



**Argonne**  
NATIONAL  
LABORATORY

*... for a brighter future*



U.S. Department  
of Energy

UChicago ►  
Argonne<sub>LLC</sub>



**Office of  
Science**

U.S. DEPARTMENT OF ENERGY

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

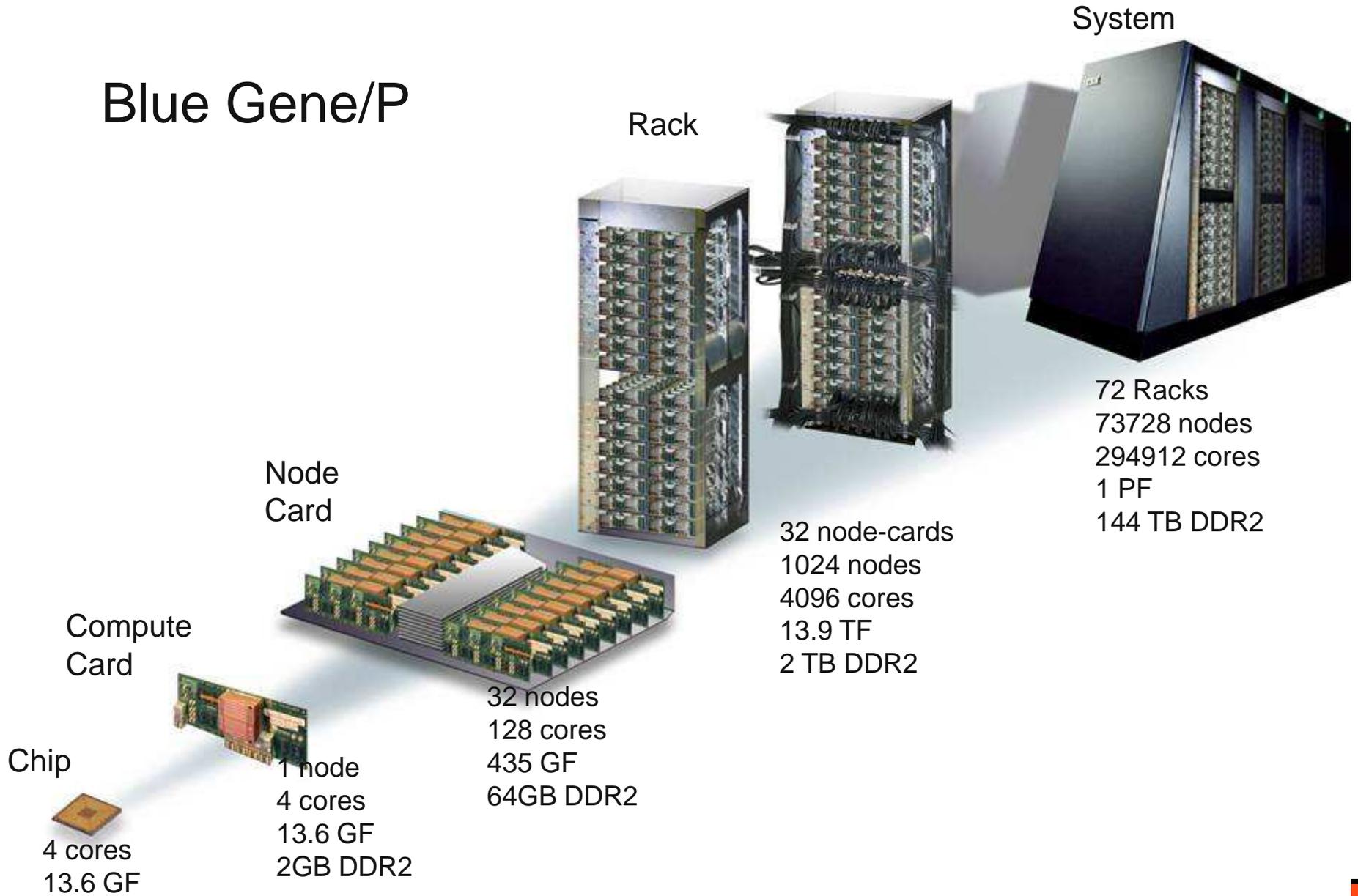
# *Blue Gene Architecture: Past, Present, and (Near) Future*

