# inSiDE

inSiDE • Vol. 2 No. 2 • Autumn 2004

## Innovatives Supercomputing in Deutschland

# Editorial

While preparing this fourth issue a number of steps were taken in German supercomputing. Operation of the most recent system at NIC was well established. The reader will see a number of applications described in this issue that make good use of this fastest European high performance computing systems. The spiral of innovation is kept running with the two other centers actively pursuing their technical refresh programs. HLRS has taken its first steps to install a vector supercomputer at Stuttgart. LRZ recently got approval of its plans for a next generation supercomputer system and has already started the process of procurement with an invitation of tenders.

The first section of inSiDE again introduces a variety of application reports. It is complemented by a short report on an MPI-debugger. This is a first step towards bringing to our users a different facet of German Supercomputing. Although Germany does not have a strong hardware industry it does have some rather interesting and successful activities in software development. Hence, one focus of inSiDE in the future will be on presenting these efforts and results.

German high performance computing is part of the European scene and hence all the three centres participate in European project activities. To show these strong German-European integration efforts we present three European Projects. DEISA is a project that aims to create a European supercomputer infrastructure backbone. On the networking side such a backbone already exists for years. The European high speed research network GEANT is briefly introduced in our project section. Despite growing network speed face to face communication is still crucial in many fields. HPC-Europa is aiming at bringing European scientists to European centres to foster co-operation between users and centre as well as supporting a European cross-fertilization inside the HPC-community.

All of the three German federal supercomputing centres are part of wider co-operations. These reflect the integration of each individual centres into its regional or topical community. In this issue we introduce these partnerships. LRZ for a long time already has established a strong co-operation among Bavarian researchers that form a network of competence for high performance computing. KONWIHR has been the role model for other institutions for long and is in its second phase of funding recently. It will certainly play an important role for the coming hardware update in Munich. NIC is a co- operation between DESY and the centre in Jülich. We present an application driven approach for collaboration in this issue. The last such network of competence was introduced 2003 when HLRS teamed up with the University of Karlsruhe in the hkz-bw to strengthen its service and development area. By extending the cooperation to Heidelberg a true network of competence is going to evolve also in Baden-Württemberg.

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





# Contents



**inSiDE**

# Nanomachine – Ion Transport by Membrane Proteins: The Molecular Mechanism

The early vision of Richard Feynman, which was widely popularized by Eric Drexler's evocative idea of a self-replicating assembler building nanoscale devices ("nanomachines") atom by atom, is nowadays coming close to reality: it is the morphogenesis and action of heterogeneous multiprotein complexes which transport material, generate energy and transduce signals to maintain the proper self-organization of cells and organisms.

Although many biomolecular nanomachines have been identified, the relations between their structural and dynamical properties ("function") are still poorly understood, mainly because molecular details are unknown. Recent breakthroughs in determination of crystal structures at the atomic resolution, however, have provided detailed insights into the molecular architecture of simple nanomachines. This is, by and large, seen in recent successful structural determination of the bacterial potassium ion channel KcsA and the water-conducting aquaporin (Nobel Prize 2003).

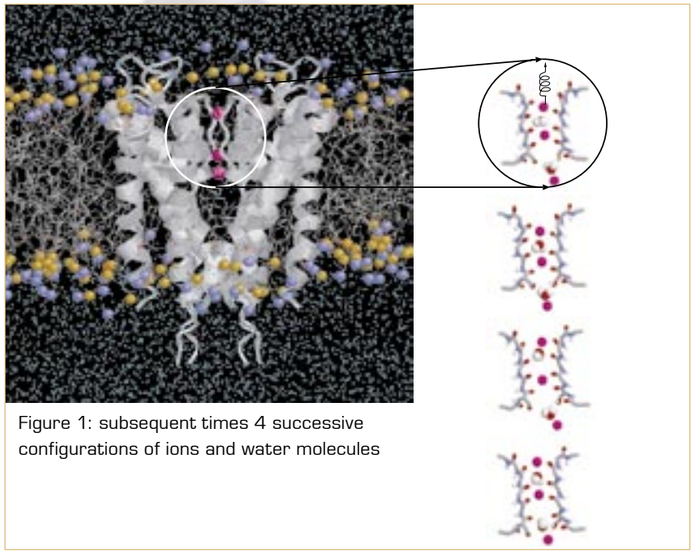


Figure 1: subsequent times 4 successive configurations of ions and water molecules

Ion channels reside in cell membranes and are an important class of biomolecular machines. They work by selecting specific ions and by catalyzing the passive diffusion through the selectivity filter.

In order to relate their structural details to their dynamic function, molecular dynamics simulation on supercomputers, like the IBM p690 cluster in Jülich, has been shown to be an appropriate tool to obtain insight into the molecular mechanisms of ion selectivity and transport.

The first picture shows the potassium channel KcsA residing inside a lipid bilayer membrane surrounded by water molecules. Molecular dynamics simulations were conducted of this model system consisting of about 30,000 atoms. The simulations were performed with the AMBER program, which uses classical forces.

The simulations permit to monitor the collective motion of ions and water molecules through the narrow selectivity filter, which is the encircled region in Figure 1. The simulations reveal that the high conductivity is based on the cooperative diffusion of ions and water molecules mediated by the charged flexible carbonyl groups lining the selectivity filter which are shown in the figures at the right hand side of Figure 1 by the red tips at the protein's backbone. The right hand side of Figure 1 shows at subsequent times 4 successive configurations of ions and water molecules passing through the selectivity filter.

The very remarkable property of this narrow filter is its high conductivity: once the outer ion is pulled out of the filter, all other ions and water molecules follow the same direction, apparently at almost no cost of energy. This becomes more obvious during the transport through very long artificial nanotubes constructed in analogy to the selectivity filter of this ion channel, which is shown in Figure 2.

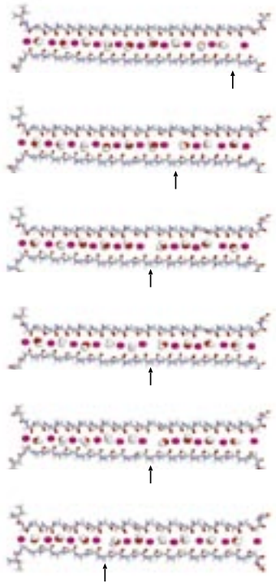


Figure 2: Very long artificial nanotubes

From top to the bottom this figure shows at subsequent times, in steps of 50 picoseconds, six successive configurations of ions and water molecules passing through the nanotube.

It is observed that the transport occurs by the hopping of vacancies from the right to the left, as indicated by the arrows. It is fascinating to speculate about a possible construction of such a kind of very efficient conducting device on the nanometer scale under experimental conditions in the laboratory.

In fact, the computer simulations of the ion conduction in such a nanotube reveal the same effects as observed for the ion channel. Therefore, since the molecular mechanism of the ion-water transport in ion channels and corresponding nanotubes is based on general physical principles, the efficiency of this type of linear organic ionic conductor has to be compared to known solid state superionic conductors. It would not be a surprise if such biomolecular conductors would be superior to the anorganic one.

In conclusion, by following the visions of Feynman and Drexler, it seems conceivable that future inventions of nanotechnological devices may be constructed based on principles copied from biomolecular machines. The discovery of the molecular transport mechanism in ion channels may currently serve only as a paradigm pointing towards a future challenge: the understanding of major biomolecular machines such as rotary ATPases providing, e.g., the necessary torque for the bacterial flagellar motor, or, translocases which are transmembrane conducting devices for protein transport.

In any case, the knowledge of such biomolecular mechanisms, which can be considered to serve as a bridge between nanotechnology and biology, cannot be achieved without the indispensible tool of massive computer simulations.

• Jean-Fang Gwan
• Artur Baumgaertner

Institut für Festkörperforschung (IFF), Forschungszentrum Jülich GmbH





Applications

# Plasma Turbulence and Fusion Energy

Applications

Magnetic fusion research aims at exploring the physical processes underlying the design and construction of fusion power plants. Inspired by the way the sun produces its vast amounts of energy, the idea is to heat up a plasma of deuterium and tritium to some 100 million degrees, keeping it confined by a toroidal magnetic field. Since the charged plasma particles basically follow the magnetic field lines, they can thus be kept away from the reactor boundary walls. For sufficiently large plasma pressure and energy confinement time, the rate of fusion reactions is large enough to both heat the plasma and allow energy to be extracted from the system.

However, small-scale plasma waves can be unstable in the presence of strong density and/or temperature gradients. The exponential growth of the unstable modes is limited by nonlinear effects which lead to an energy transfer between various degrees of freedom. Eventually, the system reaches a quasi-steady state far from thermodynamic equilibrium. This kind of turbulent dynamics is in turn associated with significant particle and heat transport. Imperfect magnetic confinement is in fact one of the main obstacles on the way to a future fusion reactor. High loss rates can only be compensated for by building larger and thus more expensive devices.

Since turbulence is an inherently nonlinear phenomenon involving a large range of space and time scales, its direct numerical simulation (DNS) provides a challenge even for the world's largest supercomputers. And plasma turbulence is even more complex and

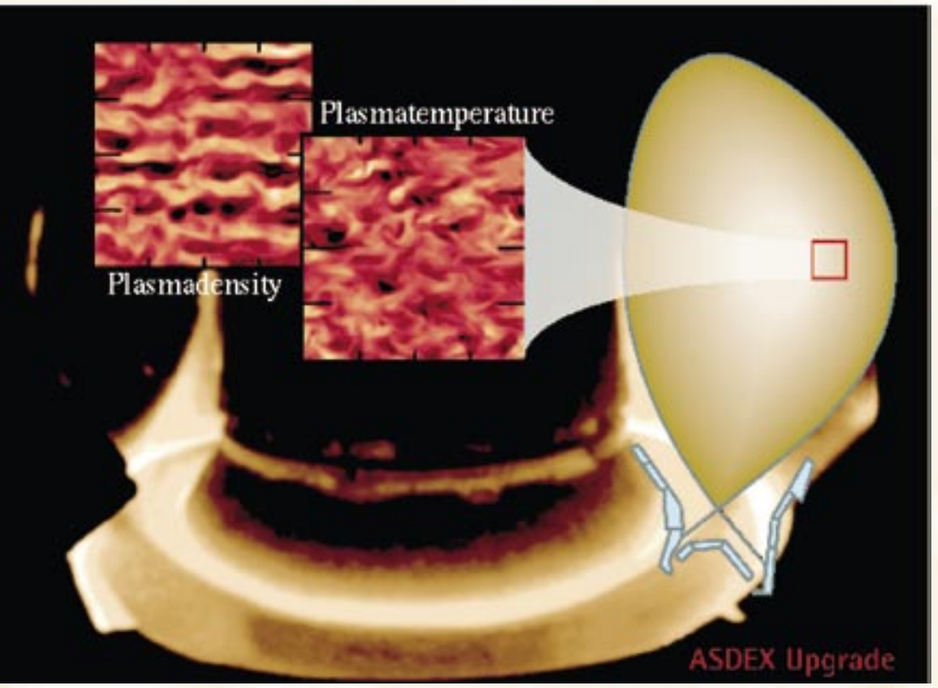



Figure 1:
Direct numerical simulations of turbulent fluctuations in the fusion experiment ASDEX upgrade

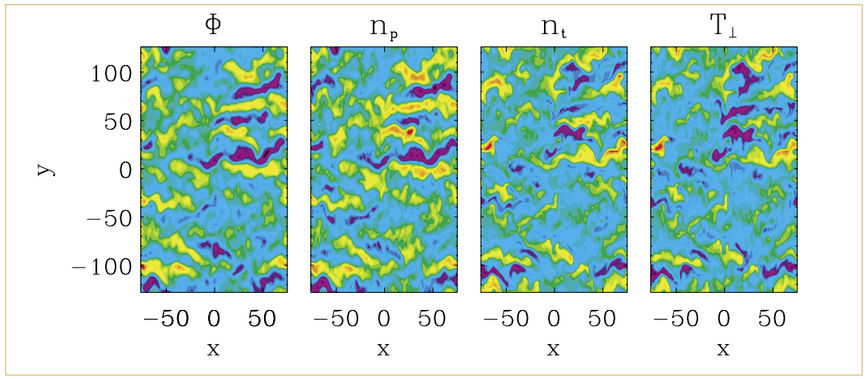



Figure 2:
Turbulent fluctuations of various plasma parameters:
A snapshot in a plane perpendicular to the magnetic field

computationally demanding than ordinary fluid turbulence. Whereas the latter is described by the Navier-Stokes equations of fluid dynamics, magnetically confined fusion plasmas are generally characterized by long mean free paths, thus violating some of the assumptions underlying standard fluid theory. This means that first-principles simulations need to be based on a gyrokinetic treatment of multiple particle species in 5-dimensional phase space. (The fast gyromotion around magnetic field lines can be removed from the basic equations analytically, reducing the number of dimensions from six to five. This motivates the label "gyrokinetic".) The DNS of plasma turbulence therefore requires a mix of efficient algorithms, careful implementations, and state-of-the-art hardware such as the Hitachi SR8000-F1 at LRZ.

