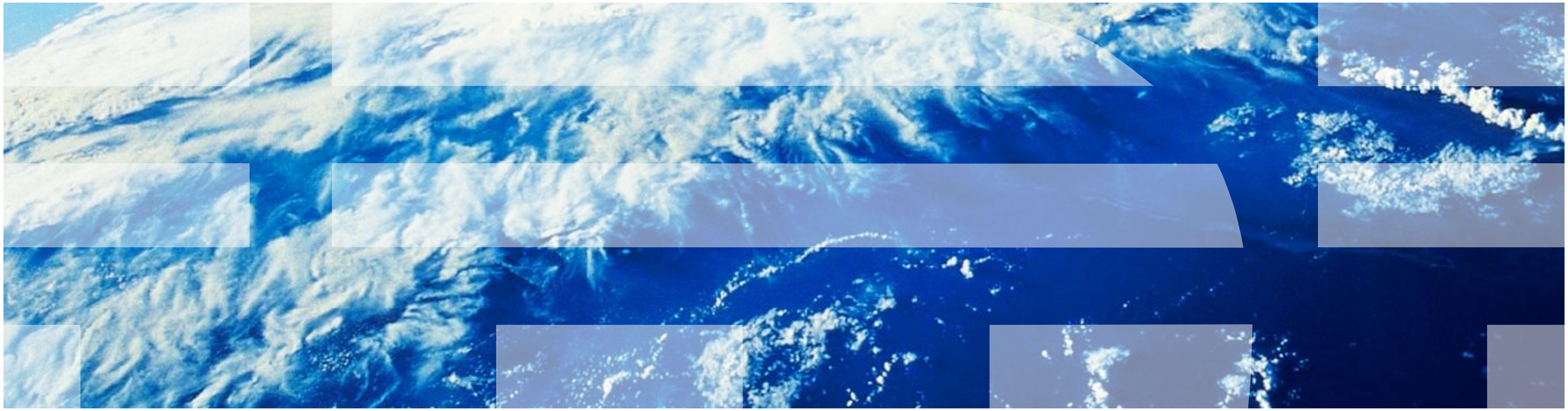
IBM Blue Gene/Q
Architecture and System Software Overview

Pascal Vezolle
vezolle@fr.ibm.com



Understanding Blue Gene

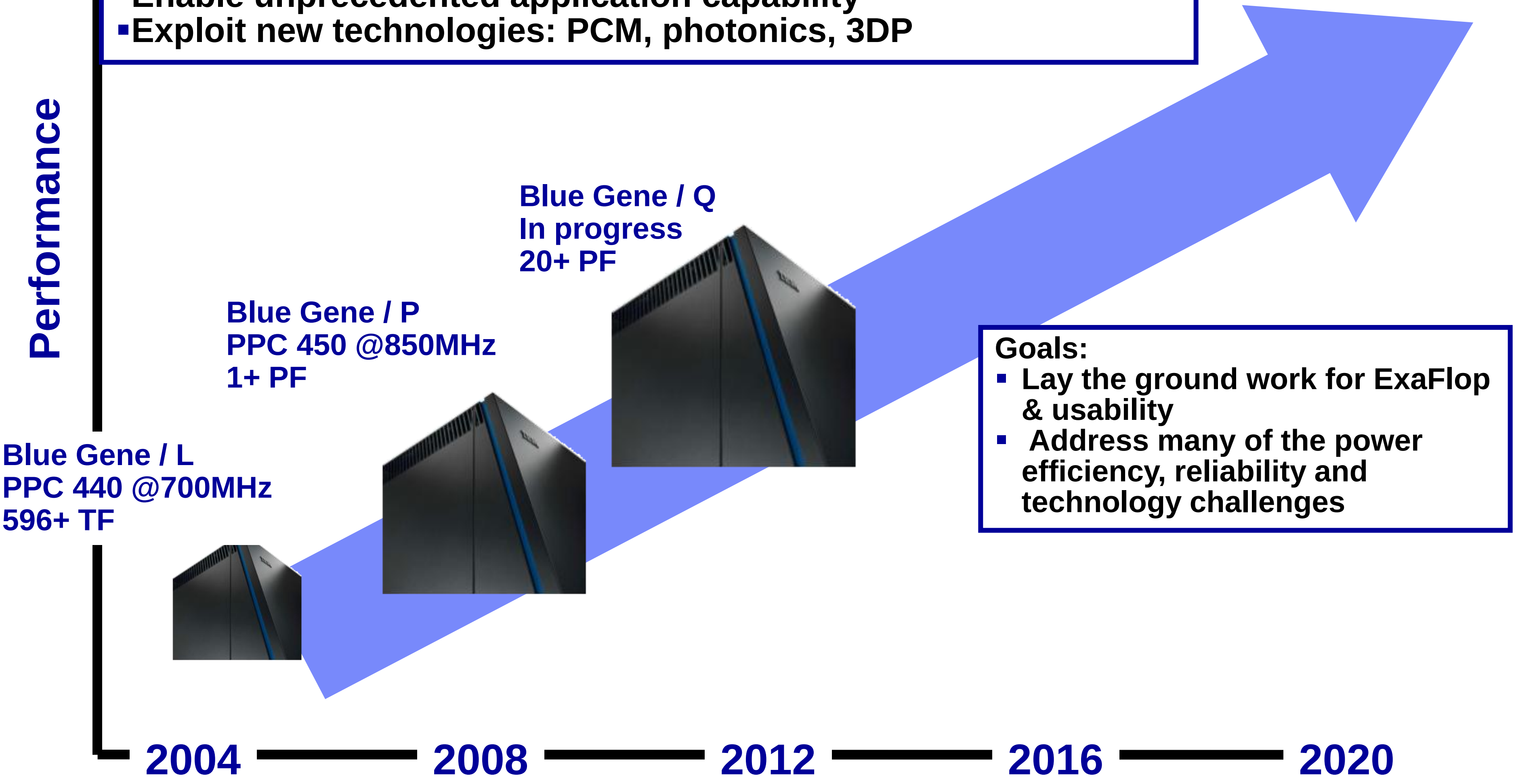
- – Hardware overview
  - • IO Node
  - • Processor
    - - BG/Q innovatives features: atomic, wake up unit, TM, TLS
  - • Network – 5D Torus
    - - partitioning

- – Software overview
  - • Programming model
  - • BG/Q kernel overview
  - • User environment

Blue Gene

Goals:

- **Three orders of magnitude performance in 10 years**
- **Push state of the art in Power efficiency, scalability, & reliability**
- **Enable unprecedented application capability**
- **Exploit new technologies: PCM, photonics, 3DP**



Blue Gene / L
PPC 440 @700MHz
596+ TF

Blue Gene / P
PPC 450 @850MHz
1+ PF

Blue Gene / Q
In progress
20+ PF

Goals:

- **Lay the ground work for ExaFlop & usability**
- **Address many of the power efficiency, reliability and technology challenges**

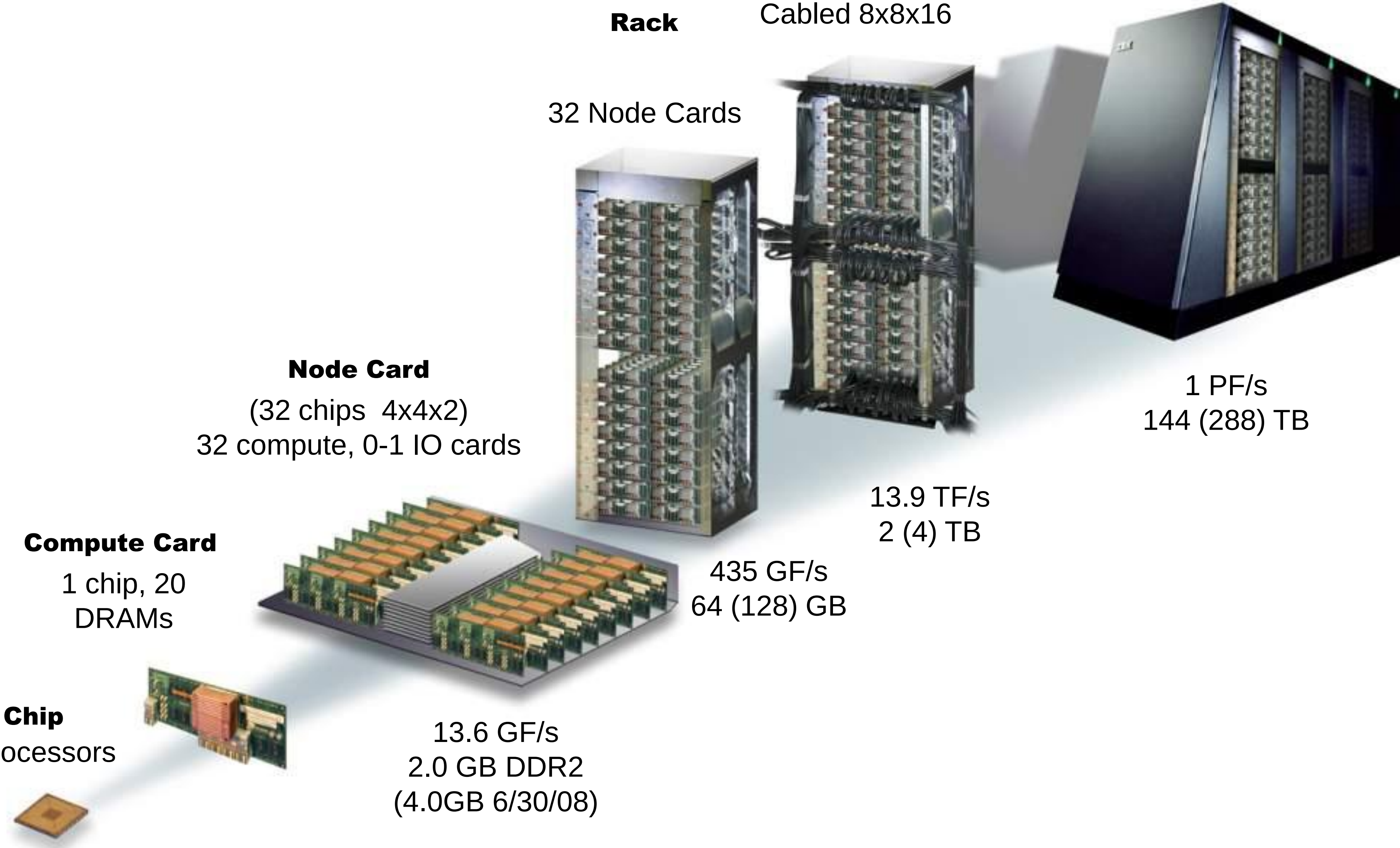
Blue Gene Evolution

- BG/L (**5.7 TF/rack**) – 130nm ASIC (1999-2004 GA)
  - Embedded 440 core, dual-core system-on-chip
  - Memory: 0.5/1 GB/node
  - Biggest installed system (LLNL): 104 racks, **212,992 cores-threads**, 596 TF/s, 210 MF/W

- BG/P (**13.9 TF/rack**) – 90nm ASIC (2004-2007 GA)
  - Embedded 450 core
  - Memoy: 2/4 GB/node, quad core SOC, DMA
  - Biggest installed system (Jülich): 72 racks, **294,912 cores-threads**, 1 PF/s, 357 MF/W
  - SMP support, OpenMP, MPI

- BG/Q (**209 TF/rack**) – 45nm ASIC+ (2007-2012 GA)
  - A2 core, 16 core/64 thread SOC
  - 16 GB/node
  - Biggest installed system (LLNL): 96 racks, **1,572,864 cores & >6M threads**, 20 PF/s, 2 GF/W,
  - Speculative execution, sophisticated L1 prefetch, transactional memory, fast thread handoff, compute + IO systems.

