Blue Gene/L Architecture

Burkhard Steinmacher-Burow
IBM Watson / Böblingen

November 2, 2004, DESY-Zeuthen

Outline

- Architecture Motivation

Given the motivation,
the architecture should seem natural and obvious.

- Architecture Overview

What is the Blue Gene/L Project?

- A 512- to 65536-node highly-integrated supercomputer based on system-on-a-chip technology: Node ASIC. Link ASIC.
- Strategic partnership with LLNL and other high performance computing centers and researchers:
  - Focus on numerically intensive scientific problems.
  - Validation and optimization of architecture based on real applications.
  - Grand challenge science stresses networks, memory and processing power.
  - Partners accustomed to "new architectures" and work hard to adapt to constraints.
  - Partners assist us in the investigation of the reach of this machine.

BG/L for Capability Computing

- December 1999: IBM Research announced a 5 year, $100M US, effort to build a petaflop/s scale supercomputer to attack science problems such as protein folding.

  Goals:

  Advance the state of the art of scientific simulation.

  Advance the state of the art in computer design and software for capability and capacity markets.

- November 2001: Announced Research partnership with Lawrence Livermore National Laboratory (LLNL).
  November 2002: Announced planned acquisition of a BG/L machine by LLNL as part of the ASCI Purple contract.

- June 2003: First DD1 chips completed.

- November 2003: BG/L Half rack DD1 prototype (512 nodes at 500 MHz) ranked #73 on 22nd Top500 List announced at SC2003 (1.435 TFlops/s).

  32 node system folding proteins live on the demo floor at SC2003
- March 3, 2004: Full rack DD1 (1024 nodes at 500 MHz) running Linpack at 2.807 TFlops/s. This would displace #23 on 22nd Top500 list.
- March 26, 2004: Two racks DD1 (2048 nodes at 500 MHz) running Linpack at 5.613 TFlops/s. This would displace #12 on 22nd Top500 list.
- May 11, 2004: Four racks DD1 (4096 nodes at 500 MHz) running Linpack at 11.68 TFlops/s. This would displace #3 on 22nd Top500 list.

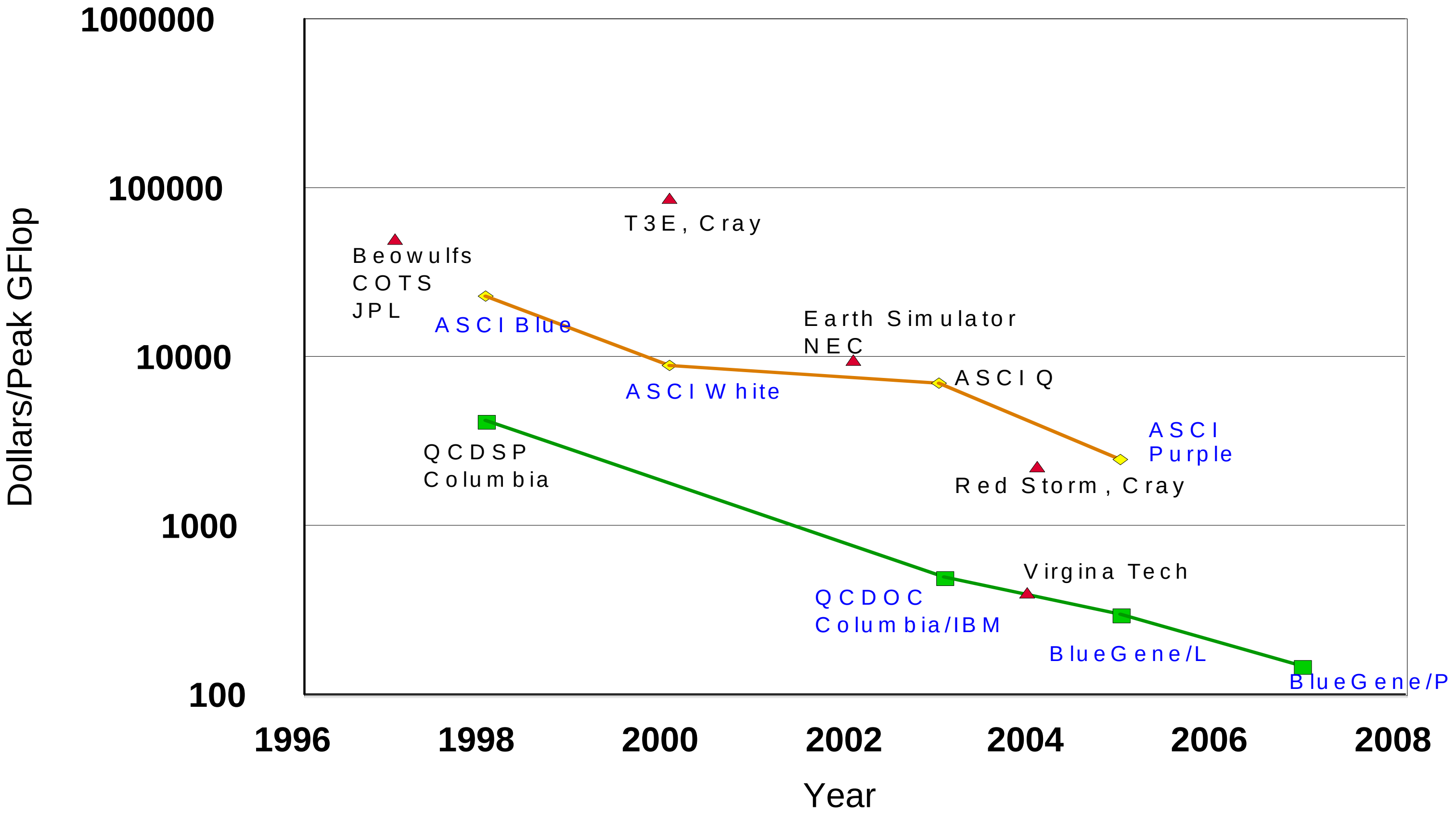
- February 2, 2004: Second pass BG/L chips delivered to Research.

  March 5, 2004 DD2 fully functional running at 700 MHz (design targets).
- June 2, 2004: 2 racks DD2 (1024 nodes at 700 MHz) running Linpack at 8.655 TFlops/s. This would displace #5 on 22nd Top500 list.

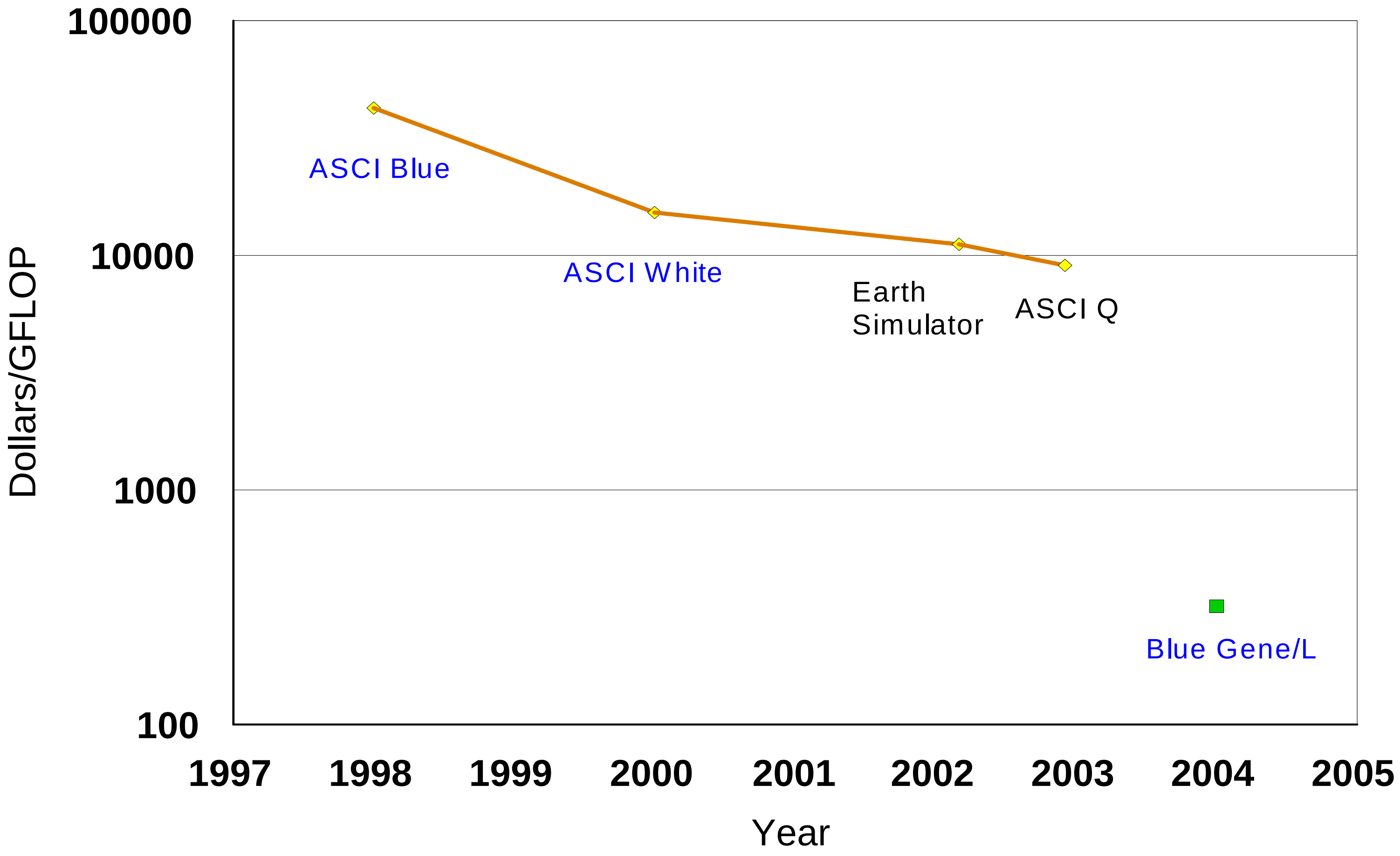
Blue Gene/L Project Motivations

- Improve Cost/Performance (Total Cost/ Time to Solution).
  - Standard familiar programming model.
- Focus on applications that can scale to large node counts.
  - Growing number of such applications.
  - Optimal design point is very different from standard capability-approach based on high-end superscaler nodes.
- Complexity and power drive many costs.
  - Can significantly reduce node-complexity and power by utilizing SOC techniques.
  - Node simplicity is critical, enough system complexity due to number of nodes.
  - Focus on low power.
  - Single chip node: reduces complexity, low power enables high density.
- Close attention to RAS (reliability, availability and serviceability).

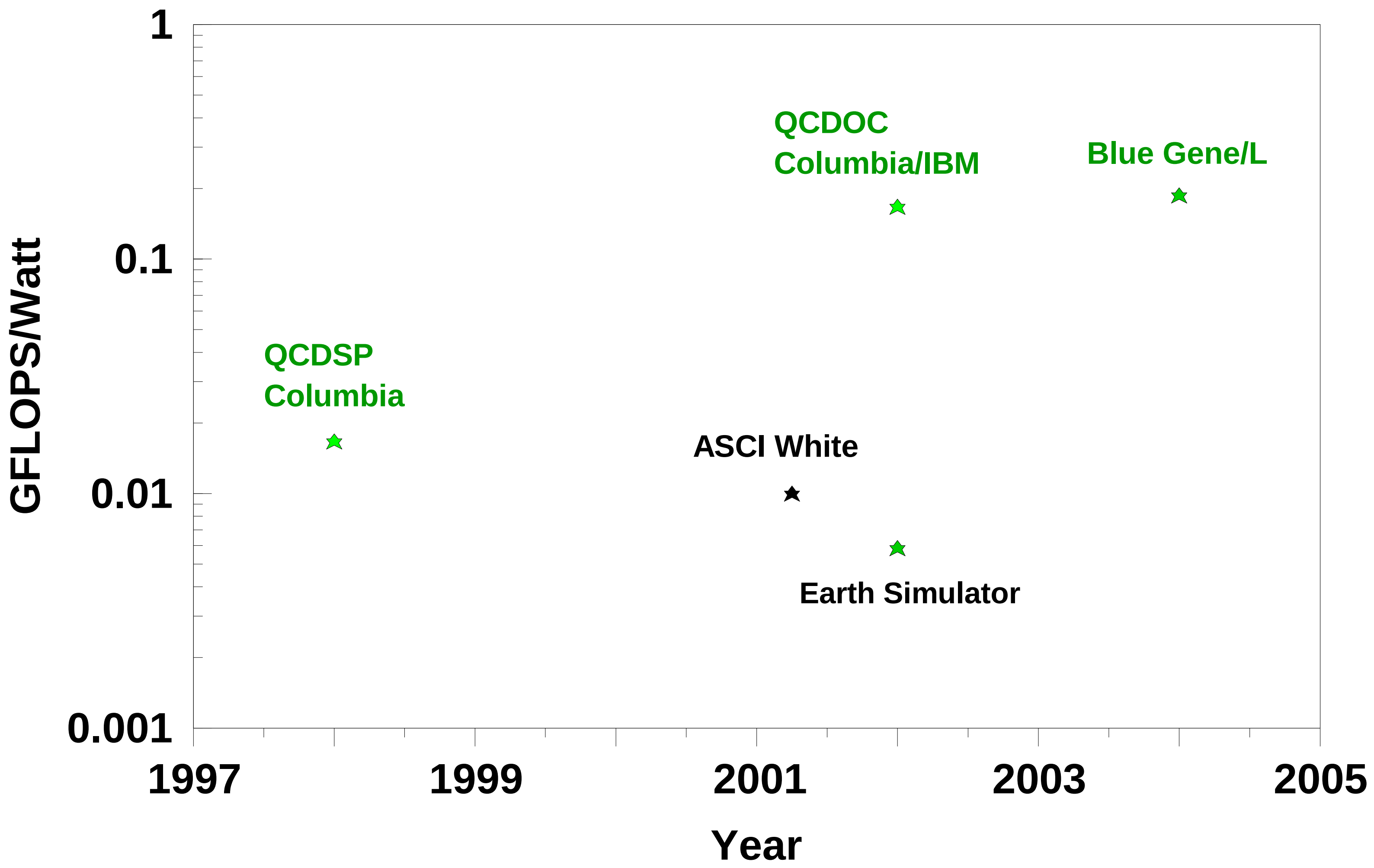
Supercomputer Price/Peak Performance



LINPACK Cost/Performance of Supercomputers



Supercomputer Power Efficiencies



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

- Within node:
  - Low latency, high bandwidth memory system.
  - Strong floating point performance: 2 FMA/cycle.
- Across nodes:
  - Low latency, high bandwidth networks.
- Many nodes:
  - Low power/node.
  - Low cost/node.
  - RAS (reliability, availability and serviceability).
- Familiar SW API:
  - C, C++, Fortan, MPI, POSIX subset, …

Reach is beyond this view. E.g. LOFAR@ASTRON.

Some Performance Results

- DGEMM:
  - 92.3% of dual core peak on 1 node
  - Observed performance at 500 MHz: 3.7 GFlops/s
  - Projected performance at 700 MHz: 5.2 GFlops/s (tested in lab up to 650 MHz)
- LINPACK:
  - 65.2% of peak on 4096 DD1 nodes (11.68 TFlops at 500MHz on 5/11/04)
  - 79.1% of peak on 1024 DD2 nodes (4,535 GFlops/s at 700 MHz on 5/12/04)
  - 75.5% of peak on 2048 DD2 nodes (8,655 GFlops/s at 700 MHz on 6/02/04)
- sPPM, UMT2000:
  - Single processor performance roughly on par with POWER3 at 375 MHz
  - Tested on up to 128 nodes (also NAS Parallel Benchmarks)
- FFT:
  - Up to 508 MFlops on single processor at 444 MHz (TU Vienna)
  - Pseudo-ops performance (5N log N) @ 700 MHz of 1300 Mflops (65% of peak)
- STREAM – impressive results even at 444 MHz:
  - Tuned: Copy: 2.4 GB/s, Scale: 2.1 GB/s, Add: 1.8 GB/s, Triad: 1.9 GB/s
  - Standard: Copy: 1.2 GB/s, Scale: 1.1 GB/s, Add: 1.2 GB/s, Triad: 1.2 GB/s
  - At 700 MHz: Would beat STREAM numbers for most high end microprocessors
- MPI:
  - Latency – < 4000 cycles (5.5 μs at 700 MHz)
  - Bandwidth – full link bandwidth demonstrated on up to 6 links

Specialized means Less General

BG/L leans towards MPI Co-Processor	BG/L leans away from General Purpose Computer
Space-shared nodes.	Time-shared nodes.
Use only real memory.	Virtual memory to disk.
No asynchronous OS activities.	OS services.
Distributed memory.	Shared memory.
No internal state between applications. [Helps performance and functional reproducibility.]	Built-in filesystem.
Reguires General Purpose Computer as Host.	

