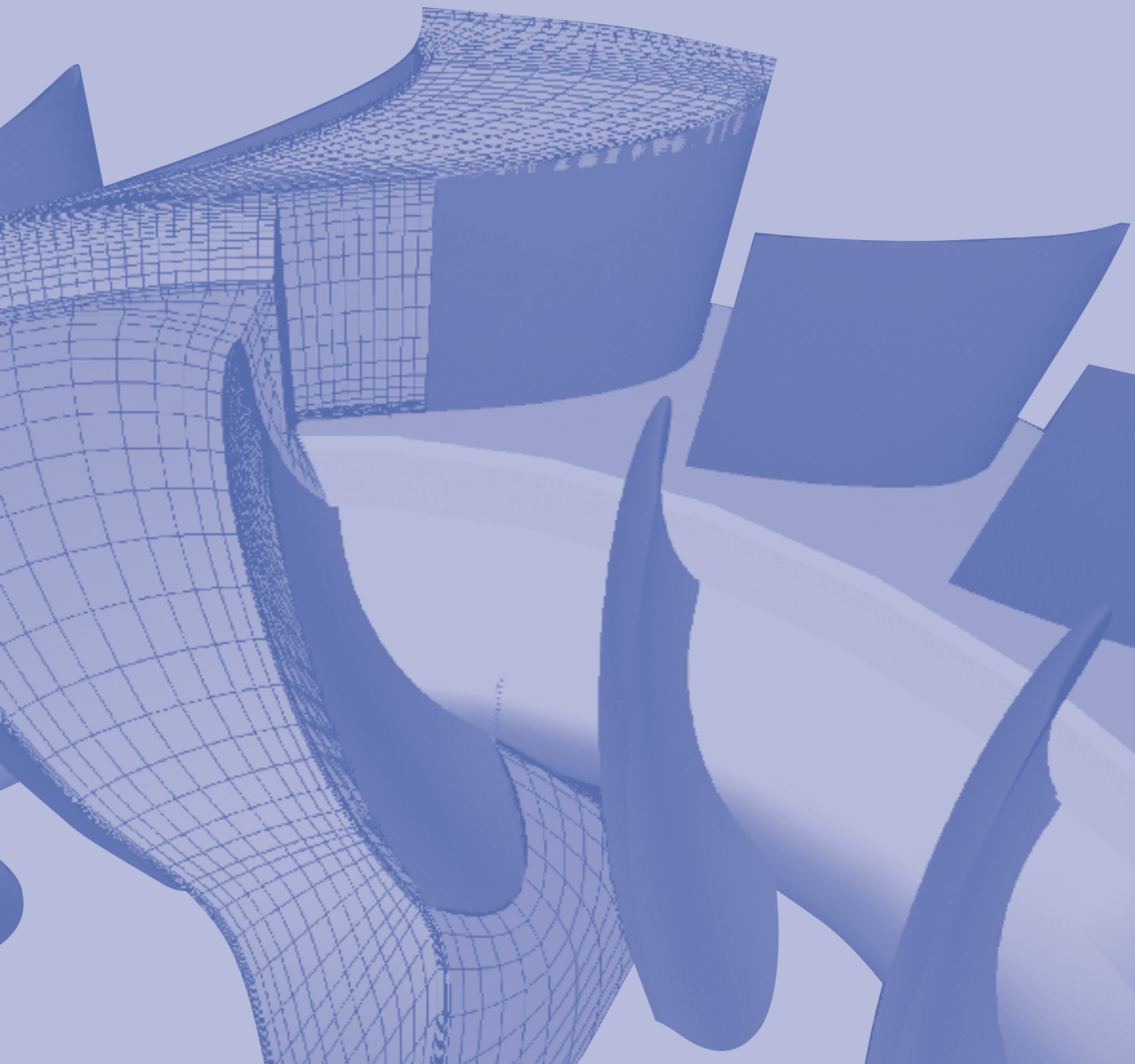
inSiDE

inSiDE • Vol. 4 No. 2 • Autumn 2006

Innovatives Supercomputing in Deutschland



Editorial

Welcome to the autumn issue of inSiDE the bi-annual German supercomputing information journal. The last months have brought interesting changes in the German supercomputing landscape. LRZ has inaugurated its new SGI system at Garching and has turned the spiral of innovation further for Germany. Details about the system can be found in this issue. With its installation the next cycle of hardware innovation of the three German national supercomputing centers (HLRS, LRZ, NIC) has been completed.

The inauguration at Garching also served as a platform to announce the strategic support for supercomputing in Germany. In her speech the German Minister of Science Annette Schavan announced her strong support for supercomputing in Germany and encouraged the community to continue its common efforts towards a national HPC strategy. As a result of these activities the three German national centers have agreed to form a new Gauss Center for Supercomputing bundling their resources. Details about the new initiative can be found in this issue.

The three new supercomputers installed in Germany in the last months and years have pushed applications to achieve new and outstanding results. Stephan Hochkeppel and Werner Hanke describe the simulation of 10^{23} electrons in a bath of magnetic correlations and how this helps to understand high-temperature superconductivity. Guido Arnold et al. describe how quantum computer simulations can be improved and present results for different architectures. Jan Wissink and Wolfgang Rodi show the potential of supercomputers for engineering in their paper on direct numerical simulation of transitional flow around turbine blades. Andreas Schäfer gives new results about hadron resonances in lattice QCD.

Continuing our effort to report on German and European projects this issue describes four projects. Two of them cover general aspects of supercomputing and deal with performance of systems. Markus Geimer et al. report on their work on scalable parallel trace-based performance analysis. Matthias Brehm et al. describe in their article how performance counters can be used to get insight into the inner state of a supercomputer. Both projects are German activities and document the level of supercomputing software research in Germany. Rainer Keller gives an overview of the int.eu.grid, a European project started in May 2006 to support large-scale interactive, parallel applications within a European Grid. At the core of such a European Grid is middleware which is dealt with by the OMII-Europe project described by Alistair Dunlop.

As always this issue includes information about events in supercomputing in Germany over the last months and gives an outlook of workshops in the field. Readers are invited to participate in these workshops.

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

Contents

The Gauss Centre for Supercomputing

Germany's leading position in the field of computational science and engineering has recently been reinforced substantially: The three German national supercomputing centres at Jülich, Garching, and Stuttgart joined forces and gave birth to the Gauss Centre for Supercomputing. The alliance of the John von Neumann Institut für Computing (NIC), the Leibniz-Rechenzentrum (LRZ), and the Höchstleistungsrechenzentrum Stuttgart (HLRS) provides one of the largest and most powerful supercomputer infrastructures in Europe.

The Gauss Centre for Supercomputing (GCS) is an alliance of the three national supercomputing centres into a virtual organization enabled by an agreement between the Federal Ministry of Education and Research (BMBF) and the State Ministries for Research of Baden-Württemberg, Bayern, and Nordrhein-Westfalen from July 2006.

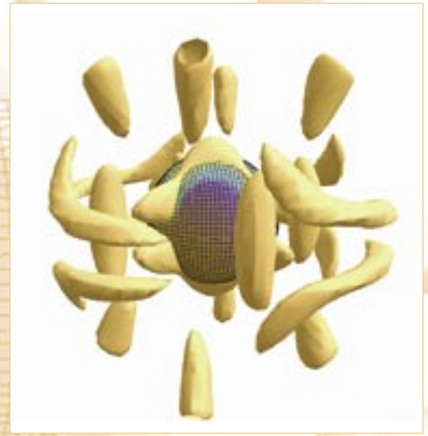


Figure 1: Simulating the Earth with a Supercomputer

The GCS will give Germany best prospects to adopt a leading role in the future European high performance computing ecosystem: Within the forthcoming Seventh Framework Programme the European Commission plans to support the building of a Europe-wide supercomputer infrastructure in the Petaflop performance range. In the future, the co-operation of the national supercomputing systems in Europe will be planned, realized and adapted to a larger structure by a concerted action of the European countries involved in supercomputing. Achim Bachem, the new Chairman of Forschungszentrum Jülich, was appointed to represent the GCS in European bodies.

The BMBF announced to support the development of highest-speed data communication between the centres by 30 Million Euros, in order to promote the scientific co-operation between the three centres and in particular between their user communities in the area of high performance computing.

The GCS offers state-of-the-art high performance computing and networking infrastructure. The LRZ recently inaugurated a 26 TeraFlop/s SGI Altix 4700 system that will be expanded to more than 60 TeraFlop/s in 2007, the HLRS offers a 12 Tera-Flop/s NEC SX-8, and the NIC can provide a 46 TeraFlop/s IBM Blue Gene/L as well as a 9 Tera-Flop/s IBM p690 Cluster. The architectures of these machines are different yet complementary. Each one favours special types of applications. The Altix can offer a huge shared memory, the SX-8 is highly efficient for vectorized codes, the Blue Gene boosts applications which scale to extreme processor numbers, and the p690 Cluster provides large-memory SMP nodes.

With strong support of the BMBF and involvement of German industry, the high-speed communication lines between the centres – for instance the 10 Gbit/s DEISA dedicated network – will be enhanced to 40 Gbit/s, later striving for 100 Gbit/s. Such a data throughput will enable completely new services and co-operative applications. Access to the resources is enabled by Grid technology, which – together with high-speed communication – will also facilitate distributed computing and data storage. Furthermore, the German universities and research laboratories and all other important computing and data centres, which have excellent communication links to the GCS through the well-established national research network (DFN), will highly benefit from the GCS acting as the major German scientific data centre and hub.

The mission of the GCS is the provision of world class supercomputing power for computational science and engineering for Germany and Europe. Another major focus of the GCS is world-leading methodical user support, education, and dissemination of best practice in simulation science. To this end, HLRS, LRZ, and NIC will synchronize and optimize their existing successful support structures within the GCS. The structures will be adapted to the specific requirements of various user communities like materials science, physics, climatology, computational biology, engineering, etc.

The GCS will be active for the further successful development of simulation methods and algorithmic and computational tools in the context of international computational science and engineering. Within the GCS, a long-term scientific programme will organize and

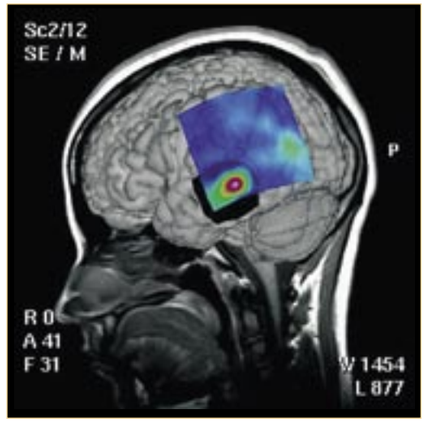


Figure 2: Creating a Virtual Brain with a Supercomputer

support this process in co-operation with universities and research institutes. The GCS will help strengthening the links to European science and engineering communities.

With the Gauss Centre for Supercomputing Germany has the unique opportunity to play a major role within the planned European Petaflop-supercomputer infrastructure, for the benefit of computational science and engineering in Europe and its applications in science and industry.

- Achim Bachem[1]
- Heinz-Gerd Hegering[2]
- Thomas Lippert[3]
- Michael Resch[4]

[1] Forschungszentrum Jülich (FZJ)
[2] Leibniz Rechenzentrum (LRZ)
[3] John von Neumann Institute für Computing (NIC)
[4] Höchstleistungsrechenzentrum Stuttgart (HLRS)

Understanding High-Temperature Superconductivity

Simulation of 10^{23} electrons in a bath of magnetic correlations

Normal metals are characterized by a finite value for their electrical resistance. The resistance is a result of the scattering of the electrons with the crystal lattice. In the beginning of the 19th century the physicist Kamerlingh Onnes found that Mercury, which at room temperature is a normal metal, has an anomalous behavior at low temperatures. Mercury discards its normal metal character below a transition temperature of approximately 4 Kelvin and displays no dissipation of electrical current, i.e. looses its resistance. The phenomenon is called superconductivity (SC). The current flow without an electrical resistance, i.e. superconductivity, is highly desirable in various fields and techniques (e.g. energy storage or ultra fast chips, etc.) or medicine applications (e.g. computer tomography). However, the very low transition temperatures of superconductors restrict the practicability for technical applications because of the complicated and expensive cooling.

Twenty years ago a Swiss-German team of physicists found a special ceramic (copper and oxygen) compound which exhibits transition temperatures in the region of about -150 degrees Celsius. At such transition temperatures, these materials can be relatively simply cooled with liquid air and, therefore have the potential for extremely exciting applications. This fact has prompted the winning of the Nobel Prize for the research team immediately after their discovery. A detailed understanding of the physical mechanism of this so-called high-temperature superconductivity (normal superconductors have a transition temperature of about -250 degrees Celsius) could be a road map to systematically find new materials with higher critical temperatures.

The behavior of the conventional superconductors (very low transition temperatures) is explained by the so-called BCS-theory where in the superconducting state the electrical current is formed by pairs of electrons. Although two electrons show a repulsive behavior due to the Coulomb force, below the transition temperature the crystal lattice ions provide an elastic medium (like a mattress) resulting in a net attractive interaction between the two electrons. The mechanism responsible for the attractive interaction between the electrons can be conceptually linked to the deformation of this mattress caused by two heavy shot-puts.

The shot-put (electron) deforms the elastic medium (ionic lattice in the solid).

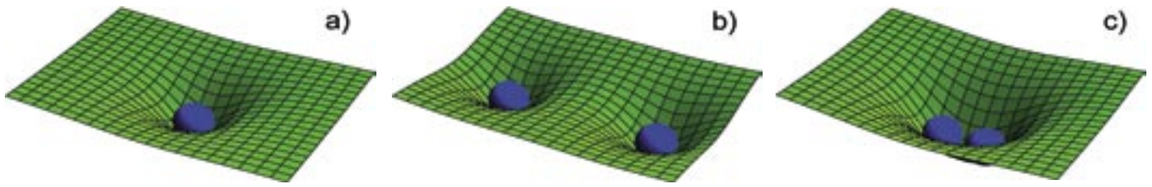


Figure 1: Illustration of the effective interaction between two electrons. The negative charged electron deforms the opposite charged ionic lattice (a). The consequence is an effective positive charge around the electron. The second electron will be attracted by this positive charge to form the so called Cooper-pair (b-c)

It is then energetically less costly to put two shot-puts (electrons) in one deformation to form the Cooper-pair.

In a microscopic picture, the high-temperature superconductors consist of parallel ordered copper-oxide planes. The electrical resistance perpendicular to the planes is several orders of magnitude larger than inside the planes. Therefore the original solid can be treated as a two dimensional object. On the basis of earlier high-performance supercomputer studies it has been shown that the pairing mechanism in the case of high-temperature superconductors must differ from the pairing mechanism of conventional superconductors in order to explain the very high transition temperatures. The result was that the electrons themselves, or more precisely the spin arrangement of the electrons is responsible for the attractive interaction between the electrons.

