ARQUITECTURA DE COMPUTADORES II

AUTORES:
David Expósito Singh
Florin Isaila
Daniel Higuero Alonso-Mardones
Javier García Blas
Borja Bergua Guerra

Área de Arquitectura y Tecnología de Computadores
Departamento de Informática
Universidad Carlos III de Madrid

Julio de 2012

TEMA 6:
COMPUTACIÓN CLUSTER

Índice

1. Introducción.
2. Arquitecturas cluster y componentes.
3. *Middleware* y *Single System Image*.
4. Ejemplos.
5. Referencias.

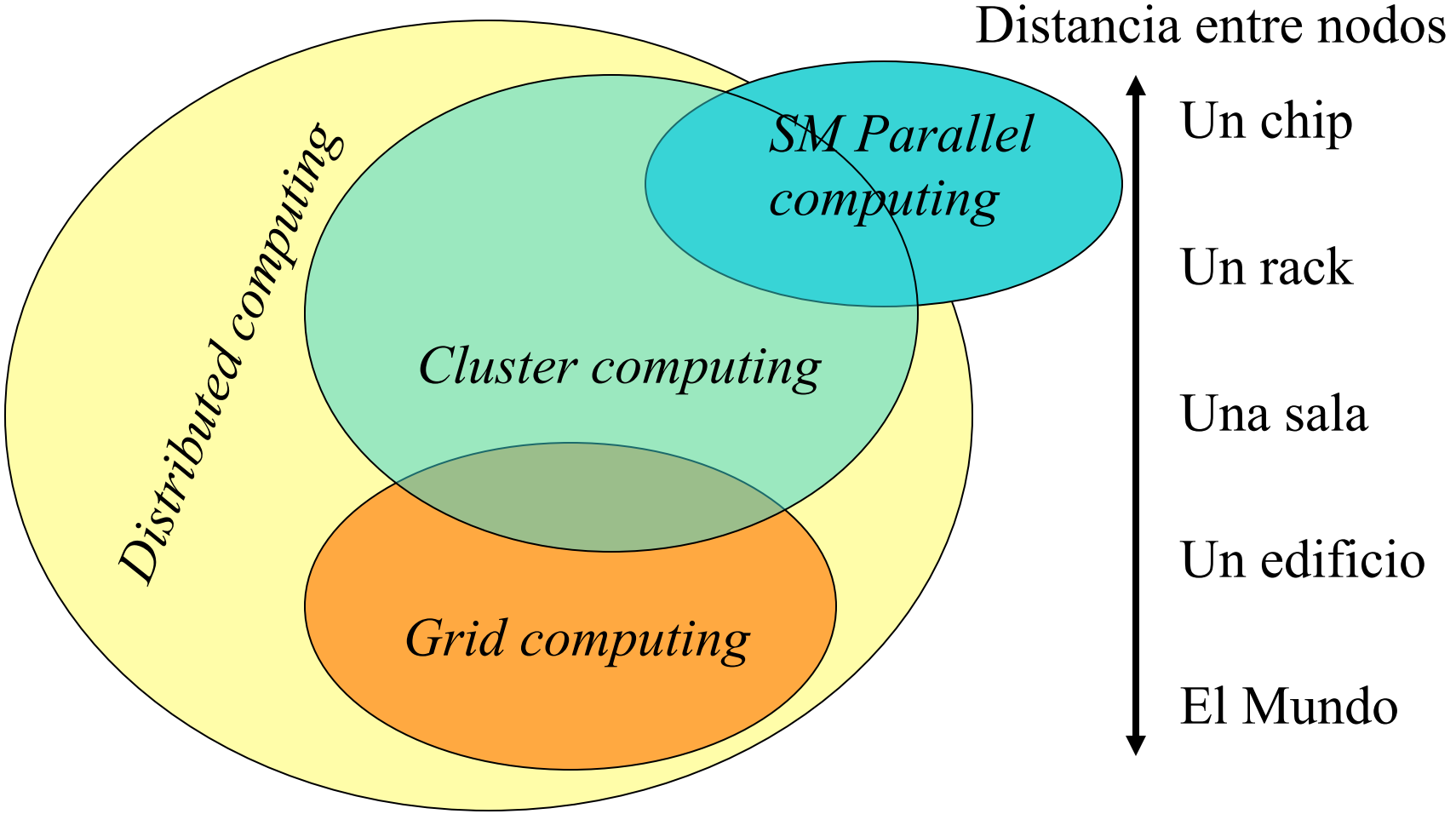
1. Introducción

- Qué es un **cluster**:

> *A **cluster** is a type of parallel or distributed processing system, which consists of a collection of interconnected **stand-alone/complete computers** cooperatively working together as a **single**, integrated computing resource.* [Buyya98]

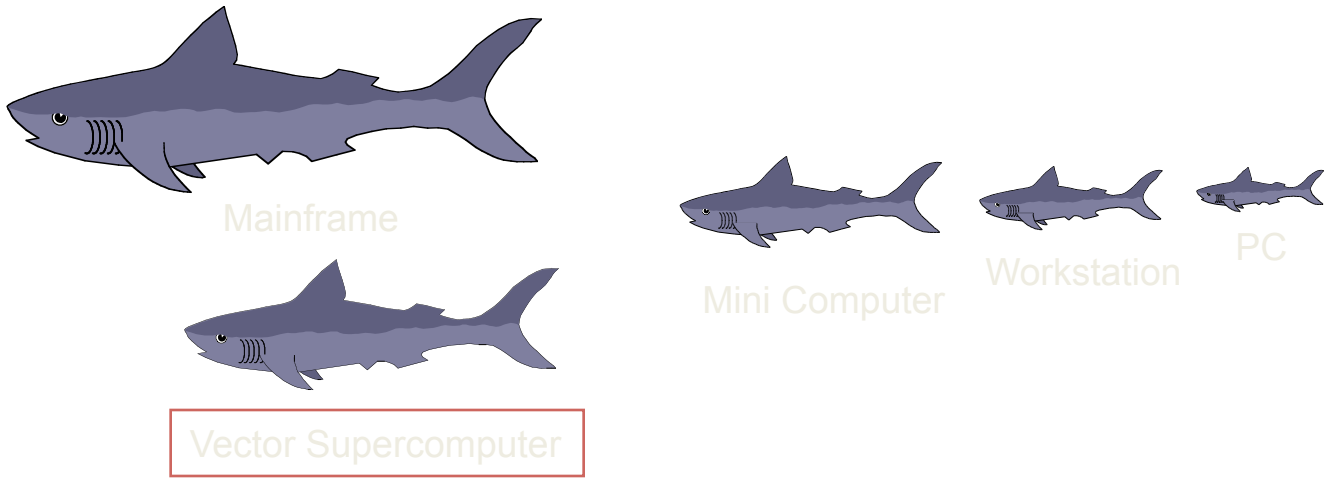
Computación cluster vs. otros

[Tony Cortés 03]



