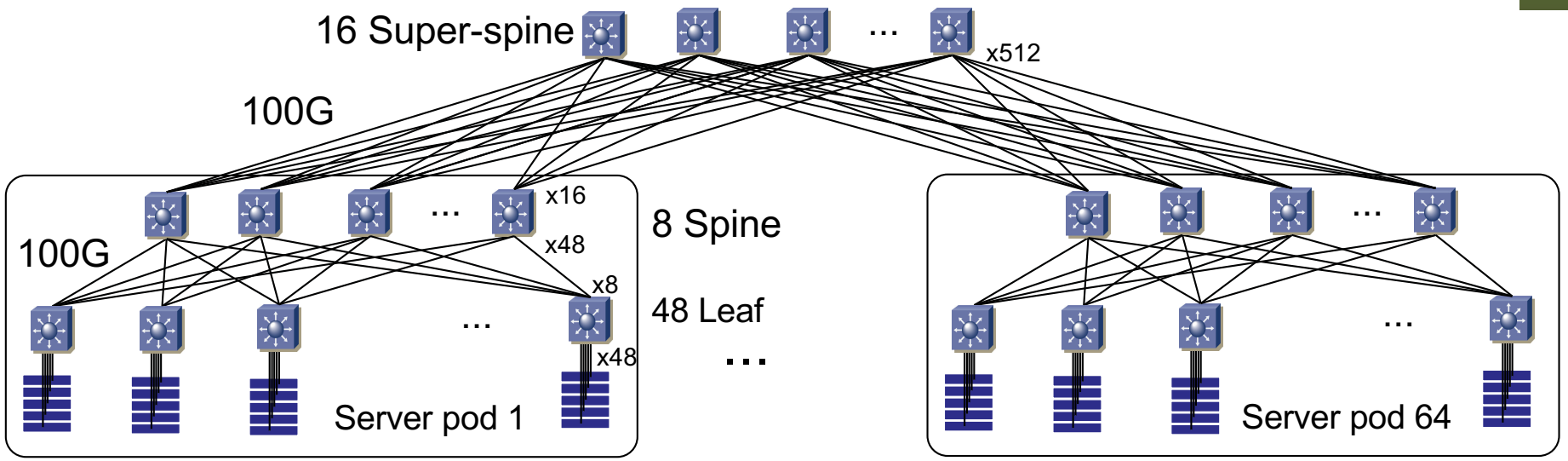
- Leaf: 48 puertos @ 10/25/50G + 16 @ 100G (ó 4 @ 400G)
- Spine: 256 @ 100G (ó 64 @ 400G)
- Super-spine: 1152 @ 100G (ó 288 @ 400G)

- 16 super-spine switches, 100G links a spines
- 64 server pods: cada server pod 8 spine + 48 leaf switches
- Cada super-spine 64x8=512 puertos @ 100G
- Cada spine switch 48 @ 100G a leaf + 16 @ 100G a super-spines

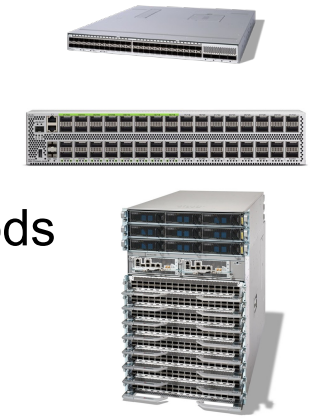


3 capas: Ejemplo

- 64 pods x 48 leaf/pod = 3.072 leaf switches
- 64 pods x 8 spine/pod = 512 spine switches
- 16 super-spines + 512 spines + 3.072 leaf = 3.600 switches
- 64 pods x 48 leaf/pod x 48 puertos/leaf = 147.456 puertos
- Conexiones entre switches: 16x512 + 64x8x48 = 32.768
- Sobre-subscripción leaf: 48x25G : 8x100G = 1.5:1
- Sobre-subscripción spine: 48x100G : 16x100G = 3:1

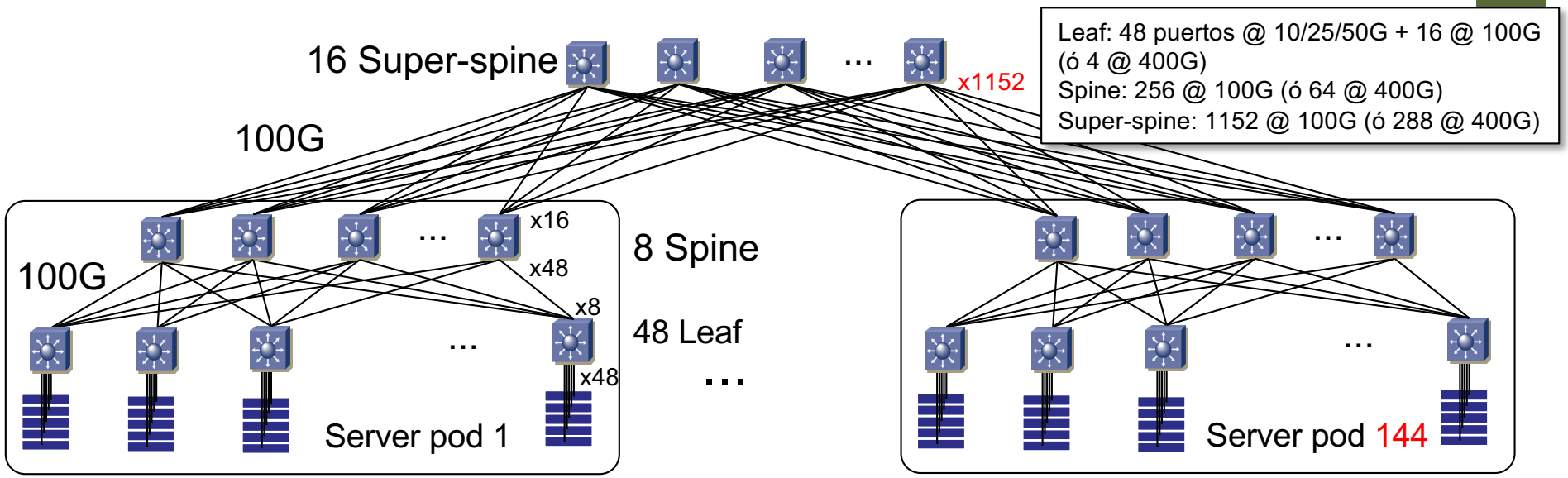


3 capas: Ejemplo



¿Ampliación?

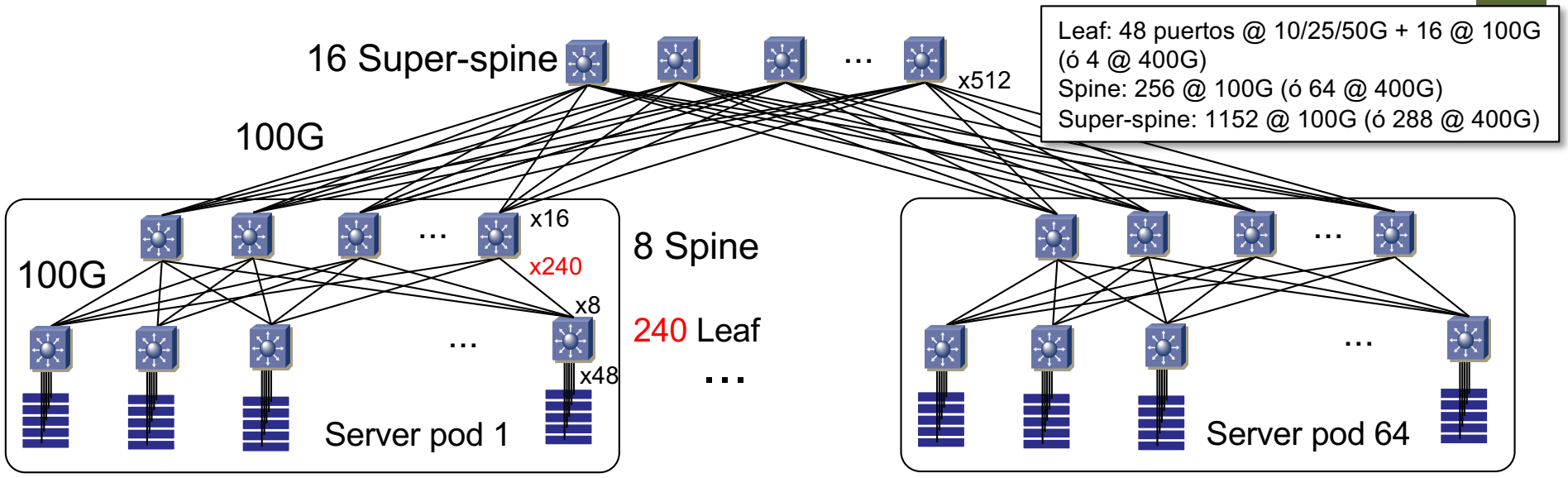
- Número de puertos (opción 1)
- Super-spines hasta 1152 puertos @ 100G → 144 server pods
- 144 pods x 48 leaf/pod x 48 puertos/leaf = 331.776 puertos
- Requiere que los super-spines sean capaces de conmutar:
- 1152x100 = 115.2 Tb/s cada uno
- ¿Bloqueo interno en esos conmutadores?