The parent copper-oxygen crystals show antiferromagnetic behavior (Neel order), that means that the magnetic moments of adjacent cooper-ions are oppositely arranged. In this antiferromagnetic scenario the copper-oxygen crystals are insulating. By doping further electrons or holes to the system the antiferromagnetic order is broken down and electrons or holes can conduct the electrical current. In the region of low temperatures, the current carriers can flip the magnetic moments of the crystal ions which can then form a so called “spin-bag” (this correspond the “deformation” in Figure 1). When “spin-bags” of two electrons overlap each other a resulting attractive interaction and thus superconductivity can occur.

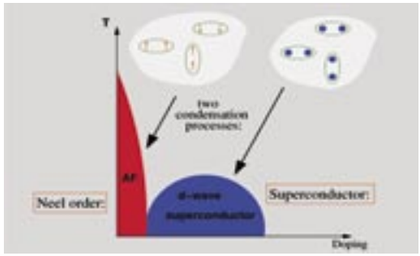


Figure 2: Generic phase diagram of high T_c-superconductors

The generic phase diagram of the high-temperature superconductors with the insulating antiferromagnetic (Neel order) and superconducting phase is depicted in Figure 2.

Nevertheless, it is not yet understood how the coherent motion of a macroscopic current of 10^{23} electrons comes about. Therefore, one tries to simulate the dynamics of the electrons with complex computer algorithms. In a first step the physics of Cooper pairs on a cluster will be taken into account in an exact way (e.g. with Quantum-Monte-Carlo methods or exact diagonalization methods). In a second step the coherent motion of the 10^{23} Cooper-pairs on the infinite lattice is simulated with clever new algorithms. Variational methods translate the microscopic physics given on the cluster to the infinite lattice. The corresponding supercomputer-simulations, in particular at the Leibniz-Rechenzentrum, have shown that the Hubbard model (simplified model for the high T_c-superconductors) exhibits superconductivity.

The remaining computational problem is to solve more realistic models which incorporate the most relevant electronic degrees of freedom of the copper-oxide orbitals. Therefore, the usage of supercomputers and the development of new computer algorithms is unavoidable.

• Stephan Hochkeppel
• Werner Hanke

Department of Theoretical Physics, University of Würzburg

Improving Quantum Computer Simulations

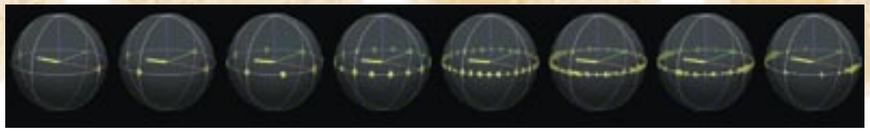


Figure 2: Quantum Fourier Transformation: the ensemble average spin vector and the cloud of measured endpoints at the end of each single experimental run. The leftmost qubit is q=1

Quantum computers would exploit the unusual behavior of the smallest particles of matter and light. Their theoretical ability to perform vast numbers of operations simultaneously has the potential to solve certain problems, such as breaking data encryption codes or searching large databases, far faster than conventional computers [1]. Promising candidates for use as quantum bits (qubits) in quantum computers are ions (electrically charged atoms), the nuclear spins of certain atoms in special designed molecules (addressed by nuclear magnetic resonance techniques) and electron spins trapped in quantum dots (serving as artificial atoms).

Despite of the impressive development during the last years, an experimental demonstration that quantum computation can solve a non-trivial problem is still lacking [2]. To be of practical use, quantum computers will need error correction, which requires at least several tens of qubits and the ability to perform hundreds of gate operations. A physically realizable quantum computer is a complex many-body quantum system. In order to exercise control over many qubits and to suppress the rate at which errors are introduced during a quantum computation (decoherence and systematic errors due to imperfections of pulse sequences), it is indispensable to understand the time-dependent behavior of the whole system. In addition, the ability to simulate realistic models of quantum computers that interact with their environment is crucial for the successful realization of scalable quantum computing hardware – which is currently the most significant barrier to building a working quantum computer.

A starting point are simulations at gate level. Each quantum operation is represented by a quantum gate which acts instantaneously on the state vector of the qubits. Since the dimension of the state vector grows exponentially with the number of qubits, the simulation of quantum computers is clearly memory bounded. For example, 3 TB of memory are required to simulate 37 qubits. Highly optimized memory access and communication patterns are needed for efficient simulation. For this purpose the “Massively Parallel Quantum Computer Simulator” described in [3] has been developed. This code exhibits very good scaling as a function of the number of processors.

Figure 1: Scaling (with fixed local system size) on IBM Regatta 690+ and Blue Gene/L using up to 1024 (8192) processors respectively. The time per quantum operation is given as a function of the global system size

Since the number of operations per CPU is constant – adding one qubit doubles the size of the state vector and therefore the number of CPUs – ideal scaling would correspond to a line parallel to the x-axis. Indeed, for large systems one finds such an asymptotic behavior.

In principle, because of its “universality”, this gate level code can also be used to simulate physical systems, such as quantum spin models or models for physical realizations of quantum computers, by writing the time evolution of the physical system as a sequence of elementary quantum gate operations [1]. However, this approach is bound to be computationally inefficient for all but nontrivial physical systems. Instead, it is much more effective to allow for additional unitary transformations that implement operations such as the time evolution under a Heisenberg Hamiltonian in an optimal manner. This extension does not affect the intrinsic performance of the code.

Within the framework of gate level simulations, it is nevertheless possible to implement a basic error model which covers a) operational errors and b) decoherence [4, 5]. The underlying idea is that every single qubit operation can be generated from plane rotations and phase shifts. This decomposition allows introducing angle and phasing errors. Computation of controlled multi-qubit gates can be reduced to effective single qubit operations that act only on the part of the state vector whose control-bit(s) are set to |1>. To study decoherence effects, a bit-flip, a phase-flip or both are introduced with probability p/3 each. The state vector remains unchanged with probability 1-p.

Applying this error model to the Quantum Fourier Transformation and Grover’s quantum search algorithm, one finds that the QFT circuit is more robust to operational inaccuracies than Grover’s algorithm on comparable scales. Critical parameters can be derived which give a first estimate of tolerable error thresholds. These results will be compared with future calculations from dynamic simulations of quantum computer devices, taking into account the full time evolution according to a time dependent Hamiltonian describing both, the system and the environment.

References

We thank the DEISA Consortium (co-funded by the EU, FP6 project 508830), for support within the DEISA Extreme Computing Initiative (www.deisa.org).

[1] **Nielsen, M.A., Chuang, I.L.**
Quantum Computation and Quantum Information, Cambridge University Press, Cambridge, 2000

[2] **Di Vincenzo, D.P.**
http://arxiv.org/abs/quant-ph/0002077

[3] **De Raedt, K., Michielsen, K., De Raedt, H., Trieu, B., Arnold, G., Richter, M., Lippert, Th., Watanabe, H., Ito, N.**
Massively Parallel Quantum Computer Simulator, to be published in Computer Physics Communication, 2006

[4] **Niwa, J., Matsumoto, K., Imai, H.**
http://arxiv.org/abs/quant-ph/0201042

[5] **Arnold, G., Richter, M., Trieu, B., Lippert, Th.**
Improving Quantum Computer Simulations, to appear in the proceedings of the AQIS06 conference

• Guido Arnold
• Marcus Richter
• Binh Trieu

John von Neumann Institut für Computing (NIC) Forschungszentrum Jülich

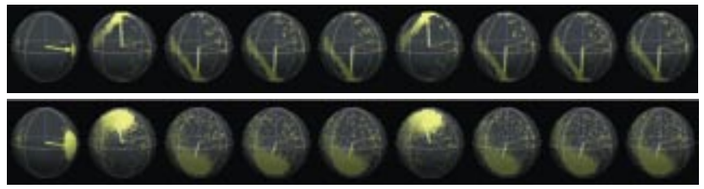


Figure 3: Grover’s quantum search algorithm: the 8 qubits encode the searched data base element k=17=10001 (the leftmost qubit is ancillary). The upper line corresponds to a subcritical error distribution while the parameter of the lower sequence is above the threshold

Direct Numerical Simulations (DNS) of Transitional Flow around Turbine Blades

Applications

Low-Pressure (LP) turbines are employed, for instance, in jet engines where they supply power to the fan and sometimes also to the first compressor stages. Recent increases in fan diameter require a higher work-output from the LP turbine at reduced rotational speed. The efficiency of a LP turbine strongly influences fuel consumption such that even small improvements have a significant effect. In order to be able to make such improvements, it is important to gain a better understanding of the behaviour of the flow around and the heat transfer to the blades of such turbines.

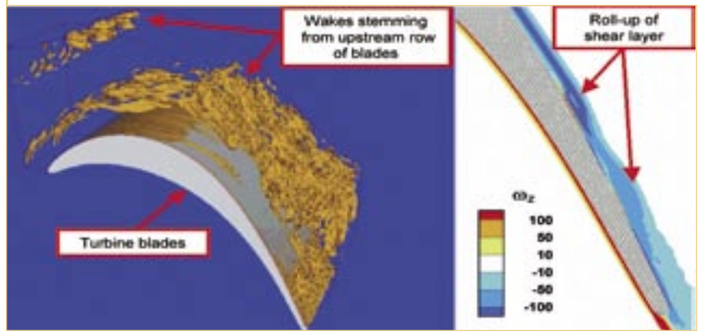


Figure 1: Case with separated-flow transition. Left: Vortical structures present in the incoming wakes above the suction side of the blade and in the near-wall turbulence at the trailing edge. The structures were made visible using the λ2-criterion. Right: contour plot at midspan of the phase-averaged spanwise vorticity near the trailing edge showing the roll-up of the separated shear layer

Because of the relatively low Reynolds number, the boundary layer flow around a LP turbine blade is often transitional. The location and length of the transition to turbulence and also the heat transfer to the blade is influenced strongly by the periodically passing wakes originating from the upstream row of blades, see Figure 1 (left part). The relatively high level of free-stream fluctuations favours two types of transition:

1) The first type is separation-induced transition. Because of the high-load on modern turbine blades, the boundary layer flow has a tendency to separate. Boundary layer separation influences the aerodynamical properties of blades, the heat transfer to blades and may even lead to mechanical failure. The separated shear layer rolls up and the rollers become unstable, transition occurs and the flow re-attaches (see Figure 1, right part). This process is strongly influenced by free-stream fluctuations and particularly wakes, which may suppress entirely separation.

2) The second, most common type is by-pass transition, where the impinging free-stream fluctuations trigger low-speed and high-speed streaks inside the otherwise laminar boundary layer. Low-speed streaks are unstable with respect to small-scale fluctuations and may give rise to the formation of turbulent spots which grow in size and merge as they are convected downstream, see Figure 2 (upper part).

In order to study both types of transition, a series of DNS was performed on the Hitachi SR8000 at LRZ. Typically, in a DNS all relevant scales of motion have to be resolved and this requires a large number of Grid points. In a Grid-refinement study of the simulation of a separating boundary layer flow along the suction side of a turbine blade with incoming wakes, Grids containing 10 million, 17 million and 25 million points, respectively, were employed. It was found that the large-scale fluctuations associated with the velocity-deficit of the wakes were sufficient to trigger a Kelvin-Helmholtz (KH) instability. As mentioned above, this instability manifests itself by the roll-up of the separated shear-layer into one or more rolls of rotating flow, see Figure 1 (right). Triggered by the small-scale fluctuations present in the wakes, transition to turbulence takes place and the rolls disintegrate. The impinging wakes were found to periodically completely suppress separation.

More recently, flow simulations of by-pass transition of the suction side boundary layer and heat transfer from the turbine blade to the outer flow were carried out on the Hitachi SR8000 using almost 100 million Grid points. The simulations had to run for a couple of months of clock-time in order to gather sufficiently converged turbulence statistics. The set-up of the simulations was selected to be largely in accord with experiments performed earlier at the University of Karlsruhe. To account for the background turbulence, besides the wakes also Grid turbulence had to be introduced at the inflow plane. The DNS was able to reproduce well the boundary layer transition scenario observed in the experiments. In the region where the suction-side boundary layer turned turbulent, the small-scale chaotic motion (see Figure 2) resulted in a significant increase in heat transfer. The increase in laminar heat transfer – that is the heat transfer from the blade along those parts were the boundary layer is laminar – caused by impinging free-stream fluctuations could, however, be only partially reproduced in the

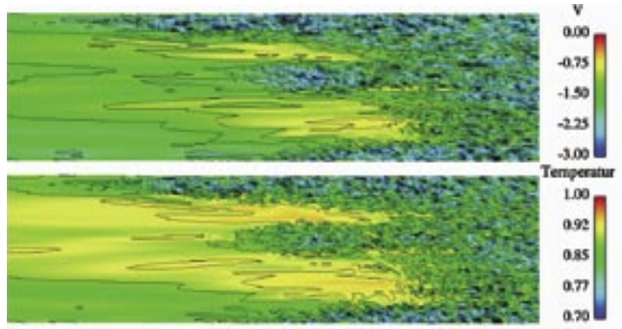


Figure 2: Case with by-pass transition. Snapshots at t = 7.35 of the flow field in a plane parallel to the suction surface of the blade. Above: contour plot of the instantaneous v-velocity showing the appearance a turbulent spot. Below: contour plot of the instantaneous temperature were small-scale temperature fluctuations can be seen in the regions were the boundary layer flow is turbulent

DNS. The cause of this discrepancy is not completely understood. An important factor might be a difference in the spectral contents of the oncoming free-stream flow specified in the DNS compared to the experiments.

To further investigate this problem, we are currently performing a more detailed study of laminar heat transfer. From experiments, it is known that free-stream fluctuations can be very effective in increasing heat transfer in the laminar boundary layer portion of a turbine blade. Further experiments have shown that to obtain such a significant increase in heat transfer, the boundary layer flow needs to be accelerating and the integral length-scale of the free-stream fluctuations needs to be about ten times the thickness of the boundary layer. The ongoing study is performed by simulating the accelerating flow over a heated flat plate with incoming free-stream fluctuations of which the integral length-scale and intensity is varied. On the next generation system at LRZ (HLRB-II) we plan to perform a Grid-refinement study and some very large-scale simulations of this flow problem.