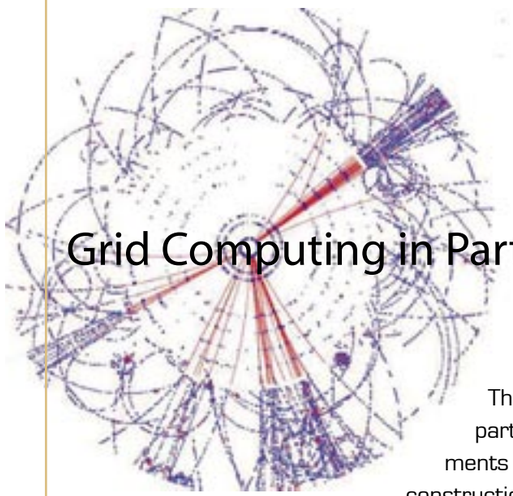
For many years, the focus of plasma turbulence research had been on length and time scales determined by the ion's dynamics. The lighter and more mobile electrons, on the other hand, have been described simply via a Boltzmann distribution. In this limit, only heat transport by the ions is allowed, whereas the particle and electron heat transport vanish exactly. The spatio-temporal scales of the electrons are smaller than those of the ions by a factor of the square root of the ion-to-electron mass ratio, i.e., by a factor of about 60 for deuterium plasmas. Despite these disparate scales, retaining the electron dynamics has a large effect on the particle and heat fluxes. It is probably needless to say that the simulations required here represent an extremely demanding computational task, pushing the employed software and hardware to their limits.

One central issue in the study of plasma turbulence is the role of structure formation. Instead of being more or less isotropic in the plane perpendicular to the magnetic field lines, turbulent eddies may be strongly elongated in the direction of the background gradients. In the presence of such "streamers", the cross-field transport may thus be significantly enhanced. Conventional estimates of turbulent transport coefficients based on Ludwig Prandtl's mixing length concept may fail to capture these surprising findings. It could be shown, however, that a secondary instability theory is able to predict the simulation results even semi-quantitatively. This may be regarded as a nice example of the interplay between simulation and theory in the realm of turbulence research. The computations have been performed using the GENE code (F. Jenko and T. Dannert, MPI für Plasmaphysik, Garching).

• Frank Jenko

Max-Planck-Institut für Plasmaphysik (IPP), Garching

Applications



# Grid Computing in Particle Physics

The next generation of particle physics experiments are currently under construction at the European Center for Nuclear Research (CERN) in Geneva. Four experiments will record data from particle collisions provided by the Large Hadron Collider (LHC). A topic of fundamental scientific interest is the search for the Higgs Boson, which - if it exists - should be produced in the proton-proton collisions at the LHC. The challenge, however, is the identification of such a reaction, since it is expected typically just once in $10^{12}$ collisions. Huge detectors are needed to measure and identify the particles produced in the collisions and many PetaBytes of data must be recorded and analyzed in order to answer questions such as the existence of the Higgs Boson or to search for other new particles.

The ATLAS collaboration, a huge international cooperation of about 1800 physicists from 34 countries, is building one of the LHC experiments. There is a strong German participation in ATLAS with 10 institutes and around 150 physicists involved in various activities ranging from construction through computing and physics analysis. When the LHC facility will be commissioned in 2007 the ATLAS experiment will produce around 2 PetaByte of data per year with an additional 1 PetaByte of simulated data per year. This huge data volume makes it impossible for all the processing and analysis to be performed at a single facility. The workload must be distributed to many regional computing centers in the participating countries. The automatic management of these data sets and the effective workload distribution pose non-trivial problems. The concepts of Grid computing developed in recent years address exactly these problems and computing in particle physics provides an ideal test case for large-scale Grid deployment.

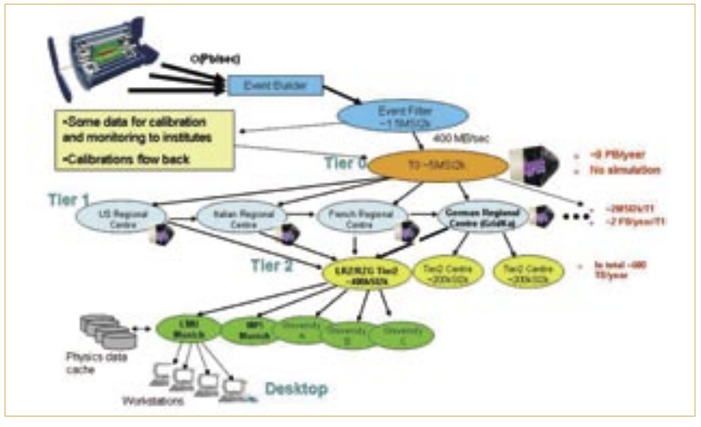



Figure 1: The Tier structure of the computing m odel

Figure 1 gives a schematic view of the Model of computing resources according to ATLAS. The model is broken down according to the Tier structure, Tier-0 represents the CERN facility while Tier-1 represents the national computing centers such as GridKa. The Tier-2 resources are to be formed from several computing centers associated to each Tier-1 center.

To aid with the development of this distributed computing infrastructure a series of test productions started in 2002. These allow us to evaluate the Computing Model, the ATLAS software suite, the Data Model and also to ensure the correctness of technical choices made for ATLAS computing. In 2004 the second large-scale test production started with one of its major aims being the use of Grid tools to their maximum extent and the assessment of their ability to process events sets on a LHC scale. The goal is to fully simulate 10 million physics events from various processes by using computational resources in approximately 100 institutes from all over the world.

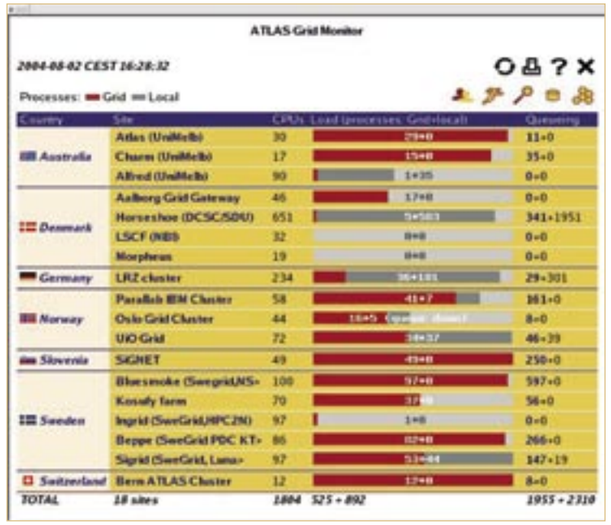



Figure 2: Snapshot of ATLAS test Production in NorduGrid

The LRZ facility in Munich is currently being used within this production as part of a pilot project involving members of LMU Munich and LRZ. At LRZ we have deployed the ARC Grid middleware, which is based on the Globus toolkit, and are contributing resources to the NorduGrid project. Figure 2 shows the web based monitoring tool for the NorduGrid resources. The ARC middleware has been adapted for use on a large variety of Linux flavors and batch systems. A particular advantage of ARC is the transparent installation on a single front-node server only; the compute nodes remain untouched. For these reasons the ARC software is well suited for installation at an already operating computer facility such as LRZ which provides services to several heterogeneous user communities and which cannot be dedicated to only one task.

For the production the full dataset is split into 100,000 jobs, each with a typical duration of 30 hours. The execution of these jobs in NorduGrid proceeds as follows:

- The NorduGrid executor contacts a central database to retrieve the information needed for the next bunch of jobs (typically 1000).
- The jobs are distributed to the sites registered in NorduGrid. The ARC resource broker takes care of a proper distribution of jobs to sites which match the requirements in terms of CPU, memory, pre-installed software, etc.
- Before a job starts on the local sites any input files needed are downloaded using Grid copy-tools as specified in the job description.
- Jobs are submitted to the local batch system (SGE in case of LRZ). During execution access to remote databases is needed.
- After the job has finished the data files produced are uploaded to a central NorduGrid storage site. The files are registered in a Grid replica catalog and made available worldwide for further processing steps.

The goal of this pilot production at LRZ is to process about 1000 jobs in order to gain experience for more continuous and larger scale productions in the future.

• Günter Duckeck

Section Physik,
LMU München

# Is there still a Need for Tailored HPC Systems in CFD or can we go with Commodity Off-The-Shelf Clusters?

Looking at the TOP500 lists of recent years, more and more tailored HPC systems have been replaced by commodity off-the-shelf (COTS) clusters which do not mainly focus on HPC requirements, but dominate the HPC market due to their (often) moderate price-performance ratio. Tailored HPC systems, in particular vector computers, have been designed to meet the numerical requirements of scientific, memory-intensive applications. Combining high sustained single processor floating point performance, high memory bandwidth and a balanced interconnection network with a mature software environment (compilers and libraries) are characteristics for the latter systems. However, owing to the low number of sold systems, they are rather expensive.

Based on two scenarios from computational fluid dynamics we will discuss – using a lattice Boltzmann flow solver – the scalability as well as the pros and cons of tailored HPC systems and COTS clusters in terms of the price-performance ratio for these types of applications.

## Introduction

Rapid advances in microprocessor technology have led to fundamental changes in the HPC market over the past decade. Commodity "off-the-shelf" (COTS) cache-based microprocessors arranged as systems of interconnected SMP nodes dominate today's TOP500 list [1] owing to their unmatched price/peak performance ratio. However, it has also been acknowledged recently that the gap between sustained and peak performance for scientific applications on COTS platforms is growing continuously [2]. Although classical vector systems can bridge this performance gap especially for memory intensive codes, they only represent a tiny fraction (3.5 %) of all current TOP500 systems [1]. The combination of high development costs and a limited market volume for HPC systems should mainly account for this trend. The manufacturing of today's single-chip vector CPUs can be carried out with the same techniques as used for commodity components, posing no further difficulty. Some authors also speculate that the ASCI program [3] has put very heavy emphasis on the use of COTS components [4].

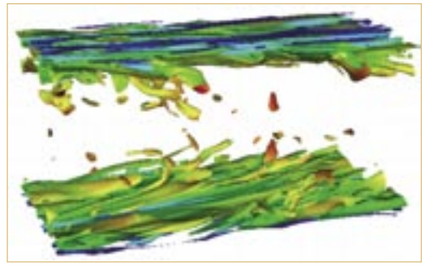
Consequently, only one manufacturer of classical vector processors has survived which, however, set a landmark with the installation of the Earth Simulator using NEC SX-6 vector technology – a "quarter Earth Simulator" will be available at HLRS from the beginning of next year. A new class of vector computers has recently been introduced with the Cray X1 which already demonstrated high sustained performance for several (vectorizable) applications [5]. It is therefore the aim of this report to discuss the performance of tailored HPC and COTS systems for two typical application scenarios and to comment on the resulting price-performance ratio.

## Application Scenarios

The two scenarios arise from computational fluid dynamics (CFD). In the first case, a given problem with fixed size should be computed as fast as possible on a parallel computer. Time-to-solution is a critical point for applications e.g. from engineering (example in Figure 1, top). In the second case, the problem size is not fixed a priori but will be scaled with the available computational power, i.e. the number of CPUs used. For the latter case, basic turbulence research can serve as an example (see Figure 1, right).

In terms of computer science, we denote these scenarios speed-up (strong scaling, fixed total size) and scale-up (weak scaling, fixed problem size per CPU).

Owing to its high scientific potential for large scale applications, we have chosen the lattice Boltzmann method (LBM) [6] for our study. This is a recent method from the area of CFD which is characterised by algorithmic simplicity owing to the explicit nature of the algorithm and equidistant Cartesian grids.



