



Argonne
NATIONAL
LABORATORY

... for a brighter future



U.S. Department
of Energy

UChicago ►
Argonne_{LLC}



A U.S. Department of Energy laboratory
managed by UChicago Argonne, LLC

The ALCF Blue Gene/P System Overview

Raymond Loy

Vitali Morozov

*Applications Performance Engineering and Data
Analytics (APEDA)*

Argonne Leadership Computing Facility

With special thanks to IBM:

Rajiv Bendale, Kirk Jordan,
Jerrold Heyman, Carlos Sosa,
Bob Walkup

DOE Leadership Computing Facility Strategy

- DOE SC selected the ORNL, ANL and PNNL team (May 12, 2004) based on a competitive peer review of 4 LCF proposals
 - ORNL will deploy a series of systems based on Cray's XT3/4 architectures @ 250TF/s in FY07 and 1000TF/s in FY08/9
 - ANL will develop a series of systems based on IBM's BlueGene @ 100TF/s in FY07 and 250-500TF/s in FY08/FY09 with IBM Blue Gene/P
 - PNNL will contribute software technology
- DOE SC will make these systems available as capability platforms to the broad national community via competitive awards (e.g. INCITE Allocations)
 - Each facility will target ~20 large-scale production applications teams
 - Each facility will also support development users
- DOE's LCFs complement existing and planned production resources at NERSC
 - Capability runs will be migrated to the LCFs, improving NERSC throughput
 - NERSC will play an important role in training and new user identification

Mission and Vision for the ALCF

Our Mission

Provide the computational science community with a world leading computing capability dedicated to breakthrough science and engineering.

Our Vision

A world-class center for computation driven scientific discovery that has:

- outstandingly talented people,
- the best collaborations with computer science and applied mathematics,
- the most capable and interesting computers and,
- a true spirit of adventure.

See <http://www.alcf.anl.gov/> for info and openings

ALCF Timeline

2004

- Formed of the Blue Gene Consortium with IBM
- DOE-SC selected the ORNL, ANL and PNNL team for Leadership Computing Facility award

2005

- Installed 5 teraflops Blue Gene/L for evaluation

2006

- Began production support of 6 INCITE projects, with BGW
- Continued code development and evaluation
- “Lehman” Peer Review of ALCF campaign plans

2007

- Increased to 9 INCITE projects; continued development projects
- Installed 100 teraflops BlueGene/P (late 2007)

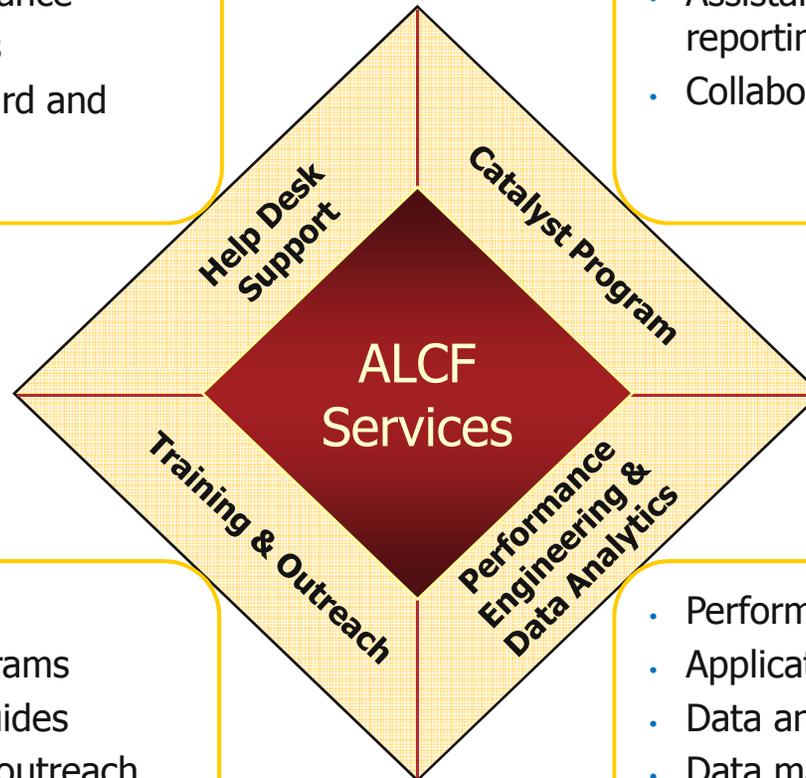
2008

- Began support of 20 INCITE projects on BG/P
- Added 557 Teraflops BG/P

ALCF Service Offerings

- Startup assistance
- User administration assistance
- Job management services
- Technical support (Standard and Emergency)

- ALCF science liaison
- Assistance with proposals, planning, reporting
- Collaboration within science domains



- Workshops & seminars
- Customized training programs
- On-line content & user guides
- Educational and industry outreach programs

- Performance engineering
- Application tuning
- Data analytics
- Data management services

ALCF's Blue Gene/P Intrepid: Ranked #3 in TOP500

- Ranked third fastest overall worldwide.
- Fastest supercomputer in the world for open science, according to the semiannual Top500 List of the world's fastest computers.
- Peak performance of 557 Teraflops (557 trillion calculations per second), speed of 450.3 Teraflops on the Linpack application used to measure speed for the rankings. [80.8% of peak]
- Currently, one-fifth of the system is in production, providing 111 million processor hours of computing time to 20 INCITE projects.
- In 2009, Intrepid will be in full production mode, providing 500 million processor hours for INCITE research.



Summary: BG/P vs BG/L

- Increased clock
 - 1.2x from frequency bump 700 MHz => 850 MHz
- Processor density
 - Double the processors/node (4 vs. 2)
- Memory
 - higher bandwidth
 - Cache coherency
 - *Allows 4 way SMP*
 - supports OpenMP, pthreads
 - DMA for torus
- Faster communication
 - 2.4x higher bandwidth, lower latency for Torus and Tree networks
- Faster I/O
 - 10x higher bandwidth for Ethernet I/O
- Enhanced performance counters
- Inherited architectures
 - double Hummer FPU, torus, collective network, barrier

Recap: Blue Gene/P key architectural improvements over BG/L

Property		Blue Gene/L	Blue Gene/P
Node Properties	Processor cores/chip	two 440 PowerPC	four 450 PowerPC
	Processor Frequency	0.7GHz	0.85GHz
	Coherency	Software managed	SMP with snoop filtering
	L1 Cache (private)	32KB I + 32KB D/proc.	32KB I + 32KB D/proc.
	L2 Cache (private)	15 line buffers	15 line buffers
	L3 Cache size (shared)	4MB	8MB
	Main Store	512 MB and 1 GB DDR	2 GB DDR2
	Main Store Bandwidth	5.6 GB/s (16B wide)	13.6 GB/s (2*16B wide)
	Peak Performance	5.6 GFLOPs/node	13.6 GFLOPs/node
Torus Network	Aggregate Bandwidth	6*2*175 MB/s=2.1 GB/s	6*2*425 MB/s= 5.1 GB/s
	Hardware Latency (Nearest Neighbor)	<1μs	<1μs
Collective Network	Aggregate Bandwidth	3*2*350 MB/s=2.1 GB/s	3*2*0.85 GB/s=5.1 GB/s
Performance Monitors	Counters	48	256
	Counter resolution (bits)	32b	64b
GFLOPS/Watt		0.22	0.33

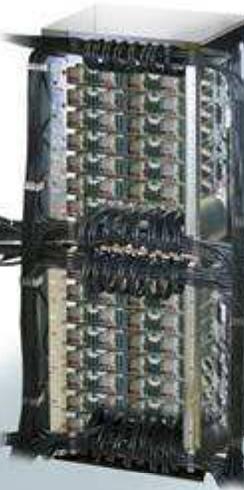
Blue Gene/P

System



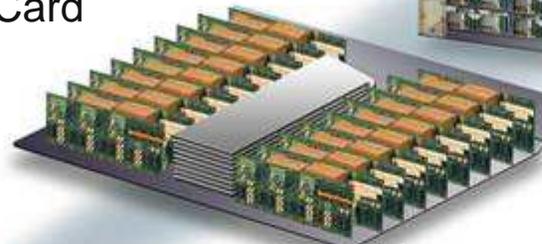
72 Racks
73728 nodes
294912 cores
1 PF
144 TB DDR2