*Raymond Loy*

*Applications Performance Engineering*

*Argonne Leadership Computing Facility*

# Blue Gene/P

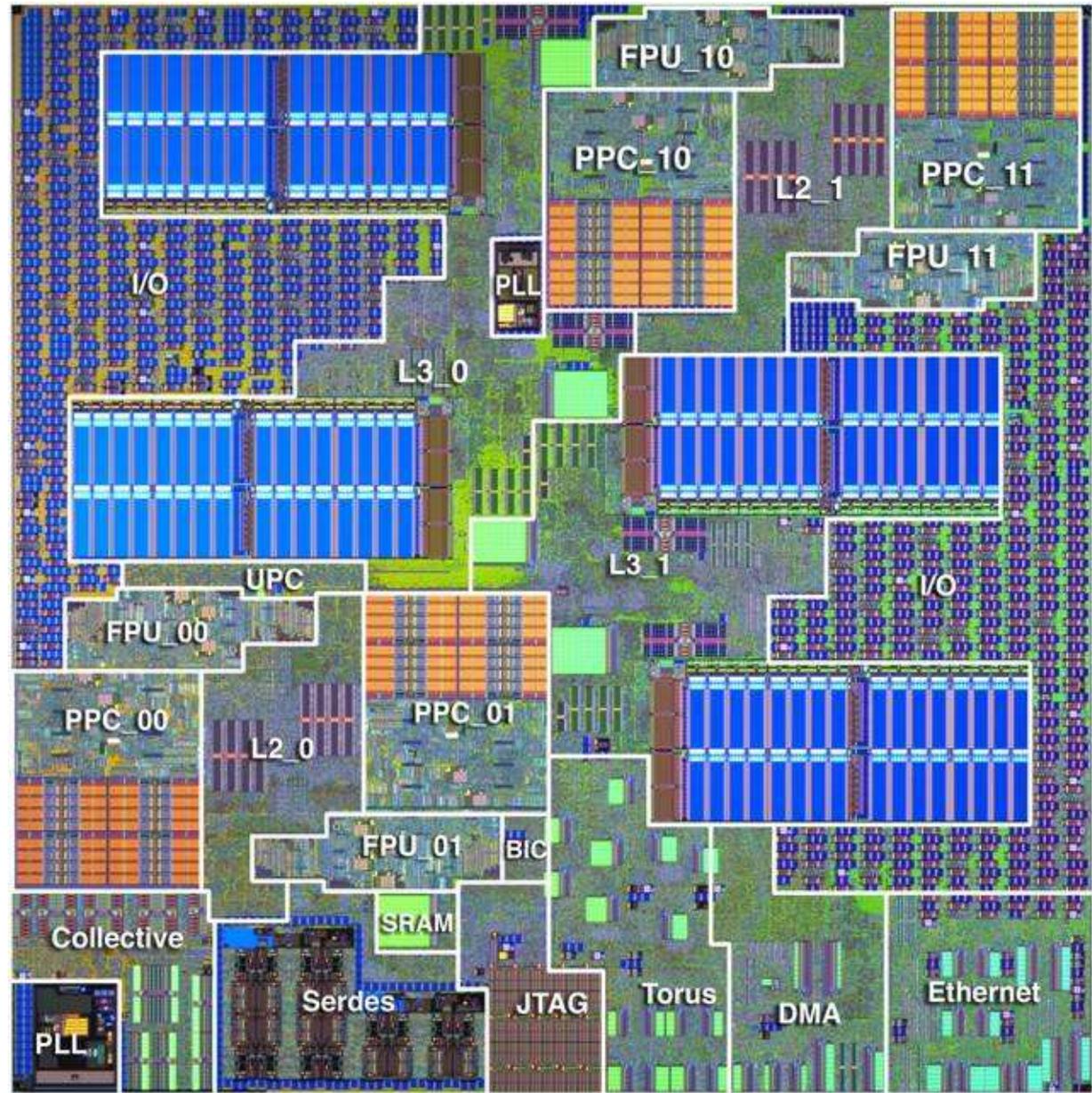


# Summary: BG/P vs BG/L

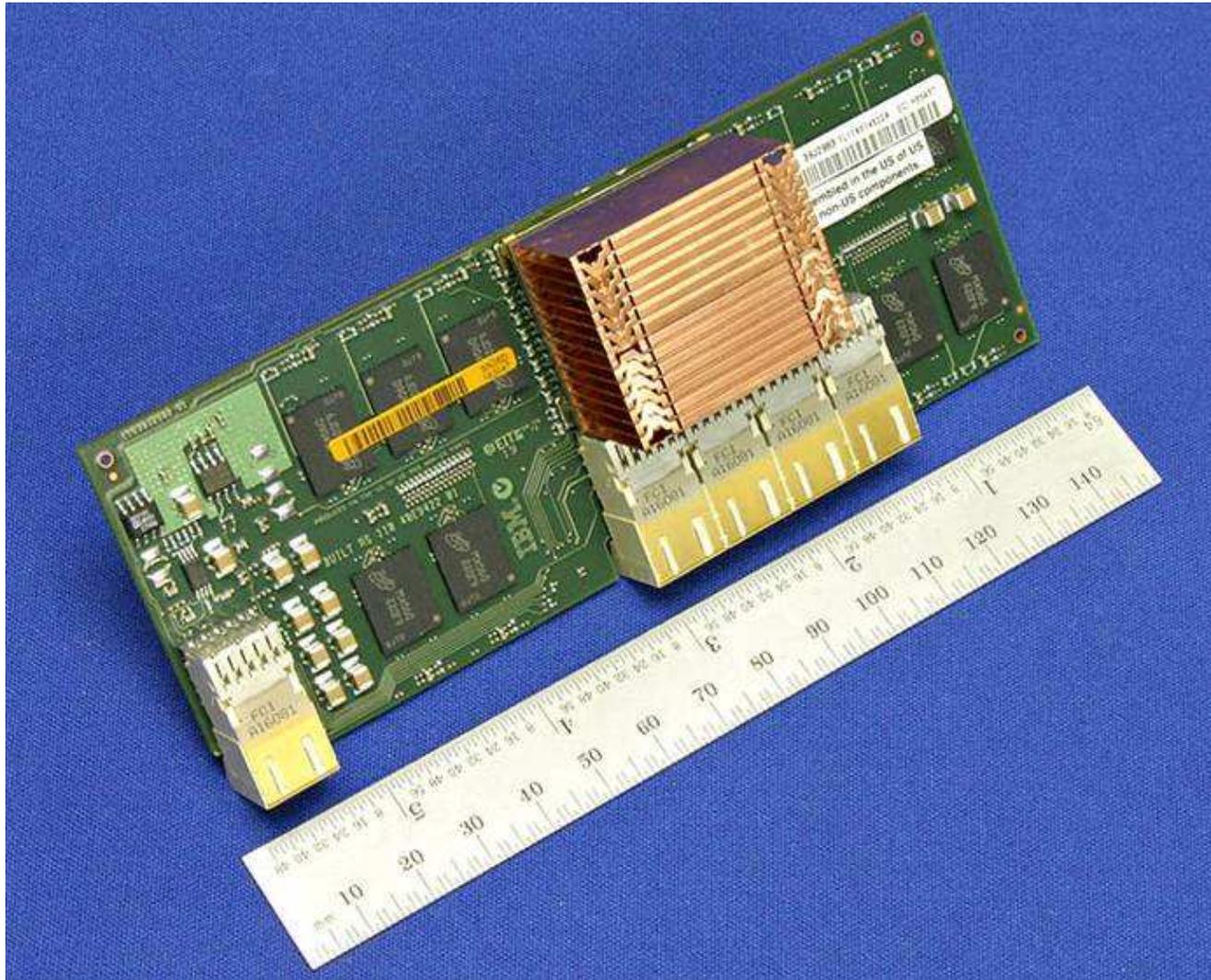
- Increased clock
  - 1.2x from frequency bump 700 MHz => 850 MHz
- Processor density
  - Double the processors/node (4 vs. 2)
- Memory
  - higher bandwidth
  - Cache coherency
    - *Allows 4 way SMP*
      - supports OpenMP, pthreads
  - DMA for torus
- Faster communication
  - 2.4x higher bandwidth, 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

**BPC chip  
DD2.1 die  
photograph**

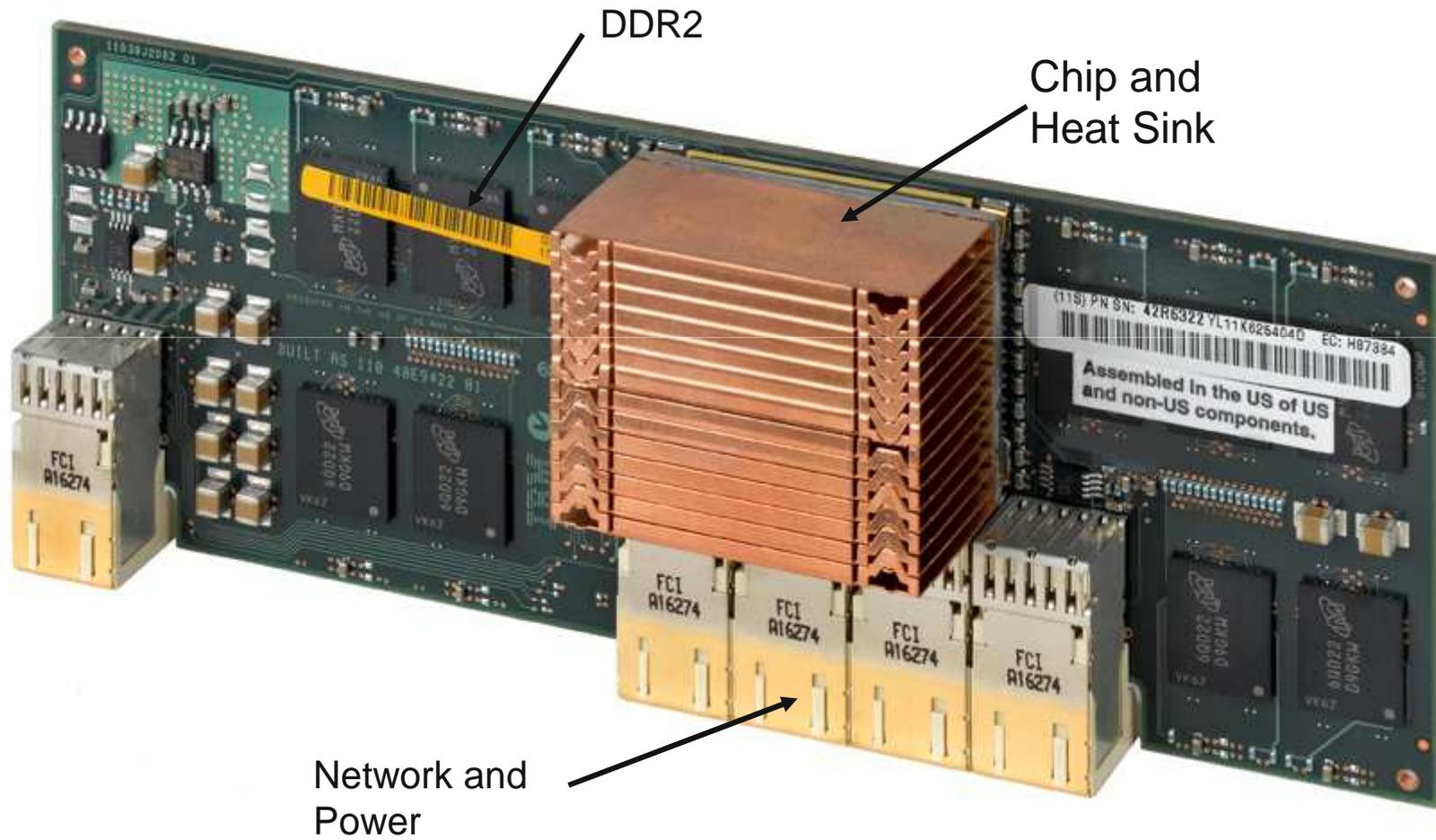
13mmx13mm  
90 nm process  
208M transistors  
88M in eDRAM



