High Performance Computing in Science and Engineering

Peter W. Haas

Höchstleistungsrechenzentrum Stuttgart

1. Introduction to HLRS

Slide 1: Title

High performance computing has gradually shifted from the realm of research into development and partially even into the production cycles of industry. High performance computers therefore have to be integrated into production environments that demand the simultaneous solution of multidisciplinary physics problems.

Slide 2: Table of Contents

Supercomputer centers can learn from these new challenges imposed by industry. The concepts of work flow and production cycle open up a new horizon for integrating systems and software into what is called a distributed „Teraflop-Workbench“ approach. Terascale storage and communication infrastructures eventually will be needed to support such an environment.

Slide 3: Höchstleistungsrechenzentrum Stuttgart

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

Slide 4: The HLRS Framework

Due to its embedding in the industrial environment of Southwest Germany’s high-technology region around Stuttgart, HLRS traditionally focuses on applications from the engineering sciences like Computational Fluid Dynamics, Combustion, Structural Mechanics, Electromagnetics and Process Engineering. However, recently HLRS has extended its portfolio to medical applications, environmental applications and has started an initiative to bring new fields of application to the supercomputer.
The services of HLRS offer complete support to user groups such as physical modeling of parallel numerics and algorithms, the embedding of commercial packages and the visualization of results either remotely or in HLRS’ virtual reality laboratory at Stuttgart.
Research groups at HLRS are involved in numerous projects targeting at the industrial use of HPC technology and at the further improvement of such technologies based on user feedback. This includes the participation in the standardization of base technologies.

Slide 5: Networks

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

The HLRS/hww computer configuration is distributed over multiple campuses. It uses state of the art equipment in a highly secure environment. The major campuses are located at the universities of Stuttgart, Heidelberg and Karlsruhe as well as within the premises of DaimlerChrysler AG, Deutsches Zentrum für Luft- und Raumfahrt (DLR), Porsche AG, and T-Systems GmbH. Most of the communication lines use dedicated fiber links with optical multiplexors that allow for an appropriate collection

of link signals, e.g. Ethernet, Fibre Channel or InfiniBand. Transmission speeds typically range between one and fourty Gbit/s.

Slide 6: Supercomputer

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

We would like to reference two important application examples first, one each for science and industry, before we go into a detailed discussion of our workbench concept.

Slide 7: Medical Application Example

For more than one decade numerical investigations of blood flow inside vessels have been performed. Due to an increase in user-friendliness of software, the quality of tomographic scanner based digital images, and computer performance these simulations are gaining more and more practical importance.

The influence of geometrical deformations of vessels, e.g. by stenotic lesions, aneurysm, or by-pass operations, are subject of these investigations. A number of special features have to be considered to calculate blood flow accurately. Mainly the transient flow behavior due to the heartbeat and the non-Newtonian flow effects, especially the dependency of the blood viscosity on the local shear rate, have to be mentioned in that respect.
The blood flow in a healthy natural junction, arteria vertebralis dextra and sinistra with ateria basilaris, needs to be compared with the flow in an artificial junction, resulting from a by-pass operation. The areas for preferred formation of stenotic lesions are identified. The shape of the vessels may differ significantly in many areas from one human being to another. A detailed geometrical representation of the particular situation is required in such cases to achieve useful results, i.e. the vessel geometry of a particular person has to be known. This is valid for pathological geometrical changes as well [1].

Slide 8: Vision: The completely integrated digital product development

The establishment of efficient, IT-based procedures in a highly complex network of local and remote manufacturing is a central target of the DaimlerChrysler EDM strategy. It requires continuous process control from early design stages via product development, product manufacturing and after-sales support, and it is a key requirement for successful placement of innovative, high-quality products in a competitive environment.
The integrated digital product development can be viewed as an assembly pipeline process by itself which involves various stages that may require high-performance computing. This is especially true for proving the feasibility of a design in a virtual reality environment, computer-aided engineering in multiple physics domains, and the integration of digital products into the digital factory [2].

2. The Teraflop Workbench Concept

Slide 9: The Teraflop-Workbench Concept

Slide 10: Workbench Concept

When HLRS started its latest request for proposals for an HPC platform it was clear from the beginning that the system offered would have to be part of a larger concept of supercomputing called the Stuttgart Teraflop Workbench [3], [4]. The basic concept foresees a central file system where all the data reside during the scientific or industrial workflow. A variety of systems of different performance and architecture are directly connected to this file system. Each of them can be used for special purpose activities. In general, one can distinguish between systems for pre-processing, high-performance computing and post-processing.