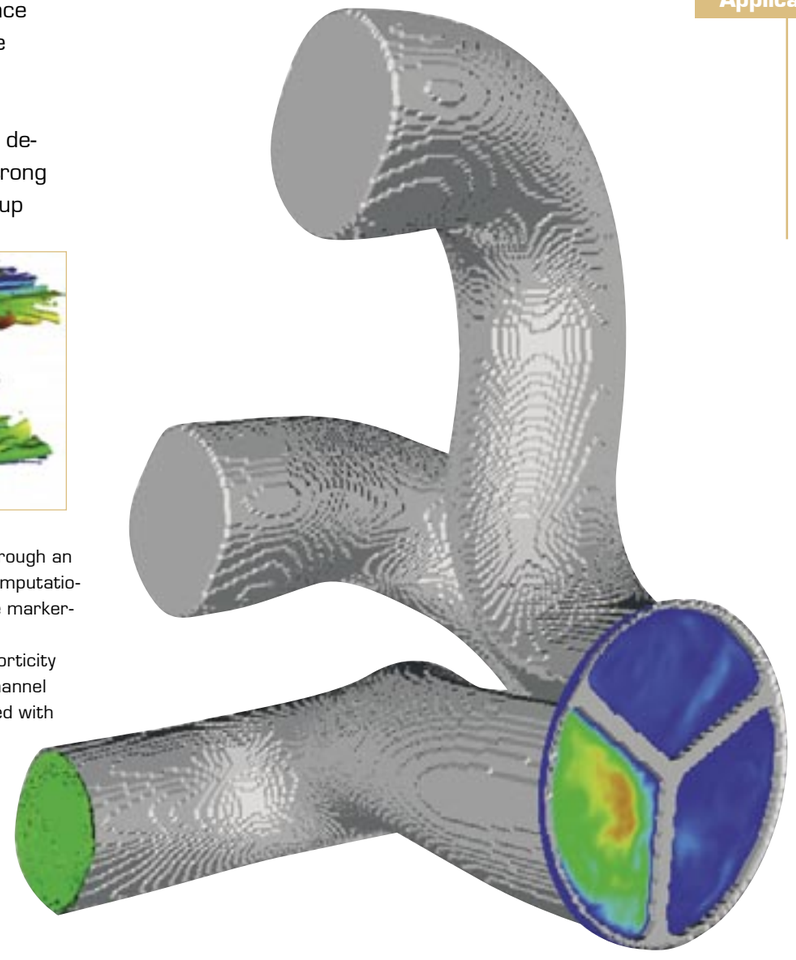


Figure 1:
(top) Large-eddy simulation of the flow through an exhaust system at Re = 300000. The computational grid is obtained from CAD data by the marker-and-cell approach.
(right) Snapshot of an isosurface of the vorticity amplitude of developed turbulence in a channel ($Re_\tau$ =180; see [10] for details). Calculated with up to 300 million voxels on the Hitachi SR8000-F1 at LRZ

## Performance results, scalability and cost effectiveness

For our survey, we have chosen two flavours of COTS clusters, one tailored HPC system with commodity CPUs and a traditional vector computer: a GBit/Xeon cluster, a Myrinet/Opteron cluster, an SGI Altix 3000 system and finally up to six NEC SX-6+ nodes. Details of the configurations can be found in Table 1.

| Architecture | 1 CPU | 2 CPUs |
|---|---|---|
| Intel Xeon | 1.9 | 3.0 |
| AMD Opteron | 2.8 | 5.7 |
| Intel Itanium2 | 5.0 | 7.2 |
| NEC SX-6+ | 38 | 74 |

Table 2:
Speed-up performance within a 2-way node given in million lattice site updates per second (MLup/s). 5 MLup/s roughly correspond to 1 GFlop/s. On the Itanium2 Altix system the two CPUs chosen for this measurement share one path to the memory

| System | GBit/Xeon | Myrinet/Opteron | SGI Altix 3000 | NEC SX-6+ |
|---|---|---|---|---|
| Basic building block | 2-way SMP node with 1 memory path | 2-way SMP node with 2 memory paths | 4-way SMP node with 2 memory paths | 8-way SMP node |
| CPU | Intel Xeon 2.66 GHz | AMD Opteron 2.0 GHz | Intel Itanium2 1.3 GHz, 3 MB L3 | 565 MHz |
| Peak performance per CPU | 5.3 GFlop/s | 4.0 GFlop/s | 5.2 Gflop/s | 9 GFlop/s |
| Memory bandwith per building block | 4.3 GByte/s | 2x 5.4 GByte/s | 2x 6.4 Gbyte/s | 8x 36.0 GByte/s |
| Interconnect | Cisco 4503 GBit Ethernet switch | Myrinet2000 | SGI NUMALink3 2x 1.6 GB/s bidiret. | NEC IXS crossbar 8 GB/s bidirectional |
| Operating system | Debian Linux 3.0 | SuSE SLES 8 Linux | Redhat AS2.1 +SGI Propack 2.4 | SUPER-UX |
| Compiler | Intel ifc 7.1 | PGI 5.0 | Intel efc 7.1 | Native NEC SX |

Table 1: Details of the platforms investigated and compilers used

An important difference between the Intel x86 and AMD x86_64 designs is the memory subsystem. While Intel still promotes (also for its new x86 EM64T architecture) bus based systems where two or four processors share one path to main memory, AMD uses a separate memory interface for each CPU providing full bandwidth within a shared memory node. Of course, the AMD design is favourable for memory bound applications concerning the scalability within the used dual processor nodes. As shown in Table 2, we find a nearly linear speed-up within the AMD node, reducing the single processor performance gap substantially compared to the Itanium2 processor, if both processors on the dual nodes are used.

Since the Opteron node provides higher aggregate bandwidth (10.8 GByte/s) – which is the critical resource for memory bound programs like LBM applications – than the Itanium 2 systems (6.4 GByte/s) there should be some room for improvement by further optimizations or advances in compiler technology.

Discussing the impact of the cluster interconnects on the performance we have to distinguish speed-up and scale-up performance: In the case of scale-up we find an almost linear performance increase with the number of compute nodes involved. If we scale in units of dual nodes (which are the basic building blocks of our clusters) we find parallel efficiencies on 64 processors beyond 80% both for GBit and Myrinet.

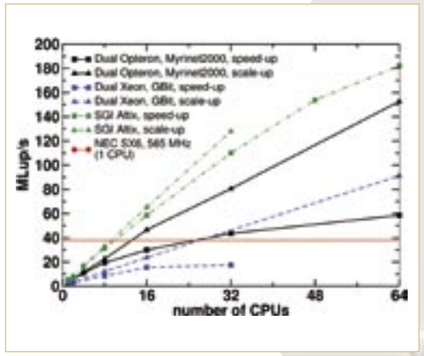



Figure 2: Scalability tests for modern cluster configurations. The domain size is 256x129x128 for speed-up and $128^3$ per processor for scale-up tests. For reference the corresponding results of a shared memory system (SGI Altix) and the single processor performance of the NEC SX-6 are given. MLup/s is the achieved number of lattice site updates per second

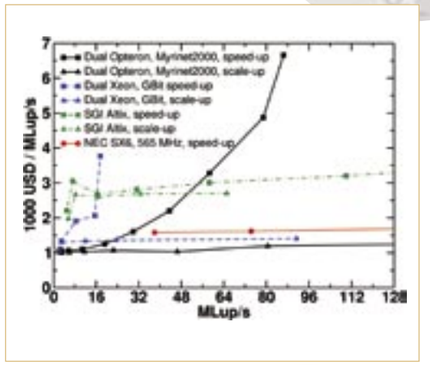



Figure 3: Costs per MLup/s to achieve a certain performance level. The domain size is 256x129x128 for speed-up and $128^3$ per processor for scale-up tests

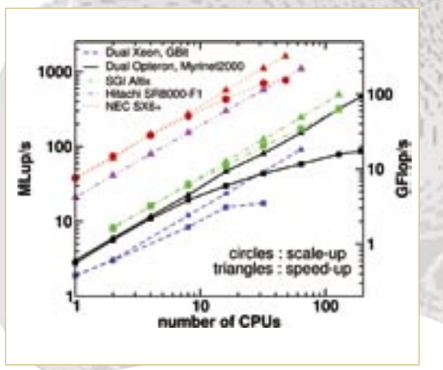



Figure 4: Comparison of the sustained performance with increasing processor count on modern cluster configurations and tailored systems. The domain size is 256x129x128 for speed-up and $128^3$ per processor for the scale-up.

The recent NEC SX-6+ only slightly outperforms the almost five years of Hitachi SR8000-F1 (see [8,9] for technical details). In the latter case, the application sees all CPUs of a node as one "virtual" CPU. Therefore the number of "virtual" CPUs (i.e. nodes) is used as metric.

The microprocessor based systems are at least one order of magnitude behind. However, it is remarkably that the performance of the SGI Altix does not degrade significantly in the speed-up case



Applications

As long as the network capacity is scaled up with the cluster size, e.g., with a fat-tree topology, we do not expect a significant breakdown for the scale-up case.

However, for the speed-up problem we recover the basic problems of cluster systems. With increasing processor count, the computational effort per processor decreases and network bandwidth and latency can impose severe limits on the achievable performance for a fixed problem to be solved. For example, on the GBit cluster reasonable scaling is not possible beyond 16 processors. Moreover, the cluster will never achieve the performance level of a single NEC vector processor for the problem size under consideration. Due to improved latencies and bandwidth, the Myrinet cluster scales significantly better but still requires roughly 24 processors to match the NEC single processor performance. The NUMALink3 interconnect of the SGI Altix provides a very high level of interconnect performance (roughly 5 times higher than the MPI bandwidth and 1/5 of its latency compared to Myrinet 2000) and thus scales very well even for the speed-up case with a parallel efficiency of roughly 80% on 64 processors (using the two processor performance as the base, ignoring the bandwidth problem within the 2-way nodes as shown in Table 2). One reason for the success of COTS clusters is their immense aggregate peak performance and high LINPACK numbers at a very moderate price level. However there is a controversial debate whether the price performance ratio of COTS clusters is still optimal if application performance is the metric. Using the performance data of Figure 2 and price estimations based on public procurements of HPC systems [7] we present a price/performance comparison in Figure 3. To achieve a high application performance, architectures like dual Xeon or dual Opteron clusters become expensive compared to vector CPUs or the SGI Altix when the problem size is kept fixed. The step at the beginning of the SGI Altix curve shows once more the bandwidth problem already mentioned. In case the system size per CPU is constant, clusters of commodity hardware show a good price to performance ratio for this kind of code.

## Conclusions

The “computer pyramid” built by COTS clusters at the base, tailored vector HPC systems at the top and clusters of shared memory systems with tailored components (e.g. high-speed interconnects) for the gap in between, seems to be the only suitable model to meet the different requirements of HPC users and codes. For grand challenge applications however, only dedicated supercomputers featuring network and memory connections with extremely high bandwidth and low latency are suitable for these kinds of CFD codes with large data sets. In turn, it is difficult or practically impossible to achieve the same performance and speed-up behaviour even with several hundreds of processors even on architectures like the SGI Altix. The insuffcient memory bandwidth reveals the weak point in the design of microprocessor based systems, in particular of SMP machines with more than one CPU per memory path.

Requiring a sustained performance of 1 TFlop/s and extrapolating the presented performance numbers assuming ideal scale-up, 20 NEC SX-6+ nodes (160 CPUs), about 320 Hitachi SR-8000-F1 nodes, roughly 1400 Itanium2 processors or at least twice as many Xeon CPUs are necessary (see also Figure 4). However, it is mainly a hypothetical question whether the problem and the hardware scale up to thousands of CPUs. Therefore, not only for scenarios with fixed problem sizes (speed-up) but also for grand challenge CFD applications (scale-up), tailored vector-parallel systems seem to be the only cost-effective answer allowing short time-to-solution.

## Acknowledgements

Part of this work is financially supported by the Competence Network for Technical, Scientific High Performance Computing in Bavaria KONWIHR. Furthermore, we would like to thank the CSAR in Manchester for providing access to their computing facilities and acknowledge the helpful cooperation with our colleagues from LSS and LSTM at the University of Erlangen.

## References

[1] **Top 500 list available at**
http://www.top500.org
(November 2003)

[2] **L. Oliker et al.**
Evaluation of cache-based superscalar and cacheless vector architectures for scientific computations,
in: Proceedings of SC2003, CD-ROM, 2003

[3] **The ASCI program**
http://www.llnl.gov/asci/

[4] **A. J. van der Steen, J. Dongarra**
http://www.phys.uu.nl/~steen/web03/overview.html

[5] **Cray Advanced Technical Workshop**
June 2004, Bologna, Italy

[6] **S. Chen, G. D. Doolen**
Lattice Boltzman method for fluid flows,
Annu. Rev. Fluid Mech. 30 (1998) 329-364

[7] **Estimated prices per processor based on public procurements in 2003/2004**
GBit/Xeon: 2000 USD; Myrinet/Opteron: 3000 USD; Altix/Itanium2: 11000 USD; NEC: 60000USD

[8] **T. Pohl, F. Deserno, N. Thürey, U. Rüde, P. Lammers, G. Wellein, T. Zeiser**
Performance evaluation of parallel large-scale lattice Boltzmann applications on three supercomputing architectures,
in: Proceedings of SC2004, CD-ROM

[9] **Höchstleistungsrechner in Bayern**
Hitachi SR8000-F1, http://www.lrz-muenchen.de/services/compute/hlrb/

[10] **P. Lammers, K.N. Beronov, F. Durst**
Direct simulation with the lattice Boltzmann method of developed turbulence in channel flows, in S. Wagner et al. (Editors): High Performance Computing in Science and Engineering, Munich 2002, Springer Verlag 2003, pp. 43-58



• Thomas Zeiser
• Gerhard Wellein

Regionales Rechenzentrum Erlangen

• Peter Lammers

High Performance Computing Center Stuttgart (HLRS)

# MARMOT – an MPI Analysis and Checking Tool



## Introduction

The Message Passing Interface (MPI) is widely used to write parallel programs using message passing ([1], [2]). Due to the complexity of parallel programming, several reasons demand an increased need for debugging of MPI programs. First, the MPI standard leaves many decisions to the implementation, e.g. whether or not a standard communication is blocking, and therefore portability between different MPI implementations is not guaranteed. Second, parallel applications get more and more complex and especially with the introduction of optimizations like the use of non-blocking communication also more error prone. Examples of incorrect usage of MPI are the introduction of irreproducibility, race conditions, deadlocks and incorrect management of resources like communicators, groups, data types and operators. MARMOT [3] is a tool to aid in the development and debugging of MPI programs by verifying the standard conformance of an MPI program automatically during runtime.

