

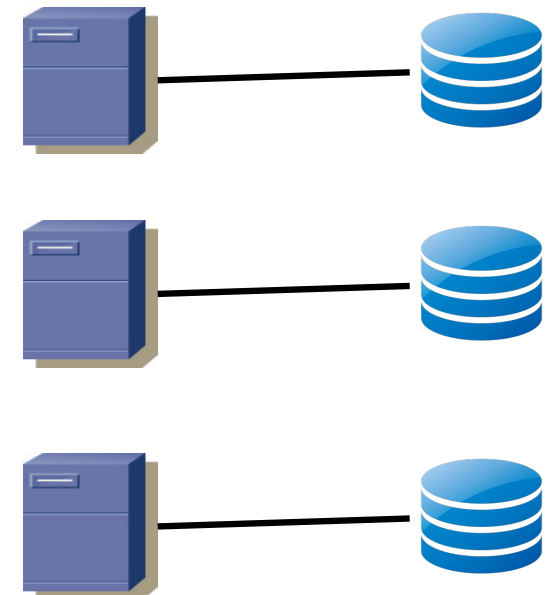
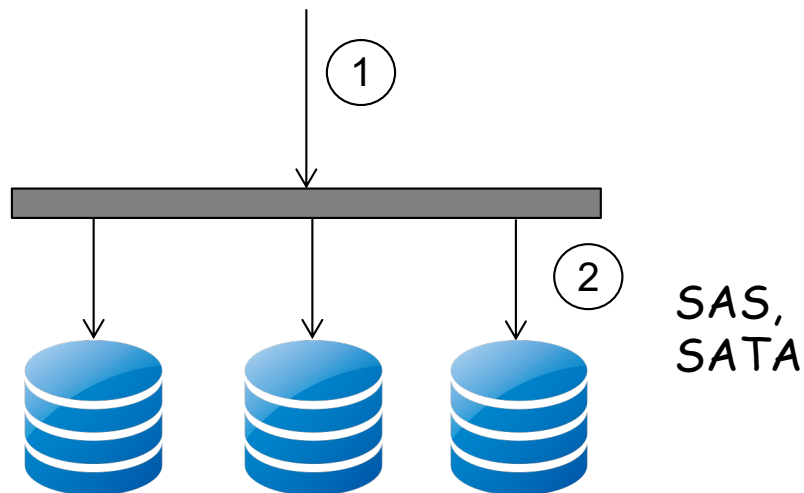


# SAN



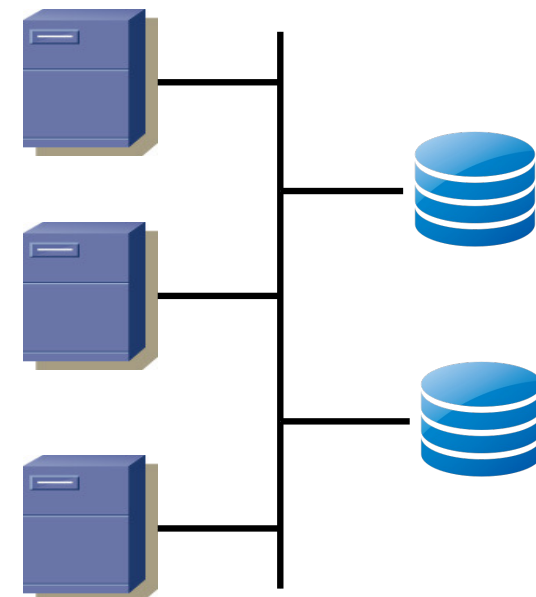
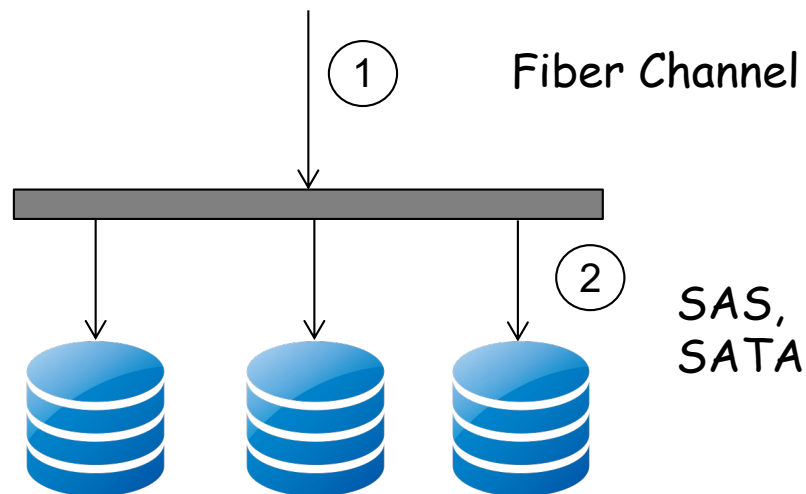
# Acceso al almacenamiento

- El acceso al almacenamiento puede ser directo
  - *Direct Attached Storage (DAS)*
  - En ese caso cada servidor necesita su sistema de almacenamiento
  - Estos sistemas de almacenamiento externos están infrautilizados
  - Pueden ser por ejemplo para llevar a cabo *backups*
  - Un backup nocturno significa que el resto del día esos discos están inutilizados



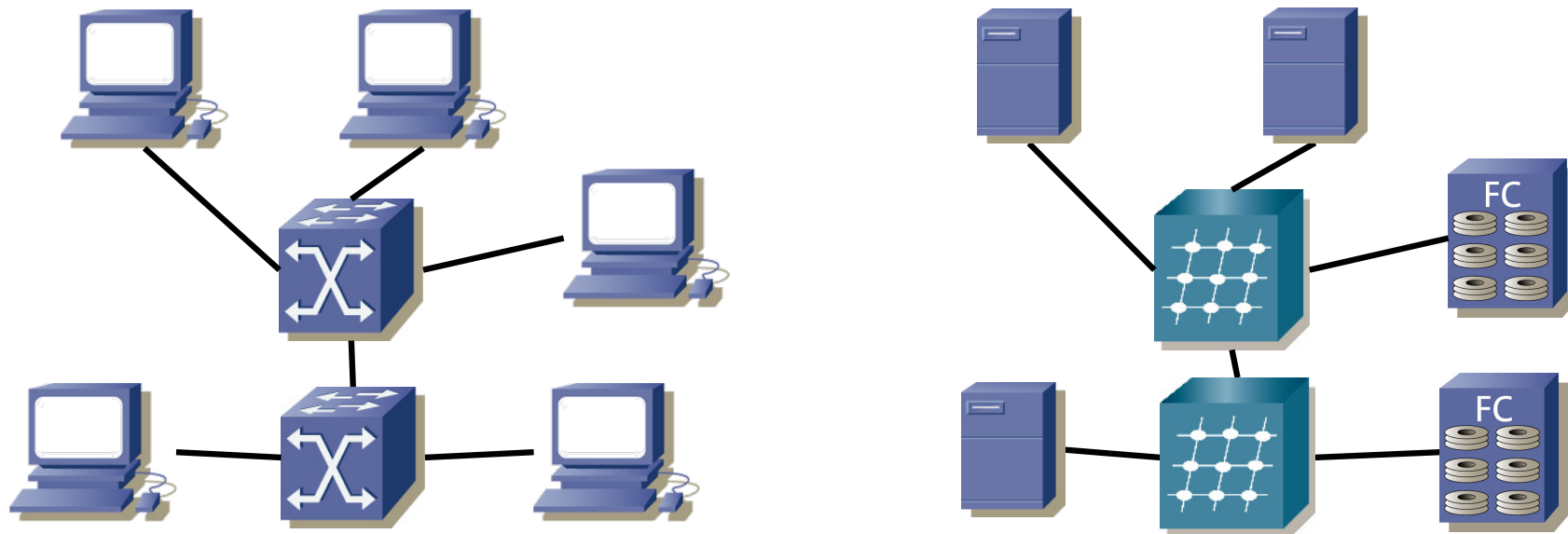
# Acceso al almacenamiento

- El acceso al almacenamiento puede ser directo
- O puede ser a través de una red
  - Hablamos en ese caso de una *Storage Area Network (SAN)*
  - O de *Network Attached Storage (NAS)*
  - Estamos hablando del *front-end* de acceso a los discos
  - El *back-end* es común que sean discos SAS o SATA



