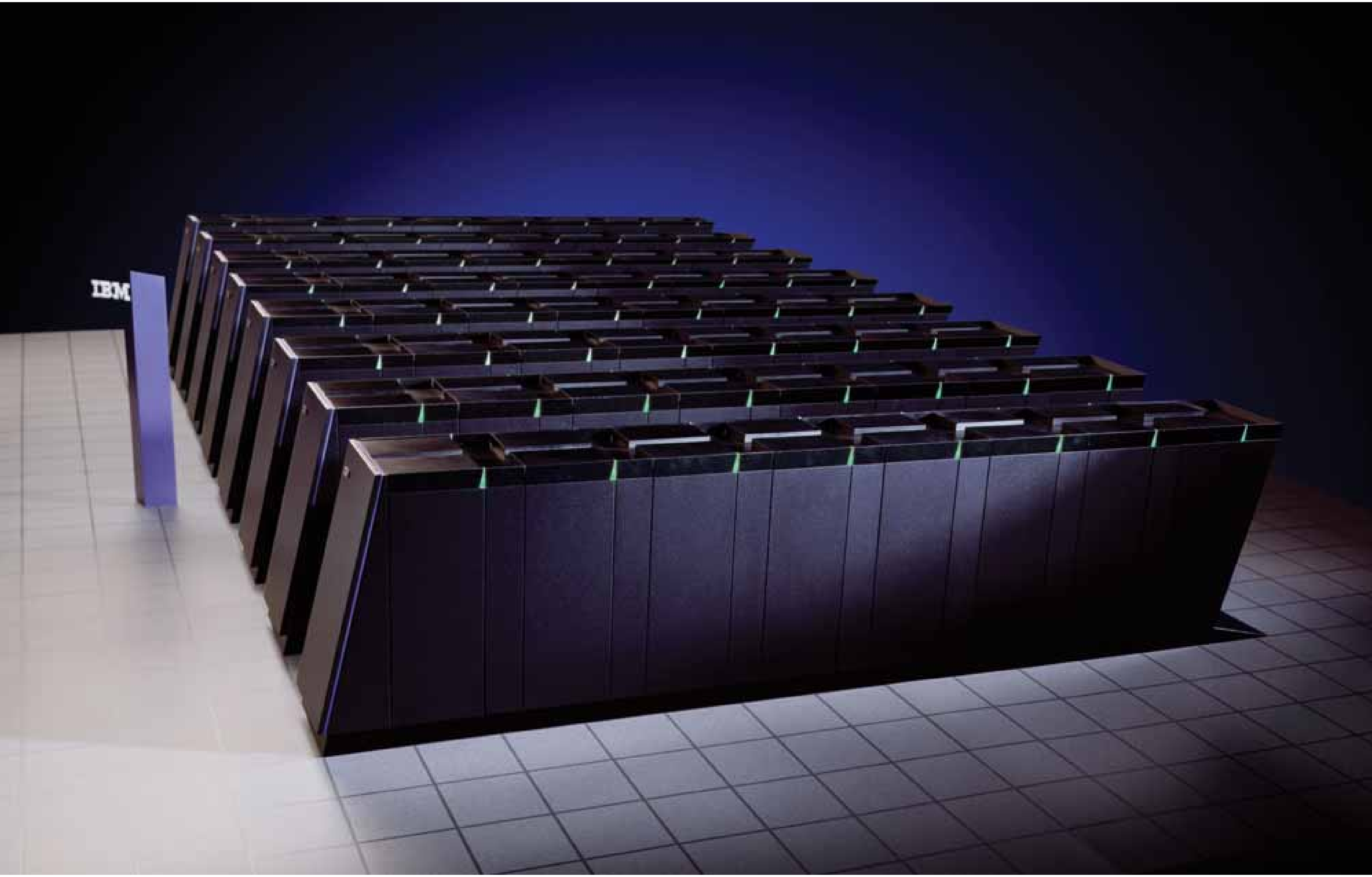
BlueGene

Luigi Brochard
EMEA HPC Architect
IBM Deep Computing

IBM
IBM
IBM



IBM

- Motivations
- Architecture
- Software
- Applications

BlueGene/L Project Motivations

- **Traditional supercomputer-processor design is hitting power/cost limits**
- Complexity and power are major driver for cost and reliability
- Optimal design point is very different from standard approach based on high-end superscalar nodes
- Integration, power, and technology directions are driving toward multiple modest cores on a single chip rather than one high-performance processor
  - Watts/FLOP will not improve much from future technologies.
- Applications for supercomputers do scale fairly well
  - Growing volume of parallel applications
  - Physics is mostly local
- But collective communications are becoming most important on large parallel systems

Performance per rack

type	POWER5	Xeon EMT
peak TF	0.7	0,6
#cpu	96	84
max memory GB	3072	1344
frequency GHz	1.9	3.6
technology um	0.13	0.13

What about a 360 TFlops system:

Floor

Cluster ~1000 SqM (extrapolation)

Power

Cluster ~15 MW (extrapolation)

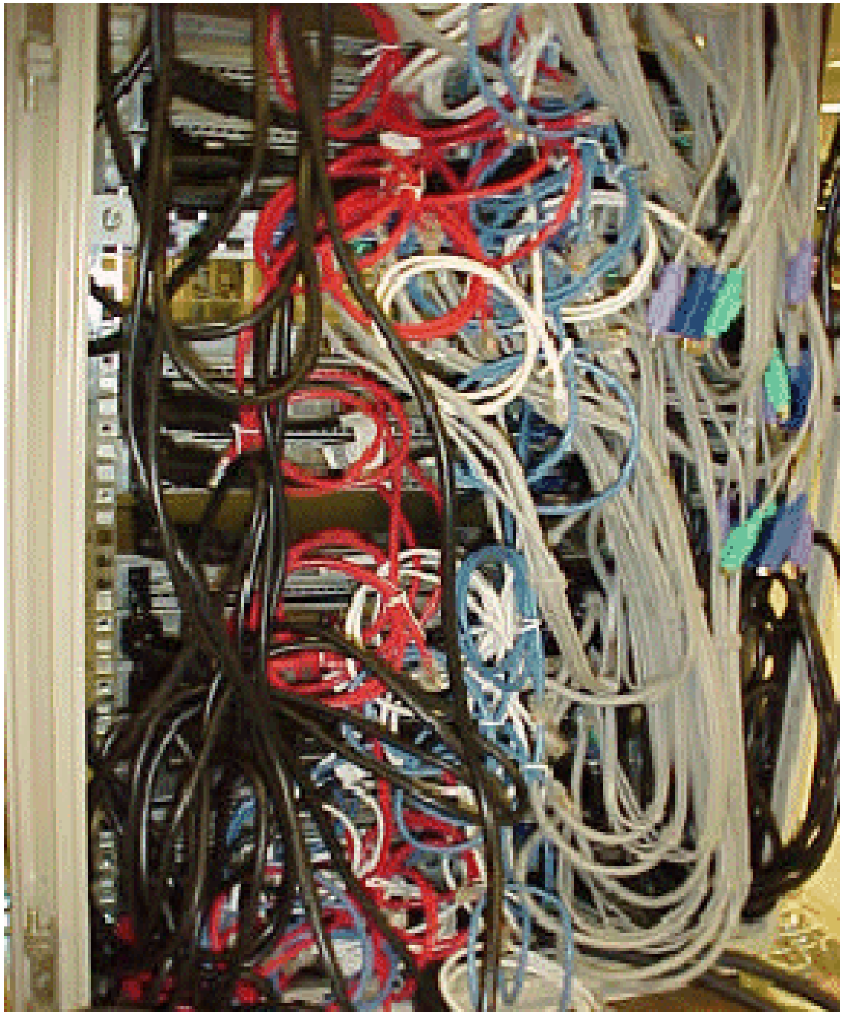
Cost of energy:

1 MW/yr = $1M

BlueGene/L Project Motivations

- Traditional supercomputer-processor design is hitting power/cost limits
- **Complexity and power are major driver for cost and reliability**
- Optimal design point is very different from standard approach based on high-end superscalar nodes
- Integration, power, and technology directions are driving toward multiple modest cores on a single chip rather than one high-performance processor
  - Watts/FLOP will not improve much from future technologies.
- Applications for supercomputers do scale fairly well
  - Growing volume of parallel applications
  - Physics is mostly local
- But collective communications are becoming most important on large parallel systems

Complexity of wiring



BlueGene/L Project Motivations

- Traditional supercomputer-processor design is hitting power/cost limits
- Complexity and power are major driver for cost and reliability
- Optimal design point is very different from standard approach based on high-end superscalar nodes
- Integration, power, and technology directions are driving toward multiple modest cores on a single chip rather than one high-performance processor
  - Watts/FLOP will not improve much from future technologies.
- Applications for supercomputers do scale fairly well
  - Growing volume of parallel applications
  - Physics is mostly local
- Collective communications are becoming most important on large parallel systems

IBM approach

- Use embedded system-on-a-chip (SOC) design
  - **Significant reduction of complexity**
    - **Simplicity is critical, enough complexity already due to scale**
  - **Significant reduction of power**
    - **Critical to achieving a dense and inexpensive packaging solution.**
    - **An absolute requirement for the future. (PowerPC 440 processor is ~1Watt)**
  - **Significant reduction in time to market, lower development cost and lower risk**
    - **Much of the technology is qualified.**
- Utilize PowerPC architecture and standard messaging interface (MPI).
  - **Standard, familiar programming model and mature compiler support.**
- Advantages in utilizing SOC technique
- Integrated and tighly coupled networks
  - **To sustain performance of applications on large number of nodes**
- Close attention to RAS (reliability, availability, and serviceability) at all system levels.
  - **One of the biggest challenges**

The BlueGene/L Project from a High Level

- A 64k-node highly integrated supercomputer
- 180–360 teraflops peak performance
- Based on embedded system-on-a-chip (SOC) technology
- Only two ASICs: node and link
- Focuses on numerically intensive scientific problems
- “Grand challenge” science projects in partnership with LLNL and high-performance computing customers
  - Validate and optimize architecture using real applications
  - Help us investigate the reach of this machine

Performance per rack

type	POWER5	Xeon EMT	BG/L
peak TF	0.7	0,6	5.8
#cpu	96	84	2048
max memory GB	3072	1344	512
frequency GHz	1.9	3.6	0.7
technology um	0.13	0.13	0.13

What about 360 TFlops system:

Floor

BG/L ~100 SqM

Cluster ~1000 SqM (extrapolation)