## BG/P Compute Card



# BG/P Compute Card



# BGP Node Card

32 Compute nodes

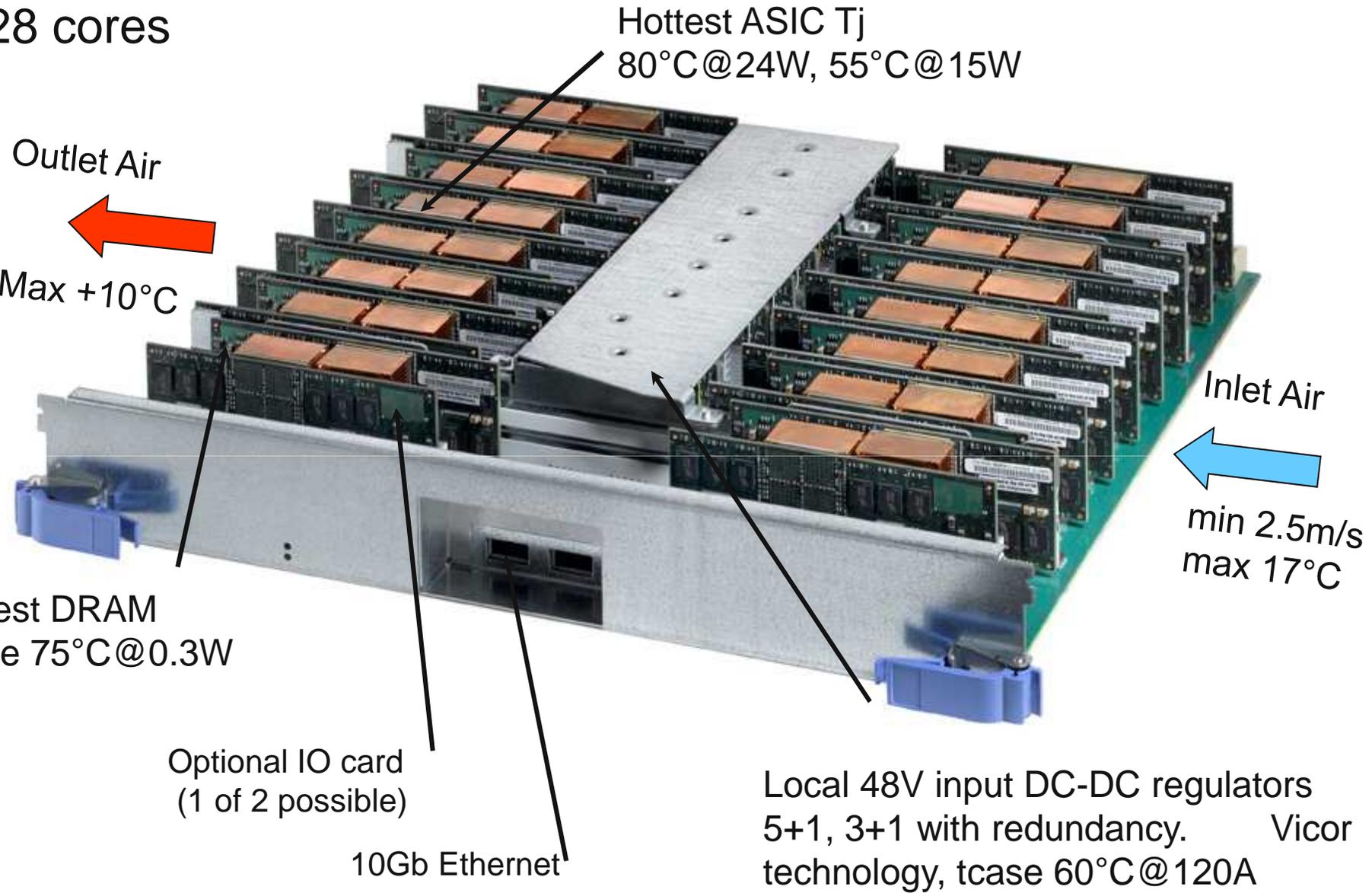


Optional IO card  
(one of 2 possible)  
with 10Gb optical link

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*



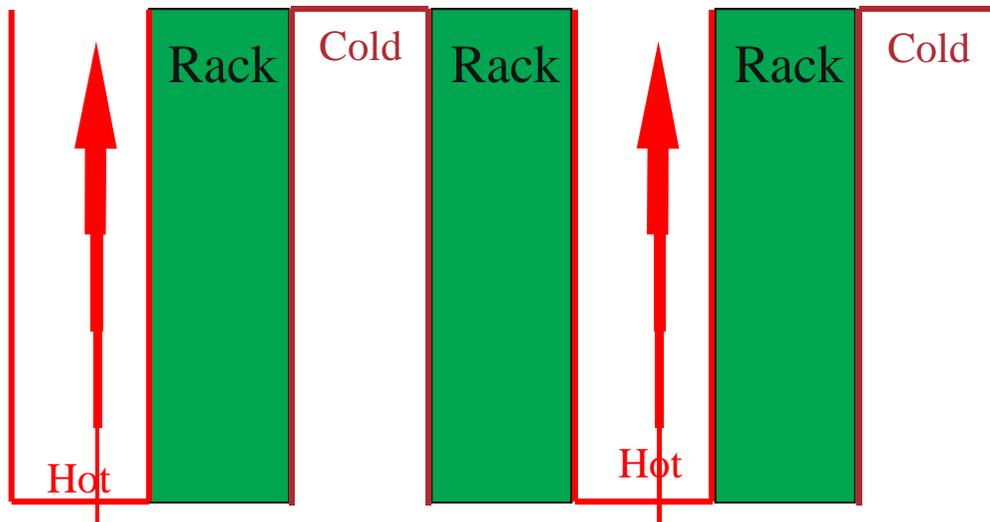
## *First 8 racks of BG/P: Covers removed*



