IBM Blue Gene Architecture

Outline

- Overview of BlueGene
- BG Philosophy
- The BG Family
- BG hardware
  - System Overview
  - CPU
  - Node
  - Interconnect
- BG System Software
- BG Software Development Environment
  - Compilers
  - Supported HPC APIs
  - Numerical Libraries
  - Performance Tools
  - Debuggers

- BG References

Blue Gene Overview

High-Level View of the Blue Gene Architecture:

Within Node

- Low latency, high bandwidth memory system
- Good floating point performance
- Low power per node

Across Nodes

- Low latency, high bandwidth interconnect
- Special networks for MPI Barriers, Global messages

Software Development Environment

- Familiar HPC “standards” - MPI, Fortran, C, C++

BlueGene/L scales to 360 TF with modified COTS and custom parts

Cabinet assembly at IBM Rochester

Compute Chip

2 processors
2.8/5.6 GF/s
4 MiB* eDRAM

(compare this with a 1988 Cray YMP/8 at 2.7 GF/s)

Compute Card
I/O Card

FRU (field replaceable unit)
25mmx32mm
2 nodes (4 CPUs)
(2x1x1)
2.8/5.6 GF/s
256/512 MiB* DDR

Node Card

16 compute cards
0-2 I/O cards
32 nodes
(64 CPUs)
(4x4x2)
90/180 GF/s
8 GiB* DDR

Cabinet

2 midplanes
1024 nodes
(2,048 CPUs)
(8x8x16)
2.9/5.7 TF/s
256 GiB* DDR
15-20 kW

System

64 cabinets
65,536 nodes
(131,072 CPUs)
(32x32x64)
180/360 TF/s
16 TiB*
1.2 MW
2,500 sq.ft.
MTBF 6.16 Days

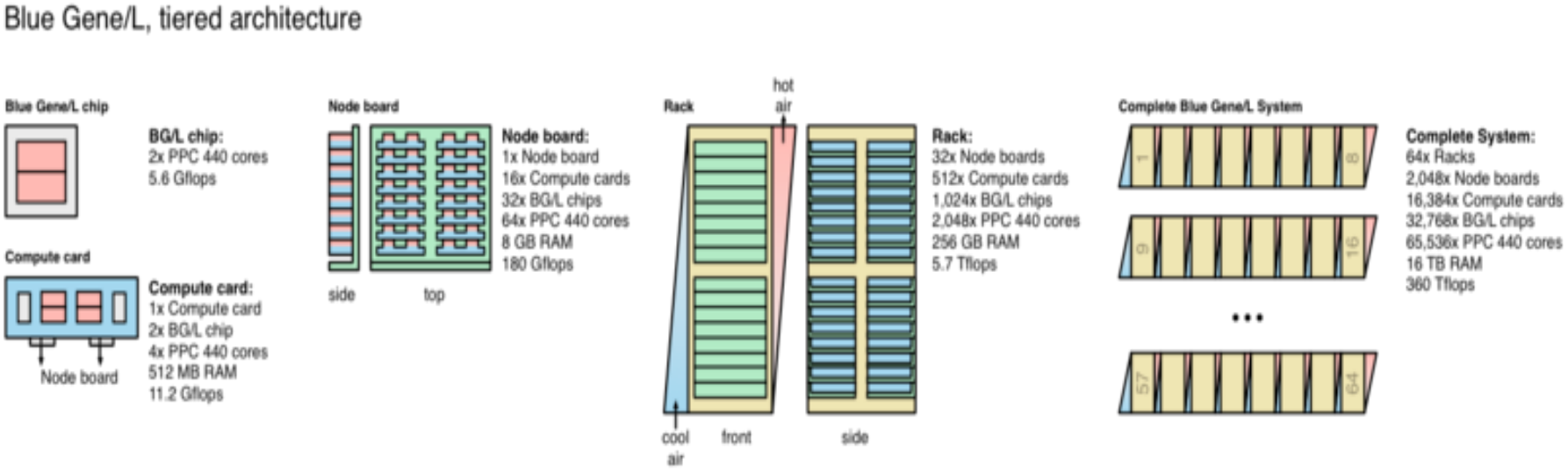
* http://physics.nist.gov/cuu/Units/binary.html

Blue Gene/L

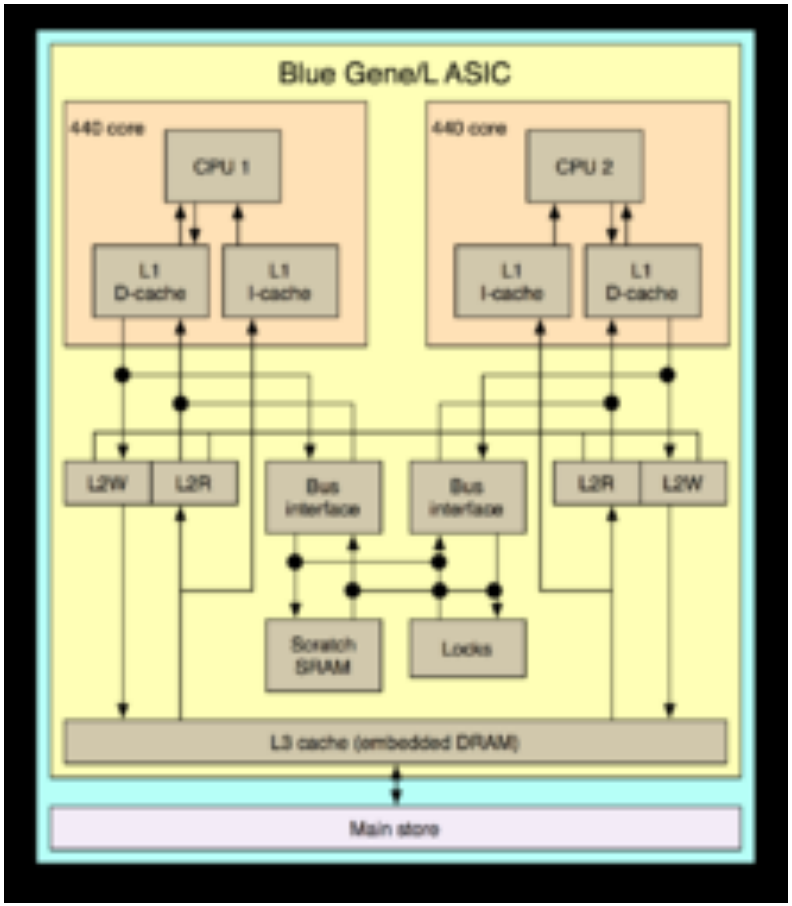
- System Architecture
  - "many" nodes – powers of 2
  - Electrically isolated partitions – allowing separate booting
- Node Architecture
  - System-on-Chip (SoC) technology
  - Dual-Core CPU
    - Low power per Flop
  - Compute, I/O, login, Admin nodes
- Interconnect Architecture
  - Multiple components
    - 3D torus for point-to-point MPI
    - Globals/Collectives network
    - Barrier network
    - Administration network
- Parallel File System
  - GPFS
- System Software
  - Lightweight OS for batch nodes
  - Linux OS for login, I/O, admin nodes,

- Software Development Environment
  - HPC Standards
  - Compute, login and I/O nodes

Blue Gene L


Blue Gene/L, tiered architecture
Blue Gene/L chip
BG/L chip:
2x PPC 440 cores
5.6 Gflops
Compute card
Compute card:
1x Compute card
2x BG/L chip
4x PPC 440 cores
512 MB RAM
11.2 Gflops
Node board
Node board:
1x Node board
16x Compute cards
32x BG/L chips
64x PPC 440 cores
8 GB RAM
180 Gflops
side
top
Rack
hot air
Rack:
32x Node boards
512x Compute cards
1,024x BG/L chips
2,048x PPC 440 cores
256 GB RAM
5.7 Tflops
cool air
front
side
Complete Blue Gene/L System
1
8
9
16
57
64
Complete System:
64x Racks
2,048x Node boards
16,384x Compute cards
32,768x BG/L chips
65,536x PPC 440 cores
16 TB RAM
360 Tflops

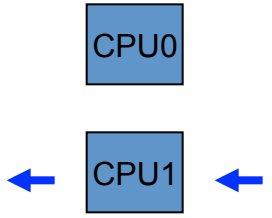
BG/L ASIC


Blue Gene/L ASIC
440 core
CPU 1
L1
D-cache
L1
I-cache
440 core
CPU 2
L1
I-cache
L1
D-cache
L2W
L2R
Bus
interface
Bus
interface
L2R
L2W
Scratch
SRAM
Locks
L3 cache (embedded DRAM)
Main store

Two ways for apps to use hardware

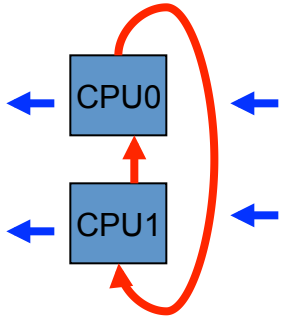
Mode 1 (Co-processor mode - CPM):

- CPU0 does all the computations
- CPU1 does the communications
- Communication overlap with computation
- Peak comp perf is 5.6/2 = 2.8 GFlops