Slide 11: Workflow Example

Provide example for Ansys simulation of electromagnetic actuator
- Preparation of actuator geometry on external CAD system and/or Ansys 10 Workbench
- import design, create input data, allocation shell/bricks with associated materials
- Generate meshes, decompose them for partitioning on parallel processors
- Transfer job to parallel machine
- Perfomance computation (interactive with error control, batch)
- Filter output data and animate (plausibility checks, e.g. with 2D designs, or expert data base)
- Output generated: 2D flow lines, 3D magnetic field and induction vector fields animated over time

Slide 12: Parallel Post-Processing

Importance of balanced Storage and Peer Communication networks
to allow for either
store & forward (interactive visualization) or
direct pipelined processing pipelines (simulation steering).
The Virtual Tour of Kiebingen water power plant (ref. to Slide 27) is an example of simulation steering.

Slide 13: Requirements

Requirements for successful steering of applications well adapted to vector systems and big clusters.

Slide 14: Final NEC Configuration at HLRS

The workbench configuration is centered around NEC's Global File System (GFS). A NEC SX-8 Parallel Vector Processor, two Asama (IA-64 shared memory system) and a cluster of Intel Nocona processors all have direct access to the same data via a Fibre Channel storage area network. In the following we briefly describe these three main systems with a focus on the SX-8.
Most of the floorspace in this chart is consumed by the 72 compute node cabinets of the NEC SX-8. Two hardware systems complement the SX-8 installation. One is an IA-64 based shared memory system that serves for pre-processing and file serving. Code named AsAmA, it consists of 2 TX-7 nodes with 32 processors each. One of the nodes is equipped with 512 GB of memory in order to allow preparation of large parallel jobs. The 200 node Intel Nocona cluster is connected by a Voltaire InfiniBand switch with a bandwidth of 10 GB/s.

Slide 15: Teraflop Workbench

Most users tend to prepare their mesh on one processor before they decompose it and transfer it to the

parallel system. Given that the main memory of the core system is 9 TB, we decided for one node with large memory to be able to prepare large jobs.
A cluster based on Intel EM64T processors and InfiniBand interconnect is added to the workbench. It serves both for post-processing/visualization and as a compute server for multi-disciplinary applications. The latter often require different types of architectures for different types of disciplines.

Slide 16: NEC SX-8

The NEC SX-8 continues the successful line of vector processors that started with the NEC SX-4 in 1996. It has done so by improving the concepts and making use of most recent production technology [5]. The key improvements are:

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

The NEC SX-8 configuration consists of 72 nodes with 8 processors each clocked at 2 GHz as well as 128 GB of shared memory. Peak processor floating point execution rate is 22 Gflop/s. Processors and memories are interconnected by a 128 x 128 IXS crossbar switch. Bidirectional link speed is 16 GB/s and it is being shared by 8 processors. Memory bandwidth is 64 GB/s per processor translating into an aggregate memory speed of 36.8 TB/s. The overall size of the memory is 9.2 TB.
Including 160 TB of disk storage connected to a Fibre Channel storage area network (SAN), the NEC SX-8 configuration requires 400m² of floor space and dissipates 750kW of electrical power.

Slide 17: HP Challenge Benchmark / TOP500

The HPC Challenge benchmark [6] aims at complementing the traditional Linpack benchmark. The 23 individual tests in the HPC Challenge benchmark do not measure the theoretical peak performance of a computer. Rather, they provide information on the performance of the computer in real applications. The tests do assess criteria that are decisive for the user such as the data transfer rate from processor to memory, the speed of communication between two processors within a supercomputer, the response times and data capacity of a network. Since the tests measure various aspects of a system, the results are not stated in the form of one single figure. In their entirety, the measurements enable an assessment of how effectively the system performs high performance computing applications. When the benchmark was run for the predecessor model of the NEC SX-8 – the SX-6 – it took the lead in 13 out of 23 categories showing the high potential that vector supercomputers still have when real performance is at stake.

