Blue Gene: il primo posto (2007)

- 64 rack,131072 proc.,32768 GB RAM, 280,6 TF/s Dept.Energy's/Nat.Nuclear Security Admin's/ Lawrence Livermore National Lab.
- 280,6 TF/s record, 101,5 TF/s sustained (7 ore)
- Application area: **Not Specified**

Blue Gene

Blue Gene

- While today's machines are amazingly fast number crunchers, many data-intensive applications are slowed because of the time it takes to simply access information from the memory chips. The Blue Gene/L design will run these applications much faster because the machine will be populated with data-chip cells optimized for data access. Each chip includes two processors: one for computing and one for communicating, and its own on-board memory. Each of the data-chip cells will work on a small part of a larger problem. This increase in data access speed will make a huge difference in the kinds of results these machines can produce and the kinds of problems they can solve.

Blue Gene

- Most biological functions involve proteins and while a protein's chemical composition is determined by a sequence of amino acids joined like links of a chain, a protein folds into a highly complex, three-dimensional shape.
- It is hypothesized that the shape of a protein is the principal determinant of its function. Arbitrary strings of amino acids do not, in general, fold into a well-defined three-dimensional structure, but evolution has selected out the proteins used in biological processes for their ability to fold reproducibly
- The level of performance provided by Blue Gene (sufficient to simulate the folding of a small protein in a year of running time) is expected to enable a tremendous increase in the scale of simulations that can be carried out as compared with existing supercomputers.

