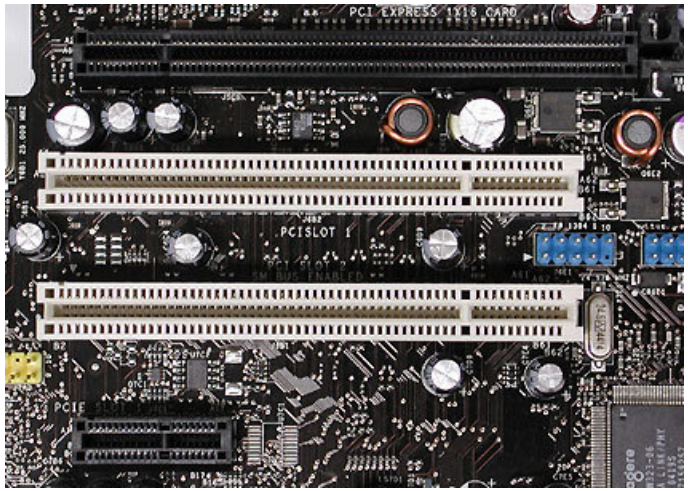
upna Universidad Pública de Navarra Nafarroako Unibertsitate Publikoa
Redes de Nueva Generación
Área de Ingeniería Telemática

PCIe IO Virtualization

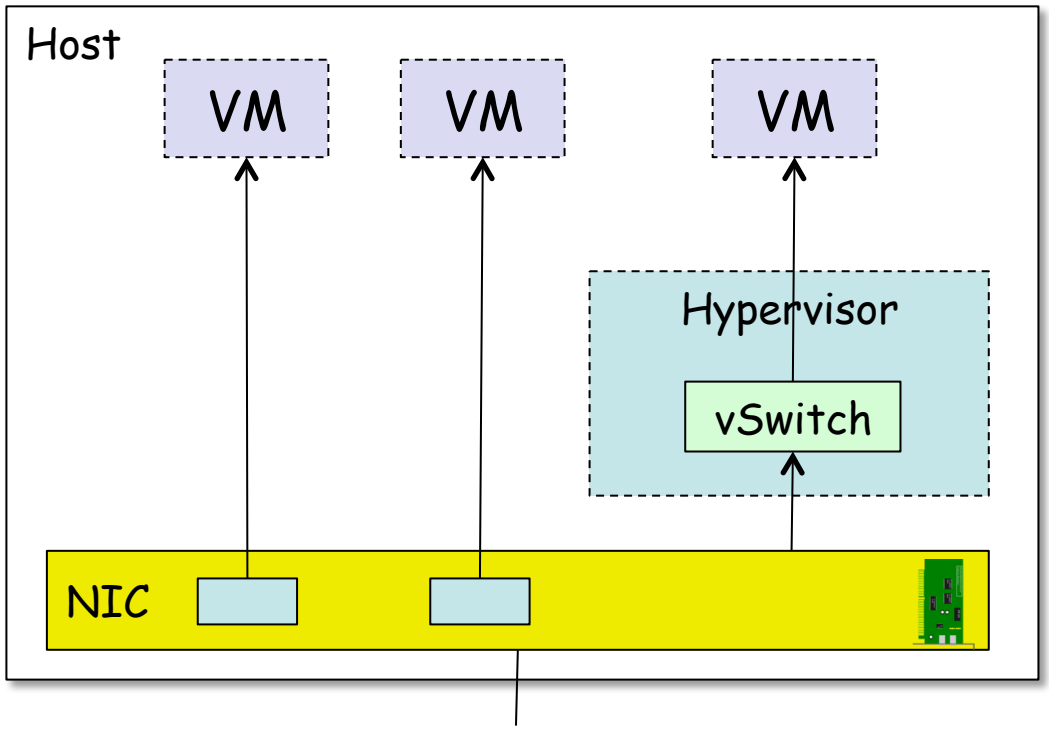
PCI Express

- PCI-SIG : *Peripheral Component Interconnect Special Interest Group* (consorcio de fabricantes)
- Interconexión para periféricos I/O
- Evolución de PCI y PCI-X
- Enlace punto-a-punto entre dos dispositivos (4 pins por “lane”)
- Un protocolo (y paquetes) para las transferencias
- Crece aumentando el número de “lanes” (pistas)
- Según la generación (1-3) entre 250MBytes/s (2Gbps) por pista y 1000 MBytes/s (8Gbps) por pista (full-duplex)



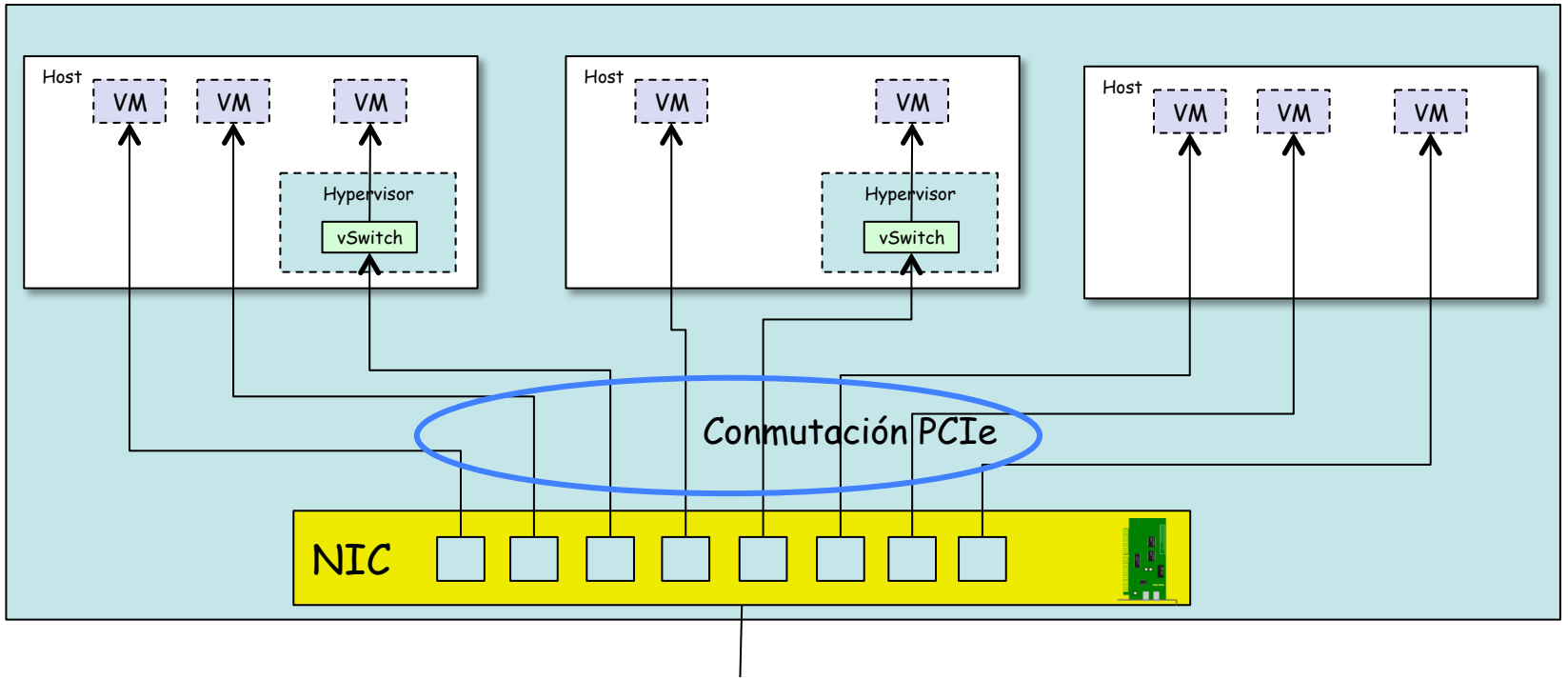
SR-IOV

- *"Single Root Input/Output Virtualization"*
- La NIC física se virtualiza (hasta centenares de instancias)
- Tiene buffers independientes para cada instancia
- La NIC virtual se asigna a la VM
- A partir de ahí el tráfico no necesita pasar por el hypervisor
- La NIC puede mover los datos por DMA a la VM
- Lo mismo con Fibre Channel