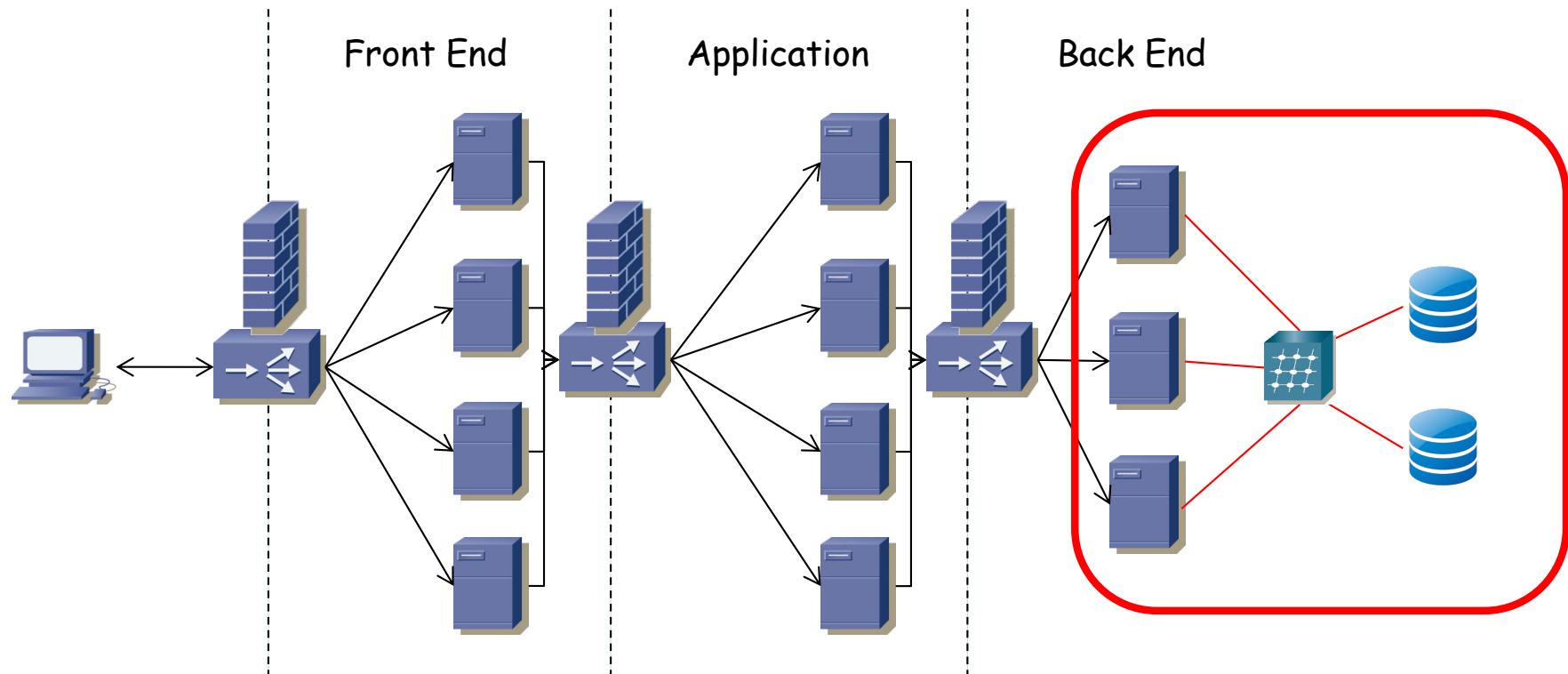
# ¿ Dónde encaja la SAN ?

- Estamos hablando de una tecnología de red
- Cuenta con sus propios conmutadores
- Se dice que forman un *fabric*
- Cuenta con su propia pila de protocolos
- El interfaz en el host es el *Host Bus Adapter (HBA)* que sería el equivalente a la NIC en una LAN



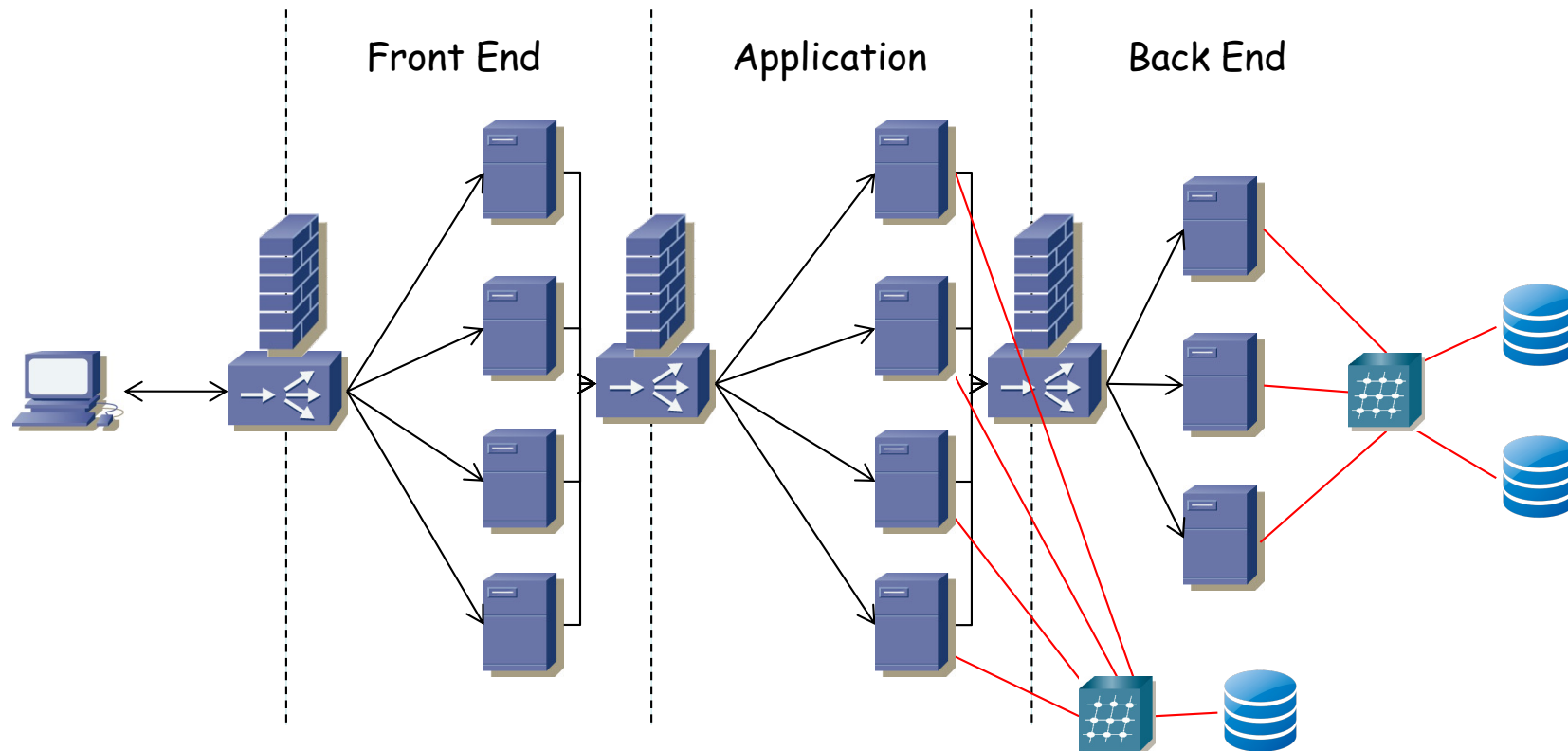
# Es decir...

- Y recordemos que lo estamos poniendo en el backend del servicio porque es donde suelen estar los datos
- Pero nada impide que los servidores de cualquier otra capa...



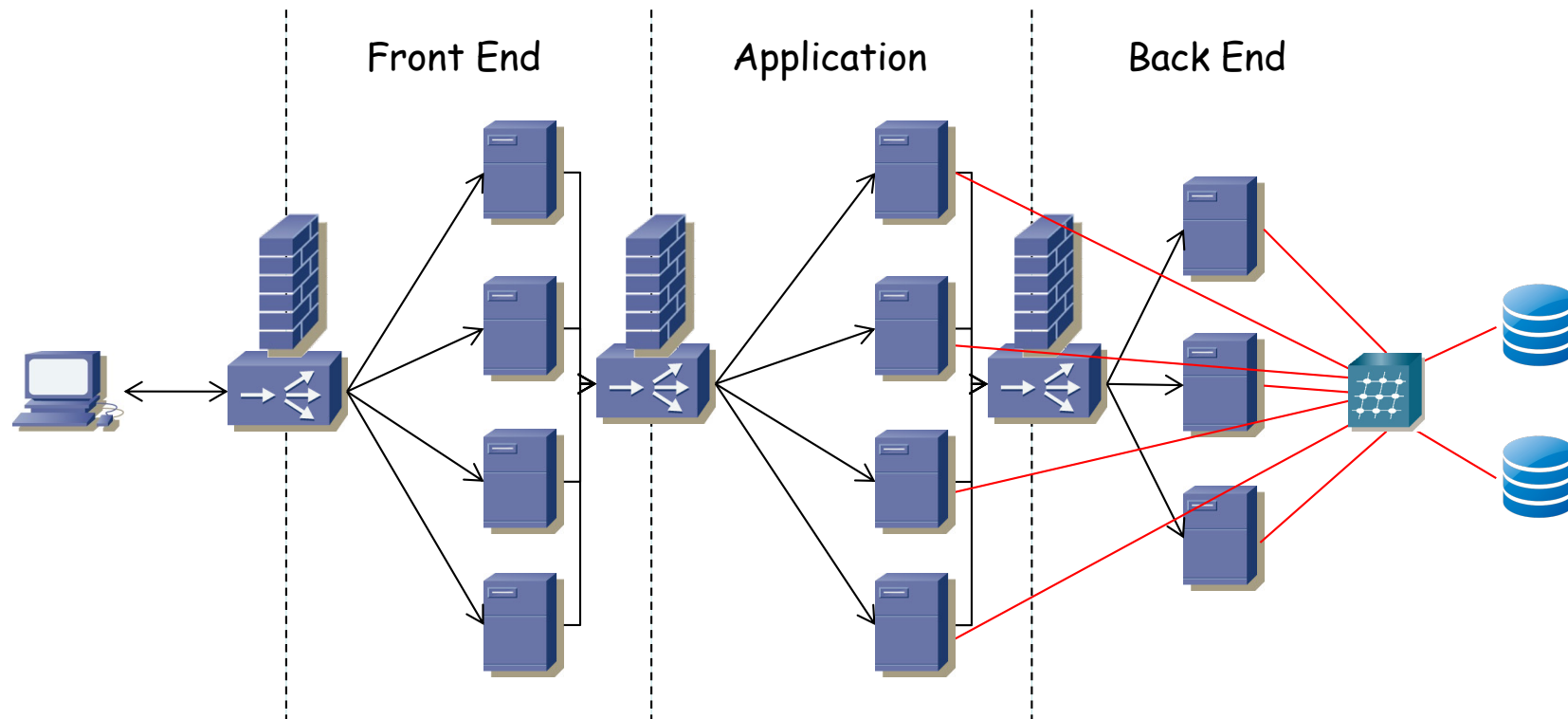
# Es decir...