A Brief History

QCDSP (600GF based on Texas Instruments DSP C31)

-Gordon Bell Prize for Most Cost Effective Supercomputer in '98
-Columbia University Designed and Built
-Optimized for Quantum Chromodynamics (QCD)
-12,000 50MF Processors
-Commodity 2MB DRAM

- QCDOC (20TF based on IBM System-on-a-Chip)

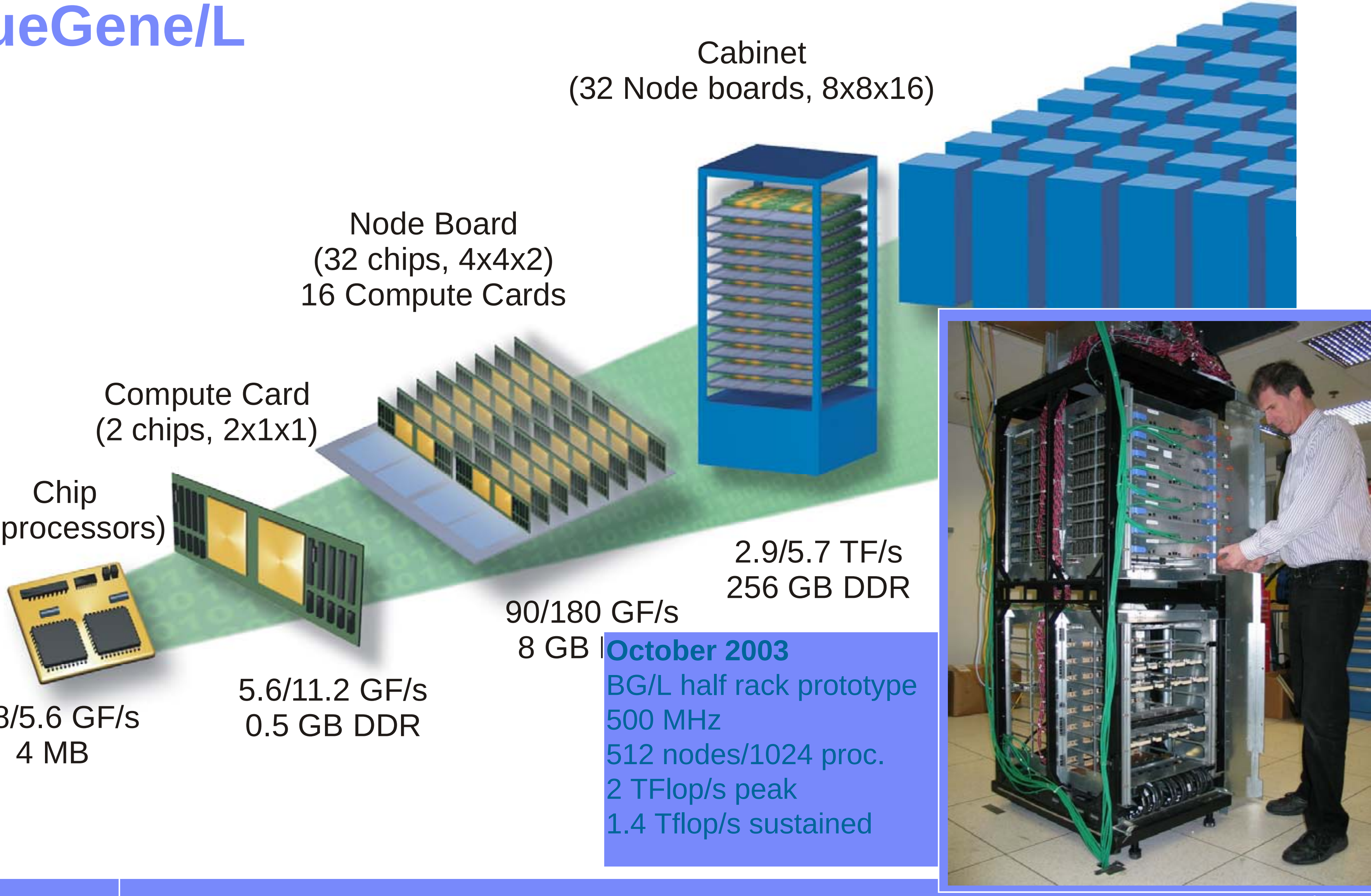
-Collaboration between Columbia University and IBM Research
-Optimized for QCD
-IBM 7SF Technology (ASIC Foundry Technology)
-20,000 1GF processors (nominal)
-4MB Embedded DRAM + External Commodity DDR/SDR SDRAM

- Blue Gene/L (180TF based on IBM System-on-a-Chip)

-Designed by IBM Research in IBM CMOS 8SF Technology
-64,000 2.8GF processors (nominal)
-4MB Embedded DRAM + External Commodity DDR SDRAM

BG/L 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)

2.9/5.7 TF/s
256 GB DDR

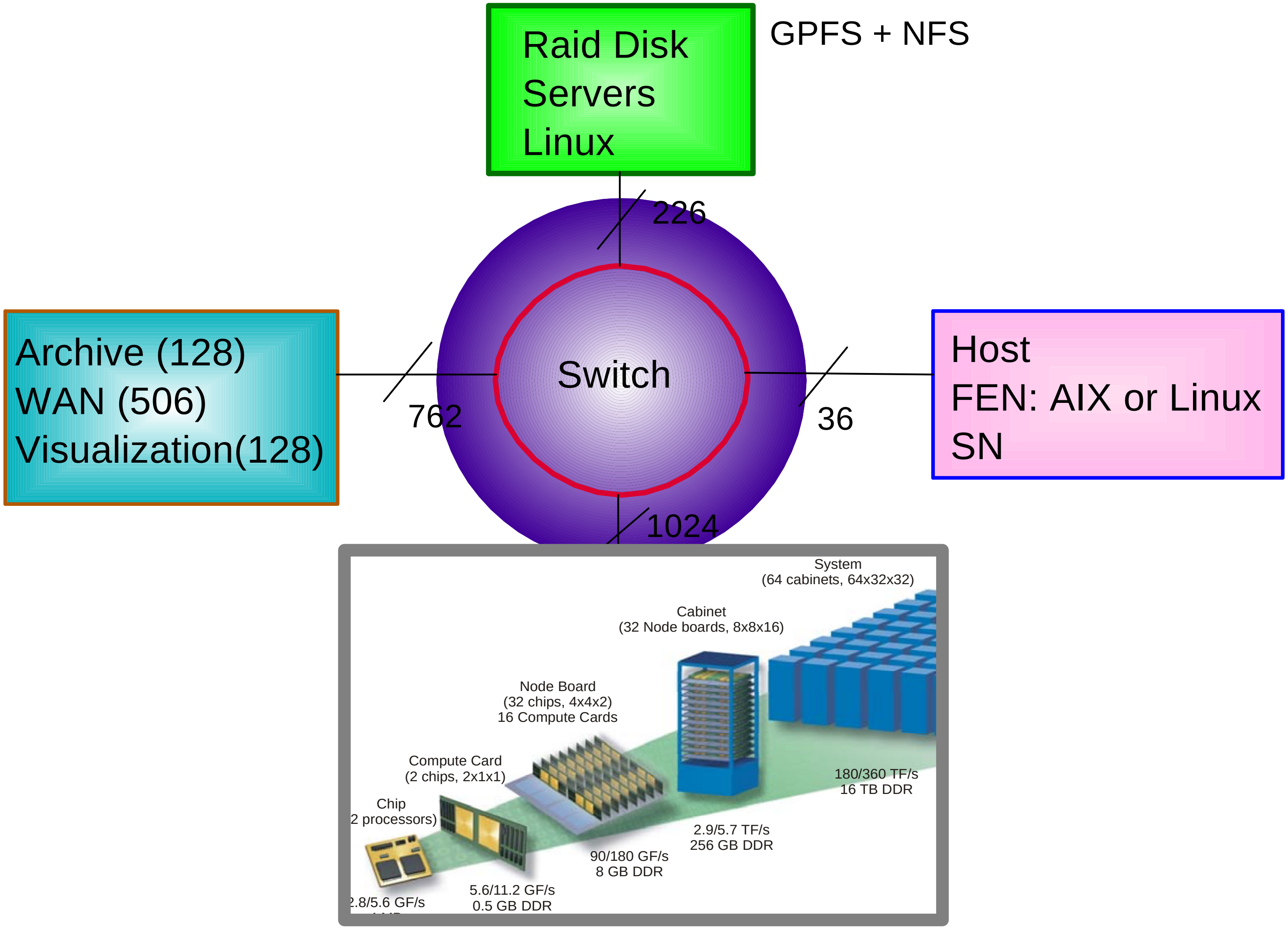
90/180 GF/s
8 GB I

5.6/11.2 GF/s
0.5 GB DDR

2.8/5.6 GF/s
4 MB

October 2003
BG/L half rack prototype
500 MHz
512 nodes/1024 proc.
2 TFlop/s peak
1.4 Tflop/s sustained

BlueGene/L System Host



Raid Disk
Servers
Linux

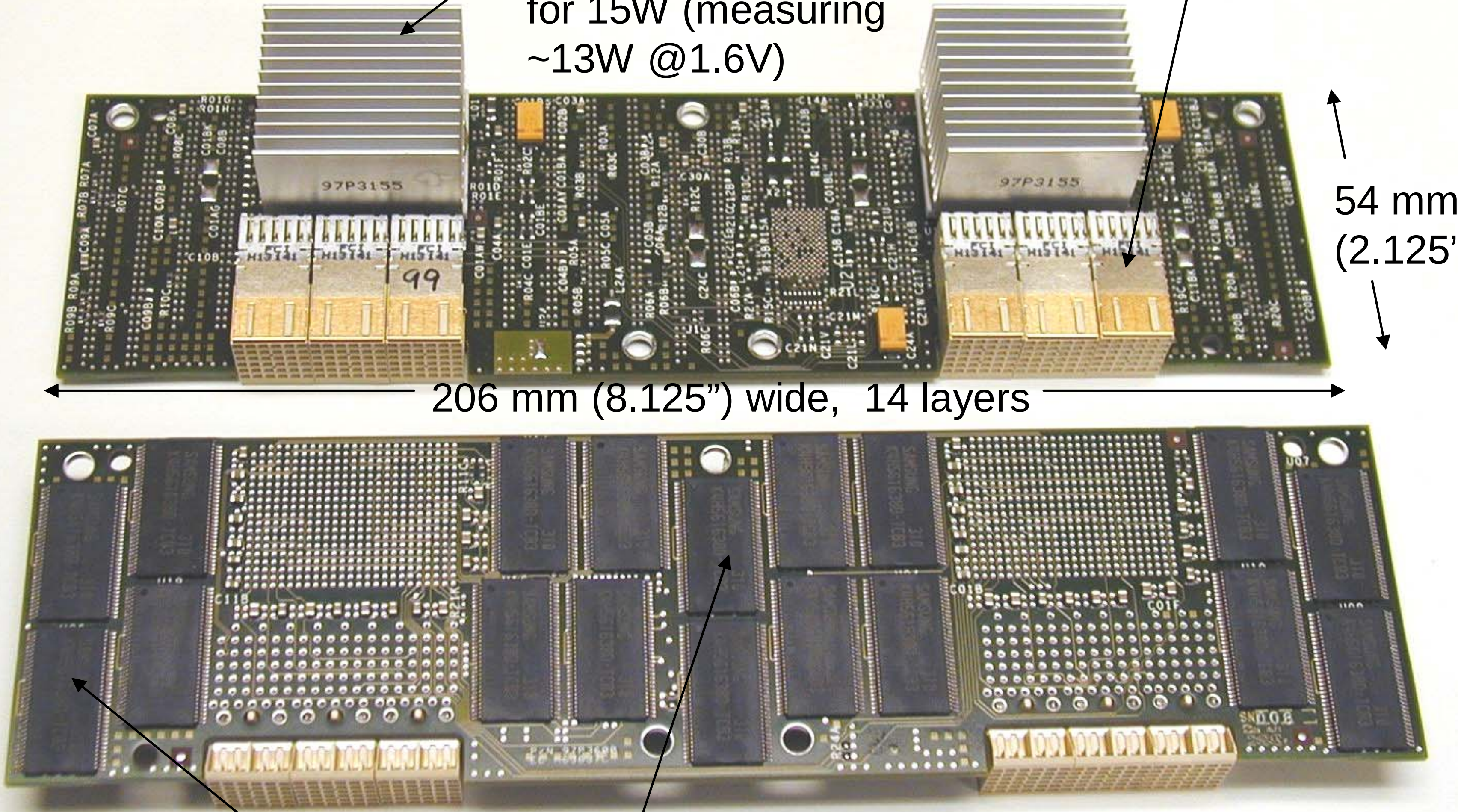
GPFS + NFS

Archive (128)
WAN (506)
Visualization(128)

Switch

Host
FEN: AIX or Linux
SN

Dual Node Compute Card



Heatsinks designed for 15W (measuring ~13W @1.6V)

Metral 4000 connector

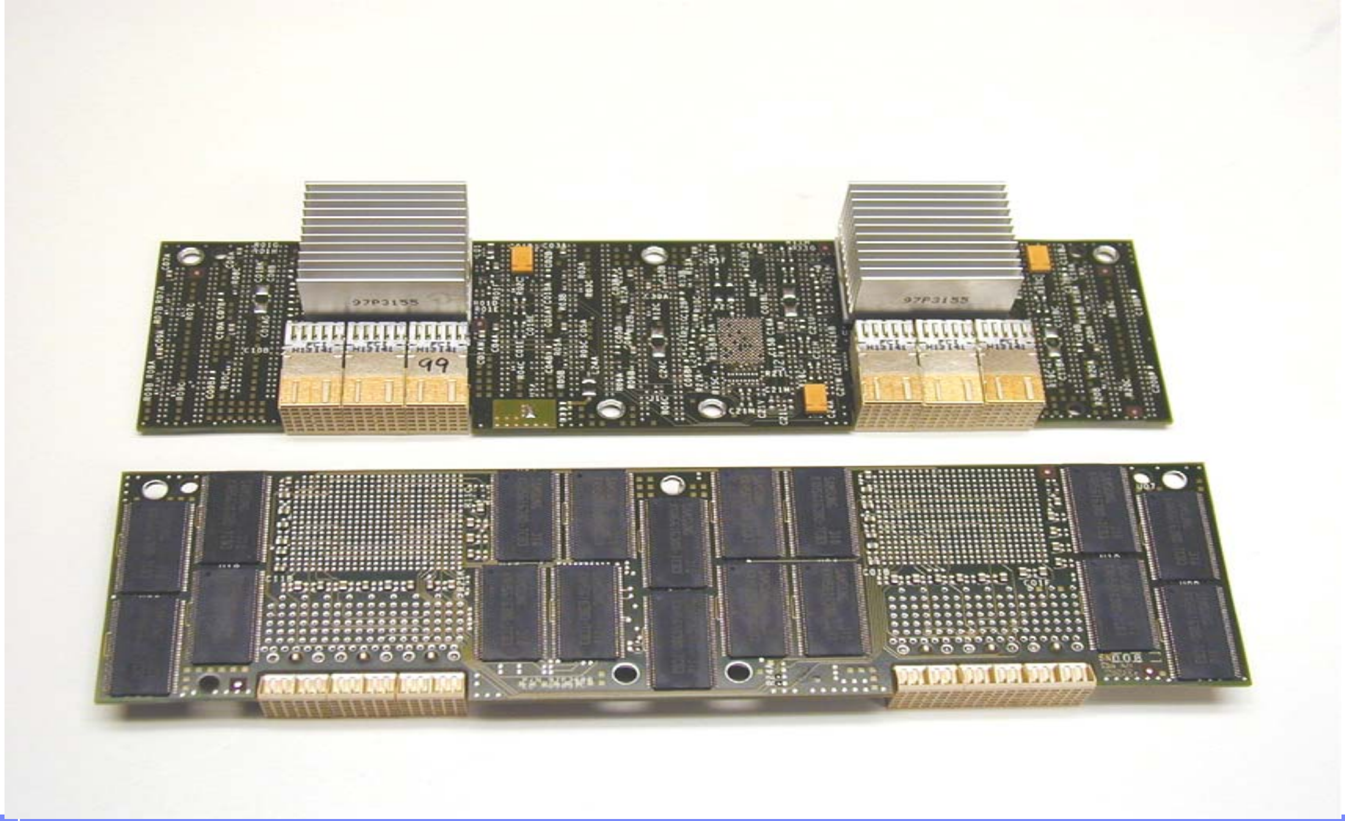
54 mm (2.125”)

