

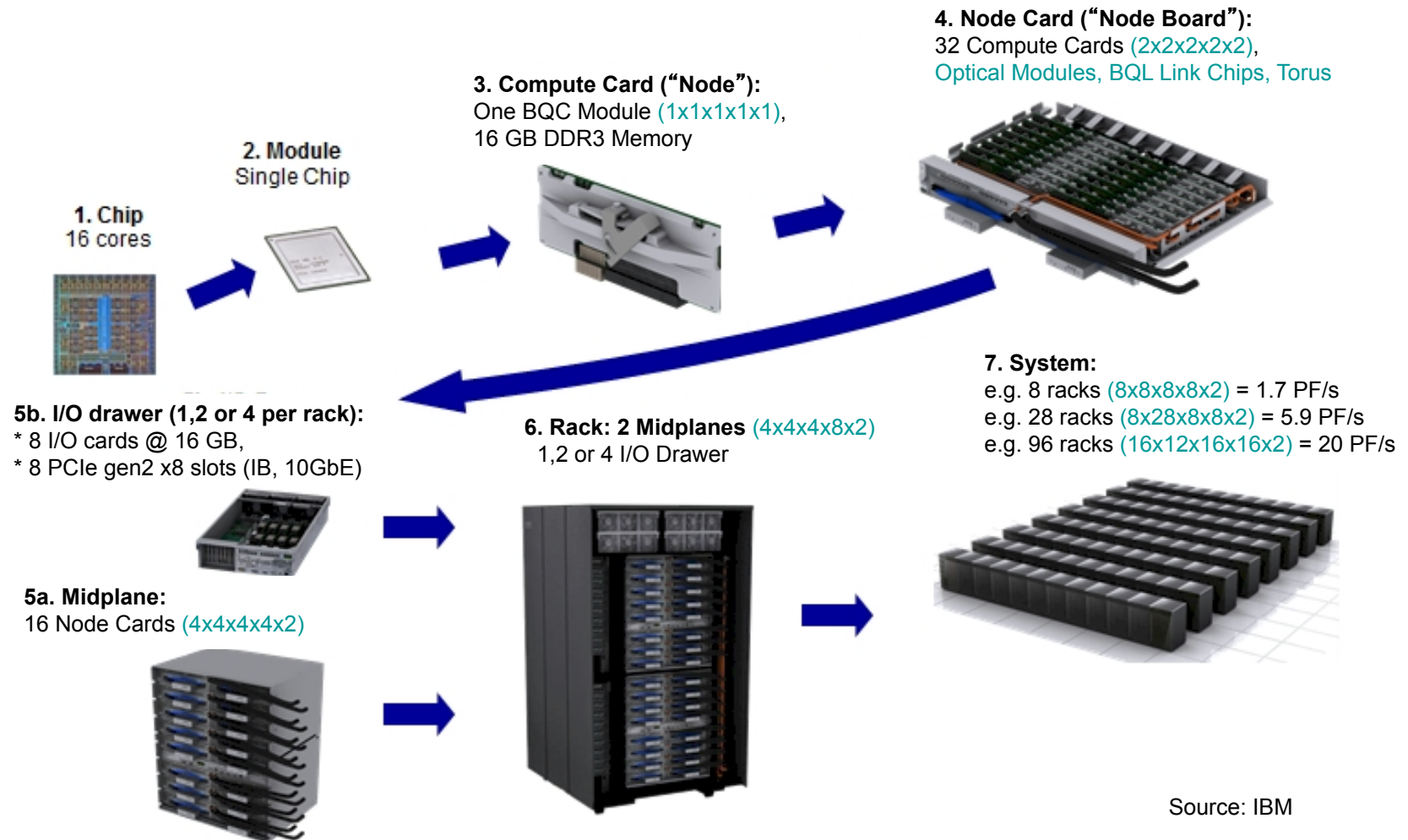
Blue Gene/Q – Hardware



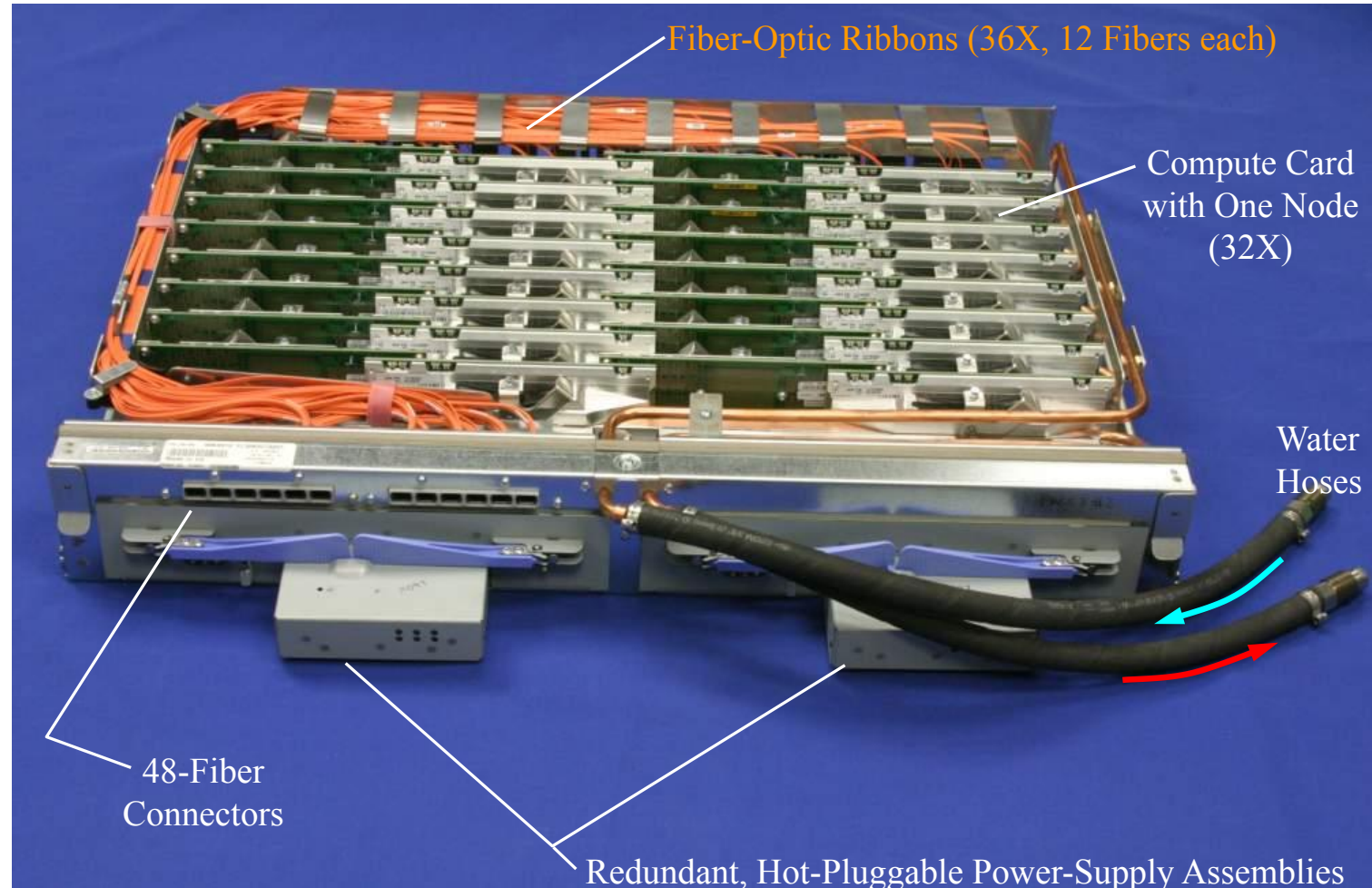
Blue Gene design goals

- System-on-Chip (SoC) design
 - *Processor comprises both processing cores and network*
- Optimal performance / watt ratio
- Small foot print
- Transparent high-speed reliable network
- Easy programming based on standard message passing interface (MPI)
- Extreme scalability ($> 1,5\text{Mi}$ cores)
- High reliability

