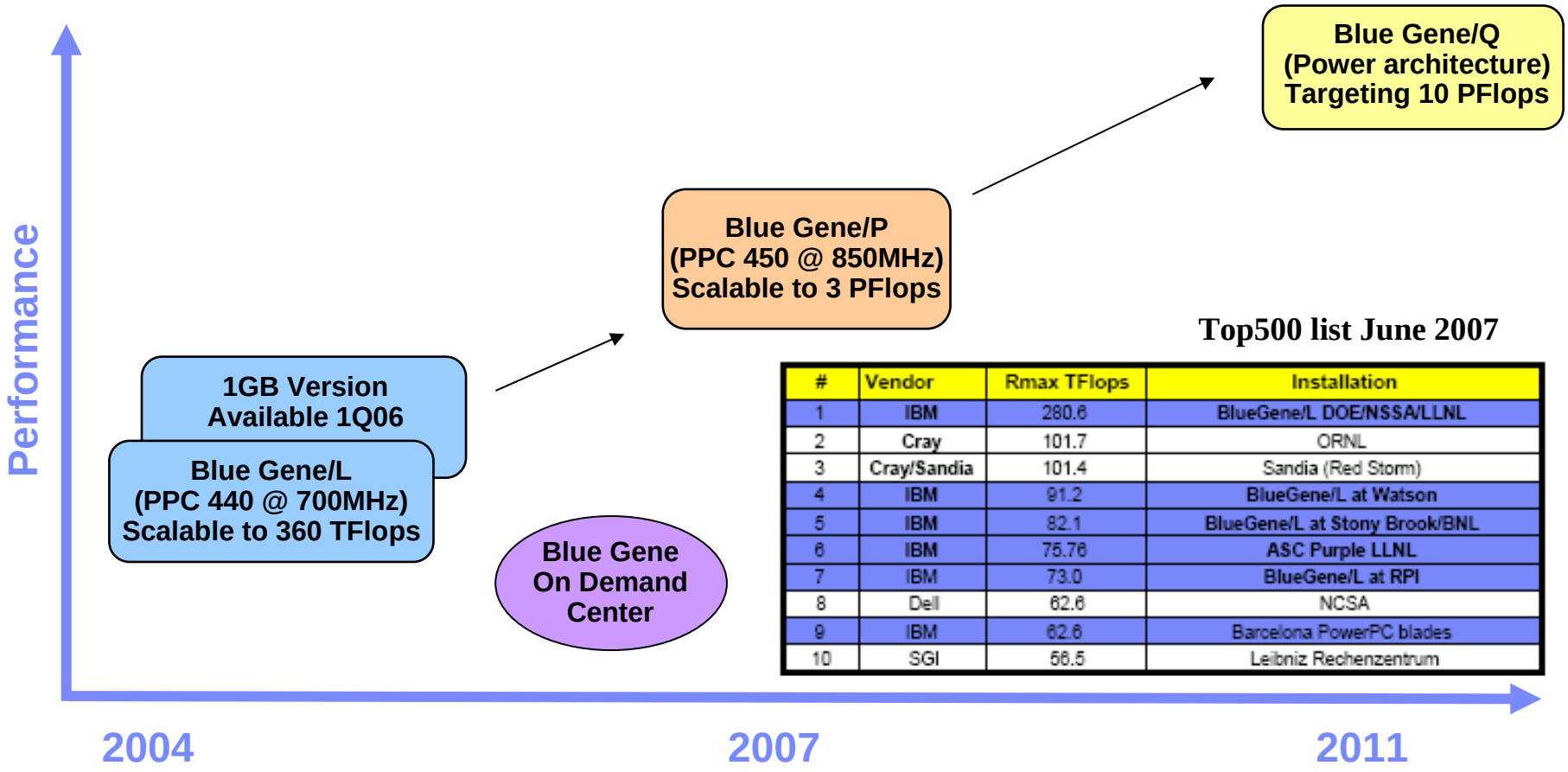

Stockholm Brain Institute
Blue Gene/L
IBM
SOUTH POLE

# *IBM® Blue Gene®/P - An Overview of a Petaflop Capable System*


Carl G. Tengwall
IBM Storage & Technology Group
tengwall@se.ibm.com

# Blue Gene Technology Roadmap




**Top500 list June 2007**

| # | Vendor | Rmax TFlops | Installation |
|---|---|---|---|
| 1 | IBM | 280.6 | BlueGene/L DOE/NSSA/LLNL |
| 2 | Cray | 101.7 | ORNL |
| 3 | Cray/Sandia | 101.4 | Sandia (Red Storm) |
| 4 | IBM | 91.2 | BlueGene/L at Watson |
| 5 | IBM | 82.1 | BlueGene/L at Stony Brook/BNL |
| 6 | IBM | 75.76 | ASC Purple LLNL |
| 7 | IBM | 73.0 | BlueGene/L at RPI |
| 8 | Dell | 62.6 | NCSA |
| 9 | IBM | 62.6 | Barcelona PowerPC blades |
| 10 | SGI | 56.5 | Leibniz Rechenzentrum |

# Blue Gene/P Architectural Highlights

- Scaled performance through density and frequency bump
  - 2x performance through doubling the processors/node
  - 1.2x from frequency bump due to technology
- Enhanced function
  - 4 way SMP
  - DMA, remote put-get, user programmable memory prefetch
  - Greatly enhanced 64 bit performance counters (including 450 core)
- Hold BlueGene/L packaging as much as possible:
  - Improve networks through higher speed signaling on same wires
  - Improve power efficiency through aggressive power management
- Higher signaling rate
  - 2.4x higher bandwidth,
  - improve latency for Torus and Tree networks
  - 10x higher bandwidth for Ethernet IO

# Blue Gene/P Architectural Highlights

- **Lightweight kernel (CNK) on Compute Nodes**
- **Linux on I/O Nodes handling syscalls**
- **Optimized MPI library for high speed messaging**
- **Control system on Service Node with private control network**
- **Compilers and job launch on Front End Nodes**

