Power and Performance Optimization at the System Level

Valentina Salapura
IBM T.J. Watson Research Center

ACM Computing Frontiers 2005

The BlueGene/L Team

The Supercomputer Challenge

- **More Performance → More Power**
  - Systems limited by data center cooling capacity
    - New buildings for new supercomputers
  - FLOPS/W not improving from technology

→ **Traditional supercomputer design hitting power & cost limits**

- **Scaling single core performance degrades power-efficiency**

The BlueGene/L Concept

- **Parallelism can deliver higher aggregate performance**
  - Efficiency is key: (deliver performance / system power)
    - Power budget scales with peak performance
    - Application performance scales with sustained performance
- **Avoid scaling single core performance into regime with diminishing power/performance efficiency**
  - Deliver performance by exploiting application parallelism
- **Focus design effort on improving efficient MP scaling**
  - e.g., special purpose networks for synchronization and communication
- **Compute density can be achieved only with low power design approach**
  - Capacity of data center limited by cooling, not floor space

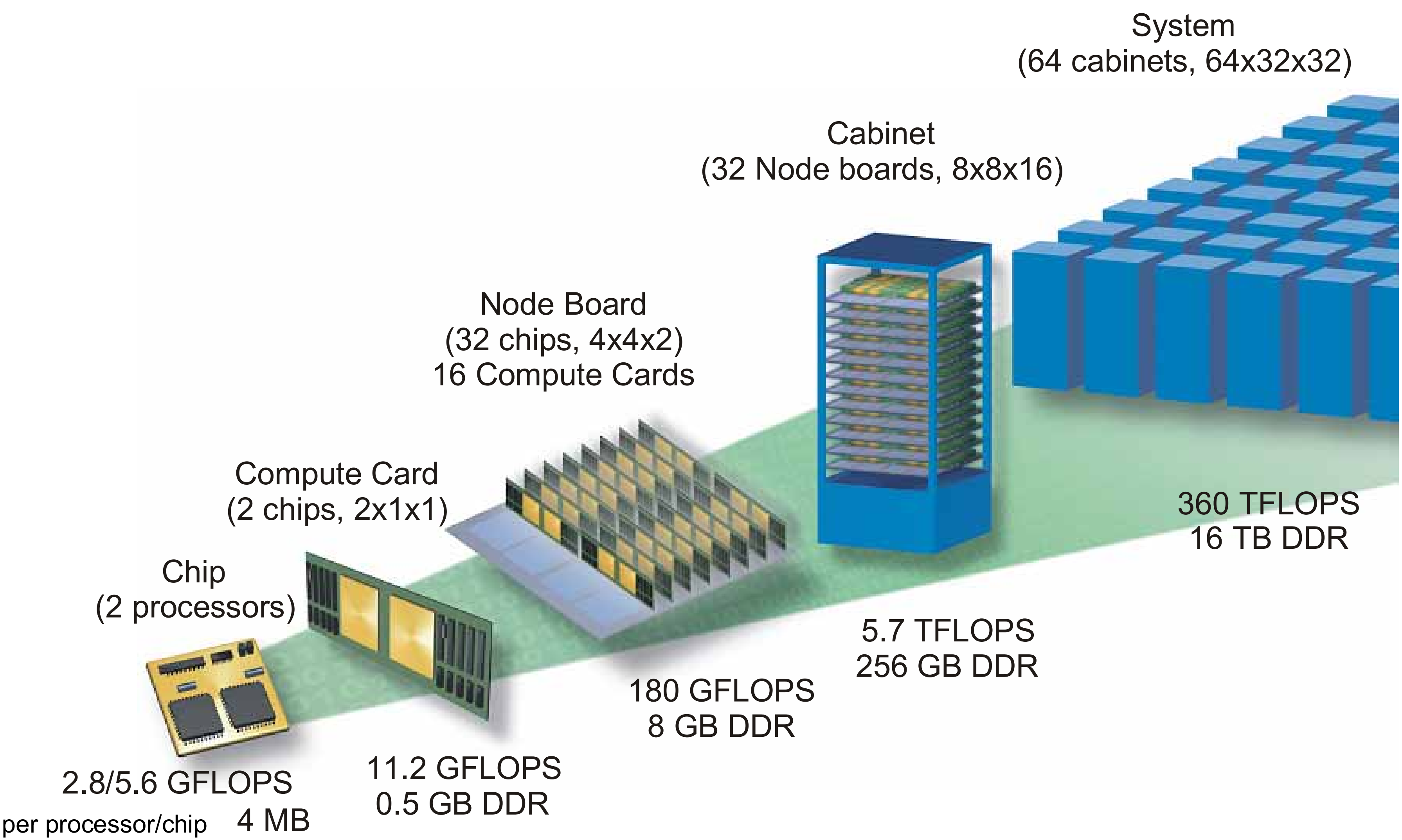
BlueGene/L Design Philosophy

- **Use standard embedded system-on-a-chip (SoC) design methodology**
- **Utilize PowerPC architecture and standard messaging interface (MPI)**
  - Standard programming model
  - Mature compiler support
- **Focus on low power**
  - Air cooling – power budget per rack 25 KW
- **Improve cost/performance (total cost/time to solution)**
  - Use & develop only two ASICs: node and link
  - Leverage industry-standard PowerPC design
- **Single-chip nodes, less complexity**
  - Enables high density

The BlueGene/L System

- **A 64k-node highly integrated supercomputer**
- **360 teraflops peak performance**
- **Strategic partnership with LLNL and high-performance computing centers**
  - Validate and optimize architecture using real applications
  - LLNL is accustomed to new architectures and experienced at application tuning to adapt to constraints
  - Help us investigate the reach of this machine
- **Focuses on numerically intensive scientific problems**
- **“Grand challenge” science projects**

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
0.5 GB DDR
180 GFLOPS
8 GB DDR
5.7 TFLOPS
256 GB DDR
360 TFLOPS
16 TB DDR

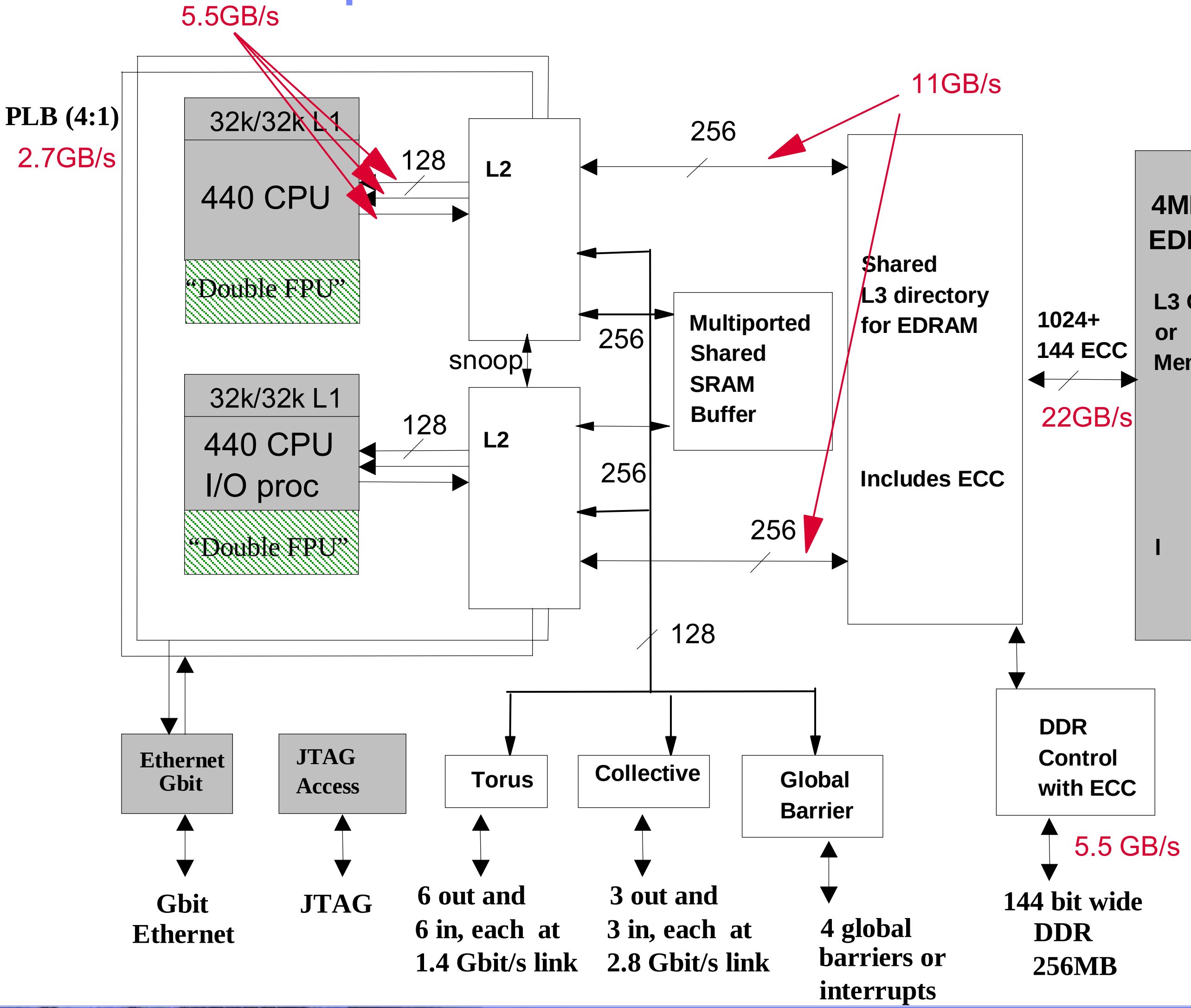
System Characteristics

- **Chip multiprocessor**
  – 2 PowerPC cores per chip
- **Data parallelism**
  – Double floating point unit for advanced SIMD operations
- **High integration**
  – 2 PowerPC cores + EDRAM cache + DDR memory interface + network interfaces on a single chip
- **High performance networks**
  – Directly on chip ➔ reduce latency
  – Multiple optimized, task-specific networks
    - Synchronization, data exchange, I/O

BlueGene/L Architecture

ACM Computing Frontiers 2005

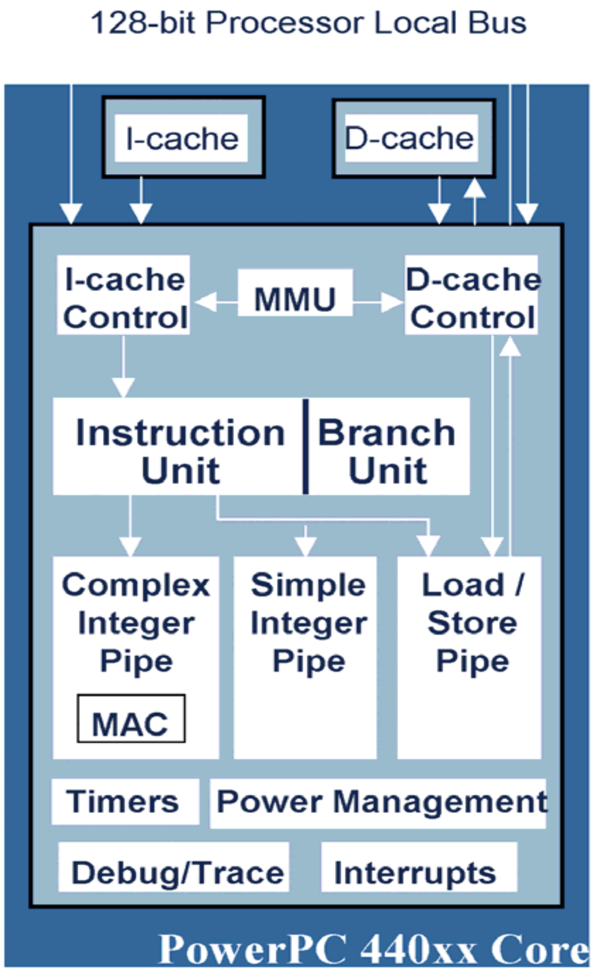
BlueGene/L Compute SoC ASIC



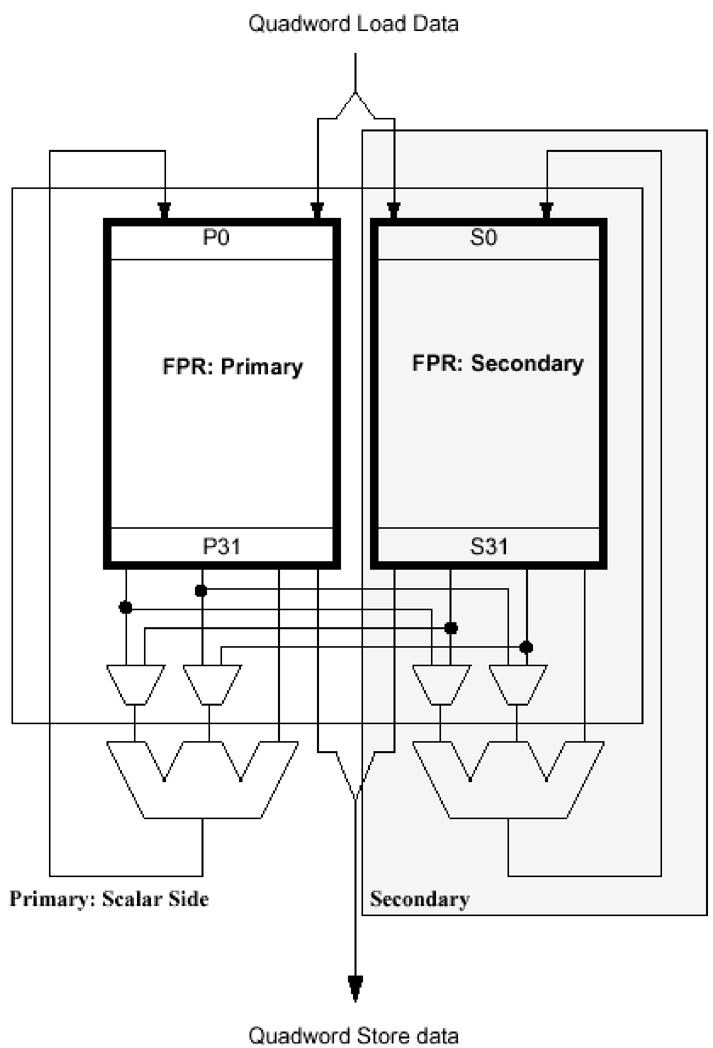
5.6GF
peak
node

PowerPC 440 Processor Core Features

- High performance embedded PowerPC core
- 2.0 DMIPS/MHz
- Book E Architecture
- Superscalar: two instructions per cycle
- Out of order issue, execution, and completion
- 7 stage pipeline
- 3 Execution pipelines
- 32 32 bit GPR
- Dynamic branch prediction
  - BHT & BTAC
- Caches
  - 32KB instruction & 32KB data cache
  - 64-way set associative, 32 byte line
- 36-bit phisical address
- 128-bit CoreConnect PLB Interface



