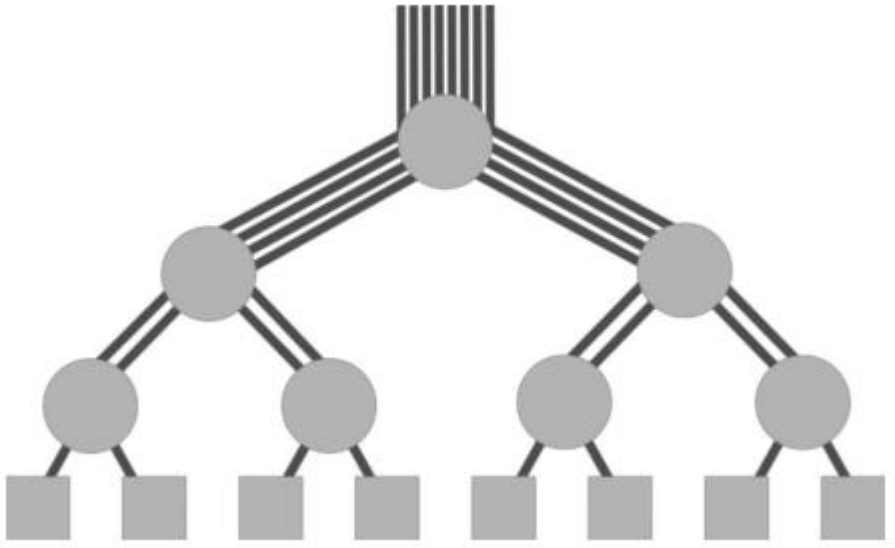
Interconexión escalable

Dept. Arquitectura de Computadores
Universidad de Málaga

Curso 2013/14

Switched fabric

- Interconectar una gran número de puntos finales (estaciones, servidores, blades, …) usando conmutadores (switches) con un número limitado de puertos (conectores).
- Usando una topología de interconexión adecuada para los switches (llamado en inglés switched fabric), se pueden conectar una gran cantidad de puntos finales.
- Topología Fat-Tree
  - Topología en árbol.
  - Los elementos finales están en las hojas del árbol
  - Un nodo dedica el mismo número de enlaces a sus hijos que a su padre
  - Red sin bloqueo

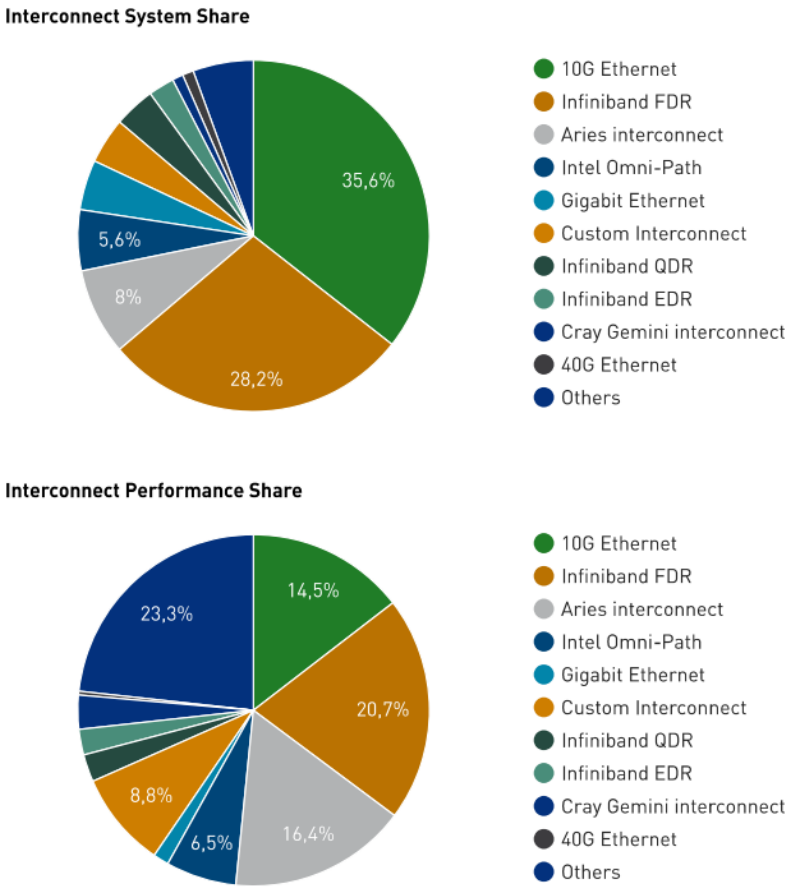


Switches

- Elemento activo de red que permite desarrollar distintas topologías de interconexión, entre ellas Fat-Tree.
- Realizan el rutado de mensajes desde cualquier puerto fuente a cualquier puerto destino usando tablas de rutado.
  - Ancho de banda agregado para permitir la conmutación simultánea en todos los puertos (sin bloqueo).
- El formato exacto de la tablas, así como su contenido y organización depende del estándar usado y del fabricante.
- El switch aporta escalabilidad el diseño de una infraestructura de interconexión permitiendo conectar nodos y otro switches a dicha infraestructura.
- El swicth maneja el tráfico en la red analizando las cabeceras de cada trama recibida y reenviándola al destino adecuado.