3 capas: Ejemplo

¿Ampliación?

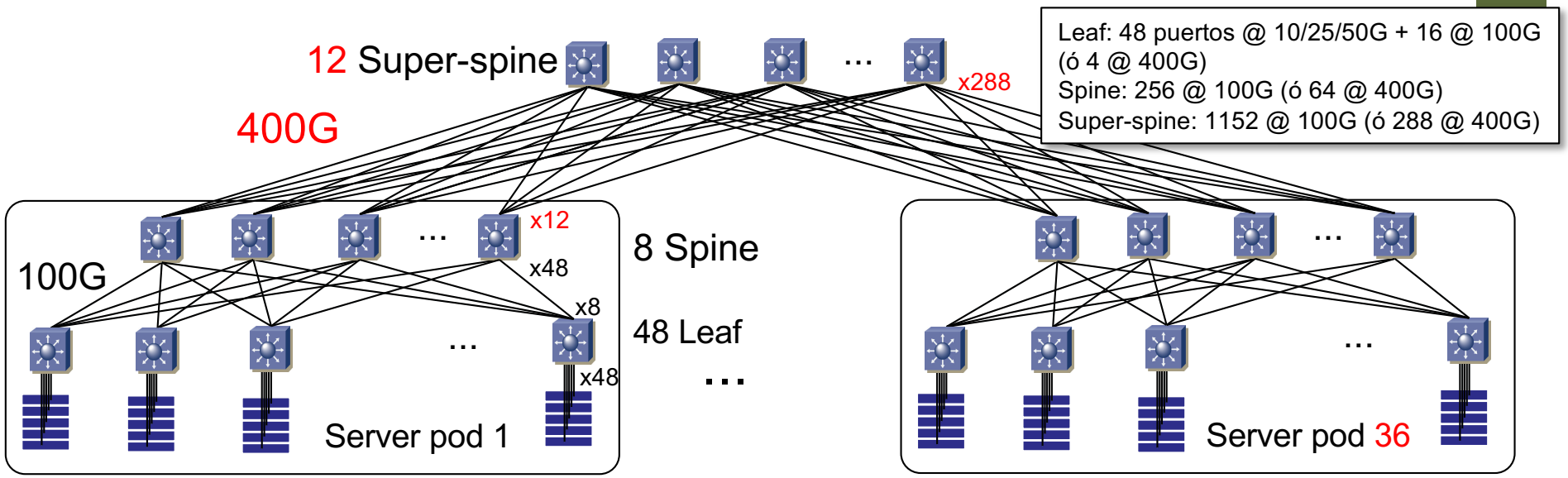
- Número de puertos (opción 2)
- Spine hasta 256 puertos @ 100G
- (256-16=240) leaf switches/pod
- 64 pods x 240 leaf/pod x 48 puertos/leaf = 737.280 puertos
- Sobre-subscripción 240x100G : 16x100G = 15:1



3 capas: Ejemplo

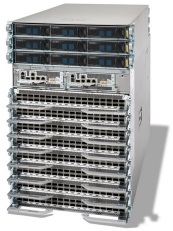
¿Ampliación?

- Reducir sobre-subscripción
- Aumentar enlaces de super-spine a 400G
- A cada spine llegan 48x100G, requiere 12x400G para 1:1
- Cada spine switch es de 64 puertos @ 100/400G
- Usar 12 @ 400G hacia super-spine + 48 @ 100G hacia leaf switches
- Cada super-spine permite 288 puertos @ 400G, 288/4= 36 pods
- 36 pods x 48 leaf/pod x 48 puertos/leaf = 82.944 puertos
- Queda la sobre-subscripción 1.5:1 en los leaf switches



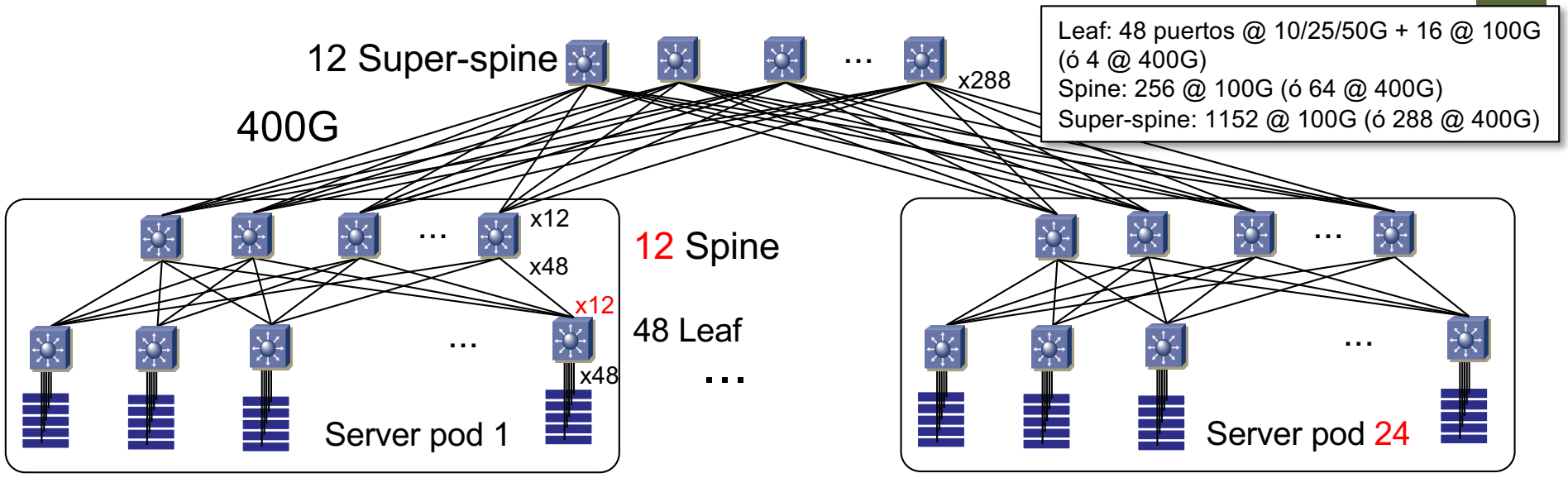
3 capas: Ejemplo





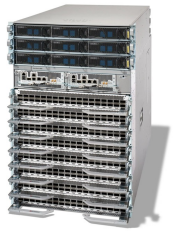
¿Ampliación?