Finding errors in MPI programs is a difficult task that has been addressed in various ways by existing tools. The solutions can be roughly grouped into four different approaches: classical debuggers, special MPI libraries and other tools that may perform a run-time or post-mortem analysis. Existing tools are afflicted with some disadvantages: they may not be freely available, they may require source code modification or language parsers, they may be limited to special platforms or language bindings, or they may be unable to catch incorrect usage of MPI, but only help to analyze the situation once the incorrect usage has produced an error like a segmentation violation.

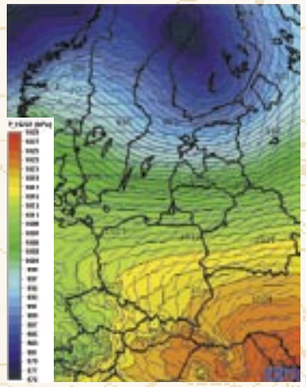


Figure 1: Meterological application

## Design of MARMOT

MARMOT is a library that has to be linked to the application in addition to the native MPI library, without requiring any modification of the application's source code. MARMOT uses the MPI profiling interface to intercept the MPI calls with their parameters for analysis before they are passed from the application to the native MPI library. As this profiling interface is part of the MPI standard, MARMOT can be used with any MPI implementation that adheres to this standard. The checking tool adds an additional MPI process for all global tasks that cannot be handled within the context of a single MPI process, like deadlock detection. Information between the MPI processes and this additional debug process are transferred using MPI. For the application, this additional process is transparent. Another global task is the control of the execution flow, i.e. the execution will be serialized if this option is chosen by the user. Local tasks are performed in a distributed way on every process of the parallel application, for example verification of resources like communicators, groups, datatypes, operators or verification of other parameters like ranks, tags, etc.

## Current Status

MARMOT supports the complete MPI-1.2 standard and provides the functionality described above. For example, when using the non-blocking call MPI_Isend the tool automatically checks if the communicator is valid, i.e. if it is MPI_COMM_NULL or if it is a communicator that has been created and registered with the appropriate MPI calls. Similarly, MARMOT also verifies if the data type is valid, i.e. if it is MPI_DATATYPE_NULL or if it has been created and registered properly. The tool also inspects the validity of the count, rank and tag parameters and the proper handling of the request argument, i.e. if an unregistered request is used or if an active request is being recycled.

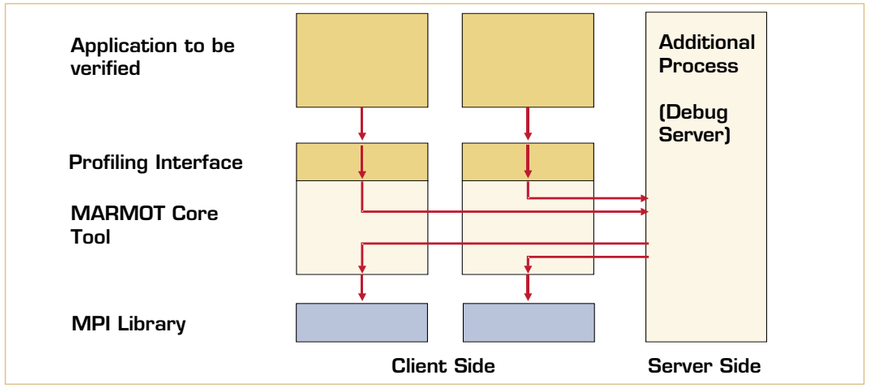



Figure 2: Architecture of MARMOT

The tool has been tested successfully with internal test suites, benchmarks, and applications from the CrossGrid project [8]. It was already ported to IA32 and IA64 clusters with MPICH [4], further to IBM, Hitachi, Cray and NEC platforms. Performance measurements show that attaching MARMOT to an application introduces an inevitable but still tolerable overhead [5, 6]. In the future we will improve MARMOT's performance, extend the number of tests and add further functionality according to the users' needs [7].

## Acknowledgements

The development of MARMOT is supported by the European Union through the IST-2001-32243 project "CrossGrid" [8].

## References

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

[2] **Message Passing Interface Forum. MPI-2**
Extensions to the Message Passing Interface, July 1997. http://www.mpi-forum.org/

[3] **MARMOT**
http://www.hlrs.de/organization/tsc/projects/marmot/

[4] **MPICH**
http://www-unix.mcs.anl.gov/mpi/mpich/

[5] **B. Krammer, K. Bidmon, M. S. Müller, M. M. Resch**
MARMOT, an MPI Analysis and Checking Tool. In Proceedings of ParCo2003, Dresden, Germany, Elsevier

[6] **B. Krammer, M. S. Müller, M. M. Resch**
MPI Application Development Using the Analysis Tool MARMOT. In M. Bubak, G. D. van Albada, P. M. Sloot, and J.J. Dongarra, editors, Computational Science - ICCS 2004, Volume 3038 of Lecture Notes in Computer Science, pp. 464 - 471, Krakow, Poland, 2004. Springer

[7] **B. Krammer, M. S. Müller, M. M. Resch**
MPI I/O Analysis and Error Detection with MARMOT. In D. Kranzlmüller, P. Kacsuk, and J.J. Dongarra, editors, Recent Advances in Parallel Virtual Machine and Message Passing Interface, Volume 3241 of Lecture Notes in Computer Science, pp. 242 - 250, 11th European PVM/MPI Users' Group Meeting, Budapest, Hungary, 2004. Springer

[8] **CrossGrid**
http://www.eu-crossgrid.org

• Bettina Krammer
• Matthias Müller

High Performance Computing Center Stuttgart (HLRS)

# DEISA – Meeting the European High Performance Computing Challenge

The provision of high performance computing resources to researchers has traditionally been the objective and mission of a few national HPC centres in Europe. Increasing global competition between Europe, USA, and Japan and growing demands for compute resources at the highest performance levels on one hand, and stagnant if not shrinking budgets combined with the need to innovate faster on the other hand, begin to show the limits of this concept. To stay competitive major investments are needed every two years – an innovation cycle that even the most prosperous countries hesitate to fund.

To advance science in Europe eight leading European HPC centres devised an innovative strategy to build a Distributed European Infrastructure for Scientific Applications (DEISA). The centres partner in building and operating a tera-scale supercomputing facility. This becomes possible through deep integration of existing national high-end platforms, tightly coupled by a dedicated network and supported by innovative system and grid software. The resulting virtual distributed supercomputer has the capability for natural growth in all dimensions without singular procurements at the European level. The investments that DEISA partners undertake in the course of their development naturally increase the capacity of the overall DEISA infrastructure. Advances in network technology and the resulting increase in bandwidth and lower latency virtually shrink the distance between the nodes in the distributed super-cluster. Furthermore, DEISA can expand horizontally by adding new systems, new architectures, and new partners thus increasing the capabilities and attractiveness of the infrastructure in a non-disruptive way.

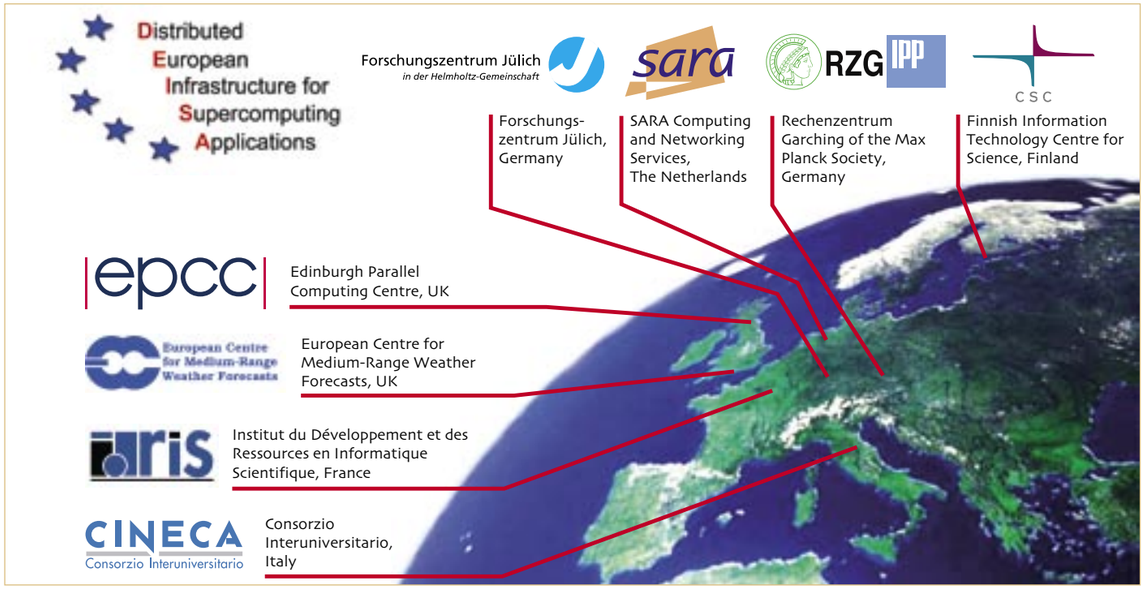



Figure 1: Map of Europe with logos of partners

Most important for the success of DEISA is the advancement of science in Europe. DEISA collaborates with leading European research groups initially from the following scientific and industrial disciplines:

- Material Sciences
- Cosmology
- Fusion Research
- Life Sciences
- Coupled CFD and CAA Applications
- Environmental Sciences.

The researchers aim to demonstrate that DEISA will enable new research results and enhanced scientific output in yet unprecedented ways. The list of disciplines listed above is not exhaustive, and new scientific initiatives are being prepared.

## Status of DEISA

Led by IDRIS-CNRS (France) the DEISA project started its activities in May 2004 with seven partners: FZJ and RZG in Germany, CINECA in Italy, EPCC and ECMWF in the UK, CSC in Finland, and SARA in the Netherlands. The DEISA infrastructure is being implemented in two phases. Initially, four "core" partners have coupled their systems using virtually dedicated 1 Gbit/s connections provided by the pan-European research network GÉANT and the national research networks RENATER (France), DFN (Germany), and GARR (Italy). Within one year other systems will be integrated into the DEISA super-cluster. At end of 2004 is combines over 4000 IBM Power 4 processors and 416 SGI processors for an aggregate peak performance of 22 teraflops.

The DEISA research infrastructure uses proven system software like AIX and Linux, GPFS and LoadLeveler, and grid middleware like UNICORE and Globus to achieve both a tight coupling of systems and access to heterogeneous resources.

DEISA will be open to collaboration with other Europe HPC centres and related initiatives world-wide. During its project span of five years it will adapt to the rapidly changing IT technology to retain its leadership position and guarantee its persistence beyond the project lifetime.

The big challenge is to demonstrate the DEISA strategy for the production of first class computational science as part of the European e-infrastructure.

For further information see the DEISA Consortium at www.deisa.org

DEISA is funded in part by the European Commission in the 6th framework program under grant number 508630.

• Dietmar Erwin

Central Institute for Applied Mathematics (ZAM)
Research Centre Jülich

• Victor Alessandrini

Institut du Développement et des Ressources en Informatique Scientifique (IDRIS)

# Network Topology and Performance Visualization for the European Research Network

When there is a failure or performance degradation in a trans-European network connection, it is rather difficult for the end-users to obtain timely and detailed information. Having access to this information is of growing interest especially for the Grid projects being established throughout Europe. Funded by the European Community, the Joint Research Activity 1 (JRA1) of the GEANT2 project is aimed at measuring the network performance of the GEANT network - which connects the national research and education networks throughout Europe – and providing meaningful information to GEANT users. Experience gained from a project at the Leibniz Supercomputing Centre is going to play a vital role in JRA1.

Besides making available high performance computing facilities, the Leibniz Supercomputing Centre also runs the Munich Scientific Network linking the Munich universities and other research institutions. This network became the starting point for the development of a network topology and performance visualisation tool called Customer Network Management (CNM). After the necessary data became available, the tool also has been provided to the customers of the German Gigabit Scientific Network (G-WiN) in April 2004. The G-WiN is the national research and education network and connects universities and research institutions throughout Germany. The aim of the DFN (German Research Network), which runs the G-WiN and funded the development of the CNM tool, is to provide a similar tool for the GEANT2 network. This work will be carried out as part of JRA1 that has been launched in September 2004.

