FORSCHUNGSZENTRUM JÜLICH GmbH
Zentralinstitut für Angewandte Mathematik
D-52425 Jülich, Tel. (02461) 61-6402

Interner Bericht

Distributed Applications in a German Gigabit WAN

Thomas Eickermann, Wolfgang Frings, Stefan Posse,*
*Gernot Goebbels**, Roland Völpel****

FZJ-ZAM-IB-9911

September 1999

(letzte Änderung: 13.09.99)

Preprint: Proceedings of the eighth IEEE International Symposium on High Performance Distributed Computing, Redondo Beach, California, August 1999

(*) Institute of Medicine
Research Centre Jülich, D–52425 Jülich, Germany
(**) Institute for Media Communication, GMD
Schloss Birlinghoven, D-53754 Sankt Augustin
(***) Institute for Algorithms and Scientific Computing, GMD
Schloss Birlinghoven, D-53754 Sankt Augustin

Distributed Applications in a German Gigabit WAN

Thomas Eickermann, Wolfgang Frings
Central Institute for Applied Mathematics
Research Centre Jülich, D–52425 Jülich, Germany
{Th.Eickermann,W.Frings}@fz-juelich.de

Stefan Posse
Institute of Medicine
Research Centre Jülich, D–52425 Jülich, Germany
S.Posse@fz-juelich.de

Gernot Goebbels
Institute for Media Communication
GMD - German National Research Center
for Information Technology
Schloss Birlinghoven, D-53754 Sankt Augustin
Gernot.Goebbels@gmd.de

Roland Völpel
Institute for Algorithms and Scientific Computing
GMD - German National Research Center
for Information Technology
Schloss Birlinghoven, D-53754 Sankt Augustin
roland.voelpel@gmd.de

Abstract

In order to prepare the upgrade of the national scientific network to Gigabit capacity in the year 2000, two testbeds have been set up in Germany. One of them, the 'Gigabit Testbed West' uses a 2.4 Gigabit/second ATM link to connect the Research Centre Jülich and the GMD – National Research Center for Information Technology in Sankt Augustin. The testbed is the basis for several application projects ranging from metacomputing to multimedia. This contribution gives an overview of the infrastructure of the testbed and the applications. As an example, the realtime-analysis and -visualization of fMRI measurements of the human brain is described in more detail.

1. Introduction

In Germany, research, science, and educational institutions are connected with each other and the rest of the internet via the so-called broadband scientific network (B-WiN). This network is operated by the DFN–Verein, an association of these institutions founded in 1984. Since 1996 this network is based on ATM–technology and allows for access capacities up to 155 Mbit/s. Extrapolations of the growth rates of the last years show that the current infrastructure will reach its limit in the next year. Therefore, an upgrade into the Gbit/s range on a national basis is planned for the beginning of the year 2000. To prepare the transition, the DFN–Verein has initiated two gigabit testbeds in Germany. The first of them is the 'Gigabit Testbed West'. It connects the Research Centre Jülich and the GMD – National Research Center for Information Technology in Sankt Augustin close to Bonn. The second one, the 'Gigabit Testbed Bavaria/Berlin', connects several institutions in Munich, Erlangen, and Berlin.

The Gigabit Testbed West is used by a couple of application projects that cover various aspects of distributed high performance computing (metacomputing) and multimedia. They can rely on a solid base of installed supercomputer capacity. Jülich is equipped with 512–node Cray T3E–600 and 512–node T3E–1200 massively parallel computers and a 10–processor Cray T90 vector–computer. An IBM SP2, a 12–processor SGI Onyx 2 visualization server, and a 8–processor SUN E500 are installed in the GMD. Each application has communication requirements that cannot be matched by the 155 Mbit/s available in the B-WiN. This presentation gives an overview over the applications as well the infrastructure of the testbed (including the network and metacomputing tools).

