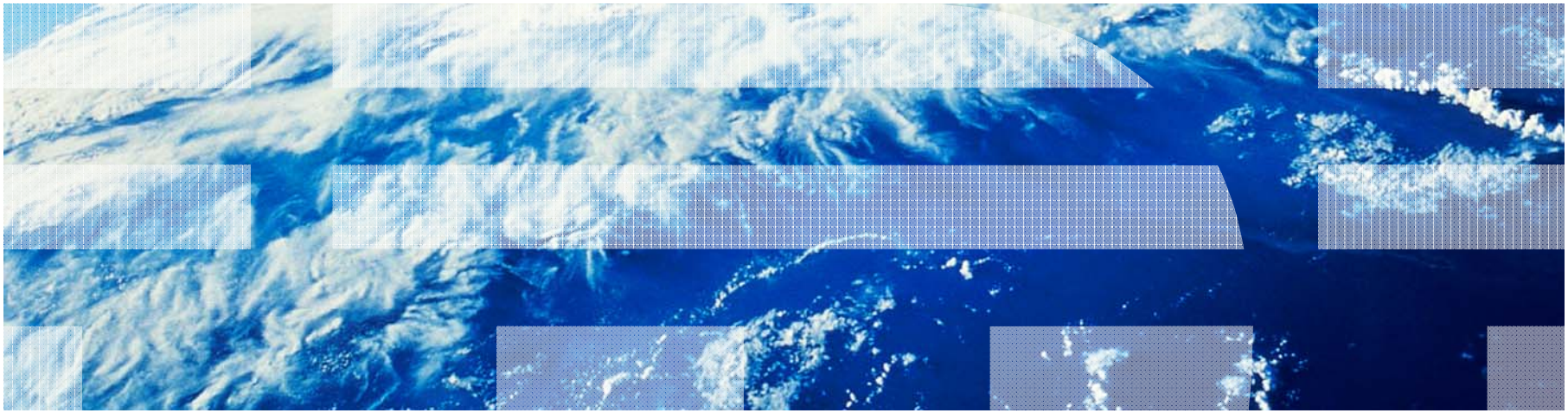
The Blue Gene/Q Compute Chip

Ruud Haring / IBM BlueGene Team



Acknowledgements

The IBM Blue Gene/Q development teams are located in
- Yorktown Heights NY,
- Rochester MN,
- Hopewell Jct NY,
- Burlington VT,
- Austin TX,
- Bromont QC,
- Toronto ON,
- San Jose CA,
- Boeblingen (FRG),
- Haifa (Israel),
- Hursley (UK).

- Columbia University
- University of Edinburgh

- The Blue Gene/Q project has been supported and partially funded by Argonne National Laboratory and the Lawrence Livermore National Laboratory on behalf of the United States Department of Energy, under Lawrence Livermore National Laboratory Subcontract No. B554331

Blue Gene/Q

- Blue Gene/Q is the third generation of IBM's Blue Gene family of supercomputers
  - Blue Gene/L was announced in 2004
  - Blue Gene/P was announced in 2007

- Blue Gene/Q
  - was announced in 2011

  - is currently the fastest supercomputer in the world
    - June 2012 Top500: #1,3,7,8, … 15 machines in top100 (+ 101-103)

  - is currently the most power efficient supercomputer architecture in the world
    - June 2012 Green500: top 20 places

  - also shines at data-intensive computing
    - June 2012 Graph500: top 2 places -- by a 7x margin

  - is relatively easy to program -- for a massively parallel computer – and its FLOPS are actually usable
    - this is **not** a GPGPU design …

  - incorporates innovations that enhance multi-core / multi-threaded computing
    - in-memory atomic operations
    - 1st commercial processor to support hardware transactional memory / speculative execution
    - …

  - is just a cool chip (and system) design
    - pushing state-of-the-art ASIC design where it counts
    - innovative power and cooling

Blue Gene system objectives

- **Massively parallel supercomputing systems**
  - Focusing on large scale scientific applications
  - … but broadening beyond (graph analysis, business analytics, …)
  - Laying groundwork for Exascale computing

- **Reduce total cost of ownership**
  - Power efficient chips
    - allow dense packaging
    - drive floor space efficiency
    - drive cost efficiency
  - Reliability
    - Long MTBF for large installations

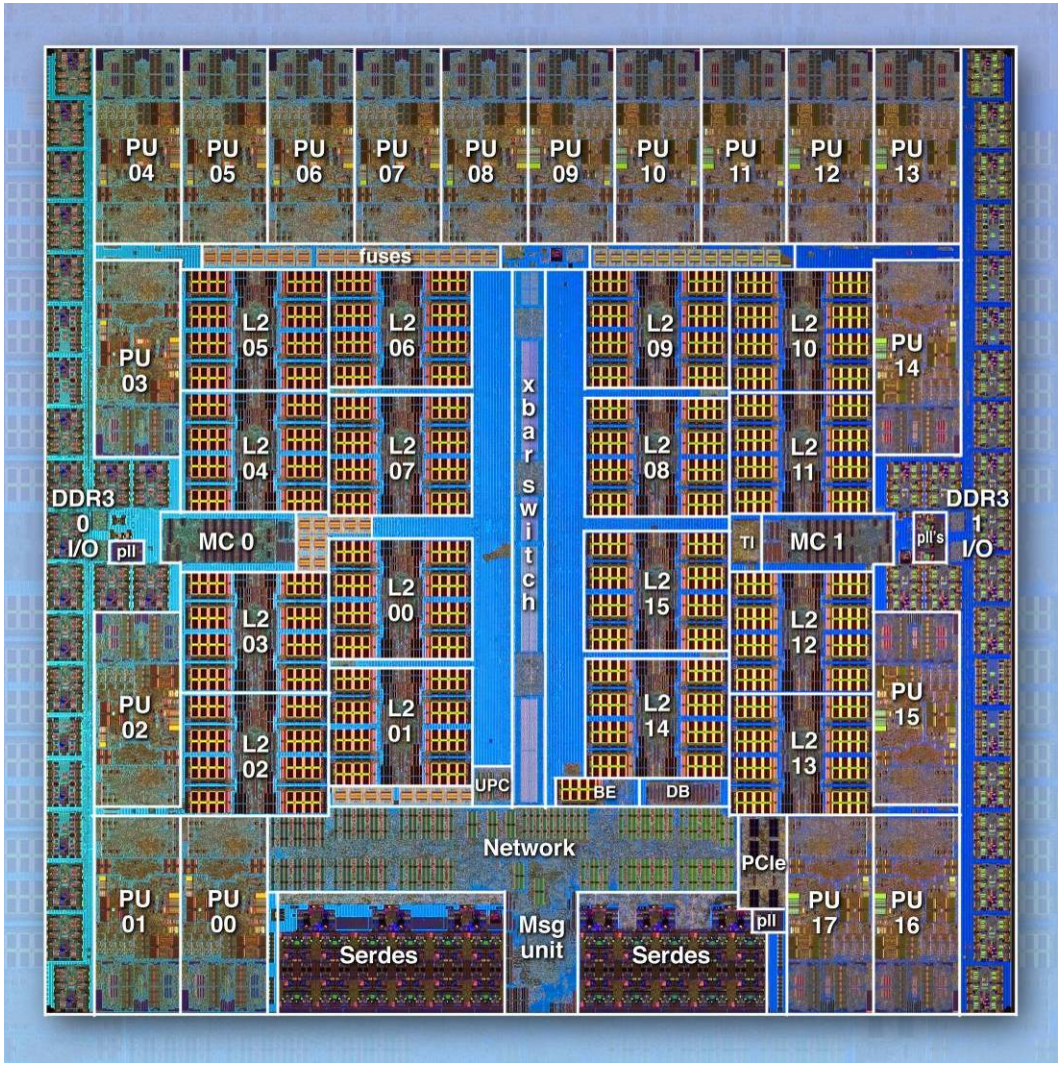
Chip design objectives:

← optimize FLOPS/Watt

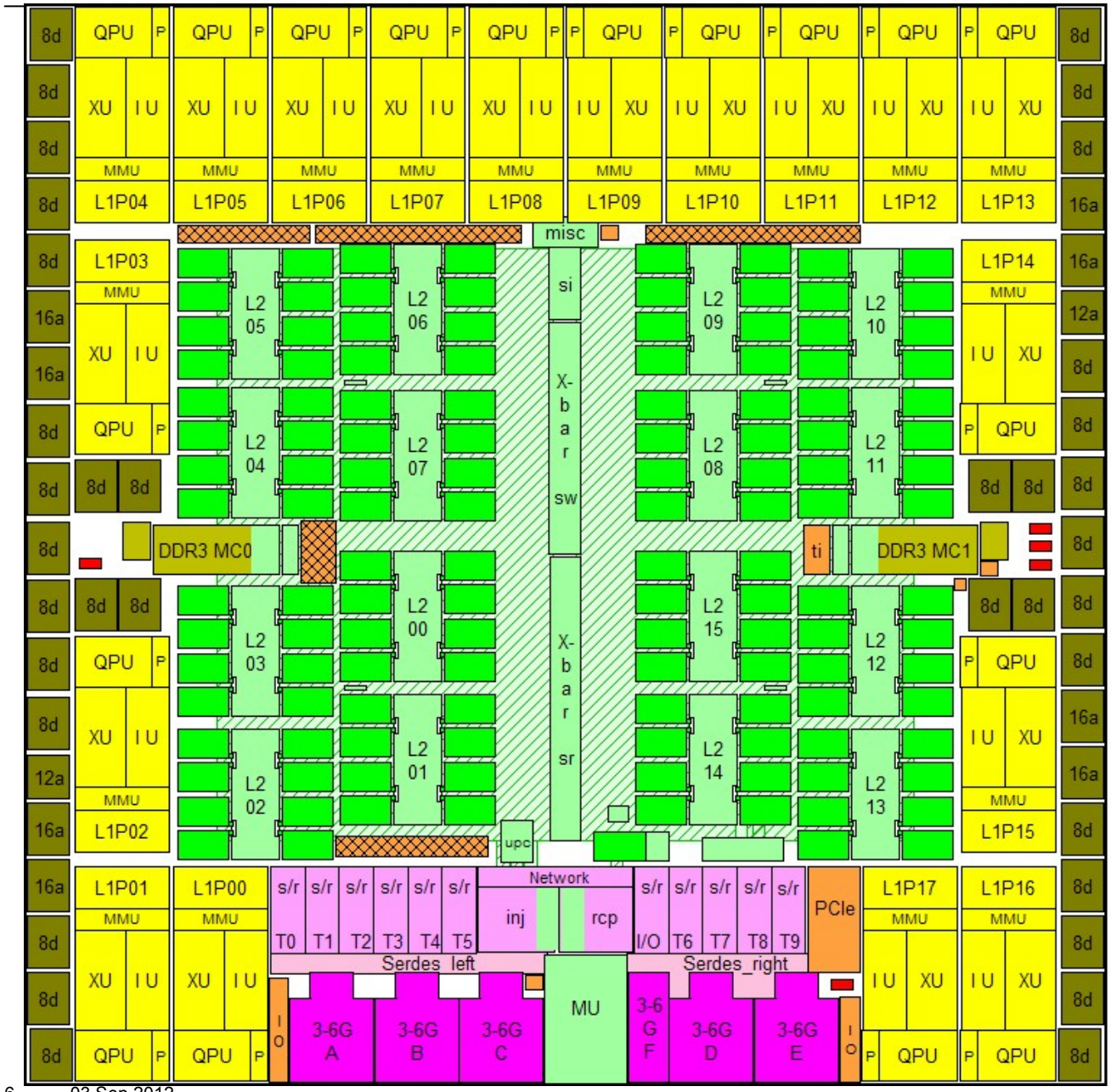
← optimize redundancy / ECC usage / SER sensitivity

BlueGene/Q Compute chip

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

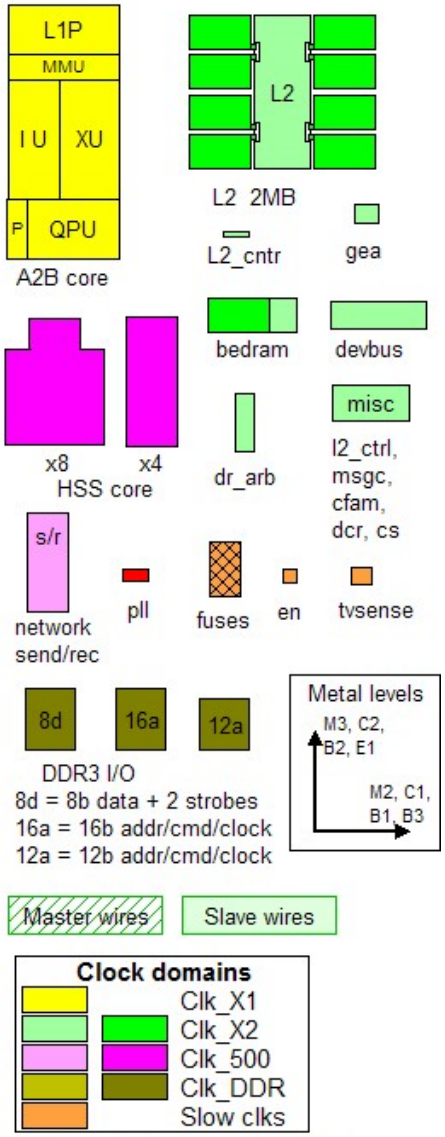


- **360 mm² Cu-45 technology (SOI)**
  - ~ 0.8V
  - ~ 1.47 B transistors
  - 11 metal layers
- **16 user + 1 service processors**
  - plus 1 redundant processor
  - all processors are symmetric
  - each 4-way multi-threaded
  - 64 bits PowerISA™
  - 1.6 GHz
  - L1 I/D cache = 16kB/16kB
  - L1 prefetch engines
  - each processor has Quad FPU (4-wide double precision, SIMD)
  - peak performance 204.8 GFLOPS @ 55 W
- **Central shared L2 cache: 32 MB**
  - eDRAM
  - multiversioned cache – supports transactional memory, speculative execution, rollback.
  - supports atomic ops
- **Dual memory controller**
  - 8–16 GB external DDR3 memory
  - 1.33 Gb/s
  - 2 * 16 byte-wide interface (+ECC)
- **Chip-to-chip networking**
  - 5D Torus topology + external link
    → 5 x 2 + 1 high speed serial links
  - each 2 GB/s send + 2 GB/s receive
  - Router logic integrated into BQC chip
  - DMA, remote put/get, collective operations
- **External (file) IO -- when used as IO chip.**
  - PCIe Gen2 x8 interface (4 GB/s Tx + 4 GB/s Rx)
  - re-uses 2 serial links
  - interface to Ethernet or Infiniband cards



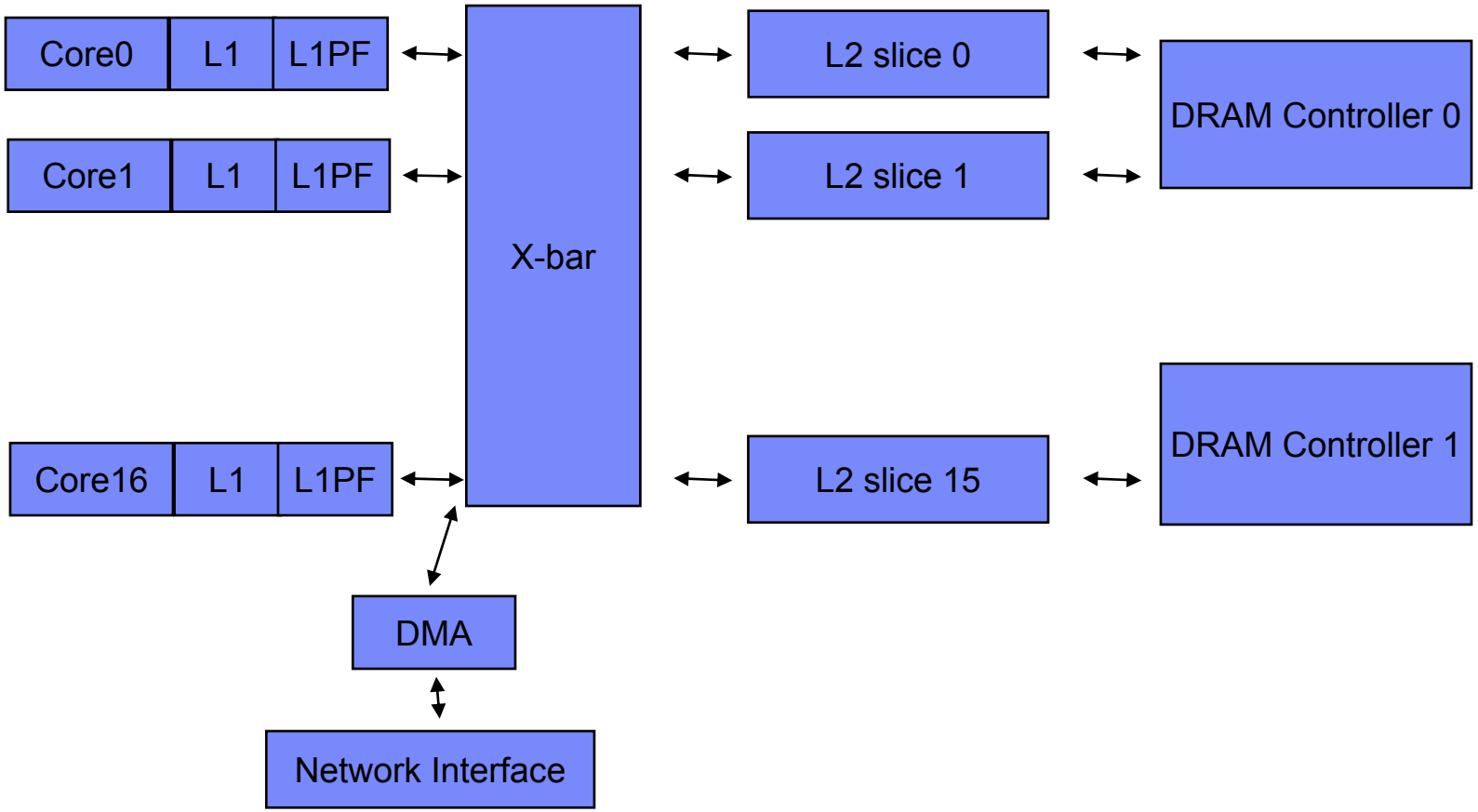
BlueGene/Q Compute Chip

chip size = 18.96 x 18.96



"balls up" view (looking at circuit side of chip)