Blue Gene/P



Chip
4 processors

13.6 GF/s
8 MB EDRAM

Compute Card

1 chip, 20
DRAMs

13.6 GF/s
2.0 GB DDR2
(4.0GB 6/30/08)

Node Card

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

435 GF/s
64 (128) GB

Rack

32 Node Cards

Cabled 8x8x16

13.9 TF/s
2 (4) TB

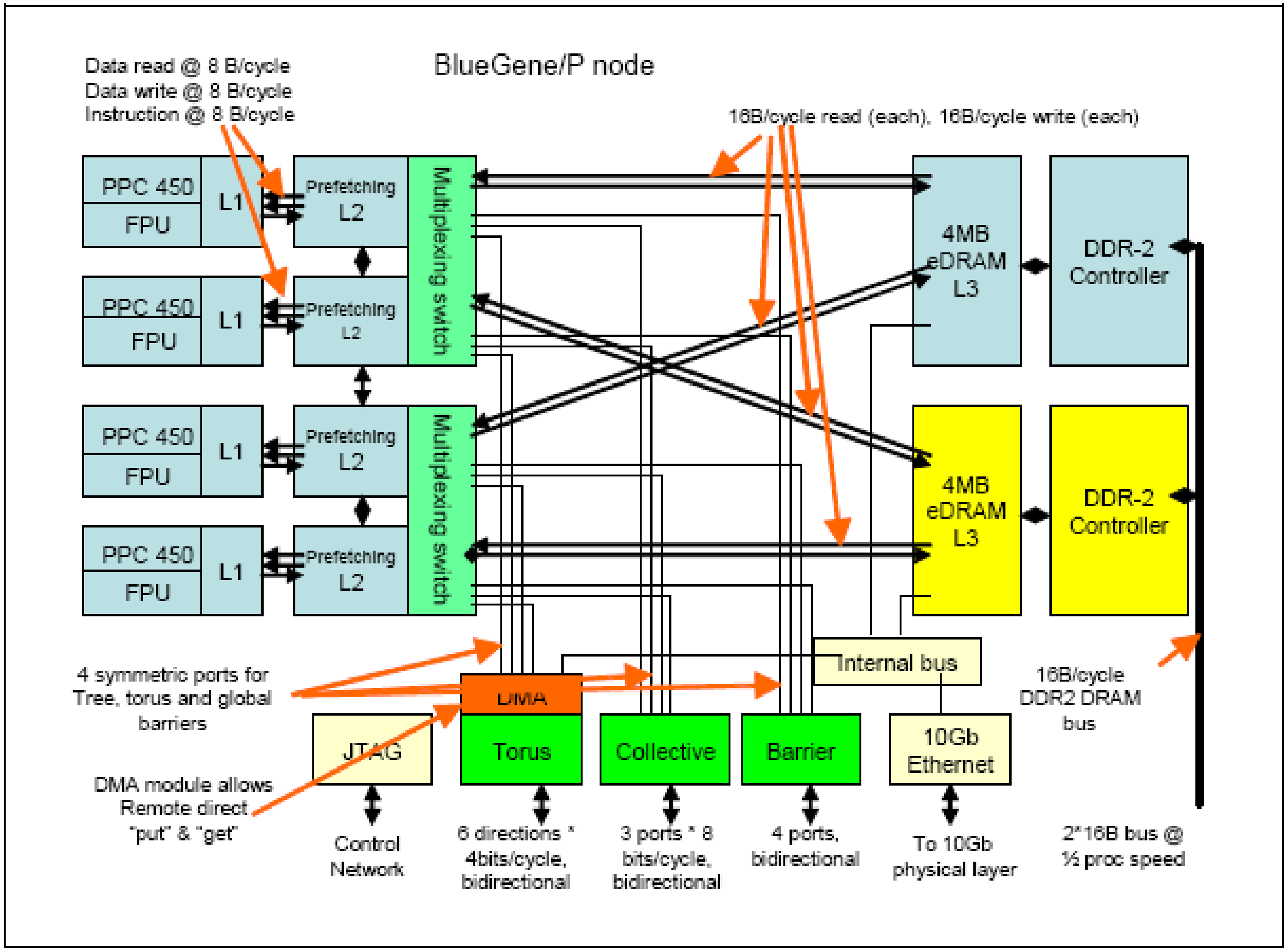
System

72 Racks, 72x32x32

1 PF/s
144 (288) TB

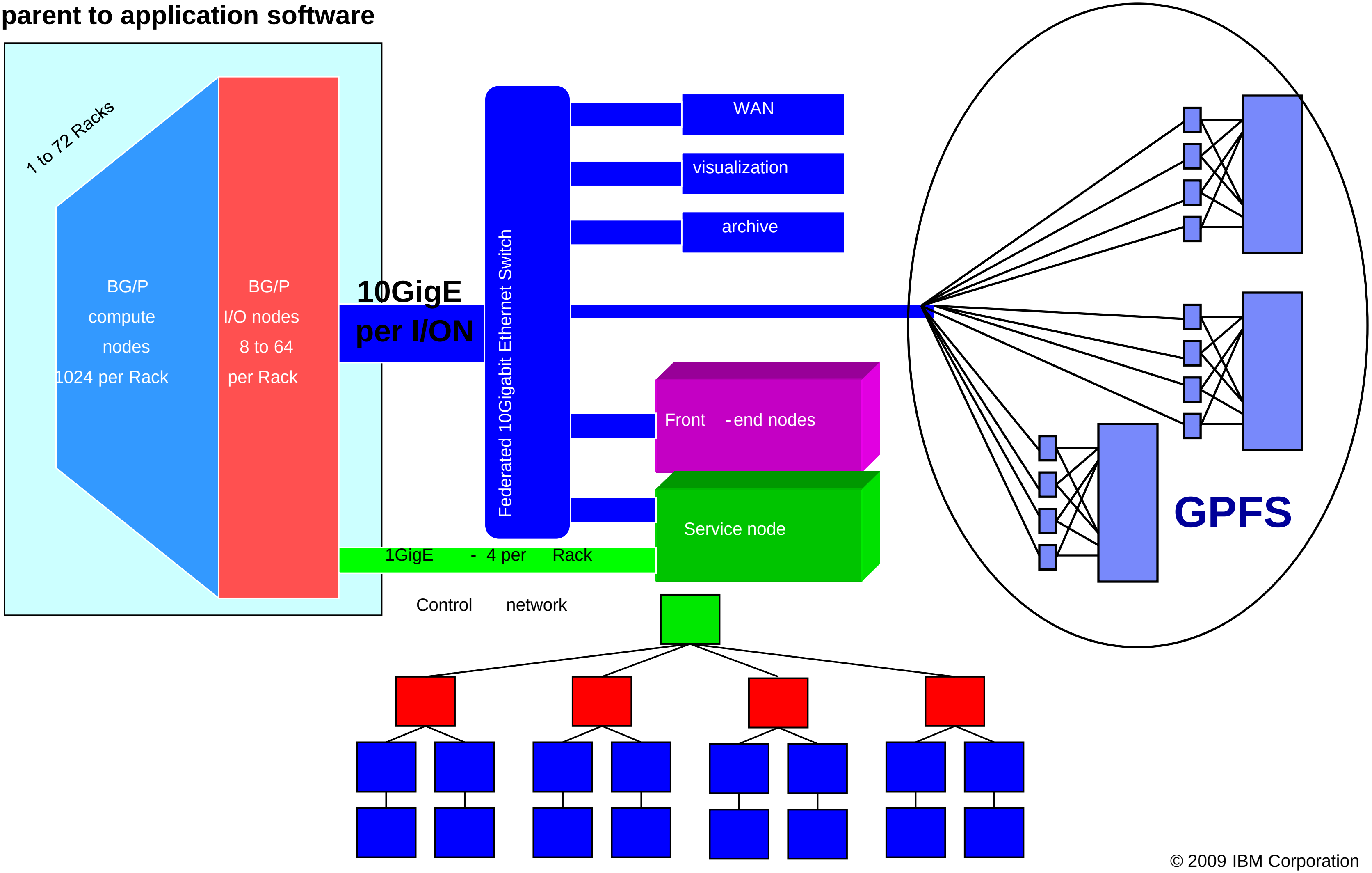
Blue Gene/P Asic

4 cores- SMP/8MB L3 shared – 0.85Ghz – 1 thread/core – 4 GBytes memory



Blue Gene/P System in a Complete Configuration

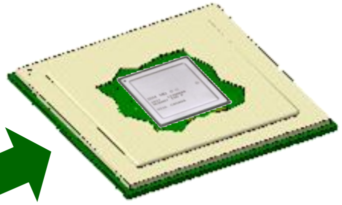
- **Compute nodes dedicated to running user application, and almost nothing else - simple compute node kernel (CNK)**
- **I/O nodes run Linux and provide a more complete range of OS services – files, sockets, process launch, signaling, debugging, and termination**
- **Service node performs system management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



Blue Gene/Q

1. Chip:
16 λP cores

2. Single Chip Module



3. Compute card:
One chip module,
16 GB DDR3 Memory,

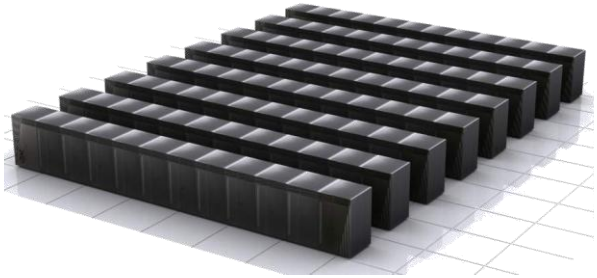
4. Node Card:
32 Compute Cards,
Optical Modules, Link Chips, Torus

