Recent Developments in Supercomputing

Th. Lippert

published in

NIC Symposium 2008,
G. Münster, D. Wolf, M. Kremer (Editors),
John von Neumann Institute for Computing, Jülich,
NIC Series, Vol. **39**, ISBN 978-3-9810843-5-1, pp. 1-8, 2008.

Recent Developments in Supercomputing

Thomas Lippert

Jülich Supercomputing Centre, Forschungszentrum Jülich
52425 Jülich, Germany
E-mail: th.lippert@fz-juelich.de

Status and recent developments in the field of supercomputing on the European and German level as well as at the Forschungszentrum Jülich are presented. Encouraged by the ESFRI committee, the European PRACE Initiative is going to create a world-leading European tier-0 supercomputer infrastructure. In Germany, the BMBF formed the Gauss Centre for Supercomputing, the largest national association for supercomputing in Europe. The Gauss Centre is the German partner in PRACE.

With its new Blue Gene/P system, the Jülich supercomputing centre has realized its vision of a dual system complex and is heading for Petaflop/s already in 2009. In the framework of the JuRoPA-Project, in cooperation with German and European industrial partners, the JSC will build a next generation general purpose system with very good price-performance ratio and energy efficiency.

1 Introduction

Supercomputers enable scientists and engineers to solve computational problems of unprecedented complexity. Today, computer simulation is established as a third category of gaining scientific insight, in addition to theory and experiment, and has been developed into a key technology for industrial product development and production optimization.

The supercomputers provided by the *Jülich Supercomputing Centre* (JSC), the former Zentralinstitut für Angewandte Mathematik, are utilized by scientists all over Germany and Europe via provision of computer time through the John von Neumann Institute for Computing (NIC). Currently, the Blue Gene/P system at JSC delivers more than 220 Teraflop/s highly scalable supercomputing power. JUGENE is the most powerful machine for free research worldwide and ranks at position 2 in the TOP500 list of November 2007. Supercomputers in this performance class are at the top of an indispensable large-scale scientific infrastructure for national and European research. The JSC strives to guarantee optimal user support, continuously developing the simulation methodology, parallel algorithms and new programming and visualisation techniques as well as carrying out intensive own research in core areas of the computational sciences.

The JSC will intensify its support, research and development activities: support and research are strengthened by establishing several simulation laboratories in the core scientific fields served by the NIC. Each simulation lab is targeted at a specific disciplinary scientific area. Currently the SLs “Plasma Physics”, “Computational Biology” and “Earth System Sciences” are in their first stage of implementation. Other labs like “Nano and Materials Science” will follow soon. While the evaluation and benchmarking of novel compute architectures, in particular in the run-up for a next procurement, have a long tradition at JSC, a novel activity of JSC will be its active participation in the design and building of next generation supercomputer systems (Sec. 4.3).

Similar activities will become an important element of the EU project PRACE, the Partnership for Advanced Computing in Europe (Sec. 3), which is coordinated by Achim Bachem at Forschungszentrum Jülich. PRACE is an initiative of 14 European countries to create a so-called tier-0 supercomputer infrastructure in the Multi-Petaflop/s range that is on one level with the US infrastructure or might even lead the worldwide supercomputing efforts for simulation science and engineering.

In order actively participate in these developments a new association was formed by initiative of the BMBF in Germany (Sec. 2). In June 2006, the Minister for Research, Dr. Schavan, announced the creation of the *Gauss Centre for Supercomputing* (GCS). The GCS-Verein has Forschungszentrum Jülich, Höchstleistungsrechenzentrum Stuttgart and Leibniz-Rechenzentrum Garching as its members. Further members are the chairs of the scientific councils, *e.g.* the NIC for the FZJ. Currently, FZJ is chairing the GCS.

In this note, I am going to review the status and the goals of PRACE and GAUSS. In the framework of PRACE, a technology platform for development of systems in the Multi-Petaflop-range will be created. Currently, consortia like PROSPECT and TALOS are being established jointly by IT-industry and user centres. They are going to support the European technology platform of PRACE. I will describe JSC's approach to Petacomputing, which is guided by our wish to take the users with us on our way to unprecedented scalability. I will finally comment on our efforts with the JuRoPA project (Jülich Research on Petaflop Architectures) to design and to build ourselves a general purpose computer system in the Petaflop-range in cooperation with our industrial partners from Germany and Europe.

2 Gauss Centre for Supercomputing

In order to continue to provide computing resources at the highest performance level for computer-based science and engineering in Germany, in June 2006, the Federal Ministry of Education and Research (BMBF) has initiated the foundation of the Gauss Centre for Supercomputing[1]. The Ministries of Innovation, Science, Research and Technology of the State of North Rhine-Westphalia, the Bavarian State Ministry of Science, Research and the Arts, and the Ministry of Science, Research and the Arts Baden-Württemberg unreservedly support the GCS. Since September 2007, the GCS is a registered association (e.V.).