Rack



32 node-cards
1024 nodes
4096 cores
13.9 TF
2 TB DDR2

Node Card



32 nodes
128 cores
435 GF
64GB DDR2

Compute Card



1 node
4 cores
13.6 GF
2GB DDR2

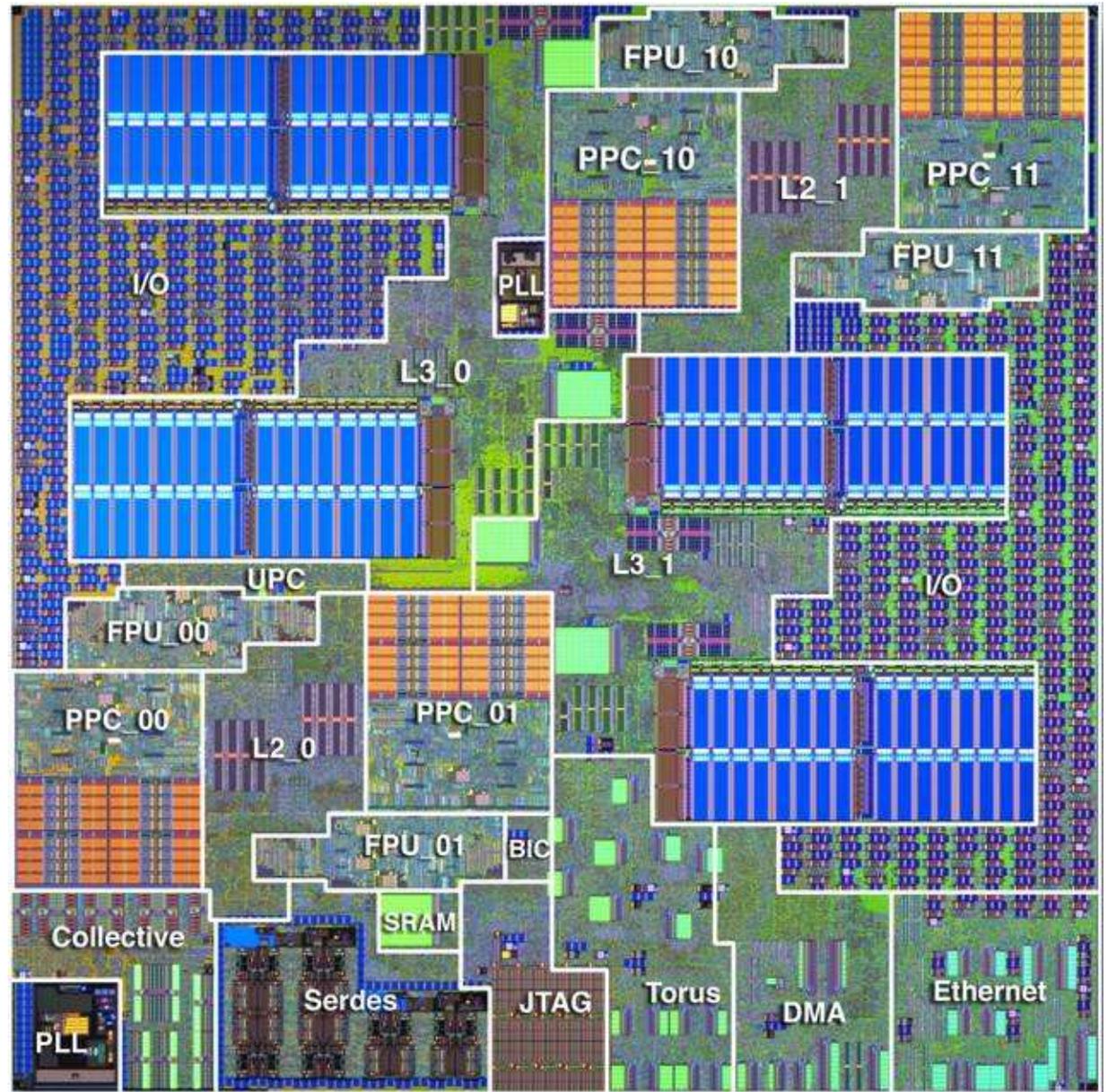
Chip



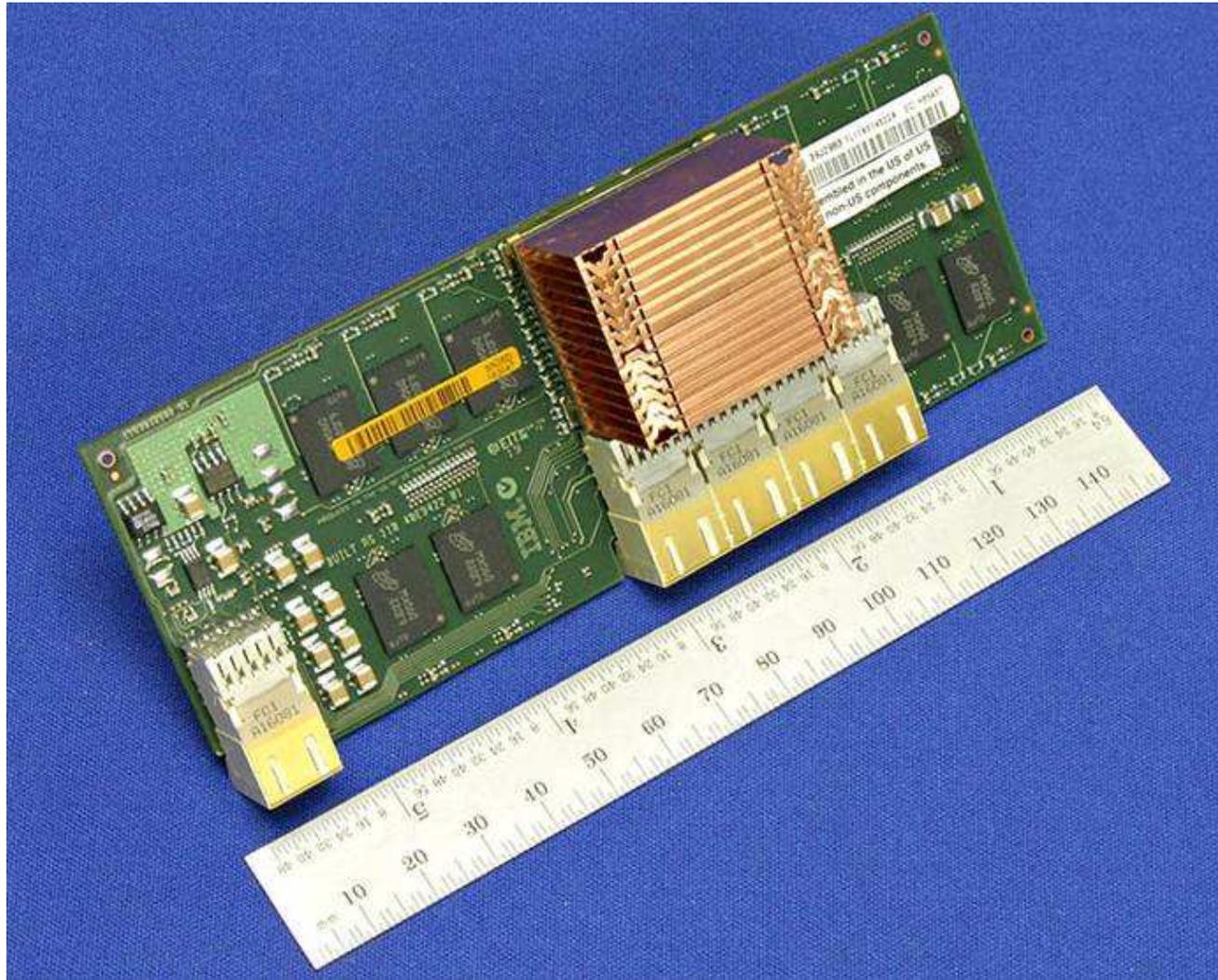
4 cores
13.6 GF

BPC chip DD2.1 die photograph

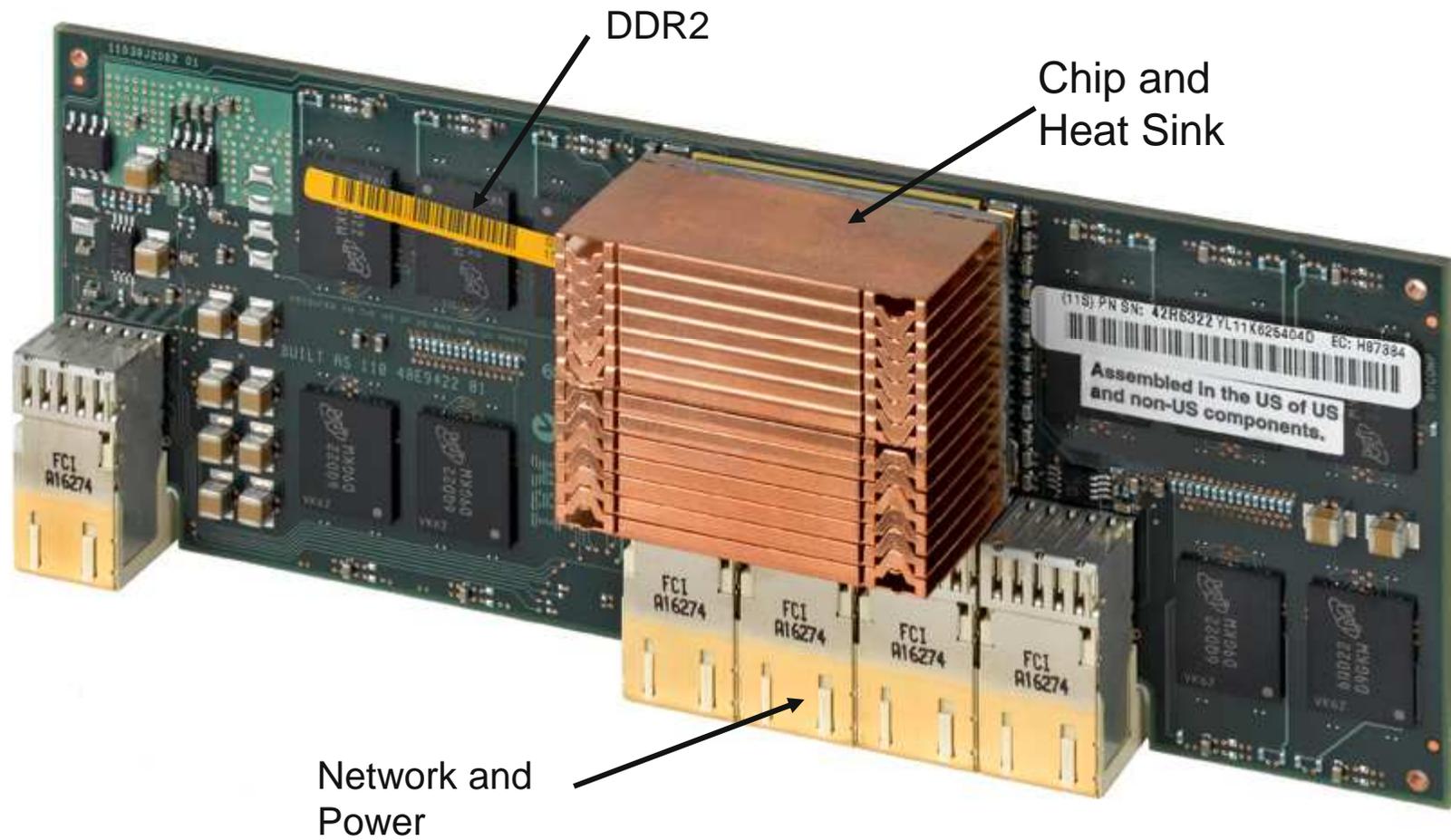
13mmx13mm
90 nm process
208M transistors
88M in eDRAM