# Blue Gene/P compared to Blue Gene/L

| Property | | BG/L | BG/P |
|---|---|---|---|
| Node Properties | Node Processors | 2* 440 PowerPC | 4* 450 PowerPC |
| | Processor Frequency | 0.7GHz | 0.85GHz (target) |
| | Coherency | Software managed | SMP |
| | L1 Cache (private) | 32KB/processor | 32KB/processor |
| | L2 Cache (private) | 14 stream prefetching | 14 stream prefetching |
| | L3 Cache size (shared) | 4MB | 8MB |
| | Main Store/node | 512MB/1GB | 2GB |
| | Main Store Bandwidth | 5.6GB/s (16B wide) | 13.6 GB/s (2*16B wide) |
| | Peak Performance | 5.6GF/node | 13.6 GF/node |
| Torus Network | Bandwidth | 6*2*175MB/s=2.1GB/s | 6*2*425MB/s=5.1GB/s |
| | Hardware Latency (Nearest Neighbor) | 200ns (32B packet)<br>1.6us(256B packet) | 160ns (32B packet)<br>500ns(256B packet) |
| | Hardware Latency (Worst Case) | 6.4us (64 hops) | 5us(64 hops) |
| Collective Network | Bandwidth | 2*350MB/s=700MB/s | 2*0.85GB/s=1.7GB/s |
| | Hardware Latency (round trip worst case) | 5.0us | 4us |
| System Properties | Peak Performance (72k nodes) | 410TF | 1PF |
| | Total Power | 1.7MW | 2.7 MW |

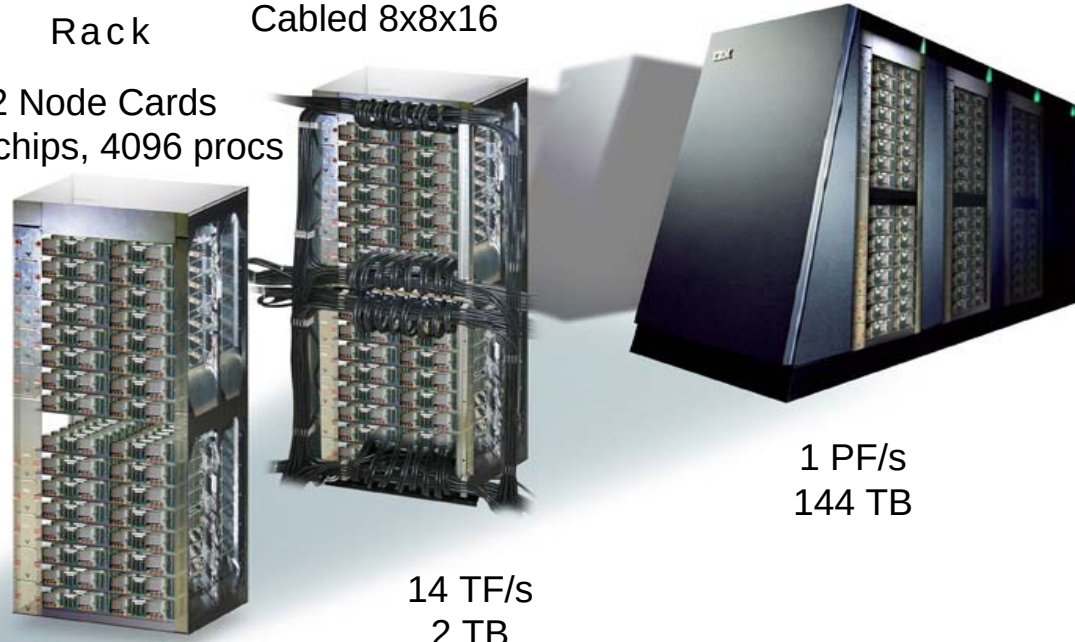
# Offering Schematic – Blue Gene/P

*Blue Gene/P continues Blue Gene's leadership performance in a space-saving, power-efficient package for the most demanding and scalable high-performance computing applications*

System
72 Racks

Cabled 8x8x16

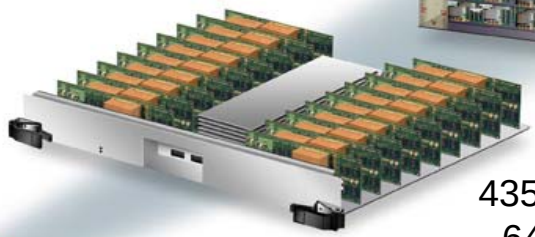
Rack
32 Node Cards
1024 chips, 4096 procs



1 PF/s
144 TB

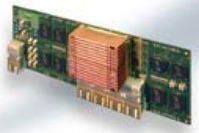
14 TF/s
2 TB

Node Card
(32 chips 4x4x2)
32 compute, 0-1 IO cards



435 GF/s
64 GB

Compute Card
1 chip, 20
DRAMs



13.6 GF/s
2.0 (or 4.0) GB DDR
Supports 4-way SMP

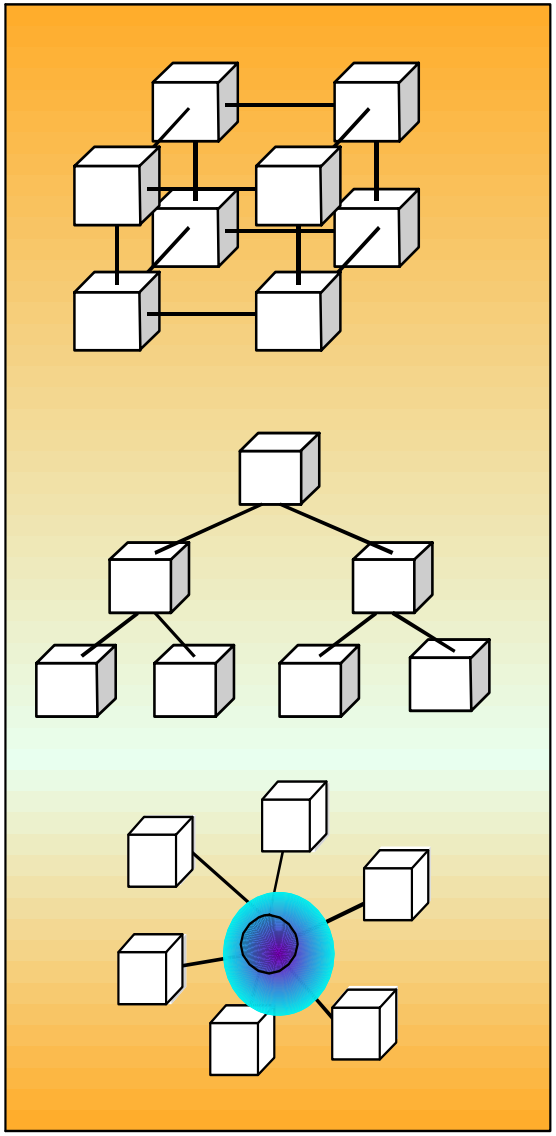
Chip
4 processors

13.6 GF/s
8 MB EDRAM

**Front End Node / Service Node**
JS21 / Power5
Linux SLES10

**HPC SW:**
Compilers
GPFS
ESSL
Loadleveler

# Blue Gene/P Interconnection Networks



## 3 Dimensional Torus

- Interconnects all compute nodes (73,728)
- Virtual cut-through hardware routing
- 3.4 Gb/s on all 12 node links (5.1 GB/s per node)
- 0.5 µs latency between nearest neighbors, 5 µs to the farthest
- MPI: 3 µs latency for one hop, 10 µs to the farthest
- Communications backbone for computations
- 1.7/3.9 TB/s bisection bandwidth, 188TB/s total bandwidth