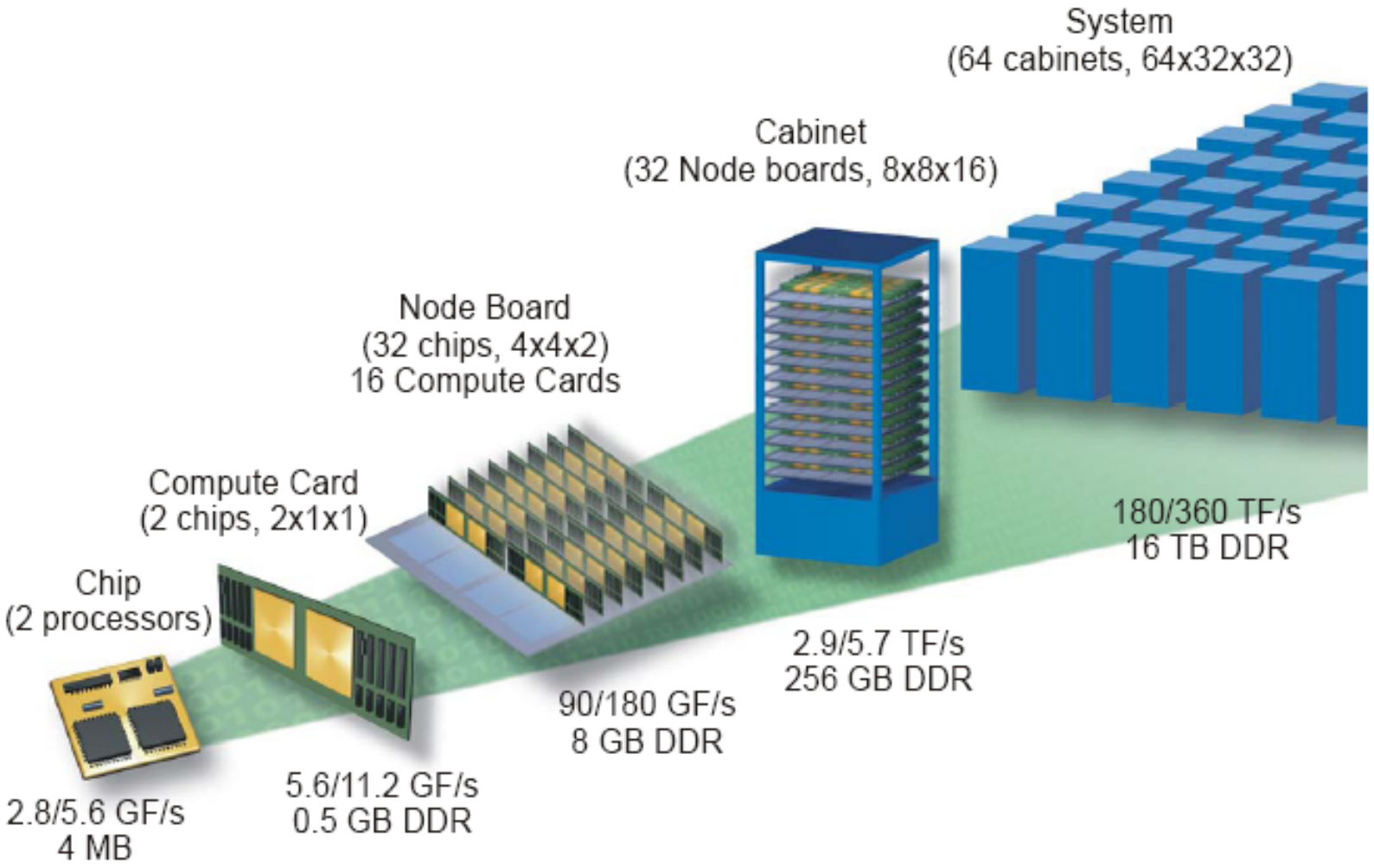
Blue Gene

- The basic software for Blue Gene will enable the system to run applications that exploit massive amounts of parallelism across chips. Particular attention has to be paid in software to high-performance, and to error recovery, as components are expected to fail during long computations. Higher-level programming environments will be developed over time, to facilitate the programming of massively parallel systems such as Blue Gene.

