JÜLICH
FORSCHUNGSZENTRUM

JUQUEEN: Blue Gene/Q - Overview

M. Stephan, J. Docter

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*

- Data Storage - JUST Fileserver

JÜLICH
FORSCHUNGSZENTRUM

Blue Gene/Q – Hardware

Blue Gene design goals

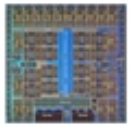
- System-on-Chip (SoC) design
  - *Processor comprises both processing cores and network*
- Optimal performance / watt ratio
- Small foot print
- Transparent high-speed reliable network
- Easy programming based on standard message passing interface (MPI)
- Extreme scalability (>1.5 mi cores)
- High reliability of all components

Blue Gene/Q

- Development: 2007-2012
- PowerPC® A2 core, 16 core/64 thread SoC, 1.6 GHz
- Memory: 16 GB/node
- Speculative execution, sophisticated L1 prefetch, transactional memory, fast thread handoff, compute + IO systems
- Biggest installed system (LLNL):
  - *96 racks*

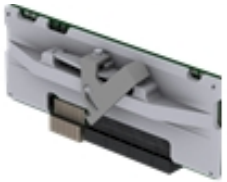
Blue Gene/Q design

1. Chip
16 cores

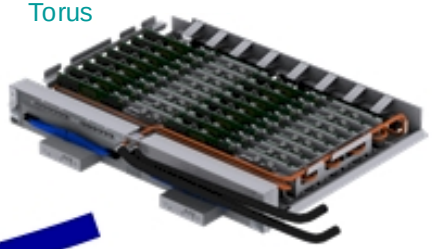


2. Module
Single Chip

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



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

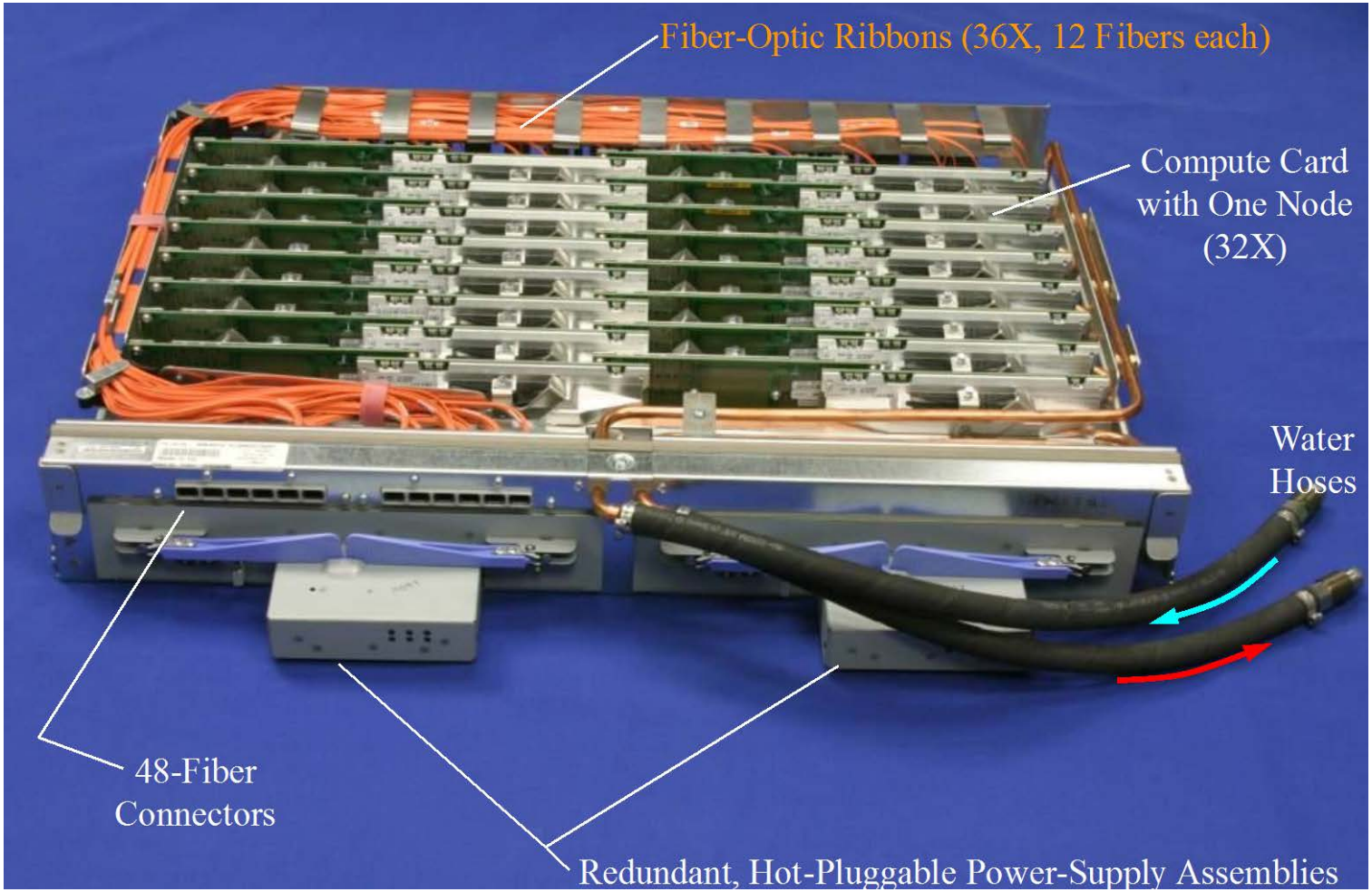
5b. I/O drawer (1,2 or 4 per rack):
* 8 I/O cards @ 16 GB,
* 8 PCIe gen2 x8 slots (IB, 10GbE)

