Arquitecturas de Computador Distribuidas

Dept. Arquitectura de Computadores
Universidad de Málaga

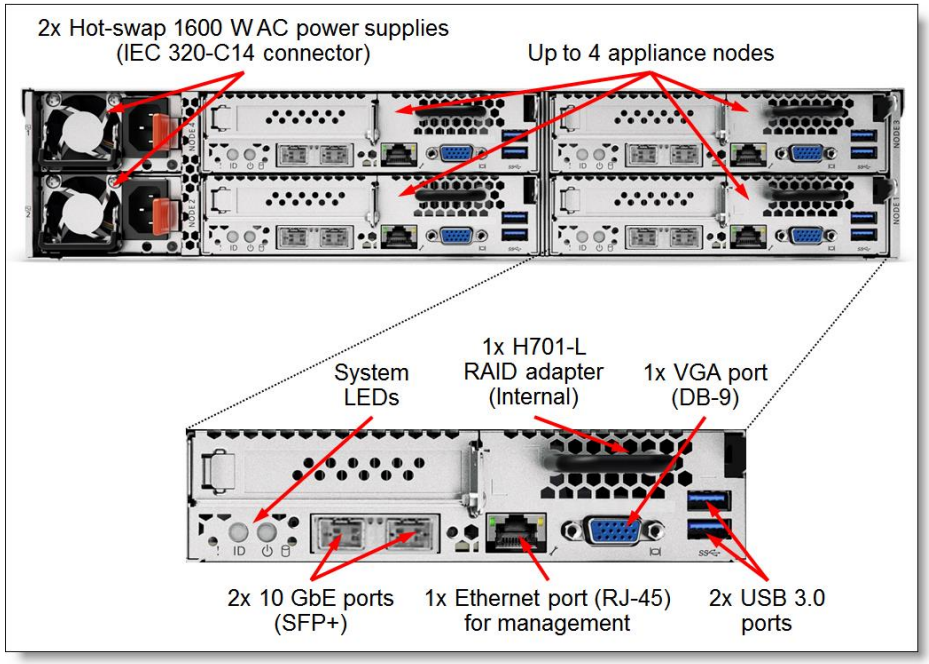
Curso 2016/2017

Definición de clúster de computación

- Un clúster es una tipo de arquitectura paralela distribuida que consiste de un conjunto de computadores independientes interconectados operando de forma conjunta como un único recurso computacional
  - Consta de nodos de computación, red de interconexión y sistema de almacenamiento
  - Uso de un cluster:
    - » Computación de alta prestaciones
      - – Nodos de computación muy potentes unidos a través de una red de interconexión de alta velocidad: cluster para computación paralela
    - » Alta productividad
      - – Incrementa el número de operaciones (procesos, transacciones, …) ejecutadas por unidad de tiempo: clúster de base de datos, granja de servidores, …
    - » Alta disponibilidad (gracias a la redundancia de componentes)
      - – En caso de fallo, el clúster puede ser todavía operacional: Failover clúster

Evolución

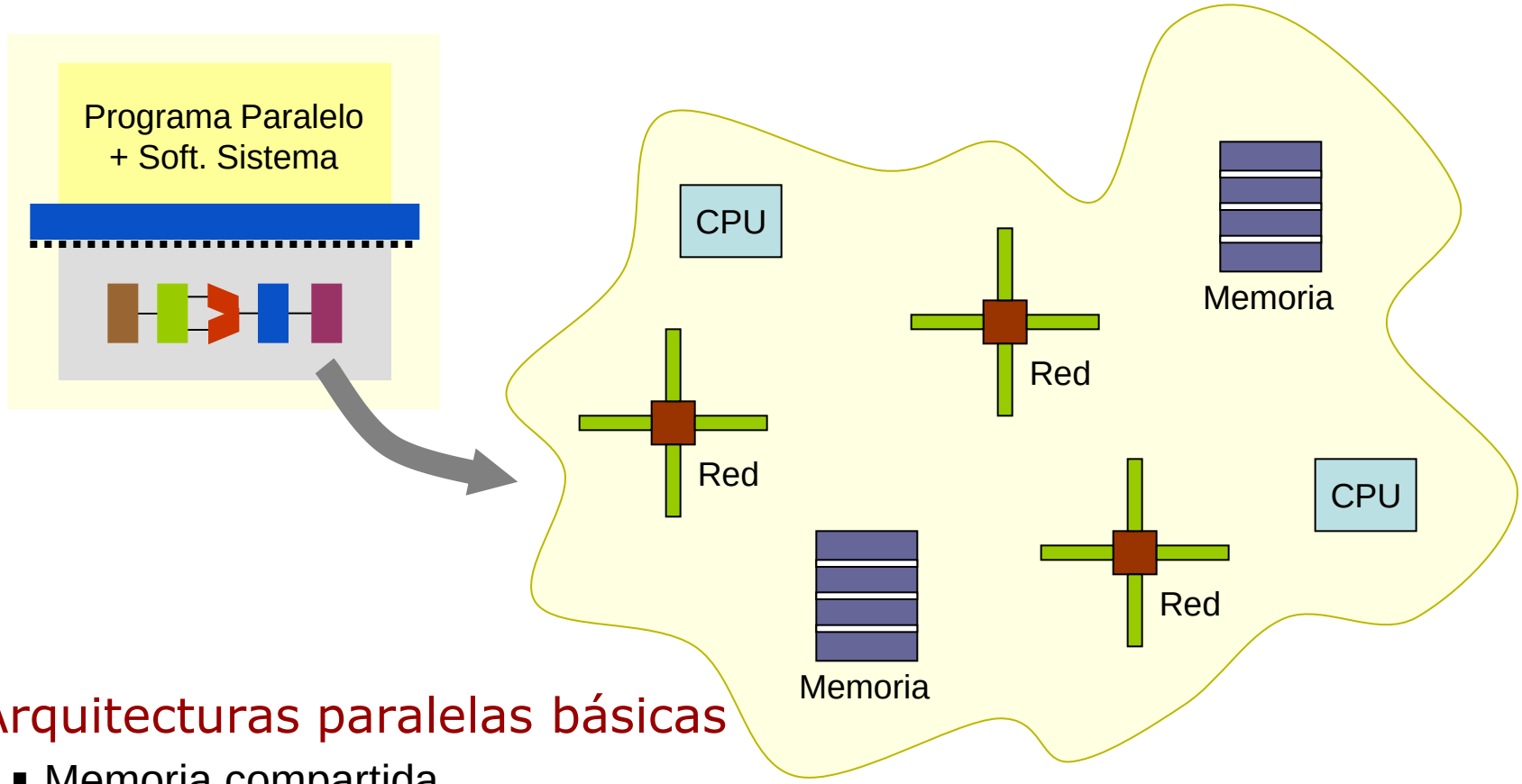
Cluster of Workstations



Two Xeon processors per node (22 cores per processor)