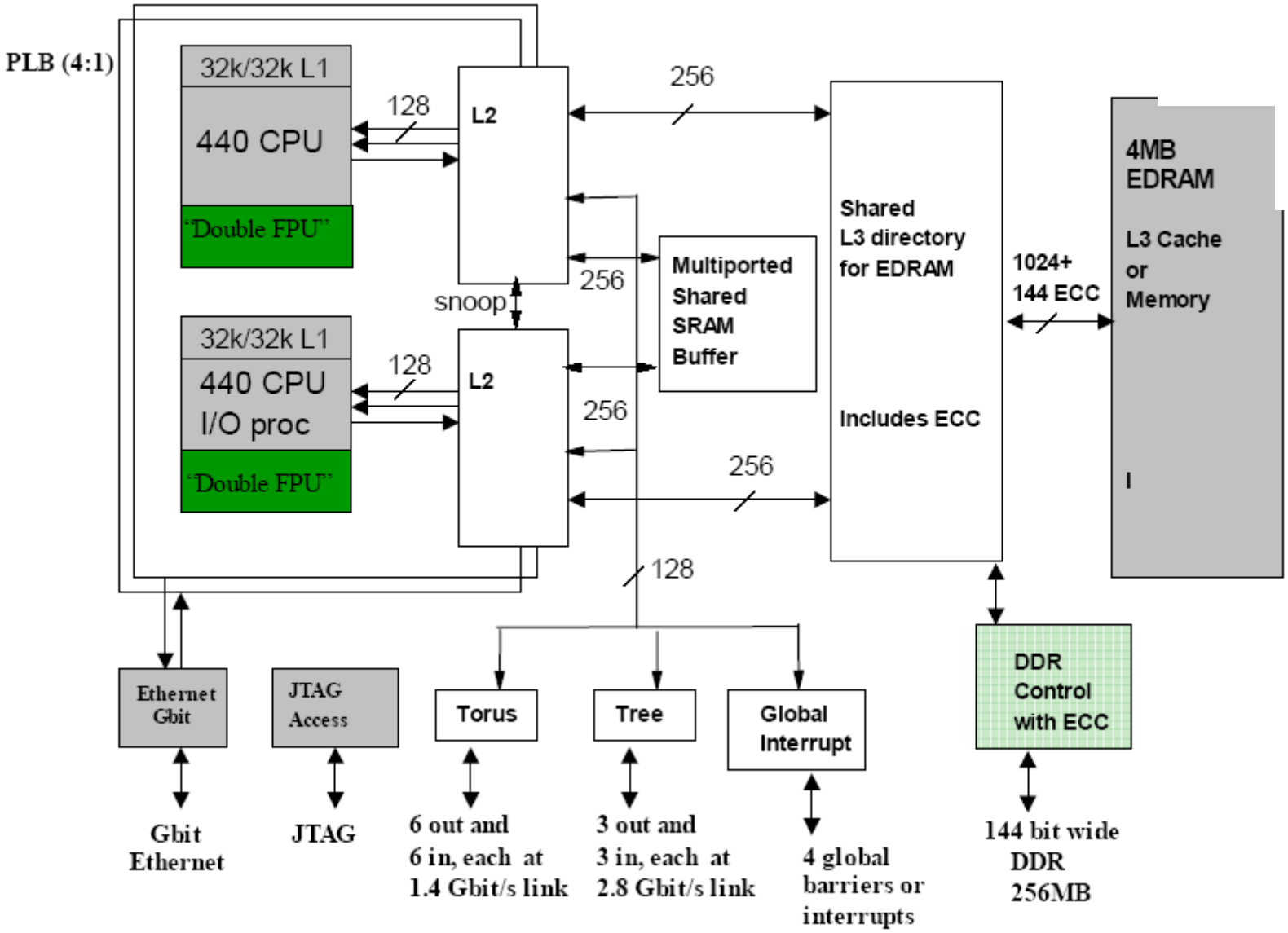
Blue Gene: curiosità

- Architettura modulare permette “small systems”: unità da 2048 processori (n°1 del 2007 ha 131072 processori)
- Il processore base è stato progettato appositamente come variante del PowerPC (FPU con istruzioni SIMD-like)
- Il clock del processore è solo 700 MHz (meno potenza, packaging più denso): più FLOPS per Watt e per m^2
- La cache L2 è più piccola della cache L1
- Chip biprocessore: si puo’ usarne uno per calcolare e uno per comunicare
- 512 MB RAM per ogni chip biprocessore
- Per battere i record si usano tutti e due per calcolare