• Jan Wissink
• Wolfgang Rodi

Institut für Hydromechanik
Universität Karlsruhe

Autumn 2006 • Vol. 4 No. 2 • inSiDE 11

Hadron Resonances in Lattice QCD

Quantum Chromodynamics (QCD) describes quarks and gluons and their interactions. It is one of the most complex theories, as it is highly non-linear and at the same time contains all the physics of relativistic quantum field theories. As a consequence, QCD has an extremely rich phenomenology. Large and very expensive high-energy accelerators and experiments are funded all over the world to investigate specific aspects of QCD (Jlab and BNL, USA; J-PARC, Japan; FAIR, Germany; ...). Unfortunately, in QCD it is only possible to obtain somewhat indirect information about quarks and gluons because they are confined: i.e., quarks and gluons cannot exist as free particles. Therefore, a solid theoretical understanding is especially important to deduce the properties of quarks and gluons from the observed complex bound states: the hadrons. To achieve this goal, the numerical (statistical) evaluation of relevant QCD-quantities on high-performance computers is an indispensable tool. For such a calculation one has to approximate the continuous space-time by a 4-dimensional space-time Grid, the so-called lattice. Here, we focus on only one aspect of lattice QCD, which was investigated by the BGR-Collaboration (Bern-Graz-Regensburg) on LRZ's Hitachi SR8000, namely, hadron resonances.

Just as an atom has excited states, so do hadrons. For example, the usual proton and neutron have orbital total angular momentum 1/2 but can be excited to states with higher angular momentum and there are also excitations with the same quantum numbers but higher masses. Each excited state is observed as a new instable hadron, called a hadron resonance. Just as one can deduce the form of the Coulomb potential from the atomic level structure, the quark-gluon interaction is linked to the spectrum of hadronic excitations. So the theory has two tasks, namely to show that one can reproduce the observed resonances and then to analyse the internal structure of these excited states. To do so, however, one has to overcome a fundamental problem: Lattice calculations are performed after the continuum theory is analytically continued to imaginary time. This operation turns the standard exponential factors exp(iEt) of quantum mechanics into exponentially decaying functions exp(-Et). If one inserts at some space-time point of the lattice a combination of quarks and gluons with the quantum numbers of, say, the proton, this trial function (called source or sink) will not be the exact quantum mechanical wave function of the proton, but rather a superposition of all possible resonance and multi-hadron states with these quantum numbers. Because the ground state has the smallest energy, all higher energy states become exponentially suppressed when propagated in time, and finally one obtains the exact many particle wave function of the ground state proton. To study resonances, one has to fight this exponential suppression. There are various methods to do so. We chose the so-called variational method: we used several (typically around 10) independent source and sink terms, propagated them in time accordingly and diagonalized a matrix of correlations. This results in certain combinations of the different source terms which represent the decoupled excited and ground states. The value of the corresponding coefficients provides information about the dominant components of the internal resonance wave function.

The computational challenges towards obtaining such results are the following:

1. Generation of a sufficiently large ensemble of independent field configurations with the correct statistical weight. The time needed for this subtask depends crucially on the parameters chosen but can easily reach 2,000 hours for 8 nodes (64 processors) of LRZ's Hitachi SR8000 and of the order of 100 configurations (for dynamical configurations).

2. Calculation of the propagators. This requires the inversion of very large sparse matrices with a typical dimension of $10^6 \times 10^6$ to $10^7 \times 10^7$.

3. Evaluation of the quark correlators which belong to the resonance of interest. This requires a convolution of the propagators.

Figures 1 to 4 show some typical results. All known hadrons are classified by a letter and a number. The letter stands for a specific combination of quantum numbers. For example π stands for the pion, e.g. the particle which binds protons and neutrons to form nuclei. The number stands for the mass of the excited state in MeV (eV is the unit of energy used in particle physics). The ground state pion has a mass

of 140 MeV and, e.g., the $\pi(1300)$ is an instable particle with the quantum numbers of the pion and a mass of roughly 1300 MeV. Figure 1 shows our results for this state.

On the x-axis we plotted the ground state pion mass squared and on the y-axis the mass of the $\pi(1300)$. The experimental values correspond to the red symbol. (The mass of the $\pi(1300)$ is not know exactly because this particle decays extremely fast.) The black and blue symbols represent mean results of our calculations, with error bars showing the expected uncertainty. Calculations are performed for several choices of unphysically large quark and consequently pion masses, in order to reduce the computational cost. The larger the quark mass, the more "classical" the dynamics and the cheaper the calculations. Thus, for a given amount of computer time one adjusts the quark masses such that one can still produce results with sufficient accuracy. These results then have to be extrapolated to the physical pion mass. So the question is whether the black symbols when extrapolated to the left are consistent with the red one. Obviously they are. The blue symbols are the results of a test calculation for a coarser lattice (a is the lattice constant). As they agree reasonably well with the black symbols we can be confident that the numerical discretization errors are small. Clearly one would like to have more CPU time to reach still smaller quark/pion masses, but the gap between the simulated pion masses and the real one has already grown rather narrow. Figures 2 to 4 show some more examples, namely the results for the ρ, ϕ, and kaon resonances, as well as the negative parity nucleon resonances. All of these results are, however, only the starting point for far more detailed studies (which constitute the main interest for experts). As we reproduce the correct mass values, the multi-particle wave-functions obtained simultaneously should be realistic, and they contain far more information than just the energy eigenvalue.

• Andreas Schäfer

Institut für
Theoretische Physik
Universität
Regensburg

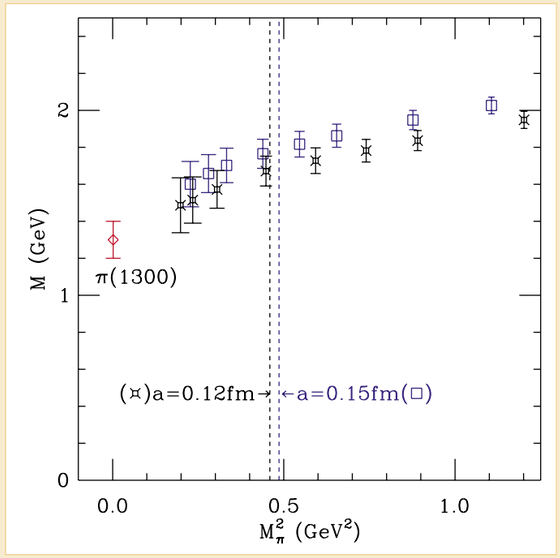
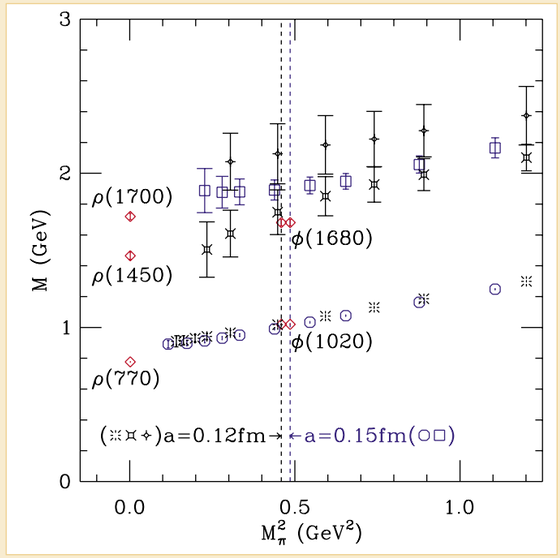
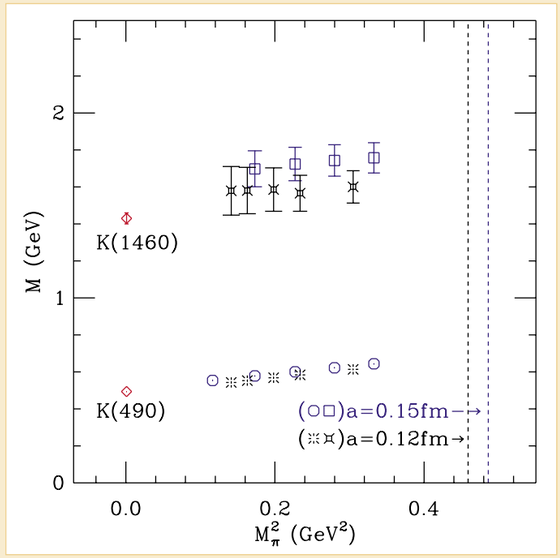


Figure 1:
Pion resonance modeling results. See the text for explanation





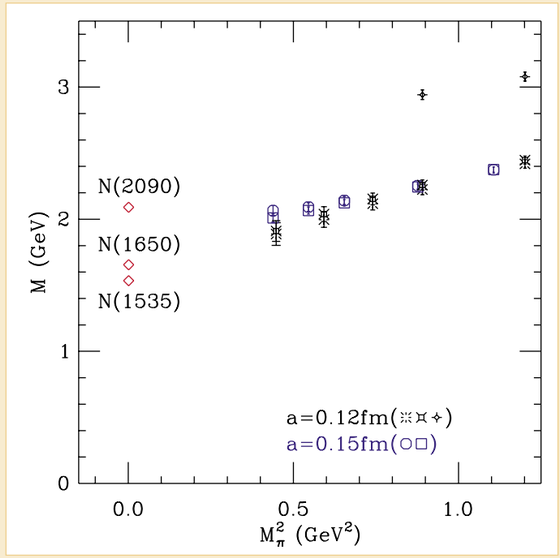


Figure 2 - Figure 4:
The same as Figure 1 for the ρ, ϕ, K and N* resonances

Scalable Parallel Trace-Based Performance Analysis

Introduction

To satisfy their increasing demand for computing power, advanced numerical simulations are required to harness larger numbers of processors offered by modern capability computing systems, such as the IBM BlueGene/L system JUBL at Forschungszentrum Jülich. Unfortunately, satisfactory speedup on many thousands of processors is extraordinarily hard to achieve. Sustained application performance is often significantly below the theoretical limit and leaves substantial room for optimization. However, tools that normally assist developers in the optimization process cease to work in a satisfactory manner when deployed on large processor counts.

Event tracing has been a well-established technique for post-mortem performance analysis of parallel applications. Time-stamped events, such as entering a function or sending a message, are recorded at runtime and analyzed afterwards with the help of software tools. In this context, automatic off-line trace analyzers, such as the EXPERT tool from the KOJAK toolset [1], can conveniently provide relevant information by automatically searching traces for complex patterns of inefficient behavior and quantifying their significance. In addition to usually being faster than a manual analysis, this approach is also guaranteed to cover the entire event trace and not to miss any pattern instances.

However, as the size of parallel systems and the number of processors used by individual applications rises, the traditional approach of sequentially analyzing a single global trace file becomes increasingly constrained by the large number of events. A new project aimed at overcoming these limitations, SCALASCA [2], started at the beginning of 2006 in a Helmholtz-University Young Investigators Group established between Forschungszentrum Jülich and RWTH Aachen University.

In this article, we outline how the pattern search can be done in a more scalable way by exploiting both distributed memory and parallel processing capabilities available on modern large-scale systems and discuss first empirical results. For a more detailed discussion, the interested reader may refer to [3].

Parallel Analysis Approach

Instead of sequentially analyzing a single and potentially large global trace file, we analyze multiple local trace files in parallel based on the same parallel programming paradigm as the one used by the application under investigation. For simplicity, we currently have restricted ourselves to handle only single-threaded MPI-1 applications. The analyzer, which is an MPI application in its own right, is executed on as many CPUs as the target application. This allows the user to run it after the target application within a single batch job, which avoids additional waiting time in the batch queue. The parallel analyzer uses a distributed memory approach, where each process reads only the local trace data that were recorded for the corresponding process of the target application. This addresses scalability specifically with respect to larger numbers of processes. Since the size of local traces can be limited by selective tracing – i.e., by recording events only for code regions and time intervals of particular interest – we assume that the local trace data can be completely held in the main memory of the compute nodes. This has the advantage of having efficient random-access to individual events, whereas this is often not the case when dealing with a global trace file.

The actual analysis can then be accomplished by performing a parallel replay of the application's communication behavior. The central idea behind this approach is to analyze a communication operation using an operation of the same type. For example, to analyze a point-to-point message, the event data necessary to analyze this communication is also exchanged in point-to-point mode between the corresponding analysis processes. To do this, the new analysis traverses local traces in parallel and meets at the synchronization points of the target application by re-enacting the original communication.

The replay-based analysis approach can be used to search for a large number of inefficiency patterns. Our current prototype supports all but one rarely significant MPI-1 pattern offered by the original sequential EXPERT tool. Two examples of these patterns are diagrammed in Figure 1. Their detection algorithms will be used to illustrate the parallel analysis mechanism below.

As an example of inefficient point-to-point communication, consider the so-called Late Sender pattern. Here, a receive operation is entered by one process before the corresponding send operation has been started by the other. The time lost is therefore the difference between the timestamps of the enter events of the MPI function instances which contain the corresponding message send and receive events. The complete Late Sender pattern consists of four events, specifically the two enter events and the respective message send and receive events.

During the parallel replay, the detection of this performance problem is triggered by the point-to-point communication events involved (i.e., send and receive). That is, when a send event is found by one of the processes, a message containing this event and the associated enter event is created. This message is then sent to

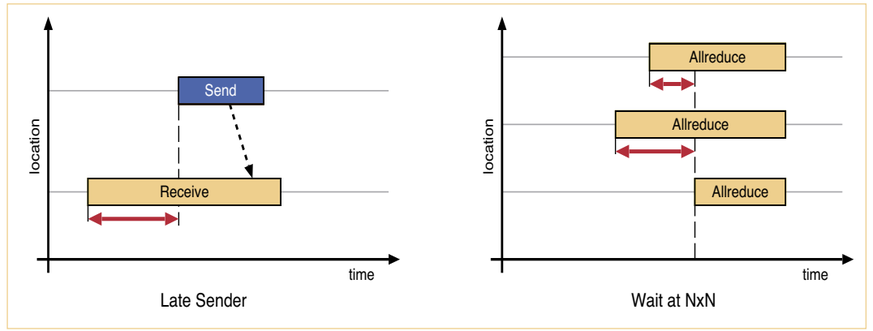


Figure 1: Two examples of inefficient program behavior; one for point-to-point communication (Late Sender) and one for collective operations (Wait at N x N)

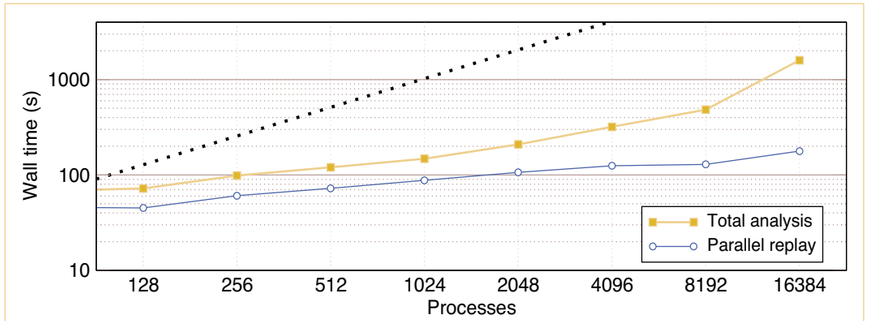


Figure 2: Wall-clock execution times for SMG2000 analysis using the new prototype at a range of scales. Linear scaling is the bold dotted line

the process representing the receiver using a point-to-point operation. To ensure the correct matching of send and receive events, equivalent tag and communicator information are used to perform the communication.

