JÜLICH
FORSCHUNGSZENTRUM

Blue Gene/Q

Hardware Overview

02.02.2015 | Michael Stephan

Mitglied der Helmholtz-Gemeinschaft

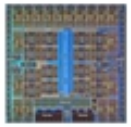
Blue Gene/Q: Design goals

- 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

Mitglied der Helmholtz-Gemeinschaft

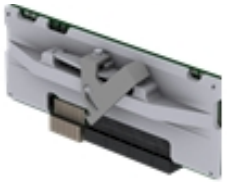
Blue Gene/Q: Design

1. Chip
16 cores

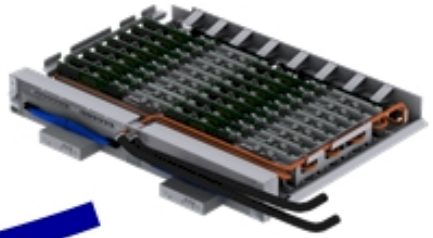


2. Module
Single Chip

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)

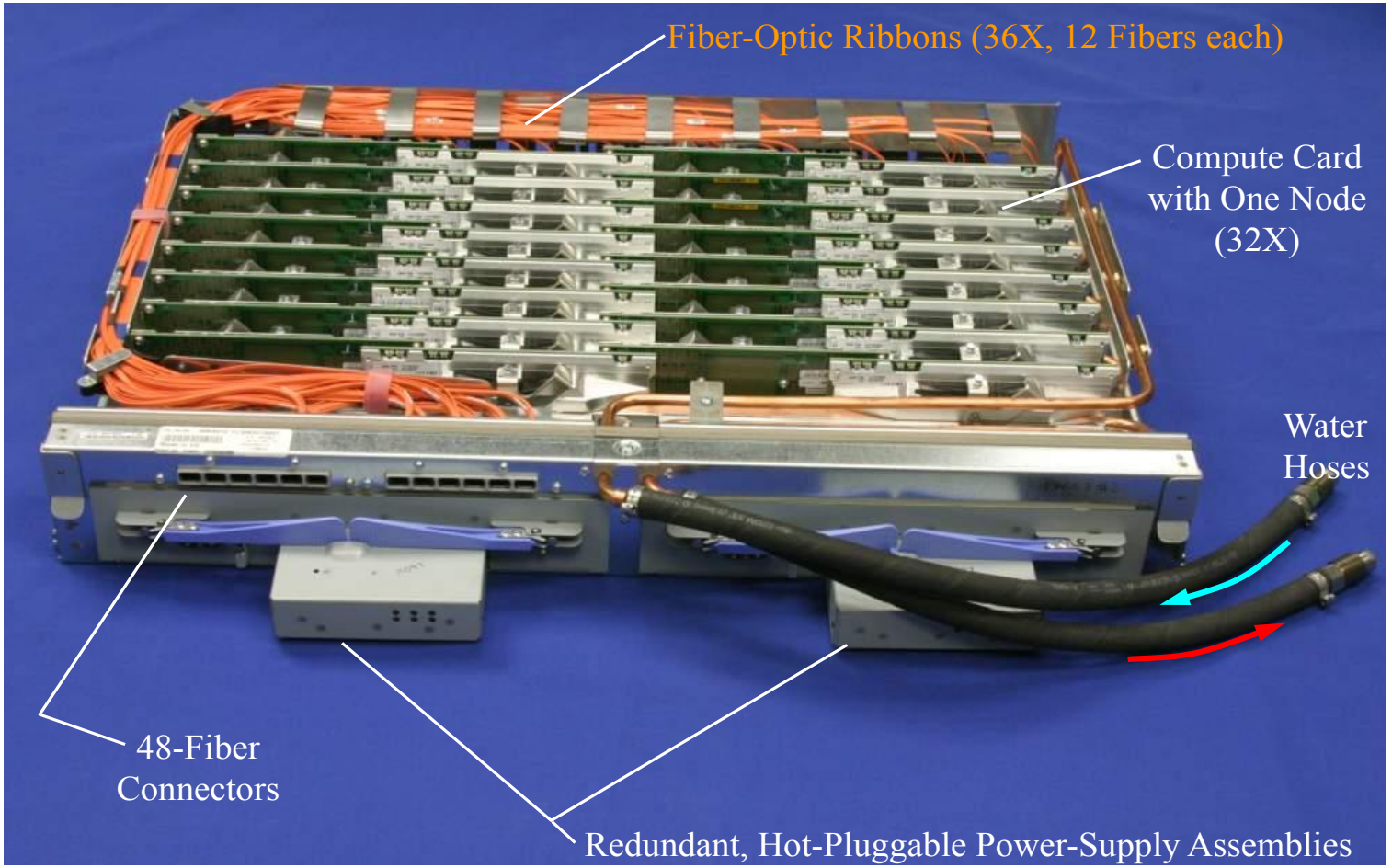
5b. I/O drawer (1,2 or 4 per rack):
* 8 I/O cards @ 16 GB,
* 8 PCIe gen2 x8 slots (IB, 10GbE)