- Y recordemos que lo estamos poniendo en el backend del servicio porque es donde suelen estar los datos
- Pero nada impide que los servidores de cualquier otra capa empleen almacenamiento SAN
- O incluso en la misma SAN (...)



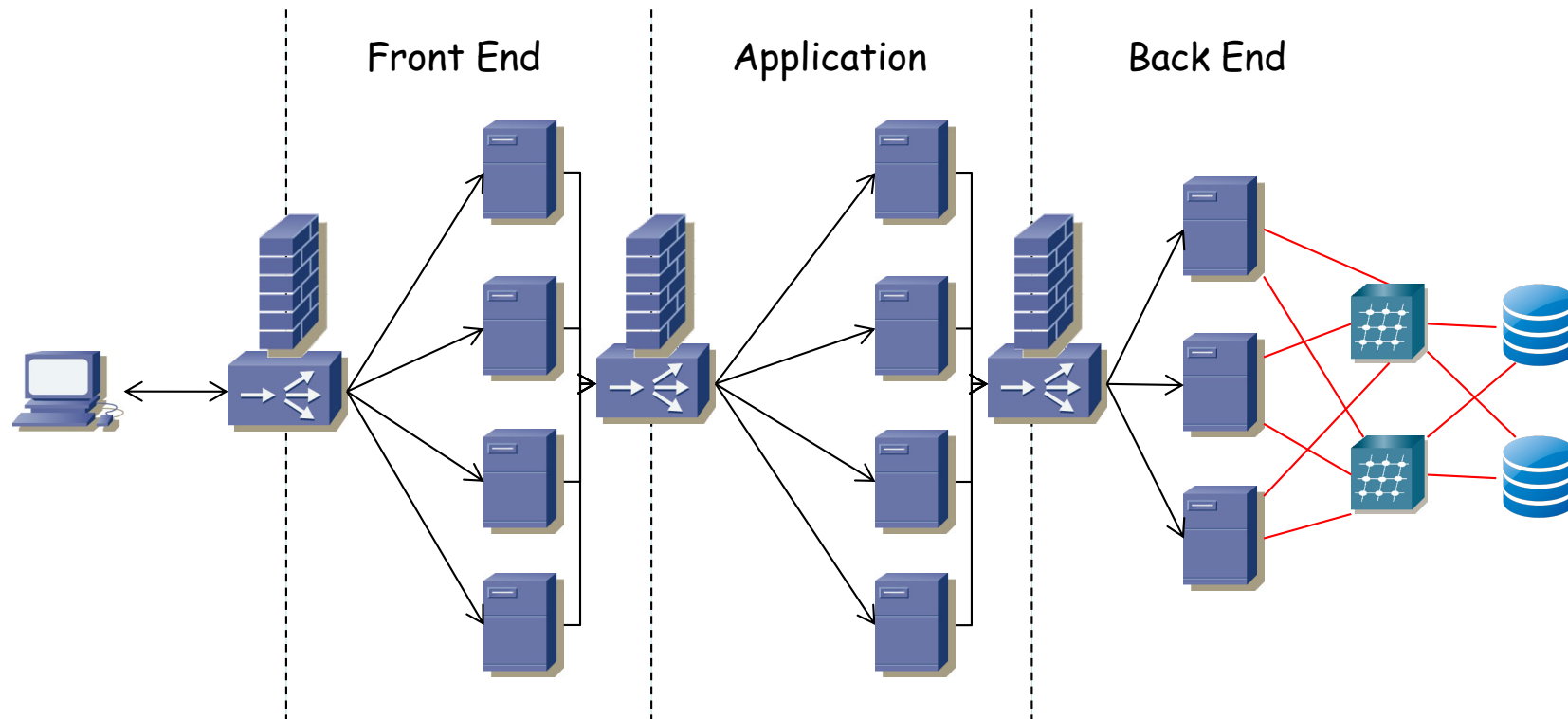
# Es decir...

- Y recordemos que lo estamos poniendo en el backend del servicio porque es donde suelen estar los datos
- Pero nada impide que los servidores de cualquier otra capa empleen almacenamiento SAN
- O incluso en la misma SAN
- ¡ Incluso en los mismos discos !



# High Availability

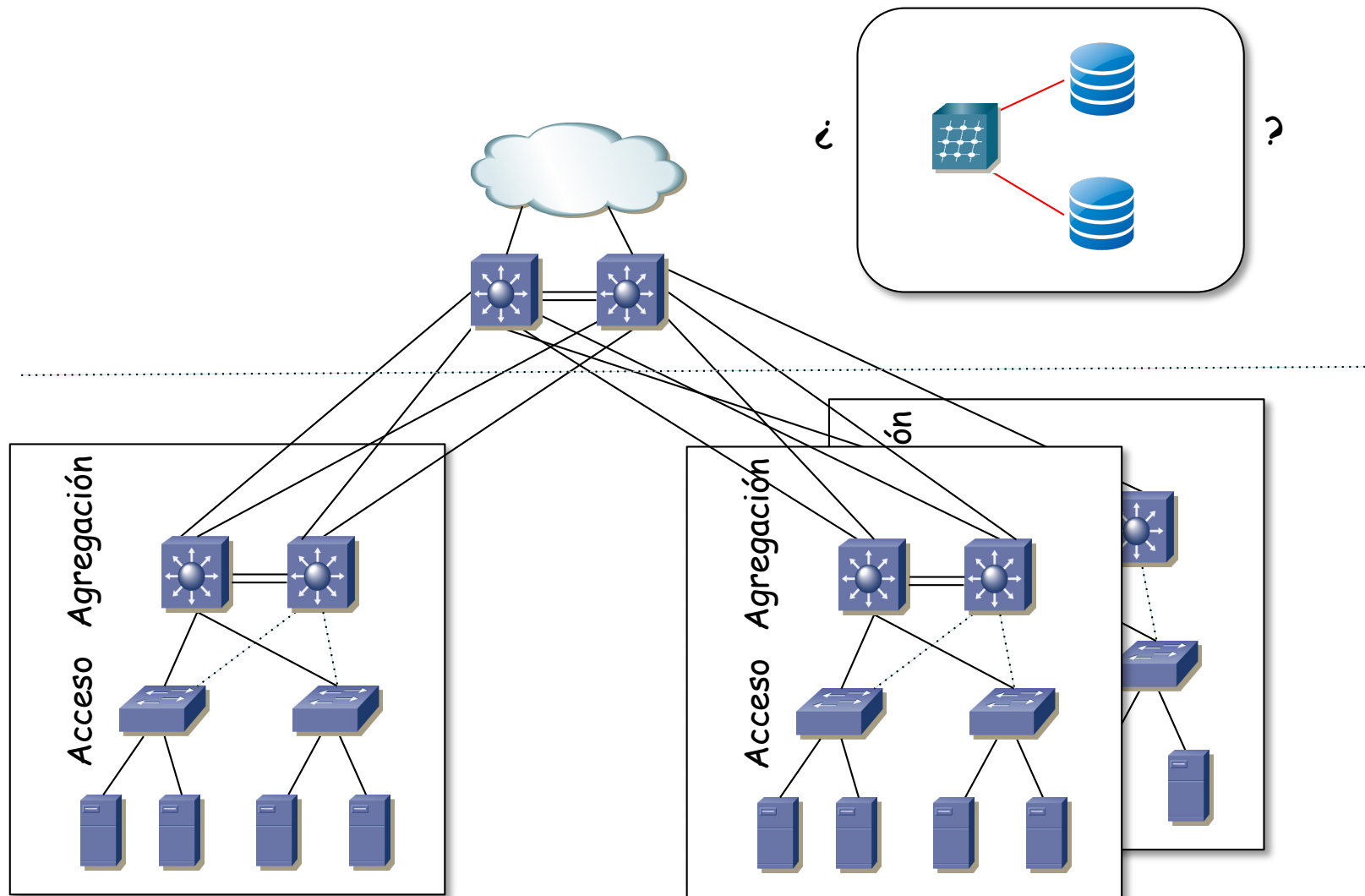
- Y no hay que olvidarse de la redundancia
- (Retiro los interfaces de los servidores de la capa de aplicación por claridad)





# Diseño

- ¿Dónde encaja esa(s) SAN(s) en la red del data center?
- Es independiente, pero hablaremos más sobre esto





# SAN

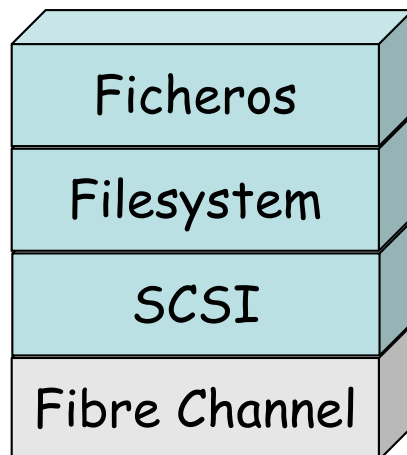


# Fibre Channel

# FC: Arquitectura

# Fibre Channel

- Desarrollo comenzado a finales de los 80s
- Lo normal es que sea sobre fibra pero puede ser sobre cobre
- “*Fibre*” es el *spelling* británico en lugar del americano para “fibra”
- Soluciona el transporte pero no fija lo que transporta
- Conmutación de paquetes
- Así, puede transportar comandos SCSI pero también IP o ATM
- Inicialmente una de sus ventajas era su alta velocidad en comparación con las tecnologías LAN de la época



