Bits & Bytes

No. 181 RZG Computer Bulletin Oct 2008

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

vip – The new IBM POWER6 supercomputer at RZG

In May 2008, the new IBM POWER6 supercomputer has been installed at RZG. It replaces the former POWER4 Regatta system *psi* which was taken out of operation in June 2008.

Hardware

The POWER6 cluster consists of 207 compute nodes with 32 processors each, 6 I/O nodes and 1 node for the Hierarchical Storage Management (HSM). So, there are 6624 processors available for computing with a total main memory of 18.5 TB. The peak performance of the system is 120 TFlops.

The POWER6 nodes are water-cooled, thus achieving a very high packing density of up to 448 processors per rack. At 600 GFlops per node, the POWER6 is three times more energy-efficient in GFlops per kW than the Power5 generation of air-cooled processors. The interconnect of the POWER6 cluster is a fast 8-plane InfiniBand network. The total disk space on the POWER6 system is about 400 TB. The permanent user data in the GPFS file systems */u* and */r* on the POWER4 Regatta cluster have been copied to the new POWER6 cluster. The detailed hardware configuration of the compute cluster is as follows:

125	compute nodes with 64 GB memory
78	compute nodes with 128 GB memory
3	compute nodes with 256 GB memory
1	login node with 256 GB memory

The name of the login node is *vip.rzg.mpg.de*.

Software

The operating system of the POWER6 cluster is AIX 5.3 with the traditional parallel programming environment (MPI, ESSL/PESSL) and the Fortran compiler xlf 11.1, C compiler xlc 9.0 and C++ compiler xlC 9.0. The batch system is LoadLeveler 3.4.3.1.

There is an important architectural change of POWER6 compared to the former POWER4 processor which has to be taken into account when setting up a LoadLeveler script: A POWER6 node is made up of 32 processors each capable of supporting two hardware threads. Thus, we have 64 logical CPUs per node, all of which can be used in "Simultaneous Multithreading" (SMT) mode. The SMT mode increases the performance of most applications significantly and is used as the default on the POWER6 nodes. It is, however, also possible to use a POWER6 node in "Single Thread" (ST) mode with 32 CPUs. Therefore, the transition from the POWER4 Regatta system to the new POWER6 system was quite smooth, and most of the user applications could easily be ported.

Like the former Regatta cluster, the POWER6 cluster *vip* is part of the European DEISA (Distributed European Infrastructure for Supercomputing Applications) project. This includes common multi-cluster GPFS file systems that are shared by all DEISA sites as well as the usage of the multi-cluster LoadLeveler batch system. Batch jobs that are submitted at one DEISA site can be executed at some other suitable DEISA site and can automatically access their data in the common GPFS file system.

Operation

Interactive access to the POWER6 system is restricted to one node (*vip.rzg.mpg.de*), and intended mainly for editing and compiling your parallel programs. To run test or production jobs, submit them to the LoadLeveler batch system, which will find and allocate the resources required for your job (e.g. the compute nodes to run your job on). A sample LoadLeveler script can be found at `www.rzg.mpg.de/computing/hardware/Power6/batch-system`. Interactive usage of the Parallel Operating Environment (poe) on the login node is not allowed.

AFS is available ONLY on the login node *vip.rzg.mpg.de*, NOT on the POWER6 compute nodes. So you can access your AFS directories and common software in AFS (e.g. the program IDL) interactively on the POWER6 login node, but not from within your batch jobs on the POWER6 compute nodes.

Ingeborg Weidl & Johannes Reetz

genius – The IBM Blue Gene/P at RZG

The two-rack Blue Gene/P system that was installed at the RZG in September 2007 was extended by a third rack in January 2008, thus increasing the number of cores to 12288 with a total memory of 6 TB. The peak performance is now 40 TFlops.

*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

Since the installation of the POWER6 system, the POWER6 I/O nodes provide the common GPFS file systems */u* and */ptmp* both for the Blue Gene/P and the POWER6 system. By default, the Blue Gene/P data are located in */u/<userid>/BlueGene* and */ptmp/<userid>/BlueGene*. The Blue Gene/P also participates in the shared GPFS file systems of the DEISA project. So, data that are produced on the Blue Gene/P at the RZG can be easily evaluated at some other DEISA site.

