Storage Networking in European Grid Projects

Peter W. Haas, Michael M. Resch
University of Stuttgart, HLRS
haas@hlrs.de, resch@hlrs.de

Abstract

High performance computing has gradually shifted from the realm of research into development and partially even into the production cycles of industry. High performance computers therefore have to be integrated into production environments that demand the simultaneous solution of multidisciplinary physics problems. Supercomputer centers can learn from these new challenges imposed by industry. The concepts of work flow and production cycle open up a new horizon for integrating systems and software into what is called a distributed "Teraflop-Workbench" approach. Tera- or rather Exascale storage and communication infrastructures will be needed to support such an environment.

1. Introduction

Based on a long tradition in supercomputing at the University of Stuttgart, HLRS was founded in 1995 as a federal center for High-Performance Computing. HLRS serves researchers at universities and research laboratories in Germany and their external and industrial partners with high-end compute power for engineering and scientific applications.

2. The HLRS Framework

2.1. Target

Due to its embedding in the industrial environment of Southwest Germany's high-technology region around Stuttgart, HLRS traditionally focuses on applications from the engineering sciences like Computational Fluid Dynamics, Combustion, Structural Mechanics, Electromagnetics and Process Engineering. However, recently HLRS has extended its portfolio to medical applications, environmental applications and has started an initiative to bring new fields of application to the supercomputer.

The services of HLRS offer complete support to user groups such as physical modeling of parallel numerics and algorithms, the embedding of commercial packages and the visualization of results either remotely or in HLRS' virtual reality laboratory at Stuttgart.

Research groups at HLRS are involved in numerous projects targeting at the industrial use of HPC technology and at the further improvement of such technologies based on user feedback. This includes the participation in the standardization of base technologies.

2.2. Organization

HLRS is a central facility of the University of Stuttgart. Its responsibility is to support supercomputing at the national level and at the university. HLRS operates computing platforms together with T-Systems, T-Systems sfr and Porsche in a joint company named hww (Höchstleistungsrechner für Wissenschaft und Wirtschaft GmbH). The universities at Heidelberg and at Karlsruhe are also shareholders of this company. The purpose of this public-private partnership is the sharing of resources in order to benefit from synergies. This allows for a broader diversification of the available architectures. At the same time common funding allows for higher budgets which in turn allow the installation of larger systems. Both research and industry benefit from this. While it is the role of HLRS to provide access to all platforms for researchers, T-Systems does marketing for industry.

2.3. HLRS Computer Configuration

The HLRS/hww computer configuration is distributed over multiple campuses. It uses state of the art equipment in a highly secure environment. The major campuses are located at the universities of Stuttgart, Heidelberg and Karlsruhe as well as within the premises of DaimlerChrysler AG, Deutsches Zentrum für Luft- und Raumfahrt (DLR), Porsche AG, and T-Systems GmbH. Most of the communication lines use dedicated fiber links with optical multiplexors that allow for an appropriate collection of link signals, e.g. Ethernet, Fibre Channel or InfiniBand. Transmission speeds typically range between one and forty Gbit/s.

The HLRS/hww compute systems provide a fair coverage of today's high performance computing architectures and implementations. The spectrum extends from Shared Memory Parallel systems (SMPs), like NEC Asama and Bull Novascale, via Parallel Clusters, like Cray Opteron and NEC Nocona, to Massively Parallel Processors (MPPs), like HP XC6000, and finally to Parallel Vector Processors (PVPs), such as NEC SX-6 and NEC SX-8. Individual system performance may range from 200 Gflop/s up to 12 Tflop/s.

The HLRS computer configuration is one of the three corner-pillars supporting the Ex@Grid framework.

3. Ex@Grid Framework

3.1. Gauss Centre for Supercomputing

The Gauss Centre for Supercomputing, www.gcfs.eu, the alliance of the John von Neumann Institute of Computing (NIC), the Leibniz-Rechenzentrum (LRZ), and the Höchstleistungsrechenzentrum Stuttgart (HLRS), provides one of the largest and most powerful supercomputer infrastructures in Europe. The German Ministry of Science and Education (BMBF) has announced to support the development of Ex@Grid, a very high-speed data communication backbone between the three centers, by 30 Million Euros. This is to promote the scientific co-operation between all major HPC sites in Germany and in particular between their user communities in the area of distributed HPC applications, virtual and augmented reality, storage systems and networking.

