Universität Karlsruhe (TH)
Rechenzentrum

HIGH PERFORMANCE SCIENTIFIC COMPUTING

Contents

Universität Karlsruhe (TH)
Rechenzentrum
Scientific Supercomputing Center Karlsruhe
Zirkel 2
D-76128 Karlsruhe

Phone: +49 721 608 8011
Fax: +49 721 32550
xc-hotline@uni-karlsruhe.de
http://www.rz.uni-karlsruhe.de/ssc

Editor: Prof. Dr. Wilfried Juling

June 2005

Scientific Supercomputing Center Karlsruhe (SSCK)

The Scientific Supercomputing Center Karlsruhe is the service unit of the University's Computing Center that cares for supercomputer customers. The SSCK provides support for experts as well as for novices in parallel computing. Our goal is to help our customers in all problems related to scientific supercomputing. Our services are not only confined to advice on how to use supercomputers efficiently but you will also get qualified help if you are looking for appropriate mathematical methods or simply having problems to login.

Since the installation of the first supercomputer in Karlsruhe in 1983 the Computing Center of the University has been continuously engaged in supercomputing. Besides the operating of the machines Karlsruhe has always been establishing expert groups for scientific supercomputing. In the 1980s and 1990s experts of the Computing Center in Karlsruhe tuned numerical libraries for vector computers as well as microprocessor based parallel supercomputers. From this time on close cooperation with other institutes of the university and industrial companies has been initiated.

At the Computing Center solvers for arbitrary systems of nonlinear partial differential equations and iterative linear solvers have been developed. Thus, the experts at the SSCK know about the needs and problems of their customers from their own experience.

The SSCK has allied with the HLRS from Stuttgart University and the IWR from Heidelberg University. The High Performance Computing Competence Center Baden-Württemberg (hkz-bw) combines HPC resources in the State of Baden-Württemberg in order to provide leading edge HPC service, support and research.

Consultation and Support

According to the mission of the Scientific Supercomputing Center Karlsruhe you will have support on different levels:

- **Hotline**
  The SSCK has set up a helpline where you rapidly get support from Monday to Friday from 9:00 to 17:00. Simply call +49(0)721/608-8011. If your problem cannot immediately be solved, it will be pursued by our experts. You can contact us as well by e-mail: xc-hotline@uni-karlsruhe.de.
- **Basic support**
  For parallel computers there are rules that should be obeyed in order to get efficient codes based on the hardware design. We will provide you with all our knowledge how to use massively parallel and vector computers efficiently.
- **Standard solutions**
  At the SSCK you will find experts who know both: the application software packages and the necessary libraries you need to solve your problem. This means that you obtain assistance from the beginning in selecting your appropriate solution and later in using it.
- **Individual solutions:**
  If you want to solve your particular problem for the first time with the aid of a high performance computer, then you will find at the SSCK even experts for modelling and state-of-the-art solvers. This unique feature is highly considered by all our customers.

The SSCK offers several ways to communicate with its customers using the HP XC6000 cluster:

- Mailing list for communication between SSCK staff and users
  xc-users-l@lists.uni-karlsruhe.de
- E-mail hotline for all inquiries, comments etc.
  xc-hotline@uni-karlsruhe.de
- Website of SSCK with lots of information on XC6000, its hardware and software environment and project submission interface
  http://www.rz.uni-karlsruhe.de/ssc

HP XC 6000 Cluster at the SSC Karlsruhe

The largest system that is currently operated by SSCK is an HP XC6000 cluster. This supercomputer is provided for projects from the universities of Baden-Württemberg requiring computing power which cannot be satisfied by local resources. The system is embedded into the infrastructure of the High Performance Computing Competence Center Baden-Württemberg (hkz-bw). Through the private-public partnership hww the system is also available to industrial customers.

The installation of the HP XC6000 Cluster began in April 2004 with a small 16 node test system named xc0. In December 2004 the first production system (xc1) with 108 HP Integrity rx2600 servers followed. After a short installation and internal test phase first customers were able to access the system by the end of 2004. Full production started on March 1, 2005. In May 2005 another six 16-way rx8620 servers were installed and integrated into the xc1 system.

By early 2006 the overall system will be upgraded to about 340 nodes, 1200 processor cores, a peak performance of 11 TFlops, 7 TB of main memory and more than 40 TB of global shared disk space.

