



IBM Systems and Technology Group

The Blue Gene/L Supercomputer

Burkhard Steinmacher-Burow
IBM Böblingen
steinmac@de.ibm.com



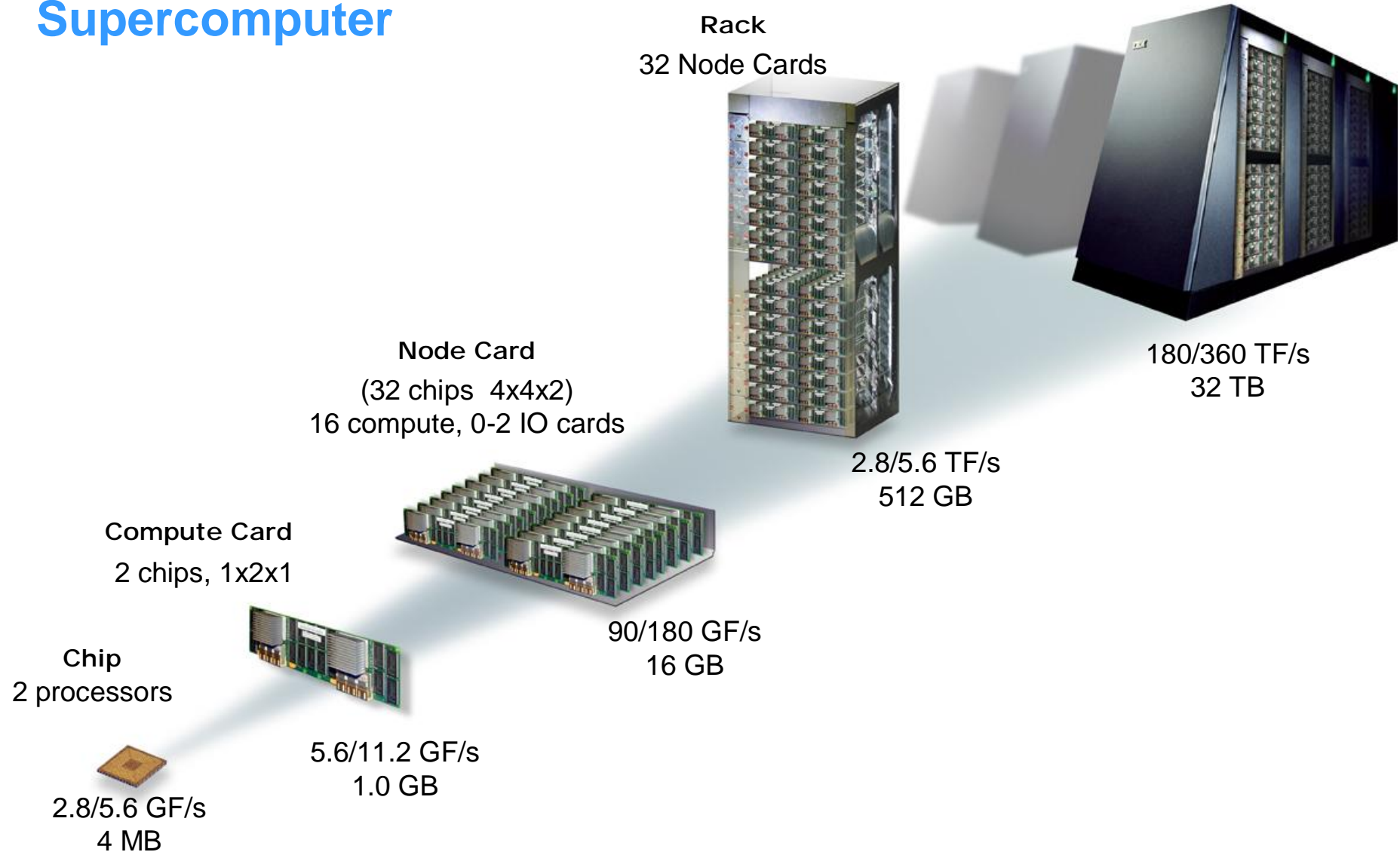
DESY-Hamburg, Feb.21, 2005

© 2005 IBM Corporation

Outline

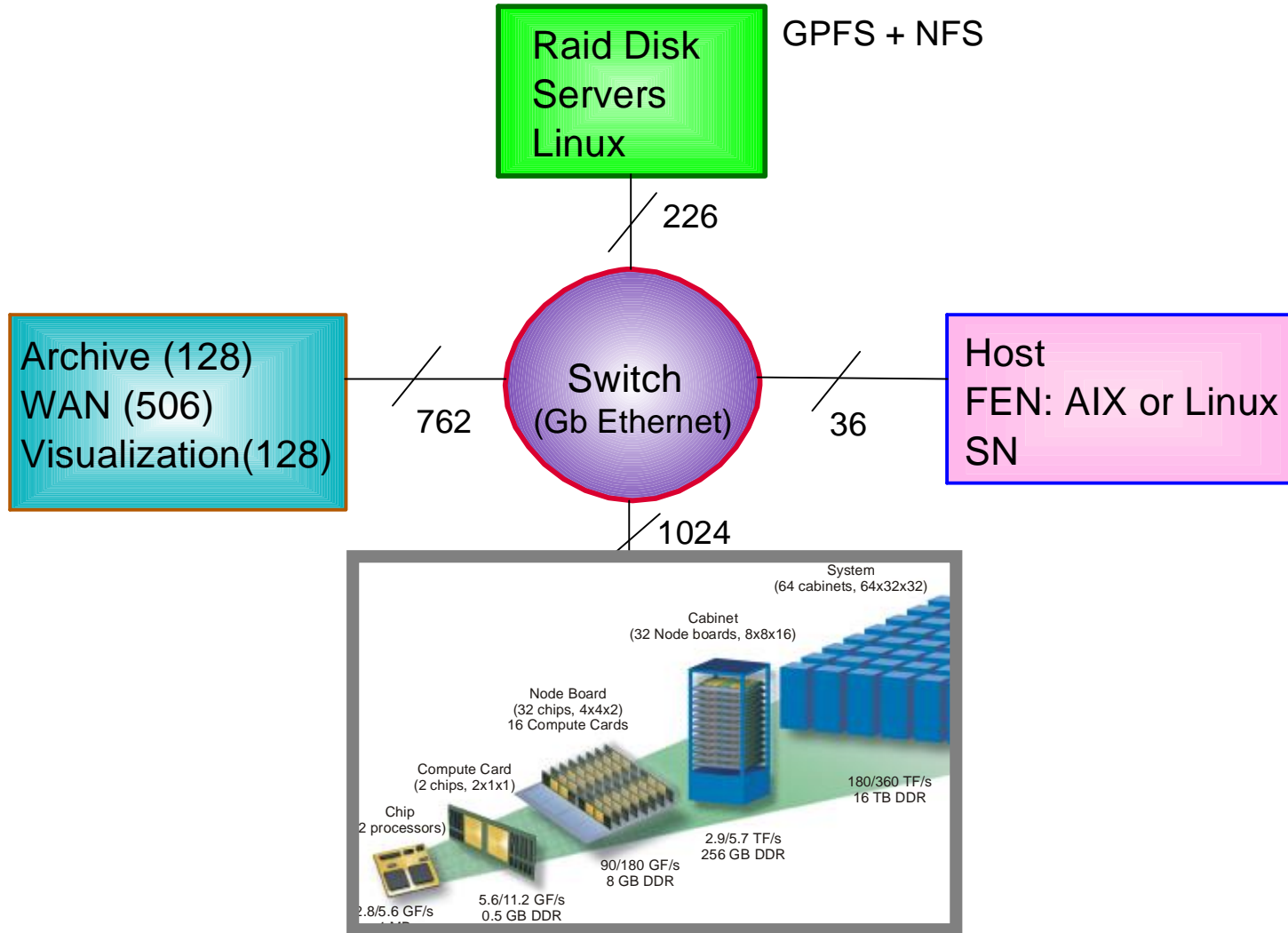
- § Introduction to BG/L
- § Motivation
- § Architecture
- § Packaging
- § Software
- § Example Applications and Performance
- § Summary

The Blue Gene/L Supercomputer



DESY-Hamburg, Feb.21, 2005

Blue Gene/L just provides processing power, requires Host Environment



A High-Level View of the BG/L Architecture: --- A computer for MPI or MPI-like applications. ---

§ Within node:

- 4 Low latency, high bandwidth memory system.
- 4 Strong floating point performance: 4 FMA/cycle.

§ Across nodes:

- 4 Low latency, high bandwidth networks.

§ Many nodes:

- 4 Low power/node.
- 4 Low cost/node.
- 4 RAS (reliability, availability and serviceability).

§ Familiar SW API:

- 4 C, C++, Fortan, MPI, POSIX subset, ...

NB All application code runs on BG/L nodes;
external host is just for file and other system services.

Specialized means Less General

BG/L leans towards MPI	BG/L leans away from General Purpose Computer
Space-shared nodes in units of $8*8*8=512$ nodes.	Time-shared nodes.
Use only real memory.	Virtual memory to disk.
No asynchronous OS activities.	OS services.
Distributed memory across nodes.	Shared memory.
No internal state between applications. [Helps performance and functional reproducibility.]	Built-in filesystem.
Requires General Purpose Computer as Host.	

Who needs a huge MPI computer?

- § BG/L has strategic partnership with Lawrence Livermore National Laboratory (LLNL) and other high performance computing centers:
 - 4 Focus on numerically intensive scientific problems.
 - 4 Validation and optimization of architecture based on real applications.
 - 4 Grand challenge science stresses networks, memory and processing power.
 - 4 Partners accustomed to "new architectures" and work hard to adapt to constraints.
 - 4 Partners assist us in the investigation of the reach of this machine.

Main Design Principles for Blue Gene/L

- § Recognize that some science & engineering applications scale up to and beyond 10,000 parallel processes.
- § So expand computing capability, holding total system cost.
- § So reduce cost/FLOP.
- § So reduce complexity and size.
 - 4 Recognize that ~25KW/rack is max for air-cooling in standard room.
 - So need to improve performance/power ratio.
This improvement can decrease performance/node, since assume can scale to more nodes.
 - 700MHz PowerPC440 for ASIC has excellent FLOP/Watt.
 - 4 Maximize Integration:
 - On chip: ASIC with everything except main memory.
 - Off chip: Maximize number of nodes in a rack.
- § Large systems require excellent reliability, availability, serviceability (RAS)
- § Major advance is scale, not any one component.

Main Design Principles (continued)

§ Make cost/performance trade-offs considering the end-use:

4 Applications ↔ Architecture ↔ Packaging

– Examples:

- 1 or 2 differential signals per torus link.
I.e. 1.4 or 2.8Gb/s.
- Maximum of 3 or 4 neighbors on collective network.
I.e. Depth of network and thus global latency.

§ Maximize the overall system efficiency:

4 Small team designed all of Blue Gene/L.

4 Example: Chose ASIC die and chip pin-out to ease circuit card routing.

Example of Reducing Cost and Complexity

§ Cables are bigger, costlier and less reliable than traces.

4 So want to minimize the number of cables.

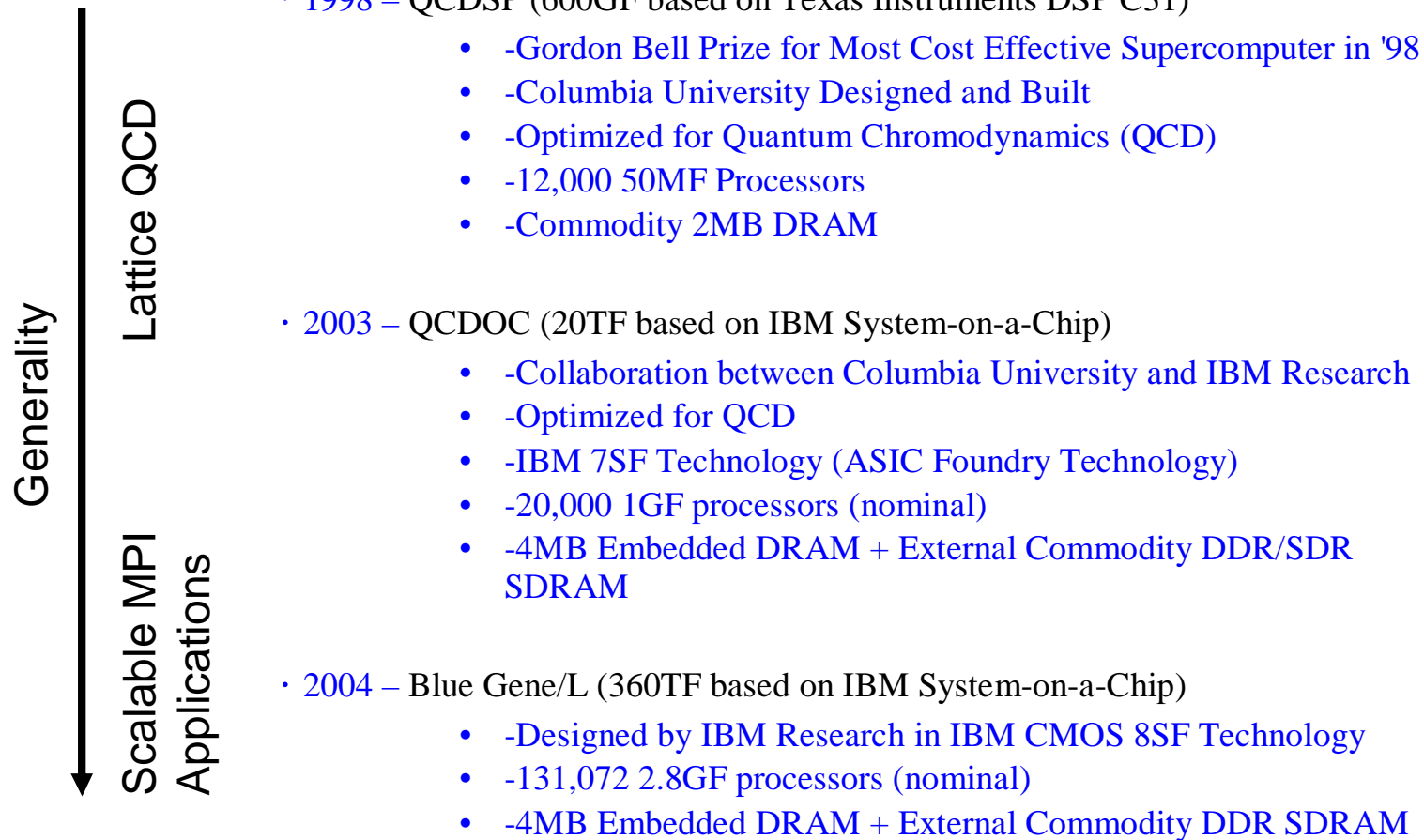
4 So:

- Choose 3-dimensional torus as main BG/L network, with each node connected to 6 neighbors.
- Maximize number of nodes connected via circuit card(s) only.

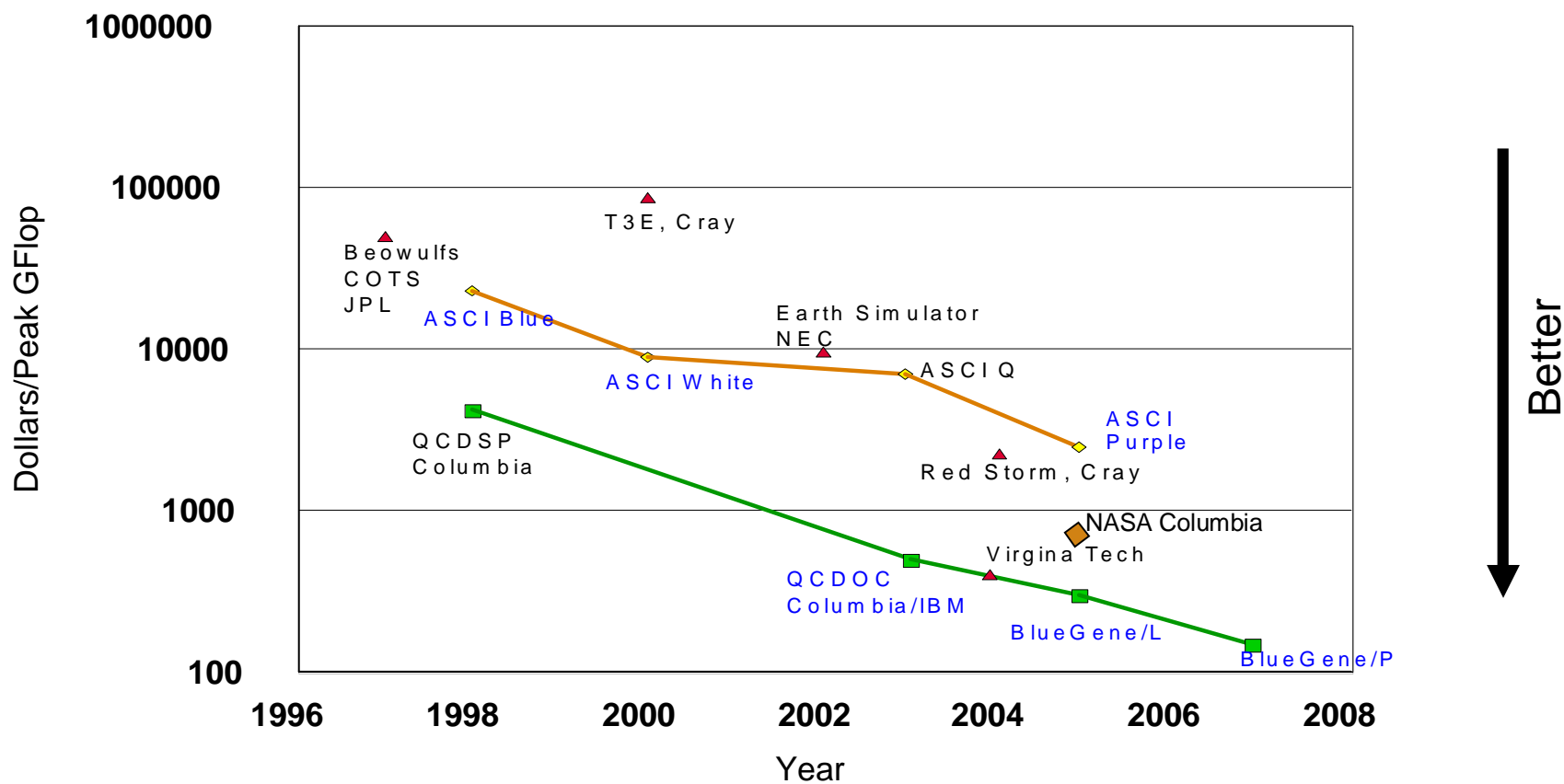
§ BG/L midplane has $8*8*8=512$ nodes.