# Fibre Channel

- Para los nodos el comité T11 del INCITS tiene estandarizado: 1GFC, 2GFC, 4GFC, 8GFC, 16 GFC, 32GFC y 128GFCp (que son 128Gbps mediante 4 canales)
- Para la conexión entre conmutadores tiene 10GFC, 20GFC, 40GFCoE (FC over Ethernet), 100GFCoE y 128GFCp

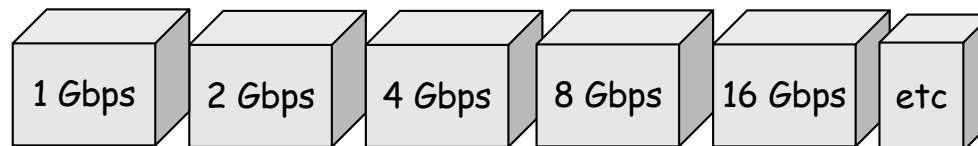
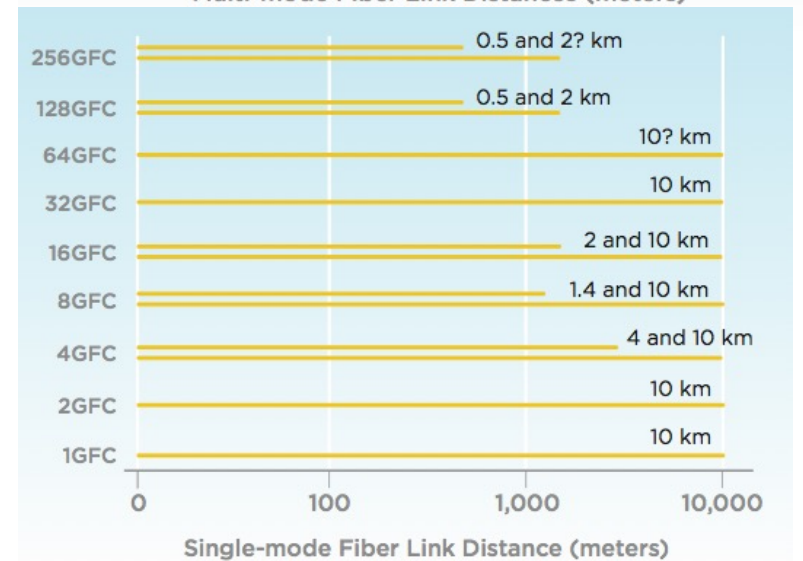
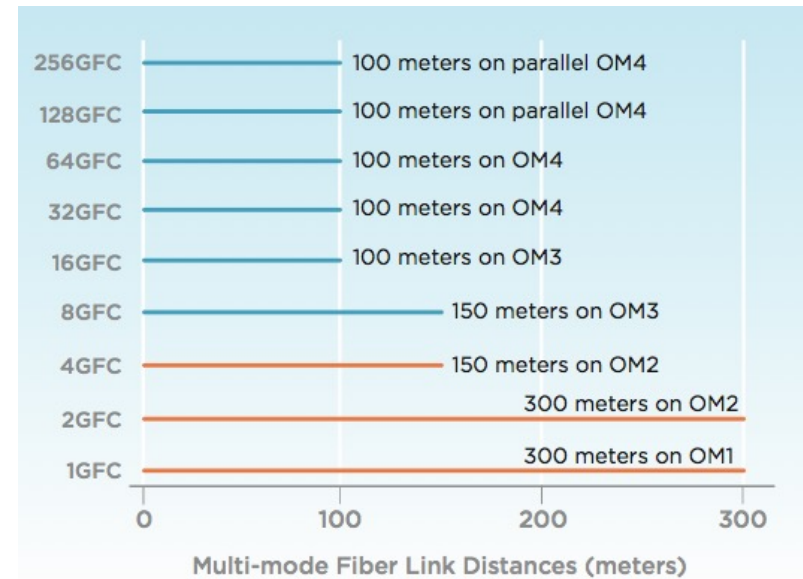
FC

| Product Naming | Throughput (Mbytes/s)* | Line Rate (Gbaud) | T11 Specification Technically Complete (Year) † | Market Availability (Year) † |
|----------------|------------------------|-------------------|-------------------------------------------------|------------------------------|
| 8GFC           | 1,600                  | 8.5 NRZ           | 2006                                            | 2008                         |
| 16GFC          | 3,200                  | 14.025 NRZ        | 2009                                            | 2011                         |
| 32GFC          | 6,400                  | 28.05 NRZ         | 2013                                            | 2016                         |
| 64GFC          | 12,800                 | 28.9 PAM-4        | 2017                                            | 2020                         |
| 128GFC         | 24,850                 | 56.1 PAM-4        | 2021                                            | 2024                         |
| 256GFC         | TBD                    | TBD               | 2025                                            | Market Demand                |
| 512GFC         | TBD                    | TBD               | 2029                                            | Market Demand                |
| 1TFC           | TBD                    | TBD               | 2033                                            | Market Demand                |

ISL  
(Inter-Switch Link)

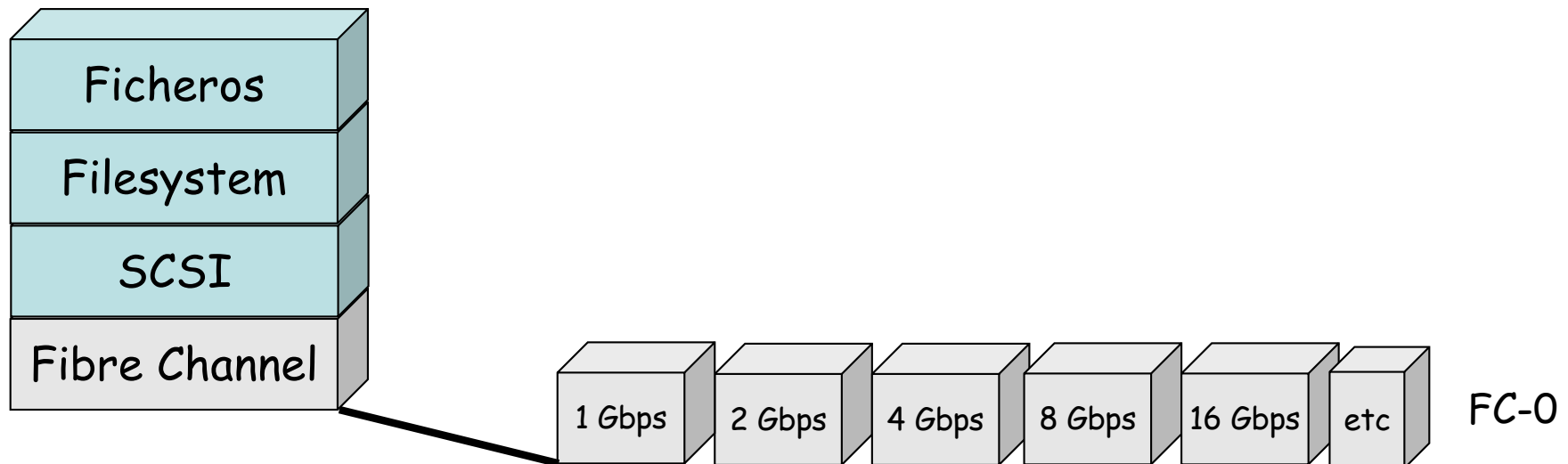
| Product Naming | Throughput (MBytes/s)* | Line Rate (Gbaud)† | Standard Technically Complete (Year)‡§ | Market Availability (Year)‡ |
|----------------|------------------------|--------------------|----------------------------------------|-----------------------------|
| 10GFC          | 2,400                  | 10.52 NRZ          | 2003                                   | 2009                        |
| 40GFCoE        | 9,600                  | 4X10.3125 NRZ      | 2010                                   | 2013                        |
| 100GFCoE       | 24,000                 | 4X25.78125 NRZ     | 2010                                   | 2017                        |
| 128GFC         | 25,600                 | 4X28.05 NRZ        | 2014                                   | 2016                        |
| 200GFCoE       | 48,000                 | 4X26.5625 PAM-4    | 2018                                   | 2020                        |
| 256GFC         | 51,200                 | 4X28.9 PAM-4       | 2018                                   | 2020                        |
| 400GFCoE       | 96,000                 | 8X26.5625 PAM-4    | 2020                                   | Market Demand               |
| 1TFCoE         | TBD                    | TBD                | TBD                                    | Market Demand               |