Ingeborg Weidl & Johannes Reetz

DEISA2

The DEISA Consortium of leading European supercomputing centres deployed and operated the Distributed European Infrastructure for Supercomputing Applications (DEISA) with a 4 years lasting support through EU FP6. The Consortium continues to support and further develop the distributed high performance computing (HPC) infrastructure and its services through the EU FP7 DEISA2 project (see www.deisa.eu) funded for three years as of May 2008. Activities and services relevant for Applications Enabling, Operation, and Technologies are continued and further enhanced, as these are indispensable for the effective support of computational sciences in the HPC area. The service provisioning model is extended from one that supports single projects to one supporting Virtual European Communities. Collaborative activities will also be carried out with new European and other international initiatives.

Of strategic importance is the cooperation with the PRACE project which is preparing for the installation of a limited number of leadership-class Tier-0 supercomputers in Europe. Both the national Tier-1 centres and the new Tier-0 centres shall be integrated into a European HPC ecosystem as proposed by the European Strategy Forum on Research Infrastructures (ESFRI).

Hermann Lederer

Archiving at RZG

RZG offers different ways to archive your data:

1. Each user has in AFS a volume in the so-called *m*-tree under */afs/ipp/m/<userid>*. There you may store large files. These files get a copy on tape after an hour and will eventually be wiped from disk after some weeks. If you access or prefetch such a file it comes back to disk. To minimize tape mounts you should have there only few, but large files. Therefore the quota is set to 1 TB while only 1000 files are allowed (as default). To archive small files please create big tar-files or zip-archives. The *m*-tree (and other project specific trees) in AFS are good for long time preservation and allow at the same time world wide access to the data.
2. TSM allows you to create archives without the need to first create tar-files. These archives also go onto tape, but you still can list their table of contents by means of TSM query commands.
3. On the POWER6 and Blue Gene/P systems a special GPFS filesystem – the */r*-tree – is visible where files by means of TSM-HSM are migrated to tape. The advantage here is the high data rate you have to this parallel filesystem, the disadvantage is that you can access it only on these HPC-machines. Also here it is necessary to create tar-files in order to limit the number of tape mounts necessary to bring the files back once they are migrated.

For long time preservation (several years) you should choose AFS because underlying HSM systems can be replaced without the users being aware of it.

Hartmut Reuter

Video Conferencing

RZG's Videoconference (VC) Group provides a central gatekeeper for VC systems following the ITU H.323 standard. Thus all registered systems inside and outside the campuses in Garching and Greifswald can be reached worldwide by so-called E.164 numbers obeying the GDS (Global Dialing Scheme). Today approx. 350 systems from all over the world are registered at our gatekeeper. The embedded proxy provides essential firewall security functions blocking all attempts to enter the LANs of IPP and RZG. Many groups at EFDA, ITER, DEISA, HGF and other institutions use this service of RZG's VC group.

In 2008 several new high definition VC systems have been installed at IPP, namely for AUG, the directorate in Garching and Greifswald, and seminar rooms of RZG and TE. The meeting rooms of IPP's board of directors (WL) and RZG have been partly reconstructed, to provide together with the new HD systems optimal communication and high acceptance by the users.

Over the last two years the hype for Green IT has reached videoconferencing, best represented by so-called Telepresence Rooms (CISCO being the most prominent provider). But also the classical VC, as used since many years by IPP and its partners is contributing a lot in this area.

This is difficult to enumerate, nevertheless the long term benefit can be estimated: Assuming 40 meetings of IPP's directorate (participants in Garching and Greifswald), furthermore roughly 20 meetings of the WL, with some 5 people per meeting not having to travel, some 300 travels back and forth between Garching and Greifswald can be spared. Alone the flights Munich–Berlin would have produced approx. 90 tons of CO_2. Assuming furthermore expenses of 750 EUR/travel more than 2 million Euros are saved every year.

Frank Hinterland & Ulrich Schwenn