DE CLUSTER A CLOUD (VISIÓN PRÁCTICA)

ARCOS

Clusters

Supercomputadores

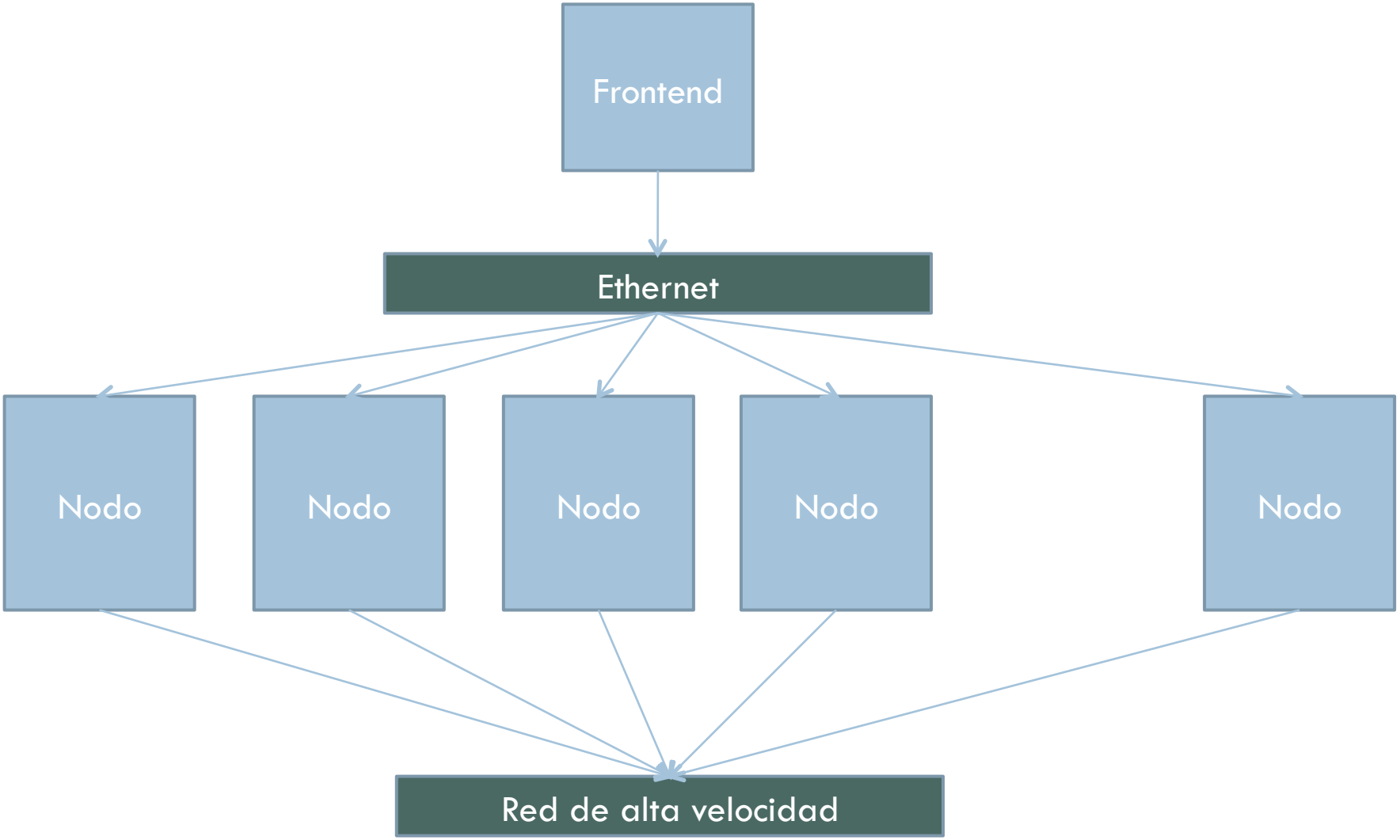
Cloud

CLUSTERING

¿Qué es un cluster?

- Supercomputador del “hombre pobre”
- “... collection of interconnected stand-alone computers working together as a single, integrated computing resource” – R. Buyya
- Cluster compuesto por:
  - Nodos
  - Red
  - SSOO
  - Cluster middleware
- Componentes estándar
- Elimina componentes caros

