JÜLICH
FORSCHUNGSZENTRUM

JUQUEEN: Blue Gene/Q - System Architecture

M. Stephan

Outline

- Blue Gene/Q hardware design
  - *Processor*
  - *Network*
  - *I/O architecture*
- Jülich Blue Gene/Q configuration (JUQUEEN)
- Blue Gene/Q software overview
  - *Programming model*
  - *Blue Gene/Q software stack*

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 (> 1,5Mi cores)
- High reliability

Blue Gene evolution

Blue Gene/L (5.7 TF/rack) – 130nm ASIC (Dev.:1999-2004)

- PPC440 core, dual core SoC, 0.5/1 GB/node
- Biggest installed system (LLNL):
  - *104 racks, 212,992 cores/threads, 596 TF/s, 210 MF/W*

Blue Gene/P (13.9 TF/rack) – 90nm ASIC (Dev.:2004-2007)

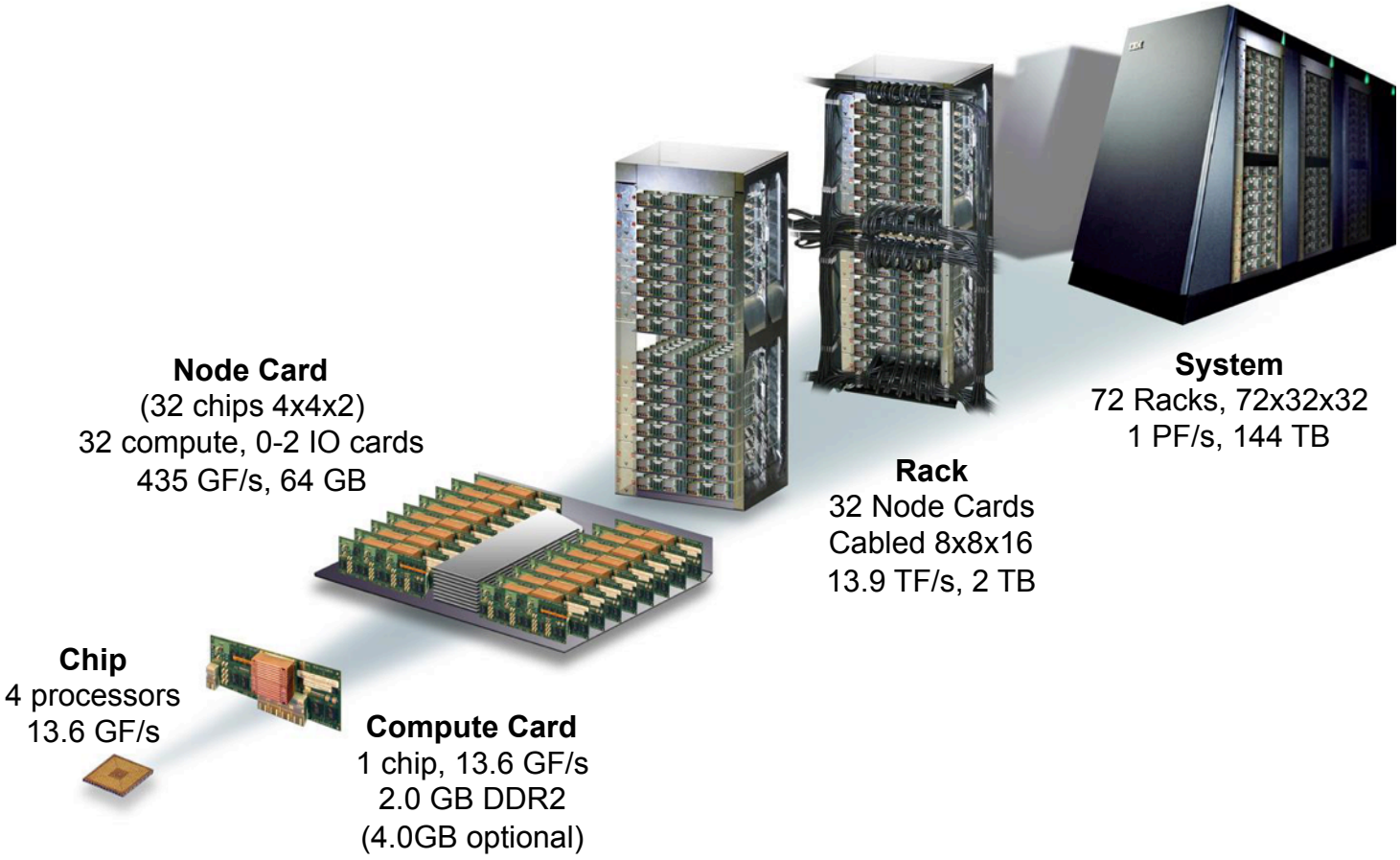
- PPC450 core, quad core SeC, DMA, 2/4 GB/node
- Biggest installed system (Jülich):
  - *72 racks, 294,912 cores/threads, 1 PF/s, 357 MF/W*

Blue Gene evolution

BG/Q (209 TF/rack) – 45nm ASIC+ (Dev.: 2007-2012)

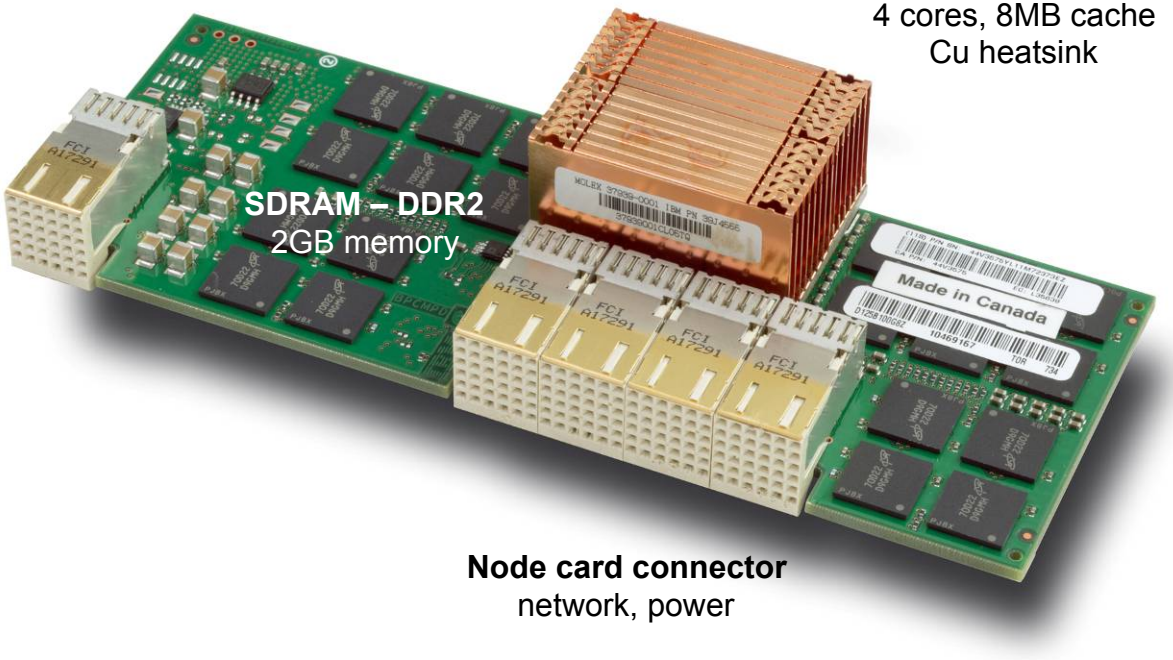
- A2 core, 16 core/64 thread SoC
- 16 GB/node
- Biggest installed system (LLNL):
  - *96 racks, 1,572,864 cores & >6M threads, 20 PF/s, 2 GF/W*
- Speculative execution, sophisticated L1 prefetch, transactional memory, fast thread handoff, compute + IO systems.