Tecnologías de interconexión

- La dos tecnologías de interconexión más populares para clusters son FRD Infiniband y 10Gb Ethernet
  - Topologías de interconexión son construidas utilizando switches



Top500: november 2016

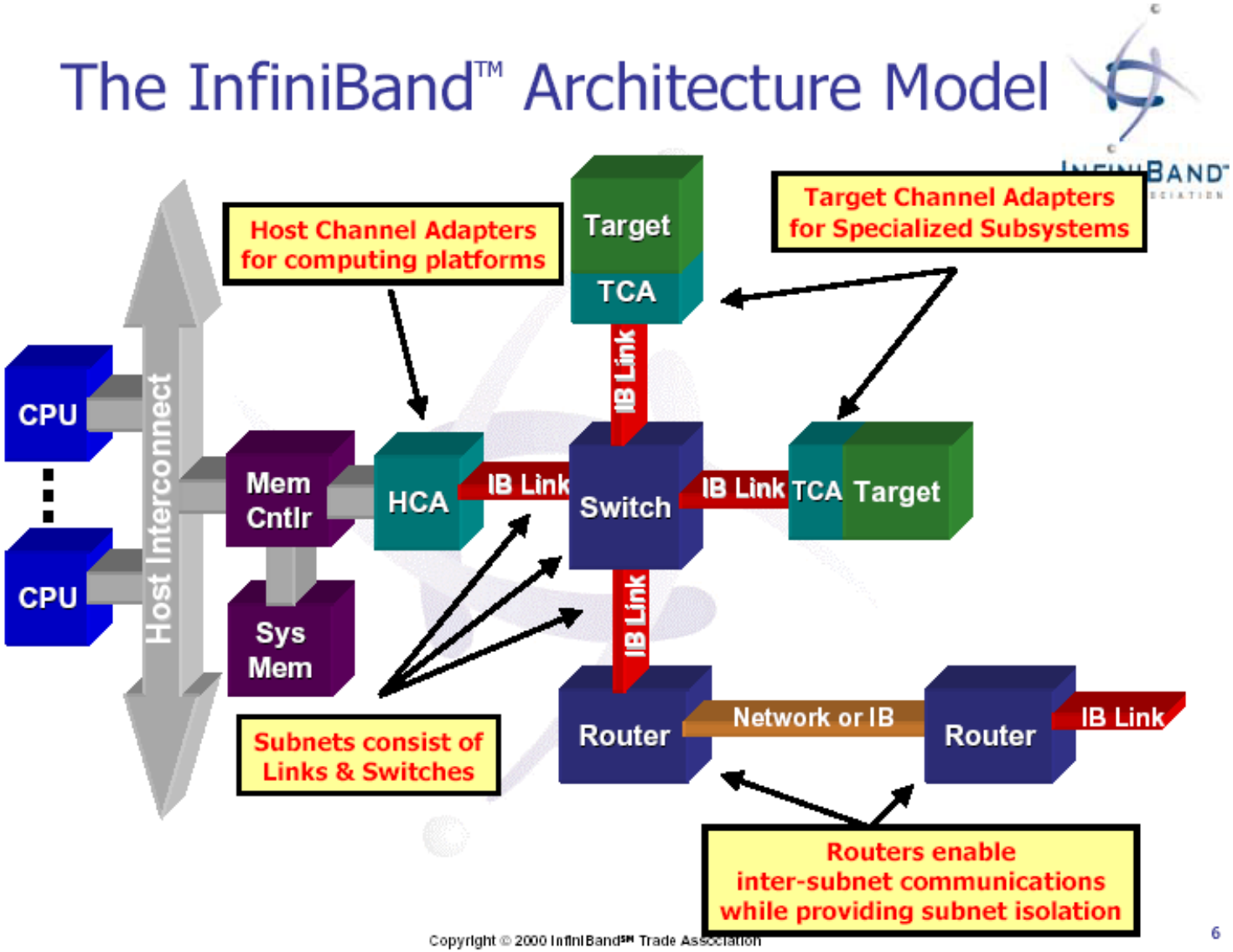
InfiniBand

- Nuevo estándar en la industria, que sustituye el bus E/S tradicional por una red de conmutadores basados en canales, que interconecta unidades de procesamiento con dispositivos de E/S
  - Motivado por el gap creciente entre la velocidad de E/S y la velocidad del procesador-memoria
  - InfiniBand pretende soportar la necesidad de comunicaciones y E/S de altas prestaciones que la era Internet exige
  - Ofrece una visión integrada de computación, interconexión y almacenamiento
  - Soportada por un consorcio de la empresas más importantes en el campo: IBM, Sun, HP-Compaq, Intel, Microsoft, Dell, …
- Configuraciones cable de cobre y fibra óptica:

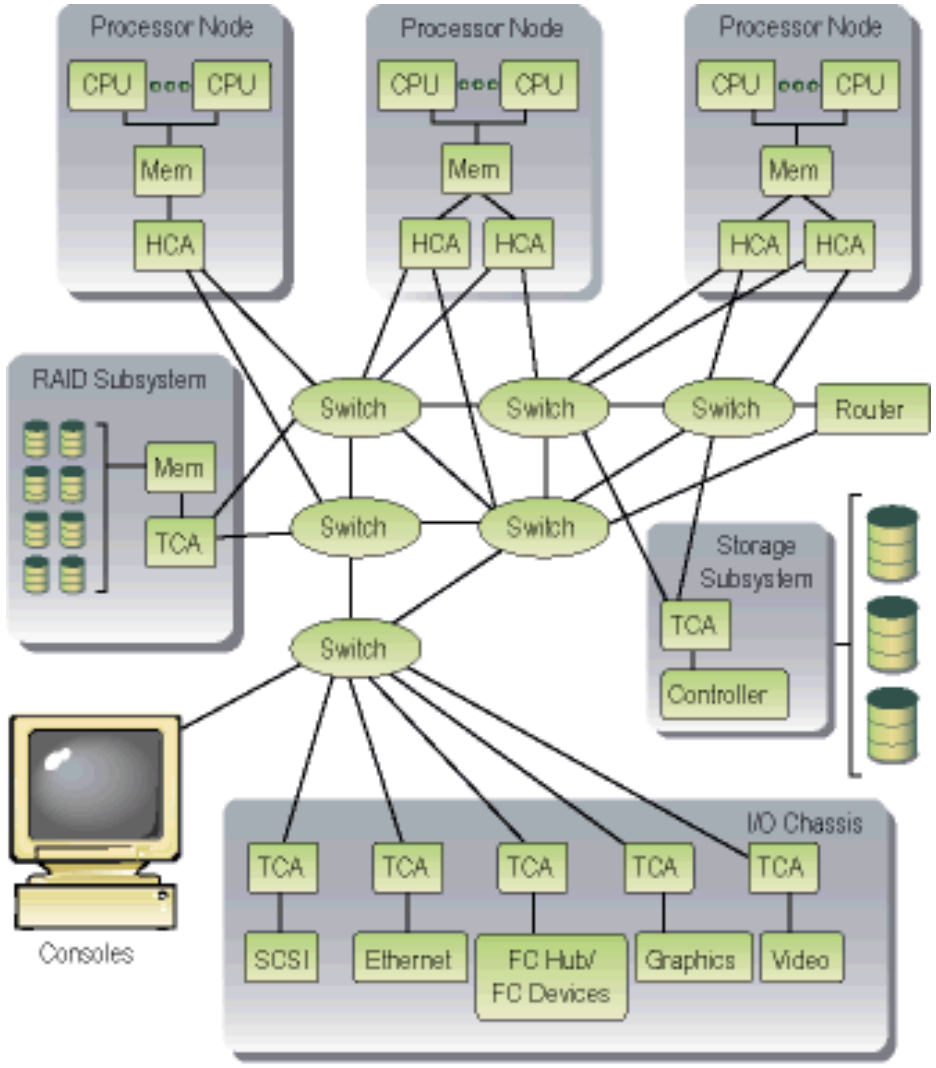
Characteristics	SDR	DDR	QDR	FDR-10	FDR	EDR	HDR	NDR
Signaling rate (Gbit/s)	2.5	5	10	10.3125	14.0625[6]	25	50	100
Theoretical effective throughput, Gbs, per 1x[7]	2	4	8	10	13.64	24.24		
Speeds for 4x links (Gbit/s)	8	16	32	40	54.54	96.97		
Speeds for 12x links (Gbit/s)	24	48	96	120	163.64	290.91		
Encoding (bits)	8/10	8/10	8/10	64/66	64/66	64/66		
Adapter latency (microseconds)[8]	5	2.5	1.3	0.7	0.7	0.5		
Year[9]	2001, 2003	2005	2007		2011	2014[7]	~2017[7]	after 2020

Wikipedia

InfiniBand


The InfiniBand™ Architecture Model
Host Channel Adapters for computing platforms
Target
TCA
Target Channel Adapters for Specialized Subsystems
IB Link
CPU
CPU
Host Interconnect
Mem Cntlr
HCA
IB Link
Switch
IB Link
TCA
Target
Sys Mem
IB Link
Router
Network or IB
Router
IB Link
Subnets consist of Links & Switches
Routers enable inter-subnet communications while providing subnet isolation
Copyright © 2000 InfiniBand℠ Trade Association
6

InfiniBand



Source: InfiniBand Architectural Overview

- Ya no hay bus E/S
- Todos los sistemas se interconectan mediante adaptadores HCA o TCA
- La red permite múltiples transferencias de datos paquetizados
- Permite RDMA (Remote Memory Access Read or Write)
- Implica modificaciones en el software de sistema

Comparación Infiniband y Ethernet

- Latencia: tiempo utilizado por un paquete de tamaño cero en ser transmitido desde un proceso en un nodo a un proceso en otro nodo atravesando NIC-switch-NIC
- Bandwith/Throughput: Velocidad máxima real alcanzada en la transmisión de paquetes a través de la red.
- N/2:the smallest packet size that reaches full network speed in one direction. Importante si las aplicaciones envían paquetes pequeños.

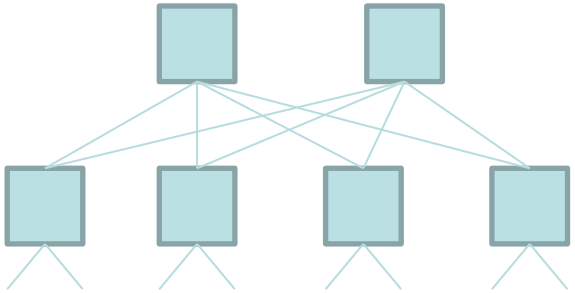
Network Solution	N/2 (bytes)	Maximum BW (MB/s)	Latency (µs)
GigE	12,300	112	47.61
10GigE	98,300	875	12.51
DDR InfiniBand	12,285	1482	1.72
QDR InfiniBand	32,765	3230	1.67

