

IBM Research

# The IBM BlueGene/L Machine

José E. Moreira  
IBM T. J. Watson Research Center

SOS7 – Durango, CO – March 4-6, 2003

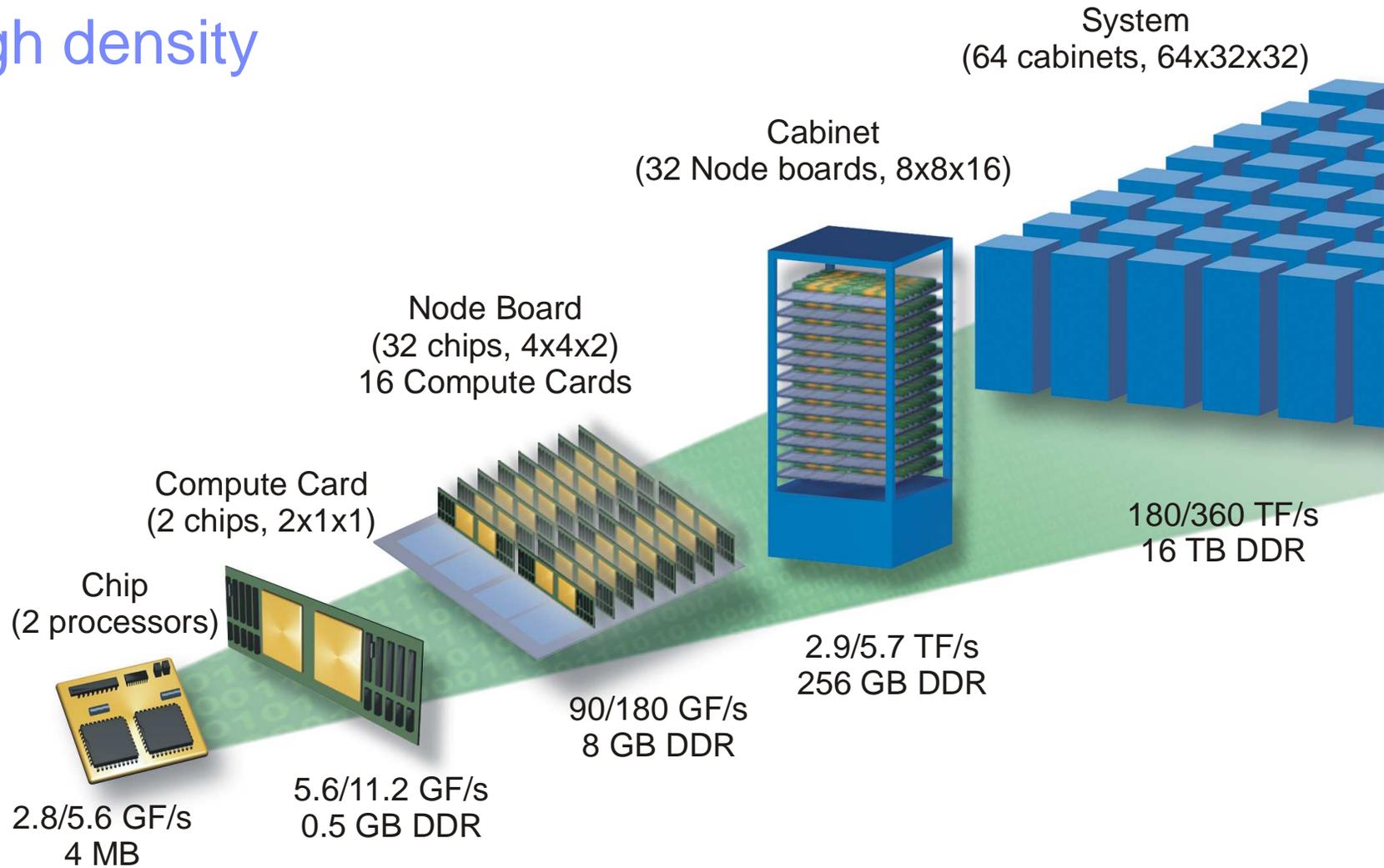
## The four questions:

1. What is unique in structure and function of your machine?
2. What characterizes your applications? Examples are: Intensities of message passing, memory utilization, computing, IO, and data.
3. What prior experience guided you to this choice?
4. Other than your own machine, for your needs what are the best and worst machines? And, why?