§ (Number of cable connections) / (all connections)
= (6 faces * 8 * 8 nodes) / (6 neighbors * 8 * 8 * 8 nodes)
= 1 / 8

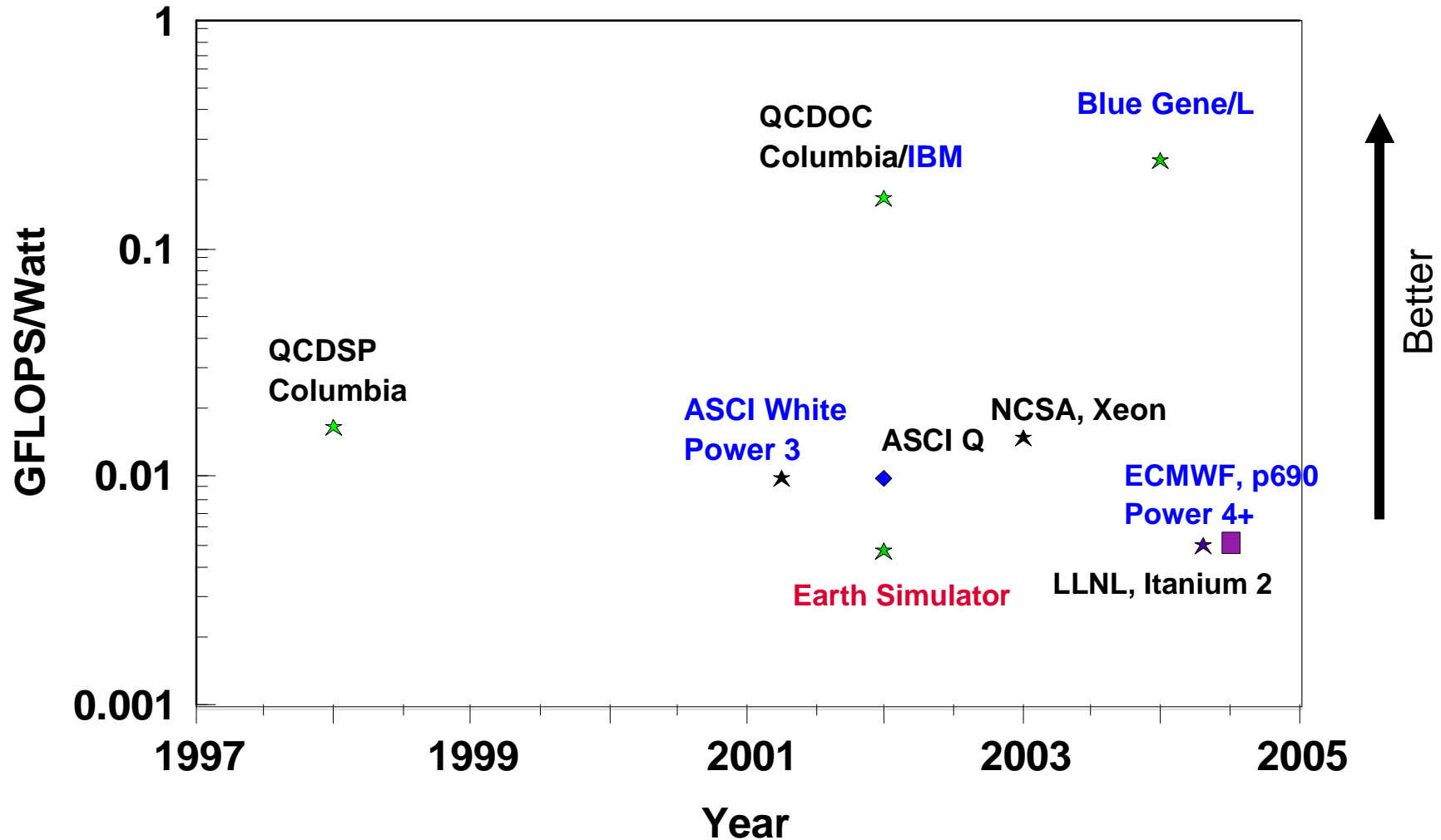
Some BG/L Ancestors



Supercomputer Price/Peak Performance



Supercomputer Power Efficiencies

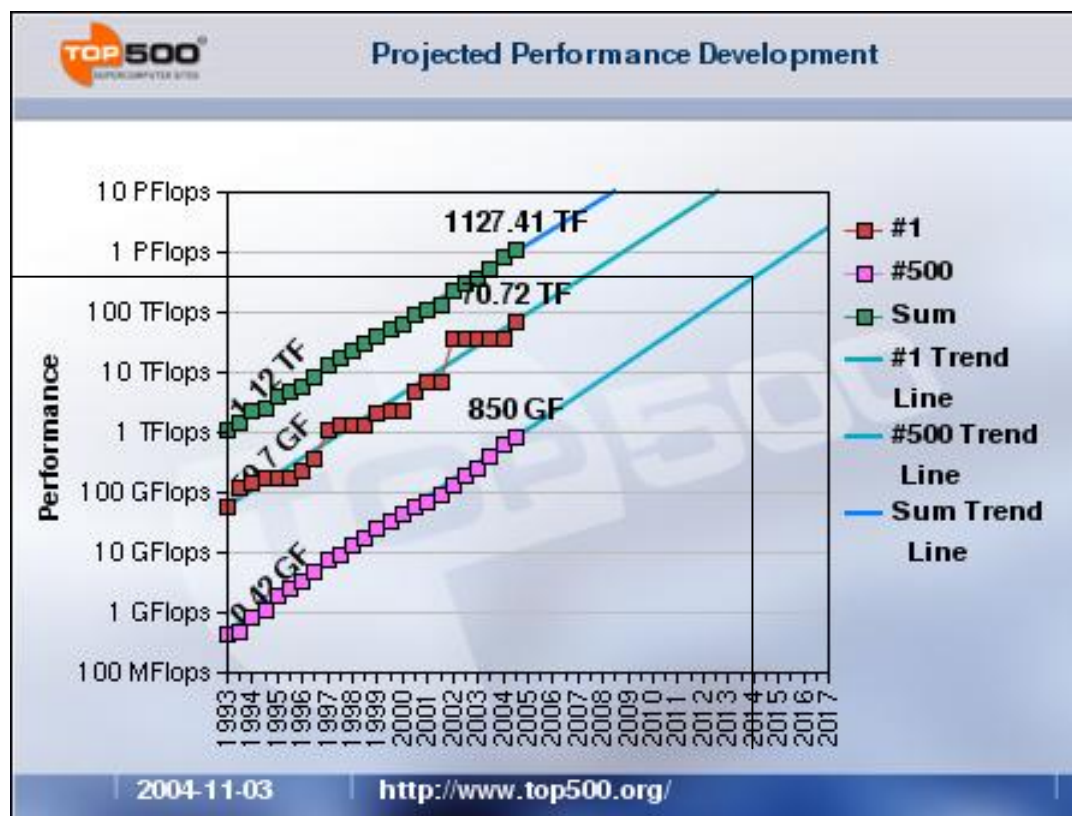


Similar space efficiency story since cooling/rack is similar across systems.

Need Very Aggressive Schedule

- Competitor performance is doubling every year!
- Year 2014 : 64K-node BG/L no longer on Top 500

250TF Linpack



2014

BG/L Timeline

- § December 1999: IBM announces 5 year, US\$100M effort to build a petaflop/s scale supercomputer to attack science problems such as protein folding. Goals:
 - 4 Advance scientific simulation.
 - 4 Advance computer hw&sw for capability and capacity markets.

- § November 2001: Research partnership with (LLNL).
November 2002: Planned acquisition of a BG/L machine by LLNL announced.

- § June 2003: First-pass chips (DD1) completed. (Limited to 500MHz).

- § November 2003: 512-node DD1 achieves 1.4TF Linpack for #73 on top500.org.
 - 4 32-node prototype folds proteins live on the demo floor at SC2003.

- § February 2, 2004: Second pass (DD2) BG/L chips achieves 700MHz design.

- § June 2004: 2rack 2048-node DD2 system achieves 8.7TF Linpack for #8 on top500.org.
4rack 4096-node DD1 prototype achieves 11.7TF Linpack for #4.

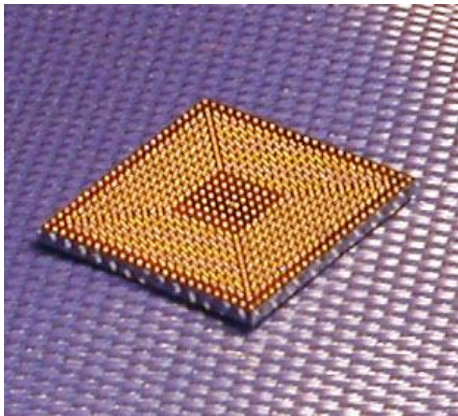
- § November 2004: 16rack 16384-node DD2 achieves 71TF Linpack for #1 on top500.org.
System moved to LLNL for installation.
eServer BG/L product announced at ~\$2m/rack for qualified clients.

- § 2005: Complete 64rack LLNL system.
Install other systems: 6rack Astron, 4rack AIST, 1rack Argonne, 1rack SDSC,
1rack Edinburg, 20rack Watson, ...

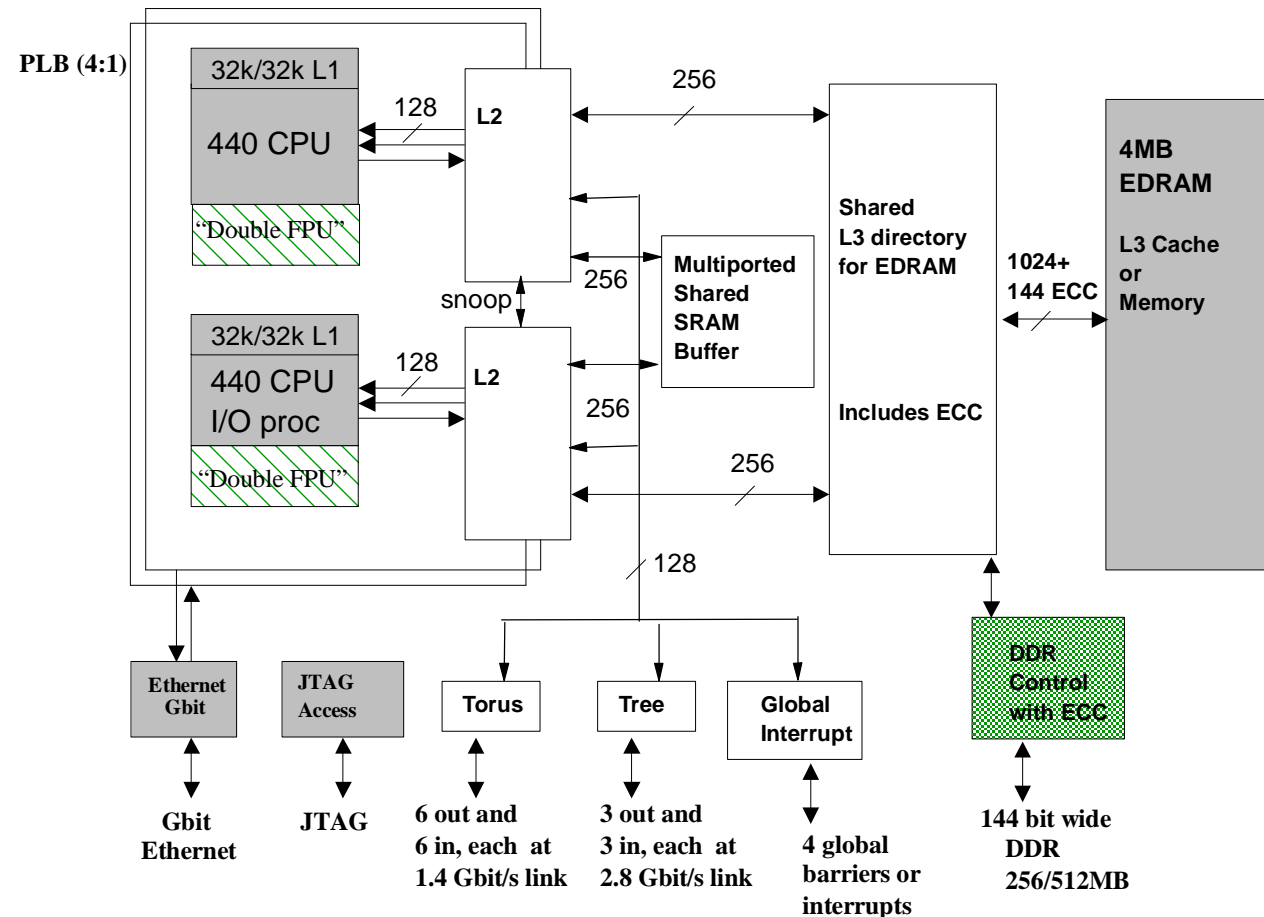
Blue Gene/L Architecture

- § Up to $32 \times 32 \times 64 = 65536$ nodes.
- § 5 networks connect nodes to themselves and to the world.
- § Each node is 1 ASIC + 9 DRAM chips.

BlueGene/L Compute ASIC

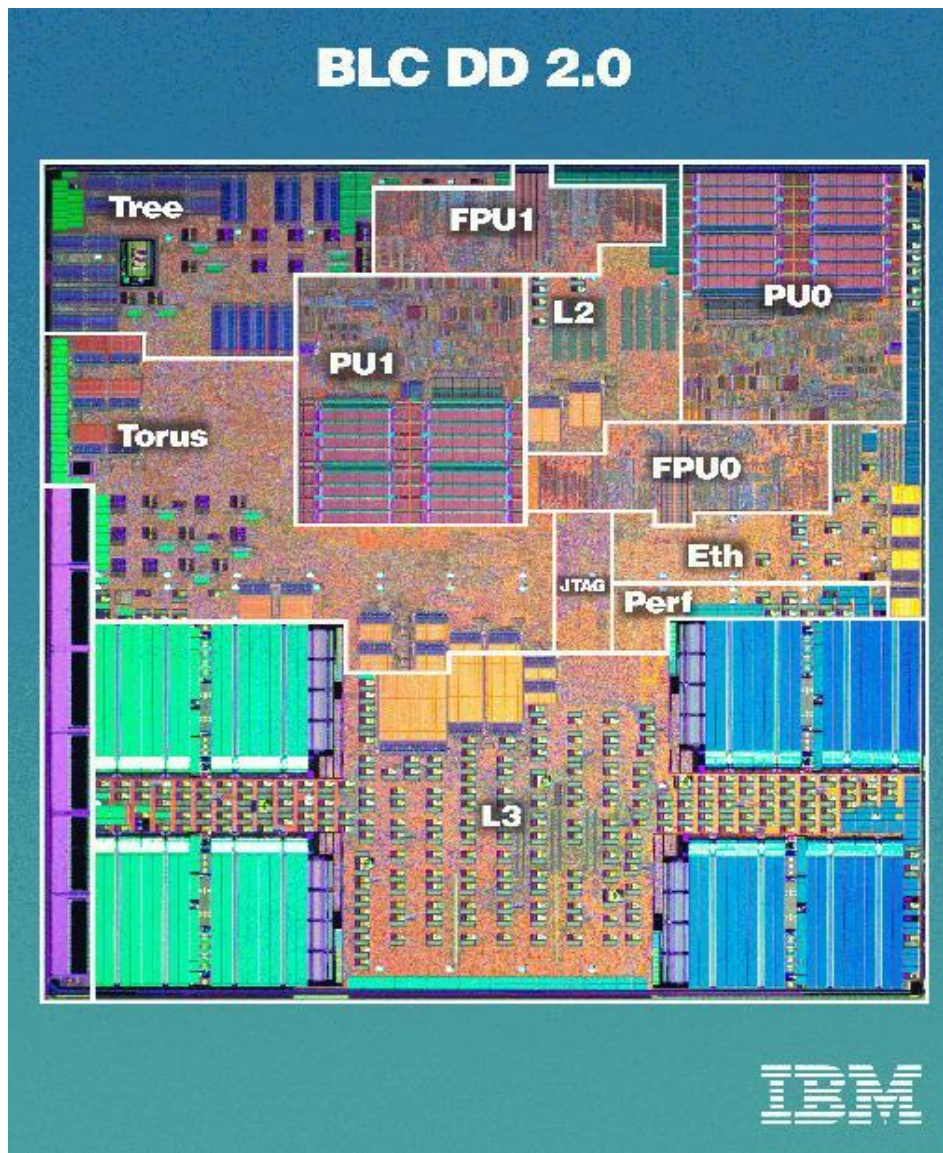


- IBM CU-11, 0.13 μm
- 11 x 11 mm die size
- 25 x 32 mm CBGA
- 474 pins, 328 signal
- 1.5/2.5 Volt

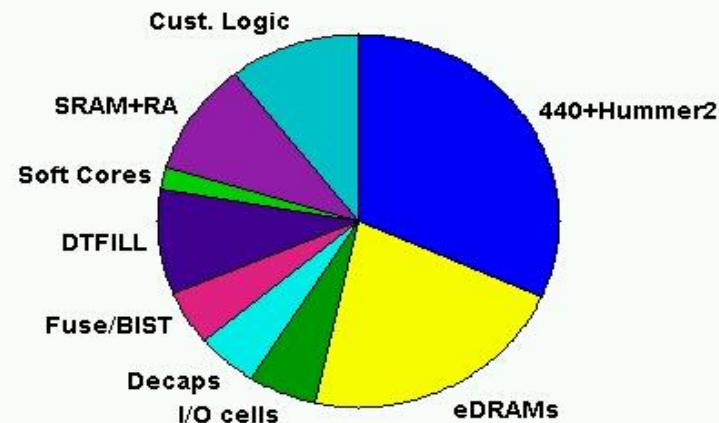


8m² of compute ASIC silicon in 65536 nodes!