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

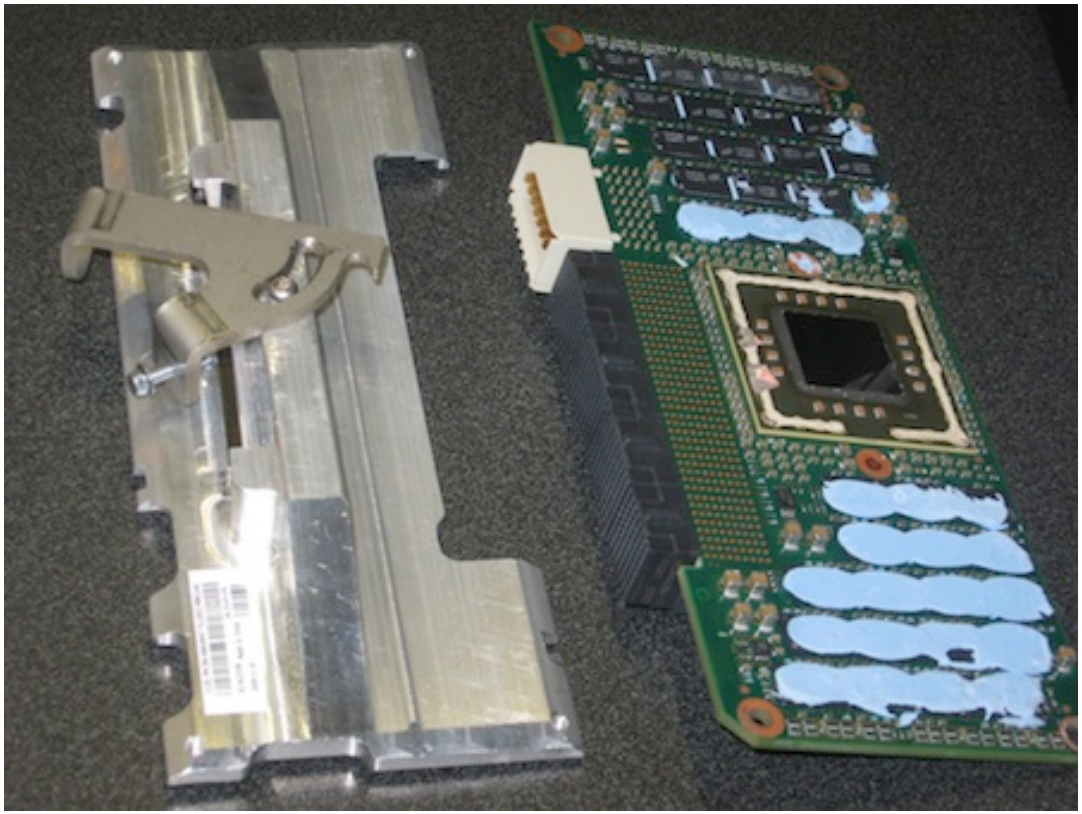
Blue Gene/Q: Node card



Source: IBM

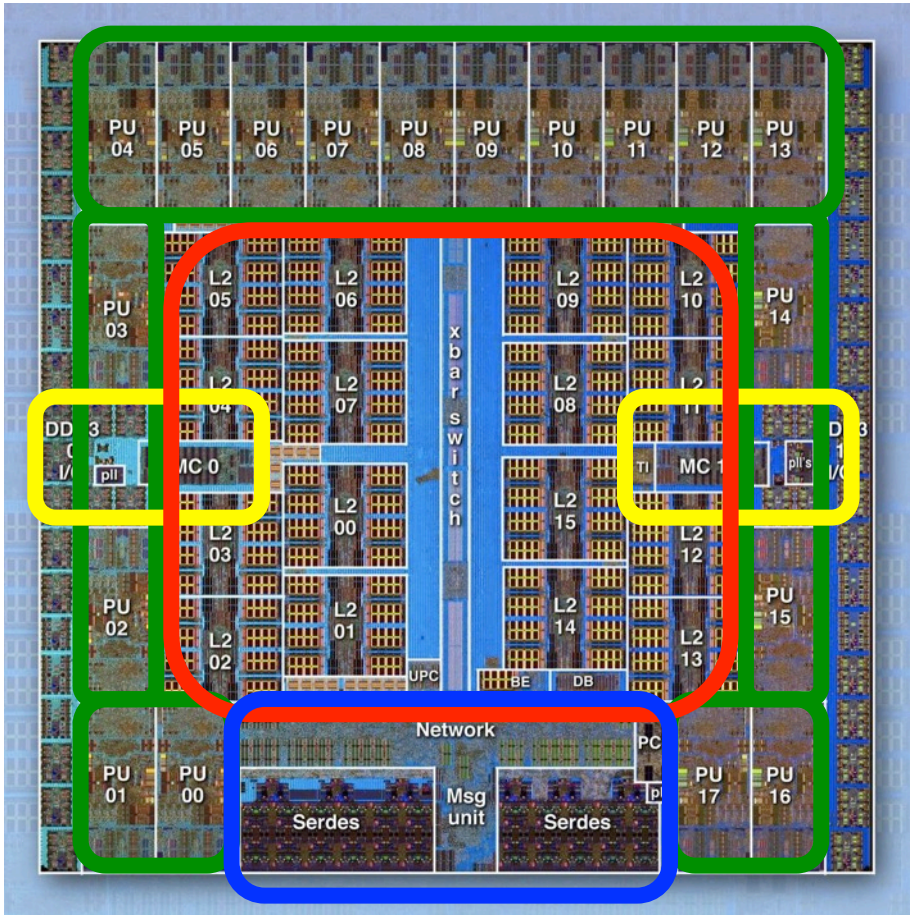
Mitglied der Helmholtz-Gemeinschaft

Blue Gene/Q: Compute card



Source: Top500.org

Blue Gene/Q: Chip tomography



16+1+1 processing units

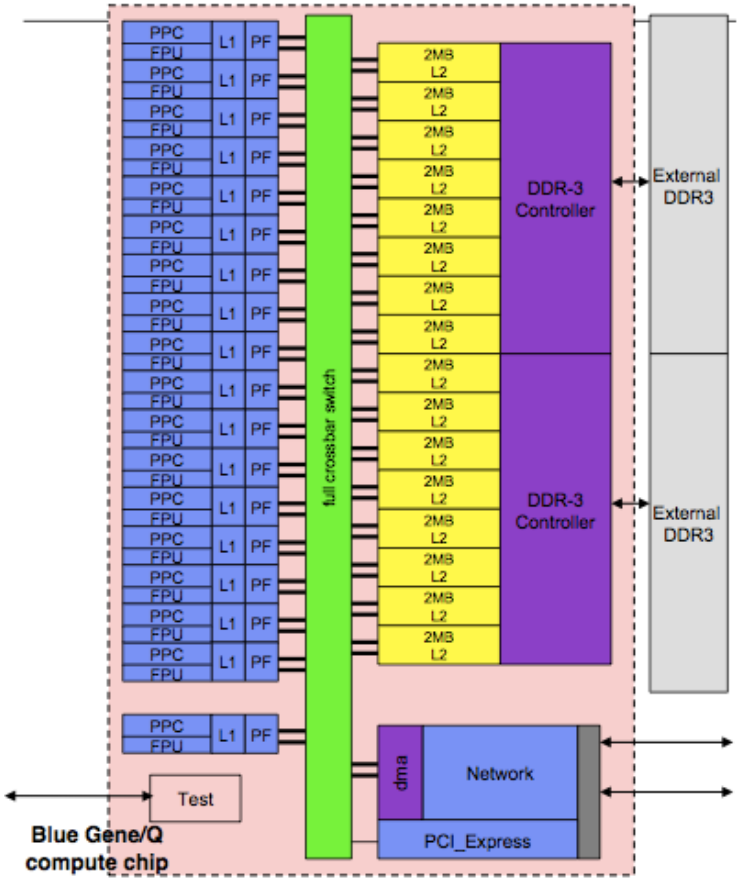