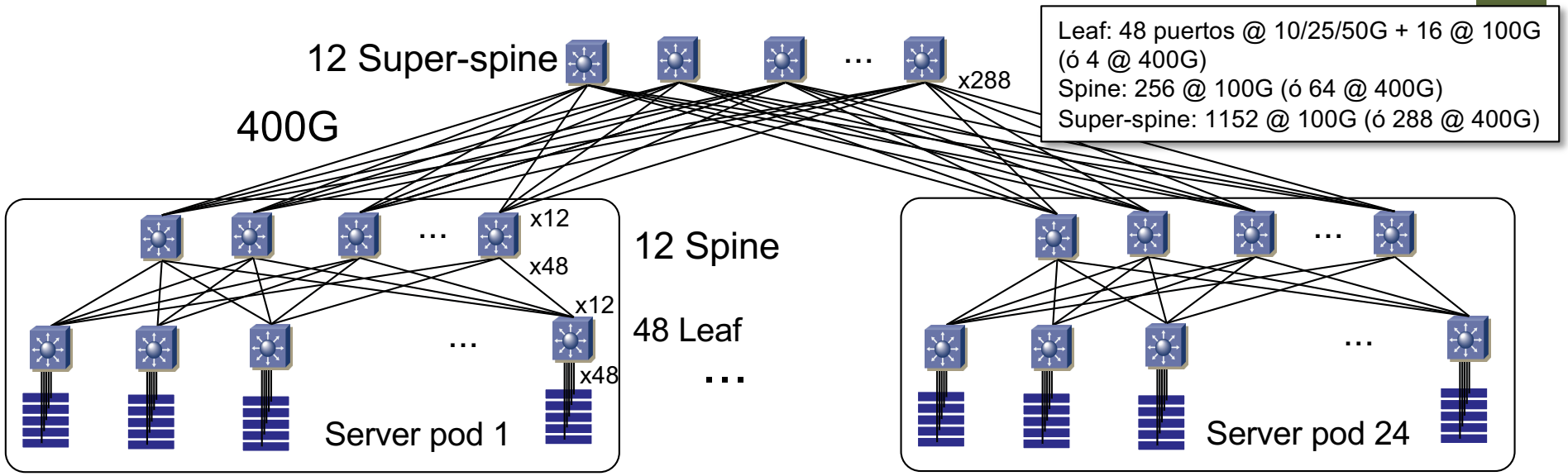
- Reducir sobre-subscripción
- Aumentar enlaces de super-spine a 400G y
- Subir de 8 a 12 uplinks @ 100G en los leaf switches
- Eso es ya 1:1 pues 48x25=12x100
- Requiere 12 spine switches por pod
- 288/12=24 pods debido a la limitación de los super-spine switches
- 24 pods x 48 leaf/pod x 48 puertos/leaf = 55.296 (puertos 25GE)
- Sin sobre-subscripción, no bloqueante



¿Problema?

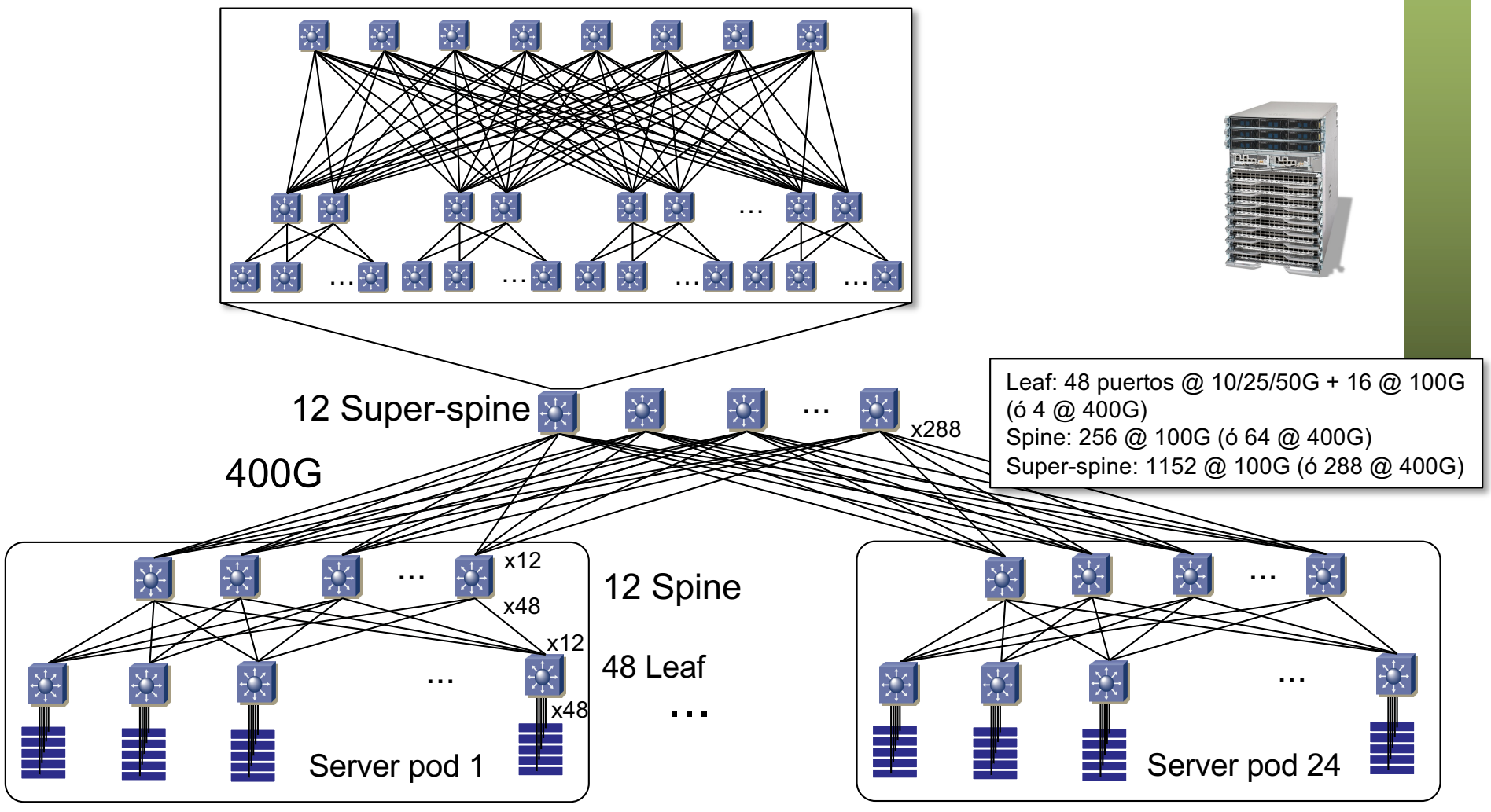
- Un data center requiere decenas de miles de puertos
- Un chip puede ofrecer en el orden de decenas o centenas
- ¿Spine switches modulares con una docena de slots?
- ¿Cómo funcionan internamente si un ASIC no puede conmutar eso?





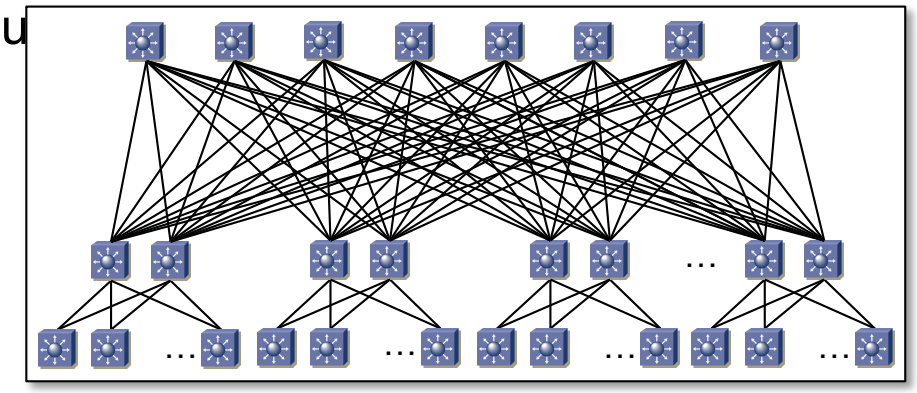
¿Problema?

