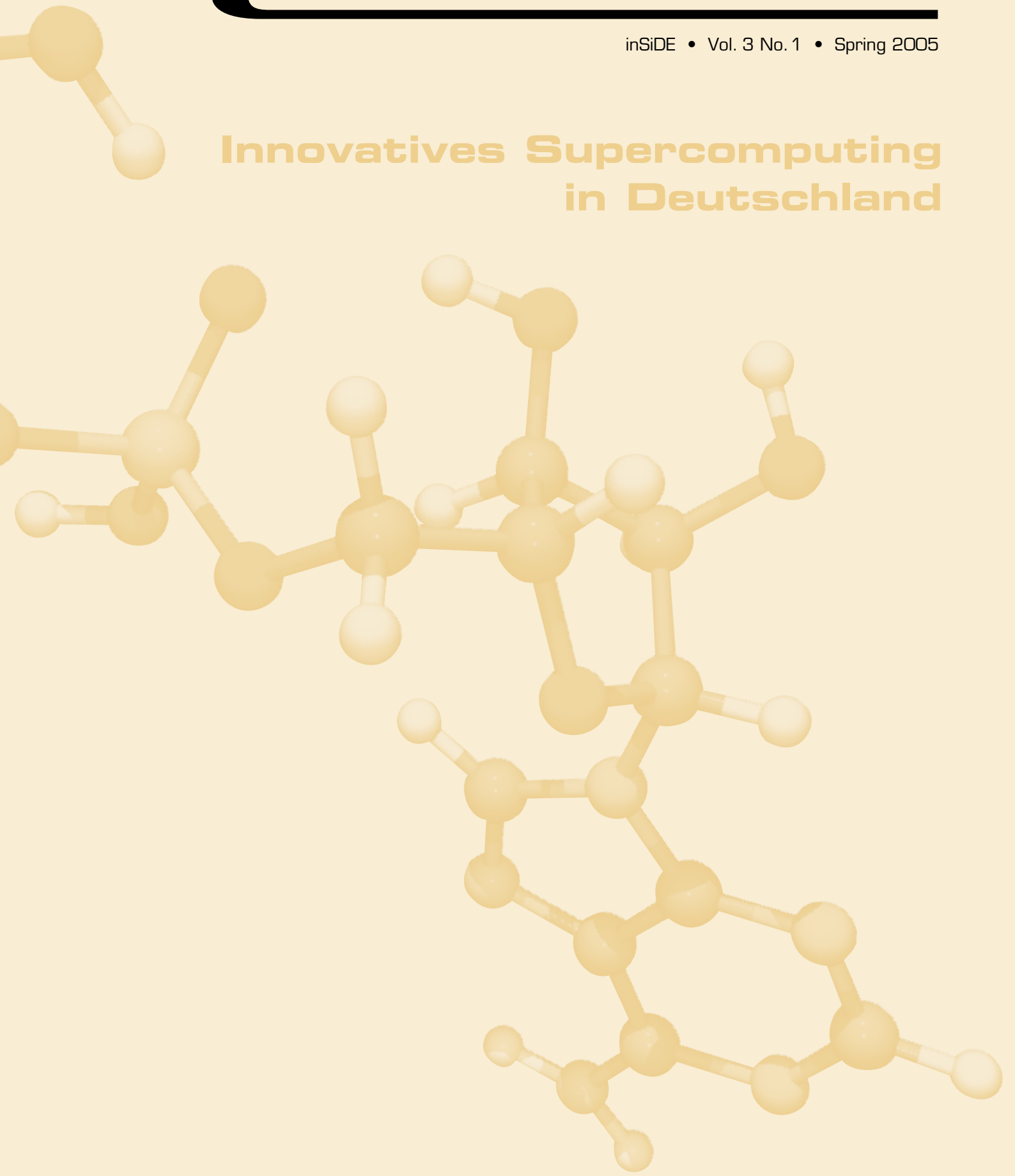
inSiDE

inSiDE • Vol. 3 No. 1 • Spring 2005

Innovatives Supercomputing in Deutschland



Editorial

The wheel has turned rapidly since the last issue of inSiDE was published and the development in German Supercomputing is going on at a fast pace. New systems are being installed, or are planned to become operational soon. After one year of successful operation of the Jülich 6.9 TF/s IBM supercomputer Jump the lead is now being taken by the recently installed NEC system at the HLRS in Stuttgart. The 72 node vector based SX-8 with its peak performance of 12.6 TF/s reaches a tremendous performance not only for the Linpack benchmark but also for real applications. First results and a description of the system can be found in this issue. Furthermore you will read a first-hand preview about the future HPC system at the LRZ in Munich named HLRB-II. With a peak performance of 70 TF/s it aims to be the fastest system in Europe and will be among the leading installations in the world in 2006 and 2007.

The first section of this issue of inSiDE covers application reports. An interesting application on the IBM supercomputer system at Jülich is presented by Paul Gibbon. He uses a parallel tree code to simulate the interaction of a laser beam with matter. Besides producing important physical results the code shows an excellent scaling behaviour on the IBM p690 cluster architecture as well as on the novel IBM Blue Gene/L system. In December 2004 the big and disastrous Sumatra earthquake, magnitude M9.3, has alerted the public worldwide and drawn attention to the science related to such phenomena. For seismologists, this event provided a unique dataset and therefore allows for better insight into the processes inside the Earth. Bernhard Schuberth and Heiner Igel from the LMU in Munich present their first findings from simulations done on the Hitachi SR 8000 at LRZ.

Novelties in high performance computing systems in Germany are presented in section two with LRZ presenting its future high-end system. It will be available to all German scientists in 2006 with a second phase to be deployed in 2007. The currently fastest system in Germany has recently become operational in Stuttgart at HLRS. A description of the system with first performance results is given. These systems help to put Germany back on the international landscape in supercomputing.

A special report on Open MPI is given in the third section of this issue. MPI has been around as a programming model for a while and MPICH has been established as a de-facto standard implementation. Users of clusters, however, experience some problems. Many software vendors do not recompile their application with every new release of MPICH or LAM/MPI, therefore the cluster administrators have to maintain several versions of each MPI library. Furthermore, most currently available implementations of MPI have to be compiled once for every network device available in the cluster. Since different compilers produce incompatible binary code, each of the available MPI libraries has to be compiled once for every available/supported compiler. Open MPI is a new implementation of the MPI-1 and MPI-2 specification, trying to solve many of the issues mentioned above.

Prof. Dr. H.-G. Hegering (LRZ)
Priv.-Doz. Dr. Dr. Th. Lippert (NIC)
Prof. Dr. M. M. Resch (HLRS)

Contents

Mesh-free Plasma Simulation with a Parallel Tree Code

Numerical simulation of hot, ionized matter poses a perennial challenge to the theoretical plasma physicist because of the virtually unlimited degrees of freedom, extreme nonlinear behaviour and vast range of length- and timescales characteristic of both natural and man-made plasmas. Traditionally, the intractability of first-principles simulation is overcome by first simplifying the problem in phase space; replacing individual particle trajectories by a smooth velocity distribution and then solving a Vlasov-Boltzmann-type equation. By rigorous application of kinetic theory, many problems can be further reduced to the magnetohydrodynamics picture – the plasma equivalent of the Navier-Stokes equations.

Whether kinetic or fluid, nearly all plasma modelling over the past four decades has relied on a spatial mesh to mediate the interplay between plasma particles and their associated electric and magnetic fields. While these models have proved highly successful, the presence of a grid ultimately places restrictions on the spatial resolution or geometry which can be considered – especially in three dimensions. Recently a new mesh-free plasma simulation paradigm has been developed which overcomes some of these limitations. Inspired by the N-body tree algorithms designed to speed up gravitational problems in astrophysics, this approach reverts to first principles by computing forces on individual particles directly, following their trajectories in a Lagrangian, "molecular dynamics" fashion [1].

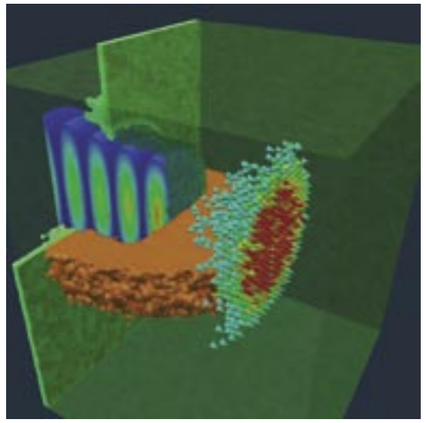


Figure 1: Mesh-free kinetic simulation of proton acceleration by a high intensity, short-pulse laser

At ZAM we have applied this technique to study particle transport in Petawatt laser interactions with solid targets – see Figure 1. In this case the laser (blue discs) is modelled as a simple momentum and heat source, drilling into the target and accelerating a substantial fraction of the plasma electrons to energies of several MeV in the process. In effect, the laser induces a multi-Megaamp current directed into the target, a feat which is only sustainable if a return current can be supplied by the cold background charge. If the target resistivity is high, an imbalance will result, setting up a DC electric field in the range of 10^{12} Vm^{-1}. This field hinders the hot electrons from passing through the target (orange cloud), but on the other hand leads to enhanced acceleration of ions from the front side of the target (arrows). Such laser-based energetic ion sources have many promising applications in areas such as isotope production, tumor therapy and advanced fusion schemes [2].

A typical investigation is set up with a total of 6 million electrons and ions placed in a "foil" with dimensions 12x12x5 µm3. The simulation of this interaction process for a 100 fs laser pulse would consume 5000 hours on a single Power4 CPU, but this reduces to around 100 wall-clock hours when run on 96 processors (3 frames) of the IBM supercomputer "Jump" at the Research Centre Jülich. Further preliminary benchmark tests with several million charges (1-25 million) demonstrate that this code scales up to at least 256 CPUs on Jump and 1024 CPUs on the new BlueGene/L architecture – Figure 2. Although slower than particle-in-cell codes (their mesh-based equivalents) parallel tree codes offer completely new possibilities in plasma simulation, particularly where collisions are important (here they are included automatically); or for modelling complex geometries; or for mass-limited systems in which artificial boundaries would severely compromise the simulation´s validity (for example atomic clusters). The generic nature of this algorithm, combined with good parallel scalability, means that it can be easily adapted to other systems dominated by long-range interactions – currently one of the research priorities at ZAM [3].