Fat-tree con switches comerciales

- La topología Fat-Tree requiere que los switches tengan un número variable de puertos en función del nivel que ocupan en el árbol.
  - No es un opción en un entorno real, donde los switches tienen configuraciones fijas de número de puertos
- Se proponen alternativas usando conmutadores comerciales
  - Clos topology
- Si el número de enlaces de los edge switches que se conectan a los nodos es el mismo que se conecta a los niveles superiores (fat-tree), entonces la red es sin bloqueo.
  - En caso contrario (más enlaces a los nodos que a siwtches de nivel superior), la red será bloqueante
    - » No todos los nodos se pueden interconectar simultáneamente
    - » Se hace para ahorrar en la red: switches y cables

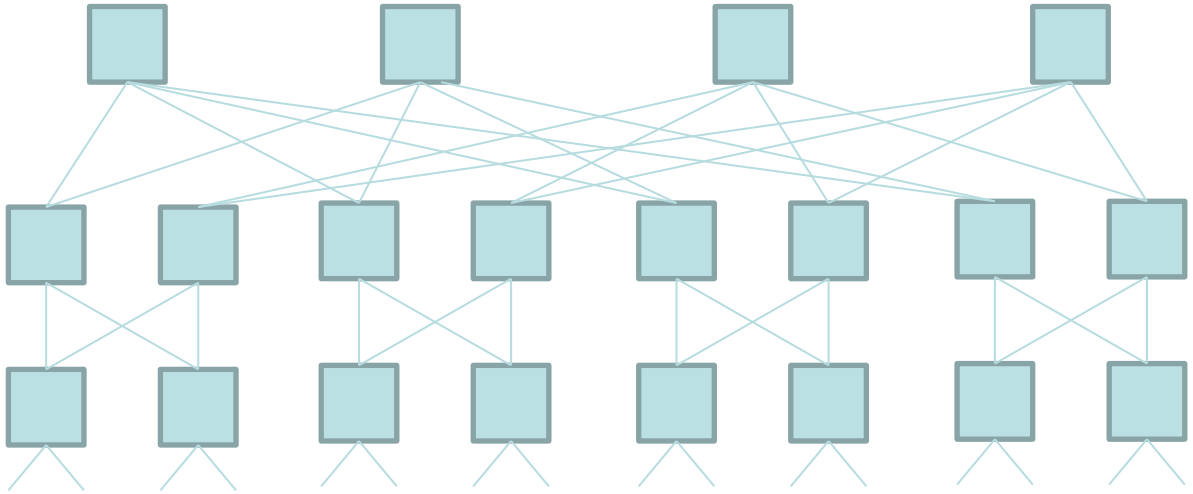
Switches de cuatro puertos (I)

- Dos niveles (sin bloqueo): conecta 8 nodos



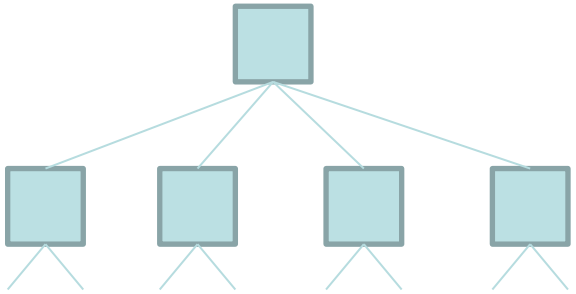
Rutas alternativas para conexiones entre nodos

- Tres niveles (sin bloqueo): conecta 16 nodos



Switches de cuatro puertos (II)

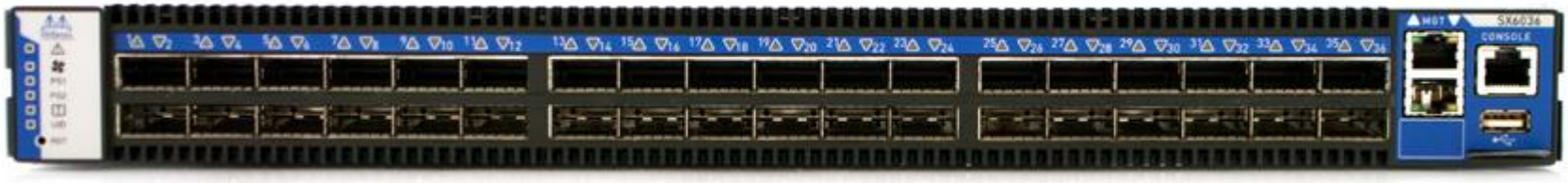
- Dos niveles con bloqueo



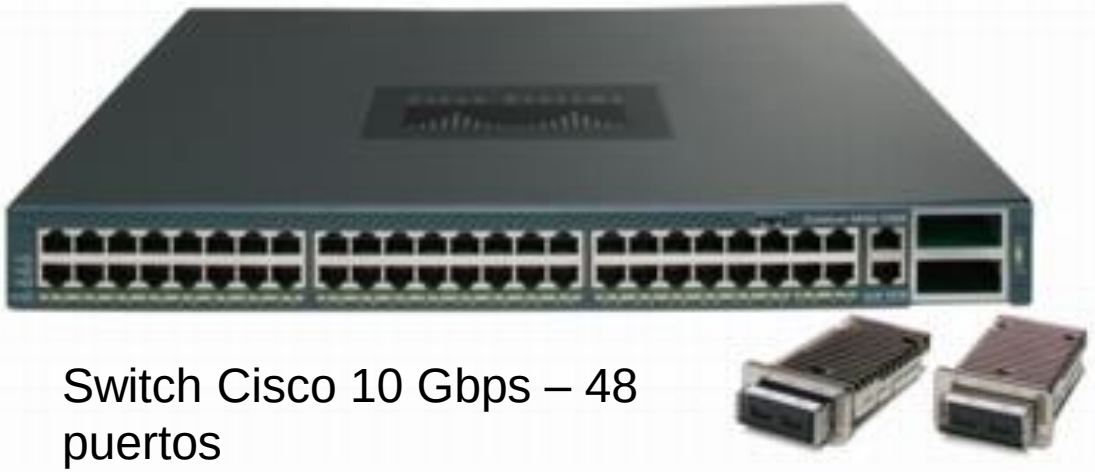
No hay rutas alternativas para conexiones entre nodos

Switches

- Un switch es un conmutador de tramas compuesto por una serie de puertos.