When the receiver reaches the receive event, the aforementioned message containing the remote constituents of the pattern is received. Together with the locally available constituents (i.e., the receive and the enter events), a Late Sender situation can be detected by comparing the timestamps of the two enter events and calculating the time spent waiting for the sender.

The second important type of communication operations are MPI collective operations. As an example of a related performance problem, consider the detection of the Wait at N x N pattern, which quantifies the waiting time due to the inherent synchronization in N-to-N operations, such as MPI_Allreduce.

While traversing the local trace data, all processes involved in a collective operation will eventually reach their corresponding collective exit events. After verifying that it relates to an N-to-N operation, accomplished by examining the associated region identifier, the analyzer invokes the detection algorithm, which determines the latest of the corresponding enter events using an MPI_Allreduce operation. After that, each process calculates the local waiting time by subtracting the timestamp of the local enter event from the timestamp of the enter event obtained through the reduction operation. The group of ranks involved in the analysis of the collective operation is easily determined from the communicator of the original collective operation.

Results

To evaluate the effectiveness of parallel analysis based on a replay of the target application's communication behavior, a number of experiments with our current prototype implementation have been performed at a range of scales. Measurements were taken on the 8-rack IBM BlueGene/L system JUBL using a dedicated partition consisting of all of the compute nodes for the parallel analyses.

Figure 2 charts wall-clock times for the analysis of ASC benchmark SMG2000 traces with a range of process numbers (the 8-fold doubling of process numbers necessitates a log-log scale

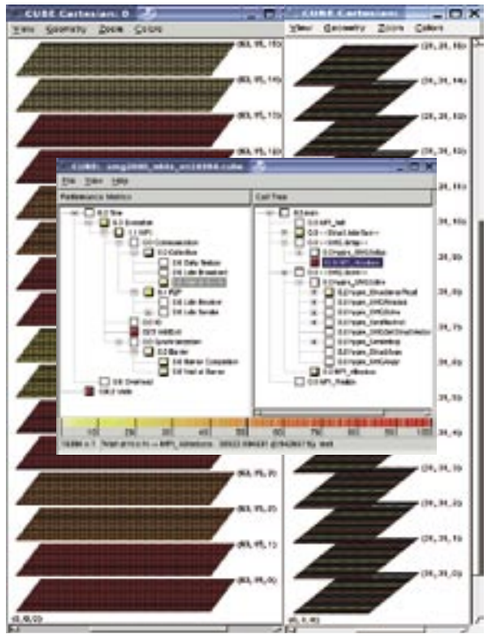


Figure 3: Analysis report for ASC SMG2000 on 16,384 processors of BlueGene/L highlighting the distribution of the Wait at N x N performance metric on the physical machine topology distribution (left) and MPI process topological distribution (right) for a particular call path

to show the corresponding range of times). The figure shows the total time needed for the parallel analysis (including trace reading and writing a complete analysis report) and the time taken by the parallel replay itself without file I/O. Due to the often considerable variation in the time for file I/O (e.g., depending on overall file-system load) the times reported are the best of several measurements.

The parallel replay for the largest set of execution traces from 16,384 SMG2000 processes, amounting to over 40,000 million events (230 GBytes of trace files), took less then 3 minutes. With the latest improvements for merging the analysis results (in comparison to [3]), which is reflected in the curve showing the total analysis time, the full analysis for 16,384 processes completed in less than 30 minutes. Although the overall analysis time is dominated by file I/O, the new approach is orders of magnitude faster than the corresponding sequential analysis carried out by the EXPERT tool, thereby enabling analyses at scales that have been previously inaccessible. A screenshot with analysis results for 16,384 processes is shown in Figure 3.

References

[1] **Wolf, F., Mohr, B.**
Automatic performance analysis of hybrid MPI/OpenMP applications,
Journal of Systems Architecture 49(10-11), pp. 421-439, 2003

[2] **http://www.scalasca.org**

[3] **Geimer, M., Wolf, F., Wylie, B.J.N., Mohr, B.**
Scalable Parallel Trace-Based Performance Analysis, Proceedings EuroPVM/MPI 2006, Springer LNCS 4192, pp. 303-312, 2006

• Markus Geimer[1]
• Felix Wolf[1,2]
• Brian J.N. Wylie[1]
• Bernd Mohr[1]

[1] John von Neumann Institut für Computing (NIC) Forschungszentrum Jülich

[2] Fachgruppe Informatik RWTH Aachen

The Inner State of a Supercomputer: Getting Insight from Performance Counters

The ranking of machine power is still based on peak performance or benchmarks but not on the actually delivered (floating point) operations over a given timescale in every day operation. However, computational workloads and communication patterns for scientific applications vary dramatically, depending in part on the nature of the problems the applications are solving. Recent works show that the characteristics of scientific applications differ significantly and the practical use of ranking or predicting system performance via single metrics and benchmarks such as High Performance LINPACK, STREAM or the HPC Challenge Benchmarks is quite limited. Only if enough information about the target applications is acquired, some simplified metrics may be combined and weighted appropriately to predict performance with reasonable accuracy. As hardware counters are ubiquitously available in modern processors, we argue that monitoring all applications in a system is an adequate way to get enough information on the system and finally achieve a good understanding of the potential of a given architecture.

Processors	128 x Itanium2 Madison 9M
Clock	1,6 Ghz
Peak per Processor	6,4 Gflop/s (4 FP Ops per cycle)
L3 Cache	6 MB
L3 Cache Line Size	128 Byte
Bandwidth to L3	32 Gbyte/s
Bandwidth to Memory, shared by two Processors	6,4 Gbyte/s (4 Byte/cycle)

Table 1: Characteristics of the Altix 37000 Bx2

In our study we monitored all applications on one of our HPC systems, an Altix 3700 Bx2 with 128 processors (see Table 1).

Samples of the most important hardware counters are taken from all processors in 5 minute intervals and are stored and subsequently processed in a database. The sampling time for each measurement was 1 second. Hence, more than 120,000 "fine grained" measurements were taken in a two month interval. We must stress the point that we measured the everyday performance of a system including badly optimized programms, test runs etc. Nevertheless the results give us deep insight into the inner state of the system.

Average Performance

The average values of some of the most important counters and their maximum increment per cycle are given in Table 2. Ratios between the counters are given in Table 3.

One of the key architectural features of the Itanium2 processor is to execute multiple instructions per clock. The burden for this explicit parallelism is put onto the compiler which encodes multiple operations for multiple functional units in every instruction. Each of the multiple operation instructions is called a bundle. Three instructions fit into a bundle, and two bundles can be executed simultaneously in each cycle. If slots can not be filled with "useful" instructions, due to

	Max Incr.	Mesurements		
Instructions Retired	per cycle	per cycle	% of Peak	per second
Instructions Retired	6	1,938	32,3 %	3.10E+09
Nops Retired	6	0,666	11,1 %	1.07E+09
Useful Instructions	6	1,272	21,2 %	2.03E+09
Floating Point Operations	4	0,617	15,4 %	9.86E+08
Stalled Cycles	1	0,545	54,5 %	8.72E+08
Back_End_Bubbles	1	0,545	54,5 %	8.72E+08
Loads Retired	4	0,250	6,3 %	4.00E+08
Stores Retired	2	0,072	1,8 %	1.15E+08
Loads+Stores, 16 Byte assumed, Bytes	48	3,866	12,1 %	6.19E+09
Loads+Stores, 8 Byte assumed, Bytes	24	1,933	6,0 %	3.09E+09
L2_References	4	0,274	6,9 %	4.39E+08
L2_Misses	0.16	0,011	6,8 %	1.70E+07
L2_Misses, Bytes	20	1,363	6,8 %	2.18E+09
L3_References	0.16	0,014	8,9 %	2.22E+07
L3_Misses	0.03	0,004	12,2 %	6.08E+06
L3_MISSES, Bytes	4	0,487	12,2 %	7.78E+08
w.r.t. Second Processor	2		24,4 %	

Table 2: Performance counters, overall average

Ratio	
Useful Instructions / Unstalled Cycles	2,79
Useful Instruction / (Loads+Stores)	3,95
FP Inst / Inst retired	0,32
FP Inst / Useful Inst	0,48
FP Inst / (Loads+Stores)	3,95
FP Inst / (L2 Misses)	57,91
FP Inst / Byte (L2_Misses)	0,45
FP Inst / (L3 Misses)	162,21
FP Inst / Byte (L3_Misses)	1,27
L3_References / L3_Misses	3,65
L2_References / L2_Misses	25,75

Table 3: Ratio of counters

dispersal constraints, NOPS are inserted. If the processor would not be stalled, six instructions per cycle could be delivered out of which two could be floating point instructions. Taking a fused multiply-add operation into account, four floating point operations per cycle could be delivered.

Our measurements show that due to the EPIC system architecture approximately two instructions are retired per cycle, but this is only one third of the maximum number. 11,4 % of all instructions are NOPS, showing that there is still room left for more instruction parallelism but the compilers, algorithms or the resource requirements are not suited to deliver this.

The "useful instructions" delivered can be computed by subtracting the NOPS from the retired instructions, leaving only 21 % of maximum. On average the system runs with approx. 1 GFlop/s per processor, or with 15 % of its peak floating point performance, a value which is very good compared to results of former RISC systems which delivered 8-10 % of peak. The reason for this relatively high value is the good exploitation of the L2 and L3 caches. The large on-die L2 and L3 caches provide a significant performance potential.

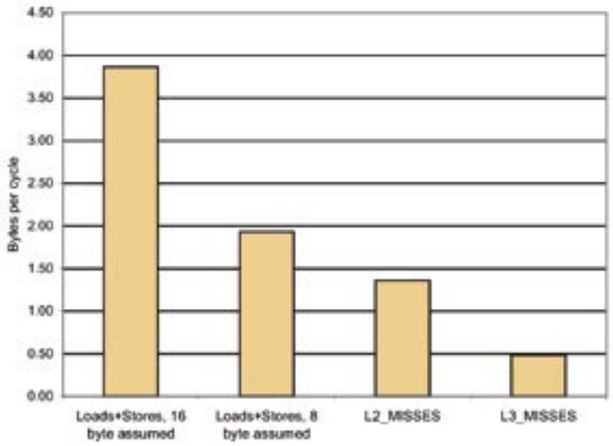


Figure 1: Transfer rates in the memory hierarchy

The processor requires functional unit stalls to assure that results are computed correctly. The Itanium2 backend has five counters for the various types of stalls in the processor backend. Each counter is associated with a given stage in execution pipeline. The Back_End_Bubbles accumulate the cycles where the instructions pipeline stalls for any reason. We must realize the fact that 54 % of all cycles are stalled.

Memory Hierarchy

With the given cache line sizes of the processor we can calculate the consumed bandwidth to the next level in the memory hierarchy, especially the data rate between memory and L3 cache. For each byte that is transferred between the memory and the L3 cache, 1,27 floating point operations are performed. Unfortunately the measured load and store counters provide no direct way for an interpretation as bandwidth achieved between L2 or L1 cache and the registers since there may be single or paired load instructions of various sizes. As an estimate we can assign 8 Byte or 16 Byte to each load/store instruction and get a rough picture for what is happening in the memory hierarchy (see Figure 1). At least 75 % of all loads and stores can be satisfied from the high levels of the memory hierarchy.

The biggest surprise in our results was to find that on average the bandwidth to memory (expressed by L3 misses in bytes per cycle) is not saturated and only 12,2 % of it are used (24,4 % if we consider that two processors share the same memory channel). In Figure 2 we plotted the floating point operations against the L3 misses of all samples. This kind of figures from the counters act like X-ray pictures which show us the backbones of the system. It is obvious that high floating-point performance has only been achieved when the L3 misses have been less than 0,6 byte/cycle. A second pattern is the linear increase of performance with L3 Misses in the lower part of the figure. Approximately 75 % of all samples fall below the inclined line. From benchmarks with known counter profiles one can deduce that this is the region where typical loops tend to reside.

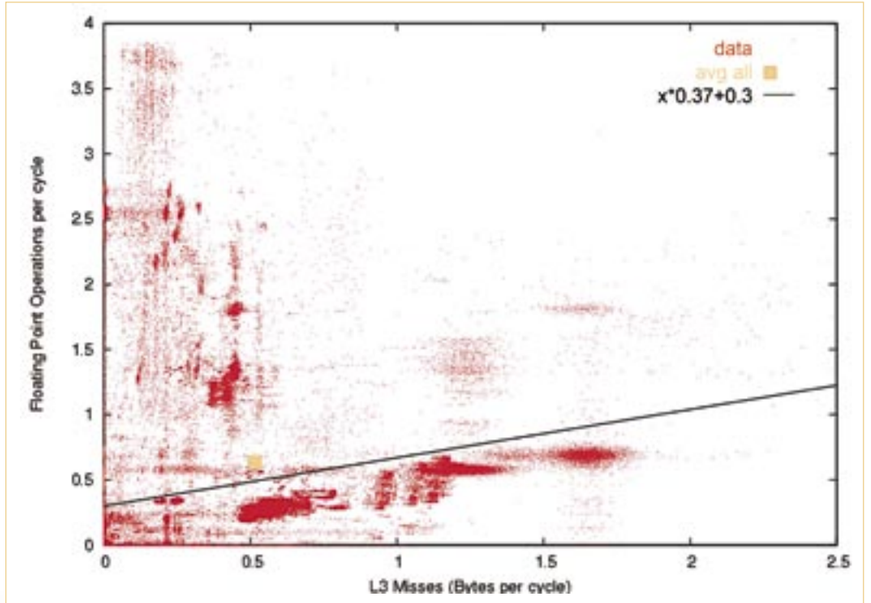


Figure 2: Floating point operations and L3 misses. The golden square indicates the average of all 120,000 measured samples (see Table 2)

Conclusions

Caches have now reached a size where many applications can draw significant advantages from them. The measurements show that overall bandwidth to memory is not fully used, therefore emphasis must be put on latency hiding, on evenly distributing the workload to memory, and improving temporal locality for nested loop execution.