5a. Midplane:
16 Node Cards

5b. IO drawer:
8 IO cards w/16 GB
8 PCIe Gen2 x8 slots

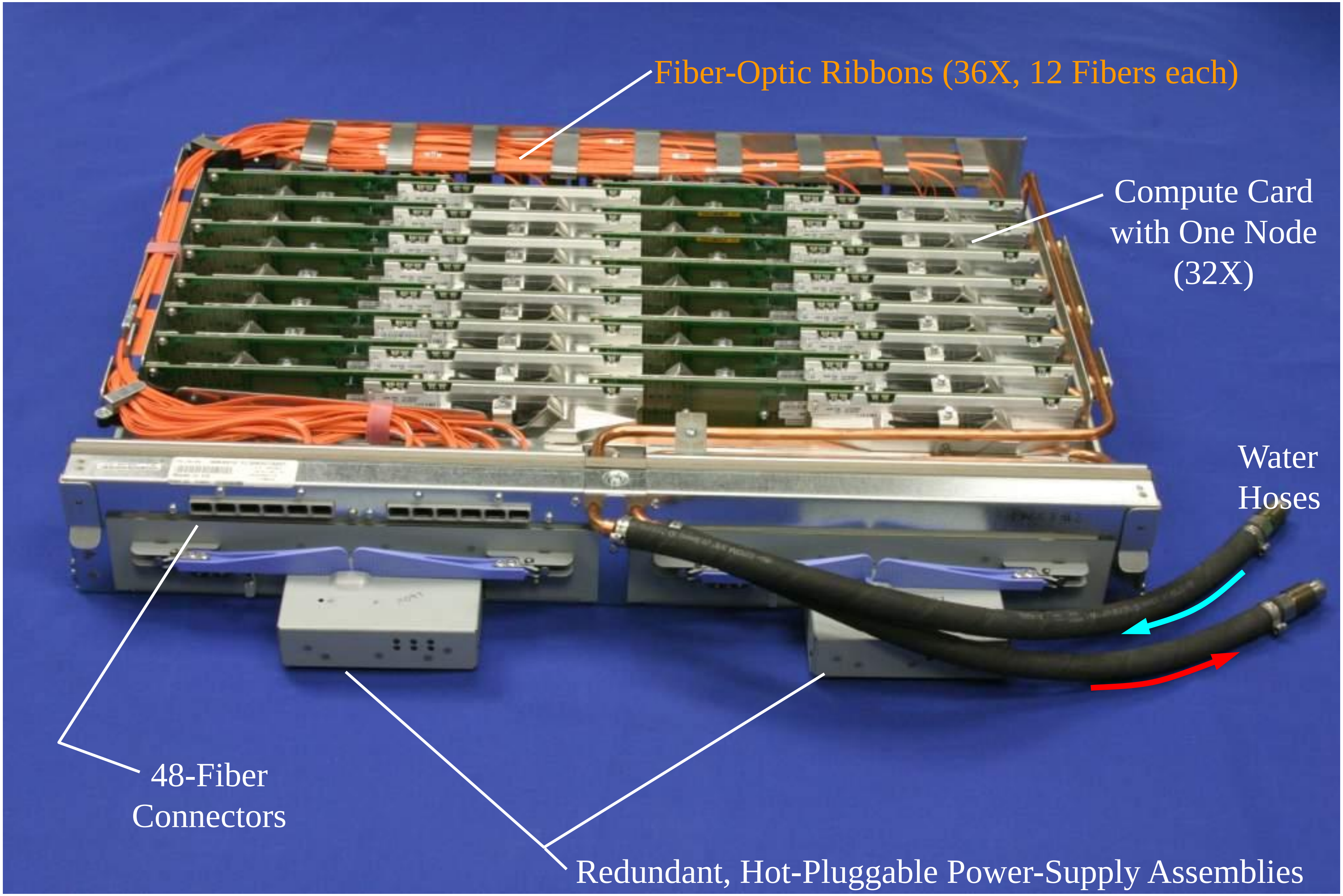
6. Rack: 2 Midplanes

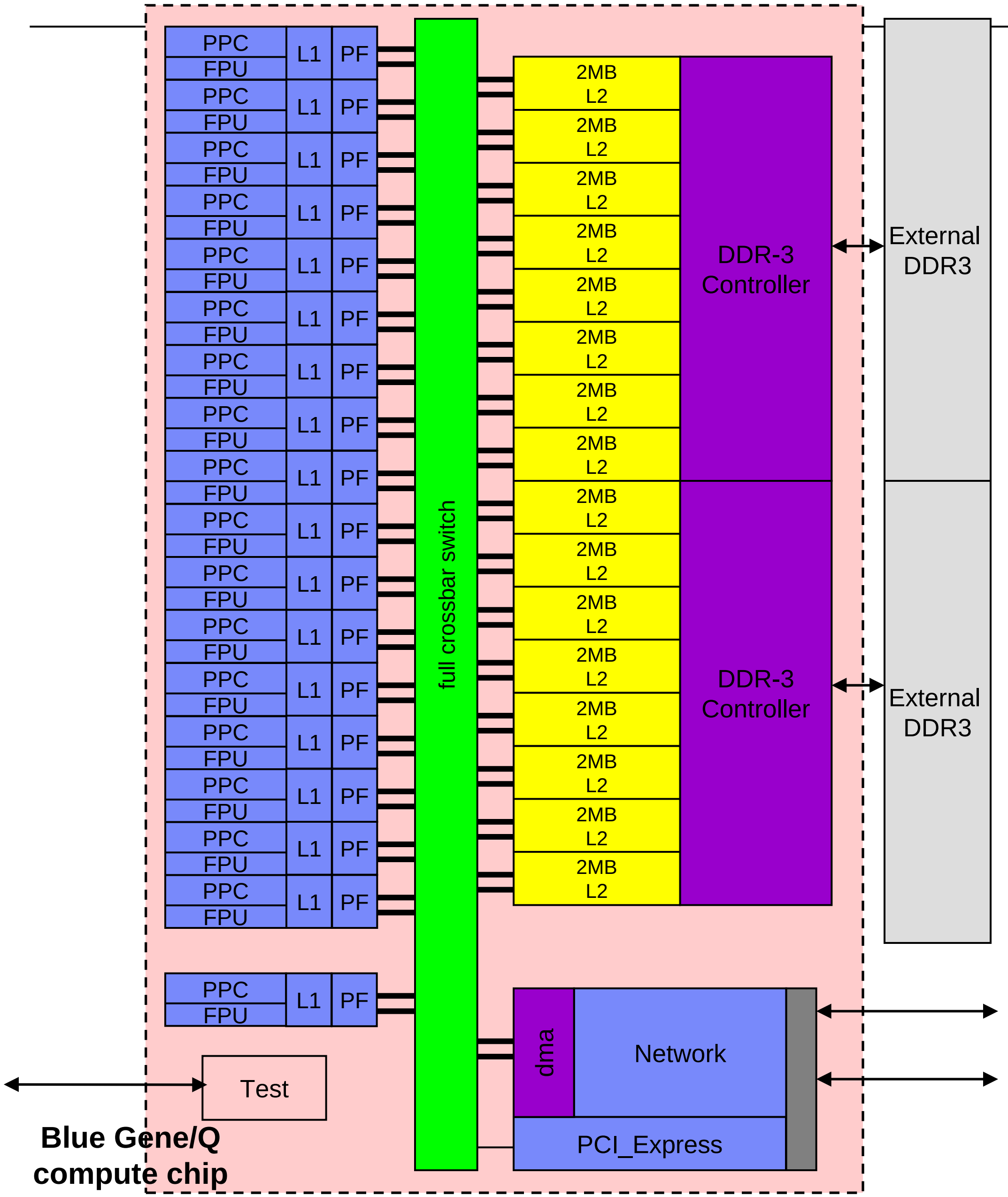
7. System:
20PF/s



- **Sustained single node perf: 10x P, 20x L**
- **MF/Watt: (6x) P, (10x) L (~2GF/W target)**

Blue Gene/Q node board: 32 compute nodes


Fiber-Optic Ribbons (36X, 12 Fibers each)
Compute Card with One Node (32X)
Water Hoses
48-Fiber Connectors
Redundant, Hot-Pluggable Power-Supply Assemblies



Blue Gene/Q chip architecture

- 16+1 core SMP

  Each core 4-way hardware threaded
- Transactional memory and thread level speculation
- Quad floating point unit on each core

  204.8 GF peak node
- Frequency target of 1.6 GHz
- 563 GB/s bisection bandwidth to shared L2

  (Blue Gene/L at LLNL has 700 GB/s for system)
- 32 MB shared L2 cache
- 42.6 GB/s DDR3 bandwidth (1.333 GHz DDR3)

  (2 channels each with chip kill protection)
- 10 intra-rack interprocessor links each at 2.0GB/s
- one I/O link at 2.0 GB/s
- 8-16 GB memory/node
- ~60 watts max DD1 chip power

2 GB/s I/O link (to I/O subsystem)

10*2GB/s intra-rack & inter-rack (5-D torus)

note: chip I/O shares function with PCI_Express

Scalability Enhancements: the 17th Core

- RAS Event handling and interrupt off-load
  - – Reduce O/S noise and jitter
  - – Core-to-Core interrupts when necessary

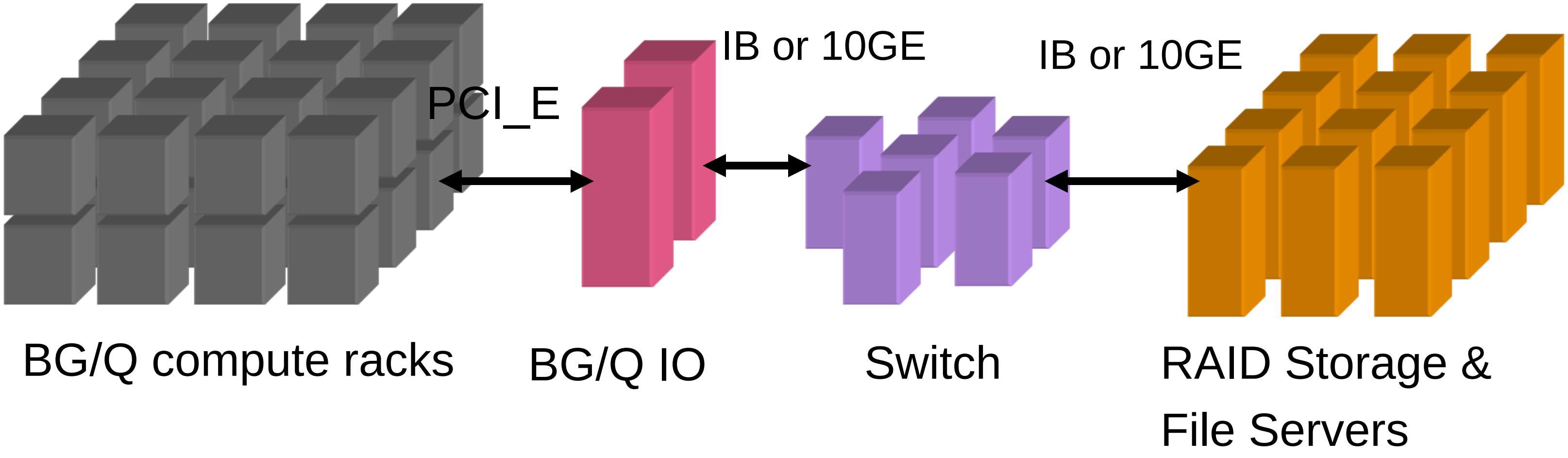
- CIO Client Interface
  - – Asynchronous I/O completion hand-off
  - – Responsive CIO application control client

- Application Agents: privileged application processing
  - – Messaging assist, e.g., MPI pacing thread
  - – Performance and trace helpers

BGQ major architecture&design changes from BGL/BGP

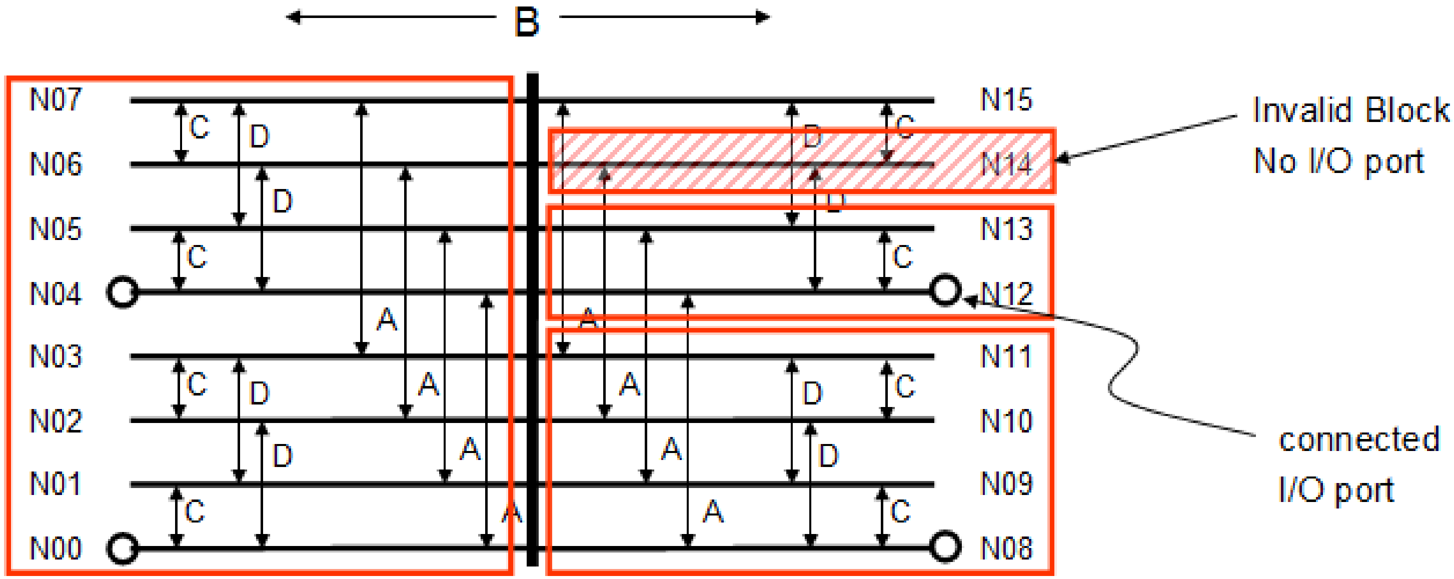
- **New Node:** Multithreading architecture
  - New voltage scaled processing core (A2) with 4-way SMT
  - New SIMD floating point unit (8 flop/clock) with alignment support: QPX
  - New “intelligent” prefetcher
  - 17th Processor core for system functions.
  - Speculative multithreading and transactional memory support with 32 MB of speculative state
  - Hardware mechanisms to help with multithreading (“fetch & op”, “wake on pin”)
  - Dual SDRAM-DDR3 memory controllers with up to 16 GB/node
- **New Network architecture**:
  - 5 D torus architecture sharing several embedded Virtual Network/topologies
    - 5D topology for point-to-point communication
    - Collective and barrier networks embedded in 5-D torus network.
  - Floating point addition support in collective network
  - 11th port for auto-routing to IO fabric
- **External, independent and dynamic I/O system**
  - I/O nodes in separate drawers/rack with private interconnections and full Linux support
  - PCI-Express Gen 2 on every node with full sized PCI slot
  - Two I/O configurations (one traditional, one conceptual)