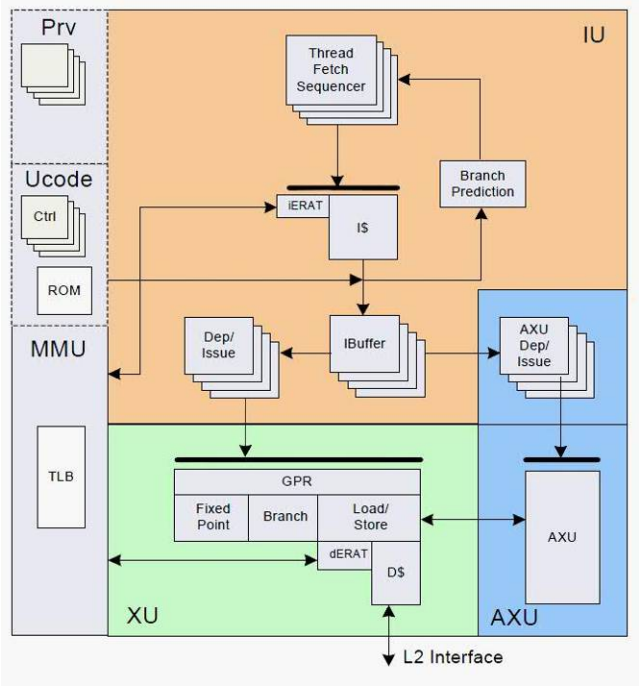
BG/Q Memory Structure

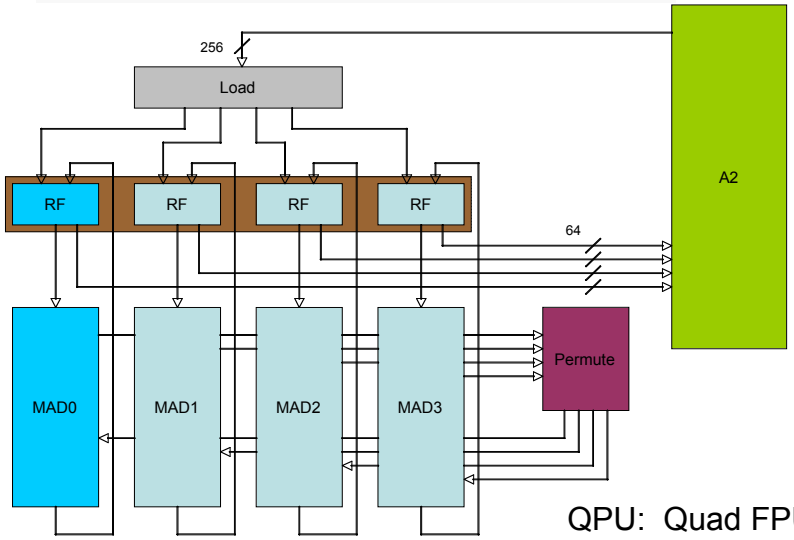


BG/Q Processor Unit

- **A2 processor core**
  - Mostly same design as in PowerEN™ chip
  - Implements 64-bit PowerISA™
  - Optimized for aggregate throughput:
    - 4-way simultaneously multi-threaded (SMT)
    - 2-way concurrent issue 1 XU (br/int/l/s) + 1 FPU
    - in-order dispatch, execution, completion
  - L1 I/D cache = 16kB/16kB
  - 32x4x64-bit GPR
  - Dynamic branch prediction
  - 1.6 GHz @ 0.8V

- **Quad FPU**
  - 4 double precision pipelines, usable as:
    - scalar FPU
    - 4-wide FPU SIMD
    - 2-wide complex arithmetic SIMD
  - Instruction extensions to PowerISA
  - 6 stage pipeline
  - 2W4R register file (2 * 2W2R) per pipe
  - 8 concurrent floating point ops (FMA)
    + load + store
  - Permute instructions to reorganize vector data
    - supports a multitude of data alignments

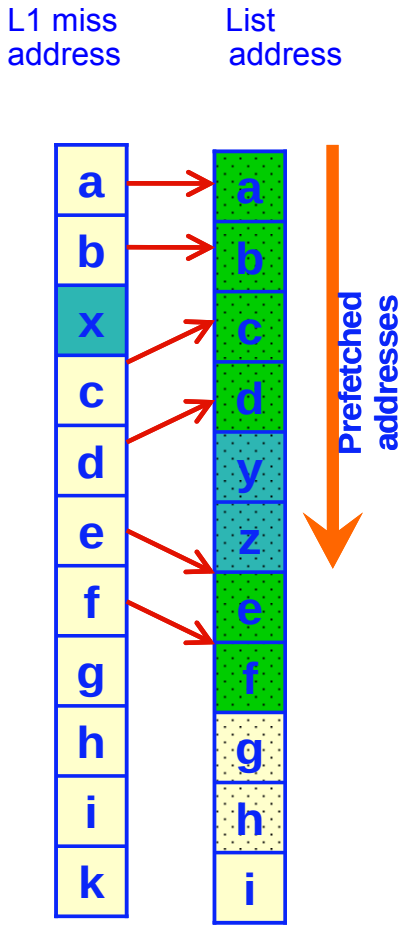




QPU: Quad FPU

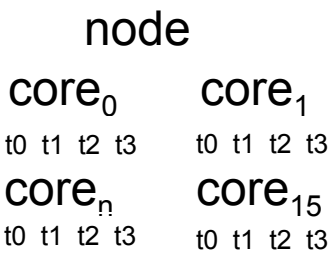
BlueGene/Q PUnit – ct.

- **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)

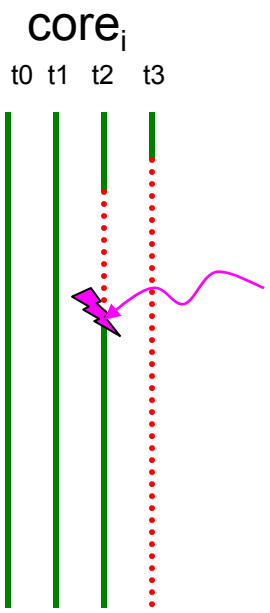


List-based “perfect” prefetching has tolerance for missing or extra cache misses

Wakeup Unit



- Allow SMT hardware threads to be suspended, while waiting for an event
- Lighter weight than wake-up-on-interrupt – no context switching, avoids software polling
- Improves power efficiency and resource utilization



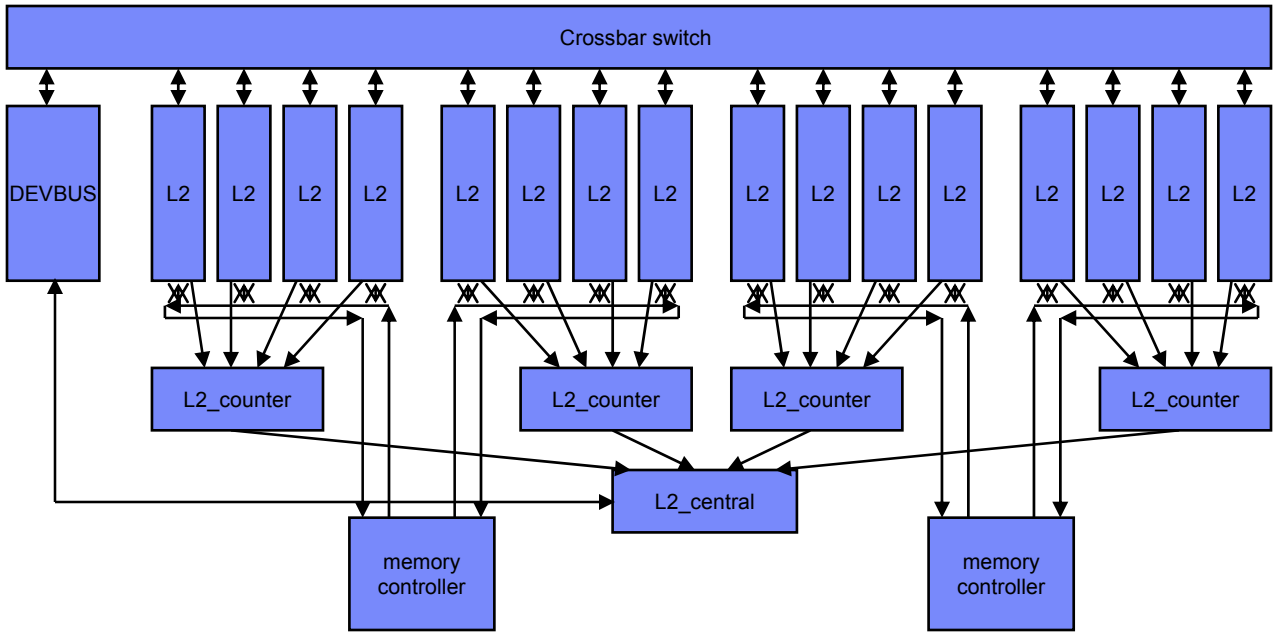
IPI (Inter Processor Interrupt)
MUI (Messaging Unit Interrupt)
L2I (L2 Invalidation)

Crossbar switch

- Central connection structure between
  - PUnits (L1-prefetchers)
  - L2 cache
  - Networking logic
  - Various low-bandwidth units
- Half frequency (800 MHz) clock grid
- 3 separate switches:
  - Request traffic -- write bandwidth 12B/PUnit @ 800 MHz (under simultaneous reads)
  - Response traffic -- read bandwidth 32B/PUnit @ 800 MHz
  - Invalidate traffic
- 22 master ports
  - PUnits
  - DevBus master -- PCIe
  - Network logic ports -- Remote DMA
- 18 slave ports
  - 16 L2 slices
  - DevBus slave -- PCIe, boot / messaging RAM, performance counters, …
  - Network logic -- injection, reception
- Peak on-chip bisection bandwidth 563 GB/s