The three German national supercomputing centres joined up to form the GCS with combined computing power of currently more than 350 teraflop/s, located in Stuttgart, Garching, and Jülich. The GCS is the largest national association for high-performance computing in Europe. It is planned to increase the overall performance of the GCS to a value larger than 1000 Teraflop/s in 2009.

The GCS members follow a common direction in their organization. The procurement of hardware will be closely coordinated, applications for computing time will be scientifically evaluated on a common basis, and software projects will be jointly developed. The GCS currently is preparing documents for a business plan and the future scientific governance of supercomputing in Germany. This task involves also the planning of a closer coordination with the German Tier-1 and Tier-2 centres in an HPC alliance of supercomputer centres.

Another important area of activities of the GCS will be the training of users and the education of young scientists. The work of specialist researchers will be supported and promoted by harmonizing the services and organizing joint schools, workshops, and conferences on simulation techniques. Methodologically oriented user support is also a major concern of the Gauss Centre.

With the establishment of the GCS and its activities, the three GCS sites have already created a high visibility worldwide and have laid the grounds to play a central role in the establishment of a European high-performance supercomputer infrastructure within the PRACE initiative.

3 Partnership for Advanced Computing in Europe

Scientists and engineers in Europe must be provided with access to leadership class supercomputing in order to remain competitive internationally and to maintain or regain leadership in the simulation sciences. In this spirit, the European Strategy Forum for Research Infrastructures (ESFRI) has identified HPC as a strategic priority concerning the creation of new research infrastructures in Europe. Following the recommendations detailed in the ESFRI Roadmap the European Commission issued a call in the 7th Framework Programme for a preparatory phase for up to 35 European research infrastructures[2].

This development goes back to the preparatory studies of HPCEUR and the High Performance Computing in Europe Taskforce (HET)[3] between 2004 and end of 2006. A consortium of 14 European countries signed a Memorandum of Understanding on April 17, 2007, in Berlin, in the presence of minister Schavan. The initiative is named PRACE, *i.e.* Partnership or Advanced Computing in Europe. Prior to this event, the founding members (HET) have submitted a project proposal, the PRACE project, to meet the above mentioned call. The project was fully granted in summer 2007 and will have its kick-off meeting end of January 2008.

PRACE aims at preparing a persistent pan-European HPC supercomputer service in form of a single legal entity, consisting of three to five tier-0 centres, similar to the US HPC-infrastructure. PRACE will be the tier-0 level of the European HPC ecosystem. The hosting centres of the planned tier-0 systems will provide the expertise, competency, and the required infrastructure for comprehensive services to meet the challenging demands of excellent users from academia and industry.

PRACE will prepare for the implementation of the infrastructure in 2009/2010. Within the project the definition and the set-up of a legal and organizational structure involving HPC centres, national funding agencies, and scientific user communities are carried out. Tasks involve plans to ensure adequate funding for the continued operation and periodic renewal of leadership systems, co-ordinated procurements, efficient use and fair access for European researchers.

PRACE will prepare the deployment of Petaflop/s systems in 2009/2010. This includes the procurement of prototype systems for the evaluation of software for managing the distributed infrastructure, the selection, benchmarking, and scaling of libraries and codes from scientific user communities, the definition of technical requirements and procurement procedures.

The PRACE Petaflop-infrastructure will require an initial investment of up to 500 Mio Euro followed by annual funds of 100 Mio Euro for upgrades and renewal in later years.

3.1 Technical Development in Europe: Consortia of Industry and User Centres

In addition to the precommercial procurement activities for the PRACE Petaflop-systems, a collaboration with the European IT-industry will be started by PRACE in order to influence the development of new technologies and components for architectures that are promising for Petaflop/s systems to be procured after 2010.

In order to achieve these goals, already early in 2007 two consortia were formed in Europe. One consortium is PROSPECT (Promotion of Supercomputers and Petacomputing Technologies), with the partners BSC, JSC, LRZ, DWD, IBM, Quadrics, Intel, ParTec, University of Heidelberg and University of Regensburg. A second one is TALOS centered on BULL and CEA/France, with the associated partners Intel and Quadrics. These consortia, and maybe more to come, will seek cooperation with the workgroup WP8 within the PRACE project in order to create a long-term technology platform for development of systems in the realm of Multi-Petaflop/s.