Clúster con componentes modernos enracables

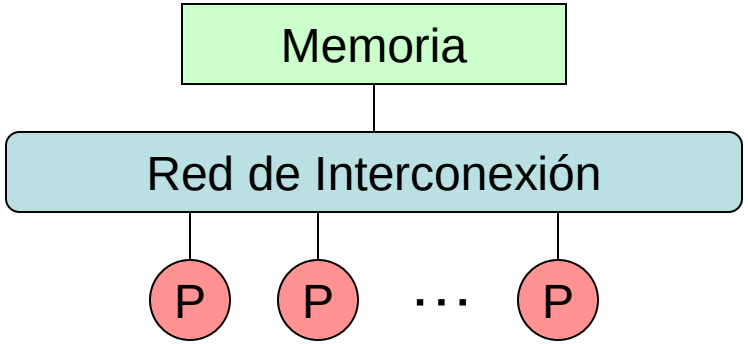
Arquitecturas Paralelas



- Arquitecturas paralelas básicas
  - Memoria compartida
  - Memoria local (pase de mensajes)

Arquitecturas de memoria compartida

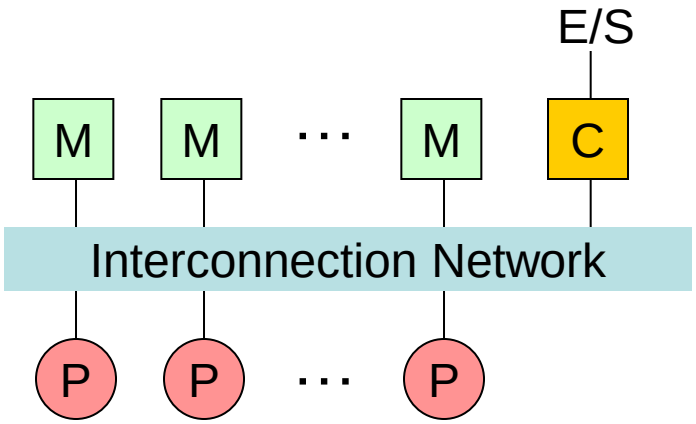
- Arquitectura de memoria compartida
  - **Todos los procesadores del sistema pueden acceder directamente a todas las posiciones de memoria**
  - Es decir, el **espacio de memoria física es único y global**



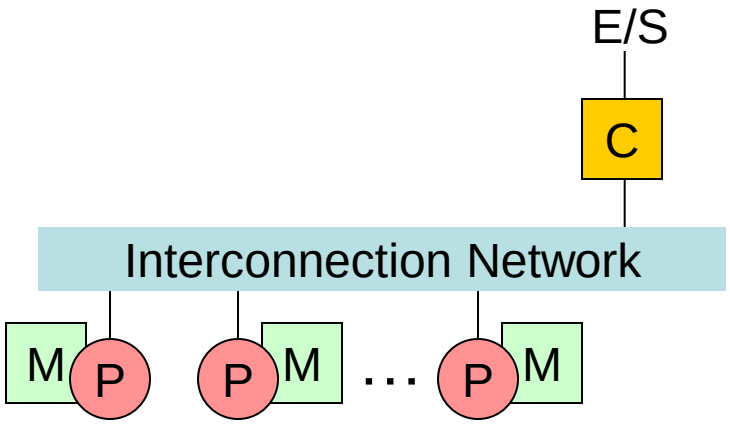
- Mecanismo de Comunicación
  - Implícito, como resultado de instrucciones de acceso a memoria (*load/store*)
  - La comunicación está integrada en el sistema de memoria

Arquitecturas de memoria compartida

- Symmetric Multiprocessor (SMP or UMA)
  - La latencia y ancho de banda a cualquier posición de memoria es a misma para todos los procesadores.
- Scalable Multiprocessor (NUMA y ccNUMA)
  - La latencia y ancho de banda depende de la zona de memoria a la que esté accediendo el procesador