# Fibre Channel



# Fibre Channel

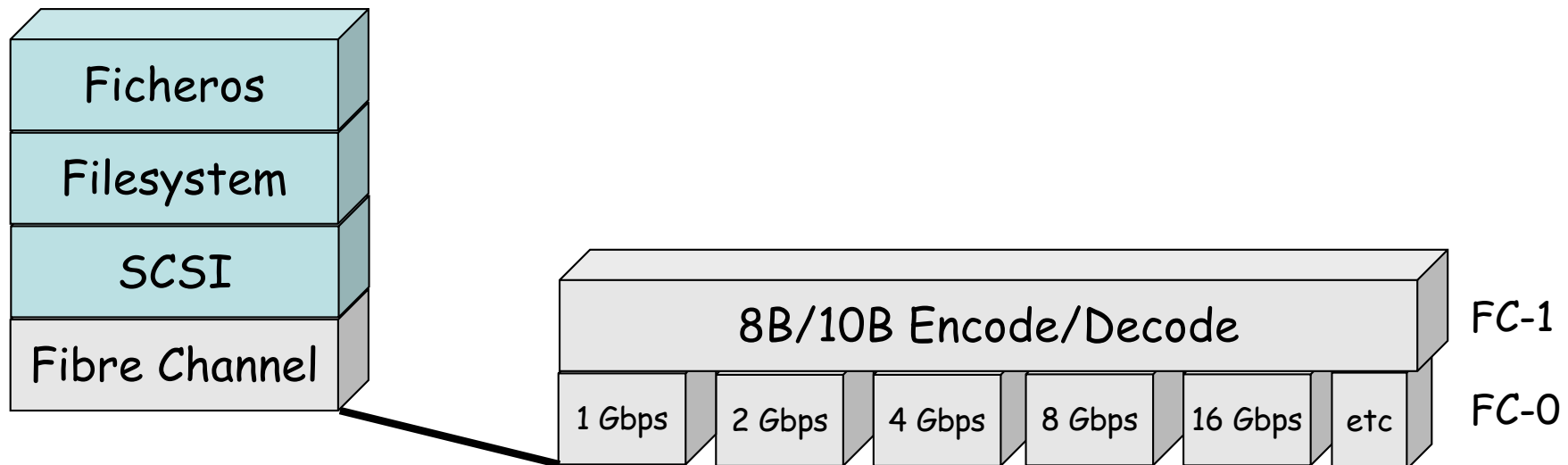
- FC-0: Capa física





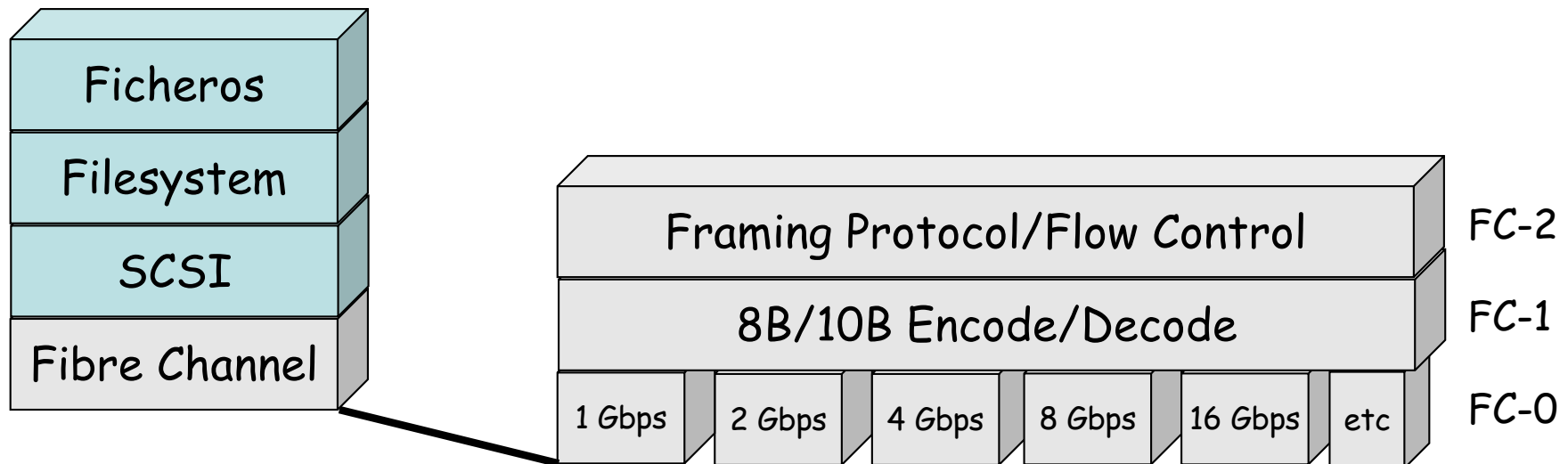
# Fibre Channel

- FC-0: Capa física
- FC-1: Codificación, sincronización, control de errores



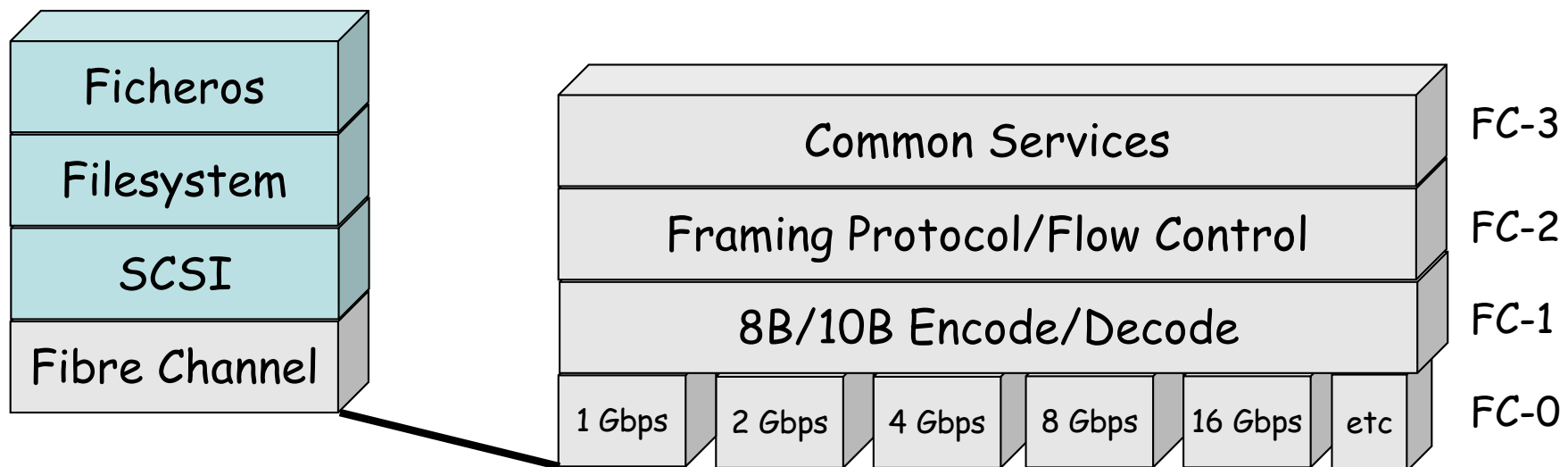
# Fibre Channel

- FC-0: Capa física
- FC-1: Codificación, sincronización, control de errores
- FC-2: Formato de trama, señalización para gestión



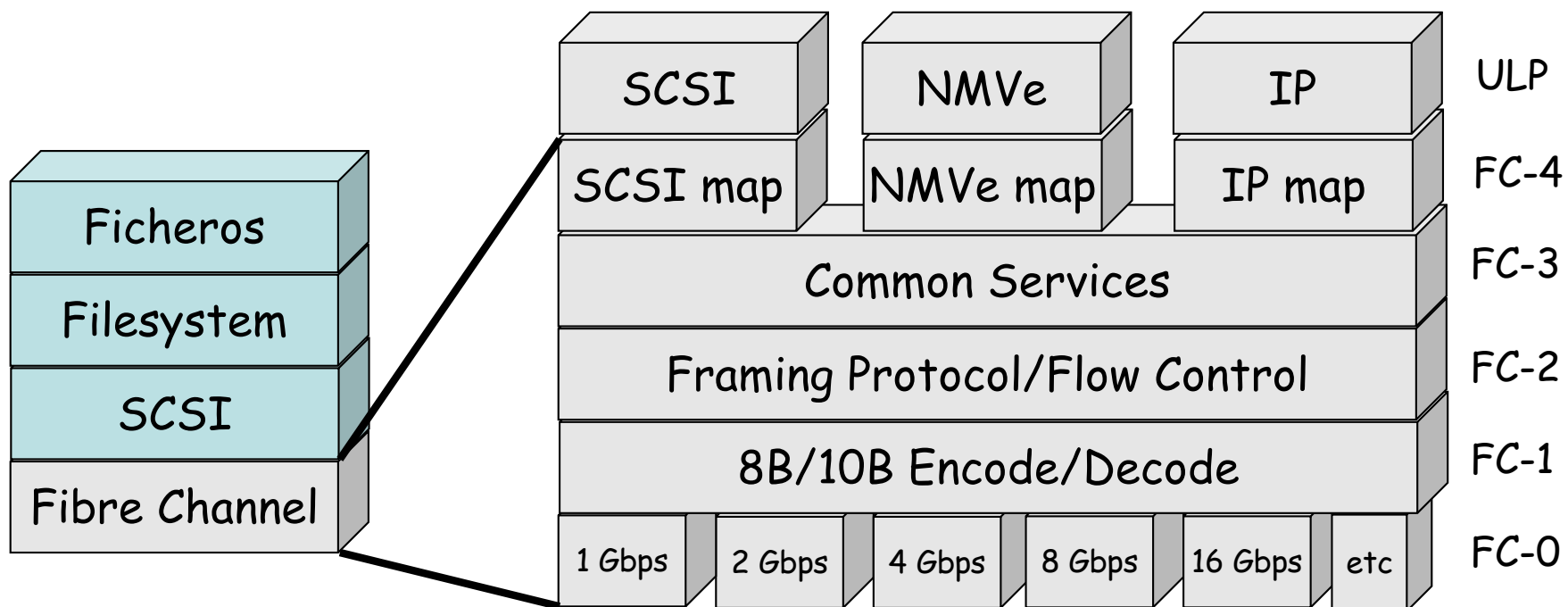
# Fibre Channel

- FC-0: Capa física
- FC-1: Codificación, sincronización, control de errores
- FC-2: Formato de trama, señalización para gestión
- FC-3: Ofrece un conjunto único de servicios aunque por debajo haya varios puertos físicos (*name, login, address manager, alias server, fabric controller, management, key distribution, time*)



