Journal of large-scale research facilities

Journal of large-scale research facilities, 1, A1 (2015) http://dx.doi.org/10.17815/jlsrf-1-18

Published: 03.06.2015

JUQUEEN: IBM Blue Gene/Q® Supercomputer System at the Jülich Supercomputing Centre

Forschungszentrum Jülich, Jülich Supercomputing Centre

Instrument Scientists:
- Jutta Docter, Jülich Supercomputing Centre, Forschungszentrum Jülich, phone: +49(0) 2461 61 6763, email: j.docter@fz-juelich.de
- Dr. Michael Stephan, Jülich Supercomputing Centre, Forschungszentrum Jülich, phone: +49(0) 2461 61 1447, email: m.stephan@fz-juelich.de

Abstract: JUQUEEN (see Figure 1) is a high-scaling supercomputer funded mainly by the Gauss Centre for Supercomputing (2015) and by Helmholtz Association (2015) and is hosted by the Jülich Supercomputing Centre (2015b). It is a 28 rack, IBM Blue Gene/Q® system combining 28,672 compute nodes through a high-speed network providing an overall peak performance of 5.9 Petaflops.

1 Gauss Centre for Supercomputing

The Gauss Centre for Supercomputing (2015) (GCS) combines the three national supercomputing centres – Höchstleistungsrechenzentrum Stuttgart (2015) (HLRS), Jülich Supercomputing Centre (2015b) (JSC), and Leibniz-Rechenzentrum (2015) (LRZ) – into Germany's foremost supercomputing institution. GCS is jointly funded by the German Ministry of Education and Science (Bundesministerium für Bildung und Forschung – BMBF) and the corresponding ministries of the three national states of Bavaria, Baden-Wuerttemberg and North Rhine-Westphalia.

2 JUQUEEN project applications

All scientists and researchers in Germany and Europe have access to JUQUEEN computing resources. Computing time allocations are granted based on scientific criteria through independent reviewers in a peer-review process on the national level

1. through the executive and allocation committees of the JSC,
2. through biannual public announcement from GCS (calls for large-scale projects), and
3. on the European level through the PRACE regular calls for projects (also biannual).

Details about the actual application process are published on the Gauss Centre for Supercomputing (2015) web site.

Figure 1: JUQUEEN at the Jülich Supercomputing Center.

Scientist of RWTH Aachen University, Forschungszentrum Jülich or German Research School for Simulation Sciences (GRS) are also qualified for applications for computing time within the Jülich-Aachen Research Alliance (2015) (JARA-HPC).

3 JUQUEEN system

3.1 System configuration

JUQUEEN, the Blue Gene/Q system hosted by the Jülich Supercomputing Centre (2015b), was installed in four steps. The final stage of expansion went into production in March 2013 and has the following configuration (Jülich Supercomputing Centre, 2015a):

- Compute system
  - 28 racks (7 rows à 4 racks) - 28,672 nodes (458,752 cores)
    - Rack: 2 midplanes à 16 nodeboards (16,384 cores)
    - Nodeboard: 32 compute nodes
    - Node: 16 cores
    - Power: 60 - 70 kW per rack (average)
  - Main memory: 448 TB
  - Overall peak performance: 5.9 Petaflops
  - Linpack: 5.0 Petaflops
  - I/O Nodes: 248 (27x8 + 1x32) each with Dual-Port 10GigE connection
  - Footage: 83m^2
- Infrastructure nodes
  - 2 service nodes
    - Processor type: Power 7, 8C, 3.55 GHz
    - Total number of processors: 16
    - Main memory: 128 GB
    - Operating system: RedHat Linux V6.6
  - 4 login nodes

* Processor type: Power 7, 8C, 3.55 GHz
* Total number of processors: 16
* Main memory: 128 GB
* Operating system: RedHat Linux V6.6
* Batch system: IBM Tivoli Workload Scheduler LoadLeveler V 5.1