Blue Gene/Q design

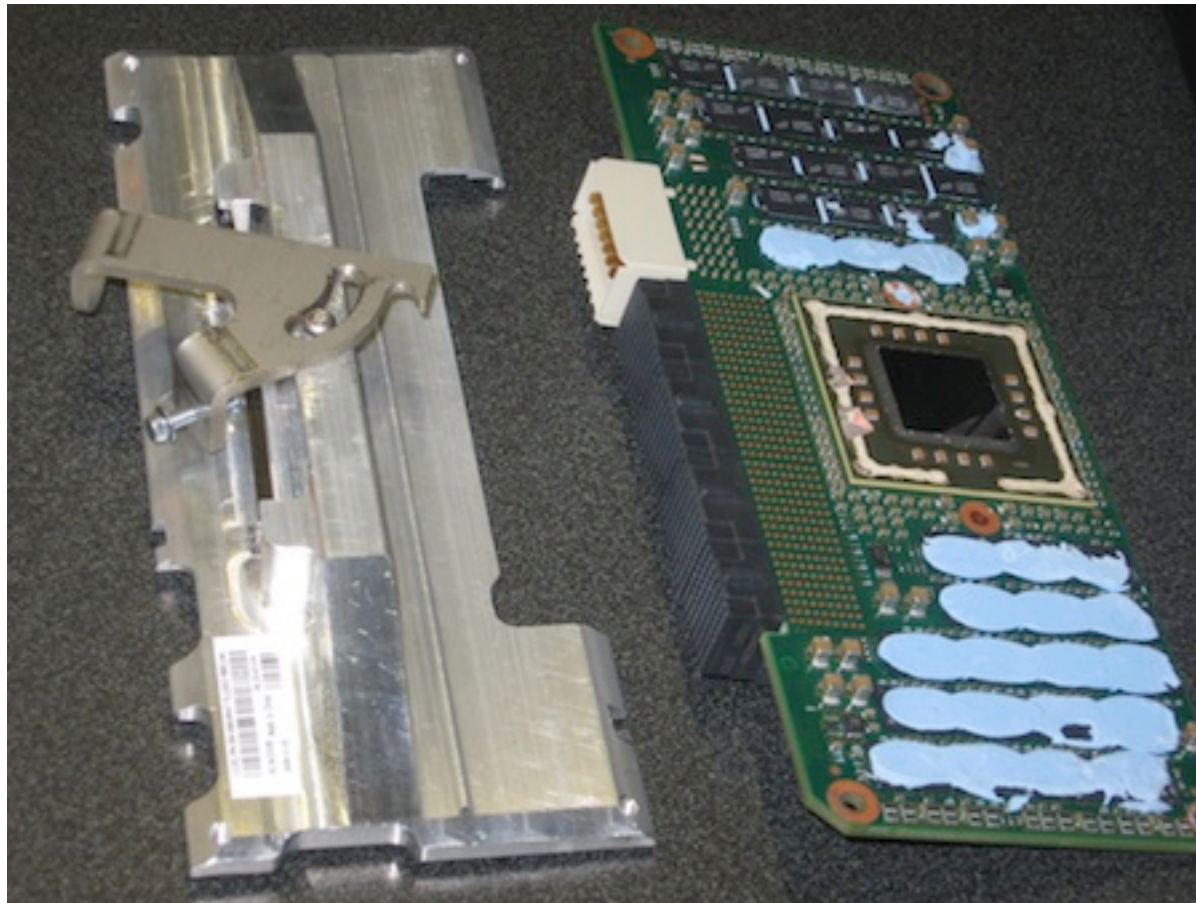


Blue Gene/Q node card



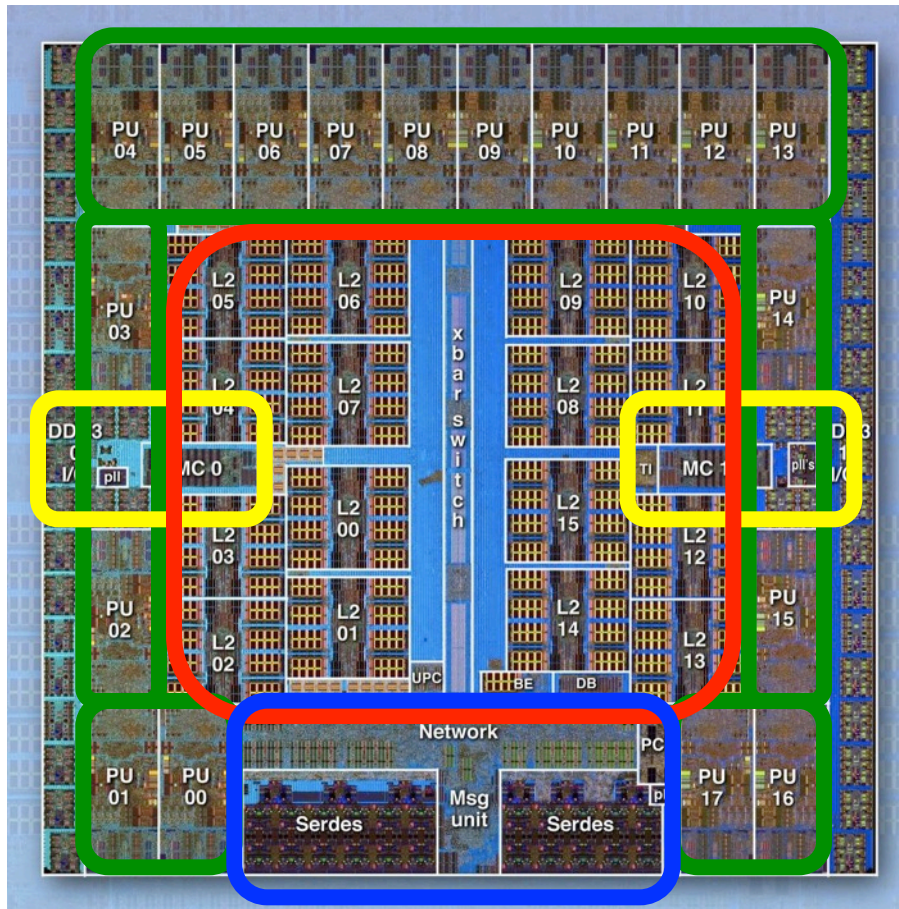
Source: IBM

Blue Gene/Q compute card



Source: Top500.org

Blue Gene/Q: Chip tomography



16+1+1 processing units

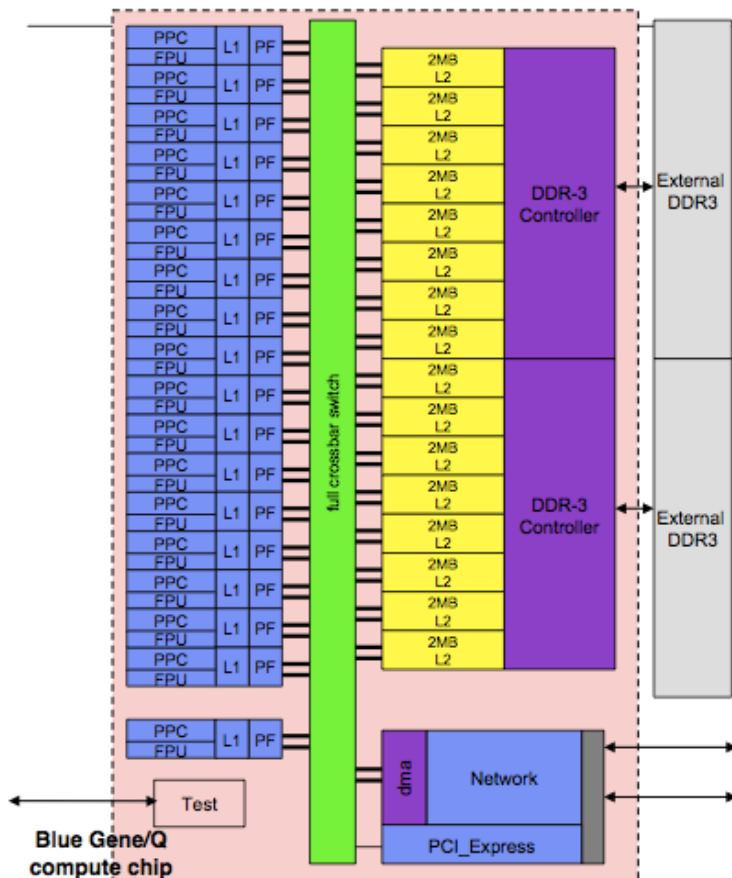
Two memory controller

L2 cache + crossbar switch

On-chip network

Source: IBM

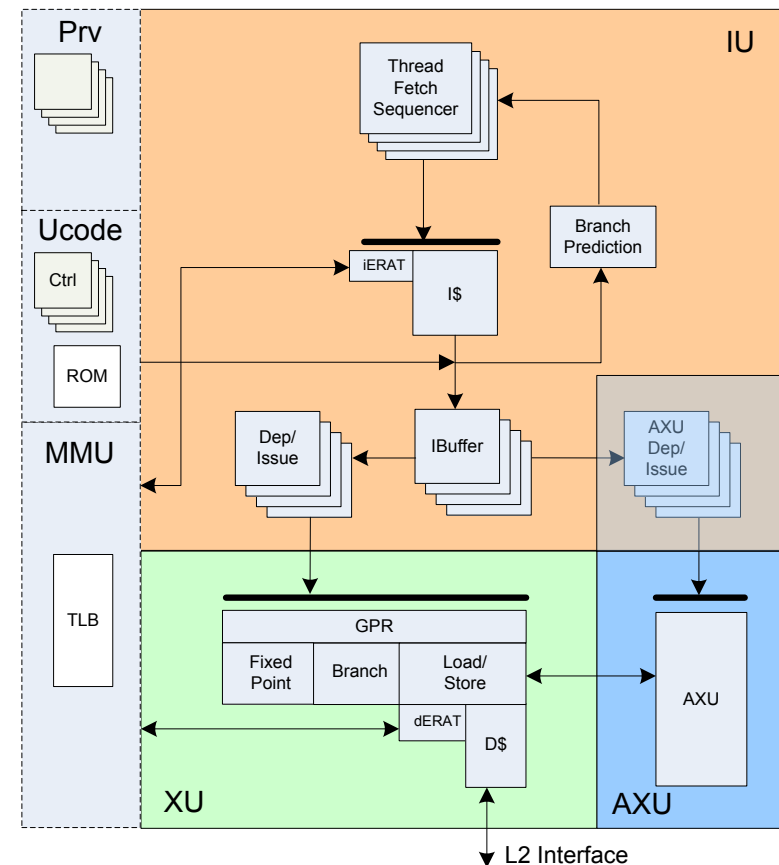
Blue Gene/Q chip architecture



- 16+1 core SMP @ 1.6 GHz
 - *Each core 4-way hardware threaded*
 - *2-way concurrent issue*
- Transactional memory and thread level speculation
- Quad floating point unit on each core
 - *204.8 GF peak node*
- 563 GB/s bisection bandwidth to shared L2
- 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 inter-processor links each at 2.0GB/s (5D-Torus)
- one I/O link at 2.0 GB/s
- 16 GB memory/node
- ~60 watts max chip power

Blue Gene/Q: A2 processor core

- Simple core
 - *designed for excellent power efficiency and small footprint*
- Embedded 64bit PowerPC compliant
- 4 SMT threads typically get a high level of utilization on shared resources
 - *Full register set for every thread*
- 1.6 GHz @ 0.74V
- AXU port allows unique BG/Q floating point unit
- One AXU (FPU) and one other instruction issue per cycle
- In-order execution