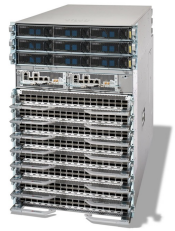
- Conmutadores multi-etapa
- Puede haber sobre-subscripción interna al conmutador



¿Problema?

- Otra alternativa es crear el data center con conmutadores sencillos
- Elementos de mucho más bajo coste
- Tenemos más control sobre la sobre-subscripción en el diseño
- Hay qu

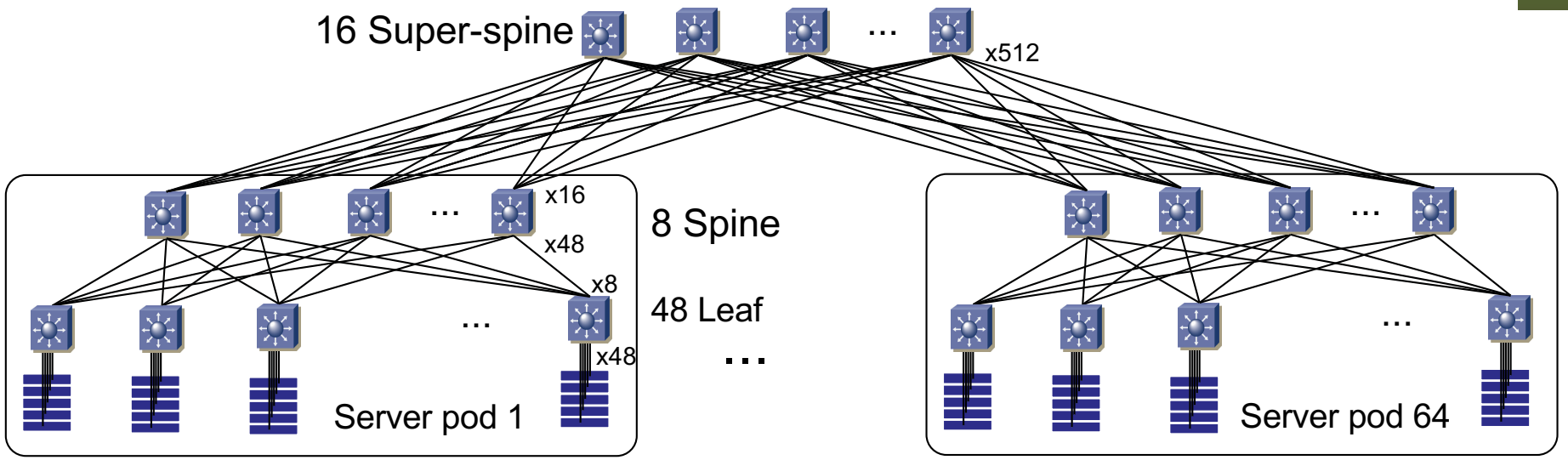




3 capas: planes

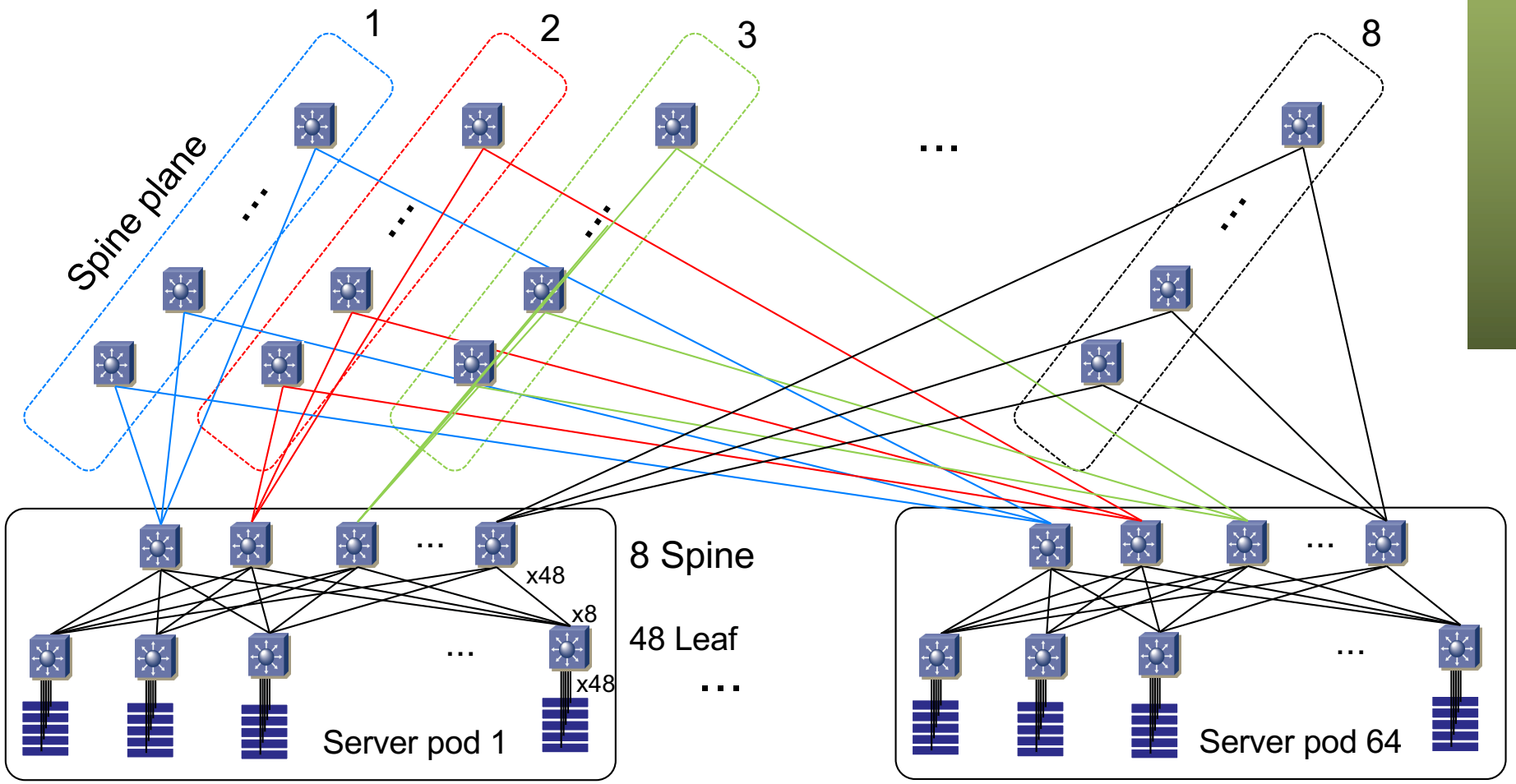
Otros diseños

- Cambiamos los super-spines por equipos más pequeños