Blue Gene/P design

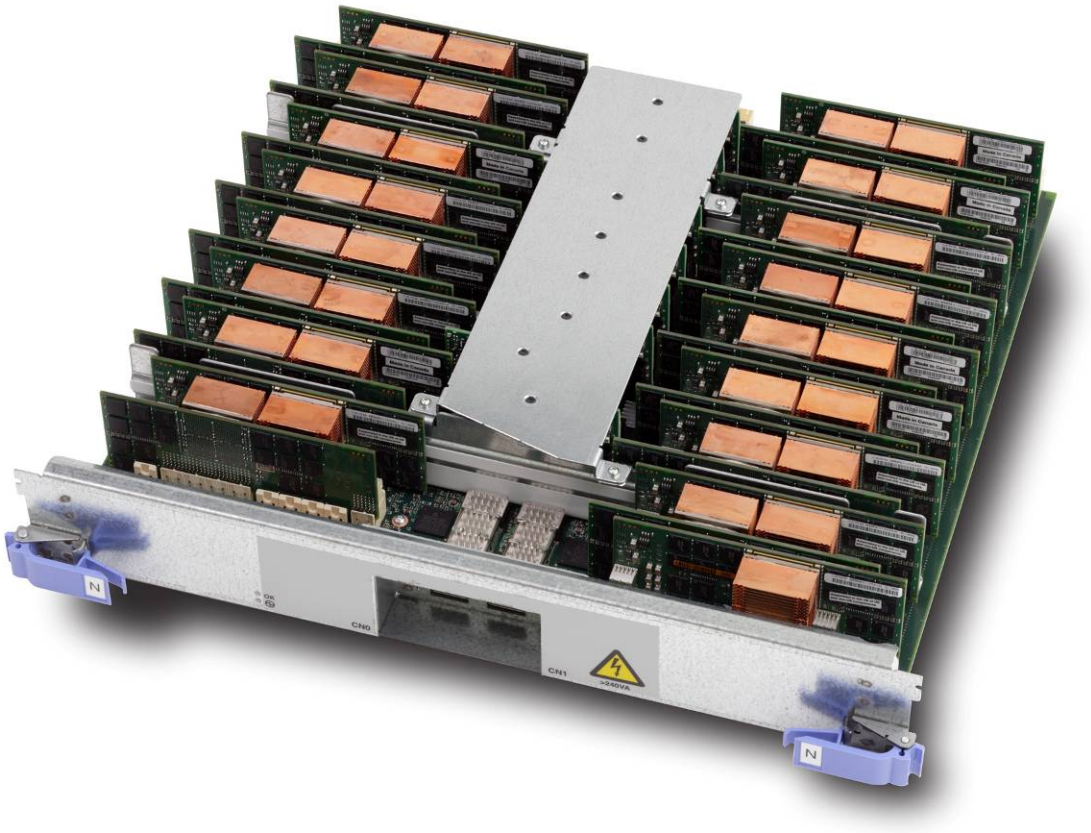


Source: IBM

Blue Gene/P compute card

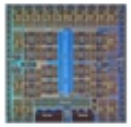


Blue Gene/P node card



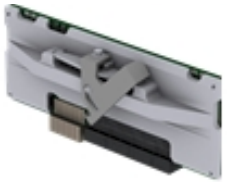
Blue Gene/Q design

1. Chip
16 cores

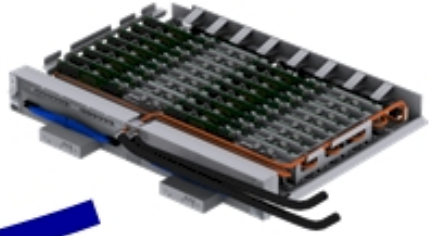


2. Module
Single Chip

3. Compute Card (“Node”):
One BQC Module (1x1x1x1x1),
16 GB DDR3 Memory



4. Node Card (“Node Board”):
32 Compute Cards (2x2x2x2x2),
Optical Modules, BQL Link Chips, Torus



5b. I/O drawer (1,2 or 4 per rack):
* 8 I/O cards @ 16 GB,
* 8 PCIe gen2 x8 slots (IB, 10GbE)

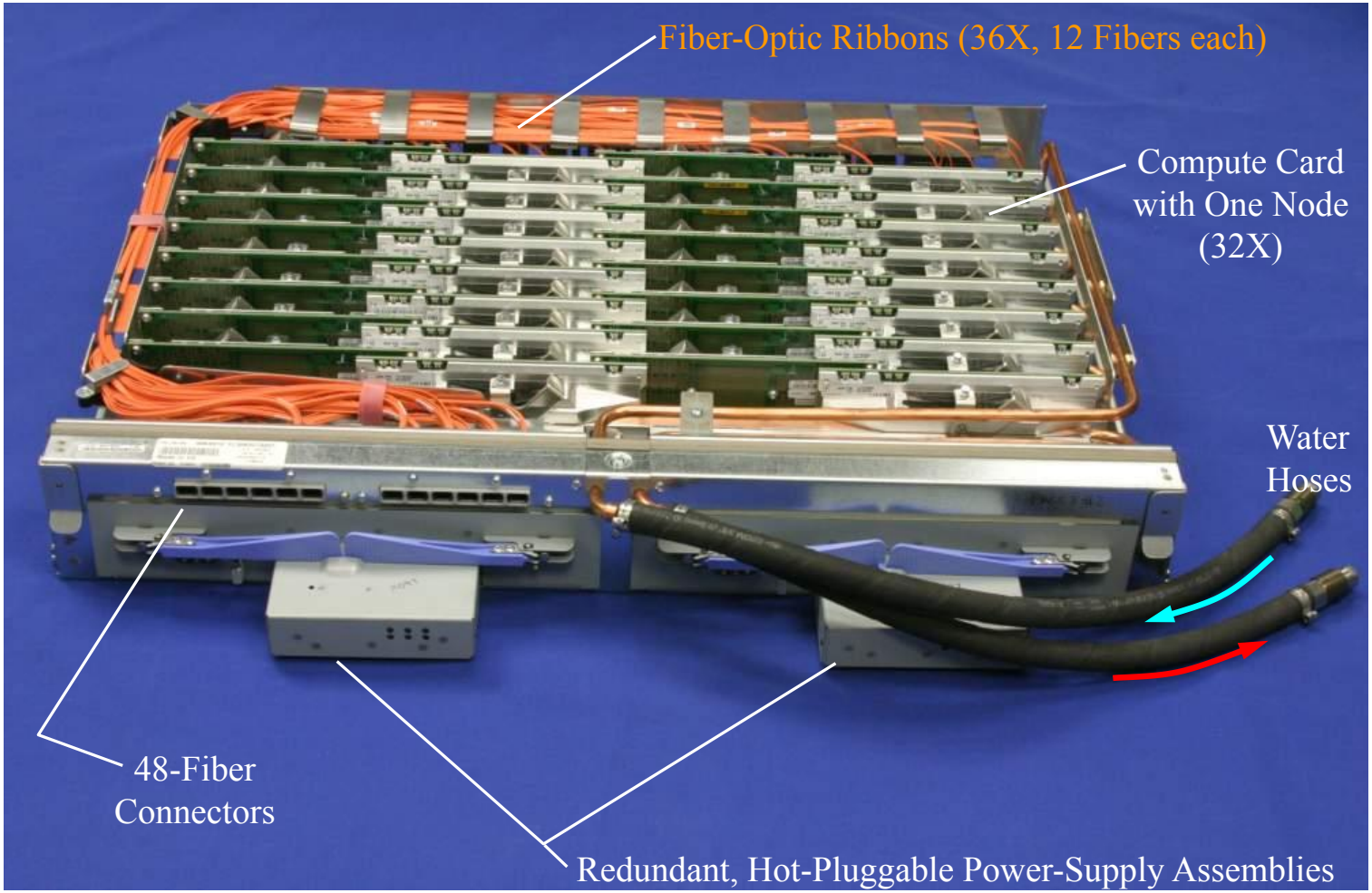
5a. Midplane:
16 Node Cards (4x4x4x4x2)

6. Rack: 2 Midplanes (4x4x4x8x2)
1,2 or 4 I/O Drawer

7. System:
e.g. 8 racks (8x8x8x8x2) = 1.7 PF/s
e.g. 28 racks (8x28x8x8x2) = 5.9 PF/s
e.g. 96 racks (16x12x16x16x2) = 20 PF/s