3. The Teraflop Workbench Project

Slide 18: Teraflop-Workbench Project

Slide 19: Teraflop Workbench Collaboration

In order to further support users and projects and in order to extend the reach of vector systems in terms of application fields HLRS and NEC have set up the Teraflop Workbench Initiative [4].
The first goal is to demonstrate the efficiency of NEC SX vector systems and their ability to deliver Teraflop/s application performance for a broad range of research and ISV codes. Secondly, together with the vector system NEC Linux clusters and SMP systems will form an environment that allows the

complete pre-processing – simulation – post-processing – visualization workflow to perform in an integrated and efficient way.

Slide 20: Methods

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

Slide 21: Areas

Selection of specific problem areas in order to probe the architectural benefits of the individual HPC systems in the Teraflop workbench.

Slide 22: Conclusions

Building blocks and design philosophy in the Teraflop framework.

Slide 23: FENFLOSS (Finite Element based Numerical FLOw Simulation System)

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

Slide 24: Projects: Fenfloss

The simulation process chain is partially established in the integrated environment. It is based on a parameterized runner design that enables a numerical optimization of axial and radial hydraulic turbines. COVISE, a distributed Collaborative Visualization and Simulation Environment developed by HLRS [7], is used as integration platform for the profile generation, the runner contour generation and the grid generation of the entire machine. The definition of boundary conditions based on the operating point, the coupled simulation around runner and guide wheel and the overall process chain will be controlled from within a virtual reality environment.

Slide 25: Fenfloss strong scaling

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

Slide 26: DEMONSTRATION in the HLRS-Cave

Scientific visualization techniques are used by scientists and engineers to understand complex simulations [7]. They comprise filtering data, generation of deduced information, mapping of data onto visual representations and their display, finally. Distributed software environments often couple simulations on remote machines with local visualizations.

Slide 27: Virtual Tour of the Kiebingen Water Power Plant

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

Kiebingen near Stuttgart.
Martin Becker shows a Covise (Collaborative VISualization Environement) of the water flow through a parameterized radial water turbine.
The animation shows:water flow lines with velocity along lines ranging from blue (slow) to red (fast)
Field in cutting plane represents the radial component of water velocity (i.e. the angular momentum being applied on the runner wheel)
In a short pause of the simulation, the inclination of the blades in the guide wheel will be modified. The simulation is being restarted leading to the following steps: generation of a new mesh, decomposition of compute problem and assignment to a number of parallel processors. After a delay of ~ 10s, first results are output by the visualization pipeline.
The radial turbine is completely parametrized: blade profile and geometry, number of blades, number of input/output water channels, etc.
It is important to notice that we are looking at an online simulation here which allows the steering of most the important design parameters by simulation steering thus leading to a new, inductive way of design.

Slide 28: BEST (Boltzmann Equation Solve Tool)

LSTM: Lehrstuhl für Strömungsmechanik, Universität Erlangen-Nürnberg
RRZE: Regionales Rechenzentrum Erlangen

Slide 29: Projects: BEST

Energy conversion in numerous industrial power devices, like automotive engines or gas turbines, is still based on the combustion of fossil fuels. In most applications, the reactive system is turbulent and the reaction progress is influenced by turbulent fluctuations and mixing in the flow. The understanding and modeling of turbulent combustion is thus vital in the conception and optimization of these systems in order to achieve higher performance levels while decreasing the amount of pollutant emission. In the last several years, direct numerical simulations (DNS), i.e. the computation of time-dependent solutions of the compressible Navier-Stokes equations for reacting ideal gas mixtures, have been one of the most important tools to study fundamental issues of turbulent combustion. Due to the broad spectrum of length and time scales apparent in turbulent reactive flows, a very high resolution in space and time is needed to solve this system of equations.

Slide 30: BEST: Scaling towards Multi-Teraflops

A detailed chemistry DNS code has been developed which exhibits an excellent scaling behavior on massively parallel systems. E.g. for a scaled problem with a constant load per processor, a parallel efficiency of 63 percent has been achieved using all 72 nodes of the NEC-SX8.

Slide 31: CCARAT (Computer Aided Research Analysis Tool)

Teachware at the Institute of Structural Mechanics (IBS)

Slide 32: Projects: CCarat - what is it ?