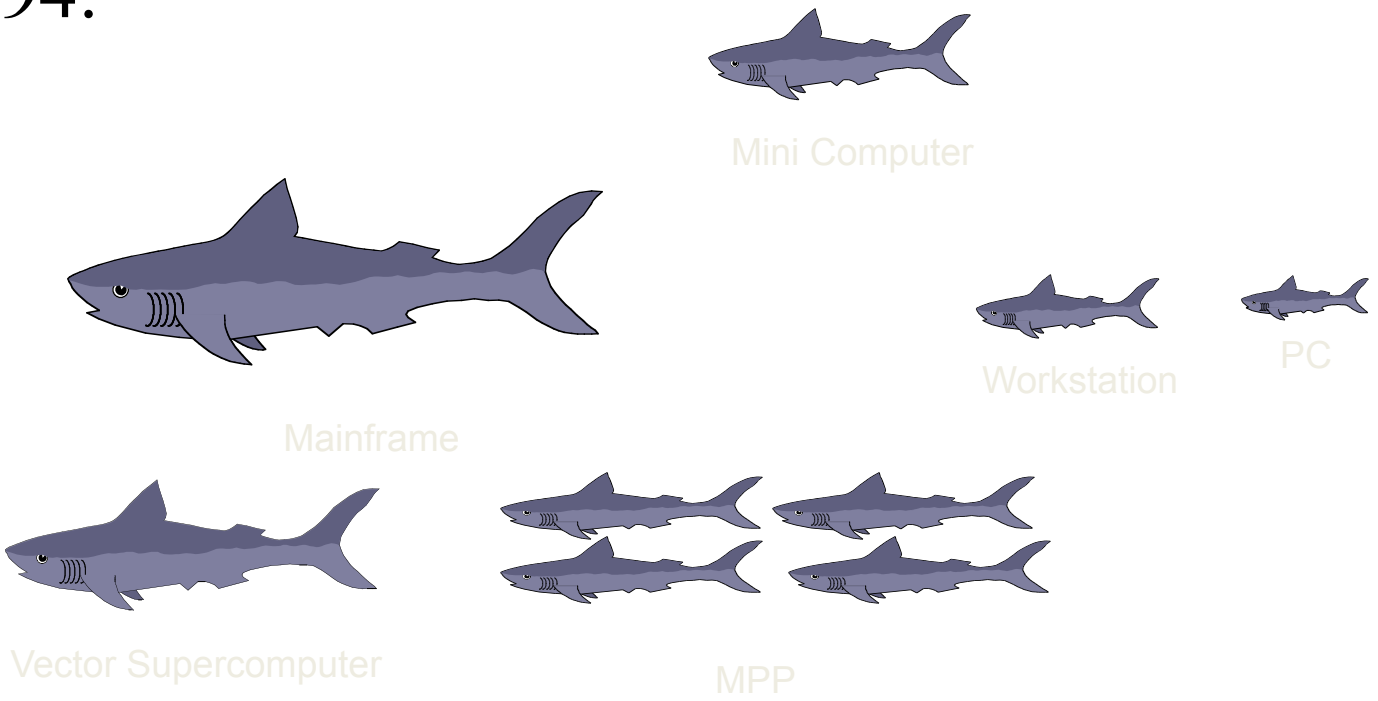
Evolución de la HPC [metáfora de Buyya]

- 1984:



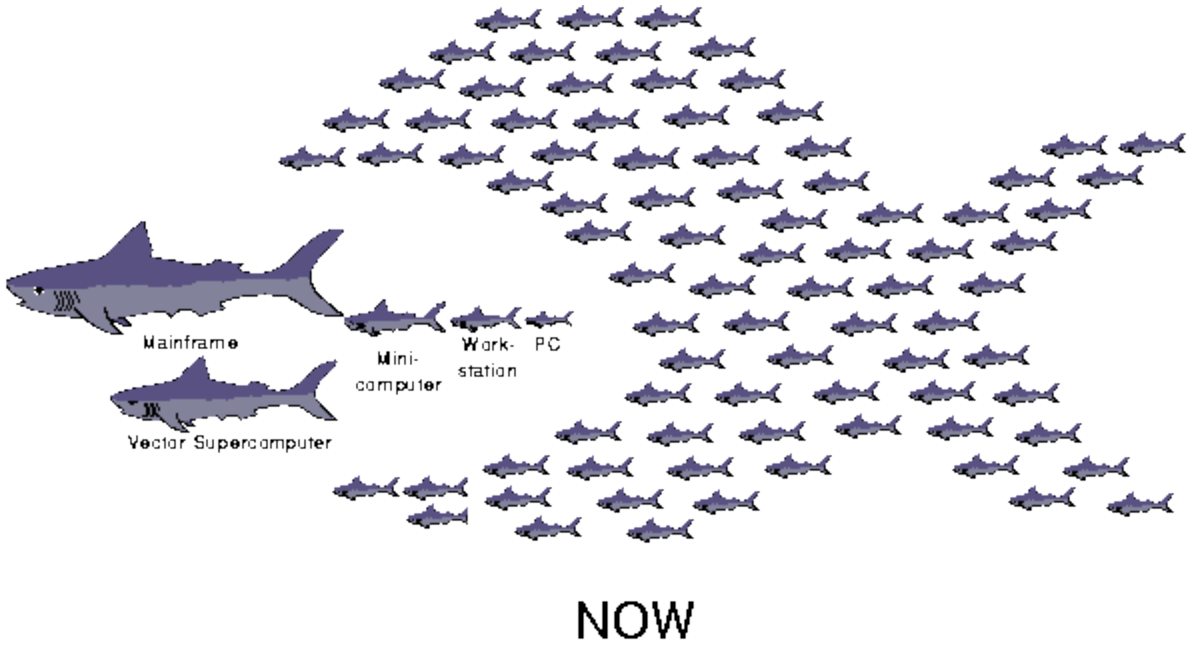
Evolución de la HPC (II) [metáfora de Buyya]