The three steps of the XC6000 installation at SSCK are:

Development and test cluster xc0 (April 2004):

- 12 two-way rx2600 nodes with 4 GB main memory
- 4 file server nodes
- Single rail QsNet II interconnect
- 2 TB shared storage

Production cluster xc1 (December 2004 / Mai 2005):

- 108 two-way rx2600 nodes with 12 GB main memory
- 6 sixteen-way nodes with 128 GB main memory (configured as 12 eight-way nodes)
- 8 file server nodes
- Single rail QsNet II interconnect
- 11 TB global disk space

Production cluster xc2 (Early 2006):

- 218 four-way nodes
  - Two sockets
  - Dual core Itanium2 processor (codenamed Montecito)
- Single or dual rail QsNet II
- 30 TB global disk space

Architecture of the XC6000 Cluster

The HP XC6000 Cluster is a distributed memory parallel computer where each node has two or more Intel Itanium2 processors, local memory, disks and network adapters. Special file server nodes are added to the HP XC6000 Cluster to provide a fast and scalable parallel file system. All nodes are connected by a Quadrics QsNet II interconnect.

The HP XC software environment is based on HP XC Linux for HPC, a Linux implementation which is compatible to Redhat AS 3.0. On top of this basic operating system a set of open source as well as proprietary software components constitute the XC software stack. This architecture implements a production class environment for scalable clusters based on open standards which enables the application programmer to port easily applications to this system.

The nodes of the HP XC6000 Cluster may have different roles and are separated into disjoint groups according to the services supplied by each node. From an end users point of view the different groups of nodes are login nodes, compute nodes, file server nodes and cluster management nodes.

- Login Nodes
  The login nodes are the only nodes that are directly accessible by end users. These nodes are used for interactive login, file management, program development and interactive pre- and post-processing. Several nodes are dedicated to this service. But the Linux Virtual Server (LVS) provides a single login to the whole cluster. LVS will distribute the login sessions to the different login nodes.
- Compute Nodes
  The majority of nodes are compute nodes which are managed by the batch system. Users submit their jobs to this batch system. A job is executed depending on its priority, when the required resources become available. Several queues with specific characteristics for development and production can be implemented.
- File Server Nodes
  Special dedicated nodes are used as servers for the HP StorageWorks Scalable File Share (HP SFS). HP SFS is a parallel and scalable file system product based on the Lustre file system. In addition to shared file space there is also local storage on the disks of each node.
- Management Nodes
  Certain nodes are dedicated to additional services within the HP XC6000 cluster. This includes resource management, external network connection, administration etc.

All the management tasks are served by dedicated management nodes. So there is a clear separation between resources that are used for management or administration and resources used for computing, i.e. the compute nodes are mostly freed from administrative tasks.

Configuration of the Phase I System xc1

The HP XC6000 cluster at the SSC Karlsruhe for the first production phase (xc1) consists of

- 108 two-way HP Integrity rx2600 nodes:
  Each of these nodes contains two Intel Itanium2 processors which run at a clock speed of 1.5 GHz and have 6 MB of level 3 cache on the processor chip. Each node has 12 GB of main memory, 146 GB local disk space and an adapter to connect to the Quadrics QsNet II interconnect.
- 6 16-way HP Integrity rx8620 nodes:
  For a transition period each of these nodes is partitioned into two units, with 8 processors, 6 GB of main memory and 500 GB local disk space. The Intel Itanium2 processors of the rx8620 nodes run at a speed of 1.6 GHz and have 64 MB of level 3 cache. Each eight CPU partition has its own connection to the QsNet II interconnect.
- 8 HP Proliant DL 360 file server nodes:
  Through a storage area network these nodes are connected to seven HP EVA5000 disk systems. This global shared storage has a capacity of 11 TB. It is subdivided into a part used for home directories and a larger part for non permanent files.

Two eight-way partitions of the rx8620 nodes are used as login nodes while the other eight-way nodes as well as the majority of the two-way nodes are compute nodes which run parallel user jobs.