Two memory controller

L2 cache + crossbar switch

On-chip network

Mitglied der Helmholtz-Gemeinschaft

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

Mitglied der Helmholtz-Gemeinschaft

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

Mitglied der Helmholtz-Gemeinschaft

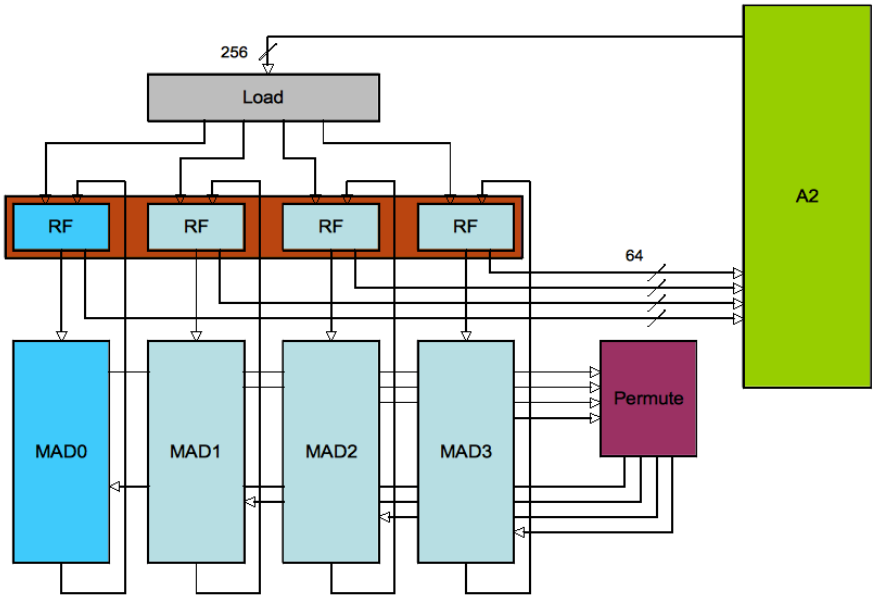
Blue Gene/Q: Advanced processor features