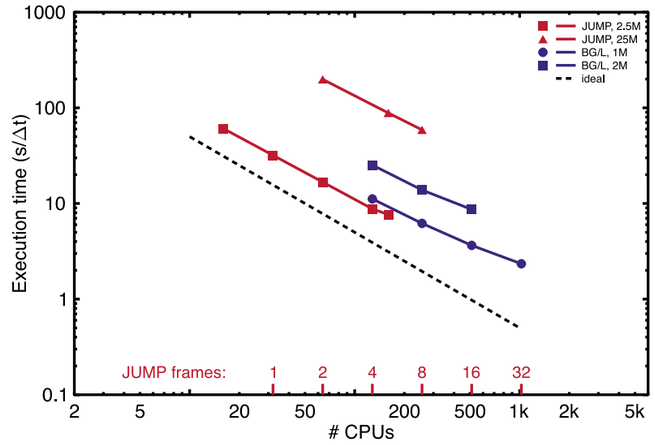


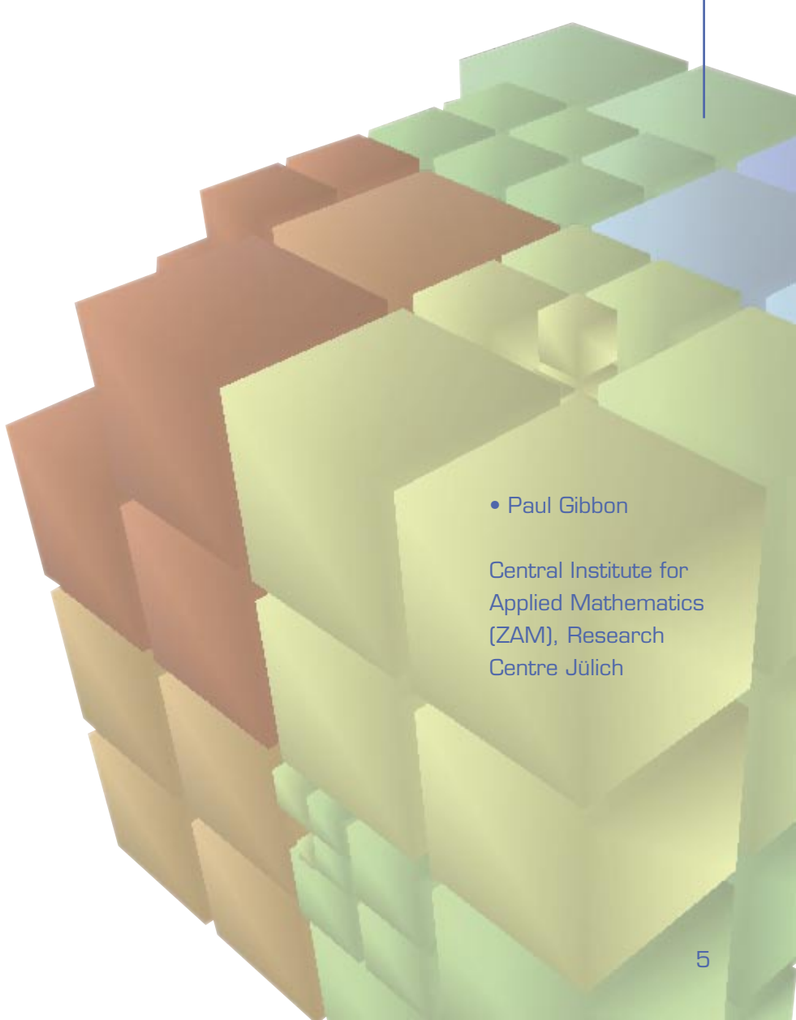
Figure 2: Scaling of the parallel tree code on Jump and BlueGene/L for spheres with various numbers of charges

References

[1] **S. Pfalzner, P. Gibbon**
Phys. Rev. E 57, 4698 (1998)

[2] **P. Gibbon**
Short Pulse Laser Interactions with Matter, Imperial College Press, London (2005)

[3] **For further information, see:**
www.fz-juelich.de/zam/cams



• Paul Gibbon

Central Institute for Applied Mathematics (ZAM), Research Centre Jülich

Challenges in Computational Seismology

In December 2004 the big and disastrous Sumatra earthquake, magnitude M9.3, has alerted the public worldwide and drawn attention to the science related to such phenomena. For seismologists, this event provides a unique high quality dataset and therefore better insight to the processes inside the Earth. One of seismology's major tasks is to resolve the structure of the Earth's interior. For decades it is known that the earth, by first approximation, can be described by a spherically symmetric model consisting of a crust, mantle, an outer and an inner core. Since then the challenge has been to extend this model, allowing for lateral variations, such as plumes (hot material arising from the core mantle boundary to the surface) or subduction zones (typically, oceanic crust forced beneath continental crust, when different tectonic plates collide, which subsequently sinks to the core).

In the last 20 years seismologists have developed a variety of techniques to solve this problem resulting in the first three dimensional images of the Earth. One approach is "seismic tomography", which is comparable to medical tomography. Seismic waves inside the earth are represented as rays penetrating the earth along various paths and recorded with seismometers at different positions all over the globe. In this way the interior of the Earth is illuminated from all directions and angles leading to the construction of a 3D image of seismic velocities inside the Earth.

An alternative approach that arose with the development of computer technology is the simulation of full 3D wave propagation using numerical techniques. Simulation of seismic wave propagation with recent supercomputers that allow for the solution of the complete wave fields through 3D structures, are currently revolutionizing seismology and related fields. Wave propagation on a planetary scale has so far predominantly been carried out using quasi-analytical approaches (e.g., spherical harmonics) and perturbation theory. Only recently the impact of 3D structures on the observed wave field is being addressed as computational power has reached a state where the simulated wave fields can be directly compared to observations.

With the tools developed in the seismology group of the LMU Munich (axi-symmetric approach [1] and spherical sections [2]) as well as a spectral element approach [3] that was developed at the California Institute of Technology, but extended and installed on the Munich supercomputer facilities of the Leibniz-Rechenzentrum Munich, a new era of global seismic data modeling is just beginning.

In the last years the spectral element method (SEM) has become one of the most important tools in computational seismology. Being a modified finite element method it provides a very flexible way of implementing geological structures, which are usually quite irregular and complex in shape. First introduced for fluid dynamics, it was further developed in the 1990's for seismological applications.

Today the SEM can be used to simulate the wave propagation in quite realistic global spherical Earth models including various features such as topography/bathymetry, laterally heterogenous velocity structures in the crust and the mantle, attenuation (anelasticity), anisotropy and also second order effects, as for example Earth's rotation, gravity or the influence of ocean water on the wave field. The advantages of the method are not only the capability of dealing with complex structures, as mentioned above, but also its high accuracy and the possibility of program implementation on parallel computers.

In this method, the model for global wave propagation is built using the "cubed sphere" approach. This is illustrated in Figure 1, where the initial cube is gradually distorted from left to right, until its six faces match the surface of the sphere. The clue in this procedure is to keep a small cube in the interior of the mesh undistorted thus avoiding singularities in the center at r=0. The applied velocity model S2ORTS is shown along the faces of the six "chunks" of the cubed sphere in Figure 2 together with a close up look of the spectral element mesh used in our simulations.

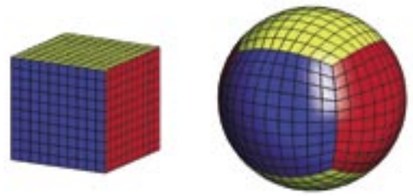


Figure 1: Creation of a global Earth model by expanding an initial cube to the sphere (i.e., cubed sphere mesh). Keeping a small central cube undistorted avoids singularities at r=0.
(Picture courtesy of Peter Danecek)

At the moment we use a model setup that resolves periods down to 20 seconds for standard applications. This setup uses 19 nodes (152 processors) of the HITACHI SR8000 supercomputer of the LRZ. The typical memory needed by our models, is between 60 to 90 GB, depending on whether attenuation is incorporated or not. The typical total runtime is 19 hours (25 including attenuation) for the calculation of a 90 minute seismogram. Figure 3 shows a simulated complex wavefield inside the Earth.

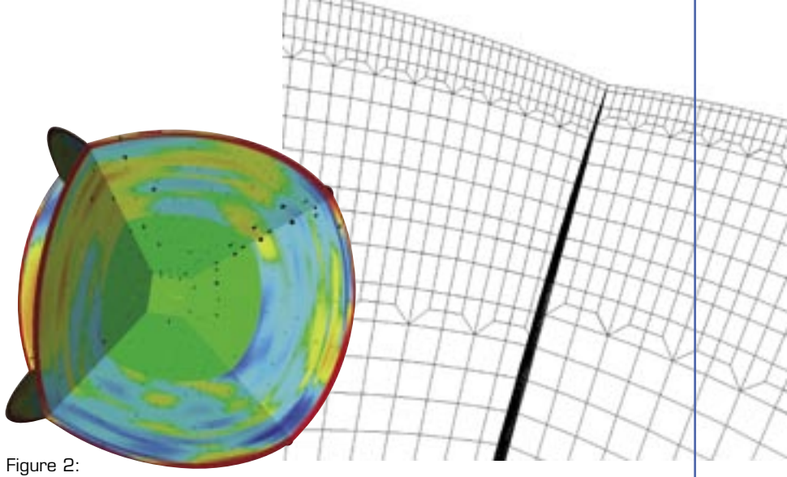


Figure 2:
(a) Cut planes along the borders of the six cubed-sphere chunks showing the three-dimensional S-wave velocity structure S2ORTS and
(b) part of the numerical mesh used in the spectral element code

The possibility of calculating wave propagation in all kind of media and structures, opens the way for many applications. Studying the amplification effects of sediment basins beneath large cities for earthquake hazard assessment, looking at the effects of several different theoretical models of earth structures on the wave field, or trying to understand the physics of faulting, just to mention a few. One big challenge in the next decade will be to create a link between seismology and geodynamics, as seismology can provide a snapshot of Earth's current state of convection. This can be used as boundary condition for modeling the dynamical behaviour of the mantle. In turn, one can use the results of geodynamical simulations (Figure 4)

and study the wave field going through those models in comparison to purely seismologically constrained models.

