inSiDE

inSiDE • Vol. 2 No. 1 • Spring 2004

Innovatives Supercomputing in Deutschland

Editorial

With this issue of InSiDE the German Federal Supercomputing Centers in München (LRZ), Jülich (NIC) and Stuttgart (HLRS) are continuing their initiative to disseminate information about the new concepts of supercomputing as are embraced by the German supercomputing research community. This issue comes with the International Supercomputer Conference 2004 at Heidelberg.

Traditionally we start with a number of application reports. This issue introduces three contributions from a variety of fields. Cameron et al. from the Max-Planck-Institute for Aeronomy describe a simulation of magneto-convection in the solar atmosphere. Irmgard Frank from the Department of Chemistry of the LMU München investigated the photochemical reaction of the retina protein rhodopsin that initiates the process of vision. Dr. Wolfgang Wenzel from the Department of Nanotechnology of the Research Centre Karlsruhe (FZK) presents a paper on measuring and explaining electronic conduction at the molecular scale. Such simulations will help to construct diodes, transistors and wires from individual molecules in the future which offers one promising avenue for further miniaturization of electronic circuits.

In our Grid section we are able to present a first description of a German e-science initiative called D-Grid. The focus of this initiative will be both on combining resources to establish a Grid infrastructure and on creating communities in e-science that are able to increase their productivity through a better usage of the Grid.

The International Supercomputer Conference has created the ISC 2004 Award to reward and promote those people who have distinguished themselves through state of the art projects in fields requiring high performance computing. Here we present a short version of the paper winning the award for the category "Requirements for HPC Systems Software".

Our section on supercomputers starts with Europe's fastest system that was recently installed by the John von Neumann Institute for Computing (NIC) in Jülich. The IBM system has a peak performance of about 9 TFlops/s and puts German high performance computing back on the international supercomputing landscape. The spiral of innovation in supercomputing in Germany will be turned one twist further in 2005 with a new supercomputer system installed at the High-Performance Computing Center Stuttgart (HLRS). The concept and basic architecture of this system are briefly described.

Finally you will again find information about the available supercomputer systems at the three centers and about workshops offered. All workshops are open to researchers from Germany. Participants from outside Germany or from industry should contact the responsible center for terms of participation. We have newly introduced a miscellany section to provide you with additional information about the three centers. By this we hope to give you more insight into what happens behind the scenes.

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

Contents

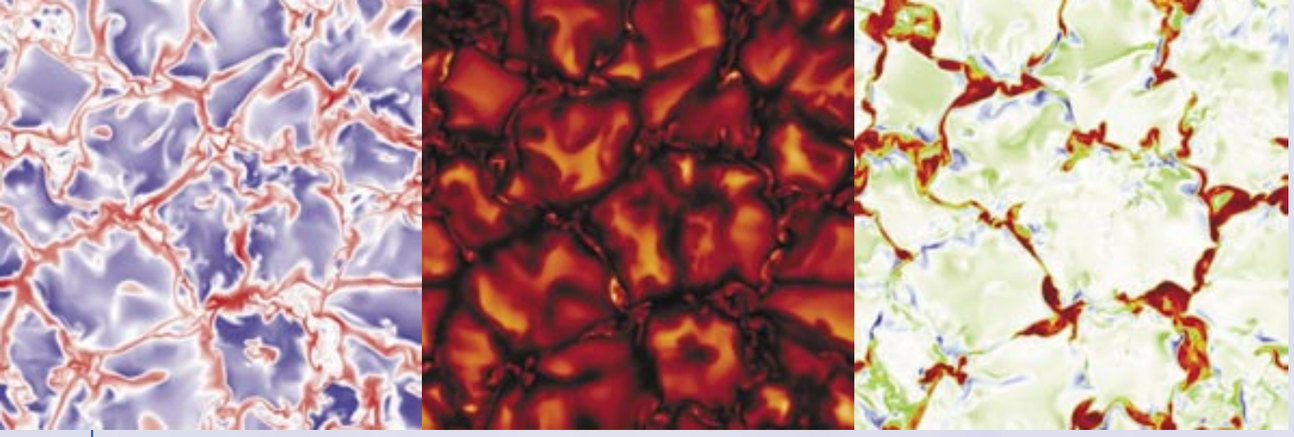


Figure 1: Vertical magnetic field (left; red: downward directed, blue: upward), brightness (middle), and vertical velocity (right; red: downflow, blue: upflow) in a horizontal cut through the 3D simulation box near the visible solar surface

The Sun – A Ball of Electrically Well-Conducting Gas

The Sun is a ball of electrically well-conducting gas (plasma). Bulk flows induce electrical currents and generate a large-scale magnetic field through self-excited dynamo action. The interaction of the vigorous convective motions of the plasma with the magnetic field leads to intermittent magnetic structure: strong concentrations of magnetic flux are embedded in almost field-free plasma. This magneto-convective interaction is most pronounced in the visible layers of the solar atmosphere, where the convective speeds become nearly sonic and the magnetic, kinetic, and thermal energy densities are all of the same order of magnitude.

Apart from being of intrinsic scientific interest, this topic has far-reaching relevance: [1] The magnetic structure affects the local emissivity of the plasma and leads to variations of the solar radiation output with potential effects on Earth's climate. [2] The continuous rearrangement and reshuffling of magnetic flux by the instationary convection represents a major source of energy for heating the solar corona to temperatures of a few million Kelvin. [3] Magneto-convective processes are relevant for the activity, variability, and non-thermal heating of other stars as well as for accretion phenomena in connection with star formation, compact objects, and active galactic nuclei.

The nonlinear interaction between magnetic field, convection and radiation is far too complex for analytical treatment, so we resort to numerical simulations based upon the fundamental equations of magneto-hydrodynamics and radiative transfer. To perform realistic 3D time-dependent simulations of compressible magneto-convection in the visible layers of the solar surface represents a considerable computational challenge. There is a wide range of length scales between the dominant scale of the convective flow pattern and the dissipation scales. The plasma is strongly stratified and even a restricted simulation has to cover a density ratio of at least 100. Solar convection is strongly affected by partial ionization effects, so that the ionization state of the most abundant species has to be monitored. Radiation is a major player in the energy budget and has to be treated non-locally by integration over rays, taking into account its frequency-dependence.

Figure 1 shows a snapshot from a simulation run. Shown is the distribution of magnetic field (left), brightness (middle), and vertical velocity (right) on a 6000 km times 6000 km piece of the visible solar surface. The brightness and velocity images show the typical signature of convective energy transport with hot, bright upflows and a network of cool downflows with velocities of several km/s. The magnetic flux is concentrated in the downflow lanes of the convection pattern and reaches field strengths of up to 2000 Gauss. The strong magnetic field leads to a significant density reduction; the resulting increase of the plasma transparency leads to light emission from deeper, hotter regions, so that the flux concentrations appear as bright features in the brightness picture. This effect leads to a slightly larger total solar radiance during times of high magnetic flux levels on the solar surface. The observed brightness on a piece of solar surface of the same size as the simulation domain is shown in Figure 2.

A three-dimensional view of magnetic field lines in a part of the simulation box is shown in Figure 3. While the magnetic flux is strongly concentrated in the network of convective downflows in the surface layers (grey area) and below, it expands strongly upward, so that a "canopy" structure evolves and the field becomes space-filling in the upper layers of the solar atmosphere.

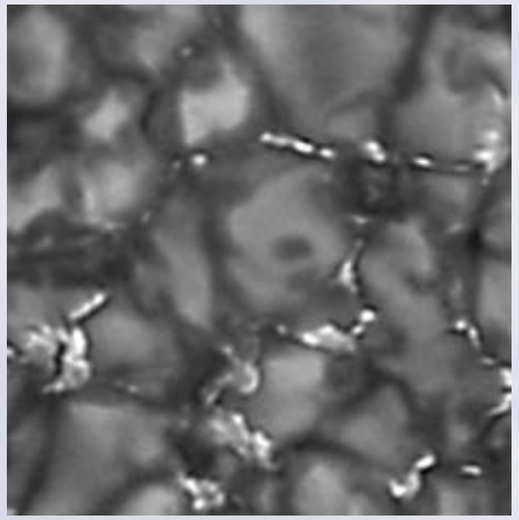


Figure 2: Observed brightness on a piece of solar surface (image taken with the 1m solar telescope of the Institute for Solar Physics/Royal Swedish Academy of Sciences on the Island of La Palma, Spain)

Apart from allowing us to study the physical processes of magneto-convection in 3D, the simulations also represent a valuable tool for the analysis of solar observations. Diagnostic quantities like intensity maps in various wavelengths, spectral line profiles, and light polarization properties are determined from the simulations and can be directly compared with the corresponding observational quantities.

