Argonne NATIONAL LABORATORY

BG/Q Architecture

Scott Parker
ALCF Performance Engineering

U.S. DEPARTMENT OF ENERGY

Evolution of the Blue Gene

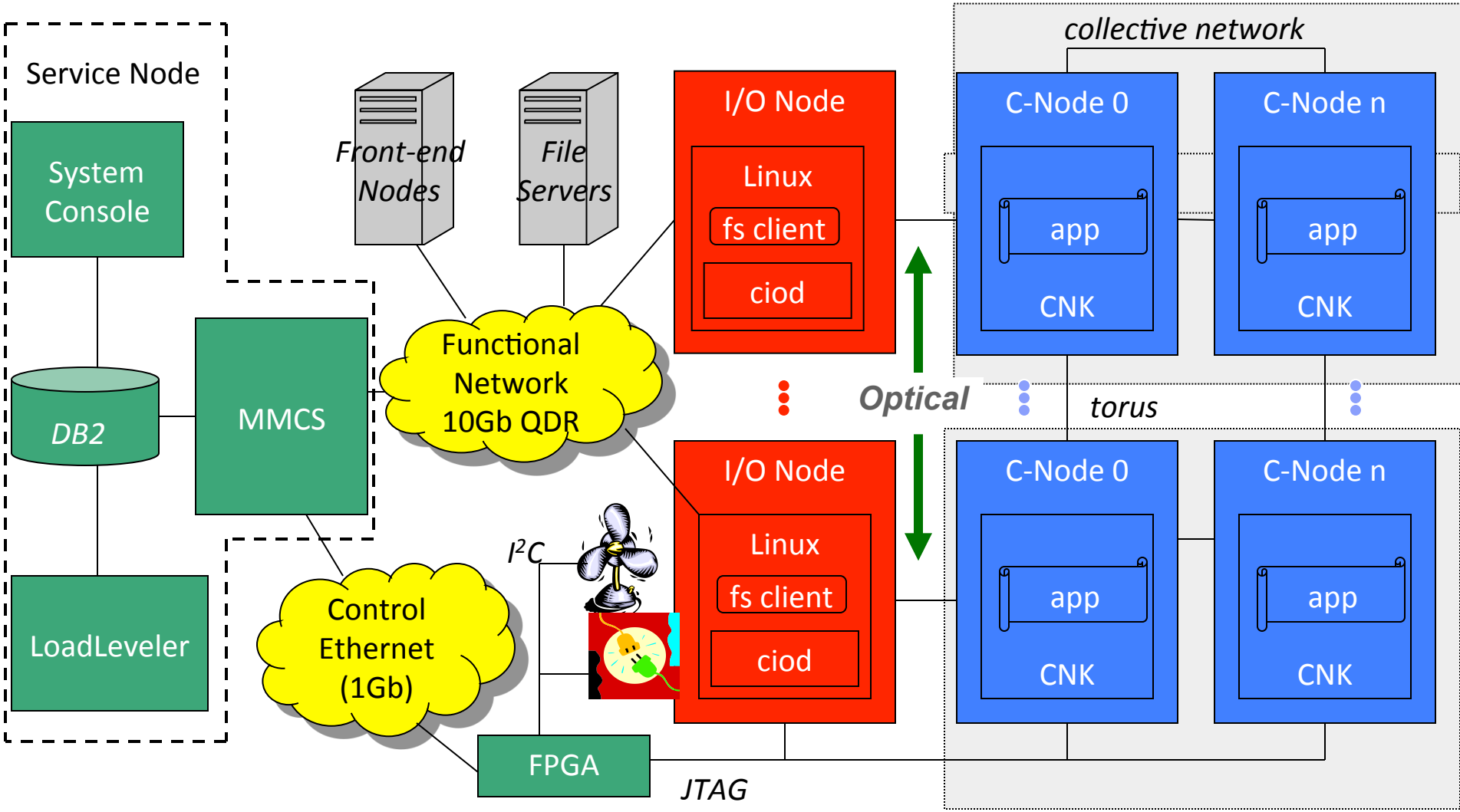
Blue Gene/L (2004)

Blue Gene/P (2007)

Blue Gene/Q (2012)

- Extends the Blue Gene Architecture
  - PowerPC Processor
  - Massive Parallelism
  - Torus Network
  - Standard programming models
- Energy Efficient – Top of Green500
- Powerful – Top of Top500
- New Features:
  - Water cooled
  - More cores, more threads