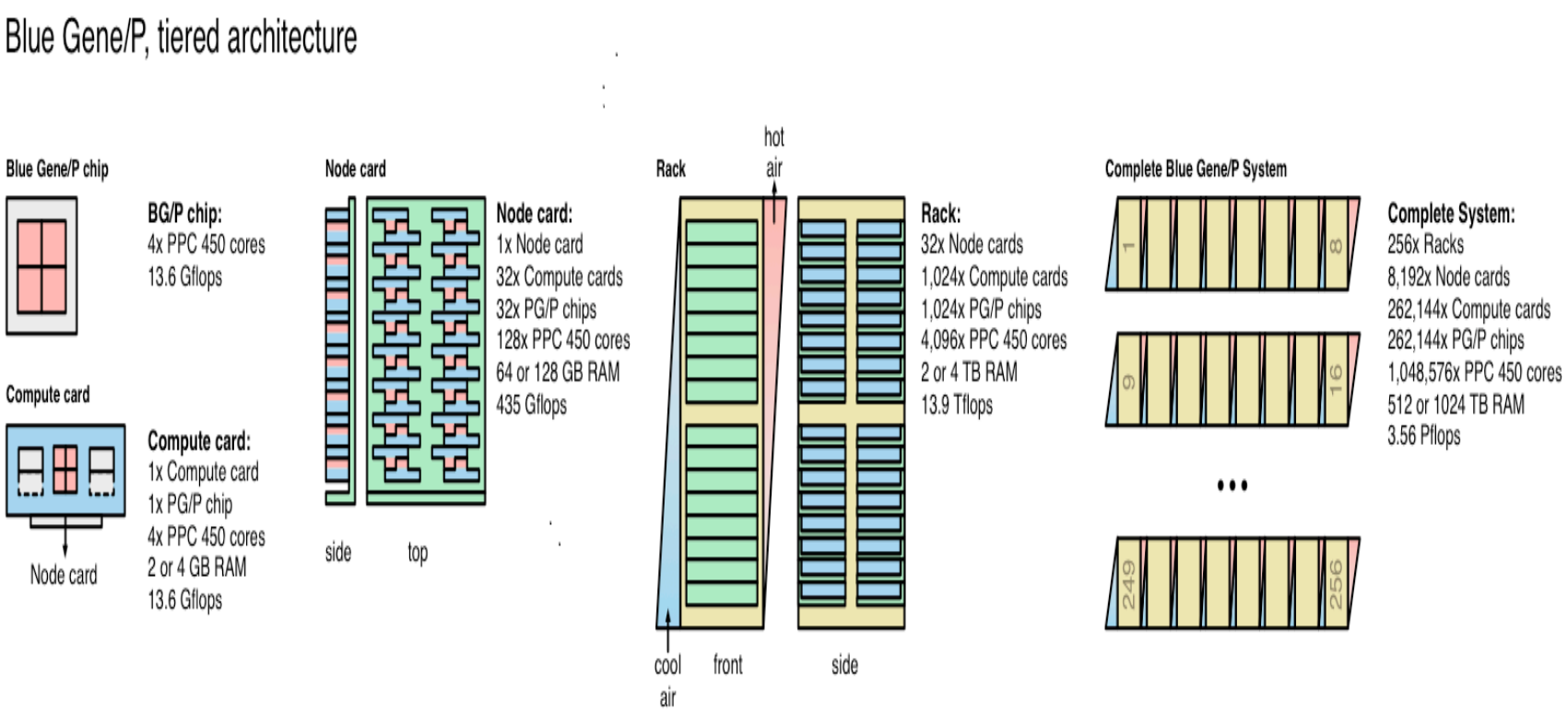
Mode 2 (Virtual node mode - VNM):

- CPU0, CPU1 independent “virtual tasks”
- Each does own computation and communication
- The two CPU’s talk via memory buffers
- Computation and communication cannot overlap
- Peak compute performance is 5.6 GFlops