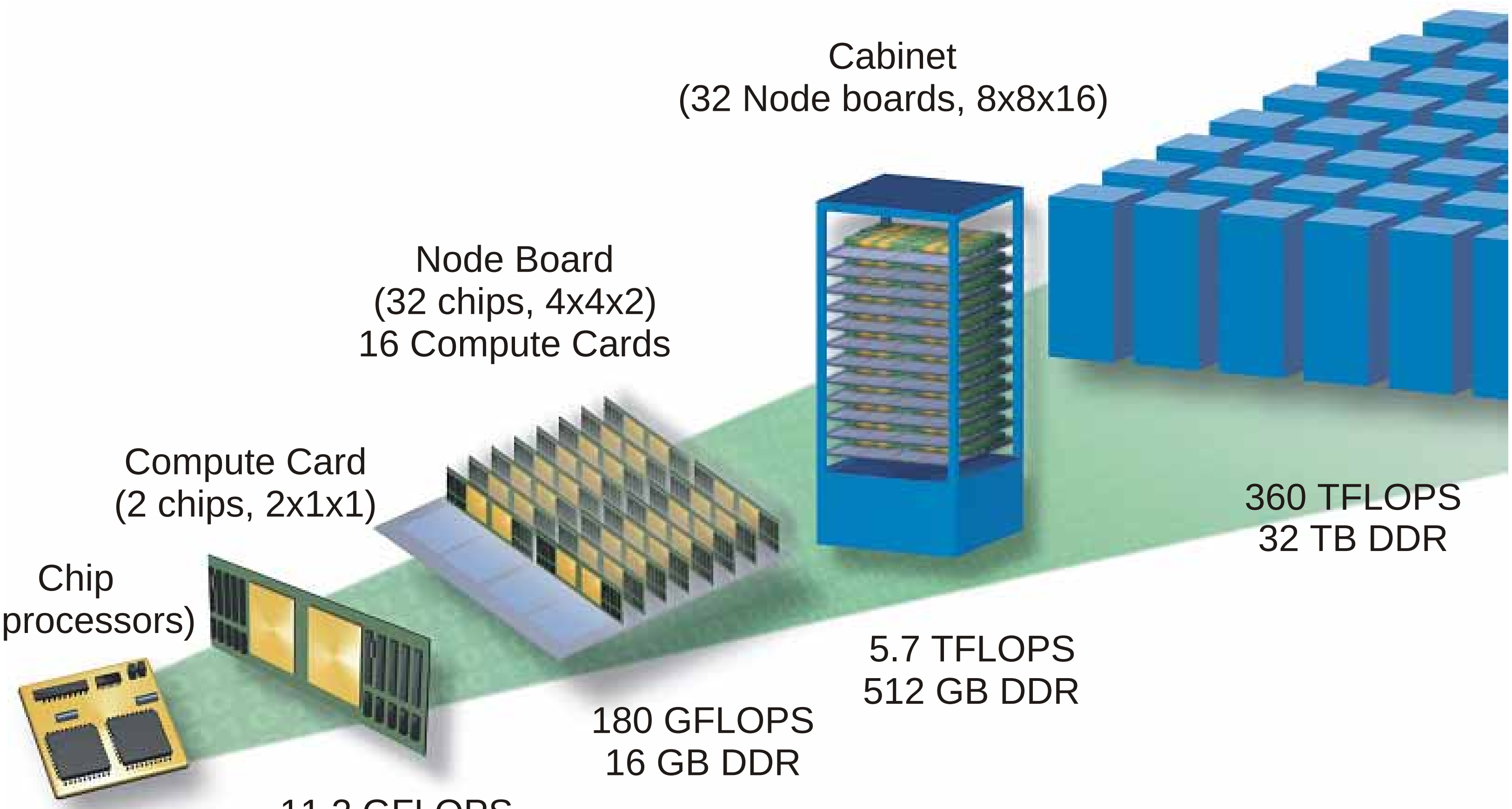
Power

BG/L ~1.5 MW

Cluster ~15 MW (extrapolation)

Architecture

BlueGene/L



System
(64 cabinets, 64x32x32)

Cabinet
(32 Node boards, 8x8x16)

Node Board
(32 chips, 4x4x2)
16 Compute Cards

Compute Card
(2 chips, 2x1x1)

Chip
(2 processors)