## Collective Network

- One-to-all broadcast functionality
- Reduction operations functionality
- 6.8 Gb/s of bandwidth per link
- Latency of one way tree traversal 1.3 µs, MPI 5 µs
- ~62TB/s total binary tree bandwidth (72k machine)
- Interconnects all compute and I/O nodes (1152)

## Low Latency Global Barrier and Interrupt

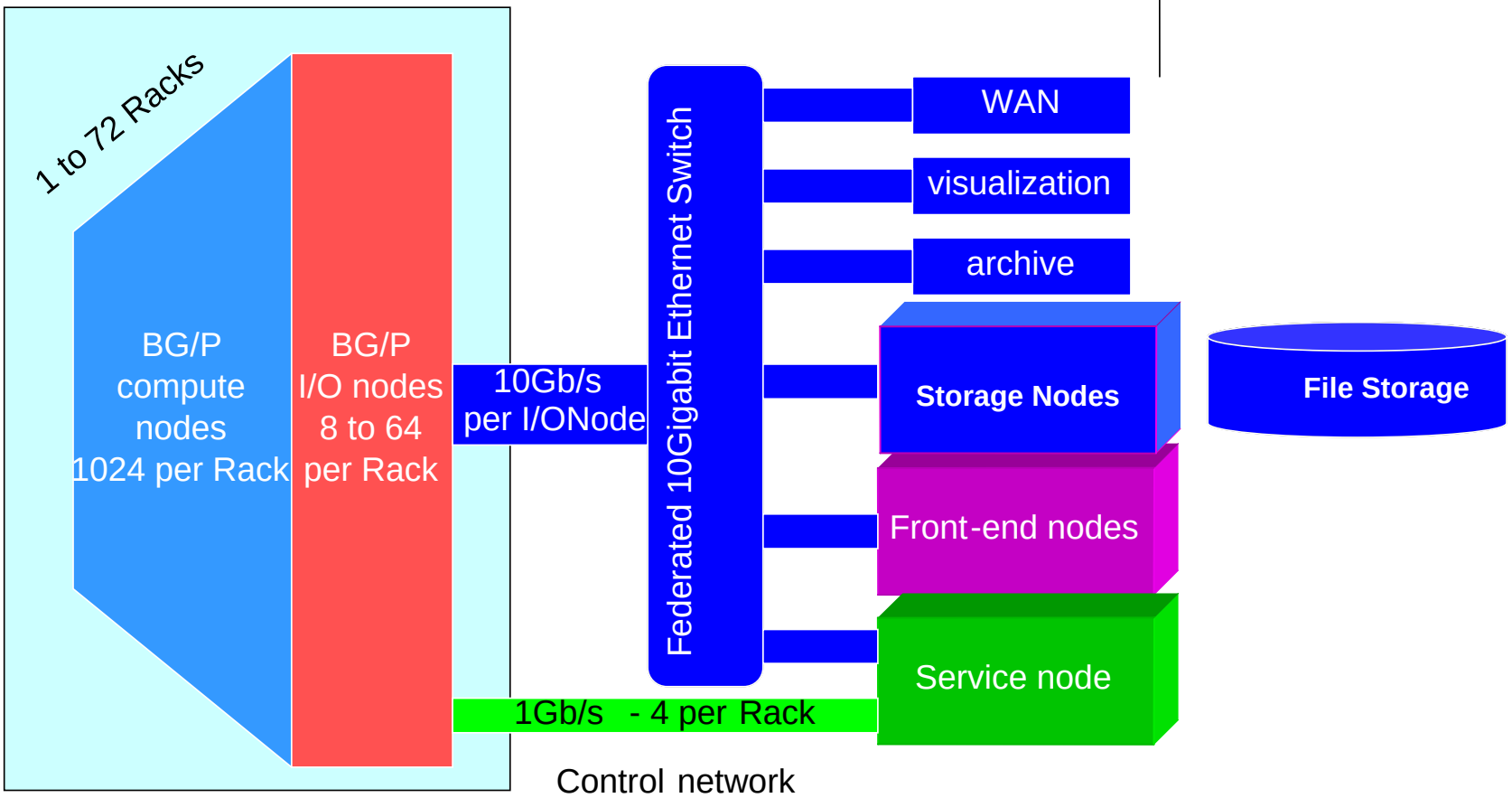
- Latency of one way to reach all 72K nodes 0.65 µs, MPI 1.6 µs

## Other networks

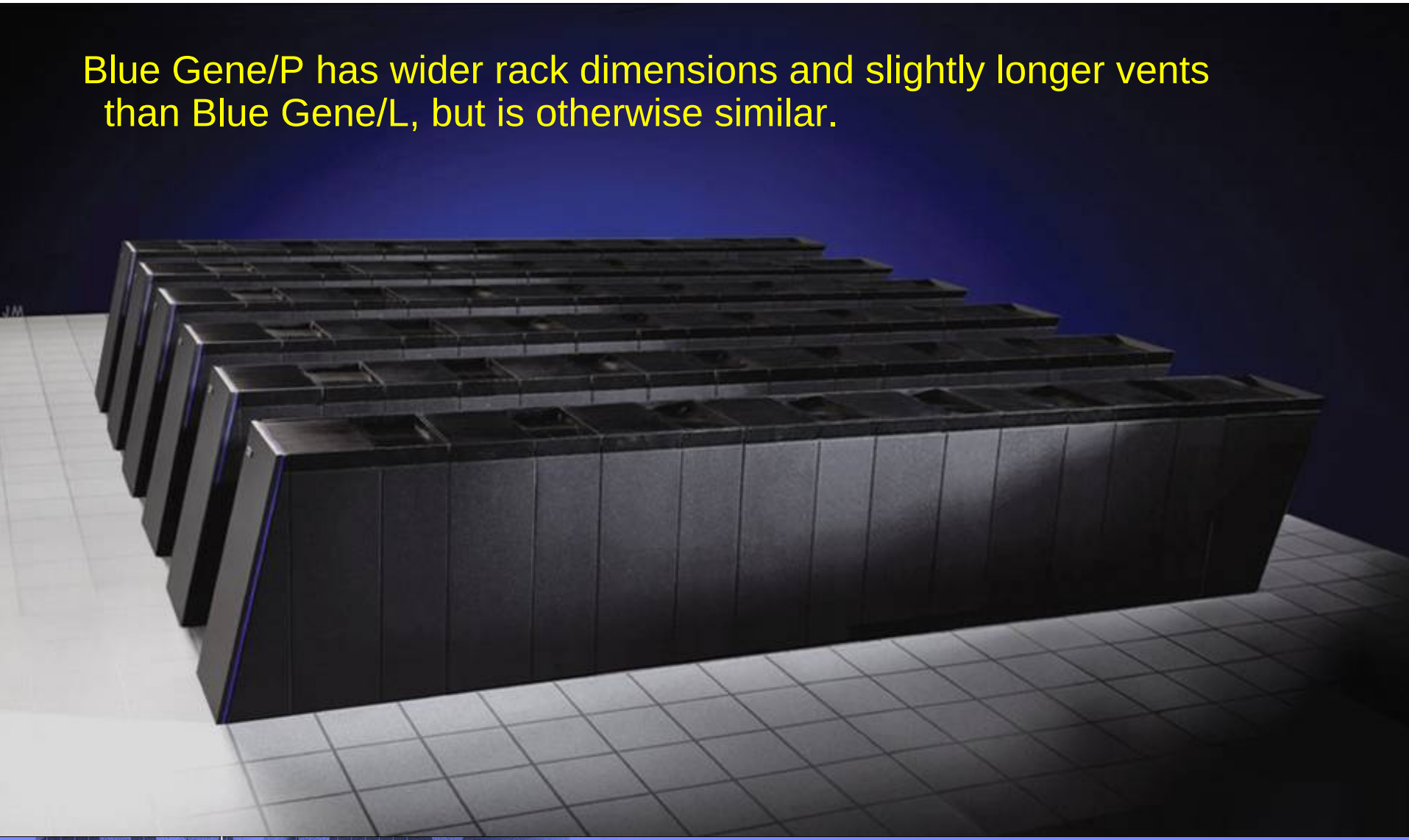
- 10Gb Functional Ethernet
- I/O nodes only
- 1Gb Private Control Ethernet
- Provides JTAG access to hardware. Accessible only from Service Node system

