Bits & Bytes

No. 176 RZG Computer Bulletin March 2004

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

IBM p690 Supercomputer

In November 2003, 17 nodes of the "Regatta" system were connected to the new, fast IBM High Performance ("Federation") switch. The operating system in this cluster was upgraded to AIX 5.2. On Nov. 20, this cluster was opened for batch processing. Two weeks later, the remaining 8 batch nodes were also connected to the "Federation" switch. There was a local GPFS filesystem */ptmp* available for batch output, while the user filesystem */u* was mounted via NFS (for a transition period) from the two I/O nodes that were still connected to the old "Colony" switch.

The final configuration of the "Regatta" system was achieved in January 2004, when the two I/O nodes were integrated into the "Federation" switch and the size of the filesystem */ptmp* was increased to 13.75 TBytes. The user filesystem */u* was locally available again and the size of */u* was increased to 4.25 TBytes (without data mirroring now). With the final configuration the inter node communication, bandwidth and latency were significantly improved system wide. I/O bandwidth could also be increased beyond 1 GByte/s for a single application.

The migrating filesystem */r* is generally available now and may be used for archiving large files. Files in */r* will be automatically written to tape when the filesystem gets filled up. As soon as the user accesses the files again, they will be automatically retrieved from tape (cf. Migrating Filesystem).

Ingeborg Weidl

Migrating Filesystem

The RZG has installed a migrating filesystem for the "Regatta" system. Data written to this filesystem will automatically be moved from disk to tape to free space on disk when necessary or from tape to disk when they are needed again by the user. This is similar to the functionality offered by the "m-tree" in AFS.

Here is some information about the functioning and usage of this filesystem:

- The migrating filesystem is called */vicepm*. It can be accessed from all "Regatta" nodes. Each user has a subdirectory */vicepm/r/<initial>/<userid>* to store his/her data. There is also a symbolic link */r* pointing to */vicepm/r*, so in practice a user with ID *smith* would work with */r/s/smith*.

- The system constantly monitors the fill level of the filesystem. When the filesystem gets filled above a certain value, files will be transferred from disk to tape, beginning with the largest files which have not been used for the longest time.

- If you access a file which has been migrated to tape, the file will automatically be transferred back from tape to disk. This of course implies a delay. You can manually force the recall of a migrated file by using the command "dsmrecall <filename>". You can recall in advance all files needed by some job with a command like "dsmrecall myfiles/*".

- You can see which files are resident on disk and which ones have been migrated to tape with the command "dsmls". Here is a sample output:

Actual Size/Byte	Resident Size/Byte	Resident Blk/kB	File State	File Name
0	0	0	r	busy.o
110080152	262144	1	m	chgcar.f
6777	6777	8	r	contcar.f
2488308	0	0	m	outcar
6777	6777	8	r	poscar
342708	342708	336	p	potcar
482116256	262144	1	m	wavecar
66692	66692	72	r	xcar.log

  File state "r" means "resident on disk", "m" means "migrated to tape", "p" means "premigrated to tape" (the file has already been copied to tape but is still present on disk and can be removed immediately if the system needs to free disk space).

- Small files will stay resident on disk, only large files can be migrated. Having small files in a migrating filesystem does not make any sense. All users are kindly asked to ensure having only large files on this filesystem. "Large" means in this context larger than 256 kB. If you have many small files which you would like to store on this filesystem, please pack them first together to a large file with a suitable tool like "tar", "cpio", "ar", "zip" or whatever. Here is a simple example of how to use "tar" to pack some

*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

small files “small000”, “small001”, etc. to a big file “big.tar”: *tar cvf big.tar small**.

- In the process of migrating, every file is written simultaneously to two different tapes. In this way, in case of a tape failure while reading back the data from the first tape, the file can probably still be read from the second tape. This happens transparently for the user. (In addition, a backup of all data, whether resident on disk or migrated to tape, will be made to a separate set of tapes in order to be able to recover the filesystem in the event of a disk crash.)