BG/Q I/O architecture



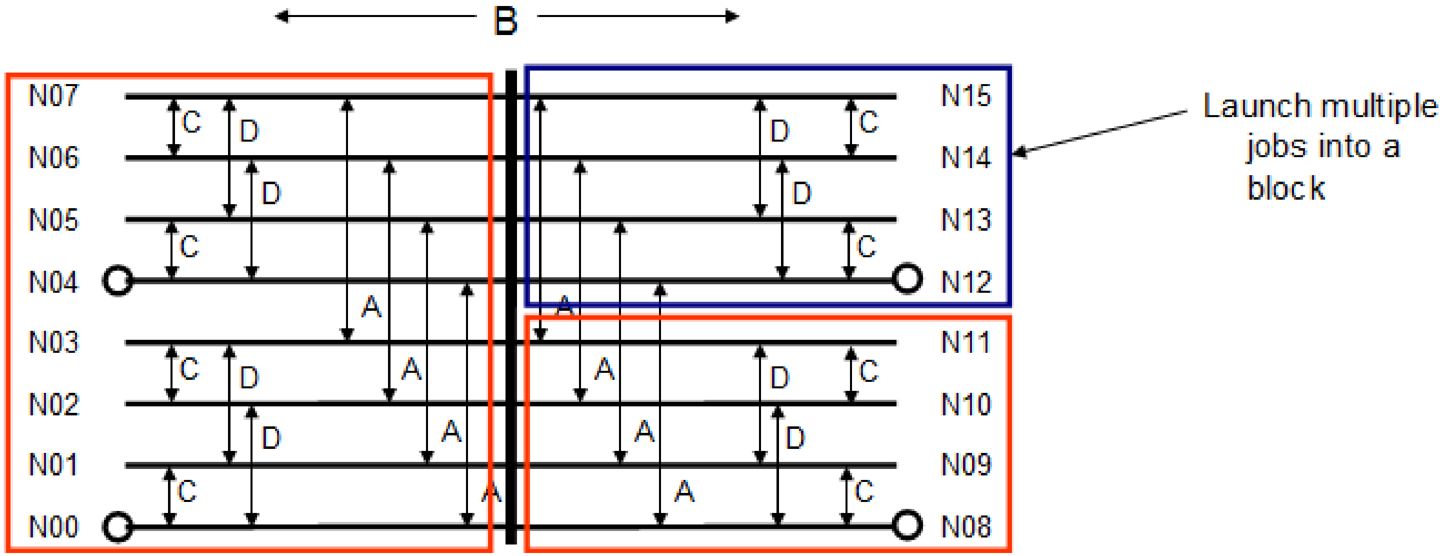
- BlueGene Classic I/O with GPFS clients on the logical I/O nodes
- Similar to BG/L and BG/P
- Uses InfiniBand switch
- Uses DDN RAID controllers and File Servers
- BG/Q I/O Nodes are not shared between compute partitions
  - IO Nodes are bridge data from function-shipped I/O calls to parallel file system client
- Components balanced to allow a specified minimum compute partition size to saturate entire storage array I/O bandwidth

I/O Requirements



- All blocks require I/O
- Only some nodeboards have I/O connections
- This restricts partitioning
- Systems can be configured with I/O up to several ports per board

Sub-block jobs

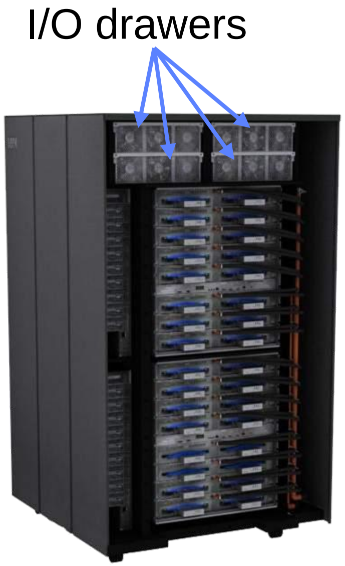


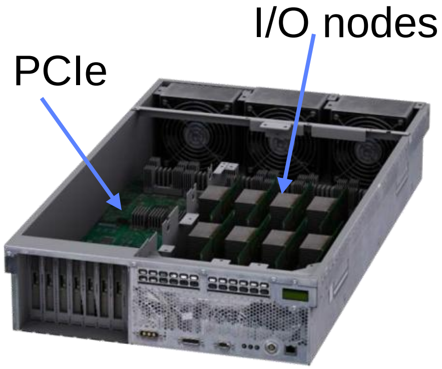
- Sub-block jobs are new in BG/Q
- A user may launch multiple jobs into the block
- A block may authorize other users so multiple users may share this block
- Allows smaller jobs without additional I/O ports

BG/Q External I/O

- **I/O Network to/from Compute rack**
  - 2 links (4GB/s in 4GB/s out) feed an I/O PCI-e port
  - Every node card has up to 4 ports (8 links)
  - Typical configurations
    - 8 ports (32GB/s/rack)
    - 16 ports (64 GB/s/rack)
    - 32 ports (128 GB/s/rack)
  - Extreme configuration 128 ports (512 GB/s/rack)

- **I/O Drawers**
  - 8 I/O nodes/drawer with 8 ports (16 links) to compute rack
  - 8 PCI-e gen2 x8 slots (32 GB/s aggregate)
  - 4 I/O drawers per compute rack
  - Optional installation of I/O drawers in external racks for extreme bandwidth configurations





I/O Blocks

- I/O Nodes are also combined into blocks
  - All I/O drawers can be grouped into a single block for administrative convenience
  - In normal operation the I/O Block(s) remain booted while compute blocks are reconfigured and rebooted
  - I/O blocks do not need to be rebooted to resolve fatal errors from I/O nodes
  - Rationale for having multiple I/O node partitions would be experimentation with different Linux ION kernels

- Can be created via genIOblock

- Locations of IO enclosures can be:
  - Qxx-Iy (in an IO rack, y is 0 - B)
  - Rxx-Iy (in a compute rack, y is C - F)

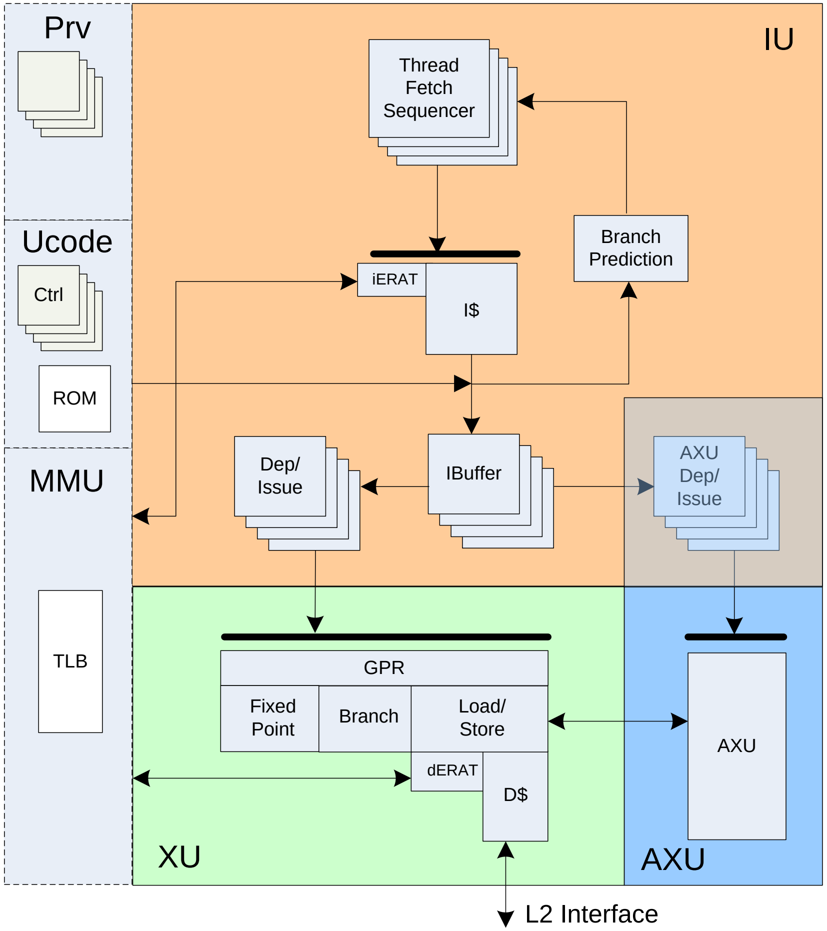
IO Node: BG/P vs. BG/Q

BG/P	BG/Q
Minimal MCP based Linux Distro	**Fully Featured RHEL6.X based Linux Distro**
Only supported IONs	Supports IONs and Log-In Nodes (LNs)
Installed via a static tar file	RPM based installation and is customizable before and after installation
Ramdisk based root filesystem	Hybrid read only NFS root with in memory (tmpfs) read/write file spaces
No persistent storage space	Per-node persistent files spaces
Rebooted frequently	**Designed to be booted infrequently**
Part of the compute block	**Independent I/O or LN block associated with mone or more compute blocks**
Only supported ethernet	**Supports PCIe based 10Gb Ethernet, Infiniband and Combo Eth/IB cards**
Image was a few hundred megabytes in size	**Each Image is 5 GB in size**
No health monitoring	**Integrated health monitoring system**

Blue Gene/Q processor

BQC processor core (A2 core)

- Simple core, designed for excellent power efficiency and small footprint.
- Embedded **64 bit PowerPC compliant**
- **4 SMT threads typically get a high level of utilization on shared resources.**
- Design point is 1.6 GHz @ 0.74V.
- AXU port allows for unique BGQ style floating point
- **One AXU (FPU) and one other instruction issue per cycle**
- **In-order execution**



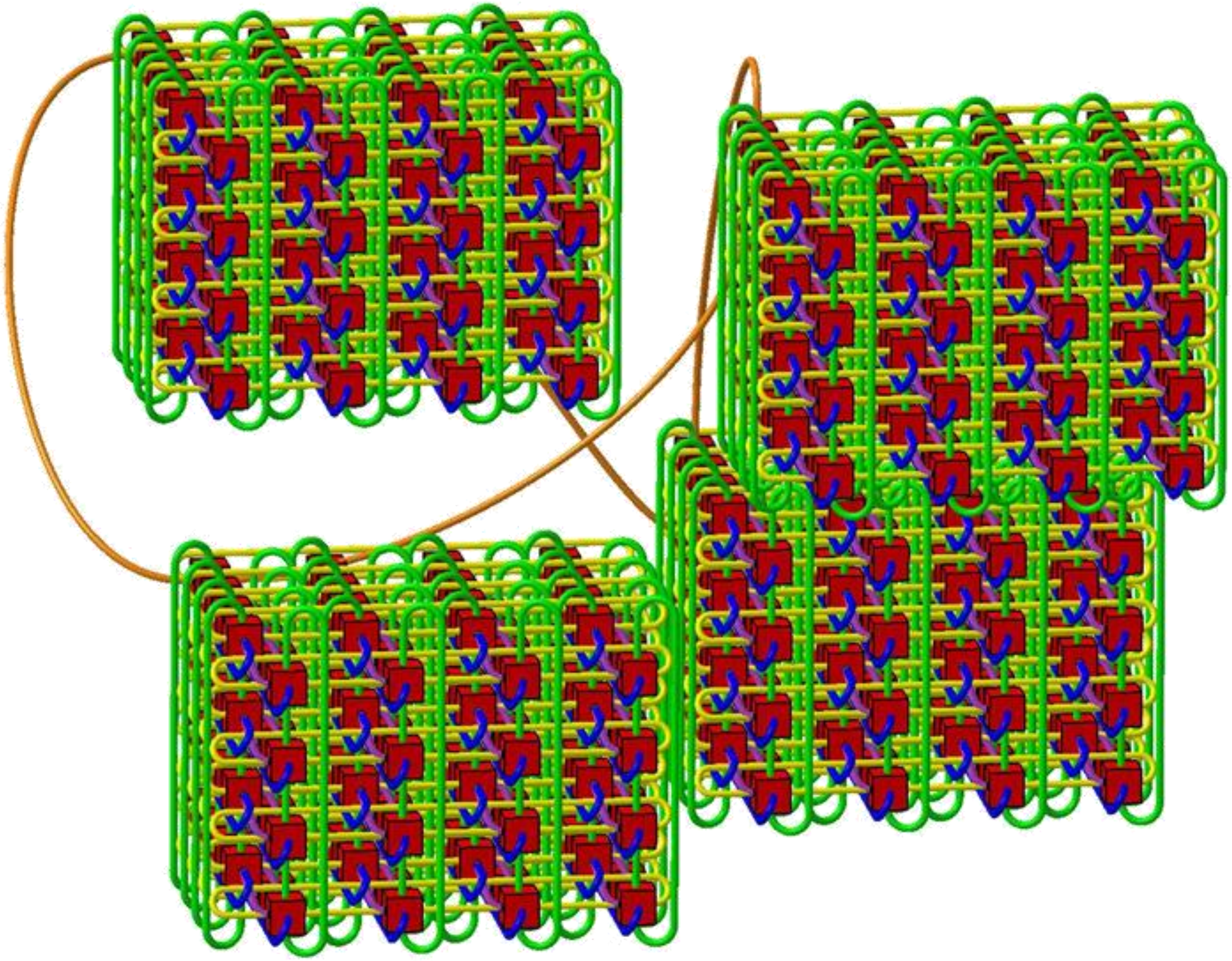
Multithreading

- Four threads issuing to two pipelines
  - Impact of memory access latency reduced

- Issue
  - Up to two instructions issued per cycle
    - One Integer/Load/Store/Control instruction issue per cycle
    - One FPU instruction issue per cycle
  - **At most one instruction issued per thread**

- Flush
  - Pipeline is not stalled on conflict
  - Instead,
    - Instructions of conflicting thread are invalidated
    - Thread is restarted at conflicting instruction
  - Guarantees progress of other threads

Blue Gene/Q Network: 5D Torus



BG/Q Networks

- Networks
  - – 5 D torus in compute nodes,
  - – 2 GB/s bidirectional bandwidth on all (10+1) links, **5D nearest neighbor exchange measured at ~1.75 GB/s per link**
  - – Both collective and barrier networks are embedded in this 5-D torus network.
  - – Virtual Cut Through (VCT)
  - – Floating point addition support in collective network

- Compute rack to compute rack bisection BW (46X BG/L, 19X BG/P)
  - – 20.1PF: bisection is 2x16x16x12x2 (bidi)x2(torus, not mesh)x 2GB/s link bandwidth = 49.152 TB/s
  - – 26.8PF: bisection is 2x16x16x16x4x2GB/s = 65.536TB/s
  - – BGL at LLNL is 0.7 TB/s

- I/O Network to/from Compute rack
  - – 2 links (4GB/s in 4GB/s out) feed an I/O PCI-e port (4GB/s in, 4GB/s out)
  - – Every Q32 node card has up to I/O 8 links or 4 ports
  - – Every rack has up to 32x8 = 256 links or 128 ports

- I/O rack
  - – 8 I/O nodes/drawer, each node has 2 links from compute rack, and 1 PCI-e port to the outside world
  - – 12/drawers/rack
  - – 96 I/O, or 96x4 (PCI-e) = 384 TB/s = 3 Tb/s.