206 mm (8.125”) wide, 14 layers

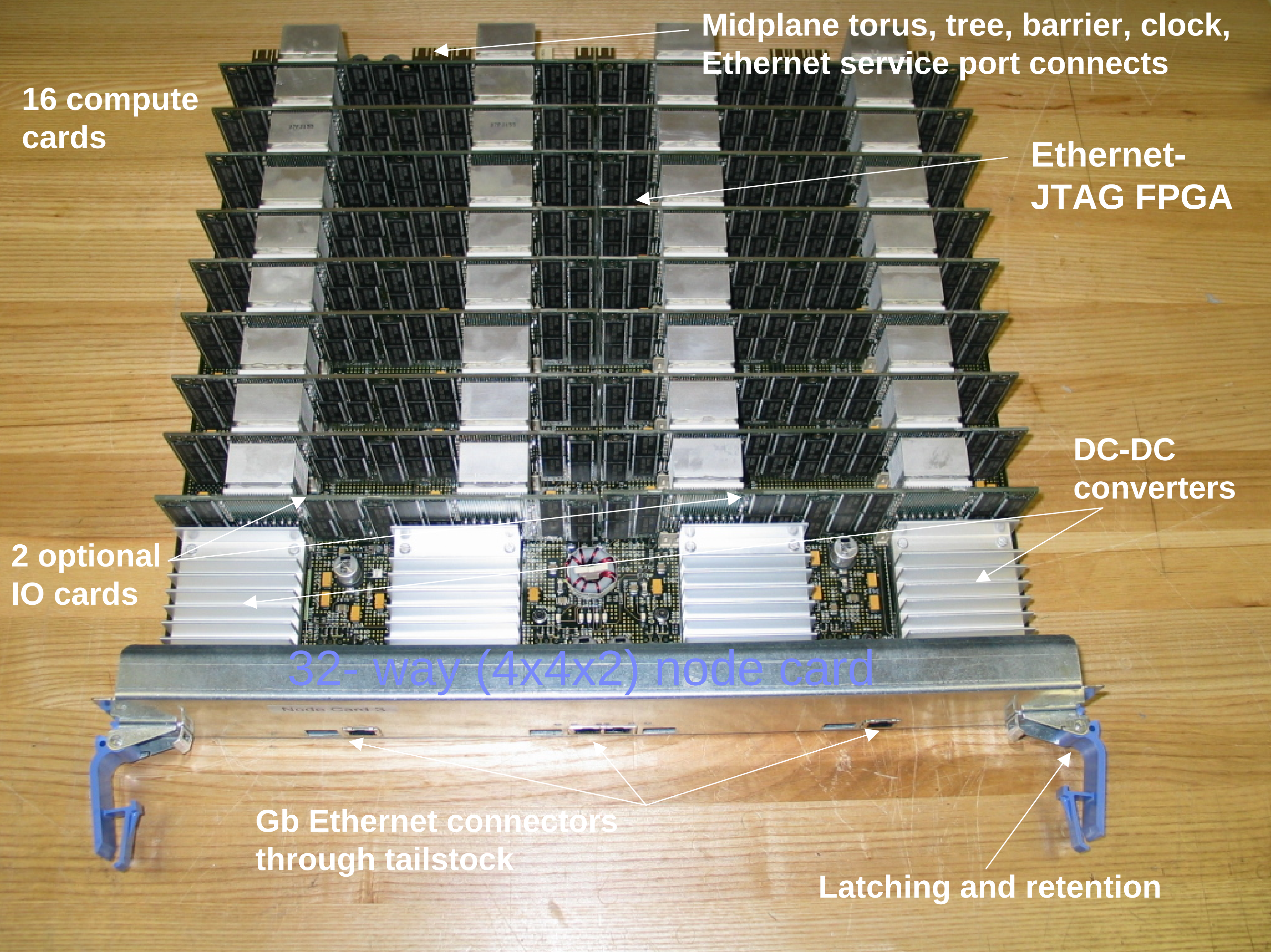
9 x 256Mb DRAM; 16B interface

I/O Card

- 2 nodes per I/O card
- FRU
- Memory can be doubled by stacking DRAMs



32- way (4x4x2) node card



Midplane torus, tree, barrier, clock, Ethernet service port connects

16 compute cards

Ethernet-JTAG FPGA

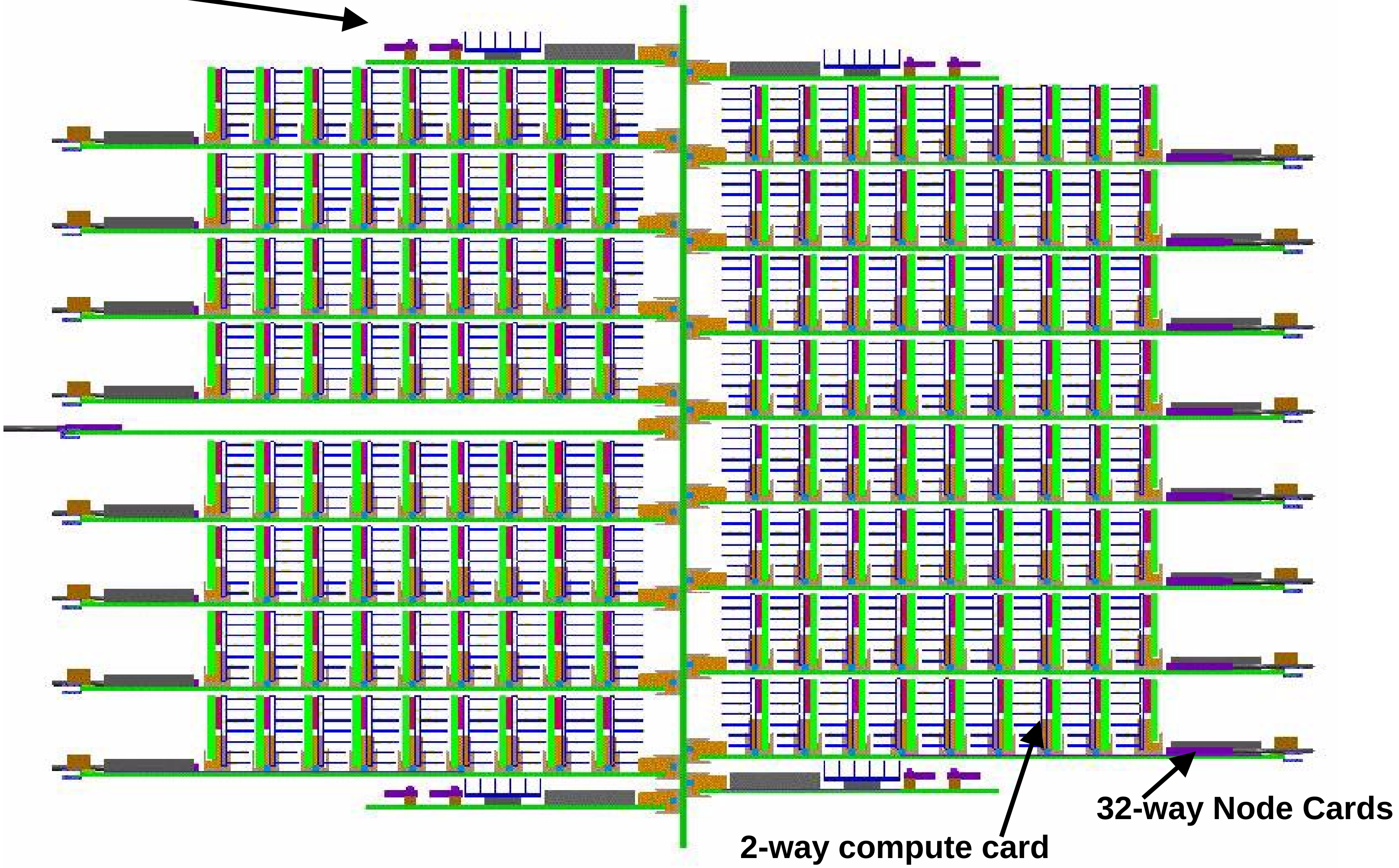
DC-DC converters

2 optional IO cards

Gb Ethernet connectors through tailstock

Latching and retention

512-way Midplane, Sideview

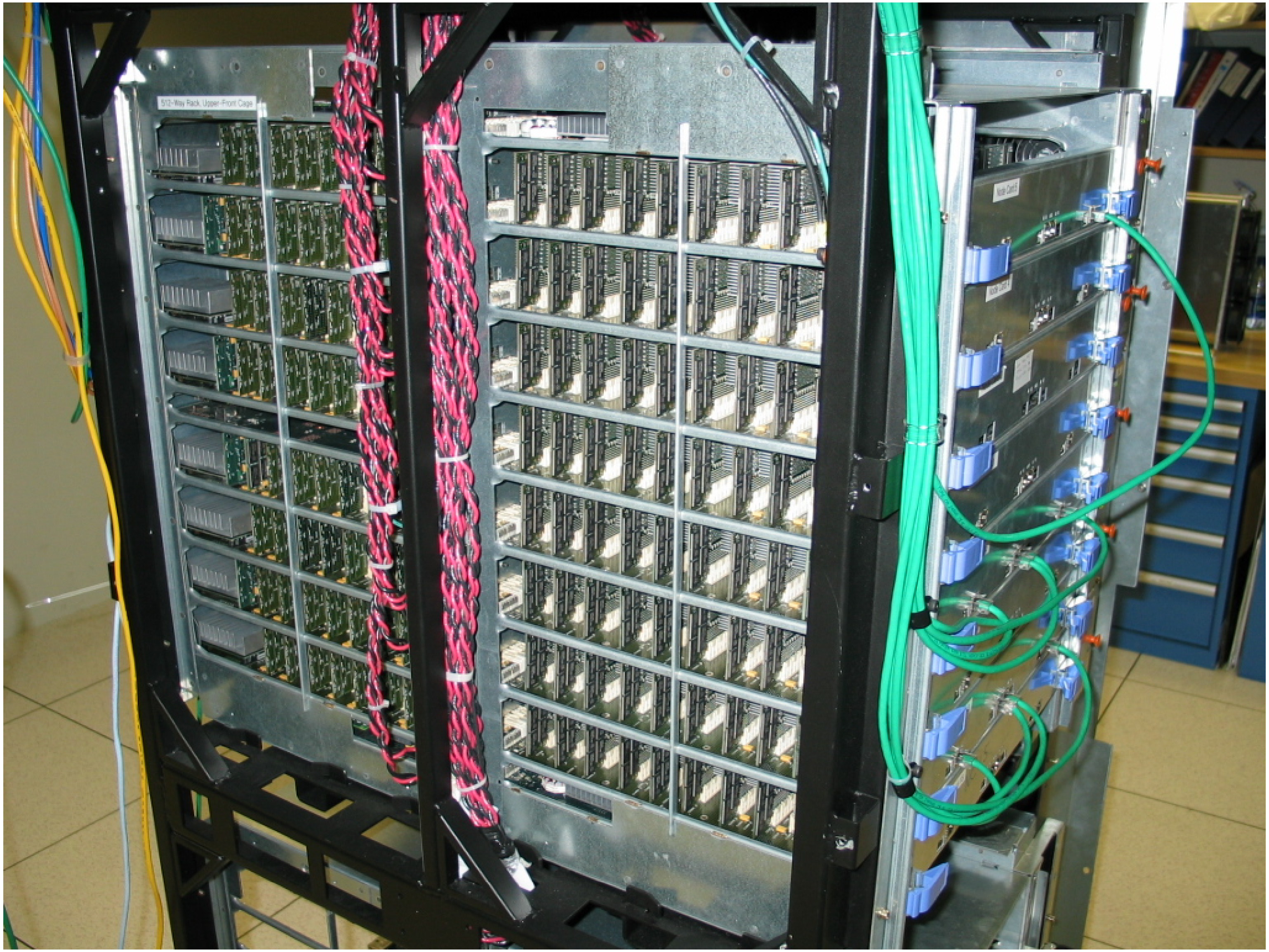


Link Cards without cables

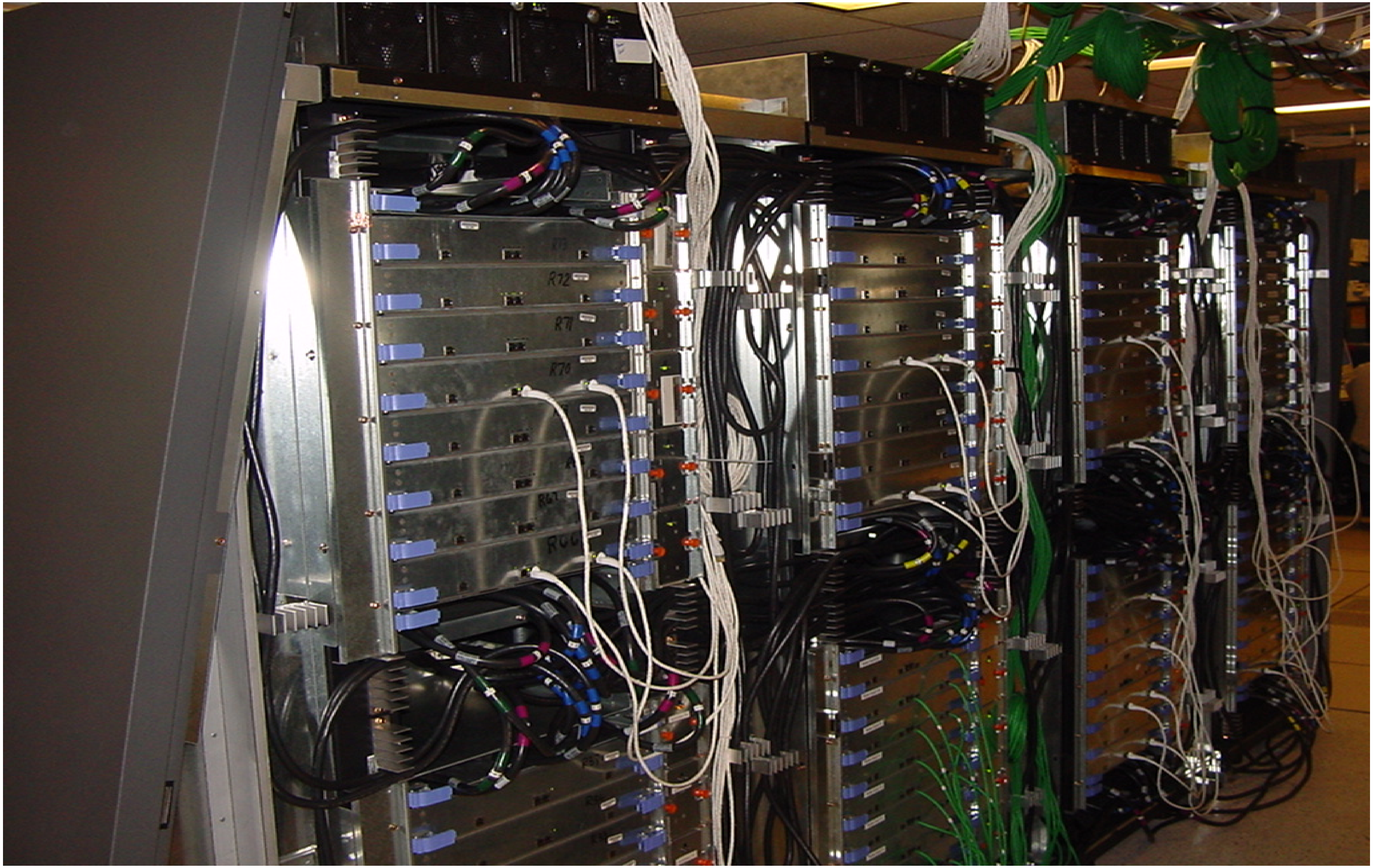
2-way compute card

32-way Node Cards

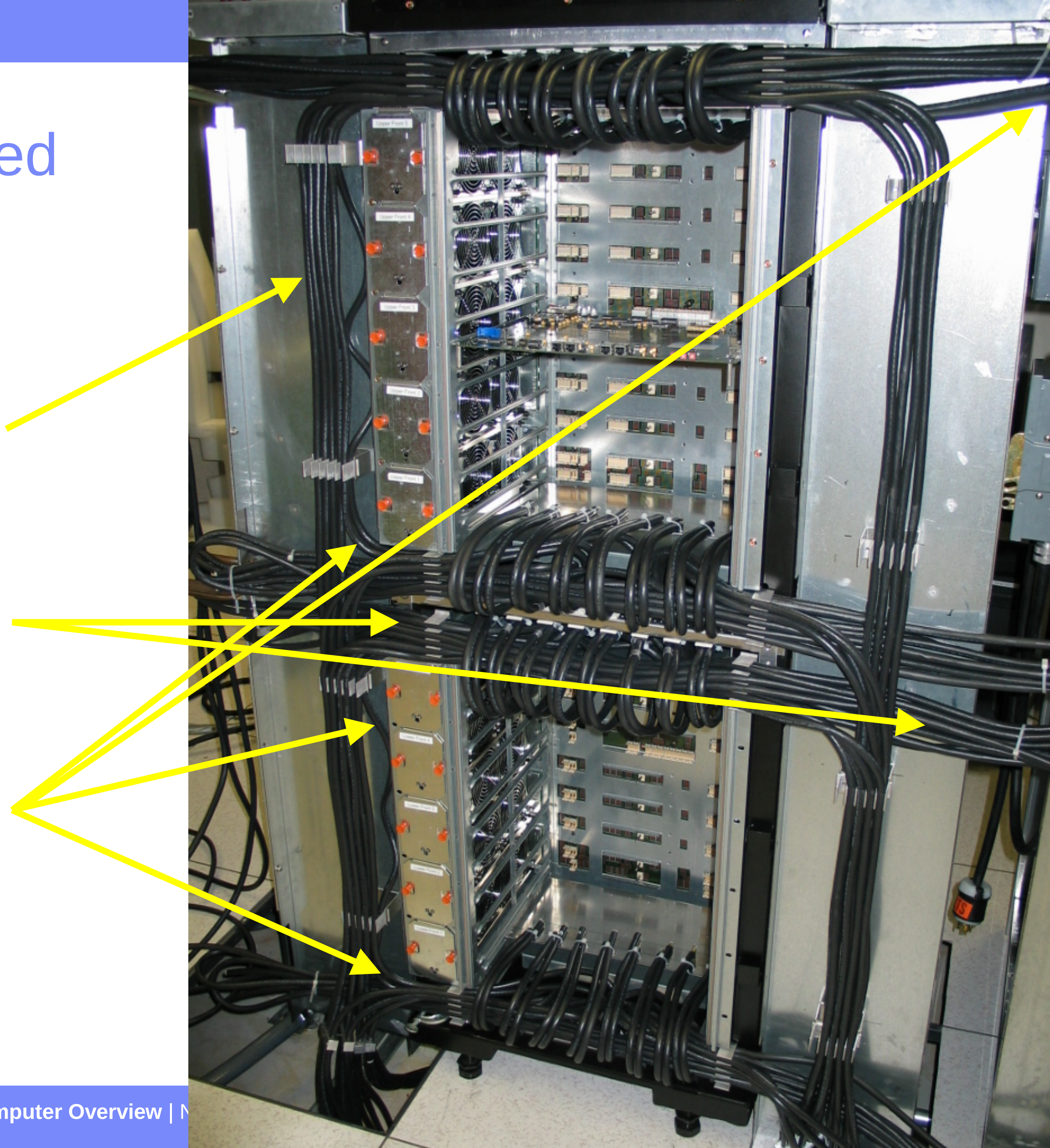
512-node midplane:



The current #4 supercomputer:



BG/L rack, cabled

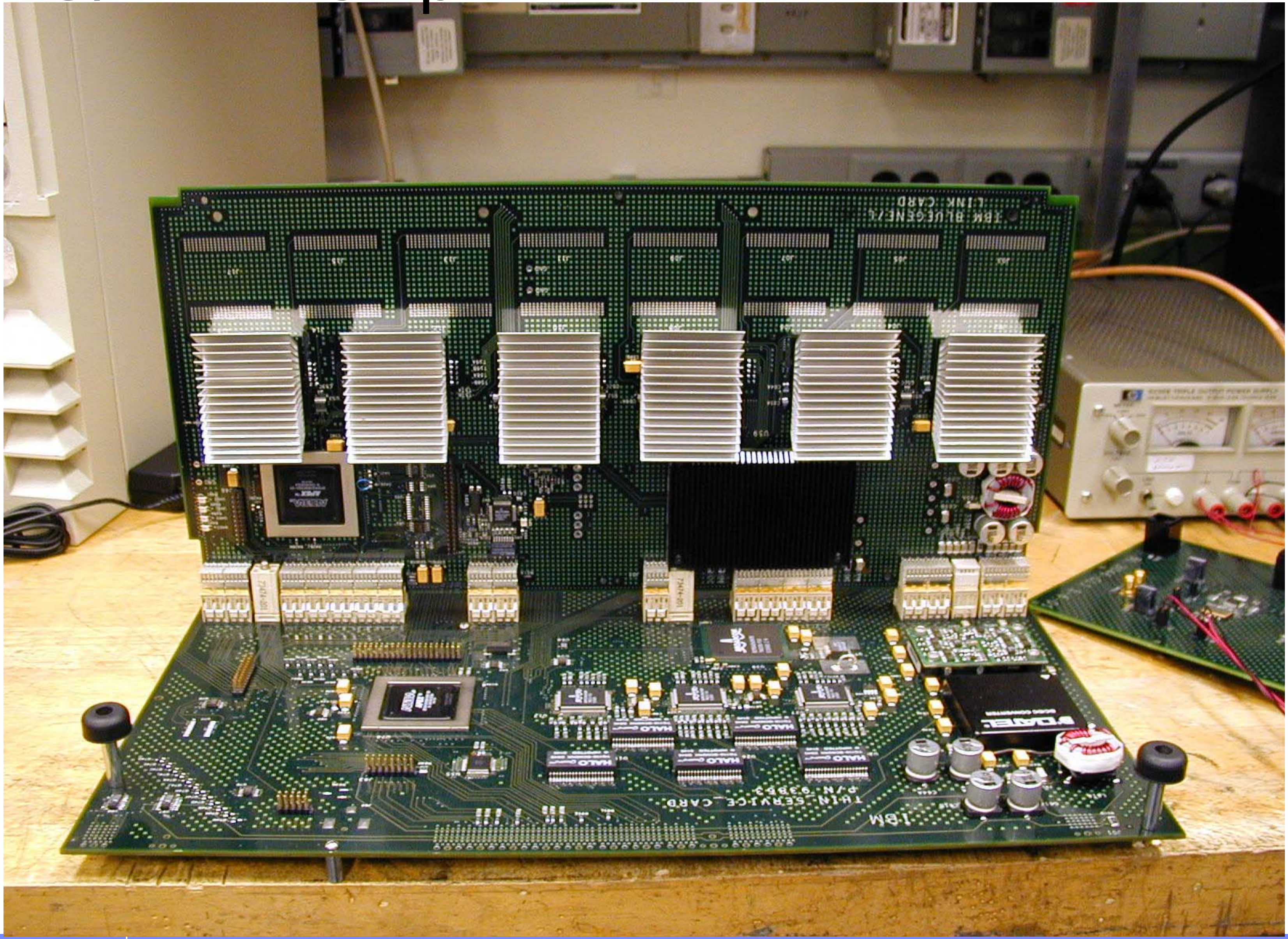


X Cables

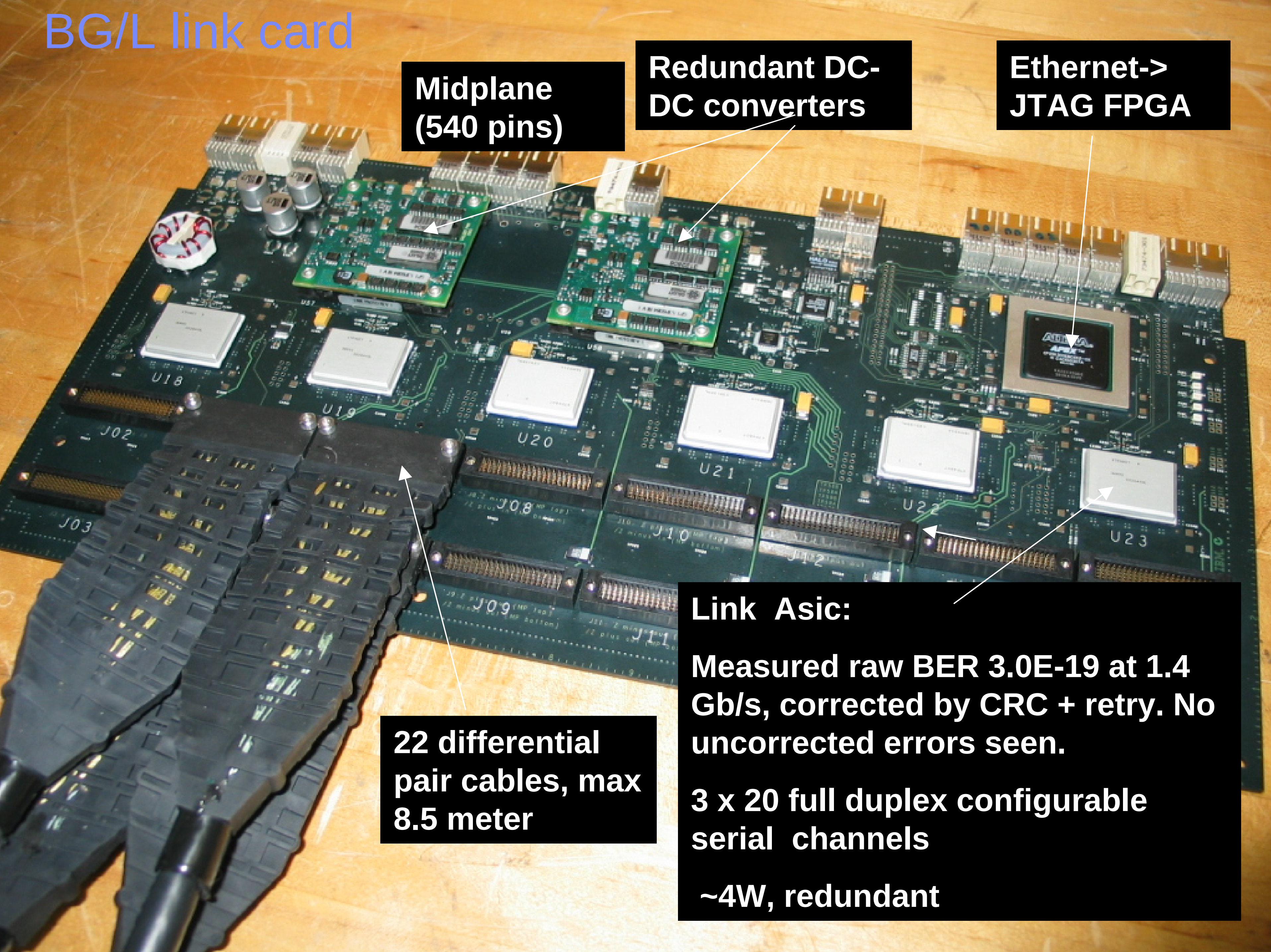
Y Cables

Z Cables

BG/L Link Chip



BG/L link card



Midplane (540 pins)

Redundant DC-DC converters

Ethernet-> JTAG FPGA

22 differential pair cables, max 8.5 meter

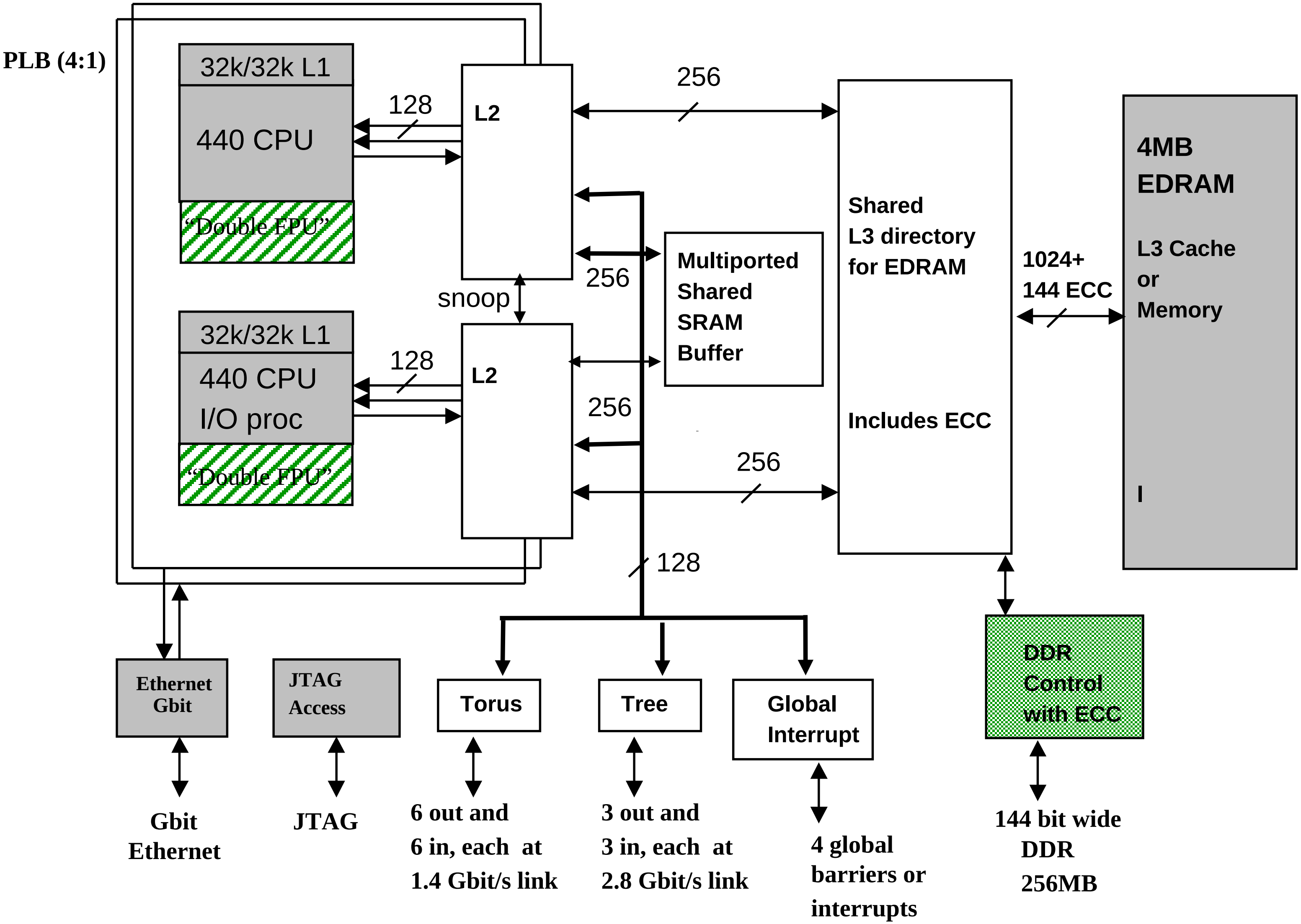
Link Asic:

Measured raw BER 3.0E-19 at 1.4 Gb/s, corrected by CRC + retry. No uncorrected errors seen.

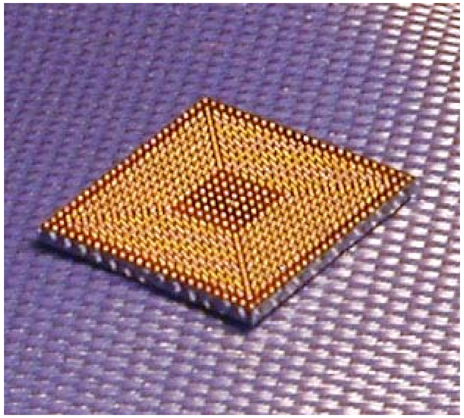
3 x 20 full duplex configurable serial channels

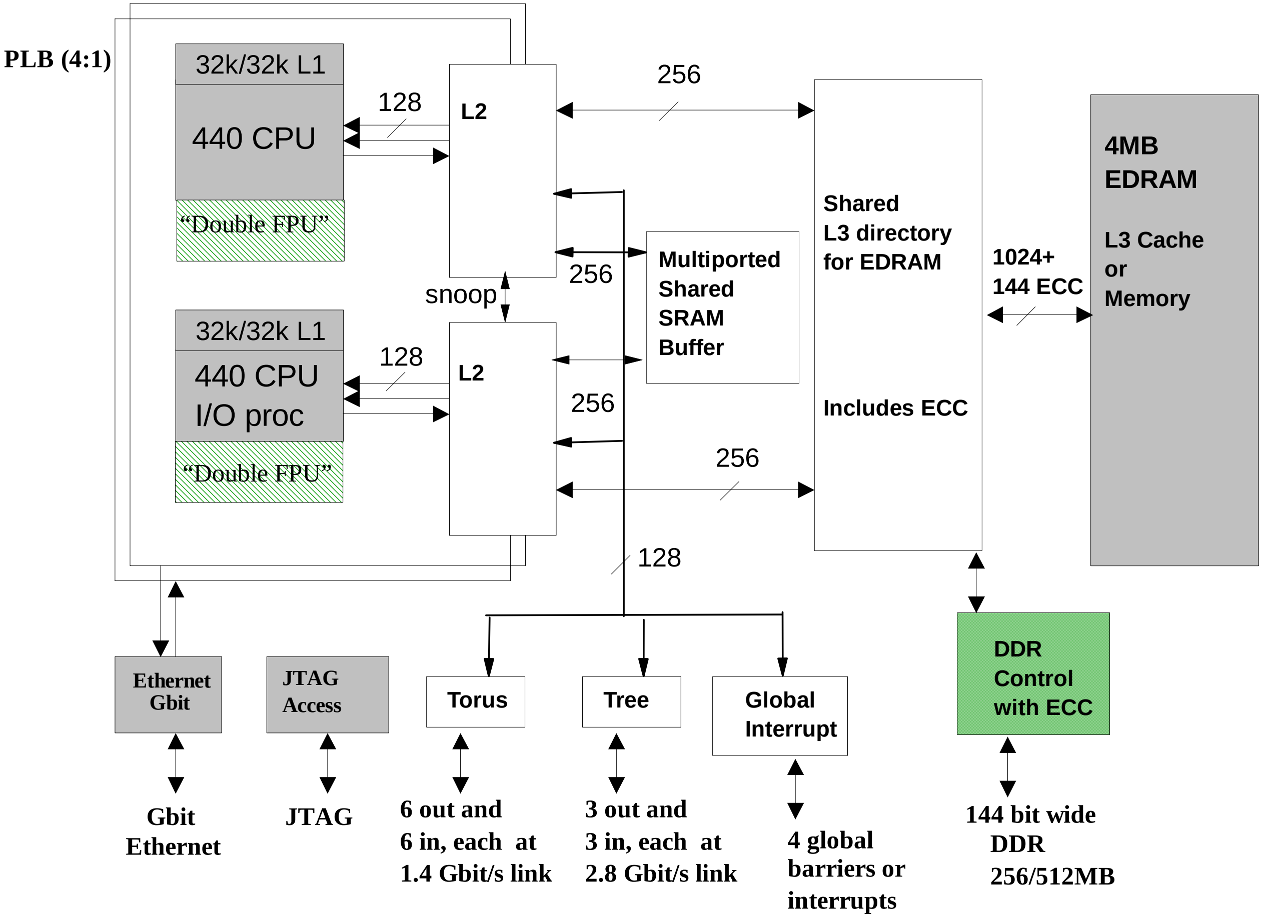
~4W, redundant

BG/L Compute ASIC



BlueGene/L Compute ASIC



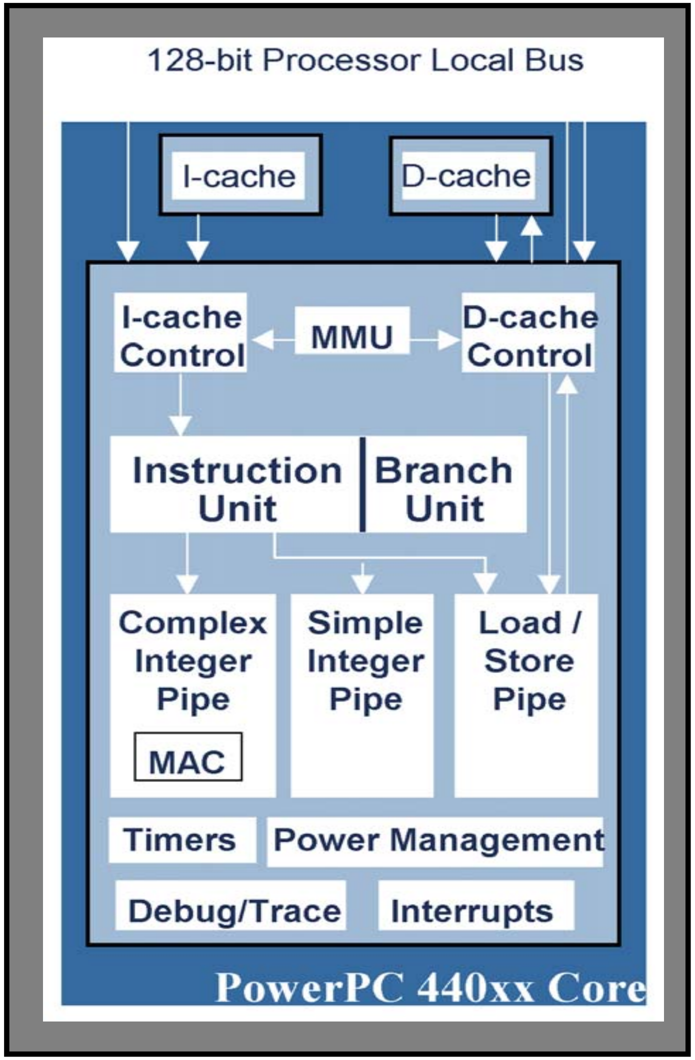


- IBM CU-11, 0.13 µm
- 11 x 11 mm die size
- 25 x 32 mm CBGA
- 474 pins, 328 signal
- 1.5/2.5 Volt

8m^2 of compute ASIC silicon in 65536 nodes!

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
  - 32KB instruction & 32KB data cache
  - 64-way set associative, 32 byte line
- 32-bit virtual address
- Real-time non-invasive trace
- 128-bit CoreConnect Interface



Floating Point Unit

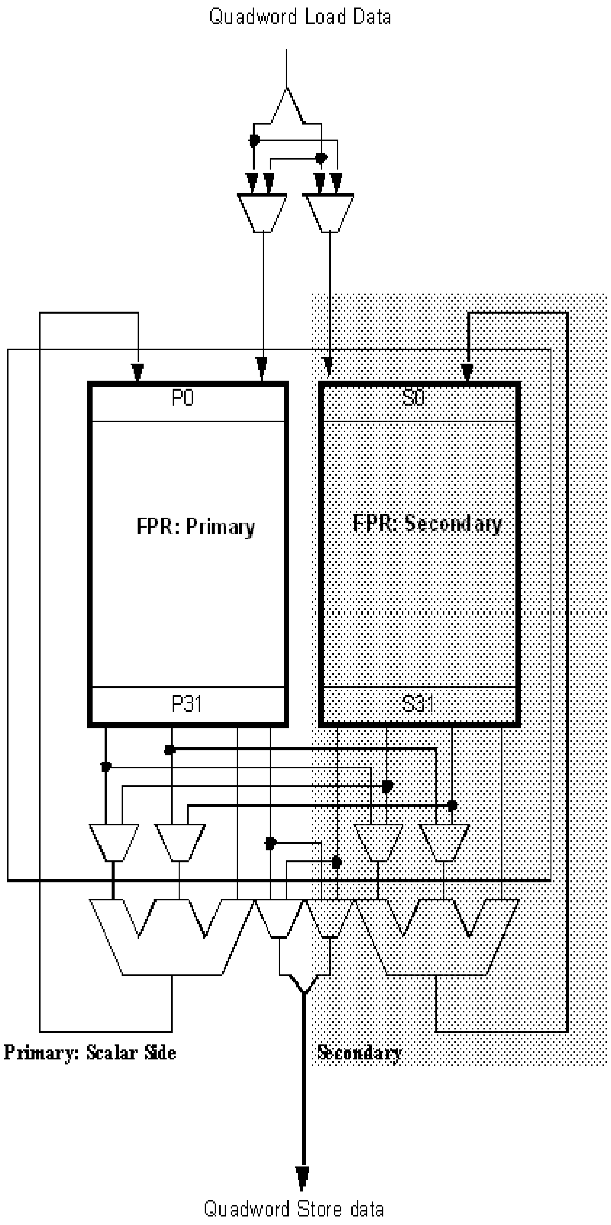
Primary side acts as off-the-shelf PPC440 FPU.