# *IBM Blue Gene/P*



*etc.*

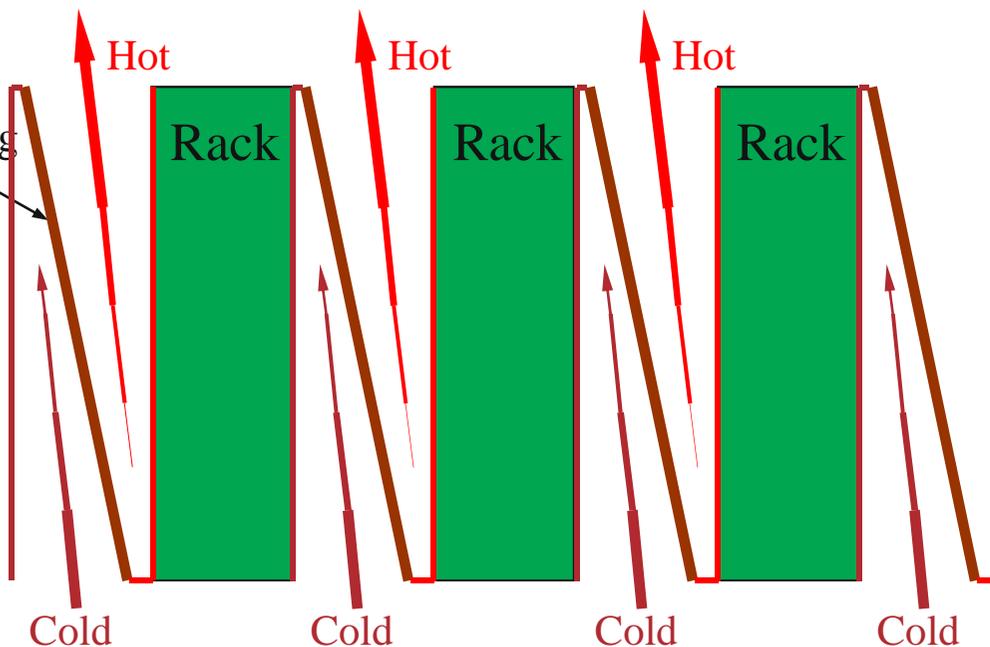


**(a) Prior Art:  
Segregated,  
Non-Tapered  
Plenums**

**(Plenum Width Same  
Regardless of Flow Rate)**

Thermal-Insulating  
Baffle

*etc.*



**(b) Invention:  
Integrated,  
Tapered  
Plenums**

**(Plenum Width  
Larger where Flow  
Rate is Greater)**

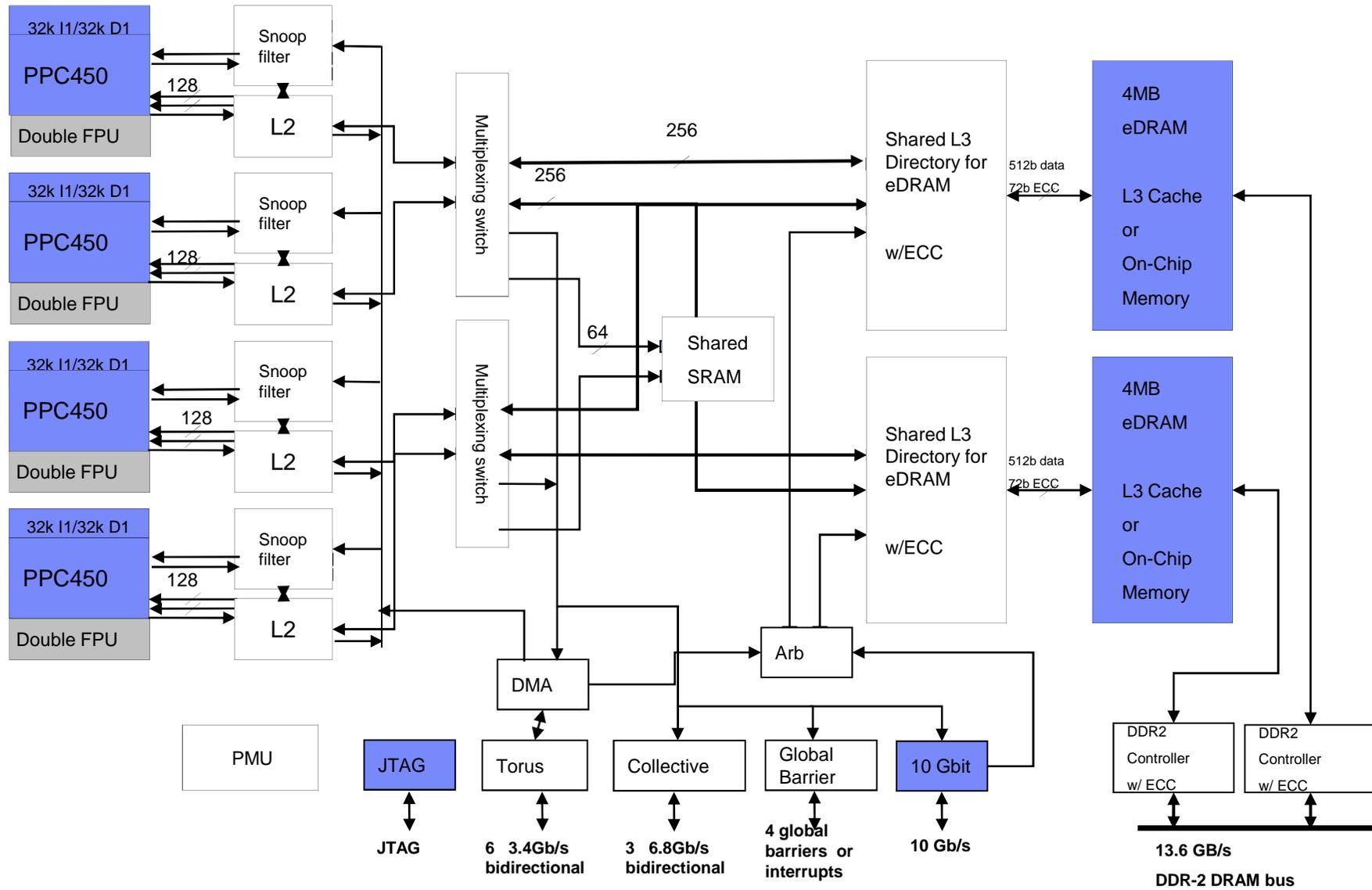
Shawn Hall 4-3-02  
02-04-03 Angled Plenums

## 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
- Green500
  - November 2007
    - *BG/P debuted taking #1-5 positions*
  - June 2011
    - *BG/P #29*
    - *BG/Q prototype rank #1*
      - 165% more efficient than Top500 #1 (Tianhe-1A)

# *Memory Subsystem*

# Blue Gene/P ASIC



# Memory System Bottlenecks

## L2 – L3 switch

Not a full core to L3 bank crossbar

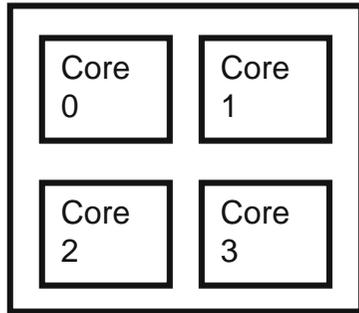
Request rate and bandwidth are limited if two cores of one dual processor group access the same L3 cache bank