Arquitectura cluster



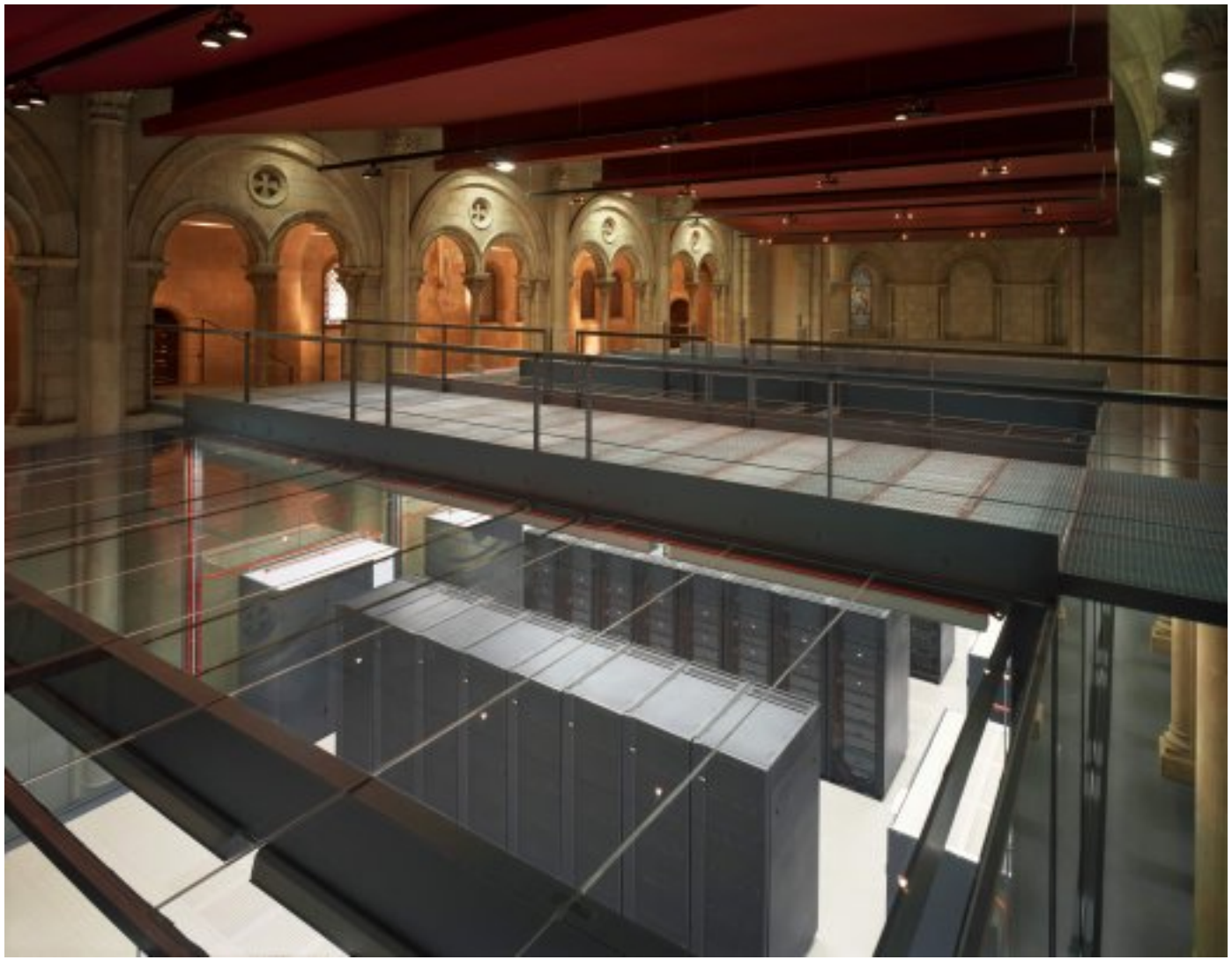
Arquitectura cluster

- Los nodos son PC Nodes are Individual PCs
  - 1 o 2 procesadores (Pentium, Athlon, Opteron, Itanium)
  - Memoria (al memos 1 GB por procesador)
  - Almacenamiento local (20 o más GB)
    - Sistema operativo
    - Bibliotecas y software
    - Espacio Swap
- Conexiones de red(es)
  - Ethernet
    - Sesiones (SSH)
    - Monitorización
    - Sistema de ficheros de los usuarios (NFS)
  - Myrinet o Infiniband o Quadrics (opcional, pero muy extendido)