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

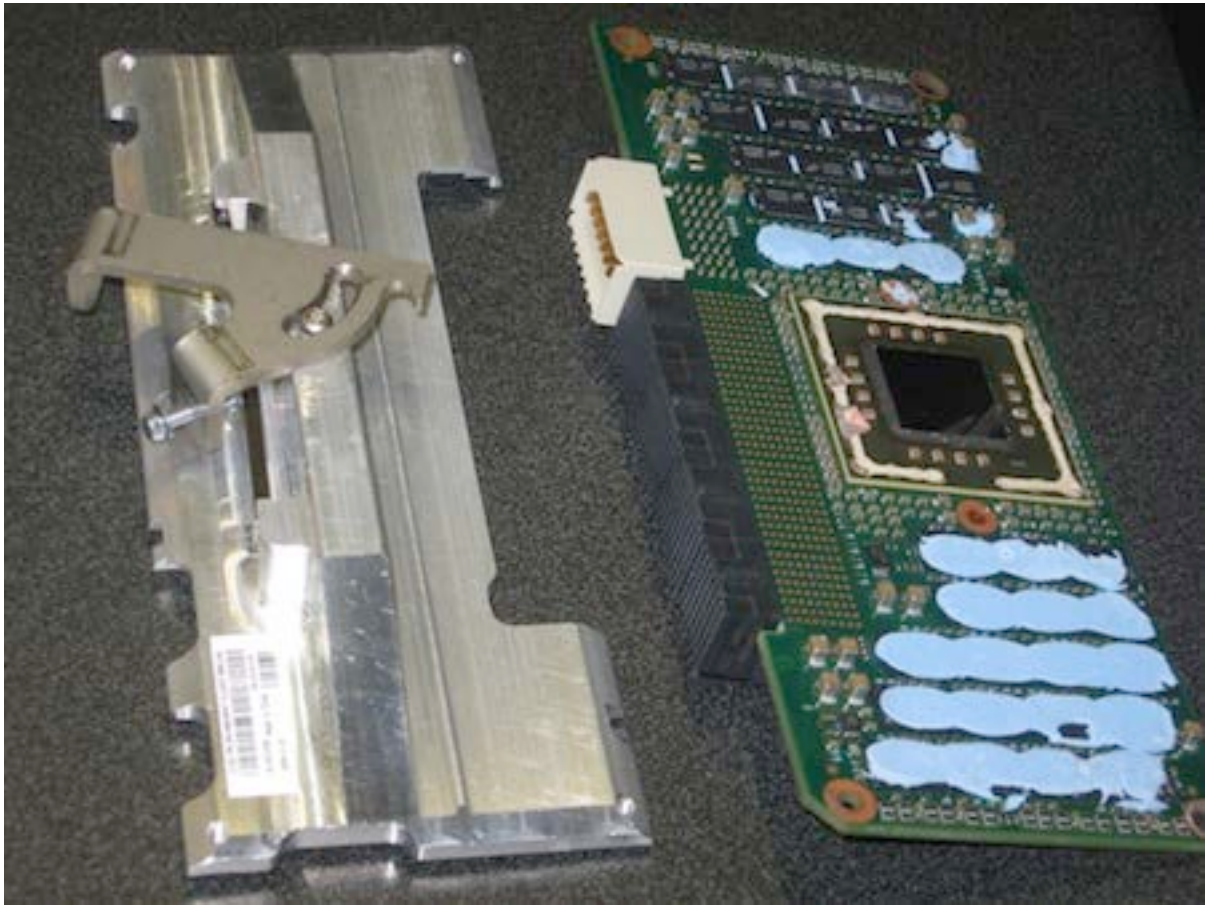
Source: IBM

Blue Gene/Q node card



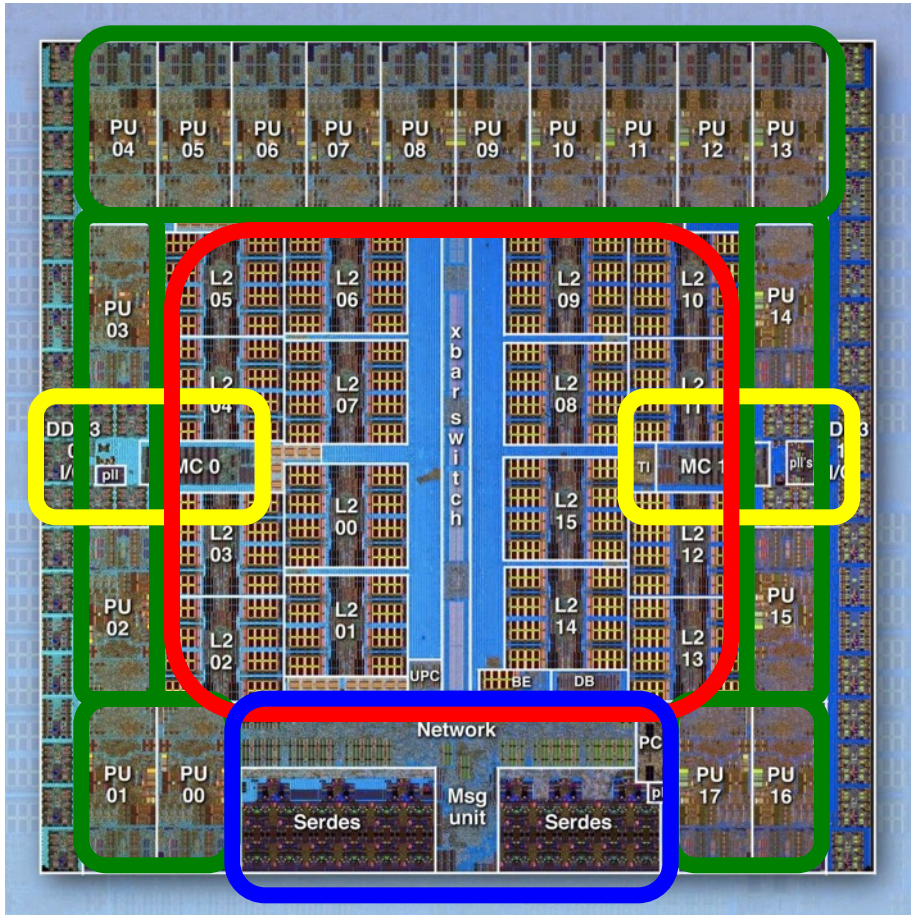
Source: IBM

Blue Gene/Q compute card



Source: Top500.org

Blue Gene/Q chip tomography



16+1+1 processing units

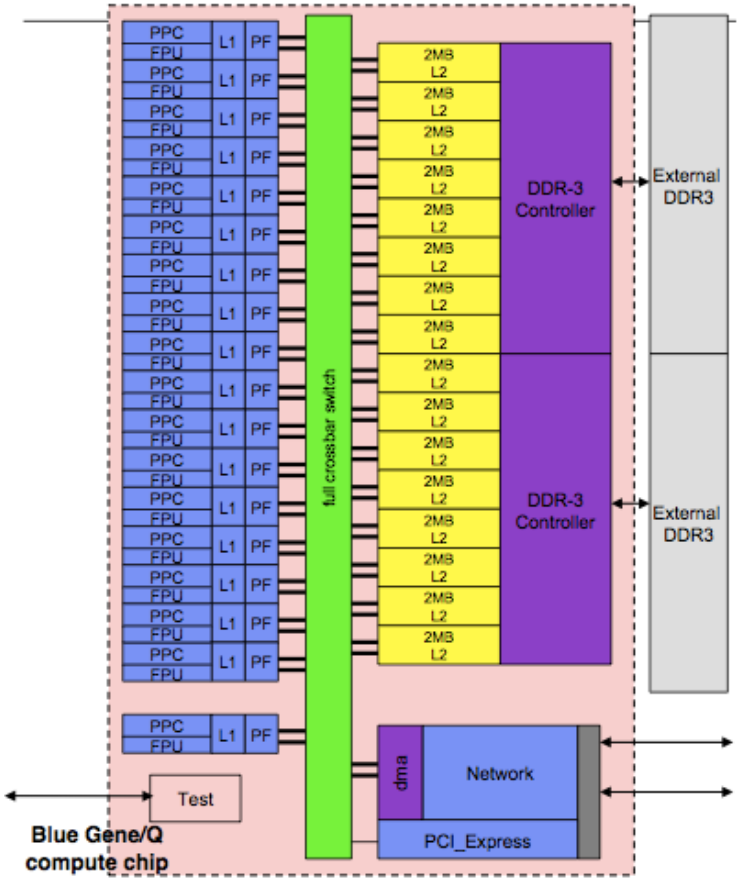
Two memory controller

L2 cache + crossbar switch

On-chip network

Source: IBM

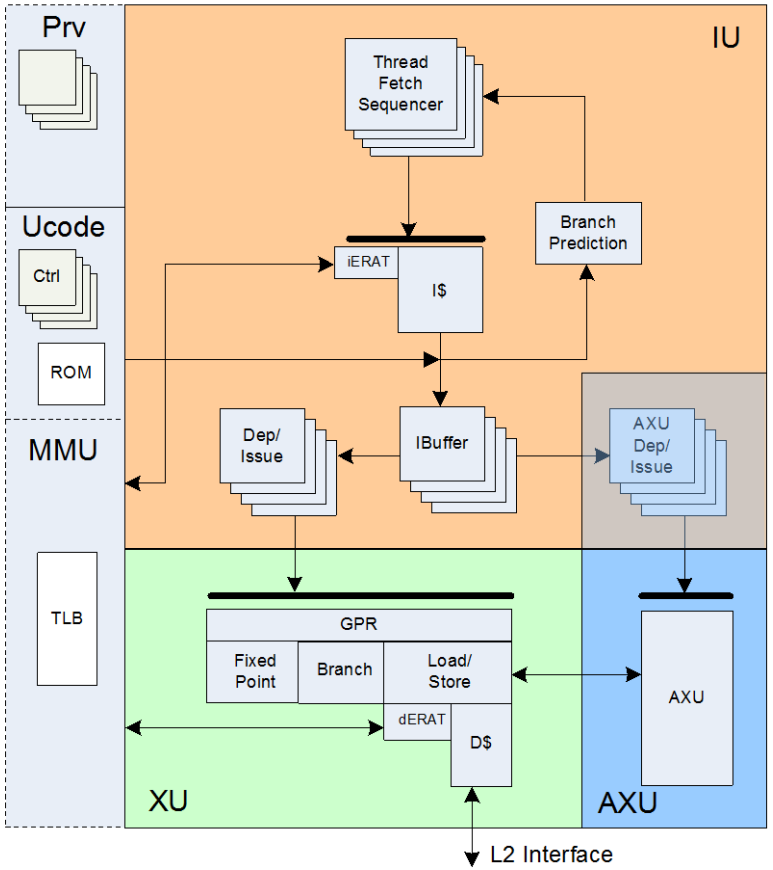
Blue Gene/Q chip architecture



- 16+1 core SMP @ 1.6 GHz
  - *each core is 4-way hardware threaded*
  - *2-way concurrent issue*
- Transactional memory and thread level speculation
- Quad floating point unit on each core (QXP)
  - *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 (node to node)
  - *each at 2.0 GB/s (5D-Torus)*
- One I/O link at 2.0 GB/s
- 16 GB memory / node
- ~60 watts max chip power consumption