- 1994:



Evolución de la HPC (y III) [metáfora de Buyya]

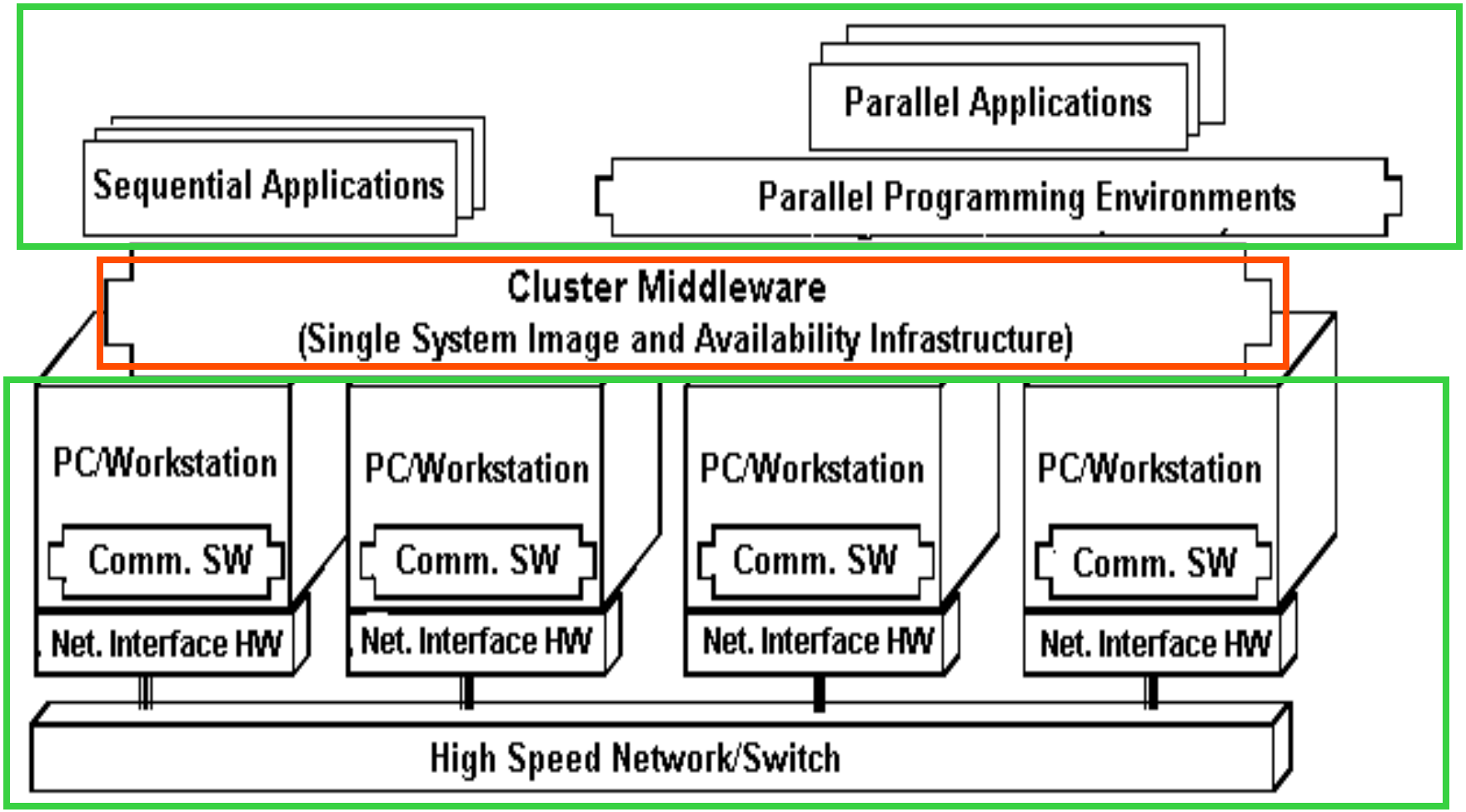
- ¡El presente y el futuro!:



Componentes de un cluster

- Hardware + Software:
  - Hardware:
    - nodos: PCs, workstations, MPs,etc.
    - red: Gigabit, Myrinet, ATM, etc.
  - Software:
    - Sistema Operativo.: Linux, Sun Solaris, IBM AIX, etc
    - Comunicación: Sockets, etc.; *Light weight protocols*.
    - Específico: *cluster middleware*:
      - Single System Image (SSI)
      - System Availability (SA)

2. Arquitecturas cluster y componentes



Hardware: nodos

- Se utilizan “computadores” –más que ‘arquitecturas’-- convencionales”:
  - PCs
  - *Workstations*
  - MP (normalmente SMP)

