JUQUEEN: The most recent Blue Gene Architecture at Jülich

Blue Gene/Q is the third generation in a line of highly scalable architectures. It is similar to the previous designs as very power-efficient processors are integrated very densely. But, due to an immense increase of parallelism at the single node level, a significant boost of performance is achieved. The network still being implemented on the chip remains a unique feature of this architecture. This is a key ingredient for being able to scale performance up to a huge number of cores.

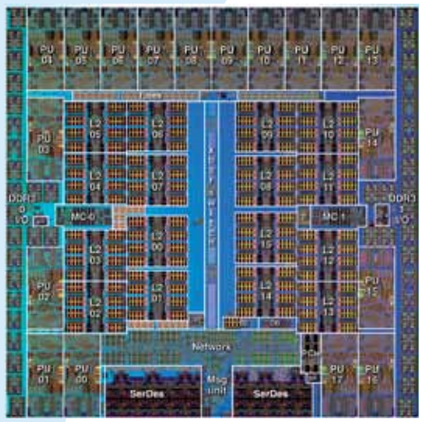

Figure 1: Physical layout of the Blue Gene/Q chip.

The Blue Gene/Q processor [1] comprises 16 cores for executing the user's application. There is an additional core for operating services. 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. Yet another, redundant, processor core is on the chip, but not available in practice; it has primarily been added to increase the manufacturing yield. With each of the cores being 4-way multi-threaded, up to 64 user threads (or processes) can run at the same time.

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.

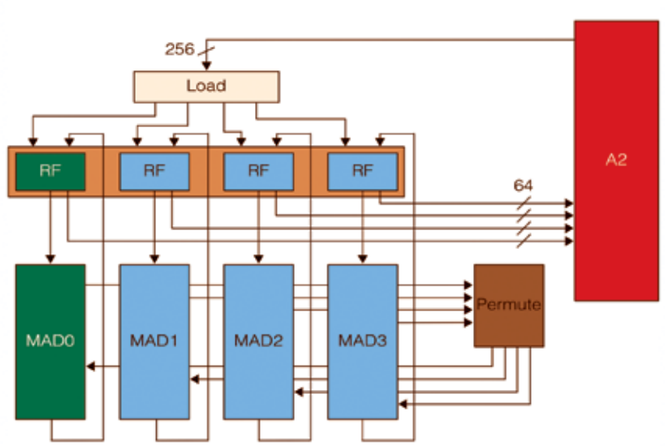

Figure 2: The Quad Floating-Point Processing Unit (QPU) is an extension of the A2 processor core to accelerate floating-point computations [1]. Operands stored in the register file (RF) are processed in 4 fused multiply-add dataflow (MAD) pipelines.

This huge amount of performance can only be exploited if it is balanced by a powerful memory subsystem. As can be seen in Fig. 1 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.

We would like to highlight several features of the memory hierarchy. To hide the latencies, which occur after an L1 cache miss because data needs to be fetched from L2 cache or external memory, the hardware supports pre-fetching. Like in other processor architectures, such pre-fetching can be triggered by an automatic detection of a sequence of contiguous set of memory addresses being accessed. Beyond this stream pre-fetching scheme there is a new scheme called list pre-fetching. This feature may be exploited by applications where a sequence of (possibly non-contiguous) addresses is accessed many times. Controlled by the user the list pre-fetch engines can record access patterns. When the execution of the relevant code section is repeated the list of recorded addresses is used for pre-fetching the data.

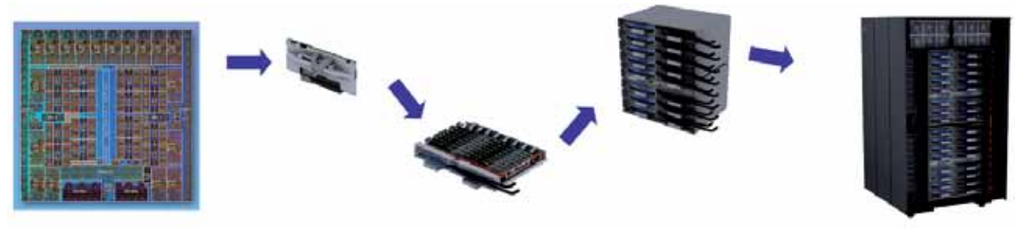

Figure 3: A Blue Gene/Q processor (left) is mounted to a heat spreader. A node-card can host 32 nodes. With 16 node-cards per midplane and two midplanes per rack the total number of nodes per rack (right) is 1,024.