PROSPECT understands itself as an interest group to foster the development of supercomputing technology in Europe. Partners within PROSPECT have started to form project groups that focus on hardware and software aspects. So far three projects are being initiated within the PROSPECT framework. One of these project is JuRoPA (Sec. 4.3), a second one is called QPACE (QCD Petacomputing Engine). Within QPACE, IBM-Böblingen, the universities of Regensburg and Wuppertal, DESY-Zeuthen and the JSC are going to build a new 3-dimensional high-speed communication network around a grid of next-generation cell processors.

A third project within PROSPECT will involve most of the PRACE partners tackling the highly relevant question of energy aware and energy efficient supercomputing. Energy issues become ever more important[4] since the need for higher performance requires to boost parallelism after frequency growth came to an end. Therefore, the individual processors necessarily must become less instead of more power consuming. The Blue Gene approach as well as the cell processor are state-of-the-art technologies concerning power efficiency. While with the new multi-core processors of Intel and AMD an important step is taken in the right direction, energy consumption still grows linearly with the number of processors, which imposes a severe limit to scalability of future supercomputer hardware. As a consequence, we have to make use of all possible ideas to reduce the waste of power.

4 Jülich Supercomputing Centre

The JSC works at the forefront of the technologically possible in HPC. Following a dual philosophy the JSC strives to provide Petaflop-class systems with highest scalability like the new Blue Gene /P together with highly flexible general purpose systems. In 2008 the JUMP general purpose system will be replaced by a very large compute cluster increasing the power of JUMP by more than a factor of 25. This cluster is currently designed and will be built together with partners from industry. The operating system software to be utilized is the German system ParaStation, developed jointly by ParTec (Munich) and JSC. Together with an online storage of 1 PetaBytes, connected to an automatized tape archive of about 10 Petabytes the JSC operates the most flexible computer complex currently available.

4.1 Dual System Complex

Already quite early in the operation cycle of JUMP, the Parateam at JSC, responsible for the scaling of user applications, realized that for the given number of users and allocation of computing time a further growth of capability computing on JUMP would have decreased the overall efficiency. This is because many users worked at their sweet spot and would not have profited from policies preferring jobs with very large processor numbers.

As a solution, the concept of a “Dual Supercomputer Complex” was conceived and described as part of a white paper of the Helmholtz Association with title “Ausbau des Supercomputing in der Helmholtz-Gemeinschaft und Positionierung im europäischen Forschungsraum”.

By installing the first Blue Gene/L system (Sec. 4.2) and boosting it up to 46 Teraflop/s, the JSC started to realize the dual concept. Blue Gene/L is only allowed to be used by highly scaling applications. Before being admitted to the machine, the scalability of a code must be proven. In this way, a decisive step towards capability computing was taken, as is shown by the performance parameters of the Blue Gene/L system. Following this philosophy, the number of accepted projects on JUBL for the granting period 2006/2007 was limited to about 25. A substantial increase of the scalability up to 16,384 processors could be reached for several applications by intensive user support during two workshops, the “Big Blue Gene Week” and the “Blue Gene Scaling Workshop” which both lasted for one week[5].

The dual system complex is supported by a common storage infrastructure which was also expanded in 2007. 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 hosts 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. In future technological developments, a LUSTRE parallel file system will join the GPFS.

4.2 Towards Petacomputing

In 2004/2005, the IBM Blue Gene technology became available. The JSC has evaluated the potential of this architecture very early as a future leadership class system for capability computing applications. A key feature of this architecture is its scalability towards Petaflop-computing based on lowest power consumption, smallest footprint and an acceptable price-performance ratio.

In early summer 2005 Jülich started testing a single Blue Gene/L rack with 2048 processors. It soon became obvious that many more applications than expected could be ported to efficiently run on the Blue Gene architecture. Due to the fact that the system is well balanced in terms of processor speed, memory latency and network performance, many applications scale up to large numbers of processors. In January 2006 the system was expanded to 8 racks with altogether 16384 processors.

The 8-rack system has successfully been in operation for almost two years now. About 4 racks or about 25 Teraflop/s are available for the users in the NIC. About 25 research projects, which were carefully selected with respect to their scientific quality, run their applications on the system using job sizes between 1024 and 16384 processors. During the Blue Gene Scaling Workshop at FZJ, experts from Argonne National Laboratory, IBM and Jülich helped to further optimize some important applications. As already stated, it could be shown that all applications considered succeeded in efficiently using all 16384 processors of the machine[5].

Computational scientists from many research areas took the chance to apply for significant shares of Blue Gene/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, FZJ decided to order a powerful next-generation Blue Gene system. In October 2007 a 16-rack Blue Gene/P system with 65536 processors was installed mainly financed by the Helmholtz Association (BMBF) and the State of North Rhine Westphalia. With its huge peak performance of 222.8 TFlop/s, Jülich's Blue Gene/P dubbed "JUGENE" is currently the most powerful supercomputer in Europe and number 2 worldwide.