# Fibre Channel

- FC-0: Capa física
- FC-1: Codificación, sincronización, control de errores
- FC-2: Formato de trama, señalización para gestión
- FC-3: Ofrece un conjunto único de servicios aunque por debajo haya varios puertos físicos (*name, login, address manager, alias server, fabric controller, management, key distribution, time*)
- FC-4: Capa de adaptación para protocolos superiores como puede ser SCSI, NMVe o IP (ULP = Upper Level Protocol)



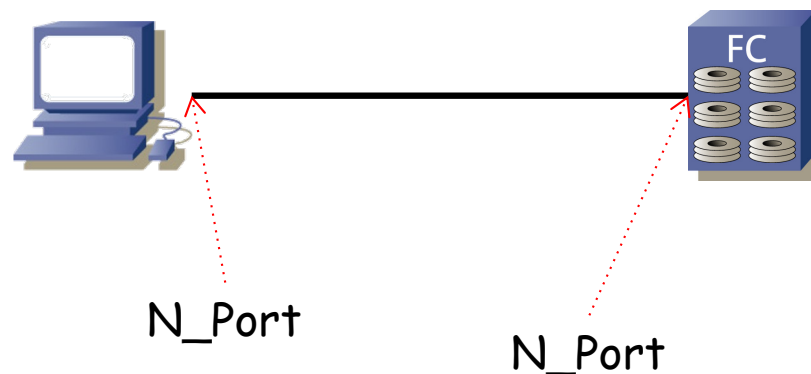
# FC: Arquitectura

# FC: Topologías

# Topologías

## Point-to-Point

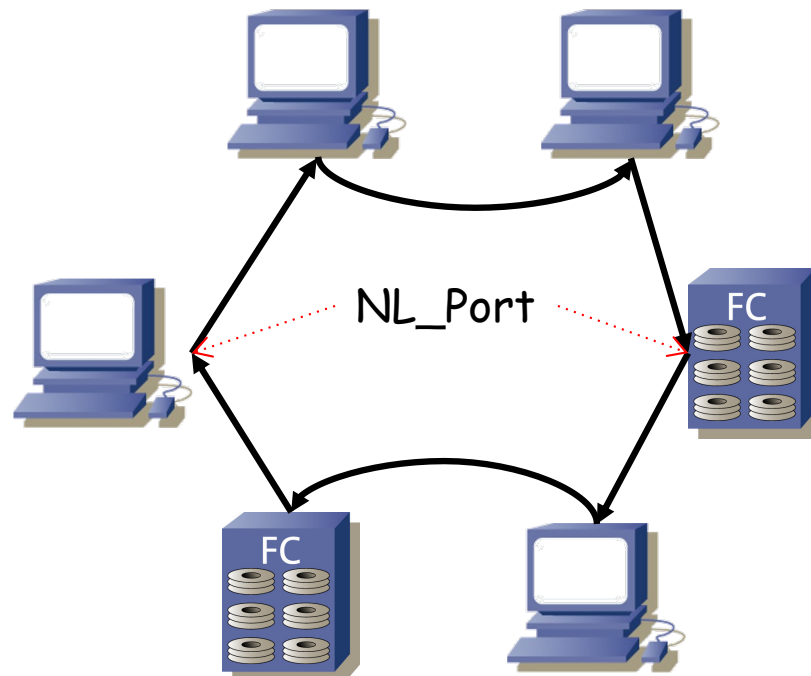
- No hace falta un conmutador o crear “una red”
- Igual que en una Ethernet, podríamos tener simplemente un enlace punto-a-punto
- Ganamos algunas de sus características técnicas pero desde luego no la de compartir el uso de los discos
- Es una conexión directa entre los puertos de 2 nodos, que se vienen a llamar “N\_Ports”



# Topologías

## Arbitrated loop (FC-AL)

- Anillo compartido
- 2-126 dispositivos pero muchos menos por rendimiento
- Los puertos se llaman “L\_Ports” (NL\_Port o FL\_Port)
- Se negocia cuál de los nodos actúa como *master*
- Un nodo establece un circuito entre dos puertos que monopoliza el anillo para poder comunicarse
- No puede hacerlo de nuevo hasta haber dado turno a todos los demás
- (...)

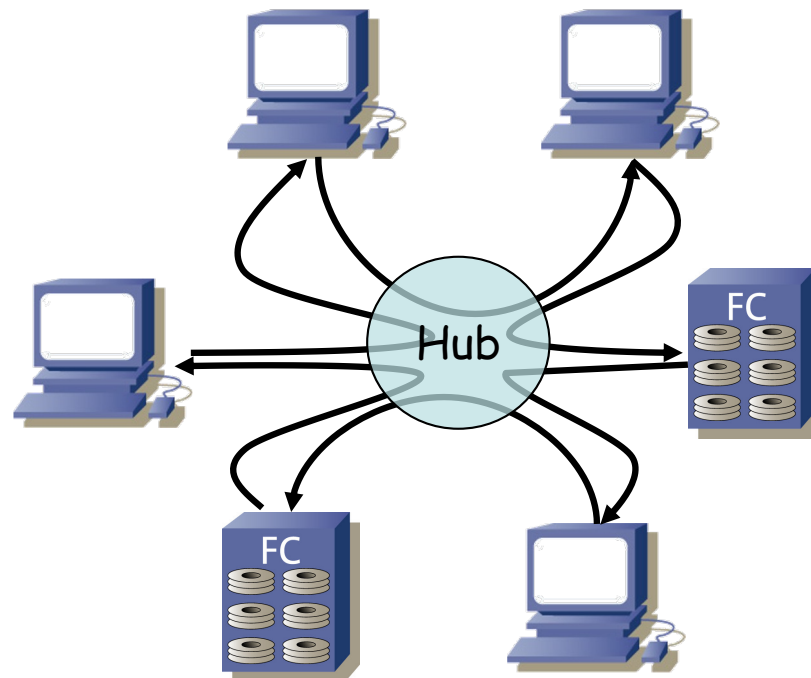




# Topologías

## Arbitrated loop (FC-AL)

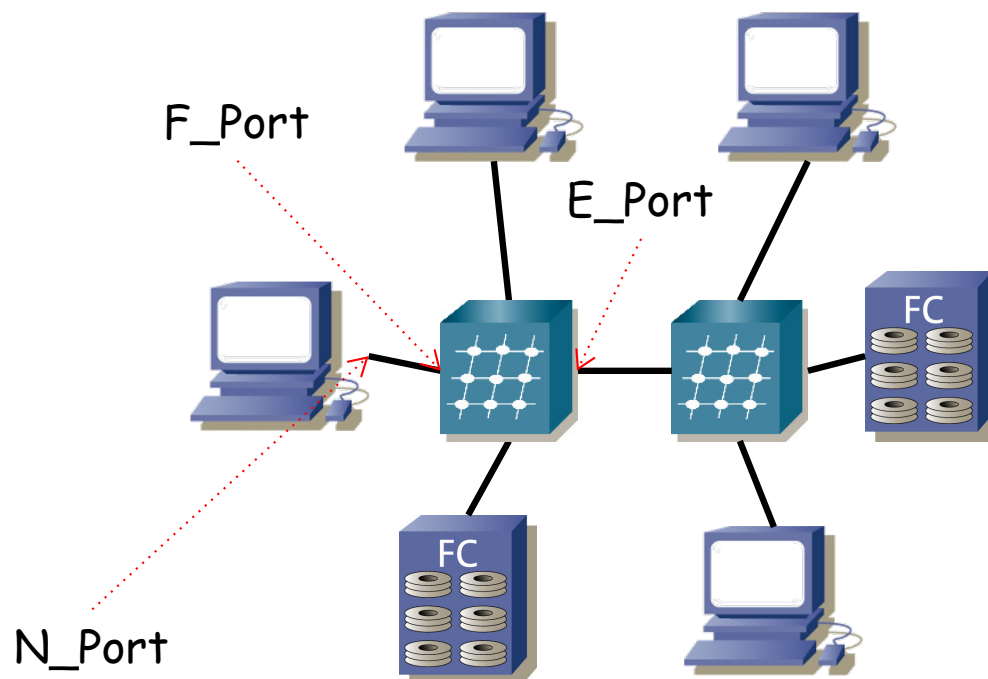
- Anillo compartido
- 2-126 dispositivos pero muchos menos por rendimiento
- Los puertos se llaman “L\_Ports” (NL\_Port o FL\_Port)
- Se negocia cuál de los nodos actúa como *master*
- Un nodo establece un circuito entre dos puertos que monopoliza el anillo para poder comunicarse
- No puede hacerlo de nuevo hasta haber dado turno a todos los demás
- También hub