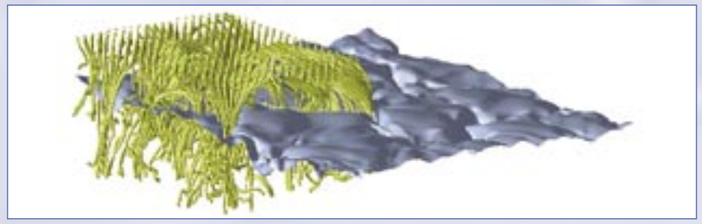


Figure 3: Magnetic field distribution in the simulation box. The grey area represents the visible solar surface, the yellow wires are magnetic field lines (drawn only in a quarter of the simulation box, for better visibility)

• Dr. Robert Cameron
• Manfred Schüssler
• Dr. Alexander Vögler

Max-Planck-Institute for Aeronomy, Kathlenburg-Lindau

First Principles Simulations on the Nanoscale

At distances of less than ten nanometers, quantum effects are important in the correct description of the structural properties of materials. With present-day supercomputers quantum chemical computations are possible for systems with up to about 1000 nuclei and 5000 electrons. The quantum mechanically delocalized electrons form a negatively charged cloud that holds the nuclei together (Figure 1). The input for a quantum chemical calculation essentially consists of an estimate of the three-dimensional arrangement of the nuclei. During the geometry optimization, the distances between the atoms are optimized according to the elementary equations in an approximate formulation, yielding molecular structures that are in very good agreement with experiment. The calculation thus provides in the output an optimized arrangement of the nuclei plus the electronic cloud that stabilizes the structure.

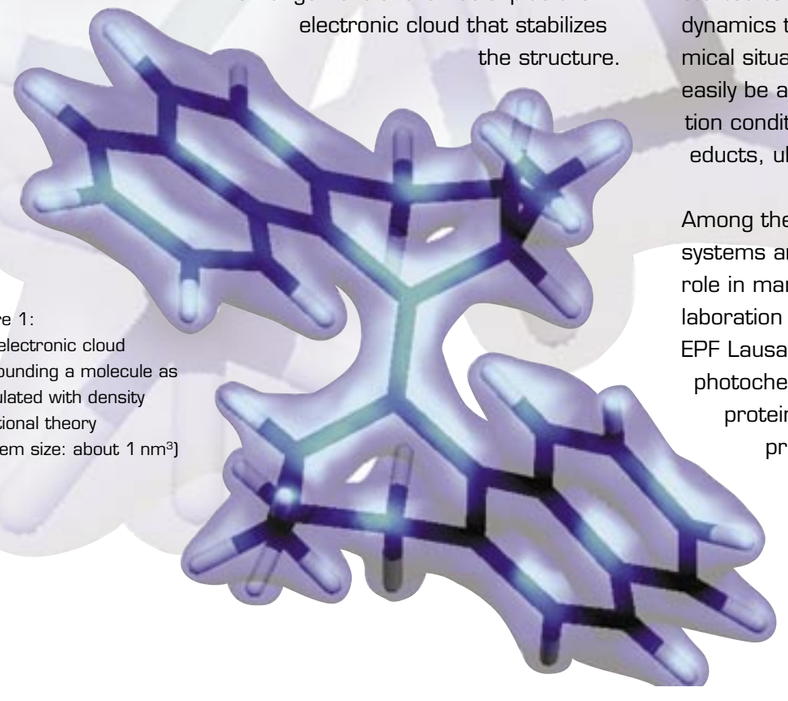


Figure 1:
The electronic cloud surrounding a molecule as calculated with density functional theory (system size: about 1 nm^3)

In the past decades major advancements have been made in increasing the accuracy of the calculations. An important step was the development of density functional theory, also called first principles theory (Walter Kohn, Nobel Prize 1998). Based on this theory, Roberto Car and Michele Parrinello devised a scheme for performing first principles molecular dynamics in 1985 (Car-Parrinello molecular dynamics, CPMD). With this approach it is possible not only to calculate an optimized molecular structure but also to model the thermal motion of the nuclei. This renders the simulation of chemical reactions feasible. The thermal motion of the nuclei may effect the breaking of bonds or the formation of new bonds provided the electronic clouds around the nuclei can interact correctly. In the past few years quantum chemists started to use first principles molecular dynamics to investigate complex chemical situations that otherwise can not easily be analysed (e.g. extreme reaction conditions, expensive and/or toxic educts, ultrafast reaction sequences).

Among the most complex molecular systems are proteins that play a key role in many biological processes. In collaboration with the ETH Zürich and the EPF Lausanne we have investigated the photochemical reaction of the retina protein rhodopsin that initiates the process of vision. This system represents a real challenge for quantum chemical approaches. One of the main problems is the size of the system (Figure 2).

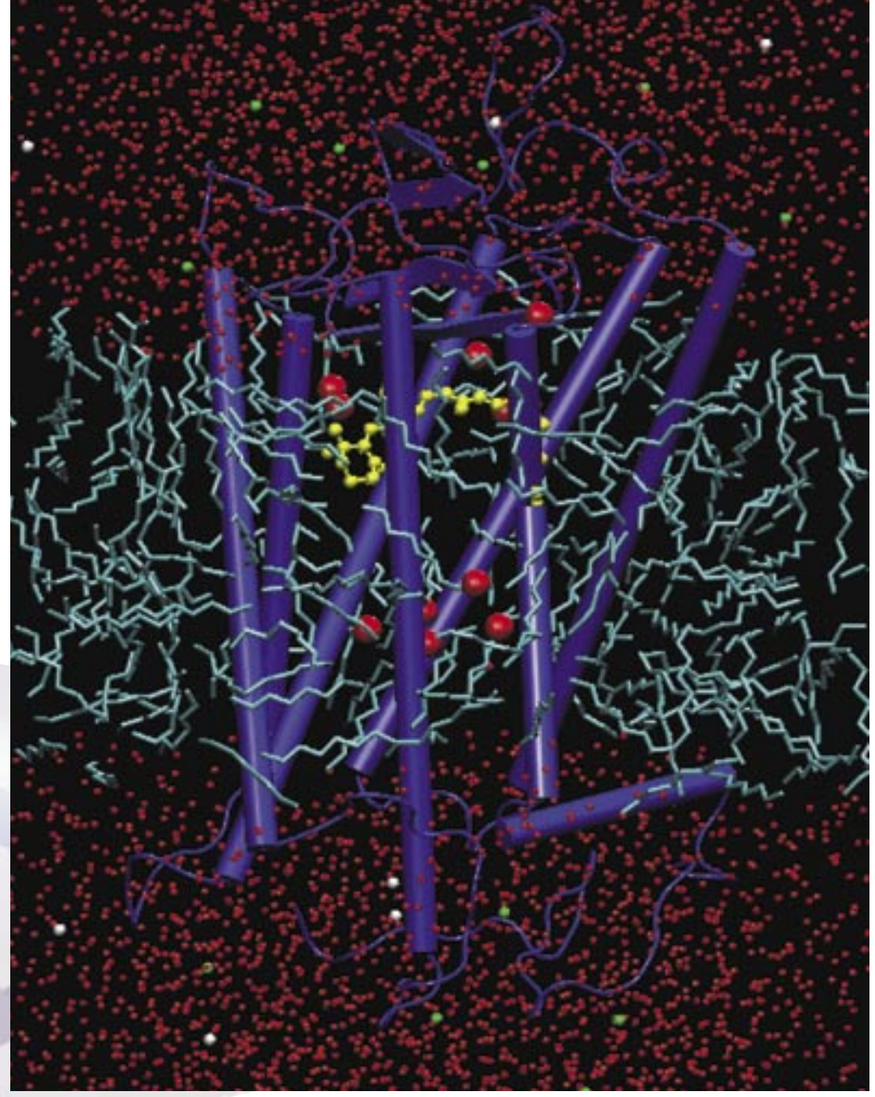


Figure 2:
Model of rhodopsin plus environment (system size: about 10 nm^3)

The membrane protein rhodopsin is stable in its active state only if it is surrounded by a suitable environment.

A model of rhodopsin in its environment contains as many as 24000 atoms, clearly exceeding the scope of present-day first principles approaches. However, the photochemical reaction is known to be confined to a small part of the protein. Hence, the rest of the protein and the surrounding can be described with a simpler classical approach, while the quantum mechanical description is restricted to the active part of the protein (QM/MM method: a combination of quantum mechanics and molecular mechanics). Using a computationally effcient extension of density functional theory for the description of photoreactions, we succeeded to simulate the full dynamics of the protein after photon absorption on the Hitachi SR8000-F1 at LRZ. The simulations show in full detail how the light energy induces a twist in the photoactive part on a timescale of less than 200 femtoseconds. Thus the system is destabilized which leads to further rearrangements in the protein, ultimately inducing a neural signal.

• Dr. Irmgard Frank

Chemistry Department,
LMU München

Measuring and Explaining Electronic Conduction at the Molecular Scale

The increase of available computational power over the past decades has been driven largely by an increase of the packing density of electronic components and the miniaturization of electronic circuitry. Extrapolating this trend, silicon based technology is predicted to reach dimensions on the nanometer scale by the year 2015 - 2020. To drive device technology to this limit, the behavior of electronic devices at the molecular or atomic scale must be investigated. The construction of diodes, transistors and wires from individual molecules offers one promising avenue for further miniaturization of electronic circuits.