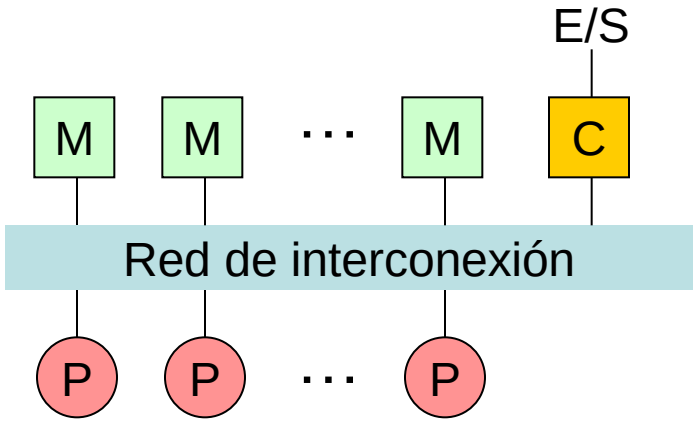
SMP or UMA



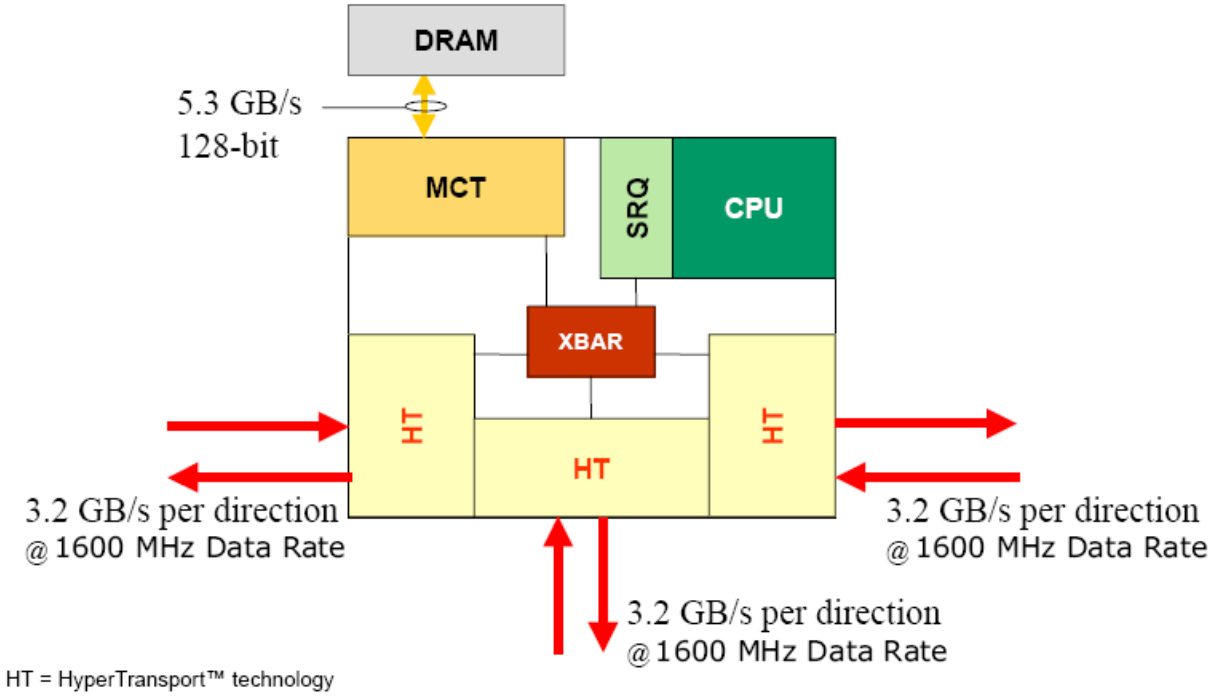
NUMA

Multiprocesadores UMA

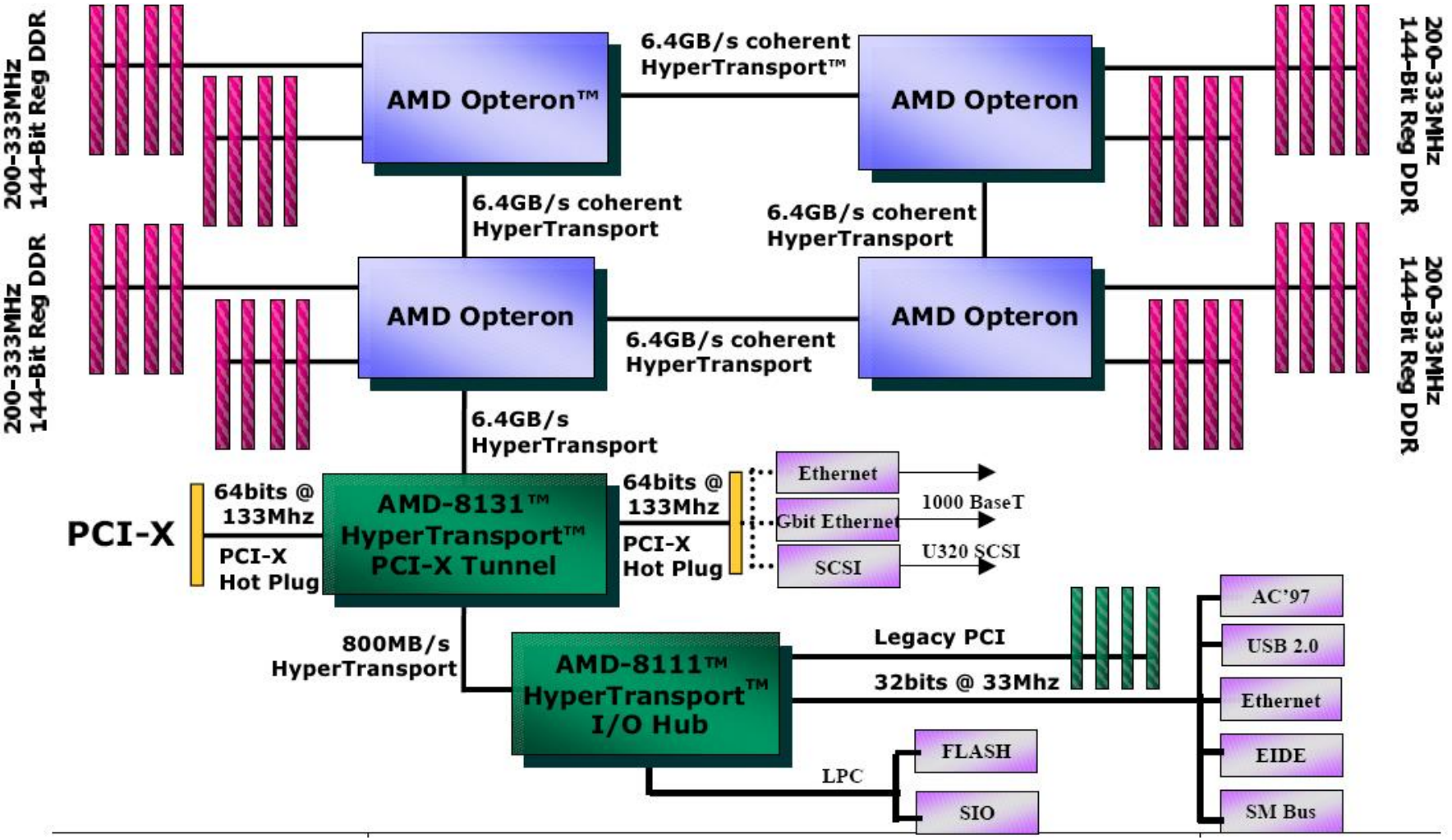
- UMA: Uniform Memory Access
- Características
  - Los procesadores y módulos de memoria se conectan entre sí compartiendo la red de interconexión
  - Cuando la red es un bus común, se llaman **multiprocesadores SMP** (**Symmetric MultiProcessor**)
  - La red limita el ancho de banda de las comunicaciones, aunque el uso de caches privadas reducen el problema
  - Las caches privadas introducen el problema de la **coherencia cache**
  - El coste es reducido pero su escalabilidad es media o baja