However, as stated above, the major goal of global seismology is to improve current Earth models and provide realistic images of our planet's interior. Today, numerical simulations are considered to be the right instrument for this full 3D wave form inversion.

Recently, an idea dating back to 1984 was rediscovered, which suggests combining simulations of wave propagation with its mathematical adjoint. One can think of this as the wave field, that is generated by sources at the receivers and traveling backward in time to the former source. Doing so, one can illuminate those parts of the numerical model that are incorrect compared to reality. Nevertheless, this procedure implies a vast amount of memory storage and even higher memory requirements during runtime. Many of those demanding computations will be necessary to obtain a reliable and high resolution image of the Earth. (For further information please visit www.geophysik.uni-muenchen.de)

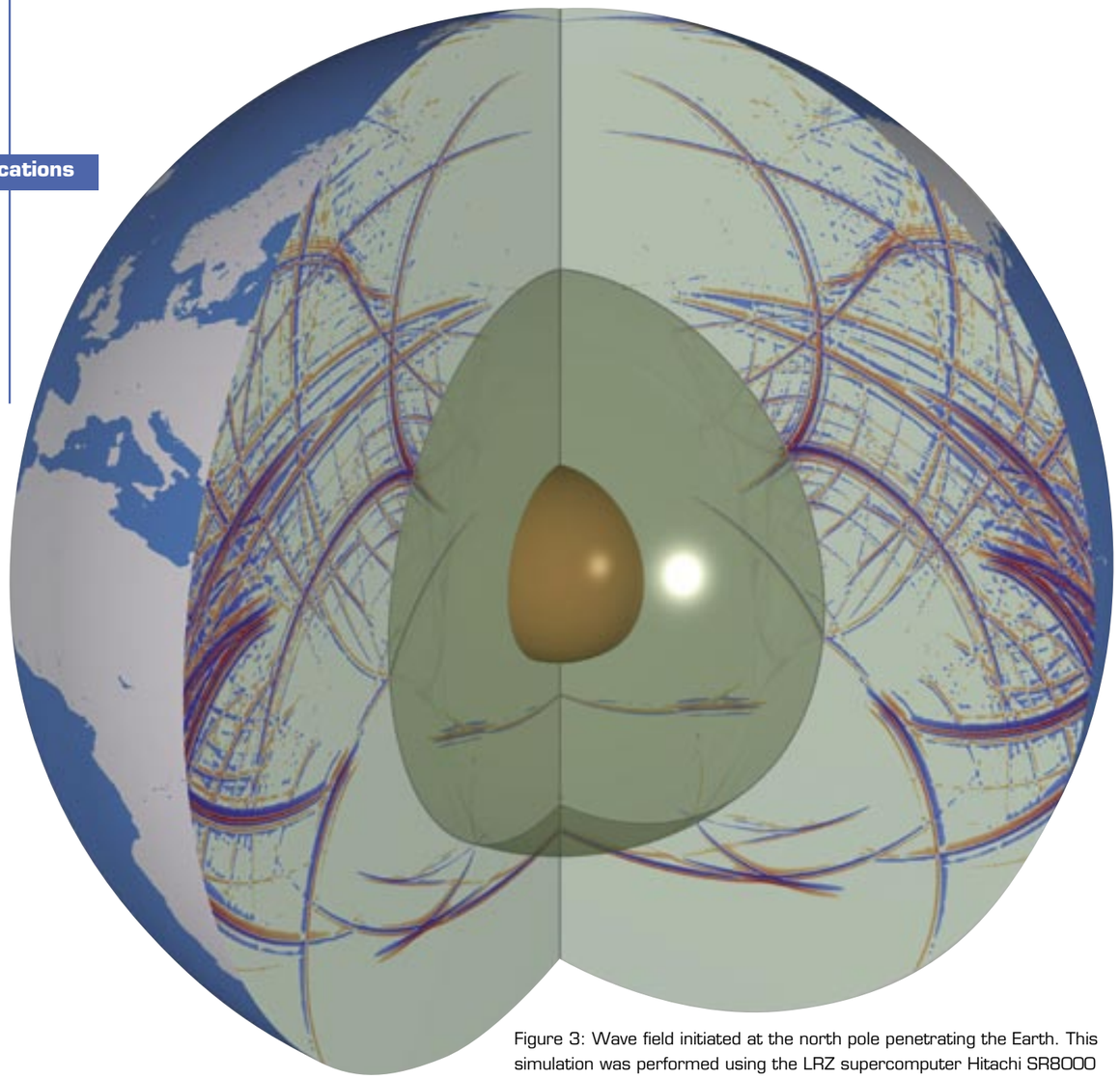


Figure 3: Wave field initiated at the north pole penetrating the Earth. This simulation was performed using the LRZ supercomputer Hitachi SR8000 with a program developed by the seismology group of the LMU Munich

Acknowledgements

We like to acknowledge funding through the KONWIHR Project, and the Leibniz-Rechenzentrum and its steering committees for providing access to the Hitachi SR8000. We also want to thank the LRZ supporting staff for their help. These projects were partly supported through the DAAD (IQN-Georisk) and the German Research Foundation.

References

[1] **H. Igel, T. Nissen-Meyer, G. Jahnke**
Wave propagation in 3D spherical sections. Effects of subduction zones. Phys. Earth Planet. Int., 132:219-234, 2002

[2] **G. Jahnke, H. Igel**
High resolution global wave propagation through the whole Earth: the axi-symmetric PSV and SH case. In EGS General Assembly, Nice, France, 2003

[3] **D. Komatitsch, J. Tromp**
Spectral-element simulations of global seismic wave propagation – I. Validation. Geophys. J. Int., 149:390-412, 2002

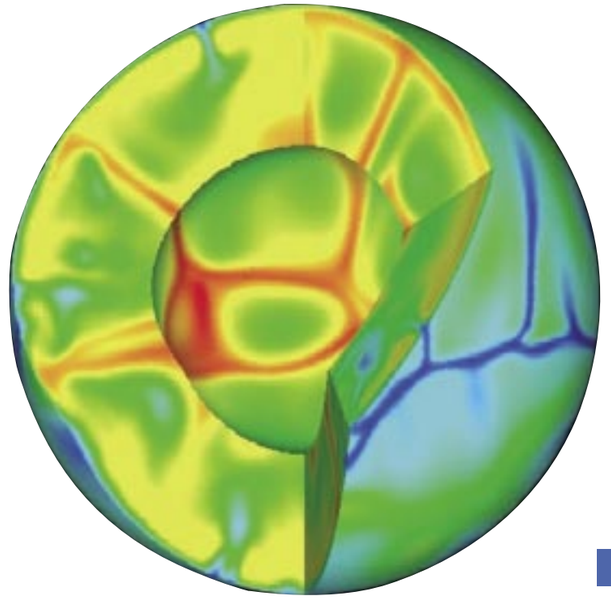


Figure 4: Temperature field inside the Earth resulting from a geodynamical mantle convection model. Numerical Simulations of wave propagation through such models is thought to better provide a link between seismology and geodynamics

• Bernhard Schuberth
• Heiner Igel

Department for Earth and Environmental Sciences – Geophysics Section
Ludwig-Maximilians-University, Munich

New High End System at Leibniz Computing Center

The Leibniz Computing Center in Munich has selected the latest generation of SGI Altix systems from Silicon Graphics to power its next generation national high end system. The new system will incorporate 3,328 dual-core Intel Itanium2 processors in its final configuration and will be capable of generating 69 trillion calculations per second, effectively boosting the computional capacity at Leibniz Computing Center (LRZ) by a factor of 30. LRZ also will deploy a 660-terabyte SGI InfiniteStorage solution to accommodate its rapidly growing stockpile of scientific data.

Named "Höchstleistungsrechner in Bayern" (HLRB-II), the new system will supersede the current Hitachi SR8000 system, which runs more than 200 projects from scientists and researchers from all over Germany requiring top performance.

The German government and the state of Bavaria share funding for the HLRB-II project, which was awarded to SGI after an international competition.

Figure 1: Under construction: The new Computer Cube (left), the institute building (middle), and lecture rooms (right) of LRZ. (Picture E. Graf, Institut für Informatik, TU München)

Installation of the system will be carried out in two stages. The first installation phase – which will become available at the beginning of 2006 – will be capable of performing 33 trillion floating point operations per second (33 TFlop/s), compared to the presently available 2 TFlop/s on the Hitachi SR8000. After the upgrade to the final configuration in 2007, this value will actually increase to 69 TFlop/s. The nodes of the system will contain up to 1024 core (512 sockets) and will be connected with NUMALink 4 technology.

The system will be installed in the upper floor of LRZ Compute Cube, a newly constructed 36x36x36 meter building which will also contain other servers and the backup and archiving facilities.

Additionally, SGI has already supplied LRZ with a 128-processor SGI Altix system equipped with 512 GB of memory and connected to an 11-terabyte SGI InfiniteStorage solution. This system recently replaced a Fujitsu-Siemens vector system when it was installed in early 2005. This 128-processor Altix and storage combination will be primarily dedicated for use in computational area of chemistry, and fluid dynamics. For preparing the migration to the new high end system, an additional dedicated 64 processor Altix system will be installed in summer 2005.