Blue Gene/Q: PowerPC A2 processor core

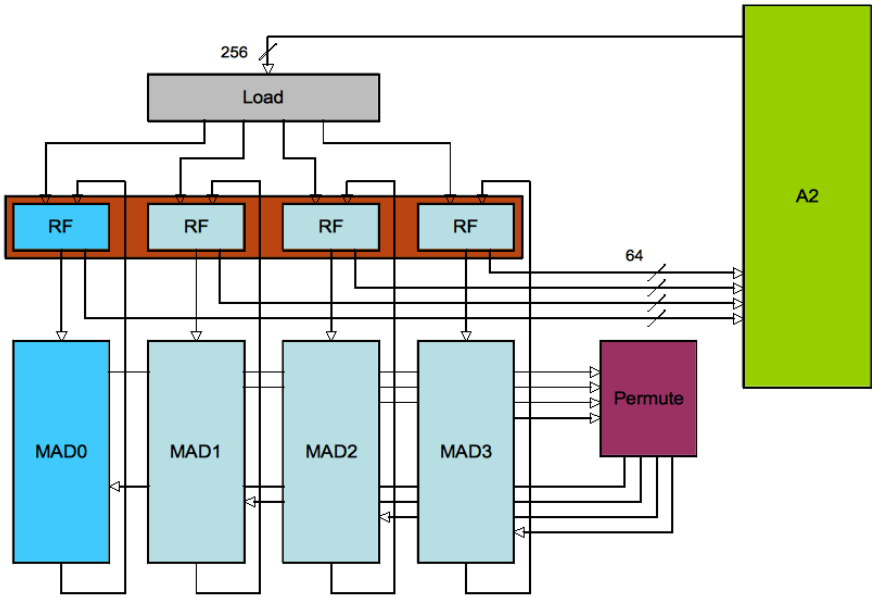
- Simple core
  - *designed for excellent power efficiency and small footprint*
- Embedded 64bit PowerPC compliant
- 4 SMT threads typically get a high level of utilization on shared resources
  - *full register set for every thread*
- 1.6 GHz @ 0.74V
- AXU port allows unique BG/Q floating point unit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



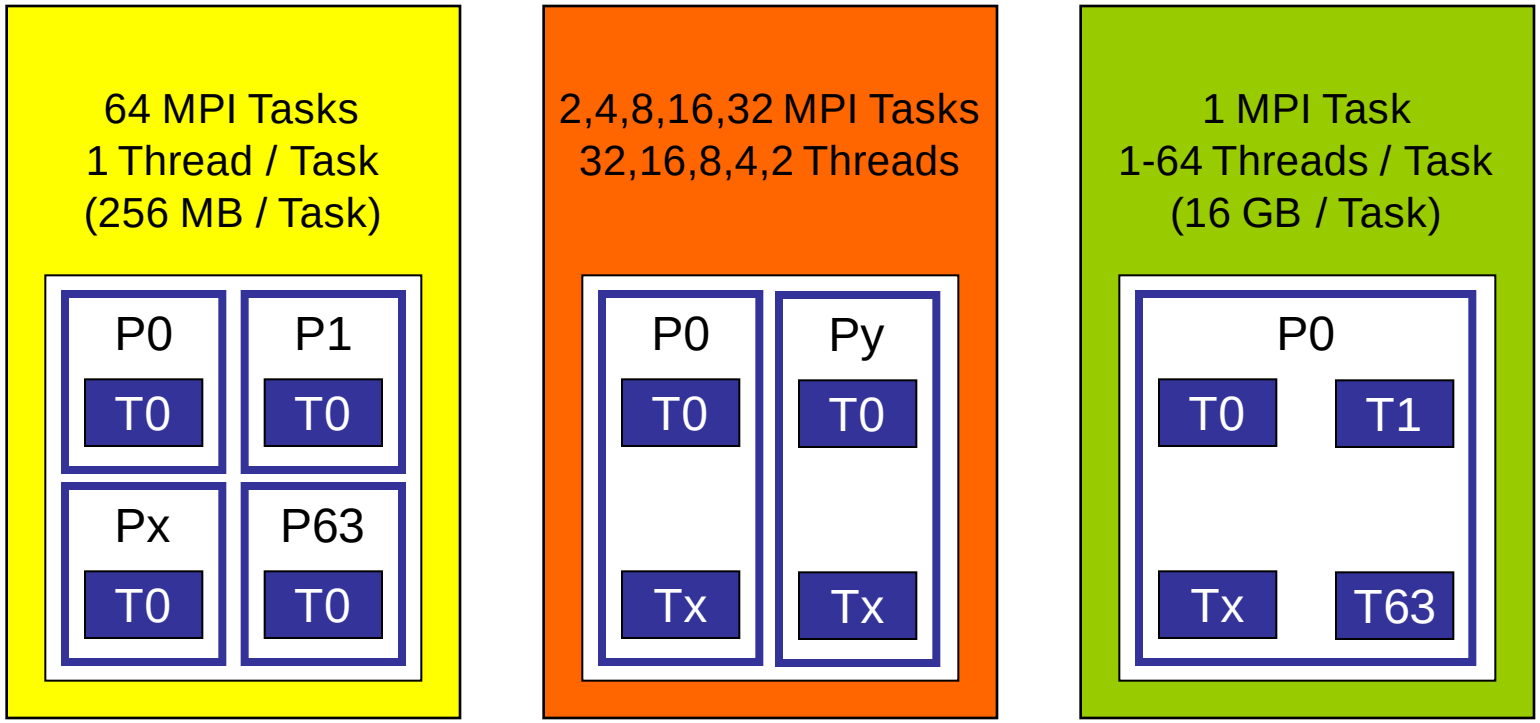
Source: IBM

Quad floating Point eXtension unit (QPX)

- 4 double precision pipelines (64 bit):
  - *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 4 FMA / cycle
  - 12.8 GFlops @ 1.6 GHz



Execution Modes in BG/Q



Blue Gene/Q Chip: the 17th core

RAS Event handling and interrupt off-load

- Reduce OS noise and jitter
- Core-to-Core interrupts when necessary

Common IO Client Interface

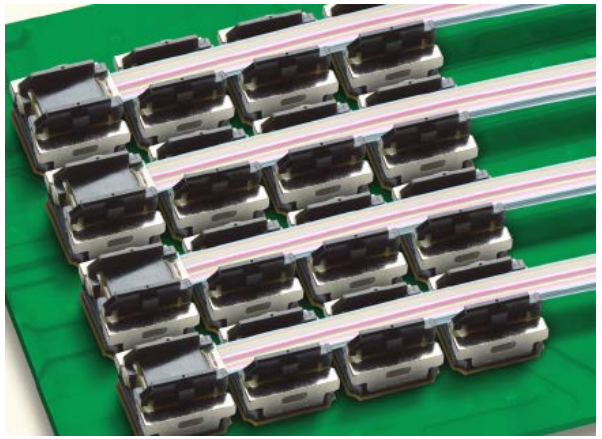
- Asynchronous I/O completion hand-off
- Responsive CIO application control client