Although it was already postulated in the mid seventies that single molecules could serve as electronic devices in principle, experimental investigation of the electronic conduction through single molecules has been possible only since the mid 1990's. Several experimental designs exist, but of particular interest are break-junction experiments, which permit the controlled covalent contact of both ends of the molecule to the electrodes and possibly the additional application of a gate voltage.

In one type of such experiments, a thin gold wire is mounted on top of a surface, which is pushed slowly upward with a piston in the center until the wire breaks (Figure 1). This setup permits the control of the gap between the electrodes to fractions of an Angstrom and individual molecules are contacted on both sides of the gap in the wire.

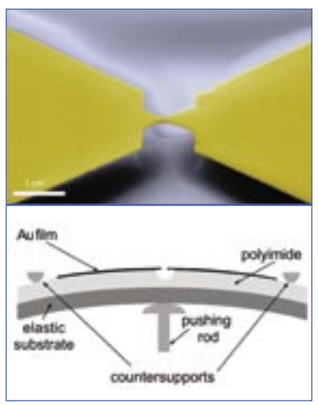
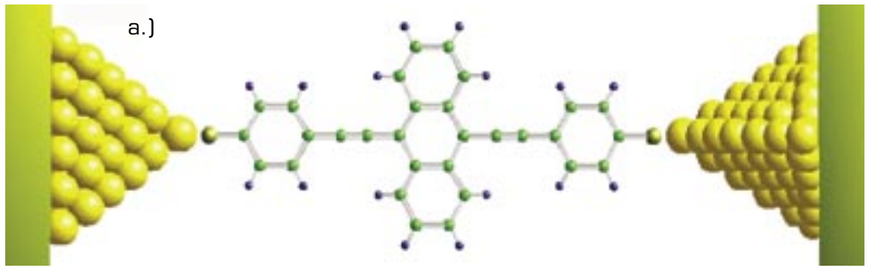


Figure 1:
Top view and schematic illustration of gold break-junction

Several interesting effects, such as the breakdown of the current in a given voltage window (negative differential conductance), asymmetric current-voltage characteristics and current enhancement through magnetic atoms in the molecule (Kondo Effect) have been observed; many still lack theoretical explanation.

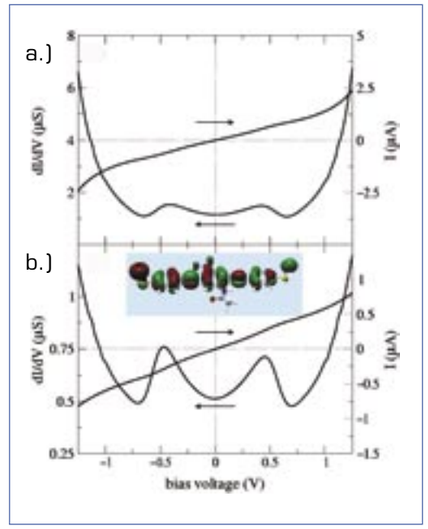
Since these effects depend strongly on the molecule used, their theoretical elucidation requires the application of high-resolution electronic structure methods (such as density functional theory or correlated methods of quantum chemistry) in conjunction with atomistic electron transport theory. These calculations, in particular quantum chemical many-body methods, require large computational resources. We have developed a massively parallel implementation of the configuration interaction method that can build and diagonalize matrices of dimension up to 10^{12} x 10^{12}. Because of the large CPU and memory requirements of a single run, this code requires the use of massively parallel architectures, where it runs scalably on up to 256 processors of the CRAY T3E or up to 128 processors of the IBM supercomputer Jump at the John von Neumann Institute for Computing (NIC). The electronic structure of the molecule varies with the applied voltage, so these calculations must be repeated for every point of the current-voltage characteristic.



Atomistic simulations have provided insights into the mechanisms of electronic conduction on the molecular and atomic scale, explaining for example, recent experimental results for the conduction properties of molecular hydrogen, which turned out to be a much better conductor than could be anticipated on the basis of simple estimates. It was also possible to semi-quantitatively reproduce and explain the current-voltage characteristics of several organic molecules that were recently investigated experimentally (Figure 2).

The inset in the bottom panel of the figure illustrates the charge distribution in the conduction channel which is responsible for the asymmetry of the current-voltage characteristic.

Alternative theoretical methods permit the prediction of novel effects under experimental conditions which are presently hard to achieve. In the limit of controlled, but weak coupling of the molecule to the electrodes, surprising features of the current-voltage characteristic, such as the complete breakdown of the current in a finite voltage window may be observable in comparatively simple systems. The fruitful interplay of experimental effort and quantitative theoretical modeling thus promises further scientific surprises and hopefully technological applications in the foreseeable future.



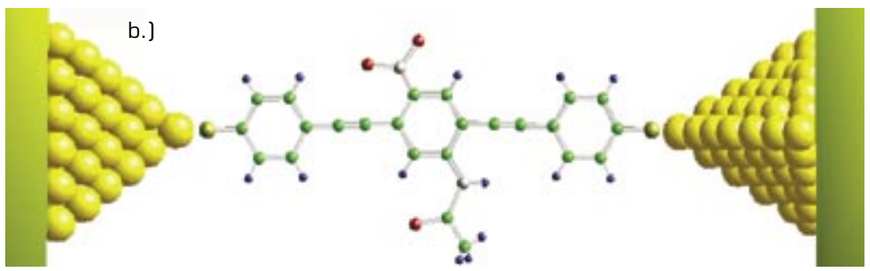


Figure 2: Organic molecules and associated computed current-voltage characteristics

• Dr. Wolfgang Wenzel

Department of Nanotechnology, Research Centre Karlsruhe (FZK)

D-Grid
Towards E-Science in Germany

When Germany started its first Grid Initiative UNICORE in the word Grid – as a description for transparent usage of distributed resources by virtual organisations – had not been coined yet. However, UNICORE and its successor project UNICORE+ certainly marked the beginning of a Grid initiative in Germany. Although it came up with a product, it was clear at the end of the project that the reach of UNICORE was too short. It only tackled the question of accessing supercomputing resources, but there was no whatsoever support for a full fledged e-science infrastructure.

This situation was addressed in the final UNICORE+ Review Meeting in December 2002 and planning for a German e-science initiative started immediately picking up on the many ideas and other projects that had grown among a vibrant Grid community in Germany over time. The code name for the initiative was chosen to be D-Grid. The founders of the initiative are faced with the following situation.

Situation in Germany

Grid computing is well established in Germany in a variety of projects. German research groups are among the most actives in European Grid Initiatives. Already in 1998 in the European framework program 5 (FP5) a French-German pilot project (METODIS) for the industrial usage of Grid tools was initiated. A number of projects followed with German researchers partially taking the lead. In FP6 German researchers were strongly involved from the start both in projects building a European research infrastructure and in using the Grid for complex problem solving.

At the same time Grid research was conducted in a number of communities. The most notable one currently is certainly high energy physics (HEP). As part of the large hadron collider (LHC) experiments of CERN the community has set up a concept for the processing of the huge pile of data generated during the experiments. At the Research Centre Karlsruhe (FZK) a Grid Centre (GridKa) was established that acts as a tier 1 centre for the LHC. Other communities like climate and weather research have established similar activities in the field.

German researchers were also rather active in international activities for a standardization of the Grid. Already the foundation of a European Grid initiative was strongly driven by German participants – although acting on a European level with colleagues from many other countries. These activities were intensified when the Global Grid Forum was established. German researchers actively participated at all levels. The importance of German Grid activities inside GGF is certainly emphasized by the fact that GGF had its March 2004 meeting in Berlin.

A Vision for e-science

Modern information society is faced with a multiplicity of possibilities for efficient communication and easy access to large sets of data and information as well as high performance computing systems. The opportunities to achieve better results – both quantitatively and qualitatively – in science and industrial development based on completely new methods have increased dramatically. At the same time complexity and thus difficulties in handling distributed and dynamic system components have increased dramatically.

Virtualisation is therefore the key word for future Grid systems. Hiding away the complexities of data – information – and compute services and thus virtualising them has to be the goal. Future Grid infrastructures will comprise thousands of geographically distributed resources that are interconnected by wide area networks. These resources will autonomously supervise and reorganize themselves whenever necessary. While providing reliable, fault tolerant and ubiquitous services of high quality they will be transparent to the user.

The goal of D-Grid is a bundling of Grid initiatives in Germany to harness the potential of synergetic effects for global, distributed and innovative collaborations in science and industry. This requires the structuring both at a horizontal and a vertical level as shown in Figure 1.

At the horizontal level synergy is achieved by bundling and a common usage of middleware, services and e-science methods across scientific communities which were so far isolated. This brings about homogenisation and standardization which allows harnessing synergy for scientific work. At the lowest level we find resources like networks, computers, programs, instruments and data archives that have to be integrated to form a basic infrastructure.