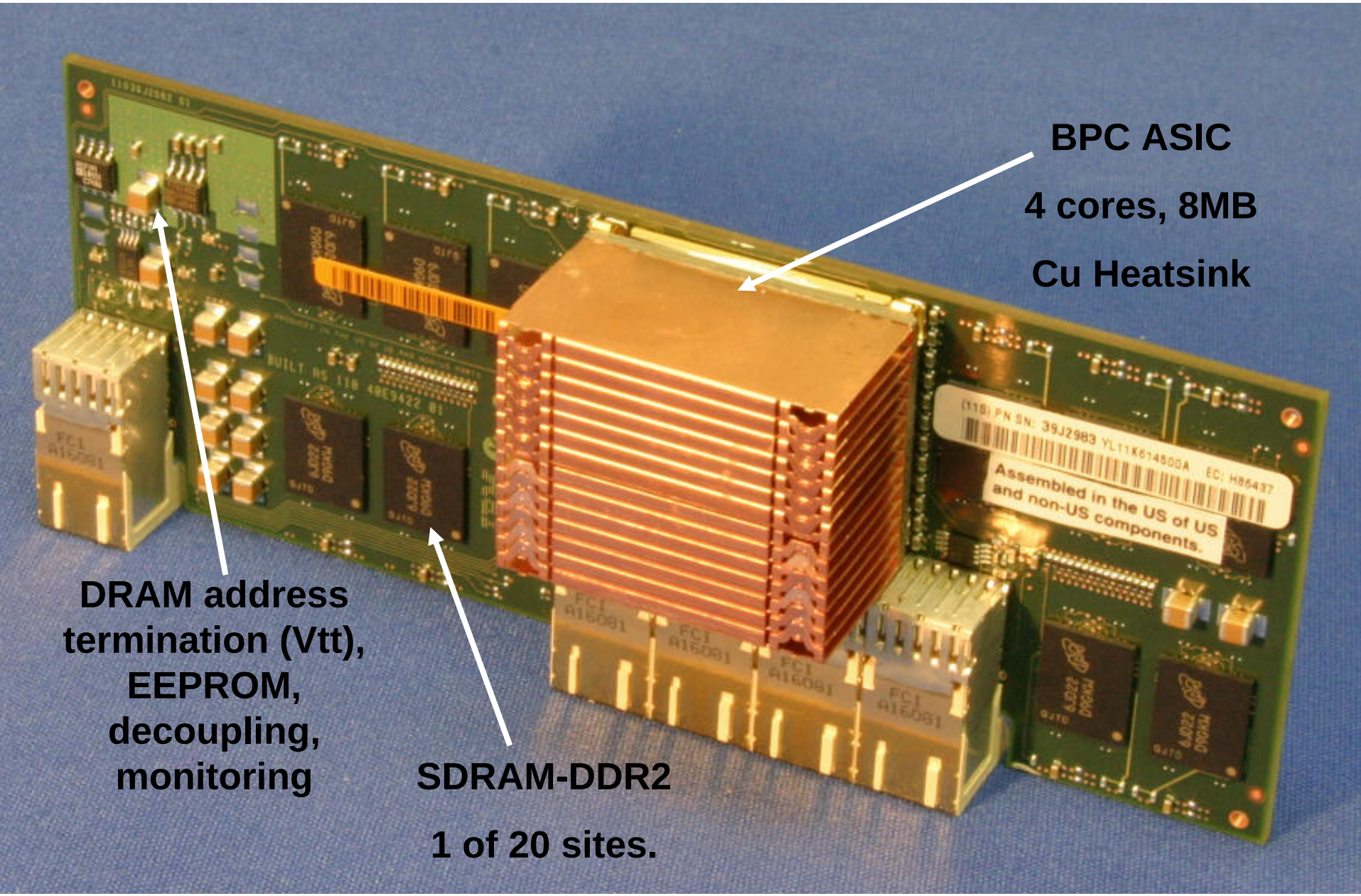
# Blue Gene/P System in a Complete Configuration