If you have any questions or want to report any problems, please contact:

Manuel Panea, <mpd@rzg.mpg.de>, Tel. 3299-1133
Ingeborg Weidl, <ifw@rzg.mpg.de>, Tel. 3299-2191

Manuel Panea

Linux Cluster Expansion

Linux clusters at RZG have become quite popular since the first rack based installations for the Fritz-Haber-Institut in autumn 2002. In early 2003, very compact blade center systems were installed as group or departmental compute servers for MPI for Astrophysics, MPI for Polymer Research, MPI for Quantum Optics, and for RZG as general compute servers. Recently a series of further institutes has decided for that technology to be installed, operated and maintained (“housed”) at RZG as group compute servers: IPP for the Tokamak physics department, the Surface physics and the Plasma physics E3 departments, MPI for chemical Physics, and a consortium of institutes as bioinformatics compute servers (MPI of Developmental Biology, MPI of marine Microbiology, MPI for Biochemistry, MPI for Computer Science). The new systems are equipped with 2.8 GHz Xeon double processors and 1.5 or 2 GB double interleave memory in the so-called “Chipkill” memory subsystem per node. (The most recently ordered system will have a clock rate of 3.06 GHz.) Most new systems have already been installed and equipped with the Sun Grid Engine Enterprise edition (SGEE) as batch system; batch classes have been individually configured according to the respective needs of the groups or departments. As a now general I/O concept the different blade centers have been connected to dedicated head-nodes (with channel bonded Gigabit-Ethernet connections) which host the individual fibre channel attached Raid Disk systems.

Hermann Lederer

Numerical Libraries

WSMP

The Watson Sparse Matrix Package, “WSMP”, (see www-users.cs.umn.edu/~agupta/wsmp.html), is a high-performance, robust, and easy to use software package for solving large sparse systems of linear equations. It uses a direct method on serial and multiprocessor workstations and on distributed-memory parallel computers with serial or multiprocessor nodes. The library is available now under AIX and Linux (see www.rzg.mpg.de/docs/libraries/wsmp.html. For correct functioning the environment variable WSMP-LICPATH should have been set by the system to */afs/rzg/@sys/lib/wsmp* to specify where the license file is located. If this is not the case the user has to do it by himself depending on the used shell in the *.profile* or *.login* files in the home directory.

Intel MKL

The Intel Math Kernel Library (Intel MKL) for LINUX (see www.rzg.mpg.de/from_external/intel/-mkl/doc/index.htm), is composed of highly optimized mathematical subroutines. It contains e.g. the Linear Algebra PACKage (LAPACK), the Basic Linear Algebra Subprograms (BLAS), and the extended BLAS (sparse). For cluster computing, it provides e.g. ScaLAPACK (Scalable LAPACK) and supports functionalities like the Basic Linear Algebra Communications Subprograms (BLACS) and the Parallel Basic Linear Algebra Subprograms (PBLAS). In order to use LAPACK and BLAS software of the Intel MKL, you must link two libraries: LAPACK and the processor specific kernel, e.g. *-L/afs/rzg/@sys/lib/mkl -lmkl_lapack -lmkl_ia32 /afs/rzg/@sys/lib/mkl/libguide.a -lpthread* for static linking, LAPACK library, Pentium, Pentium III/4 processor kernels.

Numerical Recipes

“Numerical Recipes: The Art of Scientific Computing” is the title of a series of books developed by Numerical Recipes Software and published by Cambridge University Press. Numerical Recipes is a complete text and reference book on scientific computing. Its aim is to provide general discussion, analytical mathematics, algorithmics, and actual working programs. The RZG provides within the scope of a site license the source code from the Fortran Numerical Recipes book. The location within AFS can be looked up at www.rzg.mpg.de/from_external/docs/num_recipes-.html.

Roman Hatzky