Suspension bridges with long spans may exhibit substantial oscillation amplitudes when exposed to high wind pressure. Quite different from most other engineering constructions, this behavior influences the design and proportions quite heavily. The Institute of Structural Mechanics (IBS), University of Stuttgart, has developed CCarat, a finite element code, that allows the computation of the transient interaction of incompressible viscous flows and nonlinear flexible structures. In general, the modeling uses GLS stabilized fluid elements in 2D- and 3D-configurations on a moving mesh. Sophisticated structural elements (e.g. shell, brick and beam) are available to represent material properties. CCarat is designed to use external solvers, e.g. Aztec, Spooles, Umfpack and Trilinos, which may induce performance limitations due to the iterative process.

Slide 33: CCarat Simulation of the Tacoma Bridge

IBS has conducted a 2D CCarat simulation of the failure of the Tacoma Narrows suspension bridge on November 7, 1940 [8]. This computer simulation is contrasted by a historic video documentation of that catastrophic event.

Slide 34: AIOLOS

The 3D-Combustion Simulation Code RECOM-AIOLOS is a tailored simulation environment for the modeling of industrial furnaces and boilers. The code is developed in co-operation with IVD at Stutgart University.
The physical/chemical models and numerical methods available in RECOM-AIOLOS have been carefully selected and optimized for the simulation of reacting flows:

Slide 35: Projects: Aiolos

Since coal is still one of the major energy sources worldwide, the improvement of the efficiency of power stations with coal-fired boilers is an important task. Improvements in computer performance and detailed physical models have enabled computational fluid dynamics to be a fast and economic tool for the optimization of industrial furnaces. The Aiolos simulation program developed at the Institute of Process Engineering and Power Plant Technology (IVD), University of Stuttgart, is used for the numerical calculation of three-dimensional, stationary and dynamic, turbulent reactive flows in pulverized coal-fired utility boilers [9]. In submodels treating fluid flow, turbulence, homogeneous and heterogeneous combustion, and heat transfer, equations for calculating the conservation of mass, momentum and energy are solved.

Slide 36: Performance on different systems

The implemented models are optimized for vector and parallel computers via MPI and OpenMP to achieve a high numerical efficiency. Performance on a single CPU of the NEC SX-8 is 6.3 Gflop/s, with a typical production run being executed on 4 CPUs. Execution times are an order of magnitude less compared to Itanium (11x) or Xeon 2.4-GHz processors (17x).

Slide 37: FLOWer

The CFD-tool FLOWer solves the compressible Reynolds-averaged Navier-Stokes equations on block structured grids. It is optimized for the simulation of exterior flows in the subsonic, transonic and supersonic flow regime. FLOWer is developed by the German Aerospace Center DLR with contributions from several German universities. The code is intensively used by the German aerospace industry.
FLOWer is based upon a finite volume discretization. Several central and upwind schemes are available. Turbulence can be modeled with algebraic, one or two equation models or Reynolds stress models. The time integration can be performed with explicit or implicit methods. The time integration of the flow equations is performed with an explicit multi-stage scheme and multi grid acceleration whereas the turbulence equations are solved with an implicit DDADI-method. For time accurate simulations an implicit dual time stepping method is available.
FLOWer can handle meshes with hanging nodes, grid overlap (Chimera) and moving/deforming meshes. For shape optimization, an inverse design option and an adjoint method is available. The flow solver is parallelized based on MPI and is optimized for vector computers.

Example of high performance computing with FLOWer

The prediction of the flow field around a helicopter is challenging for any CFD method. Various flow phenomena including boundary layers, flow separations, vortices and widely disparate velocities must be considered. The capabilities of FLOWer are demonstrated by a simulation of the unsteady flow

around a BO105 helicopter wind tunnel model including the model support strut. The computational mesh was created by using the overlapping grid approach. This method accounts for the moving blades and allows simplification of the mesh generation. The total grid system consists of 11.7 million grid cells. The computation of the unsteady flow took four weeks using eight processors of the NEC SX6 vector computer. The results presented in the figures show the pressure distribution on the body surface and the vortices in the flow field.

Slide 38: Problem

Rotorcraft flows rank among the most challenging applications of CFD in aviation engineering. While an attempt to numerically simulate the entire main rotor system of a helicopter calls for a multidisciplinary approach, i.e. primarily the coupling of flow and structure models, even an isolated aerodynamic analysis must cope with a wide spectrum of elementary and interactional flow problems and phenomena. Although the flow over an isolated hovering rotor is steady in a rotating frame of reference, computing this steady-state solution and thus predicting hover performance – a key issue in the design of helicopters – is not trivial at all.