- 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

Mitglied der Helmholtz-Gemeinschaft

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



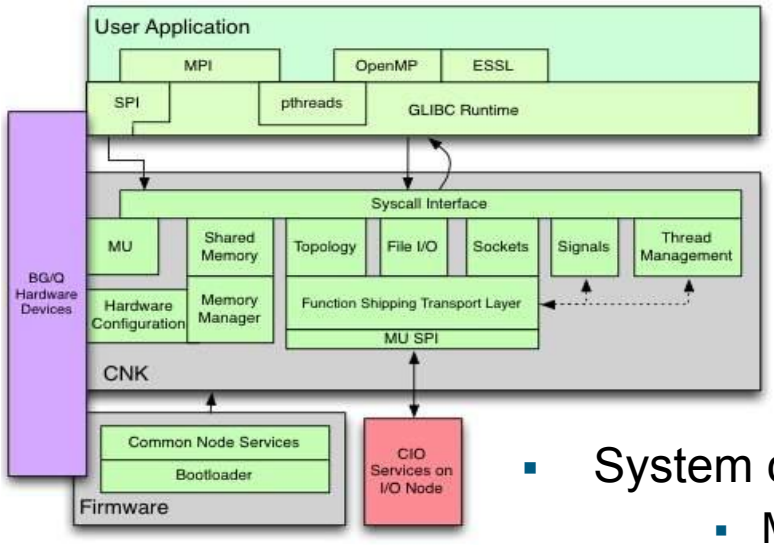
Mitglied der Helmholtz-Gemeinschaft

Blue Gene/Q: 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

Mitglied der Helmholtz-Gemeinschaft

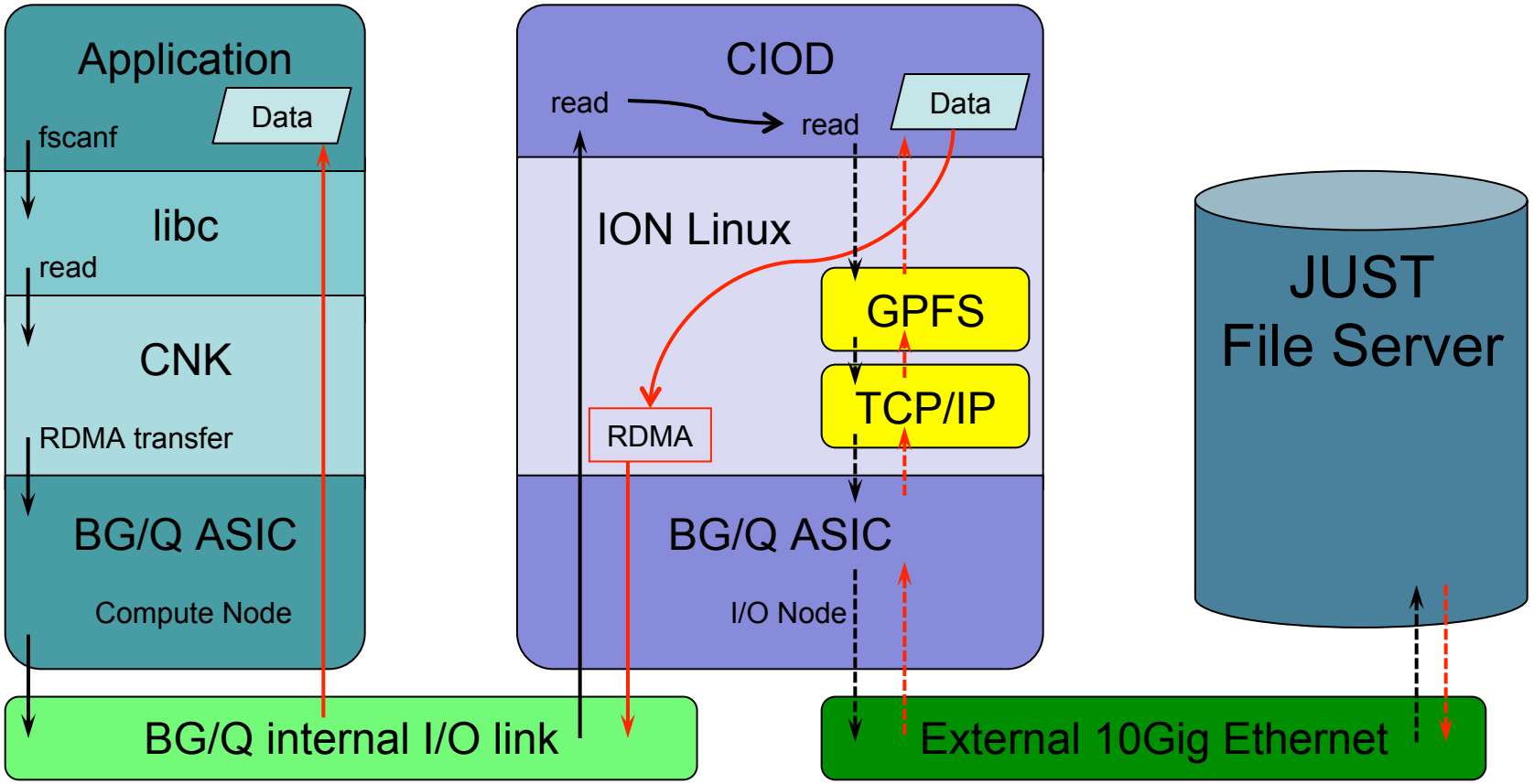
Blue Gene/Q: CNK + System Calls



- CNK (Blue Gene is NOT a cluster)
  - No “normal” Linux on the CN
  - No shell
  - No “real” task/thread scheduler
- System calls
  - Many (but not all) common Linux syscalls work on BG/Q.
  - Support glibc 2.12.2
  - Real-time signals support
  - Low overhead syscalls
    - Only essential registers are saved and restored