2. The Gigabit Testbed West

The Gigabit Testbed West started in August 1997. In the first year of operation the two locations — which are approximately 100 km apart — were connected by an OC–12 ATM link (622 Mbit/s) based upon Synchronous Digital Hierarchy (SDH/STM4) technology. This connection is provided by o.tel.o Service GmbH and uses the optical fiber infrastructure inside the power lines of the German power supplier RWE AG. In August 1998 the link was upgraded to OC–48 (2.4 Gbit/s). At the same time Fore Systems ATM switches (ASX–4000) were installed in the framework of a beta-test. They connect the local networks of the research

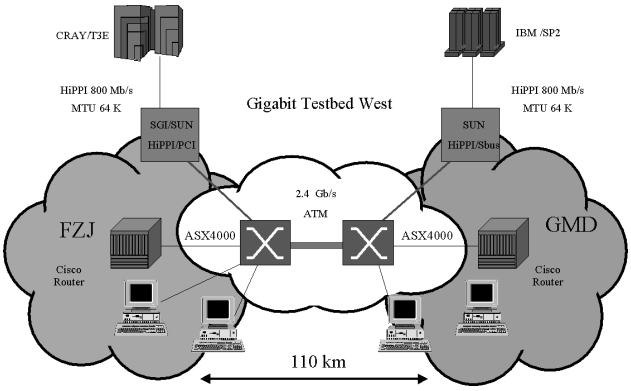

Figure 1. Configuration of the Gigabit Testbed West in June 1999. Jülich and Sankt Augustin are connected via a 2.4 Gbit/s ATM link. The supercomputers are attached to the testbed via HiPPI-ATM gateways, several workstations via 622 or 155 Mbit/s ATM interfaces.

centers to the OC–48 line. Initial stability problems that were observed during the test turned out to be related to signal attenuation and timing. Those problems have been solved and both the SDH link and the switches are in stable operation now.

The attachment of the supercomputers to the testbed imposes more problems. While 622 Mbit/s ATM interfaces are now available for all common workstation platforms, this is not the case for the supercomputers used in the project. Therefore a solution based on the traditional 'High Performance Parallel Interface' (HiPPI) was implemented. HiPPI offers a peak performance of 800 Mbit/s when a low-level protocol and large transfer blocks (1 MByte or more) are used. Even with TCP/IP communication, transfer rates of more than 430 Mbit/s are achieved within the local Cray complex in Jülich when an MTU of 64 KByte is used. The HiPPI networks of the Crays and the IBM SP2 were connected to the ATM backbone using workstations as IP gateways. Currently, an SGI O200 and a Sun Ultra 30 in Jülich and a SUN E5000 in Sankt Augustin are equipped with Fore 622 Mbit/s ATM adapters and Essential HiPPI adapters. Since the Fore ATM adapter supports large MTU sizes, IP packets of 64 KByte size can be transferred throughout the network. Besides that, 8 nodes of the IBM SP2 are equipped with 155 Mbit/s ATM adapters. First measurements show a throughput of more than 260 Mbit/s between the Cray T3E in Jülich and the IBM SP2 in Sankt Augustin. This is mainly due to the limitations of the I/O–system of the microchannel–based SP–nodes. Details are given in a separate publication [5].

3. Metacomputing Tools and Applications

A serious limitation of distributed metacomputing environments is that latency and bandwidth of the connecting network cannot compete with the performance of the internal communication paths of massively parallel supercomputers, vector supercomputers or SMP machines. Because of that, only certain classes of applications can benefit from metacomputing. One such class is represented by so-called 'coupled fields' applications. Here, two or more space- and time–dependent fields interact with each other. When the fields are distributed over the machines of the metacomputer these components are often only loosely coupled leading to moderate communication needs. A second class of applications benefits from being distributed over supercomputers of different architecture, because they contain partial problems which can best be solved on massively parallel or vector-supercomputers. For other applications, real-time requirements are the reason to connect several machines.

A key factor for the success of metacomputing in a large production environment is of course the availablity of a software infrastructure that makes the metacomputer usable for a broad range of users. Various projects deal with these problems [3, 2]. In the experimental setup of the Gigabit Testbed West, we focus on the basic tools needed for program development. These are a tool for performance evaluation and tuning of metacomputing applications and a metacomputing-aware communication library. With metacomputing–aware we mean that the communication both inside and between the machines that form the metacomputer should be efficient. Furthermore, a couple of features that are part of the the MPI–2 [6] definition are useful for metacomputing applications and have been implemented. Dynamic process creation and attachment e.g. can be used for realtime–visualization or computational steering; language–interoperability is needed to couple applications that are implemented in different programming languages. Currently, this library is available for Cray T3E, IBM SP2 and Sun Solaris. Ports to Cray T90 and SGI IRIX are under way. Details on the implementation and performance figures are given elsewhere [1].