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

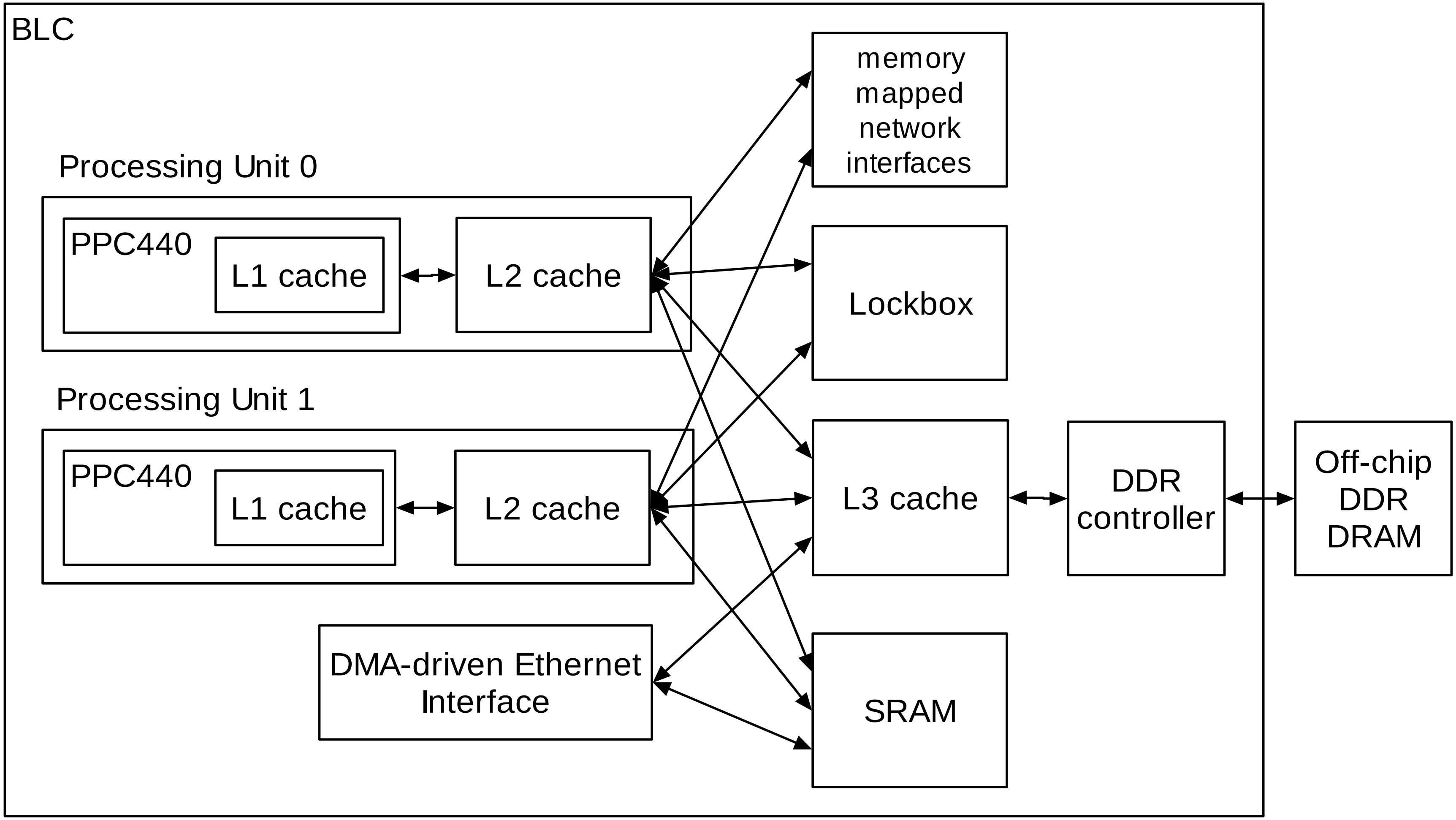
Secondary side doubles the registers and throughput.

Enhanced set of instructions for:

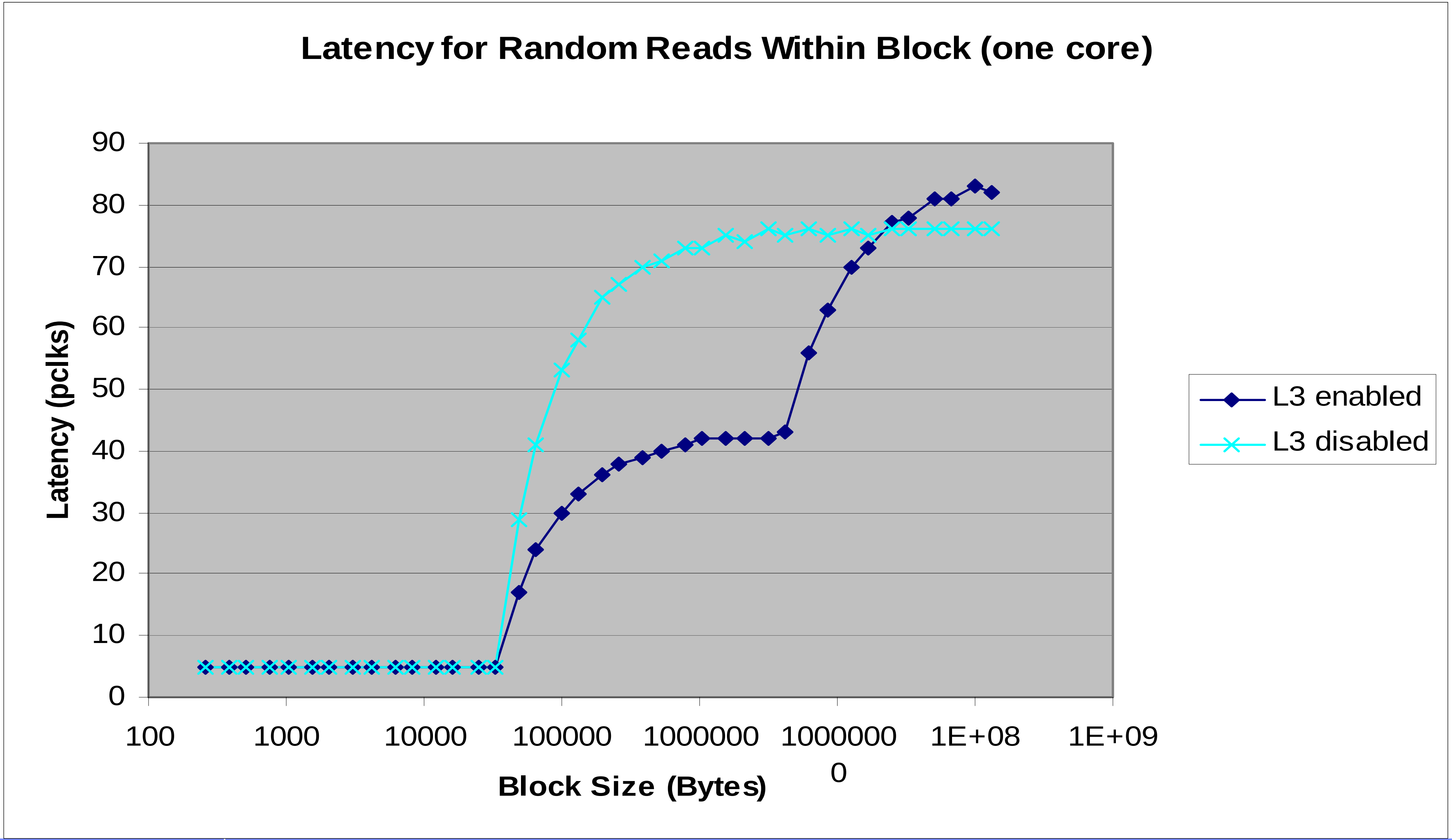
- Secondary side only.
- Both sides simultaneously:
  - Usual SIMD instructions.
    E.g. Quadword load, store.
  - Instructions beyond SIMD. E.g.
    - SIMOMD
      Single Inst. Multiple Operand Multiple Data.
    - Access to other register file.



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’

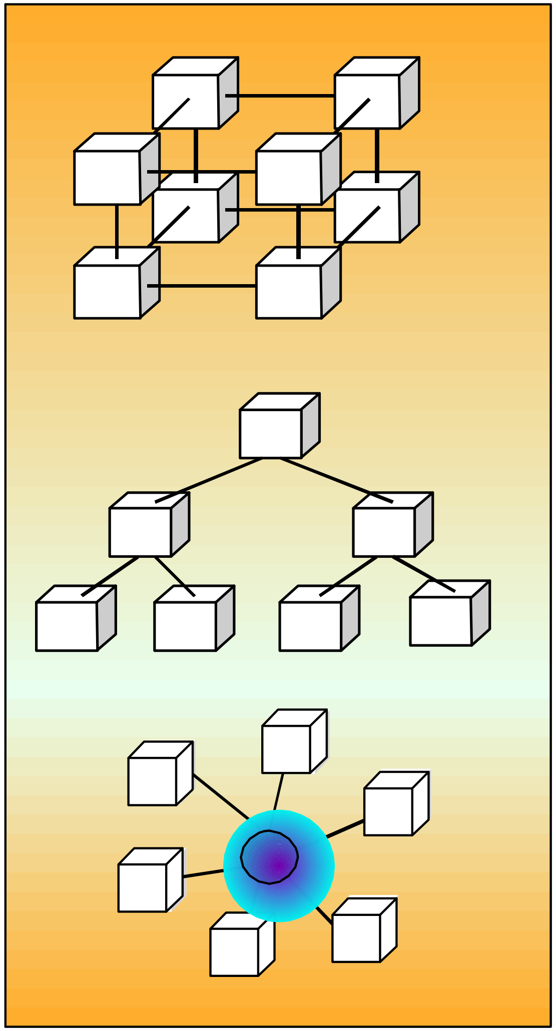
- A PPC440 core for application execution.
- A PPC440 core as communication co-processor.
- Communication library code maintains L1 coherency.

360 TeraFlops = ‘Virtual Node Mode’

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

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

BlueGene/L Interconnection Networks



3 Dimensional Torus

Interconnects all compute nodes, not I/O nodes.

Primary communications backbone for computations

Virtual cut-through hardware routing

1.4Gb/s on all 6in + 6out node links (2.1 GB/s per node)

1 µs latency between nearest neighbors, 5 µs to furthest (65536 nodes)

4 µs latency for one hop with MPI, 10 µs to the farthest (65536 nodes)

Global Tree

Interconnects all compute and I/O nodes

Communication backbone for application I/O to files, sockets, …

Specialized communication backbone for computation:

- One-to-all broadcast.
- Reduction operations across all compute nodes.

2.8 Gb/s of bandwidth per link

Latency of one way tree traversal 2.5 µs

Low Latency Global Barrier and Interrupt

Interconnects all compute and I/O nodes

Latency of round trip 1.3 µs (65536 nodes)

Ethernet

Incorporated into every node ASIC, but only active in the I/O nodes

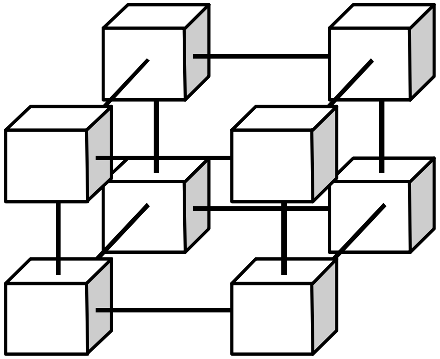
External comm. for application (file I/O, control, user interaction, etc.)

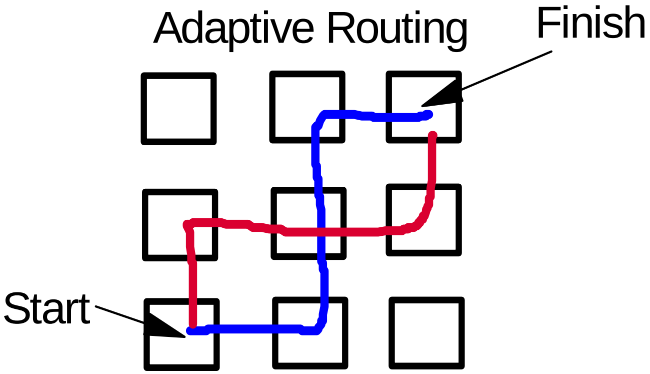
Control Network

Host access to each compute and I/O node.

Low-level boot, debug, monitorring, …

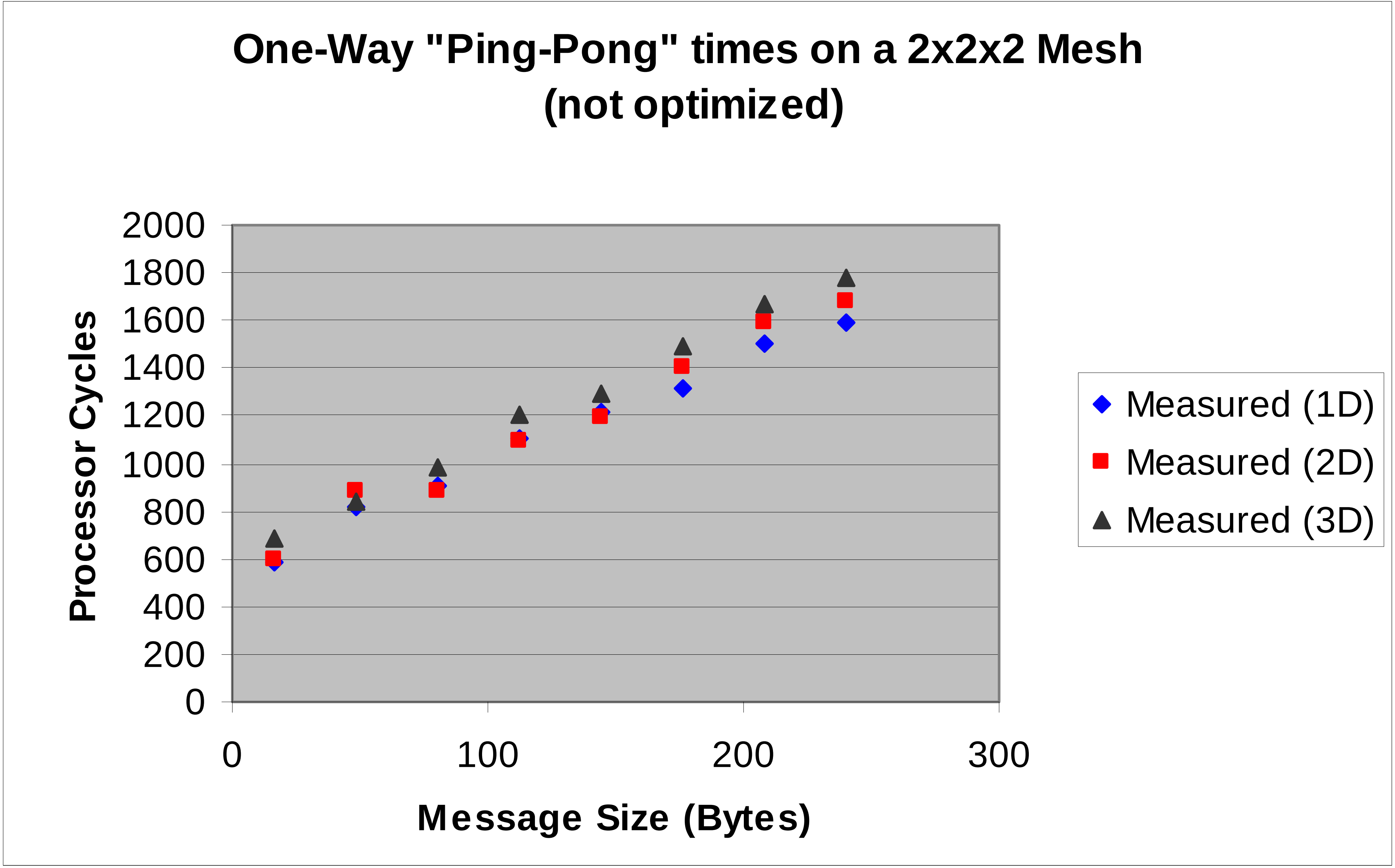
3-D Torus Network



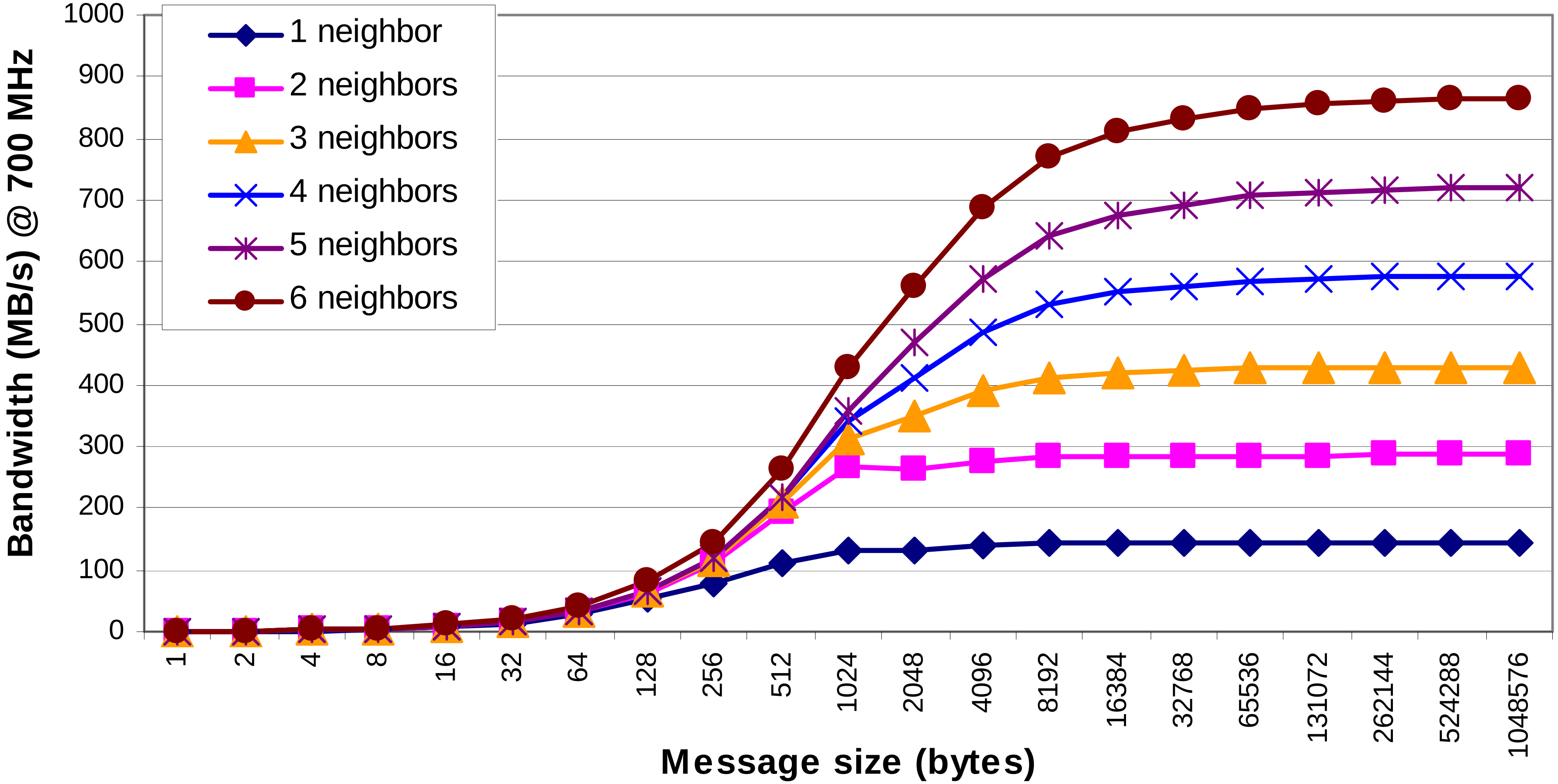


- 32x32x64 connectivity
- Backbone for one-to-one and one-to-some communications
- 1.4 Gb/s bi-directional bandwidth in all 6 directions (Total 2.1 GB/s/node)
- 64k * 6 * 1.4Gb/s = 68 TB/s total torus bandwidth
- 4 * 32 *32 * 1.4Gb/s = 5.6 Tb/s Bisectional Bandwidth
- Worst case hardware latency through 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.