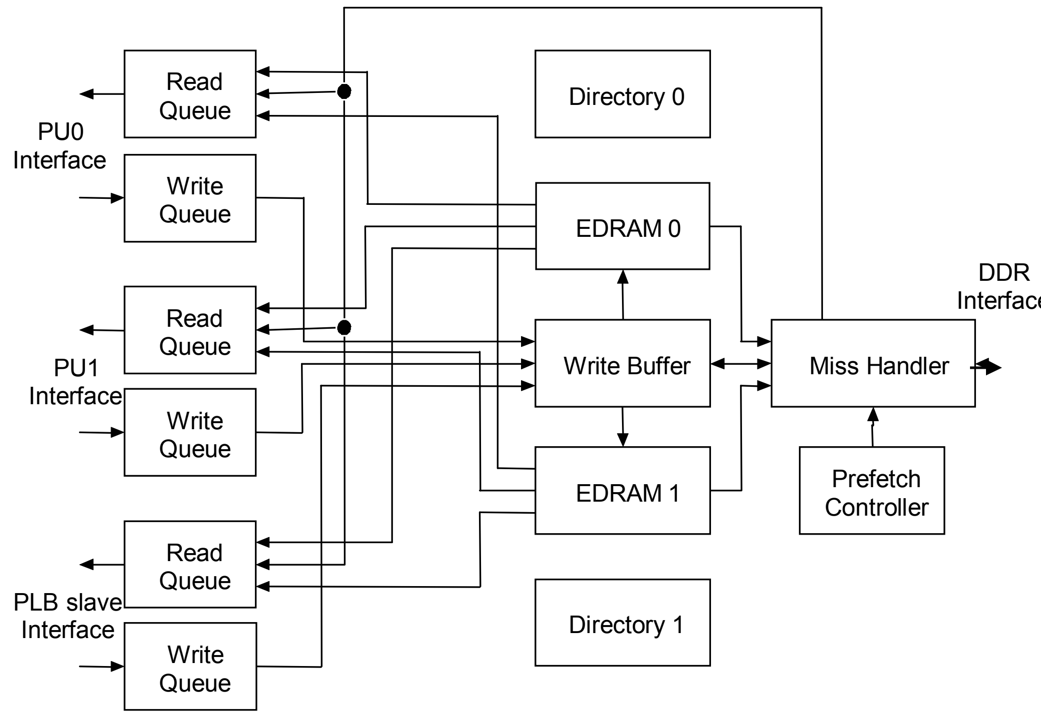
Double Hummer Floating-Point Unit



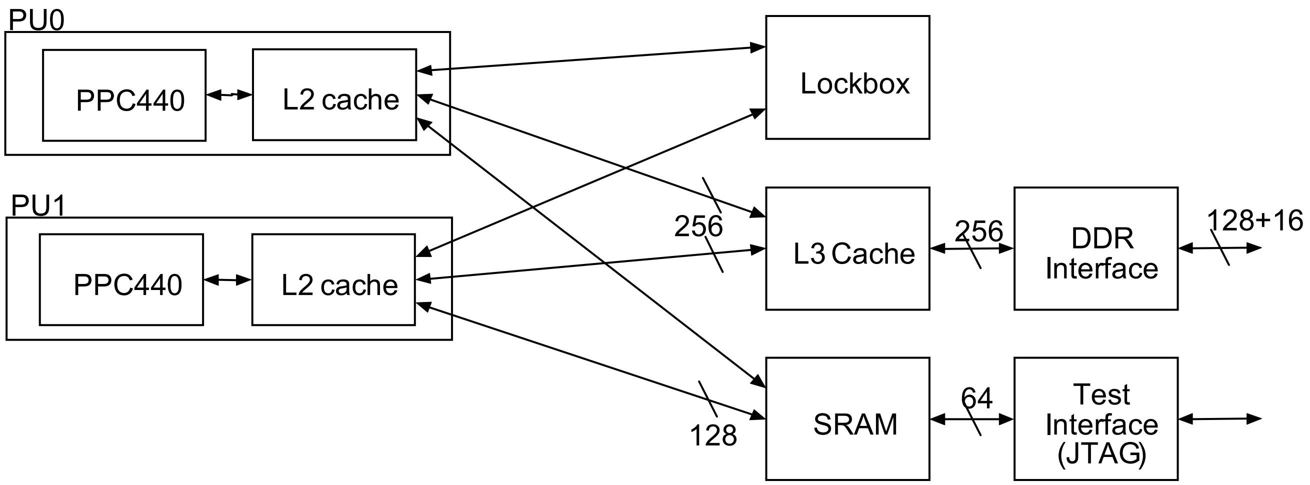
- **Two replicas of a standard single-pipe PowerPC FPU**
  - 2 x 32 64-bit registers
- **Enhanced ISA, includes instructions**
  - Executed in either pipe
  - Simultaneously execute the same operation on both sides – SIMD instructions
  - Simultaneously execute two different operations of limited types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

L3 Cache Implementation

- **On-chip 4 MB L3 cache**
- **Use EDRAM**
- **Two-way interleaved**
- **2MB EDRAM per bank, 8-way set-associative, 128-byte lines**
- **ECC protected**
- **32-byte read and write bus per core @ 350MHz**
- **2 x 64-byte EDRAM access @ 175MHz**



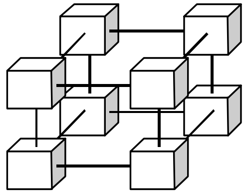
Memory Hierarchy



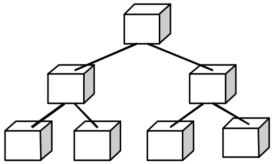
Memory Type	Latency (cycles)
L1 cache	3
L2 cache	11
L3 cache	28/36/40
Main memory	86

- **32kB D&I private cache per processor**
- **Small private L2 data prefetch caches**
  - Supports 7 streams/processor
- **On-chip 4MB L3 cache**
- **Access to main memory via L3 cache**
- **SRAM for fast exchange of control information**
- **Synchronization via lockbox semaphores**

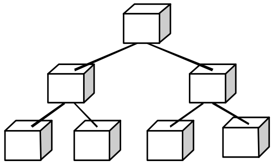
BlueGene/L Five Independent Networks



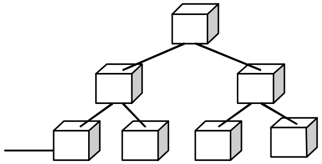
3 Dimensional Torus
- Point-to-point



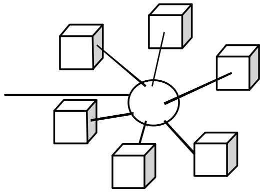
Collective Network
- Global Operations



Global Barriers and Interrupts
- Low Latency Barriers and Interrupts

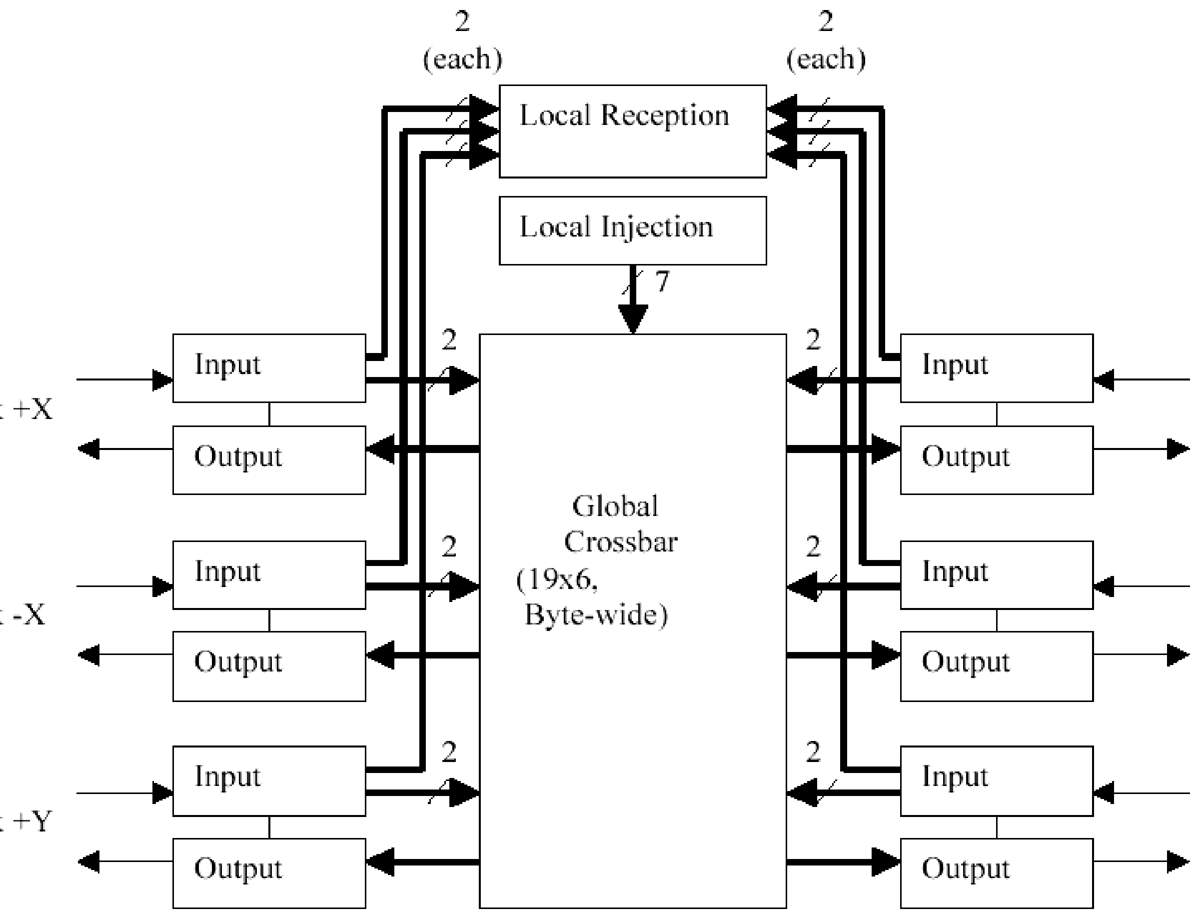


Gbit Ethernet
- File I/O and Host Interface



Control Network
- Boot, Monitoring and Diagnostics

Three-Dimensional Torus Network



Torus router

Point-to-point communication

- Nearest neighbor interconnect

Links 1 bit wide, 6 bidirectional links/chip

Per-link bandwidth 1.4Gb/s

Per-node bandwidth 2.1GB/s

Cut-through routing without software intervention

Adaptive routing

Packet length

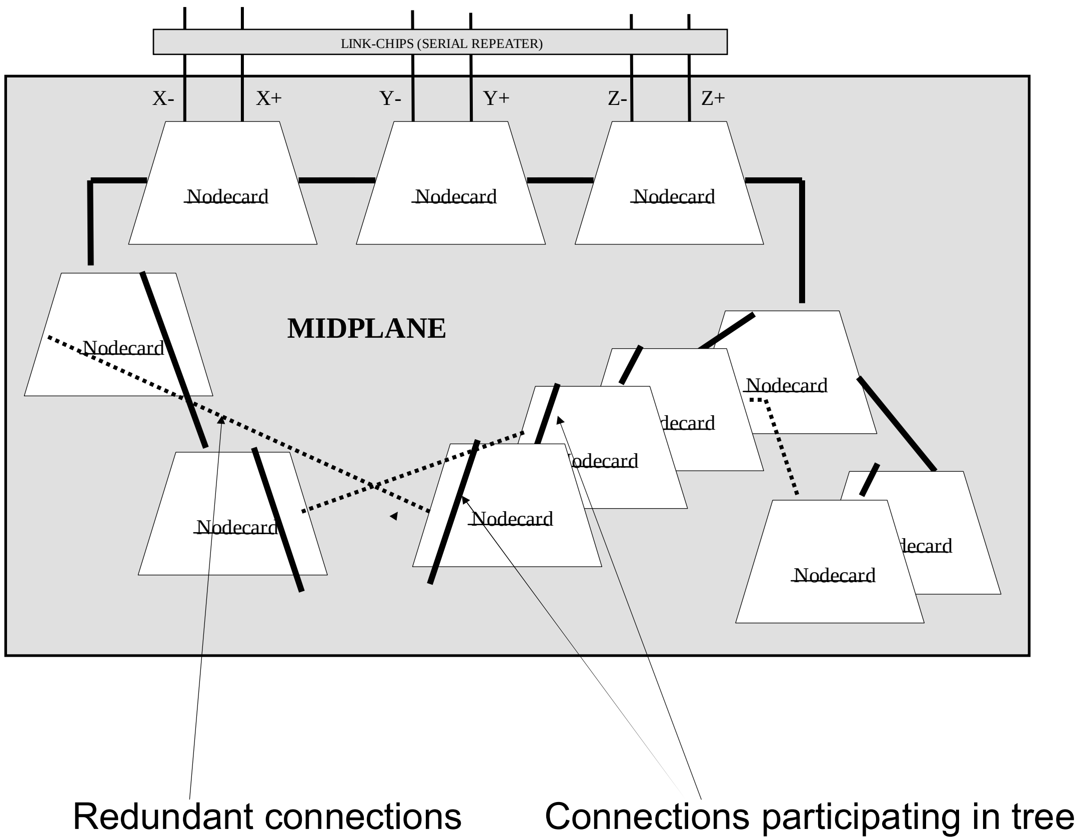
- 32–256 bytes, 4-byte trailer

Per-hop latency ~100 ns (avg.)