What follows is a short list of the application projects defined in the testbed. For all applications, the Central Institute for Applied Mathematics at the Research Centre Jülich (FZJ) and the Institute for Algorithms and Scientific Computing (SCAI) at the GMD participate by giving support in metacomputing.

- **Metacomputing Tools**

  Partners Pallas GmbH.

  Synopsis Provide the applications with basic tools: a metacomputing-aware MPI implementation (including the above mentioned parts of the MPI-2

standard) is implemtented by Pallas. The parallel tracing tool VAMPIR [7] is extended for the use with this library.

- **Transport of solutants in ground water**

  Partners Institute for Petroleum and Organic Geochemistry (FZJ).

  Synopsis Coupling of two independent programs for ground water flow simulation (TRACE) and transport of particles in a given water flow (PARTRACE).

  Communication Transfer of the 3–D water flow field from IBM SP2 (TRACE) to Cray T3E (PARTRACE) every timestep, up to 30 MByte/s.

- **Distributed computation of climate- and weather models**

  Partners Alfred Wegener Institute for Polar and Marine Research, German Climate Computing Center, and SCAI/GMD.

  Synopsis Coupling of an ocean–ice model (based on MOM-2) running on Cray T3E and an atmospheric model (IFS) running on IBM SP2 using the CSM flux coupler.

  Communication Exchange of 2–D surface data every timestep, up to 1 MByte in short bursts.

- **Analysis of magnetoenzephalography data**

  Partners Institute of Medicine (FZJ).

  Synopsis A parallel program (pmusic), that estimates the position and strength of current dipoles in a human brain from magnetoenzephalography measurements using the MUSIC algorithm is distributed over a massively parallel and a vector supercomputer to achieve superlinear speedup.

  Communication Low volume, but sensitive to latency.

- **Visualization of Realtime fMRI**

  Partners Institute of Medicine (FZJ), Institute for Media Communication (GMD).

  Synopsis, Communication See below.

- **MetaCISPAR**

  Partners Pallas GmbH, SCAI/GMD.

  Synopsis An open interface (COCOLIB) that allows the coupling of industrial structural mechanics and fluid dynamics codes is ported to the metacomputing environment.

  Communication Depends on the coupled application.

- **Multimedia in a Gigabit-WAN**

  Partners Institute for Media Communication (GMD).

  Synopsis Basic technology for transferring studio-quality digital video over ATM is examined.

  Communication E.g. 270 Mbit/s for an uncompressed D1 video stream.

4. Visualization of Realtime fMRI

As an example for a complex heterogeneous metacomputing scenario, the realtime analysis and visualization of functional magnetic resonance imaging (fMRI) measurements at the Institute of Medicine (IME) at the Research Center Jülich is briefely described here. In this application a 1.5 Tesla Siemens Vision MRI-scanner is coupled with the Cray T3E in Jülich and the SGI Onyx 2 in Sankt Augustin for online processing and visualization of human brain activity. The results are displayed on a Responsive Workbench that can either be located in Jülich or Sankt Augustin. Figure 2 illustrates this setup.

In the experimental setup, a subject (test person) is exposed to e.g. periodic visual or acoustic stimulations while his/her head is scanned by an MRI-scanner with repetition times of up to 2 seconds. The 3D image matrix is typically 64x64x16 voxels, but larger matrices can be measured at