# Topologías

## Crosspoint switched (fabric)

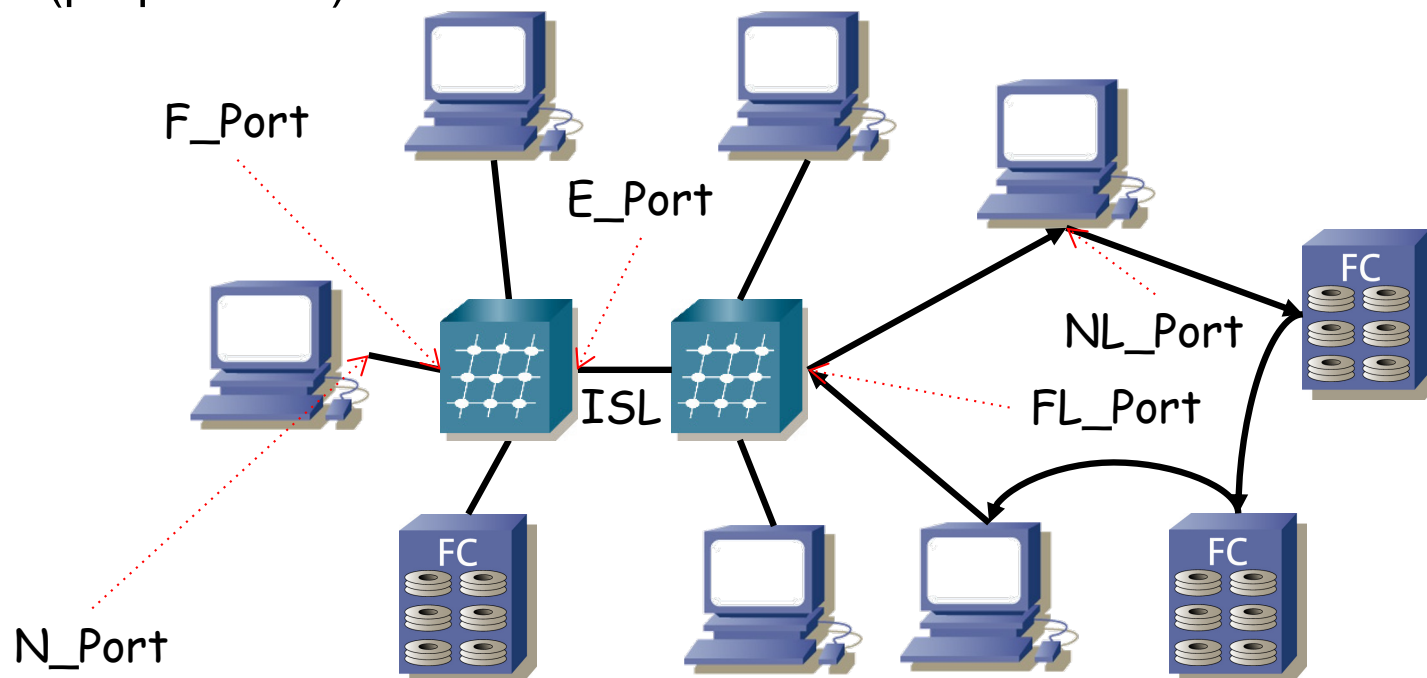
- Uno o más conmutadores, interconectando múltiples nodos
- El direccionamiento permite en teoría hasta  $2^{24}$  nodos
- “F\_Port” (Fabric Port) a los nodos, “E\_Port” (Expansion) entre switches
- ISL = *Inter-Switch Link*
- (...)



# Topologías

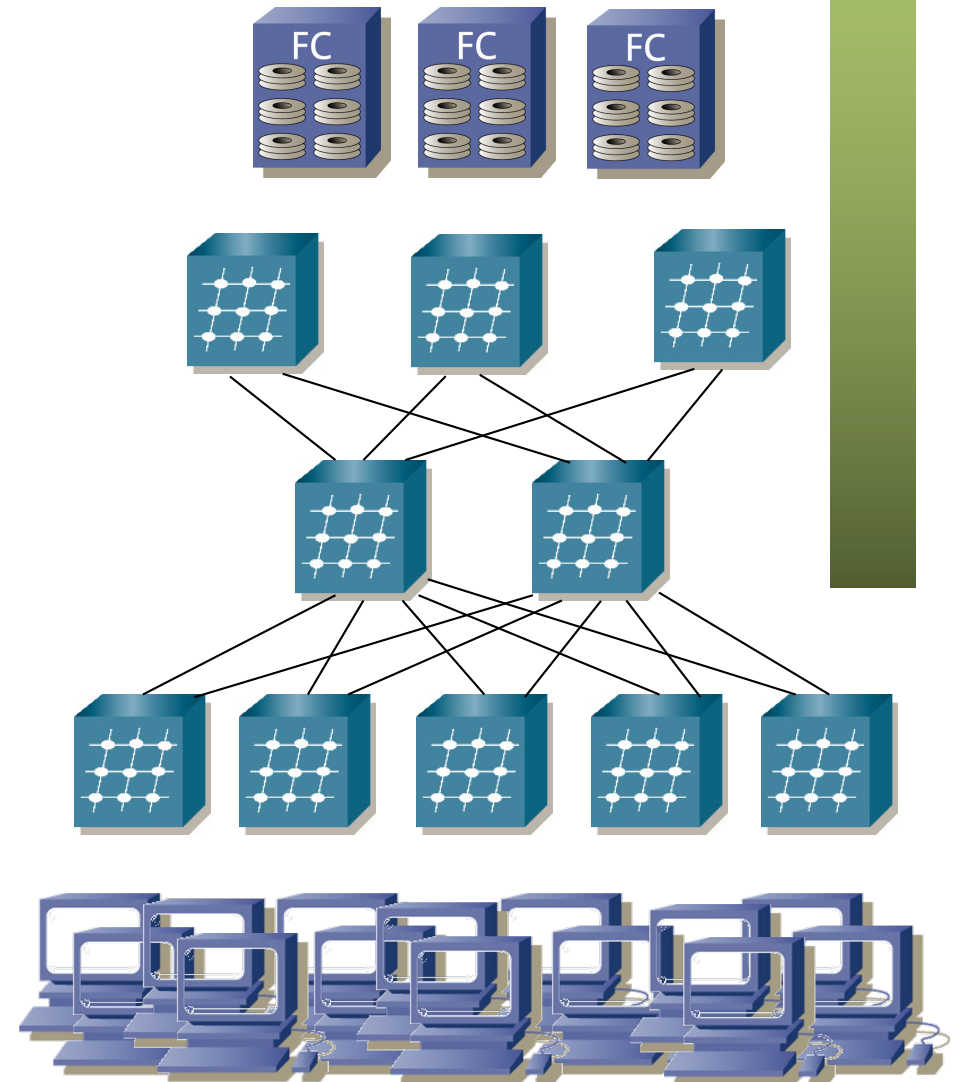
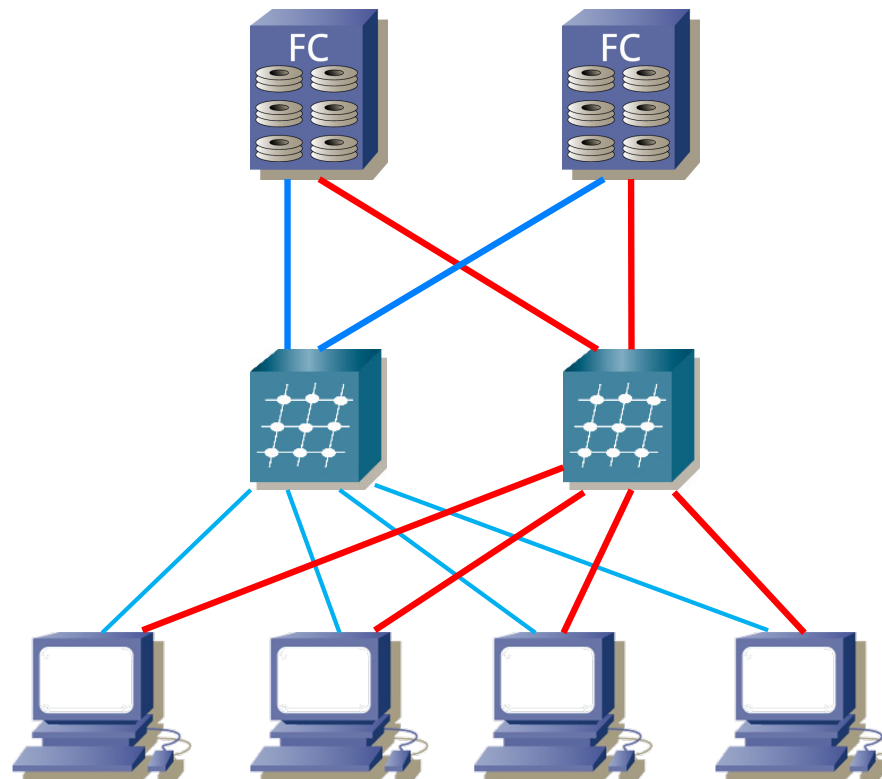
## Crosspoint switched (fabric)

- Uno o más conmutadores, interconectando múltiples nodos
- El direccionamiento permite en teoría hasta  $2^{24}$  nodos
- “F\_Port” (Fabric Port) a los nodos, “E\_Port” (Expansion) entre switches
- ISL = *Inter-Switch Link*
- “FL\_Port” para conectar a un FC-AL
- “G\_Port” puerto genérico que se comporta según lo que se le conecte
- Y otros (propietarios)



# Topologías

- Hoy en día normalmente solo fabric
- Redundancia (SAN A y SAN B)
- Topologías edge-core

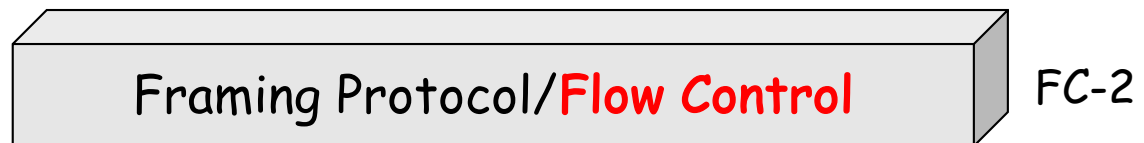
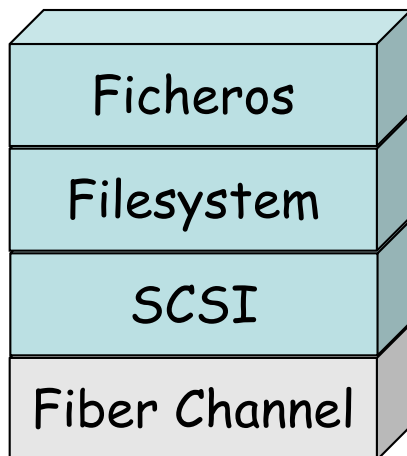