3.2 System design

The design of the Blue Gene/Q system is based on the International Business Machines Corporation (IBM) PowerPC A2 processing architecture (International Business Machines Corporation, 2013, 2015). Each processor includes 16 compute cores dedicated to the user application plus an additional core allocated to operating system administrative functions and a redundant spare core. By decoupling the execution of system services, effects of random asynchronous delays of user processes are suppressed. Such noise can significantly deteriorate the scalability of an architecture. As every core supports 4-way simultaneous multithreading (SMT), a Blue Gene/Q node can run up to 64 independent hardware threads. Every core is assisted by a 4-wide double precision floating point unit (SIMD).
The processor core architecture is relatively simple and implements a standard 64-bit power instruction set architecture. A particular feature of this core architecture is the support of an auxiliary execution unit. For Blue Gene/Q a Quad Floating-Point Processing Unit (QPU) had been developed. The QPU processes vectors of four 64-bit elements. In each clock cycle it can perform four fused multiply-add operations in parallel. Like in previous generations of Blue Gene, the vector arithmetic instructions may involve a permutation of the vector elements, which is used to implement complex arithmetics (without the need of separate shuffle operations like in other vector instruction set architectures). With each of the 16 QPUs being able to complete four multiply-add operations per clock cycle at a clock speed of 1.6 GHz, the peak performance is 204.8 GFlop/s.
This huge amount of performance can only be exploited if it is balanced by a powerful memory subsystem. As can be seen in Figure 2, a large fraction of the die space is occupied by the L2 cache, which has a capacity of 32 MBytes. Data is moved between external memory and this last-level cache by two memory controllers (MC 0 and MC 1). The L2 cache is shared by all processor cores. A central crossbar switch connects it to all cores plus the network subsystem. The cores can read from the L2 cache at an aggregate maximum bandwidth of 409.6 GByte/s.
Blue Gene/Q incorporates novel architectural advances that contribute to the system's outstanding performance and helps users to simplify programming of such high scaling many core systems.

- Hardware-based speculative execution capabilities facilitate efficient multi-threading for long code sections, even those with potential data dependencies. If conflicts are detected, the hardware can backtrack and redo the work with minimal affects on the application performance.
- Hardware-based transactional memory helps programmers avoid the potentially complex integration of locks and helps eliminate bottlenecks caused by deadlocking – when threads become stuck during the locking process. Hardware-based transactional memory helps to deliver efficient and effective multi-threading while reducing the need for complicated programming.
- The L1 pre-fetcher has the ability to run in normal stream prefetching mode which adaptively balances resources to pre-fetch L2 cache lines in response to observed memory traffic. But in addition it can also use four list-based prefetching engines to record memory access patterns in arbitrarily long code segments on a first iteration of a loop and playback this pattern for subsequent iterations. On subsequent passes, this list is adaptively refined for missing or extra cache misses and can be activated by program directives.

To interconnect the compute nodes a proprietary high-speed network in a 5D torus topology is used, providing the following advantages:

- A reduced latency of $\sim 3\mu$sec for point-to-point communication and $\sim 6\mu$sec within collectives and barrier.
- A good trade-off of nearest neighbour and bisection bandwidths.

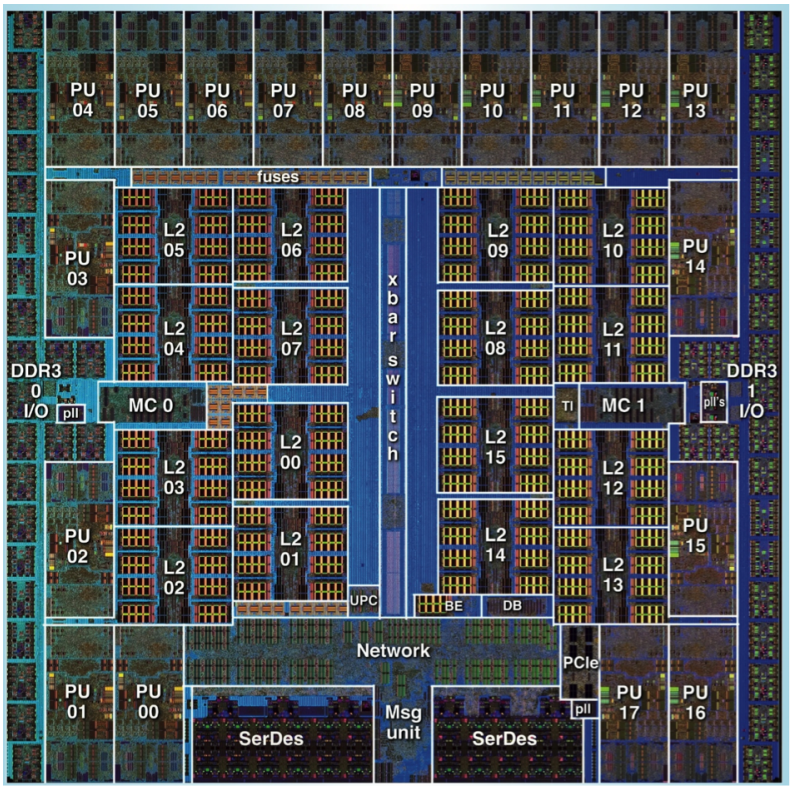

Figure 2: Physical layout of the Blue Gene/Q chip.