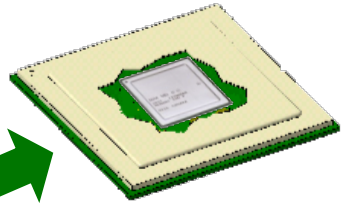
Blue Gene System Architecture



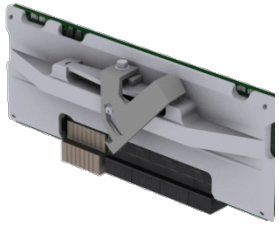
Blue Gene/Q

1. Chip:
16+2 cores

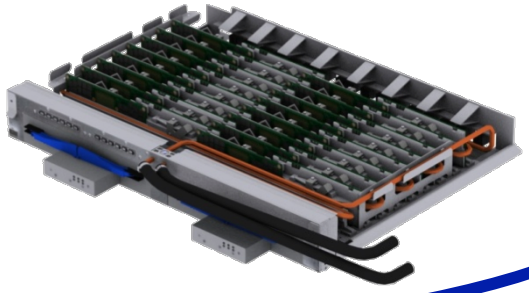
2. Single Chip Module



3. Compute card:
One chip module,
16 GB DDR3 Memory,
Heat Spreader for H_2O Cooling



4. Node Card:
32 Compute Cards,
Optical Modules, Link Chips; 5D Torus

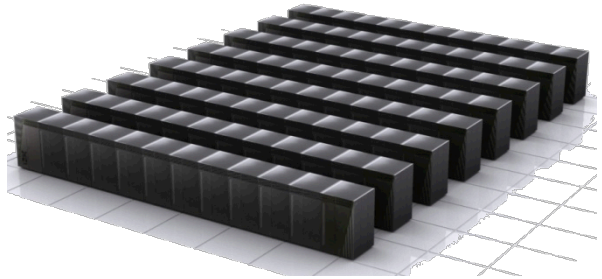


5a. Midplane:
16 Node Cards

5b. IO drawer:
8 IO cards w/16 GB
8 PCIe Gen2 x8 slots
3D I/O torus

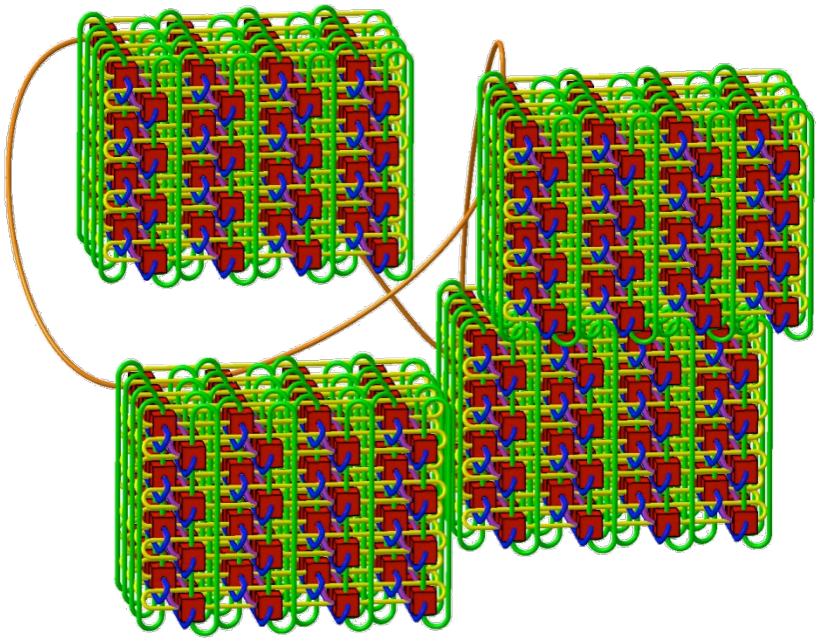
6. Rack: 2 Midplanes

7. System:
96 racks, 20PF/s



- **Sustained single node perf: 10x P, 20x L**
- **MF/Watt: (6x) P, (10x) L (~2GF/W, Green 500 criteria)**
- **Software and hardware support for programming models for exploitation of node hardware concurrency**

Inter-Processor Communication

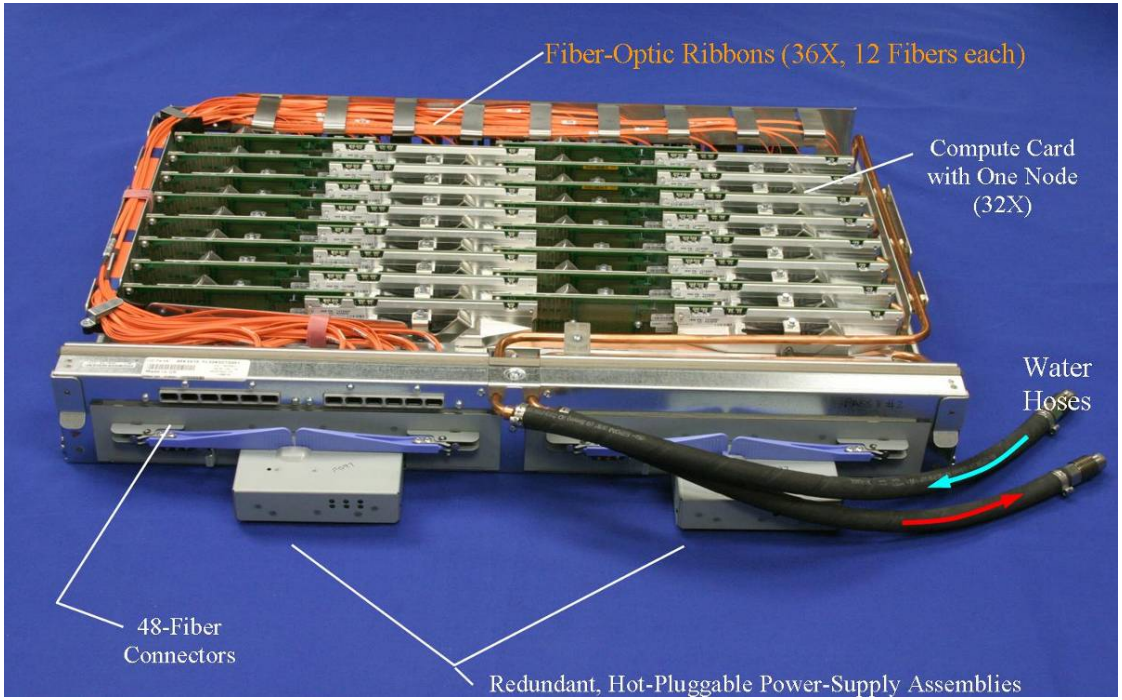


Network Performance

- **All-to-all: 97% of peak**
- **Bisection: > 93% of peak**
- **Nearest-neighbor: 98% of peak**
- **Collective: FP reductions at 94.6% of peak**

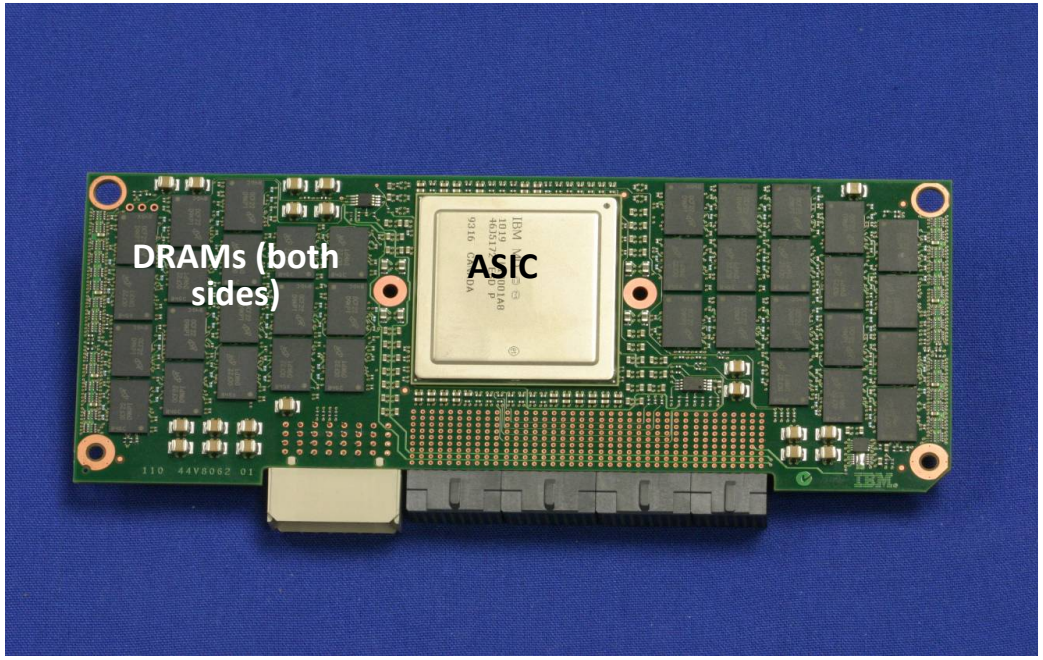
- **Integrated 5D torus:**
  - Achieves high nearest neighbor bandwidth while increasing bisectional bandwidth and reducing hops
  - Allows machine to be partitioned into independent sub machines. No impact from concurrently running codes.
  - Single network used for point-to-point, collectives, and barrier operations
  - Hardware assists for collective & barrier functions
  - Half rack (midplane) is 4x4x4x4x2 torus
- **Nodes have 10 links with 2 GB/s raw bandwidth each**
  - Bi-directional: send + receive gives 4 GB/s
  - 90% of bandwidth (1.8 GB/s) available to user
- **Hardware latency**
  - Nearest: 80ns
  - Farthest: 3us (96-rack 20PF system, 31 hops)
- **Additional 11th link for communication to IO nodes**
  - BQC chips in separate enclosure
  - IO nodes run Linux, mount file system
  - IO nodes drive PCIe Gen2 x8 (4+4 GB/s)
    ↔ IB/10G Ethernet ↔ file system & world
- **Integrate on-chip Message Unit (RDMA)**