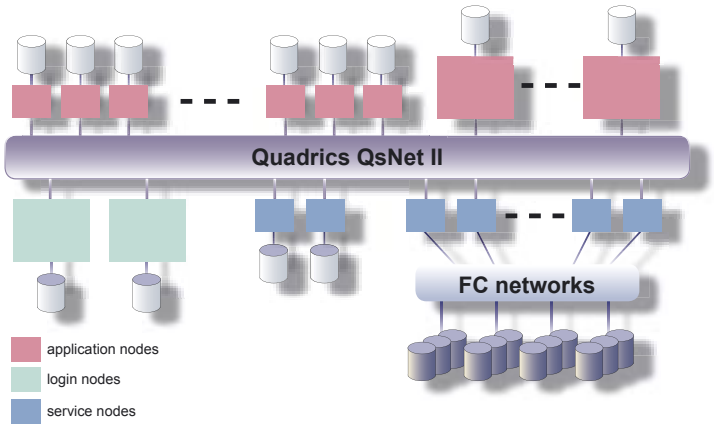


Figure 1: General layout of the HP XC6000 Cluster at the SSC Karlsruhe

The heart of the xc1 cluster is the fast Quadrics QsNet II interconnect. All nodes are attached to this interconnect which is characterized by its very low latency of less than 3 microseconds and a point to point bandwidth of more than 800 MB/s between any pair of nodes. Both values, latency as well as bandwidth, have been measured at the MPI level. The sustained bisection bandwidth of the 128 node network of the xc1 is more than 51 GB/s. These excellent performance figures of the cluster interconnect makes the xc1 ideal for communication intensive applictions.

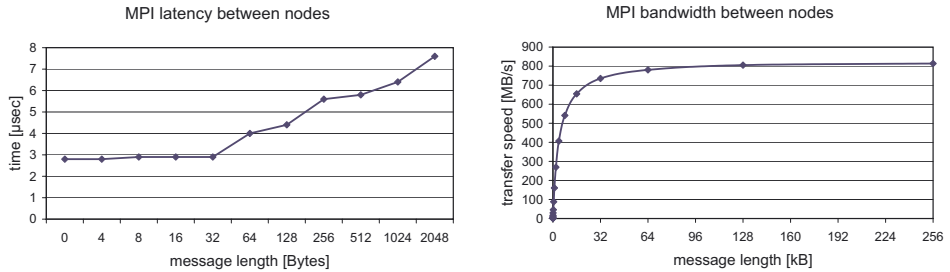


Figure 2: Performance of the cluster interconnect

The key figures of the xc1 are:

	2-way rx2600 nodes	16-way rx8620 nodes
Number of nodes	108	6
CPUs per node	2	16
Peak performance per CPU	6 GFlops	6.4 GFlops
Memory per Node	12 GB	128 GB
Local disk space per node	146 GB	1060 GB
Peak network bandwidth	1.3 GB/s	2.6 GB/s
Sustained network bandwidth	800 MB/s	1600 MB/s

Total peak performance	1.9 TFlops
Total main memory	2 TB
Total local disk space	22 TB
Shared disk space	11 TB
Bisection bandwidth	83 GB/s
Sustained bisection bandwidth	51 GB/s

The Parallel File System HP SFS

In modern compute clusters the CPU performance, the number of nodes and the available memory is steadily increasing, and the applications' I/O requirements are often increasing in a similar way. In order to avoid loosing lots of compute cycles by applications waiting for I/O compute clusters have to offer an efficient parallel file system.

The parallel file system product on HP XC6000 clusters is HP StorageWorks Scalable File Share (HP SFS). It is based on Lustre technology from Cluster Filesystems Inc. (see www.lustre.org).

Architecture of HP SFS

In addition to the standard Lustre product HP SFS includes additional software for failover and management and is restricted to certain hardware components.

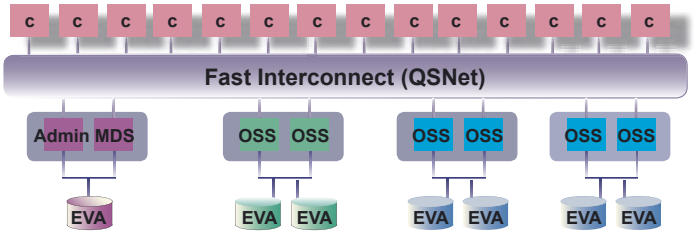


Figure 3: SSCK's HP SFS system

Figure 3 shows the structure of the HP SFS system at SSCK. The servers are configured in failover pairs: The administration node and the Meta Data Server (MDS) node build one pair and the Object Storage Server (OSS) nodes are grouped into pairs so that for each OSS there is another one that can act as its backup. If the MDS node fails a takeover of the corresponding services is done by the Admin node and vice versa.

At SSCK's HP XC6000 system eight HP SFS servers and two file systems are configured. One file system called *data*, is used for home directories and software, has 3.8 TB storage capacity, and uses two designated OSS. The other file system called *work* is used for scratch data, e.g. intermediate files between different job runs. It has 7.6 TB capacity and uses four OSS.

