# Bits & Bytes

No. 180 RZG Computer Bulletin Dec 2007

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

## Genius – the IBM Blue Gene/P system at RZG

The new Blue Gene/P system at RZG was installed in September 2007 as the first such system at a customer site world-wide. It is a massively parallel supercomputer based on the IBM system-on-a-chip technology which combines the following advantages: a scalable architecture in hardware and software with excellent energy efficiency (in FLOPS per watt) and compactness (in FLOPS per floor space). Providing these properties the Blue Gene system line scales up to the PetaFlop regime. The Blue Gene/P was put into production in October 2007 after a short phase of tuning and benchmarking.

### Hardware

The system is currently made up of two racks with 32 node boards each. Each node board is equipped with 32 quad-core compute nodes. Thus, in total we have 2048 compute nodes (8192 cores), along with 16 quad-core I/O nodes. Each compute node is made up of four 850 MHz PowerPC 450 processors with 512 MB memory each. The 2 GB memory can be shared, e.g., in a 4-way SMP mode. The total main memory of the machine is 4 TB. In addition to the general 3D torus network for inter-node communication, there are particular fast networks for global communication and synchronization. The I/O nodes are connected to the outside world via a 10 Gbit/s Ethernet network. In January 2008, the Blue Gene/P will be extended by a third rack, thus increasing the system's peak performance from 27 TF/s to 40 TF/s.

### Software

The Blue Gene/P compute nodes run a lightweight Linux kernel to execute user-mode applications only. The number of active cores within a compute node can be controlled by specifying one of the following options "Symmetric Multiprocessor Mode", "Dual Node Mode" or "Virtual Node Mode". I/O is done by the I/O nodes.

Similar to the IBM Regatta system the programming environment includes the Fortran xlf 11.1 and C++ 9.0 compilers and the ESSL library from IBM. There are also some GNU tools (emacs, gnuplot) available.

### Operation

Users can login interactively from *gate* or *psi* via ssh/slogin to *genius.rzg.mpg.de*, an 8-processor Power5 system as the frontend node of the Blue Gene/P. The frontend node is running Suse Linux SLES10. If your code is suited for running on the Blue Gene/P system you should apply for an account on the Blue Gene/P. Your userid and (Kerberos) password will be the same as on the other HPC systems at the RZG.

Jobs on the Blue Gene/P have to be submitted to the LoadLeveler batch system on *genius.rzg.mpg.de*. The LoadLeveler will allocate the resources required for your job (e.g. the compute nodes to run your job on). Two shared GPFS filesystems */u* (home directory) and */ptmp* (temporary batch output) are provided that are symmetrically accessible from all the Blue Gene/P I/O nodes as well as from the frontend node.

### First experiences

Several codes from plasma and astrophysics have been ported to Blue Gene/P, and a few run already in production mode scaling well up to 4096 or 8192 cores, respectively. Further codes, especially from materials science, are being ported. The system is well suited for applications with good scalability up to 1024 cores and beyond. A LINPACK performance of 21.9 TF/s has been achieved on 8192 cores. More detailed information on the Blue Gene/P at the RZG can be found at: www.rzg.mpg.de/computing/IBM_BGP/

Ingeborg Weidl & Johannes Reetz

## The Munich-ATLAS-Tier2 project

Collaborating research groups from the Max Planck Institute of Physics (MPP) and from the Ludwig Maximilian University (LMU) play a significant role in the ATLAS project of the Large-Hadron-Collider (LHC) experiment at the international European Centre for Particle Physics (CERN). In this experiment two beams of protons are accelerated to nearly light velocity and are focused towards each other to collide head-on. With each collision of two protons up to several thousands of new particles are generated. Various detectors will record the trajectories and energies of these generated particles. Huge amounts of data are expected, which have to be stored, processed and analyzed. The ATLAS detector, in particular, will provide data from 2008 on, and many institutes all over the

*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**

world consider to work on these data. In order to make this feasible, a hierarchical network of so-called Tier centres is being established, by which the data is shared and replicated in a reasonable fashion and the workload is distributed to the centres.