Caso real: HLRS

Caso real: HLRS

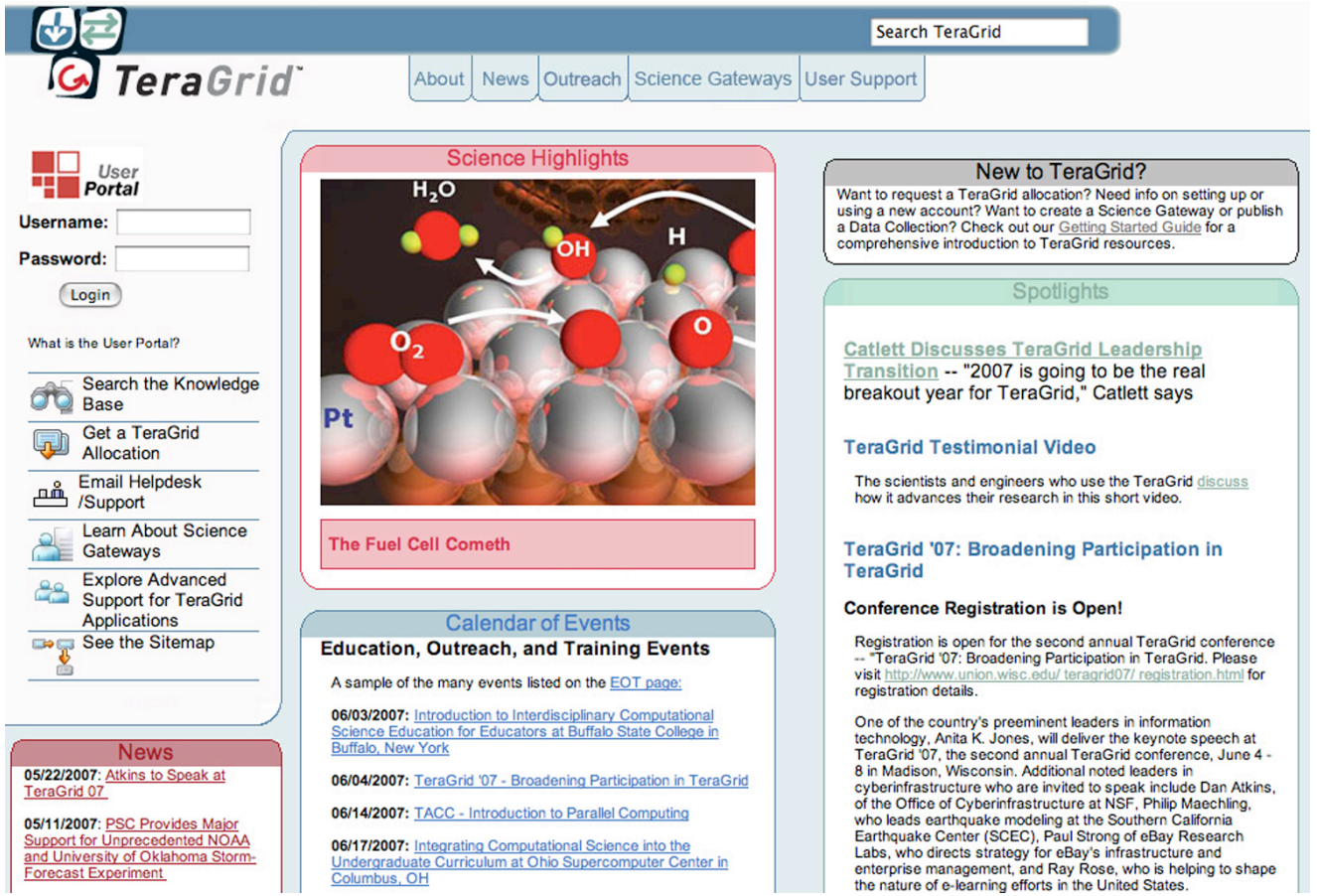
Caso real: MareNostrum



Caso real: FINISTERRAE

Caso real: Teragrid

- http://www.teragrid.org



Instalación de clusters (Rocks)

- Distribución para cluster (Red Hat)
- Permite fácil instalación y administración del cluster
- Instalación automática (PXE Boot, DHCP)
- Gestión de usuarios (cuentas, replicación,...)
- Sistema de ficheros compartido (NFS)
- Servicios del cluster
  - Gestor de tareas
  - Monitorización
- Interfaz Web
- Alternativa: Linux ->OSCAR

  Windows -> Windows 2008 HPC server

Configuración de Rocks

- Nodo Front End
  - Tiene dos interfaces de red (Ethernet)
  - Una red externa – Firewall excepto para SSH
  - Una red privada entre los nodos de cómputo
- Discos exportados via NFS
  - Distribución Rocks
  - Directorios de usuarios
- Nodos de cómputo: Net Boot y DHCP
  - Virtual (SSH y Xterm-based) KVM para monitorizar la instalación
  - Comunicación entre los nodos y el front end via SSH
  - Los usuarios solo hacen login al Front End – Compilación

Cofiguración de Rocks

- Nodos de cómputo desechables
  - En el primer arranque se instala y configura (usando RedHat's Kickstart)
  - Después del primer arranque, se pueden añadir nuevas extensiones de Rocks
  - Si un nodo falla, simplemente se reemplaza
- Ventajas
  - No hay que preocuparse por la configuración de cada nodo
  - Fácil de actualizar
  - Fácil detectar fallos en el arranque
- Desventajas
  - Tarda unos 15-20 minutos en reemplazar un nodo del cluster

Tareas (jobs)

- Tarea == aplicación de un usuario
- Hay muchos tipos de tareas:
- Batch vs. interactiva:
  - Batch: no requiere la intervención del usuario (cálculos grandes, procesamiento de datos, etc.)
  - Interactiva: requiere la interacción del usuario (interfaces gráficas)
- Secuancial vs. parallel:
  - Secuencial – las tareas solo necesitan un procesador por ejecución
  - Paralelas – las tareas necesitan más de un procesador por ejecución