Slide 39: FLOWer: Flow around Helicopter

Modern Navier-Stokes codes can provide valuable insight into local structures of the three-dimensional flow field and interactional phenomena, as needed for rotor design and verification. Over the past decade, a thoroughly validated Chimera structured grid finite volume code, based on the full Reynolds-averaged Navier-Stokes or Euler equations, has been developed by the rotary wing workgroup at the Institute of Aerodynamics and Gasdynamics (IAG) at the University of Stuttgart [10].

4. Terascale Storage & Communication

Slide 40: Terascale Storage & Communication

Slide 41: Unifies Heterogeneous File Systems

Whenever you deal with a major city's archive, you might easily run into some previous century file system, like this (1886) Shannon Filing Cabinet. We would say, it's four-way parallel with limited access per column (by looking at the slim appearance of the gentleman in front). Data Retention, Safety and Security are ways advanced over today's file systems (e.g. compare to University of Washington's University Record Management Services, urc@u-washington.edu).
And it seem's that customers were quite satisfied with record systems 100 years ago: „The Cabinet furnished the Equitable Gas Light Co., is so satisfactory that they have ordered another".

Slide 42: SNIA Shared Storage Model

For the sake of clarity, we would like to stay with the original Unix file system model located well underneath the Virtual File System Switch, as outlined by the SNIA Shared Storage Model. This is to enable multiple file systems on a given host at any time using parallel interfaces. Also the Services Subsystem will be used to incorporate parallel File System Interfaces which will be useful for layering entire file systems on top of each other.

Slide 43: File Systems for Clusters

There are quite a few cluster file system solutions on the market, like IBM's General Parallel File System (GPFS), Tivoli SANergy, ADIC StorNext, and Sun Microsystems' shared Quick File System (QFS). We normally confine ourselves to those cluster file system implementations which build upon Unix file system variants that are part of the standard Linux kernel. Currently, this applies to the GPFS/JFS stack of IBM, the Cluster Extended File System of SGI, CXFS, as well as NEC's GFS,

building upon XFS. They have been coverred in depth by our last year's workshop. I think it is in order to point out to the various principal bottlenecks that are available with commercial cluster file systems (metadata- or lock server made).

Slide 44: Lustre Solution

The Lustre Solution aims at scalability right from its inception by separating the management of meta-data and file allocation data on two distinct entitites: the Metadata Server Cluster and the Object Storage Targets.

Slide 45: Linux cluster = Lustre is a completely new start

The Lustre (concatenation of Linux Clustre) File System is Cluster File Systems' response to the ASCI PathForward RFP B514193 for a Scalable Global Secure (SGS) File System. Lustre uses the proven decomposition of file systems' functions into Meta Data Control, Client Access, and Object Storage Targets. However, the OSTs provide an additional level of abstraction over the traditional block device, comparable but exceeding the NAS file model. Security and Resource Data Bases are maintained via LDAP Directory servers. All services are centered around an IP Storage Area Network. Therefore, the architecture does allow for almost unlimited scalability wrt the number of constituent entities and also geographic extension. Lustre is supposed to constitute ASCI's next generation secure cluster file system.
The Lustre client FS will be available under the Linux VFS from the very beginning, which makes it a native file system for any Linux-based architecture. It's interesting to note that Lustre will adopt proven journaled file system technology, e.g. SGI's XFS or IBM's JFS, for physical storage.

Slide 46: Panasas: New Storage Architecture for Clusters

It's interesting to note that the original input to the ASCI PathForward File System RFP was actually drafted after Garth Gibson's specification of a Linux cluster storage at Panasas. On October 22, Bob Coyne emails his pal Ann Borrett at IBM:
Ann,
Looks like the product, for which the RFP seemed well aligned, is the winning solution.
LANL has successfully introduced a near term alternative to the ASCI Lustre solution.
Bob
Let me share a few slides on Panasas Active Scale File System, that Garth Gibson did present at the HLRS workshop. They will give you some idea of the object storage architecture, and how it helps to achieve performance, in both I/O rates and data throughput.