Features of HP SFS

The most important requirement for parallel file systems is performance. Parallel file systems typically offer completely parallel data paths from clients to disks even if the files are stored in the same subdirectory. In HP SFS data is transferred from the clients via the fast cluster interconnect to servers and from there to the attached storage. The data is striped over multiple servers and multiple storage systems.

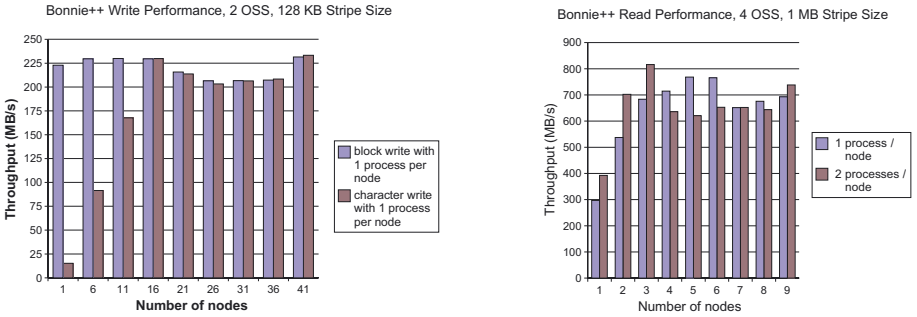


Figure 4: Block-wise write and read performance of the file system work

Figure 4 shows measurements for the sequential block-wise write and read performance of the file system *work*. A single client can reach 400 MB/s for writes and 300 MB/s for reads. On the server side the bottleneck for writes is the storage subsystem. For reads the FC adapter on the servers is the restricting hardware component. Since the file system *data* uses half the number of OSS, the throughput performance is half as high as for the file system *work*.

Besides throughput the metadata performance is important. In fact many parallel file systems are restricted to allow only tens or few hundreds of file creations per second. On the xc1 cluster at SSCK more than 5000 file creates per second have been measured. This metadata performance is shared between the two file systems *work* and *data*.

In a cluster the scalability of the file system is required: The parallel file system must work even if all clients are heavily doing I/O. Figure 5 shows that the performance stays pretty much at the same rate when lots of clients are doing I/O into the file system *data*. It also shows that character-wise I/O only has an impact on the performance of clients. This is because Lustre gathers small I/O packets into large buffers.

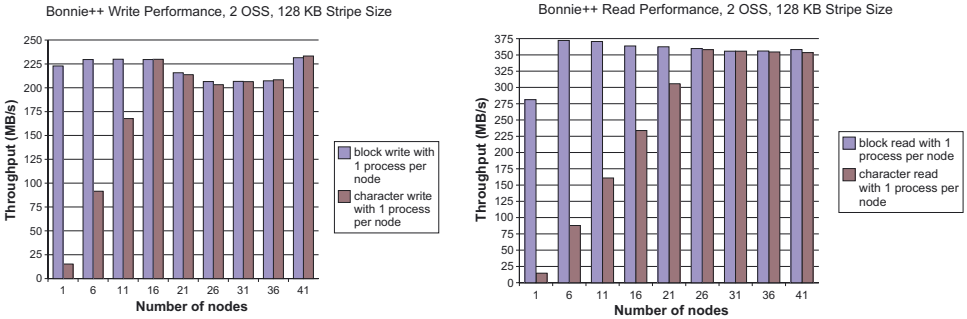


Figure 5: Character-wise and block-wise write and read performance of the file system data

Applications on the XC6000 Cluster

With its different types of nodes the XC6000 cluster at SSCK can meet the requirements of a broad range of applications:

- applications that are parallelized by the message passing paradigm and use high numbers of processors will run on a subset of the two-way rx2600 nodes and exchange messages over the Quadrics inter connect.
- applications that are parallelized using shared memory either by OpenMP or by explicit multithreading with Pthreads can run on the 8-way nodes or soon on 16-way nodes.

In the future even applications combining both parallelization paradigms may be supported.

As all nodes have at least 6 GB of main memory for each CPU the system is especially suited for applications with high memory requirements.

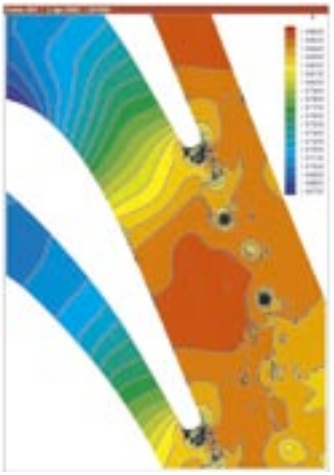
The Intel Itanium2 processors of the HP XC6000 cluster are characterized in particular by a high floating point performance and a very large cache, which is located on the processor chip. Therefore it can be accessed with very short latency and at extremely high bandwidth. Optimization of application programs for the architecture of the Itanium2 processor have demonstrated that an increase of performance by a factor of two can be reached.