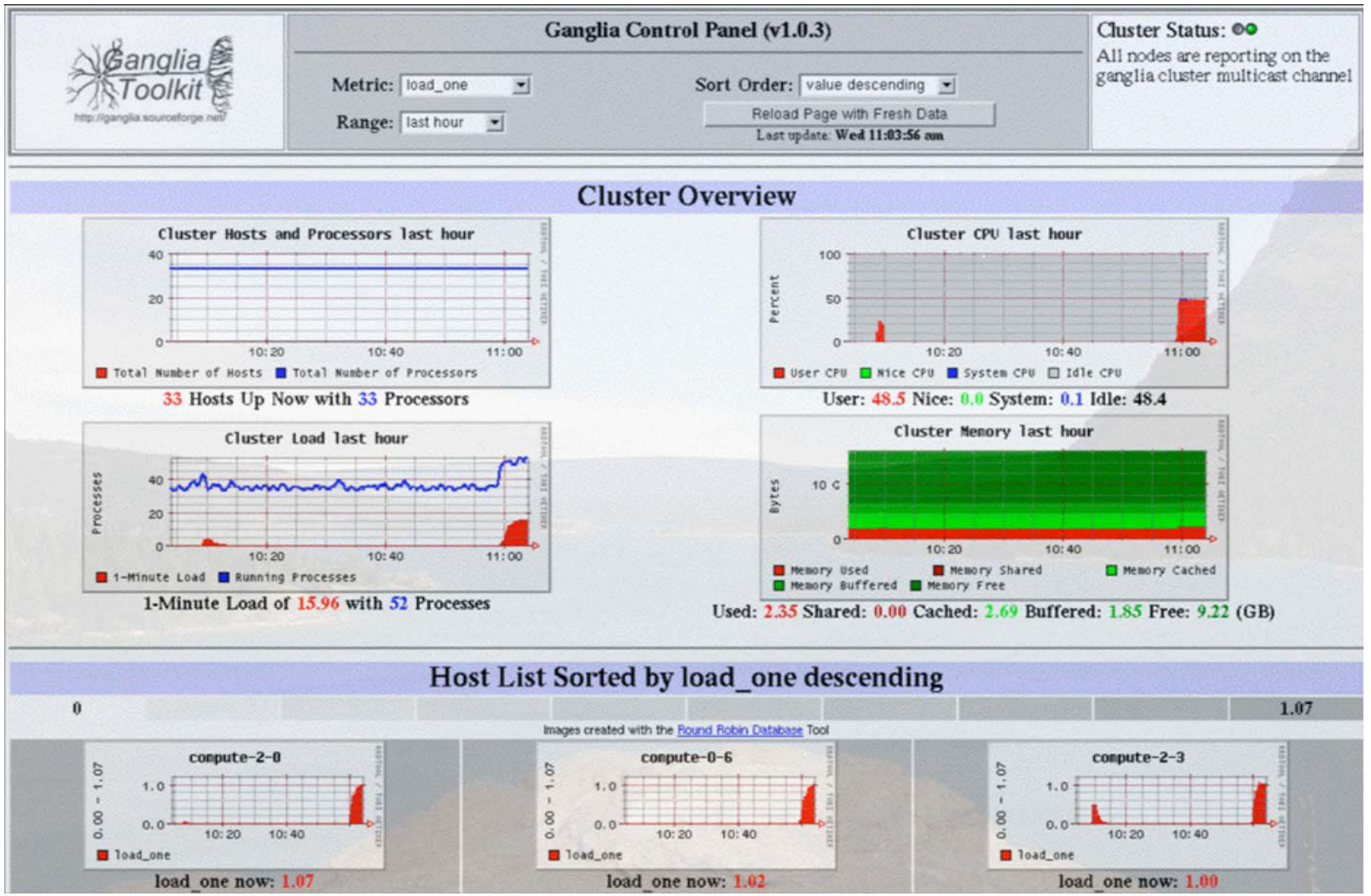
Gestor de tareas

- Interfaz de usuario para enviar y controlar tareas
- Políticas del uso de los recursos
- Optimizar la utilización de los nodos del cluster
- Información de las tareas de los usuarios
- Aspectos:
  - Reserva avanzada,
  - Checkpointing,
  - Migración de procesos,
  - Balanceo de carga,
  - Tolerancia a fallos
- Soluciones: **SGE**, **PBS** (instalada en kasukabe), **LSF**, **Condor**