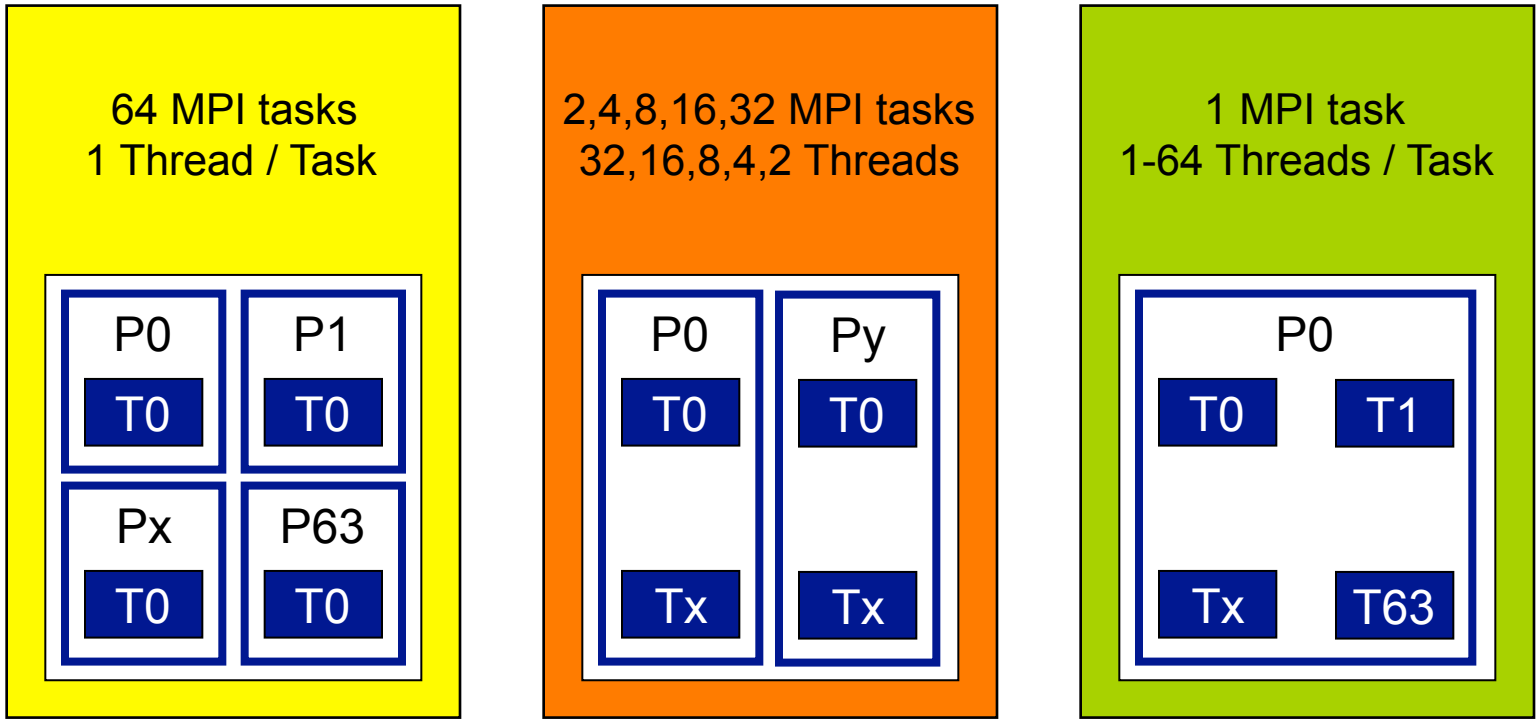
Mitglied der Helmholtz-Gemeinschaft

Blue Gene/Q: File I/O



Mitglied der Helmholtz-Gemeinschaft

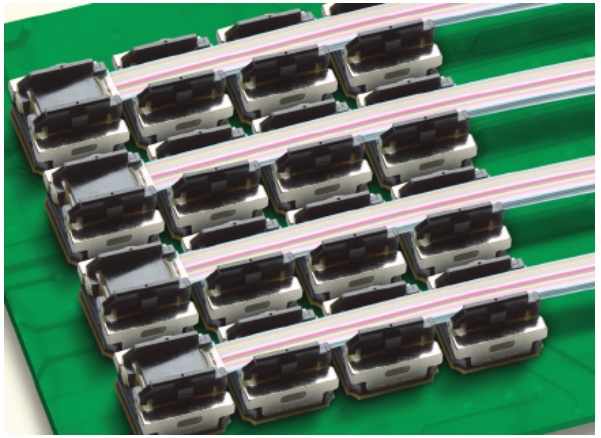
Blue Gene/Q: Execution Modes



Mitglied der Helmholtz-Gemeinschaft

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

Mitglied der Helmholtz-Gemeinschaft

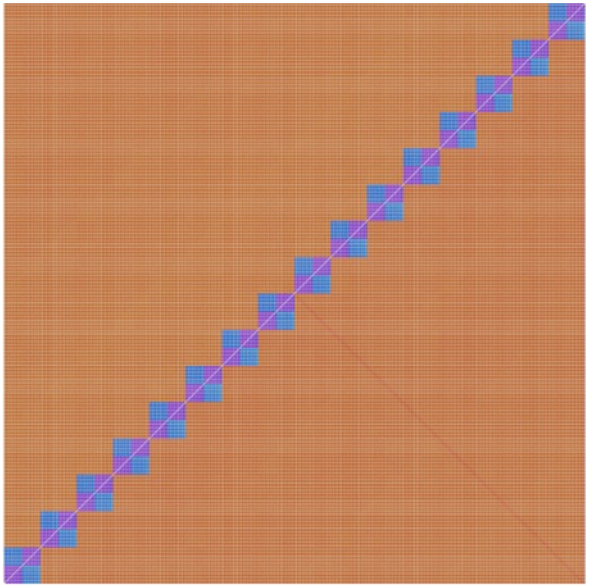
Blue Gene/Q: Torus node to MPI mapping