A screenshot demonstrating how the tool could be applied for GEANT is shown in the Figure. The network topology is depicted as a tree structure on the left as well as in the topology maps on the right. For network nodes (i.e., routers, switches) and links current information about their status and characteristics are available. The status is indicated using colored status codes. To quickly find errors in the network hierarchy, a special network tree is provided showing only elements having warning or failure status. In its current version for the G-WiN the only characteristic for the nodes is the number of forwarded IP packets, while bandwidth, throughput, and utilization are provided for the links. In addition to the display of current values in the maps, it is also possible to get access to past values for status and characteristics. For doing so, statistics showing intraday, week, or month charts are provided.

The network topology is discovered by using SNMP variables defined for the use of the proprietary Cisco Discovery Protocol. The status and characteristics are derived from other SNMP variables. The client of the CNM is written in Java using the JavaWebStart tool to launch and update the applet. For the communication between client and server, CORBA has been chosen, the server itself is written in C++ and the data used for the statistics is stored in a PostgreSQL database.

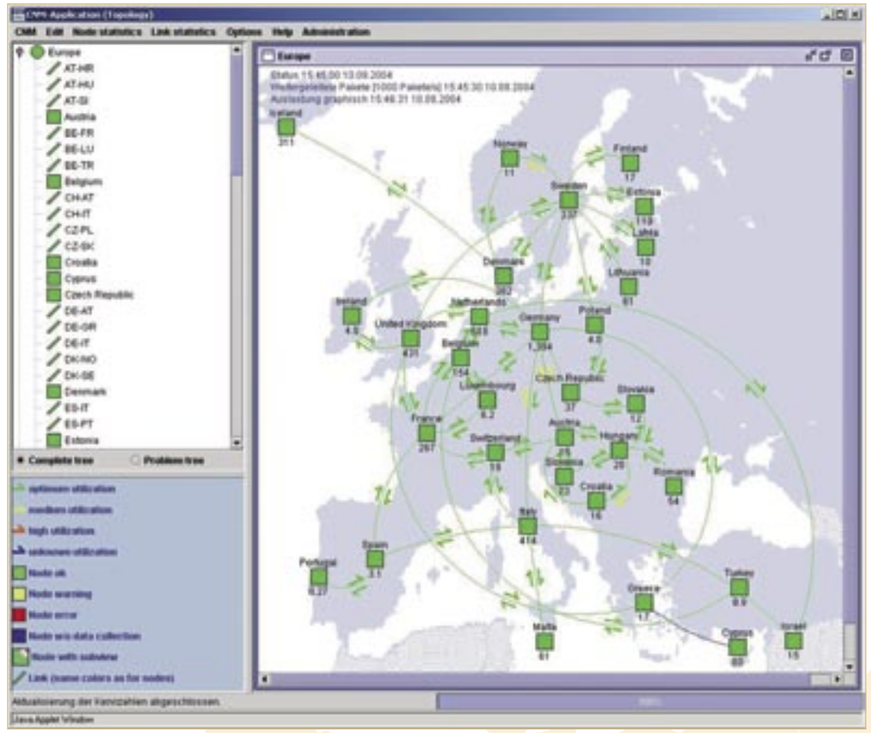



Figure 1: Customer Network Management tool for the trans-European Research Network GEANT

• Andreas Hanemann
• David Schmitz
• Matthias Brehm

Leibniz
Rechenzentrum LRZ



Projects

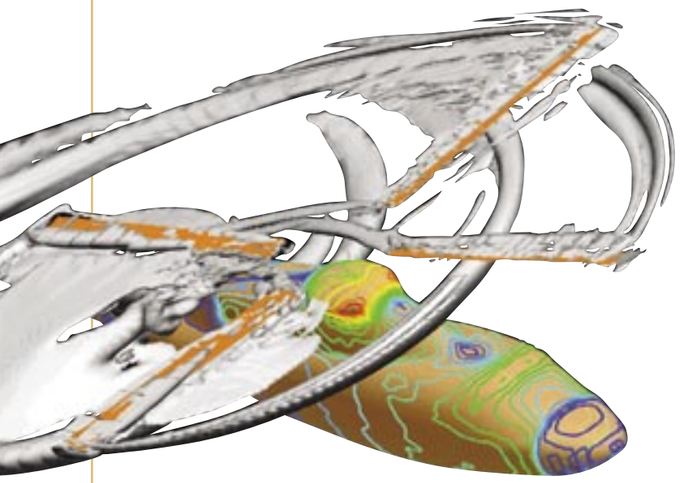


Figure 1: Aeroelastic analysis of helicopter rotor blades using HPC Aero- and Gasdynamics, University of Stuttgart

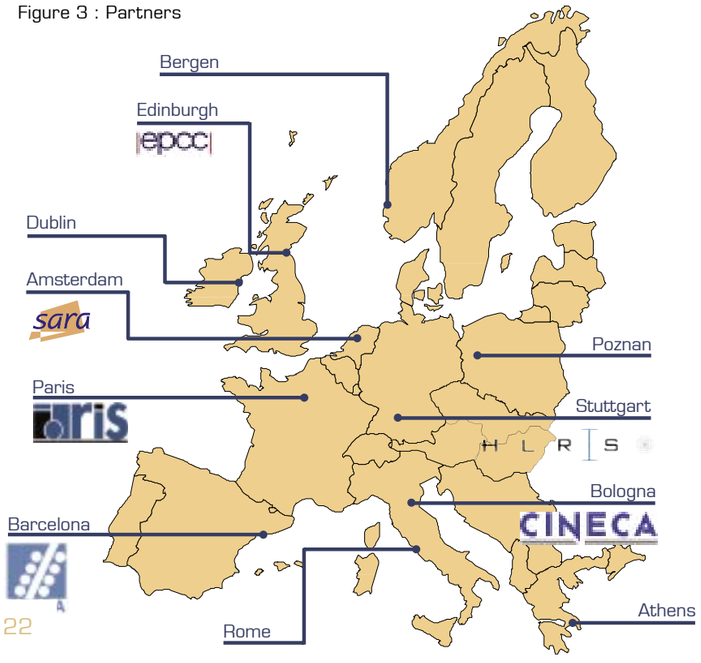


# The HPC – Europa Project

The HPC-Europa project is an EU-funded project within the 6th Framework Programme (FP6) with the focus on providing HPC services to the European research community in an innovative and coherent manner. HPC-Europa consists of several interrelated subprojects, at the core of which lies the transnational access visitor programme (TNA), accompanied by several networking and research activities.

The consortium of HPC-Europa consists of eleven leading centers working on the three parts of the project: Transnational Access (TNA), Networking Activities (NA) and Joint Research Activities (JRA).

Figure 3 : Partners




While the first and biggest part, TNA, is the visitor programme, the other two parts go hand in hand with this activity by developing new solutions for the AccessGrid video-conferencing toolset (NA), a seamless integration of performance measurement tools into the development chain on up-to-date HPC machines (JRA1) and single-point of access to HPC-facilities (JRA2).

## Transnational Access Visitor Programme

Within TNA, which builds on the successful TRACS, MINOS and ACCESS programs, around 800 scientists from European and other countries (see table1) may visit one of the affiliated universities using the HPC-resources of the six partner HPC-centers. Although the focus is mainly on high-performance computing, researchers from all scientific areas are encouraged to apply. The bjectives of HPC-Europa are to spark a fruitful collaboration of scientists from different subject areas in different countries and to increase the usage of HPC in areas not yet taking advantage of the capabilities of these computing resources.

Eligible countries are all EU-member countries plus Liechtenstein, Bulgaria, Norway, Iceland, Romania, Israel, Switerland, Turkey

It is envisaged hosting around 140 guest scientists at HLRS during the four year project period from 2004 until 2007. Of course, other universities in Southern Germany who are interested in hosting a guest are welcome to participate as well.

At the University of Stuttgart and the Max-Planck-Institute, currently 20 institutes are offering their experience with HPC and computational science to integrate guests into their working groups. All through their stay, guests will be supported by HLRS staff in case of problems. HPC-Europa funds the travel-costs, lodging, a daily allowance, as well as the compute costs for stays ranging from two to twelve weeks duration.

As accommodation, a single-room apartment at the University's guest-house is offered, which is an ideal location, at the heart of the campus in Vaihingen.

## Networking and Joint Research Activities

**Collaboration Support Infrastructure and Tools in AccessGrid**

Support is provided for the collaboration of European scientists through the usage of the AccessGrid technology, thereby allowing scientists to do distributed work on AccessGrid-enabled tools. The applications to be used and ported into AccessGrid will be HLRS' Covise collaborative visualization tool and CEPBA's Paraver performance analysis tool.

**Data Management and Portability**

Another activity is the management of complex scientific data in heterogeneous distributed computing environments. This problem will be addressed with a focus on tools and standards for efficient, reliable and transparent transfer of data between different platforms, browsing within the data and promoting best practices within the user community.

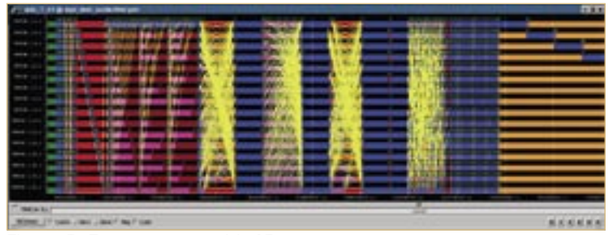


Figure 2: Paraver trace with 16 processors

**Performance Analysis Tools**

This part aims at developing tools and methodologies that will help operators and users to better measure and understand the performance behaviour of their applications, to allow them to make effective use of HPC-Europa´s infrastructure. The first objective is therefore to deliver uniform tools across all sites for advanced performance analysis. These tools, based on CEPBA's Paraver technology, will be ported within the frame of the project on all existing and future platforms. One example is the NEC-SX system at HLRS. Another activity will be fitting HRLS' PACX-MPI library with tracing and performance introspection capabilities.

**Single Point of Access**

A single point of access is developed for all centers, so that they can be used in a transparent way, regardless of the physical location of users and resources. This will be achieved by introducing two stages into the project: at first, existing solutions will be developed on a small subset of the computer center's resources, and later these will be integrated into a common access portal. Close relationships will be established with existing activities such as the Polish Progress Portal, the GridLab portal, GRASP, the UK eScience portal projects, Unicore, and the Global Grid Forum.

Further information:
http://www.hpc-europa.org

• Matthias Müller
• Rainer Keller

High Performance Computing Center Stuttgart (HLRS)

# The High Performance Computer Competence Centre Baden-Württemberg (hkz-bw)

## Formation and Scope of hkz-bw

In 2002 the Universities at Stuttgart and at Karlsruhe started negotiations about a bundling of resources in the field of supercomputing. While the High Performance Computing Centre Stuttgart (HLRS) had been a federal German supercomputing centre since 1996 the Scientific Supercomputing Centre (SSC) at Karlsruhe had provided compute power to scientists in the state of Baden-Württemberg.

The co-operation of these two centres was intended to strengthen the role of supercomputing in Germany and especially in the state of Baden-Württemberg. In 2003 the negotiations were brought to a close by a co-operation agreement between the two universities that was also signed and supported by the Ministry of Science, Research and the Arts of the State of Baden-Württemberg. In the corporate charter of the High Performance Computer Centre Baden-Württemberg (hkz-bw) the partners agreed to bundle the personnel and financial resources in the field of high performance computing so that an international competitive offering of high performance computing power can be maintained permanently.