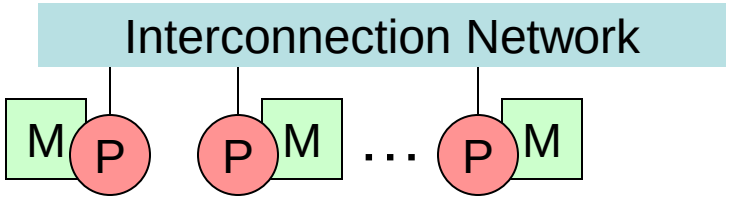
UMA: AMD Opteron



UMA: Quad AMD Opteron Workstation



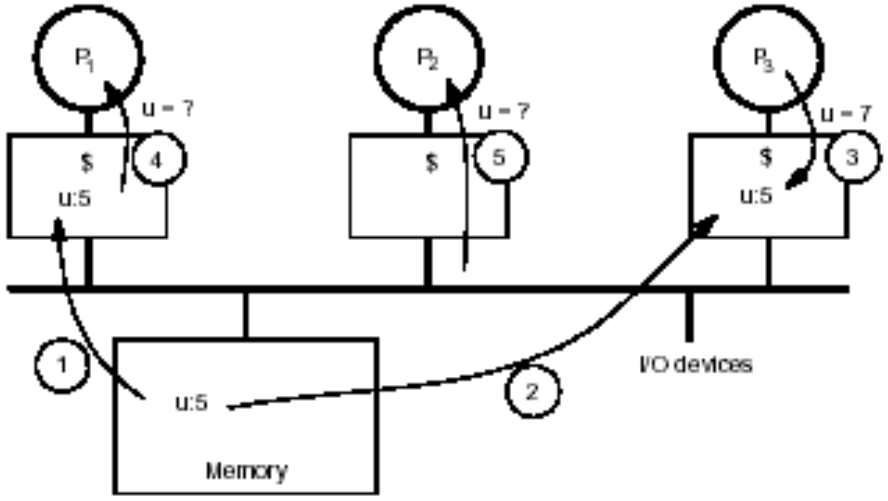
Multiprocesadores NUMA



- Los multiprocesadores UMA son poco escalables
  - La latencia de acceso a memoria es alta
  - Todo el tráfico debido a los fallos cache atraviesan la red
- Multiprocesadores escalables
  - Distribuyendo físicamente la memoria entre los procesadores permite que muchos fallos cache se resuelvan sin pasar por la red
  - Se llaman **multiprocesadores NUMA** (***Non Uniform Memory Access***) o **ccNUMA** (***Cache Coherent NUMA***)
  - El inconveniente es que la latencia de acceso a memoria es **variable**, y se requiere un **distribución** adecuada de los datos en las memorias locales
  - Las caches privadas introducen el problema de la **coherencia cache**, que puede evitarse (NUMA) o resolverse (ccNUMA)
  - El coste es alto y su escalabilidad es media (debido al tráfico extra en la red debido a la coherencia cache) o alta (si no hay coherencia hardware)

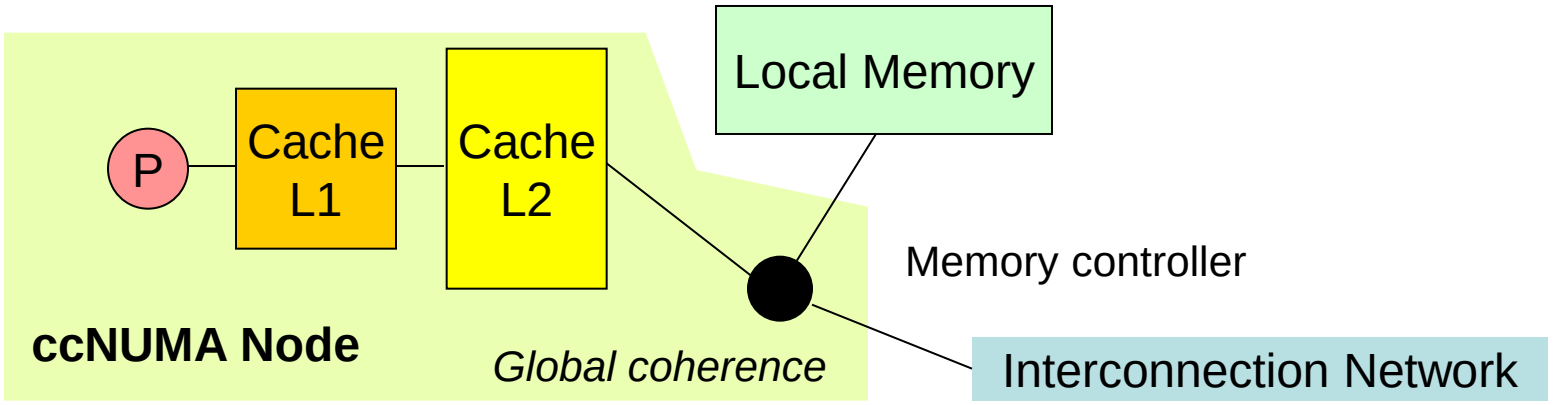
Problema de coherencia cache

- El uso de caches incrementa el rendimiento al reducir el acceso a la memoria principal del computador paralelo
- En el caso de arquitecturas de memoria compartida, todos los procesadores comparten el mismo espacio de direccionamiento.
  - Es posible que más de un procesador tenga en cache la misma dirección al mismo tiempo
- Si un procesador actualiza el contenido de esa dirección sin informar a los otros procesadores, se produce una inconsistencia de la memoria y la ejecución de una aplicación puede fallar