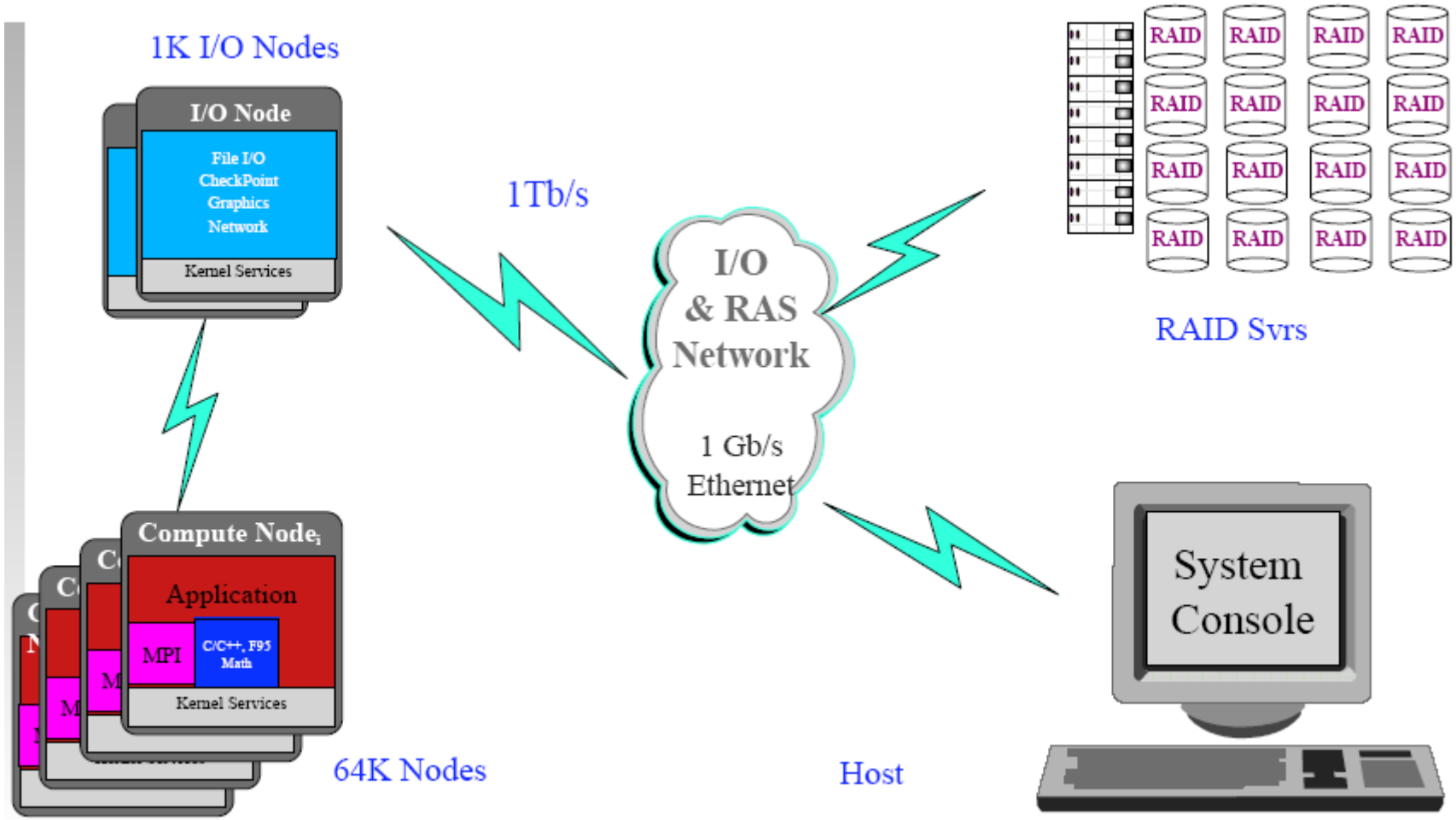
Blue Gene: architettura



Blue Gene: compute ASIC



Blue Gene: operating environment



Blue Gene System software

- Software environment includes:
  - High performance Scientific Kernel
  - MPI-2 subset, defined by users
  - Math libraries subset, defined by users
  - Compiler support for DFPU (C, C++, Fortran)
  - Parallel file system
  - System management

- External Collaborations: Boston University, Caltech, Columbi University, Oak Ridge National Labs, San Diego Supercomputing Center, Universidad Politecnica de Valencia, University of Edinburgh, University of Maryland,Texas A&M, Tech. Univ. of Vienna...

Software Available for Blue Gene

- The following High Performance Computing cluster software is now available on Blue Gene: the Engineering and Scientific Subroutine Library (ESSL) for Linux on POWER, General Parallel File System (GPFS) for Linux on POWER, and LoadLeveler for Linux on POWER. ESSL provides over 150 math subrouties that have been specifically tuned for performance on Blue Gene. GPFS is the top performing cluster-wide file system for Blue Gene, providing superior scalability and high reliability. And LoadLeveler is a job scheduler designed to maximize resource utilization and job throughput to get the most out of the available resources.

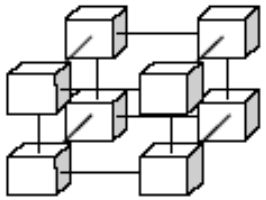
ma nelle installazioni HPC …

- Ogni nodo “processing” fa girare un lightweight kernel sviluppato appositamente: un solo processo alla volta senza process switch o demand paging
- Alcuni nodi fanno I/O e interazione con gli utenti e qui gira “Linux”

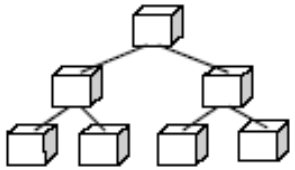
Interconnessioni

- 5 reti
- Gb Ethernet per colloquiare col mondo dei calcolatori “normali”
- Rete “di controllo” Ethernet
- 3-D torus per message passing (max 64 hops)
- A collective network (p.e. somme globali)
- A barrier network