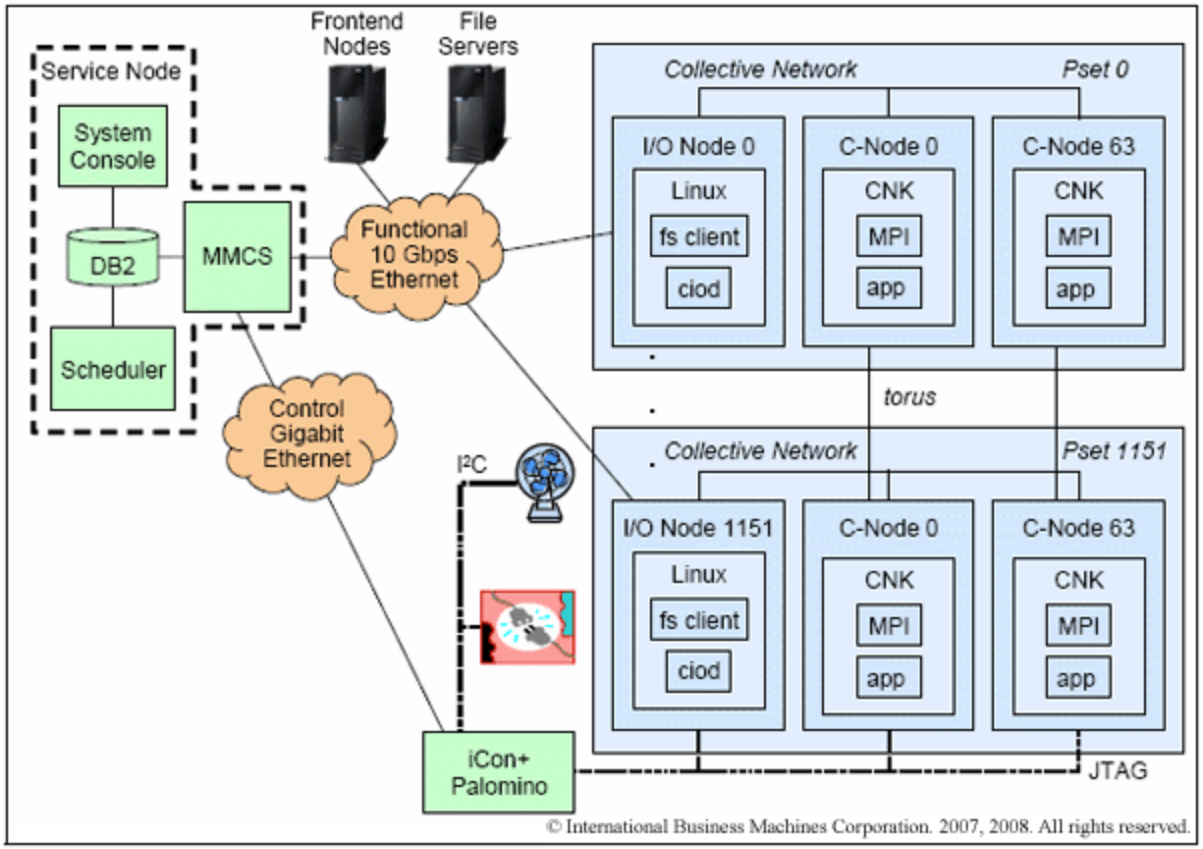


BG/P – LLNL Dawn

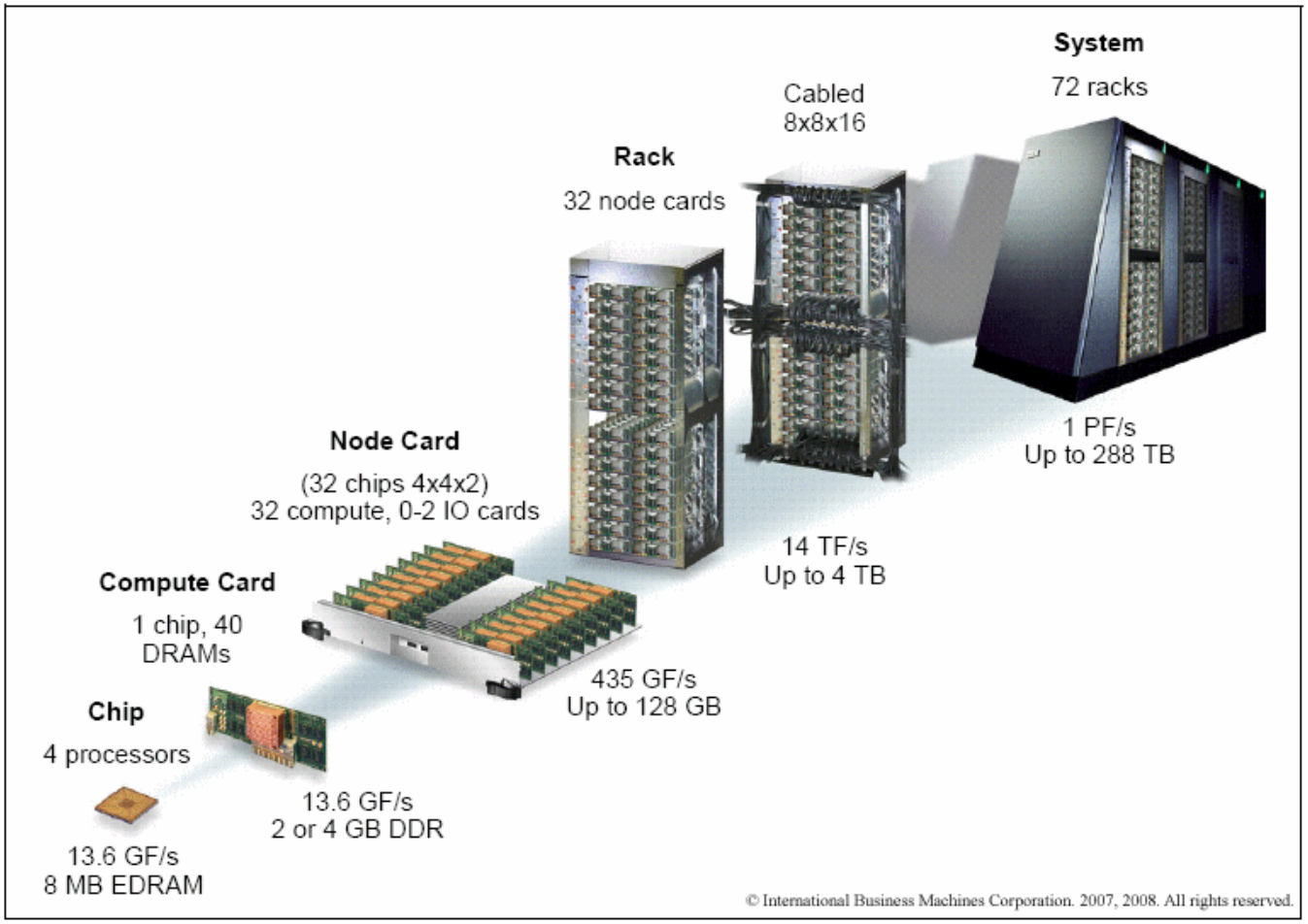
Blue Gene P


Blue Gene/P, tiered architecture
Blue Gene/P chip
BG/P chip:
4x PPC 450 cores
13.6 Gflops
Compute card
Node card
Compute card:
1x Compute card
1x PG/P chip
4x PPC 450 cores
2 or 4 GB RAM
13.6 Gflops
Node card
side
top
Node card:
1x Node card
32x Compute cards
32x PG/P chips
128x PPC 450 cores
64 or 128 GB RAM
435 Gflops
Rack
hot air
cool air
front
side
Rack:
32x Node cards
1,024x Compute cards
1,024x PG/P chips
4,096x PPC 450 cores
2 or 4 TB RAM
13.9 Tflops
Complete Blue Gene/P System
1
8
9
16
249
256
Complete System:
256x Racks
8,192x Node cards
262,144x Compute cards
262,144x PG/P chips
1,048,576x PPC 450 cores
512 or 1024 TB RAM
3.56 Pflops

BG/P General Configuration



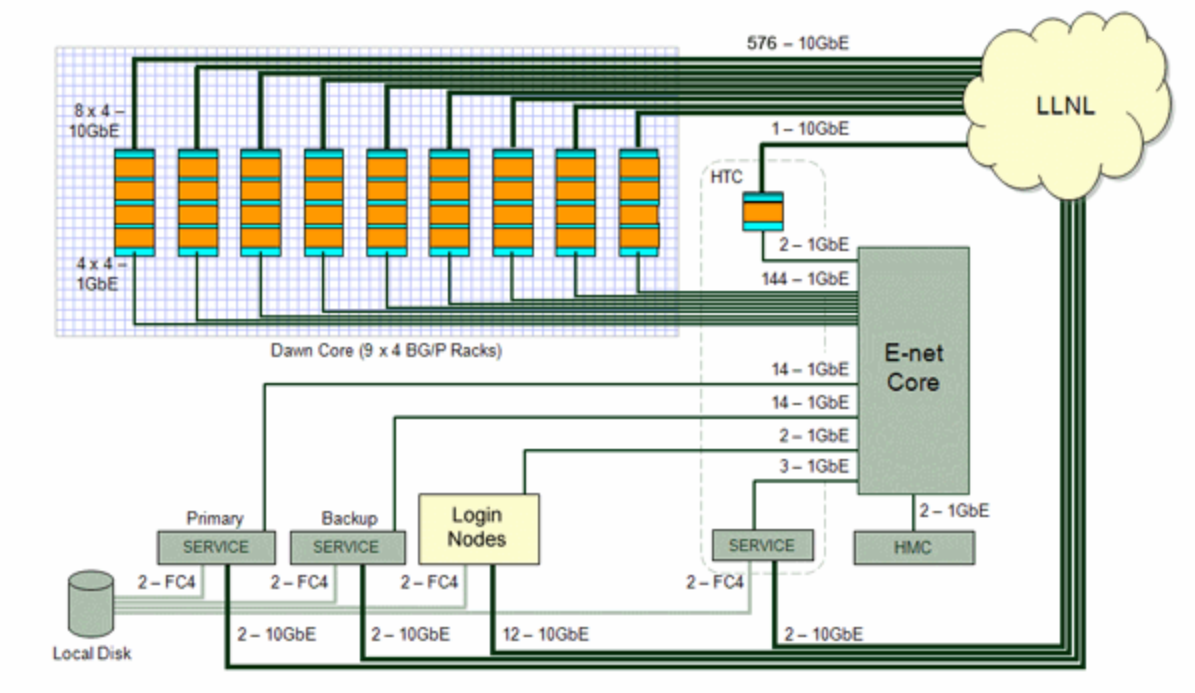
BG/P Scaling Architecture



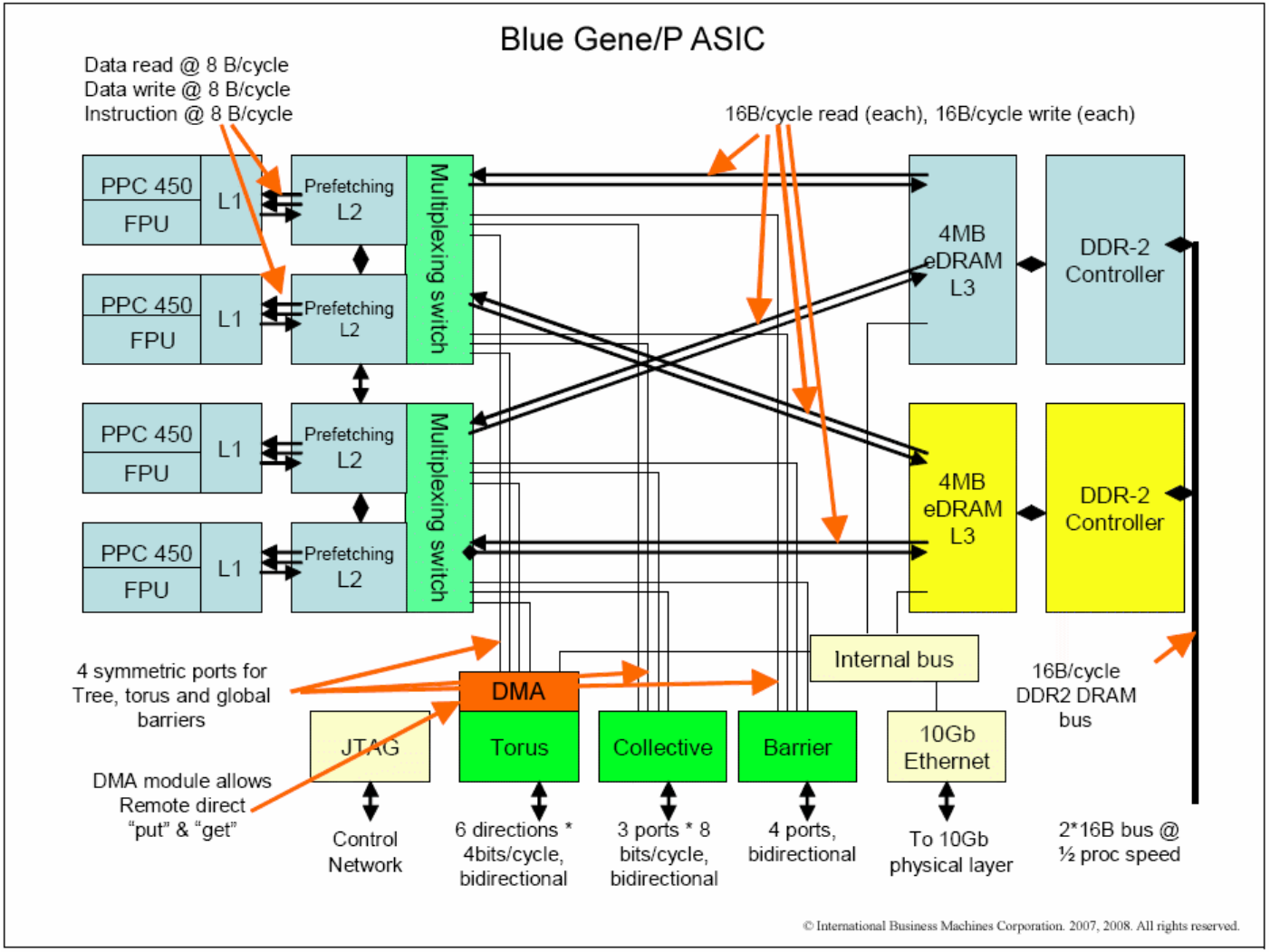
BG/P vs BG/L

Feature	Blue Gene/L	Blue Gene/P
Node		
Cores per node	2	4
Clock speed	700 MHz	850 MHz
Cache coherency	Software managed	SMP hardware
L1 cache (private)	32 KB data; 32 KB instruction	32 KB data; 32 KB instruction
L2 cache (private)	2 KB; 14 stream prefetching	2 KB; 14 stream prefetching
L3 cache (shared)	4 MB	8 MB
Memory per node	512 MB - 1 GB	2 GB - 4 GB
Memory bandwidth	5.6 GB/s (16 bytes wide)	13.6 GB/s (2x16 bytes wide)
Peak performance	5.6 Gflops/node	13.6 Gflops/node
Torus Network		
Bandwidth	2.1 GB/s (via Core)	5.1 GB/s (via DMA)
Hardware Latency (nearest neighbor)	<1 us	<1 us
Hardware Latency (worst case - 68 hops)	7 us	5 us
Collective Network		
Bandwidth	2.1 GB/s	5.1 GB/s
Hardware Latency (round-trip, worst case - 72 racks)	5 us	3 us
Functional Ethernet		
Capacity	1 Gb	10 Gb
Performance Monitors		
Counters	48	256
Counter resolution (bits)	32	64
System Properties (72 racks)		
Peak Performance	410 Tflops	1 Pflops
Area	150 m^2	200 m^2
Total Power (LINPACK)	1.9 MW	2.9 MW
GFlops per Watt	0.23	0.34