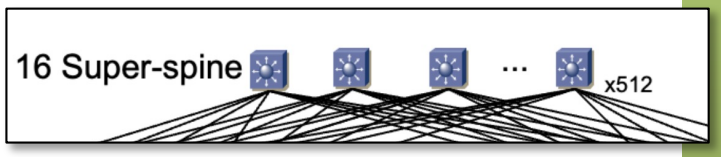
Fabric plane

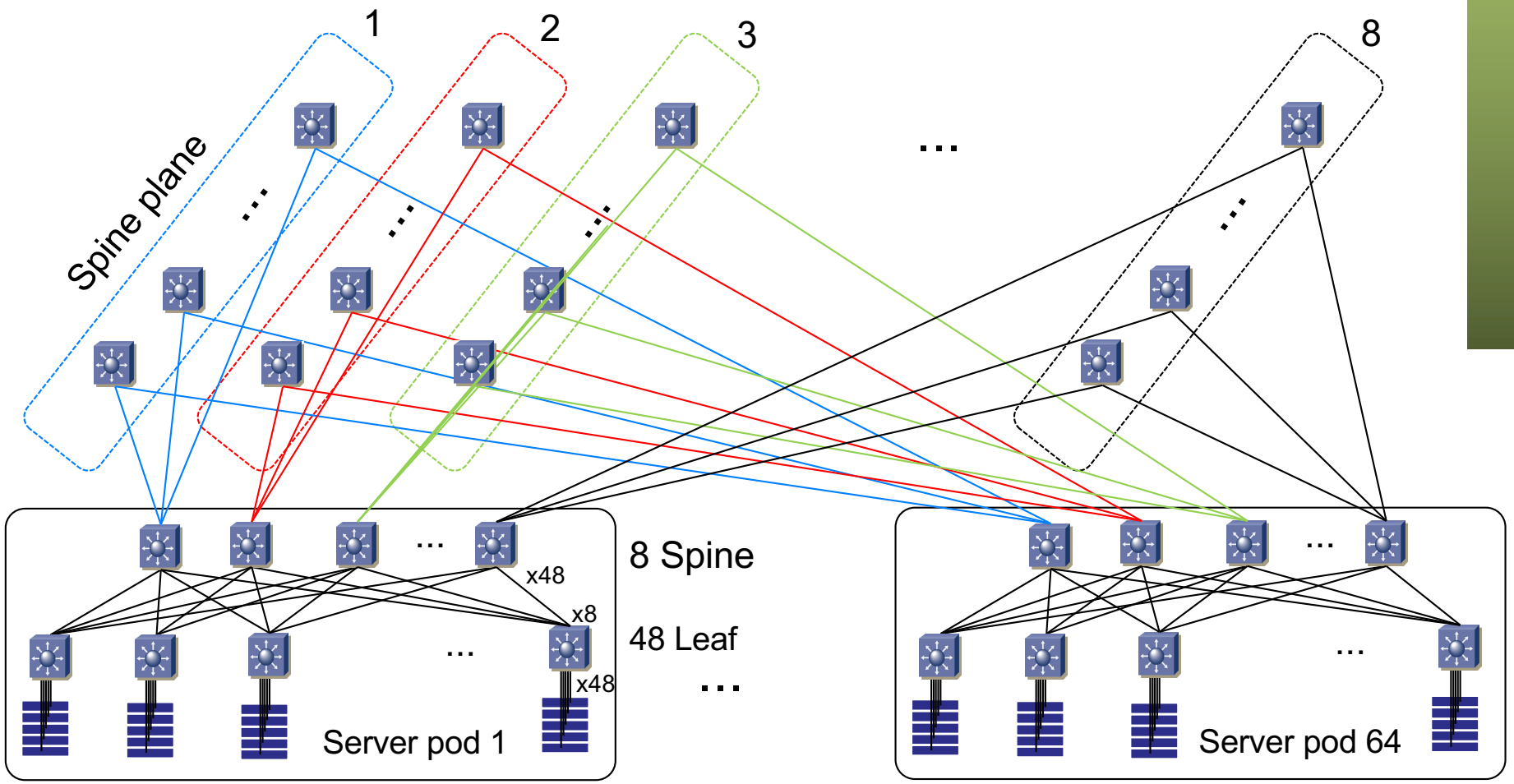
- Ejemplo: 8 spines/pod, cada uno a un “spine plane”
- Tantos planos como spine switches por pod
- Cada switch de un spine plane tiene un puerto por pod, no un puerto por cada spine switch



Fabric plane

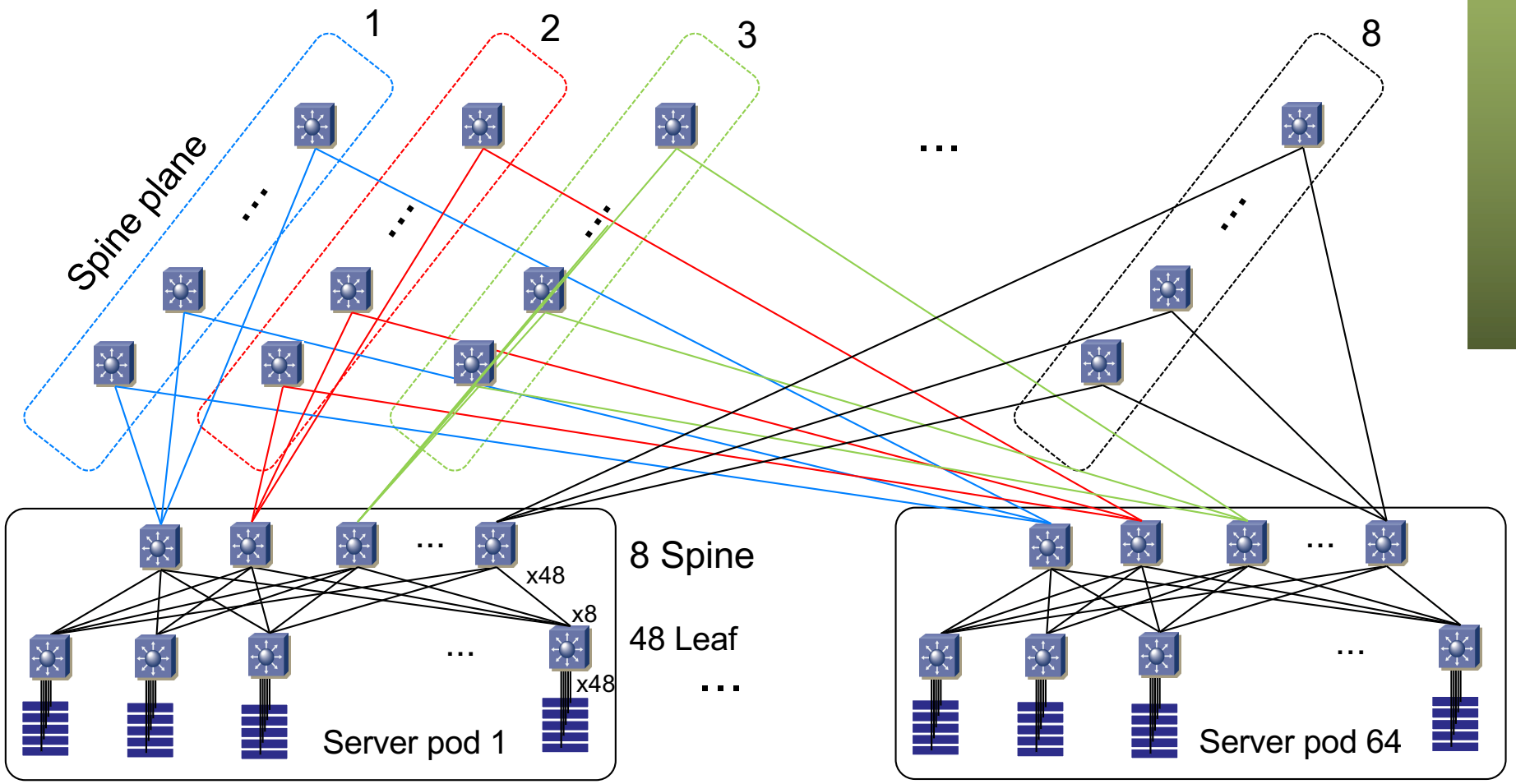
- Para tener 64 pods con 8 spine/pod no necesitamos switches de 64x8=512 puertos sino de solo 64 puertos
- Switches más económicos