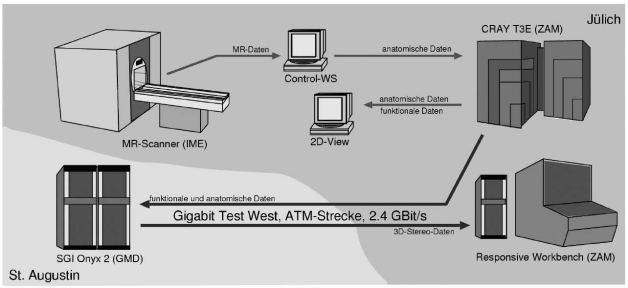

Figure 2. Setup of the fMRI experiment. The raw scanner data are transferred through a front-end workstation to the T3E where they are processed. From there, anatomical and functional brain images are transferred to either a workstation with a 2-D display or over the testbed to an Onyx 2 in the GMD. The rendered images are sent back over the testbed to a Responsive Workbench in Jülich. A 2-processor SGI Onyx 2 in Jülich will be used as frame–buffer for the workbench.

correspondingly lower temporal resolution. It is possible to identify brain activity by correlating the measured signal with a so-called reference vector which represents a convolution of the stimulation time course with a hemodynamic response function. The latter takes into account the delay and dispersion of the blood flow in response to neuronal activation. The measurement relies on the effect, that variations of the blood–oxygen saturation result in changes of the magnetic field inhomogeneities (BOLD-effect) [8]. In order to control e.g. data quality and subject movement and to optimize the experimental parameters it is highly desireable to perform the analysis of the data and the visualization within a few seconds. For this purpose a software package named FIRE (Functional Imaging in REaltime) has been developed at the IME. FIRE includes an 'RT–server' that runs on the front–end workstation of the scanner. It serves as an interface between the scanner and the 'RT–client'. The latter processes and displays the raw images obtained from the server. The RT–client is operated via a Motif–based graphical user interface (GUI) that is shown in figure 3. It is run on a standard UNIX workstation and is able to perform the basic processing steps to analyse and display the data within the acquisition time of 2 seconds. This includes the following steps:

1. The raw images are transferred from the control-workstation of the scanner to the workstation that runs FIRE. This requires a slight modification of the operating system of the Siemens MRI scanner.

2. For each voxel, the correlation between the measured signal and a fixed reference vector is calculated.

3. The results are visualized in 2–D. For those pixels of each slice, for which the correlation coefficient is larger than an adjustable clip-level, the anatomical data are overlayed with the color-coded correlation coefficient.

However, a couple of more complex tasks require supercomputer capacity in order to be performed in realtime. Therefore, the RT–client was modified such that it can delegate parts of the work to the Cray T3E in Jülich in a 'remote procedure call' like manner. The following set of modules have been implemented on the T3E using a domain decomposition of the brain.

Spatial filters a median filter is used to reduce noise in the unprocessed picture. After the processing pipeline, the data can be smoothened by an averaging filter.

3–D movement correction even small head movements of the subject tend to produce artefacts in the correlation coefficient due to the high intrinsic contrast of the MR images. Therefore it is extremely important to compensate for these movements. Here an iterative linear scheme is used.

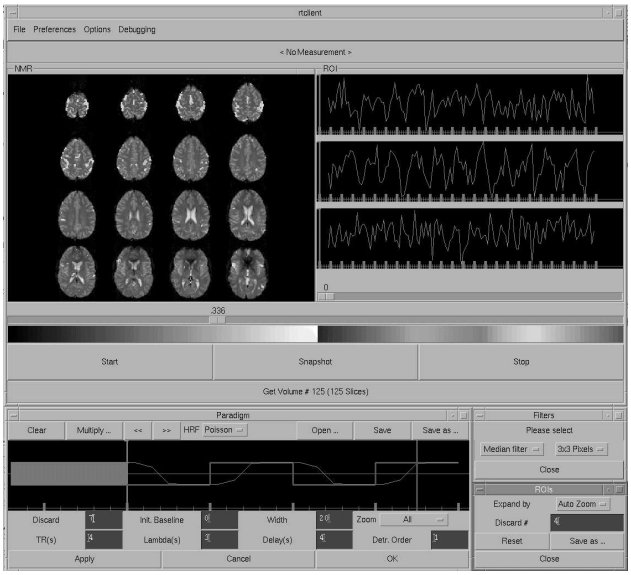

Figure 3. Control panel and 2–D display of the FIRE software. The upper left canvas shows MR-images with a color coded correlation map overlay. In the upper right part, the signal time courses of special 'regions of interest' can be displayed. In the lower panel, the stimulation time course and the modeled hemodynamic response can be specified.