BlueGene/L – System-on-a-Chip



Chip Area usage



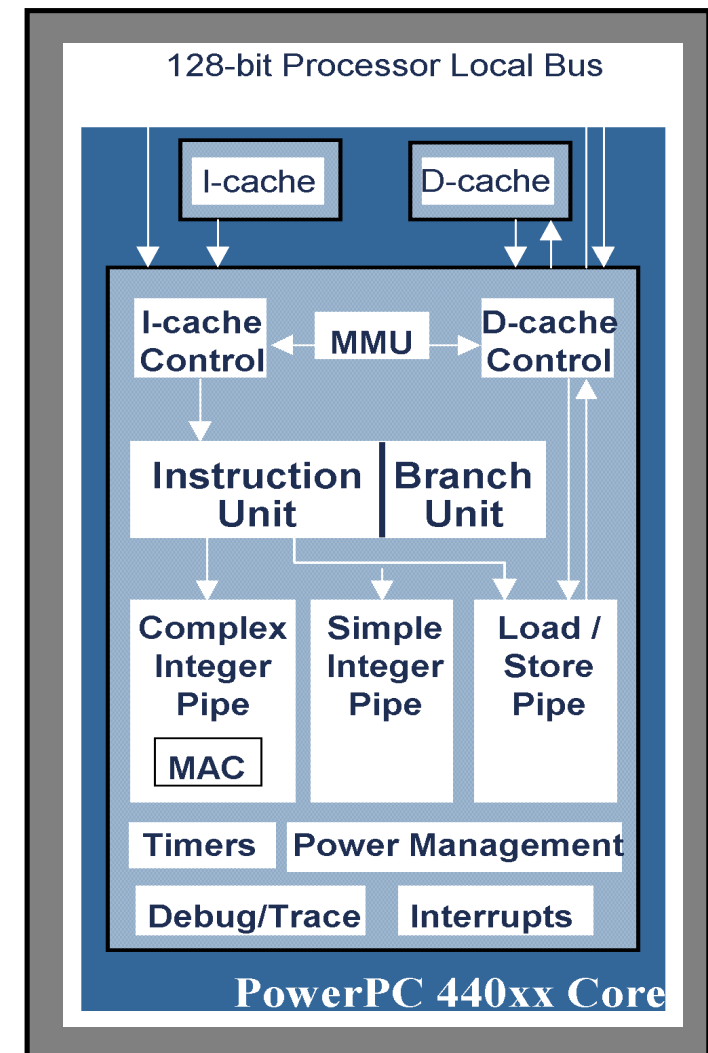
Cell Count	57M
Transistor Count	95M
Placeable Objects	1.1M
Clock Freq.	700MHz
Power Dissipation	13W
Bit Count eDRAM	38M
Bit Count eSRAM	2.6M

Main BG/L Frequencies

- § 700MHz processor.
- § Torus link is 1 bit in each direction at $2 \times 700\text{MHz} = 1.4\text{GHz}$.
(Collective network is 2 bits wide.)
- § 700MHz clock distributed from single source to all 65536 nodes, with $\sim 25\text{ps}$ jitter between any pair of nodes.
 - 4 Low jitter achieved by same effective fan-out from source to each node.
 - 4 Low jitter required by torus and collective network signalling.
 - No clock sent with data, no receiver clock extraction.
 - Synchronous data capture trains to and tracks phase difference between nodes.
- § Each node ASIC has 128+16 bits @ 350MHz to external memory.
(I.e. 5.6GB/s read xor write with ECC.)

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
- § Dynamic branch prediction
- § Caches
 - f* 32KB instruction & 32KB data cache
 - f* 64-way set associative, 32 byte line
- § 32-bit virtual address
- § Real-time non-invasive trace
- § 128-bit CoreConnect Interface



Floating Point Unit

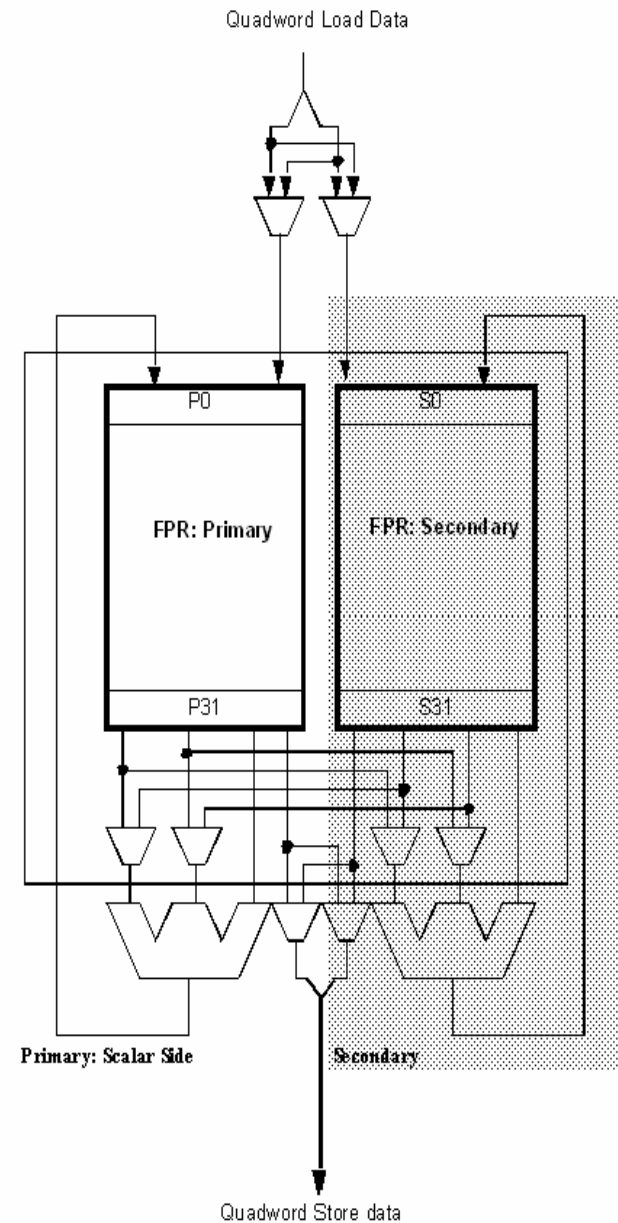
Primary side acts as off-the-shelf PPC440 FPU.

- § FMA with load/store each cycle.
- § 5 cycle latency.

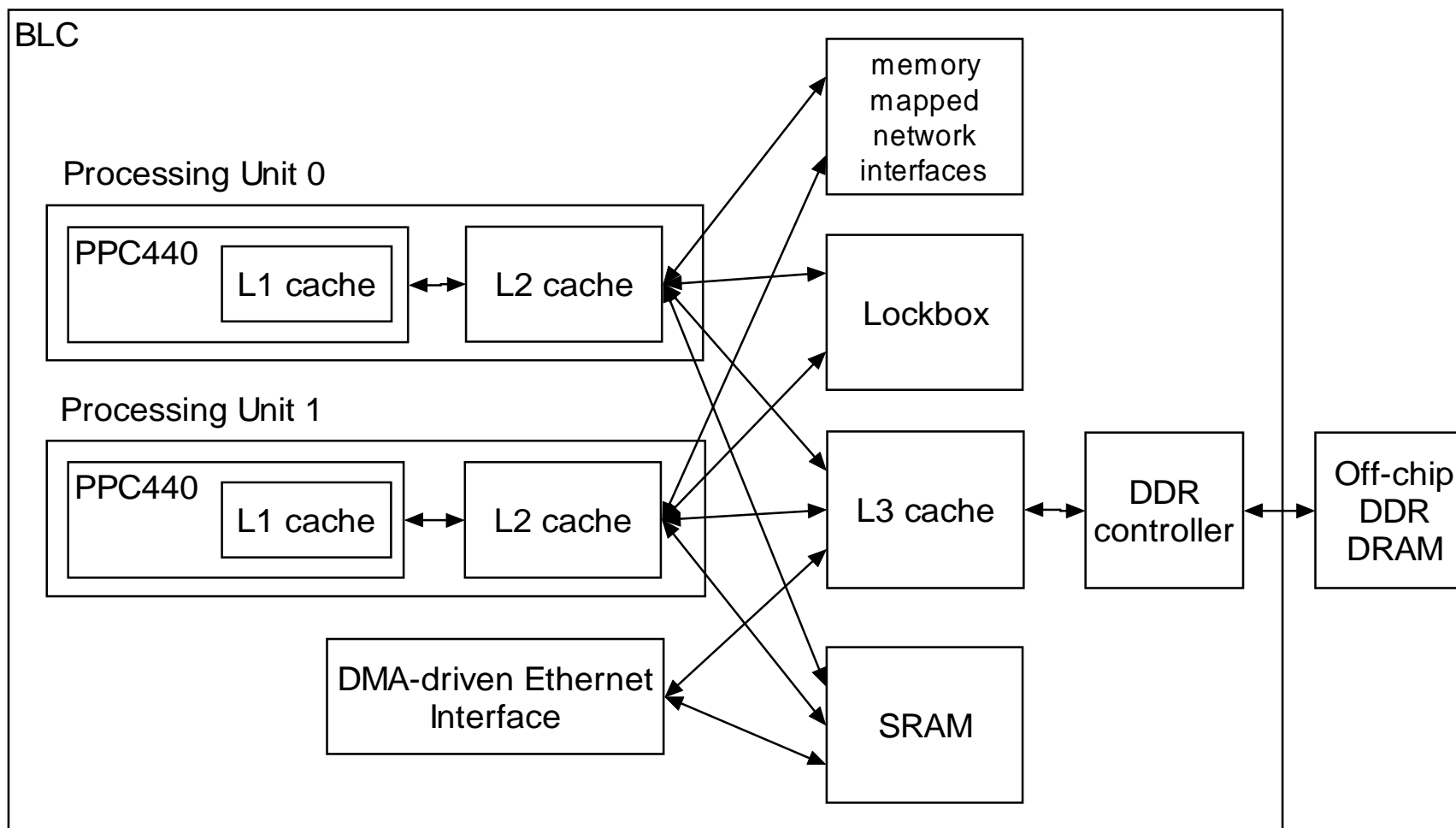
Secondary side doubles the registers and throughput.

Enhanced set of instructions for:

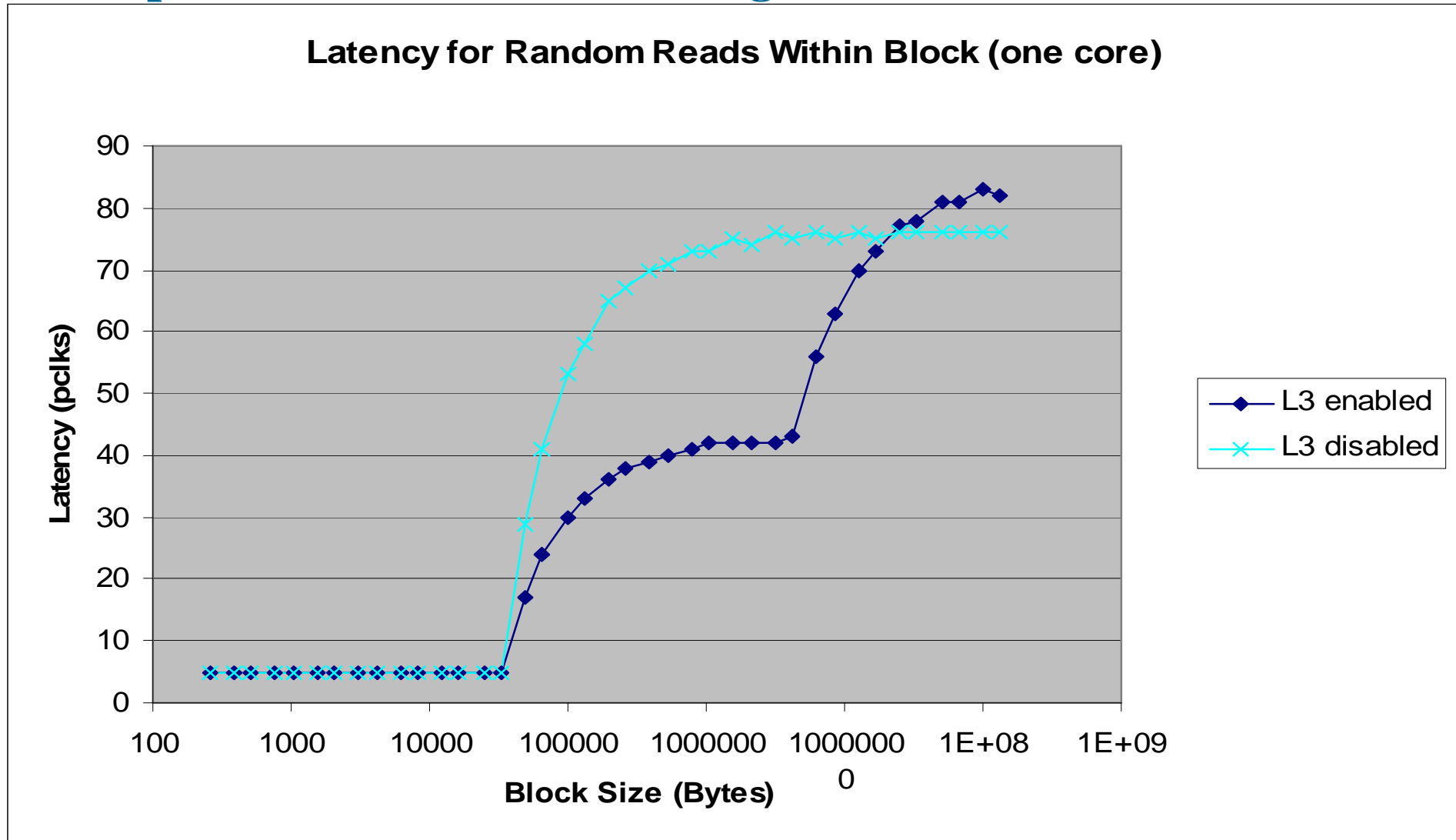
- § Secondary side only.
- § Both sides simultaneously:
 - 4 Usual SIMD instructions.
E.g. Quadword load, store.
 - 4 Instructions beyond SIMD. E.g.
 - SIMOMD
Single Inst. Multiple Operand Multiple Data.
 - Access to other register file.



Memory Architecture



BlueGene/L Measured Memory Latency Compares Well to Other Existing Nodes



180 versus 360 TeraFlops for 65536 Nodes

The two PPC440 cores on an ASIC are **NOT** an SMP!

- § PPC440 in 8SF does not support L1 cache coherency.
- § Memory system is strongly coherent L2 cache onwards.

180 TeraFlops = 'Co-Processor Mode'

- § A PPC440 core for application execution.
- § A PPC440 core as communication co-processor.
- § Communication library code maintains L1 coherency.

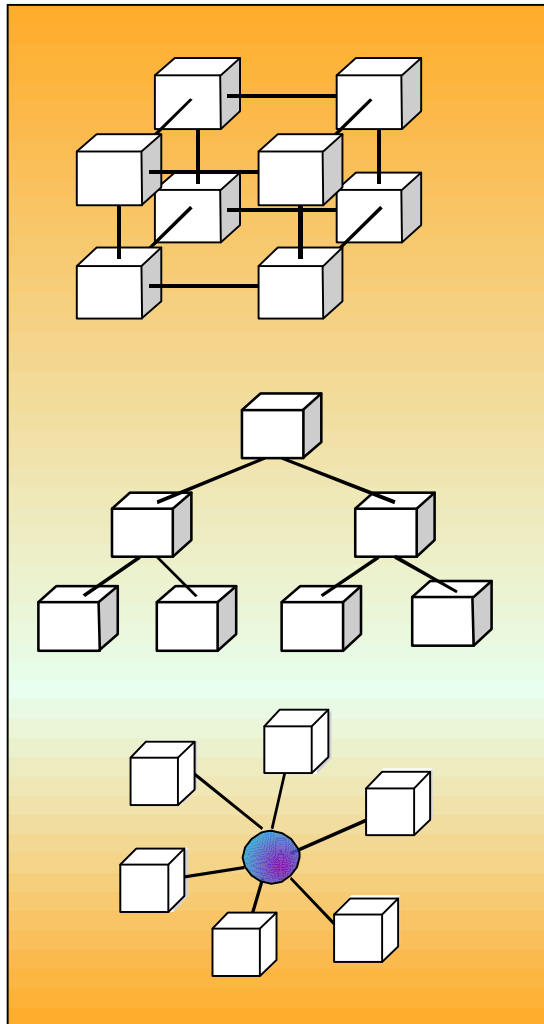
360 TeraFlops = 'Virtual Node Mode'

- § On a physical node,
each of the two PPC440 acts as an independent 'virtual node'.
Each virtual node gets:
 - 4 Half the physical memory on the node.
 - 4 Half the memory-mapped torus network interface.

In either case, no application-code dealing with L1-coherency.

Blue Gene Interconnection Networks

Optimized for Parallel Programming and Scalable Management



3-Dimensional Torus

- 4 Interconnects all compute nodes (65,536)
- 4 Virtual cut-through hardware routing
- 4 1.4Gb/s on all 12 node links (2.1 GB/s per node)
- 4 Communications backbone for computations
- 4 0.7/1.4 TB/s bisection bandwidth, 67TB/s total bandwidth

Global Collective Network

- 4 One-to-all broadcast functionality
- 4 Reduction operations functionality
- 4 2.8 Gb/s of bandwidth per link; Latency of tree traversal 2.5 μ s
- 4 ~23TB/s total binary tree bandwidth (64k machine)
- 4 Interconnects all compute and I/O nodes (1024)

Low Latency Global Barrier and Interrupt

- 4 Round trip latency 1.3 μ s

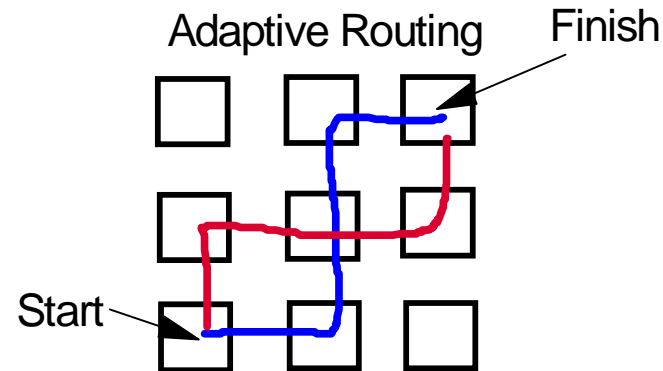
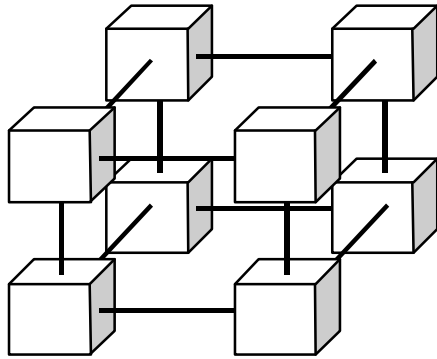
Control Network

- 4 Boot, monitoring and diagnostics

Ethernet

- 4 Incorporated into every node ASIC
- 4 Active in the I/O nodes (1:64)
- 4 All external comm. (file I/O, control, user interaction, etc.)