3D versus 5D Torus network

- BG/P
  - – the partitions are denoted by letter XYZT, for XYZ the 3 dimensions and T for core (0-3)

- BG/Q
  - – the 5 dimensions are denoted by the letters A, B, C, D, and E, T for the core (0-15).
  - – the latest dimension E is always 2, and is contained entirely within a midplane.
  - – for any compute block, compute nodes (as well midplanes for large blocks) are combined in 4 dimensions - only 4 dimensions need to be considered.

Network Performance

- Performance
  - All-to-all: 97% of peak
  - Bisection: > 93% of peak
  - Nearest-neighbor: 98% of peak
  - Collective: FP reductions at 94.6% of peak
  - No performance problems identified in network logic

BG/Q Software
& Programming model

BG/Q Software

Property		BG/Q
Overall Philosophy	Scalability	**Scale infinitely, added more functionality**
	Openness	**almost all open**
Programming Model	Shared Memory	**yes**
	Hybrid	**1-64 processes** **64-1 threads**
	Low-Level General Messaging	**PAMP, generic parallel program runtimes, wake-up unit**
	Programming Models	**MPI, OpenMP, UPC, ARMCI, global arrays, Charm++**
Kernel	System call interface	**Linux/POSIX system calls**
	Library/threading	**glibc/pthreads**
	Linking	**static or dynamic**
	Compute Node OS	**CNK, Linux, Red Hat**
	I/O Node OS	**SMP Linux with SMT, Red Hat**
Control	Scheduling	**generic and real-time API**
	Run Mode	**HPC, generalized sub-partitioning, HA with job cont**
Generalizing and Research Initiatives	OS	**Linux, ZeptOS, Plan 9**
	Tools	**HPC/S Toolkit, dyninst, valgrind**
	Financial	**Kittyhawk, InfoSphere Streams**
	Commercial	**Kittyhawk, Cloud, SLAcc, ASF**

BG/Q innovations will help programmers cope with an exploding number of hardware threads (64 per node)

- Exploiting a large number of threads is a challenge for all future architectures. This is a key component of the BGQ research.
- Novel hardware and software is utilized in BGQ to ..

a) Reduce the overhead to hand off work to high numbers of threads used in OpenMP and messaging through hardware support for atomic operations and fast wake up of cores.

b) Multiversioning cache to help in a number of dimensions such as performance, ease of use and RAS.

c) Aggressive FPU to allow for higher single thread performance for some applications. Most will get modest bump (10-25%), some big bump (approaching 300%)

d) "perfect" prefetching for repeated memory reference patterns in arbitrarily long code segments. Also helps achieve higher single thread for some applications.

Programmability

- Standards-based programming environment
  - Linux^{TM} development environment
  - Familiar GNU toolchain with GLIBC, pthreads, gdb
  - XL Compilers providing C, C++, Fortran with OpenMP
  - Totalview debugger
- Message Passing
  - Optimized MPICH2 providing MPI 2.2
  - Intermediate and low-level message libraries available, documented, and open source
  - GA/ARMCI, Berkeley UPC, etc, ported to this optimized layer
- Compute Node Kernel (CNK) eliminates OS noise
  - File I/O offloaded to I/O nodes running full Linux
  - GLIBC environment with few restrictions for scaling
- Flexible and fast Job Control
  - MPMD (4Q 2012) and sub-block jobs supported

Toolchain and Tools

- BGQ GNU toolchain
  - gcc is currently at 4.4.4. Will update again before we ship.
  - glibc is 2.12.2 (optimized QPX memset/memcopy)
  - binutils is at 2.21.1
  - gdb is 7.1 with QPX registers
  - gmon/gprof thread support
    - Can turn profiling on/off on a per thread basis

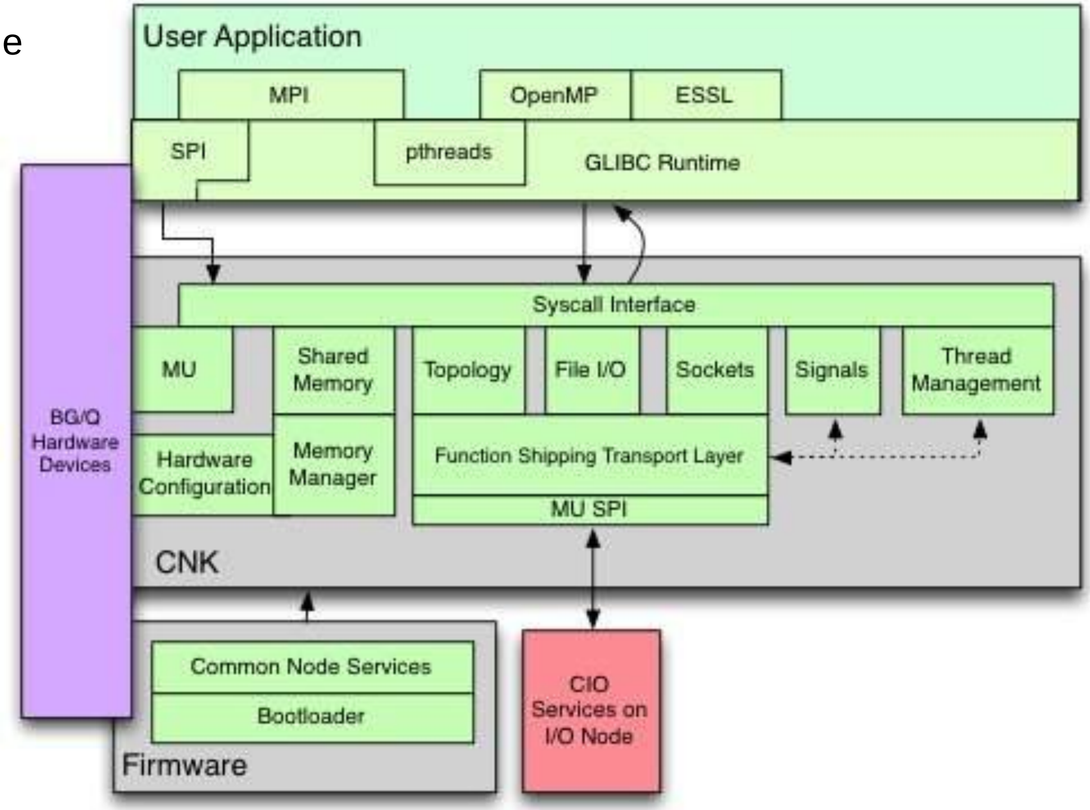
- Python
  - Running both Python 2.6 and 3.1.1.
  - NUMPY, pynamic, UMT all working
  - Python is now an RPM

- Toronto compiler test harness is running on BGQ LNs

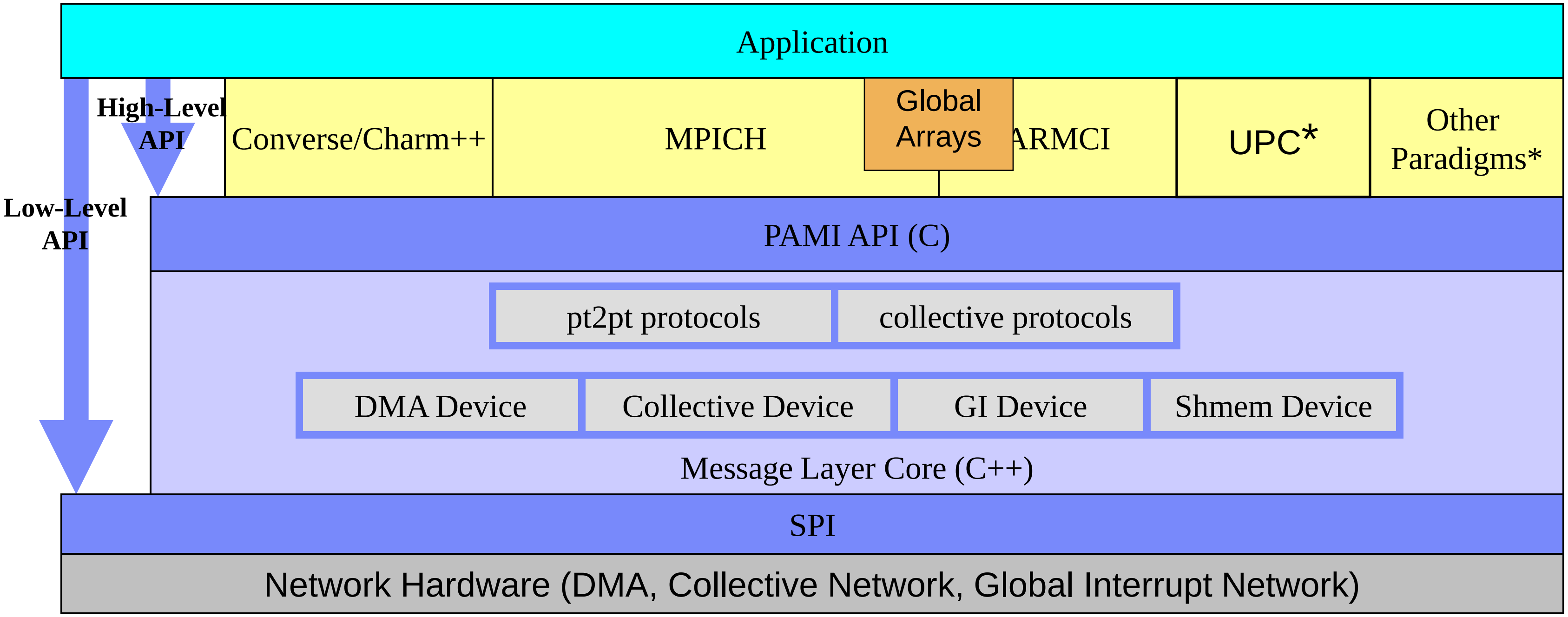
CNK Overview

Compute Node Kernel (CNK)
- Binary Compatible with Linux System Calls
  - Leverage Linux runtime environments and tools
- Up to 64 Processes (MPI Tasks) per Node
  - SPMD and MPMD Support
- Multi-Threading: optimized runtimes
  - Native POSIX Threading Library (NPTL)
  - OpenMP via XL and Gnu Compilers
  - Thread-Level Speculation (TLS)
- System Programming Interfaces (SPI)
  - Networks and DMA, Global Interrupts
  - Synchronization, Locking, Sleep/Wake
  - Performance Counters (UPC)
- MPI and OpenMP (XL, Gnu)
- Transactional Memory (TM)
- Speculative Multi-Threading (TLS)
- Shared and Persistent Memory
- Scripting Environments (Python)
- Dynamic Linking, Demand Loading

Firmware
- Boot, Configuration, Kernel Load
- Control System Interface
- Common RAS Event Handling for CNK & Linux



Parallel Active Message Interface



- Message Layer Core has C++ message classes and other utilities to program the different network devices
- Support many programming paradigms

Blue Gene/Q Job Submission

- BG/L & BG/P Single interface for job submission
  - mpirun
  - submit
  - submit_job
  - mpiexec