Application Agents: privileged application processing

- Messaging assist, e.g. MPI pacing thread
- Performance and trace helpers

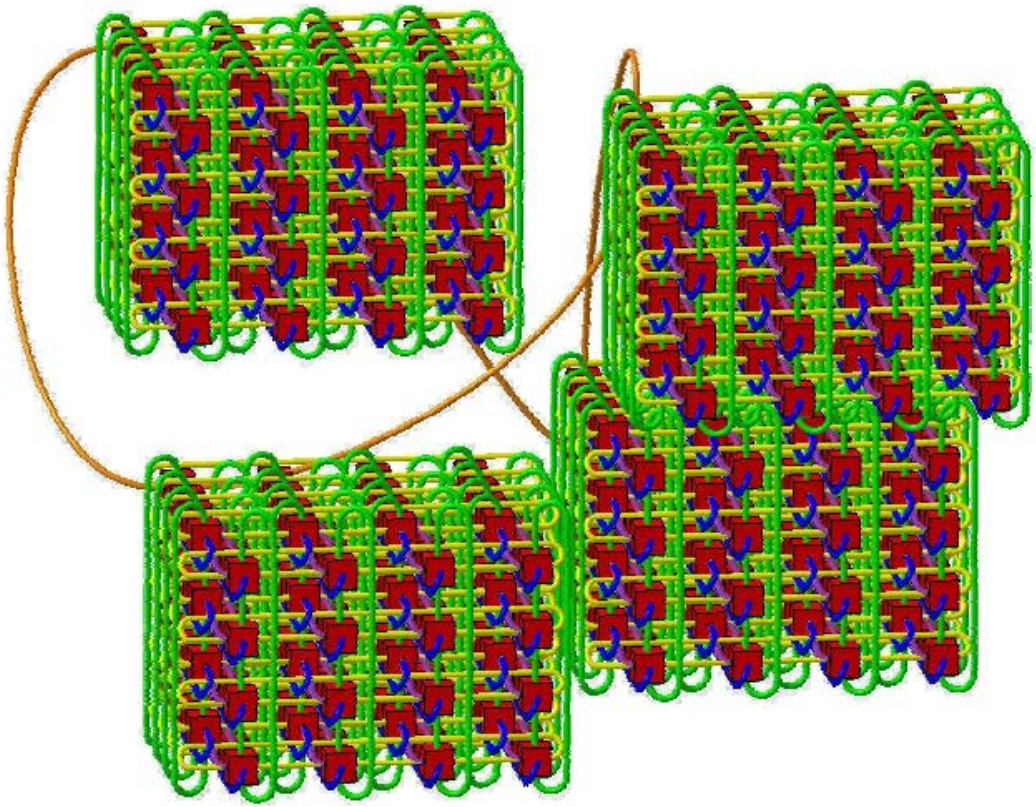
Blue Gene/Q: Network Architecture

- 11 bi-directional chip-to-chip links
  - 2 GB/s bandwidth, about 40 ns latency
- 5-dimensional torus topology (ABCDE)
  - 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 5D Torus Network



Blue Gene/Q: Torus configuration

- BG/Q nodes form a 5D torus
  - Dimension labels: ABCDE T
  - Nodecard: 2x2x2x2 x2 (smallest partition)
  - Midplane: 4x4x4x4 x2 (512 compute nodes)

    Dimensions A-D are cabled to other midplanes
  - 5th dimension stays within the node card
  - 6th dimension is the CPU# within a node

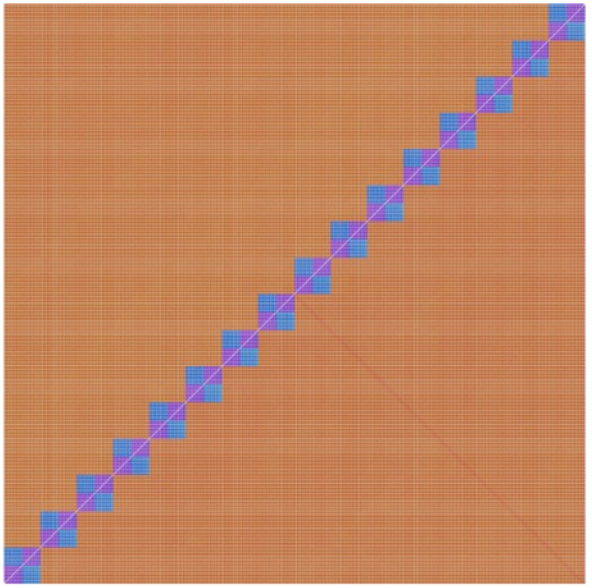
- 28 rack configuration (JUQUEEN):
  - 7 rows, 4 columns
  - Torus size in midplanes: 2x7x2x2
  - Torus size in nodes: 8x28x8x8x2

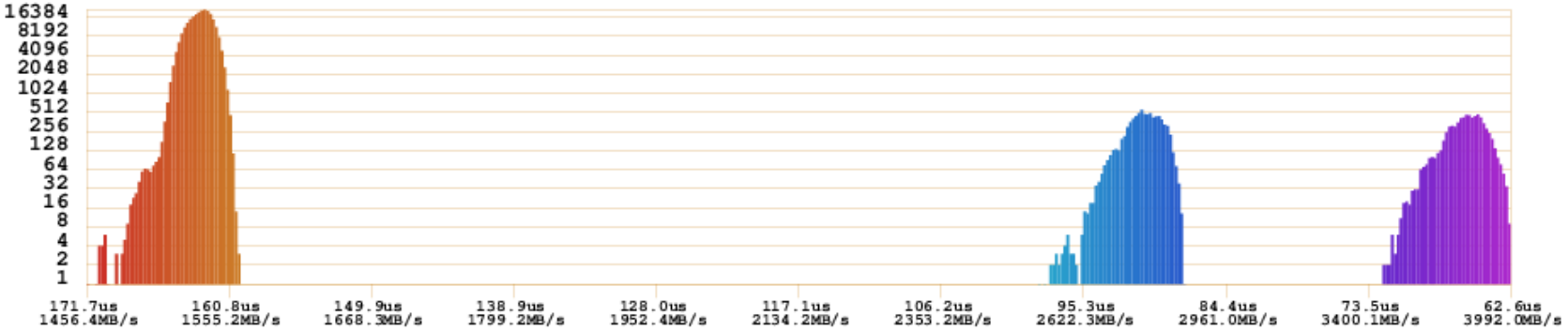
Torus node to MPI mapping

- 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
- Processes need to be placed to minimize 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 MPIX_Cart_comm_create() (IBM extension to MPICH2)
  - 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 (on a midplane)
- Bandwidth distribution
  - *Intra-node communication*
  - *Communication via link E*
  - *Communication via link A, B, C, D*