BG/P Compute Card



BG/P Compute Card



BGP Node Card

32 Compute nodes



Optional IO card
(one of 2 possible)
with 10Gb optical link

Local DC-DC
regulators
(6 required, 8 with
redundancy)

32 Compute Nodes

128 cores

Hottest ASIC Tj
80°C@24W, 55°C@15W

Outlet Air
Max +10°C

Inlet Air
min 2.5m/s
max 17°C

Hottest DRAM
Tcase 75°C@0.3W

Optional IO card
(1 of 2 possible)

10Gb Ethernet

Local 48V input DC-DC regulators
5+1, 3+1 with redundancy. Vicor
technology, tcase 60°C@120A

First BG/P Rack



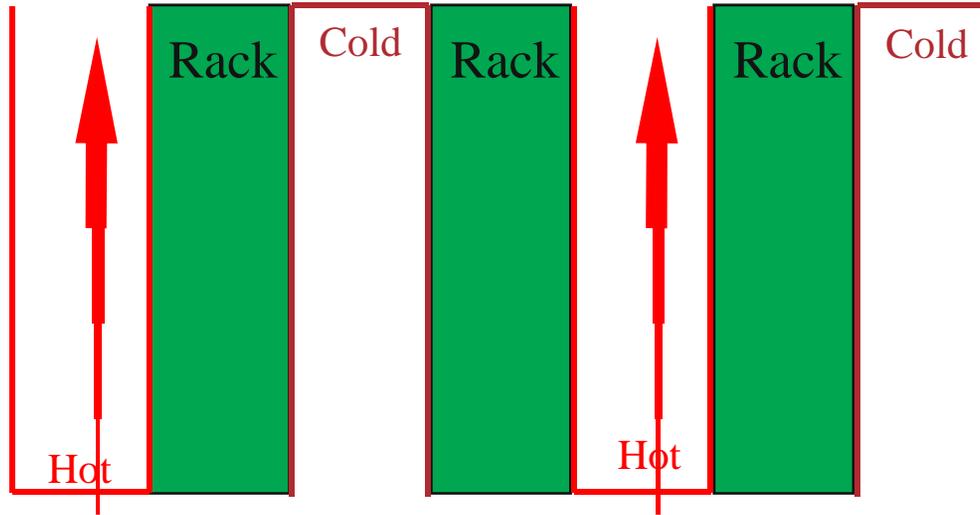
First 8 racks of BG/P: Covers removed



IBM Blue Gene/P



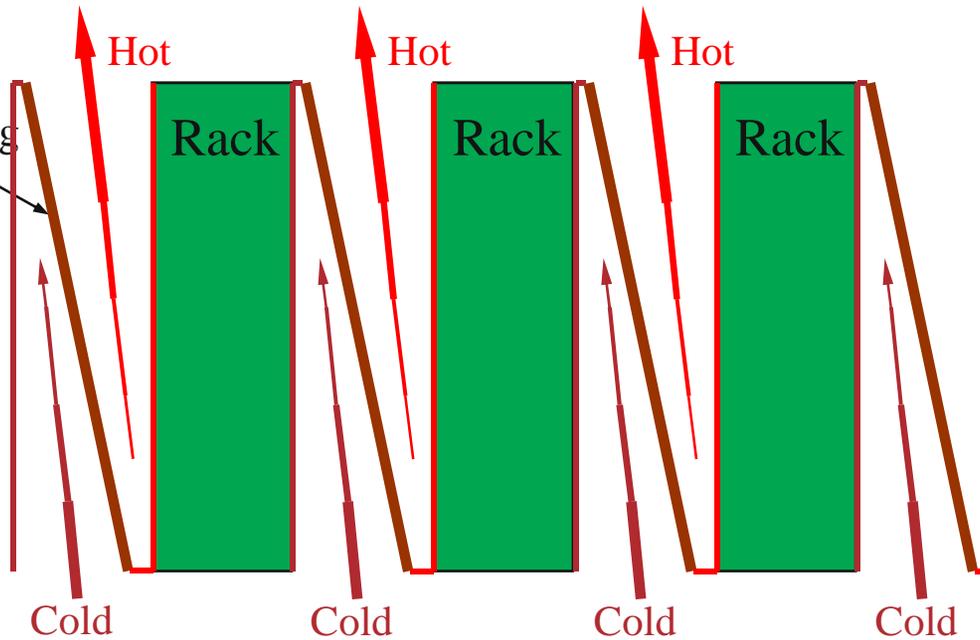
etc.



**(a) Prior Art:
Segregated,
Non-Tapered
Plenums**
(Plenum Width Same
Regardless of Flow Rate)

Thermal-Insulating
Baffle

etc.



**(b) Invention:
Integrated,
Tapered
Plenums**
(Plenum Width
Larger where Flow
Rate is Greater)

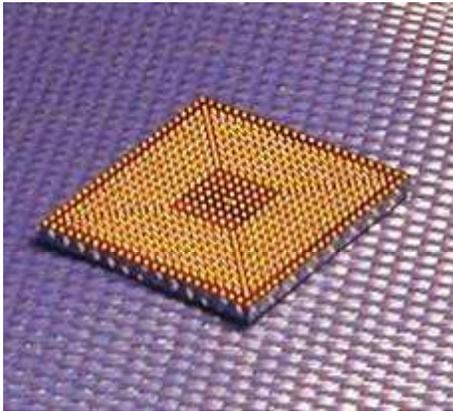
Shawn Hall 4-3-02
02-04-03 Angled Plenums

Power Efficient Computing

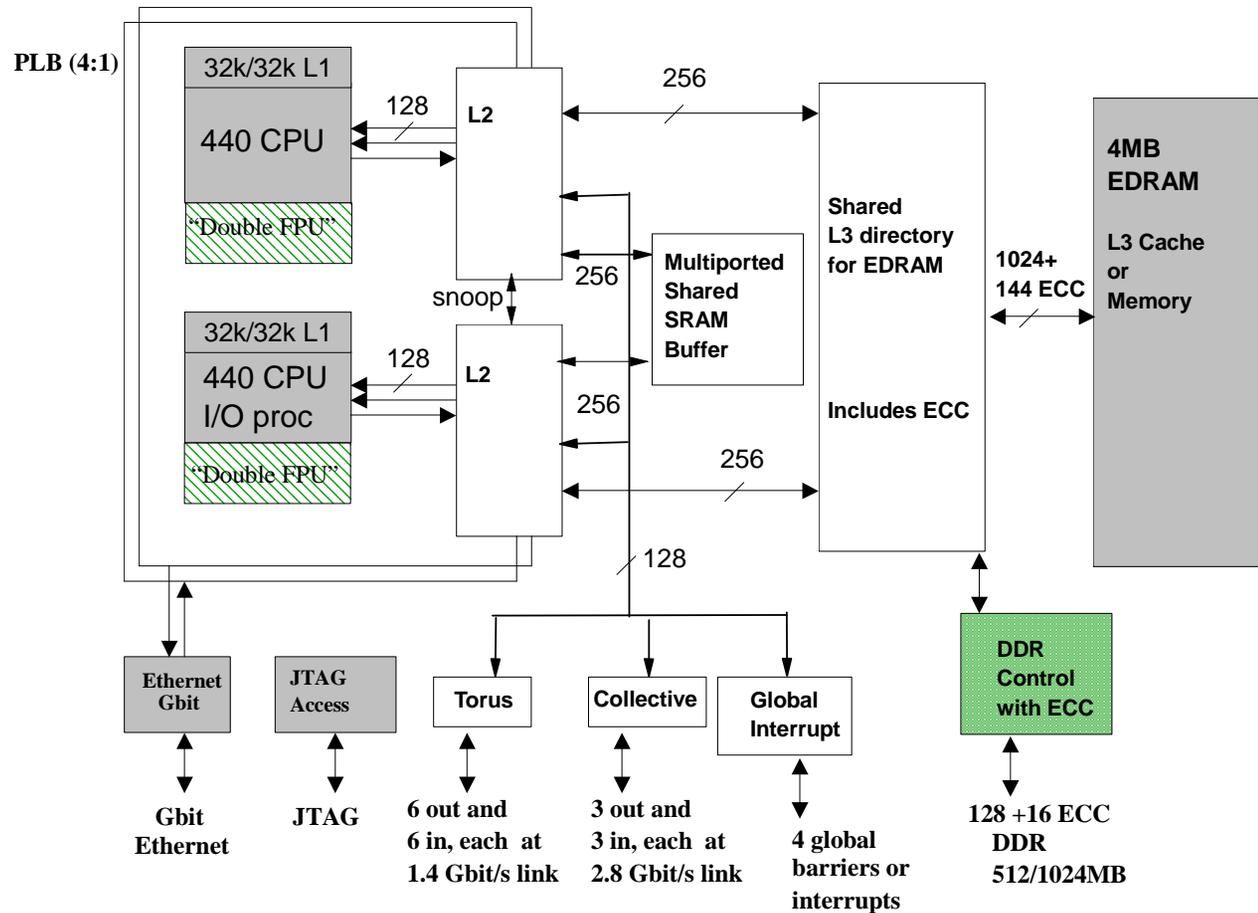
- Blue Gene/P 372 MFlops/Watt
 - Compare Blue Gene/L 210 MFlops/Watt
 - Only exceeded by IBM QS22 Cell processor (488Mflop/s/Watt)
- Single rack
 - Idle 8.6KW
 - Avg 21KW
 - Linpack 29KW