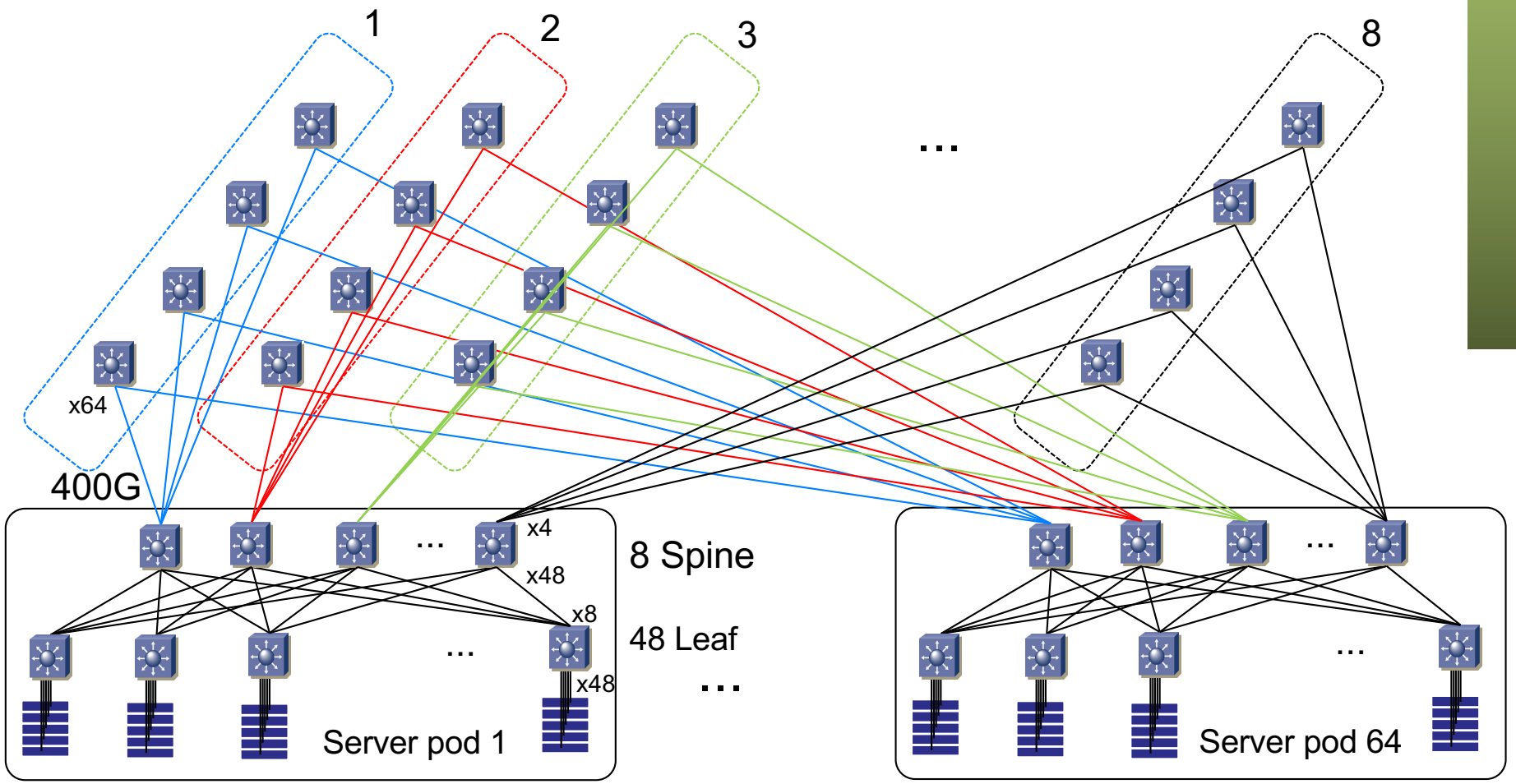
Fabric plane

- Número de switches por plano nos da el BW entre pods y la disponibilidad
- El BW entre dos pods es el nº de switches/plano x BW del enlace



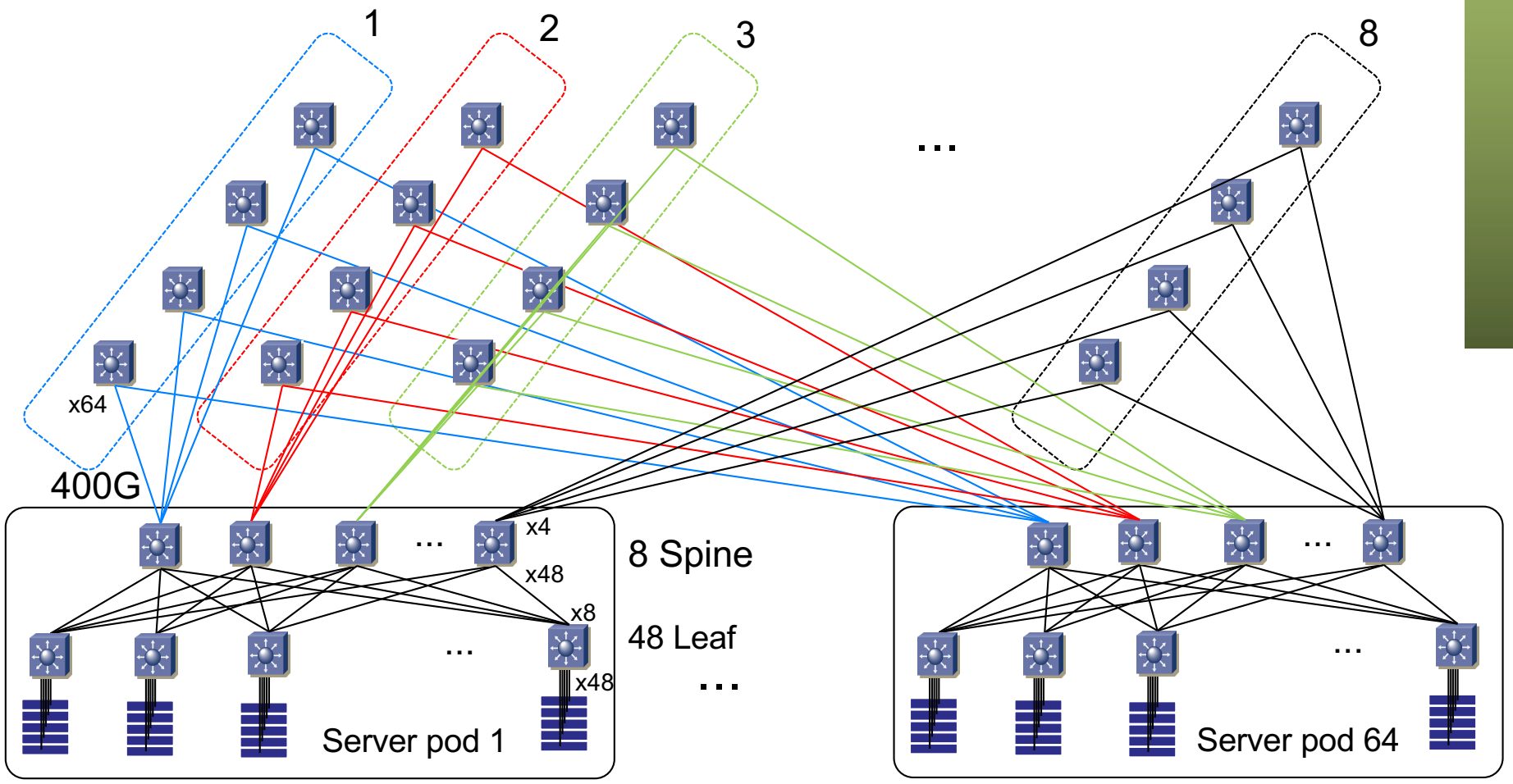
Fabric plane

- Ejemplo: con 4 switches por plano, enlaces @ 400G
- A un Spine llegan 48x100G y su uplink son 4x400G, es un 3:1
- Si subimos a 8 switches por plano tenemos 2:1 y con 16 switches por plano 1:1 (de spine a spine-plane)



Fabric plane

- Para aumentar el número de hosts aumentar el número de pods
- Por cada nuevo pod hace falta un puerto en cada switch de los planos
- Para aumentar el BW entre pods aumentar el número de switches en los planos