Also be installed in summer 2005 is a 70 node dual-core Itanium2 enhancement to LRZ's Linux compute environment. The nodes of this part of the cluster are

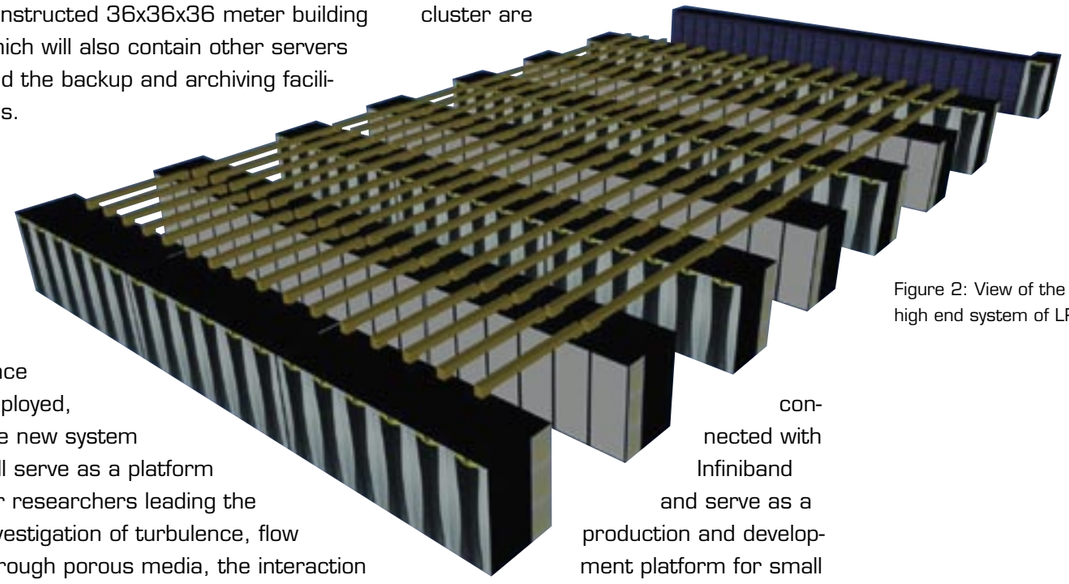


Figure 2: View of the new high end system of LRZ

connected with Infiniband and serve as a production and development platform for small and medium-sized parallel applications which are not suitable for running on the big systems.

Once deployed, the new system will serve as a platform for researchers leading the investigation of turbulence, flow through porous media, the interaction of flows and deformable structures, the generation and propagation of acoustic waves, high temperature superconductivity, the investigation of materials exhibiting memory effects with regard to their form, the analysis of chemical reactions involved in combustion and catalysis processes, and the propagation of seismic waves and earthquakes.

• Matthias Brehm

Leibniz
Rechenzentrum LRZ

The NEC SX-8: A European Flagship for Supercomputing

When in spring 2004 HLRS announced the final decision for its procurement for a supercomputer [1] the time for installation was still about a year away. The system purchased was projected to be the fastest supercomputer in Europe. At the time of announcement this was true both in terms of peak performance and sustained performance. During the last year since the publication of the article a number of things happened. The Spanish government announced the purchase of a cluster based on standard components. The French government agency CEA made a decision for another cluster with a similar architecture. Both of these systems show a larger peak performance. It remains to be seen what the actual level of performance is that can be achieved for the users of these three systems.

In this article we present the installation set up at Stuttgart including the systems that surround the SX-8 in order to create a workbench for supercomputing. First performance figures indicate that the assumptions and promises made during the procurement phase can be fulfilled. The level of performance is very good in general and for many examples even exceeds our expectations.

Introduction

In this article we want to give a description of the new system and present some of the first results that were achieved. At the time of publication of the most recent issue of inSiDE the reader will be able to check at least for further Linpack results at the top 500 webpage [2]. It may, however, be more interesting to look at Jack Dongarra's new High Performance Computing Challenge Benchmark [3]. The project that Dongarra started aims to complement the traditional linpack benchmark. The 23 individual tests in the HPC Challenge benchmark do not measure the theoretical peak performance of a computer. Rather, they provide information on the performance of the computer in real applications. The tests do not measure processor performance but criteria that are decisive for the user such as the rate of transfer of data from the processor to the memory, the speed of communication between two processors in a supercomputer, the response times and data capacity of a network. Since the tests measure various aspects of a system, the results are not stated in the form of one single figure. In their entirety, the measurements enable an assessment of how effectively the system performs high performance computing applications. When the benchmark was run for the predecessor model of the NEC SX-8 – the SX-6 – it took the lead in 13 out of 23 categories showing the high potential that vector supercomputers still have when real performance is at stake.

A Supercomputing Workbench

When the Höchstleistungsrechenzentrum Stuttgart (HLRS) started its request for proposals 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 [4]. 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. We distinguish between systems for pre-processing, simulation and post-processing (see figure 1).

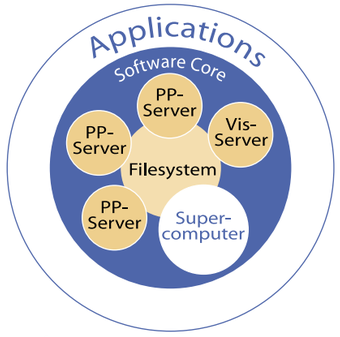


Figure 1: The Stuttgart Teraflop Workbench Concept

As described in a previous contribution here [1] the decision was made to go with NEC as the key partner to build the workbench. The schematic configuration is briefly shown in figure 2.

The concept is centred around the global file system of NEC. SX-8, Asama (IA64 shared memory system) and a cluster of Intel Nocona processors all have direct access to the same data via fibre channel. In the following we briefly describe these main three systems with a focus on the SX-8.

The SX-8 System

The SX-8 series was announced by NEC in autumn 2004 as the next step of the SX series that had been extremely successful in Japanese and European supercomputing over the last 10 years.

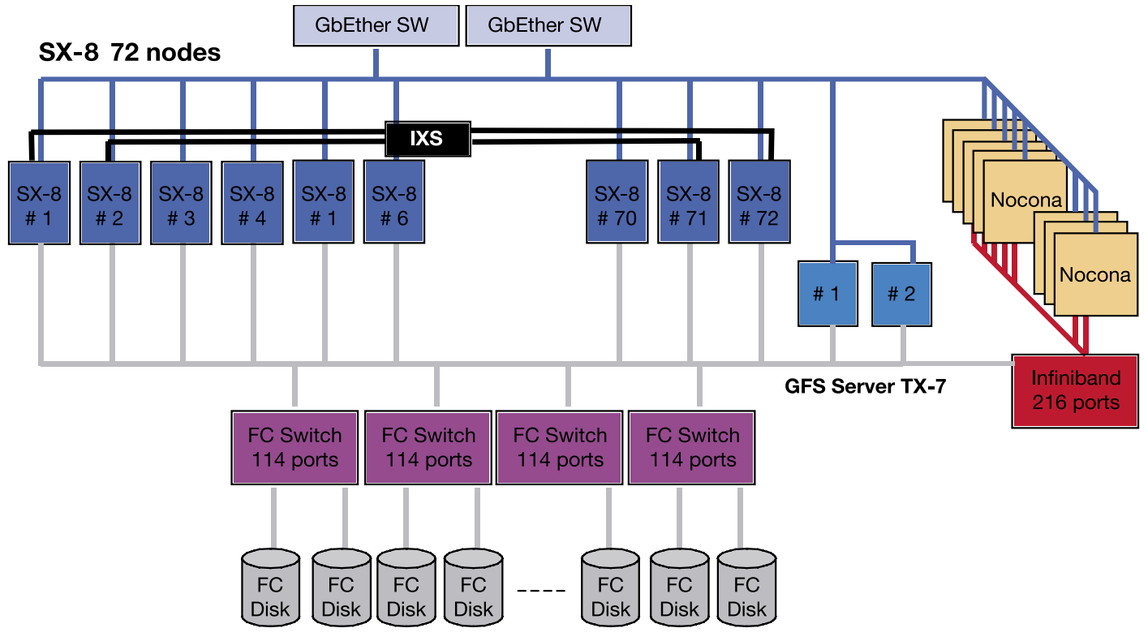


Figure 2: Technical description of the workbench concept as implemented by NEC

The SX-4 had repeatedly replaced Cray vector systems in the mid 90s. The SX-5 was seen to be the leading system in vector supercomputing in the late 90s. The SX-6 was the basis for the Earth Simulator – a system that dominated the top 500 list for nearly three years.

The SX-8 continues this successful line by improving the concepts and making use of most recent production technology. The key improvements are:

- Advanced LSI, 90nm CU, 8000 pins: This leads to high density packaging and to lower operational and investment costs for the user. The SX-8 consequently consumes 10 times less power than the SX-5
- Optical Interconnect cabling: This leads to easy installation and maintenance and reduces the costs for the user and reduces the number of parts by a factor of six compared to the SX-6
- Low Loss PCB technology, serial signalling to memory: This leads to high packing density, easy manufacturing which again reduces the costs for the user reducing the required space by a factor of four compared to the SX-6 model.

CPU

The CPU is manufactured in 90nm technology. It uses copper wiring and the vector pipes operate at 2GHz. The CPU LSI is connected via high density I/O pads to the node circuit board. Advanced technology is used for the signal transmission.

The vector unit is the most important part of the SX-8 CPU. It consists of a set of vector registers and arithmetic execution pipes for addition, multiplication, division and square root. The hardware square root vector pipe is the latest addition to the SX-8 CPU architecture and is only available on the SX-8. Each of the arithmetic pipes is four way parallel, i.e. can produce four results per cycle. A high speed load/store pipe connects the vector unit to the main memory.

The traditional vector performance of a single CPU is 16 GFLOP/s. Together with the new square root vector pipe and the scalar unit the total peak performance adds up to 22 GFLOP/s.

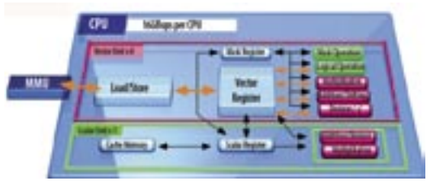


Figure 3: Schematic view of an SX-8 CPU

Node & Interconnect

Each node of the SX-8 is an 8-way shared memory system. The outstanding feature of the node is the extremely high memory bandwidth of 64 GB/s for each processor or a total of 512 GB/s for the overall node. Given the vector performance of 16 GF/s this results in a peak memory transfer rate of 4 Byte for every flop that the processor performs. With these numbers the SX-8 outperforms all its competitors in the field by a factor of about 3 with only the Cray vector systems being able to compete.