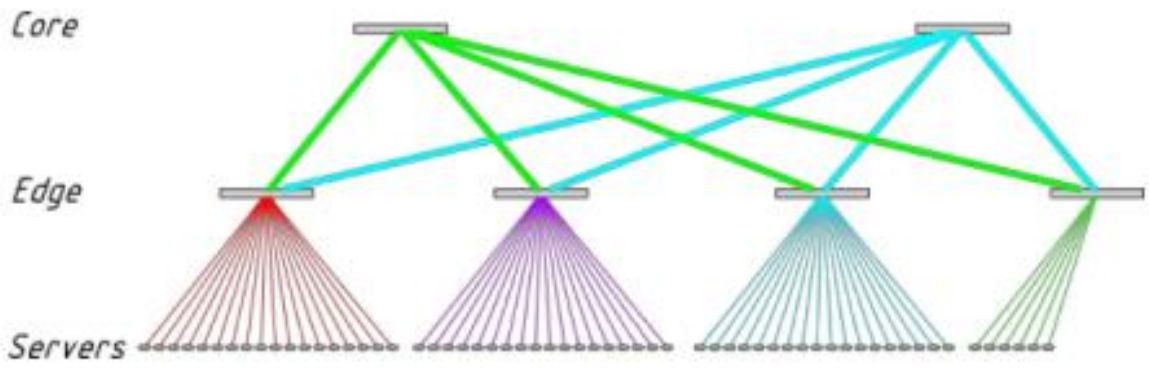


Switch Infiniband de Mellanox – 36 puertos QSFP (8Gbps)



Switch Cisco 10 Gbps – 48 puertos

Fat-tree de dos niveles (sin bloqueo)



Switches de 36 puertos

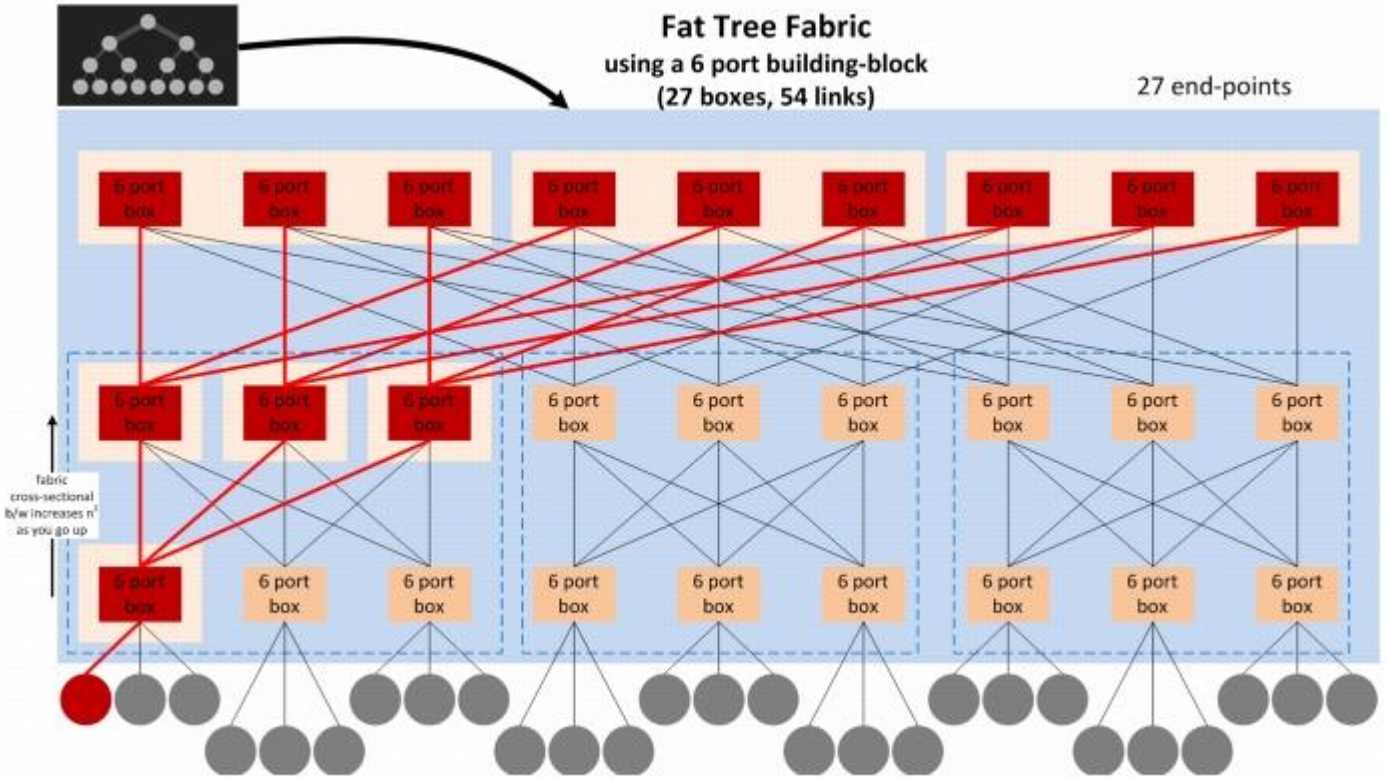
- Si dos nodos están conectados al mismo edge switch, se pueden comunicar directamente a través de él.
- En caso contrario, el intercambio de información se producirá a través de los core switches.
- En una red de dos niveles se puede conectar Pe*Pc/2 elementos finales
  - Pe: puertos del conmutador en el edge
  - Pc: puertos del conmutador en el core
    - » Si usamos switches comerciales de 36 puertos podemos conectar 648 máquinas