## What is unique in structure and function of your machine?

- What you really want to ask is: What are the defining characteristics of the BlueGene/L machine?
  1. High density
  2. Scalable interconnect
  3. Hierarchical system software

# High density



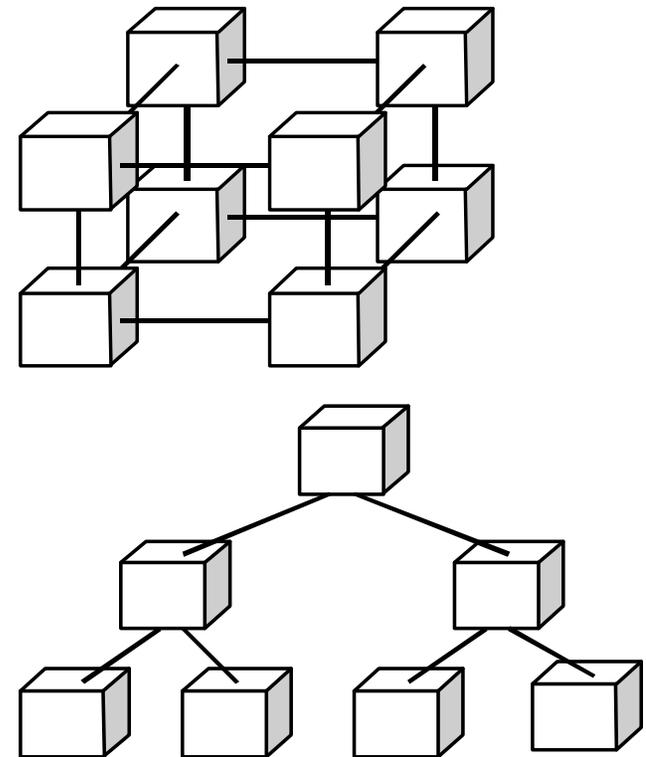
# Scalable interconnect

## 3 Dimensional Torus

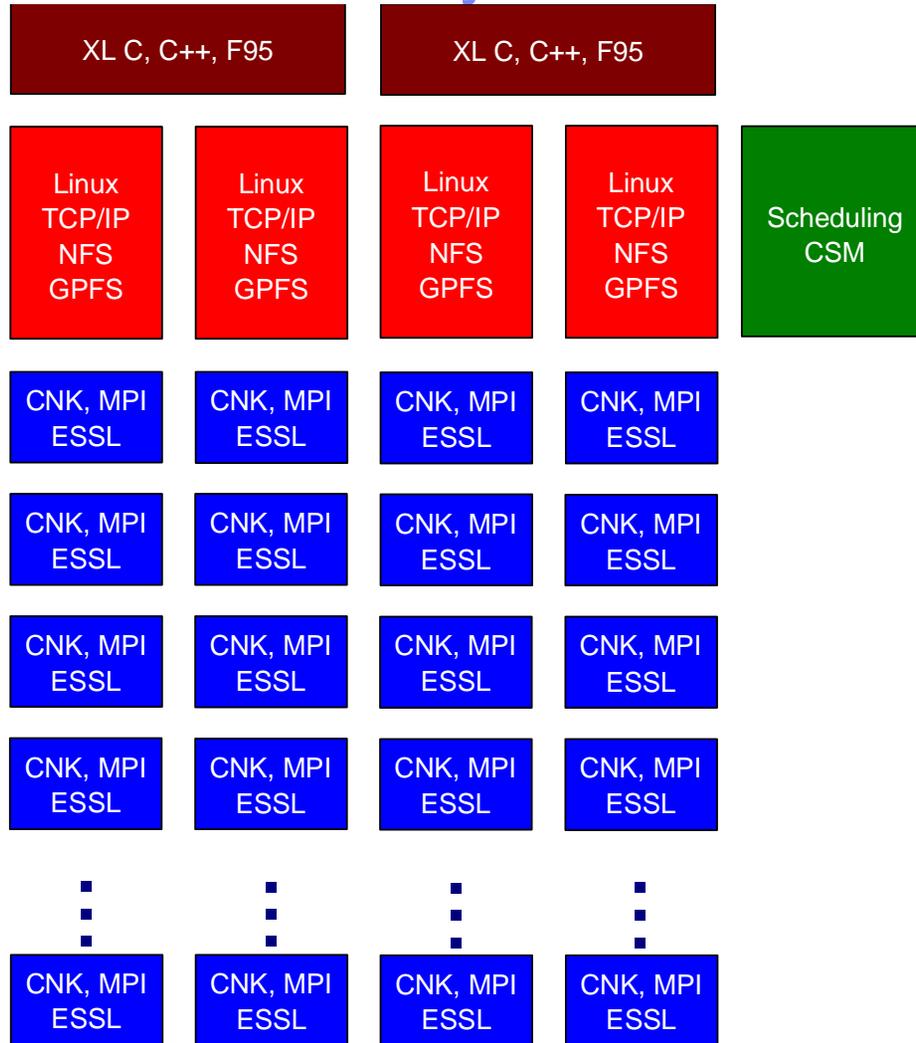
- ✓ Interconnects all compute nodes (65,536)
- ✓ Virtual cut-through hardware routing
- ✓ 1.4Gb/s on all 12 node links (2.1 GB/s per node)
- ✓ Communications backbone for computations
- ✓ 350/700 GB/s bisection bandwidth

## Global Tree

- ✓ One-to-all broadcast functionality
- ✓ Reduction operations functionality
- ✓ 2.8 Gb/s of bandwidth per link
- ✓ Latency of tree traversal in the order of 2  $\mu$ s
- ✓ Interconnects all compute and I/O nodes (1024)



# Hierarchical system software



System software is based on proven solutions to large cluster operating environments:

- OS: Linux, Compute Node Kernel
- Communication: MPI, TCP/IP
- Math libraries: ESSL
- File System: GPFS, NFS
- Compilers: IBM XL C, C++, Fortran95
- System Management: CSM
- Job scheduling: SLURM

- compute node
- I/O node
- service node
- front-end node

## What characterizes your applications?

- What you really want to ask is: What kind of applications run well on the BlueGene/L machine?