## Complete Blue Gene/P Solution

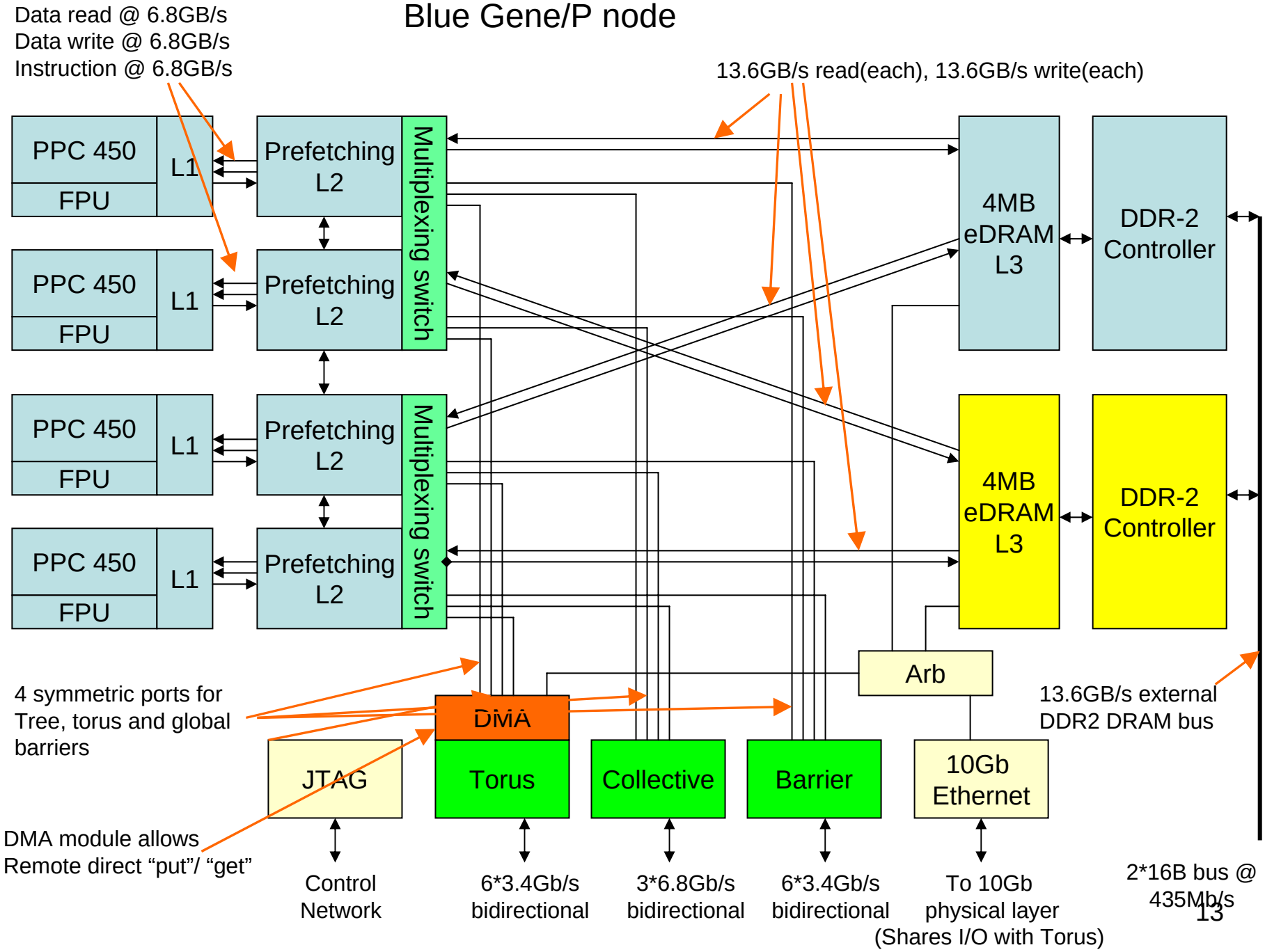


Blue Gene/P has wider rack dimensions and slightly longer vents than Blue Gene/L, but is otherwise similar.