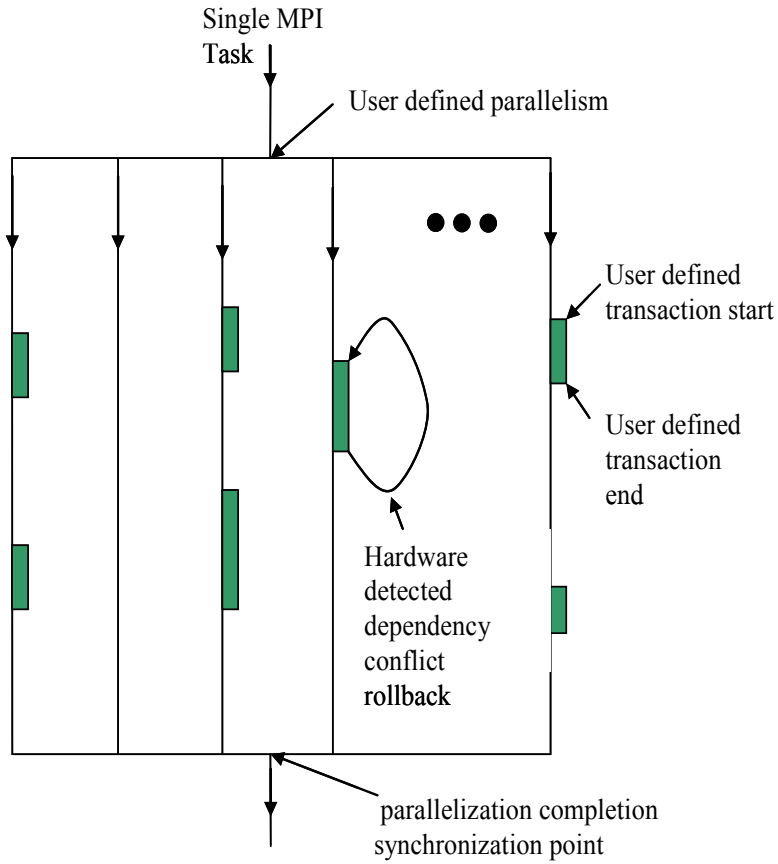
L2 structure

- **32 MB / 16 way set-associative / 128B line size**
- **Point of coherency**
- **Organization:**
  - 16 slices @ 2MB each
  - Each slice contains 8 * 2.25 Mb eDRAMs (data+ECC) plus directory SRAMs, buffers, control logic.
  - Addresses are hashed across L2 slices



Multi-versioned L2 cache

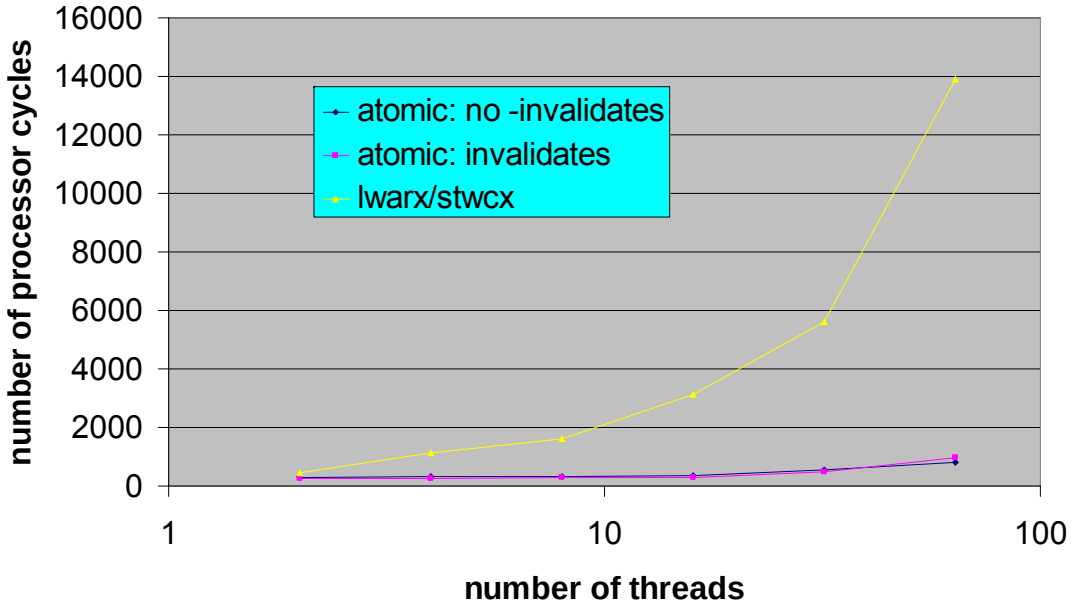
- **Multi-versioned cache**
  - Uses the “ways” of the L2 cache to store different versions of data for same address
  - 15 out of 16 ways can contain speculative content
  - Data tagged -- tags tracked by score board
  - Designed for Transactional Memory:
    - Guarantees “atomicity” of user-defined transactions
    - Eliminates need for locks
    - L2 detects and reports load/store conflicts
      -- software will need to resolve



  Speculative Execution:
    - Allows coarse grain multi-threading for long sequential code sections with (potential) data dependencies
    - L2 detects and resolves load/store conflicts according to sequential semantics
      -- software will need to re-run invalidated segment

L2 cache -- continued

- **Atomic operations**
  - Fast and scalable alternative to classic locking schemes
  - From processor perspective: load/store to aliased memory (unused physical address bit)
    - can be invoked on any 8-byte word in memory
  - 16 operations: e.g. fetch-and-increment, store-add, store-XOR, etc.
  - Some operations access multiple adjacent locations, e.g., fetch-and-increment-bounded
  - Low latency even under high contention
    - avoids L2-to-PU roundtrip cycles of lwarx/stwcx -- "queue locking"
  - Accelerates s/w operations: locking, barriers
  - Facilitates efficient work queue management, with multiple producers and consumers
  - Scales to 64 threads, enables efficient inter-core messaging



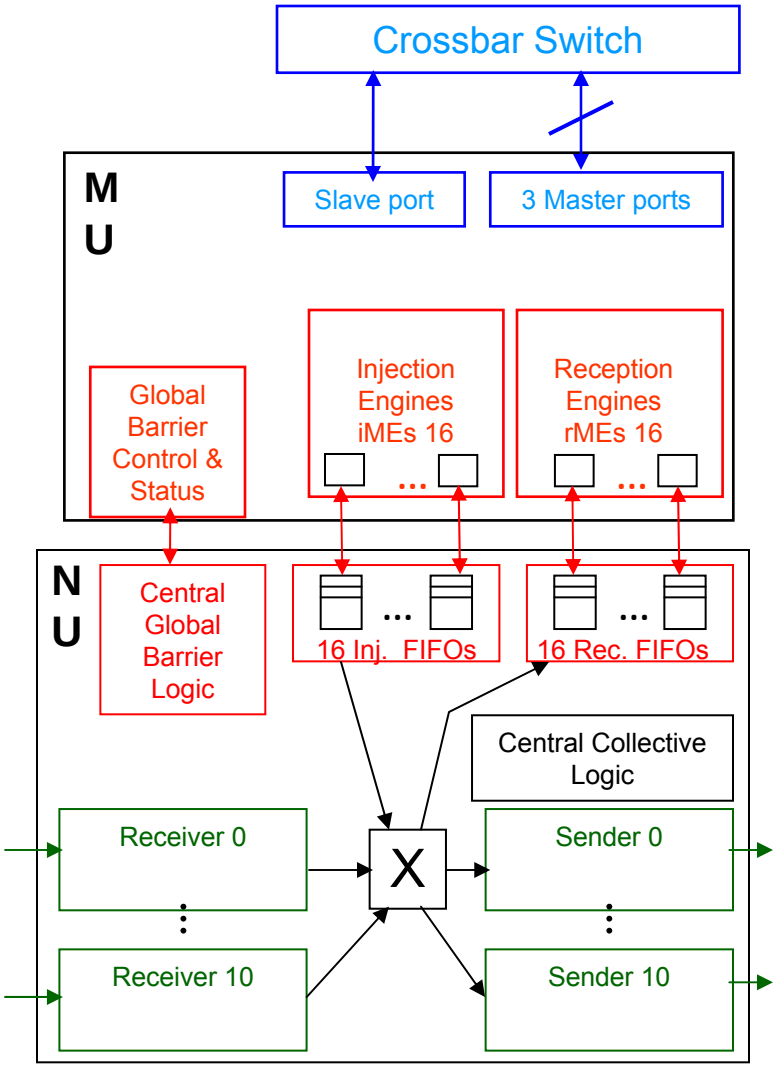
DDR3 memory interface

- L2 cache misses are handled by dual on-chip DDR3 memory controllers
  - each memory controller interfaces with 8 L2 slices
- Interface width to external DDR3 is 2 * (16B + ECC)
  - aggregate peak bandwidth is 42.7 GB/s for DDR3-1333.
- Multiple density/rank/speed configurations supported
  - currently configured with 16GB DDR3-1333 -- 72 * 2 Gb DDR3 chips driven by each BQC chip.
  - soldered onto same card
    - eliminates connector reliability issues
    - reduces driver and termination power
- Extensive ECC capability on 64B basis:
  - Double symbol error correct / triple detect (symbol = 2bits * 4 beats)
  - Retry
  - Partial or full chip kill
- DDR3 PHY
  - integrated IO blocks: 8bit data + strobe; 12 /16 bit address/command
  - integrates IO with delay lines (deskew), calibration, impedance tuning, …

Network and Messaging Unit

- Designed to permit highly efficient messaging
- Point-point, Collective, Barrier packets all flow over same network
- Functionality of network and NIC integrated onto single chip with processors
  - Only 8% of chip, including IO cells
- Message Unit: on-chip NIC/DMA
- Network Unit: on-chip router
- Tight integration between the network and MU
  - Fifo interface
- Low overhead SPI programming interface
- Hardware support for MPI collective operations: reduce and all-reduce
  - Integer add, min, max
  - Bit operations
  - Floating point add, min, max
  - Overflows, NaNs flagged by hardware