MR-IOV

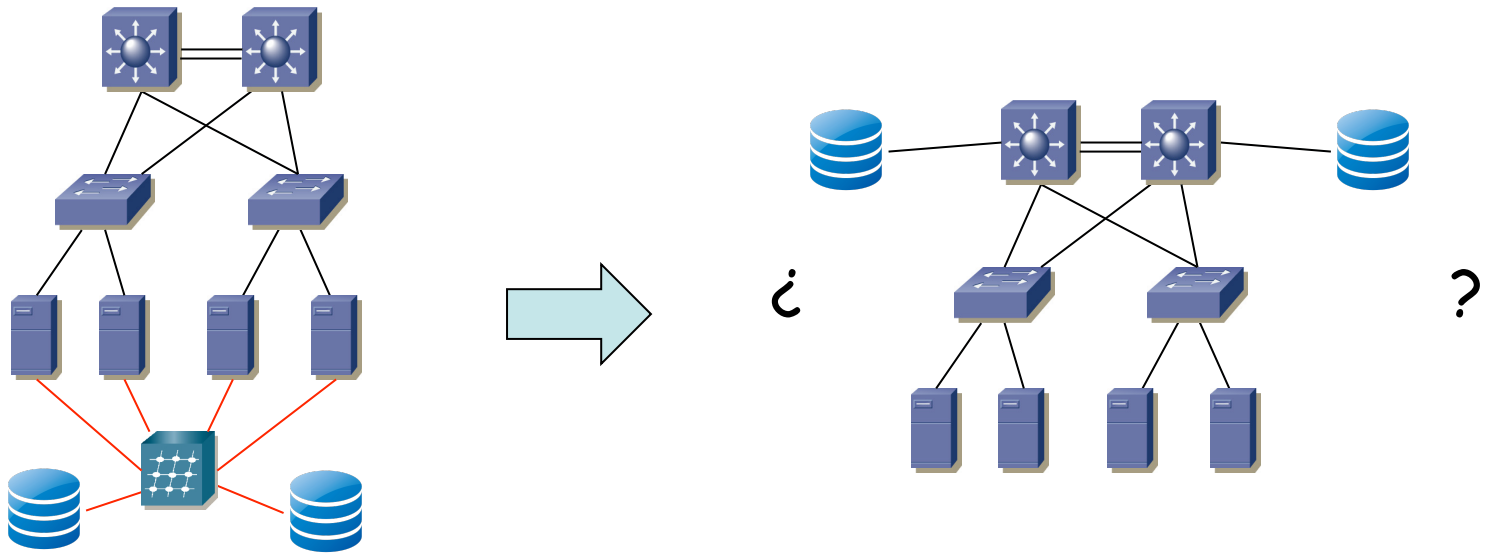
- *"Multi Root Input/Output Virtualization"*
- Por ejemplo en un entorno *blade*
- Se comparte la NIC entre diferentes hosts
- Ahorra en NICs y en consumo eléctrico
- Requiere una estructura de conmutación PCIe



I/O Consolidation

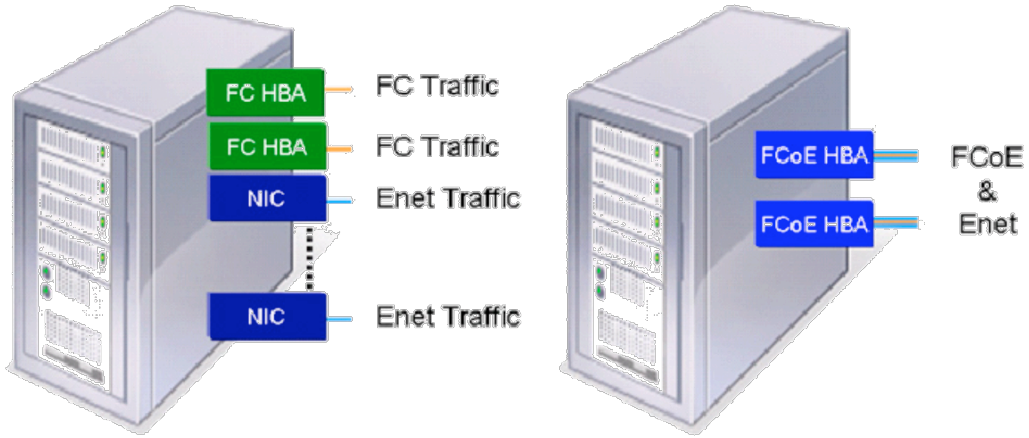
I/O Consolidation

- Consolidación: Emplear la misma infraestructura física para transportar múltiples tipos de tráfico
- En su día se produjo consolidación entre las redes de voz y datos
- Ahora entre la de datos (+voz) y la de almacenamiento
- Esto ya lo hacía Infiniband pero no ha sido popular
- Parece que finalmente van a converger sobre Ethernet (¿ o sobre IP ?)



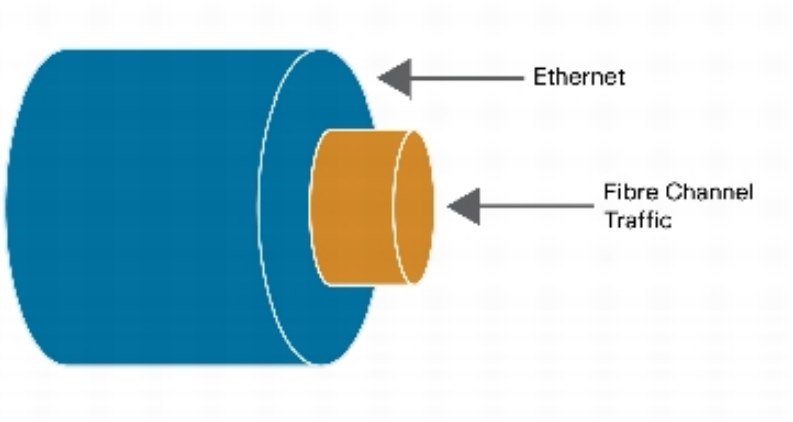
I/O Consolidation

- En lugar de NICs y HBAs tendremos solo CNAs
- CNAs (Converged Network Adapters) son NICs Ethernet que implementan por hardware parte de los protocolos
- Beneficios:
  - Simplifica el cableado y la infraestructura de red (pasamos de 2 redes a solo 1)
  - El equipo retirado nos ahorra costes de alimentación, refrigeración y espacio
- Desventajas
  - Una nueva tecnología a dominar y gestionar
  - Y hay que cambiar el hardware
  - Posibles acoplamientos
  - Mayor fragilidad