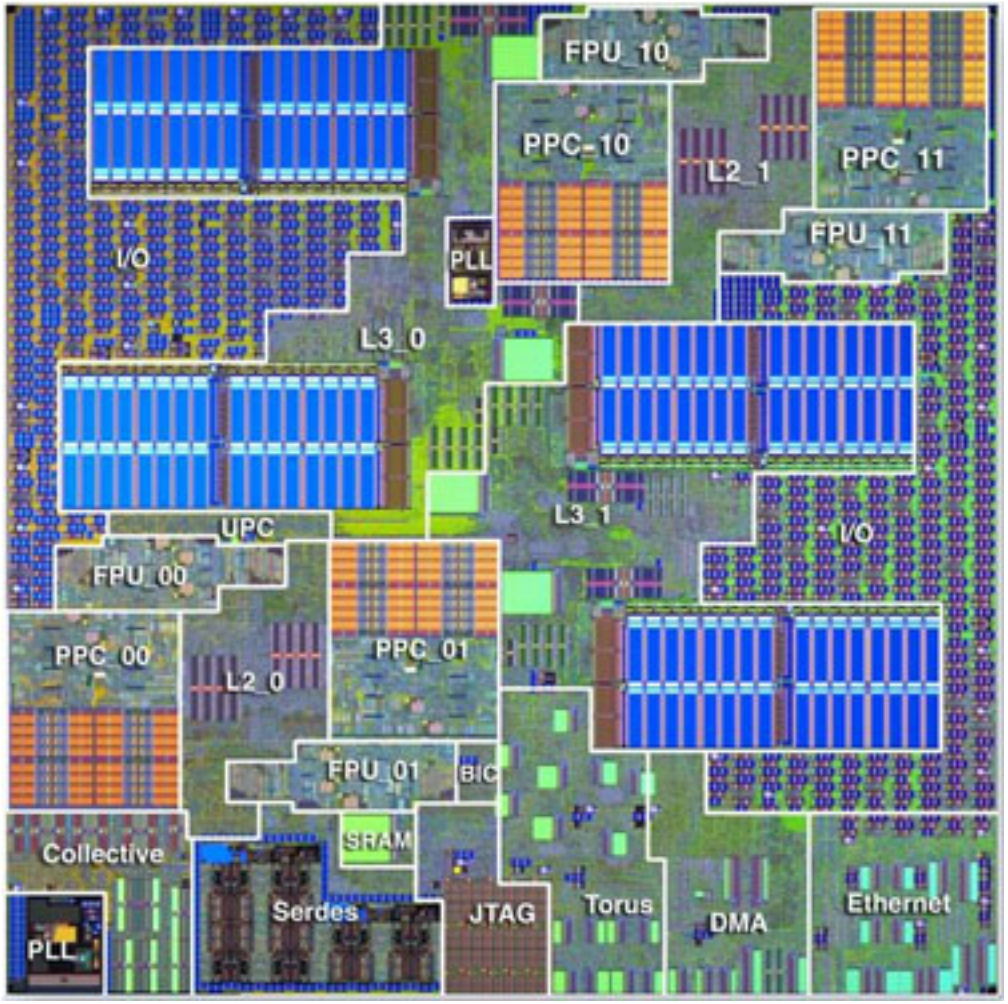
Dawn BG/P External and Service Networks



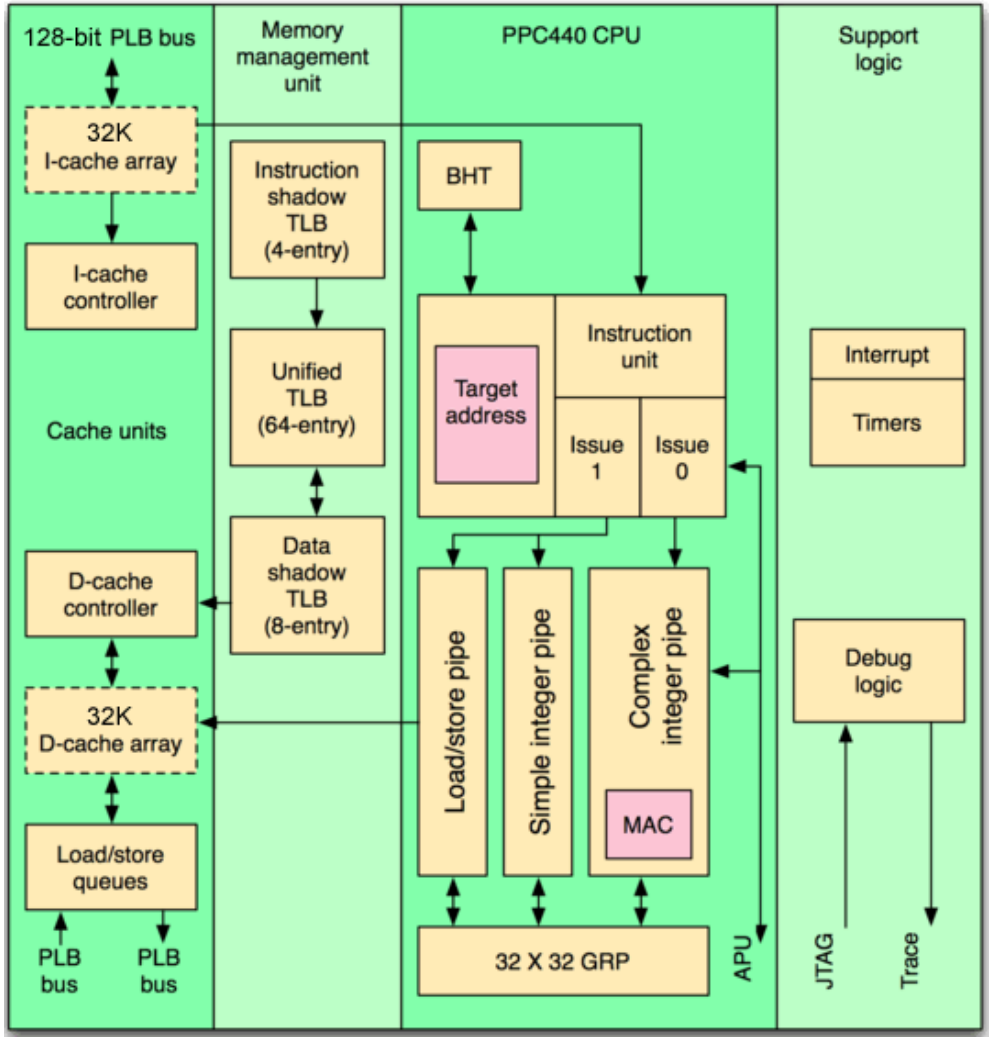
BG/P ASIC


Blue Gene/P ASIC
Data read @ 8 B/cycle
Data write @ 8 B/cycle
Instruction @ 8 B/cycle
16B/cycle read (each), 16B/cycle write (each)
PPC 450
FPU
L1
Prefetching
L2
Multiplexing switch
4MB
eDRAM
L3
DDR-2
Controller
4 symmetric ports for
Tree, torus and global
barriers
DMA
Internal bus
16B/cycle
DDR2 DRAM
bus
JTAG
Torus
Collective
Barrier
10Gb
Ethernet
DMA module allows
Remote direct
"put" & "get"
Control
Network
6 directions *
4bits/cycle,
bidirectional
3 ports * 8
bits/cycle,
bidirectional
4 ports,
bidirectional
To 10Gb
physical layer
2*16B bus @
½ proc speed
© International Business Machines Corporation. 2007, 2008. All rights reserved.