Slide 47: Object Storage Architecture

The Object Storage Architecture is an evolutionary improvement of the standard SCSI storage interface. In principle, it adds a new optional command set to the existing ones.
This new command set raises the level of abstraction: Object is a container for „related" data (i.e. file attributes, user data, data placement information)

– Storage understands how different blocks of a „file" are related

Offload most datapath work from server to intelligent storage, e.g. layout, striping etc.

The Object-based Architecture is based on data Objects, which encapsulate user data (a file) and attributes of that data. The combination of data and attributes allows an Object-based storage system to make decisions on data layout or quality of service on a per-file basis, improving flexibility and manageability. The device that stores, retrieves and interprets these objects is an Object-based Storage Device (OSD). The unique design of the OSD differs substantially from standard storage devices such as Fibre Channel (FC) or IDE, with their traditional block-based interface. By moving low-level storage functions into the storage device itself and accessing the device through a standard object interface, the Object-based Storage Device enables:

Slide 48: Object Access Example

Even thought this is not a tutorial, we can differentiate four phases in this Object Access Example:
(1) Client seeking first access to a file contacts File Manager (Meta Data Server)
(2) File Manager returns cacheable access approval (so-called capability) and scalable object map
(3) Clients repeatedly send requests (read/write) directly and in parallel to OSDs
(4) High bandwidth direct-data transfer between clients and OSDs

File and directory access management: The MDS provides the compute node with the file structure of the storage system. When the node requests to perform an operation on a particular file, the MDS examines the permissions and access controls associated with the file and provides a map and a capability to the requesting node. The map consists of the list of OSDs and their IP addresses, containing the components of the object in question. The capability is a secure, cryptographic token provided to the compute node, which is examined by the OSD with each transaction. The token describes to the OSD which object that the compute node is allowed to access, with what privileges, and for what length of time.

Slide 49: Object Storage Bandwidth

With Gigabit Ethernet we can expect a 40-MB/s contribution per network connection (Panasas Storage Blade packages one Xeon processor/RAM/single GigE with two SATA disks). What you see here is a real life lab experiment demonstrating linear scalable bandwidth for 0 < device count < 300 (~ 12 GB/s).

Slide 50: Standardization Timeline

This slide is to show you how the the initial research of the CMU NASD was transformed into the founding document of the "Object-based Storage Device" working groups in the Storage Networking Industry Association (www.snia.org/osd) and the ANSI X3 T10 (SCSI) standards body (www.t10.org). Since that time, the OSD working group in SNIA has guided the evolution of Object Storage interfaces.

The Letter Ballot for the SNIA/T10 OSD Draft Standard ended March 24, 2004.

- -

In 1997 CMU initiated an industry working group in the National Storage Industry Consortium (now www.insic.org). This group, including representatives from CMU, HP, IBM, Seagate, StorageTek and Quantum, worked on the initial transformation of CMU NASD research into what became, in 1999, the founding document of "Object-based Storage Device" working groups in the Storage Networking Industry Association (www.snia.org/osd) and the ANSI X3 T10 (SCSI) standards body (www.t10.org). Since that time, the OSD working group in SNIA has guided the evolution of Object Storage interfaces, as member companies experiment with the technology in their R&D labs. Today the SNIA OSD working group is co-led by Intel and IBM with participation from across the spectrum of storage technology companies.
Panasas has made a commitment to continue to help drive the development of standards based on the Object Storage Architecture. Rather than the traditional approach of trying to build a business based on proprietary, closed systems, Panasas believes that customers will benefit if an industry is built around the Object Storage Architecture. Therefore, we are also working with key members of the file system community, some of who are already involved with Object Storage through the ANSI X3 T10 work. Panasas has developed a standard file system that can fully exploit the capabilities of the Object-based Storage Device with the vision that an Object Storage Architecture-based parallel file system eventually becoming as ubiquitous as NFS V3 is today. To that end, Panasas has committed to working with its partners to make sure that open-source, reference implementations of the file system are available and that the file system is driven through an open standards process. This is all intended to lay the

foundation for an industry around the Object Storage Architecture that has interoperable Object-based Storage Devices, Metadata Controllers and a parallel, object-based file system.

Slide 51: Storage Tank (SAN FS)

View of the Storage Tanks at the IBM Almaden Research Center when approached on Bernal Road.