Tecnología

- Ethernet es la solución de LAN
- Fibre Channel la solución de almacenamiento en red
- Transportar los datos de LAN sobre FC no ha sido interesante por la falta de buen soporte de multicast/broadcast en FC
- SCSI requiere transporte fiable porque se diseñó para cables paralelos fiables y su recuperación ante pérdidas es lenta
- iSCSI es una opción pero tampoco se ha desplegado masivamente por el posible sobrecoste de TCP y porque no mantiene la gestión de FC
- En general los transportes sobre IP son más costosos en el hardware
- Finalmente Ethernet parece la solución ganadora, pero hay que modificarlo para evitar las pérdidas



DCB

DCB

- *Data Center Bridging*
- Modificaciones a Ethernet por parte del IEEE 802.1 para permitir la convergencia con I/O en el data center
- Entornos con un producto retardo-ancho de banda limitado, así como un limitado número de saltos

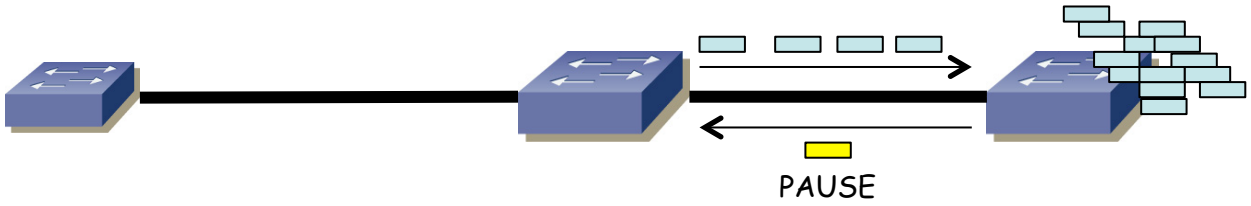
Feature	Benefit
Priority-based Flow control (PFC; IEEE 802.1 Qbb)	Provides capability to manage bursty, single traffic source on a multiprotocol link
Enhanced transmission selection (ETS; IEEE 802.1 Qaz)	Enables bandwidth management between traffic types for multiprotocol links
Congestion notification (IEEE 802.1 Qau)	Addresses the problem of sustained congestion by moving corrective action to the network edge
Data Center Bridging Exchange (DCBX) Protocol	Allows autoexchange of Ethernet parameters between switches and endpoints

Control de flujo en FC

- Fibre Channel ofrece una red sin pérdidas por congestión
- Mediante control de flujo salto a salto
- Esto le obliga a topologías simples
- El problema es que la congestión en un switch propaga el control *upstream*
- Puede afectar a flujos que no son responsables de la congestión
- En topologías complejas puede llevar a interbloqueos que degraden el rendimiento
- Así que requiere topologías simples

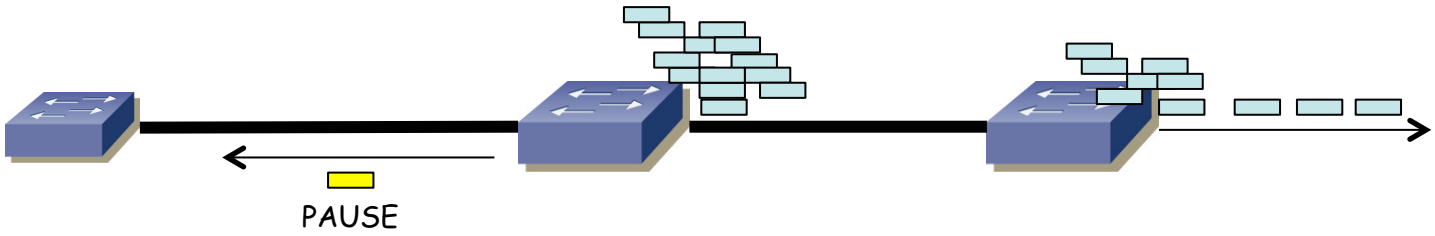
Control de flujo en Ethernet

- Inicialmente no tenía sentido salto a salto pues no había “saltos”
- Con la introducción de puentes/switches, se añade en 802.3x
- El receptor que sufre congestión envía tramas de PAUSE a la fuente
  - Son tramas de control MAC (Ethertype 0x8808)
  - Un campo indica el tiempo de pausa en tiempos de tx de 512 bits
  - Tramas enviadas a MAC multicast reservada (01:80:C2:00:00:01) que no son reenviadas
- ¿Qué sucederá a continuación?
- (…)



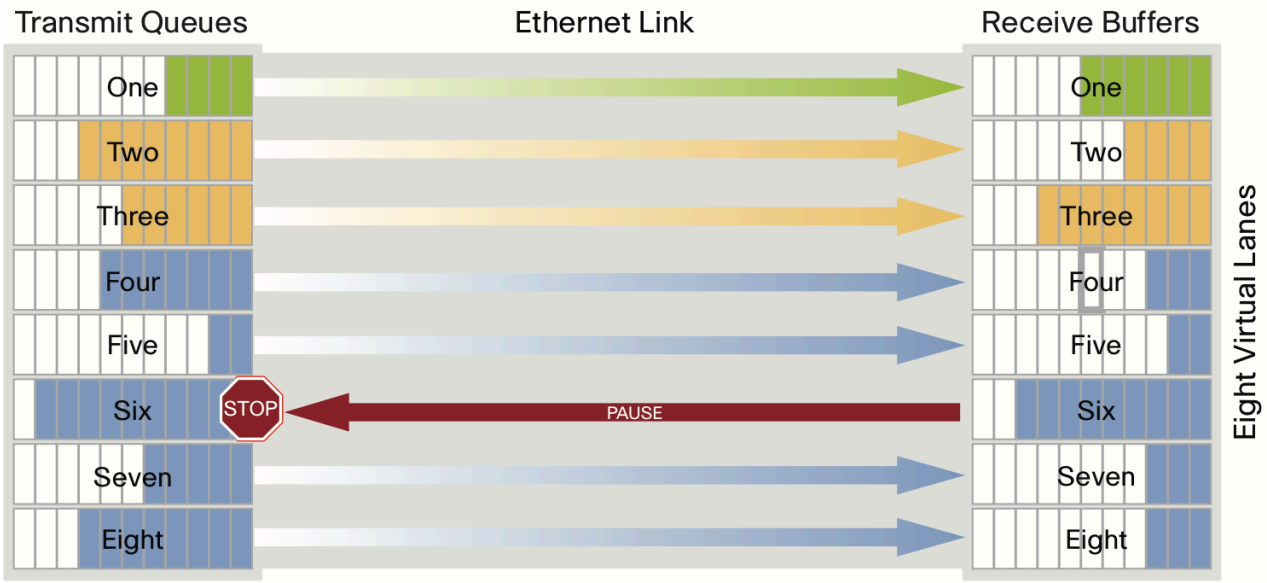
Control de flujo en Ethernet

- ¿Qué sucederá a continuación?
- El conmutador anterior dejará de enviar y con ello probablemente empiece a acumular tramas
- Con ello probablemente envíe una trama de PAUSE al conmutador anterior
- Esto no es un reenvío de la trama anterior
- El mecanismo es para cada salto
- Poco empleado, implementaciones inconsistentes



PFC

- *Priority-based Flow Control*
- IEEE 802.1Qbb (ya recogido en 802.1Q-2012)
- Permite que no haya pérdidas por congestión para los protocolos que así lo requieran
- Se hace control de flujo de forma independiente para cada clase de servicio de 802.1p