Worst case latency for 64k machine (64 x 32 x 32)

- 6.4 μs (64 hops)

Collective Network

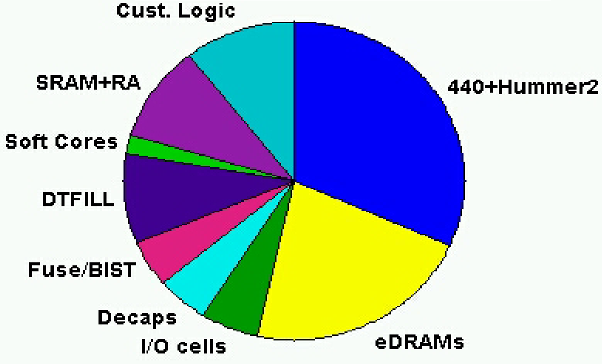


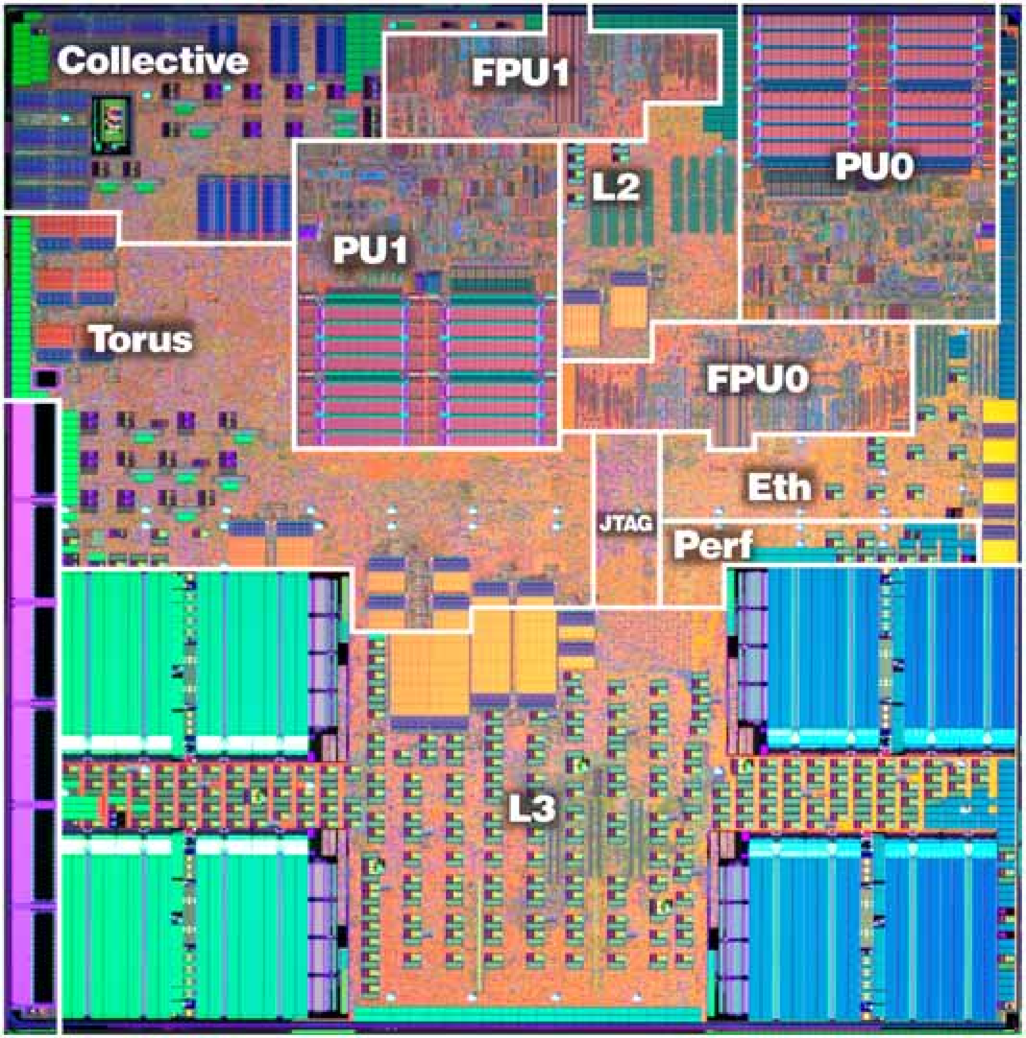
- **Global reduction support**
- **Bidirectional 3 links per node**
- **Per node bandwidth 2.1 GB/s**
- **Worst case latency (round trip) 5.0μs**
- **Efficient for collective communication**
  - For broadcast messages
  - Arithmetic reductions implemented in hardware
- **Fault-tolerant for tree topologies**
- **Connect every node to I/O node for file system**

BlueGene/L Chip Design Characteristics

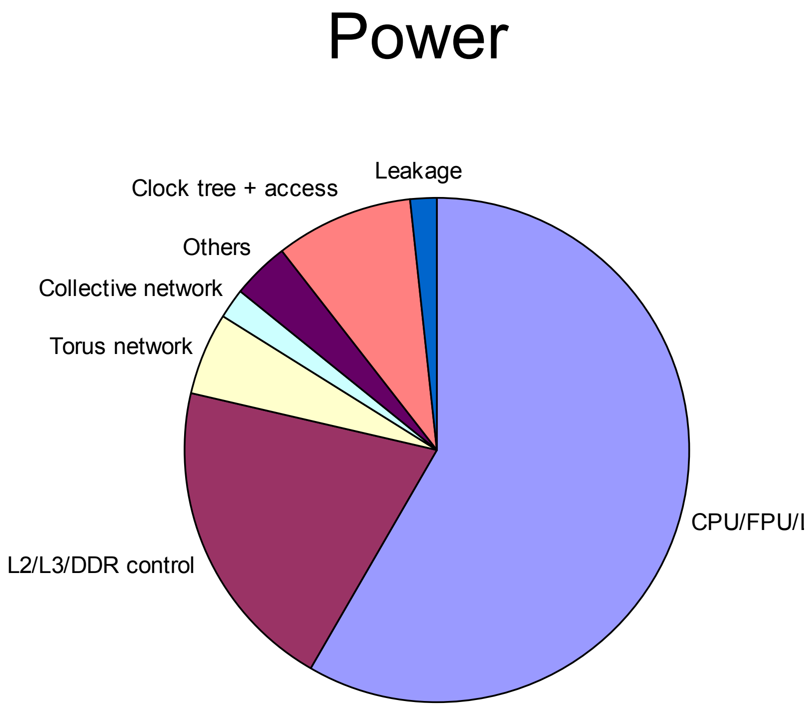
- **IBM Cu-11 0.13µ CMOS ASIC process technology**
- **11 x 11 mm die size**
- **95M transistors**
- **1.5/2.5V**
- **12.9W**
- **CBGA package, 474 pins**

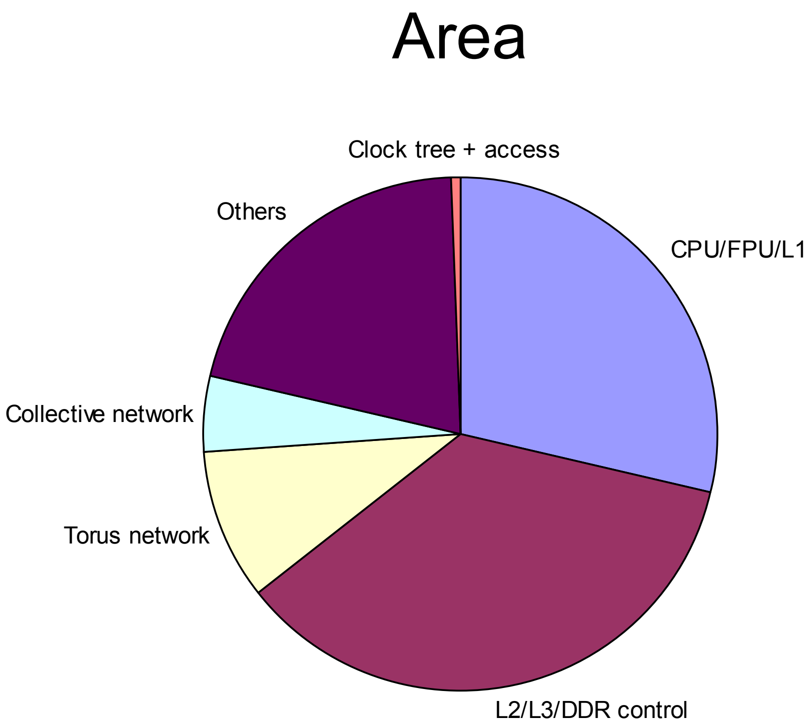
Chip Area usage





BlueGene/L Compute Chip Power and Area





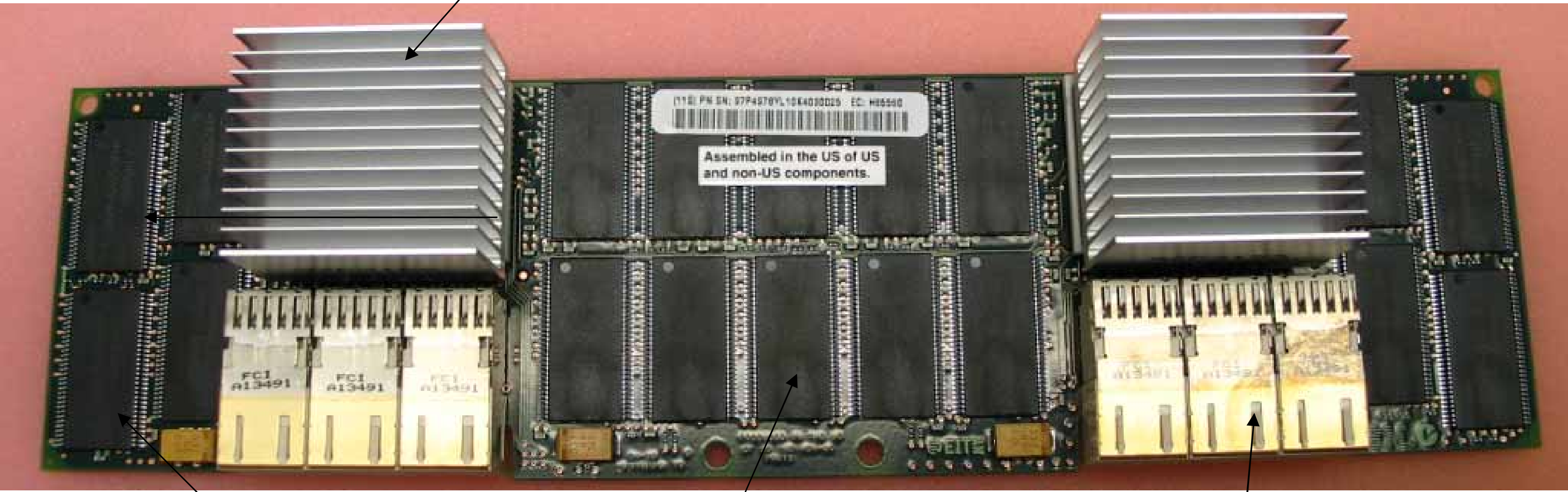
BlueGene/L System Package

ACM Computing Frontiers 2005

Dual Node Compute Card

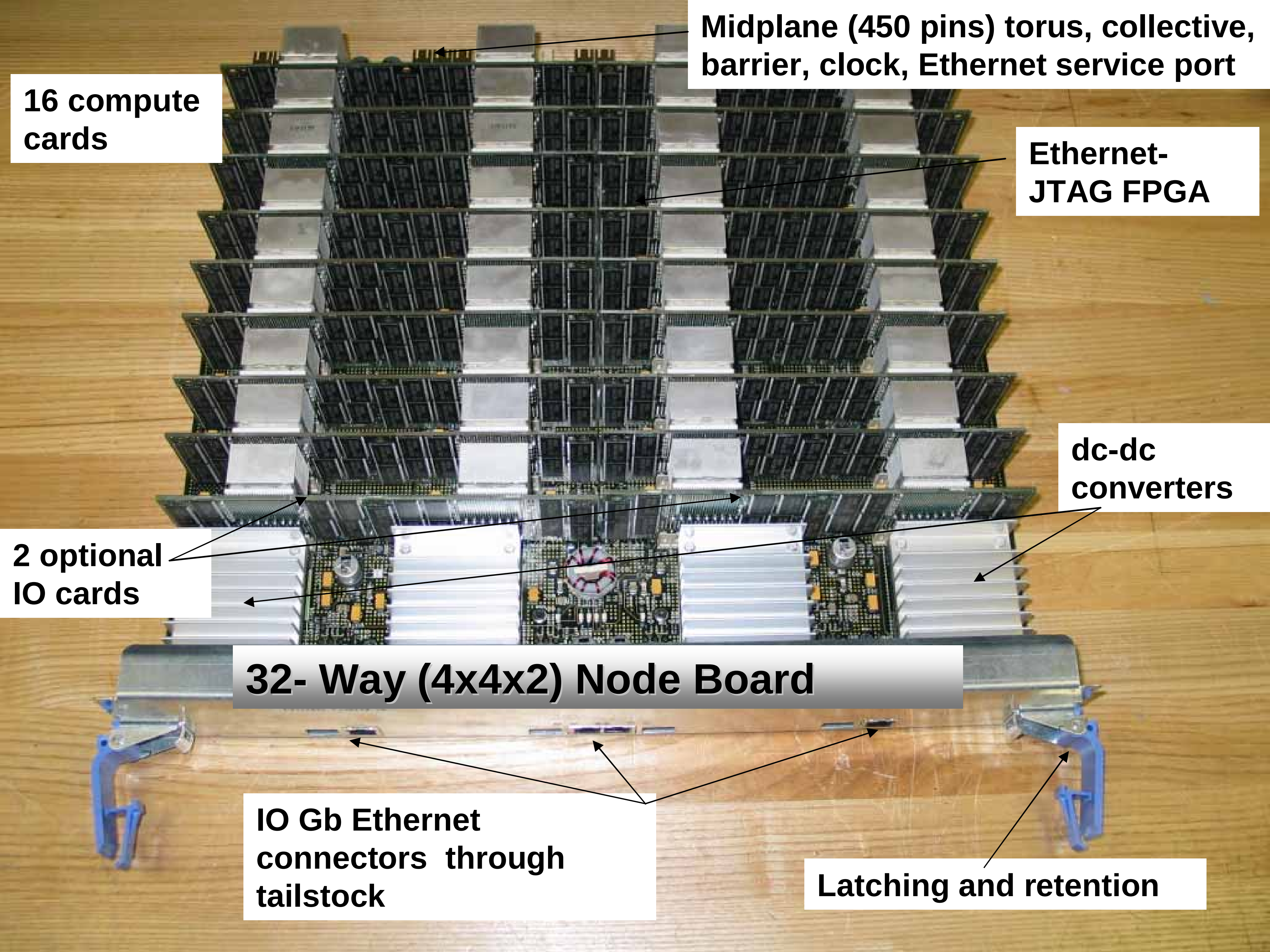
206 mm (8.125”) wide, 54mm high (2.125”), 14 layers, single sided, ground referenced