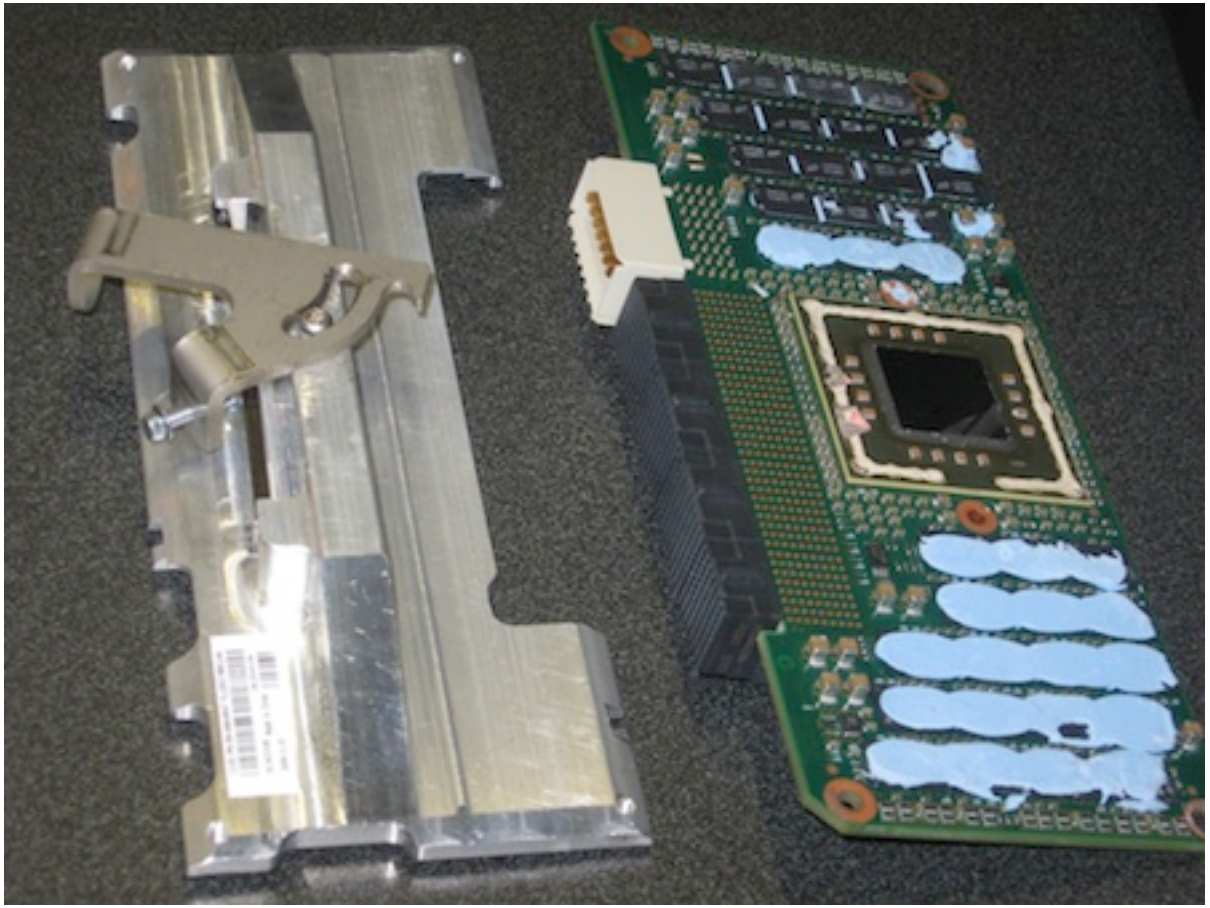
Source: IBM

Blue Gene/Q node card



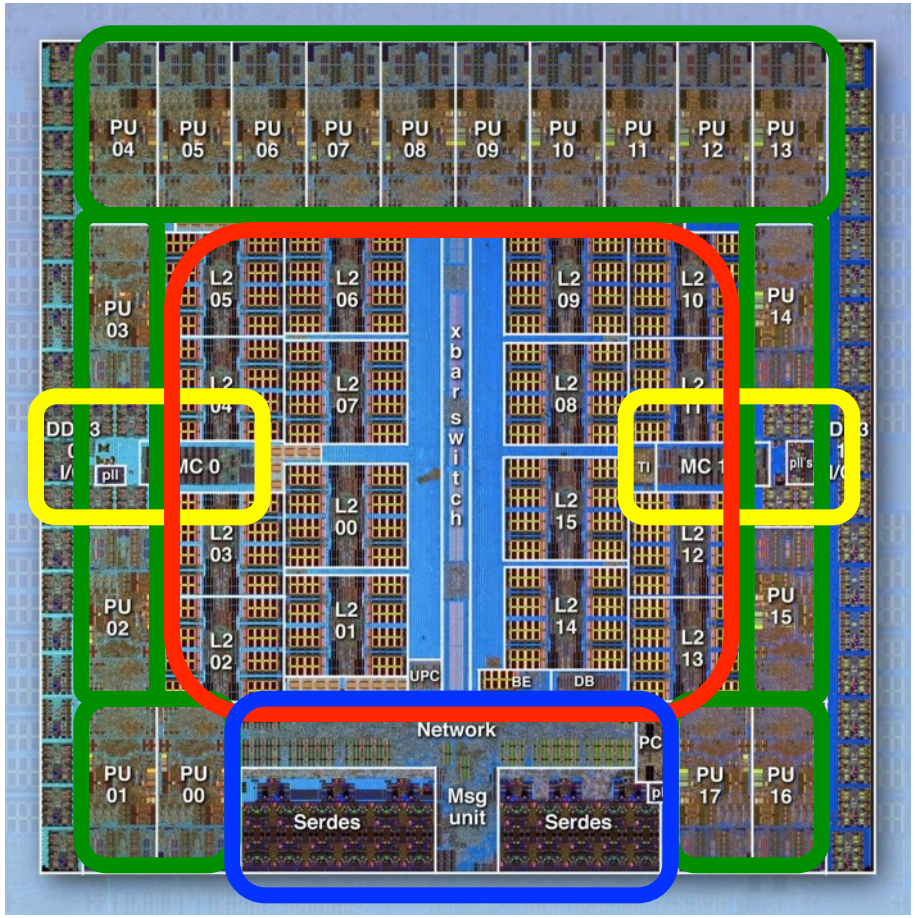
Source: IBM

Blue Gene/Q compute card



Source: Top500.org

Blue Gene/Q: Chip tomography



16+1+1 processing units

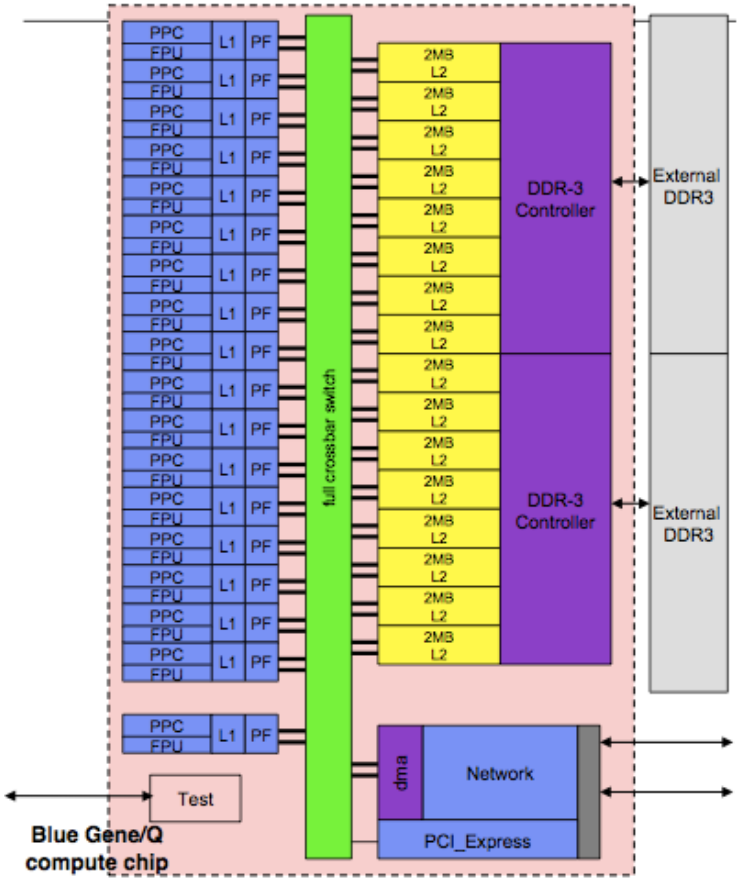
Two memory controller

L2 cache + crossbar switch

On-chip network

Source: IBM

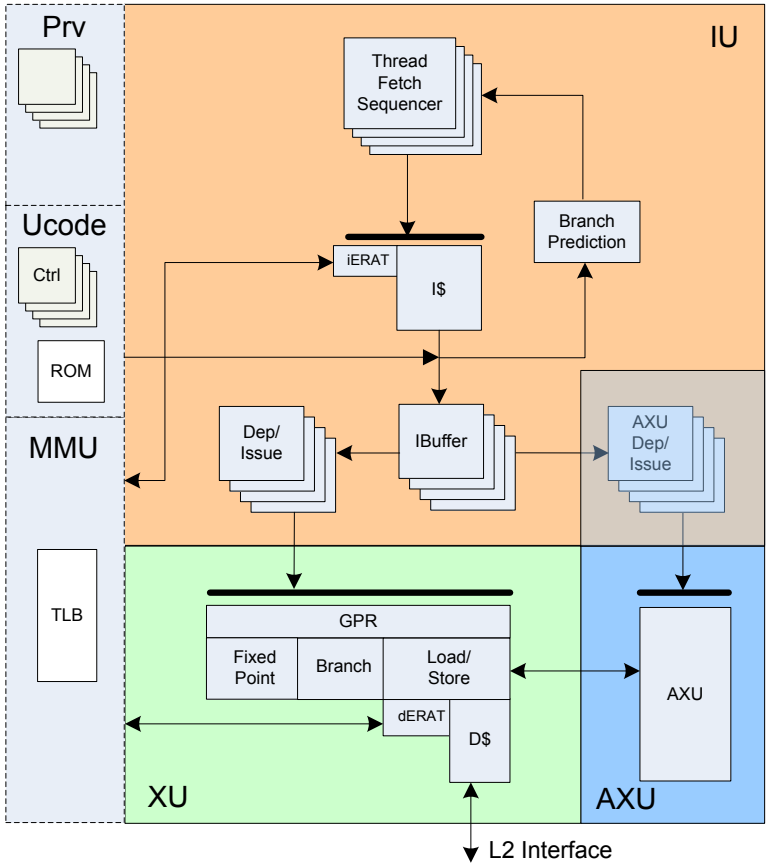
Blue Gene/Q chip architecture



- 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



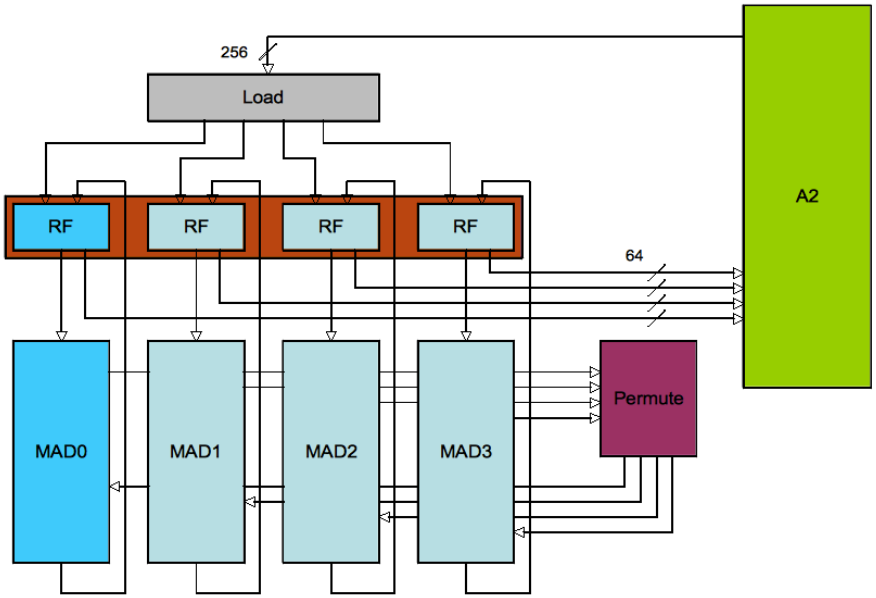
Source: IBM

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*

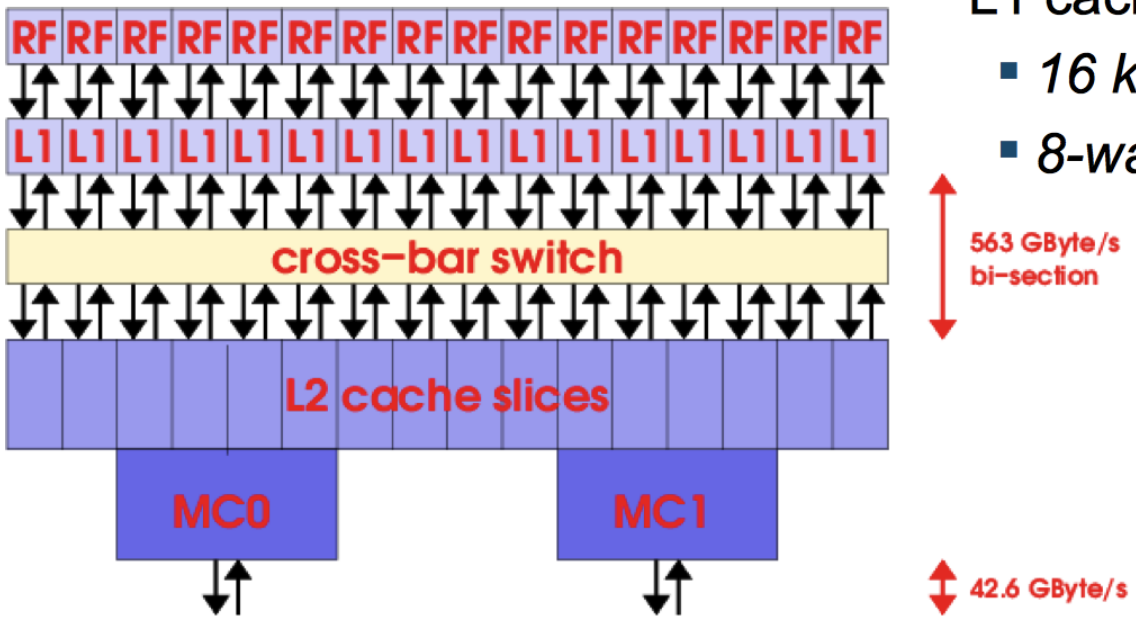
Quad floating Point eXtension unit (QPX)

- 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*



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

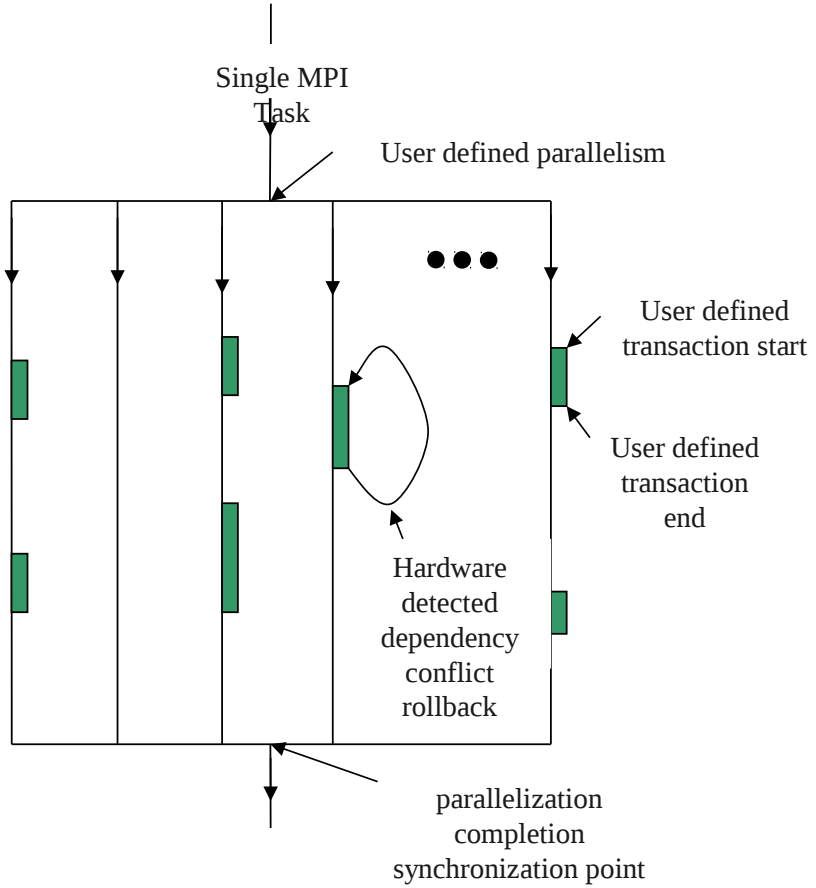
- *Cache can track state changes caused by speculative threads*
- *At the end of speculative code: invalidate or commit and/or react with software notification*
- *Use: transactional memory, thread-level speculation*

Blue Gene/Q: Thread Level Speculation

- Parallelize potentially dependent serial fragments
  - *runtime creates threads for each speculative section*
  - *threads run parallel and commit serialized if no conflict*
  - *on conflict, all threads except current master is rolled back*
- Enable by compiler flag -qsmp=speculative
- Performance governed by trade off of overhead and conflict probability
- Times to try rollback before non-sepc. execution:
  - ***export SE_MAX_NUM_ROLLBACK=N***
- Hardware Limitation: maximum of 16 domains
- Turn on logs to inspect runtime-gathered statistics for tuning: (see IBM XL Optimization Guide)
  - ***export SE_REPORT_...=...***

Blue Gene/Q: Transactional Memory

- BG/Q has support for hardware transactional memory
  - *mechanism to enable atomic operations on arbitrary set of memory locations*
  - *App needs to allow that transactions commit out-of-order*
  - *may be used to parallelize workload into collaborative but indepenent tasks on shared data*
  - *hardware detects write/read conflicts*
  - *runtime rolls back on failure*