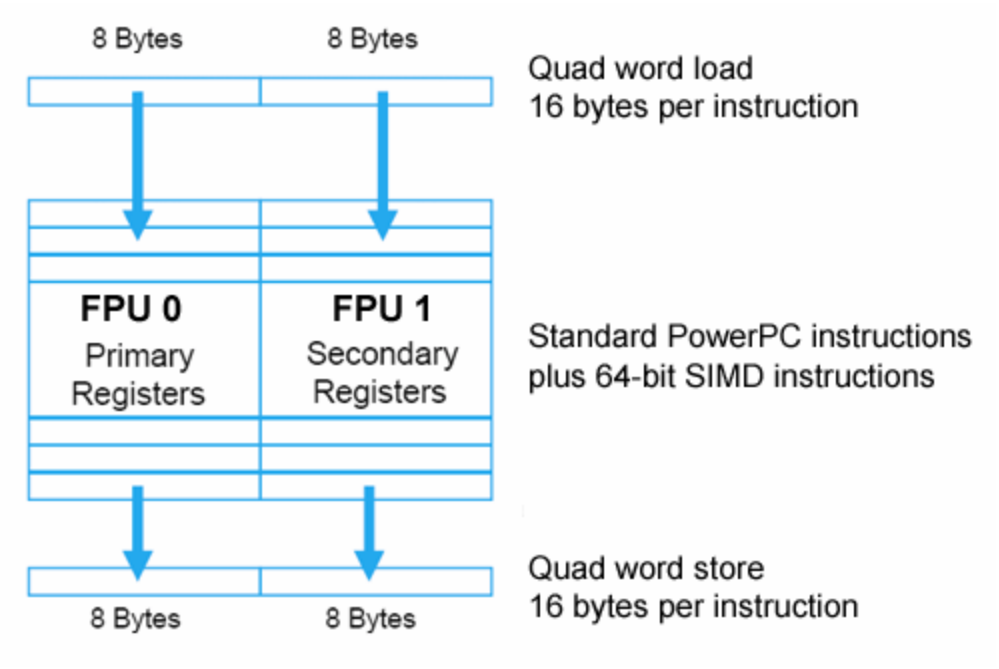
BG/P ASIC


FPU_10
PPC-10
L2_1
PPC_11
FPU_11
I/O
PLL
L3_0
L3_1
I/O
UPC
FPU_00
PPC_00
L2_0
PPC_01
FPU_01
BIC
SRAM
Collective
PLL
Serdes
JTAG
Torus
DMA
Ethernet