Heatsinks designed for 15W



9 x 512 Mb DRAM

Metral 4000 high speed differential connector (180 pins)

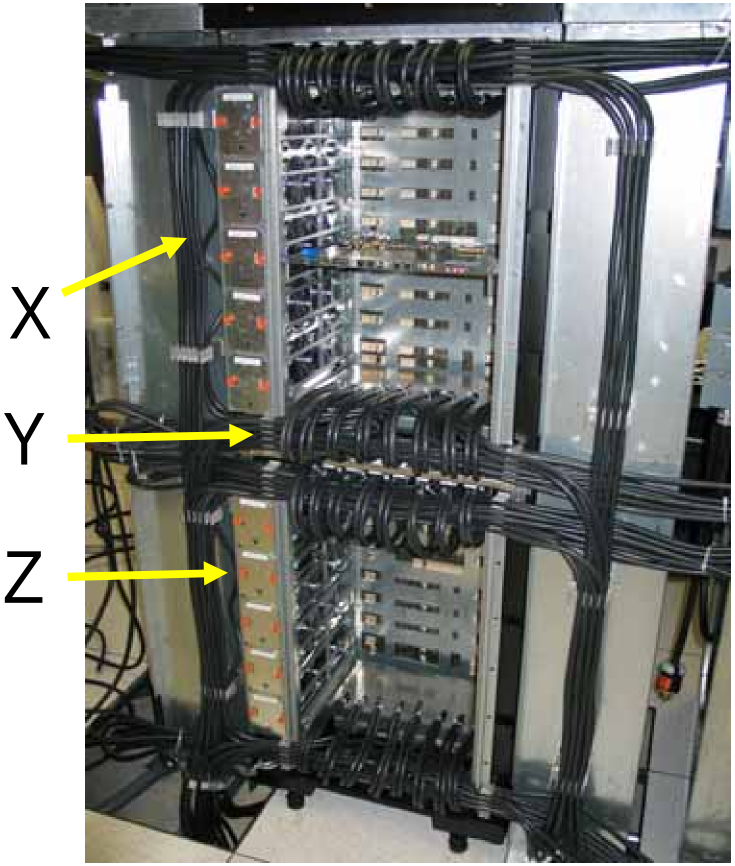

Midplane (450 pins) torus, collective, barrier, clock, Ethernet service port
16 compute cards
Ethernet-JTAG FPGA
dc-dc converters
2 optional IO cards
32- Way (4x4x2) Node Board
IO Gb Ethernet connectors through tailstock
Latching and retention

BlueGene/L Rack

512 – way (8 x 8 x 8) “midplane” (half-cabinet)

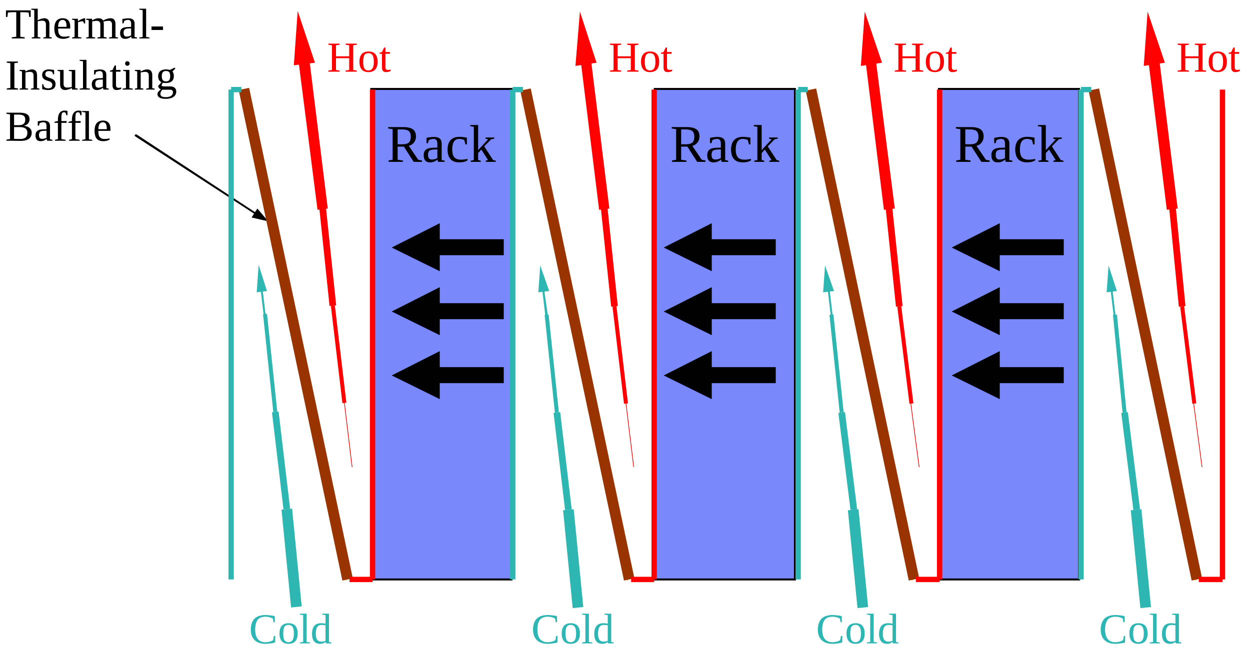
16 node boards

All wiring up to this level (>90%) card-level



Two midplanes interconnected with data cables

BlueGene/L Rack Ducting Scheme

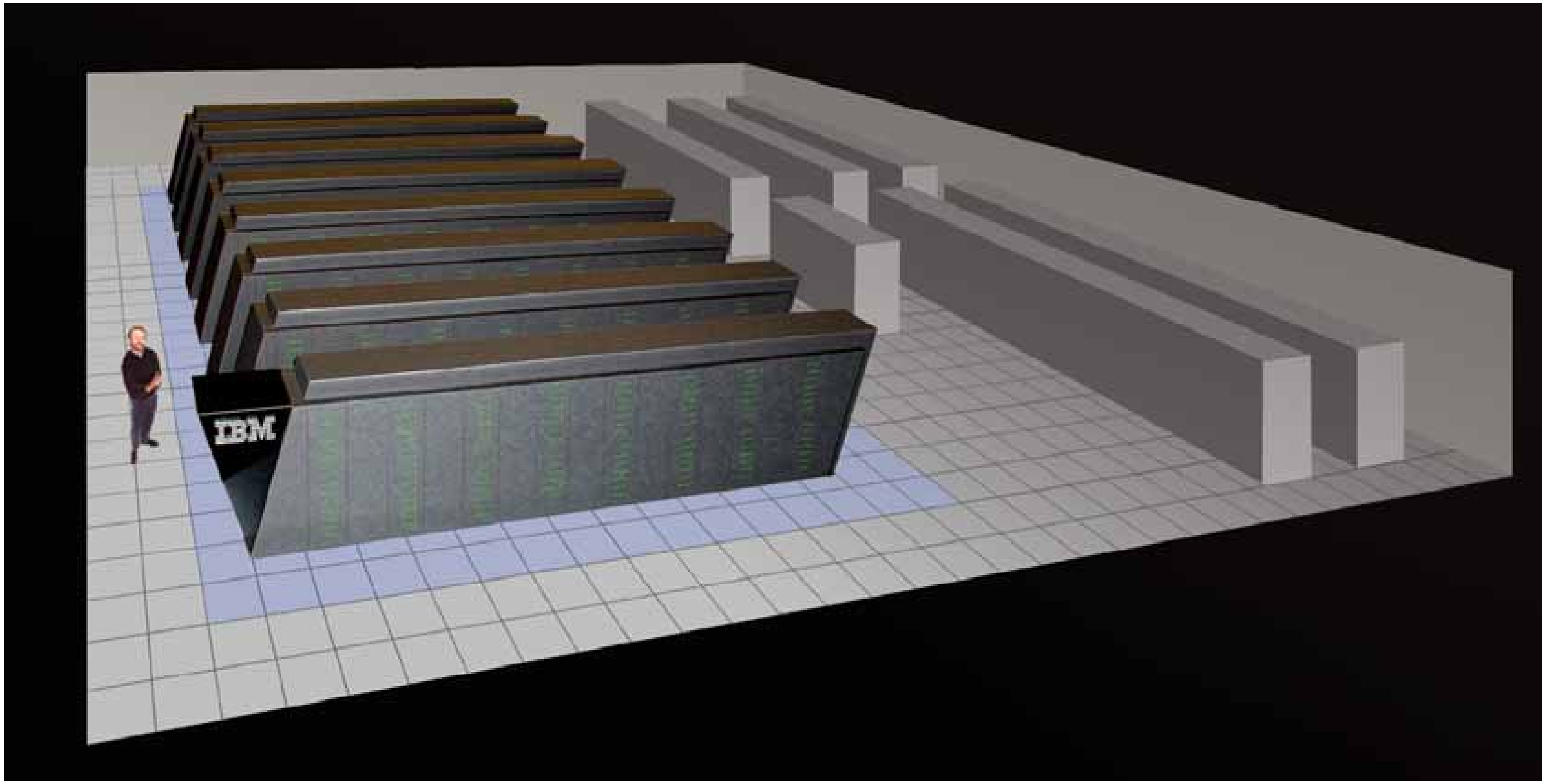


Airflow direct from raised floor

BlueGene/L 16-Rack System at IBM Rochester

16384 + 256 BLC chips. About 400 kW

BlueGene/L System



BlueGene/L System Software

ACM Computing Frontiers 2005

BlueGene/L – Familiar Software Environment

- **Fortran, C, C++ with MPI**
  - Full language support
  - Automatic SIMD FPU exploitation
- **Linux development environment**
  - Cross-compilers and other cross-tools execute on Linux front-end nodes
  - Users interact with system from front-end nodes
- **Tools support**
  - debuggers, hardware performance monitors, trace based visualization

BlueGene/L – Familiar Software Environment

- **Programmer's view: Nearly identical software stack/interface to pSeries**
  - Compilers: IBM XLF, XLC, VA C++, hosted on PPC/Linux
  - Operating System: Linux-compatible kernel with some restrictions
  - 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
  - Managed by a novel control system

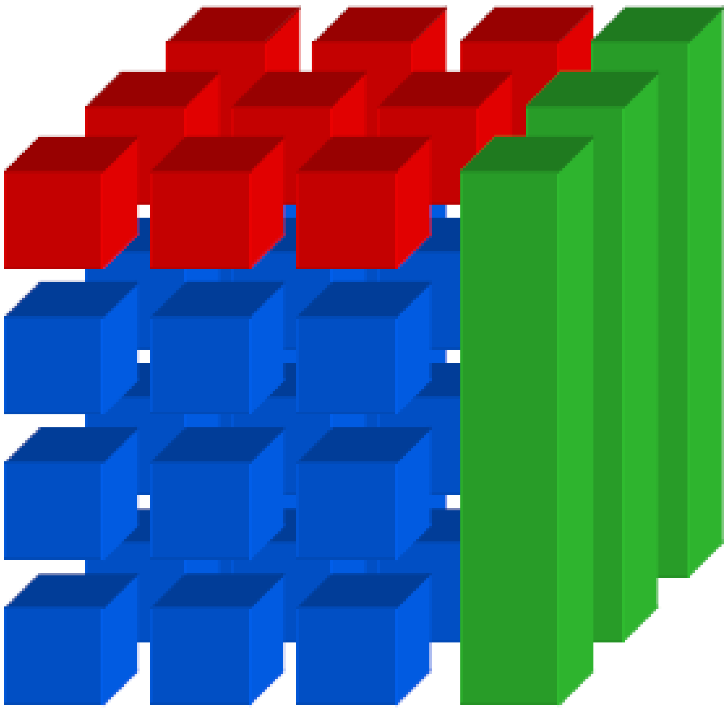
BlueGene/L System Software Main Characteristics

- **Logical partitioning (LPAR) of the system for multiple concurrent users**
  - Link chip partitions the system into logically separate systems
- **Strictly space sharing**
  - One parallel job (user) per partition of machine
  - One process per processor of compute node
- **Intra-chip communication**
  - MPI message passing programming model
- **Modes of operation**
  - Co-processor mode
    - Compute processor + communication off-load engine
  - Virtual node mode
    - Symmetric processors