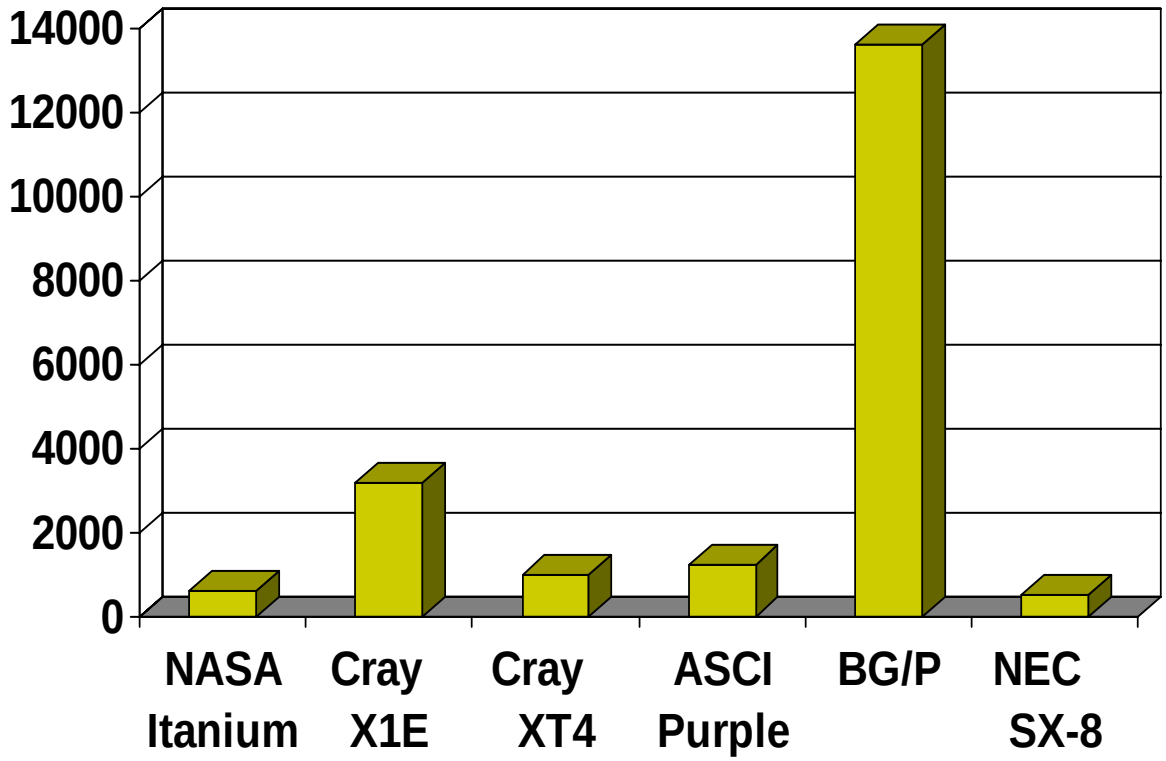




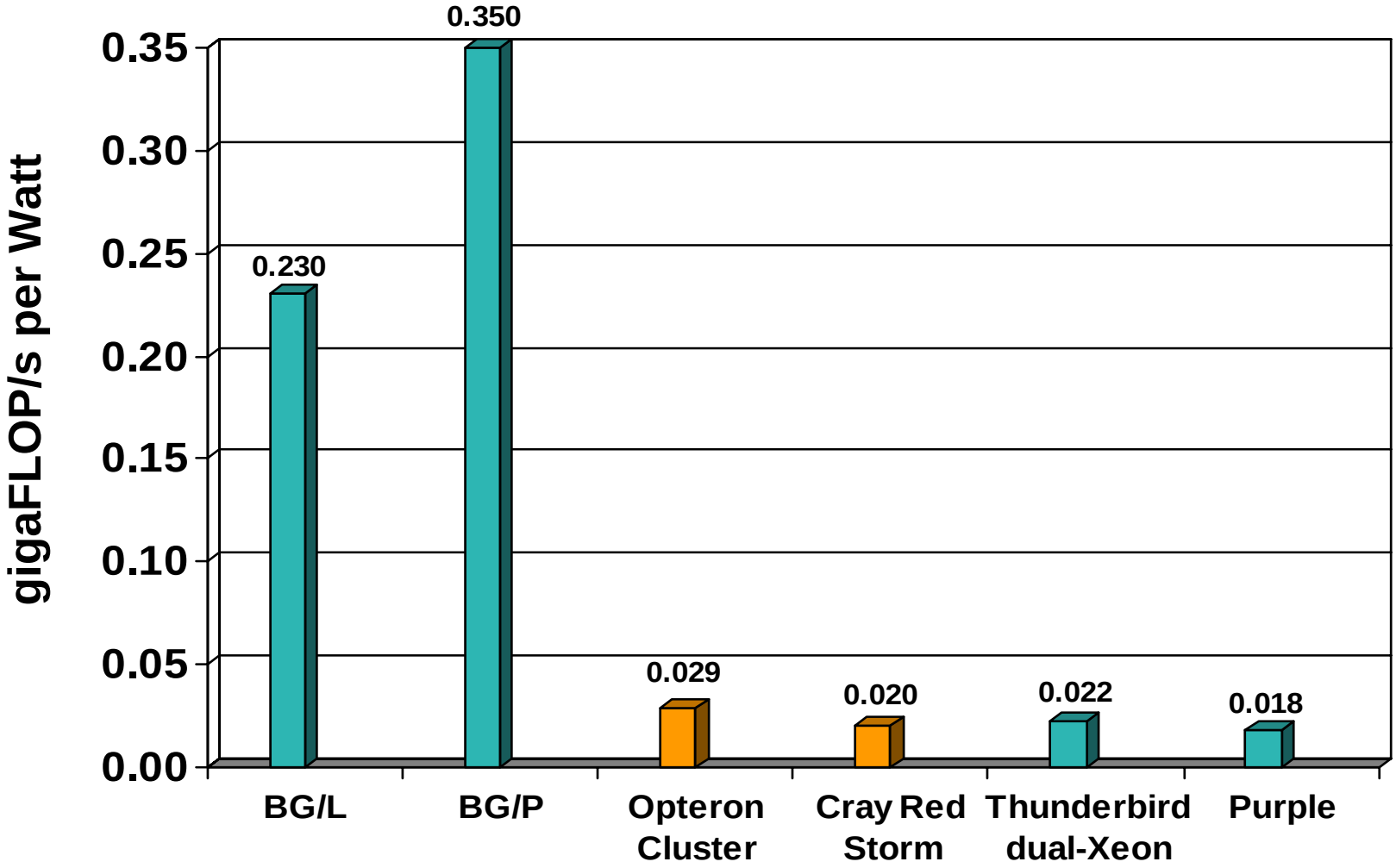
# Blue Gene/P node



# Main Memory Bandwidth per Rack

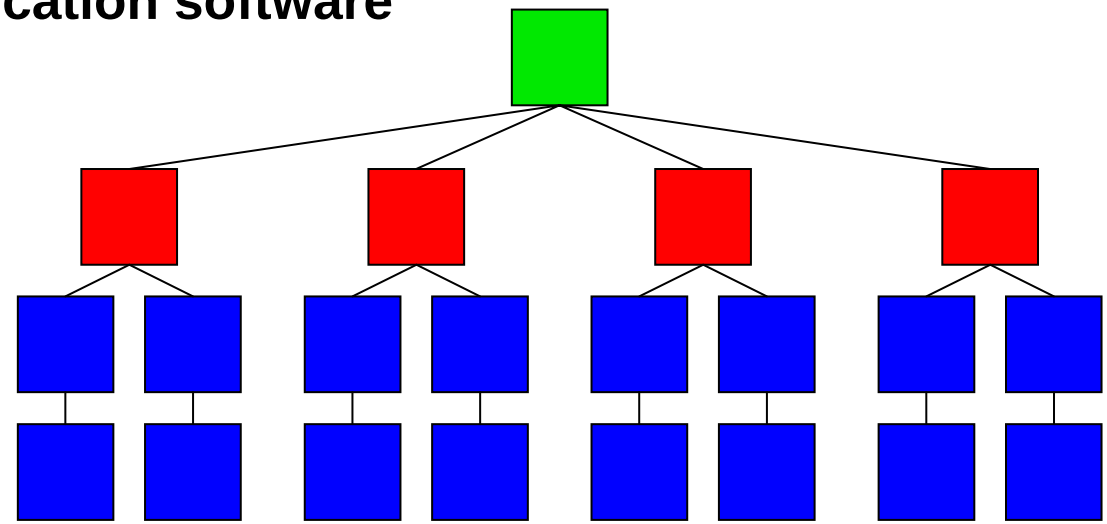


# System Power Efficiency



# Blue Gene Software Hierarchical Organization

- **Compute nodes dedicated to running user application, and almost nothing else - simple compute node kernel (CNK)**
- **I/O nodes run Linux and provide a more complete range of OS services – files, sockets, process launch, signaling, debugging, and termination**
- **Service node performs system management services (e.g., heart beating, monitoring errors) - transparent to application software**



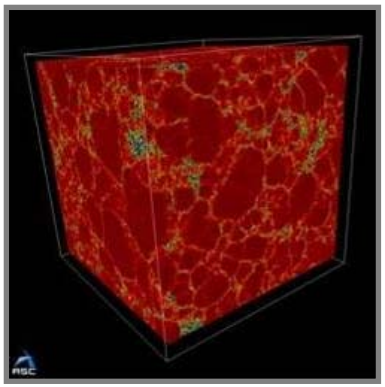
# Supporting Equipment

- Front End Nodes
  - IBM system-p 64-bit servers (not Blue Gene)
  - Used for development of application code
  - XL compilers, GNU compilers, libraries available
- Storage Nodes
  - A parallel filesystem is generally shared between to Blue Gene (the I/O nodes) and the Front End Nodes