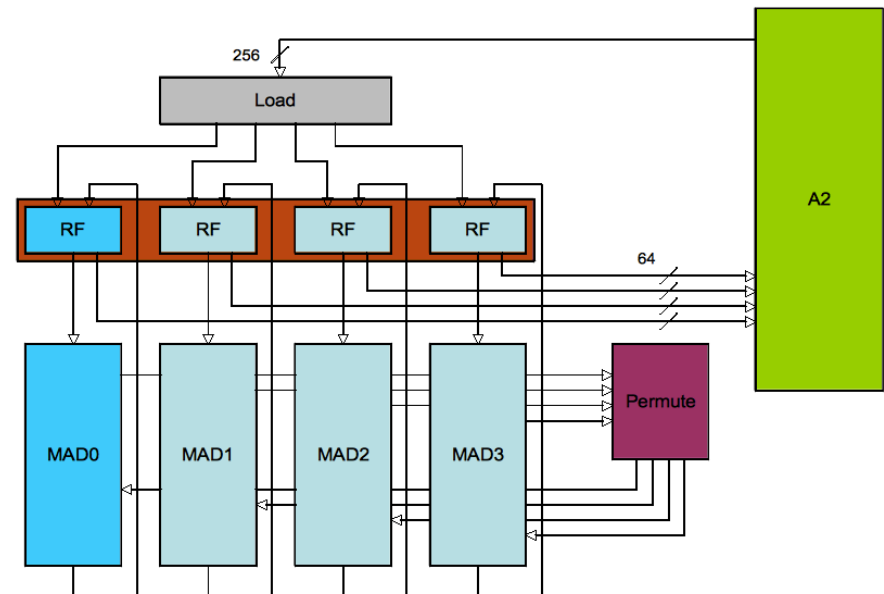
Source: IBM

Blue Gene/Q: 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*

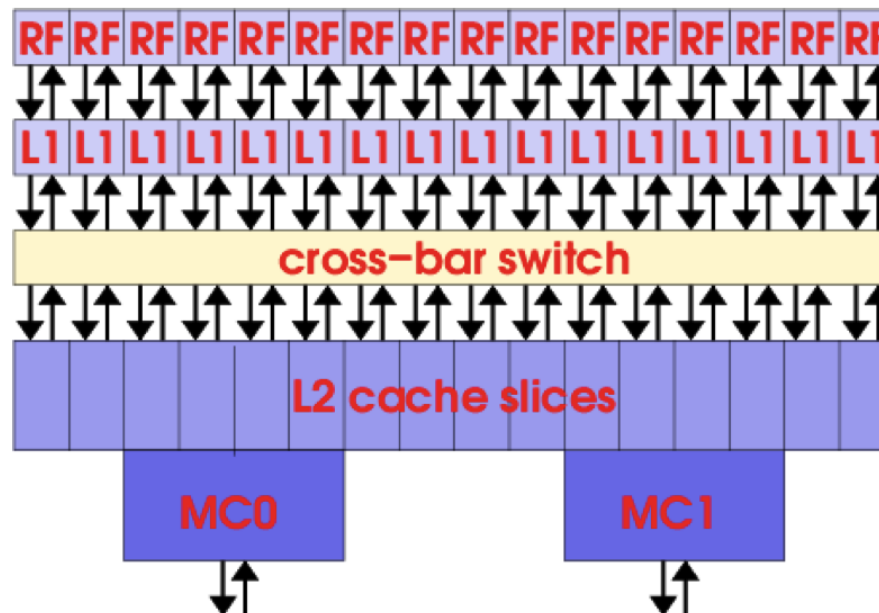
Quad floating Point eXtension unit (QPX)

- 4 double precision pipelines (64bit):
 - *scalar FPU*
 - *4-wide FPU SIMD*
 - *2-wide complex arithmetic SIMD*
- 32 x 4 x 256 bit registers
- Instruction extensions to PowerISA
- 8 concurrent floating point ops (FMA) + load + store
- Permute instructions to reorganize vector data
- Supports a multitude of data alignments
- Peak performance 4FMA / cycle
 - 12.8 GFlops @ 1.6 GHz