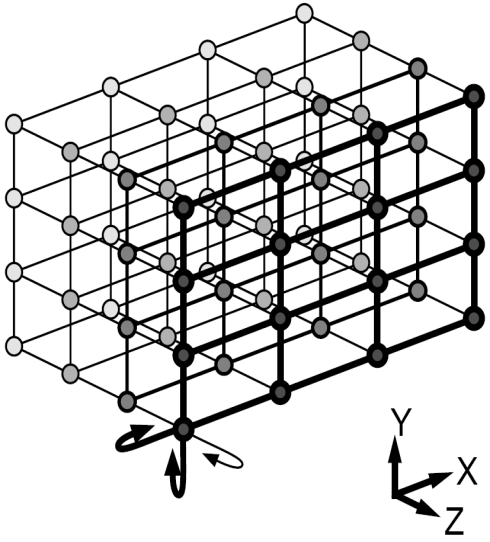
Multiprocesadores y coherencia

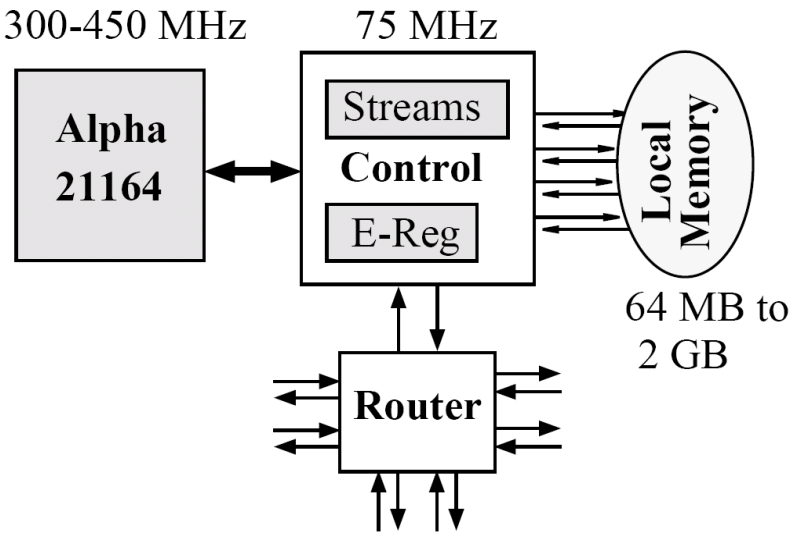
- Arquitecturas sin coherencia cache hardware
  - La memoria está distribuida físicamente entre los procesadores (NUMA)
  - Sólo los accesos a los módulos de memoria locales son cacheados
  - Ejemplos: IBM RP3, BBN Butterfly, Cray T3D/T3E, Cray T90, NEC SX-5
- Arquitecturas con coherencia cache hardware
  - Multiprocesadores con acceso uniforme a memoria (UMA)
    - » Coherencia cache *snoopy* (bus, crossbar)
    - » Bus común: Multiprocesador Simétrico (SMP)
  - Multiprocesadores con memoria distribuida (ccNUMA)
    - » Coherencia cache basada en directorios
  - Ejemplos: Cray T3E, HP Superdome y muchas configuraciones multisocket (Intel Xeon, AMD Opteron)



NUMA: Cray T3E

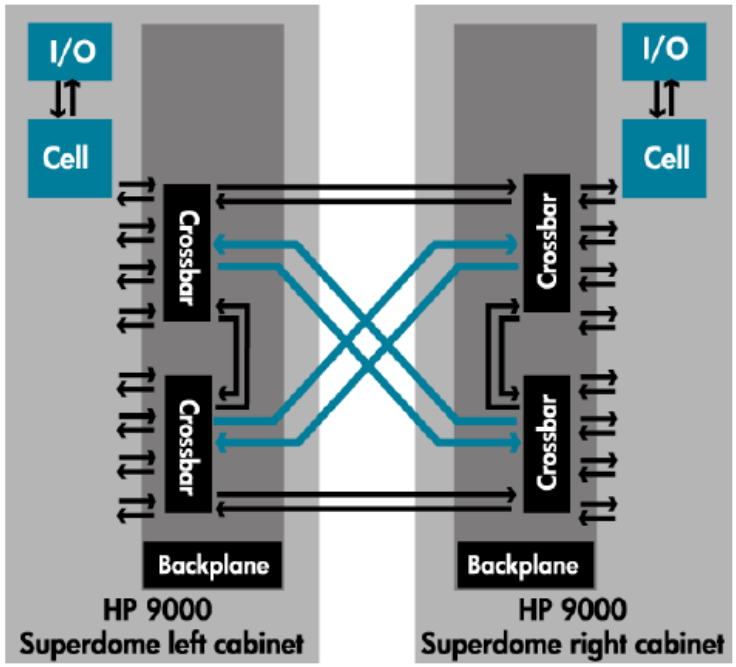
- Sucesor del Cray T3D, e introducido en 1996
  - Hasta 2048 600MHz procesadores Alpha 21164, a 600 MHz
  - Red toro 3D
  - Cada nodo tiene 256MB-2GB de memoria local DRAM
  - Instrucciones Load/Store sólo permiten acceso a la memoria física local
  - Sólo los accesos a memoria local pasan por la cache
  - Los accesos directos a memoria remota (get, put) pasan por los E-Registers