BlueGene/L Operating Environment

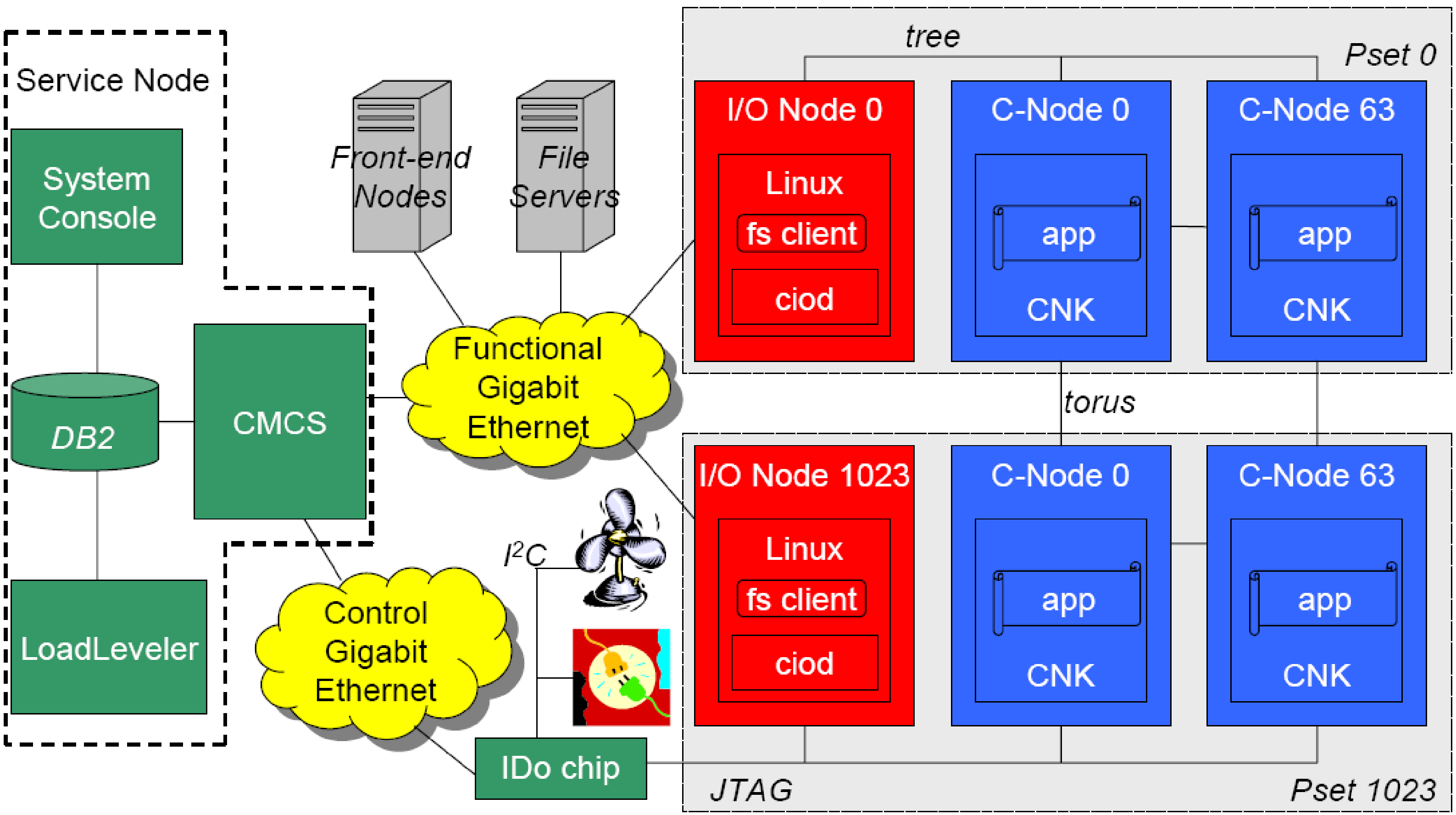
- **blrts operating system**
  - Linux compatible minimalist kernel
  - Single user single program operation
    - Minimal operating system interference
- **Virtual memory constrained to physical memory size**
  - Implies no demand paging
- **Torus memory mapped in the user address space**
  - no operating system calls needed for application communication

BlueGene/L System Software Architecture



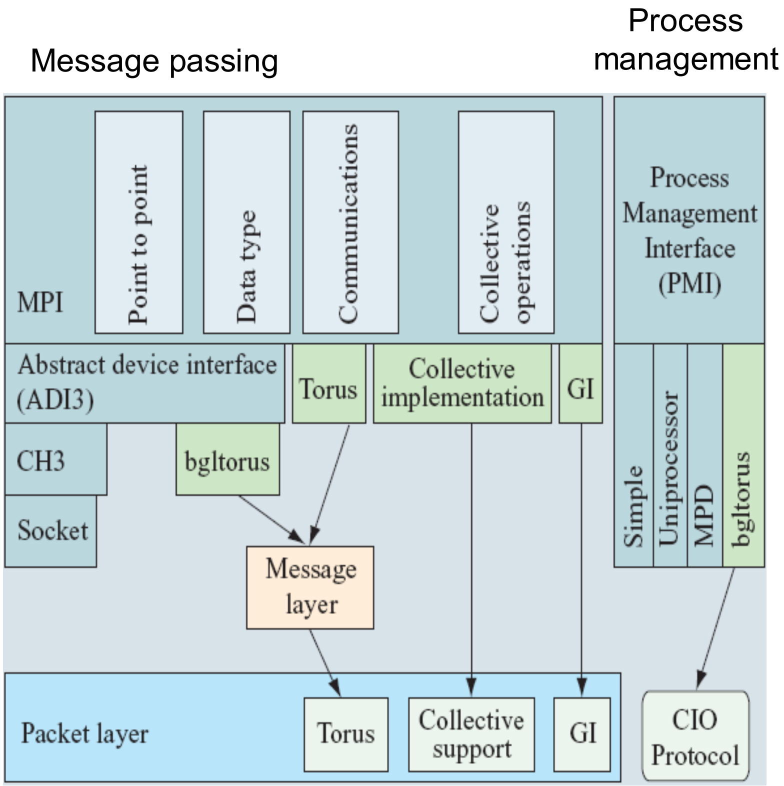
- **Compute nodes for user applications**
  - Simple Compute Node Kernel
  - Connected by 3D torus and collective network
- **I/O nodes for interaction with the outside world**
  - Run Linux
  - Provide OS services – file access, process launch/termination, debugging
  - Tree network and Gigabit Ethernet
- **Service nodes for machine monitoring and control**
  - Linux cluster
  - Custom components for booting, partitioning, configuration

Blue Gene/L System Architecture



MPI

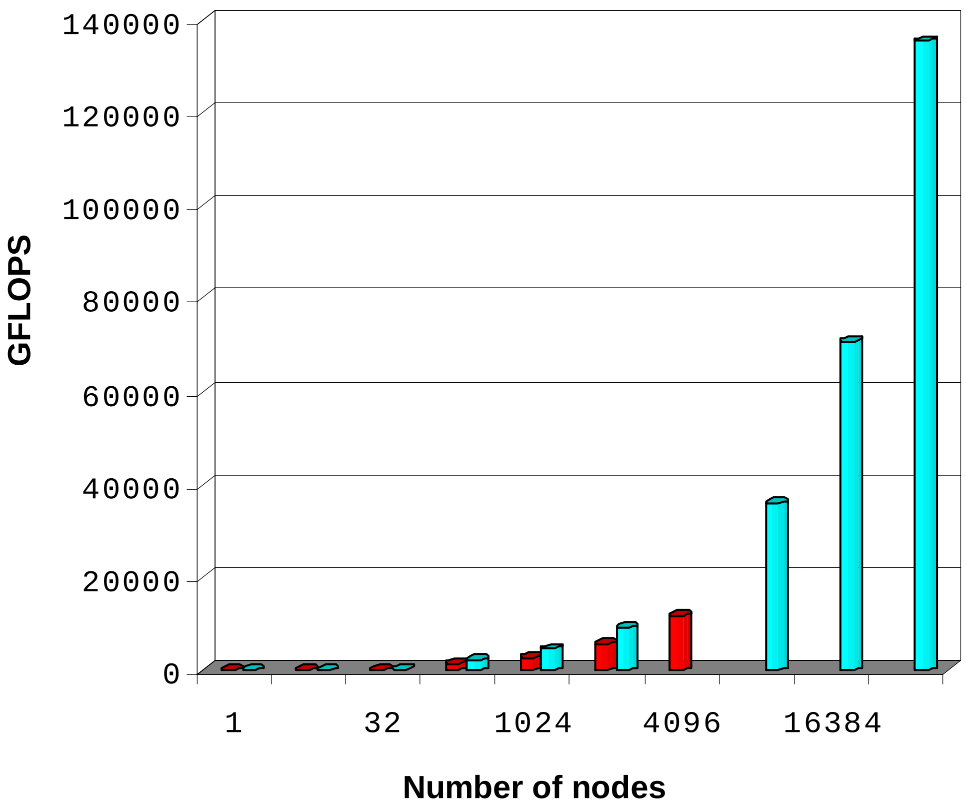
- **MPI 1.1 compatible implementation for message passing between compute nodes**
  - Only the most widely used features of MPI implemented
- **Based on MPICH2 from ANL**
- **Point-to-point**
  - Utilizes Torus
  - Implements a BlueGene/L version ADI3 on top of message layer
- **Global operations**
  - Utilizes both torus and collective network
- **Process management**
  - Use BlueGene/L's control system rather than MPICH's process managers



BlueGene/L Application Performance and Power Analysis

ACM Computing Frontiers 2005

LINPACK Performance



DD1 hardware @ 500 MHz

#4 on June 2004 TOP500 list

11.68 TFLOP/s on 4K nodes

DD2 hardware @ 700 MHz

#1 on Nov 2004 TOP500 list

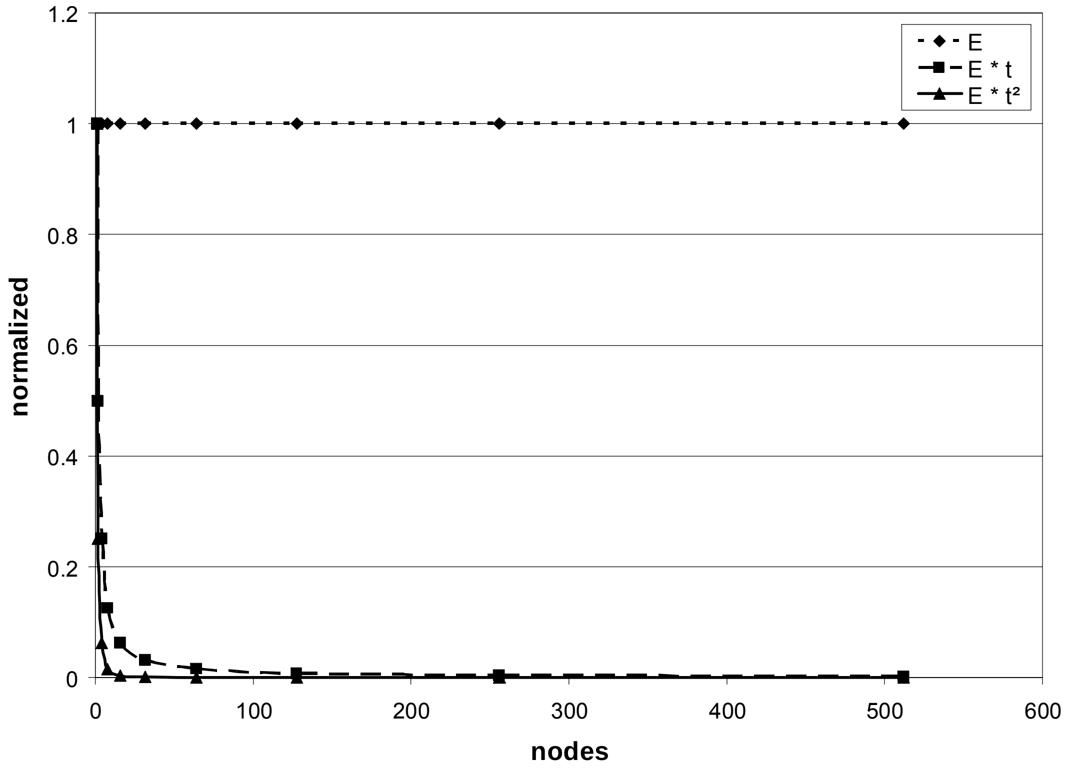
70.72 TFLOP/s on 16K nodes

77% of peak

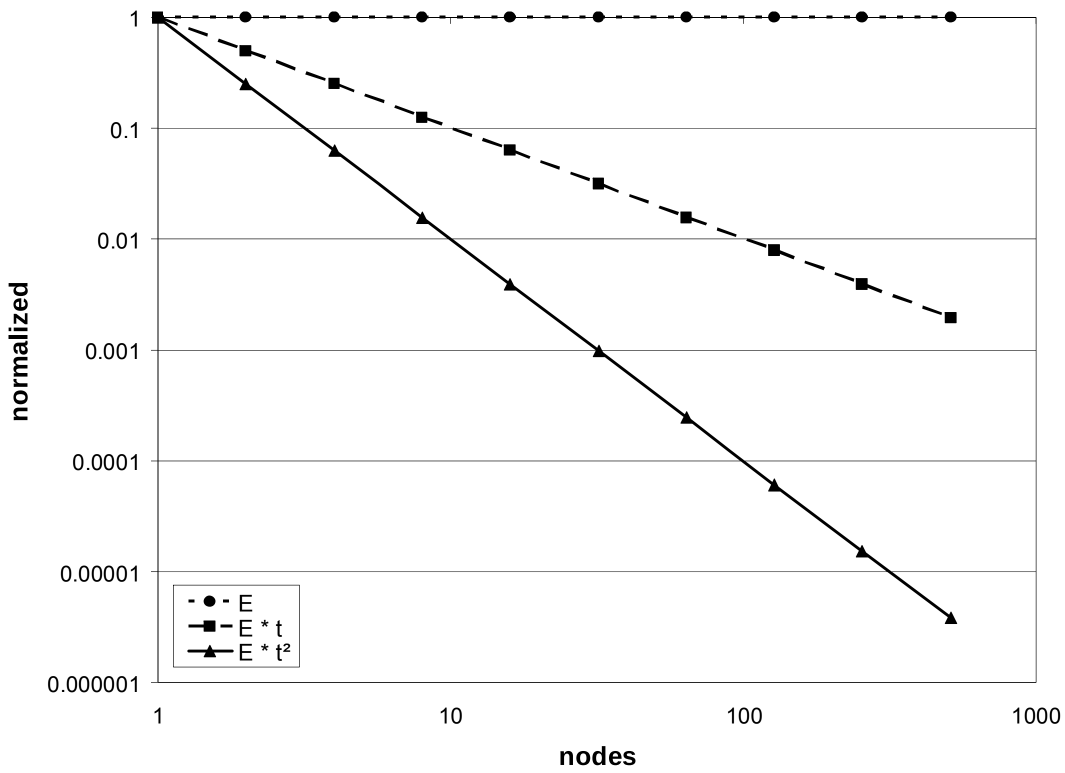
Application Performance and Power Efficiency