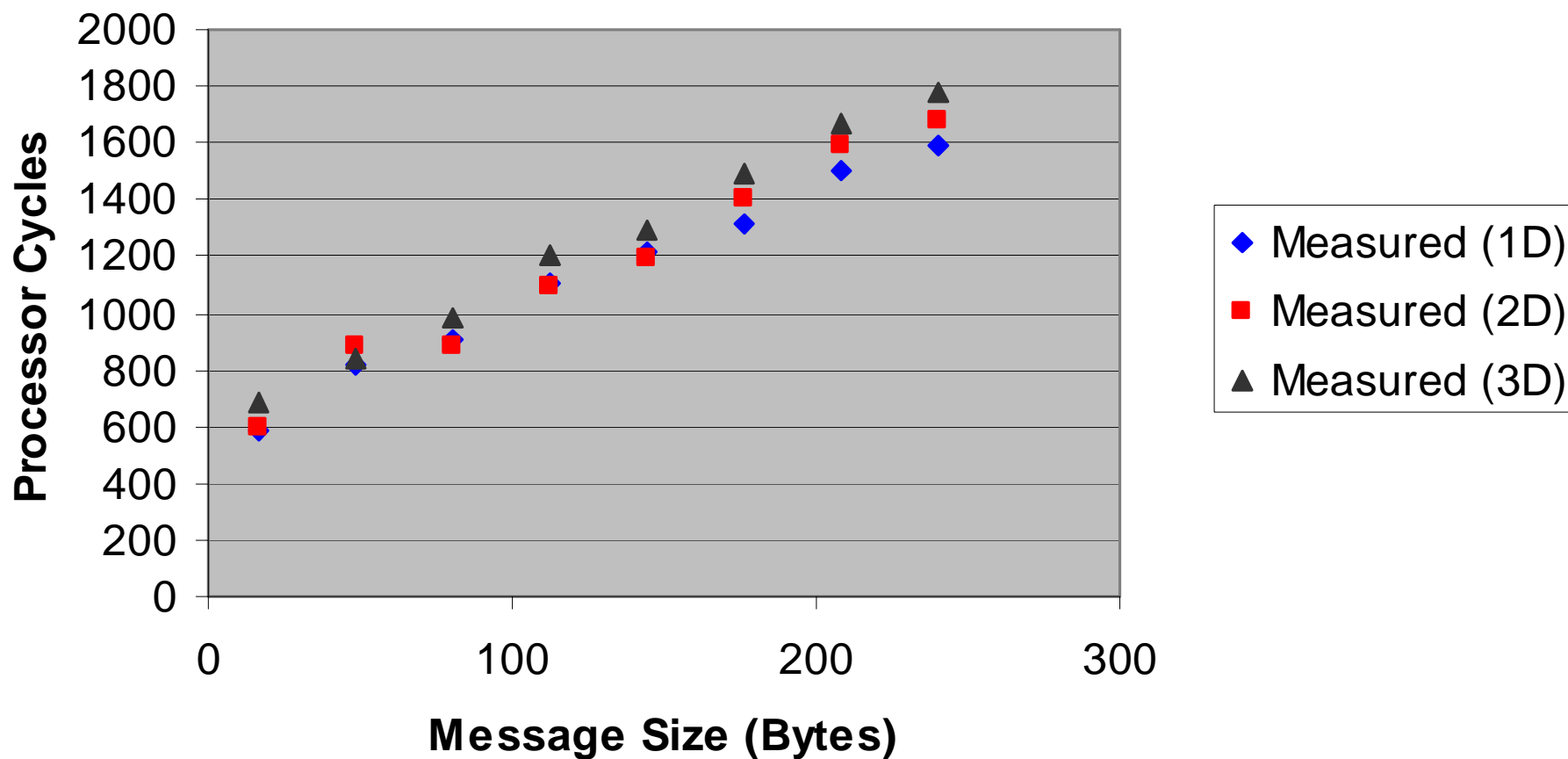
3-D Torus Network



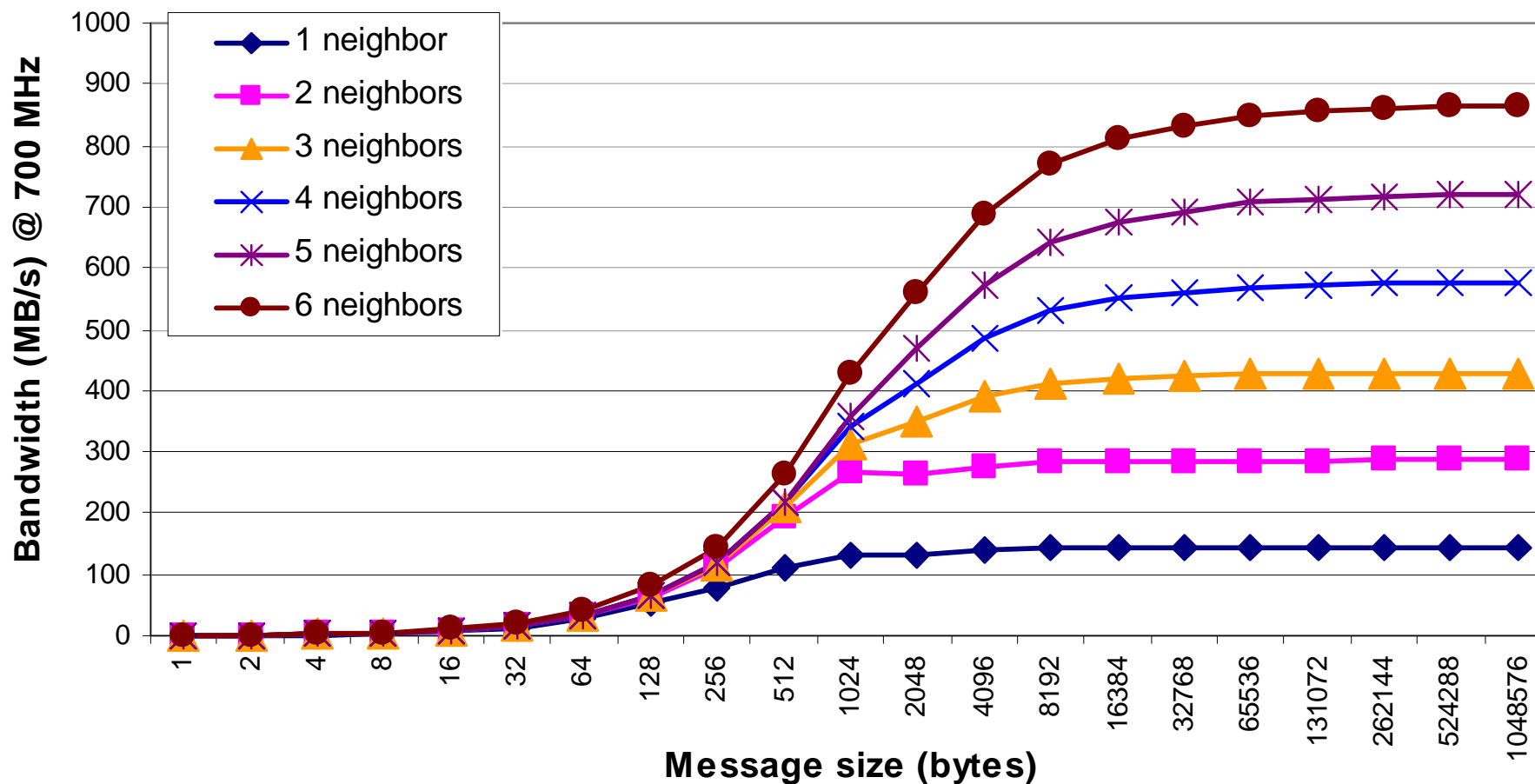
- **32x32x64 connectivity**
- **Backbone for one-to-one and one-to-some communications**
- **1.4 Gb/s bi-directional bandwidth in all 6 directions (Total 2.1 GB/s/node)**
- **$64k * 6 * 1.4Gb/s = 68 TB/s$ total torus bandwidth**
- **$4 * 32 * 32 * 1.4Gb/s = 5.6 Tb/s$ Bisectonal Bandwidth**
- **Worst case hardware latency through node ~ 69nsec**
- **Virtual cut-through routing with multipacket buffering on collision**
 - Minimal
 - Adaptive
 - Deadlock Free
- **Class Routing Capability (Deadlock-free Hardware Multicast)**
 - Packets can be deposited along route to specified destination.
 - Allows for efficient one to many in some instances
- **Active messages allows for fast transposes as required in FFTs.**
- **Independent on-chip network interfaces enable concurrent access.**

Prototype Delivers ~1usec Ping Pong low-level messaging latency

One-Way "Ping-Pong" times on a 2x2x2 Mesh (not optimized)

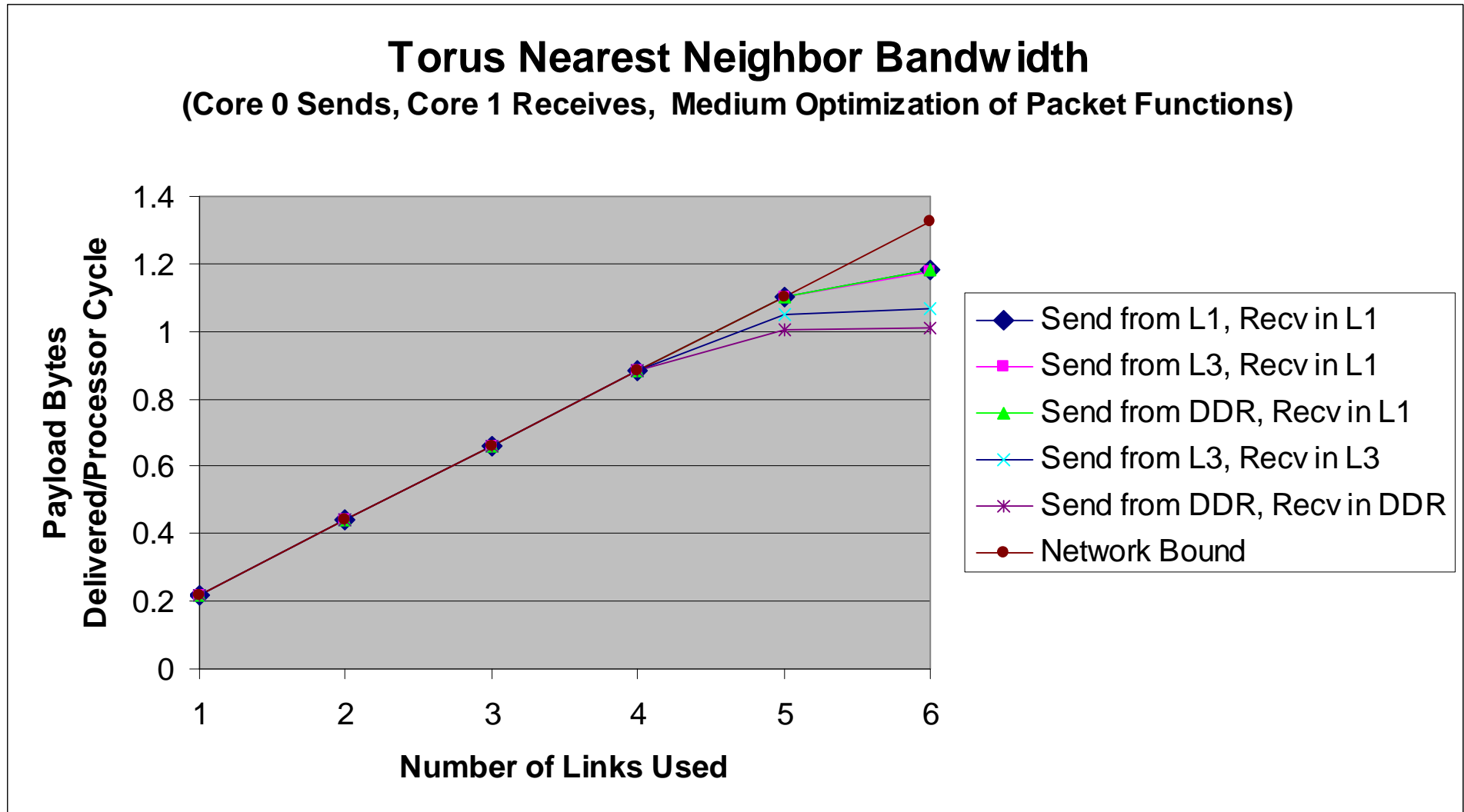


Measured MPI Send Bandwidth and Latency



Latency @700 MHz = $4 + 0.090 * \text{"Manhattan distance"} + 0.045 * \text{"Midplane hops"} \mu\text{s}$

Nearest neighbor communication achieves 75-80% of peak



Peak Torus Performance for Some Collectives

$L = 1.4\text{Gb/s} = 175\text{MB/s} = \text{Uni-directional Link Bandwidth}$

$N = \text{number of nodes in a torus dimension}$

$\text{All2all} = 8L/N_{\text{max}}$

§ E.g. $8*8*8$ midplane has 175MB/s to and from each node.

$\text{Broadcast} = 6L = 1.05\text{GB/s}$

§ 4 software hops, so fairly good latency.

§ Hard for two PPC440 on each node to keep up,
especially software hop nodes performing 'corner turns'.

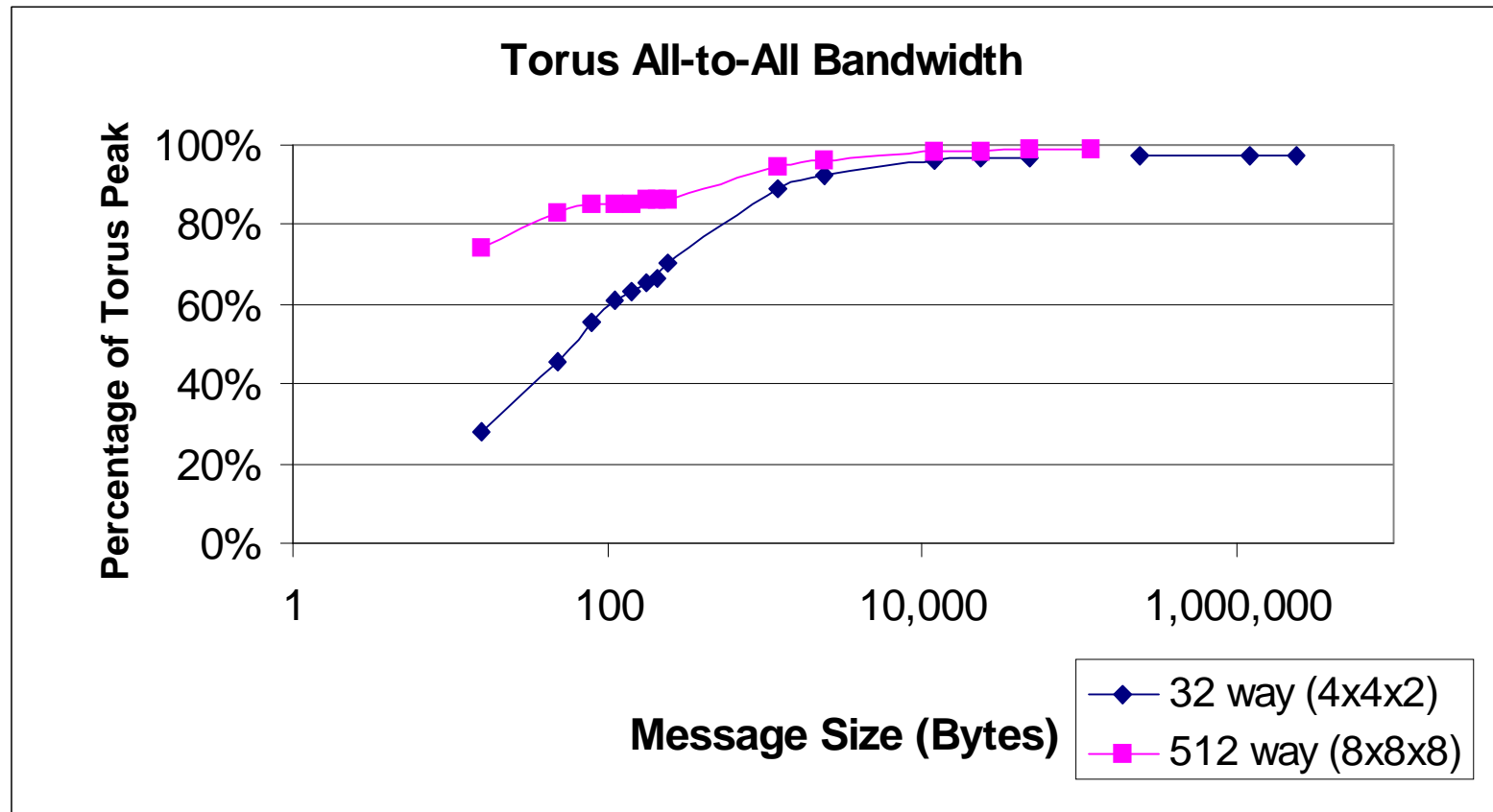
$\text{Reduce} = 6L = 1.05\text{GB/s}$

§ $(N_x+N_y+N_z)/2$ software hops, so needs large messages.

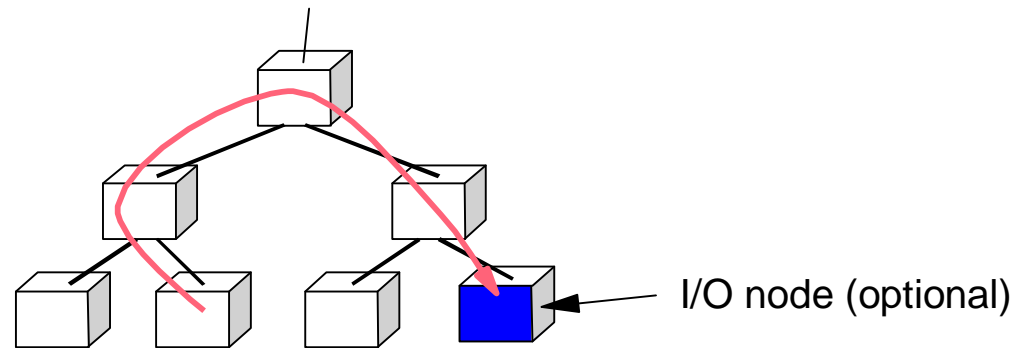
§ Very hard/Impossible for PPC440 to keep up.