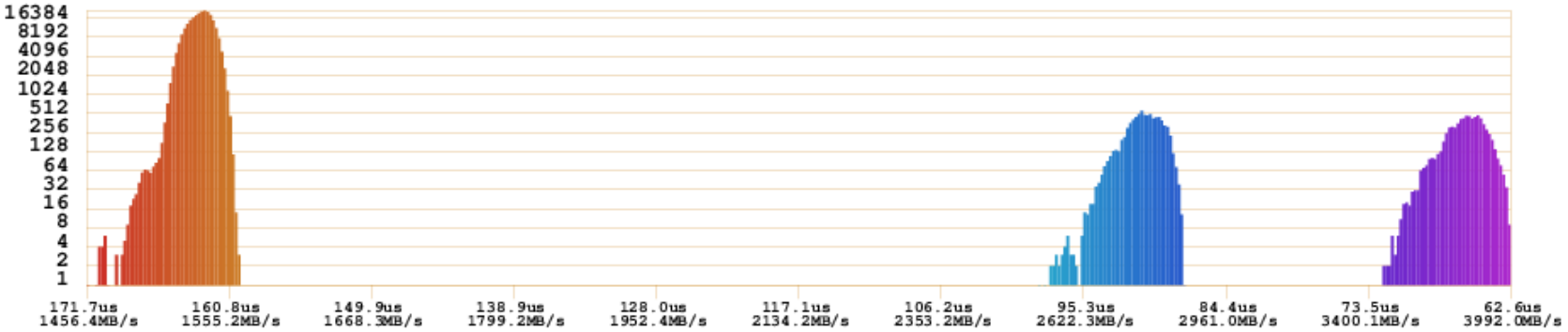
- Block's physical network topology: 6 dimensional ABCDE T
  - local on each node; shared memory communication (fast); T=0..63
  - A,B,C,D,E depend on size of block, e.g. 2×2×4×4×2 for quarter midplane (256 nodes)
- processes need to be placed to minimize maximum load on network-links
  - take advantage 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>)