2.8/5.6 GFLOPS
per processor/chip 4 MB

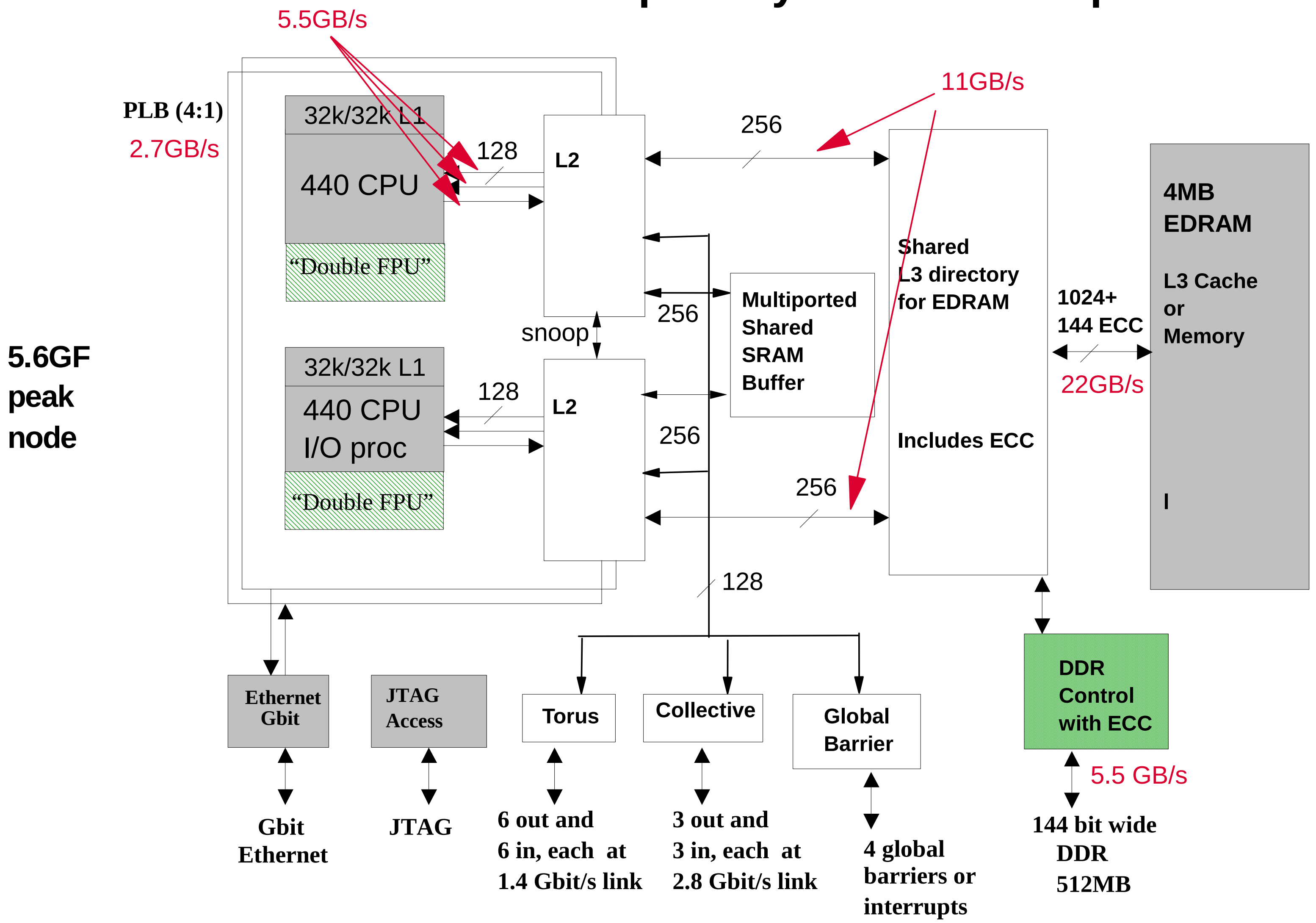
11.2 GFLOPS
1.0 GB DDR

180 GFLOPS
16 GB DDR

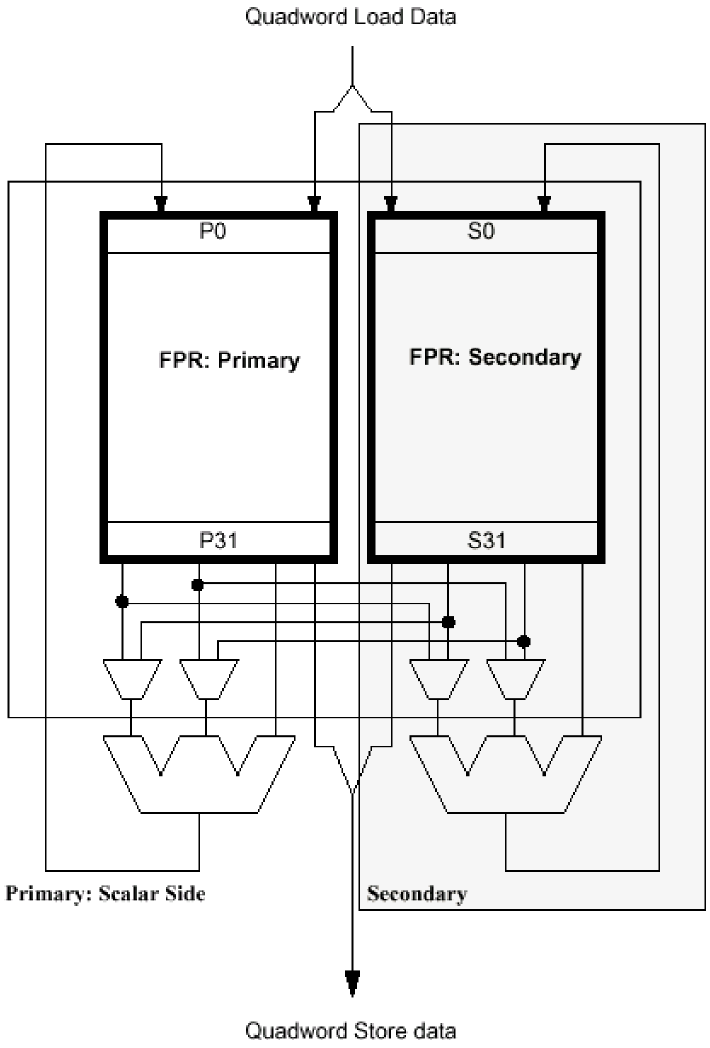
5.7 TFLOPS
512 GB DDR

360 TFLOPS
32 TB DDR

BlueGene/L Compute System-on-a-Chip ASIC



Double Hummer Floating-Point Unit

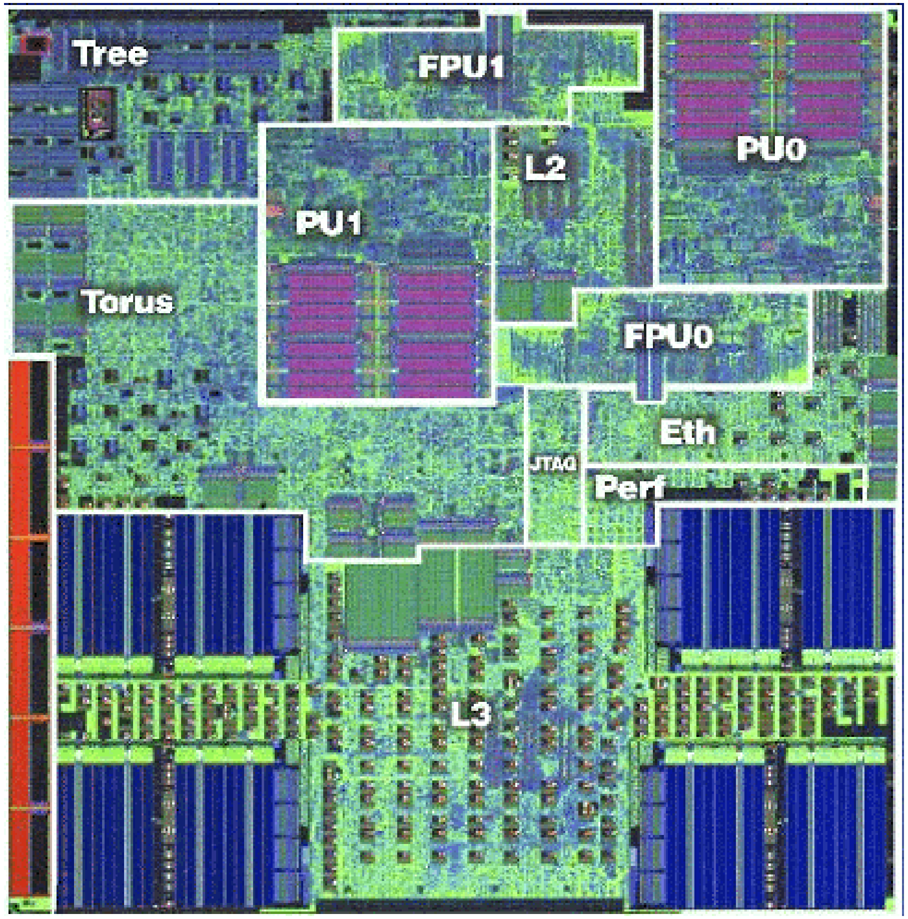


- 64-bit FPU
- Two replicas of a standard single-pipe PowerPC FPU
  - **2 x 32 64-bit registers**
- Enhanced ISA, includes instructions:
  - **Executed in either pipe**
  - **Simultaneously execute the same operation on both sides – SIMD instructions**
  - **Simultaneously execute two different operations of limited types on different data**

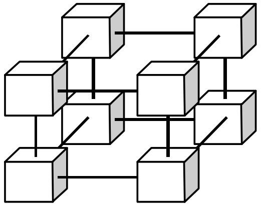
BlueGene/L Chip Physical Design and Power Measurements

- **IBM CMOS 0.13µ Cu-11 ASIC process technology**
- **123mm^2 chip, 95M transistors, ~15 Watt**

Unit	Active power (W)	Size (cells)
Clock tree + access	1.15	264k
CPU/FPU/L1	7.54	14,700k
Torus network	0.67	4,963k
Collective network	0.25	2,350k
L2/L3/DDR control	2.6	18,310k
Others	0.49	10,720k
Leakage	0.2	

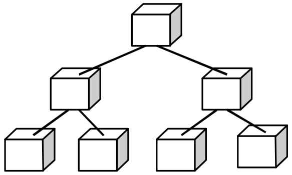


BlueGene/L - Five Independent Networks



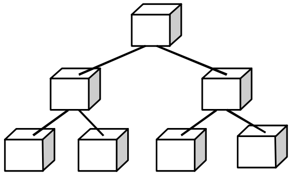
3 Dimensional Torus

- 32x32x64 connectivity
- Backbone for one-to-one and one-to-some communications
- 1.4 Gb/s bi-directional bandwidth in all 6 directions (Total 2.1 GB/s/r
- ~100 ns hardware node latency



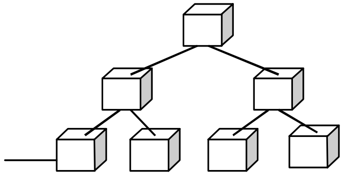
Collective Network

- Global Operations
- 2.8Gb/s per link , 68TB/s aggregate bandwidth
- Arithmetic operations implemented in tree
  - Integer/ Floating Point Maximum/Minimum
  - Integer addition/subtract, bitwise logical operations
- Latency of tree less than 2.5usec to top, additional 2.5usec to br
- Global sum over 64k in less than 2.5 usec (to top of tree)



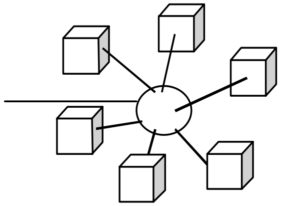
Global Barriers and Interrupts

- Low Latency Barriers and Interrupts



Gbit Ethernet

- File I/O and Host Interface
- Funnel via Global Tree network



Control Network

- Boot, Monitoring and Diagnostics

Optimizing point-to-point communication (short messages: 0-10 KBytes)