Memory Subsystem

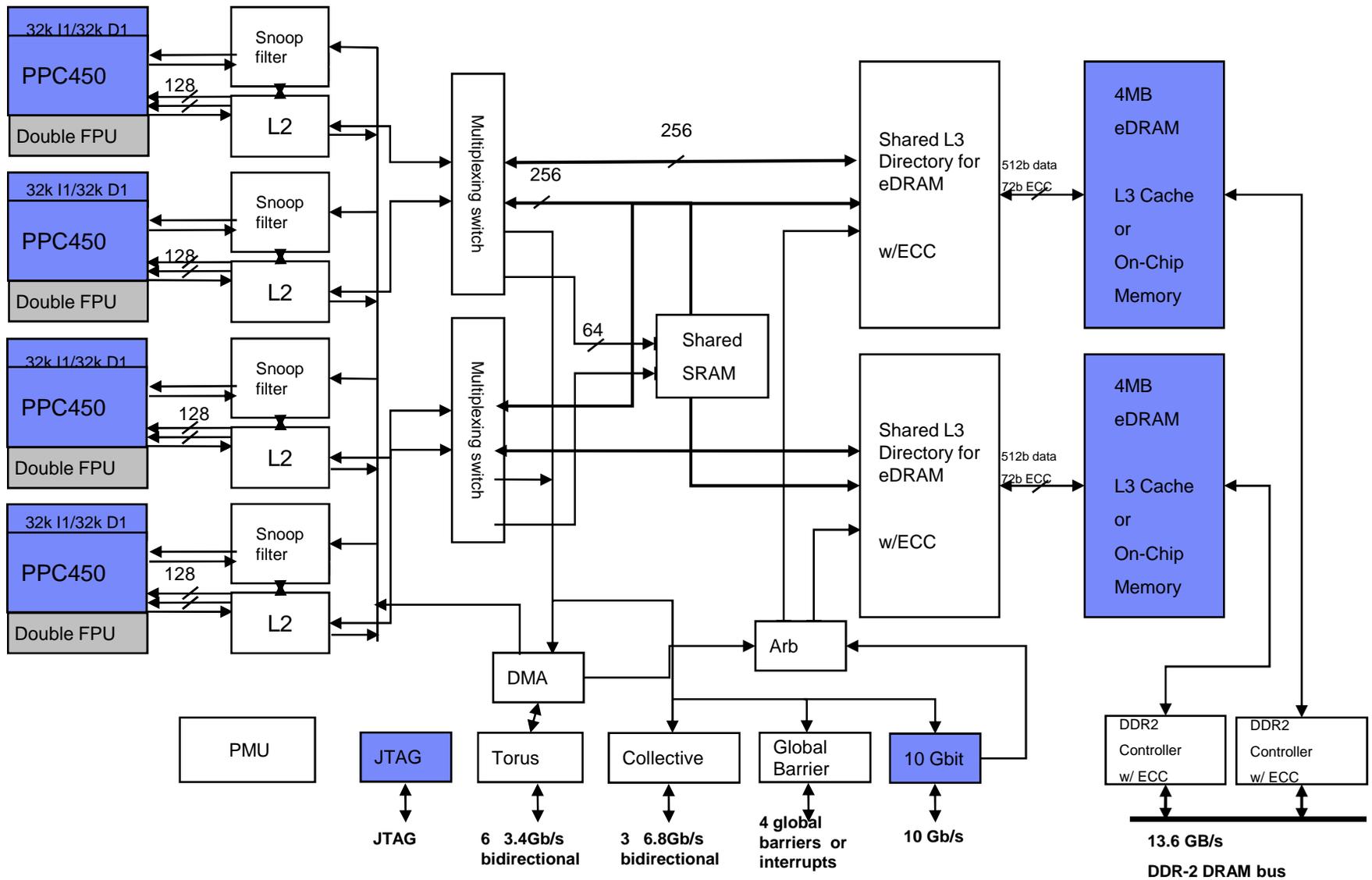
Blue Gene/L ASIC



- 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



Blue Gene/P ASIC



L1 Cache

– Architecture

- *32KB Instruction-cache, 32KB Data-cache per core*
- *32Byte lines, 64 way-set-associative, 16 sets*
- *Round-robin replacement*
- *Write-through mode*
- *No write allocation*

– Performance

- *L1 load hit => 8Bytes/cycle, 4 cycle latency (floating point)*
- *L1 load miss, L2 hit => 4.6Bytes/cycle, 12 cycle latency*
- *Store: Write through, limited by external logic to about one request every 2.9 cycles (about 5.5Bytes/cycle peak)*

L2 Cache

Three independent ports:

Instruction read

Data Read = prefetch

Data write

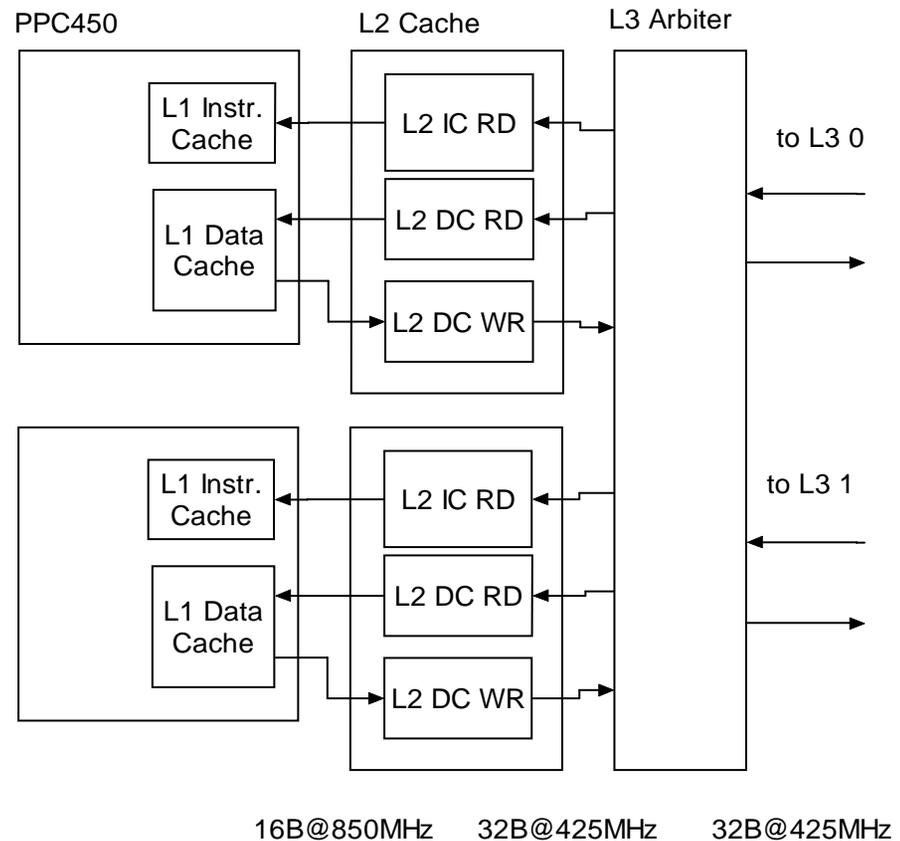
15 lines @128B
(matches L3)

L3 Arbiter

Switch function

1 read and 1 write
switched to each L3
every 425MHz cycle

L2 is only 2KB (<L1)

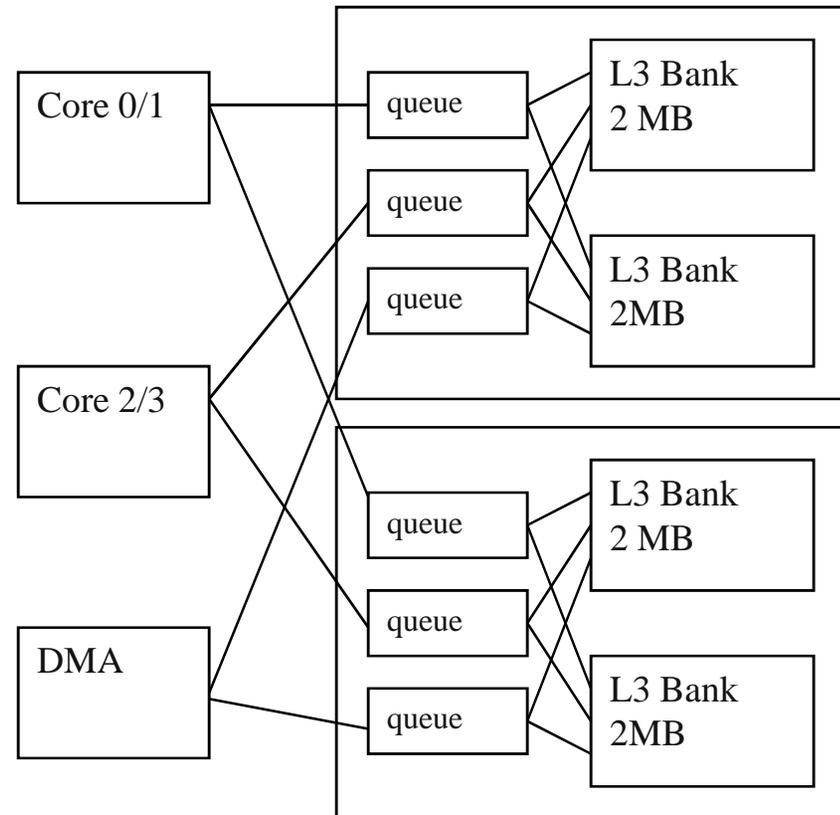


L3 Cache

4 x 2 MB embedded DRAM banks per node (8MB total), each containing:

L3 directory

15 entry 128B-wide write combining buffer



Memory System Bottlenecks

L2 – L3 switch

Not a full core to L3 bank crossbar

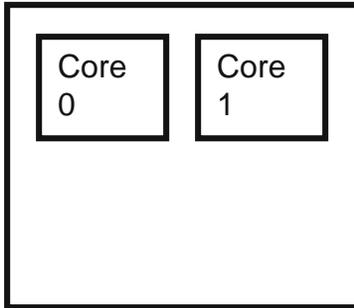
Request rate and bandwidth are limited if two cores of one dual processor group access the same L3 cache bank

Banking for DDR2

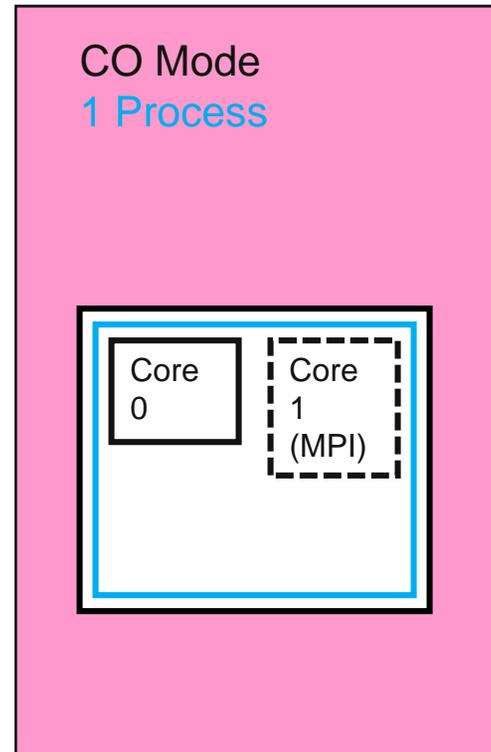
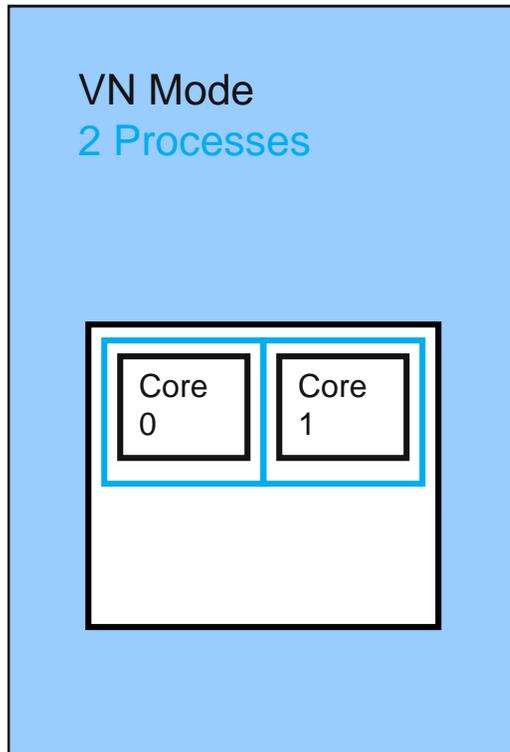
4 banks on 512Mb DDR modules

Peak bandwidth only achievable if accessing 3 other banks before accessing the same bank again

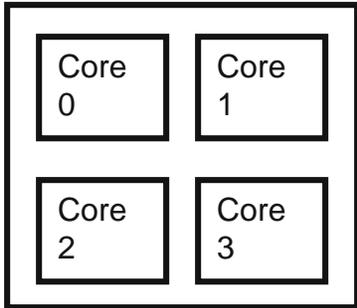
Execution Modes in BG/L



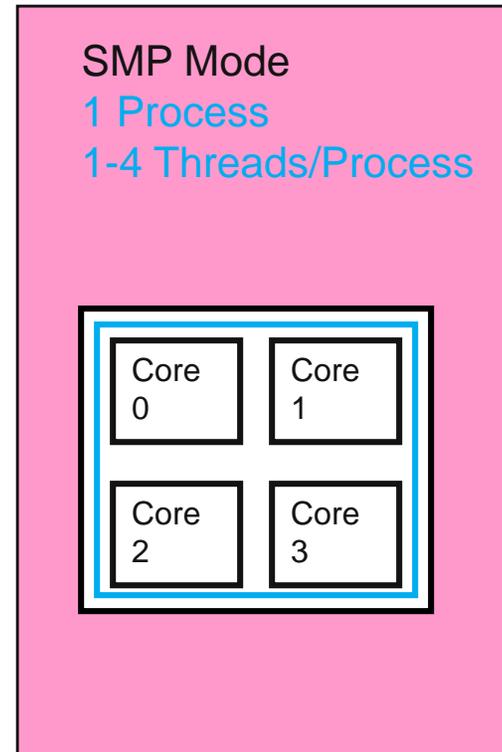
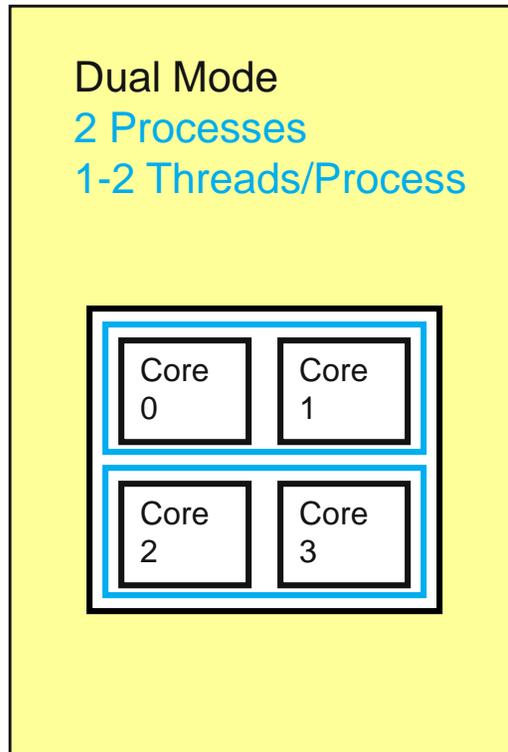
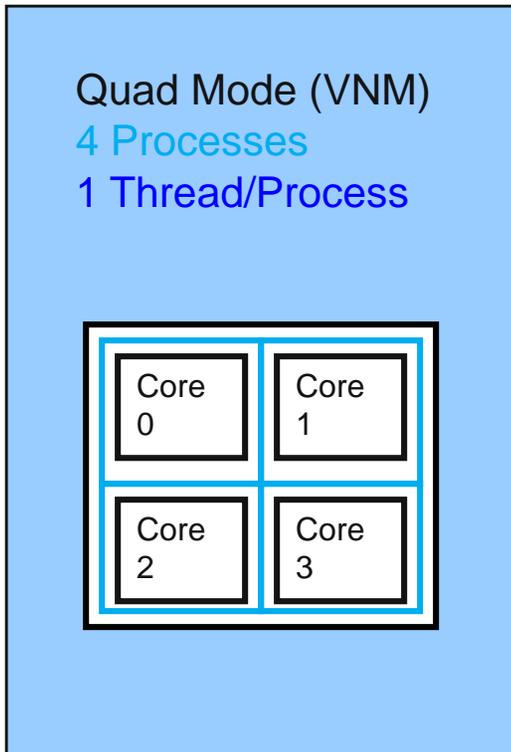
Hardware Elements Black
Software Abstractions Blue



Execution Modes in BG/P

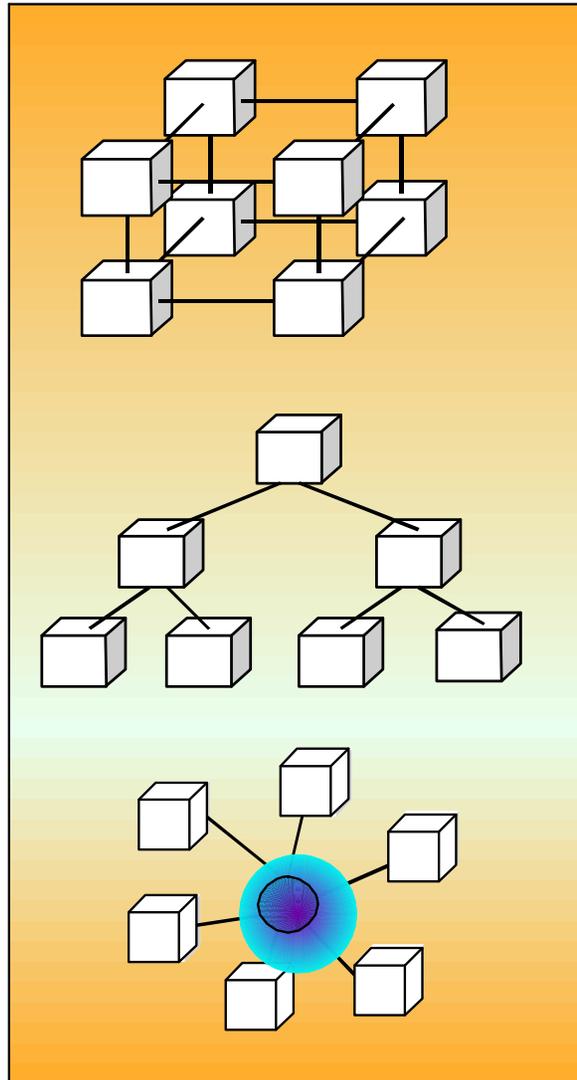


Hardware Elements Black
Software Abstractions Blue



Communication subsystem

Blue Gene/P Interconnection Networks



3 Dimensional Torus

- Interconnects all compute nodes
- Virtual cut-through hardware routing
- 3.4 Gb/s on all 12 node links (5.1 GB/s per node)
- 0.5 μ s latency between nearest neighbors, 5 μ s to the farthest
- MPI: 3 μ s latency for one hop, 10 μ s to the farthest
- Communications backbone for point-to-point
- *Requires half-rack or larger partition*