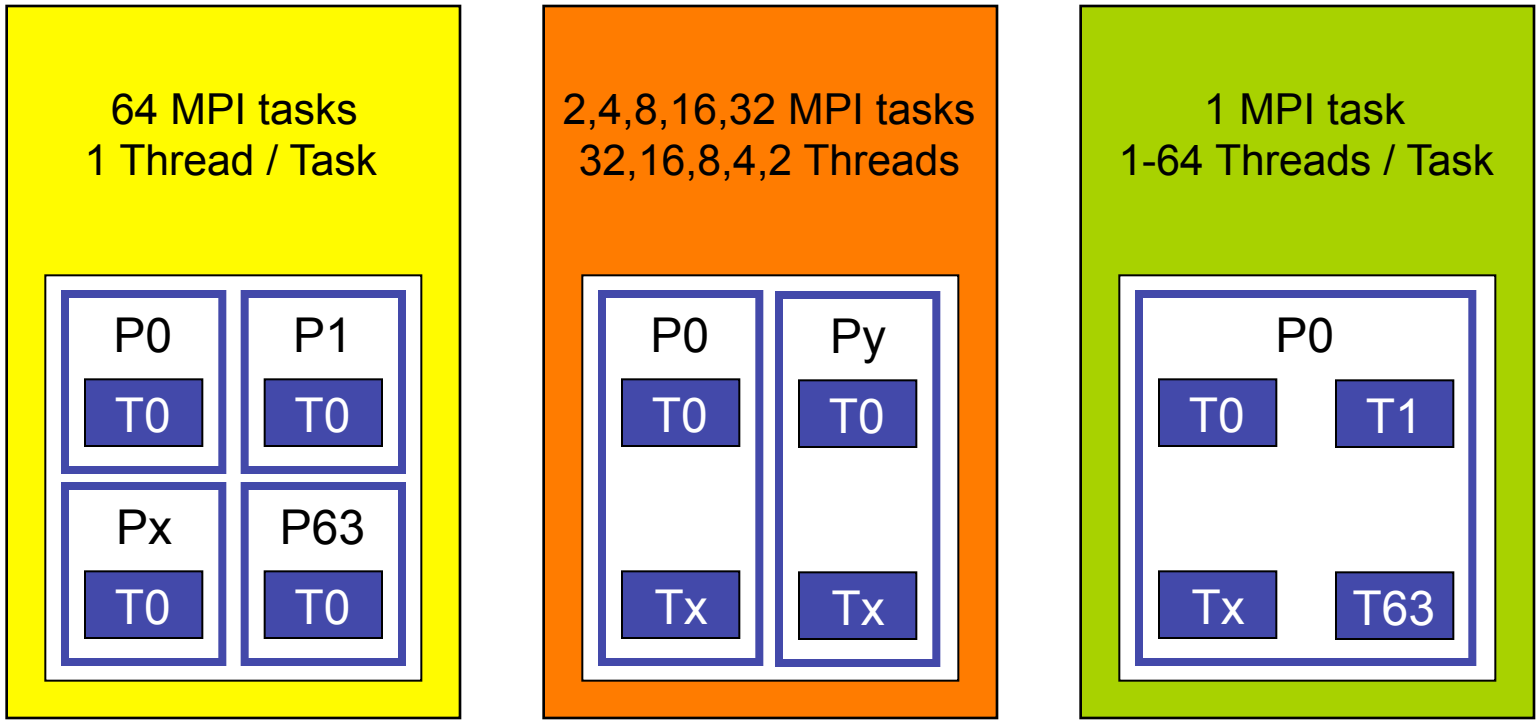
Comparison with other processors

	IBM Blue Gene/Q	Intel Xeon E5-2670	AMD Opteron 6267
#cores	16	8	16
Core clock [GHz]	1.6	2.6	2.3
SIMD width [bit]	256	256	256
Peak FP performance (DP) [GFlops]	204.8	166.4	147.2
Peak memory bandwidth [GByte/s]	42.6	51.2	51.2
Last-level cache [MiBytes]	32	20	16
Power [Watt]	55	115	115

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

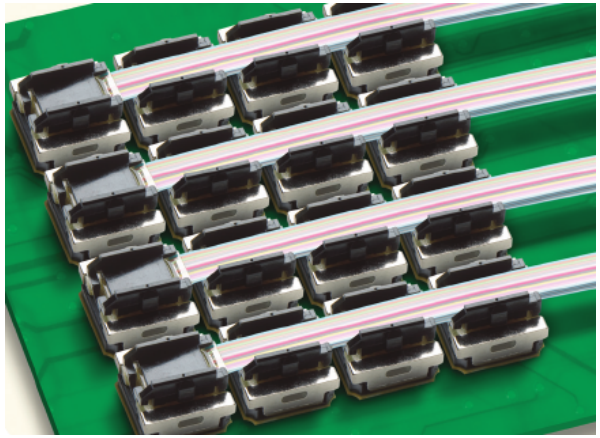
- Asynchronous I/O completion hand-off
- Responsive CIO application control client

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: System configurations

- BG/Q Nodes form a 5D torus
  - Nodecards: 2x2x2x2x2
  - Midplanes: 4x4x4x4x2
    - 4D are cabled to other midplaness
  - 5th dimension: extent 2 (stays within nodecard)
  - 6th dimension is cpu # within the node
  - Dim. labels: ABCDE T
- Different floor shapes (Rows x Cols) for a given number of racks may correspond to the same, or to different torus shapes
- This list is not complete; other configs are possible...
  - up to 16x16 = 256 racks

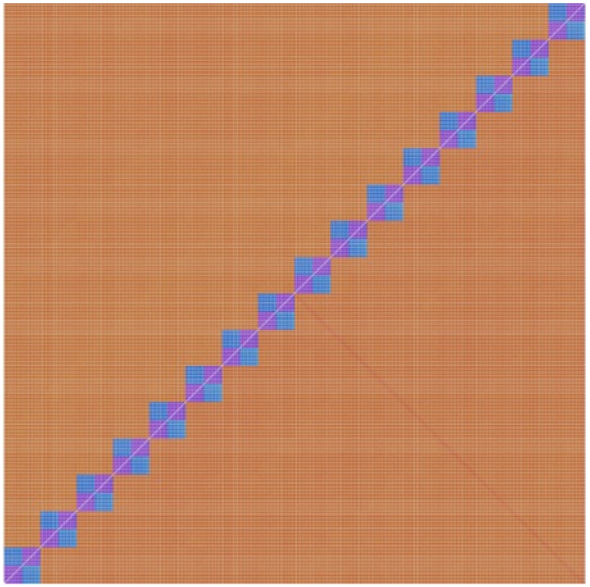
			Torus size, in nodes					Torus size, in midplanes						
Racks	*Rows*	*Col.*	A	B	C	D	E	A	B	C	D	Nodes	Cores	Threads
mid	1	1	4	4	4	4	2	1	1	1	1	512	8.192	32.768
1	1	1	4	4	4	8	2	1	1	1	2	1.024	16.384	65.536
2	1	2	4	4	8	8	2	1	1	2	2	2.048	32.768	131.072
3	1	3	4	4	12	8	2	1	1	3	2	3.072	49.152	196.608
4	1	4	8	4	8	8	2	2	1	2	2	4.096	65.536	262.144
4	2	2	4	8	8	8	2	1	2	2	2	4.096	65.536	262.144
6	1	6	12	4	8	8	2	3	1	2	2	6.144	98.304	393.216
6	2	3	4	8	12	8	2	1	2	3	2	6.144	98.304	393.216
6	3	2	4	12	8	8	2	1	3	2	2	6.144	98.304	393.216
8	1	8	8	8	8	8	2	2	2	2	2	8.192	131.072	524.288
8	2	4	8	8	8	8	2	2	2	2	2	8.192	131.072	524.288
8	4	2	8	8	8	8	2	2	2	2	2	8.192	131.072	524.288
10	5	2	4	20	8	8	2	1	5	2	2	10.240	163.840	655.360
12	3	4	8	12	8	8	2	2	3	2	2	12.288	196.608	786.432
12	2	6	12	8	8	8	2	3	2	2	2	12.288	196.608	786.432
16	2	8	16	8	8	8	2	4	2	2	2	16.384	262.144	1.048.576
16	4	4	8	16	8	8	2	2	4	2	2	16.384	262.144	1.048.576
20	5	4	8	20	8	8	2	2	5	2	2	20.480	327.680	1.310.720
24	6	4	8	12	16	8	2	2	3	4	2	24.576	393.216	1.572.864
24	2	12	8	8	12	16	2	2	2	3	4	24.576	393.216	1.572.864
28	7	4	8	28	8	8	2	2	7	2	2	28.672	458.752	1.835.008
32	8	4	8	16	16	8	2	2	4	4	2	32.768	524.288	2.097.152
32	4	8	8	16	16	8	2	2	4	4	2	32.768	524.288	2.097.152
40	5	8	8	20	16	8	2	2	5	4	2	40.960	655.360	2.621.440
48	6	8	16	12	16	8	2	4	3	4	2	49.152	786.432	3.145.728
48	4	12	8	16	12	16	2	2	4	3	4	49.152	786.432	3.145.728
48	3	16	8	12	16	16	2	2	3	4	4	49.152	786.432	3.145.728
56	7	8	8	28	16	8	2	2	7	4	2	57.344	917.504	3.670.016
64	8	8	8	16	16	16	2	2	4	4	4	65.536	1.048.576	4.194.304
64	4	16	8	16	16	16	2	2	4	4	4	65.536	1.048.576	4.194.304
72	9	8	12	12	16	16	2	3	3	4	4	73.728	1.179.648	4.718.592
80	10	8	8	20	16	16	2	2	5	4	4	81.920	1.310.720	5.242.880
96	12	8	16	12	16	16	2	4	3	4	4	98.304	1.572.864	6.291.456
96	8	12	16	16	12	16	2	4	4	3	4	98.304	1.572.864	6.291.456
96	6	16	16	12	16	16	2	4	3	4	4	98.304	1.572.864	6.291.456

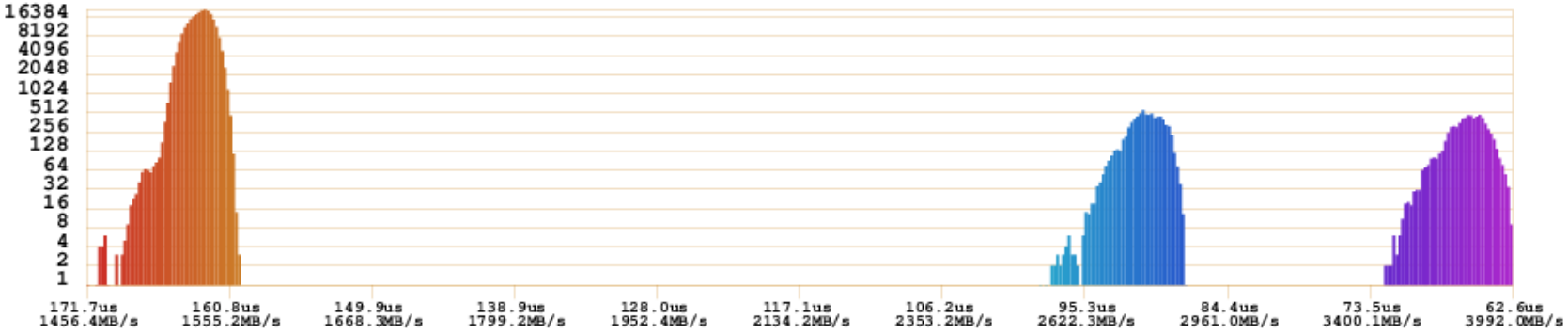
Torus node to MPI mapping

- Block's physical network topology: 6 dimensional A×B×C×D×E×T
  - *local on each node; shared memory communication (fast); T=0..63*
  - *A,B,C,D,E depend on size of block, e.g. 2×2×4×4×2 for quarter midplane*
- 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*