Memory hierarchy

- L2 cache
 - 32 MBytes
 - Organised in 16 slices
 - 16-way associative
- External memory
 - 16 GBytes DDR3
 - 2 memory controllers



- L1 cache
 - 16 kB
 - 8-way associative

563 GByte/s
bi-section

42.6 GByte/s

Blue Gene/Q: Advanced processor features

L1 pre-fetching engines

- *Stream pre-fetching*
 - *1 engine per core*
- *“Perfect” (list) pre-fetching*
 - *4 engines per core*
 - *Hardware memorizes access sequence*

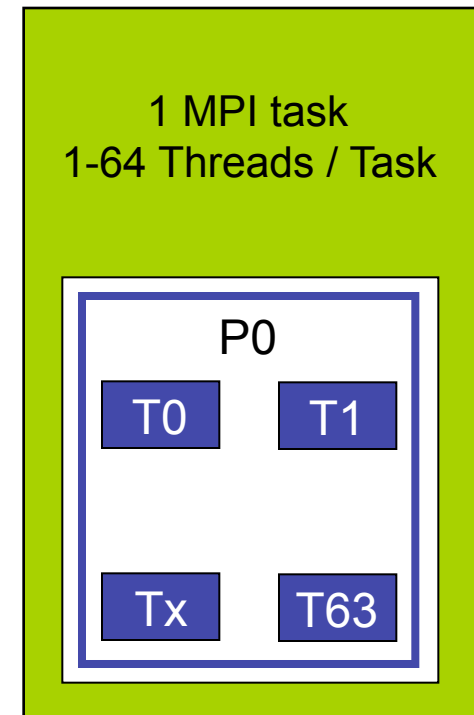
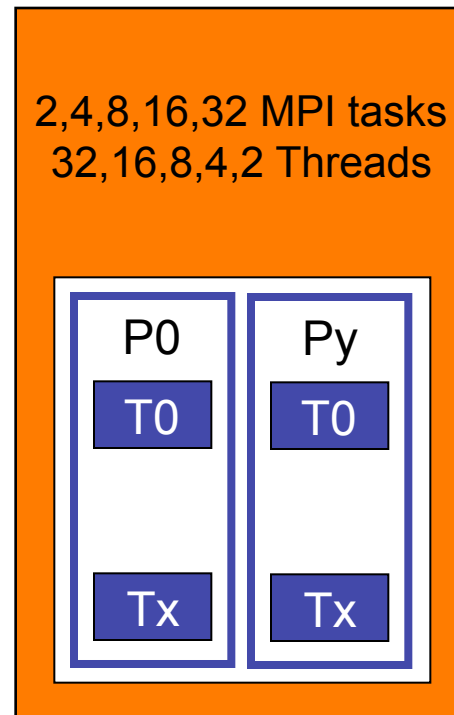
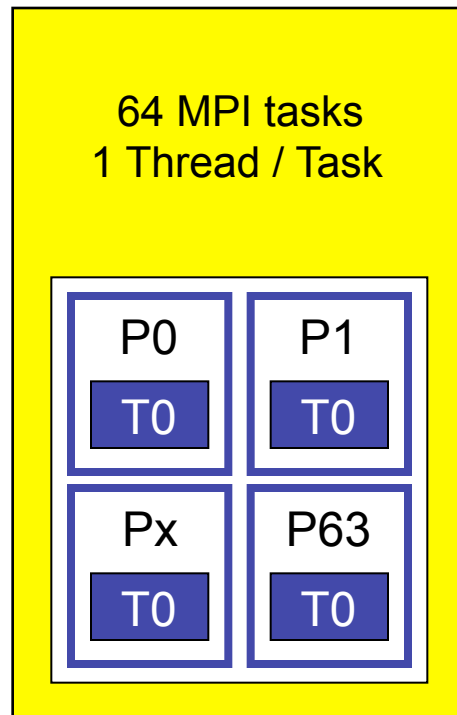
Multi-versioning L2 cache