Slide 52: Storage Tank Overview (Description)

The Storage Tank research project has been conducted at Almaden Research under the direction of David Pease. Storage Tank did lead to both a scalable global file system with security credentials comparable to Microsoft's NTFS as well as an underlying virtualization scheme. As such it may assume any role in the known file system design space.

Storage Tank, commercially known as SAN FS, offers a complete, policy-based storage management solution for heterogeneous, distributed environments. It is designed to provide I/O performance that is comparable to that of file systems built on bus-attached, high-performance storage. In addition, it provides high availability, increased scalability, and centralized, automated storage and data management. All SAN FS management functions adhere to the Storage Management Initiative Specification (SMIS) standard for distributed environments (140 operating systems supported today)

- - - - - - - - - - - - - - - - - - -

David Pease: IBM Storage Tank. Ongoing research and development

The Storage Tank research project has been conducted at Almaden Research under the direction of David Pease. Storage Tank did lead to both a scalable global file system with security credentials comparable to Microsoft's NTFS as well as an underlying virtualization scheme.

http://almaden.ibm.com/StorageSystems/file_systems/storage_tank/index.shtml
Storage Tank offers a complete, policy-based storage management solution for heterogeneous, distributed environments. It is designed to provide I/O performance that is comparable to that of file systems built on bus-attached, high-performance storage. In addition, it provides high availability, increased scalability, and centralized, automated storage and data management. Storage Tank has been renamed to IBM TotalStorage SAN FS in the course of productization at IBM Beaverton. All SAN FS management functions adhere to the Storage Management Initiative Specification (SMIS) standard for distributed environments (140 operating systems supported today):

http://www.snia.org/smi/tech_activities/smi_spec_pr/spec/SMIS_1_0_2_final.pdf

Another important related project is Distributed Storage Tank, which extends the Storage Tank technology for geographically distributed file sharing.

http://almaden.ibm.com/StorageSystems/GRID/distributed_storage_tank/index.shtml

In addition to the basic Storage Tank project, work is being done on NAS over Storage Tank, a project that adapts the Storage Tank technology for Network-Attached Storage (NAS).

http://almaden.ibm.com/StorageSystems/file_systems/NAS/index.shtml

Side Remarks: According to David Pease, GPFS, which also originated from Almaden Research, and Storage Tank are competing product developments within IBM. GPFS remains the primary choice for massively parallel systems, whereas Storage Tank is supposed to address the file serving needs of heterogeneous, distributed environments.
Parallel I/O is not so much in the scope of IBM SAN FS. However, customer requirements might very well lead to the inclusion of parallel I/O into SAN FS. As of today, IBM SAN FS does enable parallel I/O via calls to an MPIO library. Dr. Garth Gibson, Panasas, wants to promote a parallel NFS standard which would also be very interesting for use in SAN FS. David Pease is supposed to join the standardization body for IBM.

Slide 53: Data Management Strategy

Basically, any large computing center runs a computational pipeline which revolves between pre-processing (PP), high performance computing (HPC) and rendering of results (VIS). Aside from the very fast local disks of the HPC stage, we would assume a single shared data exchange that equally serves all sections of the computational pipeline. For very large data sets, parallel tapes will offer an order of magnitude cost reduction which may be used for redundant storage as well.

Slide 54: HLRS Storage Configuration

This slide is to say that a shared storage system may very well extend over 100km, and will use different networks as appropriate: e.g. 10GE for user access and metadata arbitration, 2-/4- and 10-GFC for storage access, and eventually InfiniBand for interprocessor communication.

HPSS, the High Performance Storage System, is a full implementation of the IEEE mass storage reference model version 5. It has been developed by major US national laboratories and IBM Global Services, Houston, in the time frame of 1993 to 1996. HPSS separates metadata operations from user level data transfers for the explicit sake of scalability. All services are parallel in nature and are configured within a general purpose peer IP network. Therefore, from its basic inception, HPSS constitutes an IP storage area network without any network layer or geographic distance limitations. The basic user level services right now are NFS, parallel FTP, the HPSS API and MPI-IO.