## Banking for DDR2

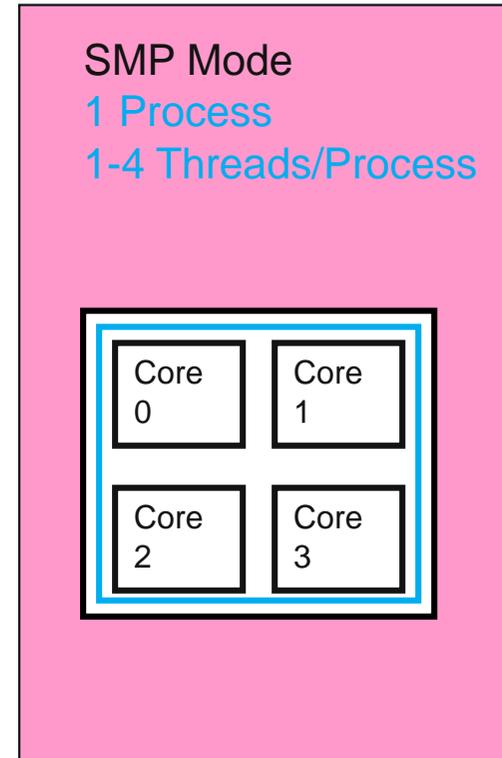
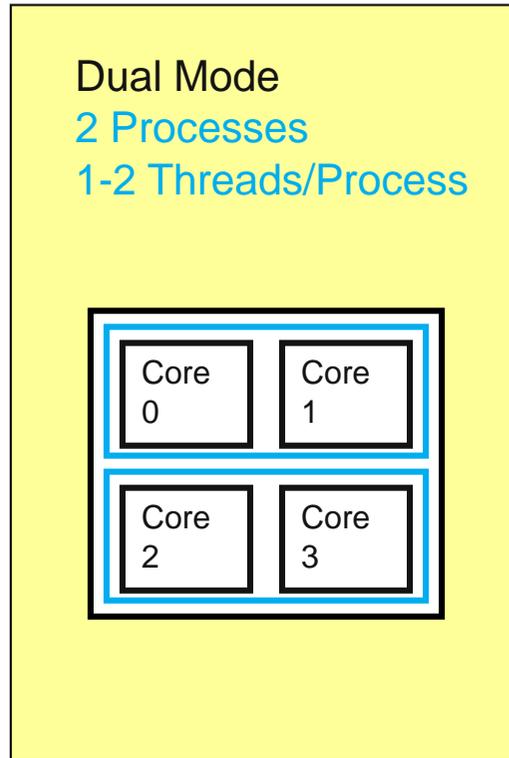
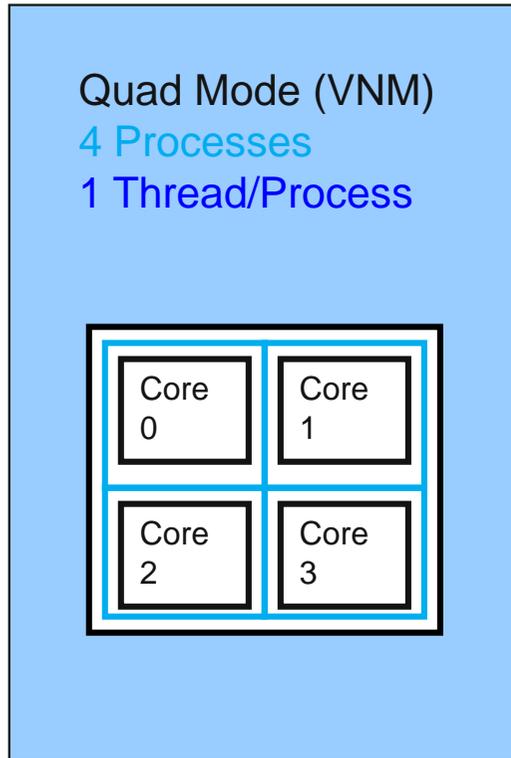
4 banks on 512Mb DDR modules

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

# Execution Modes in BG/P

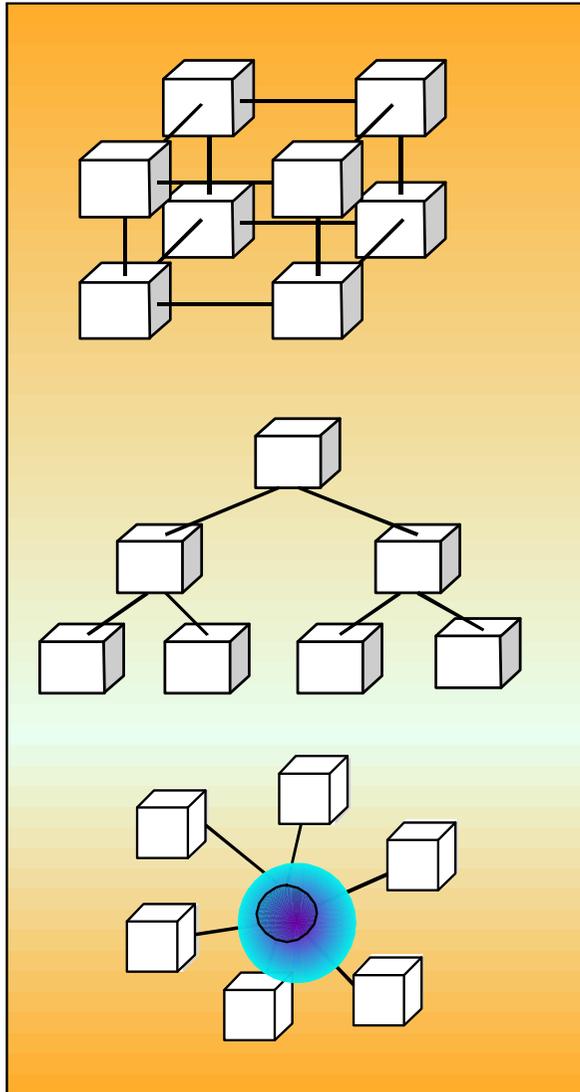


Hardware Elements Black  
Software Abstractions Blue



# *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  $\mu$ s latency between nearest neighbors, 5  $\mu$ s to the farthest
- MPI: 3  $\mu$ s latency for one hop, 10  $\mu$ 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  $\mu$ s, MPI 5  $\mu$ 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  $\mu$ s, MPI 1.6  $\mu$ 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

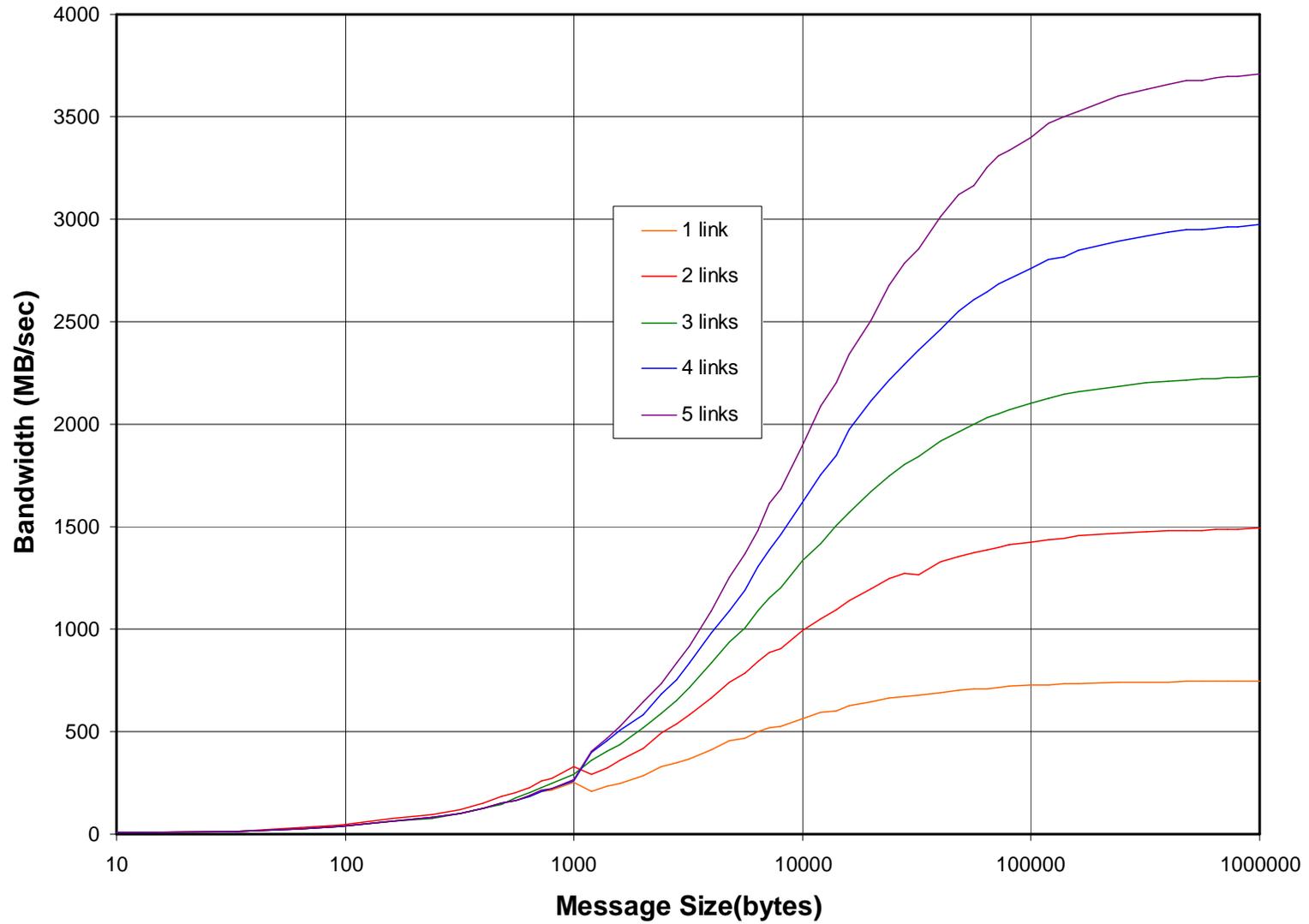
Same Network Bandwidth per Flops as BG/L