Fat-tree con más niveles

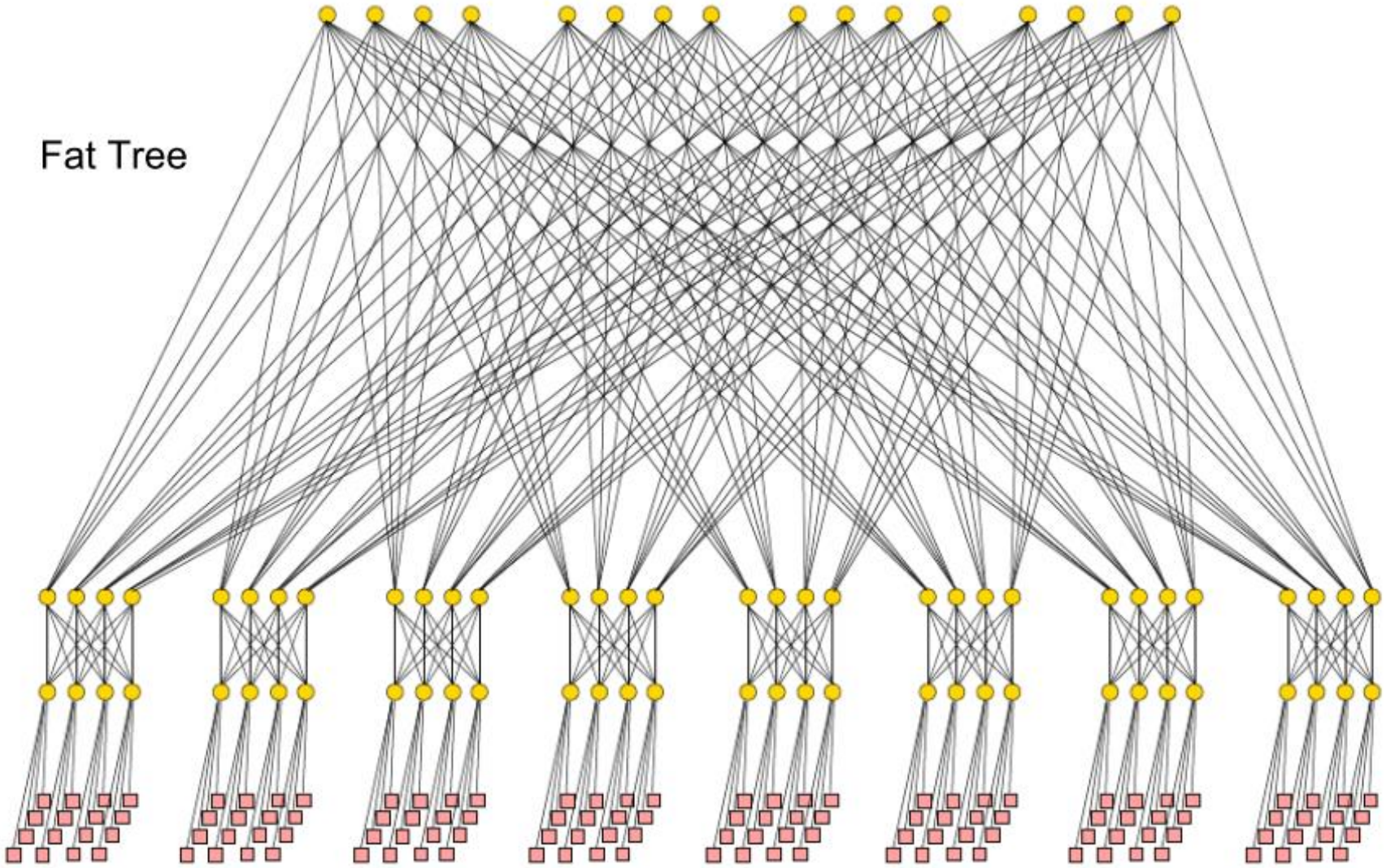
- Por cada nivel adicional se multiplica el número de nodos por P
- Tres niveles Nmax = Pe*Pc/2*Pc/2
- Número máximo de nodos con una topología en tres niveles usando switches de 36 puertos = 11664
  - Aumenta la latencia