- BG/P a single interface for job submission: **runjob**

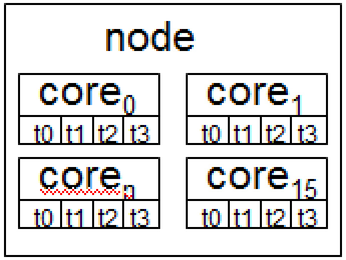
Ranks per node

- BG/L mpirun
  - supported SMP and co-processor mode
  - either 1 or 2 ranks per node

- BG/P mpirun
  - supported SMP, dual, and virtual node mode
  - either 1, 2, or 4 ranks per node

- BG/Q runjob
  - supports 1, 2, 4, 8, 16, 32, or 64 ranks per node
  - parameter is --ranks-per-node rather than --mode

Execution Modes in BG/Q per Node



node

core$_0$ t0 t1 t2 t3

core$_1$ t0 t1 t2 t3

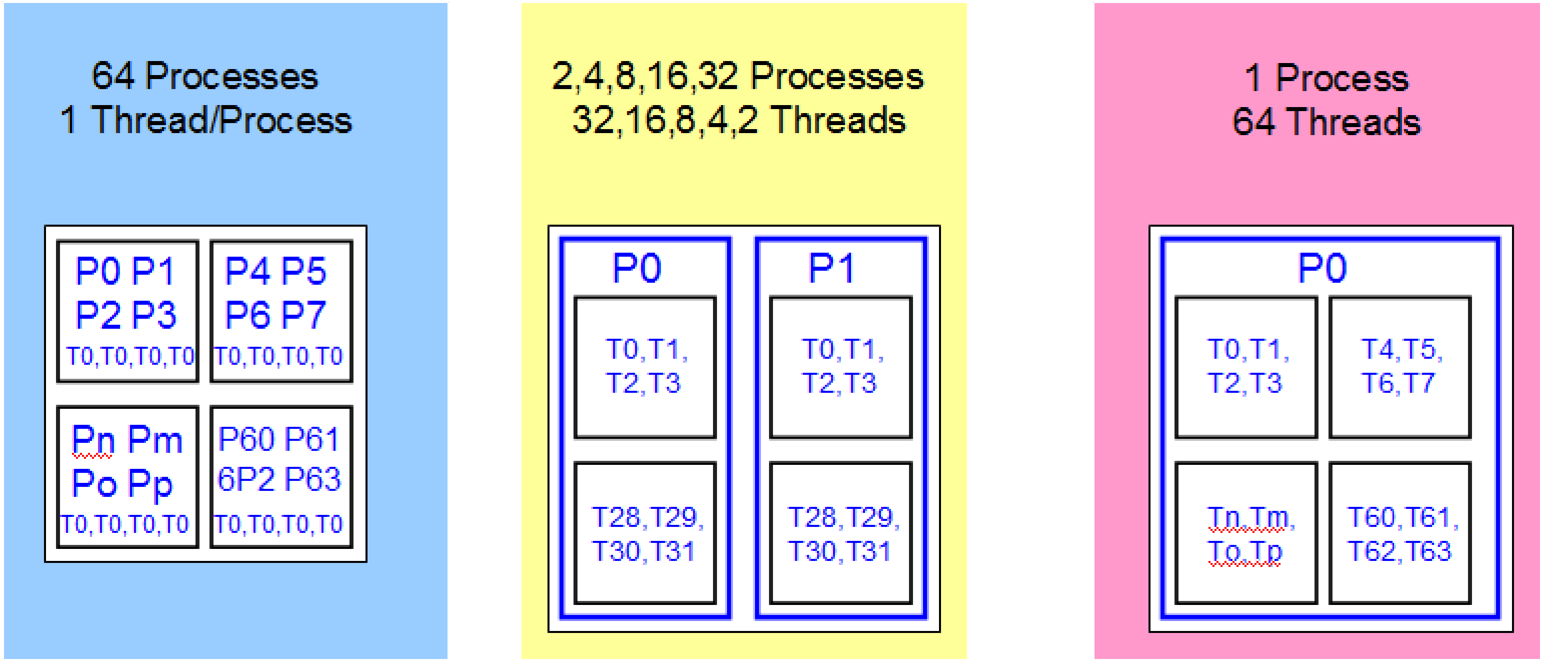
core$_n$ t0 t1 t2 t3

core$_{15}$ t0 t1 t2 t3

Hardware Abstractions Black

Software Abstractions Blue

- **Next Generation HPC**
  - **Many Core**
  - **Expensive Memory**
  - **Two-Tiered Programming Model**



64 Processes
1 Thread/Process

P0 P1
P2 P3
T0,T0,T0,T0

P4 P5
P6 P7
T0,T0,T0,T0

Pn Pm
Po Pp
T0,T0,T0,T0

P60 P61
6P2 P63
T0,T0,T0,T0

2,4,8,16,32 Processes
32,16,8,4,2 Threads

P0
T0,T1,
T2,T3
T28,T29,
T30,T31

P1
T0,T1,
T2,T3
T28,T29,
T30,T31

1 Process
64 Threads

P0
T0,T1,
T2,T3
T4,T5,
T6,T7
Tn,Tm,
To,Tp
T60,T61,
T62,T63

BG/Q MPI Implementation

- MPI-2.1 standard (http://www.mpi-forum.org/docs/docs.html)

- BG/Q mpi execution command: runjob

- To support the Blue Gene/Q hardware, the following additions and modifications have been made to the MPICH2 software architecture:
  - A Blue Gene/Q driver has been added that implements the MPICH2 abstract device interface (ADI).
  - Optimized versions of the Cartesian functions exist (MPI_Dims_create(), MPI_Cart_create(), MPI_Cart_map()).
  - MPIX functions create hardware-specific MPI extensions.

5-Dimensional Torus Network

- The 5-dimensional Torus network provides point-to-point and collective communication facilities.

- point-to-point messaging

  the route from a sender to a receiver on a torus network has the following two possible paths:

  - Deterministic routing
    - Packets from a sender to a receiver go along the same path.
    - Advantage: Latency - maintained without additional logic. However, this technique also creates
    - Disadvantage: network hot spots with several point-to-point coms

  - Adaptive routing
    - This technique generates a more balanced network load but introduces a latency penalty.

- Selecting deterministic or adaptive routing depends on the protocol used for communication – 4 in BG/Q: Immediate Message, MPI short, MPI eager and MPI rendez-vous

- environment variables can be used to customize MPI communications (c.f. IBM BG/Q redbook)

Blue Gene/Q extra MPI communicators

- **int MPIX_Cart_comm_create (MPI_Comm *cart_comm)**
  - This function creates a six-dimensional (6D) Cartesian communicator that mimics the exact hardware on which it is run. The A, B, C, D, and E dimensions match those of the block hardware, while the T dimension is equivalent to the ranks per node argument to **runjob**.

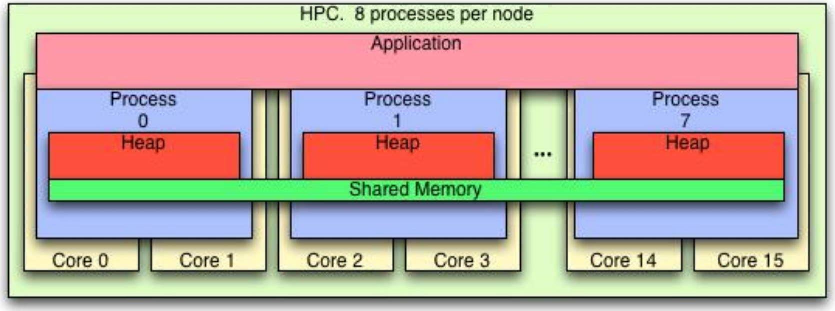
- **Changing class-route usage at runtime**
  - int MPIX_Comm_update(MPI_Comm comm, int optimize)

- **Determining hardware properties**
  - Int MPIX_Init_hw(MPIX_Hardware_t *hw);
  - int MPIX_Torus_ndims(int *numdimensions)
  - int MPIX_Rank2torus(int rank, int *coords)
  - int MPIX_Torus2rank(int *coords, int *rank)

Blue Gene/Q Kernel Overview

Processes

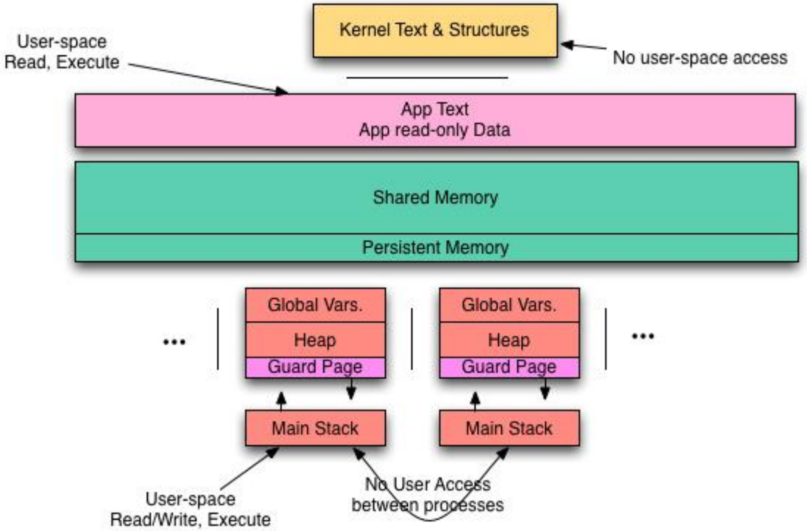
- Similarities to BGP
  - Number of tasks per node fixed at job start
  - No fork/exec support
  - Support static and dynamically linked processes
- Plus:
  - 64-bit processes
  - Support for 1, 2, 4, 8, 16, 32, or 64 processes per node
    - Numeric "names" for process config. (i.e., not smp, dual, quad, octo, vnm, etc)
  - Processes use 16 cores
  - The 17th core on BQC reserved for:
    - Application agents
    - Kernel networking
    - RAS offloading
  - Sub-block jobs
- Minus
  - No support for 32-bit processes



Threads

- Similarities to BGP
  - POSIX NPTL threading support
    - E.g., libpthread
- Plus
  - Thread affinity and thread migration
  - Thread priorities
    - Support both pthread priority and A2 hardware thread priority
  - Full scheduling support for A2 hardware threads
  - Multiple software threads per hardware thread is now the default
  - CommThreads have extended priority range compared to normal threads
  - Performance features
    - HWThreads in scheduler without pending work are put into hardware wait state
    - Snoop scheduler providing user-state fast access to:

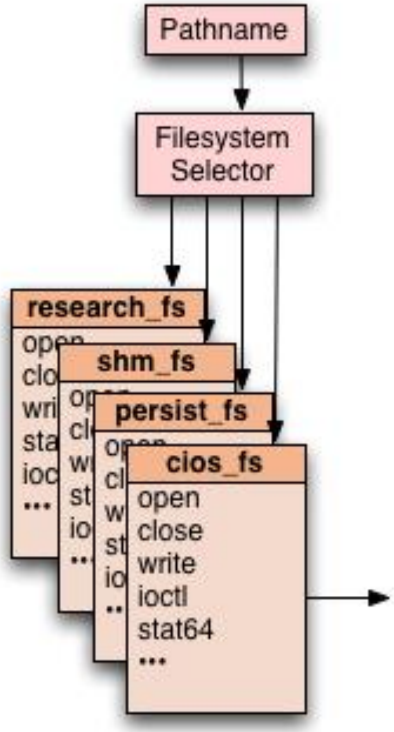
Memory



- Similarities to BGP
  - Application text segment is shared
  - Shared memory
  - Memory Protection and guard pages
- Plus
  - 64-bit virtual addresses
    - Supports up to 64GB of physical memo
    - No TLB misses
    - Up to 4 processes per core
  - Fixed 16MB memory footprint for CNK. Remainder of physical memory to applications
  - Memory protection for primordial dynamically-linked text segment
  - Memory aliases for long-running TM/SE
  - Globally readable memory
  - L2 atomics