• Matthias Brehm
• Reinhold Bader
• Richard Patra

Leibniz-
Rechenzentrum

The int.eu.grid Project

On May 1, 2006 the two-year EU-funded Interactive European Grid project [1] started. It extends on the knowledge gained within the successful CrossGrid project, which set up a Grid infrastructure in Europa. The partners in the project include 13 leading institutions in seven countries as shown in Figure 1.

Figure 1: Partners in the int.eu.grid project

Interactive MPI-Parallel Applications in the Grid

The main purpose is, as suggested by the project name, to support large-scale interactive, parallel applications within a European Grid, allowing computation on distributed compute resources, as well as interactive visualization and simulation steering on selected applications using the Migrating Desktop (MD) software. The major focus with regard to the Grid-software stack deployed is compatibility and interoperability at the middleware-level with LCG and the EGEE-II project, based on gLite-3.0.

The computational infrastructure operating at the eleven computational sites includes several production clusters based mainly on Intel Xeon and dual AMD Opteron clusters with a total of around 500 processors. The Message Passing Interface (MPI) will be used as the main parallelization paradigm for intra- and inter-cluster computing.

MPI-Development within the Project

HLRS, being a partner in the Open MPI consortium, will be responsible for the integration and support of the MPI infrastructure in the testbed and production clusters. As parallel middleware libraries, Open MPI, PACX-MPI [3] and the analysis and checking tool Marmot [4] will be supported on the platforms. The aim of Open MPI is to build a completely new MPI implementation based on the experience of several projects, such as LAM/MPI, FT-MPI, LA-MPI and PACX-MPI. It offers full MPI-2 functionality, and by a modular design, allows easy integration of different subsystems for different implementations, e.g. startup mechanisms supporting PBS, Bproc, POE and standard ssh-startup mechanisms. Other startup mechanisms for Grid-environments are being developed.

This will allow users seamless startup of parallel applications from the researchers desktop using MD on one or more of the available clusters with direct and fast interactive steering of these applications. Coupling of these computational resources will be done through the fast European Géant network.

To support researchers, these applications have to be optimized to run in such an environment; the major concern is to analyse promising parallel applications using performance analyzers such as Opt, Vampir or Paraver and guide the application developer in the process. The MPI analysis and checking tool Marmot will help the application developer in finding programming errors with regard to parameters, non-portable usage of MPI or even deadlocks. While the MPI library itself may e.g. detect non-freed request handles, in a high-performance production environment, internal parameter checking may be switched off on production clusters. Dissemination and developer support will be handled in lectures and tutorials over the AccessGrid infrastructure installed at the various sites. Effectiveness of such tutorials has already been proven with experience gained in the so-called “HPC-Europa surgeries” within the HPC-Europa project, where guest scientists at the various centers would attend a video conference on specific themes.

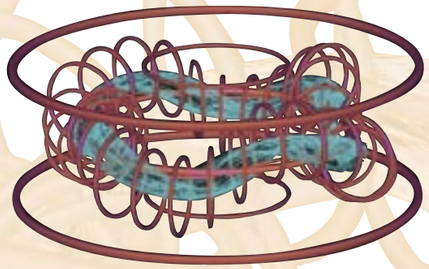


Figure 2: Results of the plasma fusion simulation

The applications developed in the project’s Joint-Research Activity are from various areas, including high-energy physics (see Figure 2), ultrasound computer tomography used in medicine as well as applications in environmental research.

Project Overview

Int.eu.grid (contract number 031857) started on the May 1st, 2006 for a total of 24 months. The 13 partners from seven countries are the Consejo Superior de Investigaciones Científicas (CSIC), Laboratório de Instrumentação e Física Experimental de Partículas (LIP), Poznan Supercomputing Center (PSNC), Forschungszentrum Karlsruhe (FZK), Universidad Autonoma de Barcelona (UAB), Akademickie Centrum Komputerowe (CYFRONET), Institut für Graphische und Parallele Datenverarbeitung (GUP), Trinity College Dublin (TCD), Centro Tecnológico de Supercomputación de Galicia (CESGA), Ústav Informatiky Slovenská Academia (IISAS), Interdisciplinary Centre for Mathematical and Computational Modelling (ICM), Instituto de Biocomputacion y Fisica de Sistemas Complejos (BIFI), and the Höchstleistungsrechenzentrum Stuttgart (HLRS). Further information may be obtained from the following sites:

References

[1] **Int.eu.grid-Webpage**
http://www.interactive-grid.eu

[2] **Gomes, J., David, M., Martins, J., Bernardo, L., Marco, J. et al.**
Experience with the international testbed in the CrossGrid project,
LNCS, Vol. 3470, pp. 98-110, Springer, 2005

[3] **Keller, R., Krammer, B., Müller, M.S., Resch, M.M, Gabriel, E,**
Towards efficient execution of MPI applications on the Grid: porting and optimization issues,
Journal of Grid Computing, Vol. 1 (2), pp. 133-149, 2003

[4] **Krammer, B., Müller, M.S.**
MARMOT – an MPI Analysis and Checking Tool,
inSiDE Vol. 2 No. 2, 2004

• Rainer Keller

Höchstleistungs-
rechenzentrum
Stuttgart

Autumn 2006 • Vol. 4 No. 2 • inSiDE

Projects

OMII-Europe – towards Grid Interoperability

Deployments of Grid infrastructures and Grid software distributions have proliferated over last few years. No longer are they regarded as elaborate networks of resources with interest confined solely to the academic and research worlds. Instead, they are now regarded as the most appropriate and sometimes the only way to solve both scientific and commercial problems. The growth in interest of Grid infrastructures has been matched by groups rushing to develop the required software distributions. This enthusiasm to develop solutions for specific user groups has resulted in separate development of a number of Grid software distributions, with the consequence that multiple "islands" of Grid infrastructures now exist with little in the way of interoperability between these "Grid islands".

OMII-Europe is an EU project which has been established to address in large this interoperability issue, by sourcing key software components for Grid applications and to ensure that these components interoperate across heterogeneous Grid middleware platforms. A key attribute of the project is it's backing of open standards for common Grid services and the associated development of robust, quality-assured components that meet these standards for the gLite, UNICORE and Globus Grid software distributions. OMII-Europe further places emphasis on the re-engineering and porting of existing tried-and-tested software components, rather than on the development of new technology.

OMII-Europe is thus in the process of becoming a repository of quality-assured service-level Grid components capable of running on existing major Grid infrastructures. OMII-Europe is also in the unique position of being able to offer impartial advice on both Grid services and Grid distributions.

Middleware Platforms

OMII-Europe has chosen three Grid middleware platforms as its primary focus, based on the observation that these three platforms are widely used in European Grid infrastructures. These are gLite, UNICORE and Globus. In addition, OMII-Europe is collaborating with Chinese partners, who will make the selected services available on the CROWN Grid distribution.

- **gLite** is a complete set of middleware components developed within the EGEE project primarily to support computation required for the Large Hadron Collider project at CERN, but with dependent users in other domains. gLite is based on the pre-web-services Globus Toolkit 2.4 and is available in open source form from CERN.

- **UNICORE** is also a set of middleware components developed within various European and German national projects. These components provide a service-oriented solution to Grid computing. UNICORE is deployed at many supercomputer sites, in particular those available through DEISA.

- **Globus** is the world-leading open-source platform for Grid computing. It is a complete set of middleware components. Its pre-web-services releases, culminating in version 2.4 (properly referred to as GT2.4 or Globus Toolkit 2.4) is widely used in major projects in Europe as well as globally. The more recent web services version of Globus Toolkit (GT4) contains the whole of the pre-web-services components, for backward compatibility, thus allowing projects to evolve over time to a web services solution.

Service-level Components

There are five basic service-level components for establishing Grids that OMII-Europe is initially focusing on making available on all the above Grid platforms. Achieving common services across these Grid distributions is the first step towards infrastructure interoperability. Not surprisingly, these services also represent some of the most advanced developments in the open standards area in Grid computing. The services are:

1. A Basic Execution Service (BES) supporting JSDL. The BES specification from GGF is complemented by the Job Submission Description Language (JSDL) specification. The intention of these specifications is to create an abstraction from the operating system and the cluster controller on which the job is to run. By deploying this component on the three platforms OMII-Europe will show interoperability between them at the level of jobs being submitted to each without alteration.

2. A Data Integration Service, specifically OGSA DAI. OGSA DAI (the Open Grid Services Architecture – Data Access and Integration Service) is currently distributed either standalone or as part of the Globus or OMII-UK distributions. The service federates data resources with different support mechanisms (relational databases, XML databases or flat file systems) allowing uniform access across these resources. OMII-Europe will show interoperability between the Grid distributions at the level of data being accessed and updated without specialization of the query.

3. A Virtual Organisation Management Service, specifically VOMS. The Virtual Organisation Management Service (VOMS) is an authorization service which is currently available for use on either gLite or Globus. The OMII-Europe project is making it available on UNICORE and testing it for interoperability across all platforms.

4. An accounting service, based on the forthcoming GGF RUS specification. The Resource Usage Service (RUS) specification from GGF is the chosen standard for accounting services in Grid solutions. The service is used to track usage in order that it can be charged. Since the specification is still in development, candidate components are under review by OMII-Europe.

5. A portal capability, specifically GridSphere. The GridSphere Portal framework provides an open-source portlet-based web portal. GridSphere enables developers to quickly develop and package third-party portlet web applications that can be run and administered within the GridSphere portlet container.

These five services above do not comprise all the required services for Grid infrastructures. They do however comprise a set of essential services which need to be common across all Grid distributions to facilitate interoperability. Furthermore, OMII-Europe is exploring the need for additional components and has thus established an activity to evaluate candidate components, including components available from its Chinese partners, and specifically from CROWNGrid.

Solutions

The solutions that OMII-Europe will advocate imply an architecture which assumes deployment of one or more of the chosen middleware platforms and then the deployment of one or more of the quality-approved components on those platforms.

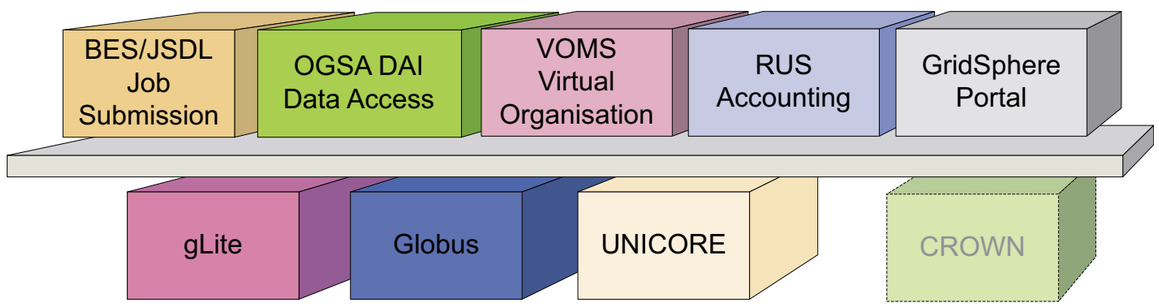


Figure 1: OMII-Europe components and platforms

Figure 1 shows each of the platforms and each of the components. Any particular site or location will have chosen one of the platforms and installed one or more of the components.

The typical application scenario will then establish, where necessary, interoperability between these platform/component deployments. For example, one may have a JSDL service deployed on UNICORE in one location and another JSDL service deployed on gLite in another location. A third location may support GridSphere which manages the workflow of submitting jobs to one or other of these services, as established by some decision making in that workflow.

This deployment anticipates a security solution which federates across the separate deployments. Both VOMS and the chosen accounting component are essential parts of this security solution. Although all three platforms, and VOMS, use X509 certificates, the extent to which the tokens used by the chosen platforms federate is an important aspect of interoperability and one on which OMII-Europe is progressively reporting.

Where to get more Information

OMII-Europe maintains a website at http://OMII-Europe.org from which more information can be obtained. In particular, there are brief introductions and links to each of the three platforms gLite, UNICORE and Globus and to the five components BES/JSDL, OGSA DAI, VOMS, RUS, and GridSphere. These pages are regularly updated by OMII-Europe and recommend, in particular, where to go to obtain implementations and documentation.

• Alistair Dunlop

The University of Southampton
UK

New National Supercomputing System at LRZ: SGI Altix 4700

After being in continuous user operation for six and a half years, the Hitachi SR8000 supercomputer located in the old LRZ building in Munich was retired from service at the end of June, 2006. It has been replaced by a considerably more powerful machine, an Altix 4700 from SGI; this system, comprised of 4096 Intel Itanium2 CPUs in its first installation phase, has been installed in the top floor of the new LRZ Compute-Cube on the Garching campus area. The performance characteristics of the new system are impressive: with a peak performance of more than 26 trillion floating point operations per second LRZ now again can provide competitive computing performance on a Europe-wide level to German scientists. The system's memory size also is gigantic: more than 17 Terabytes of RAM will enable very extensive and novel simulations to be performed.

Figure 1: SGI 4700 during installation

Advantages of the New System

Apart from the high performance the new system provides a broad spectrum of advantages, which in combination enable a very efficient usage. The most important of these are listed in the following:

1. The system is subdivided into sixteen partitions of 256 processors each; each partition has full access to its complete shared memory area of 1 TBytes size. No other system throughout Europe presently supports this shared memory size. Suitably parallelized programs can also use multiple partitions. Later in time – particularly with installation of phase 2 of the system – the size of a partition will further increase.
2. The system provides high aggregate bandwidth to memory, since each processor is assigned a memory channel of its own, where the bandwidth of a single channel is 8,5 GBytes/s. As a result, data intensive simulations can be executed efficiently. And since each processor is provided with a 6 MBytes large fast cache memory, some applications can expect a performance increase over-proportional to the number of CPUs requested.
3. The background storage needed for storing and post processing result datasets has been configured with high requirements on quantity as well as quality: For large files more than 300 TBytes of disk storage are available, a large fraction of which is provided as a single file system. The bandwidth available for transferral of data from main memory is 20 GBytes/s; this means that it is possible to write the complete memory content of the system to disks within a quarter of an hour.
4. 40 Terabytes of high-quality Network-Attached (NAS) disk Space are available for the users' home directories with program, configuration and small testing files. This disk area is also accessible from the outside world. It provides high transaction rates so as to guarantee efficient processing of small files.
5. Since the system makes use of standard Itanium2 CPUs from Intel and the widely used Linux operating environment is deployed, a large spectrum of software packages is available which can be provided without great porting effort. For those programs written by the scientists themselves a complete development environment is available, which provides a nearly seamless workflow going from the personal computer or cluster to the new supercomputing system.

Table 1 provides an overview of the relevant parameters of the new system. The configuration shown here is valid until 2007; at that point phase 2 will start, which will replace the single-core by dual-core processors with a smaller cycle time. Furthermore, further disk space and main memory will be added; as a result the system as a whole will provide nearly double the performance of phase 1.