- *¡No se desarrollan máquinas especiales para ser nodos de un cluster!*

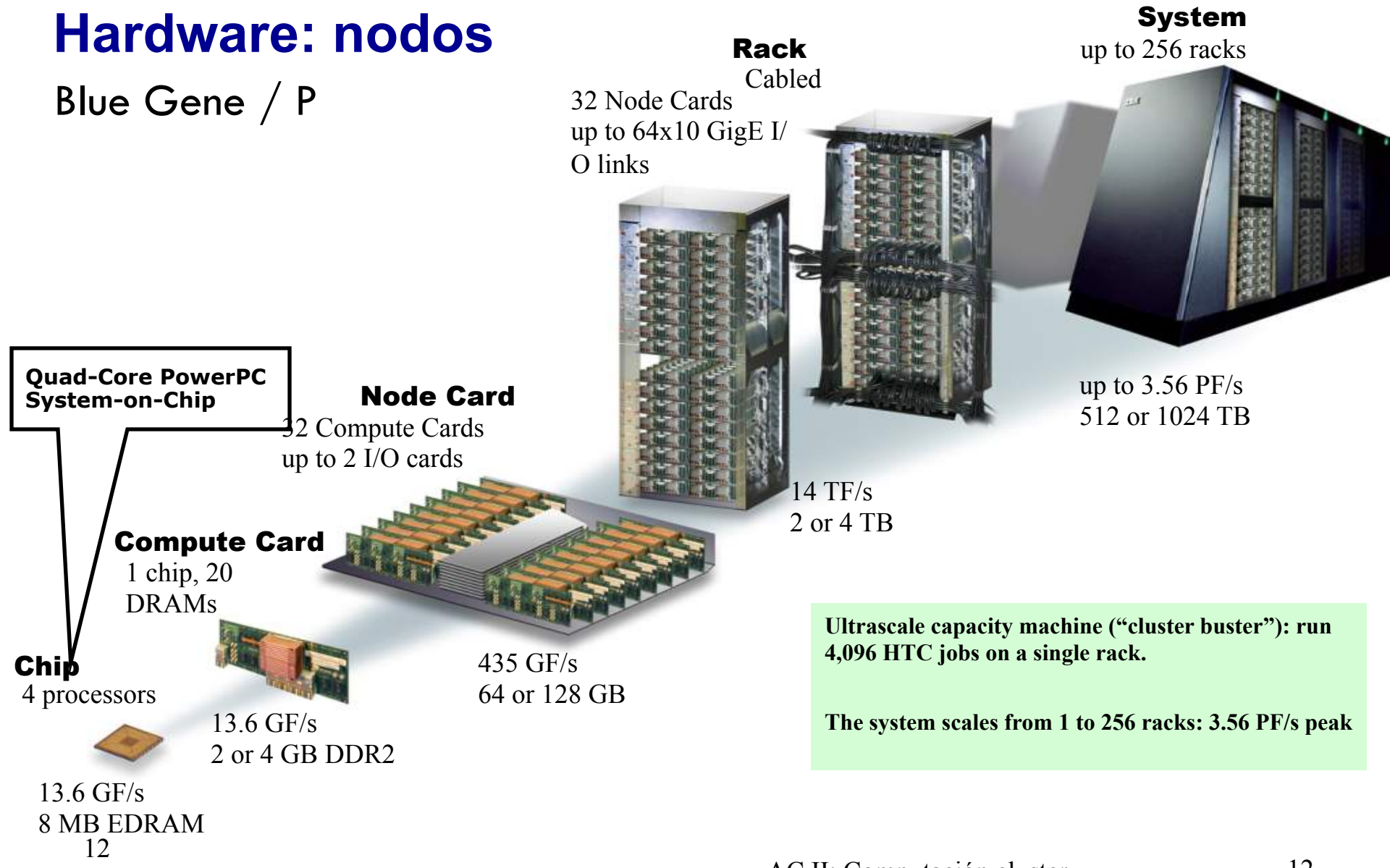
Hardware: nodos

- Ejemplos:
  - Intel Pentium, Xeon, etc.
  - Sun Sparc, Ultrasparc
  - IBM RS6000
  - Digital

- En un mismo cluster pueden coexistir diferentes arquitecturas y S.O., o ambos

Hardware: nodos

Blue Gene / P



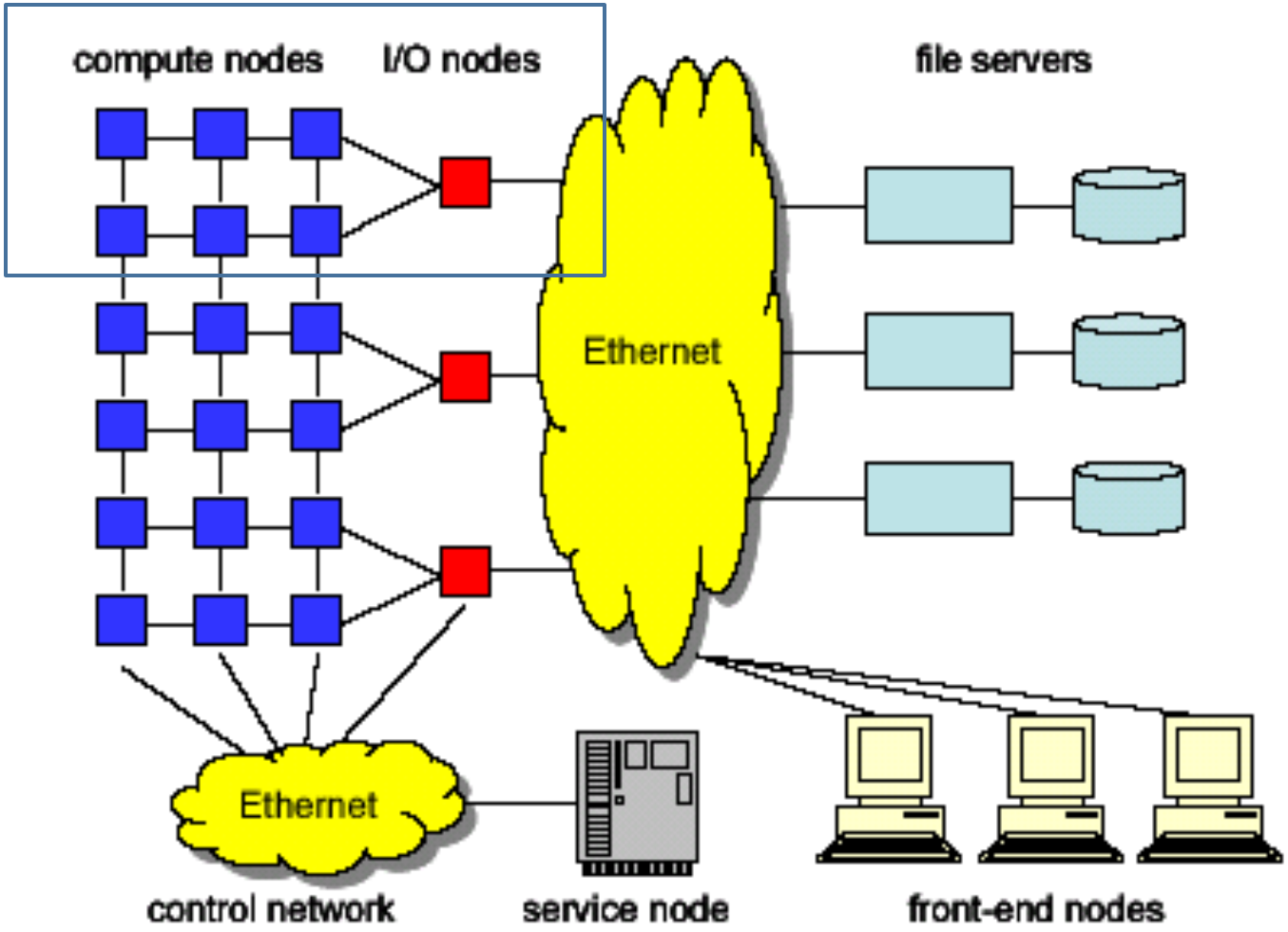
Ultrascale capacity machine ("cluster buster"): run 4,096 HTC jobs on a single rack.