Mitglied der Helmholtz-Gemeinschaft

Blue Gene/Q: Torus link bandwidth

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





Mitglied der Helmholtz-Gemeinschaft

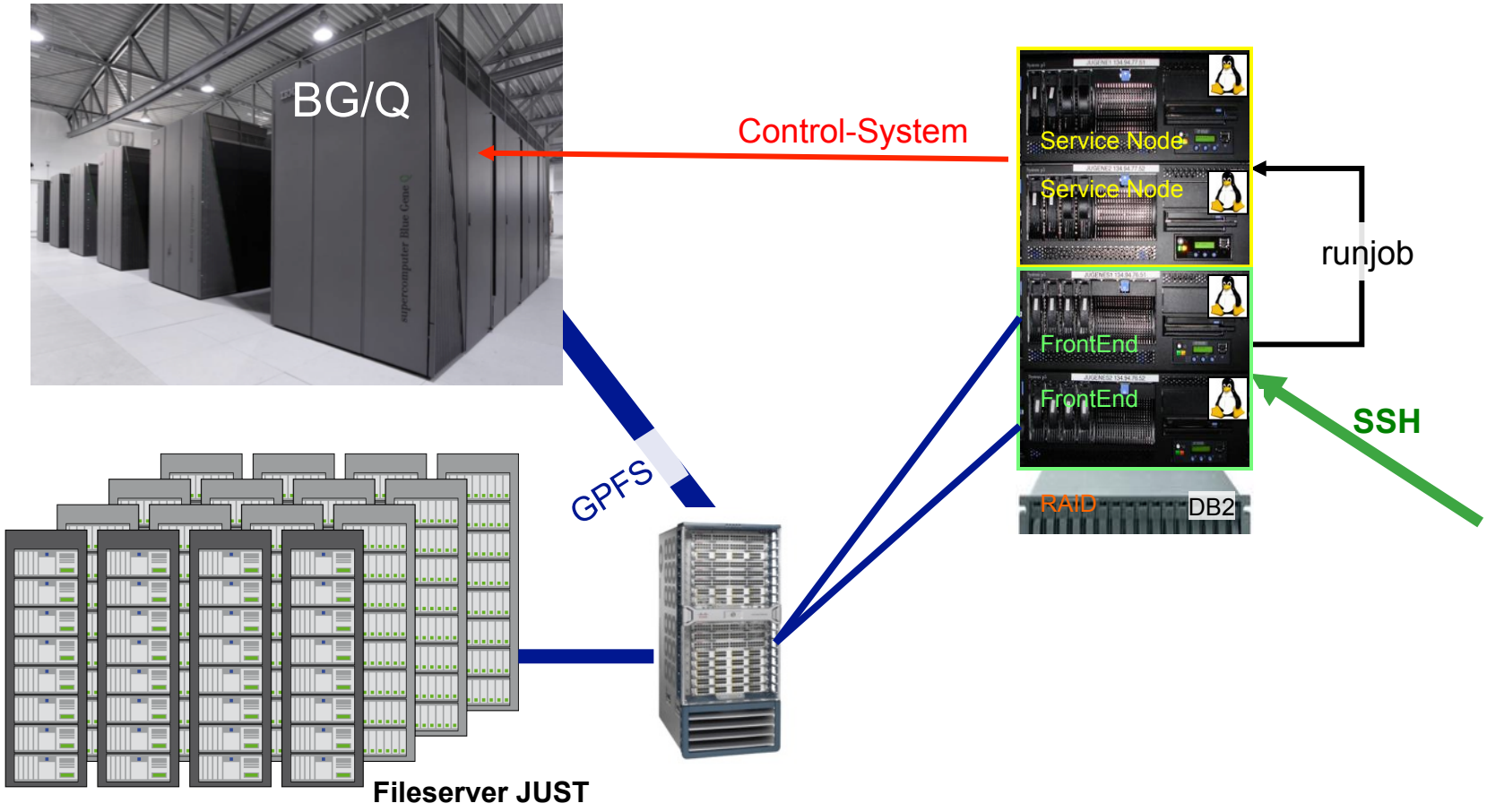
JUQUEEN Configuration

- 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)
  - 2.0 MW power consumption (10-80 kW per rack)