Ex@Grid wants to provide a complete all-optical, forward error corrected network for the high performance computing centers in Germany. This will enable new services, like distributed HPC workflows based on a common network-centric data management infrastructure, for all scientific users since the network services can be very low latency due to lack of contention as well as datagram loss.

3.2. Ex@Grid Design

In conventional packet communication networks we have learned to take the lowest cost switching option for express traffic depending on bandwidth and latency. Express traffic has been moved from "fat IP routers" to layer-2 switches over time. The next significant cost reduction will be achieved by optically bypassing very high rate express traffics on demand. Switching times for transparent optical paths are ~ 1ms.

Therefore, the design of Ex@Grid tries to follow the construction principles of intersection-free highways where the highways and feeders together form 3-dimensional add/drop systems. This can be perceived as a direct equivalent to a degree-3 ROADM system. In comparison however, optical transmission systems are not restricted with respect to the number of lanes. Up to ~ 40 different wavelengths (colors) are feasible within the ITU-100 Grid, and each of those wavelengths may be used to carry a payload signal of up to 100 Gbit/s.

4. Consolidation in the Datacenter

4.1. Datacenter Ethernet

In future, it will be very important to restrict the number of link signals and communication protocols because scalable gateways between dissimilar networks are either very expensive or prone to loss of datagrams or other forms of information. The IEEE 802.3 standardization efforts toward the Datacenter Ethernet [1] all aim at the integration of LAN, SAN and IPC into a common Ethernet-based switching technology which will enable Quality of Service (QoS) via service-specific configuration parameters. The direct provisioning of transparent optical channels between HPC applications may turn out to be a very elegant solution from the users' perspective. Here, users may select the most appropriate communication protocols for the duration of a link provisioning in a way that is best suited to their application. Thus information loss due to transmission errors or blocking inside multi-stage switching networks will be avoided at large.

4.2. IEEE 802.3ar: Congestion Management

We have tried to compose a top-level view of the most important definitions of the IEEE P802.3ar Congestion Management Task Force [2] which have been collected from information publicly available under URL:

http://www.ieee802.org/3/ar/public/index.html. First, the overview tries to address the essential differences amongst the established network cultures with respect to traffic type, preferred characteristics as well as the associated transmission latencies. Below that there are three categories, namely Parallel Links, Virtually partitioned Links and Flow control, which are supposed to enable the coexistence of today's three most prevalent traffic types - within a single Ethernet, finally. We aim at making Datacenter Ethernet technology available within Ex@Grid at a very early point in time (both within switching systems and at the local end system's interface). The final version of the IEEE 802.3ar standard is supposed to be available in September 2007. This should enable first industry products at the beginning of 2008.

5. Multicluster Parallel File Systems

5.1. Multicluster GPFS in DEISA

The Distributed European Infrastructure for Supercomputing Applications, DEISA, www.deisa.org, is a consortium of leading national supercomputing centers that currently deploys and operates a persistent, production quality, distributed supercomputing environment with continental scope. The purpose of this FP6 funded research infrastructure is to enable scientific discovery across a broad spectrum of science and technology, by enhancing and reinforcing European capabilities in the area of high performance computing. This becomes possible through a deep integration of existing national high-end platforms, tightly coupled by a dedicated network and supported by innovative system and grid software.

All high performance computing systems share data among the computing nodes with a Cluster File System, which offers users a single system data view and transparent data access. The extension of this data sharing model to a grid of geographically distributed HPC systems over a wide area network leads to the concept of a Global File System or Grid File System, which avoids data replication.

In a first step four DEISA sites, all running AIX systems, integrated their local IBM GPFS file systems via the 10-Gbit/s dedicated DEISA wide area network into a joint Grid File System, using newly added features in the GPFS software. The step from a homogeneous to a heterogeneous environment was performed by the integration of the PowerPC-Linux system at BSC as well as the SGI Altix systems at SARA and LRZ. Additional software will enable hierarchical storage management functionality on top of GPFS via TSM, thus conveying the perception of an infinite space distributed storage system.