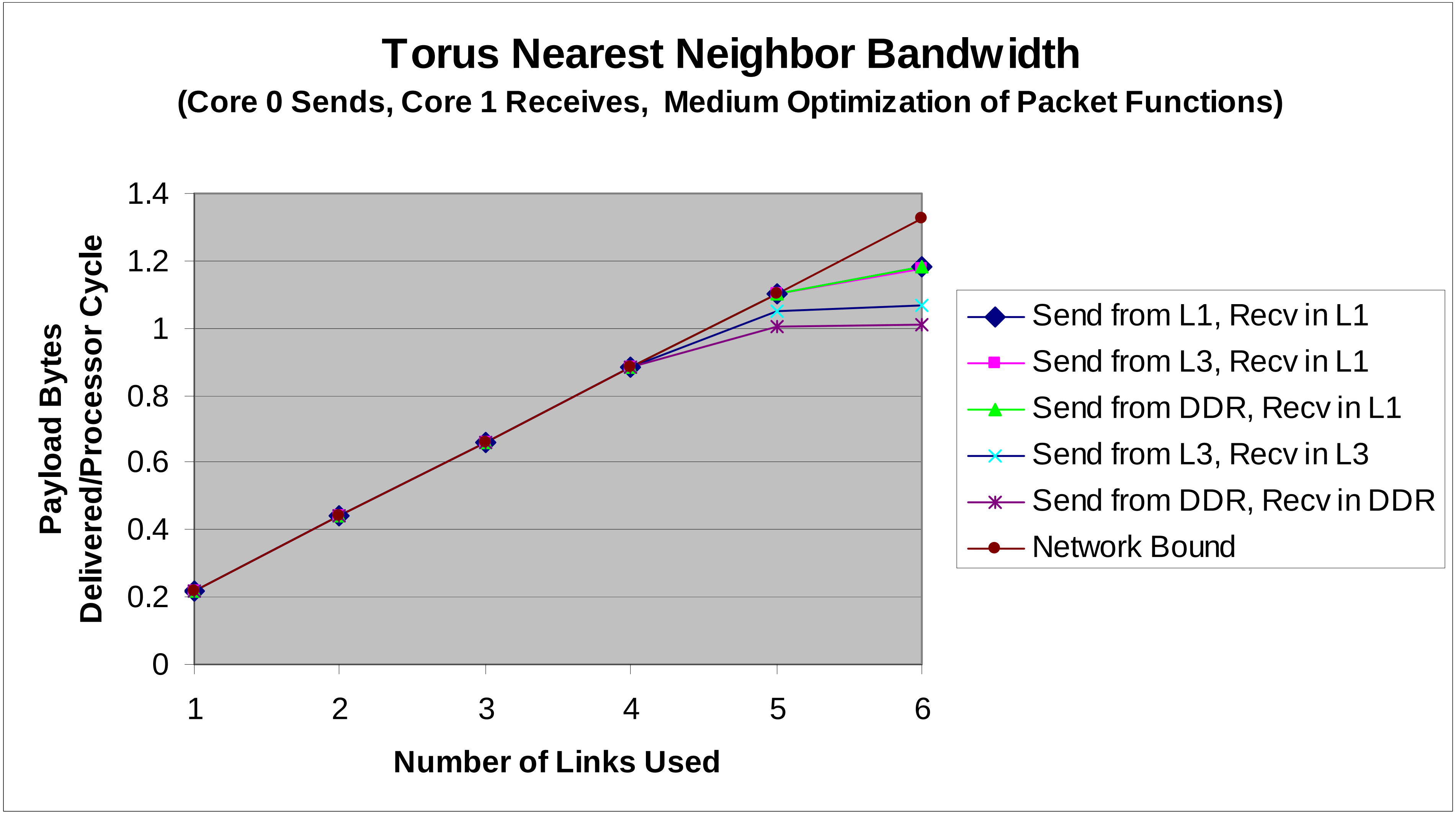
Prototype Delivers ~1usec Ping Pong low-level messaging latency



Measured MPI Send Bandwidth and Latency



Latency @700 MHz = 4 + 0.090 * “Manhattan distance” + 0.045 * “Midplane hops” μs


Torus Nearest Neighbor Bandwidth
(Core 0 Sends, Core 1 Receives, Medium Optimization of Packet Functions)
Payload Bytes Delivered/Processor Cycle
1.4
1.2
1
0.8
0.6
0.4
0.2
0
1
2
3
4
5
6
Number of Links Used
Send from L1, Recv in L1
Send from L3, Recv in L1
Send from DDR, Recv in L1
Send from L3, Recv in L3
Send from DDR, Recv in DDR
Network Bound

Peak Torus Performance for Some Collectives

L = 1.4Gb/s = 175MB/s = Uni-directional Link Bandwidth
N = number of nodes in a torus dimension

All2all = 8L/N_{max}

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

Broadcast = 6L = 1.05GB/s

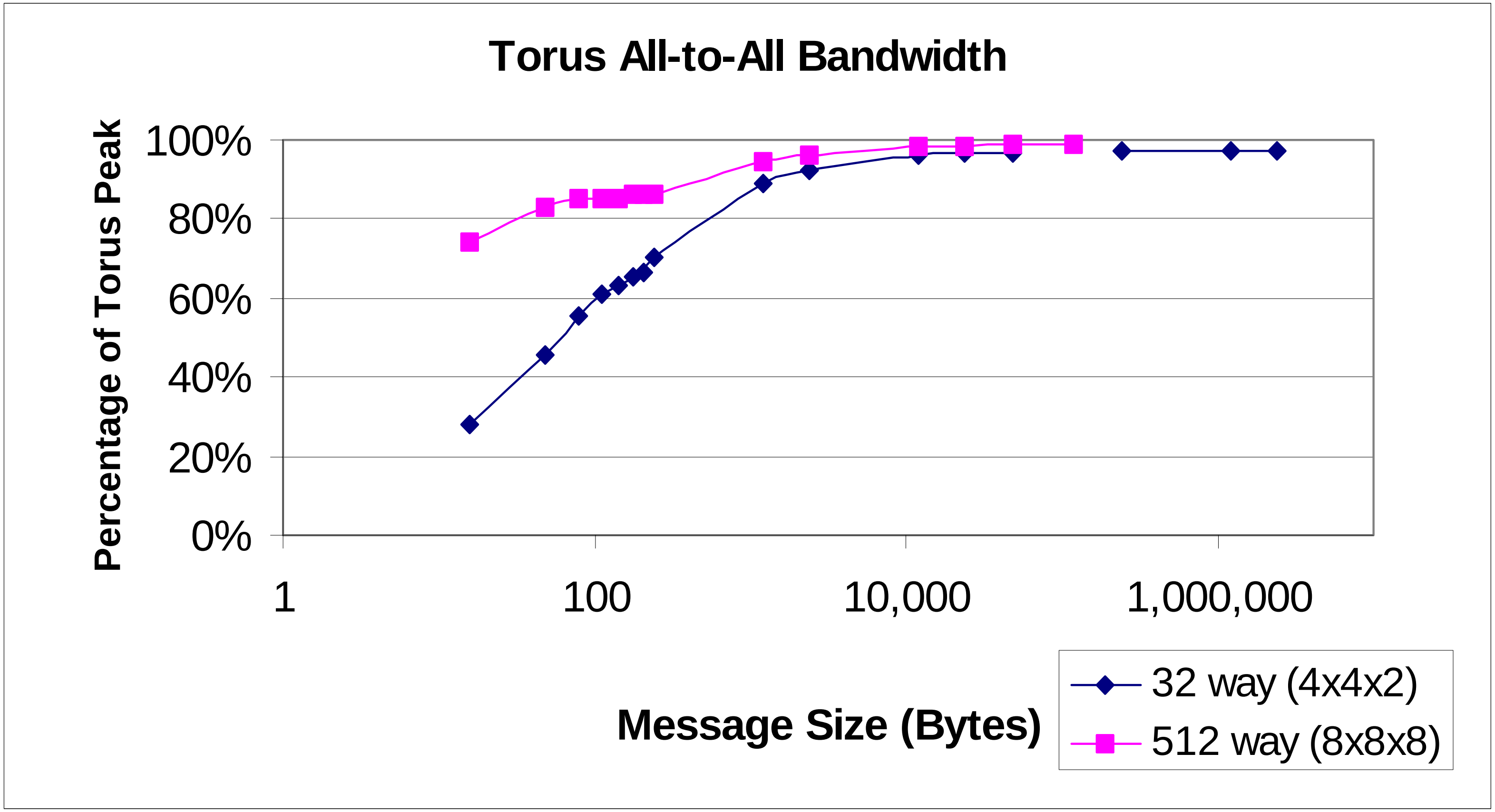
- 4 software hops, so fairly good latency.
- Hard for two PPC440 on each node to keep up, especially software hop nodes performing ‘corner turns’.

Reduce = 6L = 1.05GB/s

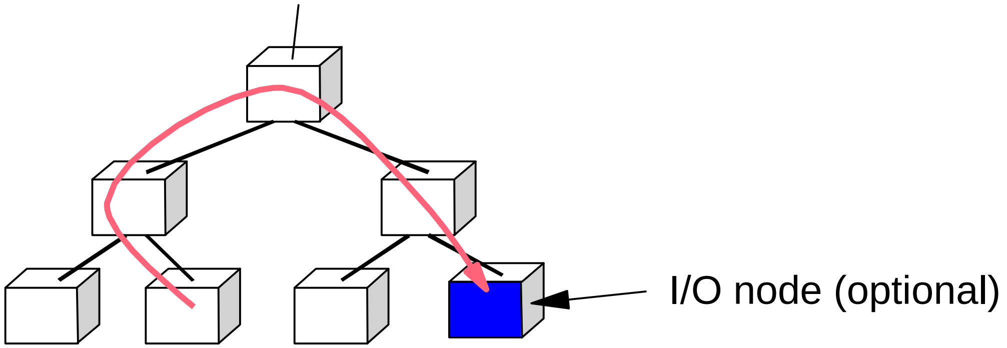
- (Nx+Ny+Nz)/2 software hops, so needs large messages.
- Very hard/Impossible for PPC440 to keep up.

AllReduce = 3L = 0.525GB/s

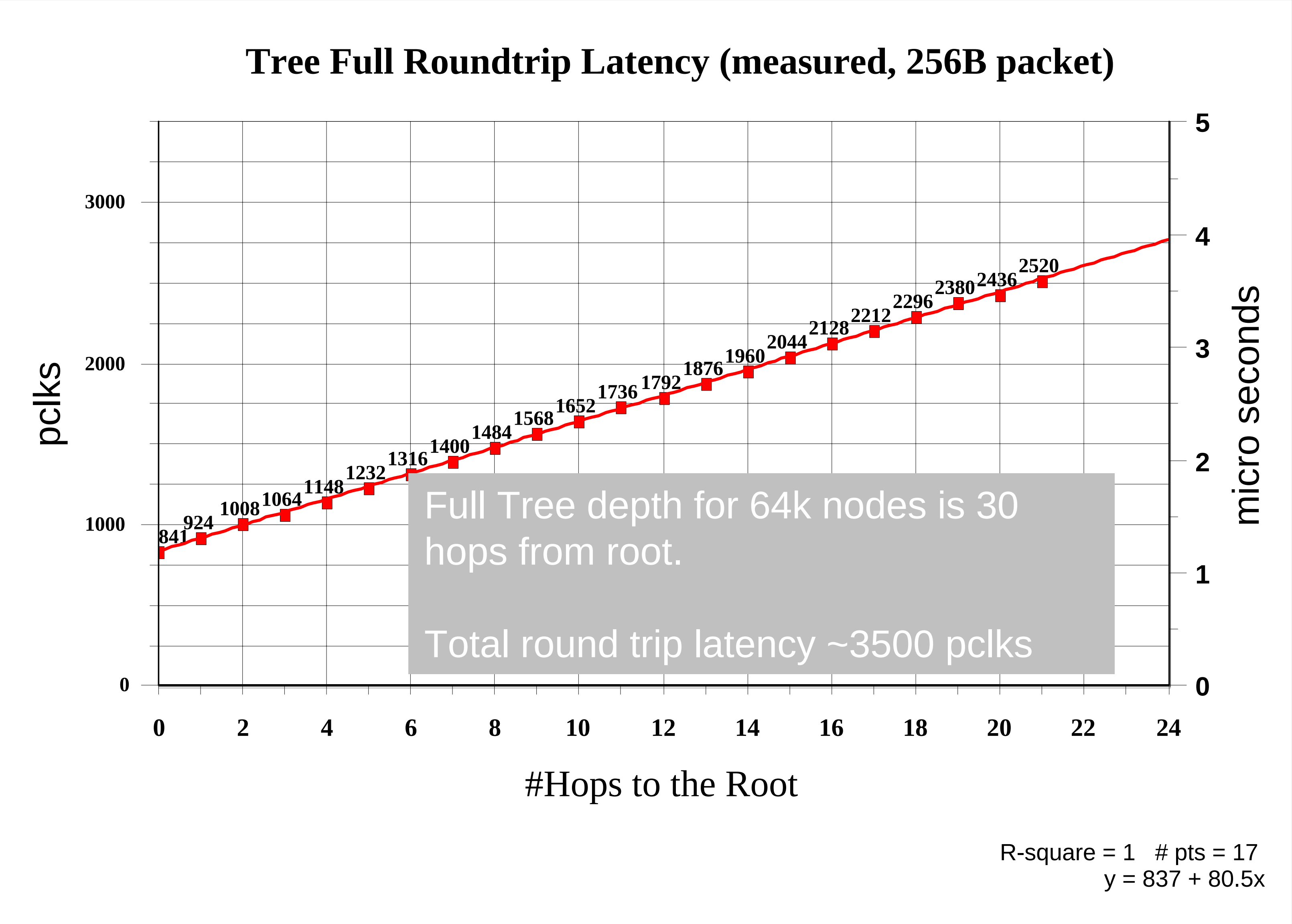
Link Utilization on Torus



Tree Network



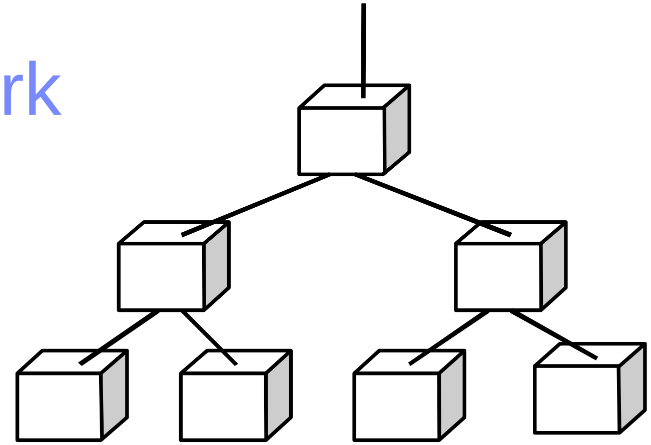
- High Bandwidth one-to-all
  - 2.8Gb/s to all 64k nodes
  - 68TB/s aggregate bandwidth
- Arithmetic operations implemented in tree
  - Integer/ Floating Point Maximum/Minimum
  - Integer addition/subtract, bitwise logical operations
- Latency of tree less than 2.5usec to top, additional 2.5usec to broadcast 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



Full Tree depth for 64k nodes is 30 hops from root.

Total round trip latency ~3500 pclks

Fast Barrier/Interrupt Network



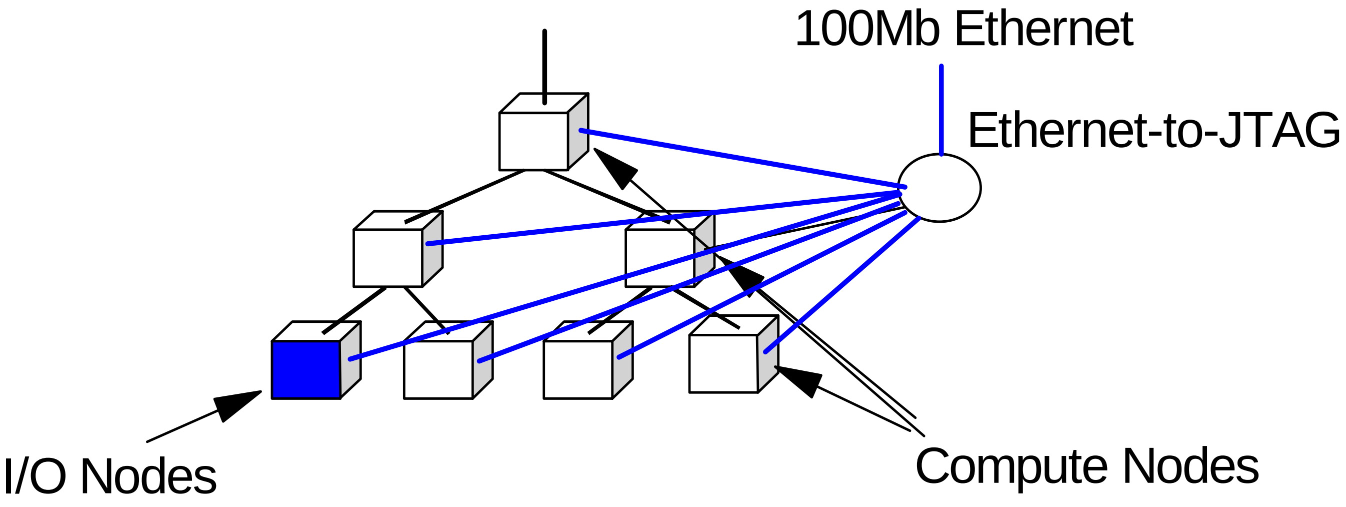
- Four Independent Barrier or Interrupt Channels
  - Independently Configurable as "or" or "and"
- Asynchronous Propagation
  - Halt operation quickly (current estimate is 1.3usec worst case round trip)

    > 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

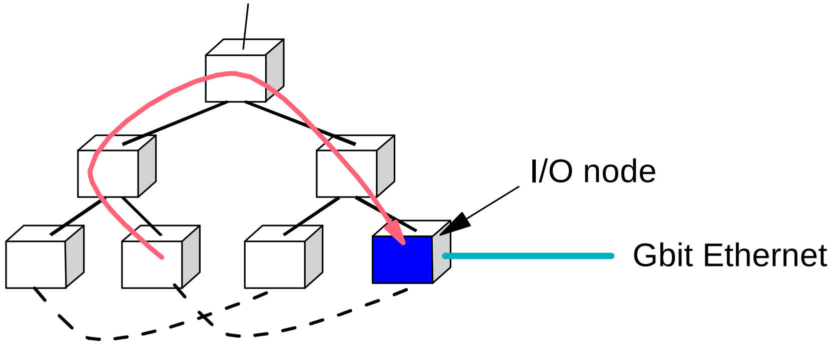
Control Network

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



Gb Ethernet Disk/Host I/O Network



Gb Ethernet on all I/O nodes

- Gbit Ethernet Integrated in all node ASICs but only used on I/O nodes.
- Funnel via global tree.
- I/O nodes use same ASIC but are dedicated to I/O Tasks.
- I/O nodes can utilize larger memory.