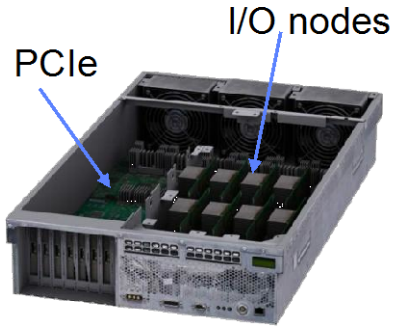




Blue Gene/Q: I/O Architecture

- **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

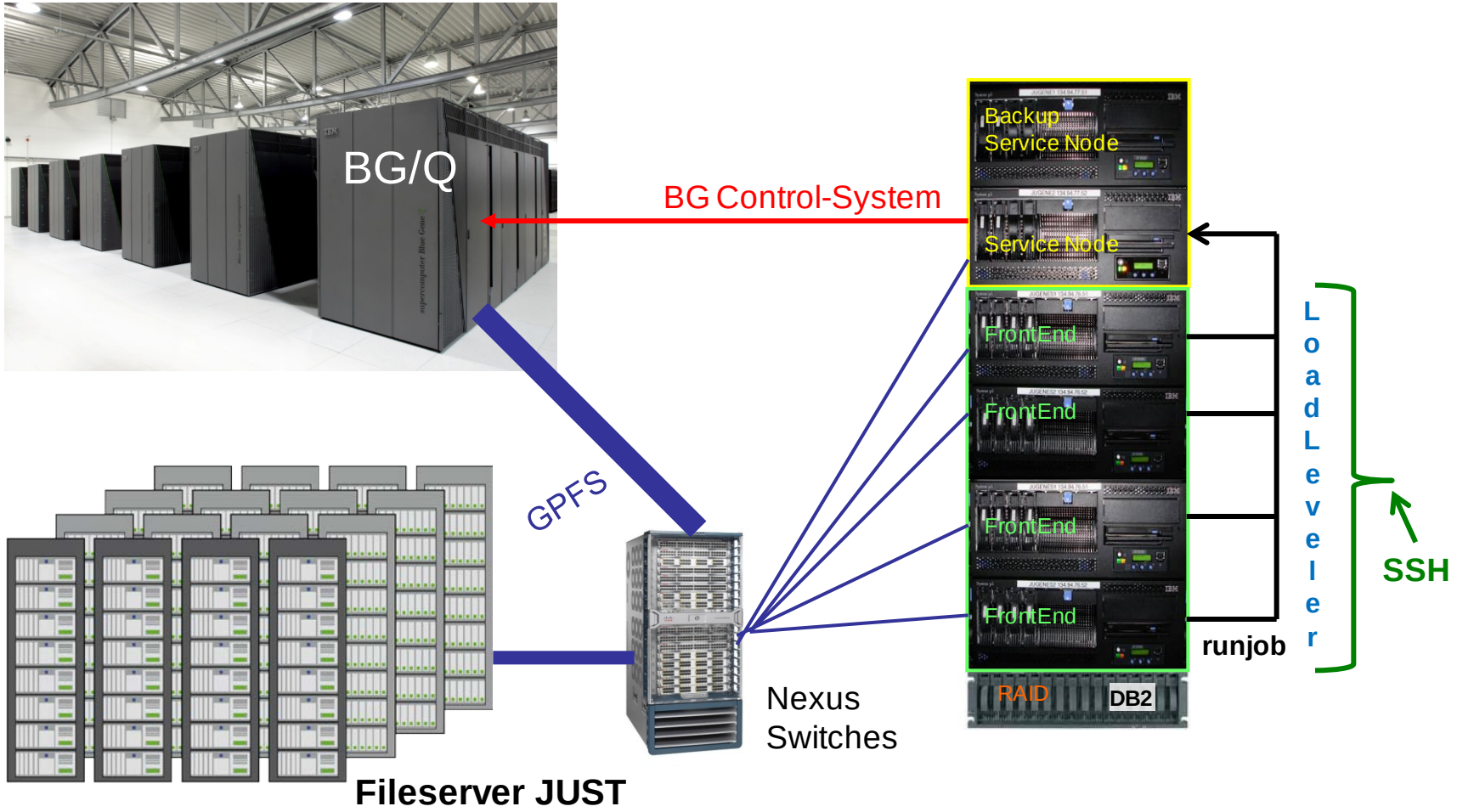
28 Racks Blue Gene/Q

- 28,672 compute nodes (16 cores, 16 GB memory)
- 458,752 cores / >1.8 mi 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 (administration)