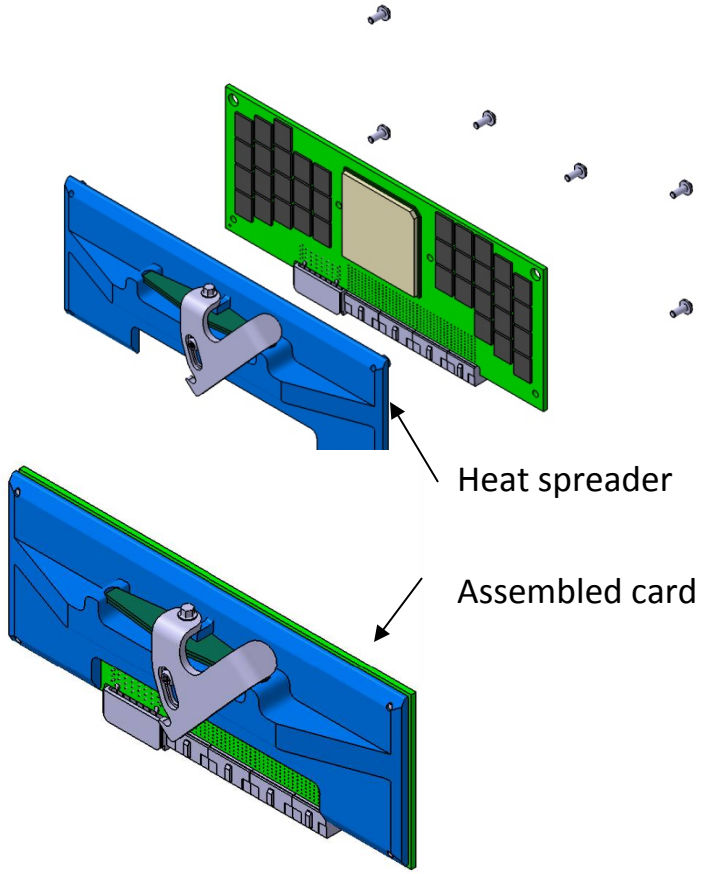
Blue Gene/Q Node Card Assembly



- High bandwidth / low latency electrical interconnect on-board
- Power efficient processor chips allow dense packaging
- 32 Compute Cards per Node Card
- Compute Node Card assembly is water-cooled (18-25ºC – above dew point)

Blue Gene/Q Compute Card

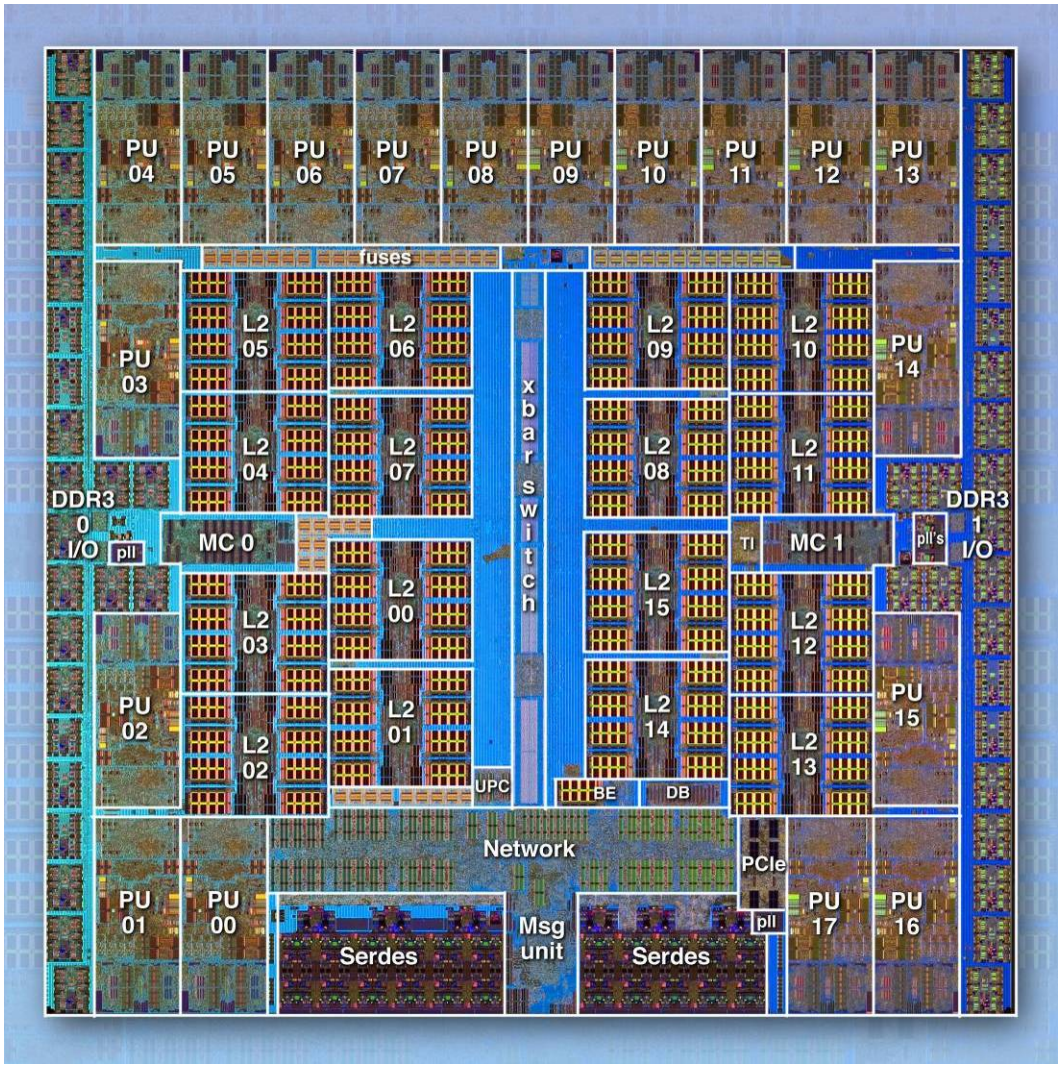




- **Node has 1 BQC chip + 72 SDRAMs (16GB DDR3)**
- **Memory is soldered on for high reliability**
- **Two heat sink options:**
  - Water-cooled → "Compute Node"
  - Air-cooled → "IO Node"