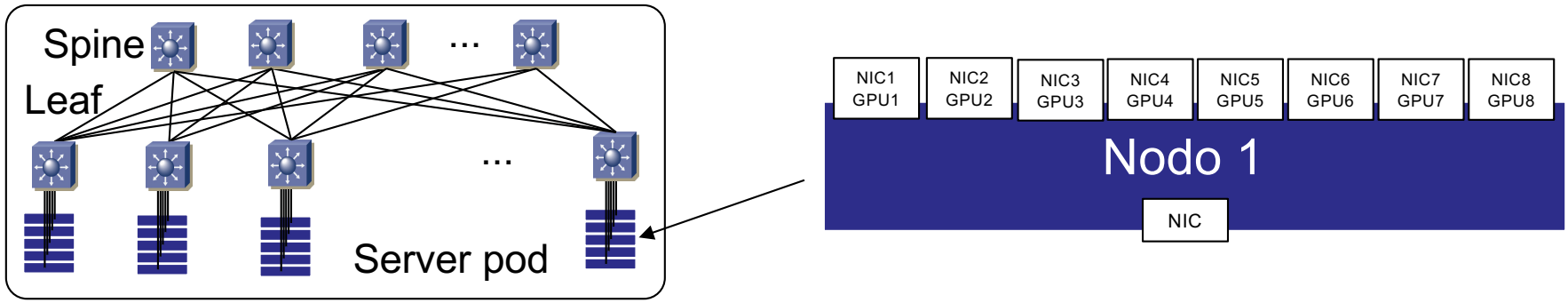


upna
Universidad Pública de Navarra
Nafarroako Unibertsitate Publikoa
Redes de Nueva Generación
Área de Ingeniería Telemática

GPU clusters

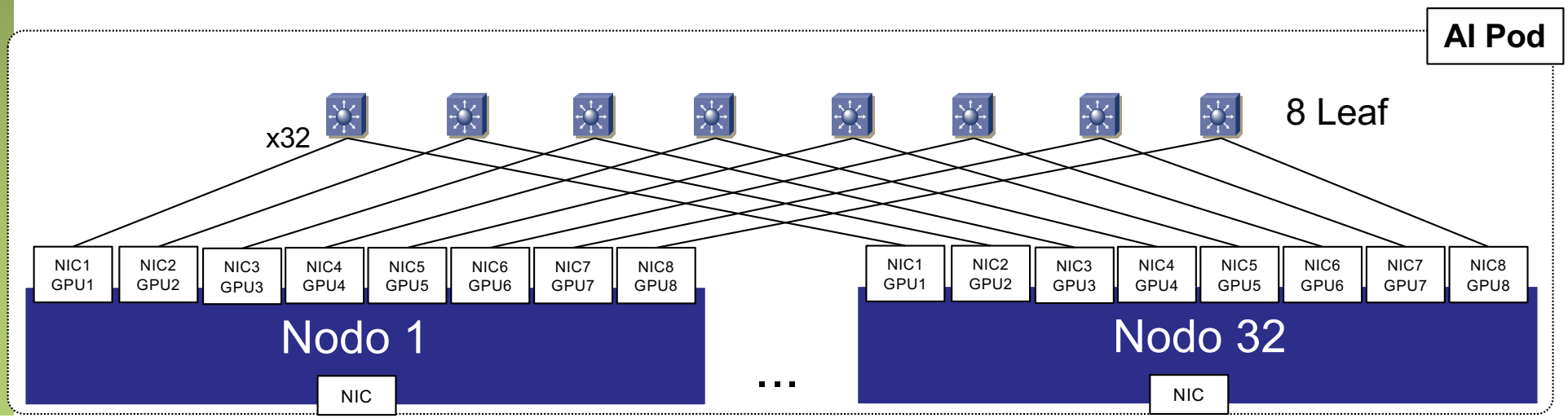
AI workloads

- Lo que hemos descrito sería el interfaz de red “normal” (front end) de los servidores
- Equipos con varias GPUs
- Networking entre GPUs sin pérdidas (back end)
- Alternativas como RoCE (RDMA over Converged Ethernet)
- Arquitectura de red recomendada puede depender de optimizaciones en la librería de comunicación entre las GPUs
- ¿Front-end y back-end mismos switches?



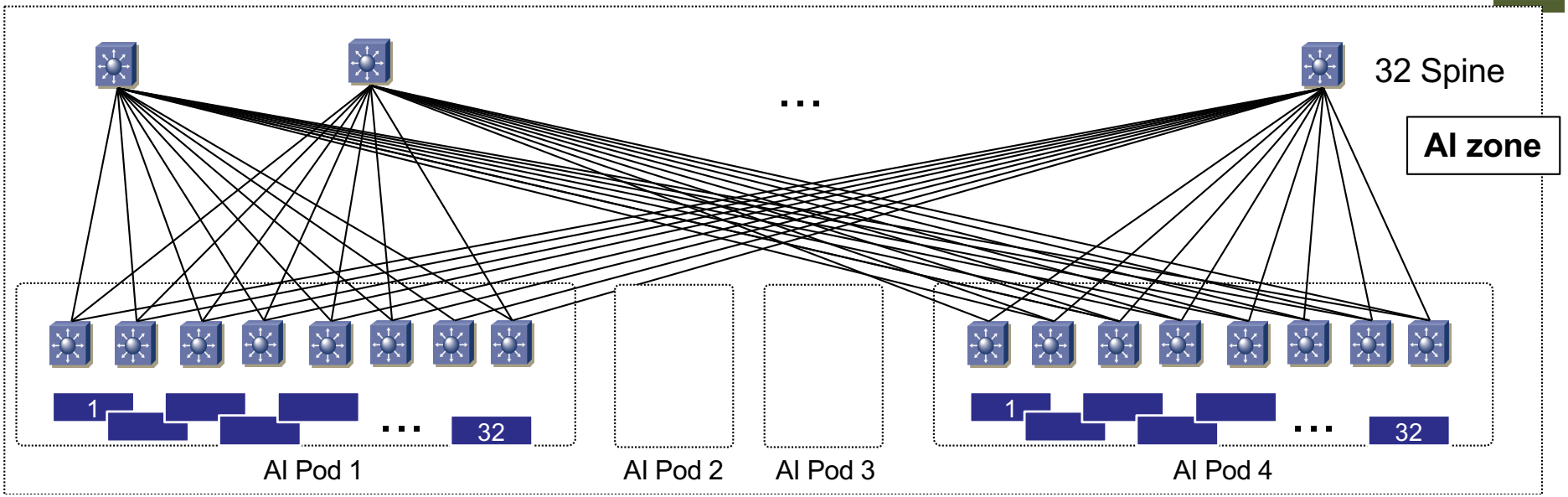
Ejemplo: AI Pod

- 32 nodos/pod x 8 GPUs/nodo = 256 GPUs/pod
- Tantos leaf switches como GPUs por nodo (8 leaf switches)
- GPU links @ 400G
- Cada leaf switch interconecta una GPU de cada nodo
- Leaf switch n interconecta las GPUn
- Cada leaf switch recibe 32 x 400G de los nodos
- GPUn se comunica con GPUm a través de spine switches