Collective Network

- One-to-all broadcast functionality
- Reduction operations for integers and doubles
- 6.8 Gb/s of bandwidth per link per direction
- Latency of one way tree traversal 1.3 μ s, MPI 5 μ s
- Interconnects all compute nodes and I/O nodes

Low Latency Global Barrier and Interrupt

- Latency of one way to reach 72K nodes 0.65 μ s, MPI 1.6 μ s

Blue Gene/P Torus Network

Logic Unchanged from BG/L, *except*

Bandwidth

BG/L:	clocked at $\frac{1}{4}$ processor rate	1Byte per 4 cycles
BG/P:	clocked at $\frac{1}{2}$ processor rate	1Byte per 2 cycles

With frequency bump from 700 MHz to 850 MHz

BG/P Links are 2.4x faster than BG/L

425 MB/s vs **175** MB/s

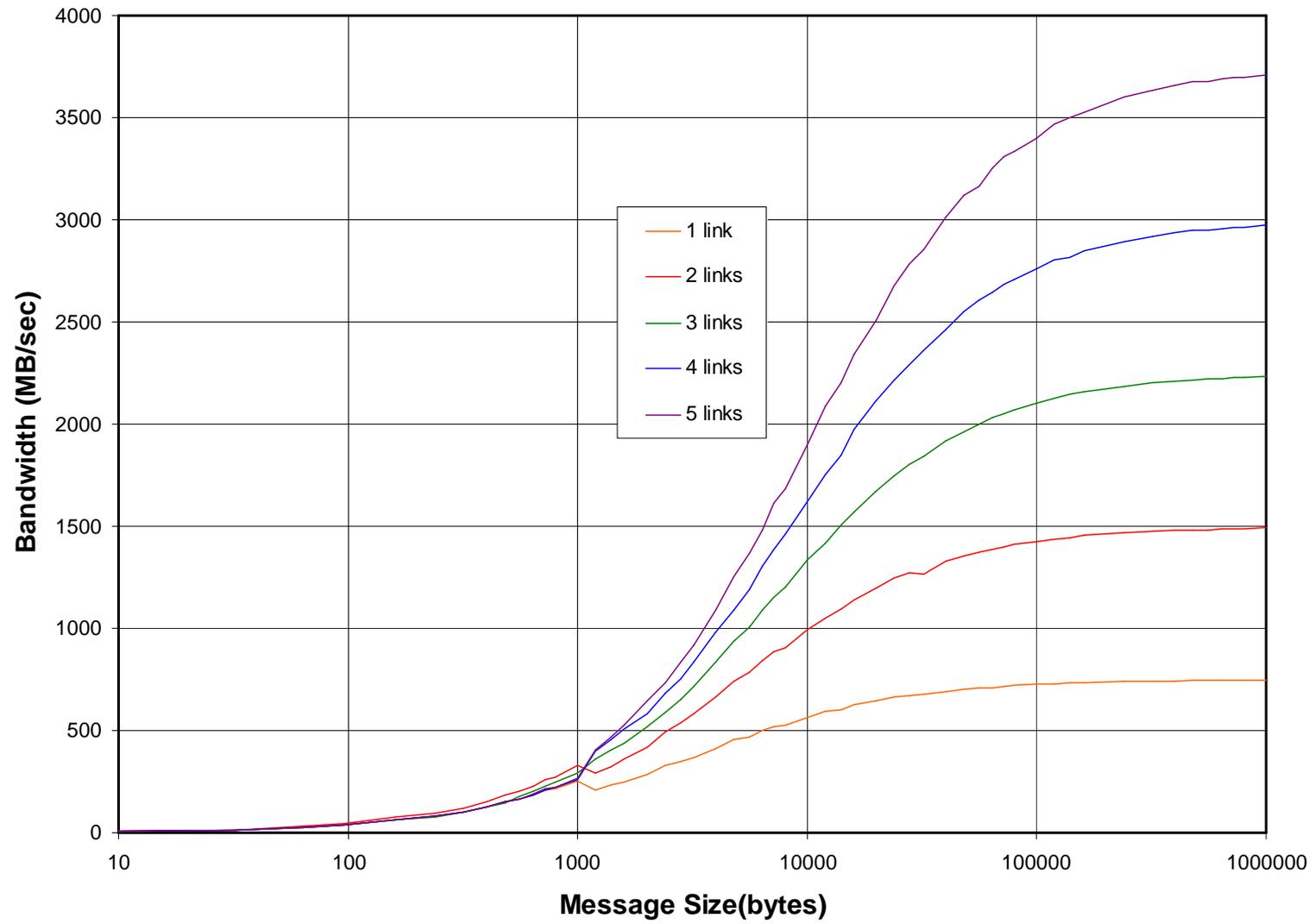
Same Network Bandwidth per Flops as BG/L

Primary interface is via DMA, rather than cores

Run application in DMA mode, or core mode (not mixed)