- **Figures of merit**
  - t -- time (delay)
    - application execution time
  - E -- energy (W/MIPS)
    - energy dissipated to execute application
  - E * t -- energy-delay [Gonzalez Horowitz 1996]
    - energy and delay are equally weighted
  - E * t^2 -- energy-delay squared [Martin *et al.* 2001]
    - metric invariant on the assumption of voltage scaling

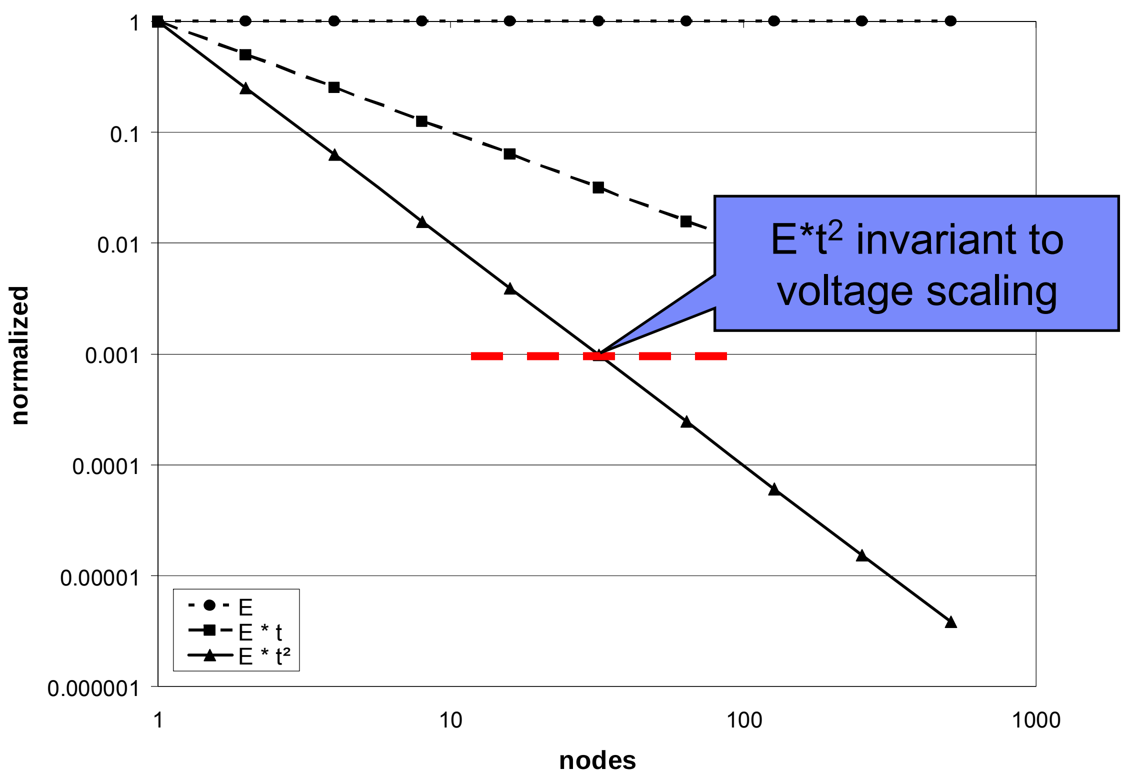
Low Power - High Performance System Concept



Low Power - High Performance System Concept (log-log)



Low Power - High Performance System Concept (log-log)

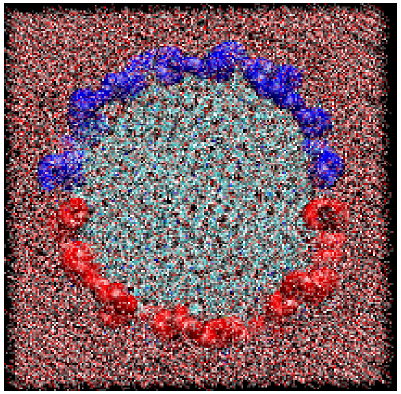


Applying Metrics to Actual Applications

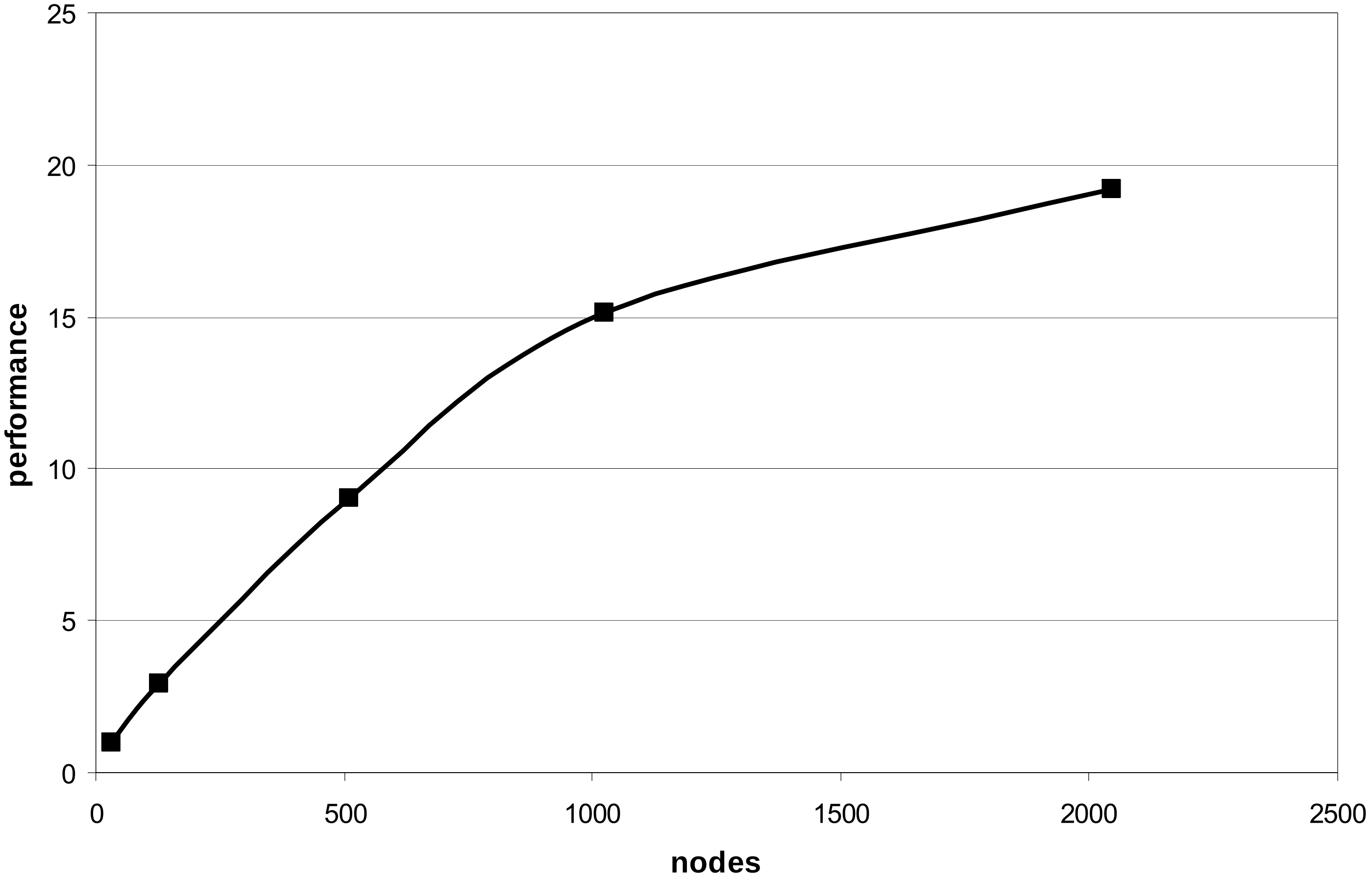
- **LINPACK highly parallel – follows 77% of peak performance**
  - Problem size matches the size of the system
  - Weak scaling
- **Many applications require constant amount of computation regardless of the size of the system**
  - Fixed sized problems
  - Strong scaling
  - More conservative performance evaluation
- **Apply metrics for several applications and problems**
  - e.g., NAMD, UMT2K

NAMD

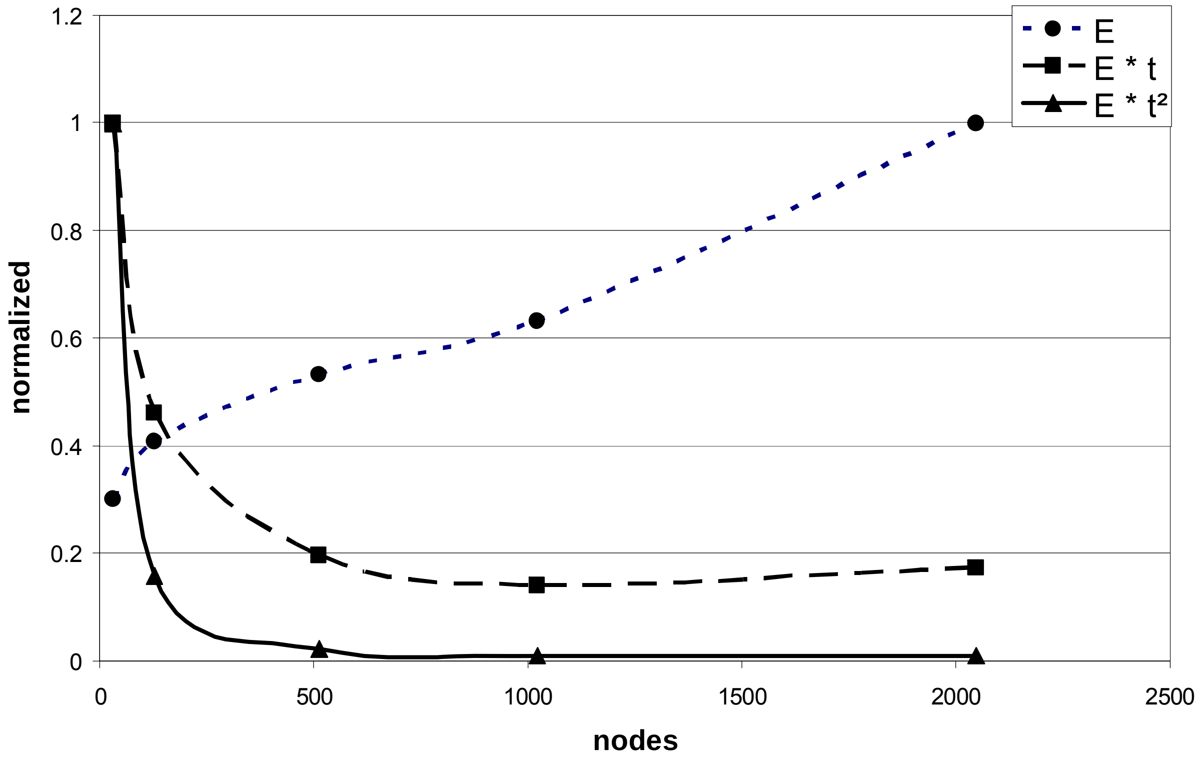
- **Parallel, object-oriented molecular dynamics code designed for high-performance simulation of large biomolecular systems**
  - Developed by the Theoretical Biophysics Group in the Beckman Institute for Advanced Science and Technology at the University of Illinois at Urbana-Champaign
- **Distributed free of charge with source code**
  - Based on Charm++ parallel objects
- **NAMD benchmark**
  - Box with one molecule of apoprotein A1 solvated in water
- **Fixed size problem on 92,224 atoms**