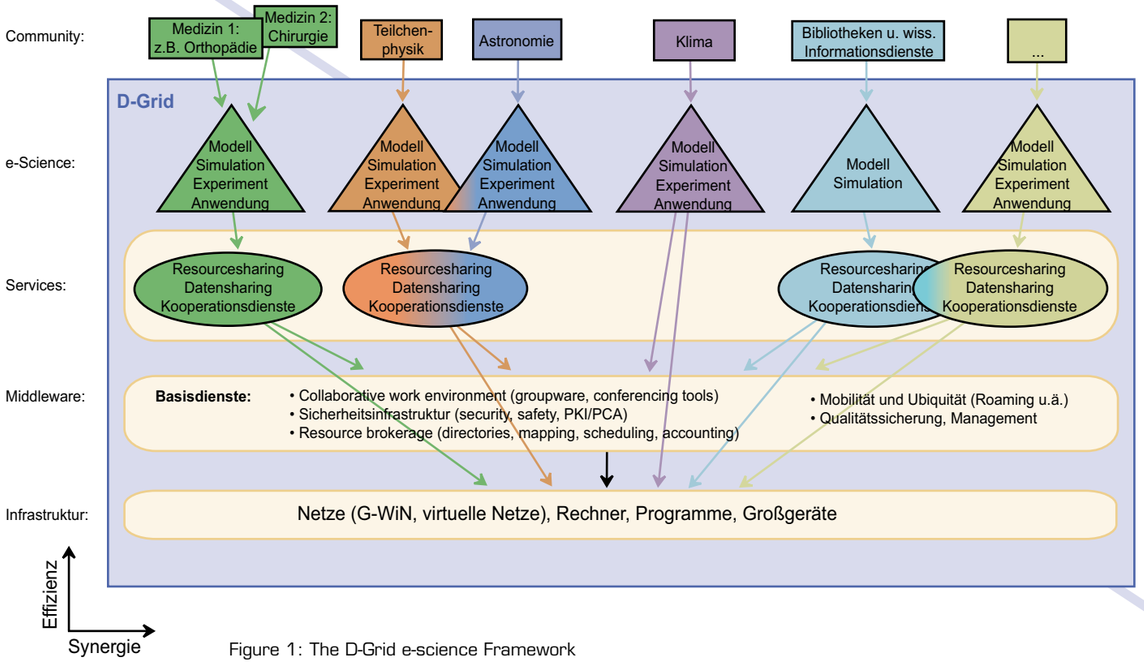


Figure 1: The D-Grid e-science Framework

At the vertical level an increase in efficiency for each scientific discipline can be seen in the creation, bundling and exploitation of services, resources and infrastructures. Figure 1 shows how this can be achieved at each layer of the conceptual view of the Grid.

Package of Measures

To achieve such an integration and exploitation of synergetic effects a set of measures will have to be taken:

- Installation of a network of competence in e-science: A well co-ordinated network of competence based on centres of excellence is the essential ingredient for a national e-science program. These centres are crystallisation points of competence in the field of Grid computing. A common head office coordinates the key responsibilities which are
  - Provision of resources and know-how
  - Development of application scenarios for science and industry
  - Research and development in Grid systems
  - Permanent integration of user requirements
- Development of Grid software components: Generic and standardized software is the basis for Grid computing in science and industry. It has to be developed in close co-operation with the international research community and with industry
- Creation of a lasting infrastructure: The core of such an infrastructure is formed by existing and improved networks, computer systems and data bases. This has to be complemented by services and a middleware layer. Although based on existing national infrastructures like the German research network DFN, the German federal HPC centre resources or the UNICORE group, this has to be extended into the international arena to be competitive
- Setting up of a number of pilot projects in e-science: Pilot projects do not only help to demonstrate the benefit of a German Grid initiative. They can also help to gradually lure more and more scientific communities into the Grid arena. Thus a true e-science space is created

Organisational Aspects

The German Grid community is based on self organisation. At plenary meetings the main issues are discussed openly and necessary decisions are made. Currently D-Grid is a group of about 45 organisations. These come from universities, research organisations and industry. Since D-Grid is an open community any research group in Germany can join by signing a memorandum of understanding and participating in the coordination process. More information on the organization and how to participate can be found at http://d-grid.de. Most of the information provided here was taken from this webpage.

The community has established a steering committee in May 2003 for the preparation of a research program and to prepare organisational concepts. Its main duties are:

- To work out a scientific vision and goals for a German Grid research program together with leading scientists in the field
- To work out a concrete R&D program focussing on infrastructures, Grid software and applications
- To act as a liaison for the federal Ministry of Science for the topics mentioned above
- To integrate various scientific fields and communities
- To work out a memorandum of understanding and lay the foundation for a work program for a German Grid and e-science initiative.

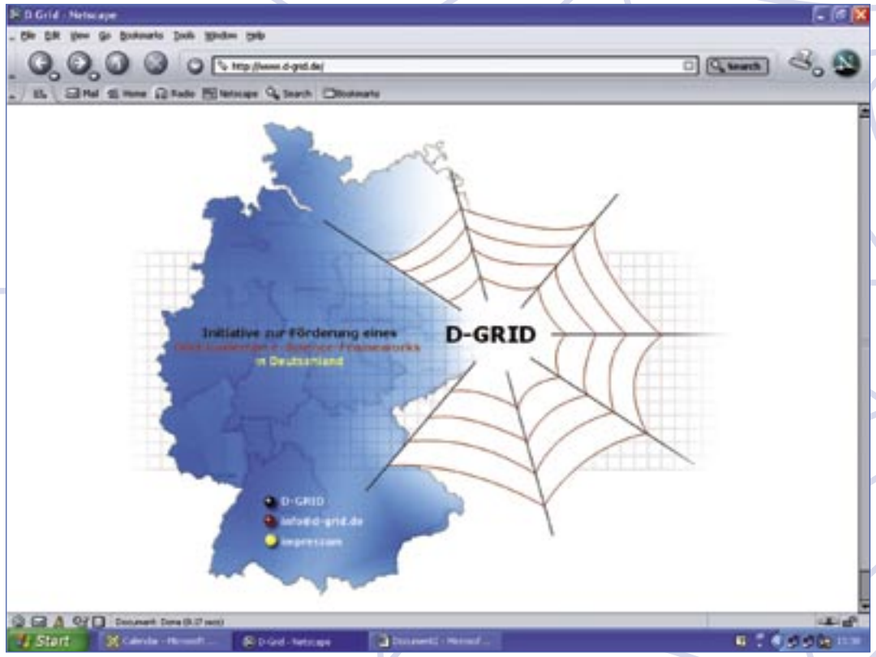
The steering committee has proposed a matrix structure for the further work on a research program. This matrix structure can also be seen as the basis for the further organisation of D-Grid.

At the horizontal level a set of working groups (WG) was established:

- Operations
- Networks
- Middleware
- Autonomic Computing/Management
- Data/Information

At the vertical level individual communities were asked to collect and formulate their requirements as well as visions for the future. Communities can centre on an application field – as is the case for the High Energy Physics (HEP) community or the Climate and Weather Research community. They can also be communities of providers like supercomputer centres or network providers.

These working groups and communities have created a work program that is finalised in June 2004. Based on this work D-Grid expects to be able to launch operation in early 2005.



• Prof. Dr. Michael M. Resch

High-Performance Computing Center Stuttgart (HLRS)

FT-MPI: Extending the MPI Specification for Process Fault-Tolerance on HPC Systems

Motivation

Today, end-users and application developers of high performance computing systems have access to larger machines and more processors than ever before. Systems such as the Earth Simulator, the ASCI-Q machines and the IBM Blue Gene consist of thousands or even tens of thousand of processors. Machines comprising 100000 processors are expected within the next few years. A critical issue of systems consisting of such large numbers of processors is the ability of the machine to deal with process failures. Based on the current experience with high-end machines, it can be concluded, that a 100000-processor machine will experience a processor failure every few minutes. While on earlier massively parallel processing systems (MPPs) crashing nodes often lead to a crash of the whole system, current architectures are more robust. Typically, the applications utilizing the failed processor will have to abort, the machine, as an entity is however not affected by the failure.

Current parallel programming paradigms for high-performance computing systems are typically based on message passing, especially on the Message-Passing Interface (MPI) [2] specification. However, the current MPI specification cannot deal with the case where one or more process failures occur during runtime. MPI gives the user the choice between two possibilities of how to handle failures. The first one, which is also the default mode of MPI, is to immediately abort the application. The second possibility is just slightly more flexible, handing the control back to the user application without guaranteeing however, that any further communication can occur. The latter mode has mainly the purpose to give an application the possibility to perform local operations before exiting, e.g. closing all files or writing a local checkpoint.

Summarizing the findings of this section, there is a discrepancy between the capabilities of current high-performance computing systems and the most widely used parallel programming paradigm. When considering machines with tens of thousand of processors, the only currently available technique to handle fault tolerance, checkpoint/restart, has its performance and conceptual limitations. However, if today's and tomorrow's high-performance computing resources were to be used as a means to perform single, large scale simulations and not solely as a platform for high throughput computing, extending the core communication library of HPC systems to deal with aspects of fault-tolerance is inevitable.