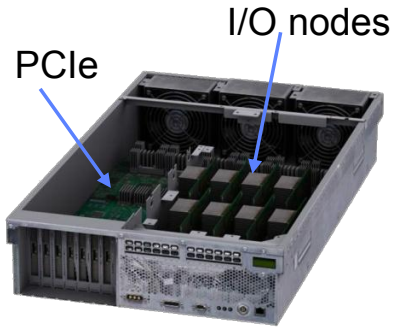




Blue Gene/Q: I/O architecture

- **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 (04/2012)

8 Racks Blue Gene/Q

- 8192 Compute Nodes (16 cores, 16 GB memory)
- 131072 cores / 0.5M treads
- 1.68 PFlop/s peak performance
- 88 I/O nodes (10GigE) ← (1x32 + 7x8)
- 400 kW power consumption (10-50 kW per rack)

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

- IBM p7 740, 8 cores (3.55 GHz), 128 GB memory
- local storage device DS5020 (16 TB)

JUQUEEN Configuration (11/2012)

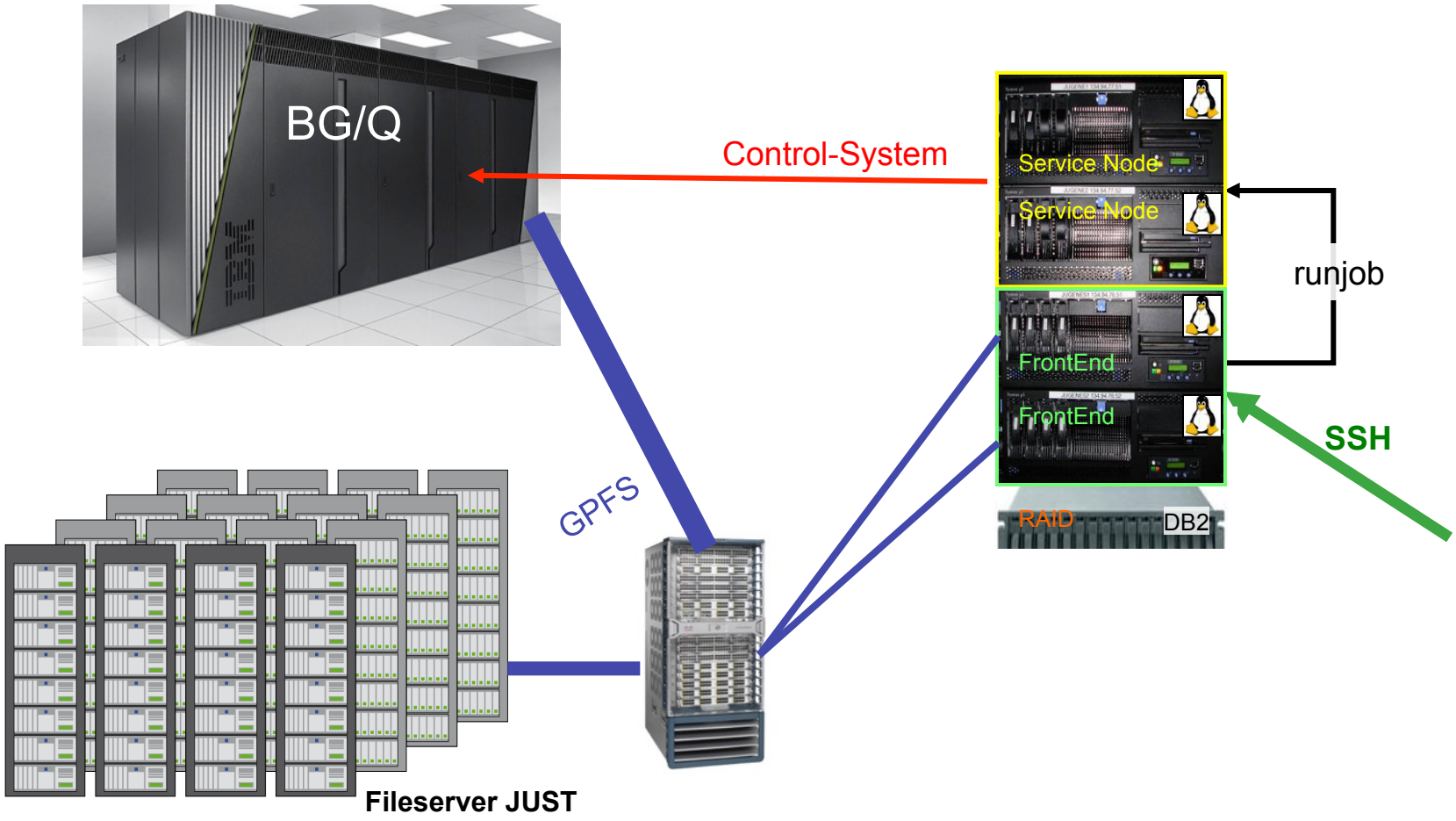
28 Racks Blue Gene/Q

- 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)
- 1.4 MW power consumption (10-50 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 DS5020 (16 TB)