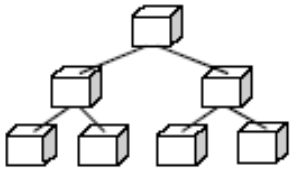
Blue Gene: interconnessioni



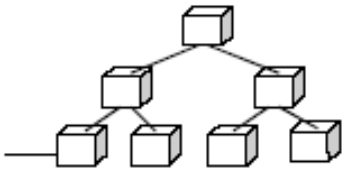
3 Dimensional Torus
- Point-to-point



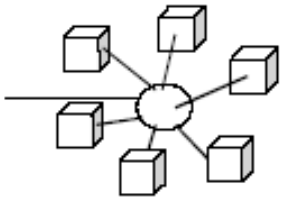
Global Tree
- Global Operations



Global Barriers and Interrupts
- Low Latency Barriers and Interrupts

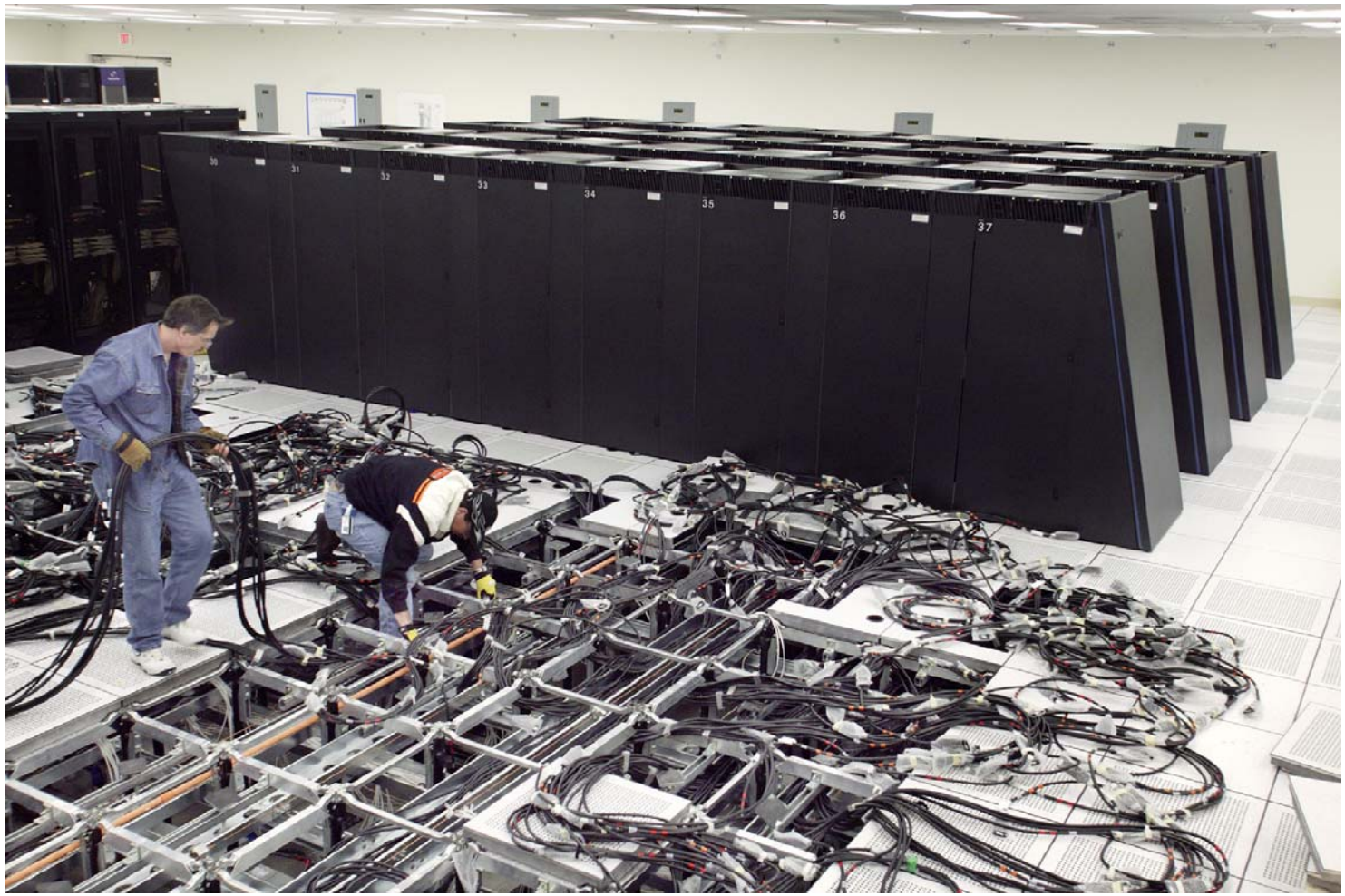


Gbit Ethernet
- File I/O and Host Interface



Control Network
- Boot, Monitoring and Diagnostics

Blue Gene: interconnessioni



Limitazioni

- Pensato per simulazioni di fenomeni fisici e altri problemi con molta località nei nodi
- Può rallentare molto se ci fossero da scambiare molti dati
- Altri MPP hanno meno TF ma sono meglio come bandwidth tra nodi e potrebbero “vincere” in certi casi

Fault tolerance

- Un guasto ogni 6 giorni
- Schede RAM saldate (!) per evitare problemi sui connettori
- Capacità di isolare le schede guaste
- Recovery parziale del processo perso tramite checkpointing