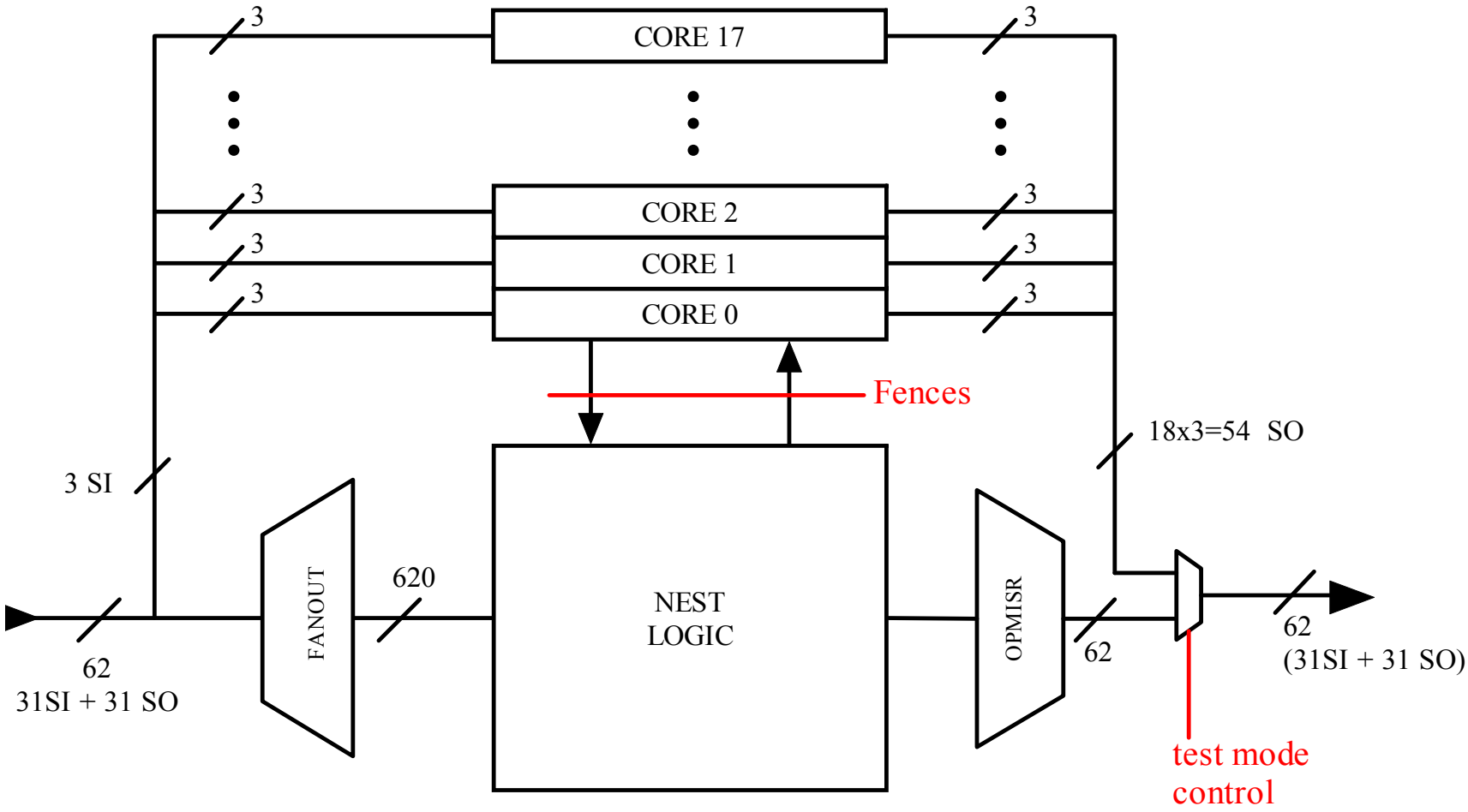
  → Single pass floating point reductions at near link bandwidth – without impacting PUnits



The 17th Core

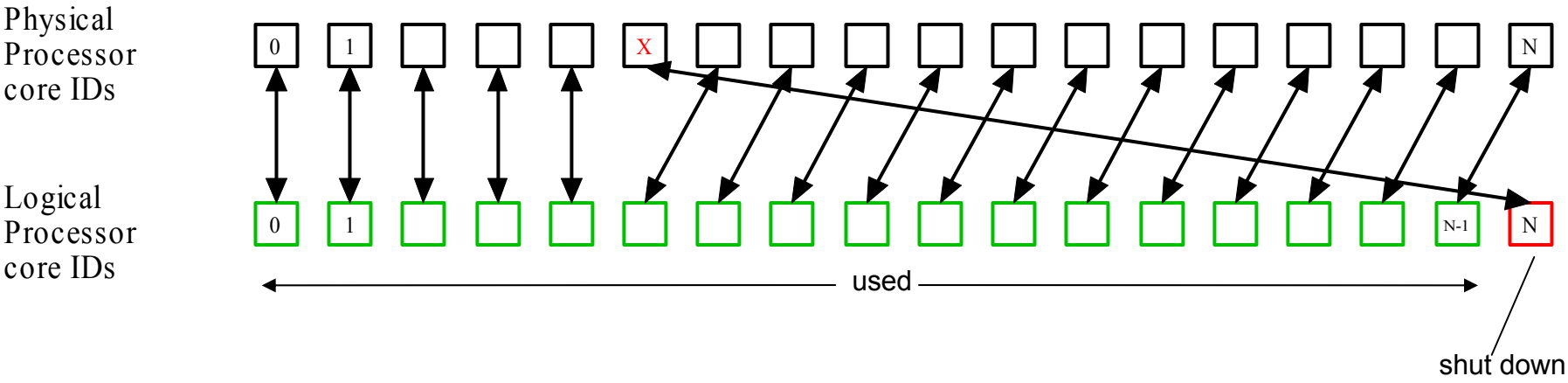
- **Provides Operating System services to the 16 user cores**
  - Reduces O/S noise and jitter on user cores
    - Helps applications run predictably / reproducibly
  - Offloads interrupt handling
  - Asynchronous I/O completion
  - Messaging assist, e.g. MPI pacing
  - Offloads RAS Event handling

Redundancy – the 18th core



Scan chain arrangement allows for simple determination of PUnit logic fails at manufacturing test

Physical-to-Logical mapping of PUnits in presence of a fail



- Inspired by array redundancy
- PUnit N+1 redundancy scheme substantially increases yield of large chip
- Redundancy can be invoked at any manufacturing test stage
  - wafer, module, card, system
- Redundancy info travels with physical part -- stored on chip (eFuse) / on card (EEPROM)
  - at power-on, info transmitted to PUnits, memory system, etc.
- Single part number flow
- Transparent to user software: user sees N consecutive good processor cores.