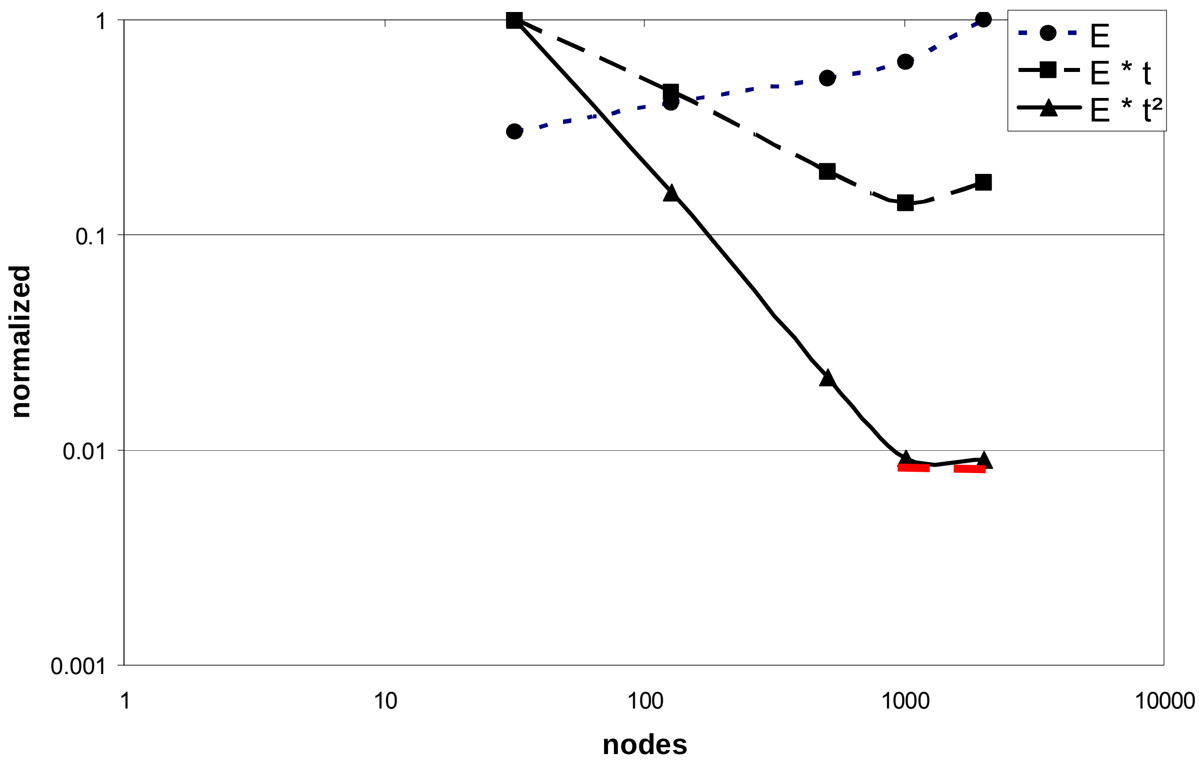
NAMD Performance Scaling



NAMD Power and Power Performance

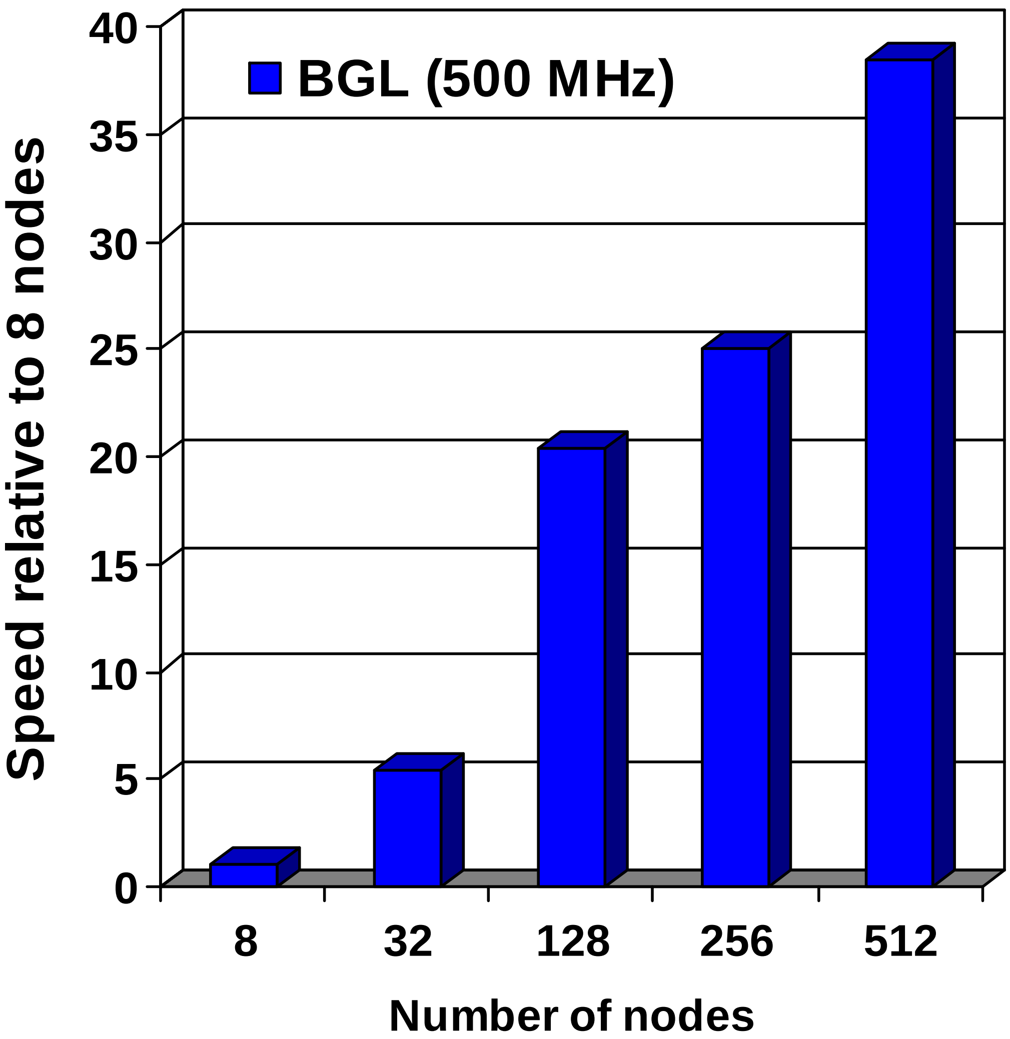


NAMD Power and Power Performance on log-log

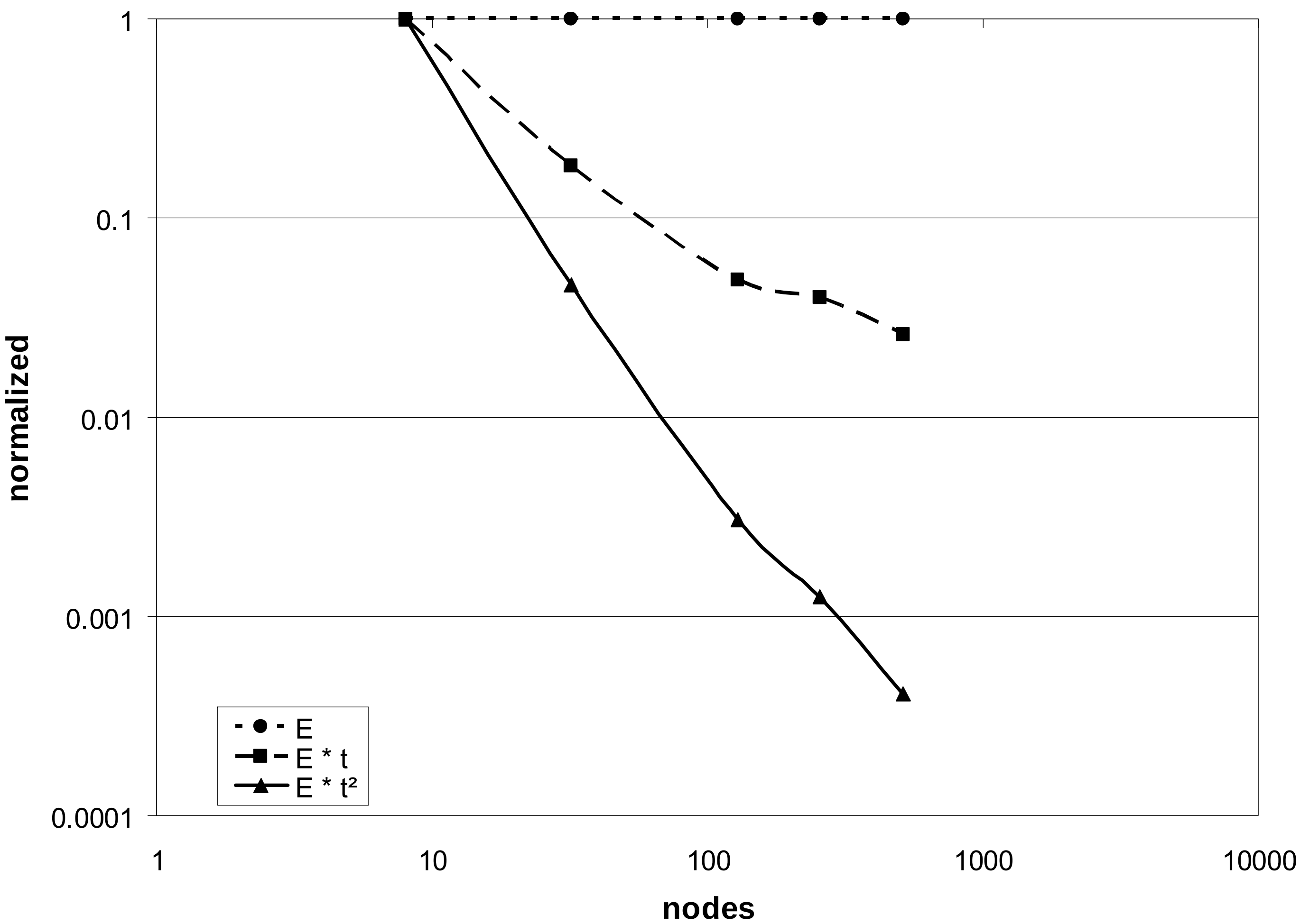


ASCI Purple Benchmarks – UMT2K

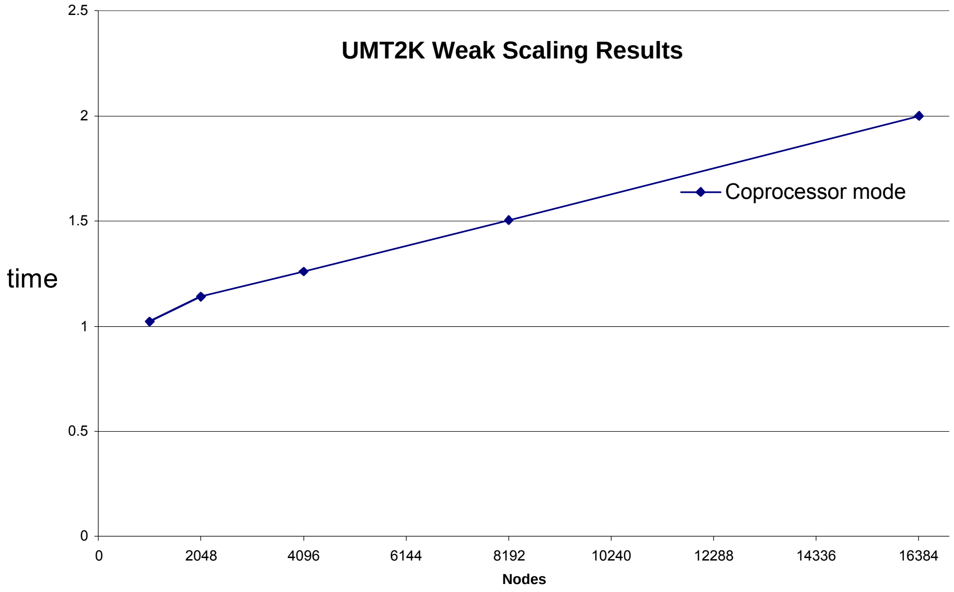
- **UMT2K – Unstructured mesh radiation transport**
- **Problem size fixed**
- **Excellent scalability up to mid-sized configurations**
  - Load balancing problems when scaling to 2000 or more nodes
  - Needed algorithmic changes in original program
  - Tuned UMT2K version scales well beyond 8000 BlueGene/L nodes