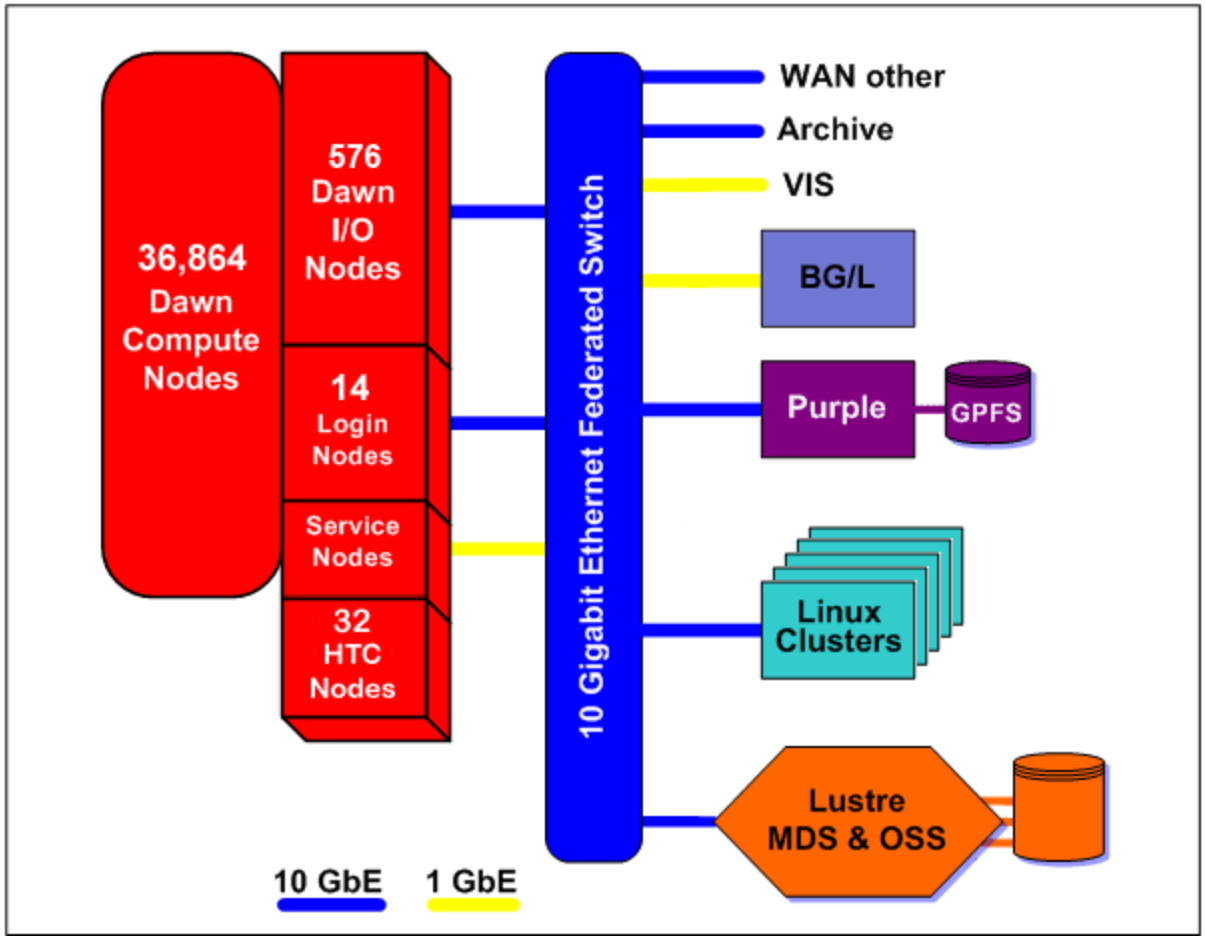
BG/P PowerPC 450 CPU



BG/P “Double Hummer” FPU

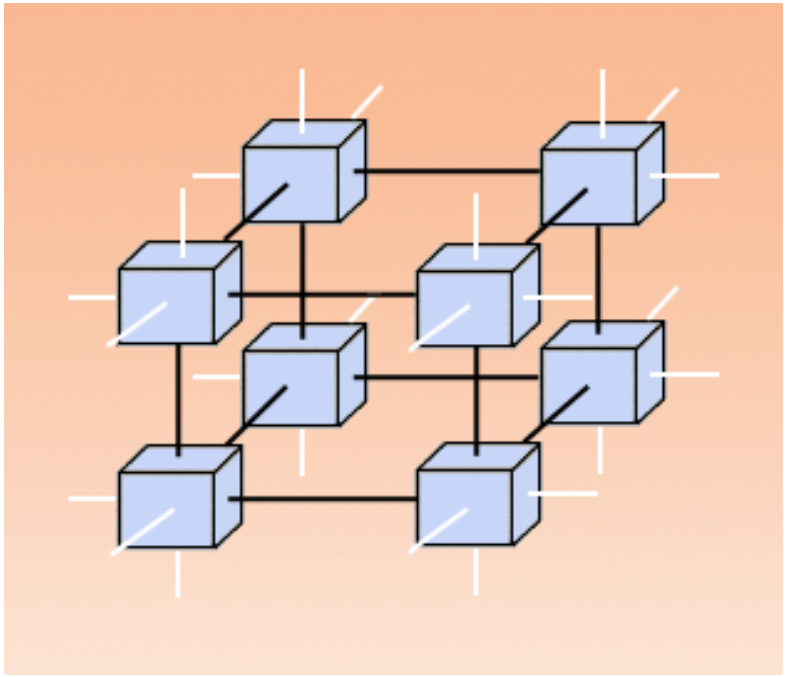


LLNL Dawn BG/P - Configuration



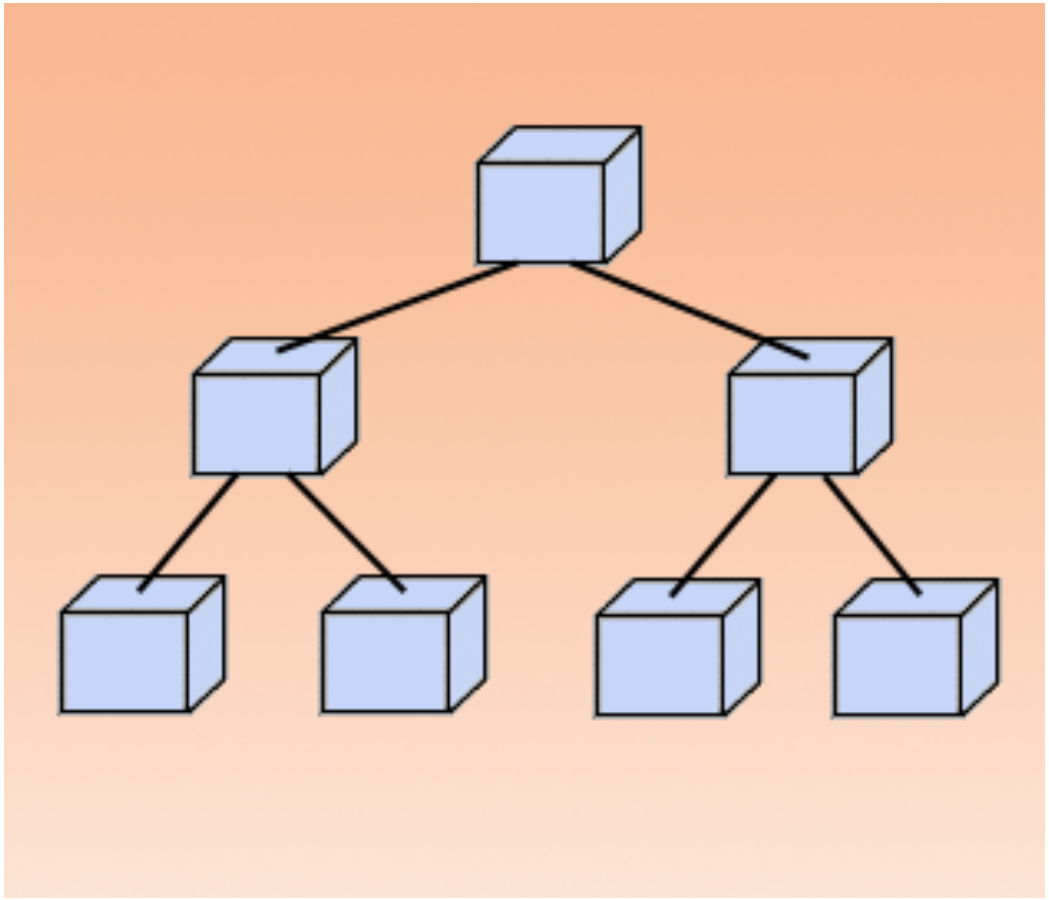
BG/P Interconnect – 3D Torus

- MPI point-to-point communications
- Each compute node connected to six nearest neighbors.
- Bandwidth: 5.1 GB/s per node (3.4 Gb/s bidirectional * 6 links/node)
- Latency (MPI): 3 us - one hop, 10 us to farthest
- DMA (Direct Memory Access) engine



BG/P Interconnect – Global Barrier/ Interrupt

- Connects to all compute nodes
- Low latency network for MPI barrier and interrupts
- Latency (MPI): 1.3 us for on
- Four links per node



BG/P interconnect – Global Collective

- Connects all

compute and I/O nodes

- One-to-all/all-to-all

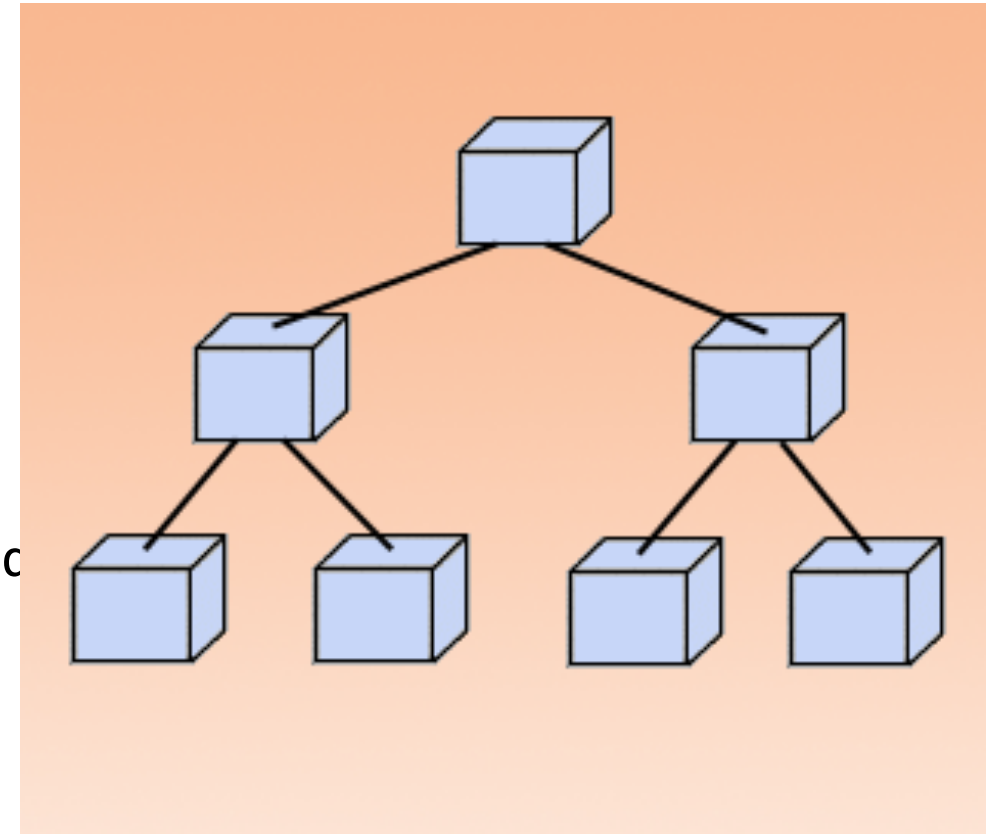
MPI broadcasts

- MPI Reduction

operations

- Bandwidth: 5.1 GB/s per noc
  node)
- Latency (MPI): 5 us for one

way tree traversal

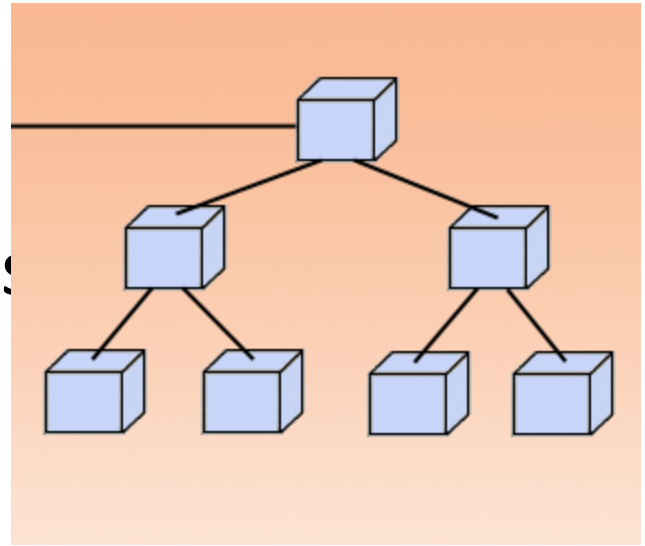


BG/P Interconnect – 10 Gb Ethernet

- Connects all I/O nodes to

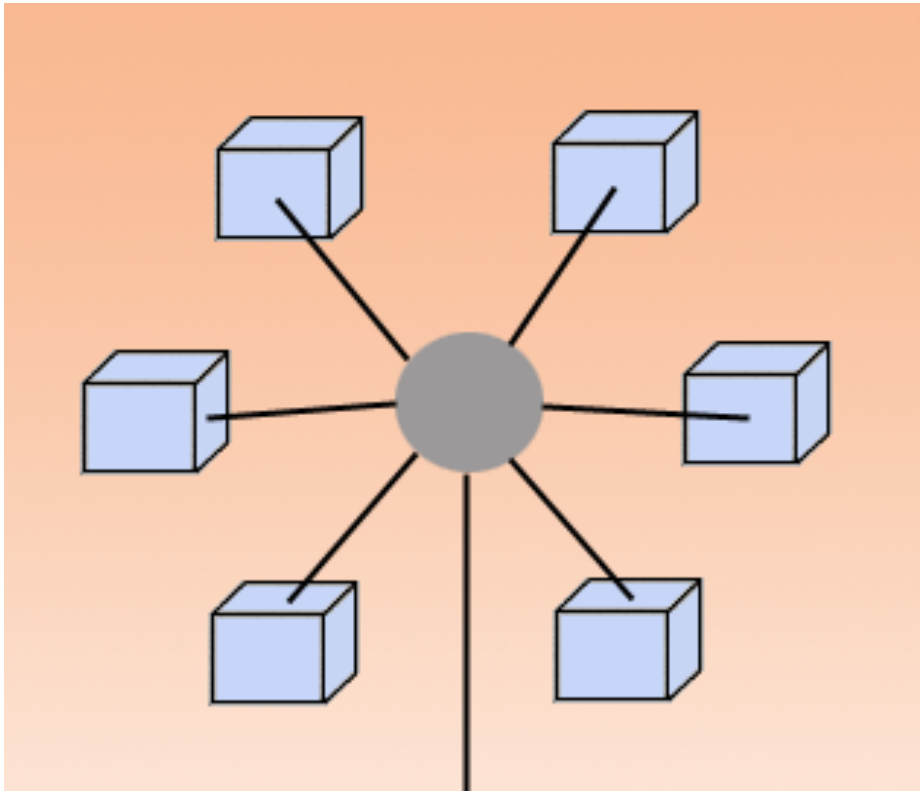
10 Gb Ethernet switch, for

access to external file systems



BG/P 1 Gb Control Ethernet

- IEEE 1149.1 interface
- Gives service node direct access to all nodes. Used for system boot, debug, monitoring.
- Provides non-invasive access to performance counters.