Extensions to the Message-Passing Interface for Process Fault–Tolerance

Handling fault-tolerance typically consists of three steps: failure detection, notification, and recovery. The only assumption the FT-MPI specification [1] makes about the first two steps is that the run-time environment discovers failures and all remaining processes in the parallel job are notified about these events.

The notification of failed processes is passed to the MPI application through the usage of a special error code. As soon as an application process has received the notification of a death event through this error code, its general state is changing from no failures to failure recognized. While in this state, the process is just allowed to execute certain actions.

The recovery procedure is considered to consist of two steps: recovering the MPI library and the run-time environment, and recovering the application. The latter one is considered to be the responsibility of the application.

The FT-MPI specification tackles answers to the following questions:

- What are the necessary steps and options to start the recovery procedure and therefore change the state of the processes back to no failure?
- What is the status of MPI objects after recovery?
- What is the status of ongoing messages during and after recovery?

The first question is handled by the so-called recovery-mode, the second by the communicator-mode, and the third by the message-mode respectively the collective-communication-mode. FT-MPI offers for each of those modes several possibilities. This allows application developers to take the specific characteristics of their application into account and use the best-suited method to handle fault-tolerance.

An Implementation of the FT-MPI Specification

The University of Tennessee implemented the specification presented in the previous section. The current version is relying on the HARNESS framework. HARNESS (Heterogeneous Adaptable Reconfigurable Networked SyStems) [1] provides a fault-tolerant, dynamic run-time environment, which is used by FT-MPI for process management and failure notification.

The currently available functionality in FT-MPI includes the full MPI-1.2 specification, as well as several sections of the MPI-2 document. Furthermore, many efforts have been invisted in optimizing the collective operations and

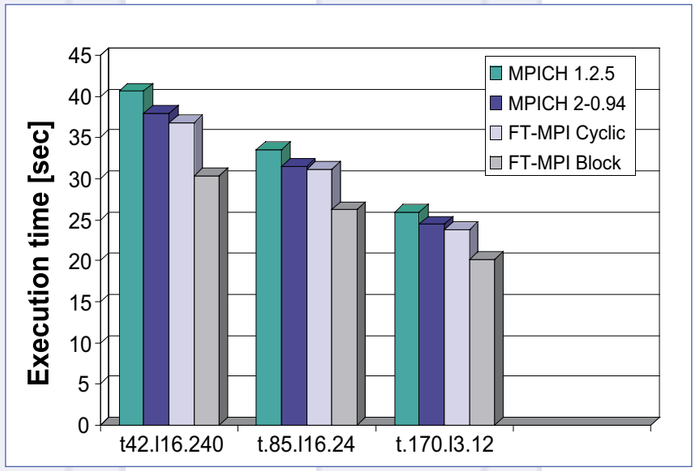


Figure 1:
Execution times for various PSTSWM benchmark problems for MPICH-1.2.5, MPICH2 and FT-MPI (cyclic and blockwise process distribution)

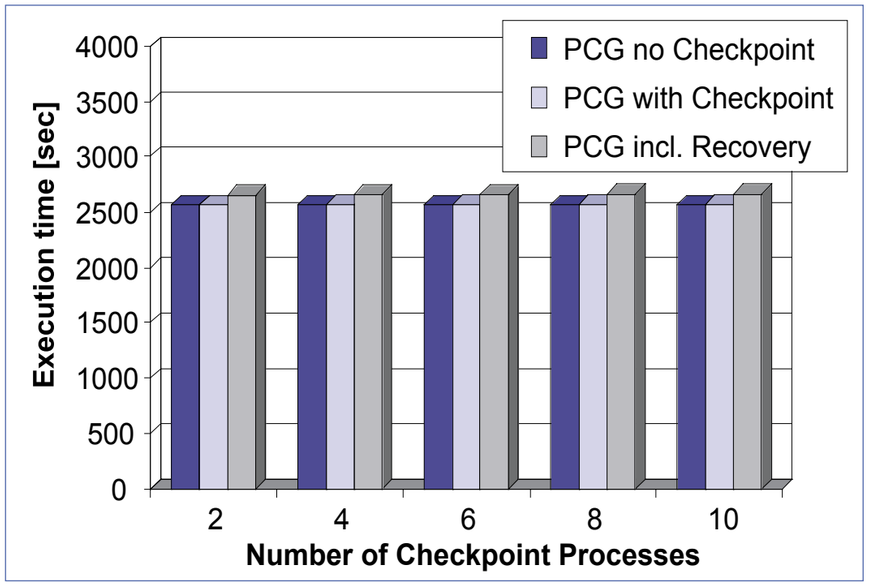


Figure 2:
Comparison of a preconditioned CG solver for various scenarios and number of checkpoint processes

the derived datatype section of MPI. The library has been tested on a wide variety of platforms, operating systems and compilers, e.g. IA-32 LINUX and Windows platforms, AMD 64-bit architectures, SUN, SGI and HP platforms.

In various benchmarks and application scenarios, FT-MPI has proven that its performance is comparable to the current portable state-of-the-art MPI libraries during a fault-free execution. This indicates, that as long as no error occurs, the new specification does not harm the performance of applications. Furthermore, the results show, that the FT-MPI specification is compatible to the current specifications of MPI, since all current MPI applications work without any modifications with FT-MPI. Figure 1 shows the execution time for various MPI implementations for three different tests of the Parallel Spectral Transform Shallow Water Code (PSTSWM) [3] benchmark.

Usage Scenarios

Simultaneously to the development of the specification and the implementation of FT-MPI, a large set of applications have been tested and benchmarked. Most of these rely on diskless checkpointing. This technique avoids writing checkpoint files by distributing additional information based on encoding techniques such as the Reed-Solomon Algorithm on other processes. When an error occurs, the application need not be restarted, since the additional information can be used to reconstruct the data of the failed process. This technique improves the performance of the application for large number of processes, compared to writing and reading checkpoint files, since it avoids typically slow file operations. Among the applications using this technique are a preconditioned conjugate gradient solver (PCG) [1], a dense matrix multiplication, and a Cholesky factorisation. Figure 2 shows the execution time for the PCG solver, using various numbers of checkpoint processes, which also indicates the number of process failures, which can be handled by the code simultaneously. The numbers indicate, that the overhead to write a diskless checkpoint is hardly measurable, and even a recovery operation introduces just a marginal overhead.

Recently, a fault-tolerant manager-worker framework has been added to the distribution. The key point of this framework is to show, that all applications, which can be written using a manager worker paradigm (e.g. many parameter studies) can easily be adapted to benefit from the features of FT-MPI.

Even with checkpoint/restart techniques based on file operations, the FT-MPI approach still offers advantages to the end-users, as shown for the parallel spectral transform shallow water code (PSTSWM). In the case errors occur, the application could re-spawn the failed processes, and reconstruct the missing data based on the last checkpoint files. The advantage compared to conventional checkpointing techniques is, that the application does not have to be stopped and restarted as an entity. Thus, the application will still finish correctly despite of any errors without requiring any user interaction.

Acknowledgments

This material is based upon work supported by the Department of Energy under Contract No. DE-FG02-02ER25536. The NSF CISE Research Infrastructure Program EIA-9972889 supported the infrastructure used in this work.

References

[1] G.E. Fagg, E. Gabriel, Z. Chen, T. Angskun, G. Bosilca, A. Bukovsky and J.J. Dongarra:
Fault Tolerant Communication Library and Applications for High-Performance Computing,
LACSI Symposium 2003, Santa Fe, October 27-29, 2003.

[2] Message Passing Interface Forum:
MPI: A Message Passing Interface Standard,
http://www.mpi-forum.org, 1995.

[3] P. H. Worley and B. Toonen:
A User's Guide to PSTSWM,
July 1995.

FT-MPI

• Dr. Edgar Gabriel
• Dr. Graham E. Fagg
• George Bosilca
• Prof. Jack J. Dongarra

Innovative Computing Laboratory,
Computer Science Departement,
University of Tennesse

Inauguration of the New IBM Supercomputer in Jülich

Figure 1: The new supercomputer in Jülich (Photo: Forschungszentrum Jülich)

On February 16, 2004 the new IBM supercomputer at the Research Centre Jülich was officially inaugurated and presented to the public. The supercomputer, nicknamed Jump (Jülich Multi Processor), is installed in the new machine room of the Central Institute for Applied Mathematics (ZAM). Jump consists of 41 nodes, each with 32 Power4+ processors (1.7 GHz) and 128 GByte shared memory. The performance of each node is 217.6 GFlop/s. Two nodes are used as login nodes as well as for I/O and data management tasks. All nodes are connected via a high-speed network, the "High-Performance Switch". The theoretical peak performance of the whole aggregate is nearly 9 TFlop/s. Actually, the LINPACK benchmark produced 5.568 TFlop/s. Thus, Jump currently is the most powerful computer in Europe and has been the first installation using this network technology in this dimension.