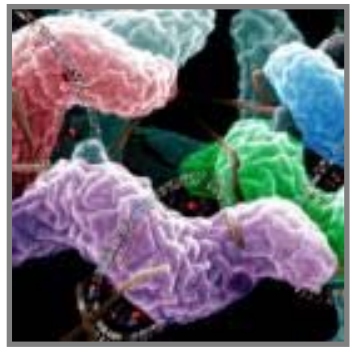
# Blue Gene Software

- Compilers
- Message Passing Library
- ESSL & MASS Libraries
- GPFS File System
- LoadLever scheduler
- HPC Toolkit
- Management software on Management Node
- Compute Node Kernel on Compute Nodes
- Linux on I/O nodes

# What is Blue Gene used for?



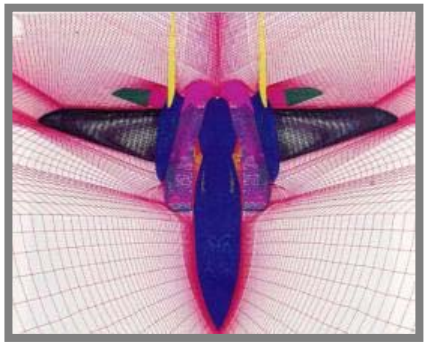
Materials Science




Pandemic Research

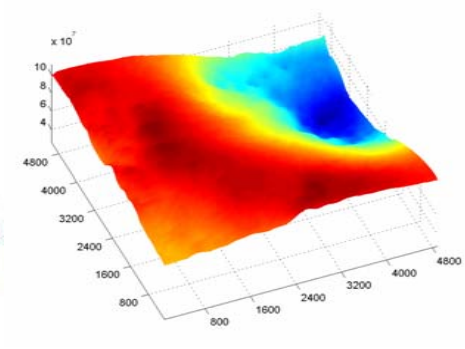

Drug Discovery

Fluid Dynamics



Climate Modeling

Financial Modeling


Geophysical Data Processing

# Applications already tested and running on BG/P

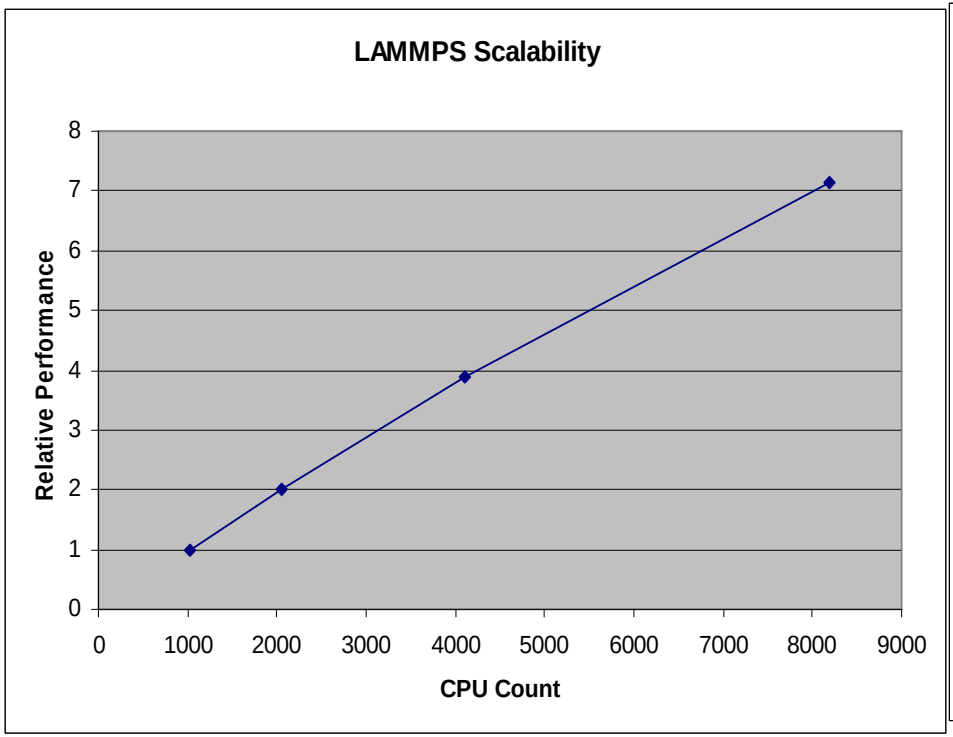
- Benchmarks
  - Linpack, HPC Challenge
  - NAS Serial, NAS OpenMP, NAS Parallel
  - PALLAS, STREAM, MPPTest
  - UMT2K, SPPM, SPHOT
- Applications
  - Physics:  SPHOT, QCD, MILC, ICEPIC
  - Weather/Climate:  HOMME, HYCOM, WRF, POP
  - Astrophysics:  FLASH
  - CFD:  Raptor, AVUS, CTH, AMR
  - Molecular Dynamics:  NAMD, LAMMPS
  - Quantum Chemistry:  CPMD, GAMESS
  - Materials Science:  ParaTEC
- All applications from BG/L will run on BG/P after recompile
  - General performance improvement of 2.4 going from BG/L to BG/P
  - Some applications have ratio >2.4 due to “superscaling”
    - SMP mode with 4 threads on BG/P can achieve better performance than BG/L nodes
- Many more applications in progress of being ported over and tested

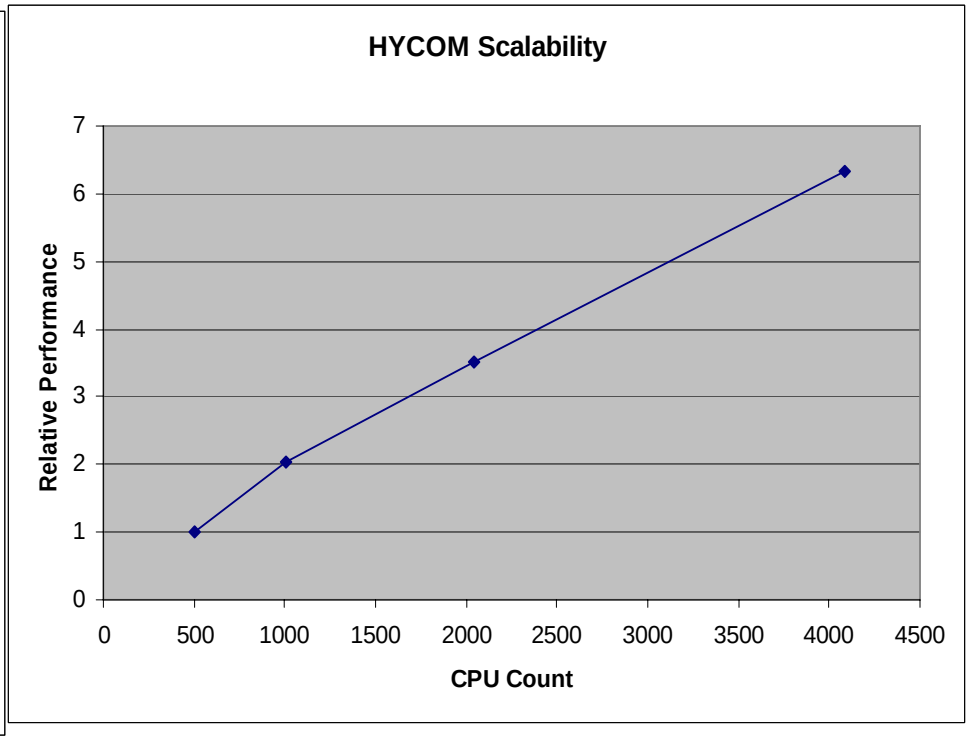
# Linpack on Blue Gene/P

- ISC2007 Top 500 submission
  - 2 Racks, 1 MPI Process / Core, 8192 MPI Processes Total
    - used basic NETLIB HPL with tweaks
    - used BG/L version of ESSL
  - 20.86 TF, 74.89% of peak
  - More performance to come