The IXS interconnect is a 128x128 crossbar switch. Each individual link has a peak bidirectional bandwidth of 16 GB/s. Again this outperforms competing networks – especially in the cluster field. However, it has to be mentioned that 8 processors share the bandwidth. Still the system has an acceptable communication balance – which is also reflected by first linpack results shown below.

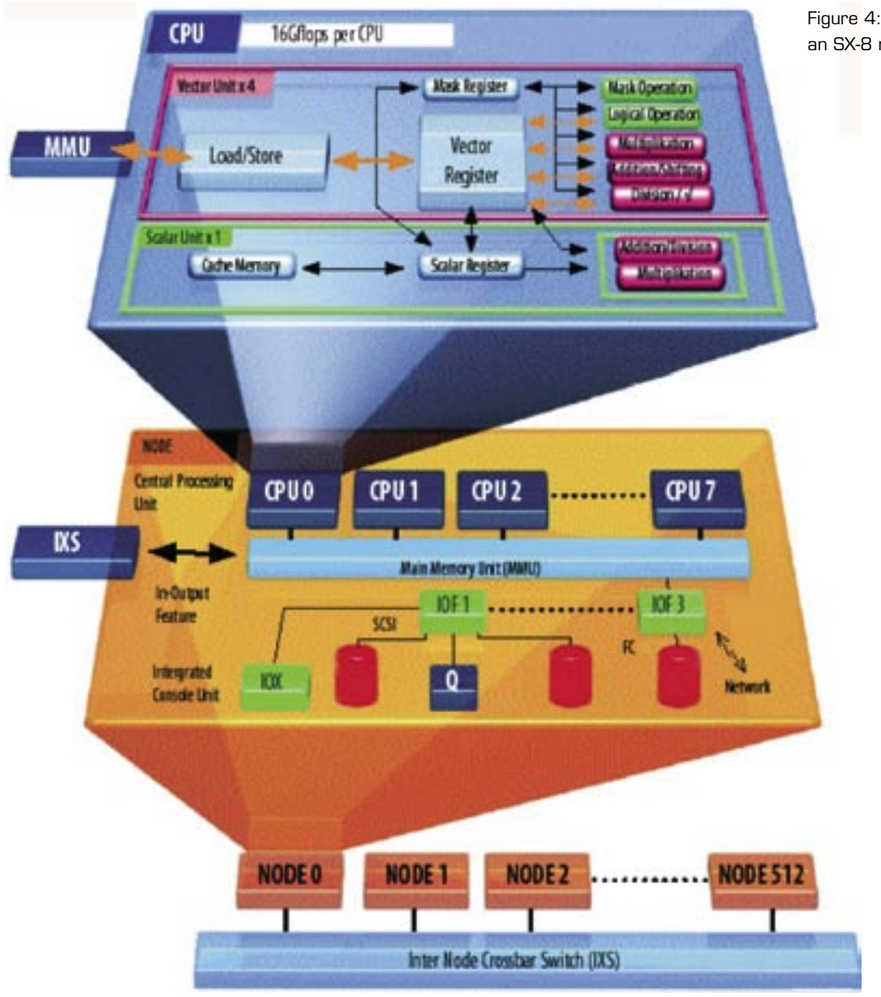


Figure 4: Schematic view of an SX-8 node

Software

The SX-8 operating system "SUPER-UX" is based on UNIX System V industry standard:

- The system provides powerful and flexible functions, which allow effective and reliable operation and administration of large scale multi node supercomputer systems
- Super UX provides all standard Unix communications protocols and technologies. An SX-8 supercomputer can be seamlessly integrated into an existing network environment
- Super UX provides the necessary functionality to manage and schedule the supercomputer resources flexibly and effectively
- The SX-8 OS supports high performance I/O, a memory file system, support for archiving systems and the functionality to manage tape libraries
- The NQSII batch processing system in conjunction with the enhanced resource scheduler ERSII provide reliable job scheduling functionality in even the most demanding environments
- The GFS global file system is the glue that holds the compute centre together, it allows a variety of client

systems to share files with a SX-8 multi node system in a transparent way via a storage area network (SAN)
- Super UX offers full checkpoint and restart functionality for batch jobs for flexible operation and maintenance. The Master-Scope software supports the management of large multi node systems.

Preprocessing/ Postprocessing

Two hardware systems complement the SX-8 installation. One is an IA64 based shared memory system that serves for pre-processing and file serving. Code named AsAmA it consists of 2 TX-7 nodes with 32 processors each. One of the nodes is equipped with 512 GB of memory in order to allow preparation of large parallel jobs. We found that still most users 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 technology is added to the workbench. It serves both for post-processing/visualization and as a compute server for multi-disciplinary applications. The later 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 GB/s.

Summary of HLRS Installation

	SX-8	AsAmA	Nocona
CPU			
Type	Vector	IA64	IA32
Clock Rate [GHz]	2	1,5	3,2
Performance [GF/s]	22	6	6,4
Node			
#Processors	8	32	2
Memory [GB]	128	256/512	1 or 2
Performance [GF/s]	176	192	12,8
System			
#Nodes	72	2	205
#Processors	576	64	410
Peak Performance [TF/s]	12,67	0,384	2,62
Memory [TB]	9,2	0,768	0,24
Workbench			
Peak Performance [TF/s]	15,674		
Memory [TB]	10,208		

Table 1: Basic installation and performance parameters

Installation Schedule

The system is set up in the newly built computer room of the Höchstleistungsrechenzentrum Stuttgart (HLRS). While construction work for the office part of the building was still under way the computer room was finished and the set up of the system began in December 2004. The first 36 nodes were rolled in at the end of January and were set up within three weeks. NEC did intensive testing before handing over the first 36 nodes to the Höchstleistungsrechenzentrum Stuttgart (HLRS) in March. The Intel EM64T cluster was set up within a few days in January together with the first AsAmA front end node.

The second 36 nodes were brought in late in March to be set up during April. The final installation of the overall system was scheduled after the deadline of this contribution. Acceptance will take about one month such that the full system could be operational by June/ July 2005.

Installation and First Results

Linpack

One of the first applications to run on any new system is the linpack benchmark that defines the ranking of any system in the top 500 list. As soon as the system was installed linpack results were measured for 72 nodes. The theoretical peak performance of the 72 nodes with their 576 processors is 9,2 TFLOP/s. Average microprocessor systems and clusters usually exhibit a level of performance of 50% -70% for the linpack benchmark. The Earth Simulator had shown 87,5% of its peak performance in the linpack benchmark. Our expectation was hence to achieve about 95% of peak performance in a linpack run on 72 nodes. The results were much better as can be seen in figure [4]. With 72 nodes (576 processors) the SX-8 achieves a sustained linpack performance of about 8,92 TF/s which is about 97% of the peak performance.

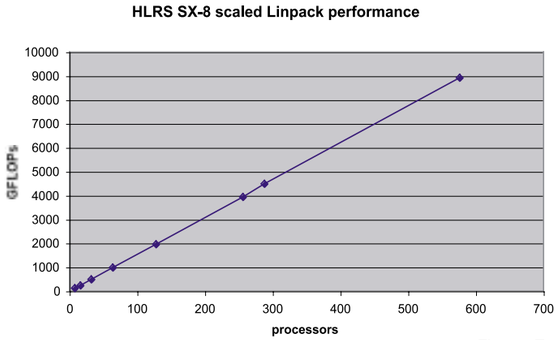


Figure 5: Linpack performance on 72 nodes (576 processors) of the SX-8

Computational Fluid Dynamics

First benchmarks were done on the system shortly after the 36 nodes were set up. The benchmark run was a Lattice-Boltzmann code called BEST and the measurements were done by Peter Lammers from the Höchstleistungsrechenzentrum Stuttgart (HLRS). First results are shown in the figure below.

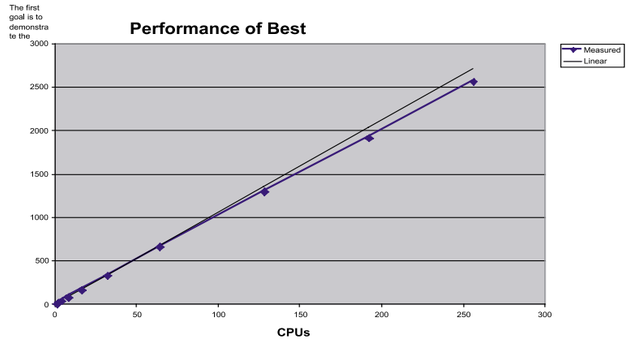


Figure 6: Performance measurements for the Lattice-Boltzmann code BEST

The figure shows not only excellent speedup – owed to the high performance IXS switch – but also demonstrates that up to 256 processors we

see an extremely high efficiency for the single processor. We achieve 13 GF/s for one processor and about 11 GF/s per processor for the 256 processor case.

The key finding here is that the system already now exceeds our expectations by far. Our hope – and part of the contract – was that for a single application we could achieve about 4 TF/s of sustained performance. The figures show that today with 288 processors we already achieve about 2,9 TF/s. The sustained performance for the full 576 processor system is expected to be at about 5 TF/s.

Considering that this is not an application kernel but a real CFD simulation case the result confirms our decision for a vector based system.

Who are the Users?

As all systems that are run by the Höchstleistungsrechenzentrum Stuttgart (HLRS) the new SX-8 vector system and the attached cluster of Intel Nocona processors is made available to a large group of users both from industry and research. Every scientist in Germany doing public research in an organization that is not directly funded by the federal government can apply for compute time on the system. A scientific steering committee continuously evaluates proposal and grants access to the system solely based on scientific merit of the proposal and on the proven necessity to use a supercomputer system. Access for European users is possible through special research projects.