BlueGene/Q Compute Chip

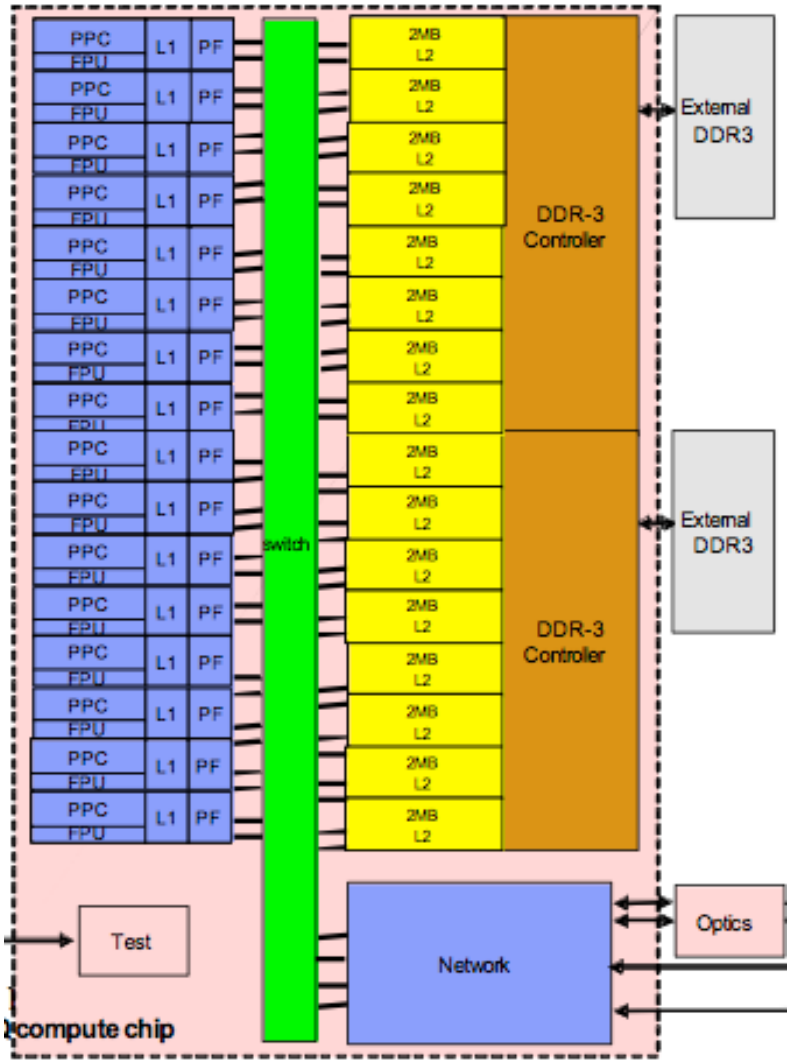
System-on-a-Chip design : integrates processors, memory and networking logic into a single chip



- **360 mm² Cu-45 technology (SOI)**
  - ~ 1.47 B transistors
- **16 user + 1 service processors**
  - plus 1 redundant processor
  - all processors are symmetric
  - L1 I/D cache = 16kB/16kB
  - L1 prefetch engines
- **Crossbar switch**
  - Connects cores via L1P to L2 slices
- **Central shared L2 cache**
  - 32 MB eDRAM
  - 16 slices
- **Dual memory controller**
  - 16 GB external DDR3 memory
  - 42.6 GB/s
- **Chip-to-chip networking**
  - Router logic integrated into BQC chip
  - DMA, remote put/get, collective operations
  - 11 network ports
- **External IO**
  - PCIe Gen2 interface

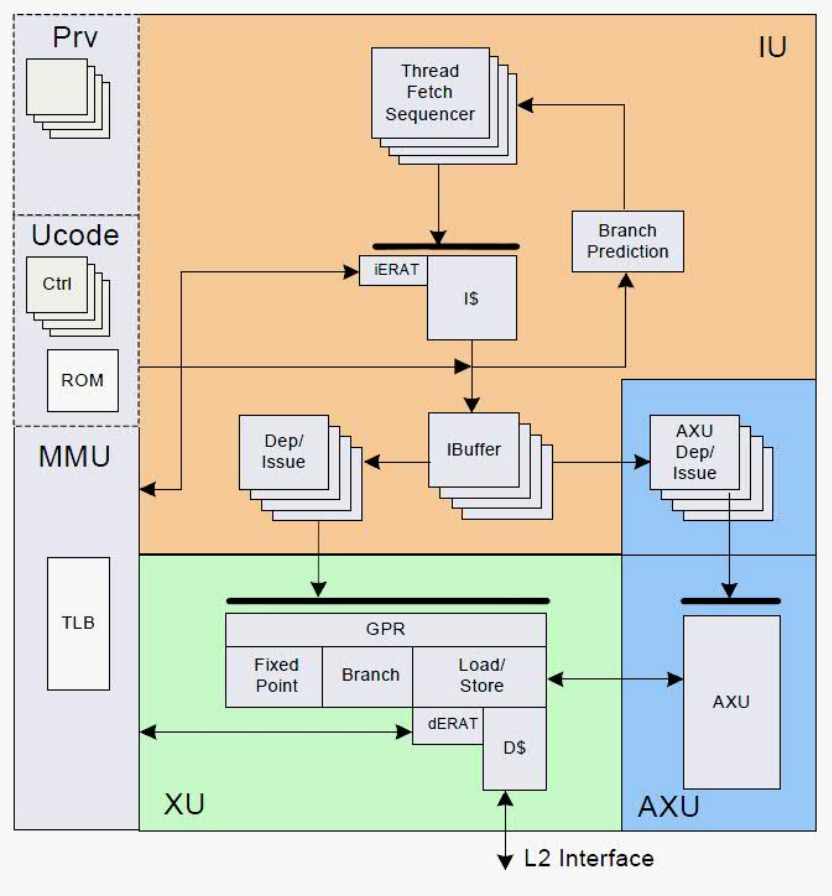
BG/Q Core

- Full PowerPC compliant 64-bit CPU (BG/P was 32-bit)
- 16 compute cores per node
- 17th core dedicated to system functions (OS and RAS)
- Each core attached directly to L1 cache & prefetcher
- 4 hardware threads per core
- 4-wide SIMD floating point unit (QPX)
- Transactional Memory & Speculative Execution
- Fast memory based atomic operations
- Stream and list based prefetching
- WakeUp Unit
- Universal Performance Counters