# Scalability – Continues Extraordinary Blue Gene Behavior

- Two Examples
  - LAMMPS - molecular dynamics
  - HYCOM - ocean modeling



# Summary

## Blue Gene is addressing six critical issues on the path to Petaflop computing

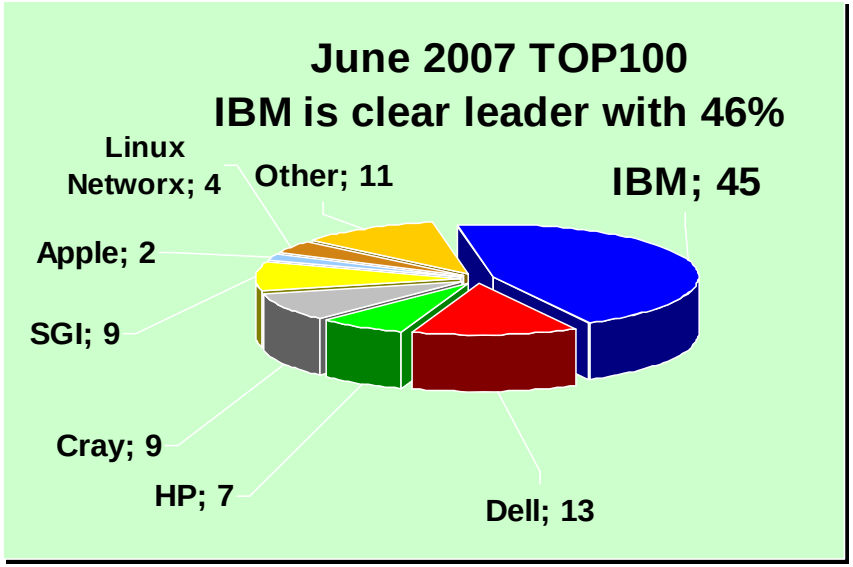
1. Power
2. Floor space
3. Cost
4. Single processor performance
5. Network scalability
6. Reliability

# Thank you!

# Top 10 reasons you need Blue Gene/P

1. Ultra-scalability for breakthrough science
   - Blue Gene/P: up to 294,912 cores, or 73,728 nodes
   - Cluster: typically 512-1024 nodes or less.
2. Highest capability machine in the world
3. Highest reliability, highest HPC MTBF/TF (10-100X), low maintenance staff
4. Low power (~4-10X), smallest footprint, lowest TCO (total cost of ownership)
5. Broad range of scientific applicability at superior cost/performance
6. High bandwidth for interprocessor communication (7.5X compared to typical clusters)
7. Low latency, high bandwidth memory system and interprocessor communication system
8. Familiar programming models: MPI, OpenMP, POSIX I/O
9. Reproducible, deterministic runs, easy to trace errors, and tune performance
10. Huge total memory bandwidth for data intensive applications such as search

# IBM supercomputing leadership

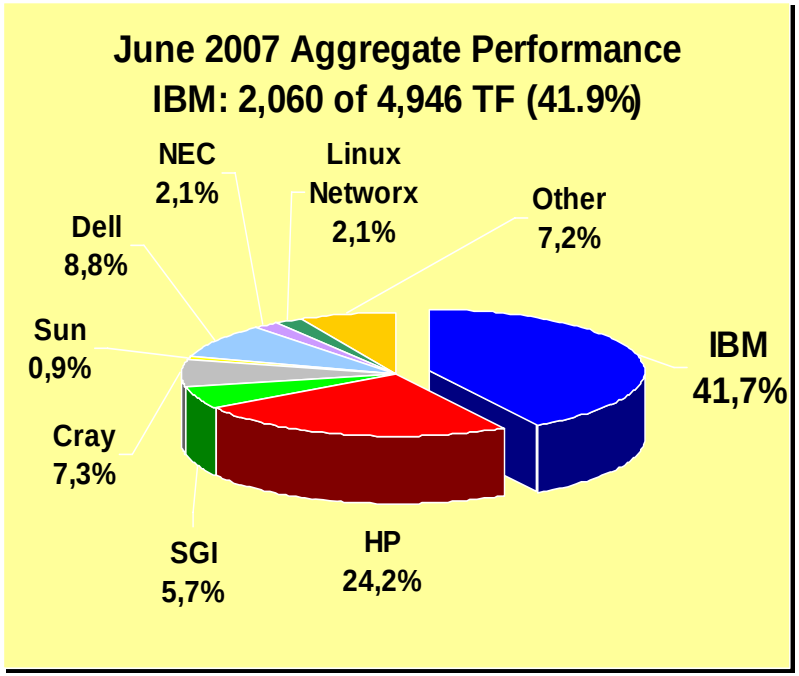



TOP500 SUPERCOMPUTER SITES

***Semiannual independent ranking of top 500 supercomputers in the world***

## IBM leads several key TOP500 metrics ...

- ✓ #1 System – LLNL – Blue Gene/L (280.6 TF)
- ✓ Most installed aggregate throughput with over 2,060 Teraflops (41.7%)
- ✓ Most in TOP10 with 6 systems (40%)
- ✓ Most in TOP100 systems with 45 (45%)
- ✓ Fastest machines in USA (BG/L)
- ✓ Fastest machine in Europe (MareNostrum)
- ✓ Fastest machine in China (Sinopec)



# More Information

- IBM Redbooks for Blue Gene **ibm.com/redbooks**
  - Application Development Guide
  - System Administration Guide
  - Performance Tools
- Open Source Communities

# Read more about it…

http://www.ibm.com/servers/deepcomputing/bluegene.html

**Sign up for newsletter**

http://www-fp.mcs.anl.gov/bgconsortium/default.htm

**Sign up for a consortium membership**

http://www.top500.org

**Latest TOP500 list**