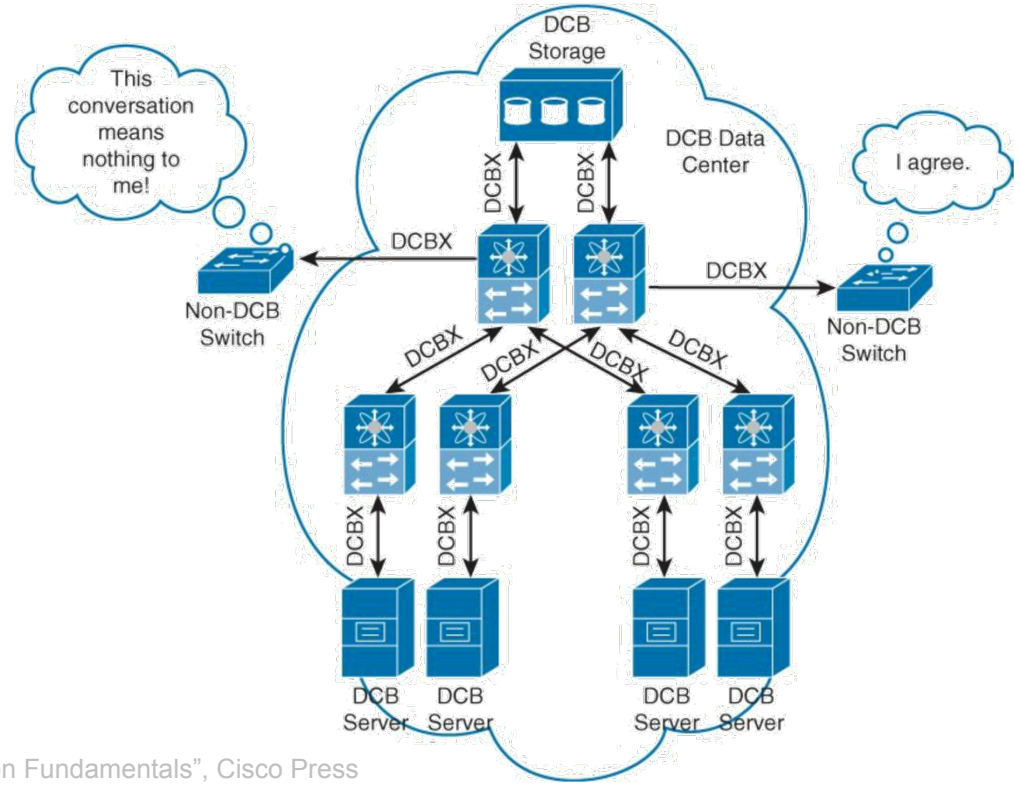
http://www.cisco.com/c/dam/en/us/solutions/collateral/data-center-virtualization/ieee-802-1-data-center-bridging/at_a_glance_c45-460907.pdf

ETS

- *Enhanced Transmission Selection*
- IEEE 802.1Qaz (ya recogido en 802.1Q-2012)
- 802.1Q definía 8 prioridades pero no cómo hacer la planificación entre ellas
- ETS no concreta el planificador a usar pero sí los requerimientos que debe cumplir
- Por defecto deberían soportar un planificador con prioridades estrictas
- También puede soportar lo que se conoce como un “credit-based shaper”, que es un token bucket para cada cola
- La tercera opción es el algoritmo ETS
  - Para cuando no hay tramas en las colas de prioridad o con “credit-based shaper”
  - Dice 802.1Q: “transmission selection is performed based on the allocation of bandwidth to that traffic class. Bandwidth is distributed among ETS traffic classes that support ETS algorithm such that each traffic class is allocated available bandwidth in proportion to its TCBandwidth”
  - Se menciona WRR como una opción, pero no se especifica más detalle

DCBX

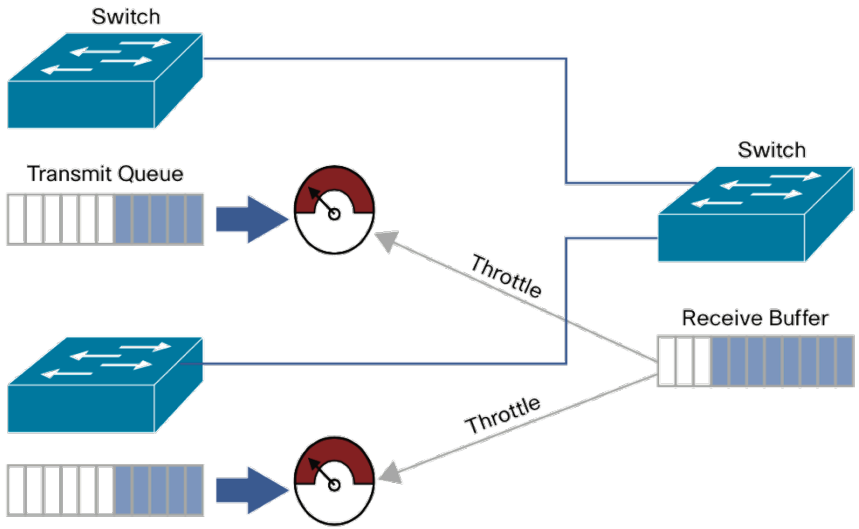
- *Data Center Bridging Exchange Protocol*
- IEEE 802.1Qaz (ya recogido en 802.1Q-2012), extensión de LLDP
- DCBX permite descubrir las capacidades de los extremos de un enlace
- Permite detectar y resolver conflictos en la configuración
- Permite la configuración de parámetros del otro extremo
- Empleado principalmente para parámetros de PFC y ETS



"Data Center Virtualization Fundamentals", Cisco Press

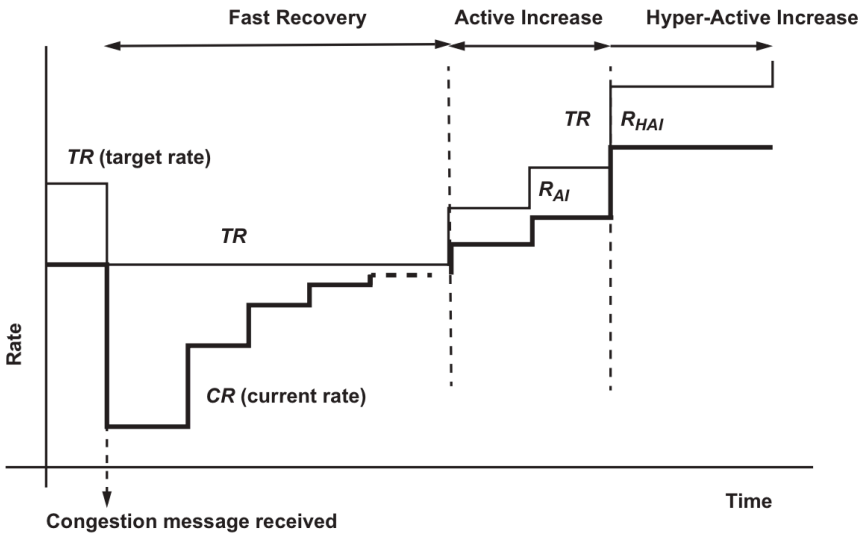
QCN

- *Quantized Congestion Notification* protocol
- IEEE 802.1Qau (ya integrado en 802.1Q-2011)
- Para dominios con BWxRTT < 5 Mbits
- Con enlaces de 10Gbps quiere decir un RTT < 0.5ms
- Es decir, data centers, backplanes, computing clusters, SANs
- Habilita la capacidad en los puentes (y hosts) de enviar señales de congestión a las estaciones finales para que limiten la tasa
- La fuente puede etiquetar las tramas con un CN-Tag (2 bytes Flow ID)
- Se le devuelve en el mensaje de CN
- Le permite identificar el flujo cuya tasa debe reducir