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

JUQUEEN Environment



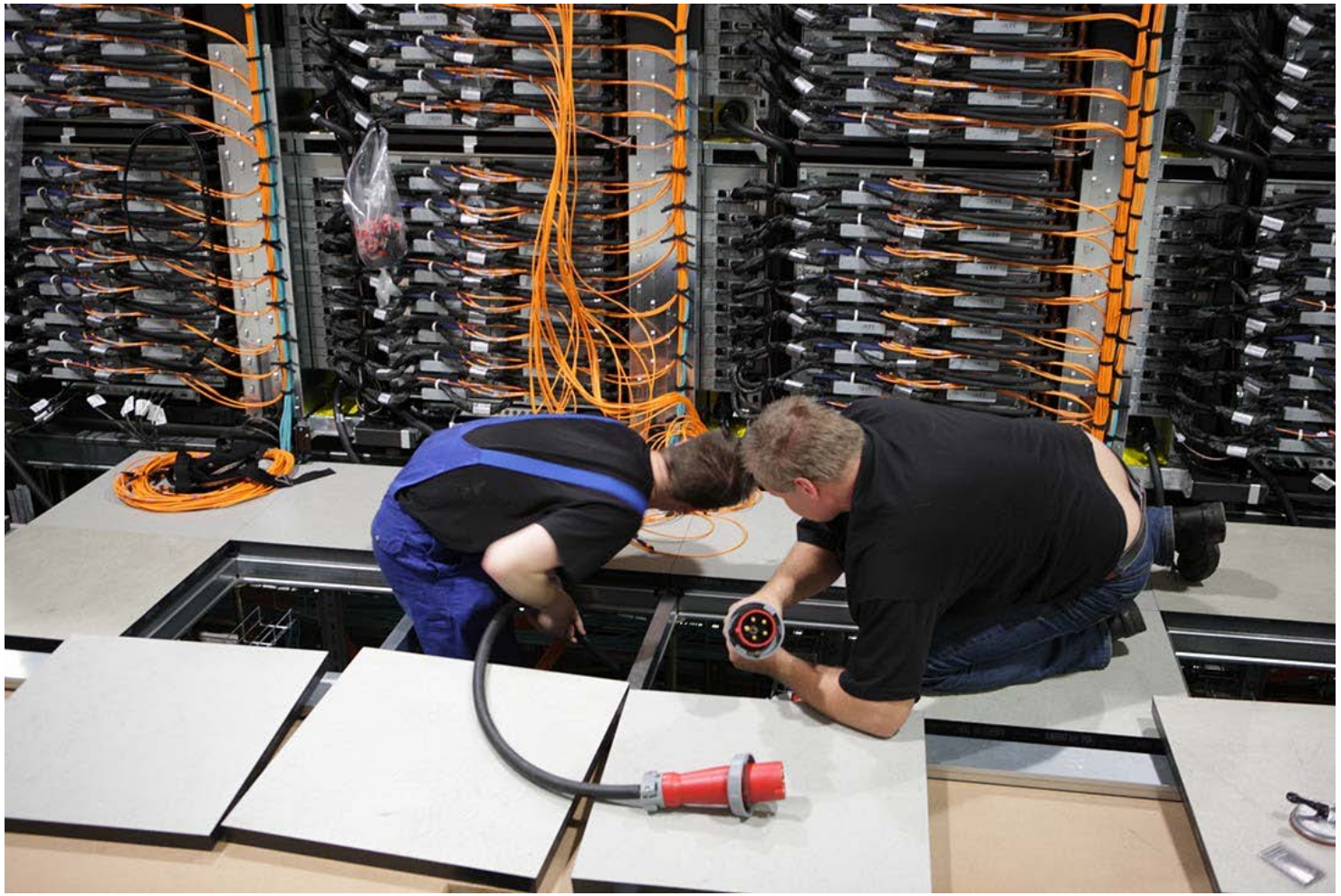
JUQUEEN: Installation pictures

JUQUEEN: Installation pictures

JUQUEEN: Installation pictures

JUQUEEN: Installation pictures

JUQUEEN: Installation pictures



JUQUEEN: Installation pictures

JUQUEEN: Installation pictures

JUQUEEN: Installation pictures

JUQUEEN: Installation pictures

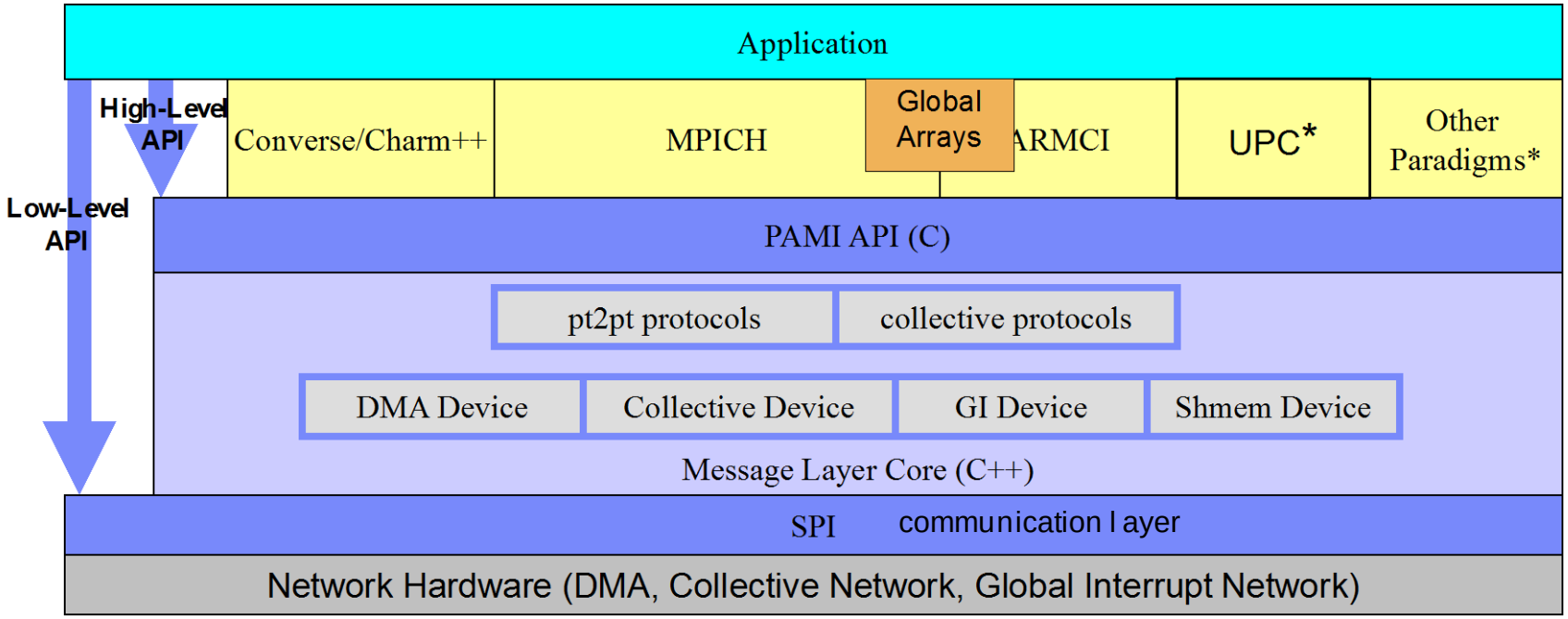

IBM
Blue Gene/P
ENE

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: PAMI
Parallel Active Message Interface

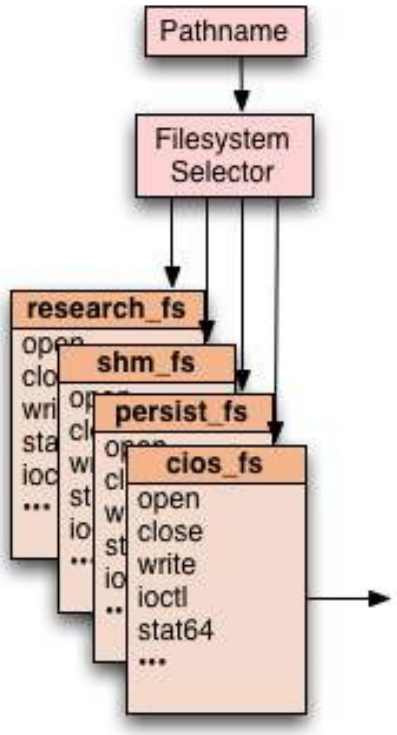


MPIX: Blue Gene/Q specific extensions to MPI

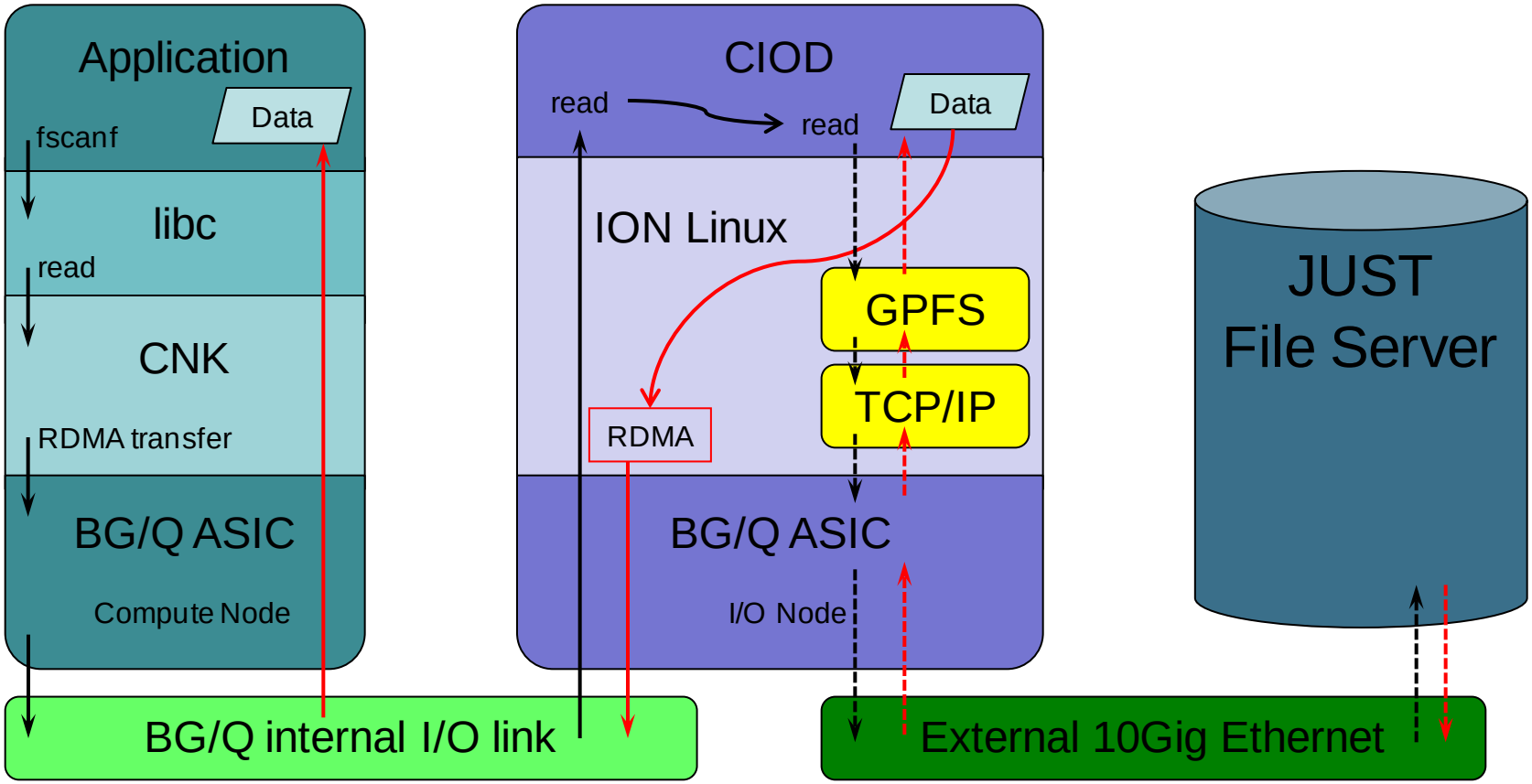
- int MPIX_Cart_comm_create (MPI_Comm *cart_comm)
  - *This function creates a six-dimensional (6D) Cartesian communicator that mimics the exact hardware on which it is run. The A, B, C, D, and E dimensions match those of the block hardware, while the T dimension is equivalent to the ranks per node argument to runjob.*
- Changing class-route usage at runtime
  - *int MPIX_Comm_update(MPI_Comm comm, int optimize)*
- Determining hardware properties
  - *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: System Calls