$\text{AllReduce} = 3L = 0.525\text{GB/s}$

Link Utilization on Torus

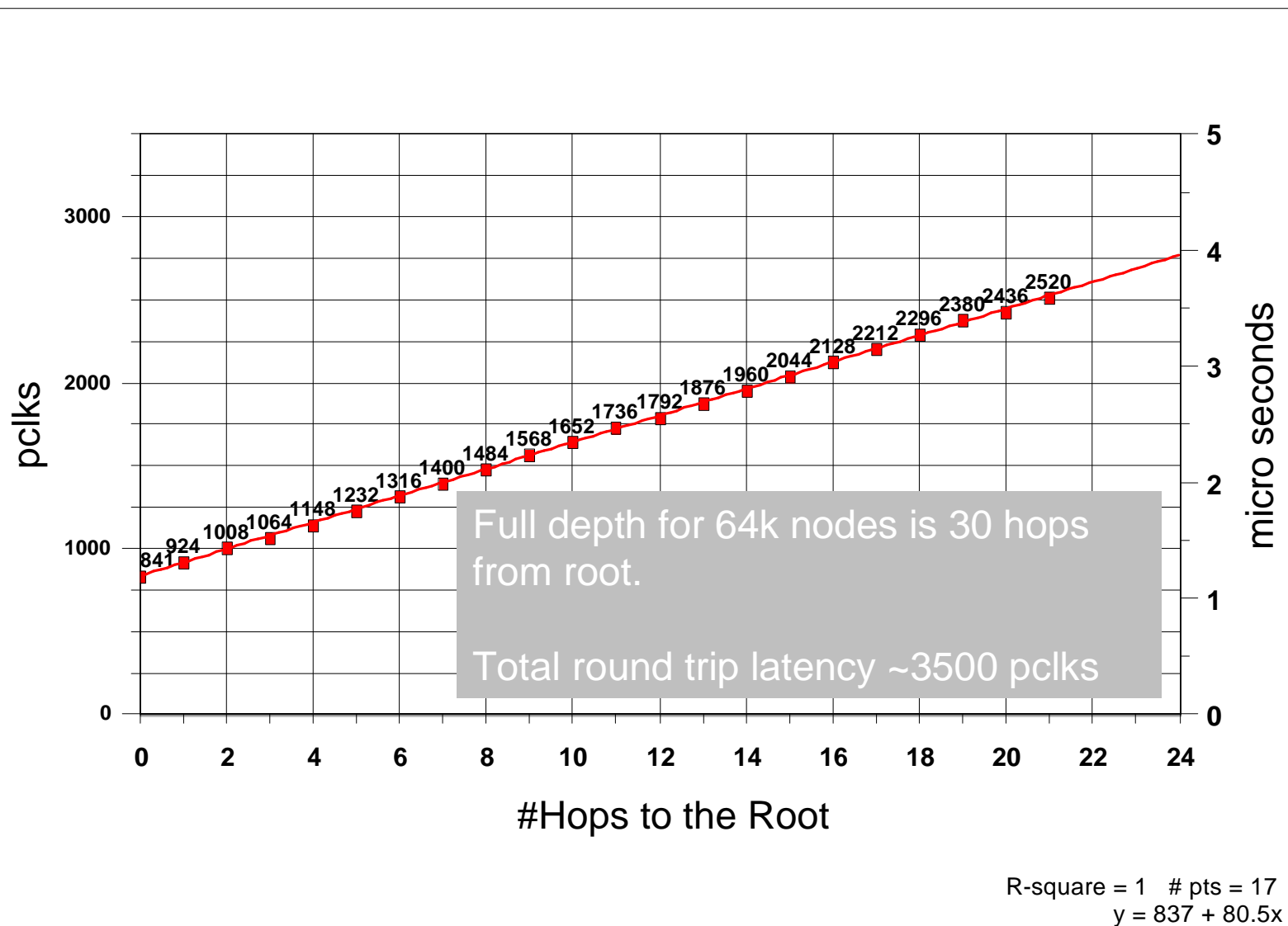


Collective Network

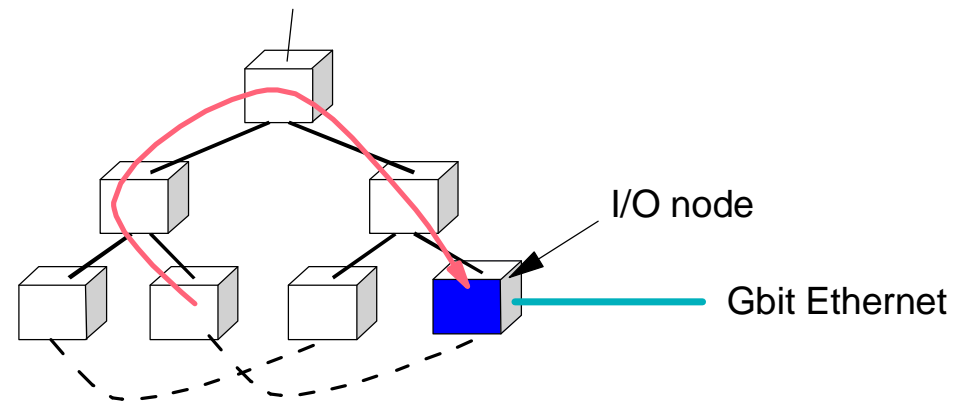


- **High Bandwidth one-to-all**
 - 2.8Gb/s to all 64k nodes
 - 68TB/s aggregate bandwidth
- **Arithmetic operations implemented in tree**
 - Integer/ Floating Point Maximum/Minimum
 - Integer addition/subtract, bitwise logical operations
- **Global latency of less than 2.5usec to top, additional 2.5usec to broadcast to all**
- **Global sum over 64k in less than 2.5 usec (to top of tree)**
- **Used for disk/host funnel in/out of I/O nodes.**
- **Minimal impact on cabling**
- **Partitioned with Torus boundaries**
- **Flexible local routing table**
- **Used as Point-to-point for File I/O and Host communications**

Collective Network: Measured Roundtrip Latency

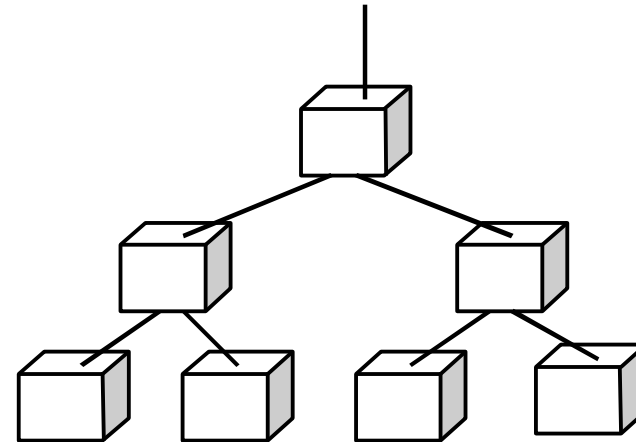


Gb Ethernet Disk/Host I/O Network



- § IO nodes are leaves on collective network.
- § Compute and IO nodes use same ASIC, but:
 - 4 IO node has Ethernet, not torus.
Minimizes IO perturbation on application.
 - 4 Compute node has torus, not ethernet.
Don't want 65536 Gbit Ethernet cables!
- § Configurable ratio of IO to compute = 1:8,16,32,64,128.
- § Application runs on compute nodes, not IO nodes.

Fast Barrier/Interrupt Network

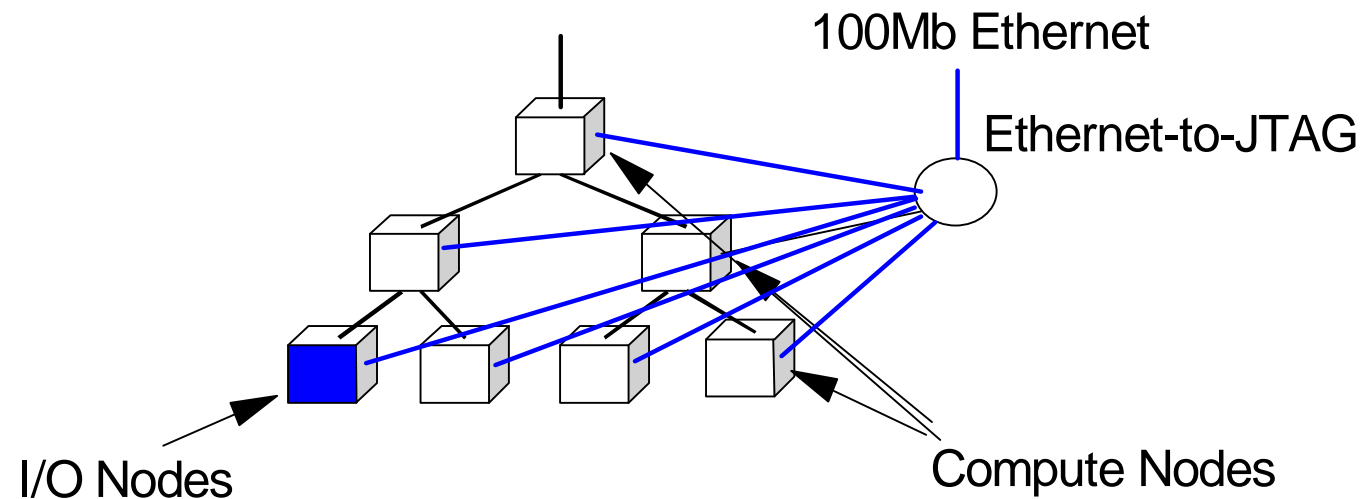


- **Four Independent Barrier or Interrupt Channels**
 - **Independently Configurable as "or" or "and"**
- **Asynchronous Propagation**
 - **Halt operation quickly (current estimate is 1.3usec worst case round trip)**
 - > **3/4 of this delay is time-of-flight.**
- **Sticky bit operation**
 - **Allows global barriers with a single channel.**
- **User Space Accessible**
 - **System selectable**
- **Partitions along same boundaries as Tree, and Torus**
 - **Each user partition contains it's own set of barrier/ interrupt signals**

Control Network

JTAG interface to 100Mb Ethernet

- **direct access to all nodes.**
- **boot, system debug availability.**
- **runtime noninvasive RAS support.**
- **non-invasive access to performance counters**
- **Direct access to shared SRAM in every node**



Control network (continued)

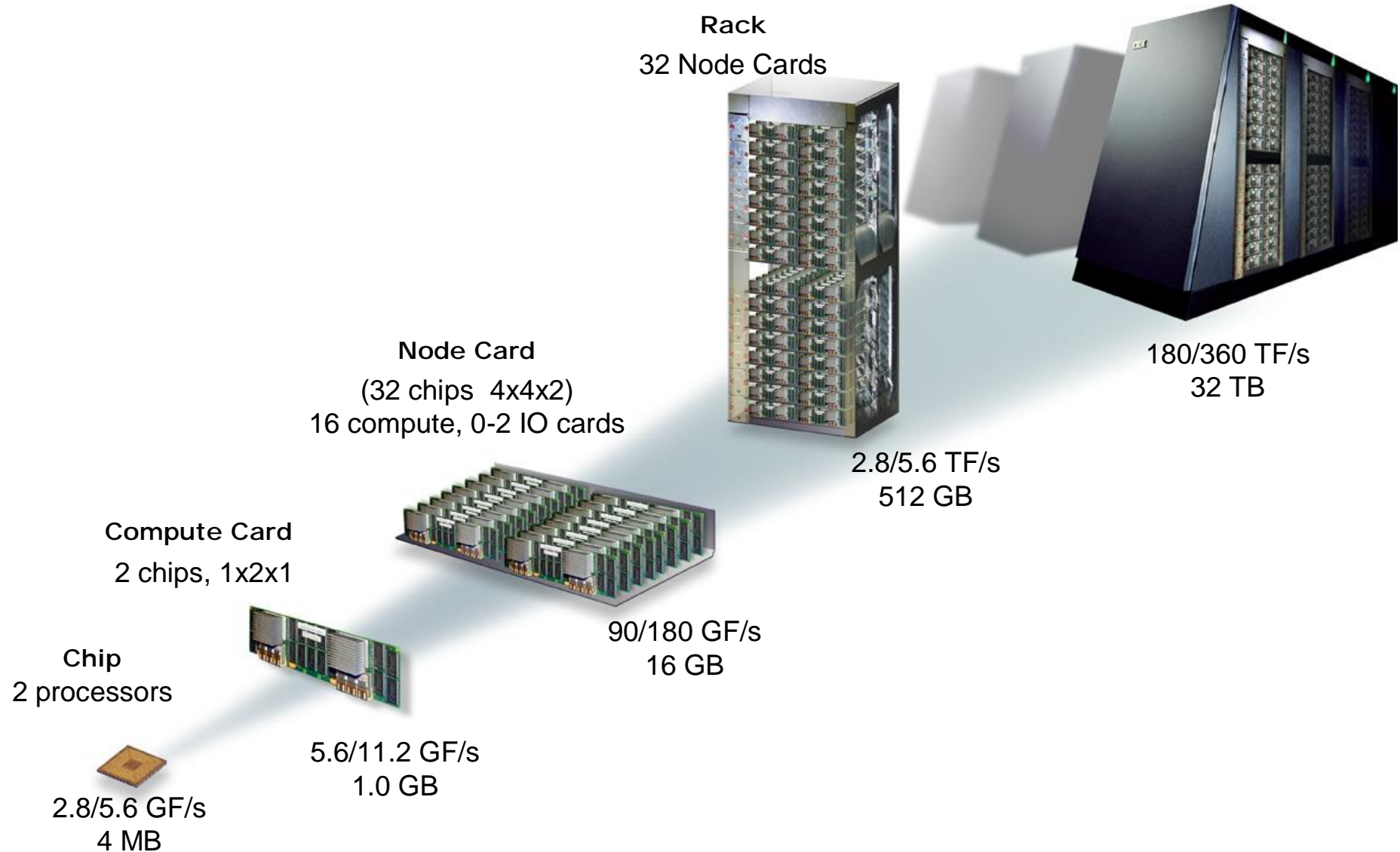
Control, configuration and monitoring:

- § Make all active devices accessible through JTAG, I2C, or other “simple” bus. (Only clock buffers & DRAM are not accessible)

- § FPGA is Ethernet to “JTAG+I2C+...” switch
 - 4 Allows access from anywhere on IBM Intranet
 - 4 Used for control, monitor, and initial system load
 - 4 Rich command set of Ethernet broadcast, multicast, and reliable pt-to-pt messaging allows range of control & speed.
 - 4 **Other than ethernet MAC address, no state in the machine!**

- § Goal is ~1 minute system boot.

Packaging

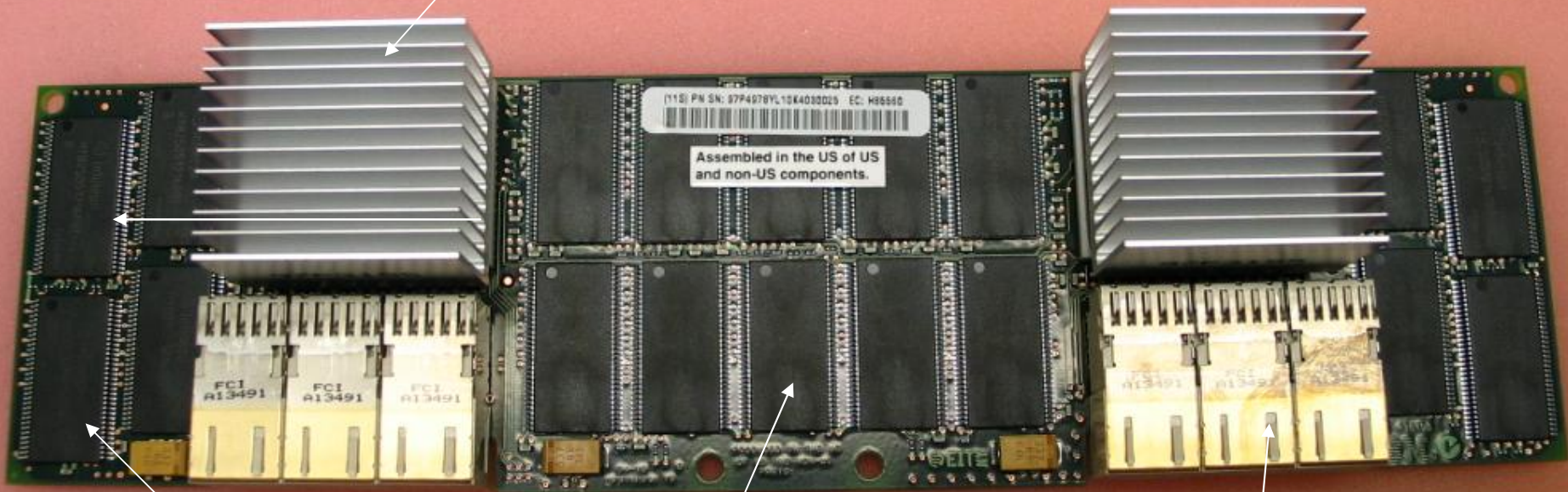


DESY-Hamburg, Feb.21, 2005

Dual Node Compute Card

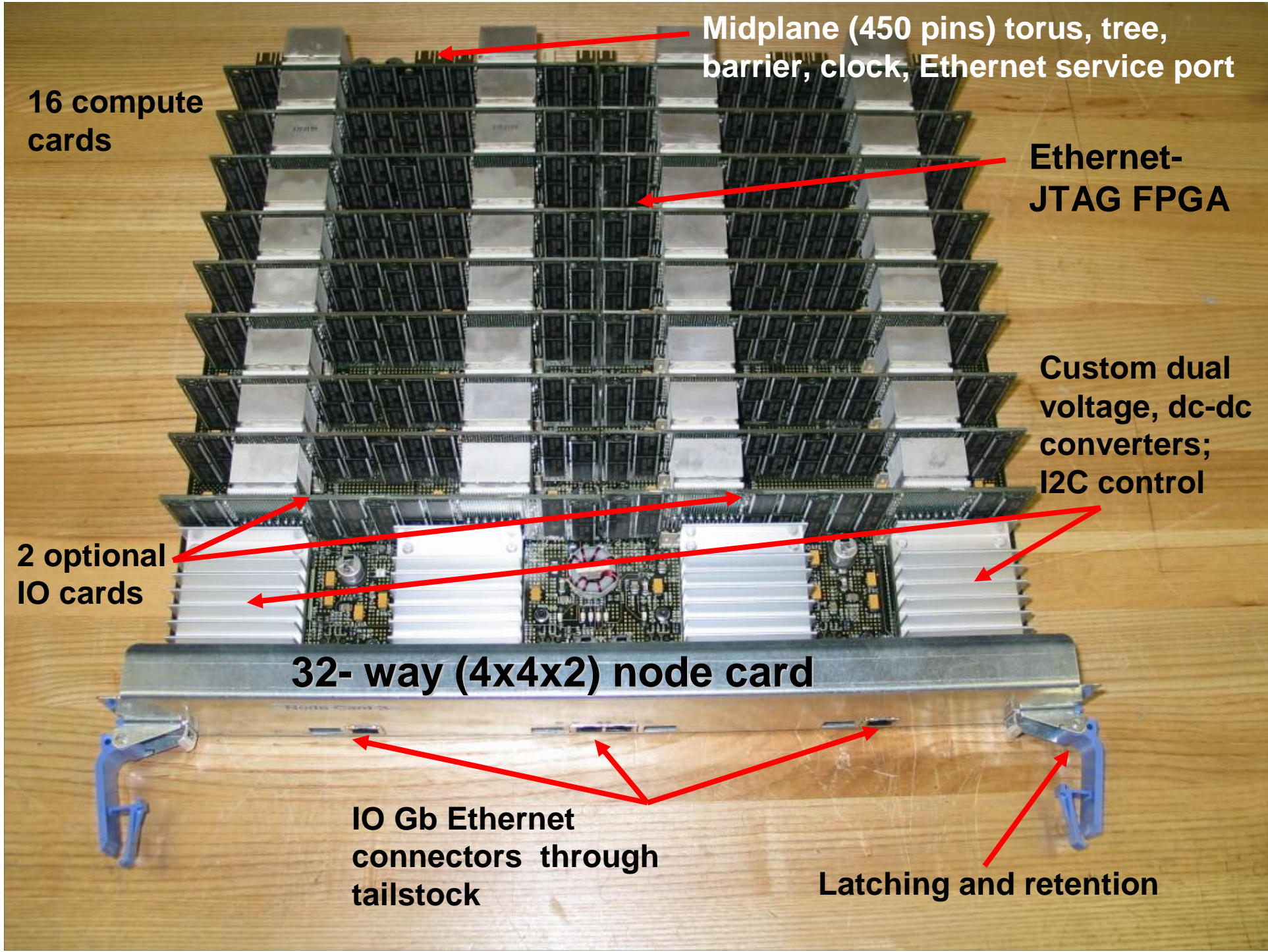
206 mm (8.125") wide, 54mm high (2.125"), 14 layers,
single sided, ground referenced