The opening ceremony for the supercomputer met with large response. In the morning, a press conference with 70 invited guests from politics and press was held in the new ZAM building. The official starting of Jump was carried out by Prof. Joachim Treusch, Chairman of the Board of Directors of the Forschungszentrum Jülich. When he pressed the red button, a contest between the new supercomputer Jump and the old CRAY T3E was started, both calculating the simulation of the ground water flow and transport of dissolved substances in the soil. The results were visualised for the audience on a projection screen during the conference. After about 35 minutes, Jump turned out to be 13 times faster than the Cray system.

Greetings were presented by Dr. Hermann Schunck from the German Federal Ministry of Education and Research, and by Ms. Dr. Waltraud Kreutz-Gers from the Ministry of Science of North-Rhine Westphalia. Both emphasized the importance of the new supercomputer for the scientific community in Germany and the necessity to have compute power of this magnitude to be competitive in world-class science. Francis Kuhlen of IBM Europe praised the good cooperation with the Forschungszentrum Jülich and wished good luck with the new system.

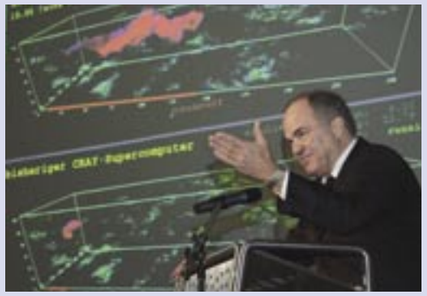


Figure 2:
Prof. Treusch during the press conference, in the background the visualised contest between the new and the old supercomputer
(Photo: Forschungszentrum Jülich)

In the afternoon of the same day, a scientific colloquium took place, which also commemorated the 100th birthday of John von Neumann. Over 300 guests accepted the invitation to this event.

The first speaker of the colloquium was Prof. Walter Kaiser, science historian of RWTH Aachen University. He presented the biography of John von Neumann, who was born on December 28, 1903. John von Neumann is widely accepted as "Father" of Scientific Computing. His research on very different scientific fields, reaching from mathematics to quantum mechanics, meteorology, ballistics, and economics (Game Theory), was always connected with elaborate computations. The development of a universal computer architecture, known as von Neumann architecture, was his outstanding contribution to the Computer Sciences. The John von Neumann Institute for Computing (NIC) is named after him to honour these achievements.

Afterwards, the president of the Hungarian John von Neumann Computer Society, Dr. Peter Bakonyi, gave his greetings and unveiled a bronze bust of John von Neumann. This bust, a work of the Hungary-born sculptor Gabriella Bollobas, has been placed in the foyer of the new technical building of ZAM, close to the new supercomputer.

In his talk, the new Director of ZAM, Priv.-Doz. Dr. Dr. Thomas Lippert, showed the perspectives of high-performance computing and, also, gave an outlook on the challenges of quantum computing. The colloquium was finished with an overview on 15 years of NIC, given by the director of the NIC Scientific Council, Prof. Dietrich Wolf. After that, the audience was invited to have a look at the new supercomputer and to renew acquaintances during the reception.

The high participation in the event shows the esteem of the new supercomputer by the German scientific community. The users are convinced by the performance of Jump: the system load reached 80 % already on the first day.

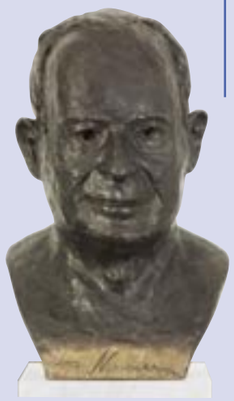


Figure 3:
Bronze bust of John von Neumann

• Dr. Sabine Höfler-Thierfeldt

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

The Next Generation Supercomputer at HLRS

The three German federal high-performance computing centres have agreed on a spiral of innovation in software systems. In 2004 NIC has started services on its new IBM system (see page 18) and the LRZ will start a procurement for a new system to be operational in 2006. In the year 2005 HLRS will turn the spiral one twist further.

In preparation for an upgrade of its HPC systems HLRS started a collaborative process with its user community as early as May 2000. A subcommittee of the steering committee was set up to work out with users and technical experts from HLRS a list of application requirements for a next system. The committee came up with both a collection of requirements from various communities and an excellent basis for a system specification that was the foundation for the preparation of a procurement.

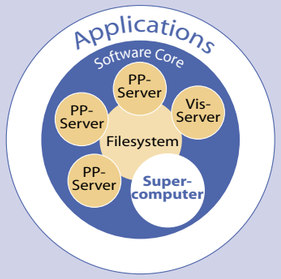


Figure 1:
HLRS Teraflop Workbench Concept

At the core of the technical specification was an overall concept that can best be described as a “Teraflop Workbench”. The pivot of the concept is a sharee file system that this allows to attach hardware for pre- and post-processing to the same file system as the supercomputer. Users work in a single data environment. Data do not have to be moved during the work flow of a simulation. The concept has to be complemented by a software core that hides away the complexity of the architecture from the user and her application (Figure 1).

The official process started in summer 2002 with a call for participation and was finished in late 2003 with a contract being signed in February 2004 at the University of Stuttgart. The key issues in the procurement on which the decision was hinging were:

- A minimum sustained performance of 2 TFlop/s for a real application picked from a set of HLRS user applications
- Balance of the architecture in terms of bandwidths, latencies and performance of individual components
- A close co-operation of the vendor with the HLRS that extends beyond the installation phase and guarantees a successful operation of the system. The key issue was support for the users in harnessing the potential of the system

Among the key questions for the HLRS was that of processor technology. Standard processors allow for a much higher peak performance for the same budget. Furthermore, users can easily migrate from their local systems to a supercomputer that is built from similar or even identical technology. On the other hand specialized processors provide a much higher sustained performance if codes can be optimized to harness their potential.

The decision was based finally on a number of benchmarks that reflect the findings of the steering committee, the user groups and the system usage on our old systems (Cray T3E, NEC SX-5). An evaluation of benchmarks showed that – in agreement with findings from experts – any code that is optimized for one type of architecture has a good chance to perform well also on the other type of architecture. It was not able that the best performing code was one that had been fine tuned over years for cache based microprocessor architectures. That code achieved highest performance on vector systems and microprocessors alike. It even outperformed codes that had been tuned for vectors over ten years.

The finding from this is that vectorization is not just a hardware architecture concept. It has to be seen as a software concept. The message HLRS has taken from this is that many codes that perform well in a cache environment – with a sustained level of performance of 15 % – can potentially perform as well on a vector system, but there with a level of sustained performance of 40-50 %.

Given the higher sustained performance of a vector processor the same level of overall performance can be achieved with a much smaller number. As scalability is something that is hardly considered by users this became one key argument for a vector based system. What can be achieved on 4000 microprocessors can also be done on 512 vector processors. And the smaller number of processors for the vector system takes away some of the burden of working on scalability.

The decision was hence made for a vector based system. A contract with NEC was signed in February 2004. The key parts of that contract are:

- NEC has installed 6 SX-6 nodes with a total performance of 540 GFlop/s in March 2004. This allows porting of codes and initial optimization during 2004
- Beginning in December 2004 NEC will start to build a complex of 64 nodes with 512 CPUs and a main memory of about 9 TByte. This will have a peak performance of about 11 TFlop/s and will achieve 4 TFlop/s for a single real application

The installation of the hardware is complemented by a co-operation of HLRS and NEC. This will define a set of applications on which to show outstanding performance on the new supercomputer.