- The thing to watch is overhead

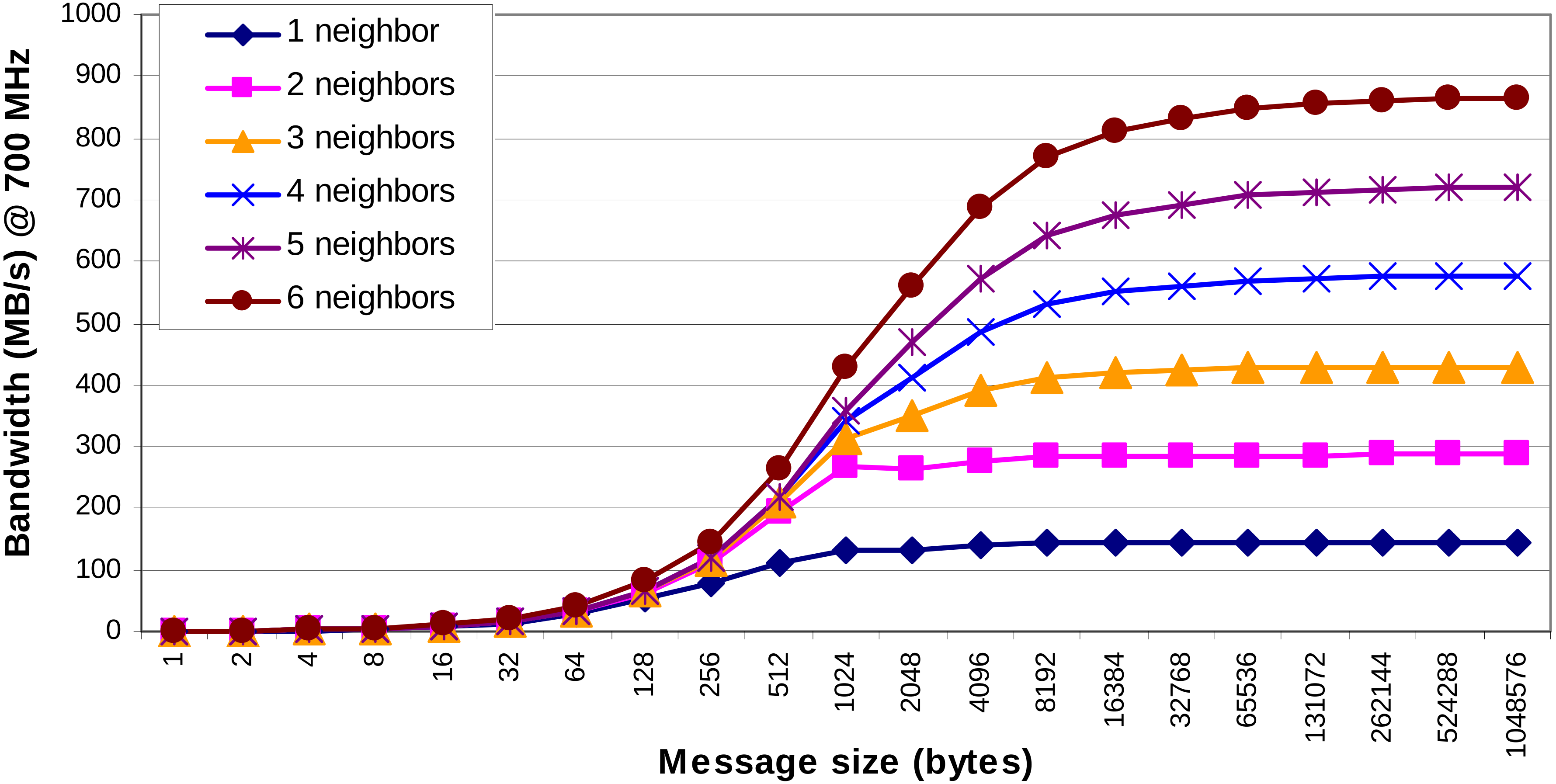
protocol	cycles	µs
short	2350	3.35
eager	4000	5.71
rendezvous	11000	15.71

- Bandwidth
- CPU load
- Co-processor
- Network load

Not a factor: not enough network traffic

- Single-packet protocol: save overhead of chopping up message into packets

Measured MPI Send Bandwidth and Latency



Latency @700 MHz = 3.3 + 0.090 * “Manhattan distance” + 0.045 * “Midplane hops” λs

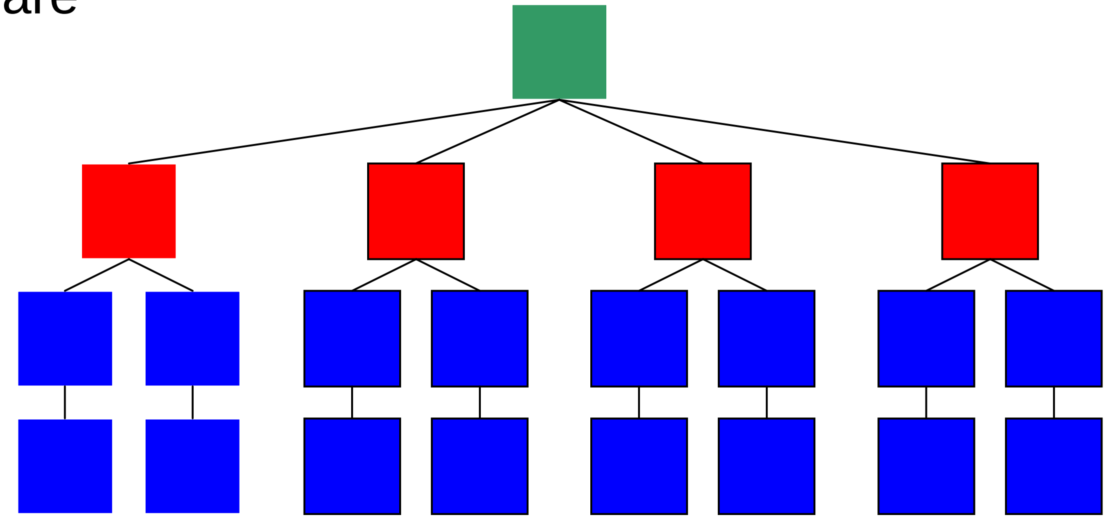
BG/L is a well balanced system

System	Memory Bandwidth GB/s Byte/Flop	Memory Latency ns cycles	Network Latency us cycles	Network Barrier 128 cpu us cycles
BG/L	2.2 0,39	110 77	3.35 2345	6,75 4725
Xeon Infiniband	2.8 0,19	140 504	4.98 17900	79,9 287640
POWER5 Federation	35 0,55	120 240	4.5 9000	29,5 59000

