No. 186
July 2013

Bits & Bytes

RECHENZENTRUM GARCHING

Garching Computing Center of the Max Planck Society and the Institute for Plasma Physics*
Boltzmannstraße 2, D-85748 Garching bei München

New Max Planck Supercomputer Hydra: Main Installation

Hermann Lederer, Markus Rampp

Hydra, deployment phase 2

In autumn 2013, the main part - phase 2 - of the new Max Planck supercomputer *Hydra* will go into operation at RZG.

Phase 1 of the new system has been in operation since autumn 2012. It comprises 612 compute nodes, each with two Intel Sandy Bridge processors (2.6 GHz, 2x8 cores per node). The nodes are connected by Infiniband (FDR14) and are attached to an I/O subsystem with 5 Petabytes of disk space, which split across two GPFS file systems. IBM Loadleveler is used as the batch system.

Phase 2 adds more than 3500 compute nodes with two Intel Ivy Bridge processors per node (2x10 cores and 64 GB RAM per node). A hundred nodes will be equipped with 128 GB memory. In addition, 350 compute nodes will host accelerator cards (two per node). In total, 676 NVIDIA *Kepler* K20X GPUs and 24 Intel *Xeon Phi* coprocessor cards will be deployed.

The new compute nodes will be integrated into the existing I/O subsystem, and Infiniband FDR14 continues to serve as the interconnect. For the operation of the Hydra system, direct ground water cooling will be applied as a cooling system with high cost efficiency.

The Hydra phase 1 (Sandy Bridge based) system will be continued as integral part of the whole system which shares a common software stack.

Accelerator and coprocessor technology

Major research organizations and high-performance computing (HPC) centers are deploying a significant amount of GPU- or coprocessor-accelerated resources worldwide (www.top500.org). Moreover, such architectures with very many, but lightweight 'cores' are expected to play a major role in future HPC systems, mostly due to their superiour energy-efficiency. RZG, in collaboration with MPG scientists has assessed the suitability of NVidia GPUs and the Intel many-core coprocessor architecture (MIC) for HPC applications of the MPG. A number of codes which are already production-ready for a GPU-accelerated cluster were identified and assessed (e.g. classical Molecular Dynamics) or were newly ported to the GPU (Astrophysics, Plasmaphysics, Eigenvalue-solver). As a result, MPG has decided to equip parts of the new Hydra system with GPUs and Xeon Phi coprocessors (MIC). Programming languages CUDA (C, FORTRAN), OpenACC, OpenCL and Intel-MIC programming tools will be supported on the new system.

Intel Software License Agreement for MPG

Hermann Lederer

As of July 1, 2013, the existing software license agreement between Intel and 27 Max Planck Institutes was extended to include all Max Planck Institutes and the GWDG. For a fixed price, the Max Planck Society can now get licenses for the Intel Cluster Studio XE software package (see hocomputer.de/intel/intel-studio-bundles/intel-cluster-studio) for all its institutes (to be ordered at hocomputer.de/mpg-intel). Additional institutes can now join for very attractive conditions. The financial management is handled by the General Administration of the Max Planck Society.

*Tel.: +49(89) 3299-01, e-mail: benutzerberatung@rzg.mpg.de, URL: http://www.rzg.mpg.de/
Editor: H.Lederer & M.Rampp, Rechenzentrum Garching am IPP, Boltzmannstr. 2, 85748 Garching

Software News

Markus Rampp

Graphical Debugger Allinea DDT

RZG has licensed the graphical debugger Allinea DDT for its Linux platforms. The tool is available on the HPC system *Hydra* and on all Linux compute clusters. The license allows to debug up to 256 parallel processes in a single session and includes support for debugging serial and parallel (MPI, threads, hybrid) applications, as well as CUDA-based codes. On the Hydra system, interactive debugging of parallel applications is supported on the development node hydra02. In addition, ddt was preconfigured to use a special 'interactive queue' which allows to run a debugging session on multiple nodes in a user-transparent way. Users simply need to load the `module/ddt`. The totalview debugger, as an alternative tool (with similar functionality) is still available for debugging programs on up to 2 nodes (max. 32 cores).

Upcoming changes to the default software environment on Hydra

In September 2013 (exact date to be announced) the version defaults for the Intel compilers and libraries as well as for the IBM parallel environment (PE) will be raised on the *Hydra* system. The new default version for Intel compilers and the Math Kernel library (MKL) will be 13.1 and 11.0, respectively (currently: 12.1, and 10.3), and the default for the PE will be 1.3 (currently: 1.2). Compiler- and PE-dependent libraries of the Hydra software stack will be updated accordingly by RZG. Besides bug fixes and functionality improvements, the latest Intel compilers and libraries deliver performance enhancements, e.g. concerning AVX vectorization on Sandy Bridge processors. The new version 1.3 of the IBM parallel environment will allow to use the Fabric Collective Accelerator (FCA) technology by Mellanox for accelerating MPI collective operations like, e.g. the `MPI_Allreduce` library call. The FCA technology and first experiences at RZG will be highlighted in a forthcoming issue of Bits & Bytes.