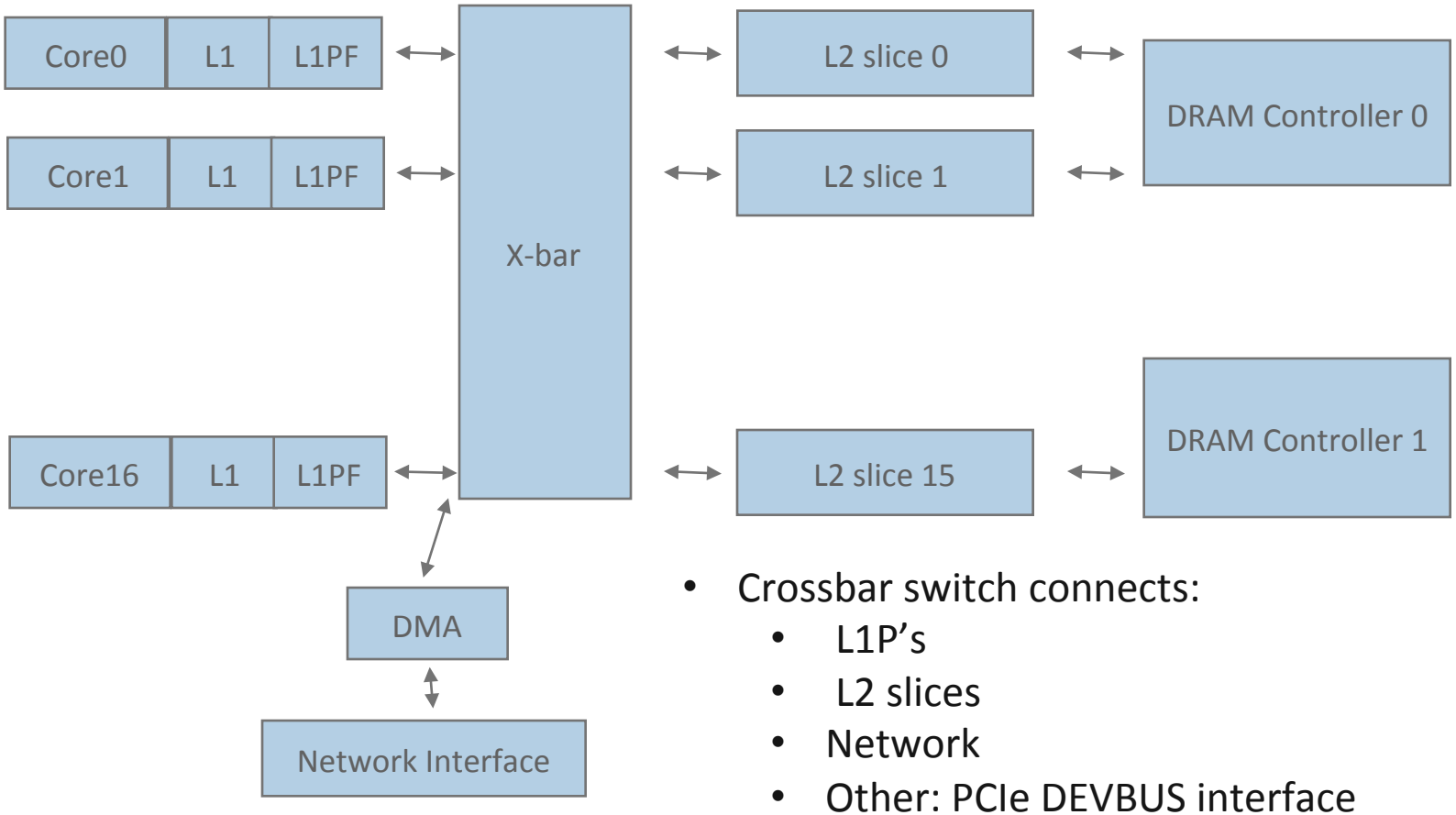


BG/Q Core

- Mostly same design as in PowerEN™ chip: Simple core, designed for excellent power efficiency and small footprint.
- Implemented 64-bit PowerISA™ v2.06
- 1.6 GHz @ 0.8V.
- 4-way Simultaneous Multi-Threading
- In-order execution
- 2-way concurrent issue 1 XU + 1 AXU
  - XU – integer, control, and memory operations
  - AXU – floating point operations
- A given thread may only issue 1 instruction per cycle
- Two threads may issue 1 instruction each cycle
- AXU port allows for unique BGQ style floating point
- 32x4x64 bit GPR
- Dynamic branch prediction
- L1 I/D cache = 16kB/16kB



BG/Q Crossbar Switch



- Crossbar switch connects:
  - L1P's
  - L2 slices
  - Network
  - Other: PCIe DEVBUS interface
- Aggregate bandwidth across slices:
  - Read: 409.6 GB/s
  - Write: 204.8 GB/s

Caches

- **L1 Cache:**
  - Data: 16KB, 8 way set associative, 64 byte lines, 32 byte load/store interface
  - Instruction: 16KB, 4 way set associative
- **L1 Prefetcher (L1P):**
  - 1 prefetch unit for each core
  - 32 entry prefetch buffer, entries are 128 bytes
  - Operates in List or Stream prefetch modes
- **L2 Cache:**
  - Shared by all cores
  - Divided into 16 slices connected via crossbar switch to each node
  - 32 MB total, 2 MB per slice, 16 way set associative
  - Supports memory speculation and atomic memory operations
  - Serves a point of coherency
  - Provides prefetch capabilities

Memory

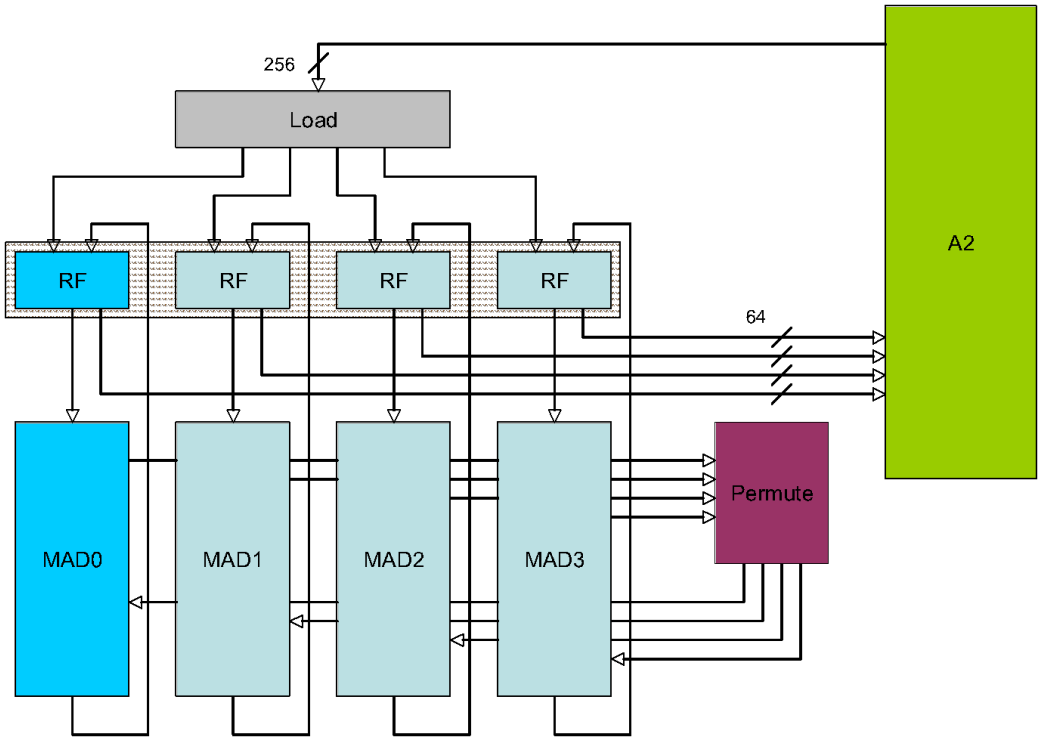
- Two on chip memory controllers
- Each connects to 8 L2 slices via 2 ring buses
- Each controller drives a 16+2 byte DDR-3 channel at 1.33 Gb/s
- Peak bandwidth is 42.67 BG/s (excluding ECC)