Blue Gene/Q Packaging Hierarchy

Node Card:
32 Compute Nodes
+ Optical Modules, Link Chips;
5D Torus

Compute Node:
Chip module, Memory

Single Chip Module

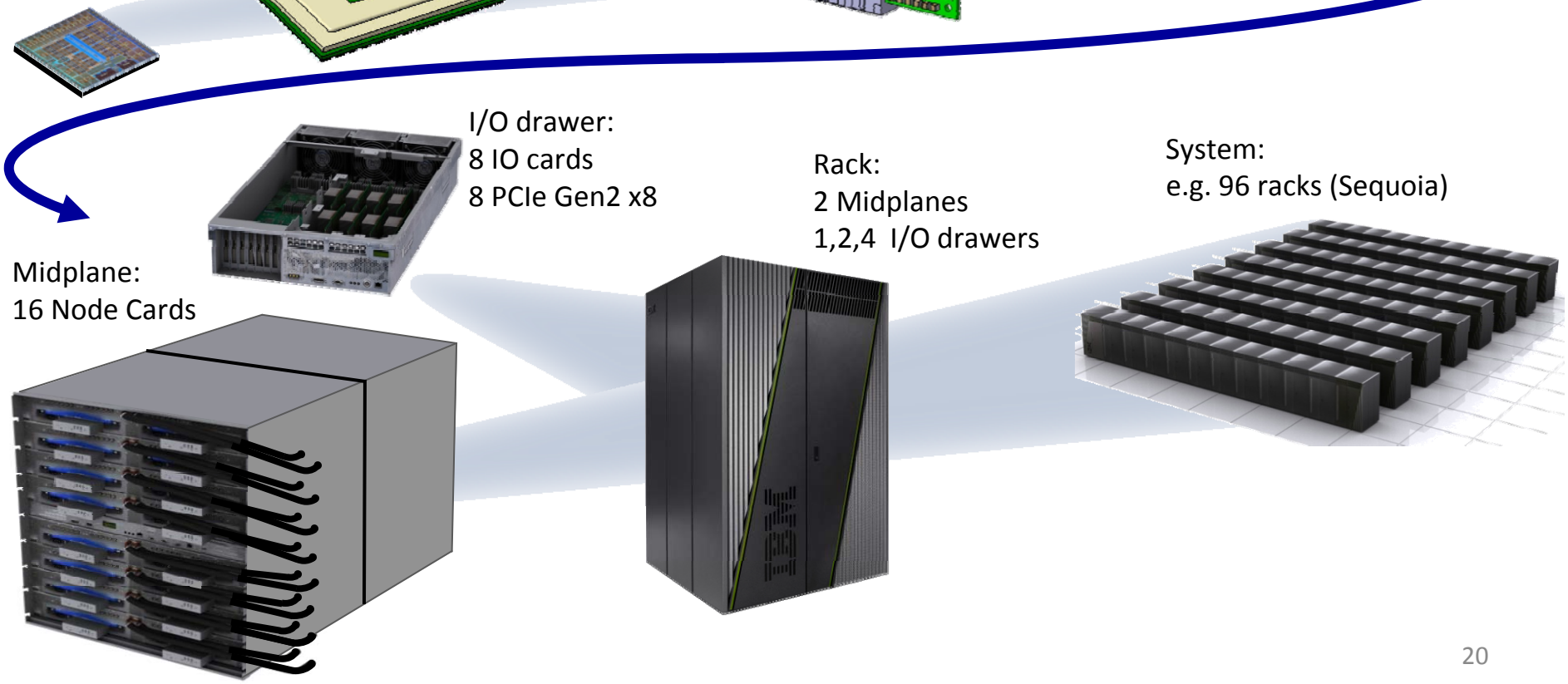
Chip
16+2 µP cores

I/O drawer:
8 IO cards
8 PCIe Gen2 x8

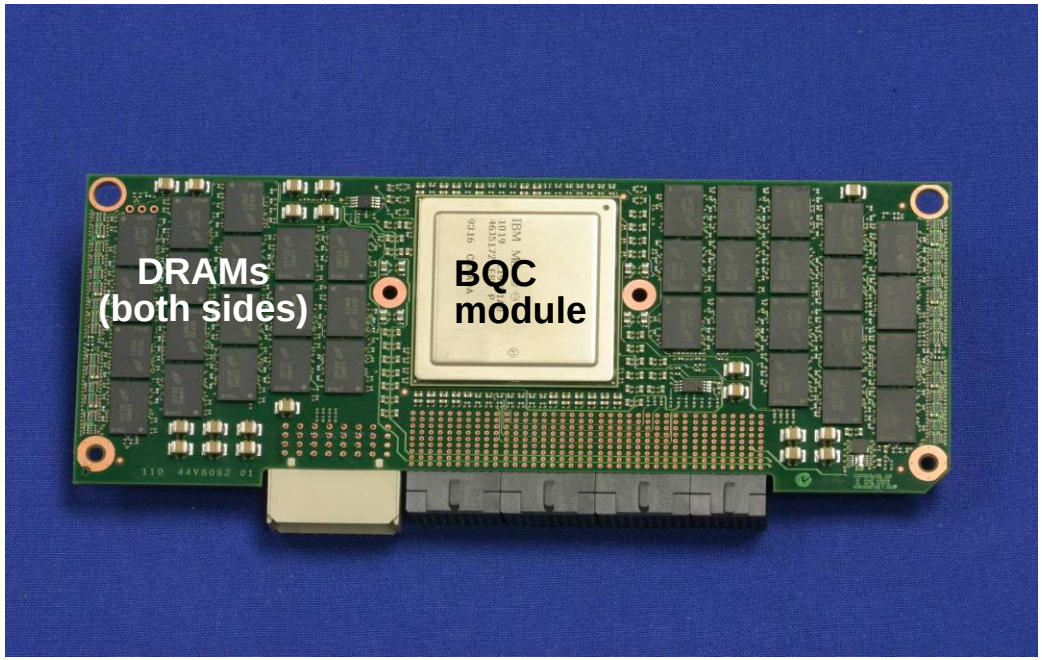
Rack:
2 Midplanes
1,2,4 I/O drawers

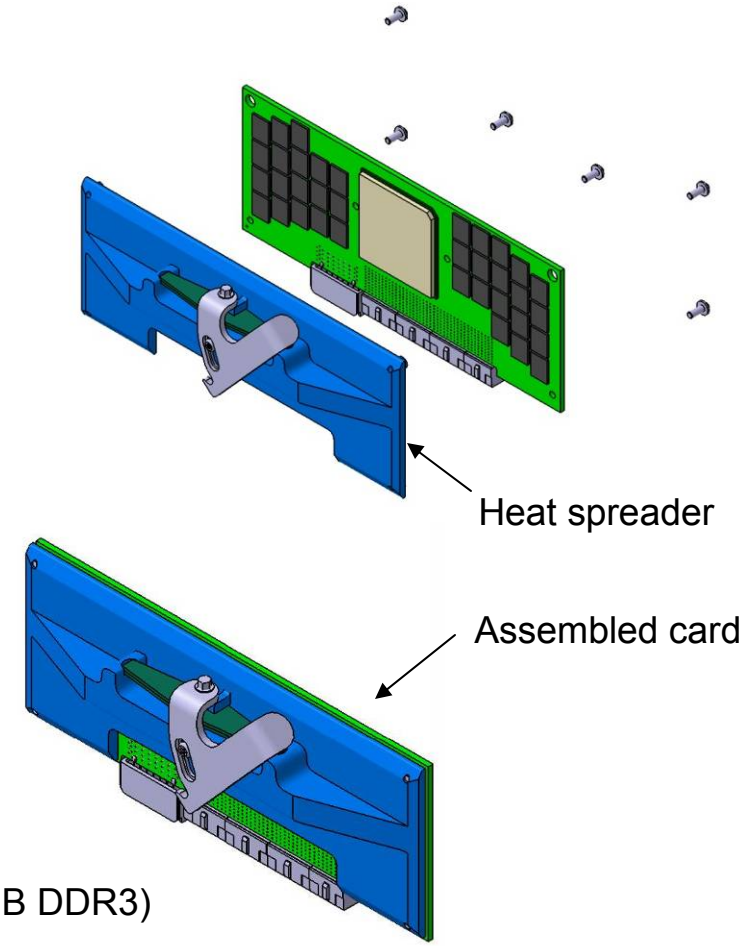
System:
e.g. 96 racks (Sequoia)

Midplane:
16 Node Cards



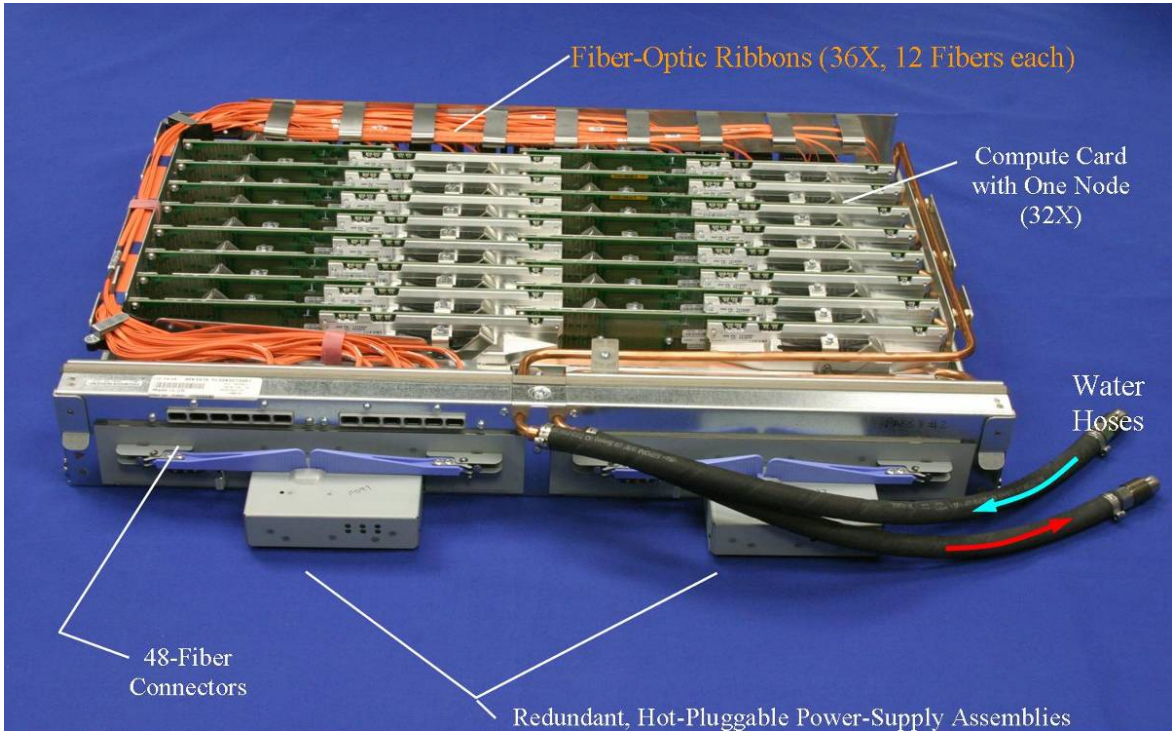
Blue Gene/Q Compute Card





- Basic FRU of a BlueGene/Q system
- ~ 179 x 69 mm card /w BQC module + 72 SDRAMs (16GB DDR3)
- ~ 72W Linpack: 56W chip + 12 W memory + 4 W communication
- Two heat sink options:
  - water-cooled → "Compute Node"
  - air-cooled → "IO Node"
- Connectors carry power supplies, JTAG etc, and 176 HSS signals (4 and 5 Gbps)

Blue Gene/Q Node Card assembly



- Power efficient processor chips allow dense packaging -- 32 Compute Nodes @ 72 W
- High bandwidth / low latency electrical interconnect on-board
- 18+18 (Tx+Rx) 12-channel optical fibers @10Gb/s
  - recombined into 8*48-channel fibers for rack-to-rack (Torus) and 4*12 for Compute-to-IO interconnect
- Compute Node Card assembly is water-cooled (18-25ºC – above dew point)
- Redundant power supplies with distributed back-end ~ 2.5 kW

Midplane

- 16 Compute Node cards
  → 512 compute nodes
- 4x4x4x4x2 (5D) Torus
- electrically interconnected

Rack

- 2 midplanes, optically connected
  → 1024 compute nodes
- 1 – 4 IO drawers on top
- Bulk power modules (480V AC -> 48V DC)