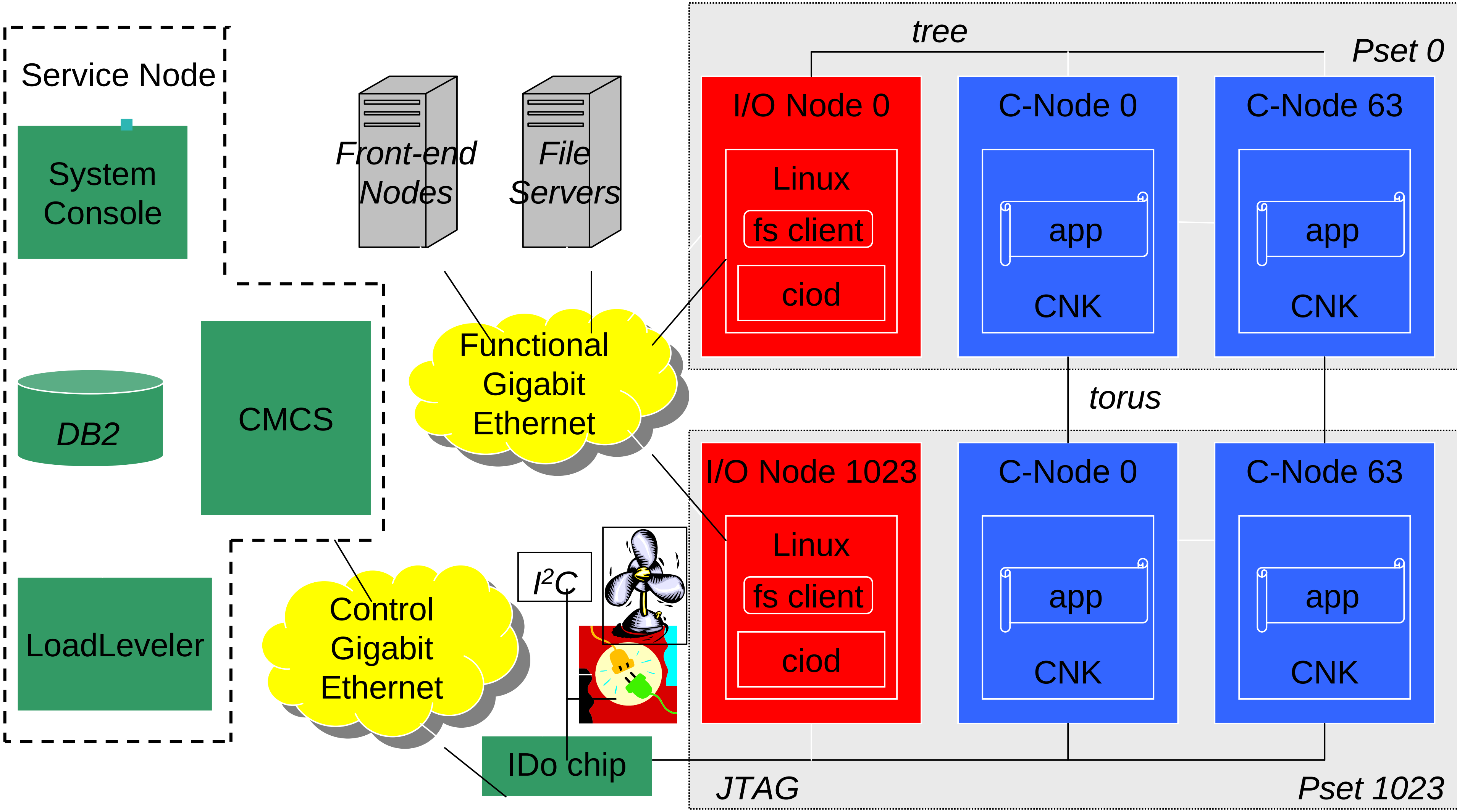
Software

BlueGene/L Software Hierarchical Organization

- **Compute nodes** dedicated to running user application, and almost nothing else - simple compute node kernel (CNK)
- **I/O nodes** run Linux and provide a more complete range of OS services – files, sockets, process launch, signaling, debugging, and termination
- **Service node** performs system management services (e.g., heart beating, monitoring errors) - transparent to application software



BlueGene/L System Architecture



BG/L Integration with pSeries

- **Programmer's view: Nearly identical software stack/interface to pSeries**
  - Compilers: IBM XLF, XLC, VA C++ compilers, hosted on PPC/Linux
  - Operating System: Linux-compatible kernel with some restrictions (not a problem for most HPC codes)
  - Message passing library: MPI
  - Math libraries: ESSL, MASS, MASSV
  - Parallel file system: GPFS
  - Job scheduler: LoadLeveler
- **System administrator's view: Look and feel of a PPC Linux cluster, managed from a PPC/Linux host, but diskless and managed by a novel control system**
  - Potential technologies to migrate to BladeCenter and clusters
    - Database-centric control system design – enhanced robustness, security
    - Ethernet based communication for control, via IDO chip