JuQueen Environment



JUQUEEN: First pictures

JUQUEEN: First pictures

JÜLICH
FORSCHUNGSZENTRUM

JUQUEEN: First pictures

JUQUEEN: First pictures

JÜLICH
FORSCHUNGSZENTRUM

JUQUEEN: First pictures

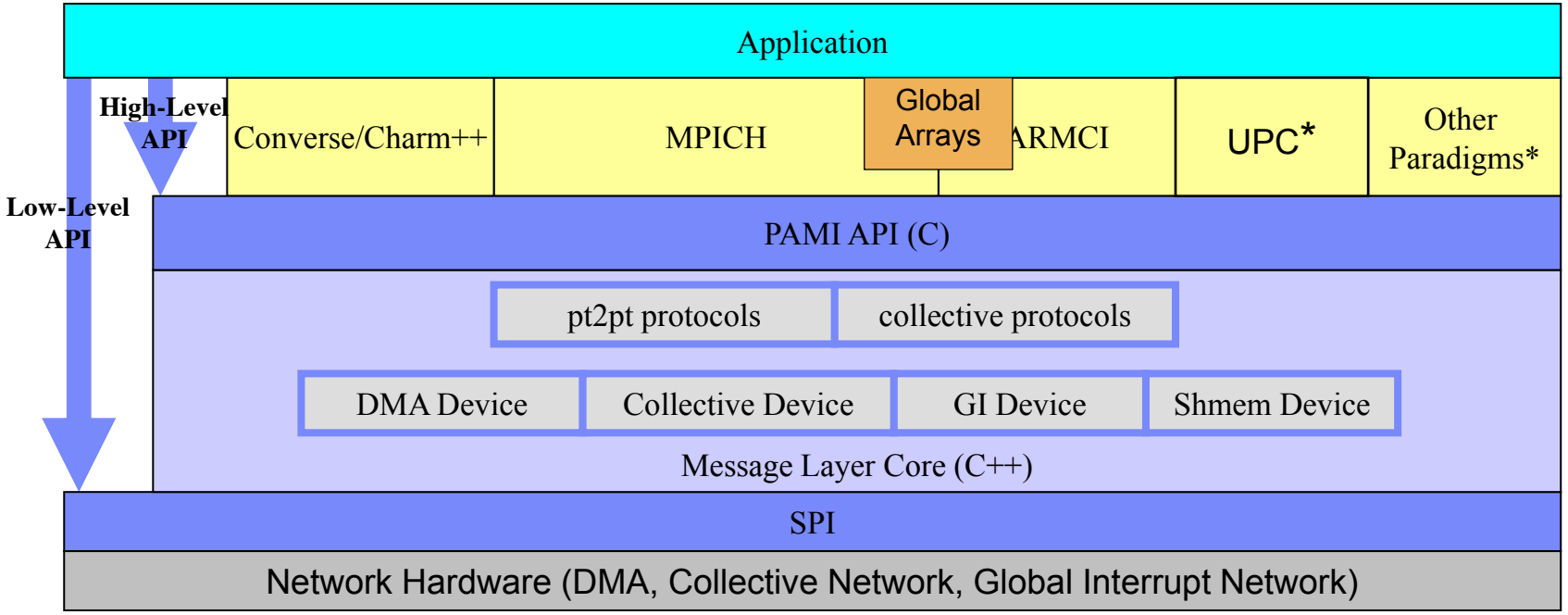
JUQUEEN: First pictures

Blue Gene/Q – Software

Blue Gene/Q: MPI Implementation

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

Blue Gene/Q: Parallel Active Message Interface

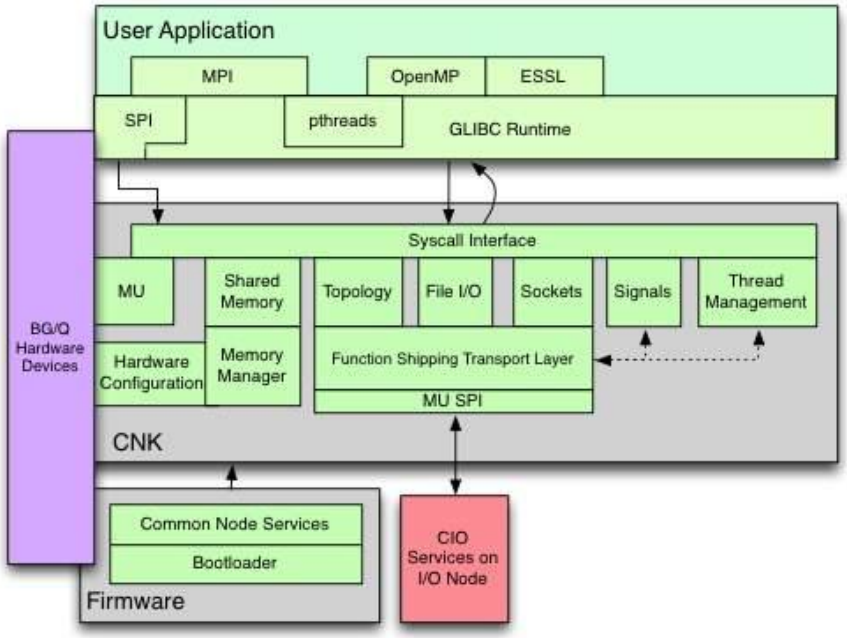


Blue Gene/Q: Extra MPI communicators

- 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)*

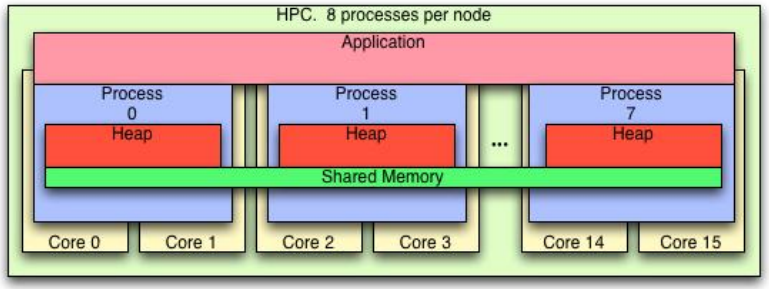
Blue Gene/Q: Compute Node Kernel (CNK)

- 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-Compiler)
- Transactional Memory (TM)
- Speculative Multi-Threading (TLS)
- Shared and Persistent Memory
- Scripting Environments (Python)
- Dynamic Linking, Demand Loading



Blue Gene/Q: Processes

- Similarities to BGP
  - Number of tasks per node fixed at job start
  - No fork/exec support
  - Support static and dynamically linked processes
  - -np supported
- 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, …, vnm)
  - 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

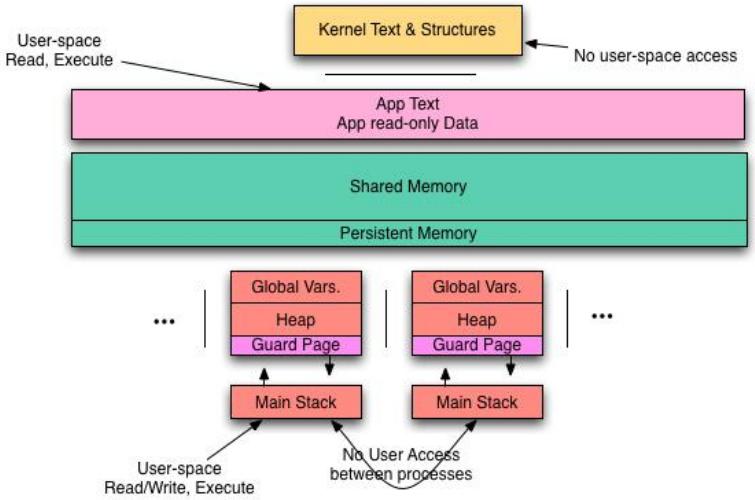


Blue Gene/Q: Threads

- 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:

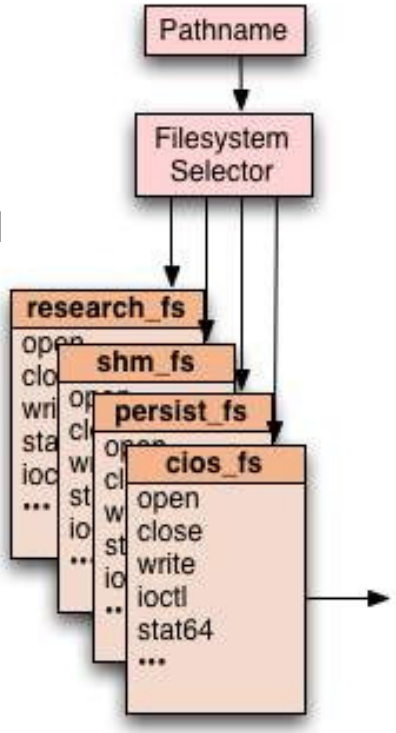
Blue Gene/Q: Memory

- 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



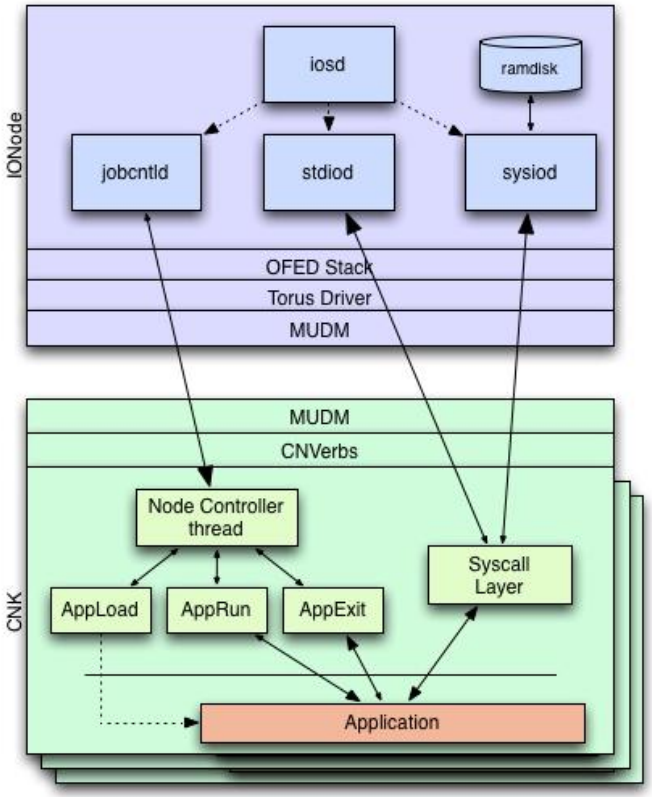
Blue Gene/Q: System Calls

- Similarities to BG/P
  - Many common syscalls on Linux work on BG/Q.
  - Linux syscalls that worked on BG/P 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
    - Different simulator platforms have different capabilities
    - Shared Memory



Blue Gene/Q: 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



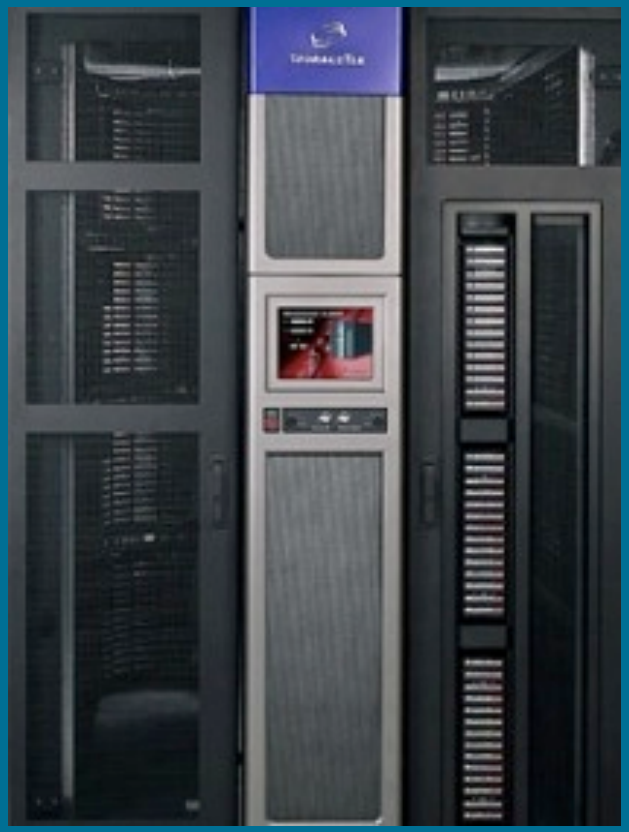
Blue Gene/Q: Debugging

- 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

Compiling MPI programs on Blue Gene/Q

There are six versions of the libraries and the scripts

- **gcc:** GNU compiler with fine-grained locking in MPICH – error checking
- **gcc.legacy:** GNU 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



Storage

Filesystems

- The user can store his datasets in three different file systems
  - *$HOME*
  - *$WORK*
  - *$ARCH*