Heatsinks
designed for 15W



9 x 512 Mb DRAM;
16B interface; no
external termination

Metral 4000 high
speed differential
connector (180 pins)



16 compute cards

Midplane (450 pins) torus, tree, barrier, clock, Ethernet service port

Ethernet-JTAG FPGA

Custom dual voltage, dc-dc converters; I2C control

2 optional IO cards

32- way (4x4x2) node card

IO Gb Ethernet connectors through tailstock

Latching and retention

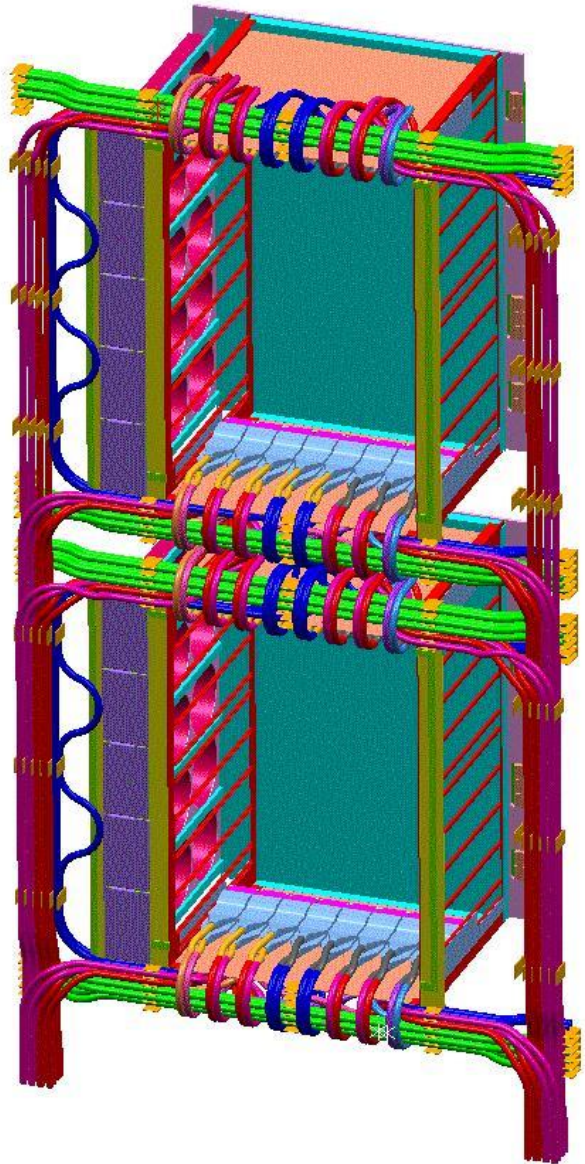
$\frac{1}{4}$ of BG/L midplane (128 nodes)

Compute cards

IO card

dc-dc
converter

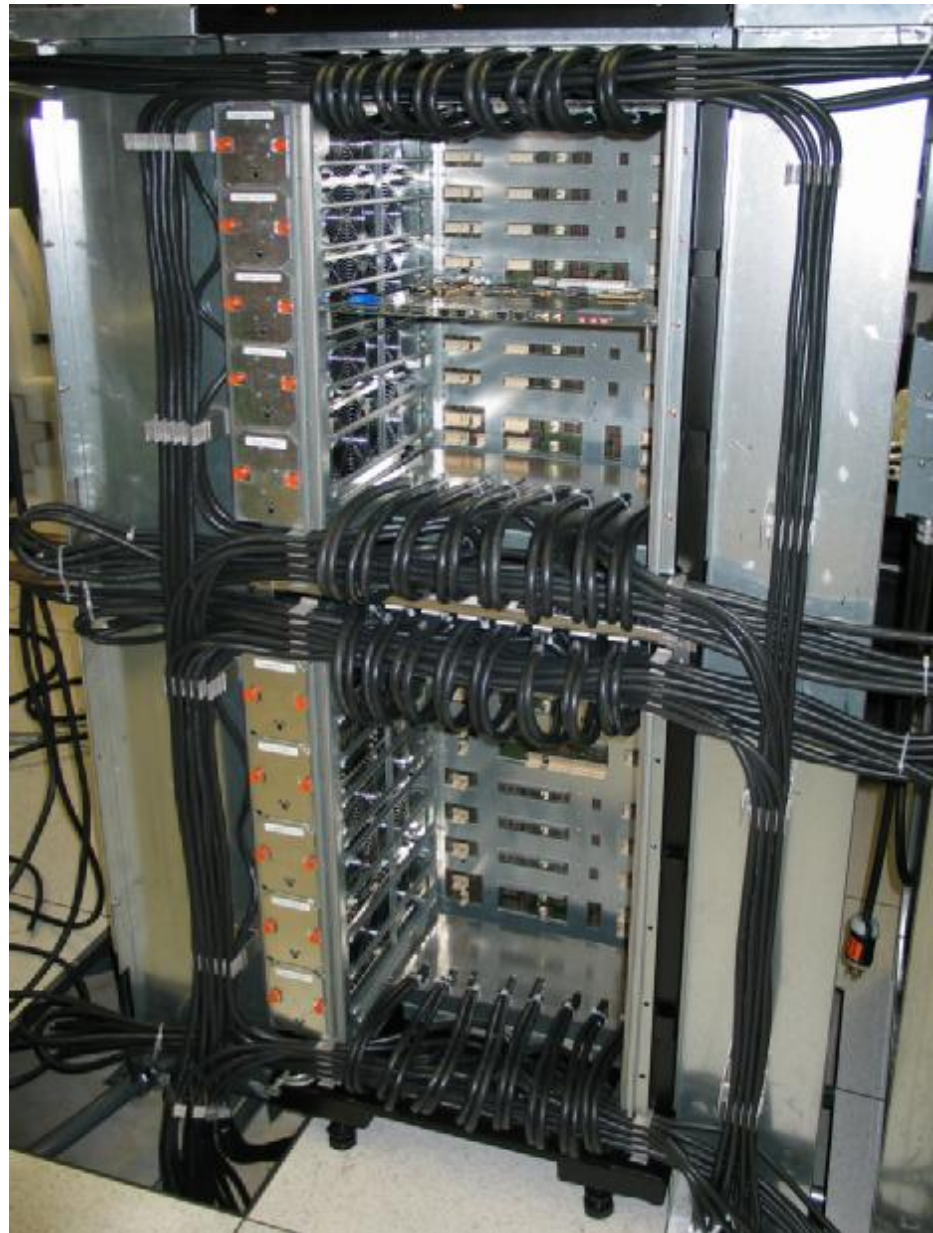
Airflow, cabling & service



X Cables

Y Cables

Z Cables



BG/L link card

Midplane
(540 pins)

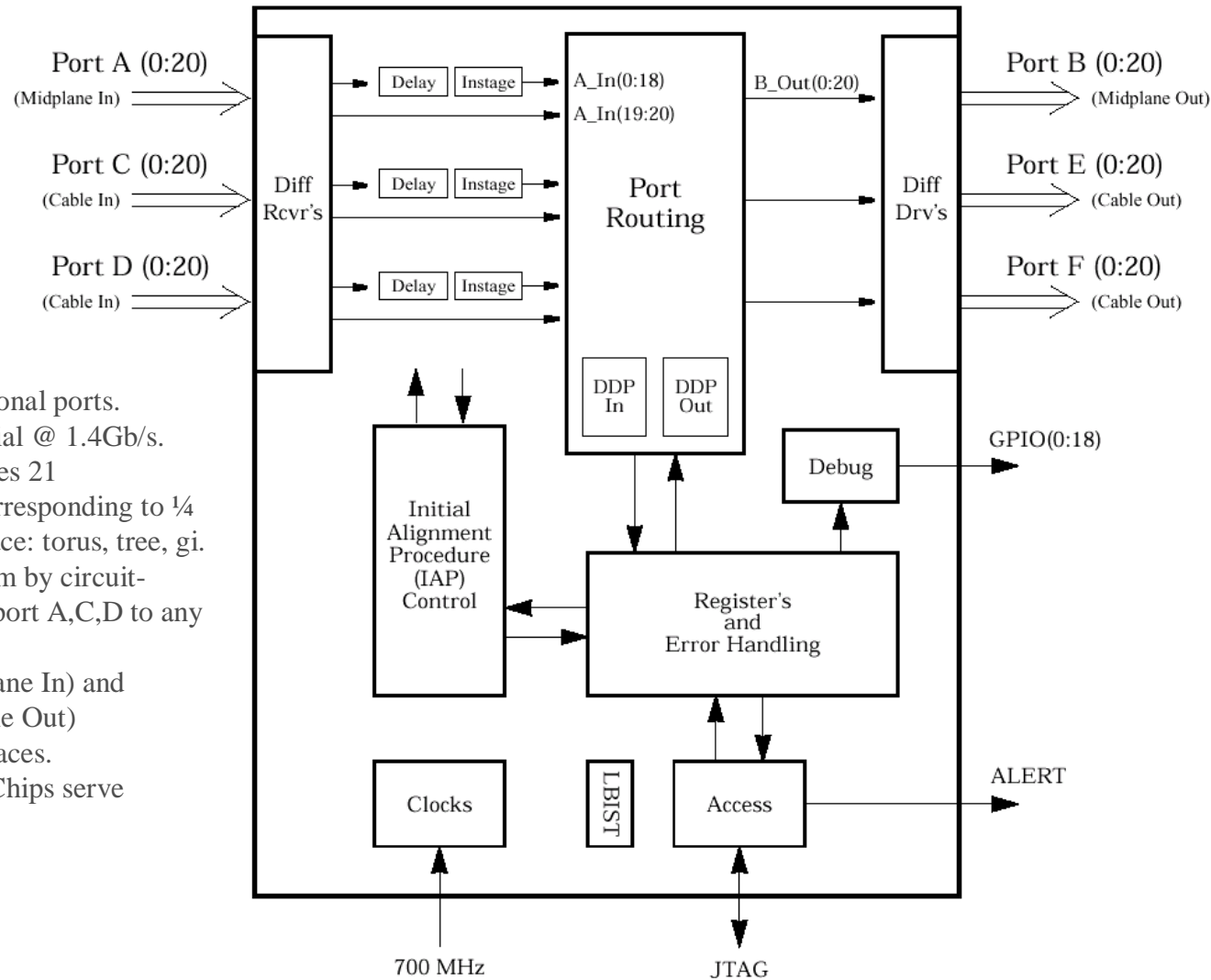
Redundant DC-
DC converters

Ethernet->
JTAG FPGA

Link ASIC
IBM CU-11, 0.13 μm
6.6 mm die size
25 x 32 mm CBGA
474 pins, 312 signal
1.5 Volt, 4W

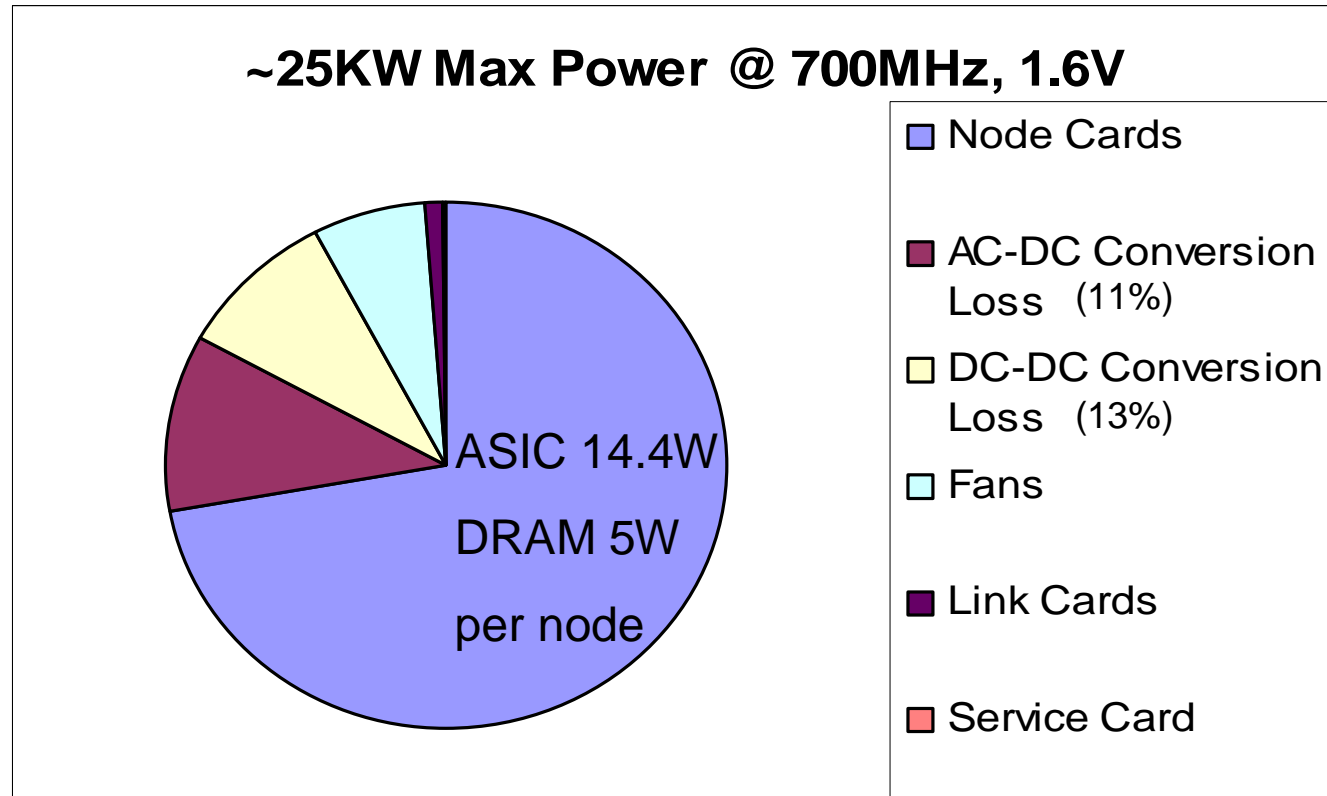
22 differential
pair cables, max
8.5 meter

BlueGene/L Link Chip : Circuit-switch between midplanes for Space-Sharing



- Six uni-directional ports.
- Each differential @ 1.4Gb/s.
- Each port serves 21 differentials, corresponding to $\frac{1}{4}$ of a midplane face: torus, tree, gi.
- Partition system by circuit-switching each port A,C,D to any port B, E, F.
- Port A (Midplane In) and Port B (Midplane Out) serve opposite faces.
- $4 \times 6 = 24$ Link Chips serve each midplane.