Primary interface is via DMA, rather than cores

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

Software product stack uses DMA mode

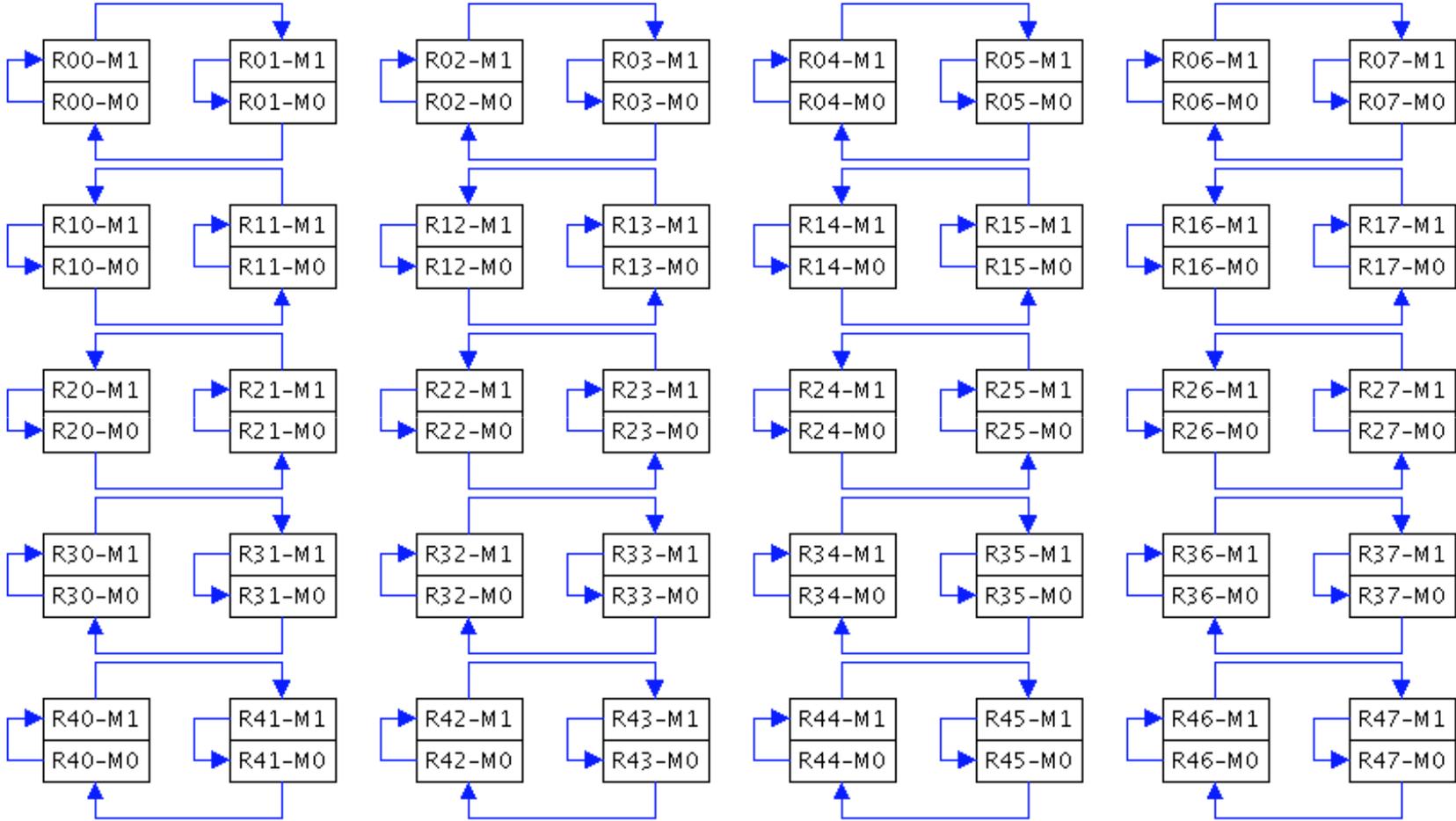
# BGP Exchange Bandwidth



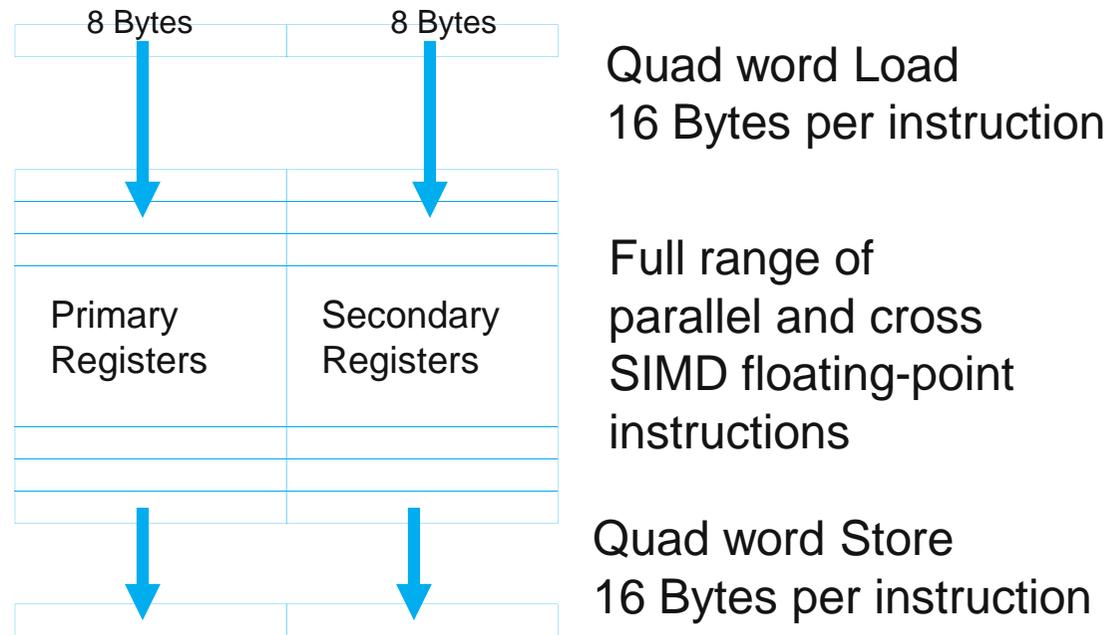
## Torus Network Limitations (ALCF)