The system scales from 1 to 256 racks: 3.56 PF/s peak

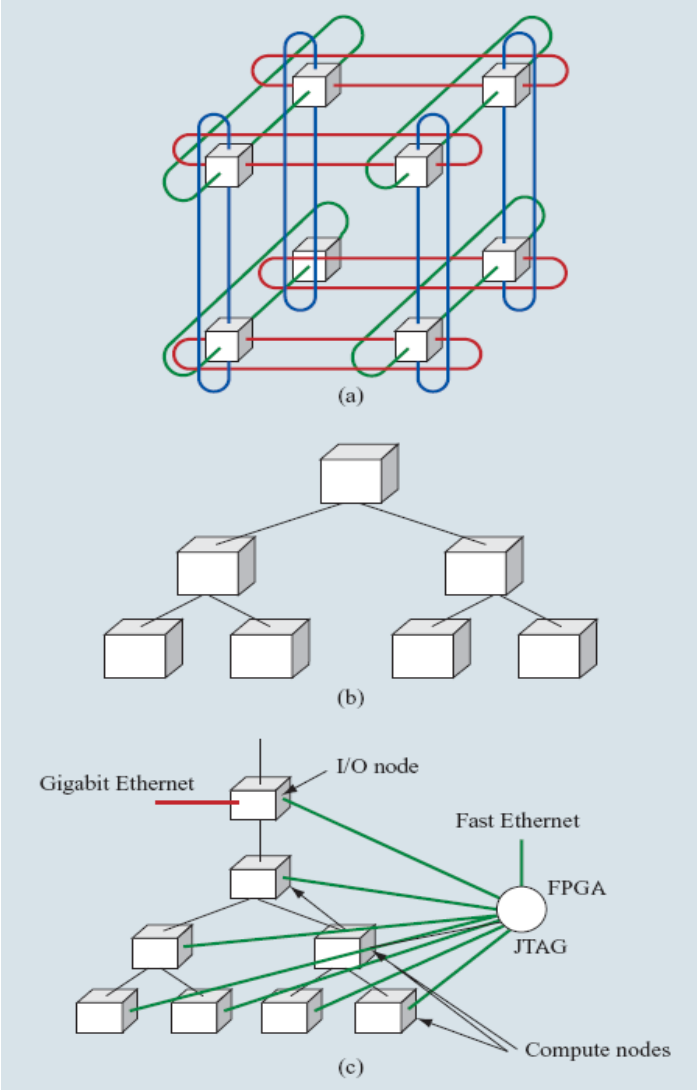
Hardware: redes

- Algunos ejemplos:
  - Ethernet
  - Fast Ethernet
  - Gigabit Ethernet
  - SCI
  - ATM
  - Myrinet
  - HIPPI
  - FiberChannel

Hardware: redes (Blue Gene /L)