Ejemplo de tarea

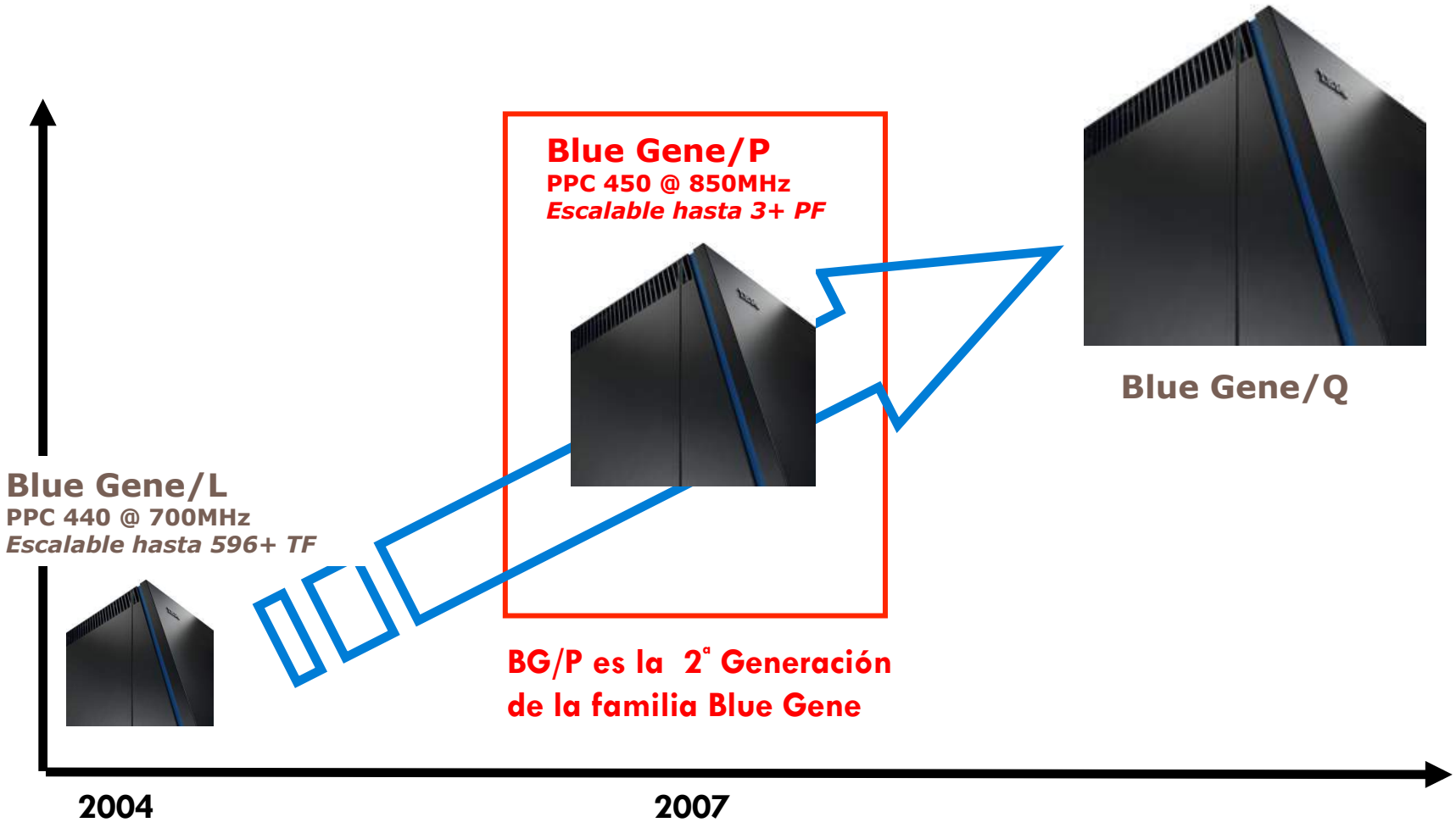
- **#PBS -N my_parallel_job**
- **#PBS -q default**
- **#PBS -l nodes=2:ppn=4:cell2**
- **#PBS -l walltime=04:00:00**
- # combine PBS standard output and error files
- **#PBS -j oe**
- # mail is sent to you when the job starts and when it terminates or aborts
- **#PBS -m bea**
- # specify your email address
- **#PBS -M John.Smith@dartmouth.edu**
- #change to the directory where you submitted the job
- **cd $PBS_O_WORKDIR**
- # include the relative path to the name of your MPI program
- **mpiexec -comm p4 ./program_name *and any arguments***
- **exit 0**

Monitorización (GANGLIA)



SUPER-COMPUTADORES (CASO BLUE GENE)

Introducción a Blue Gene
Tecnología



Blue Gene/L

Sistema

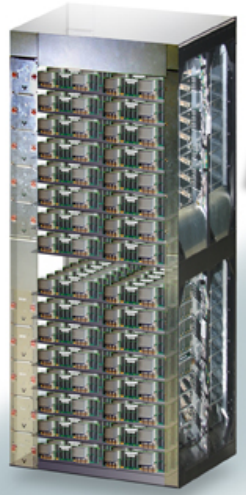
64 Racks, 64x32x32

180/360 TF/s

32 TB

Rack

32 Node Cards



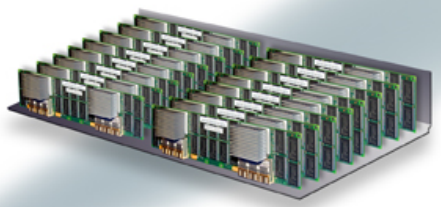
2.8/5.6 TF/s

512 GB

Node Card

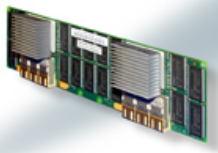
(32 chips 4x4x2)