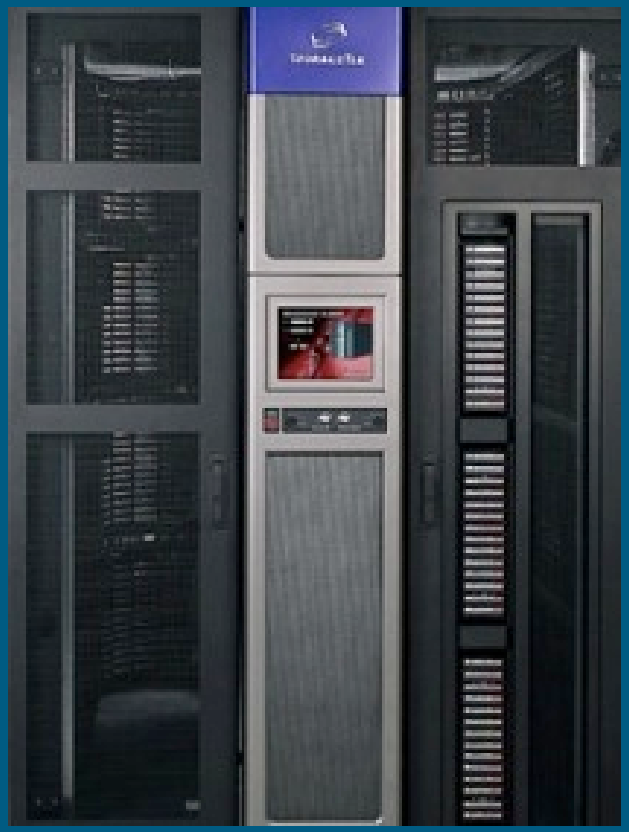
- Many (but not all) common Linux syscalls work on BG/Q *(see Application Development Redbook)*
- Support glibc 2.12.2
- Real-time signals support
- Low overhead syscalls
  - Only essential registers are saved and restored
- Pluggable File Systems
  - Allows CNK to support multiple file system behavior and types
  - Different simulator platforms have different capabilities
  - Shared Memory



Blue Gene/Q: File I/O



JÜLICH
FORSCHUNGSZENTRUM

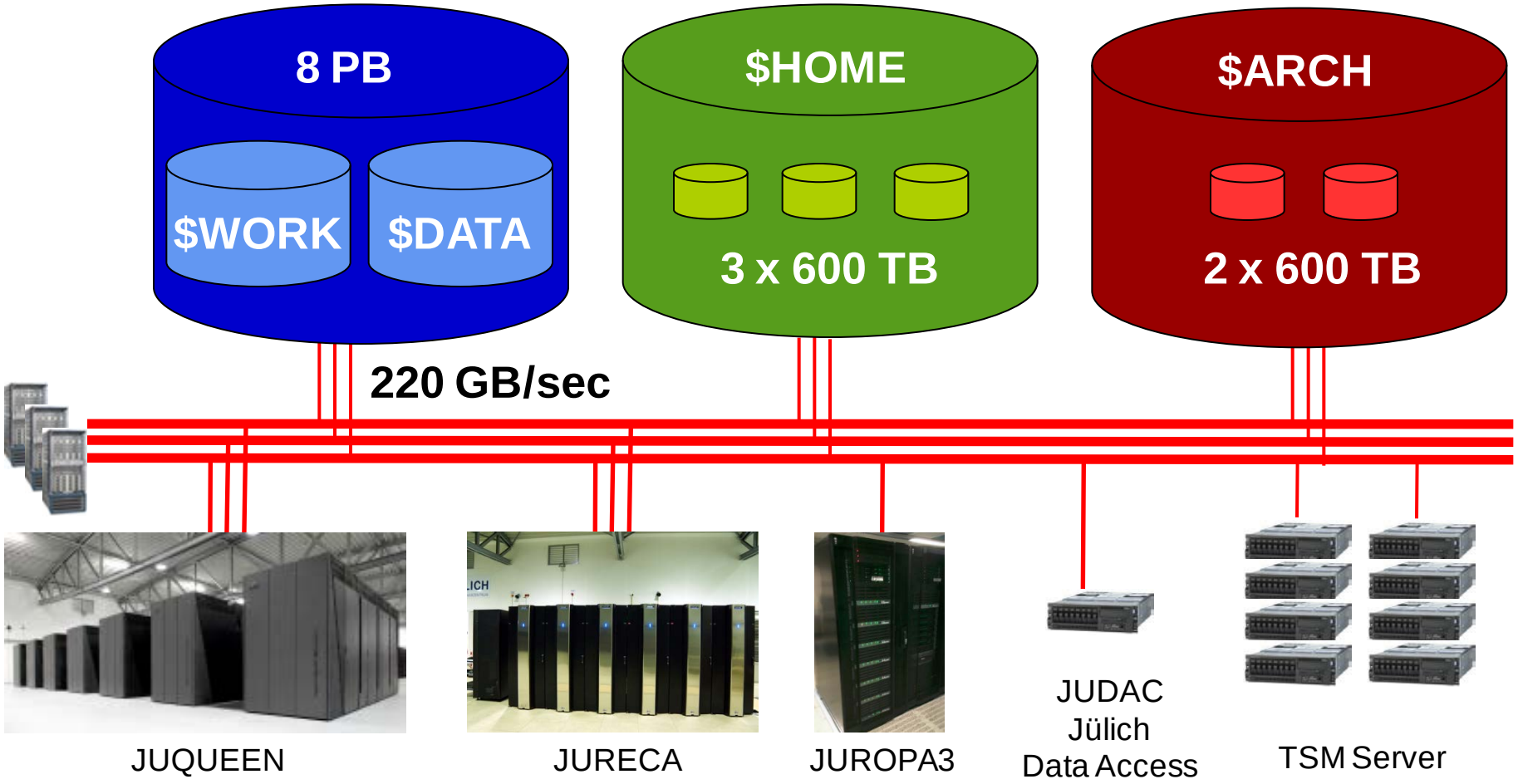


Data Storage
JUST Fileserver

GPFS - General Parallel File Systems

- The user can store his datasets in different file systems:
  - *$HOME*
  - *$WORK*
  - *$ARCH*
- It is highly recommended to access files always with the help of these variables.

Jülich Storage Server JUST for all supercomputers (Q1/2016)



Filesystems – $HOME

- Multiple filesystems: */homea, /homeb, /homec*
- Daily backup
- For regularly used files and applications
- Smaller files which are used for the current computation
- Quota: 10 TB space
  4 million files per group
- List quota information for group, user [and co-workers] with *q_dataquota [-l]*

Filesystems – $WORK

- Fastest file system (> 150 GB/s)
- No backup
- Files will be deleted 90 days after the last usage
  (access time is not updated on $WORK for performance reasons)
- Quota: 30 TB and
  3 million files per group
- List quota information with *q_dataquota [-l]*
- For large files which are used or generated during the computation
- Relevant files should be copied to $HOME or $ARCH to get backed up

Filesystems – $ARCH

- Daily backup
- All files will be migrated to tape
- NOT available on the Blue Gene compute nodes
- Large, not recently used files
- No space quota,
  max. 2 million files per group
- Usage of TAR or ZIP archives is highly recommended!

E.g.: Restore with 100 MB/s *but* up to 90s per file opening

- *1 x 200 GB → 30-40 min*
- *1000 x 200 MB → 25 h*
- *100000 x 2 MB → 104 d (3.5 month)*

Recommendations for $ARCH

- Don’t use touch to trigger recall of migrated data

  → *change of timestamp will force a new backup and a new migration even if data is used read only*

- Use **tar**-archives to reorganize a lot of small files to avoid long access times

  tar -cvf $ARCH/<archive-filename> <filelist> …

Recommendations for $ARCH

- Don't change the path of large archive entities
  by renaming or insertion/deletion of subdirectories
  → *will produce high system overhead, because*
  all data affected by change will be recalled, backed up and newly migrated

  Don't: *mv $ARCH/project $ARCH/project_old*

  Might prevent backup of really new data in the entire file system and wastes system resources like CPU time, tape storage, and tape units.