All relevant updates and changes to the RZG software environment are continuously documented for all platforms on our webpage (see www.rzg.mpg.de, 'News and Events') in a chronological list.

Storage/Archive Systems

Manuel Panea

New archive system HPSS

A new archive system called HPSS was put in operation at RZG some months ago. HPSS (High Performance Storage System) provides hierarchical storage management and services for very large storage environments. HPSS is scalable and is designed to store many petabytes of data and to use network-connected storage devices to transfer data at rates of multiple gigabytes per second. HPSS has a cluster design that combines the power of multiple computer nodes into a single, integrated storage system. By increasing the size of the cluster and by adding disks and tape libraries, HPSS is capable of storing and managing thousands of millions of files, and hundreds of petabytes at high data rates. HPSS organizes storage devices into multiple storage hierarchies and uses IBM DB2 as the metadata library for storing the identity, ownership, metadata location and status of all files and devices.

HPSS is the result of a collaborative effort by several US Department of Energy supercomputer laboratories and IBM. Today, many of the leading supercomputing centres worldwide use HPSS, including:

- LANL (Los Alamos National Lab, US)
- LLNL (Lawrence Livermore National Lab, US)
- BNL (Brookhaven National Lab, US)
- ORNL (Oak Ridge National Lab, US)
- SLAC (National Accelerator Laboratory, US)
- NCSA (National Center for Supercomputing Applications, US)
- ECMWF (European Centre for Medium-Range Weather Forecasts, UK)
- DKRZ (Deutsches Klimarechenzentrum, Germany)
- CEA (Commissariat a l Energie Atomique, France)

HPSS can be accessed via several interfaces (FTP, parallel FTP, GridFTP, HSI/HTAR, Linux VFS, GHI). At RZG, all access is currently handled via GHI (GPFS-HPSS-Interface). As such, it replaces the old TSM-based migrating filesystem:

On RZG's HPC systems, users can store data on the migrating filesystem `/r`, which can be accessed from the 'hydra' login nodes. Each user has a subdirectory named `/ghi/r/<initial>/<userid>` to store his/her

data. There is also a symbolic link `/r` pointing to `/ghi/r`, i.e. in practice a user with ID 'smith' would work with `/r/s/smith`. The system continuously monitors the usage of the filesystem. When certain capacity thresholds are exceeded, files get transferred from disk to tape, starting with the largest files which have been unused for the longest time. If (by using some program or command) a user accesses a file which has been migrated to tape, the file will automatically be transferred back from tape to disk. This of course implies a certain delay. The command will appear to hang, but it will just wait until the data is online and then continues. The command `ghi_ls` can be used to display which files are resident on disk and which ones have been migrated to tape.

Every migrated file is written to two different tapes. In this way, in case of a tape failure of the first tape, the file can still be read from the second tape. For more details, see www.rzg.mpg.de/datastorage/tsm/adsm_PSI_qa.html

For backups of desktop computers, RZG will continue to use TSM (Tivoli Storage Manager) as before.

New RZG tape library at LRZ

A new automated tape library for use by RZG's HPSS archive system was recently installed at the Leibniz Rechenzentrum (LRZ), located at the Garching campus. A tape library (also known as 'tape silo') is a large, enclosed cabinet with many slots for storing computer tapes and with several tape drives which can store huge amounts of data. One or more robot arms move the tapes from the slots to the tape drives and vice versa. The primary copy of all files stored in RZG's HPSS archive system, which includes all data in the migrating filesystem '/r' of the High Performance Computer 'Hydra', is stored in a tape library located at RZG. The new library installed at LRZ stores a second copy of all such data, thereby providing a higher level of data reliability, which is further enhanced by the physical separation of the two tape libraries.

The new library is an IBM TS3500 system. In the configuration deployed at RZG, it has 5000 slots for tapes of type LTO and 4 LTO6 tape drives connected via Fibre-Channel to a server machine. This server machine is linked to the others in the HPSS cluster at RZG by two dedicated 10 Gb/s Ethernet connections. The tapes in use are of type LTO5, with a capacity of 1.5 TB uncompressed data (or about 2 TB after compression which is done in hardware by the drives themselves), which results in a total, effective capacity of about 10 Petabytes. The library can be expanded as needed up to a total of about 20000 tape slots.

Redundant connection to the German Research Network

Klaus Desinger

The Munich/Garching Max-Planck-Institutes get their Internet connectivity through RZG's 5 Gbps link to the X-WiN (WIssenschaftsNetz) operated by the DFN (Deutsches ForschungsNetz - German Research Network). Since RZG houses the X-WiN core router for south-east Bavaria one has always benefitted from the fact that all core routers are connected redundantly with optical fibers running on fully separate tracks.

To guard against a failure of our border router or the X-WiN core router itself, RZG currently configures another 5 Gbps link from a new, redundant border router to an X-WiN core router located at Erlangen. Both lines will be used in production so that a doubling of the available bandwidth is gained in addition to the increased redundancy.