BlueGene/L Compute Rack Power



MFW (Peak)

250

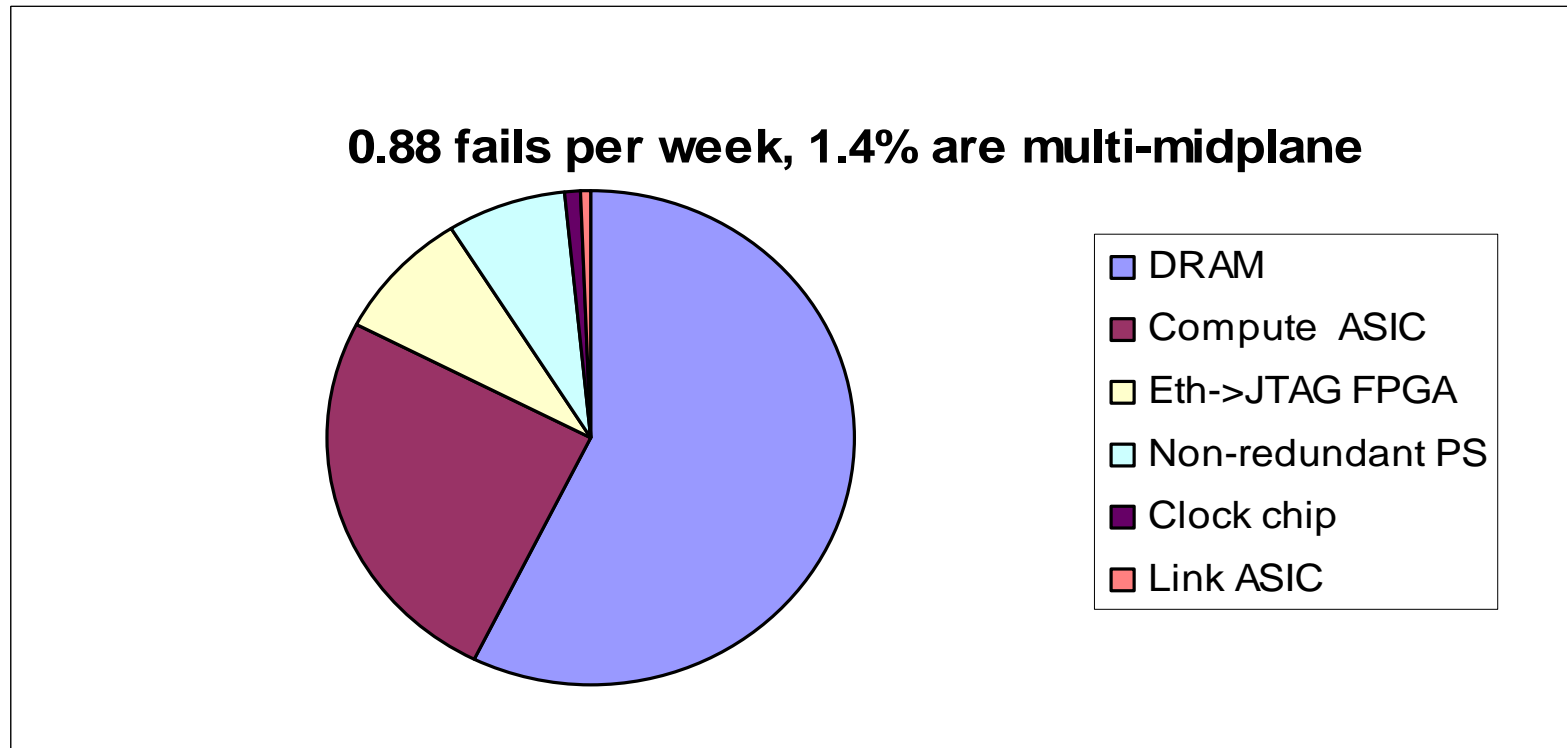
MFW (Sustained-Linpack)

172

Check the Failure Rates

- § Redundant bulk supplies, power converters, fans, DRAM bits, cable bits
- § ECC or parity/retry with sparing on most buses.
- § Extensive data logging (voltage, temp, recoverable errors, ...) for failure forecasting.
- § Uncorrectable errors cause restart from checkpoint after repartitioning (remove the bad midplane).
- § Only fails early in global clock tree, or certain failures of link cards, cause multi-midplane fails.

Predicted 64Ki node BG/L hard failure rates



Software Design Overview

§ Familiar software development environment and programming models

§ **Scalability to $O(100,000)$ processors – through Simplicity**

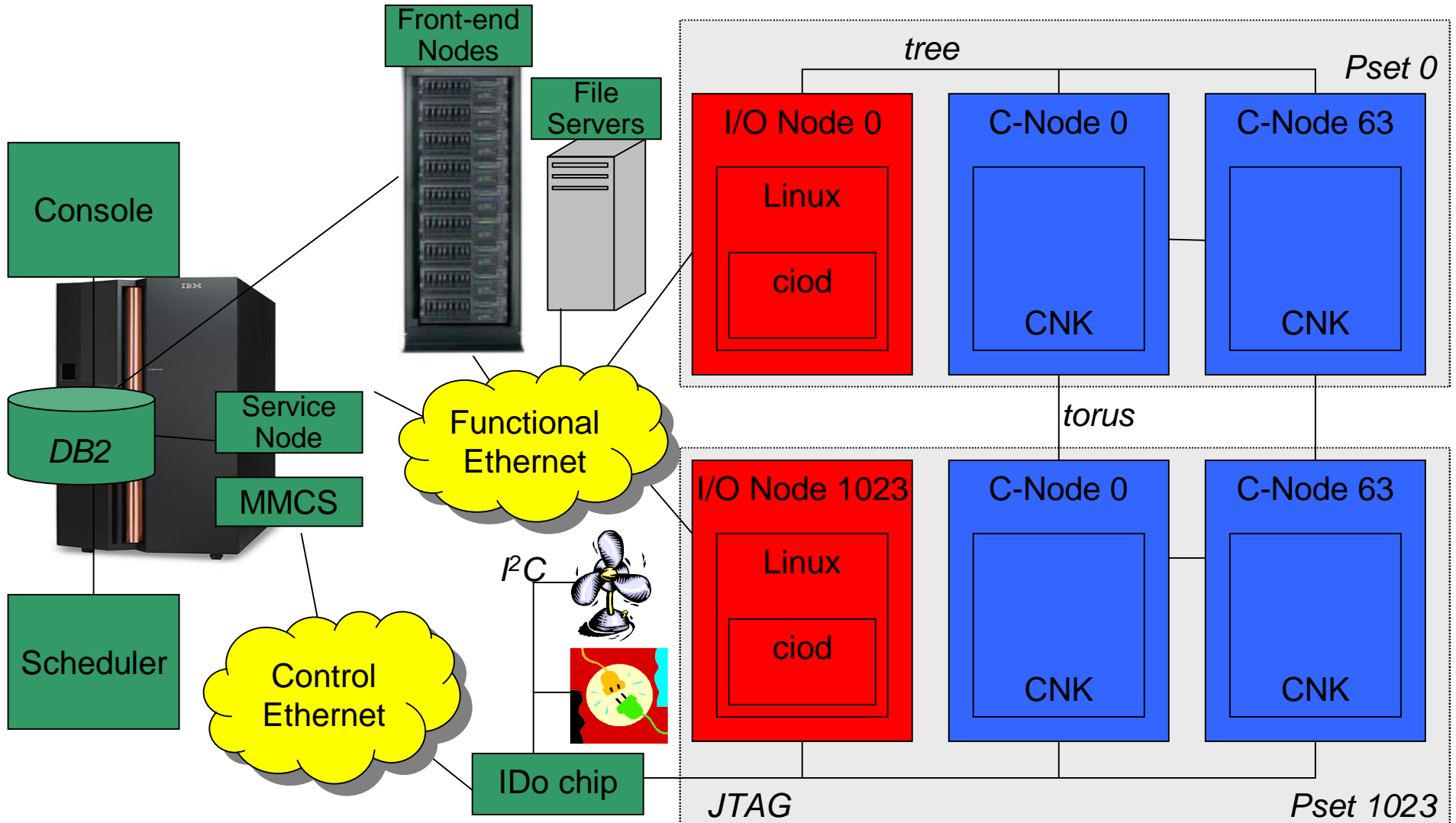
4 Performance

- Strictly space sharing - one job (user) per electrical partition of machine, one process per compute node
 - Dedicated processor for each application level thread
 - Guaranteed, deterministic execution
 - Physical memory directly mapped to application address space – no TLB misses, page faults
 - Efficient, user mode access to communication networks
 - No protection necessary because of strict space sharing
- Multi-tier hierarchical organization – system services (I/O, process control) offloaded to IO nodes, control and monitoring offloaded to service node
 - No daemons interfering with application execution
 - System manageable as a cluster of IO nodes

4 Reliability, Availability, Serviceability

- Reduce software errors - simplicity of software, extensive run time checking option
- Ability to detect, isolate, possibly predict failures

Blue Gene/L System Software Architecture



BG/L – Familiar software environment

§ Fortran, C, C++ with MPI

- 4 Full language support
- 4 Automatic SIMD FPU exploitation

§ Linux development environment

- 4 Cross-compilers and other cross-tools execute on Linux front-end nodes
- 4 Users interact with system from front-end nodes

§ Tools – support for debuggers, hardware performance monitors, trace based visualization

§ POSIX system calls – compute processes “feel like” they are executing on a Linux environment (restrictions)

Result: MPI applications port quickly to BG/L

Applications I. The Increasing Value of Simulations

§ Supercomputer performance continues to improve.

This allows:

- 4 Bigger problems.
E.g. More atoms in simulation of material.
- 4 Finer resolution.
E.g. More cells in simulation of earth climate.
- 4 More time steps.
E.g. Complete protein fold requires 10^6 or far more molecular timesteps.

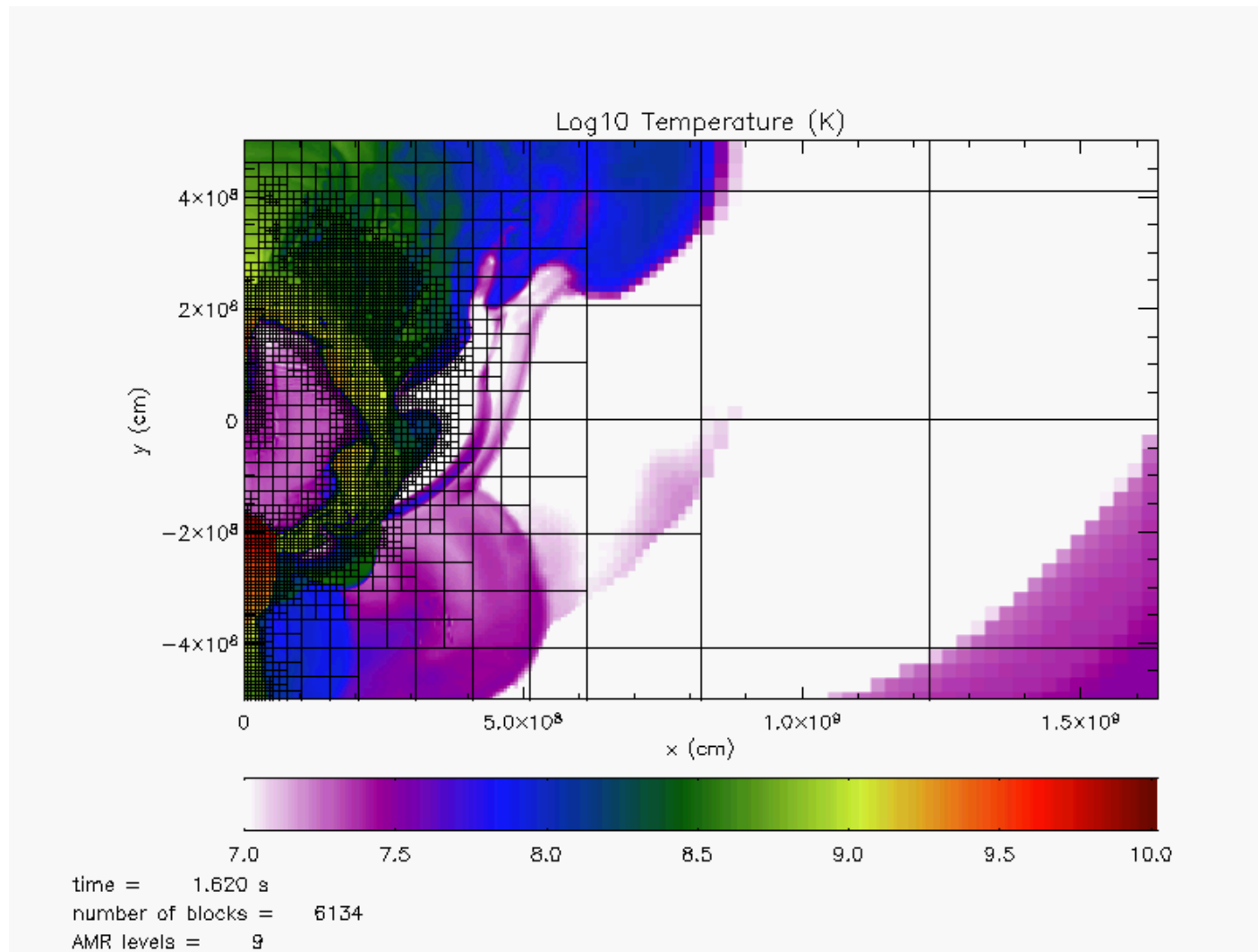
§ In many application areas, performance now allows first-principle simulations large, fine and/or long enough to be compared against experimental results.

Simulation examples:

- 4 Enough atoms to see grains in solidification of metals.
- 4 Enough resolution to see hurricane frequency in climate studies.
- 4 Enough timesteps to fold a protein.

FLASH

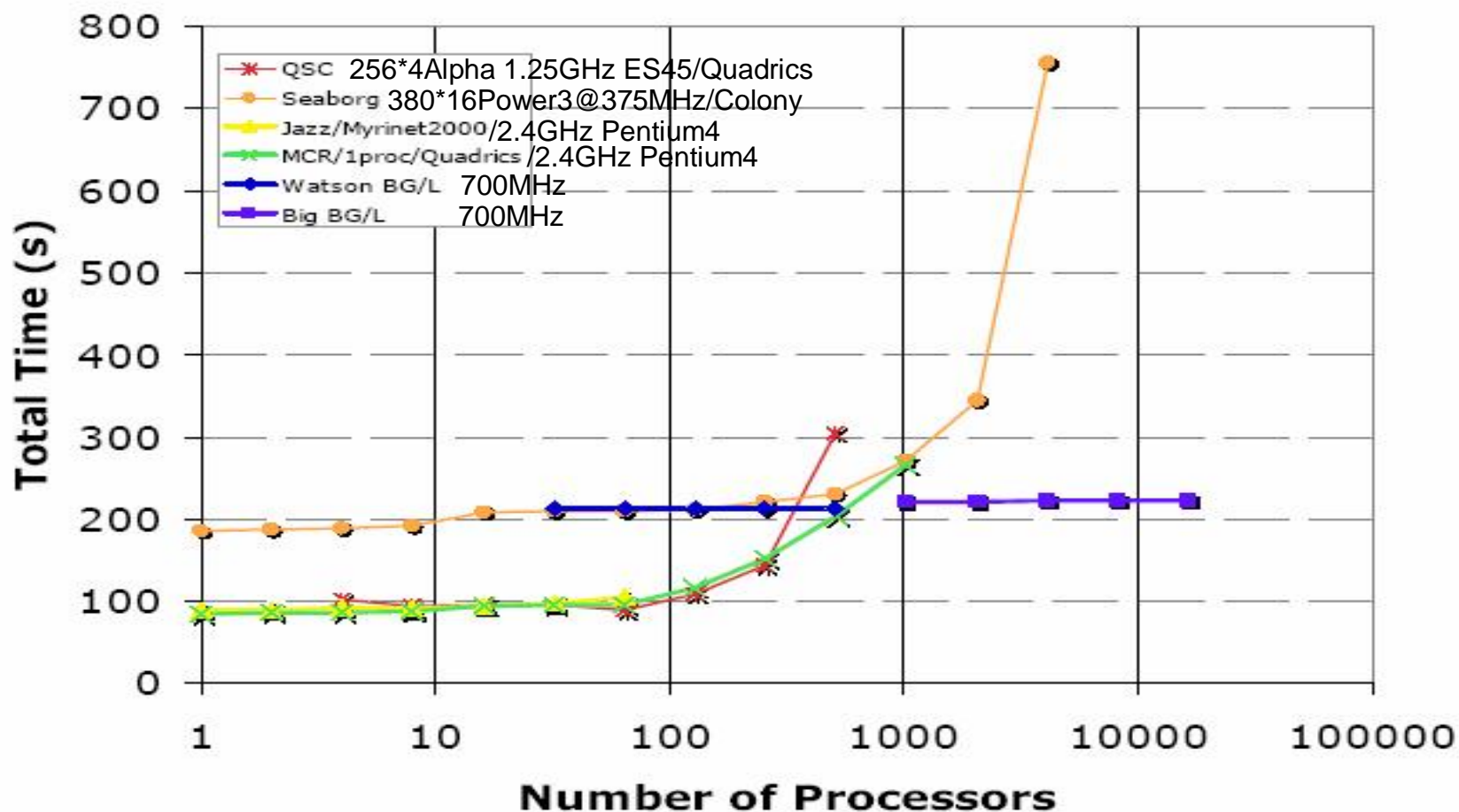
- § University of Chicago and Argonne National Laboratory,
 - 4 Katherine Riley, Andrew Siegel
 - 4 IBM: Bob Walkup, Jim Sexton
- § parallel adaptive-mesh multi-physics simulation code designed to solve nuclear astrophysical problems related to exploding stars.
- § solves the Euler equations for compressible flow and the Poisson equation for self-gravity.
- § Simulates a Type-1a supernova through stages:
 - 4 deflagration initiated near the center of the white dwarf star
 - 4 initial spherical flame front buoyantly rises
 - 4 develops a Rayleigh-Taylor instability as it expands



FLASH – Astrophysics of Exploding Stars

§ Argonne/DOE project: flash.uchicago.edu. Adaptive Mesh.

§ Weak Scaling – Fixed problem size per processor.