Processors	4,096
Peak Performance	26,2 Teraflops/s
Memory	17,2 Terabytes
Memory Bandwidth	34,8 Terabytes/s
Total Disk Space	340 Terabytes
Latency of Interconnect	1-6 microseconds

Table 1: Altix 4700 performance parameters

System Architecture

The system architecture of the Altix 4700 may be characterized as distributed shared memory architecture. This means that the globally accessible main memory is distributed among the system nodes. Memory controllers on the compute nodes enable the cache coherent access to the memory from all processors. Depending on whether accesses happen to physically local or remote memory banks one obtains varying access latencies and bandwidths. Hence, this type of architecture is also known as cache-coherent non-uniform memory access (ccNUMA). The efficient exploitation of this kind of architecture is still a challenge for the programmer of shared memory applications, but also provides considerable flexibility of usage.

System Nodes

Single system nodes of the Altix 4700 are either equipped with processing units ("compute blade") or with I/O functionality ("I/O blade"). All node types are implemented as blades, a flat system board design with integrated power and cooling units. The blades are interconnected via SGI's NUMALink 4 technology to form a shared memory system. A compute blade consists of an Itanium2 processor chip and a memory controller which connects the processor to the local memory, as well as two Numalink channels which plug into the interconnect (see following figure).

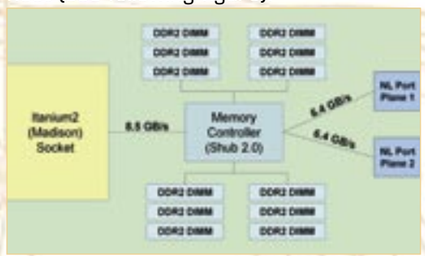


Figure 2: Structure of compute blade

The processors are clocked at 1,6 GHz and have two fused multiply-add units, resulting in a peak performance of 6,4 GFlops per CPU. Each processor is additionally equipped with 256 kBytes L2 cache as well as 6 Mbytes L3 cache. In contrast to the main memory the caches are operated at full processor frequency, thereby allowing very high application performance for programs which are designed for high re-use of data within the cache. On the system installed at LRZ, the memory banks of the batch partitions contain 4 GBytes on each compute blade; the interactive partition even provides 8 GBytes per blade.

An I/O blade consists of a cache coherence interface (TIO chip) and an ASIC which provides standard I/O functionality like PCI-X or PCI Express.

NUMALink Interconnect

The NUMALink network connects the nodes of the Altix 4700 with each other. It differentiates itself from other networking technologies by the fact that the state of the complete memory is visible to each processor at any time. Furthermore, low-latency loads of data from remote memory banks are possible. The interconnection network consists of 8-port routers, 8-port meta-routers and cable connections between the node boards with the routers, and between routers and meta-routers. Each cable is capable of providing a bandwidth of 6,4 GBytes/s (3,2 non-blocking in each direction). Routers and meta-routers are implemented as non-blocking crossbar units and provide 8 NUMALink ports (8 entry and 8 exit ports). The basic building block for a partition is a blade chassis which can hold 10 blades, of which 8 can be processor blades. The NUMALink ports are connected to the back-plane, which provides the internal connection of the 10 blade slots as well as to other blades. Figure 3 shows the topology of the chassis' backplane.

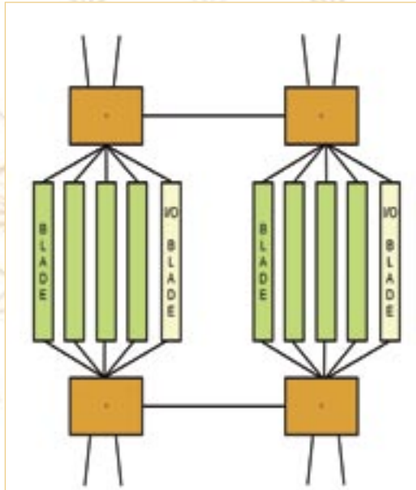


Figure 3: NUMAlink interconnect at the lowest hierarchy level

Figure 4 indicates how a 256 processor partition is connected into a powerful shared memory node via a two-tiered hierarchy of meta-routers; each blue rectangle represents a building block of 5 blades (4 processor blades). For communication across partition boundaries there are also NUMALink 4 connections available; these are however only implemented as a so-called "mesh topology" and therefore provide less bandwidth.

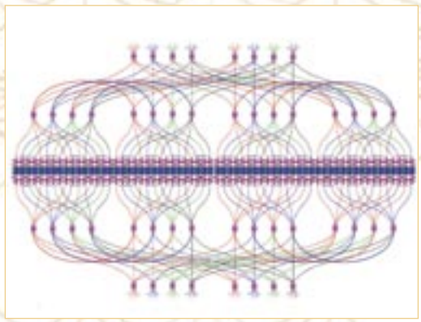


Figure 4: NUMALink topology of a 256 processor partition

Software Environment

Operating System

The Altix 4700 is operated under a standard Linux environment. The distribution used is Novell's SuSE Linux Enterprise Server (SLES 10), with add-on packages provided by SGI for the deployment of large systems in a computing centre: Apart from the HPC NUMA tools, the Message Passing Toolkit (MPT) and the Scientific Subroutine Library (SCSL) these are e.g., storage management software like the XFS file system and its cluster extension CXFS and CXVM, accounting packages and the performance co-pilot for surveillance of the system.

Compilers, Tools, and Libraries

For the generation of well-optimized binaries from Fortran, C or C++ sources the Intel compilers are provided which are capable of exploiting the explicit parallel instruction architecture of the Itanium2 processor. Furthermore, OpenMP-based code generation of shared-memory parallel programs is supported in a standard-conforming manner.

As an alternative to the SCSL implementation of linear algebra functionality (BLAS, LAPACK, and FFT) in the SCSL it is also possible to use Intel's Math Kernel Library (MKL); the latter also provides fast vectorized versions of special mathematical functions and solvers for sparse matrices. For the analysis of the run-time behaviour of MPI-parallel applications the Intel Tracing Tools are available; and for performance measurements on serial programs VTune – which in its newest release includes an Eclipse-based graphical user interface – is provided on the interactive node of the system.

For analysis and detection of programming errors debuggers from Intel, Etnus (Totalview), and Allinea (DDT) are available. Furthermore, for profiling and tuning of user's programs SGI provides additional tools.

System Usage

For the most part, the compute performance of the system will be accessible in batch mode. For this purpose, the batch queuing system PBS Pro from Altair has been licensed; for interactive work, development and testing of programs and small test runs a dedicated CPU set will be available to the users in shared mode. For the first time in LRZ's history, a significant portion of CPU time will be made available for the use by computational Grids; in particular there is a specific software environment available for the Distributed European Infrastructure for Supercomputing Applications (DEISA). Also, the D-Grid project, which is dedicated to the construction of a national Grid infrastructure, will be able to claim resources on the system.

Seamless Transition

From the beginning, LRZ has made efforts to provide a seamless transition of top-level scientific computing projects to the new system: Since July 2005 an Altix 3700 migration system with 64 processors has been in operation; on this system essentially the same software environment has been available as on the finally installed system. Many users hence have already been able to adapt their programs to the new system.

• Reinhold Bader

Leibniz-Rechenzentrum (LRZ)

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

Dr. Horst-Dieter Steinhöfer
Boltzmannstraße 1
85748 Garching bei München
Germany
Phone +49-89-3 58 31-87 79

View of „Höchstleistungsrechner in Bayern HLRB II", an SGI Altix 4700

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

System	Size	Peak Performance (GFlop/s)	Purpose	User Community
SGI Altix 4700 16 x 256 way	4,096 processors 17,5 TByte	26,200	Capability computing	German universities and research institutes
SGI Altix 4700 256 way	256 Processors 1 TByte	1,640	Capability computing	German and Bavarian universities and research institutes
SGI Altix 3700 BX2 128-way	128 Processors 512 Gbyte memory	820	Capability computing	Bavarian universities
SGI Altix 64-way	64 processors 256 GByte memory	410	Tests and porting	German universities and research institutes
Linux Cluster Intel IA64 2-way	68 nodes 136 processors 816 GByte memory	870	Capability and capacity computing	Bavarian universities
Linux Cluster Intel IA64 4- and 8-way	19 nodes 84 cores 250 GByte memory	440	Capacity computing	Munich universities
Linux cluster Intel IA32 Intel&AMD EM64T	154 nodes 192 processors 320 GByte memory	850	Capacity computing	Munich universities

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

HLRS

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

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

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

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

Contact
Höchstleistungsrechenzentrum
Stuttgart (HLRS)
Universität Stuttgart

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

View of the NEC SX-8 at HLRS

Compute servers currently operated by HLRS are

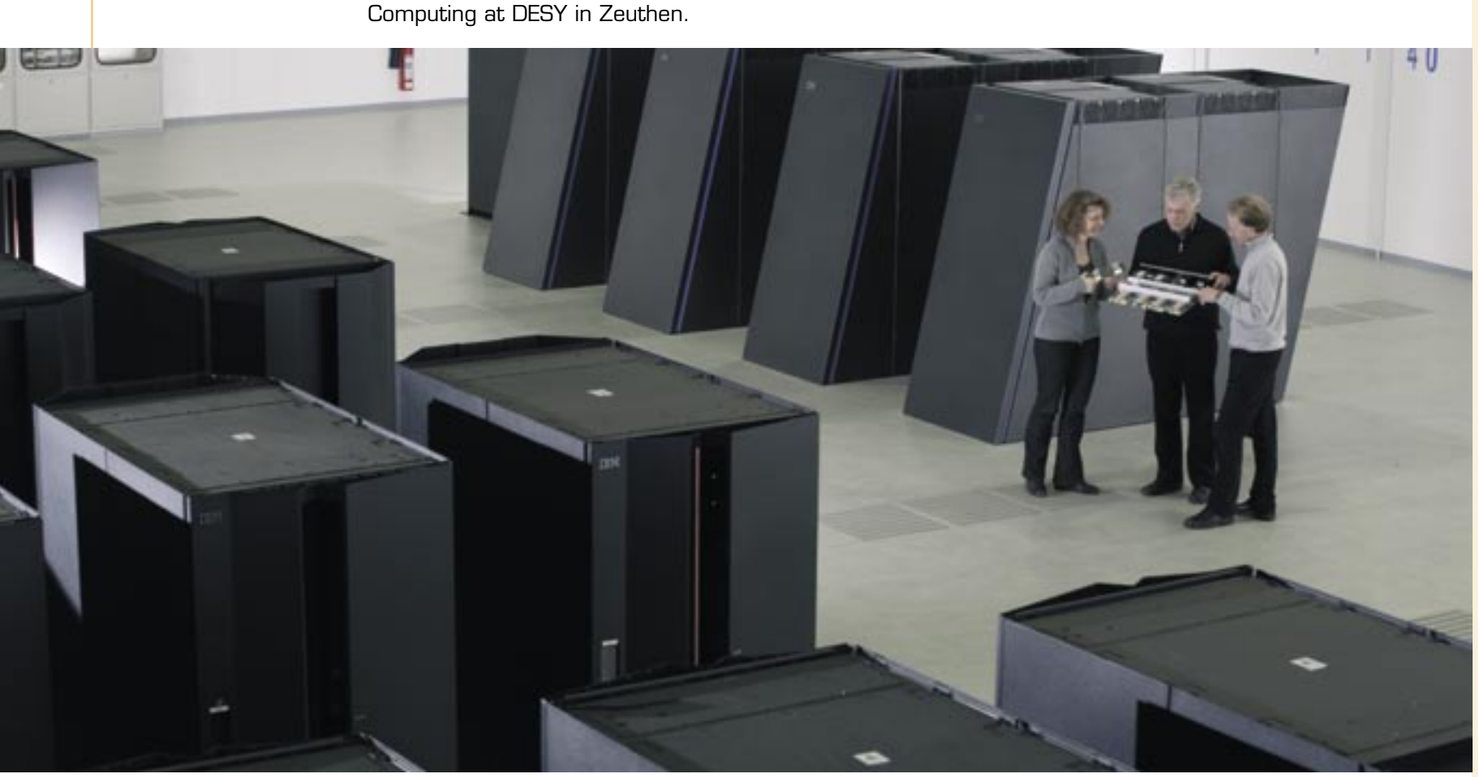
System	Size	Peak Performance (GFlop/s)	Purpose	User Community
NEC SX-8	72 8-way nodes 9,22 TB memory	12,670	Capability computing	German universities, research institutes, and industry
TX-7	32 way node 256 GByte memory	192	Preprocessing	German universities, research institutes, and industry
Intel Nocona Cluster	205 2-way nodes 336 GB memory	2,624	Capability and capacity computing	Research institutes, and industry
Cray Opteron	129 2-way nodes 512 GByte memory	1,024	Capability and capacity computing	Research institutes, and industry
Cray XD1	8 12-way nodes 96 GByte	500	Industrial development	Research institutes, and industry

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

Provision of supercomputer capacity for projects in science, research and industry in the fields of modelling and computer simulation including their methods. The supercomputers with the required information technology infrastructure (software, data storage, networks) are operated by the Central Institute for Applied Mathematics (ZAM) in Jülich and by the Center 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, two research groups exist: the group Elementary Particle Physics, headed by Karl Jansen and located at the DESY laboratory in Zeuthen and the group Computational Biology and Biophysics, headed by Ulrich Hansmann at the Research Center Jülich.

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



The IBM supercomputers "JUBL" (top) and "JUMP" (bottom) in Jülich (Photo: Forschungszentrum 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 Center Jülich and of the German Electron Synchrotron DESY, respectively:
www.fz-juelich.de/zam/nic/en
www-zeuthen.desy.de/main/html/home

System	Size	Peak Performance (GFlop/s)	Purpose	User Community
IBM Blue Gene/L "JUBL"	8 racks 8,192 nodes 16,384 processors PowerPC 440 4 TByte memory	45,875	Capability computing	German universities, research institutes, and industry
IBM pSeries 690 Cluster 1600 "JUMP"	41 SMP nodes 1,312 processors POWER4+ 5,1 TByte memory	9,000	Capability computing	German universities, research institutes, and industry
IBM BladeCenter-H "JULI"	2 racks 56+12 Blades 224 PowerPC 970MP cores + 24 Cell processors (224 + 24) GByte memory	2,240 (+4,800 SP Cell)	Capability computing	Selected NIC projects
apeNEXT (special purpose computer)	4 racks 2,048 processors 512 GByte memory	2,500	Capability computing	Lattice gauge theory groups at universities and research institutes
APEmille (special purpose computer)	4 racks 1,024 processors 32 GByte memory	550	Capability computing	Lattice gauge theory groups at universities and research institutes

Contact
John von Neumann-Institut für Computing (NIC)
Zentralinstitut für Angewandte Mathematik (ZAM)
Forschungszentrum Jülich

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

Larry Meadows

Jim Cownie

Dr. Georg Hager

Bettina Krammer

Number of participants in action

Workshop Report "Cluster OpenMP* from Intel"

OpenMP is a well-known parallel programming paradigm for shared-memory multiprocessors. Now, Intel is offering Cluster OpenMP, which extends the OpenMP programming model to clusters. The High Performance Computing Center Stuttgart (HLRS) organized and hosted a very successful joint workshop with Intel at the HLRS in Stuttgart, 19 May 2006, to assist interested participants in learning how to run their programs using Cluster OpenMP.