Filesystems – $HOME

- Daily backup
- For regular used datasets and applications
- 3 TB space and 2 million datasets per group
- Quota information with *q_dataquota*
- Urgent datasets which are used for the current computation

Filesystems – $WORK

- Fastest file system (33 GB/s)
- No backup
- Datasets will be deleted 90 days after the last usage
- Quota of 6 TB and 4 million datasets per group
- For datasets which are used or generated during the computation
- Urgent datasets should be copied to $ARCH or $HOME

Filesystems – $ARCH

- Daily backup
- All datasets will be migrated to tape
- Large, not recently used datasets
- No space quota, 2 million datasets per group
- NOT available on the Blue Gene compute nodes

- Usage of TAR or ZIP archives is highly recommended
- Restore with 100MB/s *but* up to 90s for file opening
  - *1 x 200 GB → 30 - 40 min*
  - *1000 x 200 MB → 25 h*
  - *100000 x 2 MB → 104 d (3.5 month)*

Recomendations

- Don´t use touch to trigger recall of migrated data
  - *change of timestamp will force backup and newmigration even if data is used readonly*
- Don´t change the path of large archive entities by renaming, deletion or insertion of subdirectories
  - *will produce high system overhead, because alldata affected by change will be recalled, backed up and migrated*
- Bad example:Path:
  - *mv /arch/zam/test/project /arch/zam/test/project_old*

Data transfer

- scp
  - *Too many accesses (ssh or scp) within a short amount of time will be interpreted as intrusion and leads to automatic disabling the origin system at the FZJ firewall.*
  - *For transferring multiple files in a single scp session the -r option can be used, which allows to transfer a whole directory.*

JÜLICH
FORSCHUNGSZENTRUM

Using JuQueen

Sample programs

Cross compile on **juqueenX** to generate a BG/Q executable.

Execution from within a job:

runjob --exe hello.rts [**--pwd <path>**] [**--args <args>**]

See

man runjob or **runjob –h** for options.

Sample programs and makefiles - not yet available:

/bgsys/local/samples/

-
- ….

Tools (in preparation)

- EMACS Editor
- module
- ….

Installed Libraries (/bgsys/local/lib)

Mathematical — i.gutheil@fz-juelich.de

- **essl**
- **lapack**
- **scalapack**
- **blacs**

Other

- **SIONlib**
- **(p)netcdf, HDF5** *(tbd)*
  (+ Linux version in /usr/local/lib)

Packages

- **CPMD,VASP** *(on demand only)*
- **NAMD**

Information

Message-of-today

```
********************************************************************************
* Welcome to the 4 racks Blue Gene/Q system JUGENE at FZJ!
* Information about the system, latest changes, user documentation:
*       http://www.fz-juelich.de/ias/jsc/juqueen
* For all parallel IO - use only filesystem $WORK (/work/<group>/<userid>)!
* Filesystem $WORK is not backed up and files are deleted after 90 days!
* Filesystems for $ARCH are NOT mounted on IONs.
*
********************************************************************************
```

news [-a] - list of news for installed products

LoadLeveler for Batch Jobs (V5.1)

- Job execution only possible via LoadLeveler jobs
- Submission of batch jobs from **juqueen1** or **juqueen2** (incl. **#@ bg_keywords**)
- Central Job Manager runs on service node, boots blocks for the jobs
- (Re-)uses dynamic blocks: **LLyymmddhhmmss**xx
- Front-End nodes start the job, call runjob, which talks to a runjob server on the service node

LoadLeveler Job

Required Keywords (Type & Size/Shape):

#@ job_type = bluegene

#@ bg_size = <*number of nodes*>

or

#@ bg_shape = AxBxCxD

A,B,C,D are number of midplanes in A-,B-,C-,D-direction (max. 2x2x2x2) or permutation of this

Loadleveler Job

Optional Keywords:

#@ bg_connectivity = MESH | TORUS | Either | Xa Xb Xc Xd
Xa Xb Xc Xd is equal to Torus or Mesh

#@ bg_rotate = True | False to disable permutations

Submission:

llsubmit *<cmdfile>*

llsubmit: The job "juqueen1c1.zam.kfa-juelich.de.13" has been submitted.

Job is processed by a FZJ submit filter, which associates a class name (and checks the cpu quota).

General LoadLeveler Keywords

#@ wall_clock_limit = <*HH:MM:SS*>
#@ notification = error | start | complete | always
#@ notify_user = <*valid-email-address*>
#@ input = <*some-filename*>
#@ output = <*some-filename*>
#@ error = <*some-filename*>
#@ initialdir = <*some-pathname*>
#@ environment = COPY_ALL

Job examples in **/bgsys/local/samples/LoadL**

Example 1

```
# @ error = $(job_name).$(jobid).out
# @ output = $(job_name).$(jobid).out
# @ wall_clock_limit = 01:55:00,01:50:00
# @ notification =  error
# @ notify_user  = <valid-email-address>
# @ job_type = bluegene
# @ bg_size = 2048
# @ queue
runjob  --exe myprogram  --args "456 99“
```

Default: 1 MPI Task per node

Example 2

```
# @ job_name = LoadL_Sample2
# @ error = $(job_name).$(jobid).out
# @ output = $(job_name).$(jobid).out
# @ environment = COPY_ALL;
# @ wall_clock_limit = 01:05:00, 01:00:00
# @ notification =  never
# @ job_type = bluegene
# @ bg_size = 1024
# @ queue
runjob --exe myprog.rts  --np 8000 --args "-t 1"
```

Example 3

```
# @ job_name = LoadL_Sample3
# @ error = $(job_name).$(jobid).out
# @ output = $(job_name).$(jobid).out
# @ wall_clock_limit = 01:05:00, 01:00:00
# @ notification =  error
# @ notify_user  = <valid-email-address>
# @ job_type = bluegene
# @ bg_connectivity = TORUS
# @ bg_shape = 1x1x2x2
# @ bg_rotate = FALSE
# @ queue
runjob --exe myprog --ranks-per-node 32
```

LoadLeveler Keywords not allowed

~~#@ executable =~~

~~#@ arguments =~~

Use runjob in LoadLeveler script only!

Job Classes

Class/Queue	Max. nodes	Wall Clock Limit	Priority
m016	8192	24 h	on demand only
m008	4096	24 h	
m004	2048	24 h	
m002	1024	24 h	
m001	512	24 h	
n008	256	12 h	R13
n004	128	12 h	R13
n002	64	30 min	R13
n001	32	30 min	R13
serial	0	60 min	juqueenX

Batch Job Scheduling

Backfill Scheduler

Biggest job has highest priority

Fill with short running jobs until *TOPDOG* job fits

Jobs with lower priority but **short wall clock time** get a better chance to be executed

Big Jobs

Jobs requesting full machine (m016) will run:

- after maintenance
- at least once a week

Job Status

llq [-u <userid>]

```
Id                       Owner      Submitted   ST PRI Class        Running On
------------------------------------------------------------------------------------------------
juqueen2c1.100.0         zdv087      5/8  17:33 I   50  m001
juqueen1c1.10.0          zdv068      5/8  17:38  I  50  n004
```

All jobs:

llqx

```
Loadleveler Jobs at:  Wed May  9 14:38:18 CEST 2012
JOB STEP           USER      SUB/DISP         WALL     CLASS    END     SYSPRI GP Re SIZR  SIZA WI U St BS PARTITION
juqueen1c1.53.0    zdv068    May  9 14:38:15  01:40      n001    --:--   25591432  1 y    32          Ei    Id
juqueen1c1.54.0    zdv068    May  9 14:38:16  01:40      n001    --:--   25591432  1 y    32          Ei    Id
juqueen2c1.138.0   zdv087    May  9 14:27:32  01:10   sysmall    15:37   65991499  1 y   128   128     Ei    Ru Ru LL12050914225128
juqueen1c1.50.0    zdv068    May  9 14:38:20  01:40      n001    16:18   25891432  1 y    32    32     Ei    Ru    LL12050913100601
juqueen1c1.51.0    zdv068    May  9 14:38:20  01:40      n001    16:18   25891432  1 y    32    32     Ei    Ru    LL12050913100602
juqueen1c1.52.0    zdv068    May  9 14:38:20  01:40      n001    16:18   25891432  1 y    32    32     Ei    Ru    LL12050913100603
2012-05-09 14:38 Alloc. racks:   0.22 (   0.3%), in   4 (  4 /  0) jobs, Idle jobs (m...):   0 (  0 /  0), req. (  0 /  0) racks
```

Cancel jobs

llcancel juqueen2c1.44.0

llcancel: Cancel command has been sent to the central manager.

Monitoring – llview

in preparation

Interactive Execution

not (yet) available → use LoadLeveler job

Documentation

http://www.fz-juelich.de/ias/jsc/EN/Expertise/Supercomputers/JUQUEEN/JUQUEEN_node.html

Documentation for IBM Compiler (Beta Version)
not on WEB pages confidential!!

/usr/local/BETA_compiler_doc/xlc:
/usr/local/BETA_compiler_doc/xlf:

compiler.pdf	*(Compiler Reference)*
getstart.pdf	*(Getting Started)*
langref.pdf	*(Language Reference)*
proguide.pdf	*(Optimizantion and Programming)*

Support

- sc@fz-juelich.de — application support
- bg-adm@fz-juelich.de — system administrator