16 compute, 0-2 IO cards

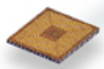


Compute Card

2 chips, 1x2x1



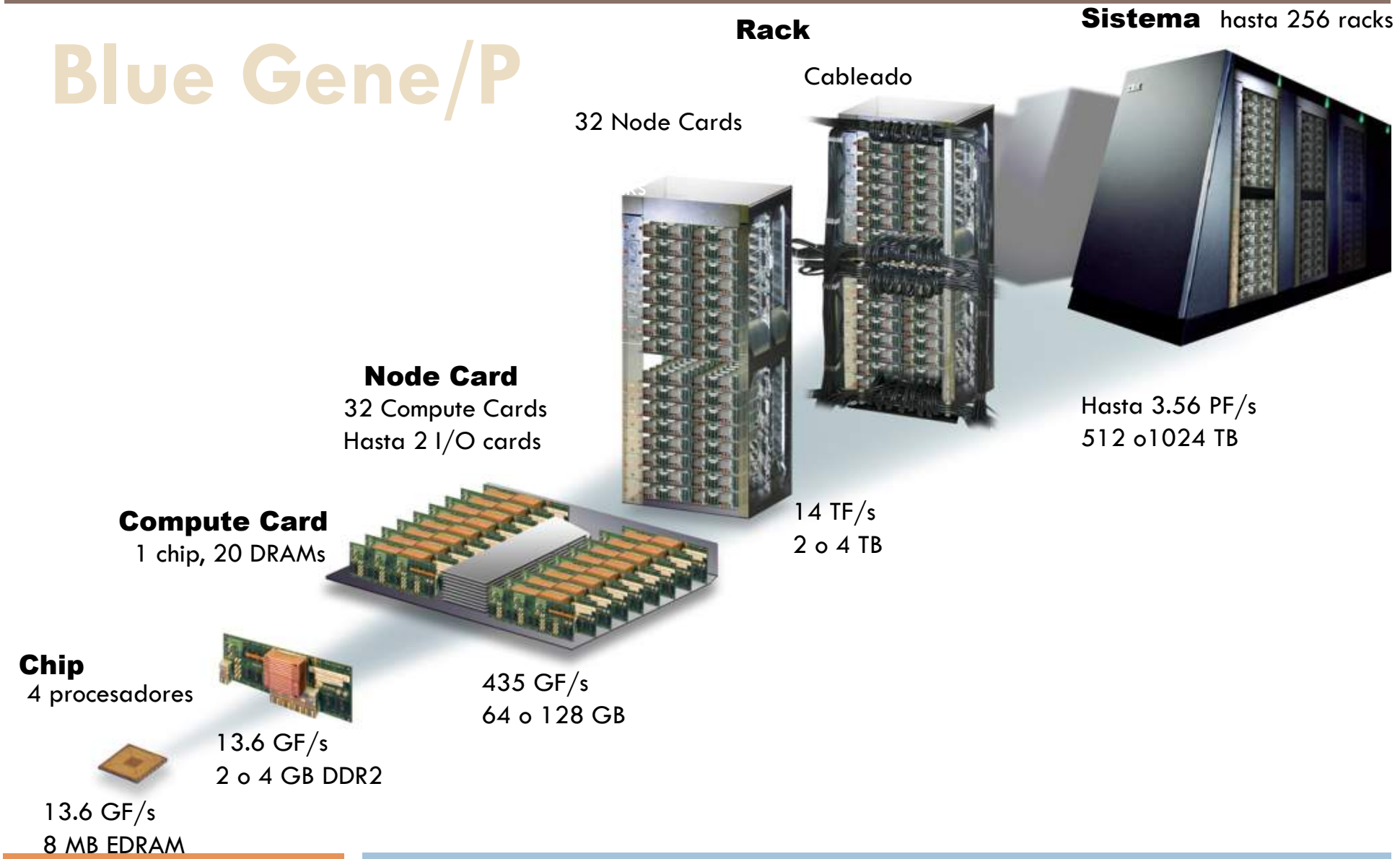
2 procesadores

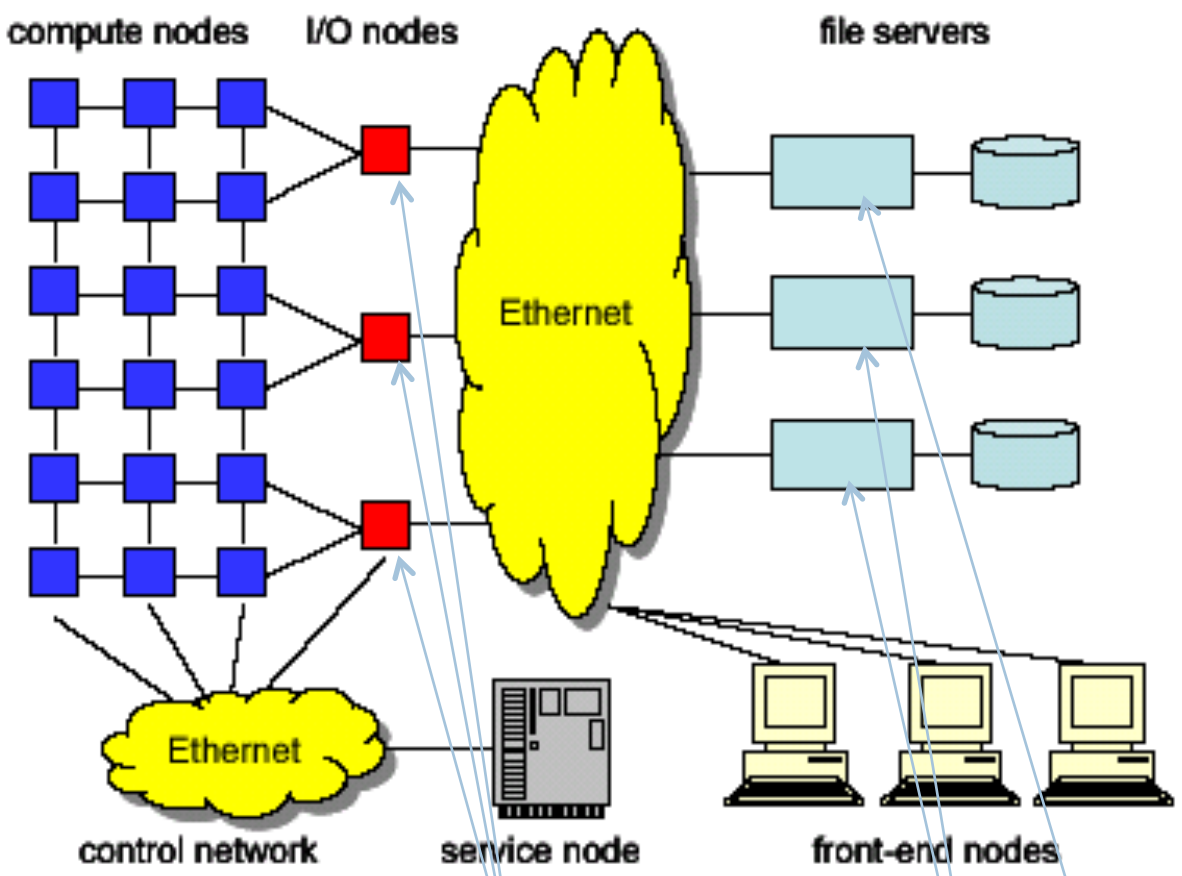


2.8/5.6 GF/s

4 MB

Blue Gene/P


Sistema hasta 256 racks
Rack
Cableado
32 Node Cards
Hasta 3.56 PF/s
512 o1024 TB
Node Card
32 Compute Cards
Hasta 2 I/O cards
14 TF/s
2 o 4 TB
Compute Card
1 chip, 20 DRAMs
435 GF/s
64 o 128 GB
Chip
4 procesadores
13.6 GF/s
2 o 4 GB DDR2
13.6 GF/s
8 MB EDRAM