The raw data remains at the Tier0 centre at Cern, Tier1 centres get preprocessed copies and distribute them to their Tier2 and Tier3 centres, where the simulations and analysis of the data are carried out. The German Tier1 centre is the GridKa facility at the FZ Karlsruhe. In Munich, a Tier2 centre was set up by the two physical institutes MPP and LMU together with the associated computing centres LRZ and RZG. The tasks of a Tier2 centre are to provide a certain amount of compute and memory resources, which are placed at the disposal of all sites taking part in the WLCG (world-wide LHC Computing Grid) community, and to establish sufficient memory bandwidth and services for the data exchange with the associated Tier1 centre. Furthermore, the Tier2 centre has to provide and support a so-called Grid-middleware system which enables all partners in the WLCG to access and use the mutual resources.

The hardware actually provided by the MPP and installed at the RZG consists of 126 worker nodes (approximately 500 CPU cores), several AFS file servers (7.2 TB non-migrating disk space and 25 TB migrating disk space) and 30 TB of dCache disk space for the efficient storage of large data amounts. The LMU at present runs 37 worker nodes (approximately 150 CPU cores) and 10 dCache pools with 40 TB in total, which are hosted by the LRZ. Both partners plan to upgrade their hardware by about 450 CPU cores and 200 TB disk space, which will, however, not be dedicated exclusively to ATLAS, but also to other Grid activities which can be managed with the same infrastructure.

The Grid middleware used by the WLCG is gLite. Besides the worker nodes on which the applications are executed, there are several service machines:

- the User Interface (UI), from which users submit their jobs to the Grid
- the Compute Element (CE), which has the task to convert Grid jobs scheduled to the respective site into local batch jobs and to submit them
- the Storage Element (SE), from which all data stored in the WLCG can be accessed via a common file catalog and data transfers are managed
- the Monbox (MON), which delivers information on the available hardware, the actual load and the software installed at the respective site

gLite provides the services to accomplish the different tasks and to control the interaction between them. A sophisticated authorization and authentication mechanism is used to control the access to the diverse resources.

The main task of the RZG in the Munich-Tier2 centre is to operate and administrate the worker nodes, the service machines and the file systems and to install and maintain the Grid middleware for MPP/RZG, while for the physicists, the emphasis lies on installing application software and testing. But, of course, a close collaboration between all four partners is essential to get and keep the system running. The Munich-Tier2 centre is embedded in the Regional Operating Centre (ROC) DECH (Deutschland/Schweiz) which means a close contact to the associated Tier1 centre and the other Tier2 centres in Germany and Switzerland to share experiences and coordinate activities.

Renate Dohmen & Christian Guggenberger

## Long Time Data Preservation at RZG

Beginning in the late eighties RZG stored data from different experiments of the IPP (W7, W7AS, Asdex, Asdex-Upgrade) and the MPE Gamma group (comptel, egret, integral). These data are stored in a secure way with two copies on either disk or tape.

Since 2005 the RZG together with GWDG has officially the obligation to provide long time storage (at least 50 years) for different projects within the MPG. This very long time is necessary to preserve mankind's cultural heritage such as video and audio records of dying out languages. Presently the following data collections exist:

| | | |
|---|---|---|
| IPP: | | |
| Asdex | (∼1980 – 1990) | 0.08 TB (33500 shot files) |
| W7AS | (1988 – 2002) | 2 TB (∼60000 shot files) |
| Asdex Upgrade | (since 1993) | >30 TB |
| MPE: | | |
| Comptel | (1991 – 2000) | 2 TB |
| Egret | (1991 – 1996) | 0.06 TB |
| Integral | (since 2001) | 8 TB |
| MPP: | | |
| Magic | (since 2003) | >60 TB |

Others (non physics):
Biblioteca Hertziana Rome, photo archive 3 TB
Kunsthist. Inst. Florence, photo archive 4 TB
MPI Psycho Linguistic Nijmegen, video/audio data 5 TB

All these archives presently are in AFS. AFS does the mapping between the filesystem location (path) and the real underlying storage in a way that allows transparently to migrate data from one storage to another without being visible to the user. This is important because the older archives have been migrated already several times as a matter of changes in tape technology, HSM-systems, and operating systems. In the future this will happen certainly many times again.

RZG is responsible only for the bit-stream preservation. Other important issues such as data formats and data-mining tools remain in the responsibility of the owners.

Hartmut Reuter