Dawn BG/P Nodes

Service Nodes:

- Primary service node:
  - IBM Power 550
  - POWER6, 8 cores @ 4.2 GHz
  - 64-bit, Linux OS
- Secondary service nodes:
  - IBM Power 520
  - POWER6, 2 cores @ 4.2 GHz
  - 64-bit, Linux OS

Dawn BG/P Login Nodes

Login Nodes: Fourteen IBM JS22 Blades

- POWER6, 8 cores @ 4.0 GHz
- 8 GB memory
- 64-bit, Linux OS

IBM Blue Gene Evolution: L to Q

Model	CPU	Core Count	Clock Speed (GHz)	Peak Node GFlops
L	PowerPC 440	2	.7	5.6
P	PowerPC 450	4	.85	13.6
Q	PowerPC A2	18	1.6	204.8

Blue Gene – L vs P: Node

Feature	Blue Gene/L	Blue Gene/P
Cores per node	2	4
Core clock speed	700 MHz	860 MHz
Cache Coherence Model	Software Managed	SMP
L1 cache (private)	32 KB/core	32 KB/core
L2 cache (private)	14 streams w prefetch	14 streams w prefetch
L3 cache (shared)	4 MB	8 MB
Memory/node	.512 – 1 GB	2/4 GB
Main memory bandwidth	5.6 GB/s	13.8 BG/s
Peak Flops	5.6 G/node	13.8 G/node

Blue Gene – L vs P:Interconnect

Feature	Blue Gene/L	Blue Gene/P
Torus		
Bandwidth	2.1 GB/s	5.1 GB/s
NN latency	200 ns	100 ns
Tree		
Bandwidth	700 MB/s	1.7 MB/s
Latency	6.0	

Blue Gene/P Application Execution Environment

SMP Mode - 1 MPI task per node

- Up to 4 Pthreads/OpenMP threads
- Full node memory available
- All resources dedicated to single kernel image
- Default mode

Blue Gene/P Application Execution Environment

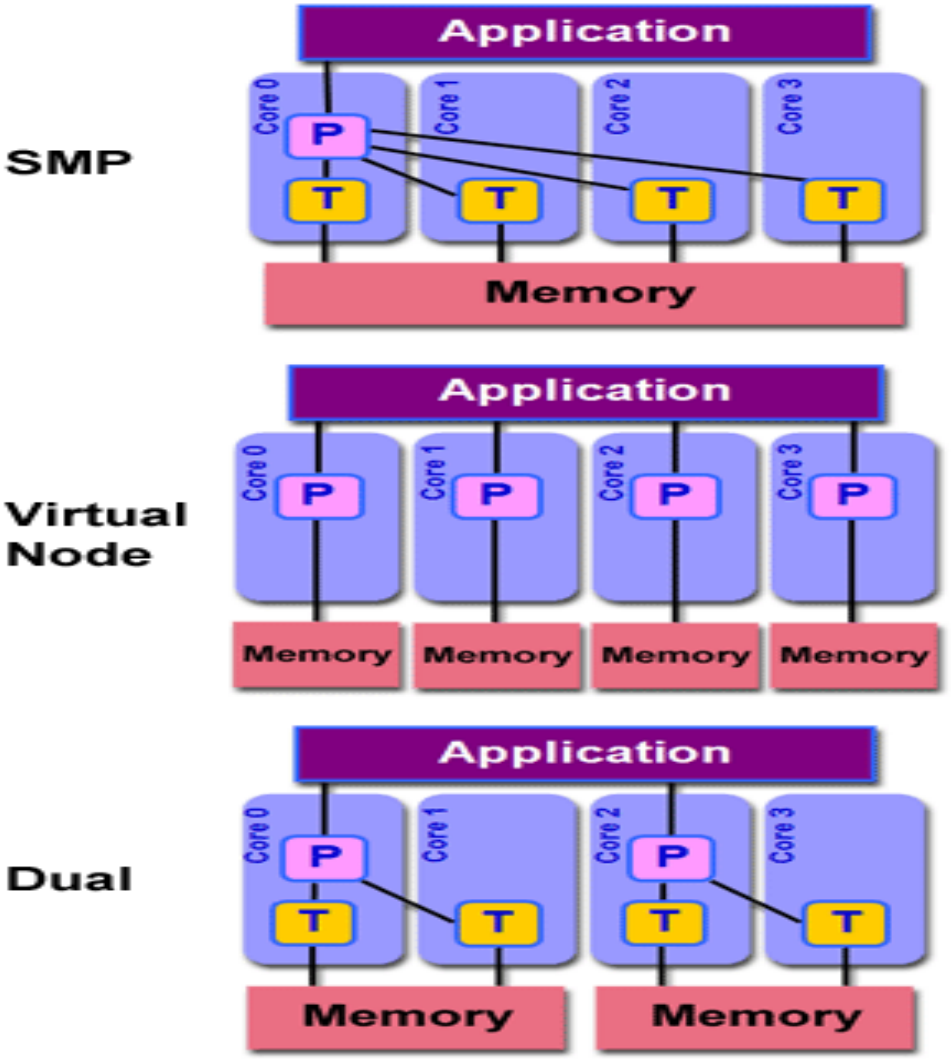
Virtual Node Mode - 4 MPI tasks per node

- No threads
- Each task gets its own copy of the kernel
- Each task gets 1/4th of the node memory
- Network resources split in fourths
- L3 Cache split in half and 2 cores share each half
- Memory bandwidth split in half

Blue Gene/P Application Execution Environment

- Dual Mode - Hybrid of the Virtual Node and SMP modes
- 2 MPI tasks per node
- Up to 2 threads per task
- Each task gets its own copy of the kernel
- 1/2 of the node memory per task

Blue Gene/P Application Execution Environment

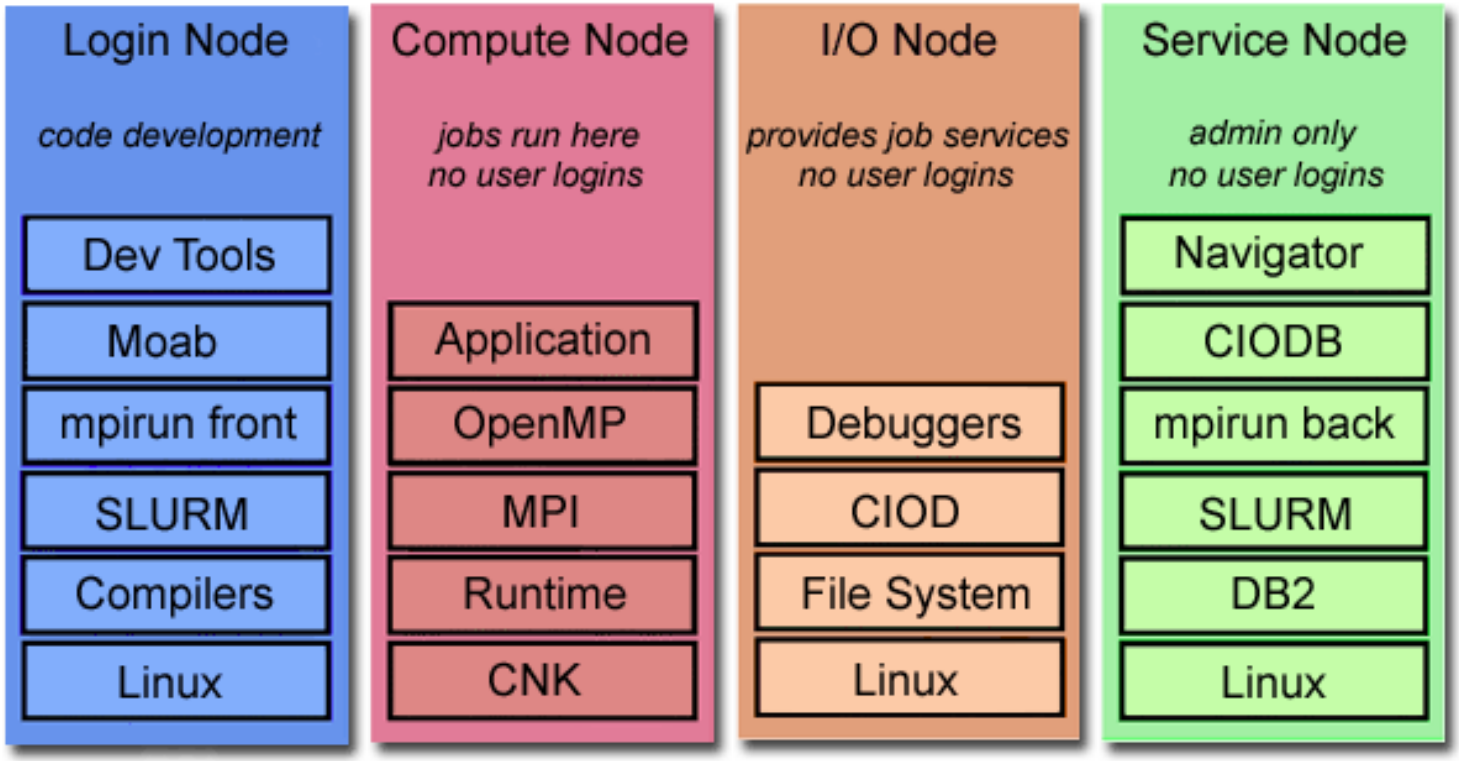


Blue Gene/P Application Execution Environment

Batch Systems:

- IBM LoadLeveler
- SLURM
  - LC's Simple Linux Utility for Resource Management
  - SLURM is the native job scheduling system on all of LC's clusters
- Moab
  - Tri-lab common workload scheduler
  - Top-level batch system for all LC clusters - manages work across multiple clusters, each of which is directly managed by SLURM