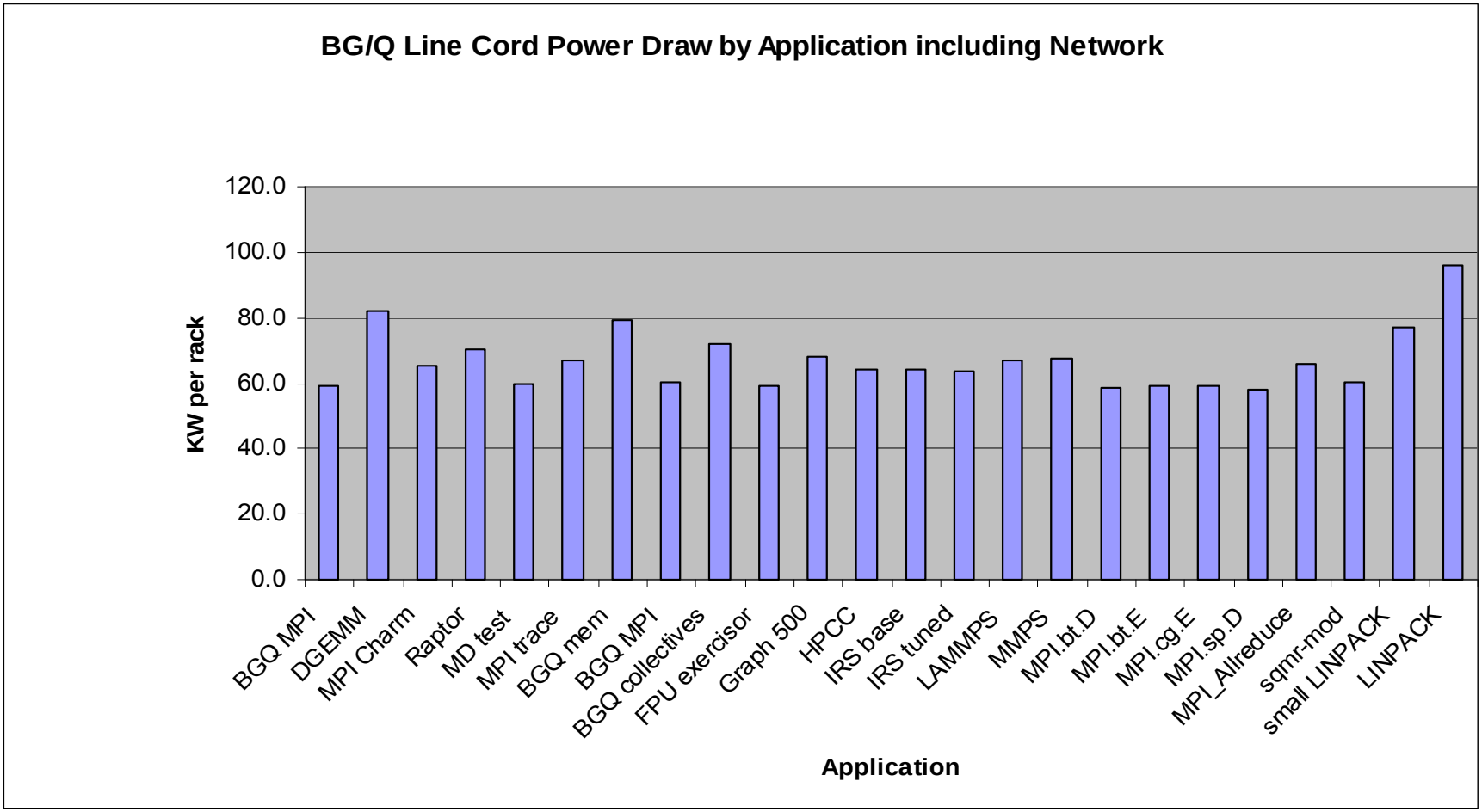
Water	**18C to 25C**
Flow	**20 gpm to 30 gpm**
Height	**2095 mm (82.5 inches)**
Width	**1219 mm (48 inches)**
Depth	**1321 mm (52 inches)**
Weight	**2000 kg (4400 lbs)** ***(including water)***
	I/O enclosure with 4 drawers **210 kg (480 lbs)**

Blue Gene/Q Sequoia installation

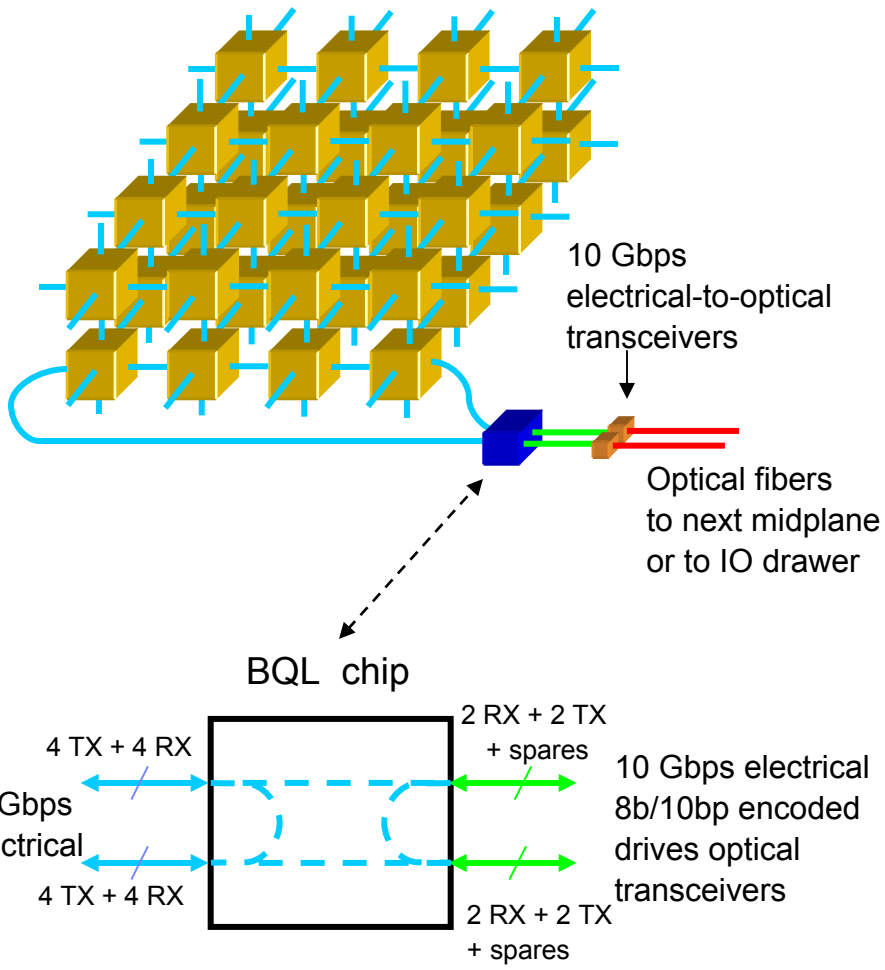
- 12 rows x 8 racks
- 98304 compute nodes in (16x16x16x12x2) 5D Torus configuration
- ~1.6 M user cores -- ~6.4M user threads
- June 2012 Top500 winner at 16.3 PFLOPS Linpack (81% of peak 20 PFLOPS)
- 7.89 MW (2.07 GFLOPS/W) ← Green500

BQC Power Measurements

From 4 rack Prototype System

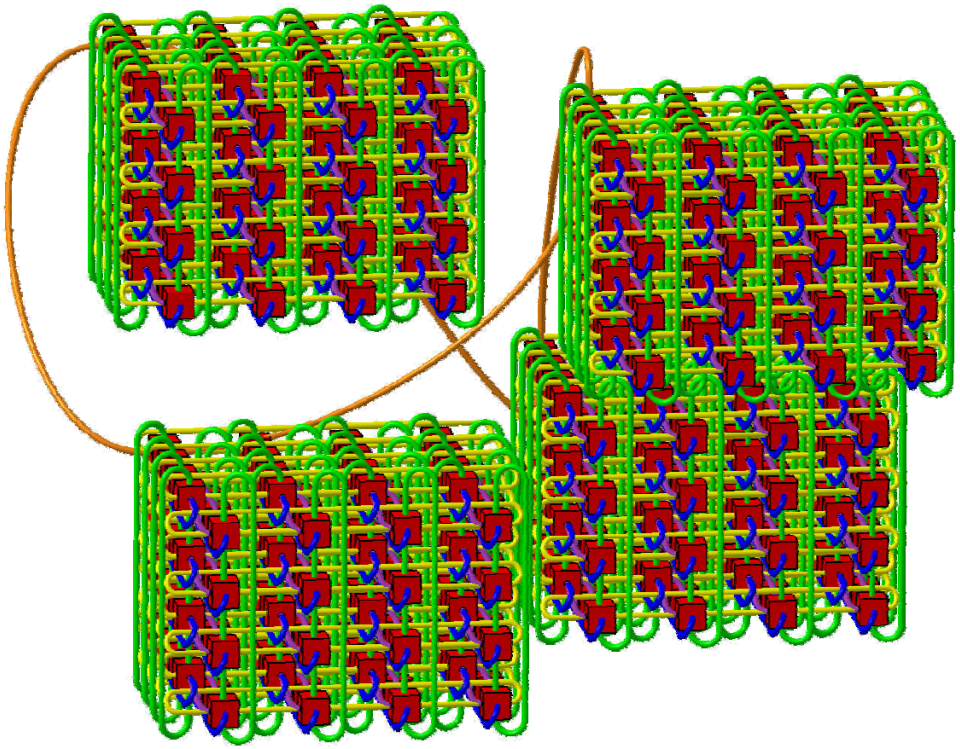


Blue Gene/Q Torus and Link chip



- **Blue Gene/Q interconnect topology is 5D torus**
  - dimensions A,B,C,D,E
  - each has + and − direction
- 512-way "midplane" is 4x4x4x4x2 hypercube
- Within midplane: each chip-to-chip link is electrical:
  4 TX + 4 RX lanes @ 4 Gbps each
  - **peak raw data rate per link: 2 GB/s TX + 2 GB/s RX**
- At edge of midplane, dimension A,B,C,D connect to BlueGene/Q Link (BQL) chip
  - E-dimension is wrapped internally
- BQL chip switches between:
    - wrap-around or
    - connection to next midplanes.
  - Set-up statically for each user partition (job)
  - User partitions do not interfere with each other
    - helps performance reproducibility
- Separately, BQL drives links to IO drawer
- BQL translates to optical transceivers
  - 4 lanes * 4 Gbps ↔ 2 lanes * 10 Gbps
  - 8b/10bp encoded
  - single error correction per byte
  - maintains **2 GB/s data rate per optical chip-chip link**
- BQL contains logic for fiber sparing