Network and Messaging Units

- **Network:**
  - Each chip has 11 network ports:
    - Each can transmit and receive at 2 GB/s
    - Total bandwidth of 44 GB/S
  - 10 links used to form a 5D torus between compute nodes
  - 1 link used to connect to IO node
  - 16 network injection FIFOs and 16 network reception FIFOs
- **Messaging Unit:**
  - Interface between the network and the BG/Q memory system
  - Supports direct puts, remote gets, and memory FIFO messages
  - Each Message Unit as 16 injection Message Engines and 16 Message Reception Engines each tied to a FIFO
  - Messaging Unit is connected to node cross-bar switch with 3 master and 1 slave port

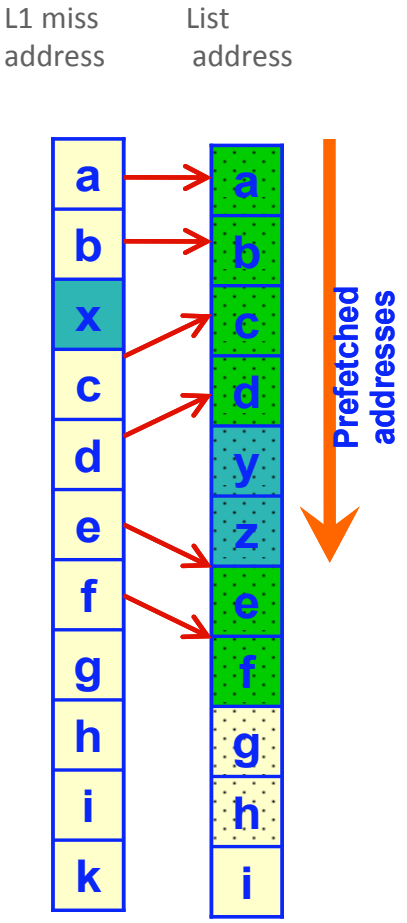
QPX Overview

- Instruction Extensions to PowerISA
- 4-wide double precision FPU SIMD (BG/L,P are 2-wide) usable as:
  - scalar FPU
  - 4-wide FPU SIMD
  - 2-wide complex arithmetic SIMD
- Attached to AXU port of A2 core – A2 issues one instruction/cycle to AXU
- 8 concurrent floating point operations (FMA) + load +store
- 6 stage pipeline
- Permute instructions to reorganize vector data
  - supports a multitude of data alignments
- 4R/2W register file
  - 32x32 bytes per thread
- 32B (256 bits) data path to/from L1 cache



BlueGene/Q L1P and WakeUp

- **L1 prefetcher**
  - Normal mode: Stream Prefetching
    - in response to observed memory traffic, adaptively balances resources to prefetch L2 cache lines (@ 128 B wide)
    - from 16 streams x 2 deep through 4 streams x 8 deep
  - Additional: 4 List-based Prefetching engines:
    - One per thread
    - Activated by program directives, e.g. bracketing complex set of loops
    - Used for repeated memory reference patterns in arbitrarily long code segments
    - Record pattern on first iteration of loop; playback for subsequent iterations
    - On subsequent passes, list is adaptively refined for missing or extra cache misses (async events)



- **Wake-up unit**
  - Will allow SMT threads to be suspended, while waiting for an event
  - Lighter weight than wake-up-on-interrupt -- no context switching
  - Improves power efficiency and resource utilization

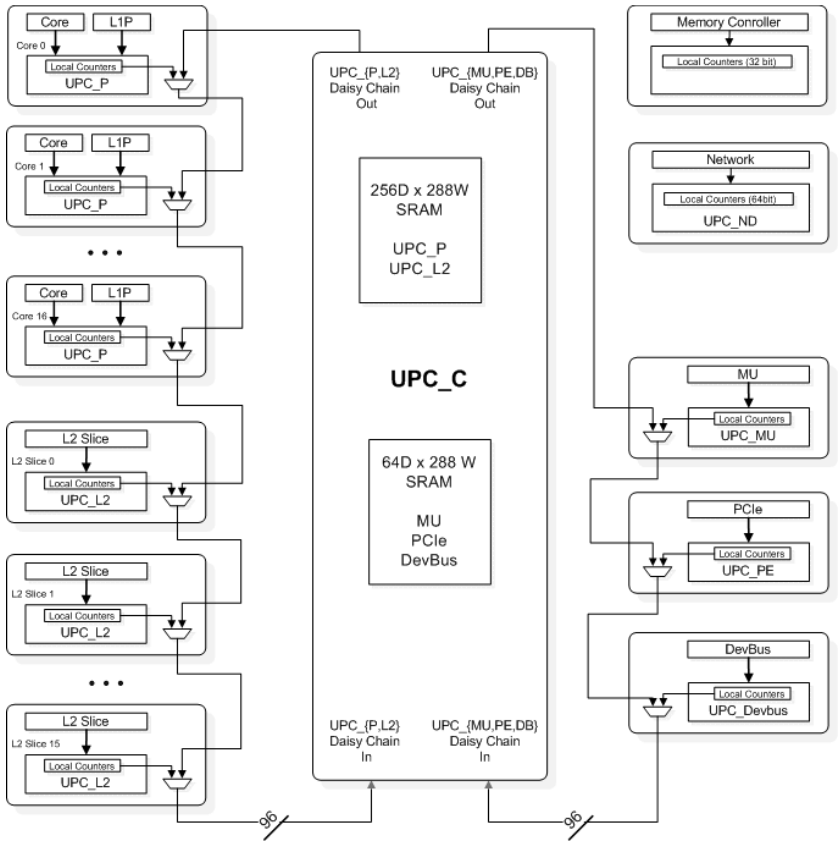
List-based "perfect" prefetching has tolerance for missing or extra cache misses

Transactional Memory and Atomics

- **L2 cache provides transactional memory and fast atomics**
- **Transactional Memory & Speculative Execution:**
  - Multi-versioned L2 cache
  - Changes kept separate from main memory state and reverted or committed
  - Tracks conflicts between threads (read-after-write, write-after-read, write-after-write)
  - Can store up to 30MB of speculative state
- **Fast Atomics:**
  - 8 byte load & store operations that can alter the value at memory address
  - Atomic use standard load & store instructions with special high order address bits
  - Operations:
    - LoadClear, LoadIncrement, LoadDecrement, LoadIncrementBounded, LoadDecrementBounded,
    - StoreAdd, StoreAddCoherenceOnZero, StoreTwin, StoreOr, StoreXor, StoreMaxUnsigned, StoreMaxSigned,