Figure 2:
The NEC SX-6 vector node

• Prof. Dr. Michael M. Resch

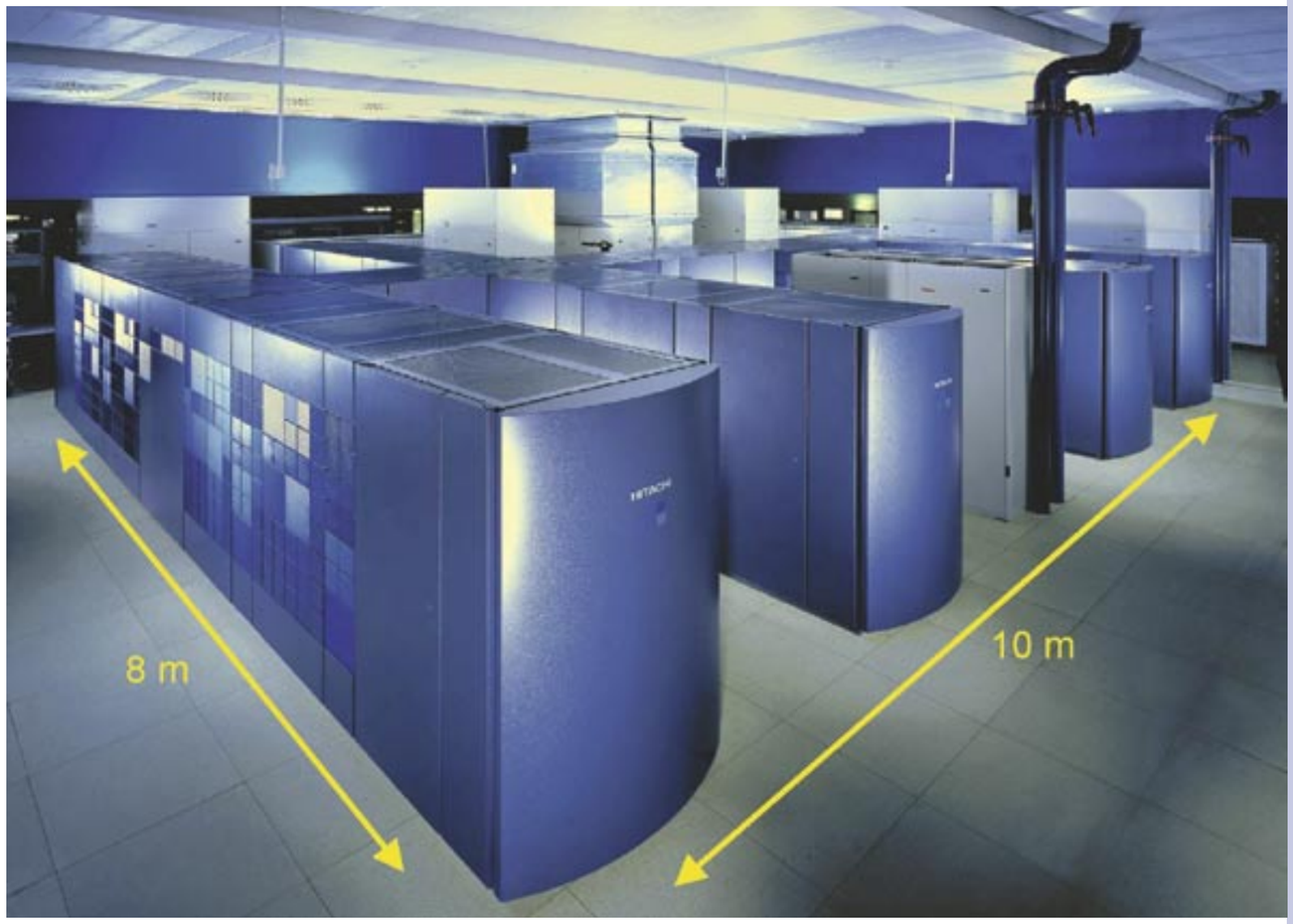
High-Performance Computing Center Stuttgart (HLRS)

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:
Leibniz-Rechenzentrum
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

System	Size	Peak Performance Performance from Memory* (GFlop/s)	Purpose	User Community
Hitachi SR8000-F1	168 8-way SMP nodes 1376 GByte memory	2016	Capability computing	German universities and research institutes
Fujitsu/ Siemens VPP700 (to be re-placed in 2005)	52 vector processors 144 GByte memory	114	Capability and capacity computing	Bavarian universities
Linux Cluster Intel IA32	147 nodes 187 IA32 processors 300 GByte memory	832	Capacity computing	Munich universities
Linux Cluster Intel IA64	17 IA64-4 way CPU 140 GByte memory	354	Capability and capacity computing	Munich universities
IBM pSeries 690 hpc	1 SMP node 8 processors POWER 4 (single core) 32 GBytes memory	42	Capacity computing	Munich universities

HLRS

Based on a long tradition in supercomputing at Stuttgart University, 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:
Prof. Dr. Michael M. Resch
Universität Stuttgart
High-Performance Computing Center Stuttgart (HLRS)
Allmandring 30
70500 Stuttgart
Germany
Phone +49 71 16 85 25 04
resch@hlrs.de
www.hlrs.de

View of the NEC SX-6 at HLRS

View of the Cray Opteron System at HLRS

Compute servers currently operated by HLRS are

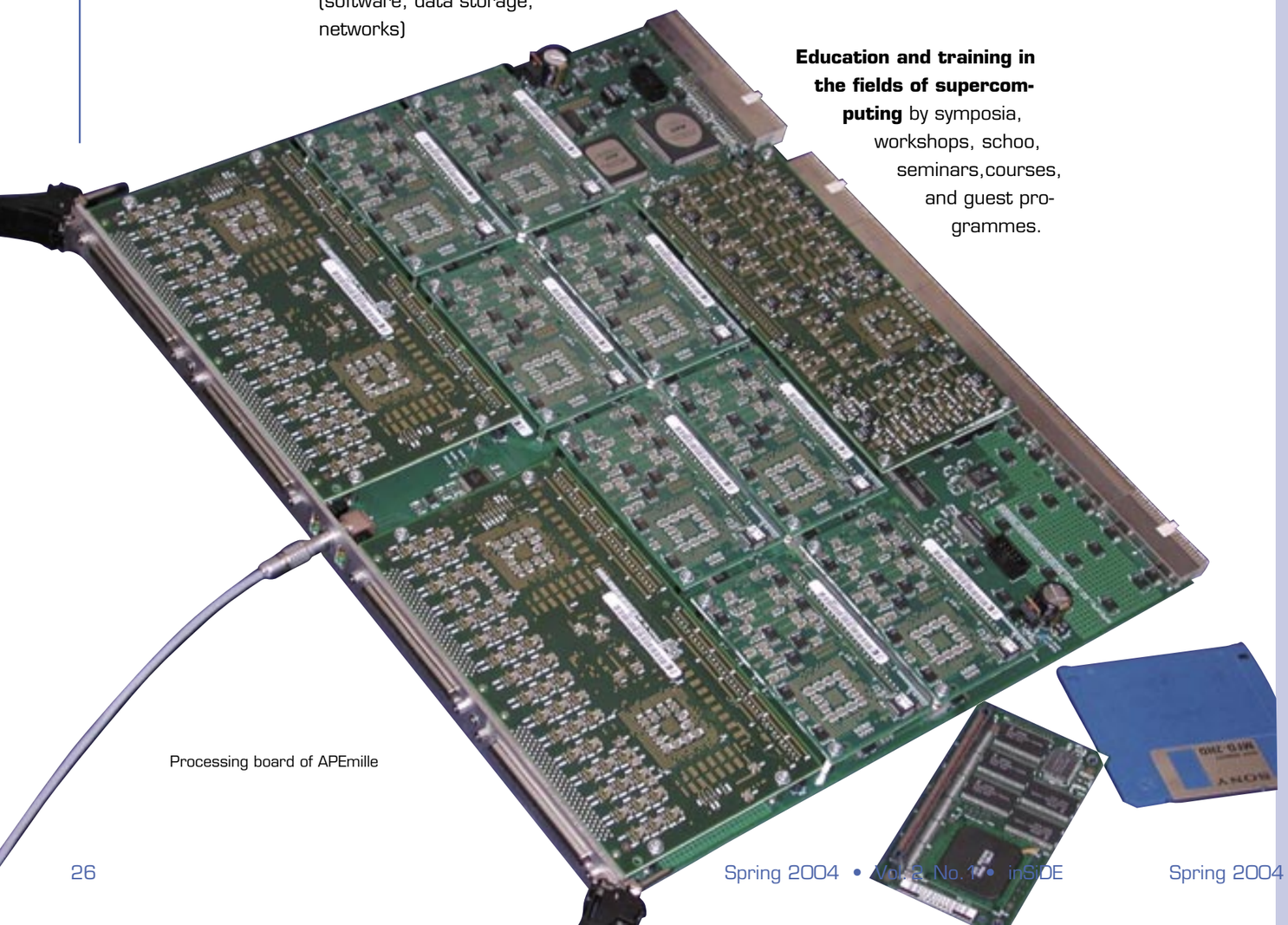
System	Size	Peak Performance (GFlop/s)	Purpose	User Community
NEC SX-6	6 8-way nodes 384 GByte memory	442	Capability computing	German universities, research institutes, and industry
Cray Opteron System	129 2-way nodes 512 GByte memory	1024	Capacity computing	German universities, research institutes, and industry
NEC TX-7	32 way node 256 GByte memory	192	Pre-processing	German universities, research institutes, and industry
IA64 Cluster	8 2-way nodes 32 GBytes memory	57,6	Capacity computing	Stuttgart University
IA32 Cluster	24 2-way nodes 48 GBytes memory	230,4	Capacity computing	Stuttgart University

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 in Germany. Its tasks are:

Nationwide 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 of 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 "Bioinformatics" is under consideration.

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



Processing board of APEmille

The following supercomputers are available for research projects of the communities mentioned below, peer-reviewed by the Resource Allocation Committee 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/ape/html/Installation/

System	Size	Peak Performance (GFlop/s)	Purpose	User Community
IBM pSeries 690 Cluster 1600	41 SMP nodes 1312 processors POWER4+ 5248 GBytes memory	9000	Capability computing	German universities, research institutes, and industry
CRAY SV 1ex	16 CPUs 32 GBytes memory	32	Capability computing	German universities, research institutes, and industry
APEmille (special purpose computers)	4 racks 1024 processors 32 GBytes memory	550	Capability computing	Lattice gauge theory groups at German universities and research institutes

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

Priv.-Doz. Dr. Dr. Thomas Lippert
Wilhelm-Jonen-Straße
52425 Jülich
Germany
Phone +49 24 61 61 64 02
th.lippert@fz-juelich.de

High-Performance Computing Courses and Tutorials

LRZ
www.lrz.de

Parallel Programming of High-Performance Computers

Date

October 4-8, 2004

Location

Leibniz Computing Center, Munich

Contents

Basic concepts of parallel programming with MPI and OpenMP:

- 1st day: Tools for program development; Processor architectures, optimization and performance analysis.
- 2nd day: Usage of HPC systems at LRZ and RRZE. Runtime optimization.
- 3rd day: MPI-1 and MPI-2 with hands-on session. Usage of VAMPIR.
- 4th day: MPI Case Studies; Introction to OpenMP with hands-on session. Debugging with Totalview.
- 5th day: Parallel optimization strategies, usage of performance libraries.