- 4 Frontend Nodes (user login) + 2 Service Nodes
  - IBM p7 740, 8 cores (3.55 GHz), 128 GB memory
  - local storage device DSC3700 (16 TB)

Mitglied der Helmholtz-Gemeinschaft

JUQUEEN Environment



JUQUEEN installation pictures

Mitglied der Helmholtz-Gemeinschaft

JUQUEEN installation pictures

Mitglied der Helmholtz-Gemeinschaft

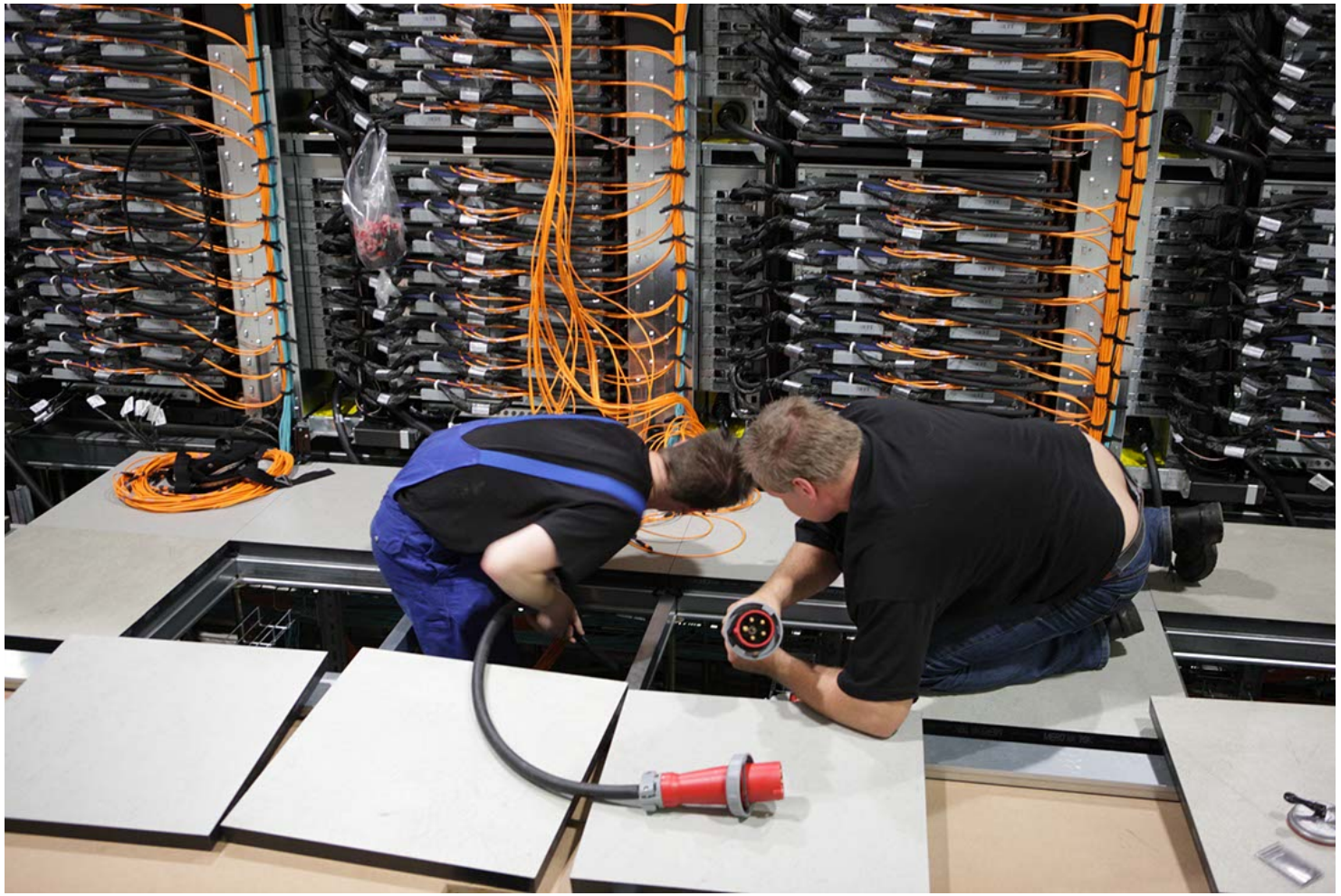
JUQUEEN installation pictures

Mitglied der Helmholtz-Gemeinschaft

JUQUEEN installation pictures

Mitglied der Helmholtz-Gemeinschaft

JUQUEEN installation pictures



Mitglied der Helmholtz-Gemeinschaft

JUQUEEN installation pictures

Mitglied der Helmholtz-Gemeinschaft

JUQUEEN installation pictures

Mitglied der Helmholtz-Gemeinschaft

JUQUEEN installation pictures

JUQUEEN installation pictures

Mitglied der Helmholtz-Gemeinschaft