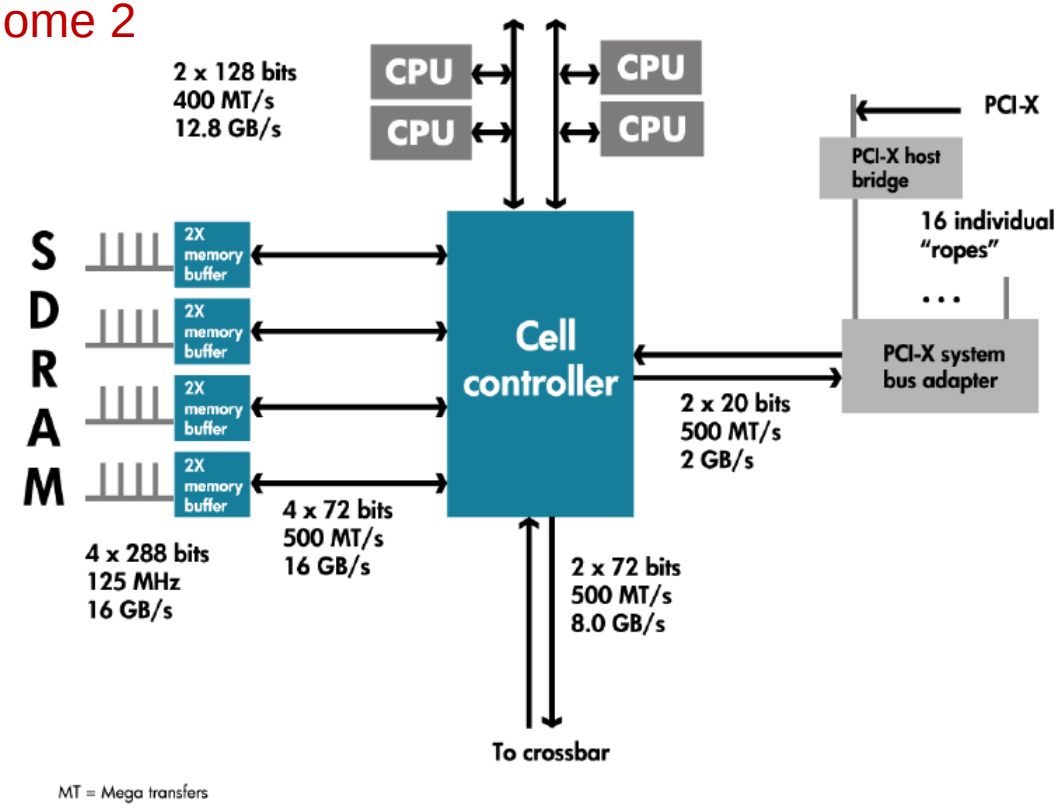




ccNUMA: HP 9000/Integrity Superdome

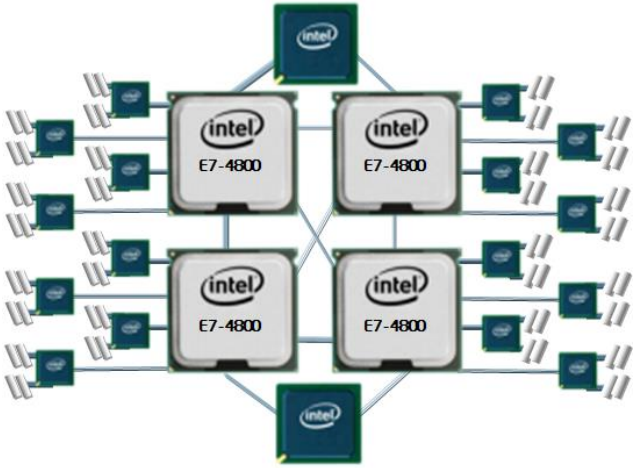
- Introducido en 2000
  - Hasta 128 procesadores Itanium2 y 2TB de memoria
  - Red multietapa basada en routers crossbar
  - Coherencia cache hardware
  - Actualizado en 2010: Superdome 2





NUMA: Intel Xeon

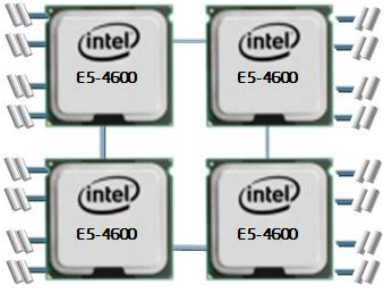
- Latencias en configuraciones multisocket



CPU	0	1	2	3
0	136	194	198	201
1	194	135	194	196
2	201	194	135	200
3	202	197	198	135

Accesos a memoria más uniformes

QPI link: enlace entre sockets

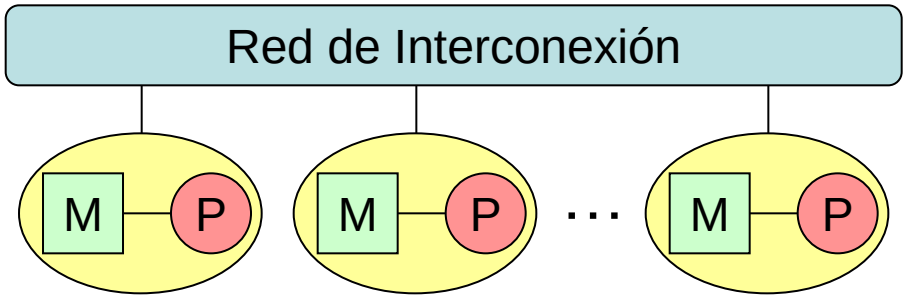


CPU	0	1	2	3
0	72	291	323	294
1	296	72	293	315
2	319	296	71	296
3	290	325	300	71

Más claramente NUMA: accesos locales muy rápidos.

Arquitecturas de memoria distribuida

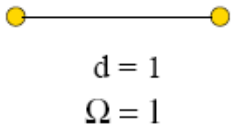
- Arquitectura de memoria privada, o pase de mensajes
  - **Los procesadores del sistema pueden acceder directamente sólo a las posiciones de memoria locales**
  - Es decir, cada procesador tiene su propio **espacio de memoria física privado**

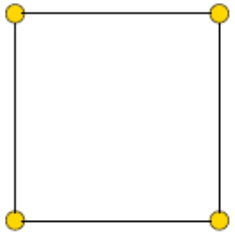


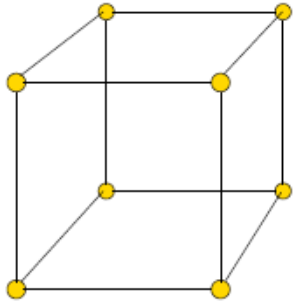
- Mecanismo de Comunicación
  - Operaciones SO explícitas (mensajes *send / receive*)
  - La comunicación está integrada en el sistema de E/S (interfaces de red)
- Dos tipos:
  - MPP (Massively Parallel Processing)
  - Clúster