The Teraflop-Workbench Initiative

In order to further support users and projects and in order to extend the reach of vector systems in terms of application fields Höchstleistungsrechenzentrum Stuttgart (HLRS) and NEC set up the Teraflop Workbench Initiative [5].

The first goal is to demonstrate the efficiency of NEC SX vector systems and that these systems can deliver Teraflop/s application performance for a broad range of research and ISV codes. Secondly, NEC Linux clusters and SMP systems will form together with the SX vector system an environment that allows to perform the complete pre-processing – simulation – post-processing – visualization workflow in an integrated and efficient way.

To show the application performance NEC and HLRS work together in selected projects with scientific and industrial developers and end users. Their codes come from areas like computational fluid dynamics, bioinformatics, structural mechanics, chemistry, physics, combustion, medical applications and nanotechnology. The Teraflop Workbench is open to new participants. An application has to demonstrate scientific merit as well as suitability and demand for Teraflop performance in order to qualify.

Industrial Usage

The whole workbench is made available for industrial usage through the public private partnership hww (High Performance Computing for Science and Industry) – a company that was set up together with T-Systems and Porsche in 1995. The systems are integrated into the security concepts of hww and are operated jointly by Höchstleistungsrechenzentrum Stuttgart (HLRS) and T-Systems.

Accounting and billing is done on a full cost model which includes investment, maintenance, software, electricity, cooling, operation staff, insurance, and overhead costs for running the public private partnership. Through T-Systems the systems are open to German Aerospace Research Centre DLR, to DaimlerChrysler and to small and medium sized enterprises.

Summary

With the new SX-8 system installed at the Höchstleistungsrechenzentrum Stuttgart (HLRS) German users have access to the best performing system in Europe which can compete with the leading installations in the world. Applications that have been optimized for leading edge processors will greatly benefit from the architecture. This will include codes that are cache optimized as first tests show. New applications will be brought to supercomputing through the Teraflop Workbench initiative which will guarantee that the full portfolio of scientific fields will turn the new system into a real scientific and industrial teraflop workbench.

The integration of various architectures into a single system turns the overall system into an excellent tool to be used in the scientific and industrial work flow and production processes. This increases the chance to open supercomputing for new scientific communities and industrial application fields.

Literature

[1] **Michael M. Resch**
"The next generation supercomputer at HLRS", inSiDE, Vol. 2, No. 1, 2004

[2] **TOP500**
www.top500.org

[3] **HPCC Benchmark**
icl.cs.utk.edu/hpcc/index.html

[4] **Michael M. Resch, Uwe Küster, Matthias Müller, Ulrich Lang**
A Workbench for Teraflop Supercomputing, SNA'03, Paris, France, September 22-24, 2003

[5] **Stuttgart Teraflop Workbench Initiative**
www.teraflop-workbench.de

• Michael M. Resch
• Matthias Müller
• Peter Lammers

Höchstleistungs-
rechenzentrum
Stuttgart (HLRS)

• Holger Berger
• Jörg Stadler

NEC

Systems

Open MPI:
A Next Generation Implementation of the Message Passing Interface

Introduction

Users of clusters are familiar with this scenario: on an average PC-cluster, they often have the choice between dozens of different MPI libraries. The decisions which they have to make include:

- Do they want to use one of the available versions of LAM/MPI [2], MPICH [3], or the library provided by their NIC vendor (e.g. Myrinet [4], Infiniband [5])?
- Shall the library use the Gigabit Ethernet device or the faster devices such as offered by Myrinet or Infiniband?
- Is the application compiled with gcc, PGI, Pathscale or the Intel compiler? The MPI library has probably to be compiled with the same compiler.

There are many reasons for this diversity: first of all, basically all vendors of HPC interconnects have their own spin-offs of a public domain library like MPICH, unfortunately being incompatible with the original version. Many ISVs do not recompile their application with every new release of MPICH or LAM/MPI, therefore the cluster administrators have to maintain several versions of each MPI library. Furthermore, most currently available implementations of MPI have to be compiled once for every network device available in the cluster. Since different compilers produce incompatible binary code, each of the available MPI libraries has to be compiled once for every available/supported compiler.

Open MPI [1] is a new implementation of the MPI-1 and MPI-2 specification, trying to solve many of the issues mentioned above. How does another MPI implementation help in this scenario? Not at all, at a first glance. However, some of the design decisions made in Open MPI help circumvent many of the problems mentioned above.

Concept and Architecture

The basic framework on which Open MPI is based is called Modular Component Architecture (MCA), and is also shown in figure 1.

The main idea behind MCA is to abstract functionality into components. The motivation behind this abstraction is usually, that one could imagine different implementations of these functions for optimal performance in various environments and scenarios. The most obvious example in this context are point-to-point operations, where depending on what network devices are available on a cluster, different implementations for the same component will exploit best the capabilities of the underlying network device. The library should therefore use a different component for a Myrinet network and for an Infiniband network.

The architecture of Open MPI is comprised of three main functional areas:

- The backbone of the component architecture, which provides management services for all other layers (e.g. finding and building components, passing run-time parameters down to components)
- Component frameworks: each major functional area in Open MPI has a corresponding framework. A detailed list of component frameworks is presented later in this section
- Modules: instances of components.

Components have to follow a certain naming scheme, which a) indicate which framework they are implementing b) the version for the specified component and c) have a unique name within the current installation. The component has to implement furthermore the API defined for this framework. In the most common case, each of the modules will be a separate dynamic shared object (.so or .DLL). On start-up of the parallel application, the MPI library checks certain directories for all available components and opens them. As an example, if the library detects during initialization the according .so files for TCP, shared memory, Myrinet and Infiniband devices, it will load each of them, without deciding at that point which one will be used later on to communicate to which process.

Before establishing a network connection to another process, the MPI library will query each of the point-to-point components, whether it should be used for the communication to this process. The components respond with a priority level. As an example, on a PC-cluster with dual processor nodes, the shared memory device will return a high prior-

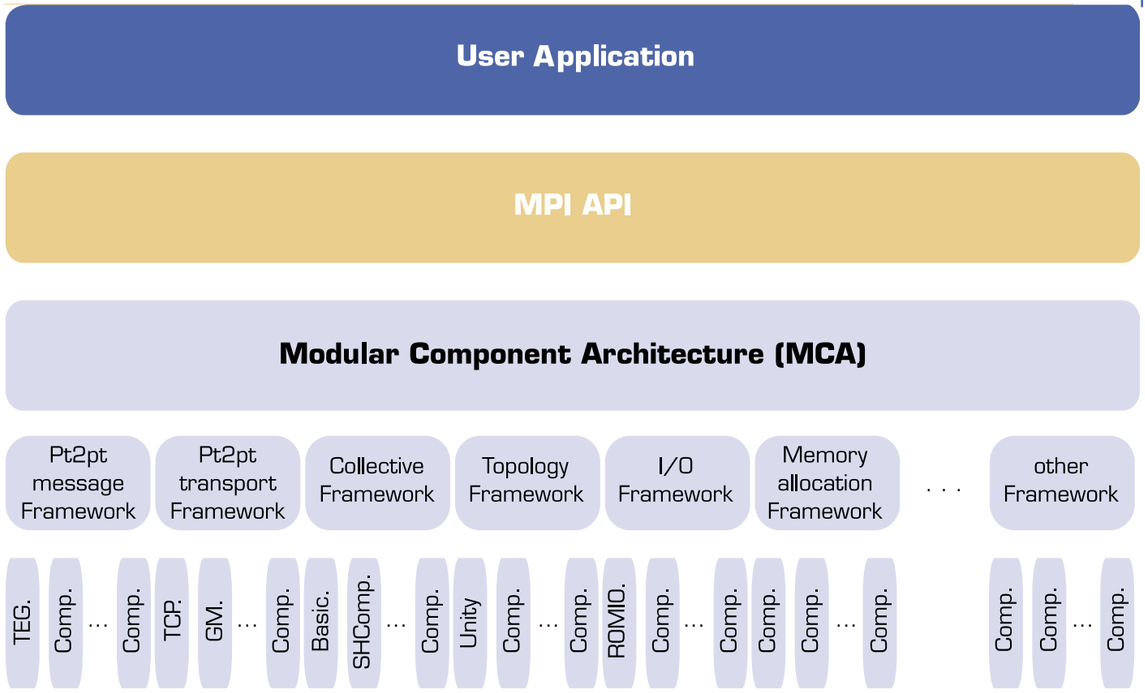


Figure 1: Modular component architecture

ity if the process to which the current process would like to establish a connection is on the same node, the Myrinet device will return a somewhat lower priority, the TCP device will indicate with an even lower priority that it could be used as well but it should not be the first choice. The Infiniband device will return a flag indicating that it should not be considered for this communications, since it did not detect any Infiniband NICs on the node. If the destination process is on another node, the priority of the Myrinet device will exceed the priority of the TCP device, while the shared memory module together with the Infiniband module will return again return a flag indicating that they should not be considered in this case.

In this concept, the library decides on a per-process basis which device shall be used for communication to this process. By being able to use different devices to different processes, Open MPI inherently avoids the situation that system administrators have to install different versions of the same library for different networks. This concept prevents furthermore, that vendors of network devices create spin-offs of the library. Their task has been reduced to the development of a point-to-point module and to building a dynamic shared object. This shared object can be installed into an existing Open MPI installation directory. Note that there is no need for vendors to publish the source code of their components; plugins can be distributed in either binary or source code packages.

Once a component has been chosen for the communication to a certain process, the component initialization routine will be called and certain function pointers set for later usage. Upon the end of the application, the module will be closed.

Open MPI defines at the current stage the following components for the MPI layer:

- Point-to-point messaging layer
- Point-to-point transport layer
- Collective operations
- Topology functions
- Parallel I/O
- Memory allocation.

Further components are defined for the runtime environment (Open RTE).