Hardware: redes

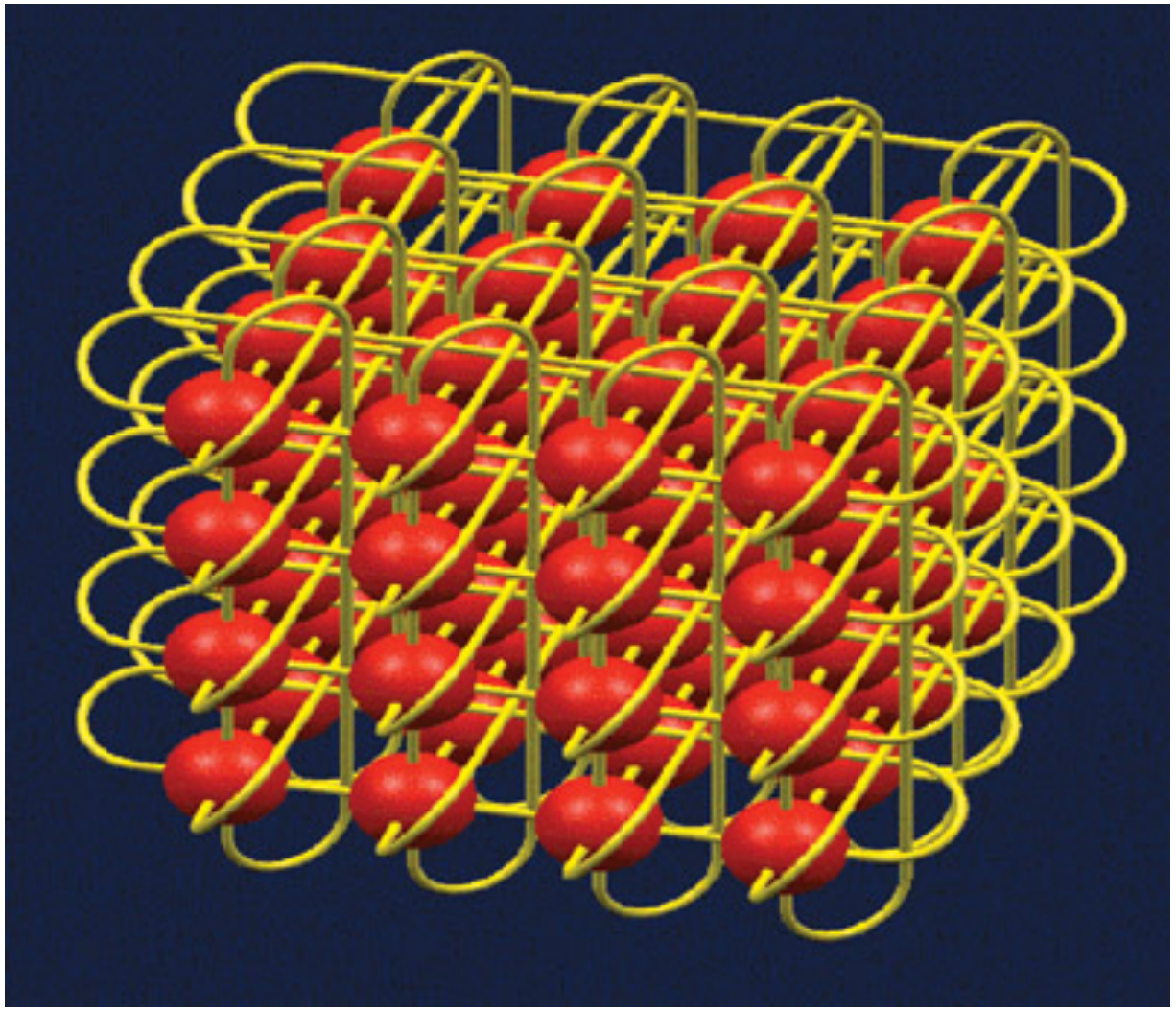


Toro 3D
- Comunicaciones

Árbol
- Comunicaciones colectivas
- E/S

Ethernet
Red de control

Hardware: redes 3D Torus



Hardware: protocolos de comunicación

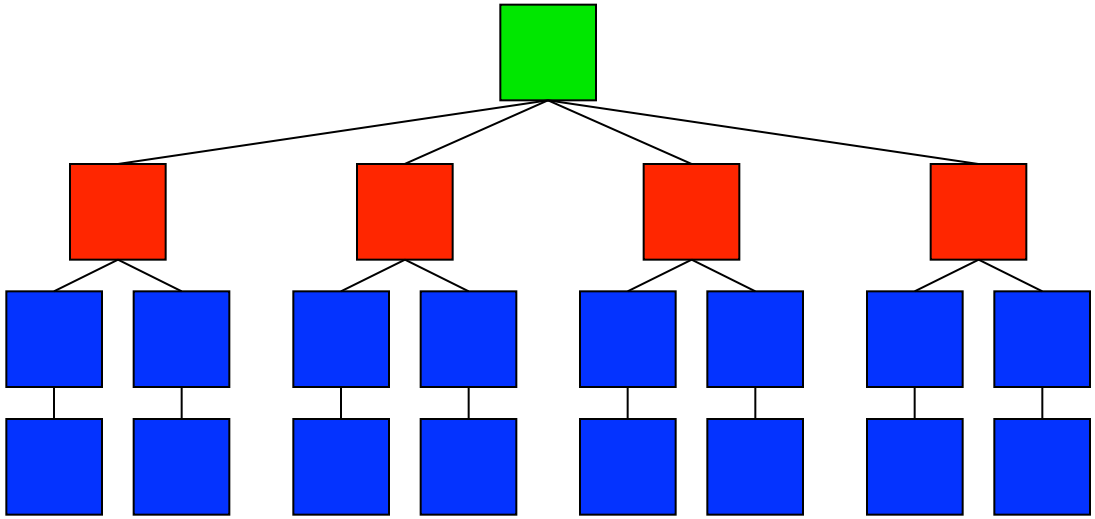
- Protocolos "tradicionales" ("pesados"):
  - TCP/IP
- Protocolos "especiales" ("ligeros"):
  - *Active Messages* (U. Berkeley)
  - VMMC (*Virtual-Memory Mapped Communication*)
  - U-net (U. Cornell)
  - XTP (U. Virginia)
  - etc.

Software

- SS. OO. habituales:
  - Linux: Beowulf
  - Solaris: Berkeley NOW
  - Windows NT: HPVM
  - etc.
- Entornos de programación:
  - Threads
  - MPI
  - PVM

Software: organización jerárquica del Blue Gene

- Nodos de cómputo: dedicados a la ejecución de aplicaciones de usuarios. Simple *compute node kernel* (CNK)
- Nodos de E/S: ejecutan Linux y proporcionan servicios del S.O: – ficheros, sockets, ejecución procesos, señales, depurado y terminación.
- Nodos de servicio: implemente los servicios de gestión del S.O. (*heart beating*, monitorización, reiniciado del nodo) transparentes a las aplicaciones.



Software: organización jerárquica de Blue Gene

- Componentes: E/S, nodos de servicio, CNK
- Processing sets (*psets): 1 nodo de E/S+ varios nodos de cómputo*
  - 8, 16, 64, 128 NCs
  - Agrupación lógica
  - Proximidad física de los componentes => comunicaciones rápidas
- *Job*: colección de N procesos de cómputo
  - Espacio de direcciones privado.
  - Comunicador MPI
  - MPI: ranks 0, N-1

CNK

- Consume 1MB
- Espacio de memoria
  - Único espacio de memoria 511/1023MB
  - Dos espacios de memoria de 255/511MB
- No memoria virtual ni páginas
- Carga en modo *push*: 1 NC lee el ejecutable del sistema de ficheros y se lo envía a los otros procesadores.
- Única imagen cargada.