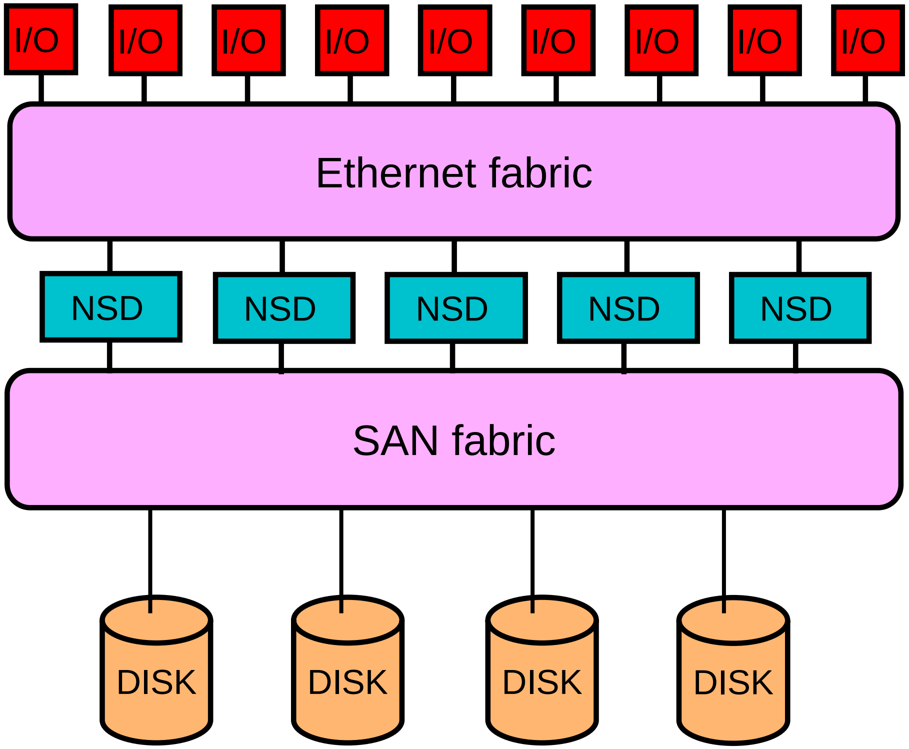
Job Scheduling in BlueGene/L

- **LoadLeveler solution**
  - BG/L specific job scheduler plugged into LoadLeveler as external scheduler
  - Working on a integrated, internal scheduler, solution
- **Job scheduling strategies can significantly impact the utilization of large computer systems**
  - Machines with toroidal topology (as opposed to all-to-all switch) are particularly sensitive to job scheduling – this was demonstrated at LLNL with gang scheduling on Cray T3D
  - BG/L scheduling strategies leveraging BG/L unique topology features can significantly enhance system utilization – from 45% to almost 90% (depends on workload)

Parallel File System for BlueGene/L

- **BlueGene/L can generate enormous I/O demand**
  - 10 GB/sec of writes per I/O-rich rack
  - 6 GB/sec of reads per I/O-rich rack
- **Serving this kind of demand requires a parallel file system**
- **GPFS is being ported on BlueGene/L**

GPFS for BlueGene/L



- GPFS solution for BlueGene/L is 3-tiered
  - First tier consists of the I/O nodes, which are GPFS clients
  - Second tier is a cluster of NSD (Network Shared Disk) servers
  - Third tier is a set of storage devices, typically Fiber Channel or iSCSI
- First-to-second tier interconnect has to be Ethernet
- Second-to-third tier can be fiber channel loop, fiber channel switch, or Ethernet (for iSCSI)
- Choice of NSD servers, SAN fabric and storage devices depends on specific requirements

Programming Models and Development Environment

- Familiar Aspects
  - SPMD model - Fortran, C, C++ with MPI (MPI1 + subset of MPI2)
    - Full language support
    - Automatic SIMD FPU exploitation
  - Linux development environment
    - User interacts with system through FE nodes running Linux – compilation, job submission, debugging
    - Compute Node Kernel provides look and feel of a Linux environment – POSIX system calls (with restrictions)
  - Tools – support for debuggers (Aetnus TotalView), MPI tracer, profiler, hardware performance monitors, visualizer (HPC Toolkit, Paraver, Kojak)
- Restrictions (lead to significant scalability benefits)
    - Strictly space sharing - one parallel job (user) per partition of machine, one process per processor of compute node
    - Virtual memory constrained to physical memory size

Math Libraries: ESSL

- **Started with small subset (of ~500 routines)**
  - Mainly dense matrix kernels – DGEMM, DGEMV, DDOT, DAXPY etc.
  - Exploiting second CPU for computation-intensive kernels
- **Using ESSL source code to drive compiler testing and exploration of complete ESSL support**
  - Status: Nearly complete functionality available using –O3 –qarch=440
  - Currently investigating SIMD FPU issues, performance enhancements
  - Expected general availability – Nov 2005
- **FFT**
  - Technical University of Vienna developing FFT library optimized for BlueGene/L – effective use of the SIMD FPU

Applications

Some Applications on Blue Gene

- Blue Matter (IBM) *
- Flash (ANL) *
- CTH (Sandia)
- MM5
- Amber7, Amber8
- GAMESS
- QMC (Caltech)
- LJ (Caltech)
- PolyCrystal (Caltech)
- PMEMD (LBL)
- **Miranda (LLNL) ***
- LSMS (ORNL)
- NIWS (NISSEI)
- HOMME (NCAR) *
- **QBox (LLNL)**
- **ddcMD (LLNL)**

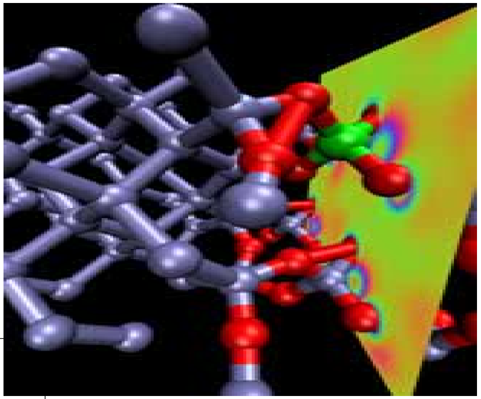
SAGE (LANL)
SPPM (LLNL)
UMT2K (LLNL)
Sweep3d (LANL)
MDCASK (LLNL)
GP (LLNL)
CPMD (IBM) *
TLBE (LBL)
HPCMW (RIST)
Paradis (LLNL)
SPHOT (LLNL)
QCD (IBM)*, QCD (BU) *
NAMD
PAM-CRASH (ESI)
Raptor (LLNL) *
Enzo (San Diego)

Lawrence Livermore National Lab Apps
SC04_BGL_Apps_LLNL.pdf

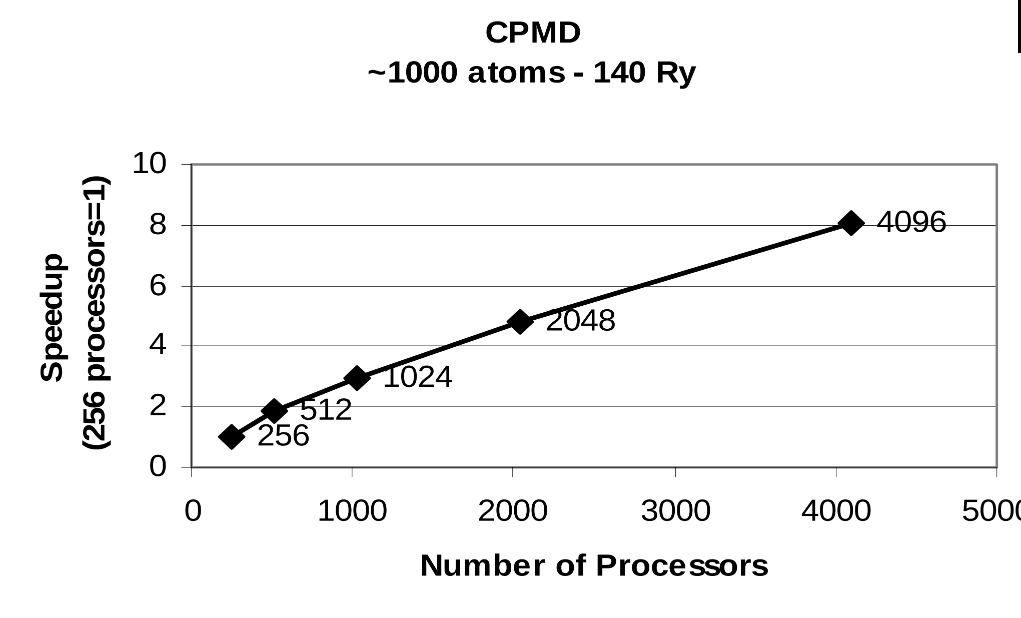
- MIRANDA: Hydrodynamics, Fluid Instability, Mixing, supernovae, Inertial Confinement Fusion
- RAPTOR: Eulerian AMR, Instabilities, Fusion Confinement, Astrophysics
- Qbox: MD, Plane Wave Pseudopotentials, DFT, nanotechnology/biochemistry
- ddcMD: Classical MD, classical metals, actinides under extreme conditions
- MDCASK: Atomic dynamics with Newtonian Mechanics and Electrostatics
- ParaDiS: Parallel Dislocation Simulator, strength of materials, AMR