Status of Open MPI

An alpha release of Open MPI was distributed to selected users end of March 2005, the public beta release is expected to be available on the website at http://www.open-mpi.org in Q2/2005. This beta release will include

- Support for all of MPI-2, except one-sided communication
- Support for multiple user-level threads
- Support for TCP, shared memory, Myrinet (gm and mx devices), Elan 4 and Infiniband
- Support for various schedulers respectively process managers.

While currently being developed by Los Alamos National Laboratories, Indiana University, the University of Tennessee and the High Performance Computing Center of Stuttgart, the goal of Open MPI is also to provide a framework which eases research in the area of Message Passing libraries. Thus, contributions of other institutions will be highly welcome. A quality assurance procedure and a proper licensing scheme will ensure, that the official Open MPI distribution will however always maintain a production-level quality.

References

[1] Edgar Gabriel, Graham E. Fagg, George Bosilca, Thara Angskun, Jack J. Dongarra, Jeffrey M. Squyres, Vishal Sahay, Prabhanjan Kambadur, Brian Barett, Andrew Lumsdain, Ralph H. Castain, David J. Daniel, Richar L. Graham, Timothy S. Woodall, "Open MPI: Goals, Concept, and Design of a Next Generation MPI Implemention", in Dieter Kranzlmueller, Peter Kacsuk, Jack J. Dongarra (Eds.), "Recent Advances in Parallel Virtual Machine and Message Passing Interface", Lecture Notes in Computer Science vol. 3241, pp. 97-104, Springer 2004

[2] Jeffrey M. Squyres and Andrew Lumsdain, "A Component Architecture for LAM/MPI", in in Jack J. Dongarra, Domenico Laforenza, Salvatore Orlande (Eds.), "Recent Advances in Parallel Virtual Machine and Message Passing Interface", Lecture Notes in Computer Science Vol. 2840, Springer 2003

[3] W. Gropp, E. Lusk, N. Doss, and A. Skjellum, "A high-performance, portable implementation of the MPI message passing interface standard", in Parallel Computing 22(6), pp. 789-828, September 1996

[4] Myricom homepage at: http://www.myricom.com

[5] For references to Infiniband see the Open IB homepage at http://www.openib.org

[6] Rainer Keller, Edgar Gabriel, Bettina Krammer, Matthias Müller, Michael M. Resch "Towards efficient execution of MPI applications on the Grid: porting and optimization issues", Journal of Grid Computing, Volume 1, Issue 2, pp 133-149, 2003

[7] Graham E. Fagg, Edgar Gabriel, Zizhong Chen, Thara Angskun, George Bosilca, Antonin Bukovsky, Jack J. Dongarra, "Fault Tolerant Communication Library and Applications for High Performance Computing", LACSI Symposium 2003, Santa Fe, October 27-29, 2003

[8] Richard L. Graham, Sung-Eun Choi, David. J. Daniel, Nehal N. Desai, Ronald G. Minnich, Craig E. Rasmussen, L. Dean Risinger, Mitchel W. Sukalski, "A Network-Failure-Tolerant Message-Passing System for Terascale Clusters", International Journal of Parallel Programming 31 (4): 285-303, August 2003.

• Edgar Gabriel

Höchstleistungsrechenzentrum Stuttgart (HLRS), Germany

• Richard L. Graham

Los Alamos National Laboratories, USA

• Jeffrey M. Squyres

Open Systems Laboratory
Indiana University, USA

• Graham E. Fagg

Innovative Computing Laboratory
University of Tennessee, USA

Leibniz Computing Center of the Bavarian Academy of Sciences (Leibniz-Rechenzentrum der Bayerischen Akademie der Wissenschaften, LRZ) in Munich provides national, regional and local HPC services. Each platform described below is documented on the LRZ WWW server; please choose the appropriate link from www.lrz.de/services/compute

Contact
High-Performance Systems Department

Dr. Horst-Dieter Steinhöfer
Barer Straße 21
80333 München
Germany
Phone +49 89 28 92 87 79
steinhoefer@lrz.de
www.lrz-muenchen.de

View of the Hitachi SR8000-F1 at LRZ

Compute servers currently operated by LRZ are given in the following table

System	Size	Peak Performance (GFlop/s)	Purpose	User Community
Hitachi SR8000-F1 8-way	168 nodes 1344 processors (+168 Service procs) 1376 GByte memory	2016	Capability computing	German universities and research institutes
SGI Altix 64-way (Q3 2005)	64 Processors 256 Gbyte memory	410	Testing and porting	German universities and research institutes
SGI Altix 128-way	128 Processors 512 Gbyte memory	820	Capacity computing	Bavarian universities
Linux Cluster Intel IA64 2-way	68 nodes 136 processors 816 GByte memory	870	Capability and capacity computing	Bavarian universities
Linux Cluster Intel IA64 4-way	17 nodes 68 processors 218 GByte memory	354	Capacity computing	Munich universities
Linux cluster Intel IA32 Intel&AMD EM64T	154 nodes 192 processors 320 GByte memory	850	Capacity computing	Munich universities
IBM pSeries 690 hpc 8-way	1 node 8 processors 32 GBytes memory	42	Capacity computing	Munich universities

A detailed description can be found on LRZ's web pages: www.lrz.de/services/compute

HLRS

Based on a long tradition in supercomputing at Universität 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 computing power for engineering and scientific applications.

Operation of its systems is done together with T-Systems, T-Systems sfr, and Porsche in the public-private joint venture hww (Höchstleistungsrechner für Wissenschaft und Wirtschaft). Through this co-operation a variety of systems can be provided to its users.

In order to bundle service resources in the state of Baden-Württemberg HLRS has teamed up with the Computing Center of the University of Karlsruhe in the hkz-bw (Höchstleistungsrechner-Kompetenzzentrum Baden-Württemberg).

Together with its partners HLRS provides the right architecture for the right application and can thus serve a wide range of fields and a variety of user groups.

Contact
Höchstleistungsrechenzentrum
Stuttgart (HLRS)
Universität Stuttgart

Prof. Dr.-Ing. Michael M. Resch
Nobelstraße 19
70500 Stuttgart
Germany
Phone +49 711 685 872 69
resch@hlrs.de
www.hlrs.de

View of the NEC SX-8 at HLRS

Compute servers currently operated by HLRS are

System	Size	Peak Performance (GFlop/s)	Purpose	User Community
NEC SX-8	72 8-way nodes 9,22 TB memory	126700	Capability computing	German universities, research institutes, and industry
TX-7	32 way node 256 GByte memory	192	Preprocessing	German universities, research institutes, and industry
Intel Nocona Cluster	205 2-way nodes 240 GB memory	2624	Capability computing	German universities, research institutes, and industry
Cray Opteron	129 2-way nodes 512 GByte memory	1024	Capability computing	German universities, research institutes, and industry

The John von Neumann Institute for Computing (NIC) is a joint foundation of Forschungszentrum Jülich and Deutsches Elektronen-Synchrotron DESY to support supercomputer-aided scientific research and development. Its tasks are:

Provision of supercomputer capacity for projects in science, research and industry in the fields of modelling and computer simulation including their methods.
The supercomputers with the required information technology infrastructure (software, data storage, networks) are operated by the Central Institute for Applied Mathematics (ZAM) in Jülich and by the Centre for Parallel Computing at DESY in Zeuthen.

Supercomputer-oriented research and development in selected fields of physics and other natural sciences, especially in elementary-particle physics, by research groups of competence in supercomputing applications. At present, research groups exist for high energy physics and complex systems; another research group in the field of "Computational Biophysics" is being established.

Education and training in the fields of supercomputing by symposia, workshops, school, seminars, courses, and guest programmes.

The IBM supercomputer "Jump" in Jülich (Photo: Research Centre Jülich)

The following supercomputers are available for research projects of the communities mentioned below, evaluated by the Peer Review Board of NIC. A more detailed description of the supercomputers can be found on the web servers of the Research Centre Jülich and of the German Electron Synchrotron DESY, respectively:

http://www.fz-juelich.de/zam/CompServ/services/sco.html
http://www-zeuthen.desy.de/main/html/home/

System	Size	Peak Performance (GFlop/s)	Purpose	User Community
IBM pSeries 690 Cluster 1600 "Jump"	41 SMP nodes 1312 processors POWER4+ 5248 GBytes memory	9000	Capability computing	German universities, research institutes, and industry
APEmille (special purpose computers)	4 racks 1024 processors 32 GByte memory	550	Capability computing	Lattice gauge theory groups at Universities and research institutes

Contact
John von Neumann Institute for Computing (NIC)
Central Institute for Applied Mathematics (ZAM)

Prof. Dr. Dr. Thomas Lippert
52425 Jülich
Germany
Phone +49 24 61 61 64 02
th.lippert@fz-juelich.de
www.fz-juelich.de/nic

High Performance Computing Courses and Tutorials

LRZ
www.lrz.de

Programming of High Performance Systems

Date

July 18-22, 2005

Location

Leibniz Computing Center, Munich

Contents

- Basic concepts in HPC and modern HPC architectures
- Parallel programming with MPI and OpenMP
- Optimization for modern processor architectures
- Intel Itanium architecture, compilers and tools
- SGI Altix architecture and tools
- Examples for the optimization of Fortran 95, C++, and I/O
- Tracing, profiling and analysis tools.

This course is organized by LRZ in collaboration with RRZE, Intel and SGI.

Course language is English, if requested.

Webpage

http://www.lrz.de/services/compute/courses/

HLRS
www.hlrs.de

Parallel Programming with MPI, OpenMP, and PETSc

Date

September 12-14, 2005

Location

HLRS, Stuttgart

Contents

The focus is on programming models MPI-1, OpenMP, and PETSc. It includes also an overview on MPI-2. Hands-on sessions (in C and Fortran) will allow users to immediately test and understand the basic constructs of the Message Passing Interface (MPI) and the shared memory directives of OpenMP. Course language is ENGLISH (if required).

Webpage

http://www.hlrs.de/news-events/events

Advanced Topics in Parallel Programming

Date

September 15-16, 2005

Location

HLRS, Stuttgart

Contents

Topics are MPI-2 parallel file I/O, OpenMP tools and tuning, hybrid mixed model MPI+ OpenMP parallelization, domain decomposition of structured and unstructured grids and with particle based applications, parallel numerics, object oriented parallel programming with C++, and Grid computing. Course language is ENGLISH (if required).