Cluster OpenMP uses a Distributed Virtual Shared Memory (DVSM) system to maintain the illusion that part of the address space of each of the processes in the program is shared; thus when a thread running in any of the processes reads memory whose address is in the sharable part of the address space it sees the correct value according to the OpenMP memory model, even if that value was written by a thread executing in a different process on a different machine in the cluster.

OpenMP's relaxed memory model means that write operations which occur in a parallel region do not need to be visible to other threads until after a synchronization operation between the writer and the reader. This allows the DVSM to defer notifications about writes until the next barrier (or locking operation); as a result many fewer messages need to be sent, and much more data can be aggregated into each message than would be the case if a sequential consistency model had to be implemented. In addition, the DVSM implementation is itself lazy; page updates are not transferred until the page is accessed; only data about page states is transferred at synchronization points.

Cluster OpenMP introduces one additional directive to those already in OpenMP: the "sharable" directive. This is used at the declaration or allocation point of variables which are accessed by more than one thread in a parallel region. It allows the compiler to allocate such variables in the part of the process' address space which is maintained consistent with the other processes in the program by the DVSM. In many cases the compiler can deduce the need to place variables in sharable space, but where this is impossible the programmer must add a sharable directive. To help to determine where sharable directives are required the Intel compiler can use interprocedural optimization to suggest which variables must be made sharable. Intel have ported the SPEComp* benchmarks to Cluster OpenMP and this required on average changes to only ~2 % of the source code lines.

While Cluster OpenMP is not a panacea (not all OpenMP codes will perform well with Cluster OpenMP), there are many codes which do perform well, sometimes even exceeding the performance of the same code on an SMP machine with identical CPUs. Cluster OpenMP is most suitable for codes which scale well with OpenMP on SMP machines, have little synchronization, and good locality. As the Cluster OpenMP DVSM protocol allows pages to be replicated in many processes, codes which have many threads accessing read only shared data work particularly well with Cluster OpenMP.

For people who have existing OpenMP codes which they want to run on a cluster, Cluster OpenMP provides an interesting alternative to rewriting the code in MPI.

The workshop was presented by Larry Meadows (an Intel principal engineer) and James Cownie (an Intel senior software engineer) who have over 40 years of HPC and parallel experience between them and who are both working on the Cluster OpenMP implementation. A talk on user experience of Cluster OpenMP was given by Dr. Georg Hager of Regionales Rechenzentrum Erlangen (RRZE). Practical exercises took place on one of the HLRS' clusters, namely a Xeon EM64T cluster with 200 dual-processor nodes. The workshop was well-attended by participants from both academia and industry.

References

All presentations from the workshop are available at the web page
http://www.hlrs.de/news-events/events/2006/openmpworkshop/html/program.html

More links
Hoeflinger, Jay P., Meadows, Larry/Intel
Programming OpenMP on Clusters,
HPCwire, Week of May 19, 2006, Vol. 15, No. 20

http://www.hpcwire.com/hpc/658711.html

* Other names and brands may be claimed as the property of others

- Jim Cownie[1]
- Larry Meadows[1]
- Bettina Krammer[2]

[1] Intel
[2] Höchstleistungsrechenzentrum Stuttgart

Prof. Dr. Wagner | Michael Heib | Christoph Gümbel | Prof. Dr.-Ing. Roland Rühle | Takayuki Sasakura | Prof. Dr.-Ing. Michael Resch

10th Anniversary of the HLRS

Celebrations held on July 28, 2006

One year after the inauguration of its new building along with the NEC SX-8 supercomputer, the High Performance Computing Center, Stuttgart (HLRS) could celebrate another special event.

On the occasion of the 10th anniversary on July 28, 2006 representatives from politics, economics, science, and administration were gathering at Stuttgart for a small celebration at its new premises.

Approximately 100 guests from around the world came to celebrate the birthday.

Stuttgart Universität's Rector, Professor Dr.-Ing. habil. Dieter Fritsch, brought with him the congratulations from the entire University. In his speech, Rector Fritsch pointed out the meaning of the HLRS to the University as being: "Close relationships with the faculties of the University – in particular computer science, mathematics, physics, mechanical engineering as well as aerospace made it possible for Stuttgart to be considered today as one of the most innovative centers in the world of high performance computing."

Takayuki Sasakura, Managing Director NEC High Performance Computing Division, congratulated the HLRS in the name of NEC. Co-operation between NEC and the HLRS began in 1996 with the SX-4. Sasakura stressed that the special concept of co-operation between industry and science excited world-wide attention and this had significantly contributed to the success of the HLRS. Today, the HLRS operates the largest SX-8 worldwide. With its Teraflop Workbench project, NEC experts and users worldwide together obtain outstanding results on its systems.

10 years of HLRS also means 10 years of hww (Höchstleistungsrechner für Wissenschaft und Wirtschaft Betriebsgesellschaft mbH). The director of hww, Michael Heib, brought birthday greetings. He emphasized the importance of public-private partnerships and congratulated HLRS on its 10 years of successful collaboration with industry.

The close solidarity with industry, which particularly distinguishes the HLRS, was emphasized also by Christoph Gümbel from Dr.-Ing Ferdinand Porsche AG. "With pride HLRS today may rank itself together with hww among the top addresses world-wide when it comes to computations and simulations in such important areas as fluid mechanics, structure mechanics, physics, and chemistry". 10 years ago, Porsche AG, one of hww's initial members, had helped the HLRS and hww to rise to become one of the most efficient centers for supercomputing in the world. As result of this trustworthy co-operation over many years, Porsche has intensively advanced the development of its new family sports car platform (Panamera) using the HLRS computers.

Professor Wagner, Universität Stuttgart, brought the congratulations of the steering committee of the HLRS on behalf of its chairman, Professor Dr. Wolfgang W. Nagel. The steering committee accompanied and supported the HLRS on its way forward with advice and actions in the past 10 years. Each computer procurement was a challenge for the future and in many meetings, often controversial, the future policy was discussed. However, at each step forward, it was shown that the decisions met were correct. Prof. Wagner pointed out that great efforts will also be necessary in the future and encouraged HLRS to con-

tinue its close collaboration with the steering committee and the scientific community as a whole.

In his commemorative address, Professor Roland Rühle, former director of the HLRS, who was intensively involved in its establishment, pointed out the highlights of 10 years of HLRS which had affected the history of the HLRS in the years of its existence. Prof. Rühle described how high performance computing developped over the years at the Universität Stuttgart and how over time the close relationship with industry turned into a stable collaboration.

The 10th anniversary party should not be concluded without a view forward. Professor Michael M. Resch, Director of the HLRS, opened his speech with a “Thank you” for all the congratulations received. Prof. Resch pointed out that also in the future, HLRS will be a strong partner for all research at the Universität Stuttgart and all over Germany. HLRS will continue and strengthen its co-operation with industry in research and development. In addition to the very strong co-operation with NEC, the HLRS is co-operating with a number of important companies in the field like Intel, CISCO, Cray, IBM, Bull, and others. Specifically interesting is the new shared research effort with Microsoft. Furthermore, the successful work in numerous projects e.g. with the European Union and the BMBF serves as an important pillar for the future work of the HLRS. Beyond that, the HLRS has also merged itself into a German initiative for a European Supercomputing Center.

Finally Prof. Resch thanked all partners – public and private – that have supported HLRS over its first 10 years and expressed the hope for another 10 years of common research, development and excellent services

Following the celebrations, the guests could inspect the new building including the computer room with its NEC SX-8. Refreshments were provided with a small lunch.

Advances in High Performance Computing and Computational Sciences

This volume contains contributions to the First Kazakh-German Advanced Research Workshop on Computational Science and High Performance Computing presented in September 2005 at Almaty/Kazakhstan. The contributions show the potential of bringing together theoretical mathematical modelling and powerful high performance computing systems. They range from computer science, mathematics and high performance computing to applications in computational fluid dynamics, combustion and industrial problems. They show a wealth of theoretical work and simulation experience with a potential to bringing together theoretical mathematical modelling and usage of high performance computing systems presenting the state of the art in computational technologies.

The 1st Kazakh-German Advanced Research Workshop, Almaty/Kazakhstan, September 25 to October 1, 2005
Series: Notes on Numerical Fluid Mechanics and Multidisciplinary Design (NNFM), Vol. 93
Shokin, Y., Resch, M., Danaev, N., Orunkhanov, M., Shokina, N. (Eds.),
Springer, Berlin, Heidelberg 2006,
XV, 224 p., Hardcover
ISBN: 3-540-33864-0

High Performance Computing in Science and Engineering ’06

This book presents the state-of-the-art in simulation on supercomputers. Leading researchers present results achieved on systems of the High Performance Computing Center Stuttgart (HLRS) for the year 2006. The reports cover all fields of computational science and engineering ranging from CFD via computational physics and chemistry to computer science with a special emphasis on industrially relevant applications. Presenting results for both vector systems and micro-processor-based systems the book allows to compare performance levels and usability of various architectures. As HLRS operates the largest NEC SX-8 vector system in the world this book gives an excellent insight into the potential of vector systems. The book covers the main methods in high performance computing. Its outstanding results in achieving highest performance for production codes are of particular interest for both the scientist and the engineer. The book comes with a wealth of coloured illustrations and tables of results.

Transactions of the High Performance Computing Center, Stuttgart (HLRS) 2006,
Nagel, Wolfgang E., Jäger, Willi, Resch, Michael (Eds.),
Springer, Berlin, Heidelberg, 2007,
Approx. 555 p., 179 illus., 126 in colour,
Hardcover
ISBN: 3-540-36165-0

9th HLRS Metacomputing and Grid Workshop

The annual Grid- and Metacomputing Workshop continues the series of workshops for the Grid community at large at the High Performance Computing Center in Stuttgart, this year for the 9th time and in conjunction of the 10-year celebration of the HLRS.

Bettina Krammer, organizer of the workshop

At the workshop, participants from several countries, like the US, Japan, and mainly European countries provided an overview on Grid activities in their countries and institutes. Especially the involvement of industry in two starting sessions on the first day provided a good overview on the involvement of large vendors such as Intel, NEC, IBM, and Cray, and also networking specialists such as Cisco and Qlogic. On the afternoon, Atos Origin and T-Systems SFR provided insights into Grid requirements and developments from an IT service providers point of view.

The large European infrastructure projects were presented in the last session on the first day.

The second day was dedicated to European Grid initiatives and research projects. Projects from the German D-Grid initiative were scheduled for the afternoon session, followed by talks on the international research projects from Japan, Ireland, and Spain.

Presentations held at the workshop, may be downloaded from the below mentioned webpage. The next annual workshop, the 10th in the series, will be held on the link:
www.hlrs.de/news-events/events/2006/metacomputing

Open MPI Workshop

Right after the EuroPVM/MPI conference, members of the Open MPI team such as Los Alamos National Labs (LANL), Sandia National Labs, Cisco, and Sun met for the regular project developer meetings, this time at the HLRS in the frame of the int.eu.grid project.

This two day workshop provided the team a forum for intensive discussions in the area of Open Runtime Environment development, distributed heterogeneous computing, integration of third-party contributions for performance analysis and fault-tolerance.

Using the AccessGrid, partners from the US, who were not able to attend in person could report on the design and status of Open RTE and the afore-mentioned fault-tolerance support.

As the next important milestone release, Open MPI-1.2 is in the finalization and stabilizing phase, this meeting provided a good occasion to get the relevant developers together. Especially with the increased interest of various industrial and scientific partners discussing the best possible implementation integrated into one development tree is important.

Link:
www.open-mpi.org

The 3rd Russian-German School on Parallel Programming

using
High Performance Computation Systems
August 28 - September 8, 2006
Novosibirsk/Russia

The present event establishes a tradition of such schools in Novosibirsk/Russia, and continues a success of the 1st and 2nd School.

Inspirational and highly professional lectures have been given by Thomas Bönisch, Bettina Krammer, and Sven Stork. A presentation of HLRS has been done by Dr. Nina Shokina. A video-conference between the HLRS and the Institute of Computational Technologies of the Siberian Branch of the Russian Academy of Sciences (ICT SB RAS) has been organized, when Uwe Küster has given a lecture on program optimization for single processor performance.

This School has been organized as two courses, allowing participation of scientists with different levels of knowledge. The basic course has included Parallel Architectures and Programming Models, MPI, Parallel Debugging, Profiling and Performance Analysis, Domain Decomposition Techniques, OpenMP. The advanced course has covered Parallel Programming Models on Hybrid Systems (MPI + OpenMP), Single Processor Optimization, Efficient Parallelization of Krylov Subspace Methods, Advanced Use of MPI, OpenMP Performance Tuning, Parallelization of Particle Based Methods.

Two lectures have been given by the Russian scientists: Prof. L. Chubarov (numerical modelling of Tsunami waves) and Prof. V. Shaidurov (numerical simulation of the Navier-Stokes equations).

The basic course has been attended by 37 participants and the advanced course – by 43 participants from Novosibirsk, Chelyabinsk, Nizhniy Novgorod, Kemerovo, Tomsk, Irkutsk, and Minsk (Belarus Republic).

Another highlight of the School has been a scientific session, where participants have had a possibility to present their own works. The talks have included Parallel Programming, Information Technologies, Laser Physics, Computational Fluid Dynamics, Application of Parallel Computations in Electrical Power System Management, Informatics.

Two books of course material have been prepared by T. Bönisch, R. Keller, S. Stork, and B. Krammer. The 3rd School has been supported by HLRS, ICT SB RAS and its director Prof. Yurii Shokin, Siberian Branch of Russian Academy of Sciences, Russian Foundation for Basic Research and Allinea Software (UK).

ParCo 2007 in Jülich and Aachen

The well-known conference Parallel Computing (ParCo) will take place from 4-7 September 2007 in Jülich and Aachen/Germany. The conference is organized by the non-profit foundation ParCo Conferences in co-operation with the Forschungszentrum Jülich and the RWTH Aachen University. ParCo 2007 marks a quarter of a century of the international conferences on parallel computing that started in Berlin in 1983. This makes ParCo the longest running series of international conferences on the development and application of high speed parallel computing technologies in Europe.