One feature of the Blue Gene architecture line is the dense integration of a large number of nodes within a single rack (see Fig. 3). 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.

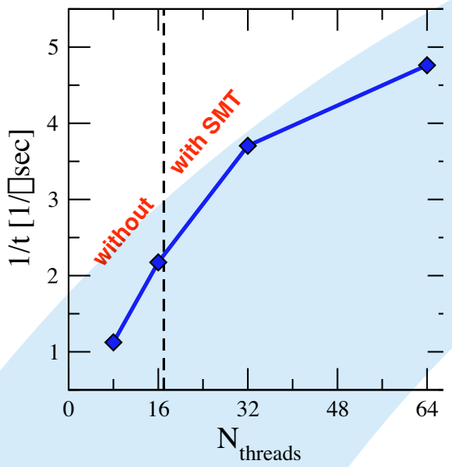

Figure 4: Inverse execution time of a fluid dynamics application running on a single node as a function of the number of threads [6]. The result shows significant gain from Simultaneous Multi-Threading (SMT).

The sustained performance obtained with the High-Performance Linpack (HPL) benchmark is similar to previous generations of Blue Gene machines, i.e. slightly more than 80% of the peak performance. This high level of device utilization in combination with a high power efficiency allowed for a significant leap on the Green500 list [4]. In this list systems are ordered according to the ratio speed achieved with the HPL benchmark versus power consumed by the system during benchmark execution. Previous number one on this list, QPACE [5], a special purpose computer operated at Forschungszentrum Jülich, achieved 0.77 GFlop/s per Watt. With Blue Gene/Q the score went up by almost a factor three to 2.1 GFlop/s per Watt.

The HPL benchmark performance of more than 80% demonstrates that it is at least for some applications feasible to obtain a high level of sustained performance, similar to previous generations of Blue Gene. However, there are challenges for the application developer to efficiently exploit the performance of this architecture. The most notable challenge is the increased parallelism at various levels. Compared to Blue Gene/P the vector units doubled their width, the number of cores and threads per node increased by a factor of four and 16, respectively. Within the Exascale Innovation Centre (EIC) the utilization of performance features has been investigated already at a stage when only prototype hardware had been available.

Fig. 4 shows the inverse execution time (which is proportional to the performance) of a fluid dynamics application based on the Lattice Boltzmann method. In case of ideal scaling the performance increases proportional with the number of threads until each core executes one thread. A significant performance increase is expected until each core executes two threads since one hardware thread can only schedule an instruction every second clock cycle. For this particular application a noticeable further increase of performance could be achieved by placing up to four threads on each core.

The first racks of Blue Gene/Q have been delivered to Forschungszentrum Jülich in April 2012. Only a few weeks later JUQUEEN consisting of eight racks could be made accessible to users for early production runs, see Fig. 5, delivering a peak performance of 1.6 PFlop/s, which is 60% more than JUGENE (72 racks Blue Gene/P, 1PFlop/s peak performance).

Even though JUGENE was an extreme reliable and stable system over the last years - achieving a user job utilization of over 90% - the system had to be shut down end of July and dismounted. Once the space, power, and cooling capacity occupied by JUGENE were freed, the extension of JUQUEEN to its final configuration with 28 racks is taking place. This translates into a more than five-fold increase of compute performance on Blue Gene available for scientific computing in Jülich, Germany and Europe.

A large fraction of the Blue Gene/Q installation at Forschungszentrum Jülich, JUQUEEN, will be available for users of the Gauss Centre for Supercomputing and PRACE.

References

[1] **Haring, R.A. et al.**
The IBM Blue Gene/Q Compute Chip, IEEE Micron, Vol. 32, No. 2, March/April 2012

[2] **Chen, D. et al.**
The IBM Blue Gene/Q Interconnection Network and Message Unit, SC '11 Proceedings, 2011

[3] **El-Harake, H. et al.**
Juniors: Prototypes for novel exascale I/O concepts, ISC'12 Poster, 2012

[4] http://www.green500.org

[5] **Baier, H. et al.**
QPACE: A QCD parallel computer based on Cell processors, PoSLAT 2009 (2009) 001

[6] **Pozzati, F.**
Porting and optimization of a Lattice Boltzmann D2Q37 code to Blue Gene/Q, Technical Report FZJ-JSC-IB-2011-06, Forschungszentrum Jülich, 2011

- Dirk Pleiter
- Michael Stephan

Jülich Supercomputing Centre (JSC), Forschungszentrum Jülich

Figure 5: 8-rack JUQUEEN at Forschungszentrum Jülich.