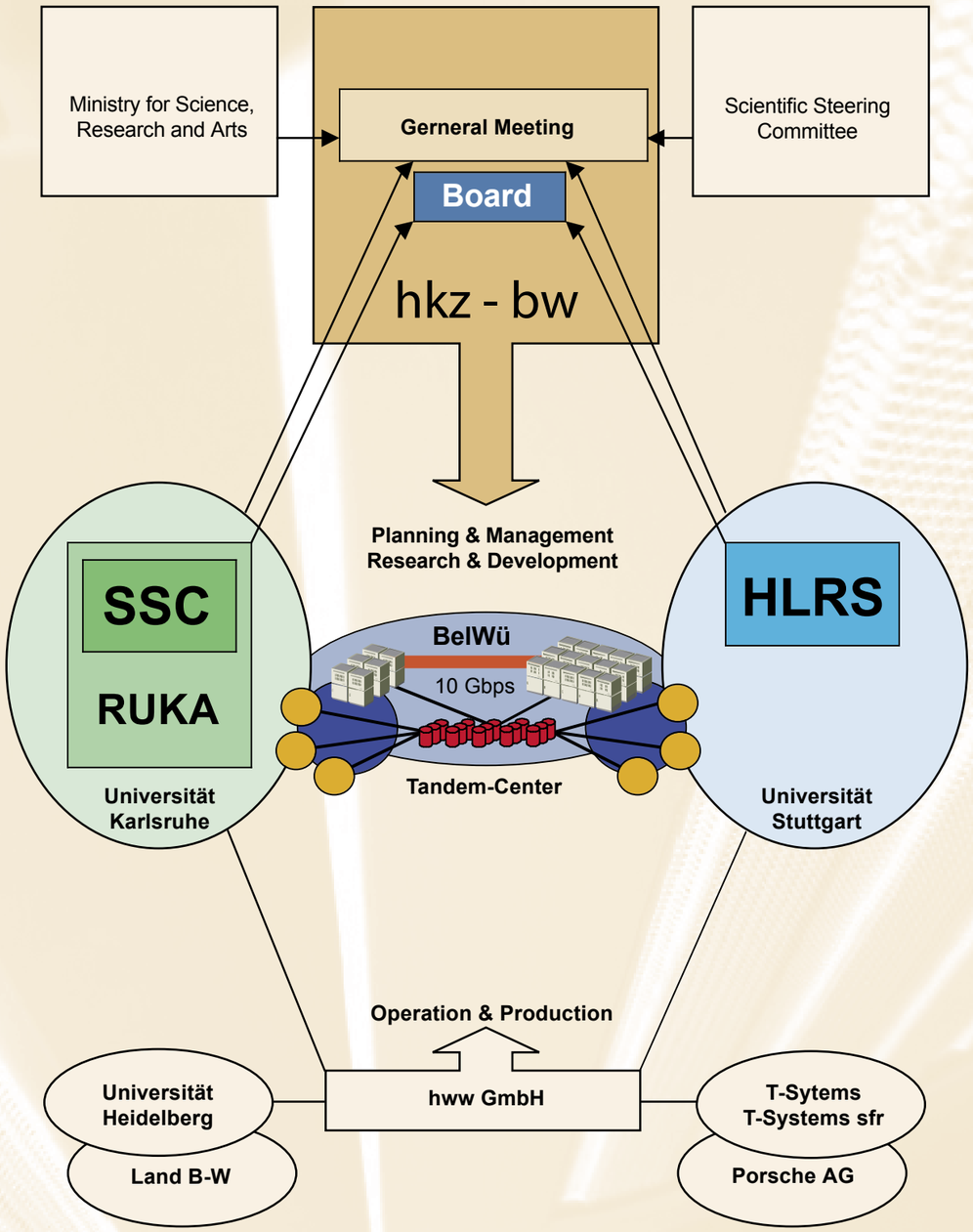
The two universities and their centres are soon to be joined by the University of Heidelberg and its Interdisciplinary Centre for Scientific Computing. This will help to strengthen the organisational ties with the grass root initiative WiR (Scientific Computing in the State of Baden-Württemberg). To extend its reach and integrate further resources in order to achieve more synergistic effects the hkz-bw is open for further partners.

## Role of hkz-bw

The hkz-bw is part of a comprehensive concept for supercomputing in the state of Baden-Württemberg and blends well in with the running activities at the HLRS and SSC as a part of a three tier framework:

- The public private partnership "High Performance Computing for Science and Industry" (hww) continues to take responsibility for basic operation of all supercomputers and mid range systems
- The hkz-bw takes responsibility to provide the basic services and aims to foster new fields of applications
- Scientific applications experts get support from mathematical experts and computer science experts in order to get the most out of any individual system.

The hkz-bw plans and organizes the utilization of appropriate computing resources for science in order to achieve a maximum efficiency for each type of application. HLRS continues to perform the tasks of a national scientific high performance computing centre by providing scientific users who need leading edge performance with outstanding computing power. SSC provides access






Cooperations

to a state-wide used platform that provides compute power at a level that can not be provided locally because of the high system requirements. Furthermore the hkz-bw takes the responsibility of coordinating the state's activities with regard networking.

## Tasks and Objectives

The hkz-bw takes over service and development tasks in the field of high performance computing. In particular these are:

- A continuous and reliable provision of science with high performance computer capacity and the appropriate high performance communication.
- The support of the users for an optimum utilization of the available computer and network infrastructure.
- The coordination and promotion of the utilization of high performance computers and networks.
- The exploration and development of new possibilities of numerical simulation and of new supercomputing applications in cooperation with the users.
- The creation and development of applications, middleware and libraries for high performance computing.
- The cooperation with and support of the research community WiR (Scientific Computing in the State of Baden-Württemberg).
- The promotion of the exchange of experience among the users as well as an active participation.
- The advanced education in the field of high performance computing.

The bodies of the hkz-bw are:

- Managing Board
- Members' Meeting
- Scientific Steering Committee.

The managing board is represented by the respective directors of the involved institutions. The hkz-bw members' meeting consists of two representatives of the concerned universities. It appoints the managing board and decides on all important affairs of the hkz-bw, especially the rules of procedure, the working program, the budget, and the admission of new members. The Scientific Steering Committee of the National High Performance Computing Centre HLRS has taken over additional tasks and now also advises the managing board of the hkz-bw on the acquisition and utilization of the high performance computers.

The utilization by scientific users is basically effected according to the regulations of the particular university respectively institution. The steering committee decides on the admission. The systems of the hkz-bw are also available to users from private enterprises against refund of costs via the industrial partners of the hww corporation.

## Additional HPC Research Promotion

Within the scope of the hkz-bw additional funds for the development of future technologies in supercomputing are provided by the State of Baden-Württemberg and the involved universities. By these funds further developments in the fields of middleware (especially scheduling and data management), programming environments, tools and application software for grid based supercomputing can be approached as well as new research areas be opened up or supported.

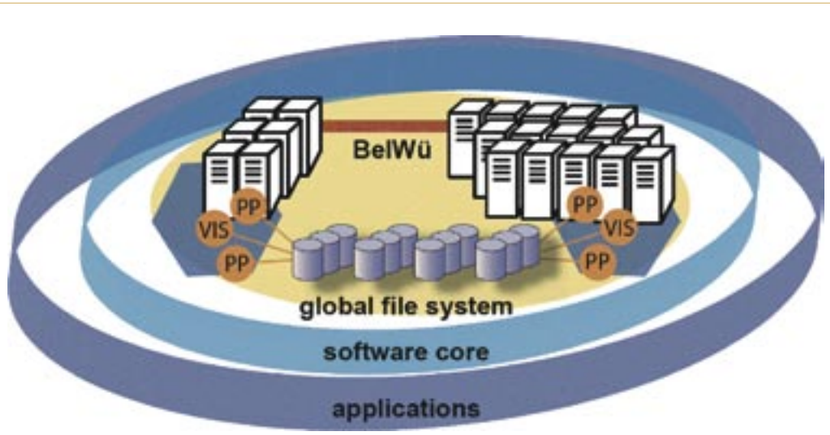
With its program "Modelling and Simulation on High Performance Computers" the Foundation of the State of Baden-Württemberg furthers application-oriented projects, which develop innovative solution methods in their particular specialist area by using the existing HPC infrastructure in the State of Baden-Württemberg. The objective is to establish the methods of HPC beyond the borders of the specialist areas over the medium and long term. The hkz-bw aims at a close cooperation with these projects in order to develop new application areas for high performance computing.

## Systems of the hkz-bw

The high performance computer systems of the hkz-bw comprise a supercomputer system of the top category and a high performance computer. Both systems have been chosen in such a way that their different architectures complement each other. So an optimum architecture can be offered for a broad application spectrum. The national high performance computer system, operated at the HLRS, is a vector system that is characterized by an extremely high storage and communications bandwidth. At the SSCK a high performance computer for the state on the basis of Intel Itanium 2 processors is operated. This system is in particular suited for applications which can profit from the large data cache. Both systems will be operated as a tightly coupled distributed system.

The computers will be directly connected on the basis of the existing network infrastructure in the State of Baden-Württemberg. Thus this installation at the hkz-bw is also to be seen as an example for an ambitious grid computing project lining up with the international initiatives like TeraGrid (USA), NPACI (USA), e-science (UK) or NAREGI (Japan).

The linking of the two systems is physically realized via the fast network connection of the state's research network BelWü with a target bandwidth of 40 Gbit/s. The logical linkage is arranged via a common .le system, which from an application's point of view only shows one name space. Within the application the communication between the systems can be performed via the PACX-MPI library developed at the HLRS, which is utilized throughout the world for coupled distributed systems in grid computing. The whole system presents itself to the user as a uniform system with a uniform access. Thereby the hkz-bw infrastructure as part of the D-Grid Initiative integrates "seamlessly" into the future research scenario.




• Michael Resch

Speaker of the
Board of the hkz-bw

# Why are Quarks Confined?

Among the German supercomputer centers, the John von Neumann Institute for Computing (NIC) is special in two aspects. Firstly, it comprises two research laboratories, the Research Centre Jülich and DESY which provide – through NIC – computer resources both on general purpose supercomputers as well as on specialized massively parallel architectures (see chapter "Centers" and then "NIC" of this issue for a description of the machines). Secondly, NIC is the only national supercomputer center that maintains research groups. One is the Complex Systems Group, led by Peter Grassberger and the other is the Elementary Particle Physics Group led by the author.

The Elementary Particle Physics group concentrates on numerical simulations in Lattice Field Theory (LFT). The main target is Quantum Chromo Dynamics (QCD) as our theory of the strong interactions, a theory that makes rather strange predictions, however. It postulates that there are fundamental constituent particles, the quarks, which interact through the gluons; but, at the same time it states that we cannot observe these particles. Instead, we can only see the bound states of the constituent quarks, the Proton, the Neutron and other hadrons in which the quarks are permanently confined.

Thus information on the theory on the fundamental level can only be obtained indirectly. How can we do this? By assuming the existence of the quarks and gluons, we can give theoretical predictions for, say, proton scattering experiments. Probing a hadron at small distances, corresponding – because of the Heisenberg uncertainty principle – to high energies, the quarks and gluons behave as almost free particles and the coupling between them is small. In this regime perturbative techniques can be applied and a number of successful theoretical predictions could be made. However, the distances where perturbation works is at most 0.2 fm while the proton radius is already 1 fm. The reason for the failure of perturbation theory is that the coupling between the quarks and gluons grows with increasing distance, eventually becoming so strong that the perturbative series in the coupling breaks down. In addition, the masses of the lightest quarks are about 10 MeV, while the mass of the proton is 1 GeV, i.e. 100 times larger. Hence, we are dealing with an unique system of an enormous binding energy impossible to describe by perturbative techniques.

In 1974 Nobel laureate Kenneth Wilson suggested that our continuum of space-time should be replaced by a 4-dimensional grid of lattice points in order to study non-perturbative phenomena. In this way, the problem can be mathematically well formulated and the complicated equations describing the interaction of the elementary particles can be solved on a computer. In the end of the 1970s Michael Creutz, following this approach, performed the first "computer experiments" in which he showed that physical quantities can be evaluated using numerical simulation techniques. The success of these early simulations in simple models was followed by a world-wide activity leading to Lattice Field Theory as being now an integral part of theoretical high energy physics including even simulations of super symmetric theories.

At present, we do not know, whether QCD can explain many fundamental questions that arise in the strong interactions, e.g.: What is the non-perturbative mechanism that forces quarks to be confined and what is the strength of the coupling of the quarks? Why is the binding energy of hadrons so enormously large? What is the nature of the finite temperature QCD phase transition and what are the properties of the quark gluon plasma that existed shortly after the big bang? Can we understand within QCD that some physical processes are not invariant under time reversal? Can we determine the internal structure of hadrons? What is the mechanism of breaking the chiral symmetry of QCD i.e. the symmetry of exchanging left- and right-handed quarks?

Lattice Field Theory is the only method we have today to give answers to the above questions. It can provide quantitative information on the theory but can also give insight into the mechanisms of the underlying physics. Of course, the most promising theoretical model is QCD, but it is as yet far from being clarified whether it is indeed the correct one. It is very important to remark that a full QCD simulation using lattice techniques will provide unambiguous answers; either the numbers emerging from such a simulation agree with experiment and QCD is established as the theory of the strong interaction, or, they disagree in which case QCD would not be correct and we have to look for different and new models that are based on completely novel and as yet unknown physical mechanisms and particles.

What is a full simulation? In principle, in such a simulation all 6 quarks of QCD have to be taken into account. However, for most of the questions it would be completely sufficient to only incorporate the three lightest quarks, the up, down and strange quarks. Unfortunately, even such simulations are very demanding and require machines that reach Teraflops performance to run for at least one year. This would mean that the whole IBM Regatta system at NIC would have to be dedicated to resolve just one problem of QCD! This is clearly unrealistic and lattice physicists today still have to resort to approximations but also work on conceptual and algorithmic developments to ease the lack of sufficient computing resources.