- 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)



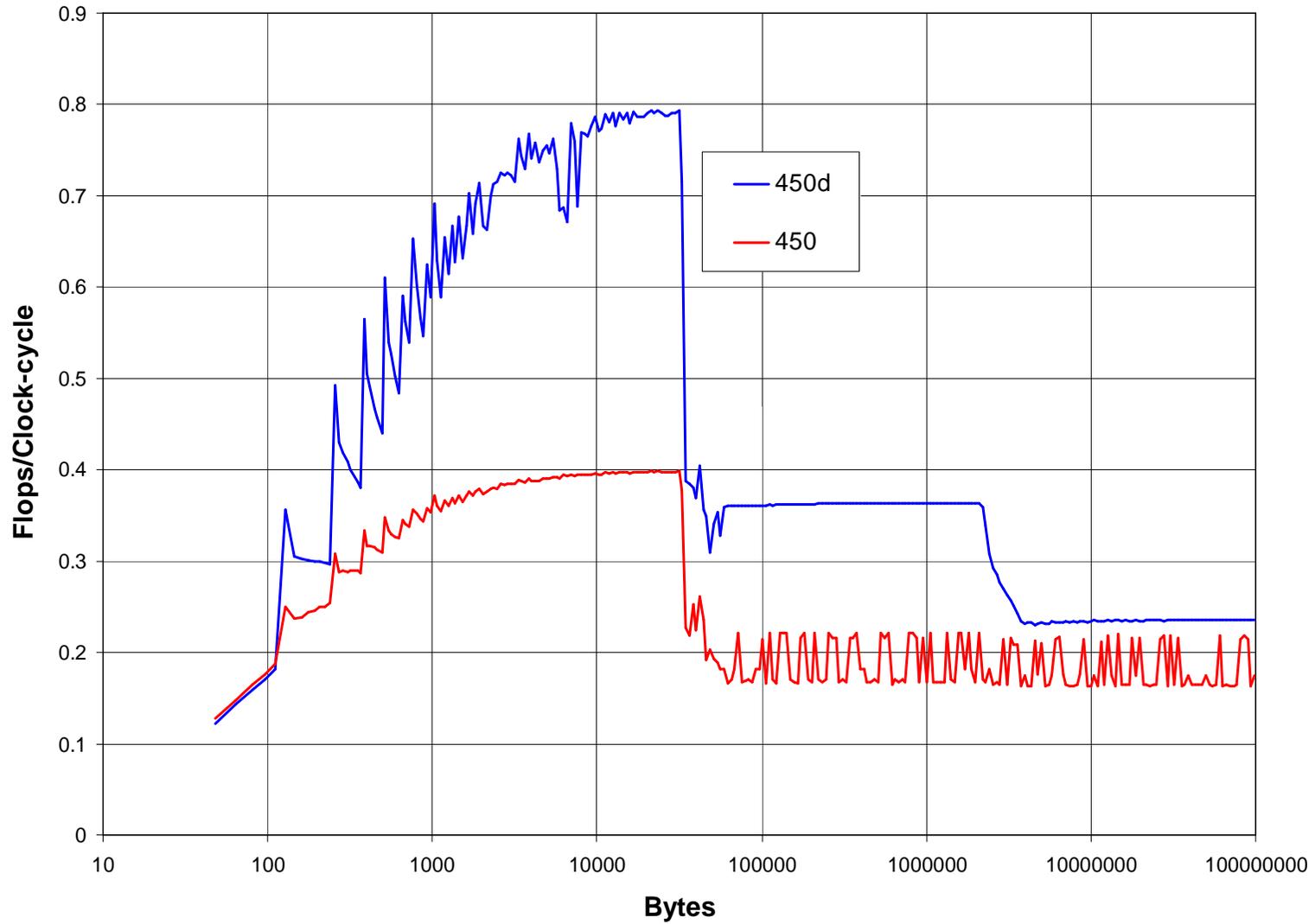
# Floating Point Unit (“Double hummer”)



8 Bytes                      8 Bytes  
Quad word load/store operations  
require data aligned on 16-Byte  
boundaries.

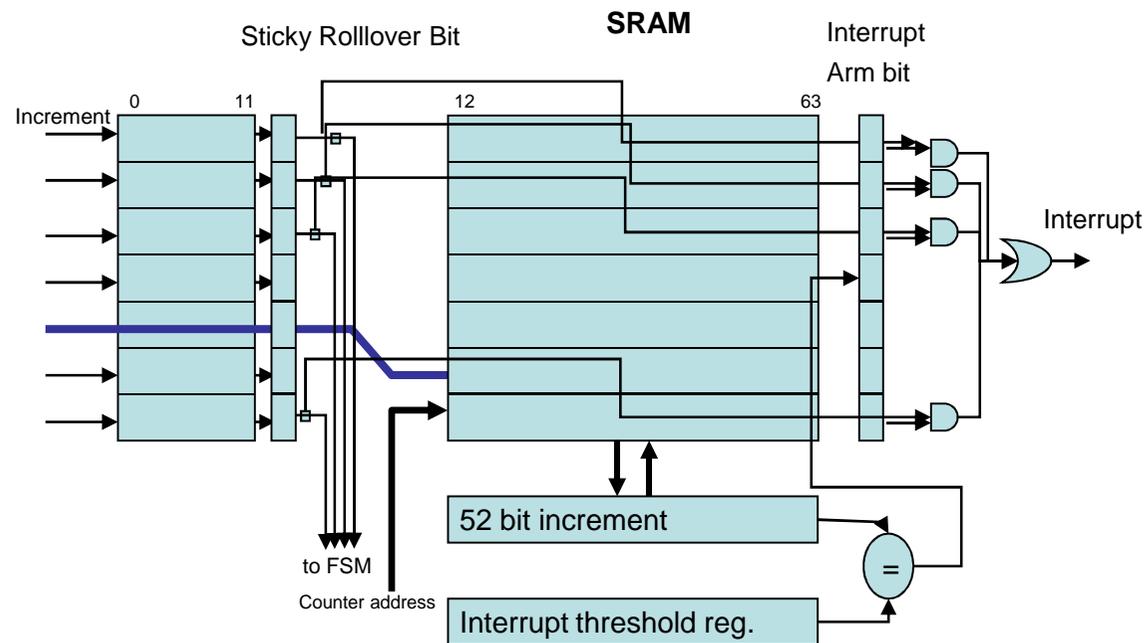
*Alignment exceptions have a time penalty*