QCN

- Al que detecta la congestión se le llama *Congestion Point (CP)*
- Debe enviar el mensaje de notificación antes de llenar el buffer pues en lo que llega el mensaje seguirá recibiendo paquetes (estimación)
- Lo envía a la dirección origen del paquete con el que toma la decisión
- Quien recibe la notificación se llama el *Reaction Point (RP)*
- Al recibir la notificación el RP reduce la tasa del flujo mediante un *rate limiter*
- No recibe indicación de que pueda aumentar la tasa de nuevo así que lo hace unilateralmente



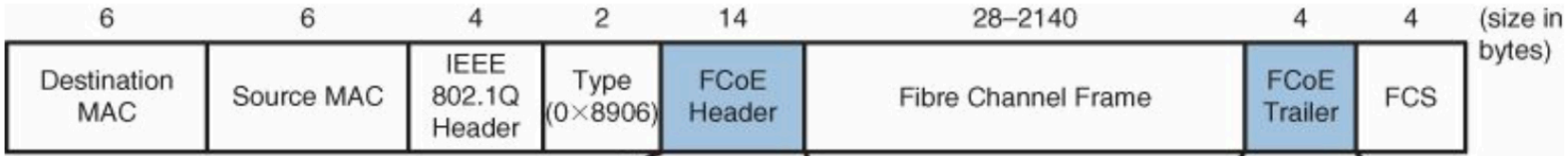
802.1Q

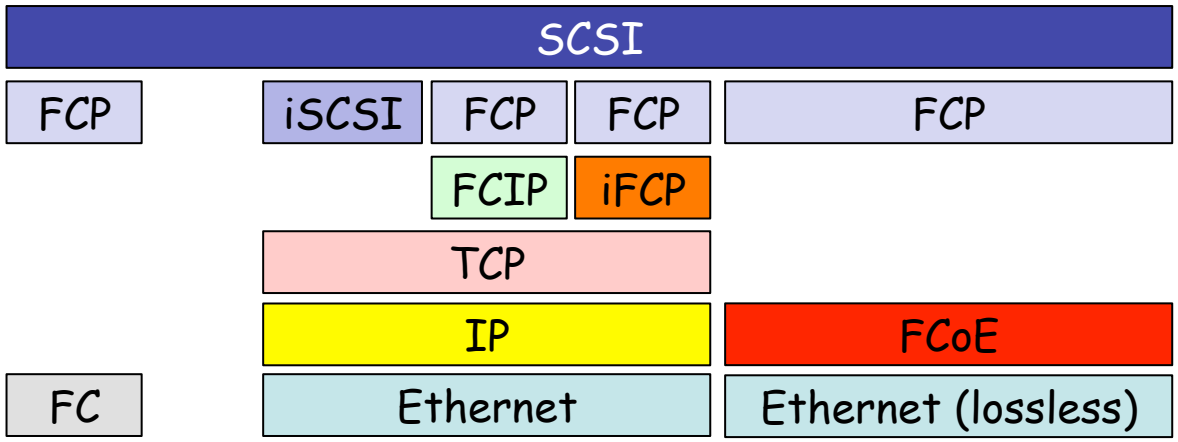
Figure 30-3—QCN RP operation

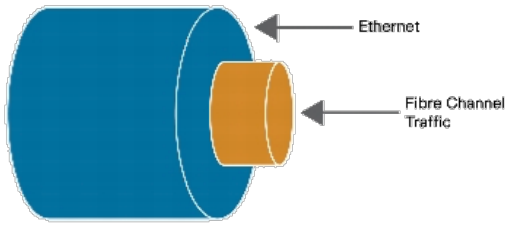
FCoE

FCoE

- *Fibre Channel over Ethernet*
- Mantiene el funcionamiento y la gestión de FC
- Estándar FC-BB-5 del grupo T11 del INCITS
- Tiene sentido a partir de 10GE, cuando está a la par con velocidades en Fibre Channel
- Requiere una Ethernet sin pérdidas (PFC)
- Requiere soporte de *jumbo frames* para transportar en una sola trama Ethernet toda la trama FC (36 bytes cabecera + hasta 2112 de datos)