Blue Gene/P Software Development Environment



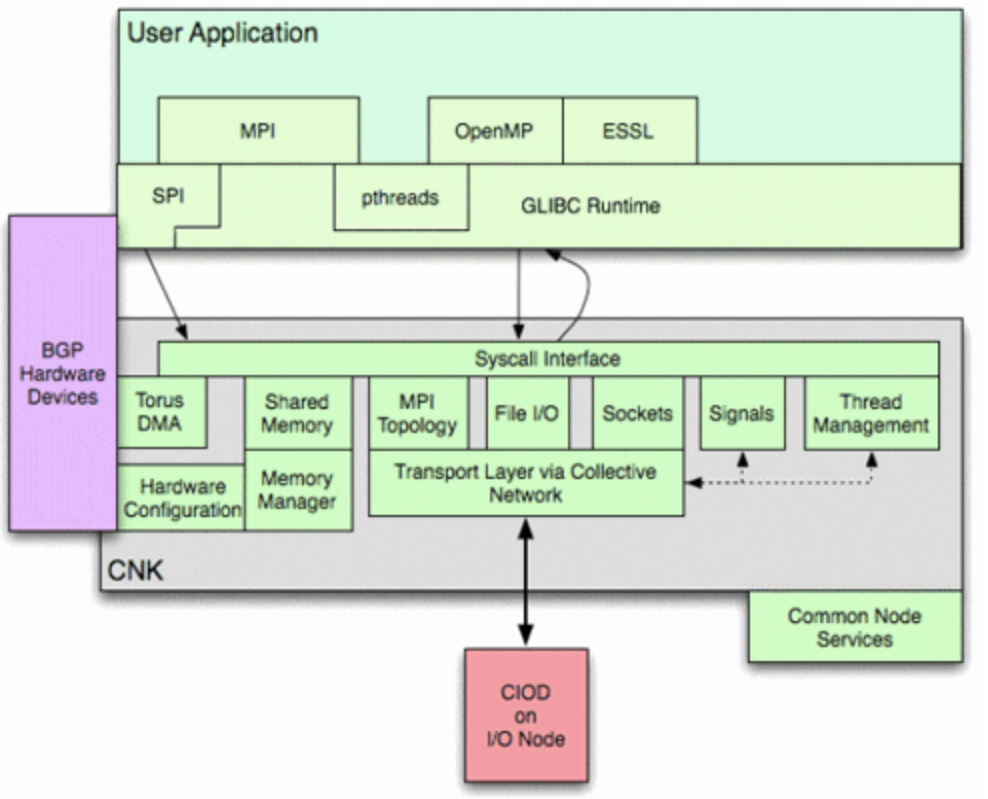
Blue Gene/P Software Development Environment – Compute Nodes

- CNK -open source,

light-weight, 32-bit Linux l kernel

  - Signal handling
  - Sockets
  - Starting/stopping jobs
  - Ships I/O calls to I/O nodes over Collective network

- Support for MPI, OpenMP, Pthreads
- Communicates over Collective network



Blue Gene/P Software Development Environment – Login Nodes

- Users must login to one of the front-end
- Full 64-bit Linux
- Cross-compilers specific to the BG/P architecture accessible
- Users submit/launch job for execution on the BG/P compute nodes
- Users can create software explicitly to run on login nodes

Blue Gene/P Software Development Environment – I/O Nodes

I/O Node Kernel:

- Full, embedded, 32-bit Linux kernel running on the I/O nodes
- Includes BusyBox "tiny" versions of common UNIX utilities
- Provides only connection to outside world for compute nodes through the CIOD
- Performs all I/O requests for compute nodes.
- Performs system calls not handled by the compute nodes
- Provides support for debuggers and tools through Tool Daemon
- Parallel file systems supported:
  - Network File System (NFS)
  - Parallel Virtual File System (PVFS)
  - IBM General Parallel File System (GPFS)
  - Lustre File System

Blue Gene/P Software Development Environment - Compilers

Compilers

- BG/P compilers are located on the front-end login nodes.
- Include:
  - IBM serial Fortran, C/C++
  - IBM wrapper scripts for parallel Fortran, C/C++
  - GNU serial Fortran, C, C++
  - GNU wrapper scripts for parallel Fortran, C, C++

Blue Gene/P Software Development Environment - Compilers

- BG/P is optimized for static executables:32-bit static linking is the default
- Executable shared among MPI tasks
- Efficient loading/memory use
- Shared libraries not shared among processes
- Entire shared library loaded into memory
- No demand paging of shared library
- Cannot unload shared library to free memory

Blue Gene/P Software Development Environment - Compilers

Support for “standard” HPC APIs

- MPI - The BG/P MPI library from IBM is based on MPICH2 1.0.7 base code
  - MPI1 and MPI2 - no Process Management functionality
- OpenMP 3.0 supported
- pThreads

Blue Gene/P Software Development Environment – Numerical Libraries

IBM's Engineering Scientific Subroutine Library. Highly optimized mathematical subroutines - ESSL

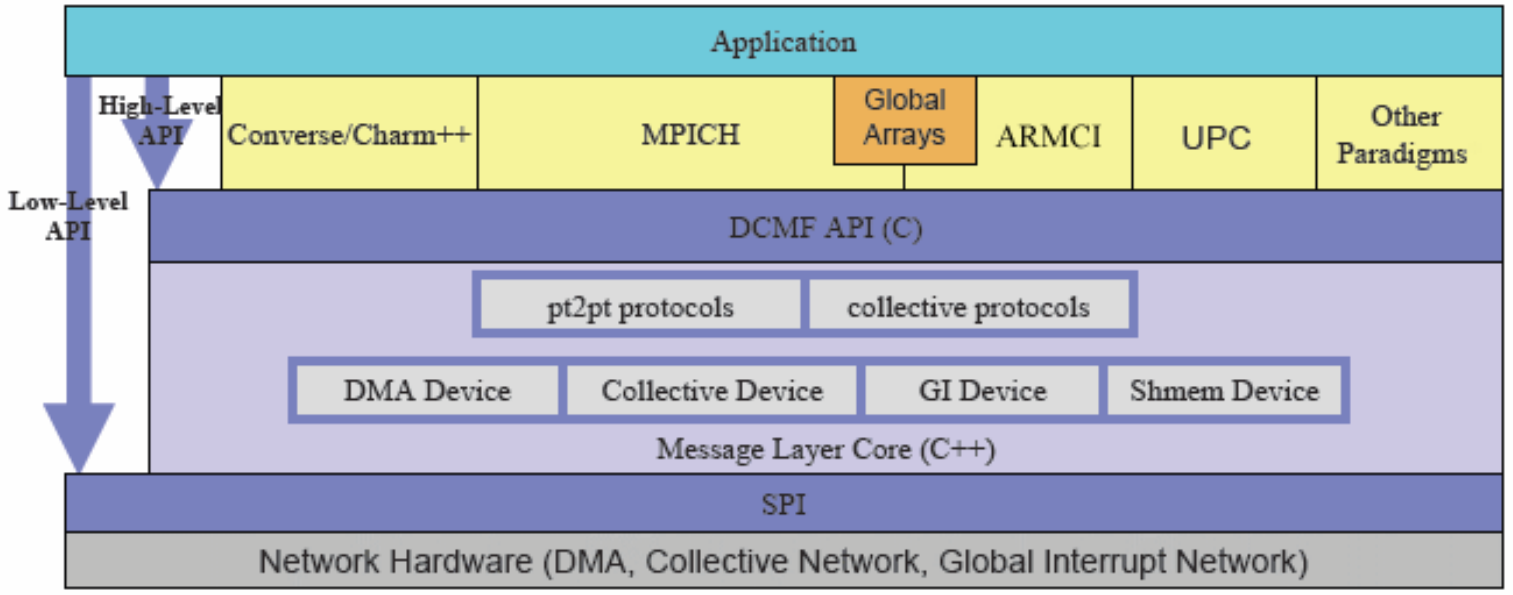
- Linear Algebra Subprograms
- Matrix Operations
- Linear Algebraic Equations
- Eigensystem Analysis
- Fourier Transforms
- Sorting and Searching
- Interpolation
- Numerical Quadrature
- Random Number Generation

Blue Gene/P Software Development Environment – Numerical Libraries

Mass, MASSV

- IBM's Mathematical Acceleration Subsystem.
- Highly tuned libraries for C, C++ and Fortran mathematical intrinsic functions
- Both scalar and vector routines available
- May provide significant performance improvement over standard intrinsic routines.

Blue Gene/P Software Development Message-Passing Environment -



Source: IBM

Blue Gene/P Software Development Environment - Debuggers

- TotalView
- DDT
- gdb

Blue Gene/P Software Development Environment – Performance Tools

	Tool	Function
IBM HPC Toolkit	mpitrace lib	Traces, profiles MPI calls
	peekperf, peekview	Tools for viewing mpitrace output
	Xprofiler	Displays program call tree
	HPM	Provides hardware counter summary
gprof		Unix debugger
papi		Hardware counter monitor
mpiP		MPI profiling lib
TAU		Performance analysis tools
openSpeed Shop		Performance analysis tools

Blue Gene/P References

- IBM BG Redbooks:
  - "IBM System Blue Gene Solution: Blue Gene/P Application Development"
  - "IBM System Blue Gene Solution:Performance Analysis Tools"
- Blue Gene HPC Center Web Sites:
  - LLNL Livermore Computing https://computing.llnl.gov
  - Argonne ALCF http://www.alcf.anl.gov/
  - Julich JSC - http://www2.fz-juelich.de/jsc/jugene