Software product stack uses DMA mode

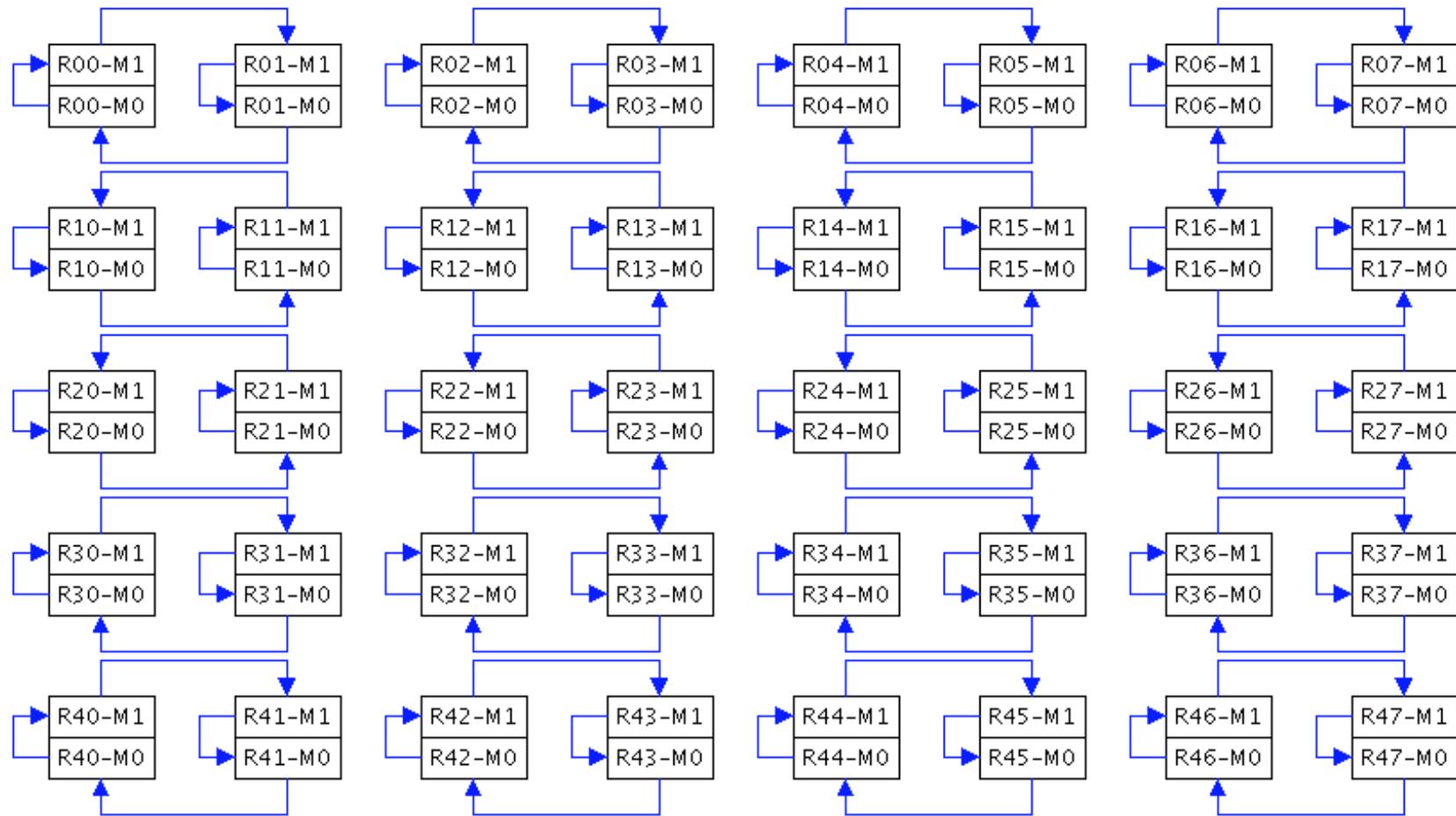
BGP Exchange Bandwidth



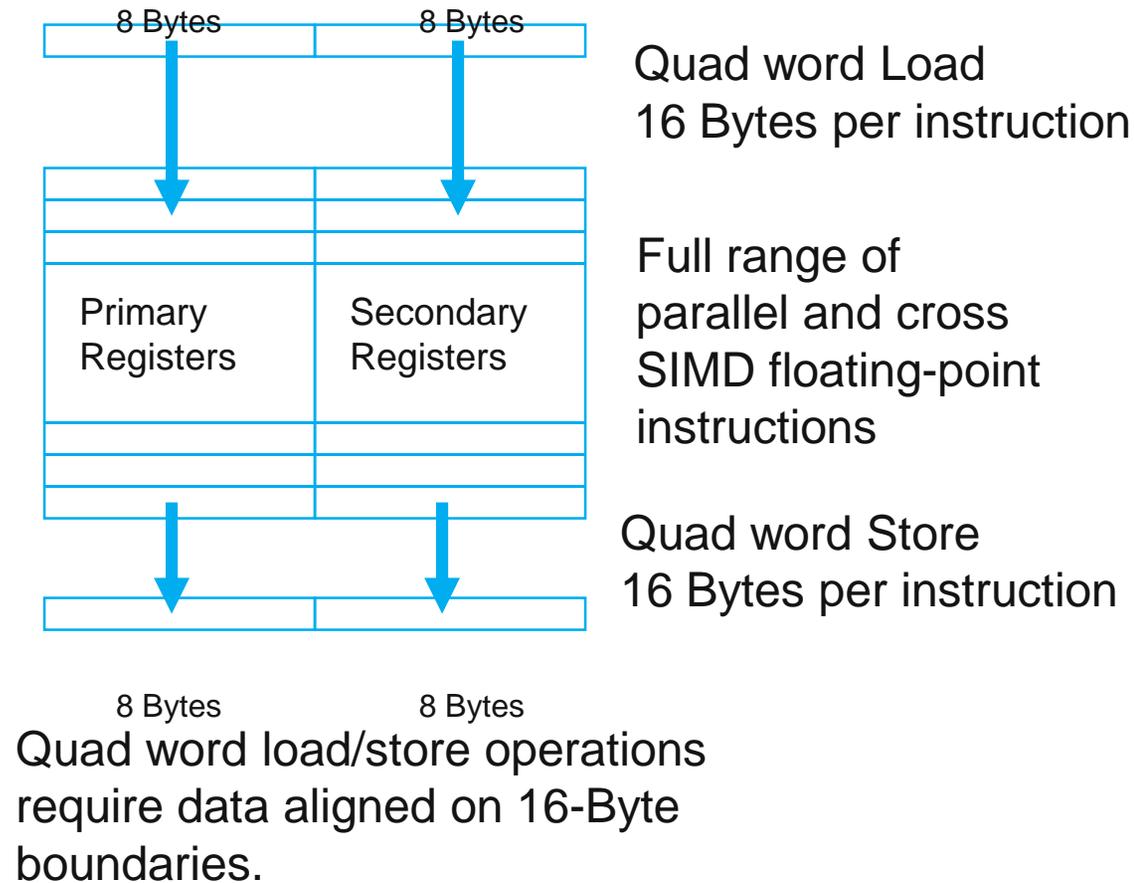
Torus Network Limitations (ALCF)

- Cabling of the ALCF BG/P
 - enables large partition configurations
 - puts some restrictions on small configurations
- Torus “Z” dimension spans pairs of racks
 - Note: half rack or more uses torus network
 - For single rack (1024 node) job, torus HW in adjacent rack is put in “passthrough” mode, looping back without its nodes participating
 - *Prevents a 1024 node job in that adjacent rack*
- Likewise
 - 4-rack (4096 node) job, prevents an adjacent 4096 node job
- Job scheduler (Cobalt) will prevent running conflicting jobs

Torus Network (Z dimension)

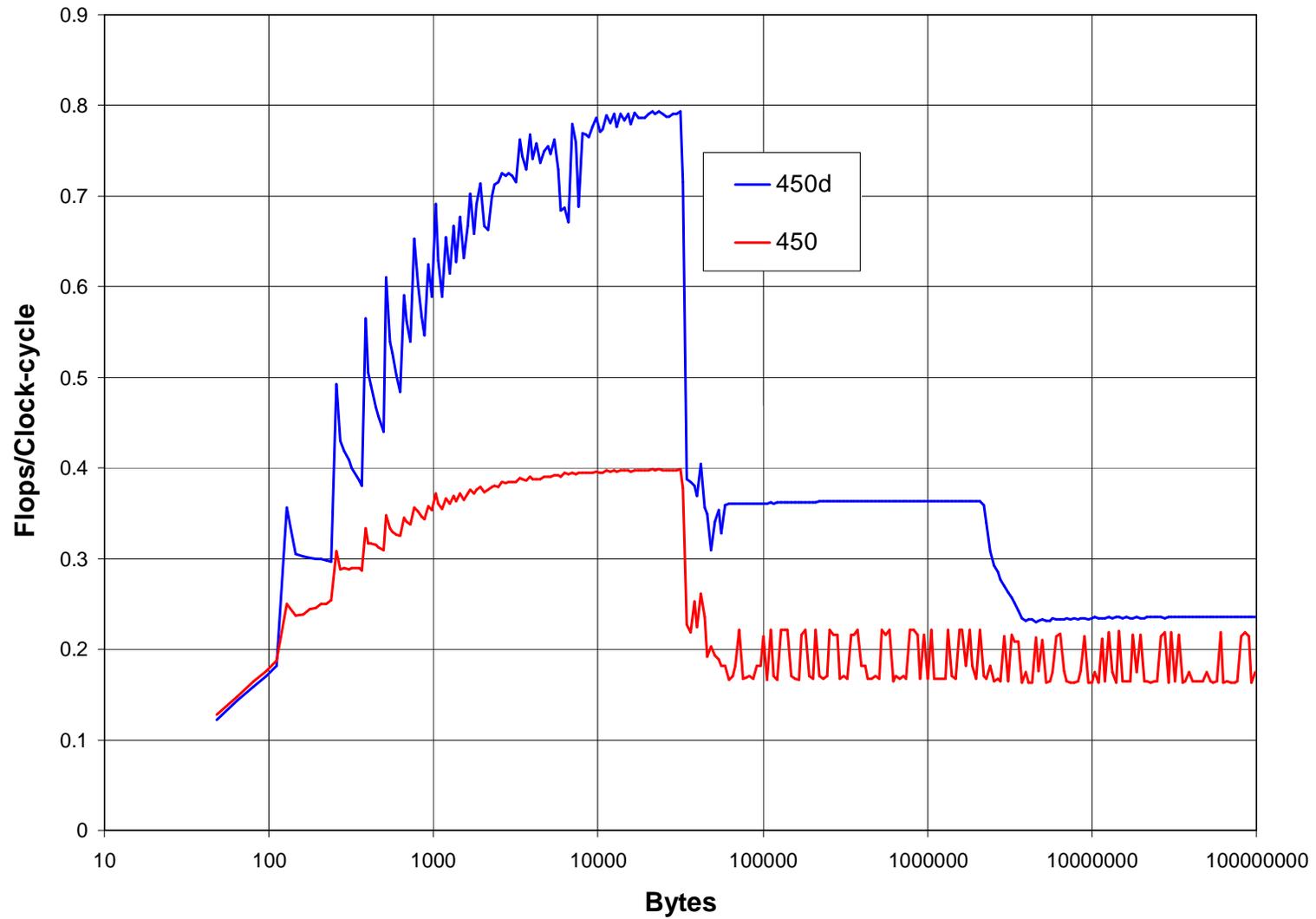


Floating Point Unit (“Double hummer”)



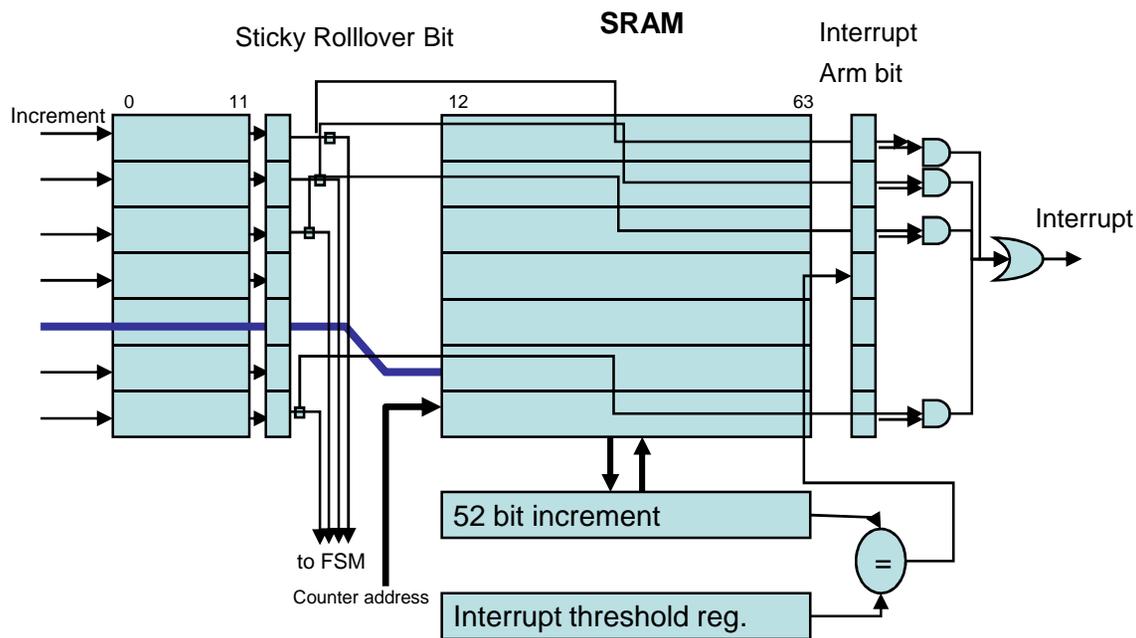
Alignment exceptions have a time penalty

BGP Daxpy Performance



Performance Monitor Architecture

- Novel hybrid counter architecture
 - High density and low power using SRAM design
- 256 counters with 64 bit resolution
 - Fast interrupt trigger with configurable threshold
 - Performance analysis is key to achieving full system potential