- *Cache can track state changes caused by speculative threads*
- *At the end of speculative code: invalidate or commit and/or react with software notification*
- *Use: transactional memory, thread-level speculation*

Blue Gene/Q ↔ Blue Gene/P

Property		Blue Gene/Q	Blue Gene/P
Node Properties	Node Processors Processor Frequency Coherency Cache size (shared) Main Store Main Store Bandwidth (1:2 pclk) Peak Performance	16*4 PowerPC® A2 1.6GHz SMP 32MB 16GB 42.6GB/s 204.8GF/node	4* 450 PowerPC® 0.85GHz SMP 8MB 2GB 13.6 GB/s 13.9 GF/node
Torus Network	Topology Bandwidth Hardware Latency (Nearest Neighbour) Hardware Latency (Worst Case)	5D 10*2*2GB/s=40GB/s 40ns (32B packet) 300?ns (512B packet) 2.6µs (31 hops)	3D 6*2*425MB/s=5.1GB/s 100ns (32B packet) 800ns (256B packet) 3.2µs(64 hops)
Tree Network	Bandwidth Hardware Latency (worst case)	6.5µs	2*0.85GB/s=1.7GB/s 3.5µs
System Properties	Area Peak Performance Total Power	~20m ² 1.7PF ~400kW	160m ² 1.008PF ~2.3MW

Execution Modes in BG/Q



Blue Gene/Q chip: 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

Blue Gene/Q: New Network architecture

- 11 bi-directional chip-to-chip links
 - *2 GB/s bandwidth, about 40 ns latency*
- 5-dimensional torus topology
 - *Dimension E limited to length 2*
- Why d -dimensional torus with large d ?
 - *High bi-section bandwidth*
 - *Flexible partitioning in lower dimensions*
- Deterministic/dynamic routing support
- 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*



