IBM BlueGene/P in Jülich: The Next Step towards Petascale Computing

When in 2004/2005 the IBM BlueGene technology became available, Research Centre Jülich (FZJ) recognized the potential of this architecture as a Leadership-class system for capability computing applications. A key feature of this architecture is its scalability towards PetaFlop Computing based on low power consumption, small footprint and an outstanding price performance ratio.

In early summer 2005 Jülich started testing a single BlueGene/L rack with 2,048 processors (inSiDE Vol. 3, No. 2, p. 18). It soon became obvious that many more applications than expected can be ported to efficiently run on the BlueGene architecture. Due to the fact that the system is well balanced in terms of processor speed, memory latency and network performance, many applications scale excellently up to large numbers of processors. Therefore in January 2006 the system was expanded to 8 racks with 16,384 processors, funded by the Helmholtz Association.

The 8-rack system has successfully been in operation for almost two years now. Today about 30 research projects, which were carefully selected with respect to their scientific quality, run their applications on the system using job sizes between 1,024 and 16,384 processors. During a BlueGene Scaling Workshop at FZJ experts from Argonne National Laboratory, IBM and Jülich helped to further optimize some important applications. It could be shown that all these applications succeeded in efficiently using all 16,384 processors of the machine.

Computational scientists from many research areas took the chance to apply for significant shares of BlueGene/L computer time to tackle unresolved questions which were out of reach before. Because of the large user demand and in line with its strategy to strengthen Leadership-class computing, Research Centre Jülich decided to order a powerful next-generation BlueGene system. In October 2007 a 16-rack BlueGene/P system with 65,536 processors was installed mainly financed by the Helmholtz Association and the State of North Rhine Westphalia. With its peak performance of 222,8 TFlop/s, Jülich's BlueGene/P – dubbed JUGENE – is currently the biggest supercomputer in Europe.

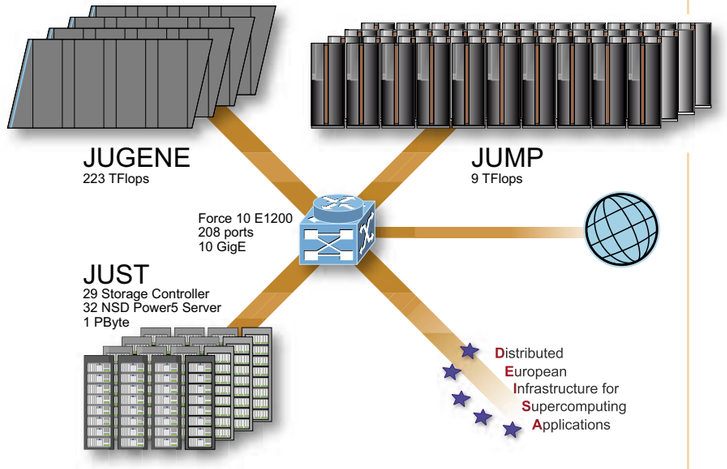
	BlueGene/L	BlueGene/P
Node Properties		
Processor	PowerPC® 440	PowerPC® 450
Processors per node (chip)	2	4
Processor clock speed	700 MHz	850 MHz
Coherency	Software managed	SMP
L1 cache (private)	32 KB per core	32 KB per core
L2 cache (private)	7 stream prefetching 2 line buffers/stream	7 stream prefetching 2 line buffers/stream
L3 cache (shared)	4 MB	8 MB
Physical memory per node	512 MB	2 GB
Main memory bandwidth	5,6 GB/s	13,6 GB/s
Peak performance	5,6 GFlop/s	13,6 GFlop/s
Torus Network		
Bandwidth	2.1 GB/s	5.1 GB/s
Hardware latency (nearest neighbour)	200 ns (32B packet) 1.6 µs (256B packet)	160 ns (32B packet) 1.3 µs (256B packet)
Global collective Network		
Bandwidth	700 MB/s	1,700 MB/s
Hardware latency (round trip worst case)	5.0 µs	3.0 µs

Table 1: BlueGene/L vs. BlueGene/P

The important differences between BlueGene/P and BlueGene/L largely concern the processor and the networks (see Table 1) while the principal build-up of BlueGene/L was kept unchanged. Key features of BlueGene/P are:

- 4 PowerPC® 450 processors are combined in a fully 4-way SMP (node) chip which allows a hybrid programming model with MPI and OpenMP (up to 4 threads per node).
- The network interface is fully DMA (Direct Memory Access) capable which increases the performance while reducing the processor load during message handling.
- The available memory per processor has been doubled.
- The external I/O network has been upgraded from 1 to 10 Gigabit Ethernet.

These improvements are also reflected by the application performance. A code from theoretical elementary particle physics, for example, on BlueGene/P runs at 31,5 % of the peak performance compared to 26,3 % on BlueGene/L. Furthermore, the increased memory of 2 GB per node will let new applications run on BlueGene/P.



JUGENE is part of the dual supercomputer complex in Jülich, embedded in a common storage infrastructure which was also expanded. Key part of this infrastructure is the new Jülich storage cluster (JUST) which was installed in the third quarter of 2007, increasing the online disk capacity by a factor of ten to around one PetaByte. The maximum I/O bandwidth of 20 GB/s is achieved by 29 storage controllers together with 32 IBM Power 5 servers. JUST is connected to the supercomputers via a new switch technology based on 10 Gigabit Ethernet. The system takes on the fileserver function for GPFS (General Parallel File System) and provides service to the clients in Jülich and to the clients within the international DEISA infrastructure as well.

With the upgrade of its supercomputer infrastructure FZJ has taken the next step towards Petascale Computing and has strengthened Germany's position to host one of the future European supercomputer centres.

• Michael Stephan
• Klaus Wolkersdorfer

Forschungszentrum Jülich (FZJ)