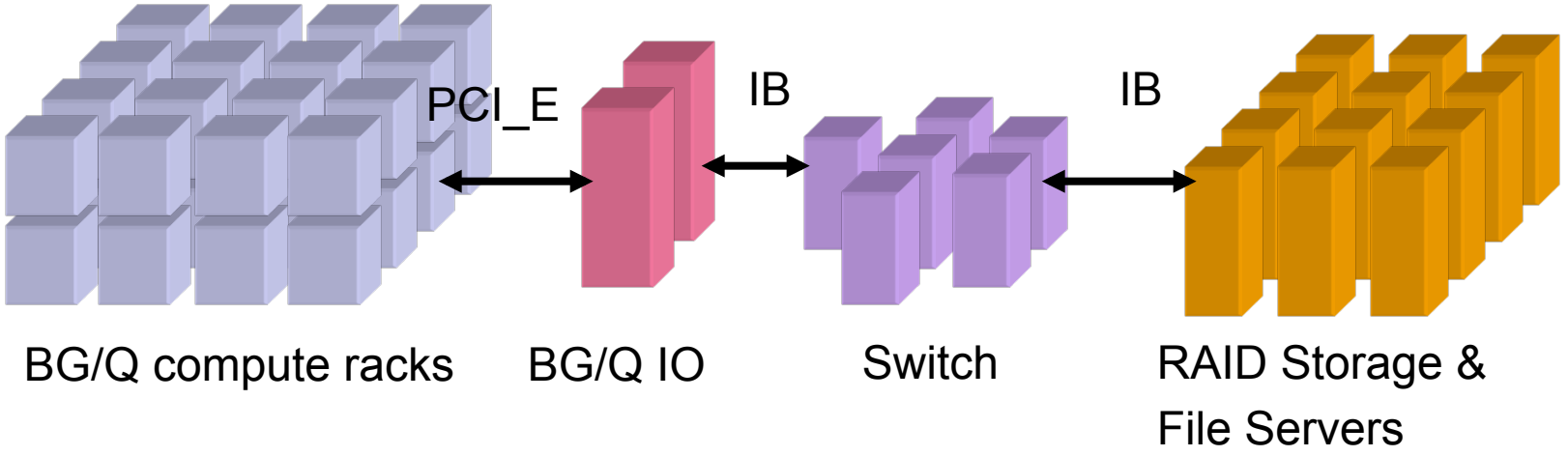
Hardware Performance Counters

- **Universal Performance Counter (UPC) unit collects hardware performance events from counters on:**
  - 17 cores
  - L1P's
  - Wakeup Units
  - 16 L2 slices
  - Message, PCIe, and DEVBUS units
- **Network Unit maintains a separate set of counters**



BG/Q IO

- IO design similar to BG/L and BG/P
- IO Nodes handle function shipped IO calls to parallel file system client
- IO nodes are not shared between compute partitions



BG I/O Max Bandwidth

	BG/L	BG/P	BG/Q
Type	1GbE	10GbE	PCI-e
BW/node	1Gb/s x2 250MB/s	10Gb/s x2 2.5GB/s	4GB/s x2
# of I/O nodes	128	64	8-128
BW/rack in BW/rack out	16GB/s 16GB/s	80GB/s 80GB/s	512GB/s@128 512GB/s@128
I/O byte/flop	0.0056	0.011	0.0048

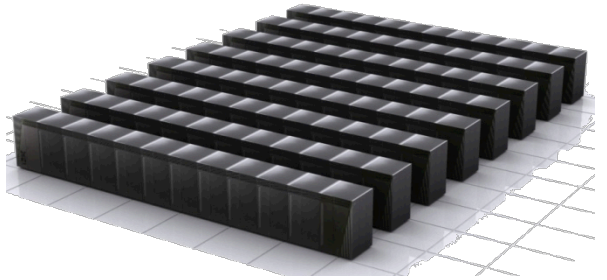
Blue Gene/Q Software High-Level Goals & Philosophy

- Facilitate extreme scalability
  - Extremely low noise on compute nodes
- High reliability: a corollary of scalability
- Familiar programming modes such as MPI and OpenMP
- Standards-based when possible
- Open source where possible
- Facilitate high performance for unique hardware:
  - Quad FPU, DMA unit, List-based prefetcher
  - TM (Transactional Memory), SE (Speculative Execution)
  - Wakeup-Unit, Scalable Atomic Operations
- Optimize MPI and native messaging performance
- Optimize libraries
- Facilitate new programming models