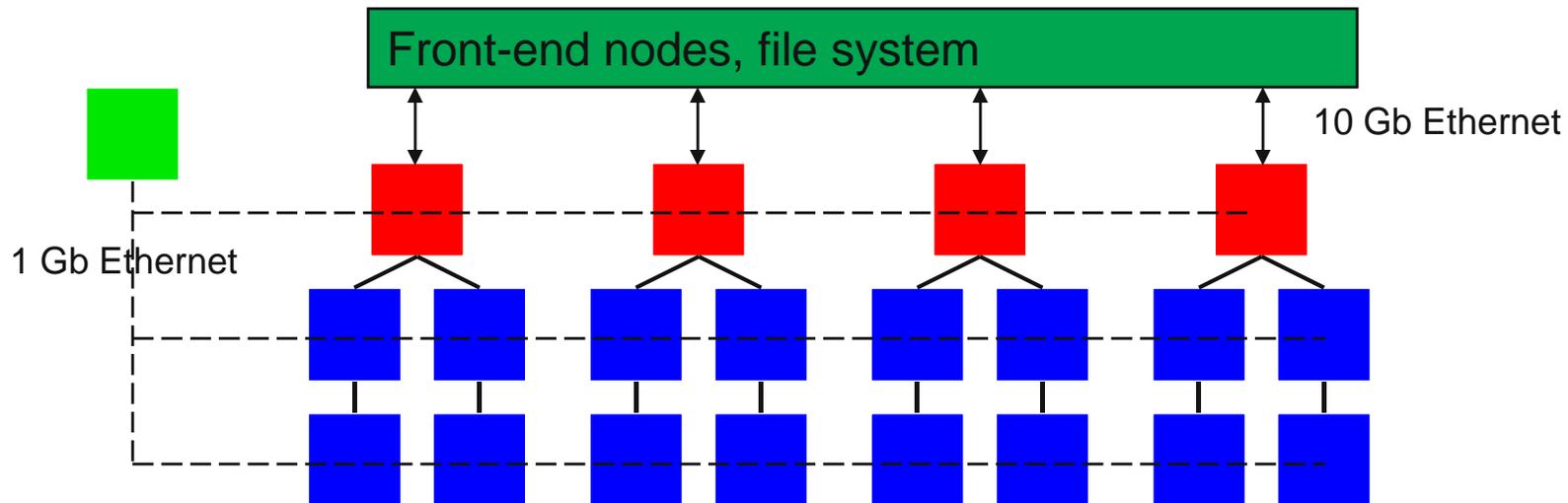
Performance Monitor Features

- Counters for core events
 - loads, stores, floating-point operations (flops)
- Counters for the memory subsystem
 - cache misses, DDR traffic, prefetch info, etc.
- Counters for the network interfaces
 - torus traffic, collective network, DMA, ...
- Counts are tied to hardware elements
 - counts are for cores or nodes, not processes or threads
- Performance monitor hardware is one unit per node;
 - Not all counters available simultaneously

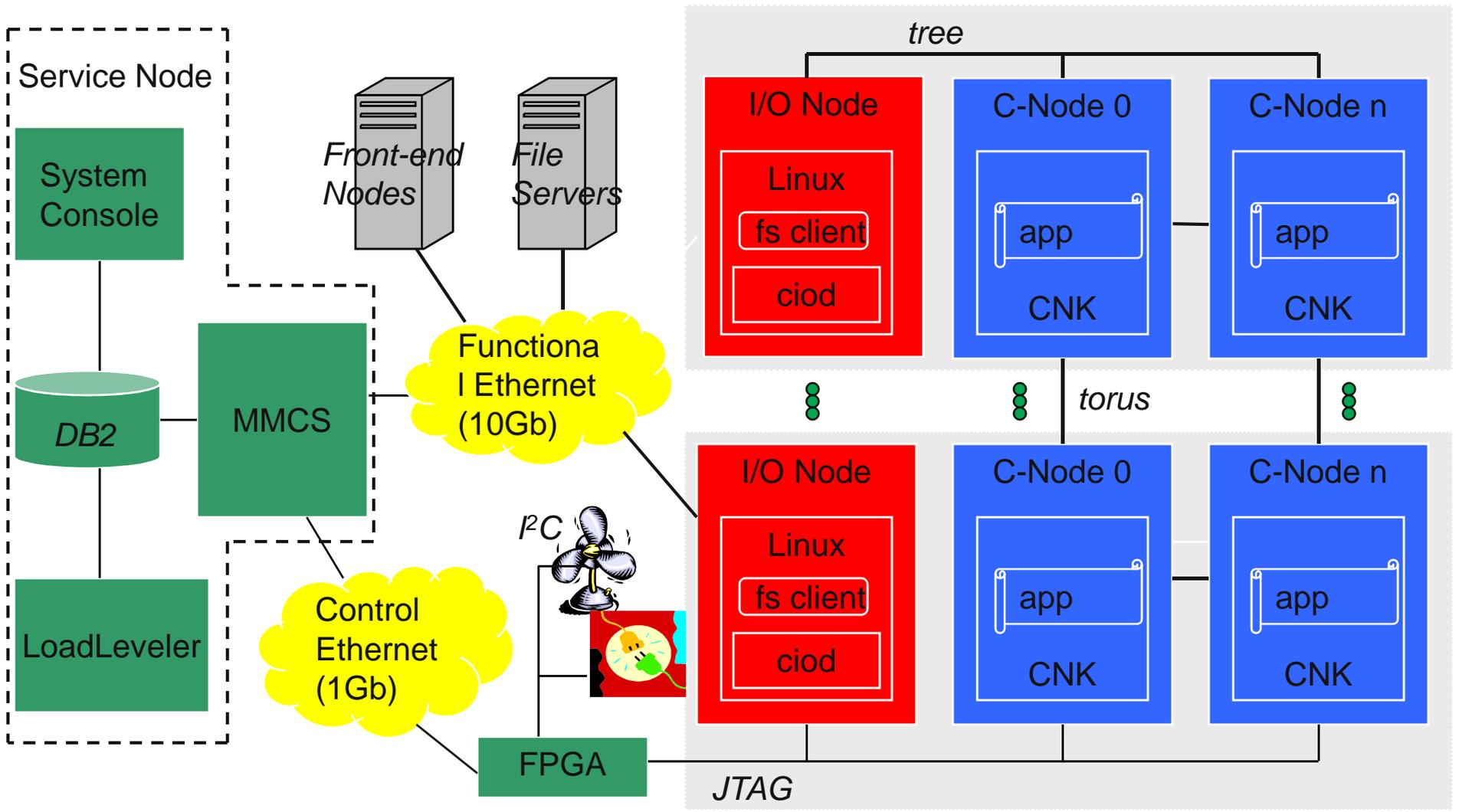
System Level

Blue Gene System Organization

- **Compute nodes** dedicated to running user application, and almost nothing else - simple compute node kernel (CNK)
 - No direct login access
- **I/O nodes** run Linux and provide a more complete range of OS services – files, sockets, process launch, signaling, debugging, and termination
 - 64:1 ratio compute:I/O nodes
- **Service node** performs system management services (e.g., partitioning, heart beating, monitoring errors) - transparent to application software (admin login only)



Blue Gene/P System Architecture



File system details

■ Surveyor

- 1 DataDirect 9550 SAN, 160TB raw storage
- 4 file servers
 - GPFS ~800 MB/s
 - PVFS ~1200 MB/s
- Each server IBM x3655 2U
 - 4 core x86_64
 - 12GB
 - 4X SDR Infiniband
 - File server \leftrightarrow SAN
 - Myricom 10Gb/s
 - File server \leftrightarrow I/O nodes, login nodes

File system details (con't)

- Intrepid
 - 100T
 - 4 DataDirect 9550 SANs, 1.1PB raw storage
 - /gpfs/home: 8 file servers (~2000 MB/s)
 - /gpfs1: 16 file servers (~8000 MB/s)
 - 500T
 - 16 DataDirect 9900 SANs, 7.7PB raw storage
 - 128 file servers (~73000 MB/s)
 - IBM x3455 1U (8GB RAM)

ALCF's Eureka Offers Breakthrough Visualization and Data Analytics

- GraphStream, Inc., Belmont, Calif., will make data analytics and visualization at Intrepid's scale possible through the world's largest installation of NVIDIA S4 external GPUs.
- Nicknamed "Eureka," the new supercomputer will allow researchers to quickly explore and visualize the massive amounts of data they produce with the ALCF's Intrepid.
- 104 dual quad-core servers containing 208 Quadro FX5600 graphics engines, 312GB of total frame buffer RAM, and 3.2TB of total system RAM
- *[Online early October]*



ALCF Machine Allocation (July 2008)

- Total 41 racks (@1024 nodes, 4096 cores)

- Surveyor
 - Single rack
 - Test and (internal) development system

- Intrepid
 - 40 racks total
 - 8 racks for INCITE production runs
 - ½ rack for small user development /debug runs
 - ~32 racks for early science applications and ongoing I/O tests

Queues (July 2008)

- Surveyor
 - Short (30 min, any size)
 - Medium (60 min, 512 or 1024 nodes)
 - Different priorities over the day

- Intrepid
 - “devel”:
 - 512 nodes and up
 - Short (60 min), Medium (1-3hr) , Long (max 6hr)
 - “prod-devel”
 - Up to 512 nodes
 - Half for short (1hr) and half for medium (3hr)

- <http://www.alcf.anl.gov/support/usingALCF/docs/scheduling.php>

Allocations

- Innovative and Novel Computational Impact on Theory and Experiment (INCITE) program
 - Projects awarded time through the INCITE program
 - Starter accounts for projects that will apply for INCITE time
 - **2009 INCITE proposals are due August 11**
 - <http://hpc.science.doe.gov>

- Director's Discretionary Accounts
 - http://www.alcf.anl.gov/support/usingALCF/docs/discretionary_allocations.php

Getting Help

- Contact us
 - support@alcf.anl.gov
- Expanded FAQ and other handy info
 - <https://wiki.alcf.anl.gov/index.php/FAQ>