Source: IBM

Blue Gene/Q: Torus configuration

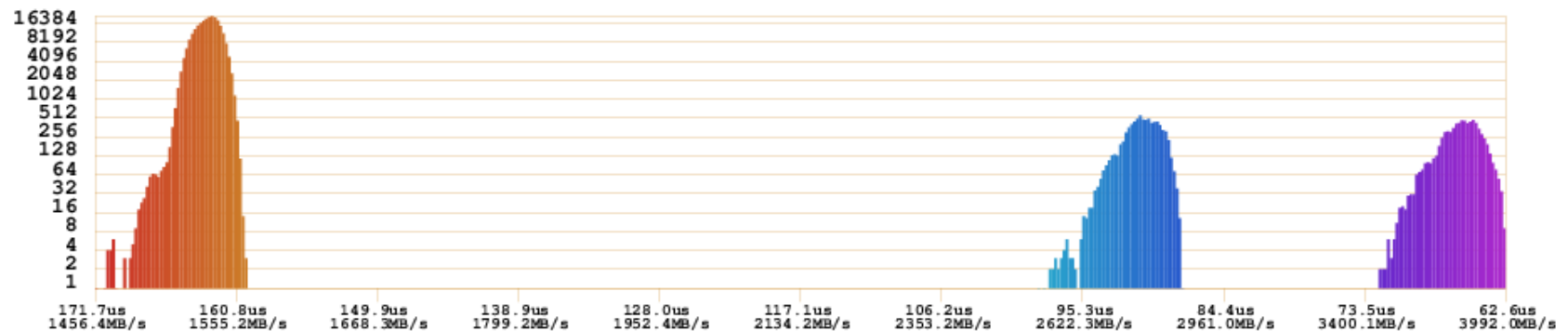
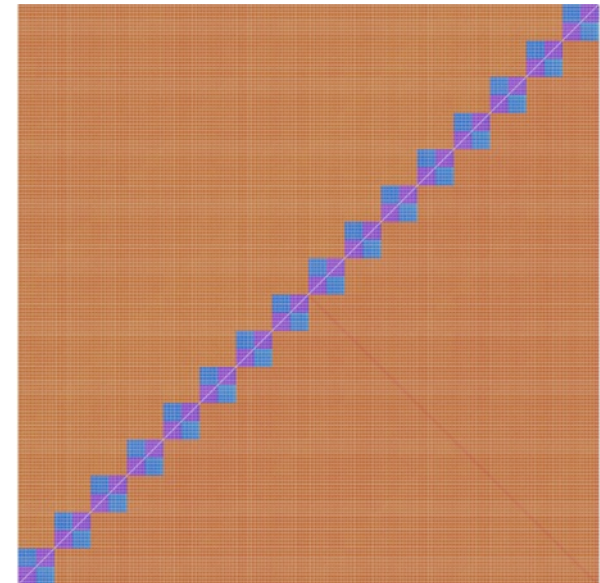
- BG/Q nodes form a 5D torus
 - *Nodecards: 2x2x2x2x2*
 - *Midplanes: 4x4x4x4x2*
 - Dimensions A-D are cabled to other midplanes
 - *5th dimension stays within the nodecard*
 - *6th dimension is the CPU# within a node*
 - *Dimension labels: ABCDE T*
- 28 Rack configuration
 - *7 rows, 4 columns*
 - *Torus size in nodes: 8x28x8x8x2*
 - *Torus size in midplanes: 2x7x2x2*

Torus node to MPI mapping

- Block's physical network topology: 6 dimensional ABCDE T
 - *local on each node; shared memory communication (fast); $T=0..63$*
 - *A,B,C,D,E depend on size of block, e.g. $2 \times 2 \times 4 \times 4 \times 2$ for quarter midplane (256 nodes)*
- processes need to be placed to minimize maximum load on network-links
 - *take advantage of logical decomposition of work*
 - *take advantage of link in E direction: double bandwidth available*
 - *e.g. domain decomposition: find best match of physical and logical dimensions*
- Three options to define mapping:
 - *By permutation of physical dimensions (--mapping DCTEBA)*
 - *Use MPI_Cart_Create() (to date: only reordering of dimensions)*
 - *By a file with a line for each process, specifying its physical position (--mapping <filename>)*

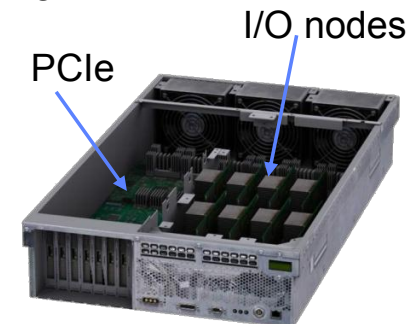
Linktest: Blue Gene torus link bandwidth tester

- All-to-all ping-pong test
- Bandwidth distribution
 - *Intra-node communication*
 - *Communication via link A, B, C, D*
 - *Communication via link E*



Blue Gene/Q: I/O architecture

- **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



JUQUEEN Configuration (01/2013)

28 Racks Blue Gene/Q

- 28672 Compute Nodes (16 cores, 16 GB memory)
- 458752 cores / 1.8M threads
- 5.88 PFlop/s peak performance
- 248 I/O nodes (10GigE) ← (1x32 + 27x8)
- 2.3 MW power consumption (10-80 kW per rack)

4 Frontend Nodes (**user login**) + 2 Service Nodes (system, database)

- IBM p7 740, 8 cores (3.55 GHz), 128 GB memory
- local storage device DSC3700 (16 TB)
 - */dbhome + /bgsys*

JuQueen Environment