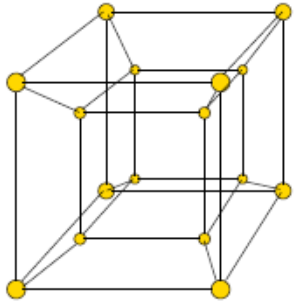
Redes de interconexión

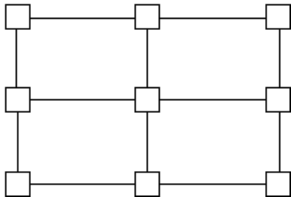
Hipercubes





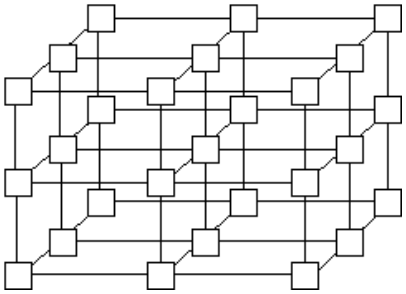


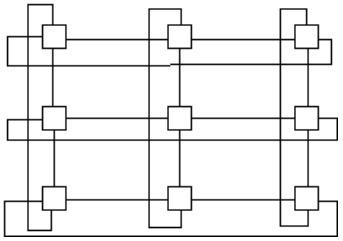




Mesh 2D

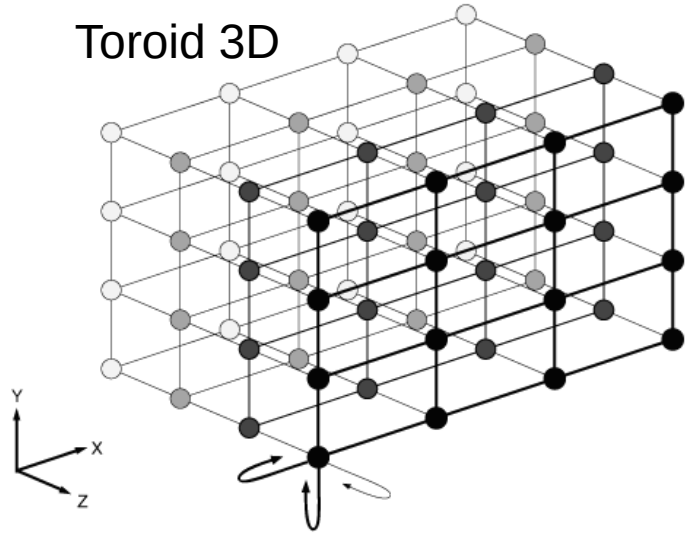
Mesh 3D





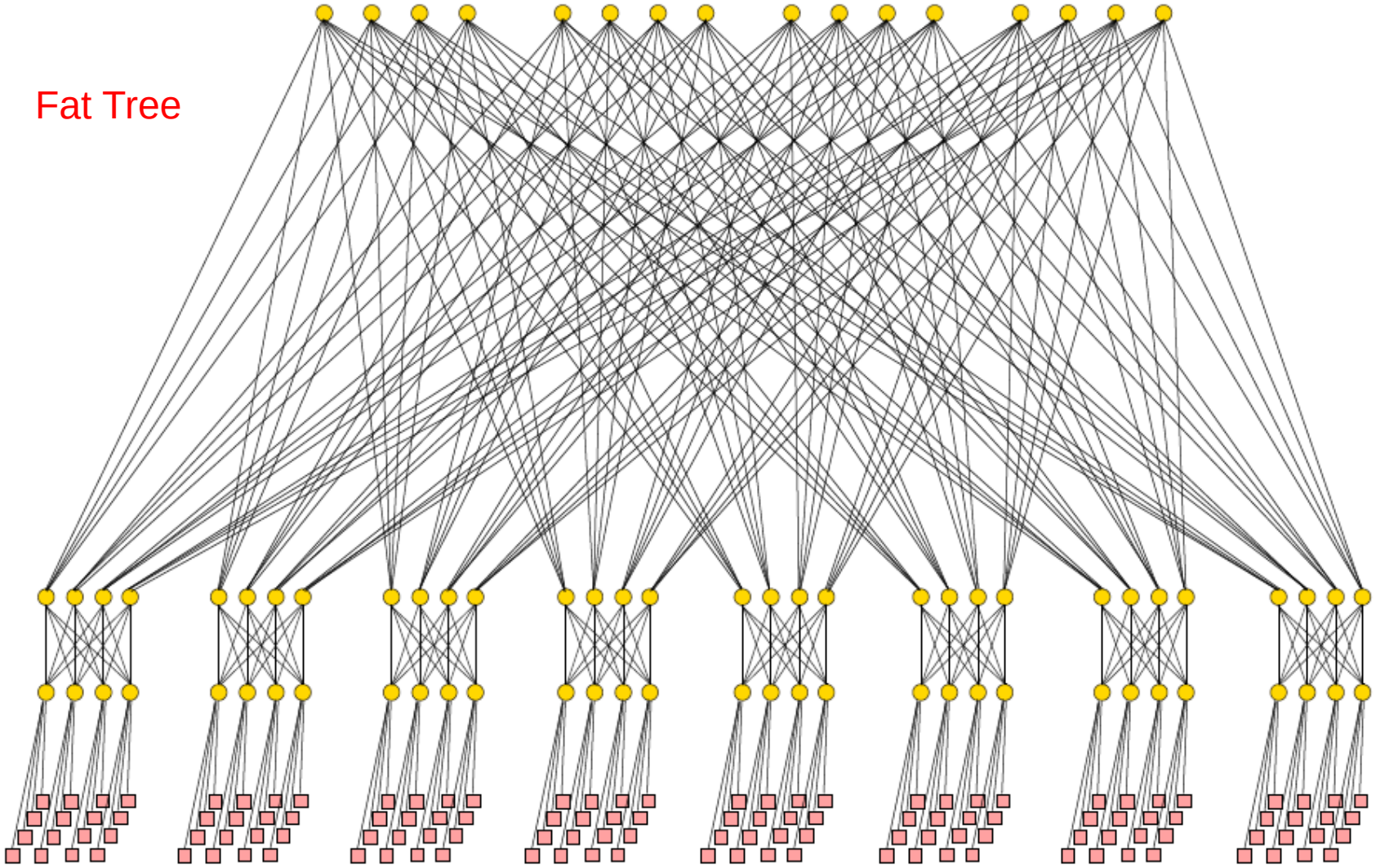
Toroid 2D

Toroid 3D



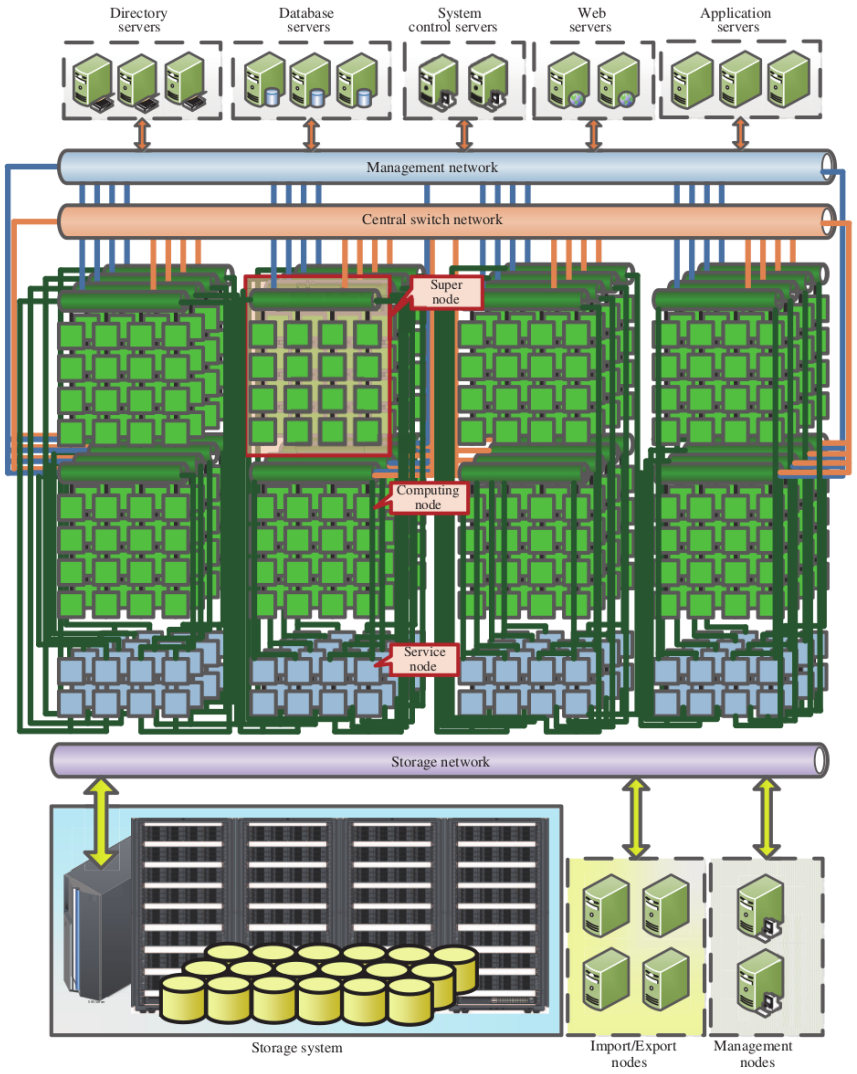
Redes de interconexión

Fat Tree



MPP: Sunway TaihuLight

- Rendimiento: 93 petaflops (Linpack bechmark)
- 40960 procesadores RISC de diseño específico
  - Cada uno con 256 cores de procesamiento y cuatro para manejo del sistema
- Interconexión jerárquica: nodos de computación, placa de computación, supernodos y armario