Nevertheless, the progress in Lattice Field Theory and the development of computer architectures that today reach the multi-Teraflops regime led to many physical results that find more

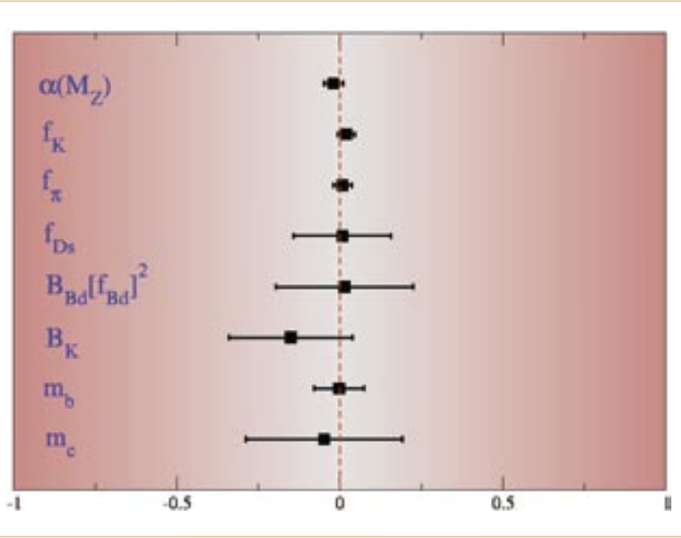



Figure 1: A comparison of experimental data and lattice results for various quantities as given by the Particle Data Group. Shown is the relative deviation between experiment and non-perturbative lattice QCD computations. We show the value of the coupling constant $\alpha_s$, several decay constants of mesons (the kaon (K), the pion , the D- and the B-meson). In addition, the masses of the charm and the bottom quark masses are displayed

and more their way into the particle data booklet, the "bible" for high energy physicists. Examples are the value of the strong coupling, the quark masses, hadronic matrix elements, form factors, the glue ball spectrum and even a random number generator. Another field where lattice results are essential is matter under extreme condition, i.e. in heavy ion collisions or neutron stars. Here, the lattice provides quantitative information about, e.g., the critical temperature, the pressure and the particle spectrum. Thus, lattice results play already now a significant and important role in the interpretation of experimental data (see Figure 1 for a comparison of lattice results and experimental data).

The physics projects discussed above could not be tackled without using high-end, powerful supercomputers. An important role is played by massively parallel supercomputers, in particular the APE (Array Processor Experiment) and the QCDOC (QCD On Chip) machines. The newest version of APE will be apeNEXT. We show in Figure 2 the main elements of the machine. It is a SPMD machine and runs asynchronously giving new challenges to the APE collaboration. The final installations will achieve 2-3Tflops for a stand alone system with price/performance ratio of 0.5Euro/Mflop (peak). These parameters will then meet the requirements formulated by an ECFA panel for the performance needs of Lattice Field Theory in the next years and the German Lattice Forum (LatFor) requirements which evaluates a need of 25 Tflops for the physics program of the German Lattice Community.

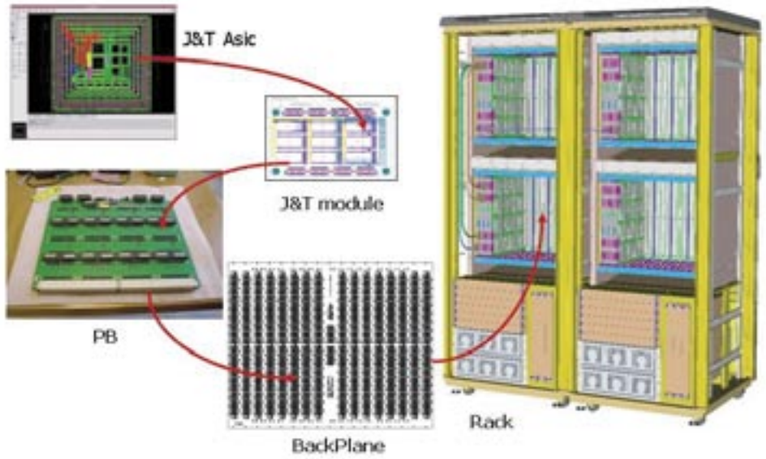



Figure 2: The design parts of the apeNEXT machine

• Karl Jansen

John von Neumann
Institute for
Computing (NIC)



Figure 1: New LRZ buildings; from left to right lecture room, institute building, computer cube

# Next-Generation National High Performance System to be Installed at LRZ

After full approval of the LRZ concept for a replacement of the presently operated national high performance computer Hitachi SR8000 by a refereeing committee of the German Research Foundation (DFG), an investment sum of 38 million Euro is now available for the procurement of the new system. The first phase of the new system will be put into user operation at the end of 2005 or early in 2006, and the projected application performance increase compared to the SR8000 should be at least a factor of 10. In mid 2007 the total performance of the system will be doubled with the phase 2 of the installation.

Simultaneously, LRZ welcomes the Bavarian government's ongoing support for the competence network for scientific high performance computing (KONWIHR), which is considered an essential aid toward efficient and application-driven usage of HPC systems, thereby assuring an internationally competitive position for research in Germany.

The new system, called "Höchstleistungsrechner in Bayern II" (HLRB-II), is to be installed in the new LRZ building in Garching near Munich, the shell of which is already nearly complete; the figure below gives an artistic impression of the new building. The HLRB-II will be located in the top floor of the 36x36x36 meter cube shown at the right.

Figure 2: New 36x36x36 meter cube for the computer equipment

• Matthias Brehm

Leibniz
Rechenzentrum LRZ

# Efficient Utilization of High Perfor mance Computers – Bavarian Competence Network for Technical and Scientific High Performance Comp uting

Many problems from the natural and engineering science can be described by mathematical and physical models the solution of which requires the use of computers. However, problems such as the simulation of turbulent flows, questions from physics and bio-sciences or the prediction of earthquakes require outstanding computational resources. The endeavour of combining state-of-the-art computer performance with efficient numerical algorithms and physical models and their application to the solution of technical and scientific problems is known as scientific high performance computing (HPC). Moreover, the efficient use of modern supercomputers is closely connected with extensive and qualified user training and support with the purpose of propagating the techniques of high performance computing.

As a follow-up to previous efforts, the Competence Network for Technical and Scientific High Performance Computing in Bavaria (KONWIHR, http://konwihr.in.tum.de/) was established by the Federal State of Bavaria in May 2000 with a total budget of about four million EUR for four years to expand the deployment of HPC technology within the scientific and industrial research communities. This goal is achieved by promoting scientific co-operation and putting emphasis on the interdisciplinary aspects of numeric and algorithmic research, as well as providing funding for training courses, technical meetings, workshops and conferences that help to transfer HPC technology into new application areas and ensure long-term viability of the development efforts.

Through KONWIHR, about 25 to 30 larger scientific projects and numerous short evaluation projects from various areas of research have been funded. To ensure high scientific quality, every project is reviewed annually by an external advisory board with experts from universities, research institutes and industry. The KONWIHR Offices in Munich and Erlangen are in charge of coordinating the activities and serve as contact points. The working groups integrate mathematicians, experts in computer science and engineers as well as natural scientists; personnel from both academic institutions in Bavaria and industrial enterprises are involved.

In this concept, the support projects located at the computing centers of the Bavarian Academy of Science in Munich (LRZ) and the computing center of the University of Erlangen-Nuremberg (RRZE) play an important role as they provide a local and persistent center of excellence for general scientific computing aspects, including HPC support (programming, debugging, optimization of codes), HPC training (lectures and tutorials for students and users) and information about the latest developments in HPC (new architectures, programming models, available resources, global trends, etc).

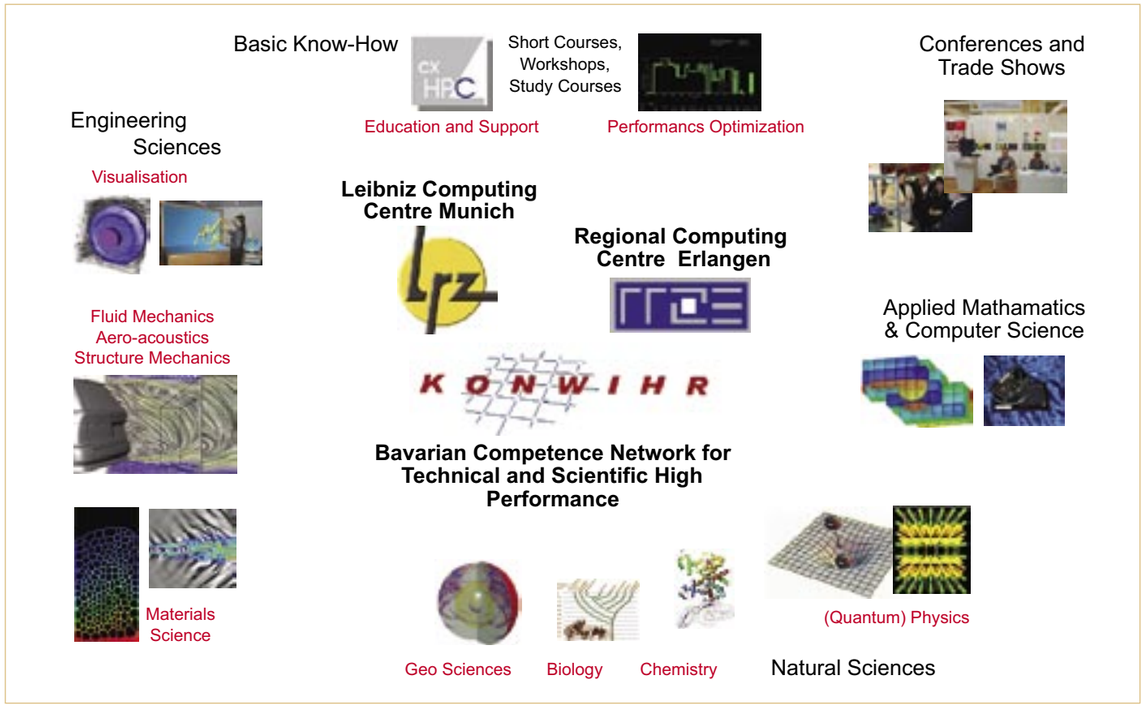
In its recommendations on “the future use of HPC systems”, the German Scientific Council points out that a “network of people” and “centers of excellence” should complement all future HPC acquisitions; therefore it is expected that the existing HPC centers of excellence at LRZ and RRZE will be strengthened, and that a KONWIHR follow-up program will be initiated alongside the next-generation national high performance system (HLRB II) scheduled for installation the new LRZ facilities at the end of 2005.


• Matthias Brehm
Leibniz-Rechenzentrum München

• Thomas Zeiser
Regionales Rechenzentrum Erlangen

• Arndt Bode
TU München
(Sprecher des KONWIHR)

• Franz Durst
LSTM University of Erlangen
(stellv. Sprecher des KONWIHR)

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 (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 replaced 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 nodes 140 GByte memory | 354 | Capability and capacity computing | Munich universities |
| IBM pSeries 690 hpc | 1 SMP node 8 processors Power 4 (single core) 32 GByte memory | 42 | Capacity computing | Munich universities |

# HLRS

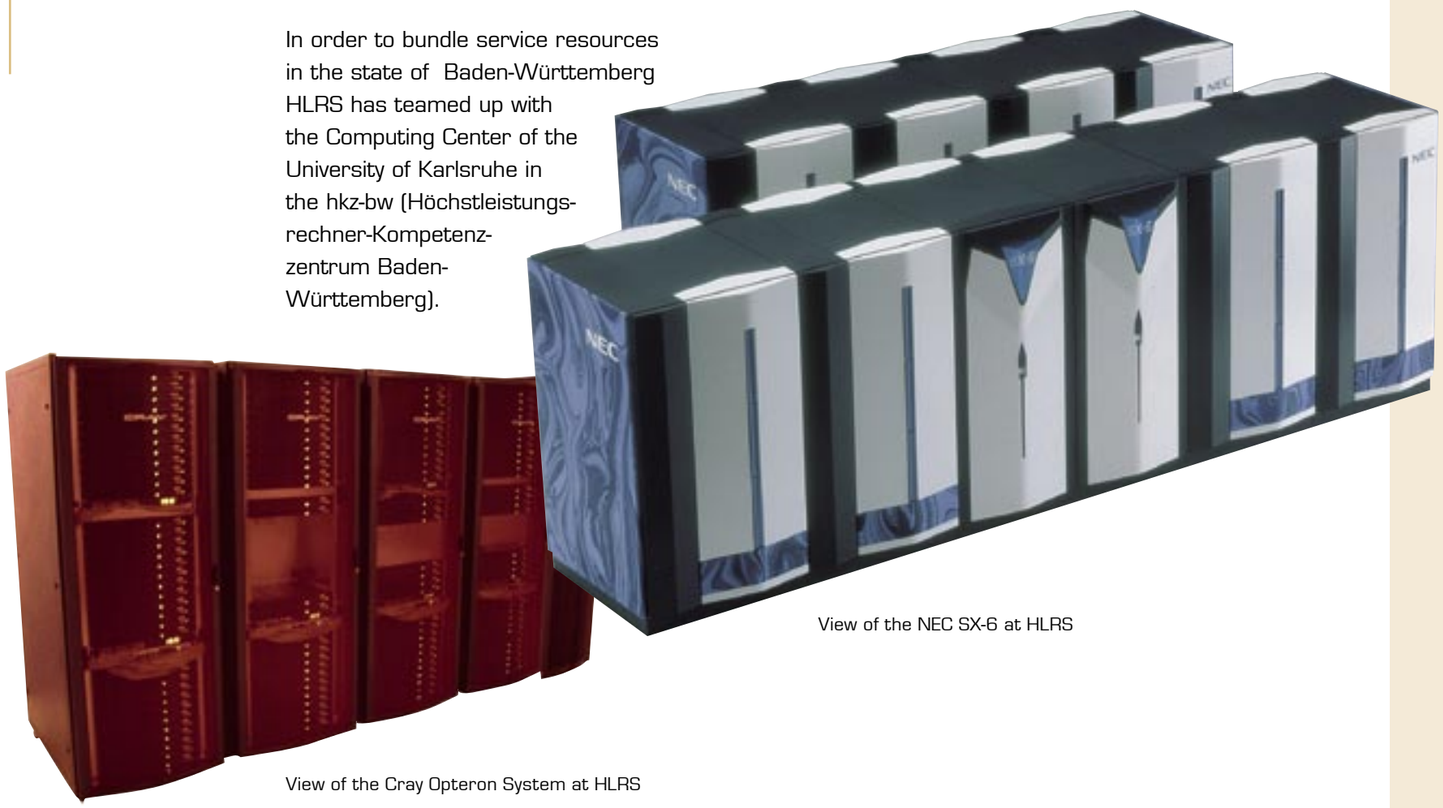
Based on a long tradition in supercomputing at Universität Stuttgart, HLRS was founded in 1995 as a federal center for High-Performance Computing. HLRS serves researchers at universities and research laboratories in Germany and their external and industrial partners with high-end computing power for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