Fat tree con switches de 6 puertos

- 3 niveles: puede conectar 54 nodos

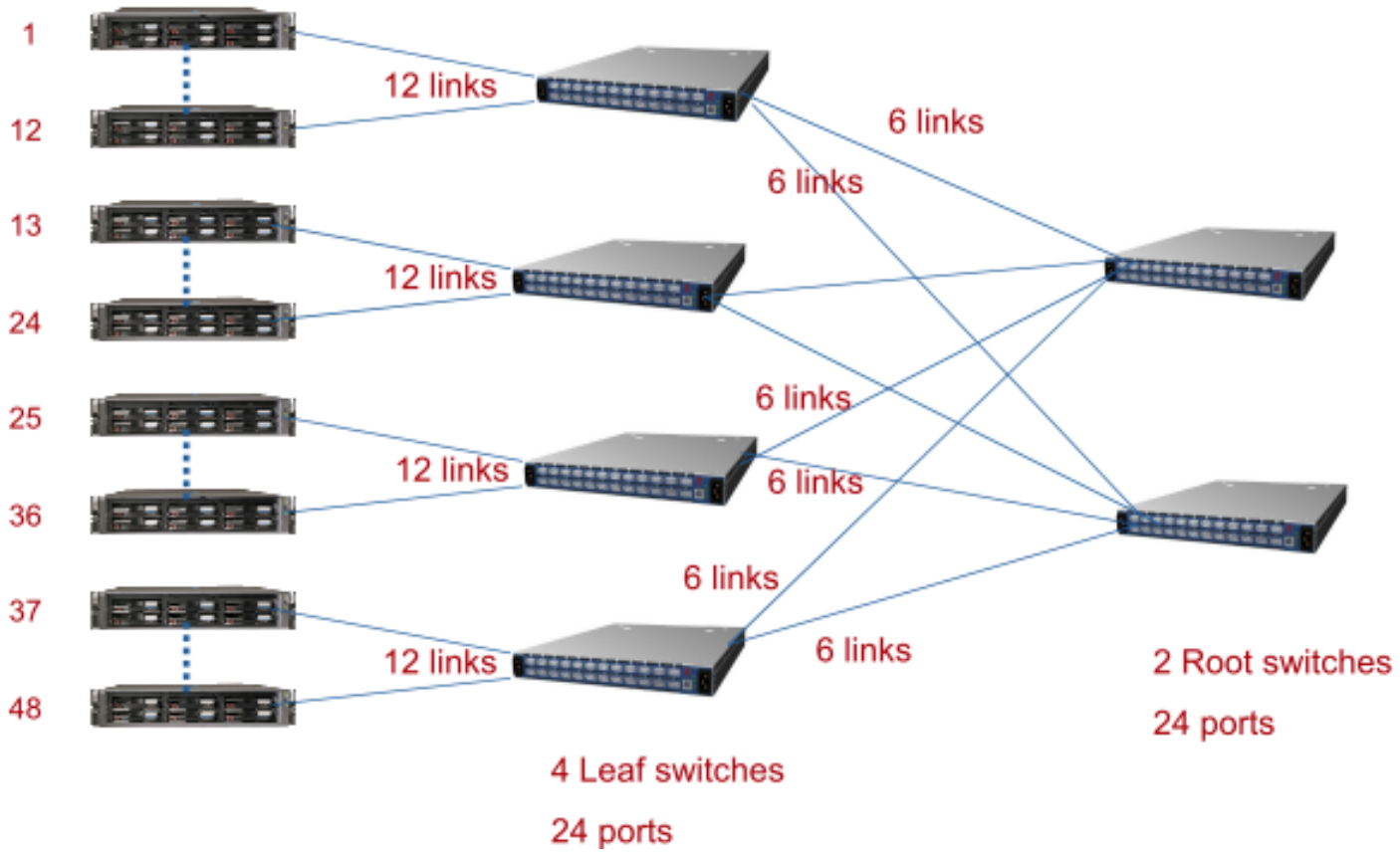


Fat-Tree con switches de 8 puertos



Ejemplo real

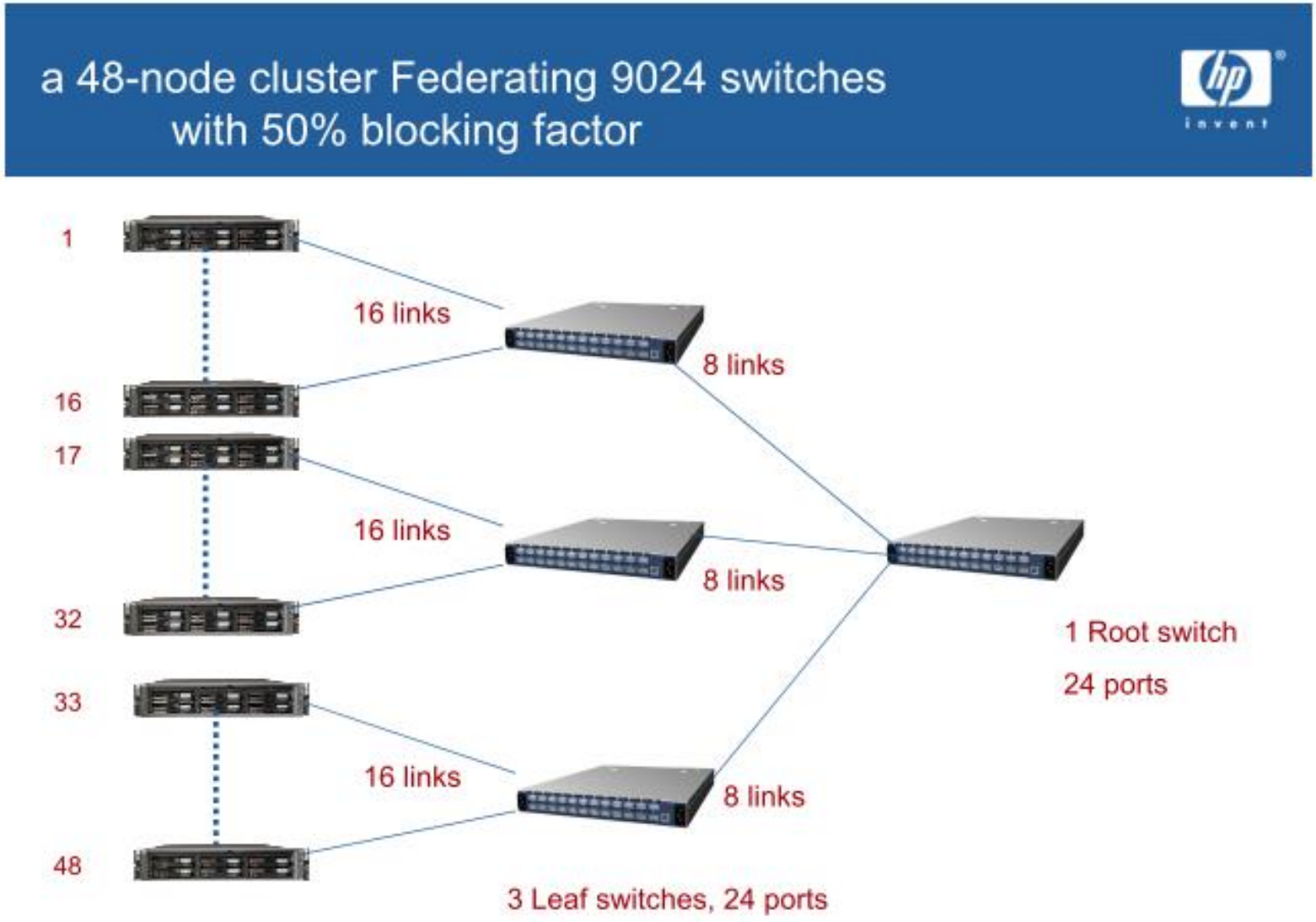
a 48-node cluster Federating 9024 switches
using full Non-blocking bandwidth