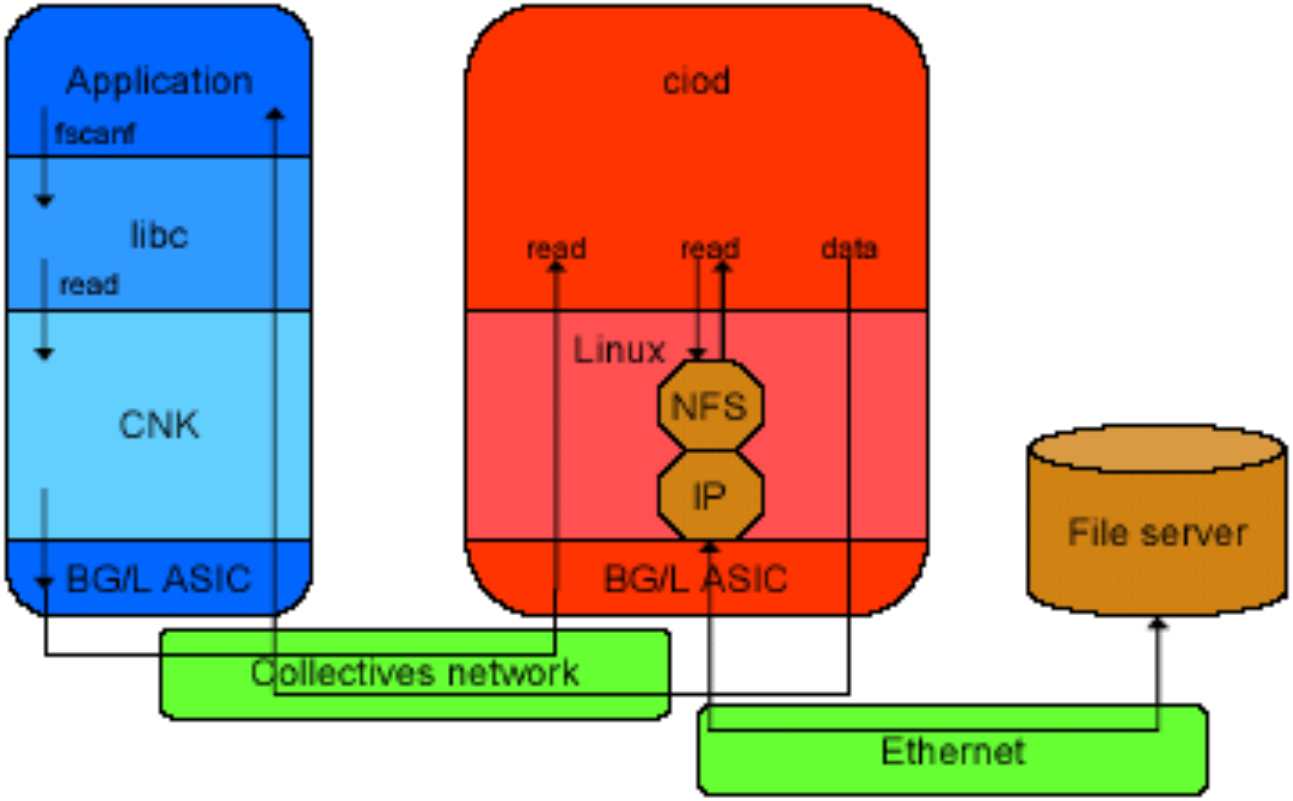
CNK

- No planificación (único proceso).
- No gestión de memoria (No se usa la TLB)
- No hay servicios de E/S locales.
- Nivel de usuario hasta que:
  - Se realiza llamada al sistema.
  - Interrupciones hardware: timer (solicitada por la aplicación), eventos anormales.
- Llamada al sistema
  - Simple: Se gestiona localmente: alarmas, obtener hora, medida tiempo.
  - Compleja: enviada a los nodos de E/S.

Nodo de E/S

- Niveles completos del protocolo TCP/IP
- Sistemas de ficheros soportados: NFS, GPFS, Lustre, PVFS
- Proceso principal: Control y demonio de E/S (CIOD)
- Ejecución de un trabajo
  - Job manager envía solicitud al nodo de servicio.
  - Nodo de servicio contacta con el demonio de E/S (CIOD)
  - CIOD envía el ejecutable a todos los procesos en el *pset*

Llamadas de sistema



Function Shipping from CNK to CIOD.

Software

- Compiladores:
  - Java
  - C, C++
- Otros:
  - depuradores
  - herramientas para análisis de rendimiento
  - herramientas de visualización
  - etc.

4. *Middleware*

- Qué es el *middleware* de un cluster:
  - Interface entre las aplicaciones y el hardware con su S.O.
  - Capas de *middleware*:
    - **SSI** (*Single System Image*)
    - **SA** (*System Availability*):
      - detección y recuperación frente a errores
      - tolerancia a fallos

SSI

- Concepto:
  - *Single System Image* (SSI) es la ilusión que presenta un conjunto de recursos como uno solo y más potente.
  - SSI hace aparecer al cluster como una máquina única para el usuario y sus aplicaciones.
  - *¡Un cluster sin SSI no es un cluster!*