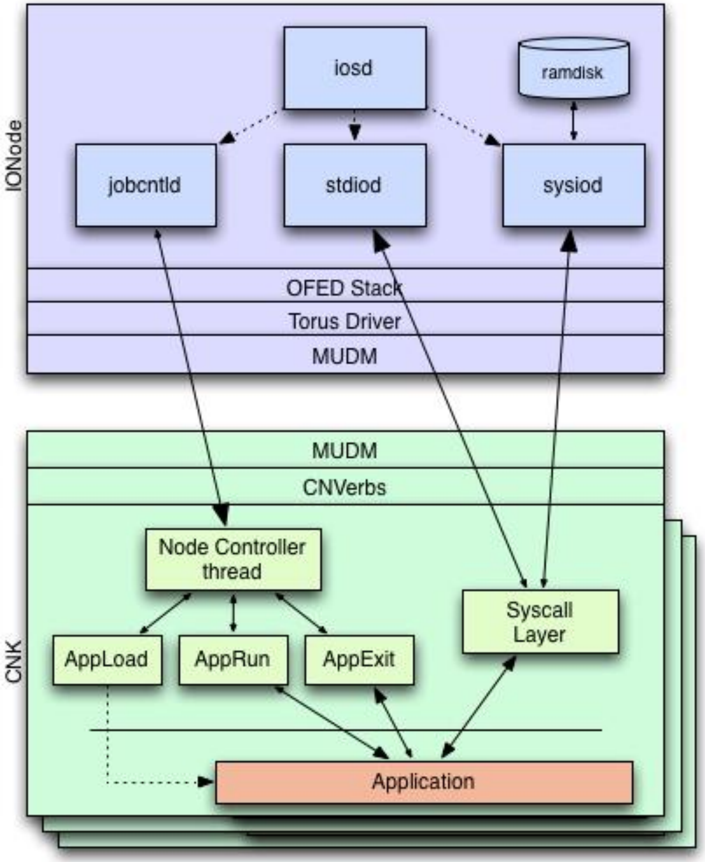
System Calls

- Similarities to BGP
  - Many common syscalls on Linux work on BG/Q.
  - Linux syscalls that worked on BGP should work on BGQ
- Plus
  - Support glibc 2.12.2
  - Real-time signals support
  - Low overhead syscalls
    - Only essential registers are saved and restored
  - Pluggable File Systems
    - Allows CNK to support multiple file system behaviors and types
    - File systems:Si



I/O Services

- Similarities to BGP
  - Function shipping system calls to ionode
  - Support NFS, GPFS, Lustre and PVFS2 filesystems
- Plus
  - PowerPC64 Linux running on 17 cores
  - Supports 8192:1 compute task to ionode ratio
    - Only 1 ioproxy per compute node
    - Significant internal changes from BGP
  - Standard communications protocol
    - OFED verbs
    - Using Torus DMA hardware for performance
  - Network over Torus
    - E.g., tools can now communicate between IONodes via torus
  - Using "off-the-shelf" Infiniband driver from Mellanox
  - ELF images now pulled from I/O nodes, vs push



Debugging

- Similarities to BGP
  - GDB
  - Totalview
  - Coreprocessor
  - Dump_all, dump_memory
  - Lightweight and binary corefiles
- Plus
  - Tools interface (CDTI)
    - Allows customers to write custom debug and monitoring tools
  - Support for versioned memory (TM/SE)
  - Fast breakpoint and watchpoint support
  - Asynchronous thread control
    - Allow selected threads to run while others are being debugged

BG/Q Application Environment

Compiling MPI programs on Blue Gene/Q

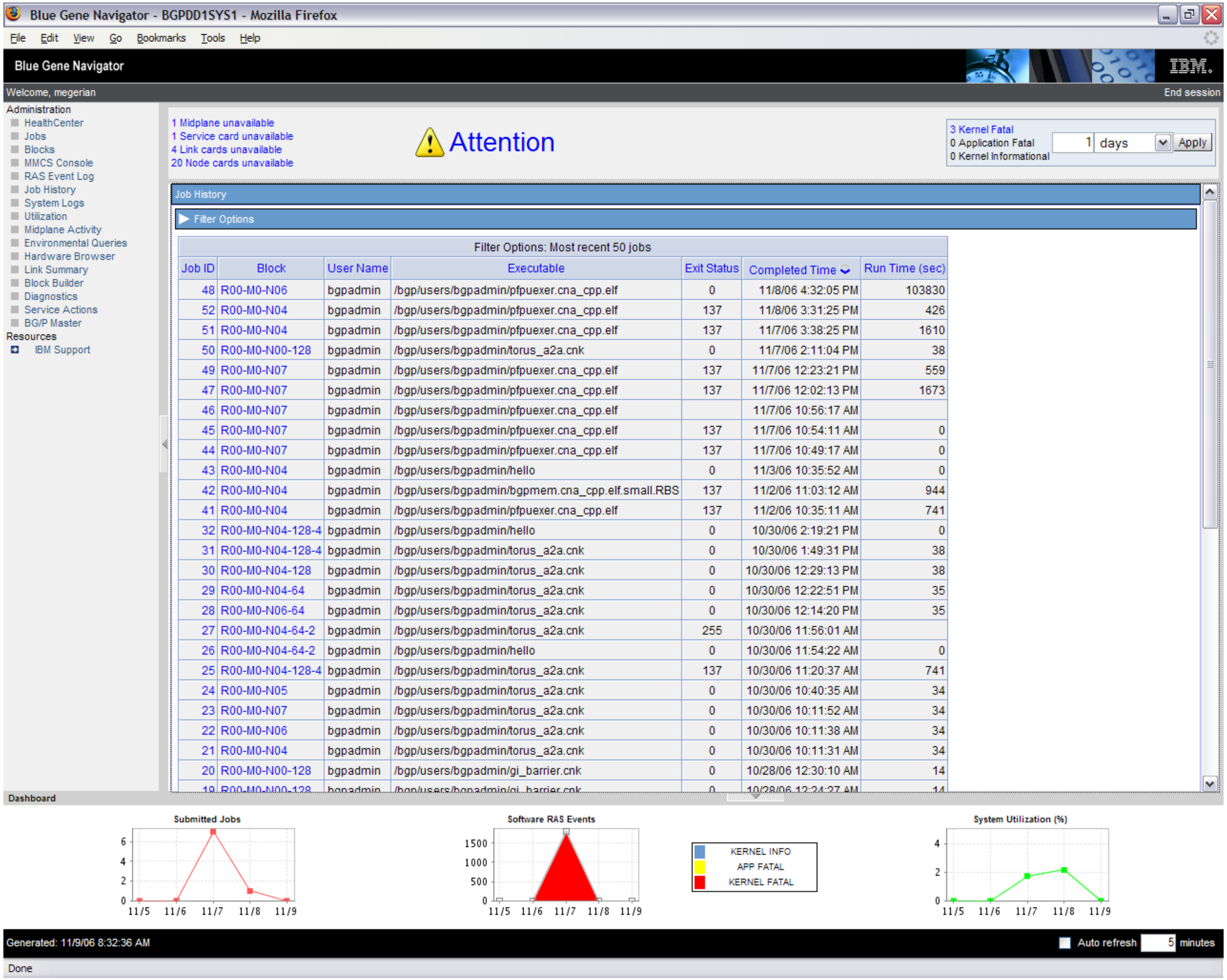
There are six versions of the libraries and the scripts

- **gcc**: GMU compiler with fine-grained locking in MPICH – error checking
- **gcc.legacy**: GMU with coarse-grained lock – sligthy better latency for single-thread code
- **xl**: PAMI compiled with GNU – fine-grained lock
- **xl.legacy**: PAMI compiled with GNU – coarse-grained lock
- **Xl.ndebug**: xl with error checking and asserts off
  ⇒lower latency but not as much debug info
- **xl.legacy.ndebug**: xl.legacy with error checking and asserts off

Control/monitoring

- **Provides the Blue Gene Web Services**
  - Getting data (blocks, jobs, hardware, envs, etc.)
  - Create blocks, delete blocks, run diags, etc.
- **A web server**
- **Runs under BGMaster**
  - Should run as special bgws user for security
- **New for BG/Q**
  - Had Navigator server in BG/P (Tomcat)
  - Tomcat in BG/L

Blue Gene Navigator



1 Midplane unavailable
1 Service card unavailable
4 Link cards unavailable
20 Node cards unavailable

Attention

3 Kernel Fatal
0 Application Fatal
0 Kernel Informational

1 days Apply

Job History

▶ Filter Options

Filter Options: Most recent 50 jobs

Job ID	Block	User Name	Executable	Exit Status	Completed Time	Run Time (sec)
48	R00-M0-N06	bgpadmin	/bgp/users/bgpadmin/pfpuexer.cna_cpp.elf	0	11/8/06 4:32:05 PM	103830
52	R00-M0-N04	bgpadmin	/bgp/users/bgpadmin/pfpuexer.cna_cpp.elf	137	11/8/06 3:31:25 PM	426
51	R00-M0-N04	bgpadmin	/bgp/users/bgpadmin/pfpuexer.cna_cpp.elf	137	11/7/06 3:38:25 PM	1610
50	R00-M0-N00-128	bgpadmin	/bgp/users/bgpadmin/torus_a2a.cnk	0	11/7/06 2:11:04 PM	38
49	R00-M0-N07	bgpadmin	/bgp/users/bgpadmin/pfpuexer.cna_cpp.elf	137	11/7/06 12:23:21 PM	559
47	R00-M0-N07	bgpadmin	/bgp/users/bgpadmin/pfpuexer.cna_cpp.elf	137	11/7/06 12:02:13 PM	1673
46	R00-M0-N07	bgpadmin	/bgp/users/bgpadmin/pfpuexer.cna_cpp.elf		11/7/06 10:56:17 AM	
45	R00-M0-N07	bgpadmin	/bgp/users/bgpadmin/pfpuexer.cna_cpp.elf	137	11/7/06 10:54:11 AM	0
44	R00-M0-N07	bgpadmin	/bgp/users/bgpadmin/pfpuexer.cna_cpp.elf	137	11/7/06 10:49:17 AM	0
43	R00-M0-N04	bgpadmin	/bgp/users/bgpadmin/hello	0	11/3/06 10:35:52 AM	0
42	R00-M0-N04	bgpadmin	/bgp/users/bgpadmin/bgpmem.cna_cpp.elf.small.RBS	137	11/2/06 11:03:12 AM	944
41	R00-M0-N04	bgpadmin	/bgp/users/bgpadmin/pfpuexer.cna_cpp.elf	137	11/2/06 10:35:11 AM	741
32	R00-M0-N04-128-4	bgpadmin	/bgp/users/bgpadmin/hello	0	10/30/06 2:19:21 PM	0
31	R00-M0-N04-128-4	bgpadmin	/bgp/users/bgpadmin/torus_a2a.cnk	0	10/30/06 1:49:31 PM	38
30	R00-M0-N04-128	bgpadmin	/bgp/users/bgpadmin/torus_a2a.cnk	0	10/30/06 12:29:13 PM	38
29	R00-M0-N04-64	bgpadmin	/bgp/users/bgpadmin/torus_a2a.cnk	0	10/30/06 12:22:51 PM	35
28	R00-M0-N06-64	bgpadmin	/bgp/users/bgpadmin/torus_a2a.cnk	0	10/30/06 12:14:20 PM	35
27	R00-M0-N04-64-2	bgpadmin	/bgp/users/bgpadmin/torus_a2a.cnk	255	10/30/06 11:56:01 AM	
26	R00-M0-N04-64-2	bgpadmin	/bgp/users/bgpadmin/hello	0	10/30/06 11:54:22 AM	0
25	R00-M0-N04-128-4	bgpadmin	/bgp/users/bgpadmin/torus_a2a.cnk	137	10/30/06 11:20:37 AM	741
24	R00-M0-N05	bgpadmin	/bgp/users/bgpadmin/torus_a2a.cnk	0	10/30/06 10:40:35 AM	34
23	R00-M0-N07	bgpadmin	/bgp/users/bgpadmin/torus_a2a.cnk	0	10/30/06 10:11:52 AM	34
22	R00-M0-N06	bgpadmin	/bgp/users/bgpadmin/torus_a2a.cnk	0	10/30/06 10:11:38 AM	34
21	R00-M0-N04	bgpadmin	/bgp/users/bgpadmin/torus_a2a.cnk	0	10/30/06 10:11:31 AM	34
20	R00-M0-N00-128	bgpadmin	/bgp/users/bgpadmin/gi_barrier.cnk	0	10/28/06 12:30:10 AM	14
19	R00-M0-N00-128	bgpadmin	/bgp/users/bgpadmin/gi_barrier.cnk	0	10/28/06 12:24:27 AM	14

