Bits & Bytes

No. 178 RZG Computer Bulletin May 2005

Computing Center of the Max Planck Society and the Institute for Plasma Physics*

New architectures doubling compute power at RZG

We had to face two severe bottlenecks in the main compute resources at RZG: on the IBM p690 based supercomputer with HPS switch for capability computing, and on the Intel Xeon based Linux compute farms (so-called BladeCenters) for capacity computing. Therefore scientists supported the installation of additional resources and the following new architectures could be installed:

- An IBM p575 Power5 based compute cluster, consisting of 8-way SMPs, to expand the p690 based supercomputer by a capacity computing component
- SUN AMD Opteron based 64 bit Linux clusters with Infiniband switch to introduce the 64 bit Linux technology with a high performance interconnect
- An SGI Altix 3700 Bx2 Server with NUMAlink dedicated for MPI for Astrophysics

The IBM p575 system

The new IBM p575 compute cluster is based on single core Power5 8-way SMPs (clock rate 1.9 GHz). Especially for memory demanding applications like Wien2k the sustained performance of one SMP is up to three times the performance of 8 Power4 processors (with 1.3 GHz) in an IBM Regatta node at RZG. The same 8-way Power5 systems will be the building blocks for the new ASC (formerly ASCI) Purple 100 TeraFlop system to be installed at Lawrence Livermore National Laboratory. RZG is one of the first customers world-wide that has obtained such systems after general availability (GA). The new compute systems consist of 86 8-way nodes (688 processors) and mainly enhance the compute power for the Fritz-Haber-Institute in Berlin and the MPI for Metals Research in Stuttgart, departments with memory intensive applications. The new p575 systems are seamlessly integrated into the p690 environment with uniform batch access. Further information can be found under www.rzg.mpg.de/computing/IBM_P5/.

The Opteron Clusters

Hardware

In early 2004, a performance comparison has been done for five parallel applications on 16 processor configurations, both on a Power4 based p690 node and dual Intel Xeon nodes in so-called BladeCenters (see Bits & Bytes No. 177). On average 16 Xeon processors with clock rate 2.8 GHz and Gigabit Ethernet connection equal the performance of half a Regatta node (16 processors Power4, 1.3 GHz) and therefore shows a significant price-performance advantage for small processor configurations. The competing processor range of a Linux cluster could be improved by a better performing interconnect than Gigabit Ethernet. Together with a demand for support of larger memory sizes than supported by the 32 bit technology, RZG together with several Max Planck Institutes has decided for 64 bit AMD Opteron based clusters with Infiniband interconnect technology. The partitioned clusters contain 113 dual nodes, 2.2 GHz and five 4-way nodes (quad Opterons), 2.4 GHz with a total of nearly 250 processors.

System environment

The operating system of the Opteron cluster is the 64 bit Linux distribution coming with the SuSE Linux Enterprise Server 9 for the AMD64 architecture. The kernel- and userspace utilities for management of the Infiniband subnet are a product of Mellanox Technologies. They provide Infiniband users with the Verbs Application Programmers Interface (VAPI). On top of VAPI several MPI implementations have been deployed. The most leading MPI implementation over Infiniband, MVAPICH (based on MPI-1 semantics), developed in the Network-Based Computing Laboratory at Ohio State University, was installed. Principally the compilers from GNU and the Intel suite are provided.

As the Opteron clusters are the first machines at RZG equipped with the AMD 64 bit technology, a new SYSNAME, `amd64_sles9`, has been introduced in RZG's AFS cell. Our goal is to minimize the number of software parts in `amd64_sles9` — only commercial products, and software which needs to be customized to fit the RZG users' needs (like Heimdal/Kerberos5), shall be placed there. Additionally, the scripts icc (C), icpc (C++), f95i (f95), their MPI pendant mpicci (C), mpiCCi (C++), mpif95i (f95) respectively and the MPI scripts for the GNU compilers mpiccg (C), mpiCCg (C++) are installed.

*Max-Planck-Institut für Plasmaphysik, Boltzmannstraße 2, D-85748 Garching bei München, tel.: +49(89) 3299-01, e-mail: benutzerberatung@rzg.mpg.de, URL: http://www.rzg.mpg.de/
Editorial: Dr. Roman Hatzky, Tel. -1707

However, there are two restrictions using MPI over Infiniband at the moment: It is **not possible to link statically** and it is **only possible to run 64 bit MPI code**. We received commitments that these restrictions will be eliminated by end of this year.