Slide 55: Generic Set of Client Interfaces

What is the generic set of client interfaces needed, both for systems and applications?
- 10G Ethernet LAN PHY for interactive traffic and control
- 10G FibreChannel for Inter Switch Links (ISLs) for access to storage devices
- 4 x InfiniBand for interprocessor communication
- maybe STM-64c for cutting through carrier payloads

Long-distance communication lines are governed by a different set of requirements, depending on the service offered by the provider. Service offerings may either include fixed bandwidth (e.g. Ethernet or SDH interface), fixed wavelength (e.g. an optical channel), or eventually an entire dark fiber. Given rental costs of dark fibers in the order of 1€/km over 15 years, a user-operated dark fiber link is definitely preferable for high-bandwidth applications.
Electro-optical transponders are available under the XFP Multi Source Agreement [17], that are able to directly transfer most existing 10-Gbit/s LAN PHYs over distances of 80-120km of dark fiber – depending on the fiber type. This is possible, because modern XFP transponders do employ Electronic Dispersion Compensation (EDC) which allows for an automatic electronic linearization of optical properties of the underlying fiber link. Therefore, the 10-Gbit/s Ethernet, the 10-Gbit/s Fibre Channel and maybe even the four-way InfiniBand (4xIB) MAC layers may be transferred directly over dark fiber at the expense of a single XFP module per communication endpoint.
Passive Coarse Wavelength Division Multiplexers may be used to enable up to eight XFP links over a single dark fiber - still at very low cost.
Higher link data rates will require more sophisti-cated compensation methods for the optical path.

Slide 56: Transmission Impairment Mitigation: FEC

Even though we do not have any client network interface today that exceeds the 10-Gbit/s regime, it is interesting to run communication links at higher data rates. Forward Error Correction (FEC) may be used to compensate for transmission bit errors and thus control the increase of network buffer sizes as well as to adapt to time-varying loss mechanisms, e.g. Polarization Mode Dispersion. Also, directly modulated distributed feedback laser diodes will perfectly work for a 40-Gbit/s data stream, and therefore their cost, including FEC, may be averaged out over four time division multiplex channels.

Slide 57: Reduction of TCO

How can we reduce costs by leaving out unnecessary components?
Leave out unnecessary system components whenever possible, e.g. use direct modulation of DFB lased diodes.

Slide 58: Transparent Switching of 10G Channels

Most electro-optical transceivers according to the XFP Multi Source Agreement exhibit a pull-in range from 9.9 - 11.3 GBd. This allows for covering a broad range of link alphabets with a single serial circuit. With the same token, serial crossbar circuits do exist that allow to build up a CrossConnect functionality for cutting through carrier payloads, or resort to just-in-time switching of packet trains by evaluating an in- or outband control header [18]. Also, CrossConnects may be used to build random multiplex groups, e.g. at 40-Gbit/s rates. In that respect, a CrossConnect may enable multiple parallel network fabrics over long-distance links, where otherwise costly (and lossy) protocol gateways would be required.

Slide 59: High Performance at its time (CWDM at 300BC)

Arie van Praag's CWDM animation

First click: There is a program, called 0-to-42, sponsored by SWR3 Radio here in Southern Germany.
Second click: It's all about getting people ready for a 42-km Marathon run.
Third Click: According to Plutarch, a Greek soldier conducted this run 490 BC in order to inform the citizens of Athens of a glorious victory over the Persians. As anyone knows, the soldier died immediately after he had delivered his message. Only 200 years later, there was an entirely new communication technology.
Fourth Click: We find Fire Towers positioned at regular intervals. First illumination sequence.
Fifth Click: Three more illumination sequences. They are perfectly enabled for optical communication using a limited set of colors (that were achieved by doping the flames with selected halogens).
Sixth Click: Final screen.
Today we would call this „wireless and CWDM". Even the distance between towers is not so far off today's 80-km spans in optical Metro networks.

With this historic excursion I would like to draw your attention back to the future. Don't wait another 2300 years in order to adopt technology that is already available.

Slide 60: BelWü 40-Gbit/s State Research Network

Redundant star-shaped state science network based on DWDM-links with 40-Gbit/s granularity.
Two Carrier routers will switch legacy traffic. Storage access and IPC are either crossconnected or handled by dedicated switching devices. In that respect, CrossConnects do enable multiple parallel network fabrics over long-distance links, where otherwise costly (and lossy) protocol gateways would be required.

Slide 61: Conclusion & Acknowledgement

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

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