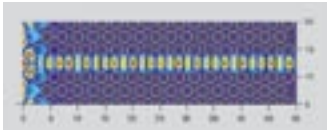
Only a few months after start of operation of the XC cluster at the SSCK user projects from different research fields have been initiated including:

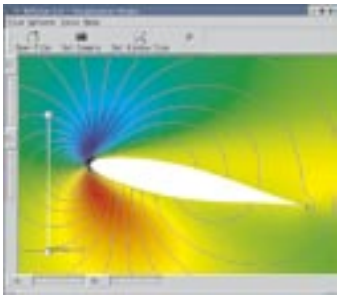
- Numerical Analysis
- Chemistry
- Life Sciences
- Computational Fluid Dynamics
- Structural Mechanics
- Computer Science
- Electrical Engineering
- Solid State Physics

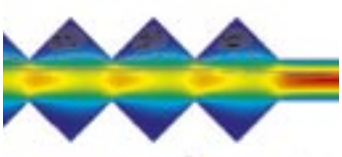
Instead of developing and porting own programs software from independent software vendors (ISV) may be used on XC6000 clusters. Many ISV codes are available today and SSCK is working together with HP, to increase the list of codes that are enabled for XC6000 clusters.

Programs that are available for the XC6000 cluster include CFX-5, Fluent, StarCD, MSC.Nastran, ABAQUS and LS-DYNA. More information, on which codes have been installed already at the SSCK and at which conditions they can be used are published on the SSCK website.









Program Development Environment

Since the operating system of the XC6000 is Linux and many parts of the XC system software are based on open source products a rich set of program development tools is available or can be installed within the XC environment.

Compilers
Different versions of the Intel C/C++ and Fortran compilers, the latest GNU C/C++ and Fortran compilers as well as the NAG Fortran95 compiler have been installed on xc1. All compilers can compile programs that are parallelized with the message passing paradigm. The Intel compilers and the NAG compiler also support the OpenMP standard.

Parallelization Environments
For distributed memory parallelism HP MPI is available. It supports the full MPI 1.2 specification but also includes many functions of MPI 2. MPI IO is supported for the parallel file system HP SFS and shows high parallel performance. Specific interface libraries allow object code compatibility with MPICH.

Debuggers
Besides the serial debugger idb form Intel and the GNU debugger gdb the Distributed Debugging Tool (DDT), a parallel debugger from Allinea Ltd. is available on xc1. DDT supports a graphical user interface and is able to debug serial and MPI-parallelized as well as thread-parallelized programs.

Performance Analysis Tools
The distribution of communication and computation can be visualized and analyzed with Intel's tracecollector and traceanalyzer package. Some other performance anlysis features are offered by the runtime environment of HP MPI.

In order to do a detailed analysis of resource usage different profiling tools and interfaces to the performance counters of the Itanium2 processor are being studied by the SSCK and will be made available to end users. This includes tools like Intel's VTune, oprofile, perfmon and others.

Mathematical Libraries
Numerical subprogram libraries that have been installed on the xc1 cluster include Intel's Mathematical Kernel Library (MKL), HP's mathematical library MLIB, the NAG C and Fortran libraries as well as the linear solver package LINSOL from SSCK.

Tuned implementations of well established open source libraries like BLAS, LAPACK, ScaLAPACK or Metis are part of MKL or MLIB.

More detailed information on the software environment can be found on the web (see http://www.rz.uni-karlsruhe.de/ssc/software).

High Performance Technical Computing Competence Center (HPTC^3)

Together with Hewlett Packard and Intel, the SSCK has established the High Performance Technical Computing Competence Center (HPTC^3). The main goals of this partnership are the further development of the HP XC systems to make it even more usable and reliable for the customers of the SSCK. This includes development, testing and early deployment of advanced tools supporting the application development process. In close cooperation with end users and independent software vendors the HPTC^3 will extend the portfolio of application packages that is available on HP XC6000 clusters. The tuning and optimization of applications to reach highest possible performance is another challenging goal of HPTC^3.

Projects that have been started in the framework of HPTC^3 include:

- Integration of LDAP into the XC software environment
- Improvements of high availability of critical resources of the XC cluster
- Tuning of application codes
- Early test of new tools in a production environment
- Organization of workshops