For batch usage the Sun Grid Engine batch environment has been installed on the Opteron cluster www.rzg.mpg.de/docs/linux/sge.html. The node for login, interactive work (not compute intensive!) and batch submission is riologin.opt.rzg.mpg.de. Further interactive logins to other nodes in the cluster are currently not allowed.

SGI Altix 3700 system

For MPI for Astrophysics a new SGI Altix 3700 Bx2 system with 64 Itanium 2 processors operated at 1.6 GHz with an SGI NUMAlink 4 interconnect has been put into operation. The system comprises a peak performance of 200 GFlop/s with a main memory of 128 GB and is used especially for highly scalable shared-memory parallel codes (based on the OpenMP programming model) for supernova simulations.

Hermann Lederer

DEISA - Supercomputing at European scale under way

One year after project start, the European DEISA project (www.deisa.org) is about to enter production state for the first four “core sites”, CINECA (Italy), FZ Jülich (Germany), IDRIS (France) and RZG (Germany). Also by May 1, 2005, DEISA has been expanded from 8 to 11 full partners, now additionally including BSC (Barcelona), HLRS (Stuttgart), and LRZ (Munich). For the inauguration of the production state, the first annual DEISA Symposium was held on May 9 and 10 in Paris. Leading computational scientists from all major scientific fields in natural sciences gave keynote presentations, among them Prof. K. Lackner from IPP and Prof. M. Parrinello from ETHZ. Positive feedback was also given from US computational scientists and representatives of the Teragrid project, the US counterpart of DEISA. With Teragrid a collaboration on arising new technologies has been started, especially in the area of global file systems, which have been a key element for DEISA from the very beginning. As a milestone for enabling the production state, DEISA’s Global File System has reached network speed across Europe.

The DEISA Global File System

Among the four DEISA “core-sites”, IBM’s Multi-Cluster (MC) Global Parallel File System (GPFS) has been set up, the world’s first real production deployment of MC-GFPS. Each site provides its own GPFS file system which is part of the common “global” file system. The current wide area network of priorized 1 Gbit/s bandwidth among the DEISA core sites can already be fully exploited by the global file system. This could be confirmed by several benchmarks, which showed I/O rates of more than 100 Mbytes/s, and also with a real application. The resource demanding plasma physics turbulence simulation code TORB from IPP and CRPP was executed at the different core sites, using direct I/O to the MC-GPFS, with the disks being physically located hundreds of kilometers away from the compute nodes.

The DEISA Extreme Computing Initiative

With entering production mode soon, DEISA has started the Extreme Computing Initiative. A first call for Expressions of Interest for challenging computational projects has been launched on April 1, ending on May 30 (for details see www.deisa.org). An application task force has been created for support. In the fore-field of production mode, Joint Research Activities had already started to prepare leading applications for suitable usage within DEISA. Here the TORB code for gyrokinetic simulations was optimized and expanded by Roman Hatzky. On the large IBM system of ECMWF (UK), which recently has decided to also contribute resources to the DEISA pool, extreme computing could be demonstrated. On that IBM Regatta based system (comprised of 1.9 GHz Power4+ processors and IBM High Performance Switch), TORB achieved a speedup of 1680 on 2048 processors, leading to a parallel efficiency of 82 %, with an overall sustained performance of 1.3 Teraflop/s.

Hermann Lederer

No virus-check on AFS-volumes

In the last time it sometimes happened that a virus made its way into the AFS home-directory of a user. This can happen on infected windows-machines, where the user has AFS mounted and a token for AFS. It is also possible to put viruses on other people’s home directories, if public write access was granted. In this case no AFS token is required.

Although this is an unwanted situation, it is not possible to check the user-home-directories for viruses by the system. The users by themselves should scan with their anti-virus-software their AFS-home-directories regularly.

If the RZG gets knowledge of viruses in the AFS of a user, he/she will be contacted directly, informing about that fact and asking for instructions. Due to Datenschutz it is not possible for RZG to scan or even to remove viruses from the directories of a user.

Andreas Schott