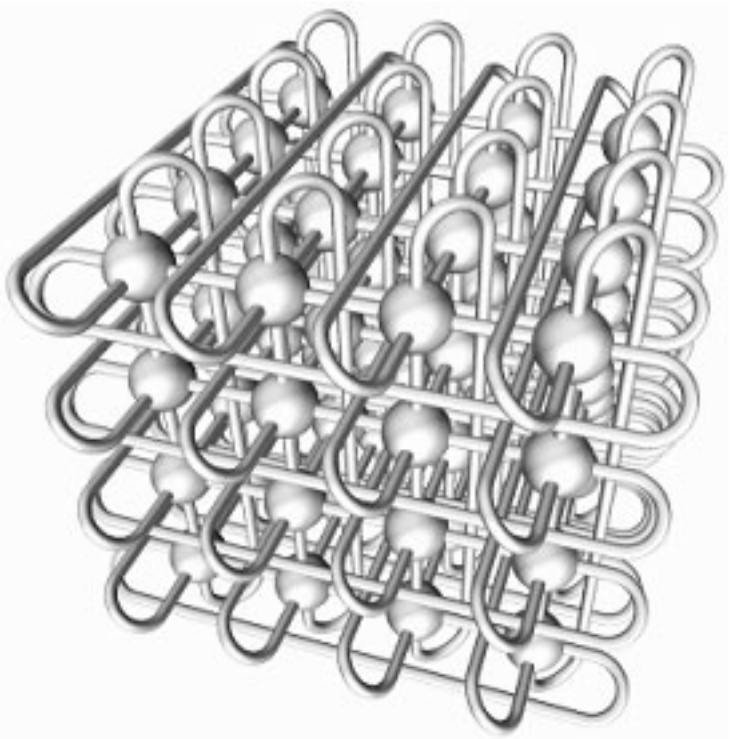


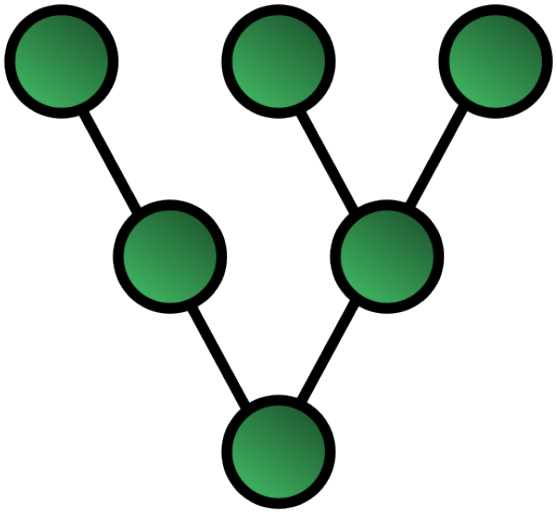
GPFS/PVFS/Lustre disponible en los nodos de E/S

Servidores de sistemas de ficheros

Topología

- Nodos de cómputo: Toro 3D
- Nodos de E/S: Árbol





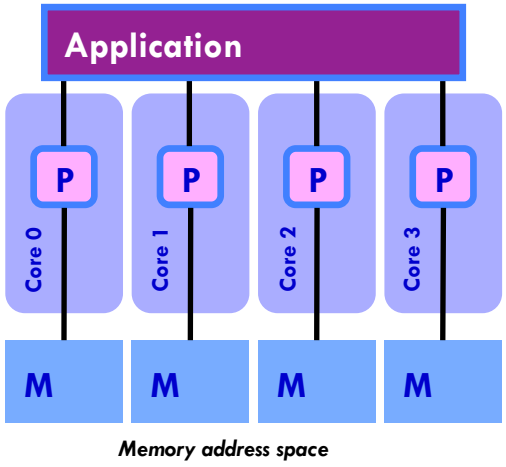
Sistema operativo

- Nodos de cómputo: SO dedicado
- Nodos de E/S: SO dedicado
- Nodos de servicio: sistemas comerciales convencionales
- Nodos cabecera (front-end): compilación, depuración, envío de trabajos
- Servidores de ficheros: no son específicos para BG (PVFS y GPFS)

BG/P Job Modes allow Flexible use of Compute Node Resources

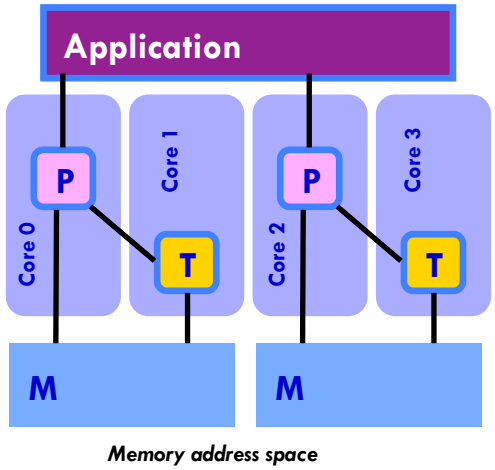
Quad Mode

- Also called Virtual Node Mode
- All 4 cores run 1 process each
- No threading
- Each process gets ¼ node memory
- MPI/HTC programming model



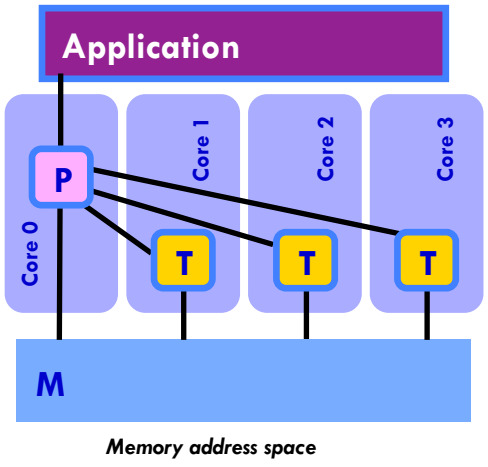
Dual Mode

- 2 cores run 1 process each
- Each process may spawn 1 thread on core not used by other process
- Each process gets ½ node memory
- MPI/OpenMP/HTC programming model



SMP Mode

- 1 core runs 1 process
- Process may spawn threads on each of the other cores
- Process gets full node memory
- MPI/OpenMP/HTC programming model



Resumen

- El SO de cada nodo está muy atado al hardware de BG.
- Parte de CPU con poca frecuenta, incremento del paralelismo.
- Separación de cada entidad consigue que BG sea:
  - Simple
  - Robusto
  - Alto rendimiento
  - Escalable
  - Extensible
- Problemas: rango de usos limitado

Lecturas avanzadas

- **Designing a highly-scalable operating system: the Blue Gene/L story -** Proceedings of the 2006 ACM/IEEE conference on Supercomputing
- www.research.**ibm**.com/**bluegene**/