Dedicated DMA controller for transfer to/from Memory

Configurable ratio of Compute to I/O nodes

- I/O nodes are leaves on the tree network

Design for Reliability and Availability

- Philosophy
  - Make system design choices to improve the reliability of the most failure-prone units
    - Soldered DRAMS - no DIMM connectors
    - Press fit card-card and cable connectors - no surface mount
    - Reliability grade 1 IBM ASICs
    - All component failure rate data reviewed – part of component choice
- Features
  - Redundant bulk supplies, power converters, fans, DRAM bits.
  - ECC or parity/retry with sparing on most buses.
  - Extensive data logging (voltage, temp, recoverable errors, …) and failure forecasting.
  - Only fails early in global clock tree, or certain failures of link cards, affect more than one rack.
- Result for 65536 nodes:
  - Total predicted FRU fails in 10 days without redundancy: 5.3
  - Total predicted FRU fails in 10 days with redundancy: 0.63
    - Counts only those types of redundancy with no performance impact

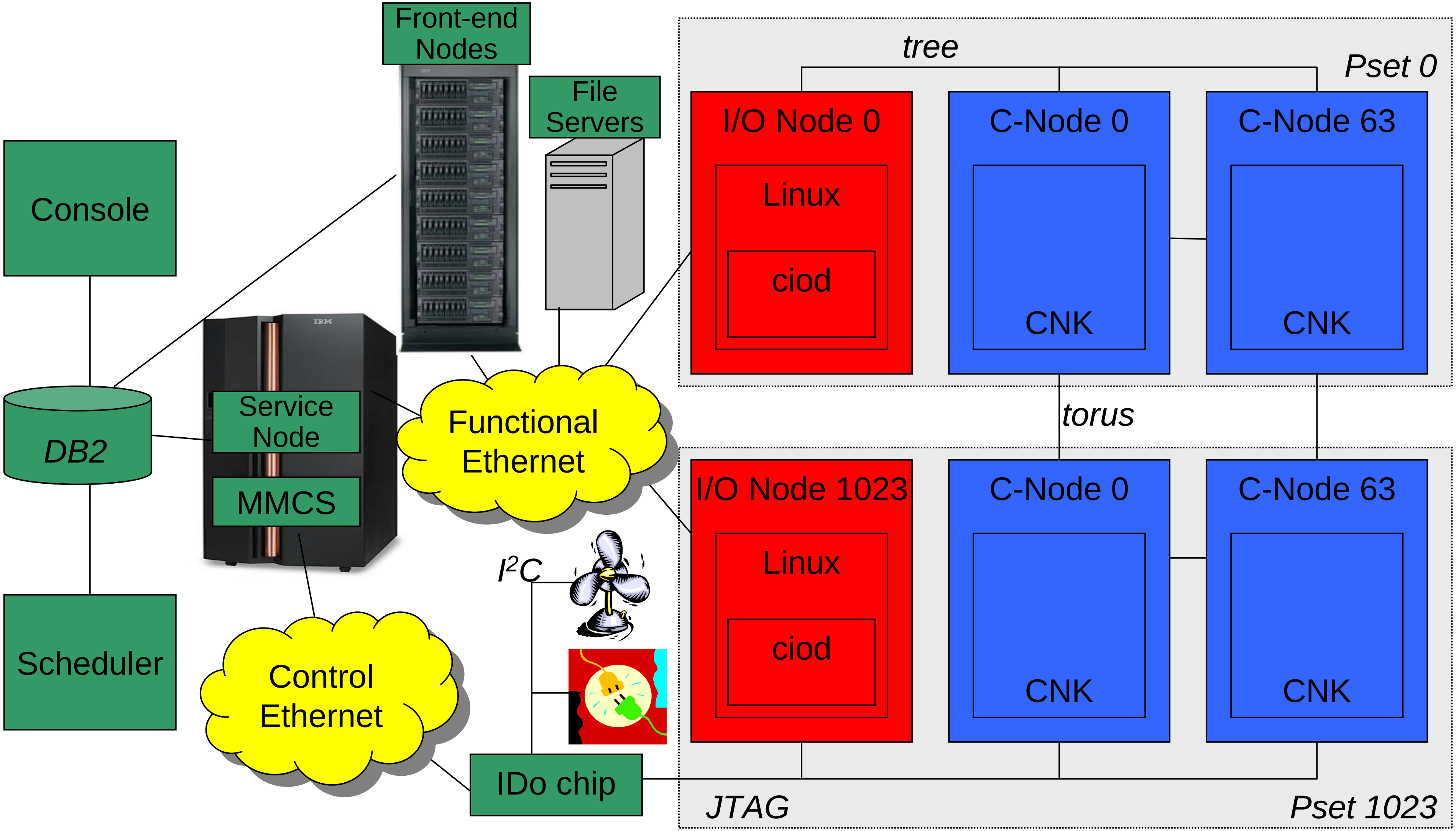
High Performance and High Reliability

- On-chip
  - Parity on torus and tree dataflow
  - ECC on SRAM modules (1 bit correct, 2 bit detect), threshold counters on errors
  - ECC on EDRAM L3 cache
  - Parity on all data buses between 440 cores and other functional units.
  - Parity on 440 L1 ICACHE and DCACHE
  - Parity on state is checked where possible including on global interrupts
- DDR main memory interface
  - 24 bit ECC (24 bits in 2 transfers): 4 bit correct
  - 4 Spare bits: can be substituted for any other 4 bit data or ECC block
- AC-DC power supplies
  - 6+1 redundancy. If any one supply in a rack drops out, the remaining six can handle the load.
- DC-DC power supplies
  - 3+1 redundancy for the node card 2.5V and 1.5V domains
  - 1+1 redundancy for the link card 1.5V domain
- Fans
  - 2+1 redundancy in each fan module. If one fan drops out, the other two speed up to compensate.
- Cable links (tree, torus and interrupt)
  - Parity added to the group of 17 synchronous bits traveling together on each cable
  - 1 spare synchronous bit and 1 spare asynchronous bit per cable
- All synchronous links (torus and tree)
  - Packet CRCs
    - Automatic hardware retransmission of corrupted packets between compute ASICs
    - Catches up to 5 random bit errors per packet, up to 24 continuous stream errors
  - * 32 bit cumulative link CRC stored at driver and receiver of each link (requires software check)

Software Design Overview

- Familiar software development environment and programming models
- **Scalability to *O(100,000)* processors – through Simplicity**
  - 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
  - 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
  - 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 for debuggers, hardware performance monitors, trace based visualization
- POSIX system calls – compute processes “feel like” they are executing on a Linux environment (restrictions)

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 = 2PPC440*700MHz*2FMA/cycle*2Flops/FMA.
- 5.6TFlops/rack.

- 512MB/node DDR memory
- 512GB/rack

- 175MB/s = 1.4Gb/sec torus link = 700MHz*2bits/cycle.
- 350MB/s = tree link

Next: QCD on BG/L

The BlueGene/L supercomputer and QCD

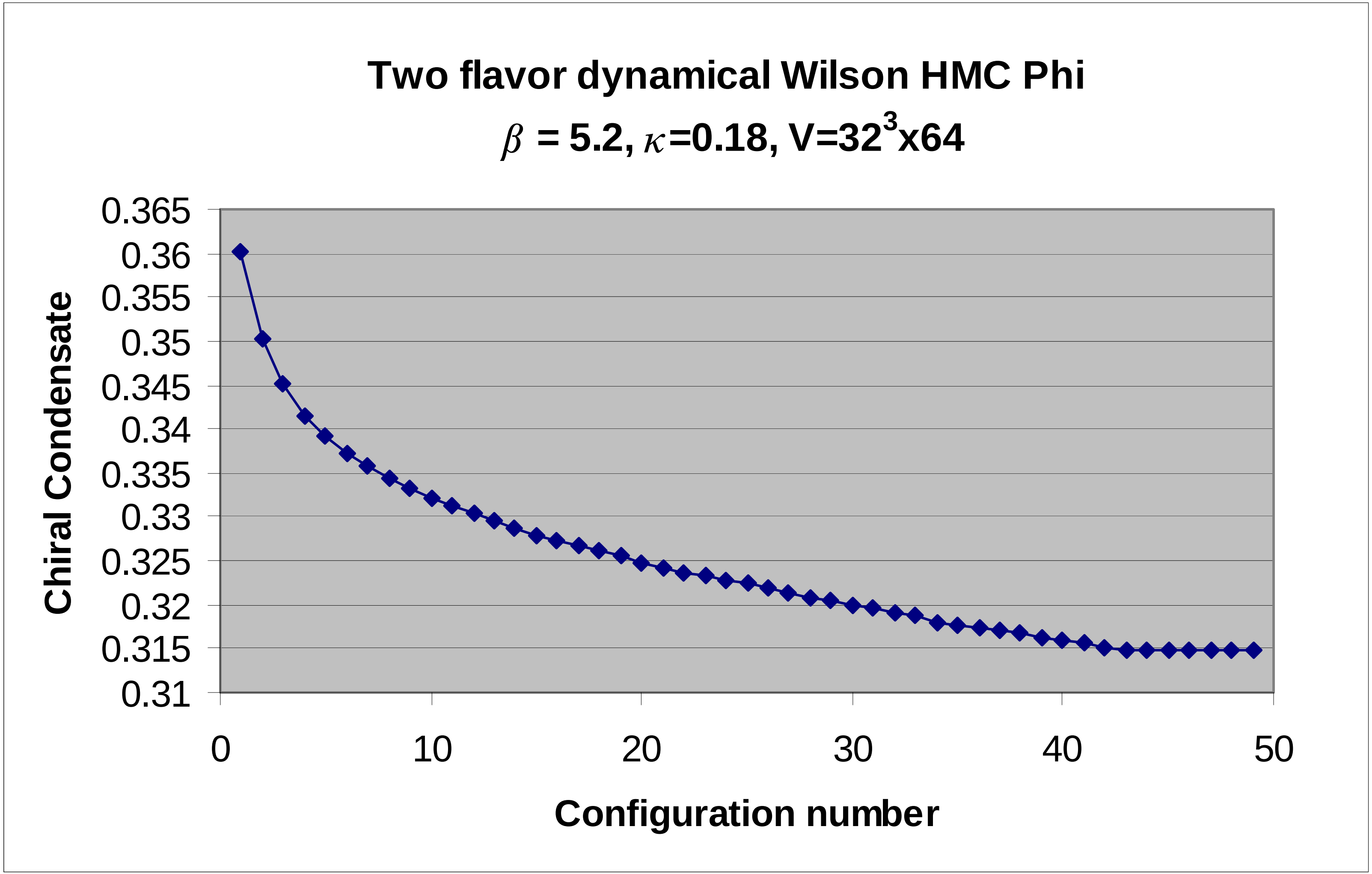
IBM Watson research lab

G. Bhanot, D. Chen, A. Gara, J. Sexton and P. Vranas

The 1 sustained-Teraflops landmark

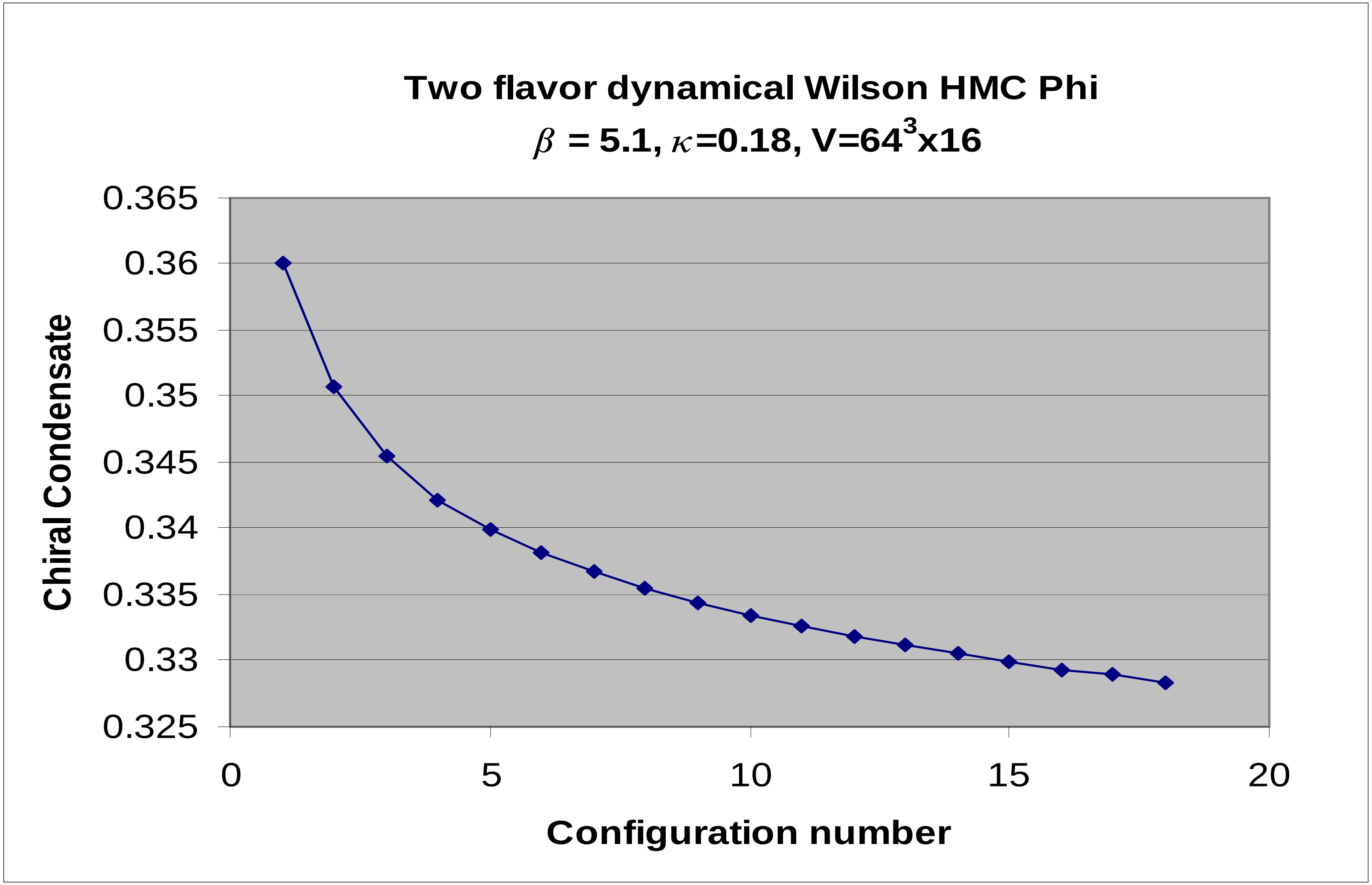
1 sustained-Teraflops for 8.5 hours on 1024 nodes (1 rack)