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
  - Virtual Cut-Through routing
  - Hardware assists for collective and barrier functions
  - FP addition support in network
  - RDMA
    - Integrated on-chip Message Unit
- 2 GB/s raw bandwidth on all 10 links
  - each direction -- i.e. 4 GB/s bidi
  - 1.8 GB/s peak user bandwidth (protocol overhead)
- 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

Design Challenges

- **Area is the enemy**
  - 16 cores (A2 + QPX + L1P) + 1 helper core + 1 redundant spare
  - enough cache per core / per thread
  - high bandwidth to external memory
  - high speed communication
  - leads to LARGE chip: 18.96x18.96 mm
    → redundant core significantly helps yield
- **Power is the enemy**
  - SOC design (processors, memory, network logic) reduces chip-to-chip crossings
  - 2.4 GHz PowerEN™ core design is run at reduced speed (1.6 GHz), reduced voltage (~0.8V)
    - reduced voltage reduces both active power and leakage power
    - speed binning → all chips run @ 1.6 GHz, with voltage adjusted to match speed sort.
  - Deployed methodologies/tools to keep power down
    - Architecture/logic level: clock gating
    - Processor cores: re-tuned for low power
    - Power-aware synthesis; power-recovery steps
    - Physical design: power-efficient clock networks
- **Soft Errors are the enemy**
  - Sensitivity to SER events affects reliability for large installations – such as BlueGene/Q

  → Needs vigilance in data protection and circuit design/use.
    - DDR3, L2 cache, network, all major arrays and buses ECC protected
    - Minor buses, GPRs, FPRs: parity protected, with recovery
    - Stacked / DICE latches

Design Challenges -- continued

And the enemy is us…

- **Methodology Complexity**
  - Processor cores originated in a high-speed custom design methodology
  - Rest of the chip implemented as ASIC
    → Required merging of different clocking/latching, timing and test methodologies

- **Logic verification**
  - On-chip memory sub-system (transactional memory, speculative execution)
  - Full-chip POR sequence, X-state (… inherited "proven" logic)
    → Extensive use of cycle simulation / hardware accelerators / FPGA emulator

- **Test pattern generation**
  - Again, mixed chip / mixed methodologies
  - Full chip models
  - turn-around time is becoming a bottleneck

Blue Gene Characteristics

	BG/L	BG/P	BG/Q
Compute Nodes			
Processor	32-bit PowerPC 440	32-bit PowerPC 450	64-bit PowerPC (A2 Core)
Processor Frequency	700 MHz	850 MHz	1.6 GHz
Cores	2	4	16+1
Peak Performance (per Node)	5.6 GF	13.6 GF	204.8 GF
Coherency	Software Managed	SMP	SMP + Speculation
L1 Cache (per Core)	32 KB	32 KB	16/16 KB
L2 Cache (prefetch per Core/Thread)	14 stream	14 stream	16 stream + List-based
L3 Cache size (shared, per Node)	4 MB	8 MB	32 MB
Main Store/Node *(same for I/O Node)*	512 MB or 1 GB	2 GB or 4 GB	16 GB
Main Store Bandwidth	5.6 GB/s (16B wide)	13.6 GB/s (2*16B wide)	43 GB/s
Torus Network			
Topology	3D	3D	5D
Bandwidth	6*2*175 MB/s = 2.1 GB/s	6*2*425 MB/s = 5.1 GB/s	40 GB/s
Hardware Latency *(Nearest Neighbor)*	200 ns (32B packet) 1.6 μs (256B packet)	100 ns (32B packet) 800 ns (256B packet)	80 ns (32B packet) 640 ns (256B packet)
Hardware Latency *(Worst Case)*	6.4 μs (64 hops)	5.5 μs (64 hops)	3 μs (31 hops)
Per Rack			
Peak Performance	5.7 TF	13.9 TF	209 TF
Sustained Performance *(Linpack)*	4.6 TF	11.9 TF	~170+ TF
Power (peak)	~20 kW	~32 kW	~100 kW
Power Efficiency	0.23 GF/W	0.37 GF/W	2.09 GF/W

15x

3.2x

7.8x

15x

5.6x

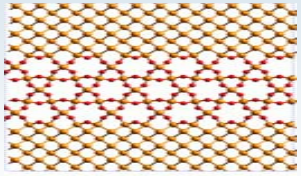
Examples of Applications Running on Blue Gene

Developed on L, P; many ported to Q

Application	Owner	Application	Owner	Application	Owner
CFD Alya System	Barcelona SC	DFT iGryd	Jülich	BM: SPEC2006, SPEC openmp	SPEC
CFD (Flame) AVBP	CERFACS Consortium	DFT KKRnano	Jülich	BM: NAS Parallel Benchmarks	NASA
CFD dns3D	Argonne National Lab	DFT ls3df	Argonne National Lab	BM: RZG (AIMS,Gadget,GENE, GROMACS,NEMORB,Octopus, Vertex)	RZG
CFD OpenFOAM	SGI	DFT PARATEC	NERSC / LBL	Coulomb Solver - PEPC	Jülich
CFD NEK5000, NEKTAR	Argonne, Brown U	DFT CPMD	IBM/Max Planck	MPI PALLAS	UCB
CFD OVERFLOW	NASA, Boeing	DFT QBOX	LLNL	Mesh AMR	CCSE, LBL
CFD Saturne	EDF	DFT VASP	U Vienna & Duisburg	PETSC	Argonne National Lab
CFD LBM	Erlangen-Nuremberg	Q Chem GAMESS	Ames Lab/Iowa State	MpiBlast-pio Biology	VaTech / ANL
MD Amber	UCSF	Nuclear Physics GFMC	Argonne National Lab	RTM – Seismic Imaging	ENI
MD Dalton	Univ Oslo/Argonne	Neutronics SWEEP3D	LANL	Supernova Ia FLASH	Argonne National Lab
MD ddcMD	LLNL	QCD CPS	Columbia U/IBM	Ocean HYCOM	NOPP / Consortium
MD LAMMPS	Sandia National Labs	QCD MILC	Indiana University	Ocean POP	LANL/ANL/NCAR
MD MP2C	Jülich	Plasma GTC	PPPL	Weather/Climate CAM	NCAR
MD NAMD	UIUC/NCSA	Plasma GYRO (Tokamak)	General Atomics	Weather/Climate Held-Suarez Test	GFDL
MD Rosetta	U Washington	KAUST Stencil Code Gen	KAUST	Climate HOMME	NCAR
DFT GPAW	Argonne National Lab	BM:sppm,raptor,AMG,IRS,sphot	Livermore	Weather/Climate WRF, CM1	NCAR, NCSA

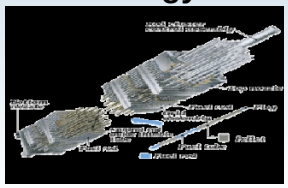
Accelerating Discovery and Innovation in:

Materials Science



Silicon Design

Energy

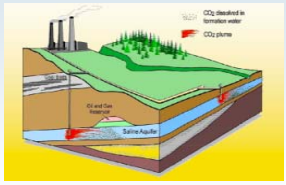


Next Gen Nuclear

Engineering

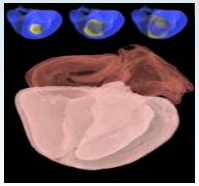
High Efficiency Engines

Climate & Environment



Oil Exploration

Life Sciences



Whole Organ Simulation

Conclusions

- The Blue Gene/Q Compute chip is the building block for a power-efficient supercomputing system that can scale to tens of PetaFLOPS.

- Hardware
  - BQC introduces architectural innovations to enable multithreaded / multicore computing
  - Hardware support to facilitate efficient multi-threading: atomic ops, wake-up unit
  - Cache structure supports Speculative Execution and Transactional Memory
  - On-chip networking logic allows dense, high-bandwidth chip-to-chip interconnect, with hardware assist for collective functions and RDMA
  - Achieves over 200 GFLOPS peak in a power-efficient fashion
    → 2.1 GFLOPS/W Linpack performance -- #1 in Green500 Nov 2010 – June 2012

- Software
  - Processors are homogeneous, implement standard PowerISA (plus SIMD extensions)
    - Compilers are available that leverage the on-chip hardware assists for multithreading
  - User cores fully available and quiet -- system functions offloaded to 17th core.
  - Supports open standards:
    - Parallel processing: MPI
    - Multi-threading: OpenMP

    … and allows for many other programming models

- Applications:
  - are in scale-up (Livermore, Argonne, Cineca, Columbia/Edinburgh, Juelich, KEK, EDF…)