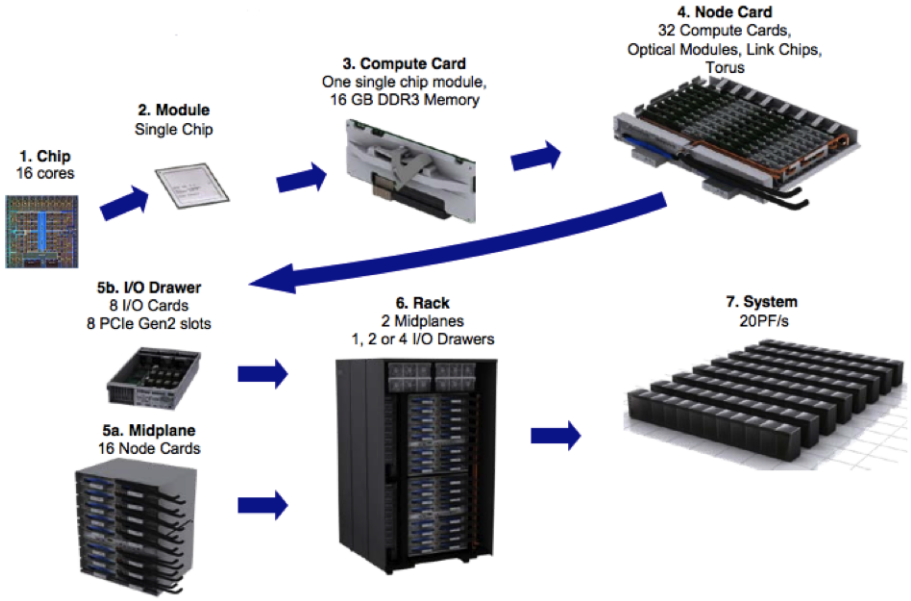

Figure 3: Blue Gene/Q packaging hierarchy.

- Extremely flexible partitioning into independent, non-interfering sub-machines is possible
- The hardware includes direct support for MPI collective reduce and all-reduce operations so that single pass floating point reductions can be executed with near link bandwidth
- With flexible configurability in its network Blue Gene/Q can spare out failed lasers without disrupting a running application
- The new networking hardware supports off-loading of the I/O traffic from the compute cores

One unique feature of the Blue Gene architecture line is the dense integration of a large number of nodes within a single rack. Figure 3 shows the packaging hierarchy of the system. Unlike in previous generations of Blue Gene, where air was used to remove the heat generated by the compute nodes, in the new generation of machines the nodes are directly connected to a liquid cooling system. 90% of the heat originating from the compute nodes in the system is directly taken away by the water, the remaining fraction, coming mainly from the power supplies, is still moved out of the rack by air. Engineered with fewer moving parts and built in redundancy, Blue Gene/Q has proven to be extreme reliable. Designed with a small footprint and low power requirements, Blue Gene/Q was ranked as the number-one most energy-efficient supercomputer in the world by the Green500 in Nov. 2011 (Green500, 2011).

References

Gauss Centre for Supercomputing. (2015). *Gauss Centre for Supercomputing e.V. (GCS).* Retrieved 05.05.2015, from http://www.gauss-centre.eu/

Green500. (2011). *The Green500 list.* Retrieved 05.05.2015, from http://www.green500.org/

Helmholtz Association. (2015). *Helmholtz-Gemeinschaft Deutscher Forschungszentren e.V. (HGF).* Retrieved 05.05.2015, from http://www.helmholtz.de/

Höchstleistungsrechenzentrum Stuttgart. (2015). *Höchstleistungsrechenzentrum Stuttgart (HLRS).* Retrieved 05.05.2015, from http://www.hlrs.de/

International Business Machines Corporation. (2013). The IBM Blue Gene/Q project. *IBM Journal of Research and Development*, *57*(1/2). http://dx.doi.org/10.1147/JRD.2012.2220487

International Business Machines Corporation. (2015). *IBM Blue Gene/Q information.* Retrieved 05.05.2015, from http://www.ibm.com/systems/technicalcomputing/solutions/bluegene/

Jülich-Aachen Research Alliance. (2015). *Jülich-Aachen Research Alliance – High-Performance Computing (JARA-HPC).* Retrieved 05.05.2015, from http://www.jara.org/en/research/jara-hpc/

Jülich Supercomputing Centre. (2015a). *JSC - JUQUEEN user information.* Retrieved 05.05.2015, from http://www.fz-juelich.de/ias/jsc/EN/Expertise/Supercomputers/JUQUEEN/JUQUEEN_node.html

Jülich Supercomputing Centre. (2015b). *Jülich Supercomputing Centre (JSC).* Retrieved 05.05.2015, from http://www.fz-juelich.de/ias/jsc

Leibniz-Rechenzentrum. (2015). *Leibniz-Rechenzentrum der Bayerischen Akademie der Wissenschaften (LRZ).* Retrieved 05.05.2015, from http://www.lrz.de/