Generated: 11/9/06 8:32:36 AM

□ Auto refresh 5 minutes

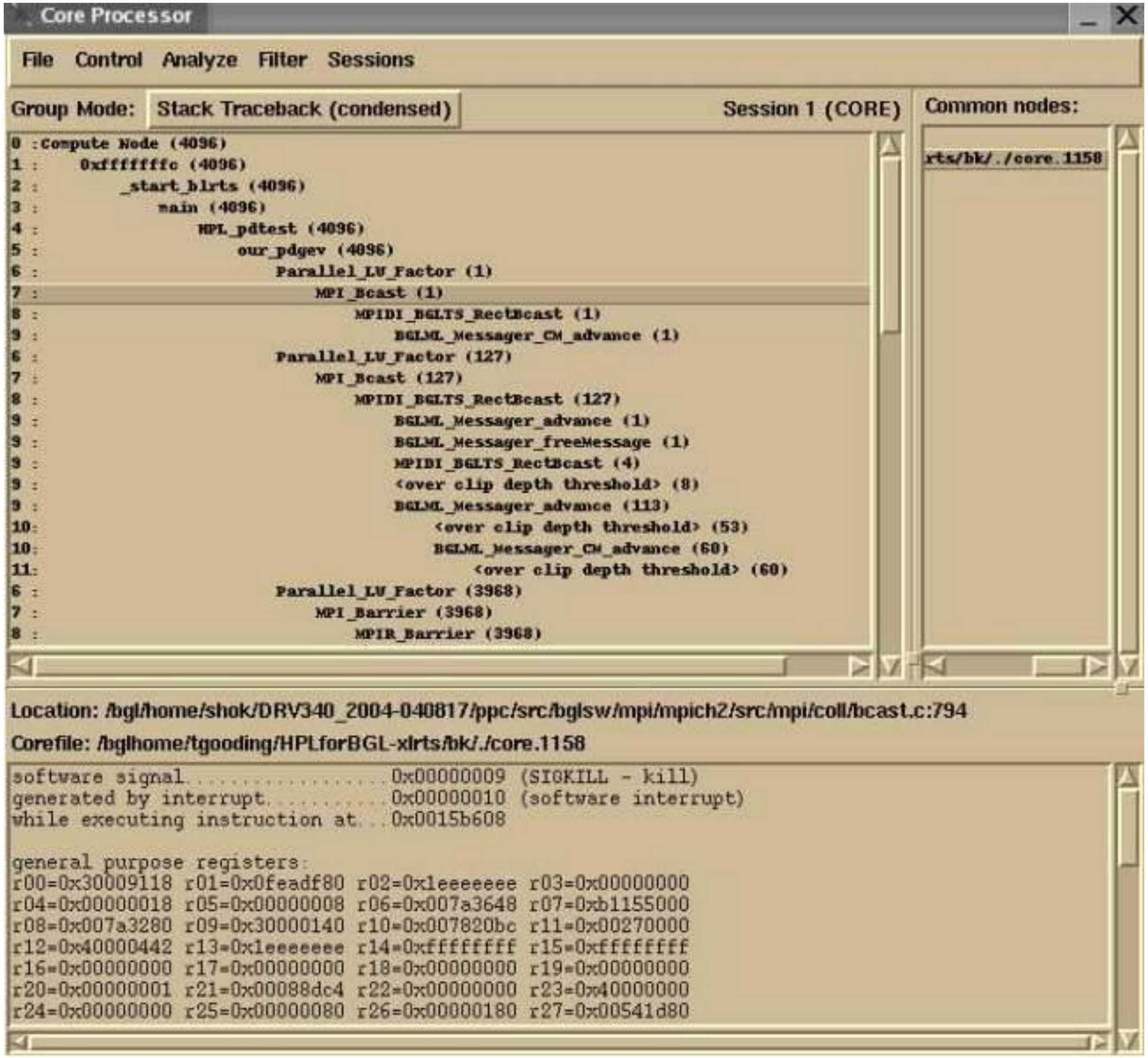
Done

Debugging – Batch scheduler

- Debugging
  - Integrated Tools
    - GDB
    - Core Files + addr2line
    - coreprossecor
    - Compiler Options
    - Traceback functions, memory size kernel, signal or exit trap, …
  - Supported Commercial Software
    - Totalview
    - DDT (Alinea) ?

- Batch scheduler
  - IBM LoadLeveler
  - SULRM
  - LSF?

Coreprocessor GUI



Group Mode: Stack Traceback (condensed)

Session 1 (CORE)

```
0 :Compute Node (4096)
1 :    0xfffffffc (4096)
2 :        _start_blrts (4096)
3 :            main (4096)
4 :                HPL_pdtest (4096)
5 :                    our_pdgev (4096)
6 :                        Parallel_LU_Factor (1)
7 :                            MPI_Bcast (1)
8 :                                MPIDI_BGLTS_RectBcast (1)
9 :                                    BGLML_Messager_CM_advance (1)
6 :                        Parallel_LU_Factor (127)
7 :                            MPI_Bcast (127)
8 :                                MPIDI_BGLTS_RectBcast (127)
9 :                                    BGLML_Messager_advance (1)
9 :                                    BGLML_Messager_freeMessage (1)
9 :                                    MPIDI_BGLTS_RectBcast (4)
9 :                                    <over clip depth threshold> (8)
9 :                                    BGLML_Messager_advance (113)
10:                                        <over clip depth threshold> (53)
10:                                        BGLML_Messager_CM_advance (60)
11:                                            <over clip depth threshold> (60)
6 :                        Parallel_LU_Factor (3968)
7 :                            MPI_Barrier (3968)
8 :                                MPIR_Barrier (3968)
```

```
rts/bk/./core.1158
```

Location: /bgl/home/shok/DRV340_2004-040817/ppc/src/bglsw/mpi/mpich2/src/mpi/coll/bcast.c:794

Corefile: /bglhome/tgooding/HPLforBGL-xlrts/bk/./core.1158

```
software signal..................0x00000009 (SIGKILL - kill)
generated by interrupt...........0x00000010 (software interrupt)
while executing instruction at...0x0015b608

general purpose registers:
r00=0x30009118 r01=0x0feadf80 r02=0x1eeeeeee r03=0x00000000
r04=0x00000018 r05=0x00000008 r06=0x007a3648 r07=0xb1155000
r08=0x007a3280 r09=0x30000140 r10=0x007820bc r11=0x00270000
r12=0x40000442 r13=0x1eeeeeee r14=0xffffffff r15=0xffffffff
r16=0x00000000 r17=0x00000000 r18=0x00000000 r19=0x00000000
r20=0x00000001 r21=0x00088dc4 r22=0x00000000 r23=0x40000000
r24=0x00000000 r25=0x00000080 r26=0x00000180 r27=0x00541d80
```

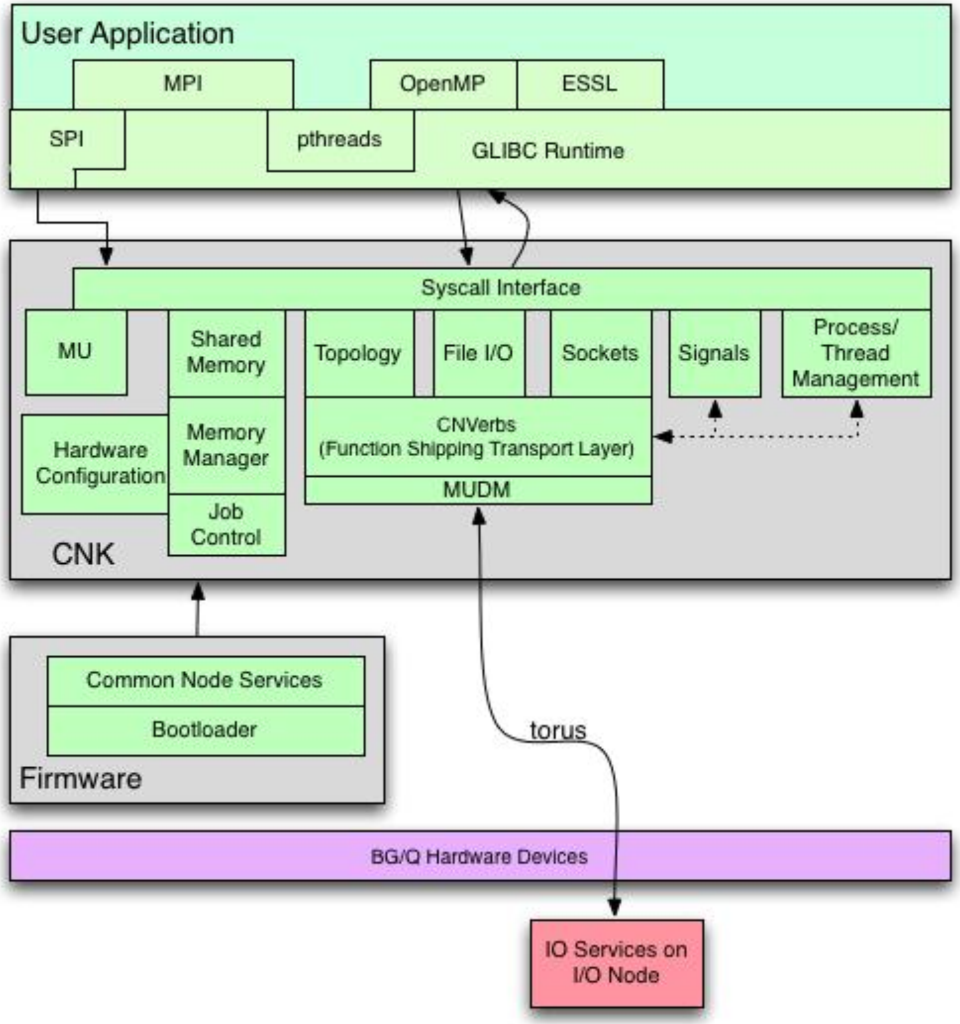
Performance Analysis

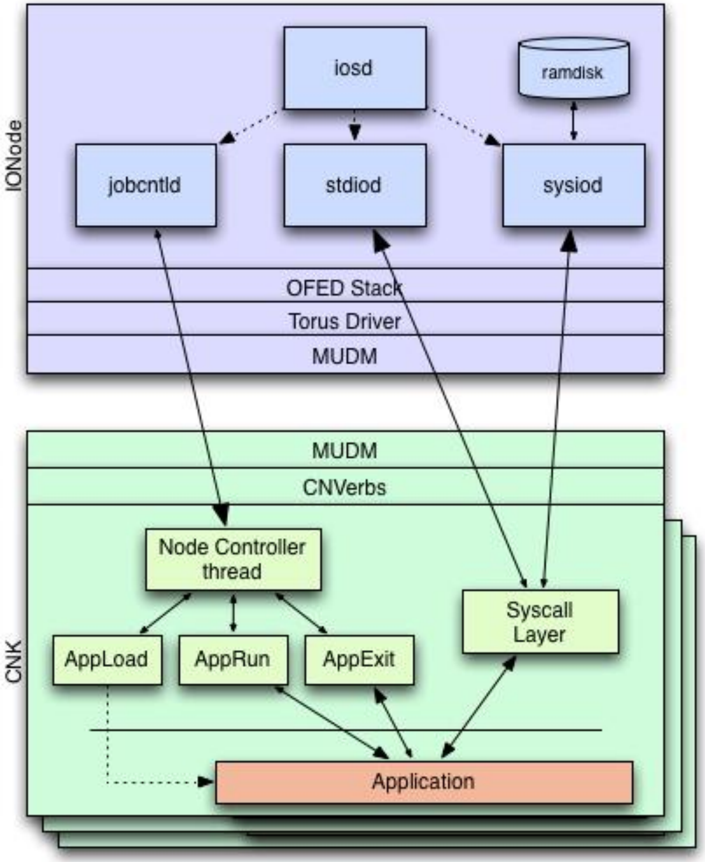
- Profiling
  - GNU profiling, vprof with command line or GUI
- IBM HPC Toolkit, IBM mpitrace library
- Major Open-Source Tools
  - Scalasca
  - TAU
  - mpiP http://mpip.sourceforge.net
  - …

IBM mpi trace library – HPC Toolkit

- Mpi timing summary
- Communication and elapsed times
- Heap memory used
- Mpi basic informations: #calls, message sizes, # hops
- Call-stack for every MPI function call
- Source and destination torus coordinates identification for point-to-point messages
- Unix-based profiling
- BG/Q Hardware-counters
- Event-tracing

Thanks, Questions?


User Application
MPI
OpenMP
ESSL
SPI
pthreads
GLIBC Runtime
Syscall Interface
MU
Shared Memory
Topology
File I/O
Sockets
Signals
Process/ Thread Management
Hardware Configuration
Memory Manager
CNVerbs (Function Shipping Transport Layer)
MUDM
Job Control
CNK
Common Node Services
Bootloader
Firmware
torus
BG/Q Hardware Devices
IO Services on I/O Node


IONode
iosd
ramdisk
jobcntld
stdiod
sysiod
OFED Stack
Torus Driver
MUDM
MUDM
CNVerbs
CNK
Node Controller thread
Syscall Layer
AppLoad
AppRun
AppExit
Application