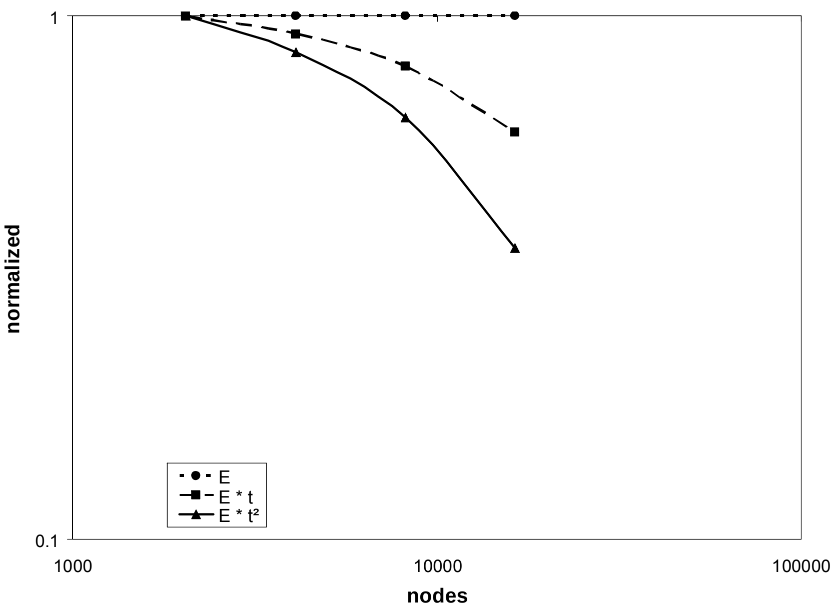


UMT2K Power and Power Performance



Recent UMT2K Runs Demonstrate Good Performance



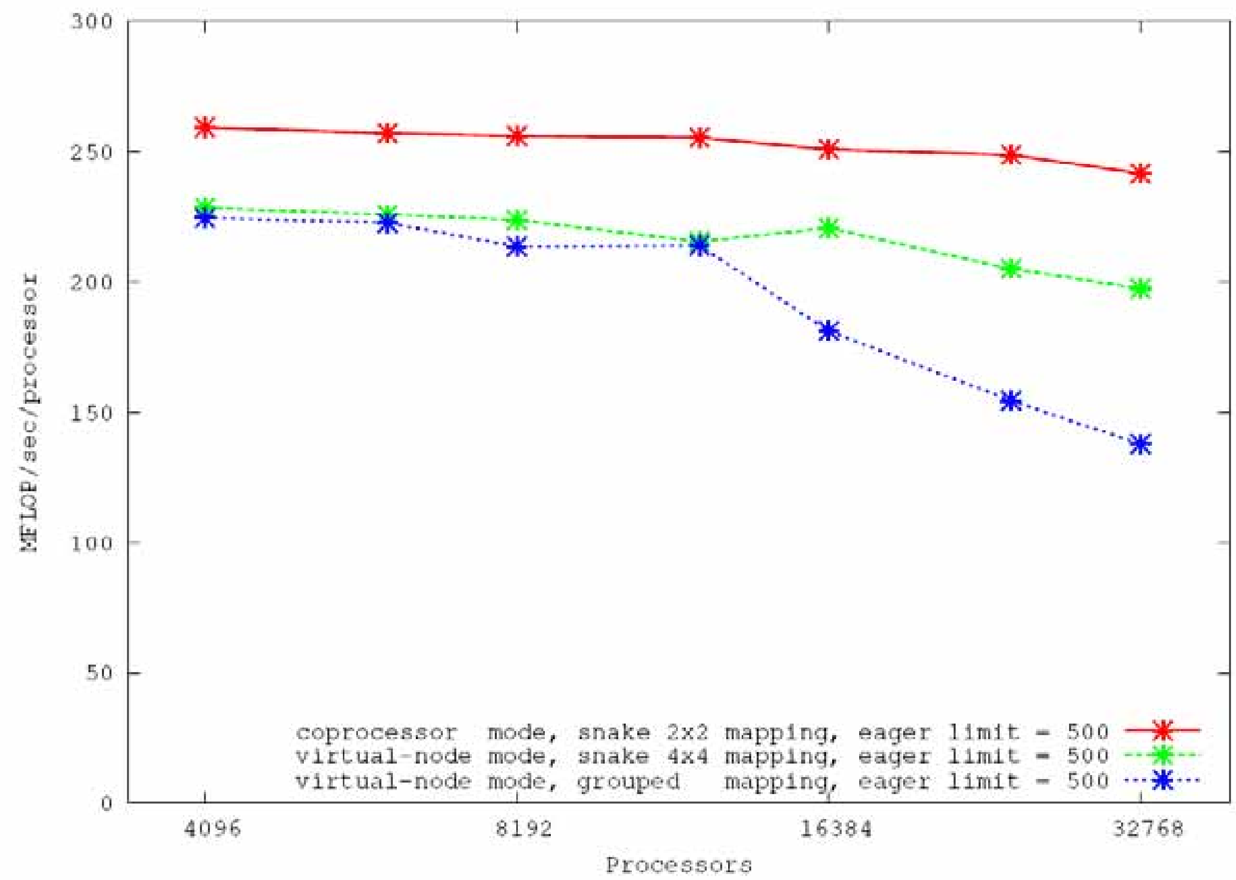


Nearest neighbor communication consistent
Load balance unchanged from 1K to 8K

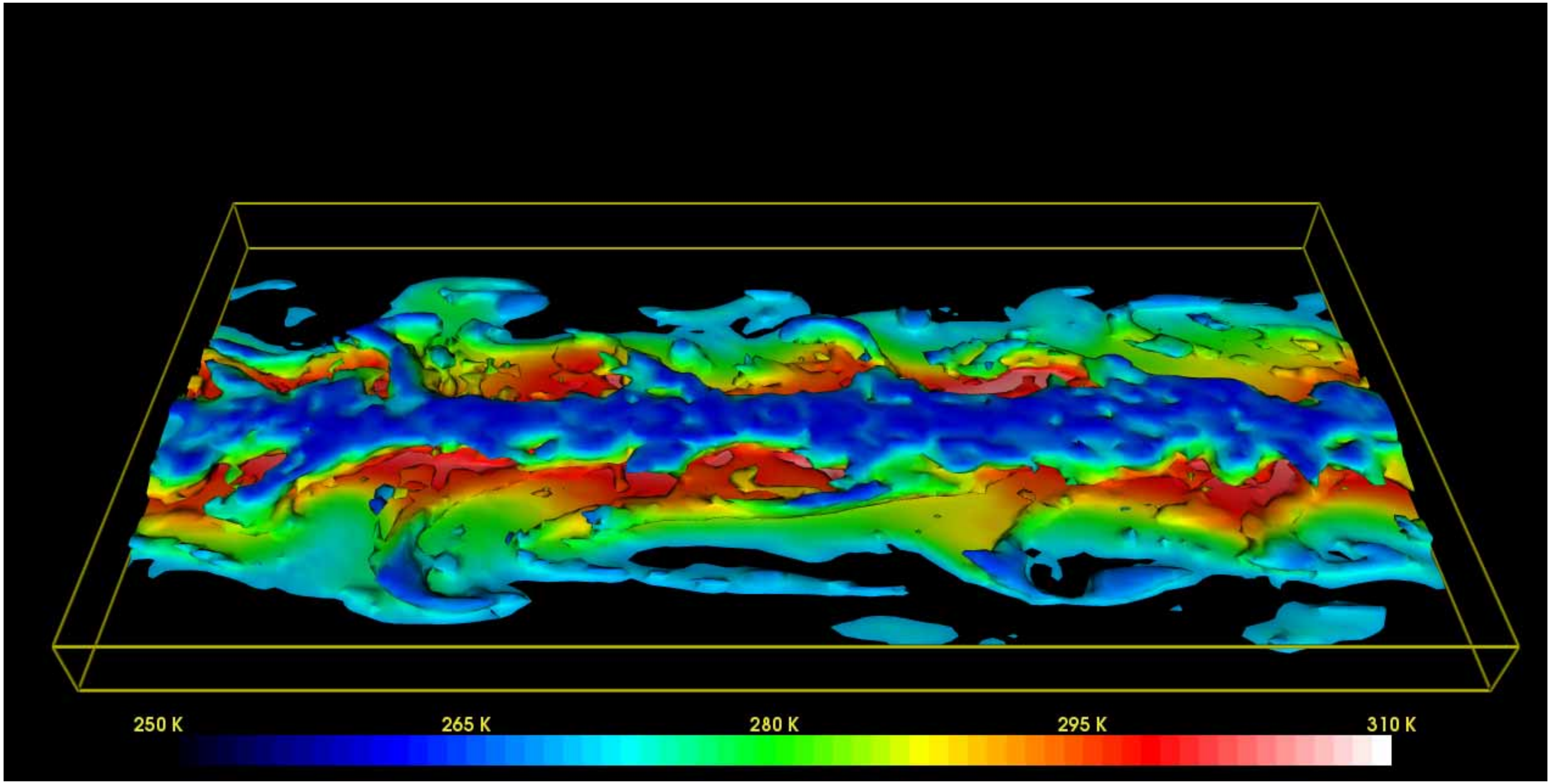
HOMME

- **National Center for Atmospheric Research**
  - Cooperation NCAR, Boulder and IBM
- **Moist Held-Suarez test**
  - Atmospheric moist processes fundamental component of atmospheric dynamics
    - Most uncertain aspect of climate change research
  - Moisture injected into the system at a constant rate from the surface
- **Importance of problem**
  - Moist processes must be included for relevant weather model
    - Formation of clouds and the development and fallout of precipitation
  - Requires high horizontal and vertical resolution
    - Order of 1 kilometer
  - Key to a better scientific understanding of global climate change

HOMME – Strong Scaling



HOMME – Visualization



BlueGene/L-Tuned Applications

- **Amber7**: Classical molecular dynamics used by AIST and IBM Blue Matter.
- **Blue Matter**: (IBM: Robert Germain et al) Classical molecular dynamics for protein folding and lipids.
- **CPMD**: (Car-Parrinello (ab initio) quantum molecular dynamics by IBM) Strong scaling of SiC 216 atoms & 1000 atoms.
- **ddcMD**: (LLNL: Classical molecular dynamics; Fred Streitz, Jim Glosli, Mehul Patel)
- **Enzo:** (UC San Diego) simulation of galaxies, has performance problem on every platform.
- **Flash**: (University of Chicago & Argonne) Collapse of stellar core and envelope explosion. Supernova simulation.
- **GAMESS**: Quantum Chemistry
- **HOMME**: (NCAR, Richard Loft) Climate code, 2d model of cloud physics.
- **HPCMW** (RIST): Solver for finite elements
- **LJ** (Caltech): Lennard Jones molecular dynamics
- **LSMS**: (Oak Ridge National Lab: Thomas Schulthess and Mark Fahey) First principles Material Science.
- **MDCASK**: (LLNL: Classical molecular dynamics; Alison Kutoba, Tom Spelce)
- **Miranda** (LLNL: instability/turbulence; Andy Cook, Bill Cabot, Peter Williams, Jeff Hagelberg)
- **MM5** from NCAR: meso-scale weather prediction
- **NAMD**: Molecular Dynamics
- **NIWS** (Nissei): Financial/Insurance Portfolio Simulation
- **PAM-CRASH**: (ESI) Automobile crash simulation.
- **ParaDis**: (LLNL: dislocation dynamics;Vasily Bulatov, Gregg Hommes)
- **Polycrystal**: (Caltech) material science
- **Qbox**: Quantum Chemistry, ab initio quantum molecular dynamical calculation.
- **Quarks** (Boston University, Joe Howard)
- **Raptor** (LLNL: instability/turbulence; Jeff Greenough, Charles Rendleman)
- **QCD**: (IBM Pavlos Vranas) sustained 1 TF/s on one rack. 19% uni efficiency.
- **QMC**: (Caltech) Quantum Chemistry
- **SAGE**: (LANL: SAIC's Adaptive Grid Eulerian Code) AMR hydrodynamics. Heat and radiation transport with AMR.
- **SPHOT**: (LLNL) 2D photon transport
- **SPPM**: **S**implified **P**iecewise **P**arabolic Method. 3-D gas dynamics on a uniform Cartesian grid.
- **Sweep3d**: (LANL) 3-d neutron transport
- **TBLE**: magnetohydrodynamics
- **UMT2K**: (LLNL) photon transport 3d Boltzmann on unstructured grid

BlueGene/L Performance and Density

Performance Metric	Single Rack Blue Gene
Peak Teraflops (Virtual Node mode)	5.73
Peak Teraflops **(**Coprocessor mode)	2.86
Linpack Teraflops	4.53

Metric	ASCI White	ASCI Q	Earth Simulator	BG/L
Memory/Space (GB/m^2)	8.6	17	3.1	***140***
Speed/Space ($GFlops/m^2$)	13	16	13	***1600***
Speed/Power (GFlops/kW)	12	7.9	4	***300***

BlueGene/L - Paradigm Shift for Supercomputers

- **Aggregate performance is important**
  - Not performance of individual chip
- **Simple building block**
  - High integration on a single chip
    - Processors, memory, interconnect subsystems
  - Low power → allows high density packaging
  - Cost-effective solution

→ As a result, breakthrough in compute power
  - Per Watt
  - Per square meter of floor space
  - Per dollar

→ BlueGene/L enables
  - New unparalleled application performance
  - Breakthroughs in science by providing unprecedented compute power

BlueGene/L on the Web

www.research.ibm.com/bluegene

The Blue Gene/L project has been supported and partially funded by the Lawrence Livermore National Laboratories on behalf of the United States Department of Energy, under Lawrence Livermore National Laboratories Subcontract No. B517552.

System Features		BG/L
Node Properties	Node Processors	2 × PowerPC440
	Processor Frequency	700MHz
	L1 Cache (private) I+D	32+32KB/processor
	L2 Cache (private)	14 (7) stream prefetching
	L3 Cache size (shared)	4MB
	Main Store	256MB/512MB/1GB
	Main Store Bandwidth	5.6GB/s
	Peak Performance	5.6GF/node
Torus Network	Bandwidth (per node)	6*2*175MB/s=2.1GB/s
	Hardware Latency (Nearest Neighbor)	200ns (32B packet) 1.6µs (256B packet)
	Hardware Latency (Worst Case)	6.4µs (64 hops)
Collective Network	Bandwidth (per node)	3*2*350MB/s=2.1GB/s
	Hardware Latency (round trip worst case)	5.0µs

BlueGene/L at a Glance

Attribute	Details	Benefits
Processor	PowerPC 440 700MHz; two per node	Low power allows dense packaging; better processor-memory balance
Memory	512 MB SDRAM-DDR per node	
Networks	3D Torus - 175MB/sec in each direction Collective Network – 350MB/sec; 1.5 usec latency Global Barrier/Interrupt Gigabit Ethernet (machine control and outside connectivity)	Special networks speed up internode communications; designed for MPI programming constructs; improve systems management
Compute Nodes	Dual processor; 1024 per rack	Double FPU improves performance
I/O Nodes	Dual processor; 16 per rack (additional nodes optional)	Strengthens systems management
Operating Systems	Compute Node – Lightweight proprietary kernel I/O Node – Embedded Linux Front End and Service Nodes – SuSE SLES 9 Linux	Kernel tailored to processor design; industry-standard distribution preserves familiarity to end user
Performance	Peak per rack (virtual node mode) – 5.73 teraflops Peak per rack (coprocessor mode) – 2.86 teraflops Linpack per rack – 4.53 teraflops	Highest available performance benefits capability customers
Power	28.14 kW power consumption per rack (maximum) 208 VAC 3-phase; 100 amp service per rack	Low power draw enables dense packaging
Cooling	Air conditioning 8 tons/rack (minimum) 2800 CFM (compute rack); 350 CFM (power supplies)	Low cooling requirements enable extreme scale-up
Acoustics	9.0 LwAD and 8.7 LwAm	
Dimensions (includes air duct)	Height – 1958 mm Width – 915 mm Depth – 915 mm Weight – 750 Kg	Design allows "brickwall" layout for better floor space utilization