June 2004

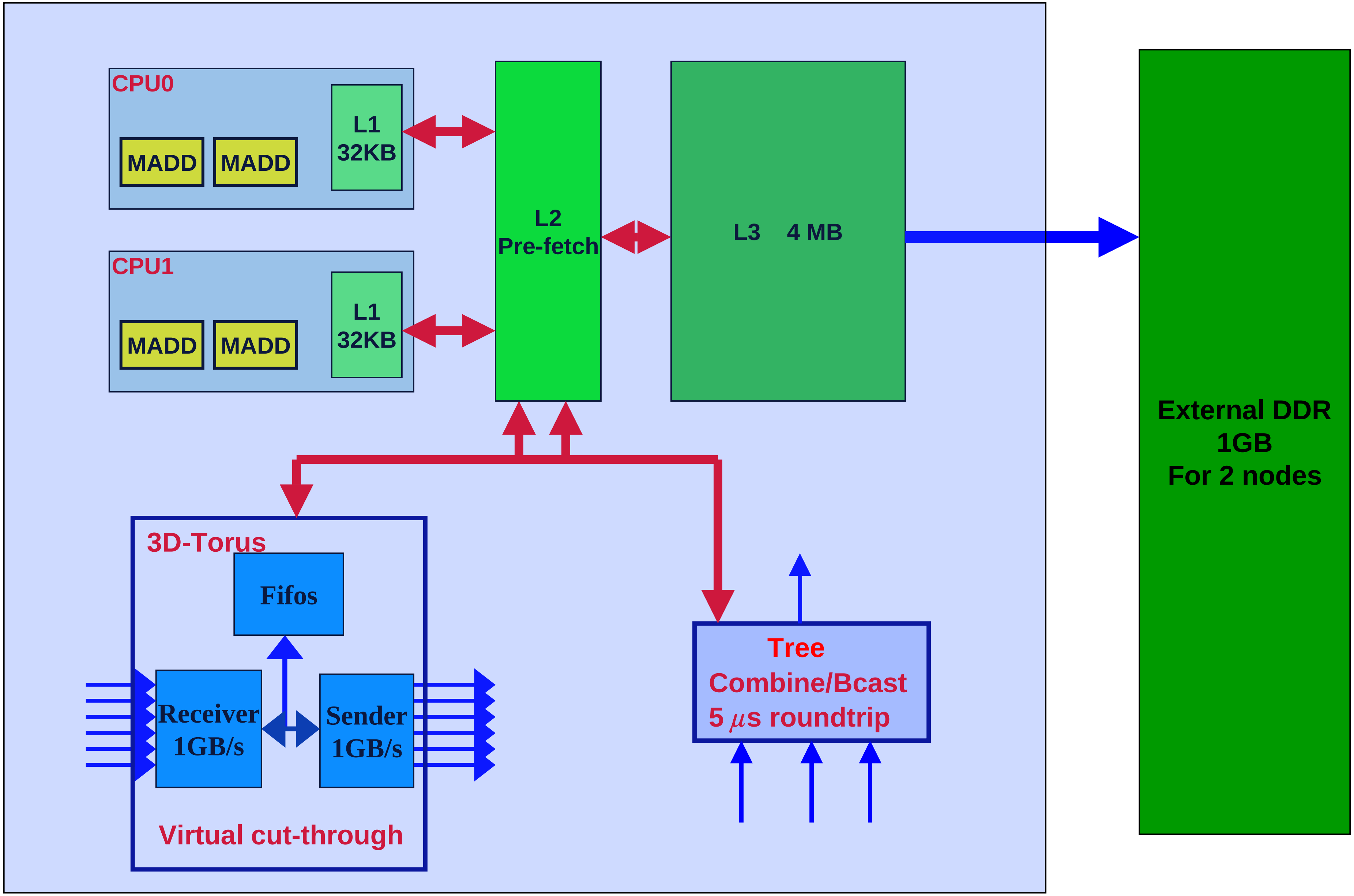


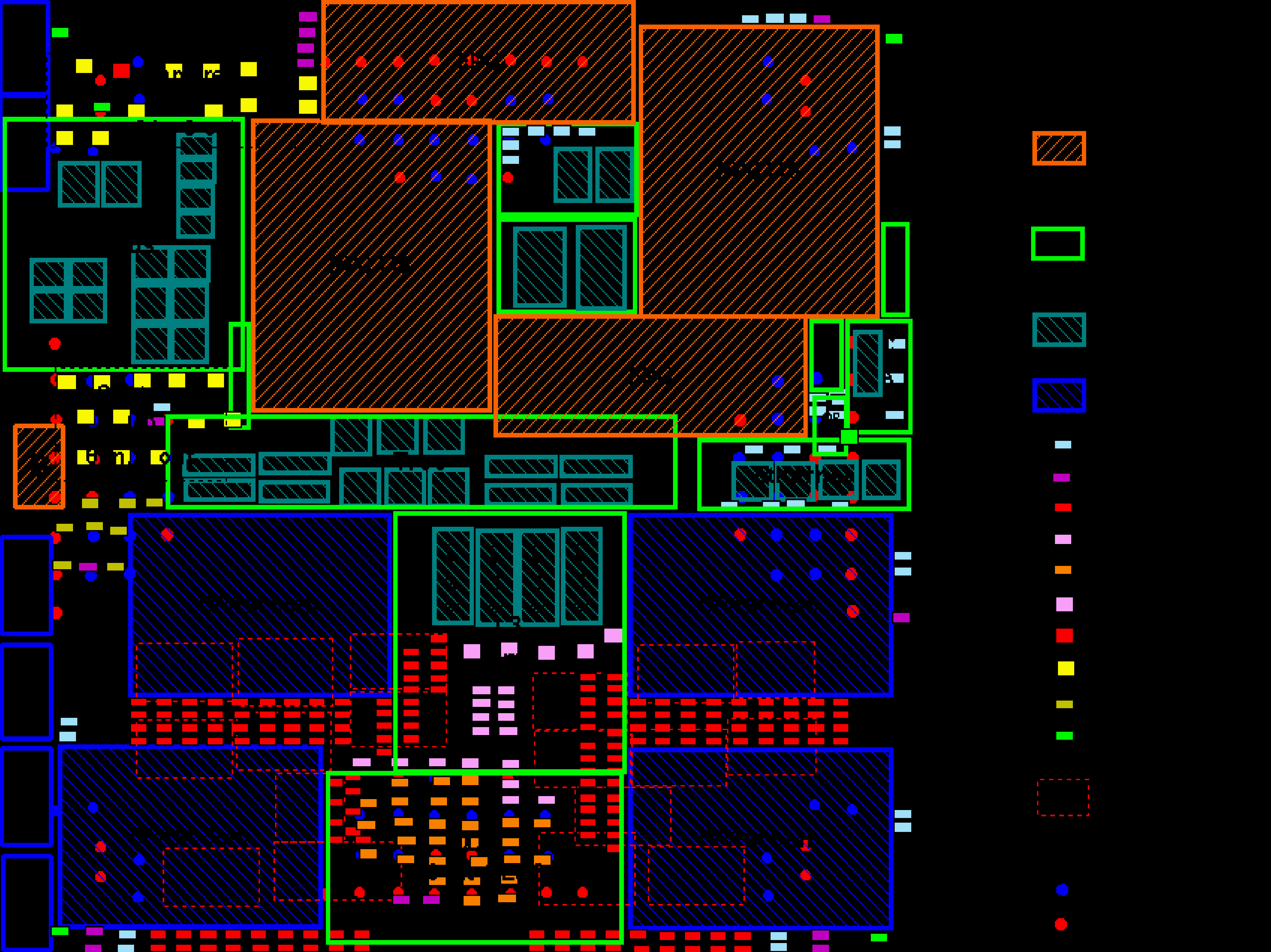
2 sustained-Teraflops for 3 hours on 2048 nodes (2 racks)

June 2004



One chip hardware





BlueGene/L at IBM

- Currently 2K nodes at IBM-Watson and 16K at IBM-Rochester.
- Final installation at LLNL in 2005 with 64K nodes and peak speed of 360 TFlops.
- Final installation at IBM-Watson in 2005 with 20K nodes and peak speed of 112 Tflops.
- Major application in production is MD for life sciences (protein folding).
- Lattice QCD has been one of the hardware diagnostic and validation tools.
- From these hardware diagnostic efforts there are some interesting numbers presented here.

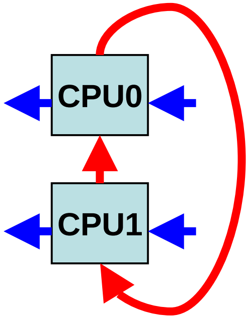
The dynamical QCD simulation

- Less than 10% of the code consumes more than 90% of the cycles. That code is an innermost kernel that simulates the fermionic degrees of freedom.
- Optimizing that kernel (D slash) is very important.
- All degrees of freedom are on a fixed 4-dimensional lattice.
- All communications are usually nearest neighbor and global sums of single floating point number per node.
- If the number of lattice points of each direction is doubled then the calculation will take 2^8 to 2^{10} longer!
- Therefore, to do a calculation with half as small lattice spacing, or twice as long linear dimension, QCD requires a computer that is faster by about a factor of 1000.
- When this factor is crossed a new generation of QCD calculations is signaled:

  Better control over the statistical and systematic errors.

  New problems become accessible.

QCD on the hardware



Virtual node mode:

- CPU0, CPU1 act as independent “virtual nodes”
- Each one does both computations and communications
- The 4-th direction is along the two CPUs
- The two CPU’s communicate via common memory buffers
- Computations and communications can not overlap.
- Peak compute performance is then 5.6 GFlops

Optimized Wilson $\not{D}$ in virtual node mode

- Inner most kernel code is in C/C++ inline assembly.

- Algorithm is similar to the one used in CM2 and QCDSP:
  - Spin project in the 4 “backward” directions
  - Spin project in the 4 “forward” directions and multiply with gauge field
  - Communicate “backward” and “forward” spinors to nn
  - Multiply the “backward” spinors with gauge field and spin reconstruct
  - Spin reconstruct “forward” spinors

- All computations use the double Hummer multiply/add instructions.
- All floating computations are carefully arranged to avoid pipeline conflicts.
- Memory storage ordering is chosen for minimal pointer arithmetic.
- Quad Load/store are carefully arranged to take advantage of the cache hierarchy and the CPUs ability to issue up to 3 outstanding loads.
- Computations almost fully overlap with load/stores. Local performance is bounded by memory access to L3.
- A very thin and effective nearest-neighbor communication layer interacts directly with the torus network hardware to do the data transfers.
- Global sums are done via a fast torus or tree routines.
- Communications do not overlap with computations or memory access.
- Small local size : Fast L1 memory access but more communications
  Large local size: Slower L3 memory access less communications.

Wilson kernel node performance

Spin-projection and even/odd preconditioning

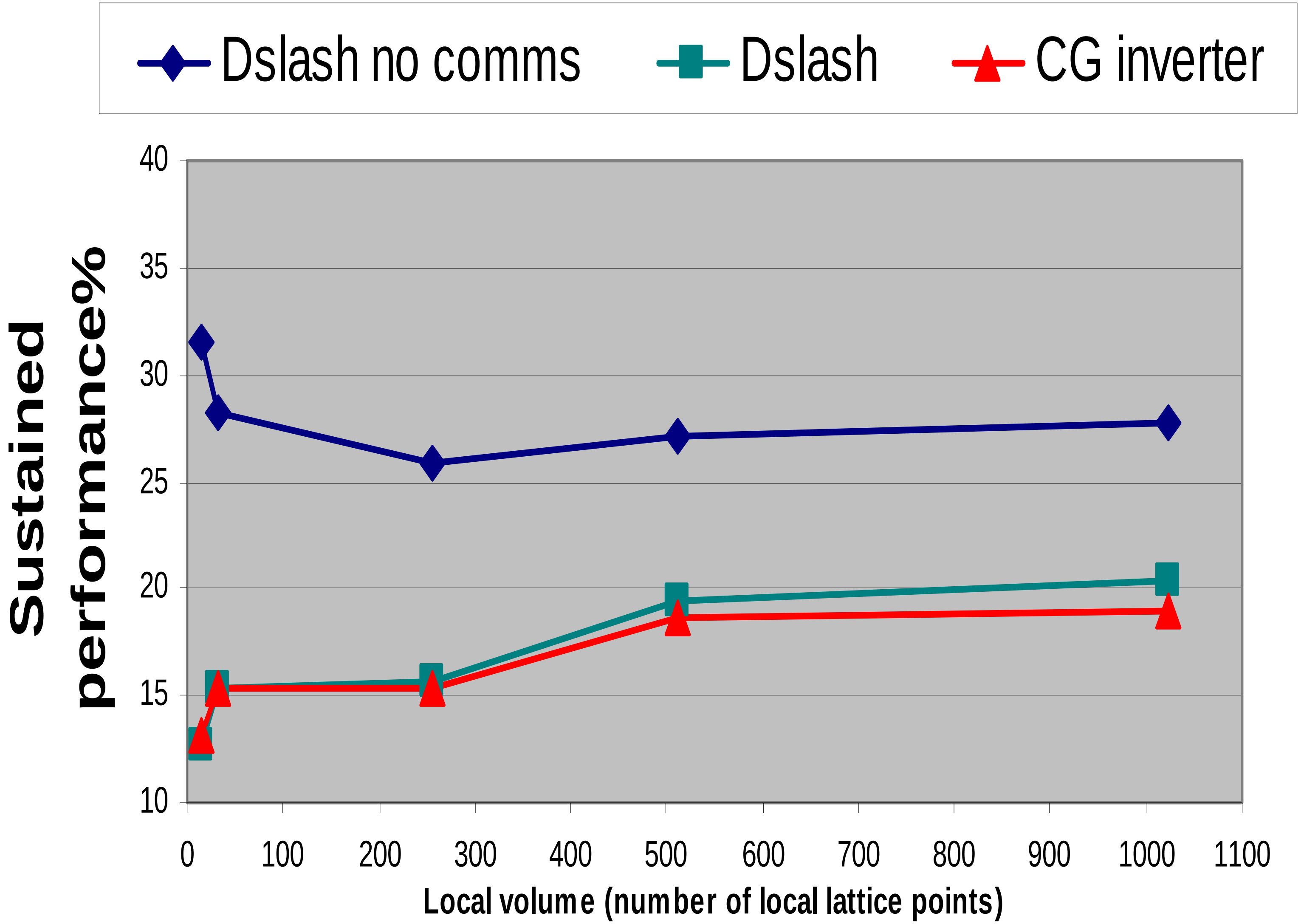
Numbers are for single chip with self-wrapped links

Full inverter (with torus global sum)

%of peak	2^4	4x 2^3	4^4	8 x 4^3	8^2 x 4^2	16 x 4^3
$\not{D}$ no comms	31.5	28.2	25.9	27.1	27.1	27.8
$\not{D}$	12.6	15.4	15.6	19.5	19.7	20.3
Inverter	13.1	15.3	15.4	18.7	18.8	19.0

QCD CG Inverter - Wilson fermions

1 core in torus loopback



Weak Scaling (fixed local size)

Spin-projection and even/odd preconditioning.

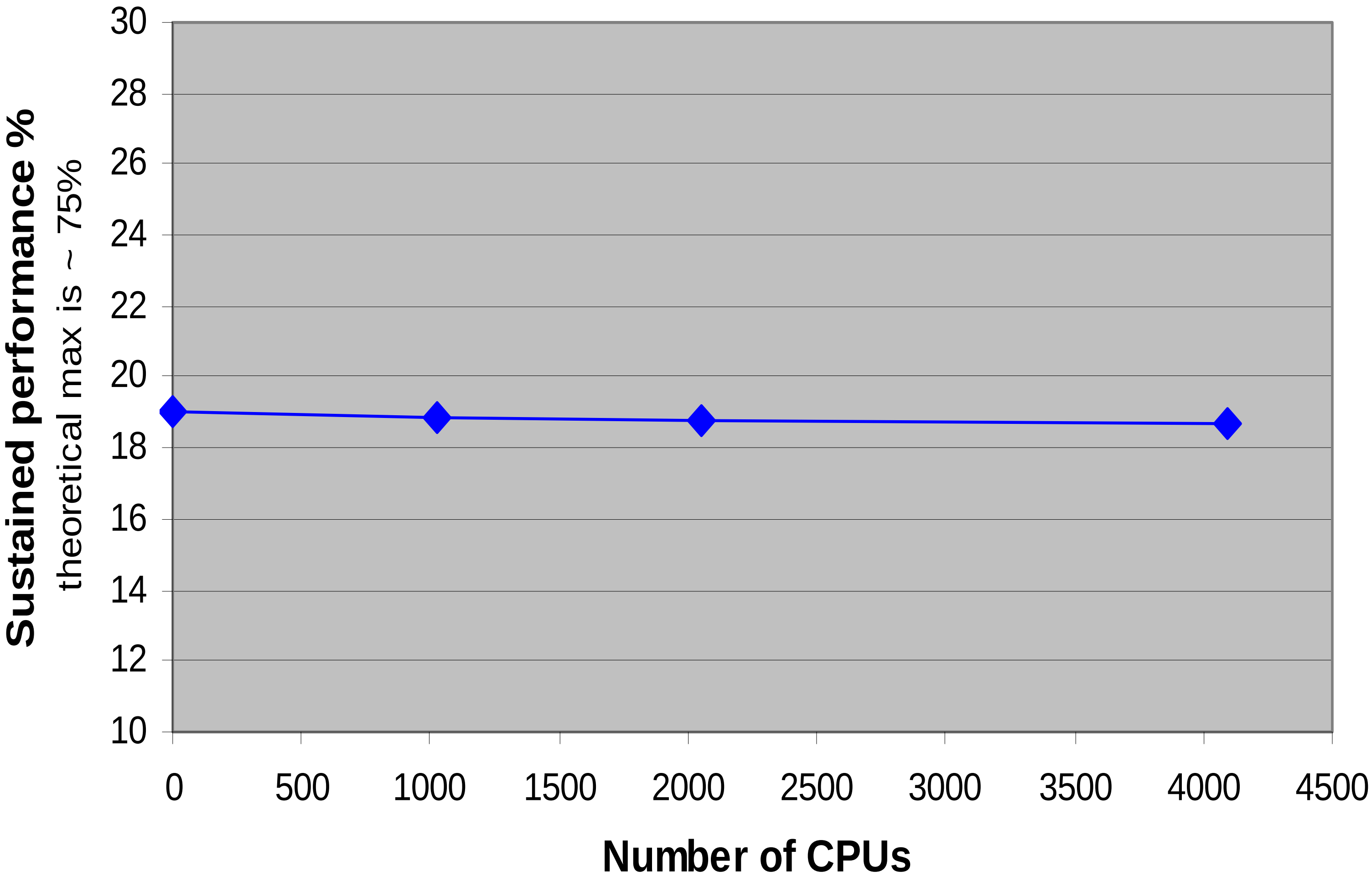
Full inverter (with torus global sum)

16x4x4x4 local lattice. CG iterations = 21.

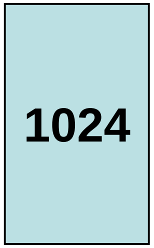
Machine	½ chip	midplane	1 rack	2 racks
Cores	1	1024	2048	4096
Global lattice	4x4x4x16	32x32x32x32	32x32x64x32	32x64x64x32
% of peak	19.0	18.9	18.8	18.7

QCD CG Inverter - Wilson fermions

21 CG iterations, 16x4x4x4 local lattice



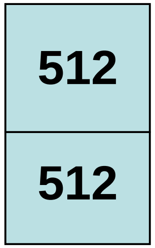
Some interesting machine configurations



1024

Nodes: 8 x 8 x 16 x 2
Local : 4 x 4 x 4 x 16
= 32 x 32 x 64 x 32

Zero temperature dynamical



512

512

Nodes: 8 x 8 x 8 x 2
Local : 4 x 4 x 4 x 4
= 32 x 32 x 32 x 8

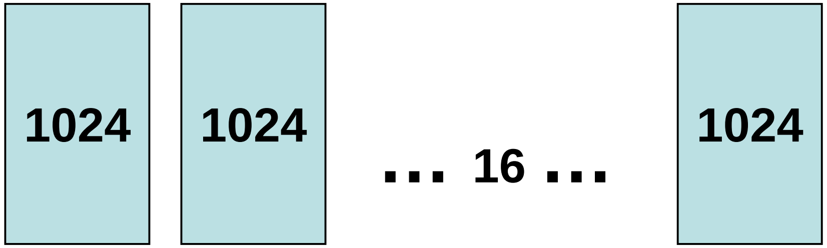
Thermo – hot start

Nodes: 8 x 8 x 8 x 2
Local : 4 x 4 x 4 x 4
= 32 x 32 x 32 x 8

Thermo – cold start

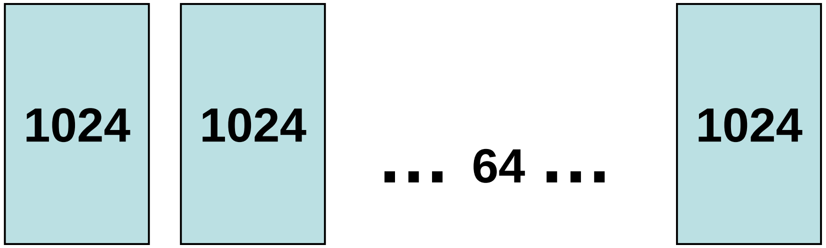
Dream machines

IBM Watson: 20 racks = 112 Teraflops peak



Nodes: 16 x 32 x 32 x 2
Local : 4 x 2 x 2 x 32
= 64 x 64 x 64 x 64

LLNL: 64 racks = 358 Teraflops peak



Nodes: 64 x 32 x 32 x 2
Local : 2 x 2 x 2 x 32
= 128 x 64 x 64 x 64

Nodes: 64 x 32 x 32 x 2
Local : 2 x 4 x 4 x 64

= 128 x 128 x 128 x 128
= Well, almost…

Full QCD physics system

- The physics code (besides the Wilson Dslash) is the Columbia C++ physics system (cps).
- The full system ported very easily and worked immediately.
- The BG/L additions/modifications to the system have been kept isolated.

Acknowledgements

We would like to thank the QCDOC collaboration for useful discussions and for providing us with the Columbia physics system software.

Conclusions

- QCD crossed the 1 sustained-Teraflops landmark in June 2004.
- In the next year, because of analytical and supercomputer developments dynamical QCD will likely get to L/a = 32 at physical quark masses and perhaps even to L/a ~ 64 and maybe even more…