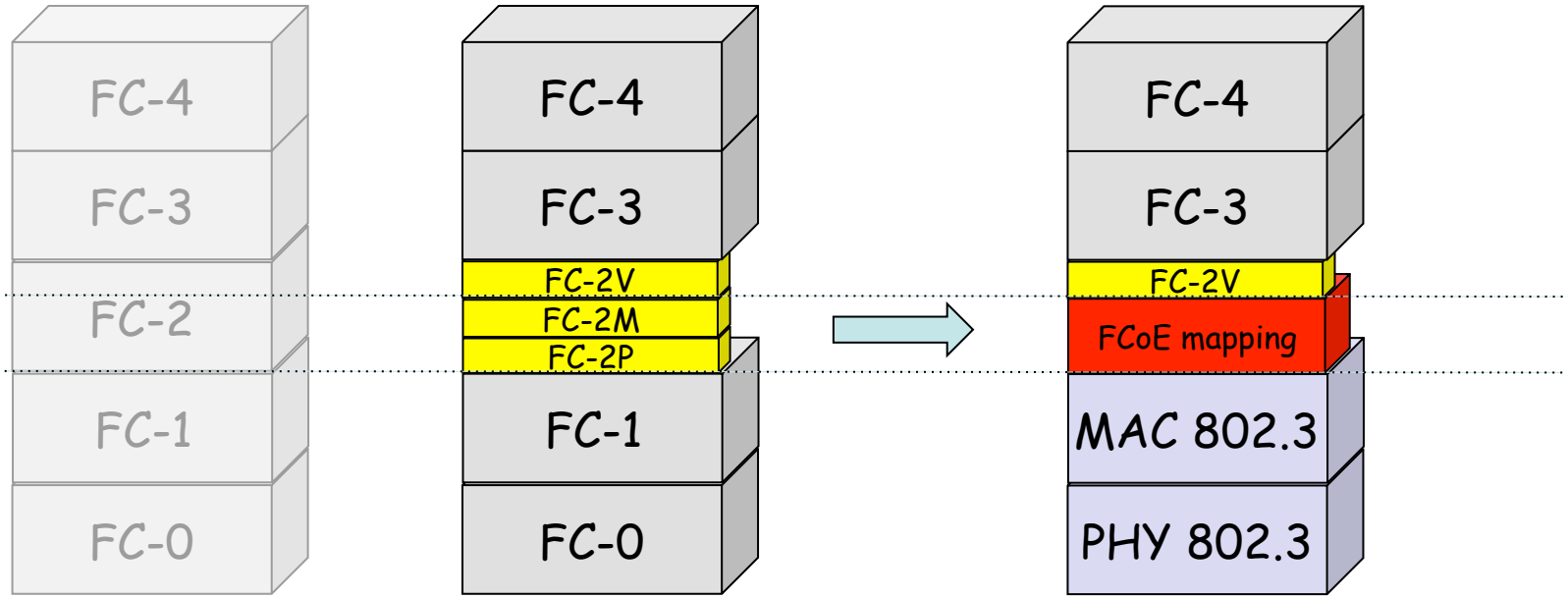






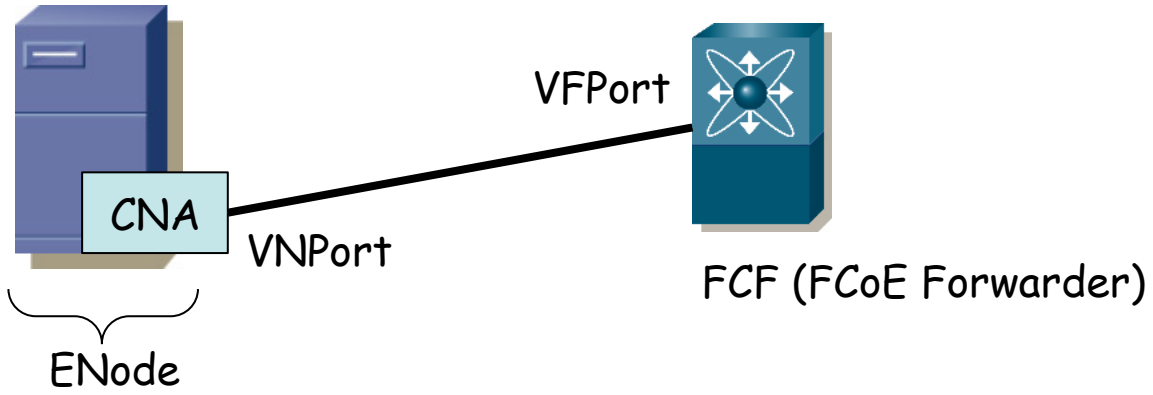
FCoE Mapping

- FC-0 y FC-1 son la capa física, sincronización, errores
- FC-2 se encarga del formato de la trama, señalización y gestión
  - FC-2V define el interfaz con FC-3 y las funciones que se le ofrecen, independientemente del FC-1
  - FC-2M define la multiplexación para cuando hay múltiples FC-2V sobre un FC-2P o viceversa
  - FC-2P implementa el control de flujo, se sustituye por el de PFC
- Mantener FC-2V hace FCoE transparente para el sistema operativo



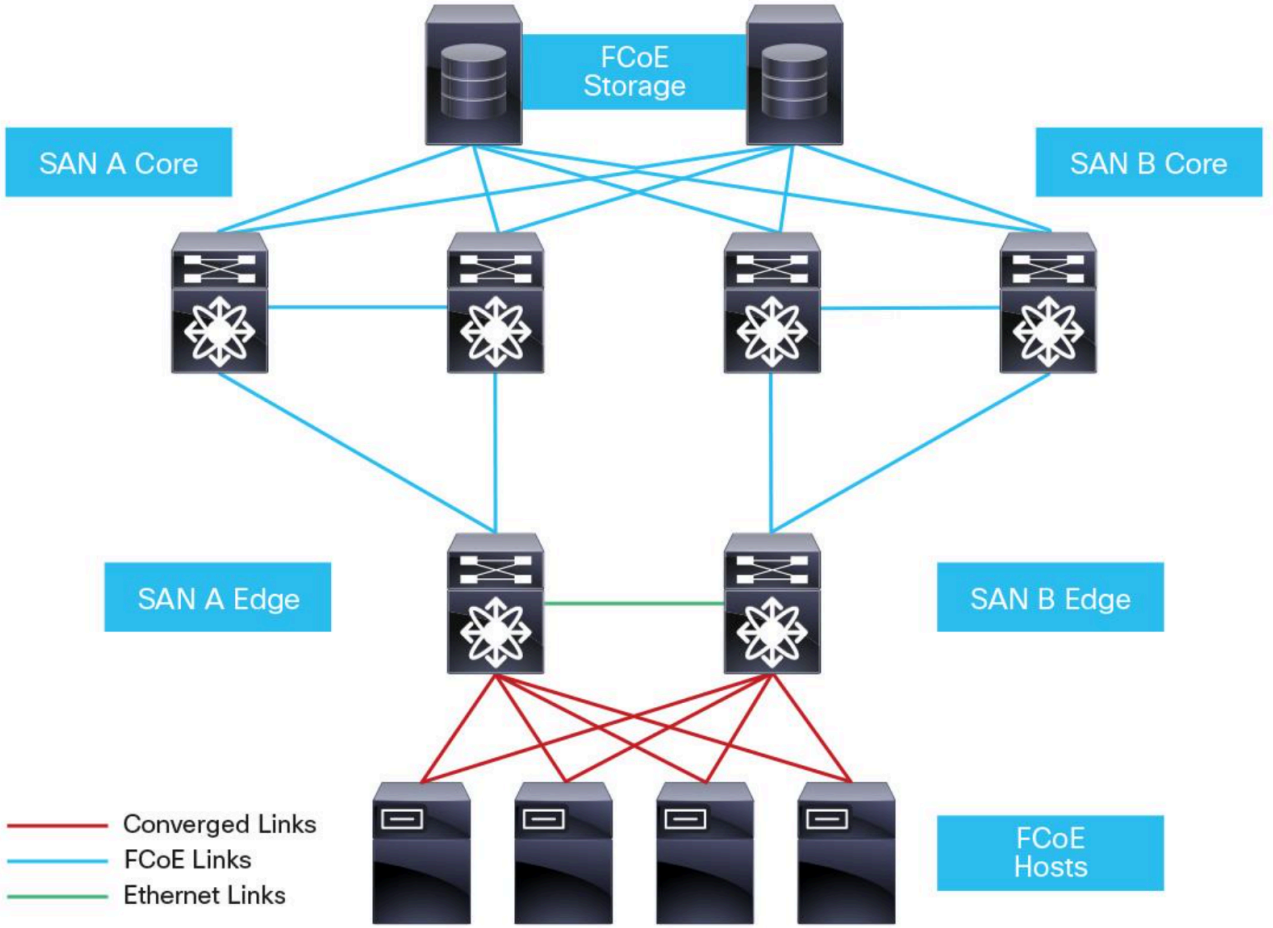
FCoE y FIP

- Un enlace nativo FC es punto-a-punto
- Ethernet es multi-acceso
- FCoE emplea el FCoE Initialization Protocol (FIP) para convertir la Ethernet multi-acceso en un conjunto de enlaces punto-a-punto virtuales
- Transporta también los servicios de descubrimiento y login de Fibre Channel
- FCoE emplea el Ethertype 0x8906 mientras que FIP el 0x8914
- FCoE es el plano de datos, FIP el de control