# 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



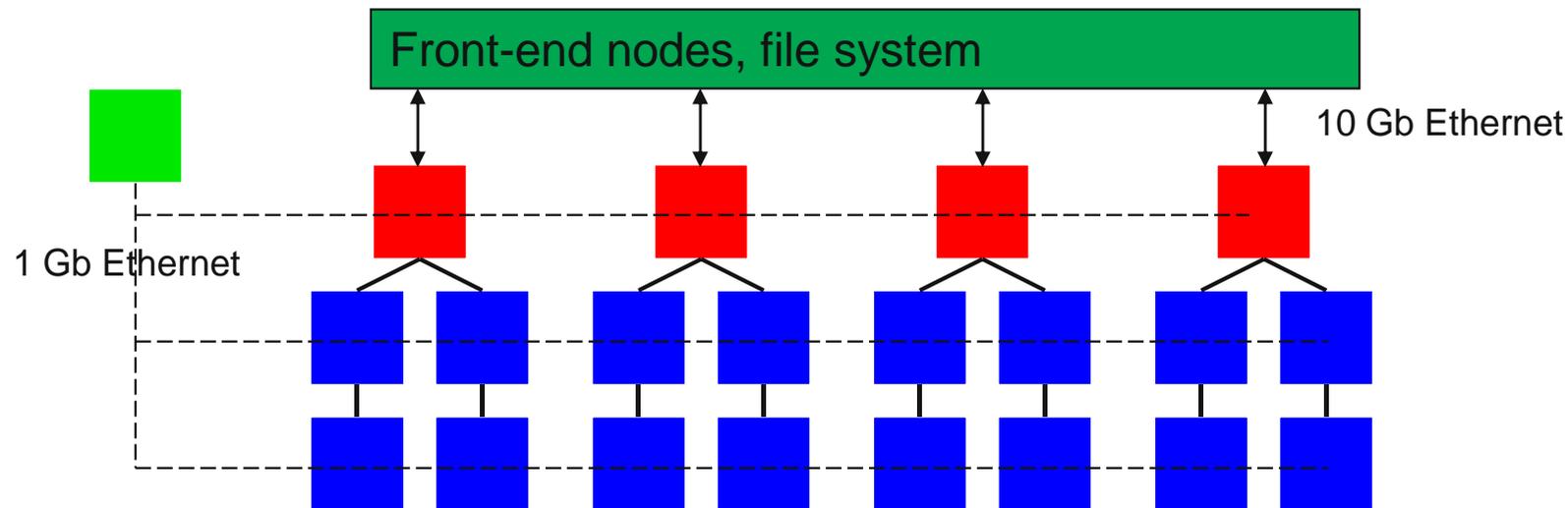
## *Performance Monitor Features*

- 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)



## *Programming models and development environment*

### ■ Familiar methods

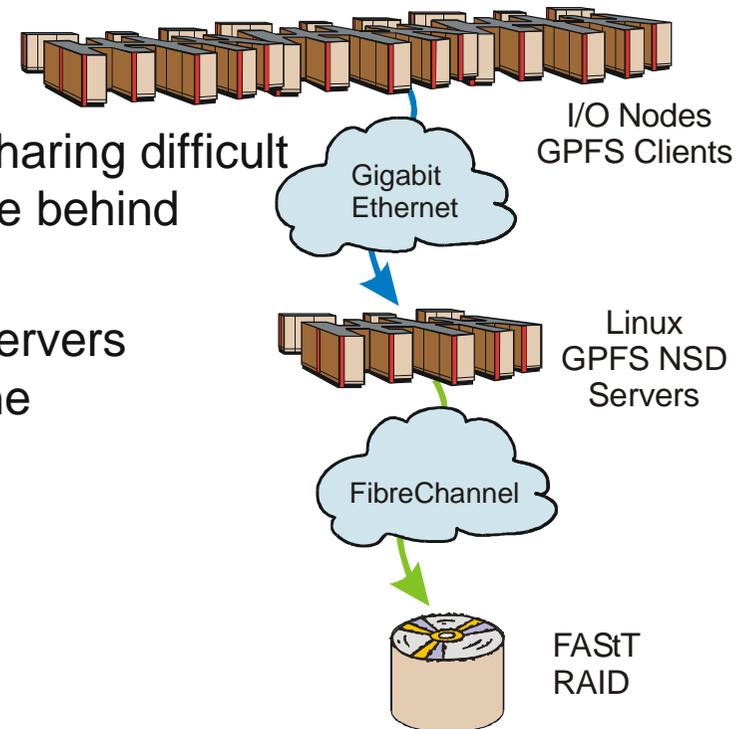
- SPMD model - Fortran, C, C++ with MPI (MPI1 + subset of MPI2)
  - *Full language support with IBM XL and GNU compilers*
  - *Automatic SIMD FPU exploitation (limited)*
- Linux development environment
  - *User interacts with system through front-end nodes running Linux – compilation, job submission, debugging*
  - *Compute Node Kernel provides look and feel of a Linux environment*
    - POSIX routines (with some restrictions: no fork() or system())
    - BG/P adds pthread support, additional socket support
  - *Tools – support for debuggers, MPI tracer, profiler, hardware performance monitors, visualizer (HPC Toolkit), PAPI*

### ■ Restrictions (which lead to significant benefits)

- *Space sharing - one parallel job per partition of machine, one thread per core in each compute node*
- *Virtual memory is constrained to physical memory size*

## General Parallel File System (GPFS) for Blue Gene

- Blue Gene can generate enormous I/O demand (disk limited)
  - BG/P IO-rich has 64 10Gb/rack – 80GB/sec
- Serving this kind of demand requires a parallel file system
- NFS for file I/O
  - Limited scalability
  - NFS has no cache consistency, making write sharing difficult
  - Poor performance, not enough read ahead/write behind
- GPFS runs on Blue Gene
  - GPFS clients in Blue Gene call external NSD servers
  - Brings traditional benefits of GPFS to Blue Gene
    - *I/O parallelism*
    - *Cache consistent shared access*
    - *Aggressive read-ahead, write-behind*



## File system details

### ■ Surveyor

- 1 DataDirect 9550 SAN, 160TB raw storage
  - 320 \* 500GB SATA HDD
- 4 file servers
  - GPFS ~600 MB/s
  - PVFS ~1050 MB/s
- Each server IBM x3655 2U
  - 2 dual-core x86\_64 (2.6 GHz)
  - 12GB RAM
  - 4X SDR *Infiniband*
    - File server  $\leftrightarrow$  SAN
  - *Myricom* 10Gb/s
    - File server  $\leftrightarrow$  I/O nodes, login nodes

## File system details (con't)

### ■ Intrepid

- /gpfs/home
  - 4 DataDirect 9550 SANs total 1.1PB
  - 24 file servers IBM x3655 (~2000 MB/s)
- /intrepid-fs0
  - 16 DataDirect 9900 SANs total 7.5 PB raw storage
    - Each with 480 \* 1TB SATA HDD
  - 128 file servers (~62000 MB/s)
    - IBM x3455 (8GB RAM)
- Networks
  - 4X SDR *Infiniband* (File server  $\leftrightarrow$  SAN)
  - *Myricom* 10Gb/s (File server  $\leftrightarrow$  I/O nodes, login nodes)

# *The Next Generation ALCF System: BG/Q*

- DOE has approved our acquisition of “Mira”, a 10 Petaflops Blue Gene/Q system. An evolution of the Blue Gene architecture with:
  - 16 cores/node
  - 1 GB of memory per core, nearly a TB of memory in aggregate
  - 48 racks (over 780k cores)
  - 384 I/O nodes (128:1 Compute:I/O)
  - 32 I/O nodes for logins and/or data movers
  - Additional non-I/O login nodes
  - 2 service nodes
  - IB data network; 70 PB of disk with 470 GB/s of I/O bandwidth
  - Power efficient, water cooled
- Argonne and Livermore worked closely with IBM over the last few years to help develop the specifications for this next generation Blue Gene system
- 16 Projects Accepted into the Early Science Program
- Applications running on the BG/P should run immediately on the BG/Q, but may see better performance by exposing greater levels of parallelism at the node level

## *ALCF-2: Blue Gene/Q (Mira)*

### *The story so far*

#### **Jan 2009**

- CD0 approved

#### **Jul 2009**

- Lehman Review (CD1/2a) passed

#### **Jul 2010**

- Lehman Review (CD2b/3) passed

#### **Aug 2010**

- Contract approved

#### **2011**

- BG/Q Early Science Program begins

## **ALCF-2: Blue Gene/Q (Mira)**

### ***What's next?***

#### **Mid 2011**

- Early Access System
  - *Approximately 128 nodes + 1 I/O node*
  - *Located at IBM, leased for ALCF use*

#### **Spring 2012**

- T&D System delivery
  - *1-2 racks , 128:1 compute:I/O node ratio (Same as Mira)*

#### **2012**

- Mira delivery expected

#### **2013**

- Mira acceptance

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