5.2. Network-centered parallel HSM systems

The High Performance Storage System (HPSS) is a parallel data management software that provides both fast access to data as well as services for very large storage environments. It has been designed with high performance computers and data handling systems in mind and can easily adapt to the throughput and capacity increases required in this field [3], [4].

HPSS may be of interest in situations having present and future scalability requirements that are very demanding in terms of total storage capacity, file sizes, data rates, number of objects stored and number of users. The focus of HPSS is the network. Data storage and archiving is distributed across a high performance network with user selectable quality of service. Hence users may access their data randomly and directly irrespective of their hierarchical grouping. HPSS is one of the very few storage systems that allow control of the communication parameters as part of the service class concept.

HLRS has been actively working in the area of layering parallel file systems via extensions to Data Management APIs, e.g. the parallel version of XDSM DMAPI. This has helped to raise the acceptance level for interworking scalable parallel file systems at HPC sites. IBM has announced extensions to its GFPS and HPSS file systems that would allow for a layering of GPFS on HPSS under a common name space. Basically, user level agents will be placed on the front end file system (i.e. GPFS) that exploit HSM control information in order to automatically migrate or stage user data files or entire file objects, respectively.

At leading HPC sites, there has been a holistic view of the global file system idea for a next generation supercomputer center for some time. This idea was (and still is) that all storage requests should be directed towards a single, site-wide storage repository. However, a closer look to the technological implications of the computer memory hierarchy will reveal, that there is still sufficient need and justification for more than just a single parallel file system.

5.3. Projected mass storage system at HLRS

The projected mass storage system at HLRS [5], e.g., is going to implement a layering of GPFS on HPSS which will definitely suffice for the pre- and post-processing stages in our HPC workflow as well as for most of our scientific users. This is either by PNFS exports or GPFS mounts on the client systems. However, the NEC SX-8 parallel vector system will maintain its local NEC GFS production file system, and rather employ local HPSS data movers in user space in order to transfer HPSS files into the parallel vector's main memory and vice versa. Of course, this procedure greatly reduces interdependencies between compute and storage subsystems. Even more importantly, it helps to mitigate the huge discrepancies in performance that tend to develop over time between successive generations of HPC systems.

6. The Teraflop Workbench Concept

When HLRS started its latest request for proposals for an HPC platform it was clear from the beginning that the system offered would have to be part of a larger concept of supercomputing called the Stuttgart Teraflop Workbench [6], [7]. The basic concept foresees a central file system where all the data reside during the scientific or industrial workflow. A variety of systems of different performance and architecture are directly connected to the file system. Each of them can be used for special purpose activities. In general, one can distinguish between systems for pre-processing, high-performance computing and post-processing.

The concept is centered around NEC's Global File System (GFS). NEC SX-8, Asama (IA-64 shared memory system) and a cluster of Intel Nocona processors all have direct access to the same data via a Fibre Channel storage area network.

6.1. Workflow Example

Most users tend to prepare their mesh on one processor before they decompose it and transfer it to the parallel system. Given that the main memory of the core system is 9 TB, we decided for one node with large memory to be able to prepare large jobs.

A cluster based on Intel EM64T processors and InfiniBand interconnect is added to the workbench. It serves both for post-processing/visualization and as a compute server for multi-disciplinary applications. The latter often require different types of architectures for different types of disciplines. The 200 node cluster is connected by a Voltaire InfiniBand switch with a bandwidth of 10 Gbit/s.

6.2. Fenfloss

The Institute for Hydraulic Machinery (IHS) of the University of Stuttgart and HLRS commonly work on the Finite Element Numerical FLOw Simulation System (Fenfloss), an integrated environment to design and shape hydraulic turbines. Computer tools will be used to perform a design for each specific water power plant. Numerical simulations warrant a high quality of the final design and optimize the overall efficiency. Insights gained during the analysis of flow simulations will thus immediately lead to design modifications.

The simulation process chain is partially established in the integrated environment. It is based on a parameterized runner design that enables a numerical optimization of axial and radial hydraulic turbines. COVISE, a distributed Collaborative Visualization and Simulation Environment developed by HLRS [8], is used as integration platform for the profile generation, the runner contour generation and the grid generation of the entire machine. The definition of boundary conditions based on the operating point, the coupled simulation around runner and guide wheel and the overall process chain will be controlled from within a virtual reality environment.