The aim of the conference is to give an overview of the state-of-the-art of the developments, applications and future trends in high performance computing for all platforms. The conference addresses all aspects of parallel computing, including applications, hardware and software technologies as well as languages and development environments. Special emphasis will be placed on the role of high performance processing to solve real-life problems in all areas, including scientific, engineering and multidisciplinary applications and on strategies, experiences and conclusions made with respect to parallel computing. Extended abstracts of at least 1,000 words should be submitted in electronic form by March 4, 2007.

The conference organizers plan the conferences such that a maximum opportunity is created for delegates to meet and interact with fellow researchers. The informal nature of the conferences allows for easy interaction of the delegates. Also planned are an industrial session and an exhibition of equipment. Further information can be found at the web page www.fz-juelich.de/parco2007.

Blue Gene/L Scaling Workshop at NIC

Figure 1: JUBL

For the first time, the John von Neumann Institute for Computing (NIC), IBM, and the Blue Gene Consortium are jointly preparing and sponsoring a common Blue Gene/L Scaling Workshop, which will take place on December 5-7, 2006 in Jülich.

The purpose of this workshop is to provide a selected number of consortium members and interested researchers the opportunity to run scaling tests of codes ported to the Blue Gene architecture. The Research Centre Jülich will provide computer time on its 8-rack Blue Gene/L system, software and support personnel over a three day period to accomplish this task.

During the preparation phase of the workshop applicants had to present a running BG/L code, which should scale up well to at least 1 rack. Furthermore, they had to submit a nomination form before October 11, which was peer reviewed for final selection. Criteria for a successful application were the expected scientific impact of the runs, the confidence that the code will scale up to 8 racks and the certainty that the Blue Gene/L infrastructure at Jülich (operating system, compiler, libraries) will be sufficient to run the program.

During the workshop members selected will be paired up with assigned advisors who will assist in administrative information (log on, moving data, storing data) and will provide scaling support. Advisors will come from Argonne National Laboratory, IBM, and Jülich. Since the intent is to provide substantial time on the 8-rack system (8 hours minimum per applicant), the number of selected applicants has been restricted to only 6.

More information on this workshop can be found at:
www.fz-juelich.de/zam/bgl-sws06

DEISA Tutorial at NIC

The Executive Committee of the European infrastructure project DEISA (Distributed European Infrastructure for Supercomputing Applications) decided in 2006 to perform 6 training sessions on DEISA activities between mid 2006 and early 2008 to enable a fast development of user skills and knowledge to utilize the DEISA infrastructure efficiently.

The first tutorial was organized by IDRIS and took place in Paris on July 3-5, 2006.

The second one was perfomed in Jülich at the John von Neumann Institute for Computing (NIC) on October 23-25, 2006. About 25 scientists from academic institutions and industrial companies involved in high performance computing coming from a large number of European countries followed the invitation. Again, the session was divided into two parts: the global description of the infrastructure and the usage of its major software components and the dedicated topic Performance and Portability of High Performance Computing Applications. The presentations went along with intensive discussions between the participants and an international group of experts. All participants got a test account on the DEISA infrastructure to access it for evaluating its features and behaviour. This allows them to evaluate the potential relevance of this distributed platform for their applications, and possibly to submit later scientific projects for execution in full production mode. More details on DEISA and on the tutorial programme can be found at:
www.deisa.org

Extrapolating the success of the first training sessions in Paris and in Jülich, the community is now looking forward to the third training event, to be held in Barcelona/Spain in February 2007.

Views from the new LRZ: The buildings, the rooms, and the technical infrastructure*

Inauguration of the New Leibniz Computing Centre (LRZ) of the Bavarian Academy of Sciences and Humanities

On July 21, 2006 the Federal Minister for Education and Research Dr. Annette Schavan, the Bavarian Prime Minister Dr. Edmund Stoiber, and the Bavarian Minister for Science, Research and the Arts Dr. Thomas Goppel inaugurated the new building of the Leibniz Computing Centre in Garching near Munich (budgeted at 45 Million Euros. At the same time the new supercomputing system "Höchstleistungsrechner in Bayern II" HLRB II, budgeted at ca. 38 Million Euros) started operation.

Figure 1: Pressing the button

More than 500 representatives from politics, science, research, industry, and press attended the opening ceremony in which Chairman of the Board of Directors of the LRZ, Prof. Heinz-Gerd Hegering, Prime Minister Dr. Stoiber, Dennis McKenna, Chief Executive Officer and President of SGI, Prof. Dietmar Willoweit, President of the Bavarian Academy of Sciences and Humanities, and Federal Minister Dr. Schavan jointly pressed the red button to symbolically start operation of the new system (see Figure 1 from left to right).

Director Prof. Hegering in his presentation outlined the new challenges for the role of LRZ as a university as well as a high performance computing centre: high availability, recentralization, consolidation, and virtualization of services, support of the IT processes in the universities, hosting and housing of systems, intensification of the HPC activities, and research and development in the fields of Grid computing, Grid and network management.

In her address Dr. Annette Schavan emphasized that the availability of computing systems of the highest performance class and the knowledge about their efficient use are essential in international research competition. The continuous upgrade and extension of compute power is inevitable to keep pace with other countries. She also stressed the fact that the three national high performance centres in Germany have agreed on further close co-operation. The Federal Government in Berlin will help funding a high performance interconnect between the three national centres in Jülich, Stuttgart, and Munich.

Bavaria's Prime Minister Dr. Stoiber commented that the Garching Campus, the new location of LRZ, is one of the most innovative places in Germany. High performance computing is a key technology without which no high ranking achievements in other areas can be done. Cutting edge science and research are necessary to create new products and materials – and finally – new jobs.

After the ceremony most participants took the opportunity to visit the impressive building and its computing equipment. The newly erected construction is the most modern and – due to its 5,700 m^2 floor space and 3,200 m^2 system floor space – the largest computing centre building in Germany. The building, which was designed by architects Herzog & Partners, is architecturally divided into three major parts:

Compute Cube: National supercomputing system, servers for network and IT infrastructure, the Linux cluster, and the extensive data backup and archiving facilities as well as the expensive infrastructure providing electricity and cooling are housed in a remarkable, cube-shaped building measuring 36 meters in each dimension. Within this area increased security measures are enforced, and hence no public access is possible. Electrical power of nominally 4,8 MVA can be provided. The cooling of the national supercomputing system requires an air throughput of 400,000 m^3/hour. Sophisticated security, fire extinguishing equipment and uninterruptible power supplies protect the technical infrastructure. Further extensions due to future demands from additional cooling, electricity or compute systems can be performed.

Institute Building: The offices and labs of the 170 LRZ staff members are located in the institute building. The ground floor also provides public access (reception, PC-equipped work places for students, offices for help desk and telephone hotline). The three upper stories house the staff offices, labs, conference rooms, and the central control room.

Lecture and Conference Building: The new building provides a lecture room with 120 seats and a number of seminar, conference, and training rooms. This area of intense public access will mainly be used by students and scientists of the universities in Munich, for LRZ presentations (e.g., workshops for computational sciences, DEISA, D-Grid), for series of lectures, or for short-term courses.

*Photos: Christoph Rehbach, Atelier Hohenwart, Fuchstal

Autumn 2006 • Vol. 4 No. 2 • inSiDE

High Performance Computing Courses and Tutorials

LRZ
www.lrz.de

Object Oriented Programming with Fortran 95/2003

Date
December 6, 2006
Location
Garching near Munich, New LRZ Building
Contents
The new Fortran standard 2003 offers new features which provide support for object oriented programming. However, it is indeed possible to implement many important design patterns using Fortran 95 only. This course has the purpose of giving an introduction on how to use the object-oriented features of Fortran at the 95 and 2003 levels without getting in the way of good application performance; furthermore also a discussion of the usage of the C interoperability features in Fortran 2003 is provided.
Webpage
http://www.lrz.de/services/compute/courses/#OOFortran

Introduction to the Usage of the HLRB II (SGI Altix 4700)

Date
December 7, 2006
Location
Garching near Munich, New LRZ Building
Contents
- SGI Altix architecture
- Programming environment
- Programming models
- OpenMP and MPI usage
- Intel compilers
- Scientific libraries
- Tools
- Optimization and tuning
- Performance counters
- Batch system PBS
- Software

Webpage
http://www.lrz.de/services/compute/courses/#IntroHLRB2

Parallel Programming of High Performance Systems

Date
February 19-23, 2007
Location
Garching near Munich, New LRZ Building
Contents
Day 1
- Introduction to HPC
- Processor and system architectures
- Programming paradigms and languages
- Memory hierarchy and caches
- HPC systems at LRZ and in Germany

Day 2
- Intel Itanium architecture
- Intel Compiler, tools and libraries
- SGI Altix architecture and tools
- Basics of optimization
- Performance counters

Day 3
- Parallel programing with MPI (incl. hands-on sessions)

Day 4
- Parallel programing with MPI
- Parallel programing with OpenMP (incl. hands-on sessions)

Day 5
- Advanced examples for optimization
- Intel tracing and threading tools
- Tuning of I/O

Webpage
http://www.lrz.de/services/compute/courses/#ParallelProgramming

NIC
www.fz-juelich.de/nic

Parallel Programming with MPI, OpenMP, and PETSc

Date
November 27-29, 2006
Location
Research Centre Jülich, NIC/ZAM
Contents
The focus is on programming models MPI, OpenMP, and PETSc. Hands-on sessions (in C and Fortran) will allow users to immediately test and understand the basic constructs of the Message Passing Interface (MPI) and the shared memory directives of OpenMP. This course is organized by NIC/ZAM in collaboration with HLRS.
Presented by Dr. Rolf Rabenseifner, HLRS
Webpage
http://www.fz-juelich.de/zam/neues/termine/mpi-openmp

CECAM Tutorial "Programming Parallel Computers"

Date
January 22-26, 2007
Location
Research Centre Jülich, NIC/ZAM
Contents
This tutorial provides a thorough introduction to scientific parallel programming. It covers parallel programming with MPI and OpenMP. Lectures will alternate with hands-on exercises.
Webpage
http://www.cecam.fr/index.php?content=activities/tutorial

Education in Scientific Computing

Date
August 6 - October 12, 2007
Location
Research Centre Jülich, NIC/ZAM
Contents
Guest Students' Programme "Scientific Computing" to support education and training in the fields of supercomputing. Application deadline is April 30, 2007.
Webpage
http://www.fz-juelich.de/zam/gaststudenten

HLRS
www.hlrs.de

Parallel Programming with MPI, OpenMP, and PETSc

Date
February 12-15, 2007
Location
Dresden, ZHR
Contents
The focus is on programming models MPI, OpenMP, and PETSc. Hands-on sessions (in C and Fortran) will allow users to immediately test and understand the basic constructs of the Message Passing Interface (MPI) and the shared memory directives of OpenMP. The last day is dedicated to tools. This course is organized by ZIH in collaboration with HLRS.
Webpage
http://www.hlrs.de/news-events/external-events

Introduction to Computational Fluids Dynamics

Date
March 5-9, 2007
Location
University of Kassel
Contents
Numerical methods to solve the equations of Fluid Dynamics are presented. The main focus is on explicit Finite Volume schemes for the compressible Euler equations. Hands-on sessions will manifest the content of the lectures. Participants will learn to implement the algorithms, but also to apply existing software and to interpret the solutions correctly. Methods and problems of parallelization are discussed. This course is organized by University Kassel, HLRS, and IAG, and is based on a lecture and practical awarded with the "Landeslehrpreis Baden-Württemberg 2003" (held at University Stuttgart).
Webpage
http://www.hlrs.de/news-events/external-events

Iterative Linear Solvers and Parallelization

Date
March 12-16, 2007
Location
Stuttgart, HLRS
Contents
The focus is on iterative and parallel solvers, the parallel programming models MPI and OpenMP, and the parallel middleware PETSc. Thereby, different modern Krylov Subspace Methods (CG, GMRES, BiCGSTAB ...) as well as highly efficient preconditioning techniques are presented in the context of real life applications. Hands-on sessions (in C and Fortran) will allow users to immediately test and understand the basic constructs of iterative solvers, the Message Passing Interface (MPI) and the shared memory directives of OpenMP. This course is organized by Kassel University , HLRS, and IAG.
Webpage
http://www.hlrs.de/news-events/events

NEC SX-8 Usage and Programming

Date
March 19-20, 2007
Location
Stuttgart, HLRS
Contents
The first day is focused on vectorizing and parallelizing on NEC SX-8, the second day is dedicated to parallel I/O.
Webpage
http://www.hlrs.de/news-events/events

C++ for Scientific Computing

Date
March 26 - April 5, 2007
Location
Stuttgart, HLRS
Contents
This introduction to C++ is taught with lectures and hands-on sessions. This course is organized by HLRS and Institute for Computational Physics.
Webpage
http://www.hlrs.de/news-events/events

inSiDE

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

Publishers

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

Editor

F. Rainer Klank, HLRS klank@hlrs.de

Design

Julia Schlatterer schlatterer@hlrs.de

Authors

Dr. Guido Arnold g.arnold@fz-juelich.de
Prof. Dr. Achim Bachem a.bachem@fz-juelich.de
Dr. Reinhold Bader reinhold.bader@lrz-muenchen.de
Dr. Matthias Brehm matthias.brehm@lrz-muenchen.de
Jim Cownie james.h.cownie@intel.com
Dr. Alistair Dunlop a.dunlop@omii.ac.uk
Dr. Markus Geimer m.geimer@fz-juelich.de
Prof. Dr. Werner Hanke hanke@physik.uni-wuerzburg.de
Prof. Dr. Heinz-Gerd Hegering hegering@lrz.de
Stephan Hochkeppel hochkepp@physik.uni-wuerzburg.de
Rainer Keller keller@hlrs.de
Bettina Krammer krammer@hlrs.de
Prof. Dr. Dr. Thomas Lippert th.lippert@fz-juelich.de
Larry Meadows lawrence.f.meadows@intel.com
Dr. Bernd Mohr b.mohr@fz-juelich.de
Richard Patra rchard.patra@lrz-muenchen.de
Prof. Dr. Michael Resch resch@hlrs.de
Dr. Marcus Richter m.richter@fz-juelich.de
Prof. Dr. Wolfgang Rodi rodi@uka.de
Prof. Dr. Andreas Schäfer andreas.schaefer@physik.uni-regensburg.de
Binh Trieu b.trieu@fz-juelich.de
Dr. Jan Wissink wissink@ifh.uni-karlsruhe.de
Prof. Dr. Felix Wolf f.wolf@fz-juelich.de
Dr. Brian J.N. Wylie b.wylie@fz-juelich.de

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