Webpage

http://www.hlrs.de/news-events/events

Iterative Linear Solvers and Parallelization

Date

September 26-30, 2005

Location

University of Kassel

Contents

The focus is on iterative and parallel solvers, the parallel programming models MPI and OpenMP, and the parallel middleware PETSc. Thereby, different modern Krylov Subspace Methods (CG, GMRES, BiCG-STAB ...) as well as highly efficient preconditioning techniques are presented in the context of real life applica tions. Hands-on sessions (in C and Fortran) will allow users to immediately test and understand the basic constructs of iterative solvers, the Message Passing Interface (MPI) and the shared memory directives of OpenMP. This course is organized by University Kassel, HLRS, and IAG.

Webpage

http://www.hlrs.de/news-events/external-events/

Introduction to Computational Fluid Dynamics

Date

October 10-14, 2005

Location

HLRS, Stuttgart

Contents

Numerical methods to solve the equations of Fluid Dynamics are presented. The main focus is on explicit Finite Volume schemes for the compressible Euler equations. Hands-on sessions will manifest the content of the lectures. Participants will learn to implement the algorithms, but also to apply existing software and to interpret the solutions correctly. Methods and problems of parallelization are discussed.

This course is organized by HLRS, IAG, and University Kassel, and is based on a lecture and practical awarded with the "Landeslehrpreis Baden-Württemberg 2003" (held at University Stuttgart)

Webpage

http://www.hlrs.de/news-events/events

NIC
www.fz-juelich.de/nic

User Course "The IBM Supercomputer in Jülich: Programming and Usage"

Date

July 4-5, 2005

Location

NIC/ZAM, Research Centre Jülich

Contents

This course gives an overview on the IBM supercomputer "Jump" in Jülich. Especially new users will learn how to program and use this system efficiently. The following topics are discussed in detail: system architecture, usage model, compiler, tools, monitoring, MPI, OpenMP, performance optimisation, mathematical software, and application software.

Webpage

http://www.fz-juelich.de/zam/neues/termine/IBM-Supercomputer

Education in Scientific Computing

Date

August 8-October 14, 2005

Location

NIC/ZAM, Research Centre Jülich

Contents

Guest Students' Programme "Scientific Computing" to support education and training in the fields of supercomputing. Application deadline was April 30, 2005; the event is already fully booked.

Webpage

http://www.fz-juelich.de/zam/gaststudenten

Parallel Programming with MPI, OpenMP, and PETSc

Date

November 28-30, 2005

Location

NIC/ZAM, Research Centre Jülich

Contents

The focus is on programming models MPI, OpenMP, and PETSc. Hands-on sessions (in C and Fortran) will allow users to immediately test and understand the basic constructs of the Message Passing Interface (MPI) and the shared memory directives of OpenMP. This course is organized by NIC/ZAM in collaboration with HLRS.

Presented by Dr. Rolf Rabenseifner, HLRS

Webpage

http://www.fz-juelich.de/zam/neues/termine/mpi-openmp

Miscellany

HLRS Personnel

Prof. Wolfgang Nagel was elected chairman of the steering committee of the HLRS in October 2004. Prof. Nagel was a member of the steering committee since the setting up of HLRS in 1996 and is hence familiar with the framework and activities of HLRS. He is the director of the Center for High Performance Computing Dresden (ZHR) and holds a chair for Computer Architecture at the Department of Computer Science of the Technical University of Dresden.

Springer: Transactions of the HLRS

HLRS held its 8th Results and Review Workshop in October 2004. The best papers of this workshop are published as Transactions of the High Performance Computing Center Stuttgart (HLRS) by Springer as "High Performance Computing in Science and Engineering 2004 – Transactions of the High Performance Computing Center Stuttgart (HLRS)"

Workshop Reports

8th HLRS Metacomputing & Grid Computing Workshop

Researchers and industrial experts from all over the world met from March 9-11 at the HLRS in Stuttgart for the eighth time in a row to discuss current issues in usage of supercomputers in Grids environment. The workshop was collocated with a workshop on industrial Grid computing organized by the European project GridCoord (www.gridcoord.org). The participants discussed Grid activities in Japan, South Korea, the USA, the Netherlands, France, Spain, and Germany. The vivid exchange of views and research results helped to strengthen the ties between the various Grid communities and will be continued next year.

2nd Russian-German Advanced Research Workshop on Computational Science and HPC

The 2nd Russian German Research Workshop on Computational Science and High Performance Computing took place in Stuttgart from 14th to 16th of March. As the 1st workshop in Akademgorodok it was arranged by Prof. Yurii Shokin (member of the presidium of the Russian Academy of Sciences http://www.ras.ru/ and director of the Institute of Computational Technologies of the Siberian Branch of the Russian Academy of Sciences http://www.ict.nsc.ru) and Prof. Michael Resch (Director of the High Performance Computing Center Stuttgart http://www.hlrs.de and chair for high performance computing at the University of Stuttgart http://www.uni-stuttgart.de). Thirty Russian, Kazakh, and German scientists discussed a broad range of topics spanning from theoretical to application problems and covering multiple disciplines like the Discretization of the Navier-Stokes Equations, Analysis of Vortices, Relativistic Systems, Turbulent Flows, Computational Aero Acoustics, Airbag Simulation, Molecular and Dissipative Particle methods, Meteorological Flood Forecast, Web-based Frameworks, Software Engineering for Numerical Methods, Water Turbine Simulation, Image Registration in Medical Applications, Visualization, Hardware Properties and Performance Analysis.

Long term target of the collaboration between the Russian Academy of Sciences and the HLRS is the fruitful interaction of the two scientific cultures. More information on the workshop and the collaboration can be found at http://www.grc-hpc.de

The next workshop is planned to be held in Almaty/Kazakhstan in September 2005.

The workshop was sponsored by the German Research Foundation (DFG).

Contacts

- Prof. Michael Resch (resch@hlrs.de)
- Dr. Nina Shokina (shokina@hlrs.de)
- Uwe Küster (kuester@hlrs.de)

2nd HLRS-NEC Teraflop Workbench Workshop

The second teraflop workshop was held in Stuttgart at March 17 and 18. This workshop is part of the Teraflop-Workbench cooperation between NEC and HLRS. With more than 50 participants from Europe, Asia and the USA, it brought together scientists, application developers, international experts and hardware designers to discuss the present and future of supercomputing. Speakers like Jonathan T. Carter (LBNL), Toshiyuki Imamura (University of Electro-Communications, Tokyo), Yoshiyuki Miyamoto (NEC Research), and Phillip Schlatter (ETH Zürich) talked about performance and applications on vector systems. At the second day experts like Jack Dongarra (ICL, University of Tennessee), Ryutaro Himeno (Riken, Japan), Satoru Tagaya (NEC) and Gerhard Wellein (RRZE, University Erlangen) discussed future architectures in supercomputing.

During the afternoon sessions computational scientists presented their current work. Many projects were able to show the first experiences on the new NEC SX-8 systems currently at installation at HLRS. Altogether they presented some of the results of the first project year, showing different solutions to achieve the application performance that is required to gain new insights in the field of computational science.

The 4th hww/HLRS Workshop on Scalable Global Parallel File Systems and HNF Europe 2005 Spring Meeting

April 4th and April 5th, 2005, HLRS, Stuttgart, Germany

This year, HLRS hosted and organized for the forth time the annual Workshop on Scalable Global Parallel File Systems. This two day event with nearly 90 participants was opened on Monday morning by the director of HLRS, Prof. Dr. Michael Resch. The key note speech was given by Dr. Peter Braam, founder and CEO of Cluster File Systems Inc. and also the mind and the spirit behind Lustre, a fast evolving and meanwhile often used Cluster File System. In the following presentations different cluster and SAN file systems were introduced. Other talks of the first day covered the topic of Grid File Systems.

The first day was completed by the workshop dinner, which was served in the beautiful Gasthof Lamm in Schlat after a tour to the Hauff Museum of the Prehistoric World in Holzmaden.

On the second day, the first of two concurrent tracks addressed the evolvement of Fibre Channel and other Storage Network technologies. In addition, the improvements in different HSM Systems and also the performance of different Lustre deployments were presented. The second track was dedicated to high performance networking. This included the presentation of future connection technologies as well as the future developments on the physical link layer. In addition, this track saw the formation of the OpenIB Users Group Europe. This formation session covered several Infiniband issues from physical layer and kernel problems to storage and message passing. To extend these discussions, there are already plans to extend next years meeting by an additional day, dedicated to Infiniband and its development only.

HLRS acknowledges the support of our sponsors who made the successful event possible.

inSiDE

inSiDE is published two times a year by
The German National Supercomputing
Centers HLRS, LRZ, and NIC

Publishers
Prof. Dr. Heinz-Gerd Hegering, LRZ
Prof. Dr. Dr. Thomas Lippert, NIC
Prof. Dr. Michael M. Resch, HLRS

Editor
F. Rainer Klank, HLRS
klank@hlrs.de

Design
Katharina Schlatterer
kschlatterer@hlrs.de

Authors
Holger Berger
hberger@ess.nec.de
Matthias Brehm
brehm@lrz.de
Graham E. Fagg
fagg@cs.utk.edu
Edgar Gabriel
gabriel@hlrs.de
Paul Gibbon
p.gibbon@fz-juelich.de
Richard L. Graham
rlgraham@lanl.gov
Heiner Igel
heiner.igel@geophysik.uni-muenchen.de
Peter Lammers
lammers@hlrs.de
Matthias Müller
mueller@hlrs.de
Michael M. Resch
resch@hlrs.de
Bernhard Schuberth
bernhard.schubert@geophysik.uni-muenchen.de
Jeffrey M. Squyres
jsquyers@open-mpi.org
Jörg Stadler
jstadler@hpce.nec.com

If you would like to receive inSiDE regularly, send an email with your postal address to klank@hlrs.de