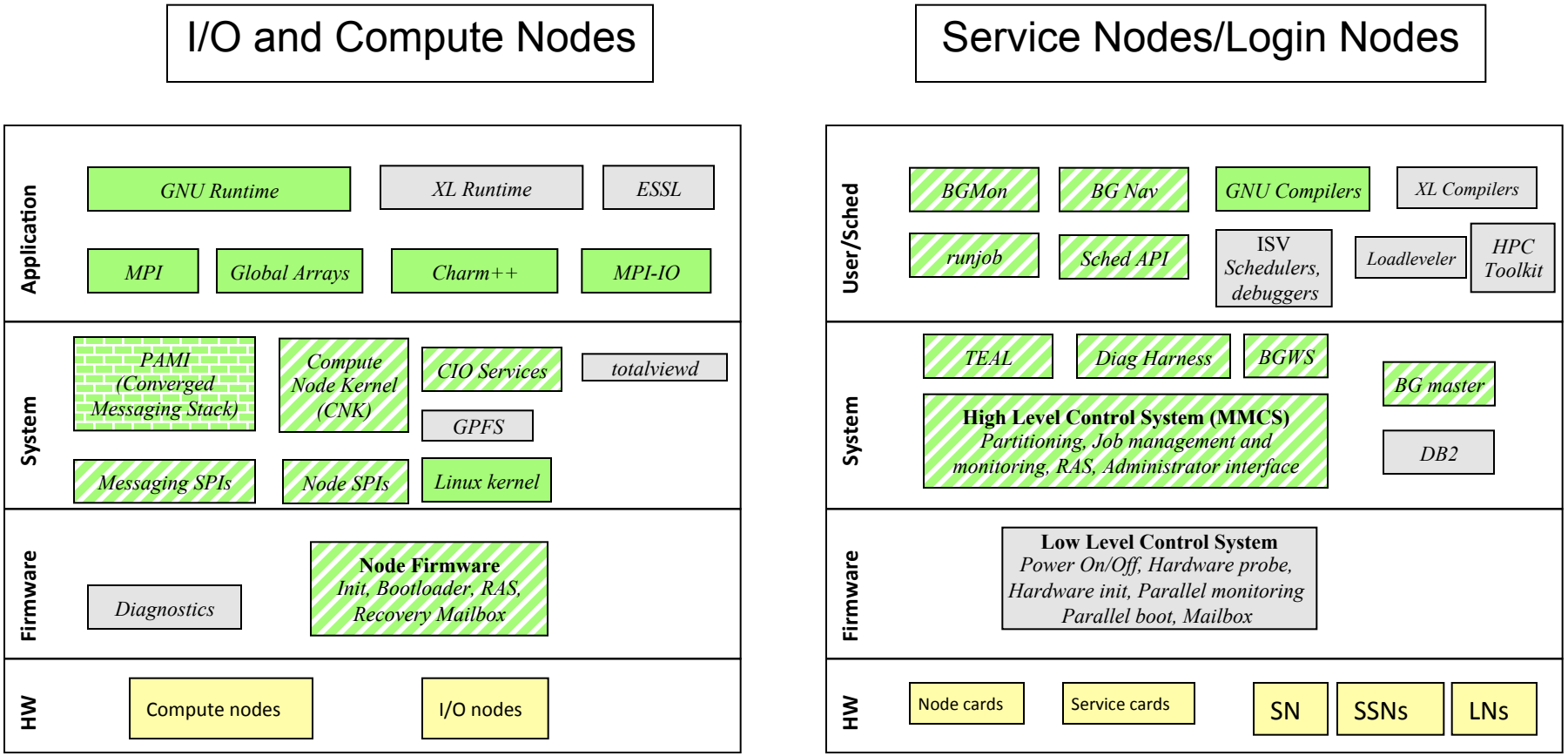
Blue Gene Q Software Innovations

- Standards-based programming environment
  - Linux™ development environment
    - Familiar GNU toolchain with glibc, pthreads, gdb
  - Red Hat on I/O node
  - XL Compilers C, C++, Fortran with OpenMP 3.1
  - Debuggers: Totalview
  - Tools: HPC Toolkit, PAPI, Dyinst, Valgrind, Open Speedshop
- Message Passing
  - Scalable MPICH2 providing MPI 2.2 with extreme message rate
  - Efficient intermediate (PAMI) and low-level (SPI) message libraries, documented, and open source
  - PAMI layer allows easy porting of runtimes like GA/ ARMCI, Berkeley UPC, etc,
- Compute Node Kernel (CNK) eliminates OS noise
  - File I/O offloaded to I/O nodes running full Linux
  - GLIBC environment with a few restrictions for scaling
- Flexible and fast job control – with high availability
  - Integrated HPC, HTC, MPMD, and sub-block jobs
  - Noise-free partitioned networks as in previous BG

- **New for Q**
  - Scalability Enhancements: the 17th Core
    - RAS Event handling and interrupt off-load
    - Event CIO Client Interface
    - Event Application Agents: privileged application processing
  - Wide variety of threading choices
  - Efficient support for mixed-mode programs
  - Support for shared memory programming paradigms
  - Scalable atomic instructions
  - Transactional Memory (TM)
  - Speculative Execution (SE)
  - Sub-blocks
  - Integrated HTC, HPC, MPMD, Sub-blocks
  - Integrated persistent memory
  - High availability for service nodes with job continuation
  - I/O nodes running Red Hat



BG/Q Software Stack Openness



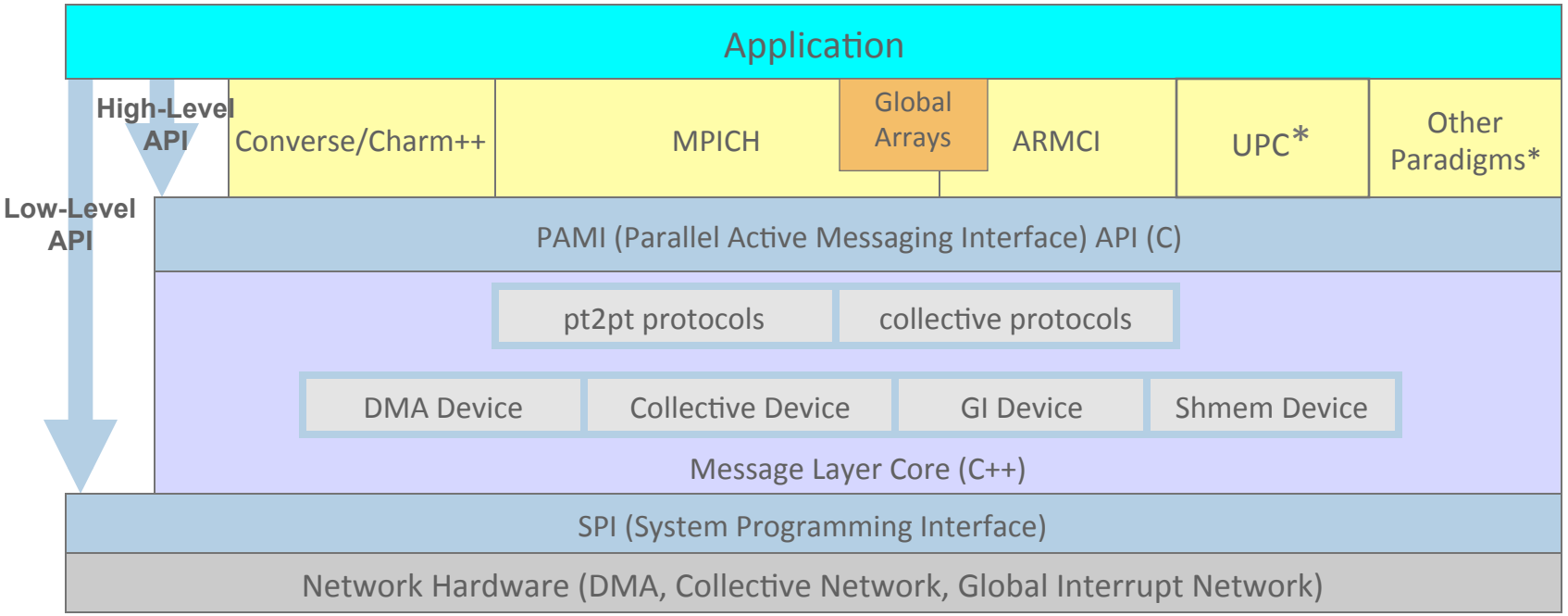
New open source reference implementation licensed under CPL.

New open source community under CPL license. Active IBM participation.

Existing open source communities under various licenses. BG code will be contributed and/or new sub-community started..

Closed. No source provided. Not buildable.

Parallel Active Message Interface



- **Message Layer Core has C++ message classes and other utilities to program the different network devices**
- **Support many programming paradigms**
- **PAMI runtime layer allows uniformity across IBM HPC platforms**

Overview of BG/Q: Another step forward

Design Parameters	BG/P	BG/Q	Improvement
Cores / Node	4	16	4x
Clock Speed (GHz)	0.85	1.6	1.9x
Flop / Clock / Core	4	8	2x
Nodes / Rack	1,024	1,024	--
RAM / core (GB)	0.5	1	2x
Flops / Node (GF)	13.6	204.8	15x
Mem. BW/Node (GB/sec)	13.6	42.6	3x
Latency (MPI zero-length, nearest-neighbor node)	2.6 μs	2.2 μs	~15% less
Bisection BW (32 racks)	1.39TB/s	13.1TB/s	9.42x
Network Interconnect	3D torus	5D torus	Smaller diameter
Concurrency / Rack	4,096	65,536	16x
GFlops/Watt	0.77	2.10	3x