Ejemplo de topologías

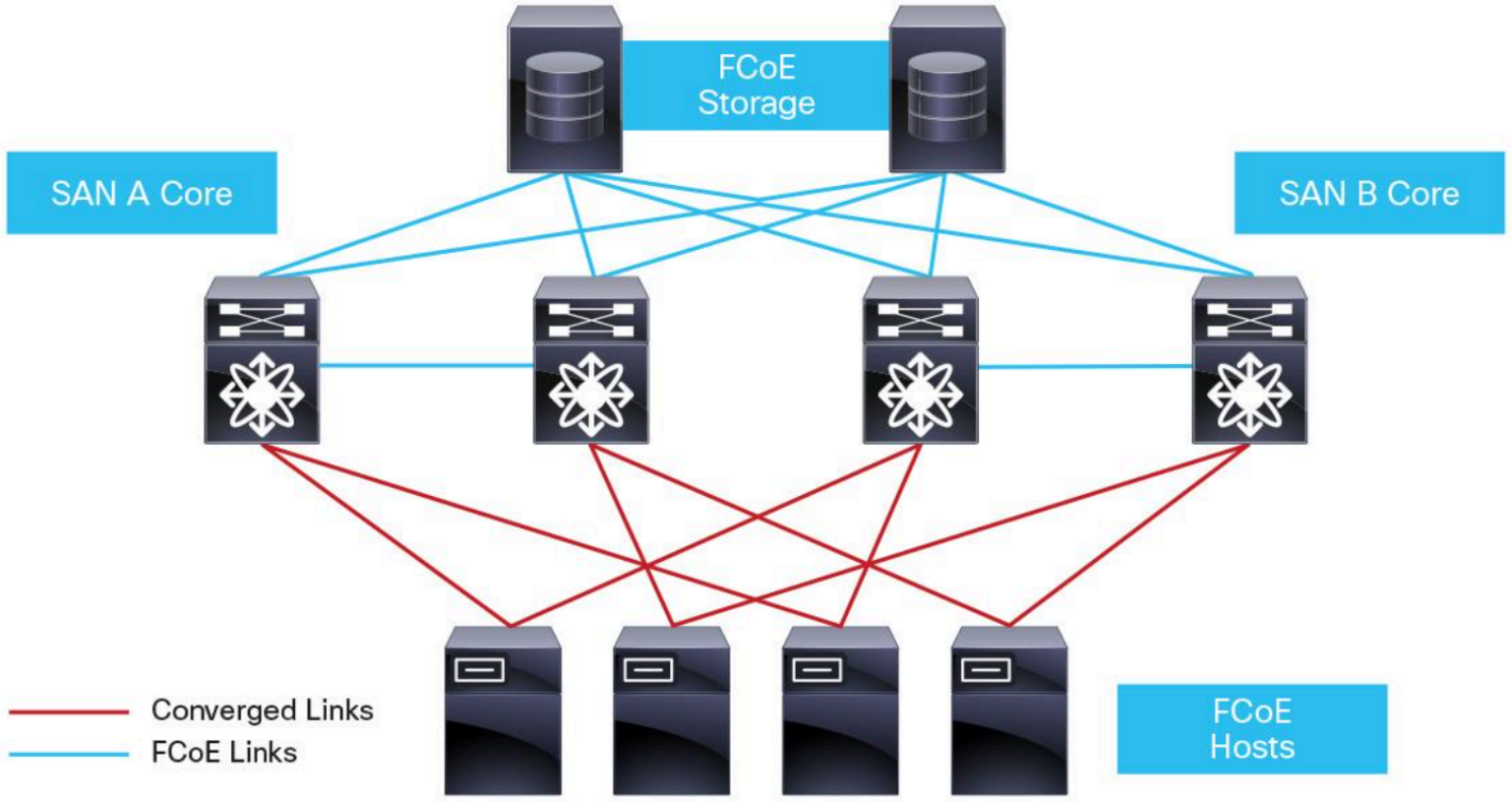
- Core-Edge



http://www.cisco.com/c/en/us/products/collateral/storage-networking/mds-9700-series-multilayer-directors/guide-c07-732733.pdf

Ejemplo de topologías

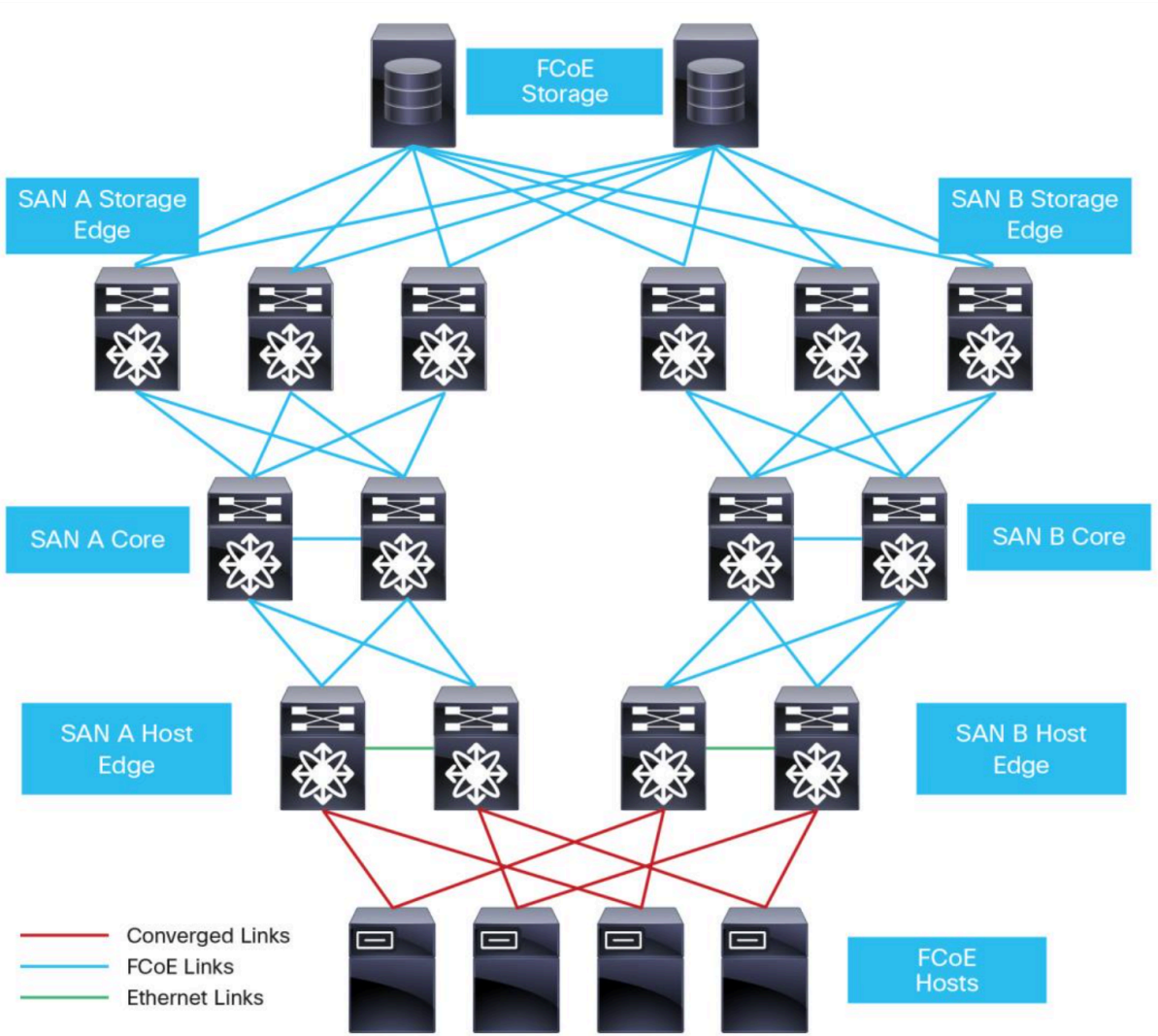
- Collapsed Core



http://www.cisco.com/c/en/us/products/collateral/storage-networking/mds-9700-series-multilayer-directors/guide-c07-732733.pdf