Operation of its systems is done to ge-ther 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 (Höchstleistungs-rechner-Kompetenz-zentrum 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 +49 711 685 2504
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 | 129 2-way nodes 512 GByte memory | 1024 | Capability computing | German universities, research institutes, and industry |
| NEC TX-7 | 32 way node 256 GByte memory | 192 | Preprocessing | German universities, research institutes, and industry |
| IA64 Cluster | 8 2-way nodes 32 GByte memory | 57,6 | Capability computing | Universität Stuttgart |
| IA32 Cluster | 24 2-way nodes 48 GByte memory | 230,4 | Capability computing | Universität Stuttgart |

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 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 "Bioinformatics" is under consideration.

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

The new supercomputer 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/ape/html/Installation/
http://www-zeuthen.desy.de/main/html/home/

| 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 GByte memory | 32 | 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 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
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

### Application Performance Optimization

**Date**

March 8 - 9, 2005

**Location**

Regionales Rechenzentrum Erlangen

**Contents**

- Techniques of performance optimization for modern microprocessors
- Tools for application performance analysis

Lecturers from LRZ, RRZE, CS University of Erlangen

## HLRS
## www.hlrs.de

### Iterative Linear Solvers and Parallelization

**Date**

February 21 - 25, 2005

**Location**

HLRS, Stuttgart

**Webpage**

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

**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.

### Industrial Grid Workshop

**Date**

March 7 - 9, 2005

**Location**

HLRS, Stuttgart

**Webpage**

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

**Contents**

The workshop brings together industrial users, software developers and political decision makers. They will discuss industrial aspects of Grid computing and the potential of the Grid to respond to these.

### 8th HLRS Metacomputing & Grid Computing Workshop

**Date**

March 9 - 11, 2005

**Location**

HLRS, Stuttgart

**Webpage**

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

**Contents**

The workshop brings together leading scientists in the field of Metacomputing and Grid computing. Researchers from the US, Asia and Europe discuss current issues and future directions.

### 2nd Advanced Research Workshop on Computational Science and HPC

**Date**

March 14 - 16, 2005

**Location**

HLRS, Stuttgart

**Webpage**

http://www.grc-hpc.de/EN/index.php

**Contents**

The workshop brings together Russian and German scientists to discuss issues of computational science and technologies. A focus will be on simulation, mathematical methos and computational tools.

### 2nd HLRS-NEC Teraflop Workbench Workshop

**Date**

March 17 - 18, 2005

**Location**

HLRS, Stuttgart

**Webpage**

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

**Contents**

The Teraflop Workbench concept is an approach of supercomputing that puts data at the center of a virtual architecture.
The workshop brings together users that integrate their applications with this virtual workbench and achieve highest performance on leading edge vector and cluster architectures.

### hww File System Workshop on Scalable Global Parallel File Systems

**Date**

April 4 - 5, 2005

**Location**

HLRS, Stuttgart

**Webpage**

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

**Contents**

The workshop brings together leading researchers as well as vendors in the field of parallel file systems. Innovative research approaches and commercial solutions are presented and discussed.

### Introduction to Computational Fluid Dynamics

**Date**

April 4 - 5, 2005

**Location**

University of Kassel

**Webpage**

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

**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 commercial software and to interpret the solutions correctly. Methods and problems of parallelization are discussed.

## NIC
## www.fz-juelich.de/nic

### User Course "Parallel Programming with MPI, OpenMP and PETc"

**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 NIC/ZAM in collaboration with HLRS.

presented by Dr. Rolf Rabenseifner, HLRS

## NIC/ZAM
### Education in Scientific Computing

**Date**

August 8 - October 14, 2005

**Location**

NIC/ZAM, Research Centre Jülich

Application Deadline: April 30, 2005
www.fz-juelich.de/zam/gaststudenten

# Miscellany



## Workshop Report "Capability Computing with Blue Gene/L"

Blue Gene/L, IBM's new and extremely powerful high-end computer architecture for capability computing, was the subject of a joint workshop organised by IBM Deep Computing and the John von Neumann Institute for Computing at the Research Centre Jülich, 7 - 8 September 2004.

What makes Blue Gene/L so attractive are its scalability and its low requirements for power, cooling and floor space. In the near future, a leadership-class Blue Gene/L system with a peak performance of 360 TeraFlops provided by 65,536 special IBM PowerPC processors will be installed at the Lawrence Livermore National Laboratory. The integration of such a machine into a production environment as well as the enormous number of processors is a highly motivating challenge to computing centres and programmers.

The aim of the workshop was to get a first assessment of the computational capabilities of Blue Gene/L and to find out which types of applications best match its architecture. Speakers from the IBM developer group gave insight into the hardware and software architecture of Blue Gene/L. Scientists from a selected set of prospective application areas including biophysics, quantum chemistry, astrophysics, and quantum field theory discussed their specific computational requirements and IBM specialists described the opportunities that Blue Gene/L offers to these fields and reported on first impressive experiences. Additional talks provided a general overview as to the options of future high-end machines including the extension of general-purpose processors by FPGAs.

The workshop has confirmed that Blue Gene/L, an existing system on the pathway to leadership class, is delivering performance values one or two orders of magnitude ahead of the development of Moore's Law for a wide class of grand-challenge applications.

All presentations of the workshop are available at the web page
http://www.fz-juelich.de/zam/bgl

R. Esser, FZJ/ZAM, 02461-61-6588

## Miscellany HLRS Personnel Matters:

Prof. Dr.-Ing. Egon Krause from the Department for Aerodynamics and Fluid Dynamics of the RWTH Aachen has stepped down as the chairman of the steering committee of the HLRS in October. Prof. Krause held this position since 1996. His activities were crucial when setting up the center and establishing it as a scientific beacon in Europe for computational science and engineering. Prof. Krause will remain in close contact with the center.

Dr.-Ing. Edgar Gabriel from the Innovative Computing Laboratory of the University of Tennessee, Knoxville, has joined HLRS in September 2004. He will be heading a newly created group on cluster computing.

## Others:

July 2004: HLRS and partners initiate the European Project GridCoord that aims at developing visions and roadmaps enabling Europe to play a leading role in Grid technologies and applications.

November 2004: HLRS and NEC set up their collaboration in cluster computing and install a 400 processor Infiniband based cluster. The collaboration aims at developing software and integrating various hardware platforms in a teraflop workbench.

November 2004: Springer and HLRS publish "High Performance Computing in Science and Engineering 2004, Transactions of the High Performance Computing Center Stuttgart (HLRS)". The book gives an excellent overview of the state of the art in Supercomputing on vector-systems and scalar systems.

## New volume of High Performance Computing in Science and Engineering

German leading-edge research groups in the field of scientific computing present their outstanding projects using the High Performance Computer in Bavaria (HLRB), Hitachi SR8000-F1, one of the top-level supercomputers for academic research in Germany. The projects address modelling and simulation in the disciplines Biosciences, Chemistry, Chemical Physics, Solid-State Physics, High-Energy Physics, Astrophysics, Geophysics, Computational Fluid Dynamics, and Computer Science. The authors describe their scientific background, their resource requirements with respect to top-level supercomputers, and their methods for efficient utilization of the costly high-performance computing power. Contributions of interdisciplinary research projects that have been supported by the Competence Network for Scientific High Performance Computing in Bavaria (KONWIHR) complete the broad range of supercomputer research and applications covered by this volume.

High Performance Computing in Science and Engineering, Munich 2004 Transactions of the Second Joint HLRB and KONWIHR Result and Reviewing Workshop, Springer-Verlag Berlin Heidelberg New York, 2004. ISBN 3-540-44326-6

# inSiDE

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

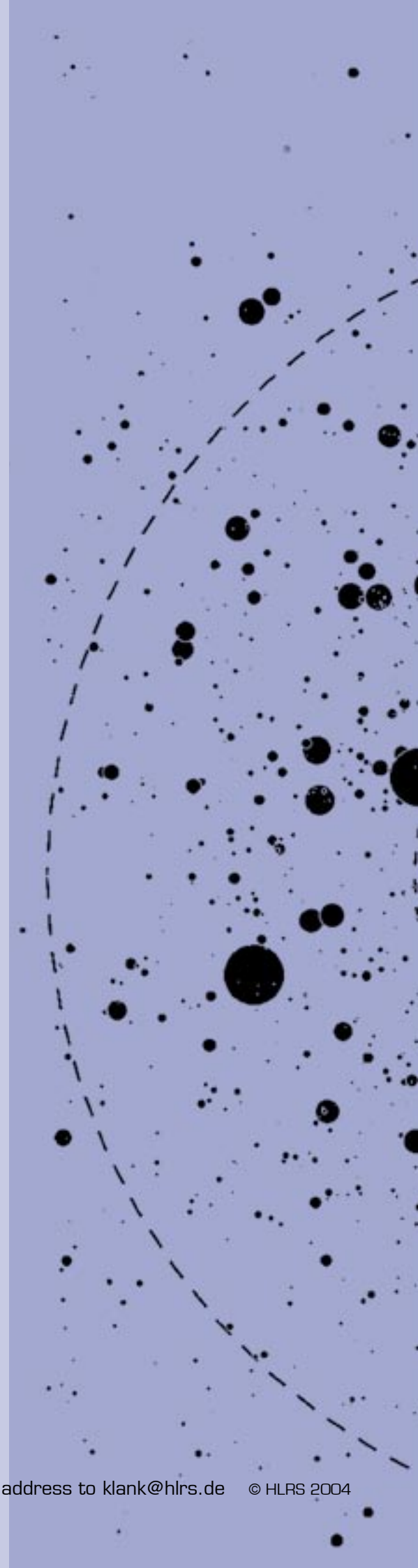
**Publishers**
Prof. Dr. Heinz-Gerd Hegering, LRZ
Priv.-Doz. Dr. Dr. Thomas Lippert, NIC
Prof. Dr. Michael M. Resch, HLRS

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

**Design**
Katharina Schlatterer
schlatterer@hlrs.de



**Authors**
Prof. Victor Alessandrini
va@idris.fr
Artur Baumgaertner
a.baumgaertner@fz-juelich.de
Prof. Arndt Bode
bode@in.tum.de
Dr. Matthias Brehm
brehm@lrz.de
Dr. Günter Duckeck
guenter.duckeck@physik.uni-muenchen.de
Prof. Franz Durst
franz.durst@lstm.uni-erlangen.de
Dietmar Erwin
d.erwin@fz-juelich.de
Jean-Fang Gwan
j.f.gwan@fz-juelich.de
Andreas Hanemann
hanemann@lrz.de
Karl Jansen
karl.jansen@desy.de
Frank Jenko
Jenko@ipp.mpg.de
Bettina Krammer
krammer@hlrs.de
Rainer Keller
keller@hlrs.de
Peter Lammers
plammers@hlrs.de
Dr. Matthias Müller
mueller@hlrs.de
Prof. Michael Resch
resch@hlrs.de
David Schmitz
schmitz@lrz.de
Dr. Gerhard Wellein
gerhard.wellein@rrze.uni-erlangen.de
Thomas Zeiser
thomas.zeiser@rrze.uni-erlangen.de



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