Data Transfer - Hints

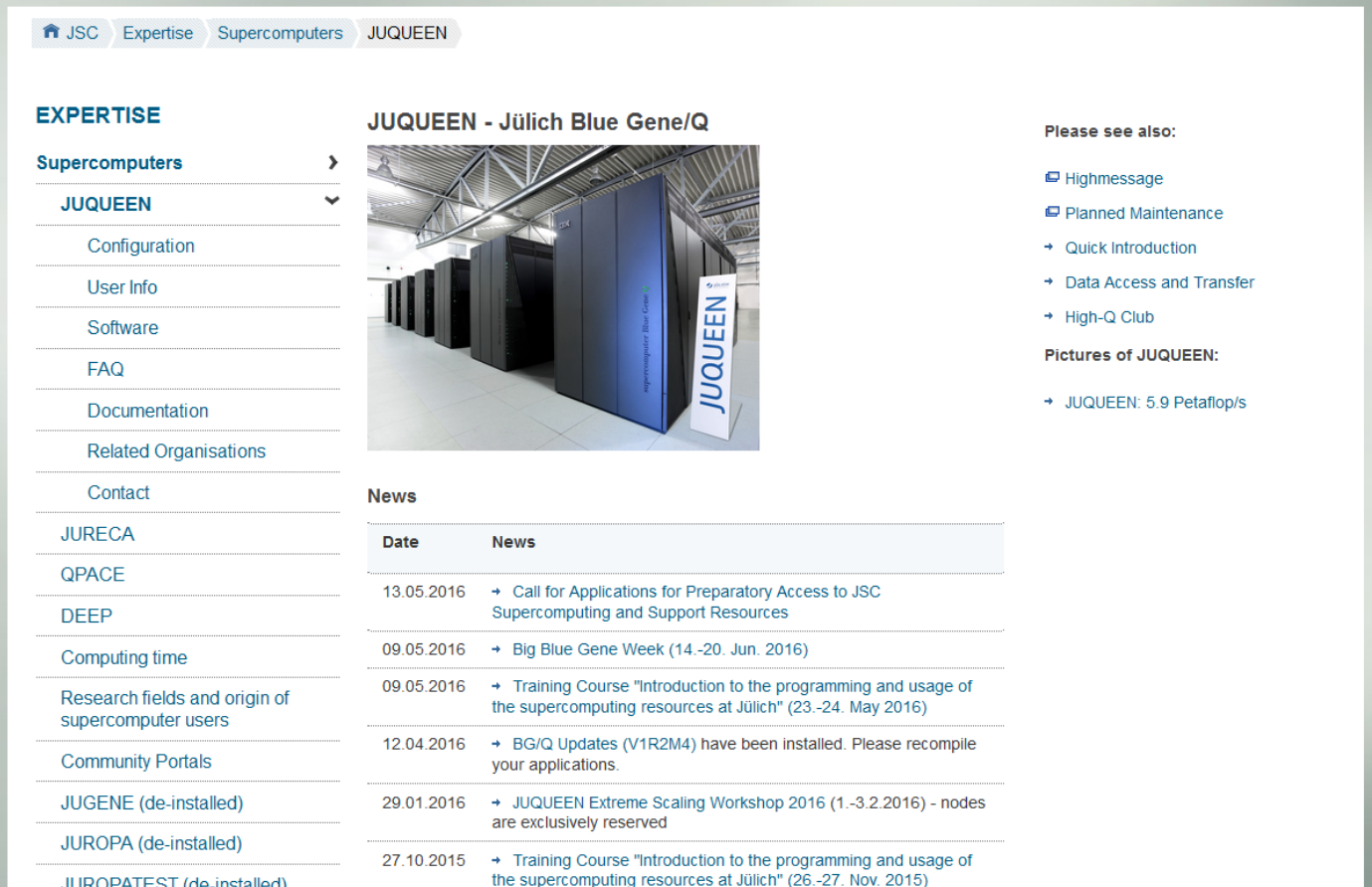
- Use **JUDAC@fz-juelich.de** for data transfer (Linux Server)
- Too many accesses (ssh or scp) within a short amount of time will be interpreted as intrusion and leads to automatic disabling the origin system at the FZJ firewall.
- To transfer multiple files in a single session use "**scp –r**", which transfers a whole directory

  Example:

  scp –r $HOME/project <myuserid>@<mysystem>:<mydir>

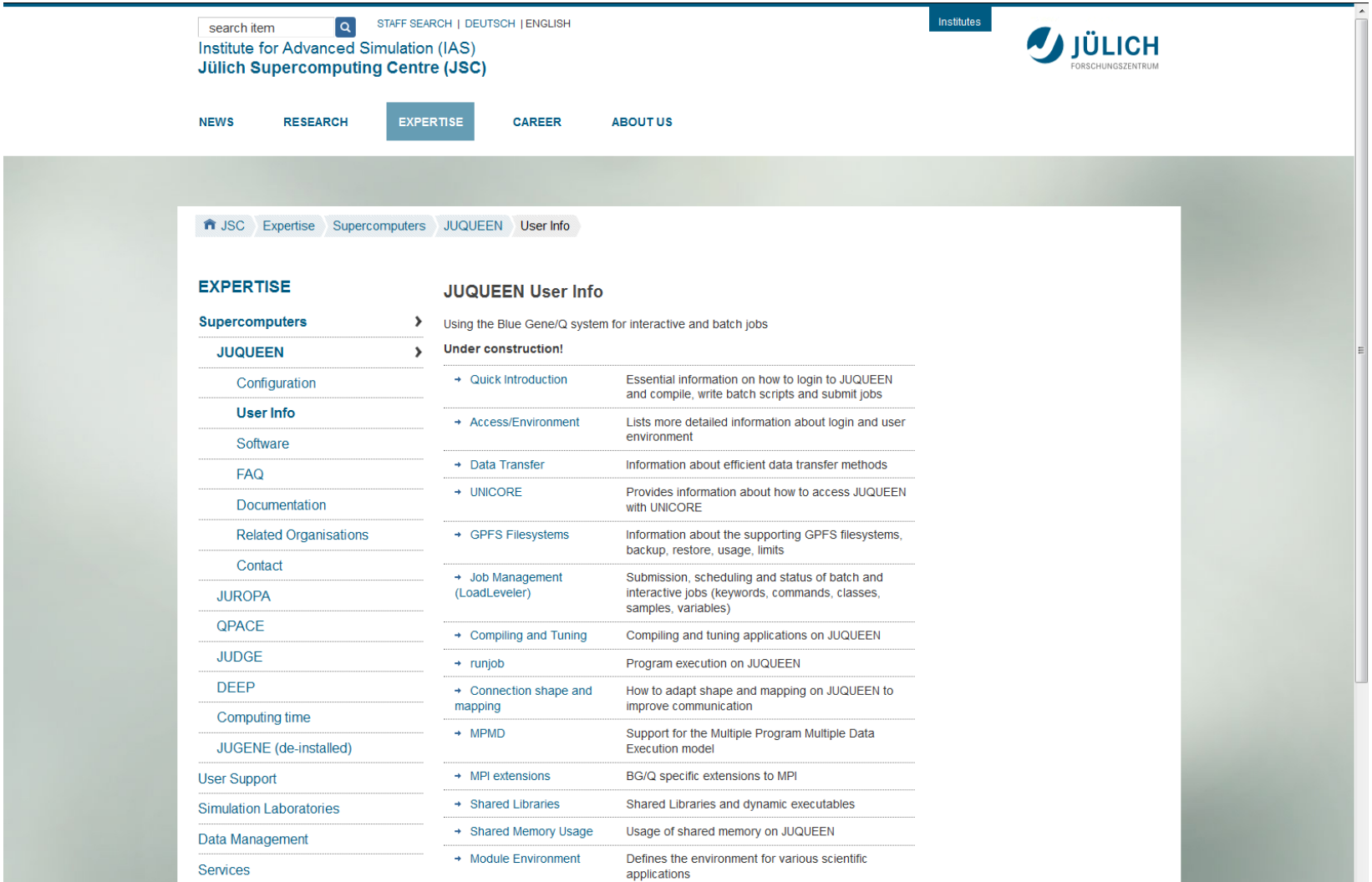
Documentation

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



Documentation – User info

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



Support Addresses

- sc@fz-juelich.de — application support *(ticket is generated)*
- bg-adm@fz-juelich.de — system administrator

Questions?