  1. High computation to memory ratio: 5.6 Gflops and 256 Mbytes
  2. Communication locality: 2.1 Gbytes/sec of nearest neighbor bandwidth/node, 10 Mbytes/sec of bisection bandwidth/node
  3. High computation to I/O ratio: 5.6 Gflops and 16 Mbits/sec
- For example:
  1. Numerical linear algebra, finite differences, and molecular dynamics are examples of codes that run well
  2. FFT (limited by bisection bandwidth), reservoir simulation (limited by memory), and seismic (limited by I/O) are examples of codes that will not run so well

## What prior experience guided you to this choice?

- Experience with QCD machines, particularly QCDOC
  - ▼ 8,000-node machine at Columbia University
  - ▼ 12,000-node machine at Brookhaven National Laboratory
- Successful machines with nearest-neighbor interconnect, for example Cray T3D and ASCI Red
- Experience with managing Linux clusters – success with managing 1000-node clusters
- Experience with developing system software for BlueGene/C
  - ▼ Lightweight kernel
  - ▼ Separation between compute and system nodes

## Other than your own machine, for your needs what are the best and worst machines?

- My need is to develop system software, so all other machines are bad :-)
- Cray X1/X2, ASCI Red Storm, ASCI Purple, Blue Planet, Intel clusters are all likely to become significant scientific and engineering resources!
- More computing is always better. There is no wrong answer!