Interconexión en fat-tree con bloqueo

- Para hacer las redes más baratas, se reduce el número de switches, incrementando el número de enlaces dirigidos a los nodos.
- Si el número de enlaces de los edge switches que se conectan a los nodos es el mismo que se conecta a los niveles superiores, entonces la red es sin bloqueo.
- Si embargo si la proporción es diferente, el red tendrá bloqueos: dos conexiones que podrían seguir caminos separados ahora comparten algún enlace.
  - Por ejemplo en la proporción 2:1 (dos veces más enlaces a los nodos),el factor de bloqueo será de dos.
- Con un factor de bloqueo de Bl, un edge switch tiene Pe*Bl(Bl+1) enlaces a los nodos y una red puede tener 2*Bl/(Bl+1) nodos adicionales.
- Sin impacto en la latencia

Ejemplo

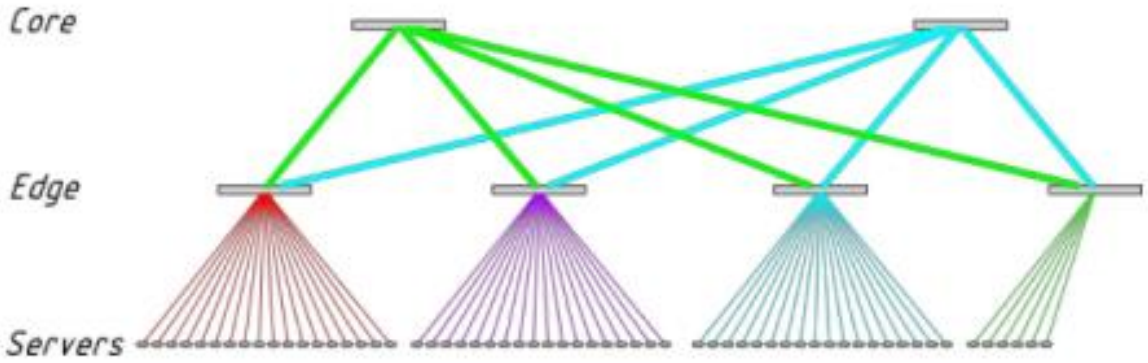

a 48-node cluster Federating 9024 switches
with 50% blocking factor
1
16
17
32
33
48
16 links
16 links
16 links
8 links
8 links
8 links
1 Root switch
24 ports
3 Leaf switches, 24 ports

Switch modular

- Solución integrada para la interconexión de gran número de nodos.
  - Rack con switches comerciales
- Los módulos suelen estar interconectados internamente usando Fat-tree de dos niveles.
- Varios switches modulares se pueden interconectar, a su vez, en una red Fat-Tree de dos niveles
  - Cuatro niveles en total contando con los dos niveles internos.
  - Con switches de 864 puertos se pueden interconectar hasta 373.248 nodos

QLogic 12800
18–864 Port, 40Gbps
Infiniband, (QDR)

Tolerancia a fallos



- La topología de la red es tolerante a fallos si éstos ocurren en los routers ubicados en el core.
  - Aunque el fallo de uno de ellos convierte a la red en bloqueante
- Sin embargo, el fallo de un switch del edge impide que los host conectados a ese switch se puedan comunicar
  - Posible solución dotando a los nodos de doble puerto que requiere ,además, replicar el número de switches en el edge (dual-plane connection)

Configurador Fat-Tree

- http://www.mellanox.com/clusterconfig/

Ejemplo: Fat-tree ethernet

- Necesidad de uso de Spanning Tree para evitar bucles en las conexiones
  - La red será tolerante a fallos pero tendrá bloqueos