# FC: Topologías

# FC: Servicios

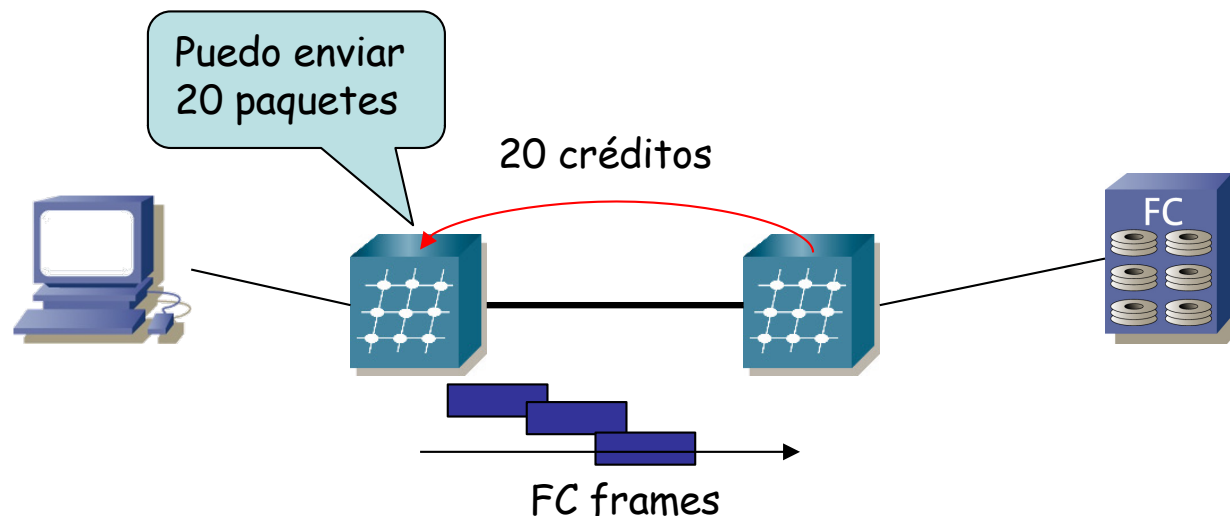
# Service Classes

- Diferentes “clases de servicio” (Class 1, Class 2, etc)
- Con circuitos virtuales (dedicados o compartidos) o datagramas
- Con ACKs y NACKs o sin ellos
- Con garantía de orden o no
- Con control de flujo en cada salto y/o extremo a extremo
- Lo habitual es clase 3 para almacenamiento
  - Sin conexión, no garantía entrega en orden
  - Best effort
  - Sin ACKs ni NACKs
  - Flow control salto a salto



# Flow control

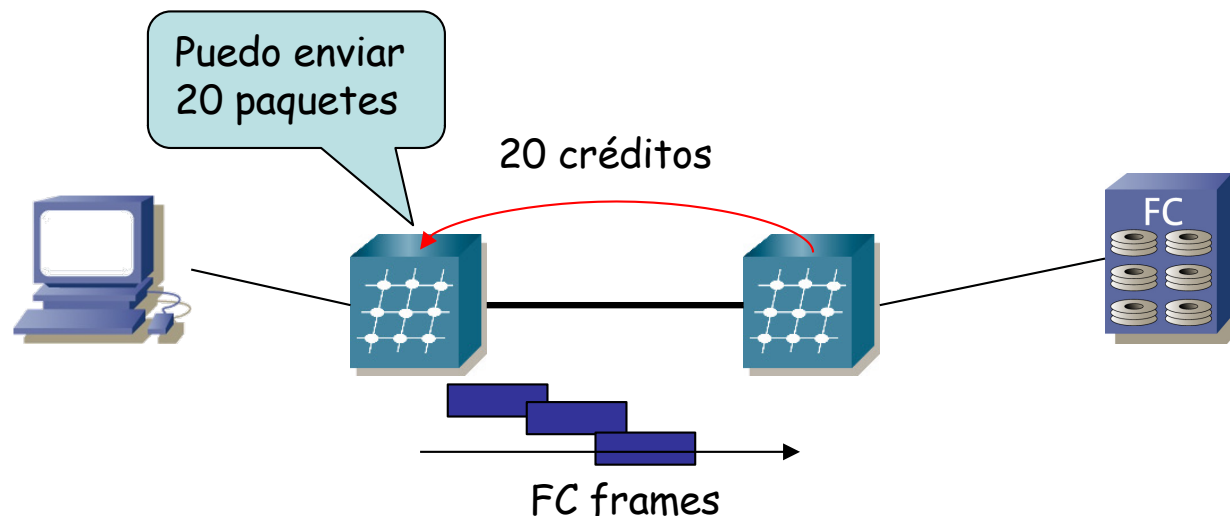
- Se busca un escenario sin pérdidas
- No se envía un paquete si el receptor no tiene espacio para almacenarlo
- Ventana deslizante
- Con flow control salto a salto el receptor es el otro extremo del enlace
- Receptor ofrece una cantidad de "créditos" (*buffer-to-buffer credits*)
- Transmisor puede enviar esa cantidad de paquetes
- Hasta que reciba un nuevo anuncio de créditos ( $R\_RDY$ ) (porque el siguiente salto haya vaciado buffers)
- Problemas de *slow drain*





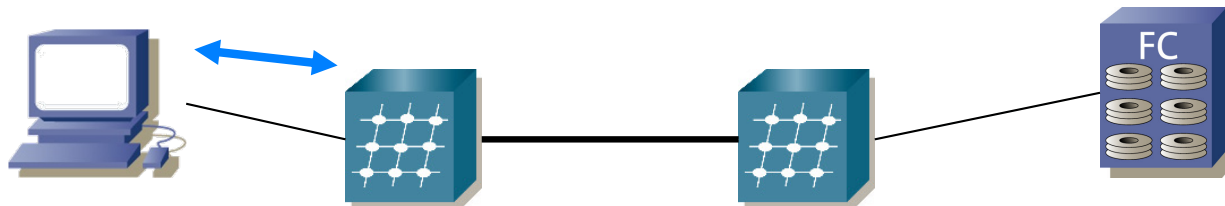
# Flow control

- Misma cuenta que en todo mecanismo de ventana deslizante
- Trama de 2000 bytes a 1Gb/s tiene un tiempo de transmisión de 16us
- Enlace de 100Km tiene un tiempo de propagación de unos 500us
- Se pueden enviar más de 30 paquetes antes de que el primero alcance el otro extremo de los 100Km
- Un R\_RDY todavía tardará al menos otros 500us
- Necesitamos al menos 60 créditos para saturar el enlace



# Login

- Cada dispositivo tiene el equivalente a las direcciones MAC de Ethernet:
  - WWNN: World Wide Node Name (identifica al dispositivo)
  - WWPN: World Wide Port Name (identifica al puerto)
- Cada switch tiene un 'Domain ID'
- Conectar un N-Port a F-Port no garantiza comunicación
- El nodo debe hacer login en el *fabric* (en el switch)



# Login

## FLOGI: Fabric Login

- Es un intercambio de paquetes: FLOGI (Fabric Login)
- Cada nodo obtiene del fabric un FCID (con el Domain ID del switch)
- Establece créditos
- El fabric enruta el tráfico hacia el FCID

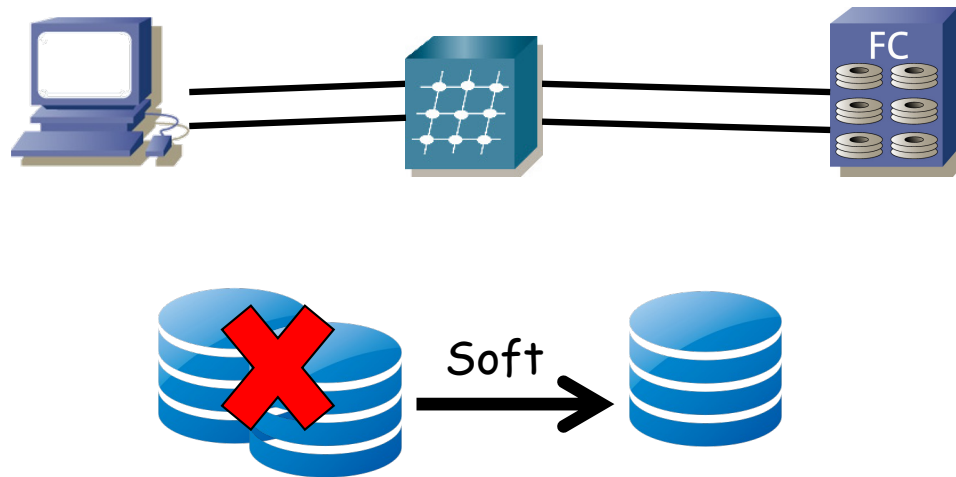
## PLOGI: Port Login

- Login en el target
- Permite establecer los créditos para flow control si no hay *fabric*
- Name server en el switch registra el mapeo entre WWPN y FCID



# Multipath

- Permite caminos redundantes
- Varios HBAs en host y varios interfaces en sistema de almacenamiento
- FC dispone del protocolo FSPF (*Fibre Channel Shortest Path First*)
- No bloquea caminos alternativos sino que puede repartir carga
- O pueden servir como caminos redundantes ante fallos
- El sistema operativo del host requiere software que le permita ver los dos accesos al volumen como un solo volumen



# FC: Services

# Fibre Channel