Ejemplo de topologías

- Edge-Core-Edge

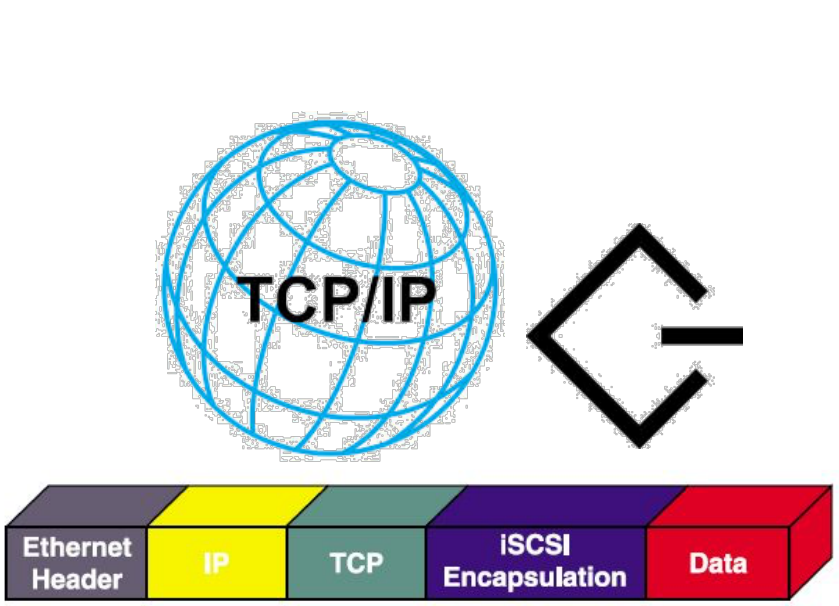


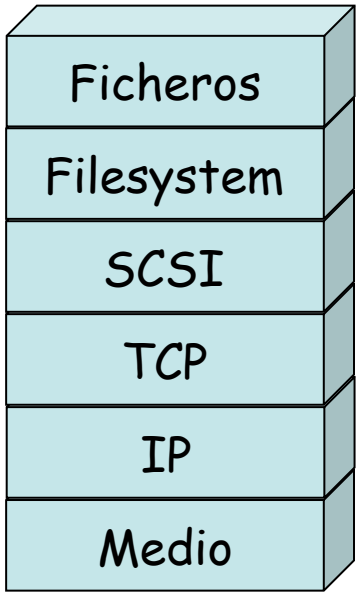
http://www.cisco.com/c/en/us/products/collateral/storage-networking/mds-9700-series-multilayer-directors/guide-c07-732733.pdf

Otros transportes de SCSI

iSCSI

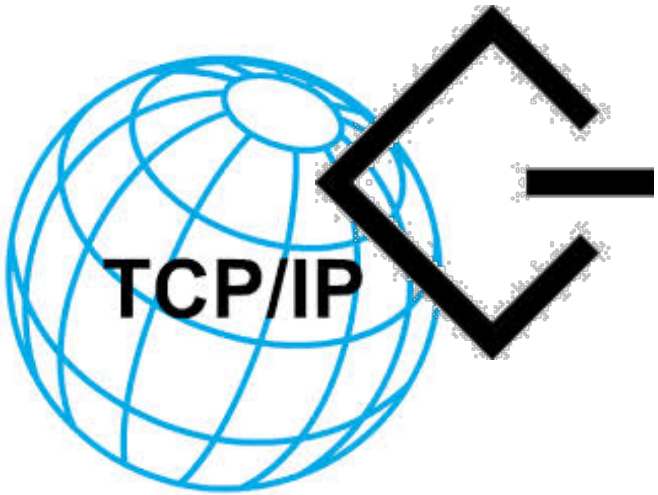
- RFC 7143 *"Internet Small Computer System Interface (iSCSI) Protocol (Consolidated)"*
- Transporte de comandos SCSI sobre una (o varias) conexiones TCP
- Permite crear una SAN atravesando una red IP
- Es una alternativa de bajo coste a Fibre Channel
- El coste puede estar en el rendimiento (retardo, pérdidas, congestión)
- Si por debajo empleamos Ethernet no ha sido interesante hasta que Ethernet ha alcanzado las velocidades de FC

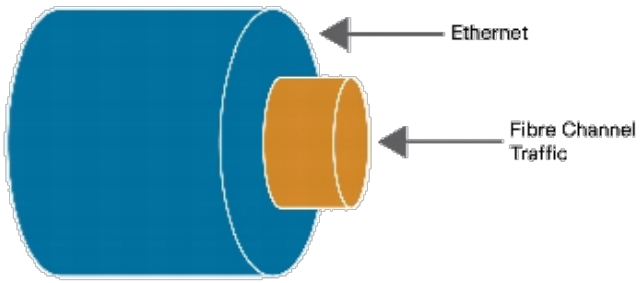




iSCSI vs FCoE

- Ethernet ya tiene velocidades comparables a FC
- Pero hay otra alternativa que es FCoE
- FCoE permite una integración nativa con equipamiento FC pues transporta sus tramas
- iSCSI no transporta las tramas FC sino comandos SCSI
- Hace falta una pasarela para integrar interfaces iSCSI con FC, con sus problemas de escalabilidad y puntos de fallo
- La gestión cambia entre FC e iSCSI
- iSCSI puede ser más apropiado en un escenario *greenfield*





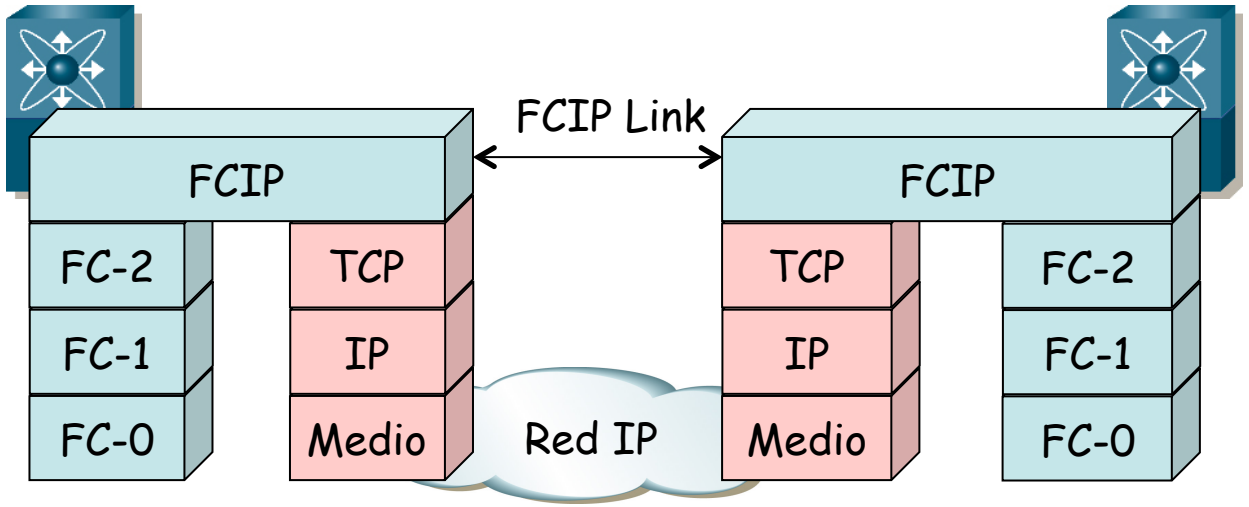
Otras soluciones

FCIP

- RFC 3821 “Fibre Channel Over TCP/IP (FCIP)”
- Es transparente para los equipos FC, solo clases 2, 3 y 4
- Interconexión de islas SAN a través de una red IP; una (o más) conexiones TCP
- Por ejemplo interconexión de SANs en diferentes DCs para replicación de datos
- Los equipos suelen permitir ajustar los parámetros de IP y TCP (timer de retransmisiones, control de flujo, control de congestión, etc)
- También diversos “hacks” para acelerar el protocolo FC

iFCP

- RFC 4172 “iFCP – A Protocol for Internet Fibre Channel Storage Networking”
- Solo clases 2 y 3



FC vs FCoE vs iSCSI

- ¿Retardo en la red por las cabeceras?
  - Despreciable en comparación con los tiempos de acceso al disco
- ¿Gestión y configuración en los hosts?
  - Similar en FC y FCoE
  - Superior en iSCSI pues muchas veces hay que activarlo en el HBA y además hay que configurar IP
- ¿Gestión y configuración en la red?
  - Similar entre FC e iSCSI
  - Superior en FCoE por configurar DCB