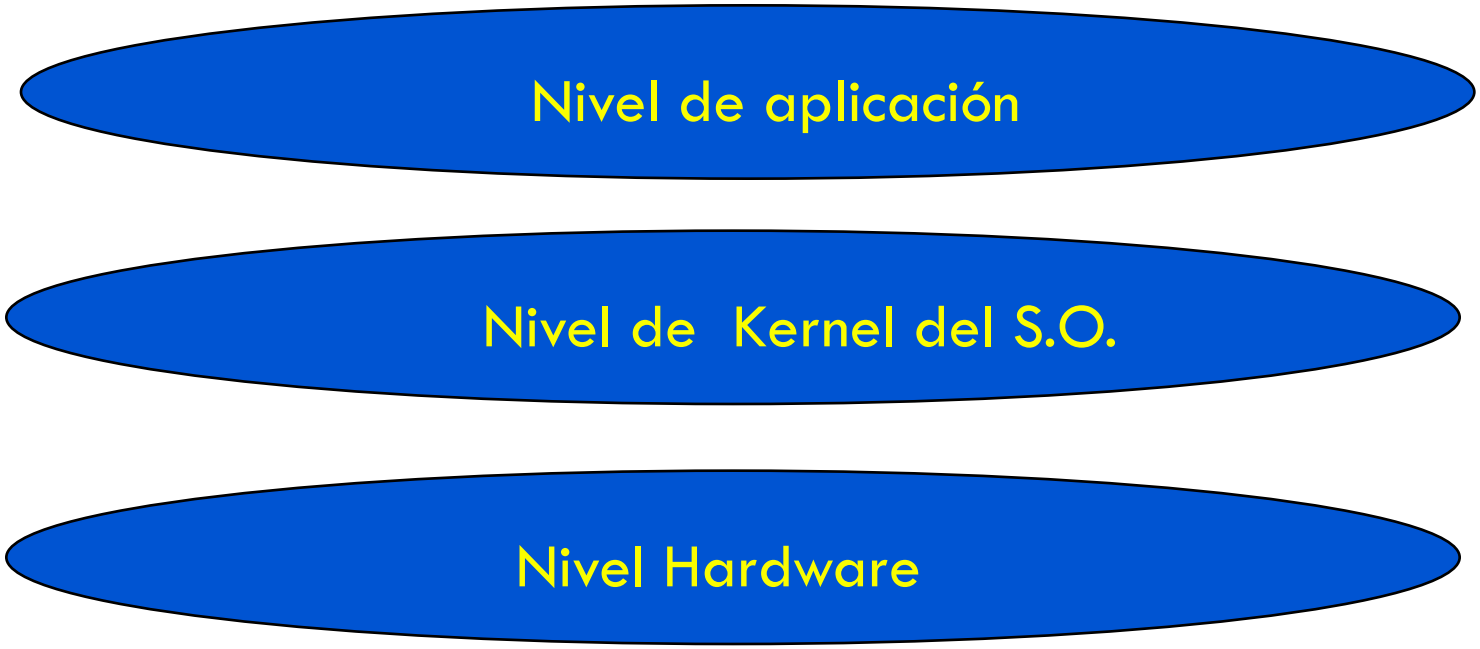
SSI: beneficios

- Ventajas del empleo de SSI:
  - Se pueden usar los recursos del sistema de manera transparente.
  - Migración de procesos y equilibrado de carga entre los nodos.
  - Mejora la fiabilidad y disponibilidad de recursos.
  - Mejora el tiempo medio de respuesta y el rendimiento.
  - Simplifica la gestión del sistema.
  - Proporciona independencia del hardware.

SSI: servicios

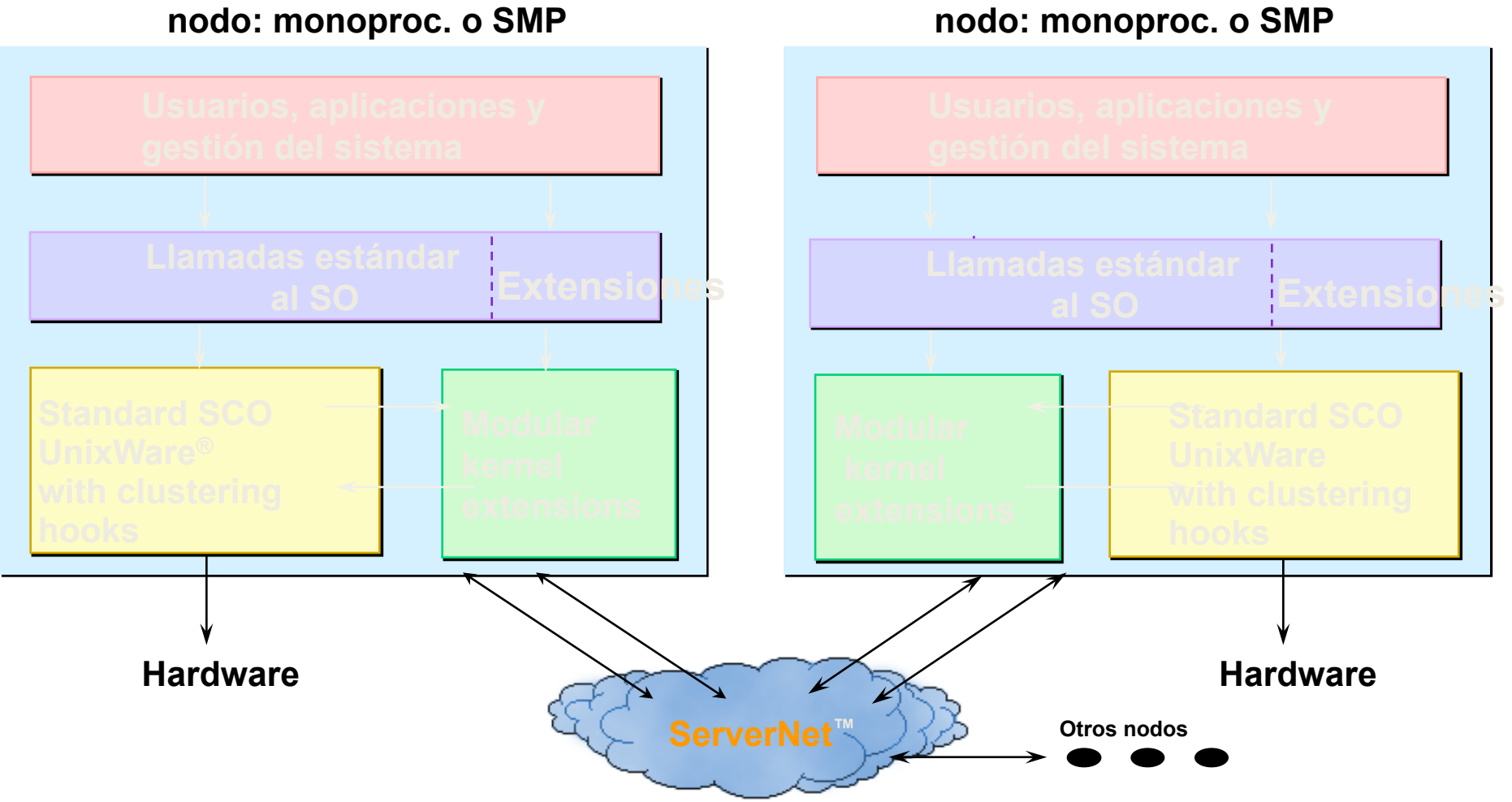
- Entre otros, son deseables los siguientes: único...
  1. punto de entrada
  2. jerarquía de archivos
  3. punto de control
  4. espacio de memoria
  5. gestor de trabajos
  6. interfaz de usuario
  7. espacio de E/S [SIO]
  8. espacio de procesos [SPP]

SSI: a qué nivel implementarlo



SCO NonStop® Cluster for UnixWare

http://www.sco.com/products/clustering/



Referencias

- Sitios web:
  - *IEEE Task Force on Cluster Computing (TFCC)*:
    http://www.ieeetfcc.org
  - Sitio de Rajkumar Buyya
    http://www.csse.monash.edu.au/~rajkumar
  - Y los señalados sobre clusters particulares...
- Libros:
  - R. Buyya, Ed., *High Performance Cluster Computing*, **Vol. 1** *System and Architecture*, **Vol. 2** *Programming and Applications*, Prentice Hall PTR, 1998.

Referencias (y II)

- G.F.Pfister, *In Search of Clusters,* 2nd Edition, Prentice Hall, 1998.
- Th. L. Sterling, J.Salmon, D. J. Becker, D. F. Savarrese, *How to Build a Beowulf*, The MIT Press, 1999.
- D.H.M. Spector, *Building Linux Clusters*, O'Reilly & Associates, 2000