Cluster: Tianhe

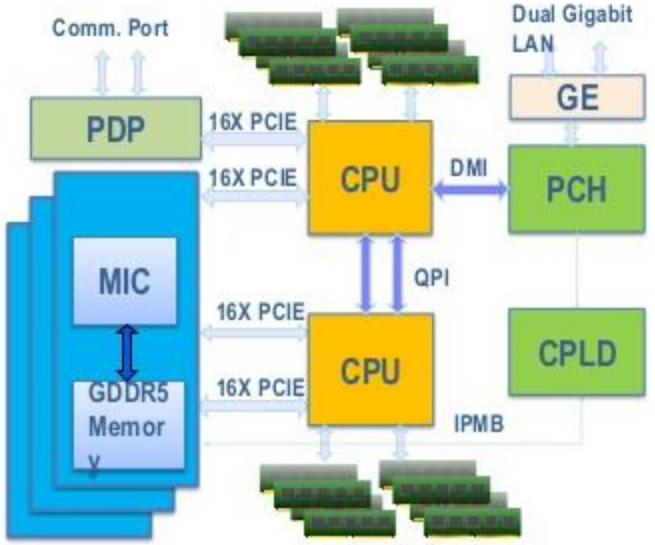
- Compute node

Compute Node

HPCL

■ Neo-Heterogeneous Compute Node

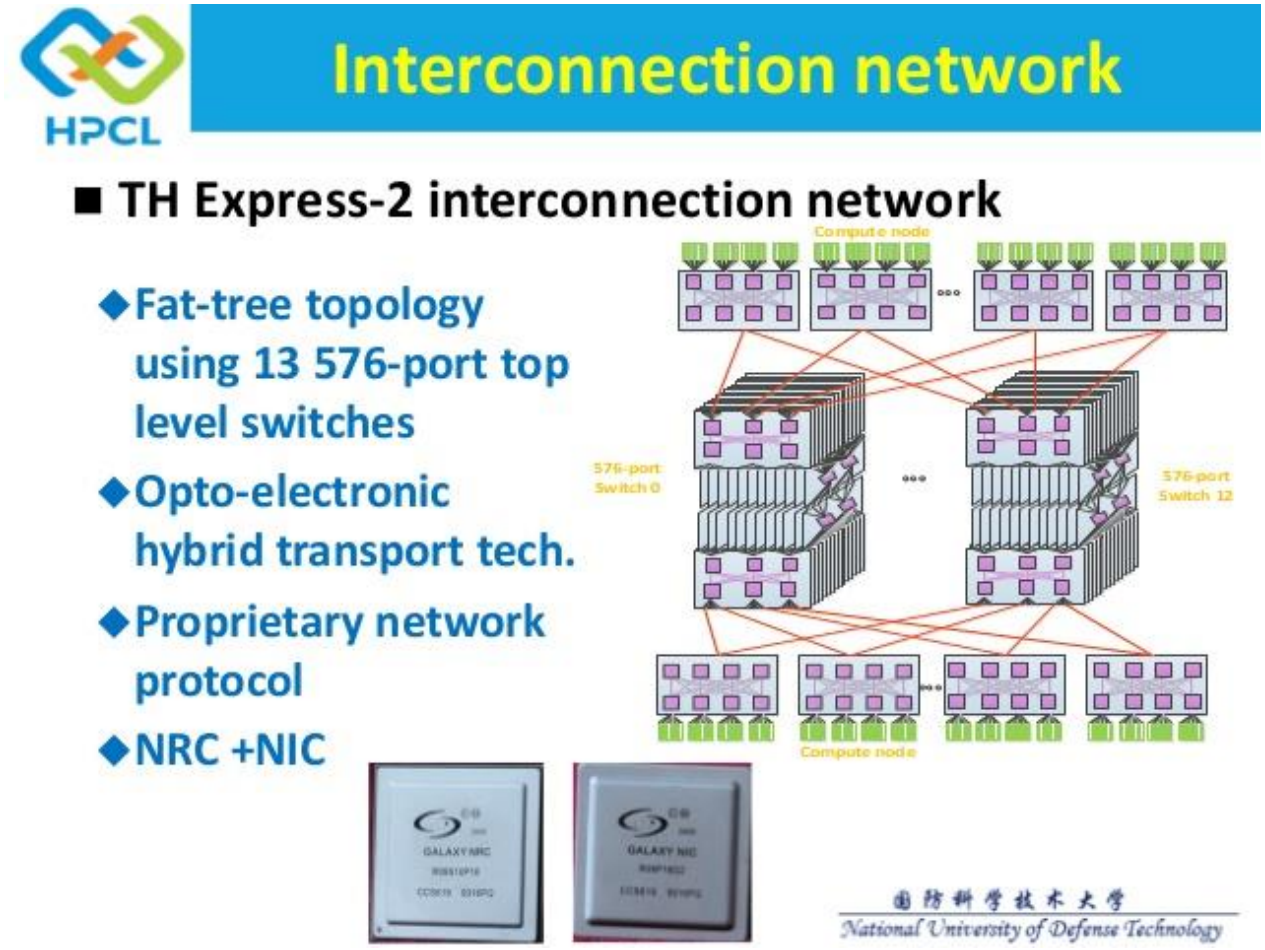
- ◆ Similar ISA, different ALU
- ◆ 2 Intel Ivy Bridge CPU + 3 Intel Xeon Phi
- ◆ 16 Registered ECC DDR3 DIMMs, 64GB
- ◆ 3 PCI-E 3.0 with 16 lanes
- ◆ PDP Comm. Port
- ◆ Dual Gigabit LAN
- ◆ Peak Perf. : 3.432Tflops



國防科學技術大學
National University of Defense Technology

Cluster: Tianhe 2

- Red de interconexión TH-Express 2



- Rendimiento: 34 petaflops (Linpack benchmark)

Computadores más potentes

- TOP500: https://www.top500.org/
  - Lista de noviembre de 2016