The important differences between Blue Gene/P and Blue Gene/L mainly concern the processor and the networks (see Table 1) while the principal architecture of Blue Gene/L was kept unchanged. Key features of Blue Gene/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.

All these improvements are well reflected by the application performance. A code from theoretical elementary particle physics, (dynamical overlap fermions), on Blue Gene/P runs at 38% of the peak performance compared to 26.3% on Blue Gene/L. Furthermore, the increased memory of 2 GB per node will allow new applications to be executed on Blue Gene/P. JUGENE has become part of the dual supercomputer complex of JSC. It will be complemented soon by the successor of the JUMP system, a general purpose system based on cluster technology (see Sec.4.3).

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.

4.3 Building a Supercomputer

In the future, the JSC will play an active role in designing and building its supercomputers, in particular as far as general purpose systems are concerned. Using component-oriented cluster technology JSC will be able to directly address its users' requirements in the design and build-up process and to go for the best possible price-performance ratio and lowest energy consumption at the earliest possible implementation date.

	Blue Gene/L	**Blue Gene/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 (256 B packet)	160 ns (32B packet) 1.3 μs (256 B packet)
Global collective network		
Bandwidth	700 MB/s	1700 MB/s
Hardware latency (round trip worst case)	5.0 μs	3.0 μs

Table 1. Blue Gene /P vs. Blue Gene /L.

At this stage, JSC is in the process to design the successor of JUMP. It is planned to realize a highly flexible general purpose cluster system comprising 2048 compute nodes to be installed end of 2008. Each node will host 8 cores of the next generation Intel HPC processor Nehalem. The high-speed interconnect will be based on QSnet$^{\text{III}}$ of the British-Italian IT-company Quadrics. As cluster operating system the German software ParaStation will be used which is proven to scale beyond 1000 nodes.

To this end, a research project group was formed named JuRoPA (Jülich Research on Petaflop Architectures). The core partners of JuRoPA are the JSC, ParTec (Munich)[6], Intel (Brühl)[7], and Quadrics (Bristol)[8]. At a later stage, a provider for the hardware which is not yet selected will join the group.

The ParaStation cluster operating system is being developed in cooperation between ParTec and JSC since several years. Together with the European interconnect technology Quadrics which will again become the state-of-the-art technology in the field of cluster computing with its next generation QSnet$^{\text{III}}$ the JSC will demonstrate that Europe is back in supercomputing technology and can build leadership-class supercomputers at lowest costs and high energy efficiency.

5 Concluding Remarks

In the last two years, coordinated efforts in Europe and Germany have led to a variety of new developments in the ever growing field of simulation science and engineering, and in particular in the field of supercomputing. With PRACE, the Partnership for Advanced

Computing, Europe is on the right track to create a Petacomputing infrastructure which can be competitive with US installations. The GCS, the Gauss Centre for Supercomputing, brings the German National Supercomputer Centres in Garching, Jülich and Stuttgart together, forming the largest supercomputing complex in Europe and striving to host the first European supercomputer centre with Petaflop capability.

The JSC, the Jülich Supercomputing Centre, has realized its vision of a dual supercomputing complex both in terms of hardware and computer time provision, which is coordinated by the John von Neumann Institute for Computing. With 223 Teraflop/s, the most powerful supercomputer for free research worldwide is accessible by the NIC community. Together with the successor of the general purpose system JUMP, a cluster being designed and built by JSC, ParTec, Intel and Quadrics, and planned to be in operation end of 2008, German and European scientists can work at the forefront of simulation science.

Acknowledgments

Th.Lippert thanks N. Attig, A. Bachem, J. Docter, N. Eicker, Th. Eickermann, R. Esser, W. Grich, S. Hfler-Thierfeldt, B. Orth, S. Krieg, M. Stephan, and K. Wolkersdorfer for kind contributions to this note.

References

1. `http://www.gauss-centre.eu`.
2. `http://cordis.europa.eu/esfri/`.
3. `http://www.hpcineuropataskforce.eu`.
4. `http://www.green500.org`.
5. Wolfgang Frings, Marc-André Hermanns, Bernd Mohr, and Boris Orth (Editors), *Report on the Jülich Blue Gene/L Scaling Workshop 2006*, `http://www.fz-juelich.de/jsc/docs/printable/ib/ib-07/ib-2007-02.pdf`, FZJ-ZAM-IB-2007-02.
6. `http://www.cluster-competence-center.com`.
7. `http://www.intel.com/jobs/germany/`.
8. `http://www.quadrics.com/quadrics/QuadricsHome.nsf/DisplayPages/Homepage`.