CPMD - Alessandro Curioni, Salomon Billeter, Wanda Andreoni

Developed at IBM Zurich and other Universities
from Car Parinello method for Molecular Dynamics
Uses Plane Wave Basis functions, FFT, MPI_Collectives
Demo – Si/SiO2 Interface **% peak ~ 60 % VNM**
Ongoing project : IBM/LLNL PdH Hydrogen Storage



Ab-initio Molecular Dynamics

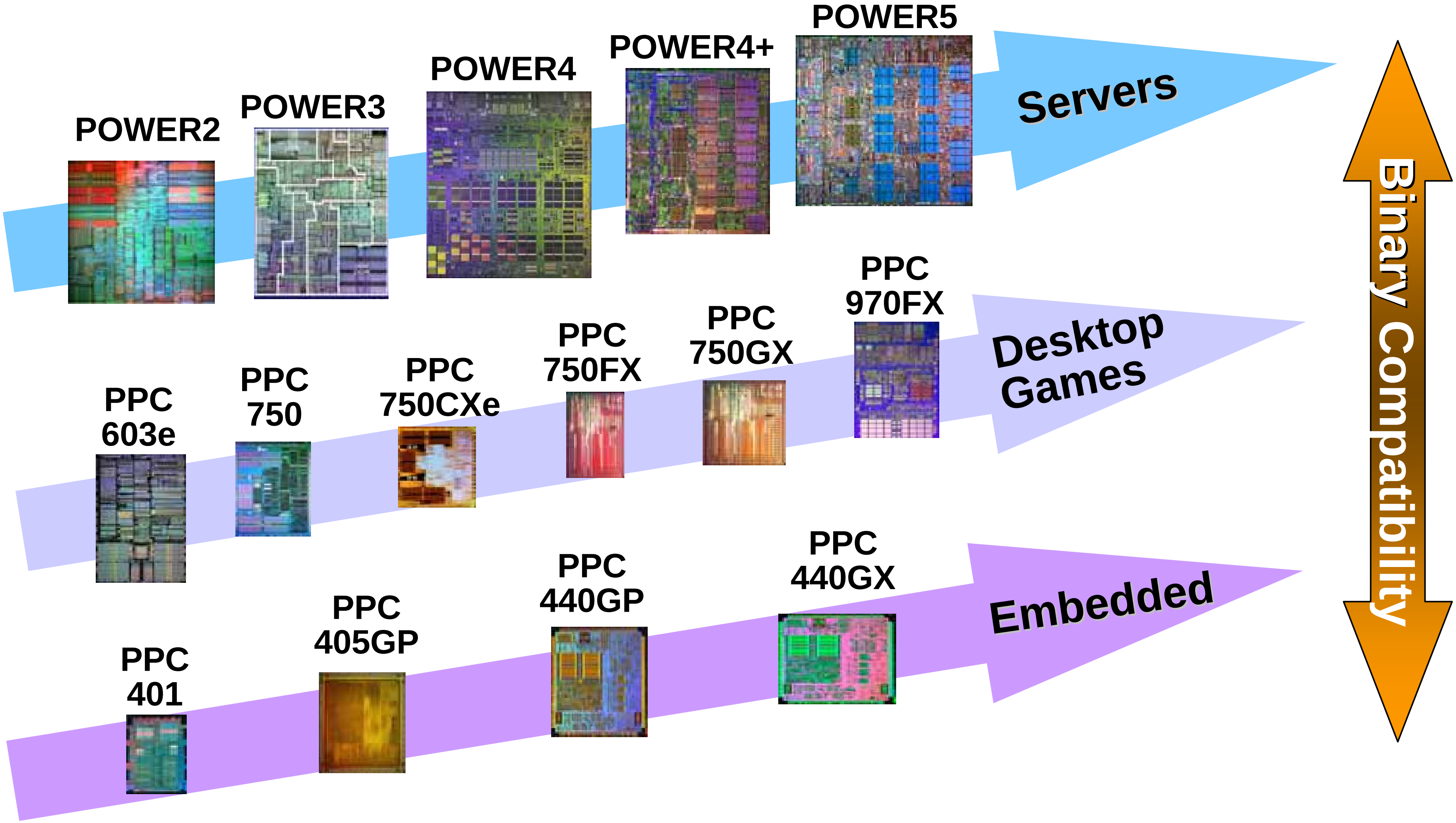


10 sec/step on 2048 BG/L
25 sec/step on 1400 Xeon Cluster at LLNL.

Top 500 List as of Nov 8, 2004 (top 10)

1.	IBM/DOE, USA	IBM BG/L DD2	32768 procs	70.7 TF/s
2.	NASA/Ames, USA	SGI Altix 1.5 GHz	10160 procs	51.8 TF/s
3.	Earth Simltr.,Japan	NEC	5120 procs	35.8 TF/s
4.	Barcelona SCC, Spain	IBM eServer JS20	3564 procs	20.5 TF/s
5.	LLNL , USA	INTEL Itanium 2	4096 procs	19.9 TF/s
6.	LANL, USA	Convex, ASCI Q	8192 procs	13.8 TF/s
7.	Virginia Tech, USA	Apple, X Server	2200 procs	12.2 TF/s
8.	IBM, USA	IBM BG/L DD1	8192 procs	11.6 TF/s
9.	NAVOCEANO, USA	IBM P655	2944 procs	10.3 TF/s
10.	NCSA, USA	DELL Xeon	2500 procs	9.8 TF/s

POWER : The Scaleable Architecture



POWER2

POWER3

POWER4

POWER4+

POWER5

Servers

PPC 603e

PPC 750

PPC 750CXe

PPC 750FX

PPC 750GX

PPC 970FX

Desktop Games

PPC 401

PPC 405GP

PPC 440GP

PPC 440GX

Embedded

Binary Compatibility