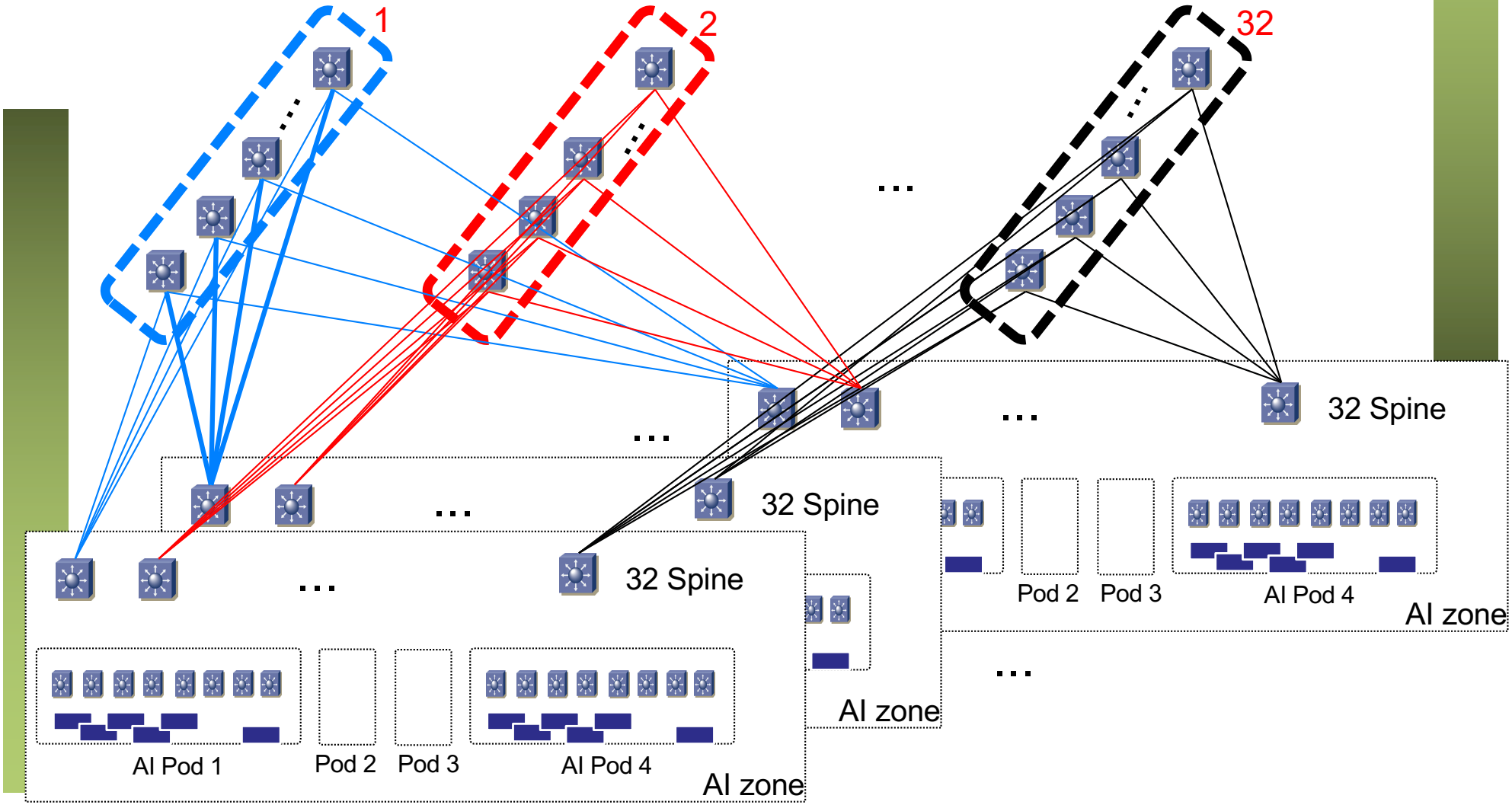
Ejemplo: AI Zone

- 4 AI Pods
- 4 pods/zone x 256 GPUs/pod = 1024 GPUs/zone
- Interconexión con 32 spine switches (tantos como total de leaf)
- Cada leaf switch a cada spine switch
- 8 switches/pod x 4 pods = 32 puertos en cada spine
- Leaf switch:
  - 32 puertos @ 400G hacia 32 GPUs (una por nodo)
  - 32 puertos @ 400G hacia spines (1:1)
  - Leaf switch de 64 puertos @ 400G



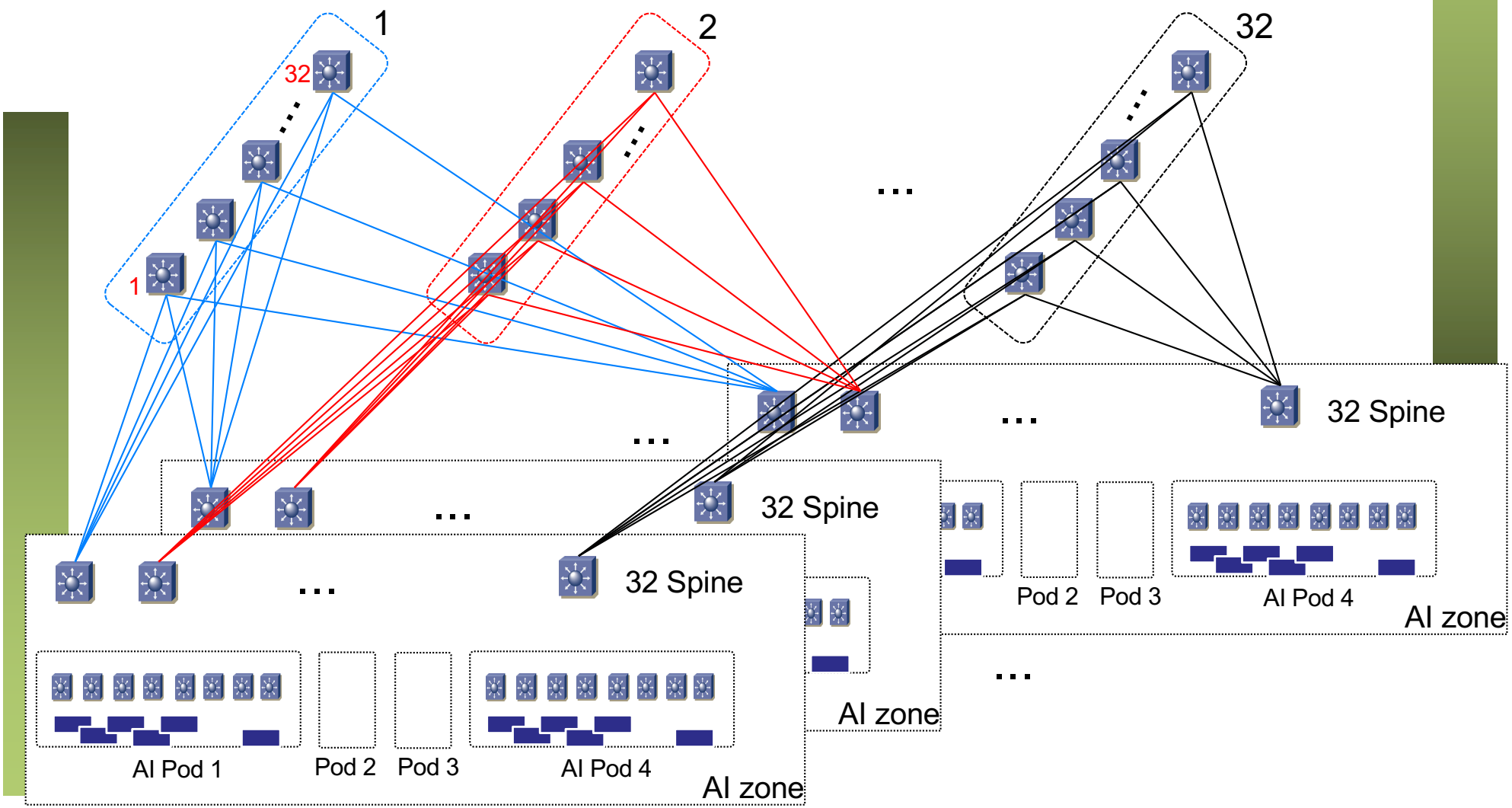
Ejemplo: Interconexión

- Mediante 32 spine planes
- Cada spine switch a todos los switches del mismo spine plane



Ejemplo: Interconexión

- A cada spine llegan 32 x 400G de los leaf
- Ponemos 32 switches en cada plane, puertos 400G (1:1)



Ejemplo: Escala

- Spine plane: switches con tantos puertos como AI zones
- Por ejemplo, switches 64 puertos @ 400G → 64 AI zones
- 64 zones x 1024 GPUs/zone = 65.536 GPUs