Fenfloss exhibits strong scaling with the number of nodes and has been found to reach a 50-percent efficiency on the NEC SX-8.

6.3. Demonstration in the HLRS Cave

Scientific visualization techniques are used by scientists and engineers to understand complex simulations [8]. They comprise filtering data, generation of deduced information, mapping of data onto visual representations and their display, finally. Distributed software environments often couple simulations on remote machines with local visualizations.

Virtual Reality techniques (VR) complement visualization methods to improve the comprehension of complex content and spatial relationships. Stereo projection rooms are used by groups of experts to enter a three dimensional virtual world consisting of visualized data or geometric representations of engineering parts. In such environments users are able to perceive interrelationships of complex structures and navigate with them. Interactions such as inserting particles into a flow field become possible, as can be seen in an MPEG movie of a virtual reality demonstration of a complete water power plant at Kiebingen near Stuttgart.

6.4. Virtual Tour of Kiebingen Water Power Plant

A Covise (Collaborative VISualization Environment) is used to demonstrate the water flow through a parameterized radial water turbine by means of simulation steering.

The animation shows: water flow lines with velocity along lines ranging from blue (slow) to red (fast). The field in the ISO cutting plane represents the radial component of water velocity (i.e. the angular momentum being applied on the runner wheel)

In a short pause of the simulation, the inclination of the blades in the guide wheel will be modified. The simulation is being restarted leading to the following steps: generation of a new mesh, decomposition of the compute problem and assignment to a number of parallel processors. After a delay of ~ 10s, first results are output by the visualization pipeline.

The radial turbine is completely parameterized: e.g. concerning the blade profile and geometry, number of blades, number of input/output water channels, etc.

It is important to notice that we are looking at an online simulation here which allows the control of most of the important design parameters by simulation steering, thus leading to a new, inductive way of design - even in very complex environments.

7. Conclusion

The HLRS Teraflop Workbench Project has resulted in a robust, scalable high performance computing environment that allows for the seamless integration of new systems and software over time. Applications from the engineering sciences like Computational Fluid Dynamics, Combustion, Structural Mechanics, and Process Engineering have been found to provide a good, if not excellent match to the available architectures. New users and new application fields have been brought to the supercomputer already. Also, research in storage and communication systems has enabled a new I/O culture that will enable a geographically distributed version of the HLRS Teraflop Workbench in the near future.

Acknowledgement

We wish to thank the many developers within the Teraflop Workbench Collaboration who have created this unique platform and provided helpful comments on this paper. This work was in part performed by Alcatel SEL, Deutsches Zentrum für Luft- und Raumfahrt, European HPC Technology Centre, Nokia Siemens, Institute of Aerodynamics and Gasdynamics, Institute for Hydraulic Machinery, Institute of Process Engineering and Power Plant Technology, Institute of Structural Mechanics (all institutes at the University of Stuttgart), NEC, and finally the HLRS team. Special thanks are due to the HLRS Visualization Department for providing the virtual reality animations.

References

[1] IEEE Datacenter Ethernet Call for Interest (CFI), http://www.ieee802.org/3/cfi/0304_1

[2] IEEE 802.3ar Congestion Management Working Group, http://www.ieee802.org/3/ar

[3] "HPSS System Administration Guide: High Performance Storage System Release 4.1", IBM, November 1998.

[4] Watson, R., "High Performance Storage System Scalability: Architecture Implementation and Experience", Proceedings of the 22nd IEEE / 13th NASA Goddard Conference on Mass Storage Systems and Technologies, pp. 145-159, April 11-14, 2005, Monterey, http://www.hpss-collaboration.org/hpss/about/watsonr_highperf.pdf

[5] 6th HLRS/hww Workshop on Scalable Global Parallel File Systems, HLRS, Stuttgart, April 16-18, 2007.
URL: http://java.hlrs.de/corga-hwws-2007

[6] Resch, M., Küster, U., Müller, M., Lang, U., A Workbench for Teraflop Supercomputing, SNA’03, Paris, France, September 22-24, 2003.

[7] Stuttgart Teraflop Workbench Initiative.
URL: http://www.teraflop-workbench.de

[8] Wössner, U., Scientific Visualization and Virtual Reality, High-Performance Computing and Communication, HLRS, Stuttgart, July 2005.