HLRS
www.hlrs.de

Parallel Programming with MPI and OpenMP

Date

August 4-6, 2004

Location

Dresden, ZHR

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 ZHR in collaboration with HLRS

Iterative Linear Solvers and Parallelization

Date

September 13-17, 2004

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, BiCGSTAB ...) as well as highly efficient preconditioning techniques are presented in the context of real life applications. 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 Uni Kassel, HLRS, IAG, and WIR.

HLRS Parallel Programming Workshop

Date

October 11-15, 2004

Location

Stuttgart, HLRS

Contents

The focus is on programming models MPI and OpenMP, domain decomposition, parallelization with PETSc, and optimization. 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)

Parallel Programming with MPI and OpenMP

Date

November 29 - December 1, 2004

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 ZAM/NIC in collaboration with HLRS

NIC
www.fz-juelich.de/nic

User Course "Programming and Usage of the System IBM pSeries 690 Cluster 1600" Common Presentation of ZAM and IBM

Date

July 8 - 9, 2004

Location

NIC/ZAM, Research Centre Jülich

User Course "Introduction to Parallel Programming with MPI and OpenMP"

Date

August 3 - 6, 2004

Location

NIC/ZAM, Research Centre Jülich

User Course "Efficient Programming in C++"

Date

August 16 - 20, 2004

Location

NIC/ZAM, Research Centre Jülich

NIC/ZAM Guest Student Program 2004 Education in Scientific Computing

Date

August 2 - October 8, 2004

Location

Research Centre Jülich

Colloquium on Parallel Processing Principal Topic: Applications in Engineering

Date

October 11, 2004

Location

RWTH Aachen, Computer Centre

Parallel Programming with MPI and OpenMP

Date:

November 29 - December 1, 2004

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 ZAM/NIC in collaboration with HLRS

Miscellany

HLRS Personnel Matters

Dr. Ulrich Lang, deputy director of HLRS and head of department for visualization of the HLRS has accepted a position as full professor for computer science and director of the computing center at the University of Cologne.

Others

November 2003: HLRS and Indiana University together with its international partners won the HPC Challenge at Supercomputing 2004 Phoenix/AZ for bringing together 28 supercomputers at 22 locations on five continents and running a DNA analysis of arthropods.

March 2004: HLRS continues its partnership with Cray. Both sides have agreed to replace the old Cray T3E-512/900 by an Opteron based system. The system has 256 AMD Opteron CPUs linked by Myrinet with a peak performance of 1 TFlop/s.

May 2004: HLRS has presented results of its co-operation with Microsoft at this year's Windows Hardware Engineering Conference (WinHEC) 2004 at Seattle (May 4-7, 2004). In his key note address Bill Gates invited Uwe Wössner to give a 3D live demo of simulation and visualization of a water turbine based on results of the Institute of Fluid Mechanics and Hydraulic Machinery of the University of Stuttgart.

Report on NIC Winter School 2004

From February 29th to March 6th, 2004 NIC/ZAM presented a winter school on the topic "Computational Soft Matter: From Synthetic Polymers to Proteins" at the Gustav-Stresemann-Institut (GSI) in Bonn. The program aimed at graduate students, postdocs and young researchers in the areas of theoretical/computational physics, chemistry and biophysics. It was prepared by Norbert Attig (NIC/ZAM, Jülich), Kurt Binder (University of Mainz), Helmut Grubmüller (MPI for Biophysical Chemistry, Göttingen) and Kurt Kremer (MPI for Polymer Research, Mainz). The lectures were presented by 15 outstanding scientists from German and international universities and research institutes, who work at the forefront of computational sciences.

Interested young scientists had to submit an application for participation together with a letter of recommendation and a short summary of their field of interest. From these applications 80 participants were selected by their scientific background and excellence of their proposal.

The interdisciplinary program was very attractive and the lectures were presented on a high level. They were complemented by more than 40 poster contributions of the participants, demonstrating their expectation to play an active role in this scientific school. The program also included an excursion to the Research Centre Jülich, where additional presentations on soft matter research in Jülich, on supercomputing at NIC and on parallel programming models and software tools were given.

Of course, also the new IBM supercomputer and the associated new machine room were introduced by a guided tour.

The lecture notes of the review talks and the poster abstracts were published in NIC's own publication series. Online versions are available under http://www.fz-juelich.de/nic-series

New Director at NIC/ZAM

Since January 1st, 2004, NIC/ZAM has a new director. Priv.-Doz. Dr. Dr. Thomas Lippert, born in 1960, accepted an offer of the Research Centre Jülich. Lippert studied physics, in particular theoretical elementary particle physics in Würzburg and Wuppertal, where he received his PhD in 1993 under the supervision of Prof. Klaus Schilling. He received a second PhD from the University of Groningen in 1998, where he worked in computer science together with Prof. Nicolay Petkov. In November 2001 he finished his Habilitation at the University of Wuppertal. His scientific interests include lattice gauge theory, numerical and parallel algorithms as well as cluster computing.

Survey of Computational Physics

In May 2004, Physik Journal 3 (2004), Nr. 5, published six review style articles on Computational Physics by leading German scientists, which refer also to the usage of the German national supercomputing centres. Thomas Lippert has contributed to the series with an overview of his own reasearch field, lattice quantum chromodynamics.

inSiDE

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

Publishers

Prof. Dr. Heinz-Gerd Hegering, LRZ
Dr. Rüdiger Esser, NIC
Prof. Dr. Michael M. Resch, HLRS

Editor

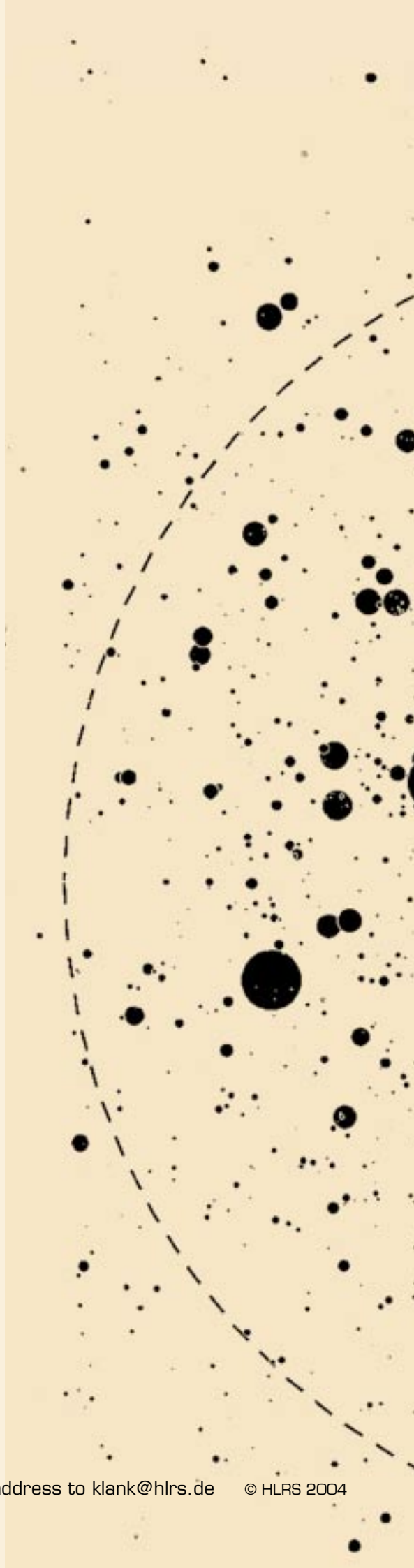
F. Rainer Klank, HLRS/RUS
klank@hlrs.de

Design

Katharina Schlatterer
kschlatterer@rus.uni-stuttgart.de

Authors

George Bosilca
gborsilca@utk.edu
Dr. Robert Cameron
cameron@linmpi.mpg.de
Prof. Jack J. Dongarra
dongarra@utk.edu
Dr. Graham E. Fagg
fagg@cs.utk.ed
Dr. Irmgard Frank
frank@cup.uni-muenchen.de
Dr. Edgar Gabriel
egabriel@cs.utk.edu
Dr. Sabine Höfler-Thierfeldt
s.hoefler-thierfeldt@fz-juelich.de
Prof. Dr. Michael Resch
resch@hlrs.de
Manfred Schüssler
msch@linmpi.mpg.de
Dr. Alexander Vögler
voegler@linmpi.mpg.de
Dr. Wolfgang Wenzel
wenzel@int.fzk.de



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