HOMME

§ National Center for Atmospheric Research Program

- 4 John Dennis, Rich Loft, Amik St-Cyr, Steve Thomas, Henry Tufo, Theron Voran (Boulder)
- 4 John Clyne, Joey Mendoza (NCAR)
- 4 Gyan Bhanot, Jim Edwards, James Sexton, Bob Walkup, Andii Wyszogrodzki (IBM)

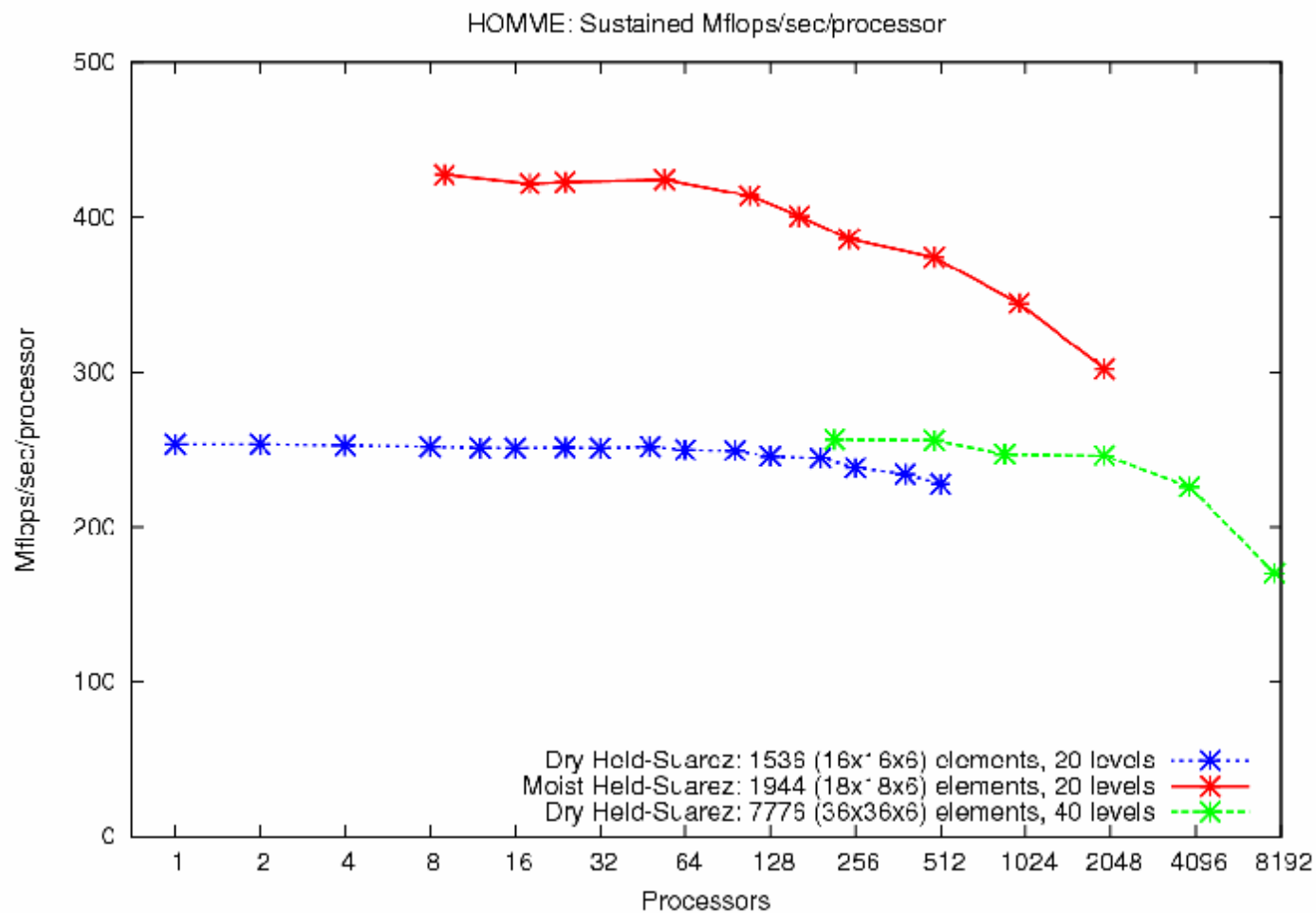
§ Description:

- 4 The moist Held-Suarez test case extends the standard (dry) Held-Suarez test of the hydrostatic primitive equations by introducing a moisture tracer and simplified physics. It is the next logical test for a dynamical core beyond dry dynamics.
- 4 Moisture is injected into the system at a constant rate from the surface according to a prescribed zonal profile, is advected as a passive tracer by the model, and precipitated from the system when the saturation point is exceeded.

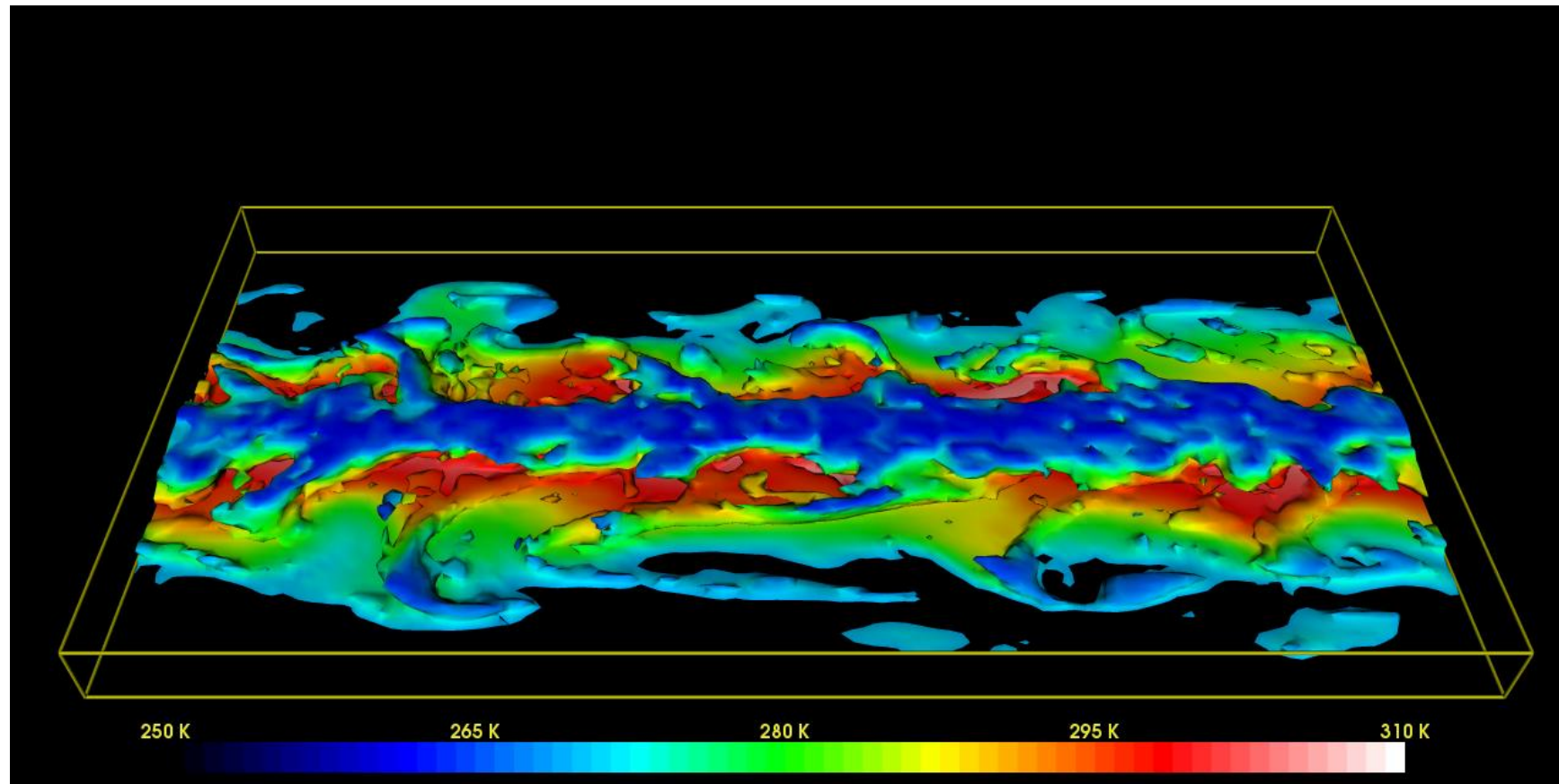
HOMME: some details

- § The model is written in F90 and has three components:
 - 4 dynamics, physics and a physics/dynamics coupler.
- § The dynamics has been run on the BG/L systems at Watson and Rochester on up to 7776 processors using one processor per node and only one of the floating-point pipelines.
- § The peak performance expected from a Blue Gene processor for the runs is then 1.4 Gflops/s.
- § The average sustained performance in the scaling region for the Dry Held-Suarez code is ~200-250 MF/s/processor (14-18% of peak) out to 7776 processors,
- § The Moist Held-Suarez code it is ~ 300-400 MF/s/processor (21-29% of peak), out to 1944 processors.

HOMME: Strong Scaling

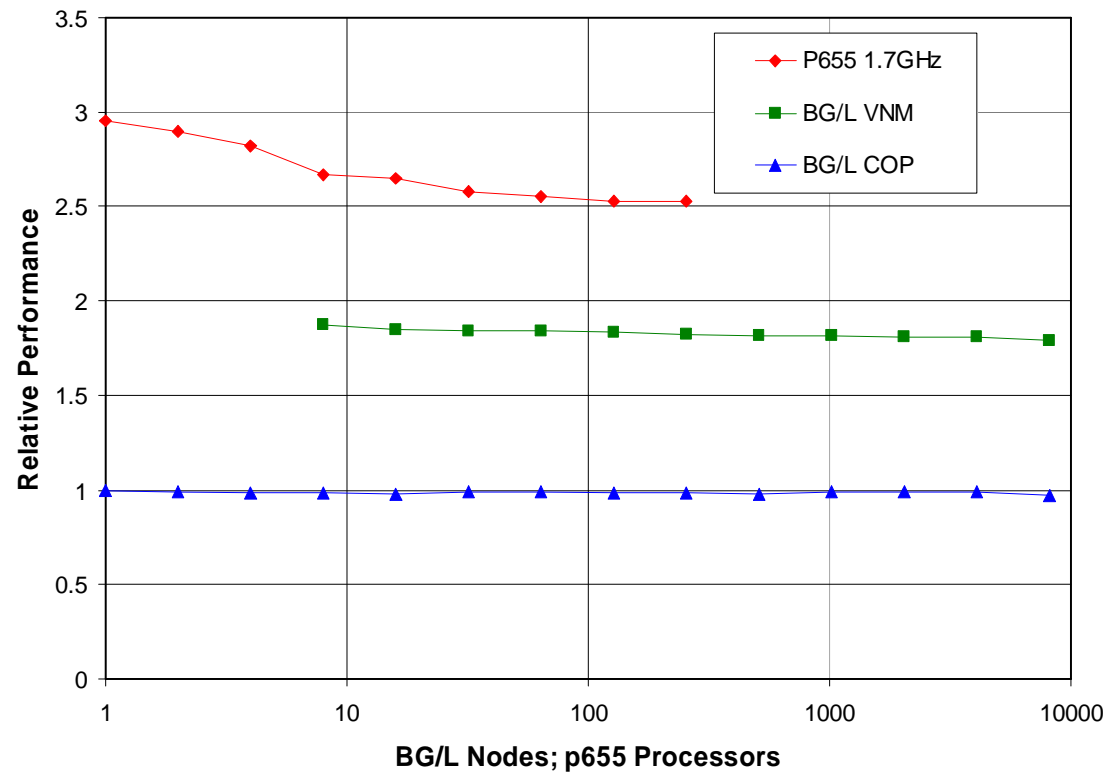


Homme: visualisation

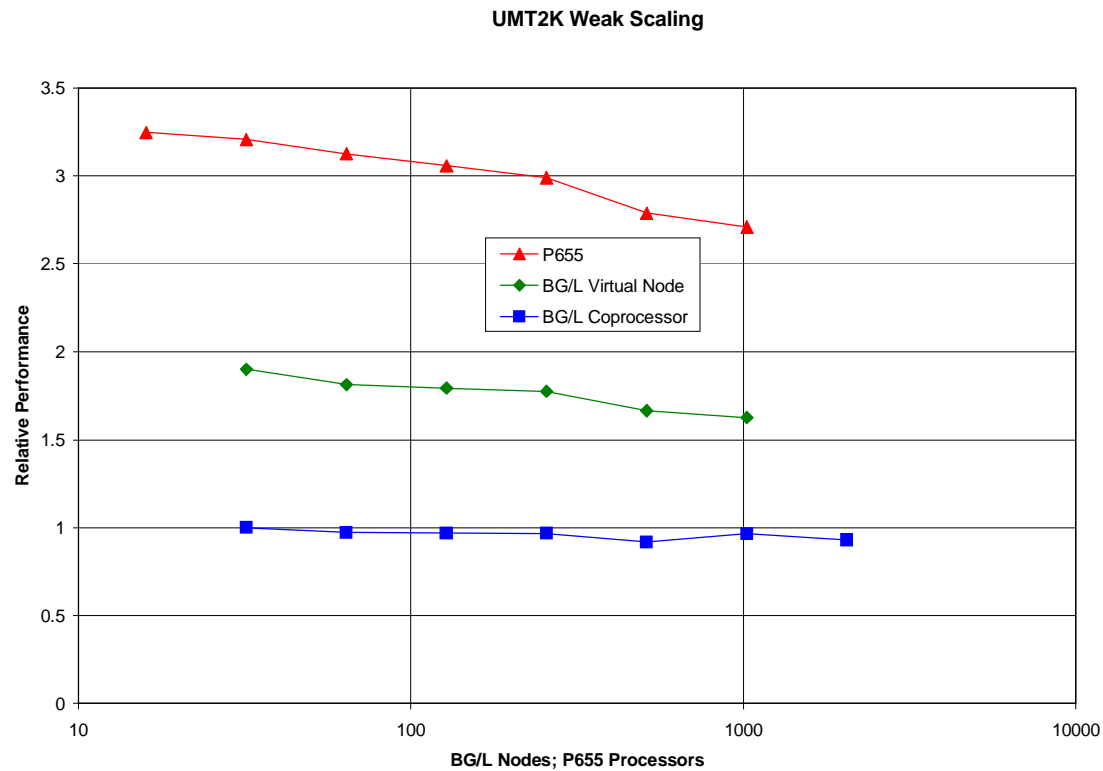


sPPM: ASCI 3D gas dynamics code

sPPM Scaling (128**3, real*8)

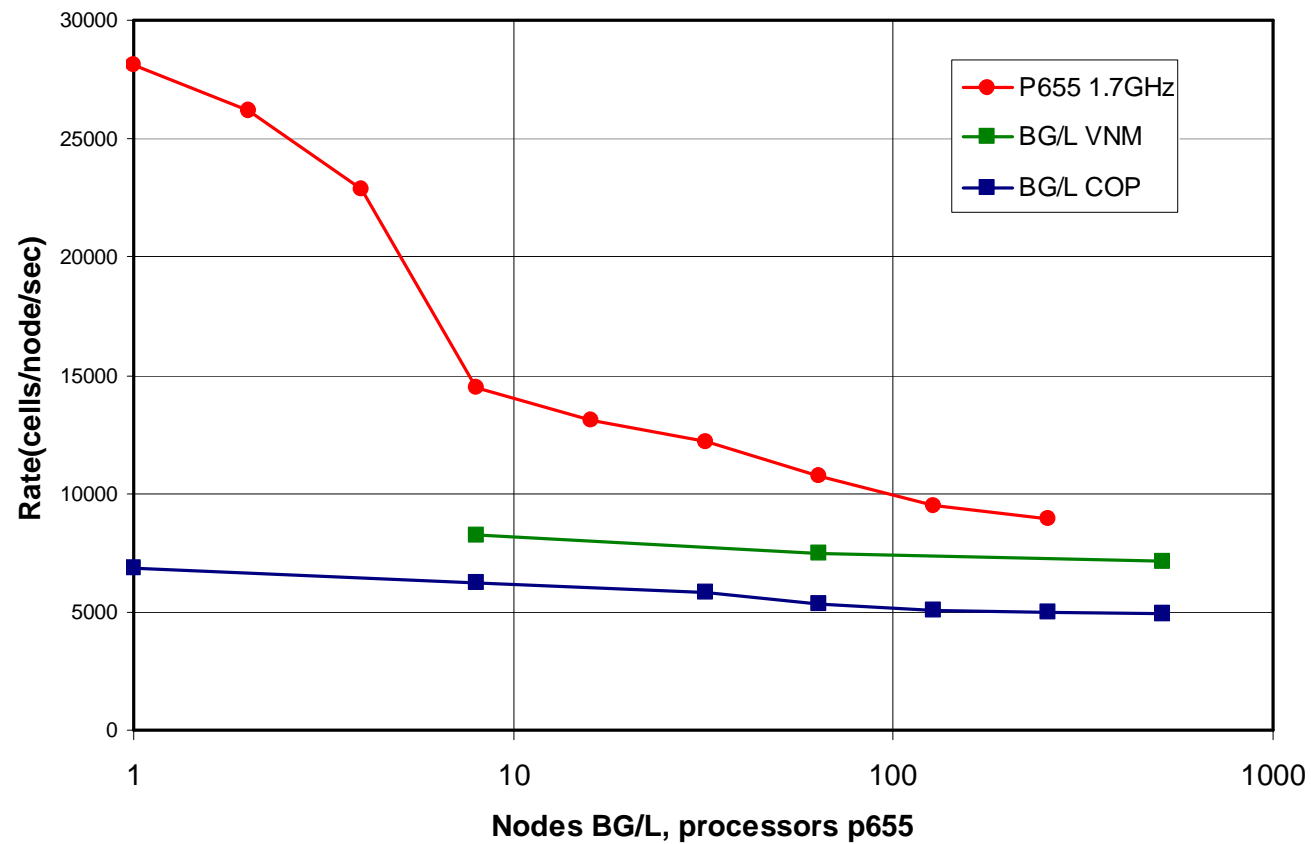


UMT2K: Photon Transport



SAGE: ASCI Hydrodynamics code

SAGE Scaling (timing_h, 32K cells/node)



Applications II. For On-line Data Processing

ASTRON's LOFAR is a very large distributed radio telescope

- § 13000 small antennas.
 - 4 In 100 stations.
 - 4 Across Netherlands, Germany
- § No physical focus of antennas, so raw data views entire sky.
- § Use on-line data processing to focus on object(s) of interest.
 - 4 Example:
Can change focus instantly.
So can buffer raw data and trigger on event.
- § 6 BG/L racks at center of on-line processing.
 - 4 Sinking 768 Gbit ethernet lines.
- § lofar.org



SUMMARY: BG/L in Numbers

- § Two 700MHz PowerPC440 per node.
- § 350MHz L2, L3, DDR.
- § 16Byte interface L1|L2, 32B L2|L3, 16B L3|DDR.
- § $1024 = 16*8*8$ compute nodes/rack is 23kW/rack.
- § 5.6GFlops/node = $2\text{PPC440} * 700\text{MHz} * 2\text{FMA/cycle} * 2\text{Flops/FMA}$.
- § 5.6TFlops/rack.
- § 512MB/node DDR memory
- § 512GB/rack
- § 175MB/s = 1.4Gb/sec torus link = $700\text{MHz} * 2\text{bits/cycle}$.
- § 350MB/s = tree link

SUMMARY: The current #1 Supercomputer

- § 70.7TF on Linpack Benchmark is 77% of 90.8TF peak.
- § 16 BG/L racks installed at LLNL.
- § 16384 nodes.
- § 32768 PowerPC440 processors.
- § 8 TB memory.
- § 2m² of compute ASIC silicon!

Before end of 2005:

- § Increase LLNL to 64 racks.
- § Install ~10 other customers:
SDSC, Edinburgh, AIST, ...

THE END

For more details:

Special Issue: Blue Gene/L,
IBM J. Res. & Dev. Vol.49
No.2/3 March/May 2005.