Detrending the measured signal often includes slow baseline drifts. A compensation using a few detrending-vectors can compensate for that.

Reference vector optimization (RVO) the sensitivity of the correlation procedure depends on the quality of the model of the hemodynamic response. While the overall behaviour is quite well known, there are variations in delay and duration of the response. In the workstation-only version of FIRE, these parameters can be adjusted manually between the measurements. On the T3E, a fully automatic least-squares fit of delay and duration is performed for each voxel during the measurement. The procedure rasters the parameter space to find the global minimum. The RVO not only improves the sensitivity but could also allow to study differences in these parameters between brain regions and temporal changes of these parameters during repetitions of the stimulation.

The use of each module is optional and can be controlled

during runtime via the GUI of the RT–client. In table 1 the time spent by a T3E-600 to process a 64x64x16 image using these modules is listed for various number of processors (PEs). As the measurements show, a reasonable speedup is achieved for up to 128 PEs. The most time consuming module is the RVO. Here further optimizations are planned for the near future (e.g. the resolution of the grid can be reduced and the solution refined using a conjugate gradient method). We expect that it will then be possible to run the whole set of modules on a mid–range parallel computer.

number of PEs	filter	motion corr.	RVO	total time	speedup
1	0.18	1.55	109.27	111.00	1
2	0.09	0.91	54.65	55.65	2.0
4	0.05	0.56	27.36	27.97	4.0
8	0.03	0.46	13.74	14.23	7.8
16	0.02	0.35	6.93	7.30	15.2
32	0.02	0.33	3.51	3.86	28.7
64	0.03	0.35	1.85	2.22	50.0
128	0.03	0.34	1.00	1.37	81.1
256	0.04	0.40	0.59	1.01	110.5

Table 1. Time spent for processing a 64x64x16 image on the Cray T3E for various number of PEs. All times are given in seconds. Larger images take more time, but achieve better speedups.

To estimate the total delay between the MR–scan of an image and the display of the correlation map on the 2–D GUI, several other processes have to be taken into account. The RT–server receives the data approximately 1.5 seconds after the scan (for a 64x64x16 image). The data transfers and the exchange of control messages between the RT–server, the T3E and the RT–client sum up to 1.1 seconds. Another 0.6 seconds elapse after the data has arrived at the client until it actually appears on the screen. When 256 PEs are used on the T3E, this leads to a total delay of less than 5 seconds. Such a short delay is not required for the control of typical experiments. However, it enables new opportunities for neuroscience research like bio–feedback (the subject watching his own brain in action).

In the current implementation of the software the sequence of processing the images remains the same, no matter wether the calculations are performed by the RT–client or the T3E. The drawback of this simple approach is that we make no use of the possibility to pipeline the work. In particular, a new image is requested from the RT–server only after the processing and displaying of the previous one is completed. Therefore, the throughput of the application (the rate at which it can process images) is the sum of the delays in the RT–client and the T3E, which is 2.7 seconds in

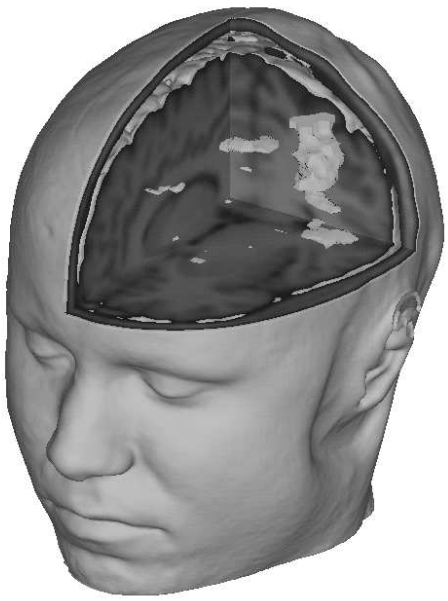

Figure 4. A human head generated from MRI data using AVS. The light areas are regions of the brain that are activated by moving the right hand.

the above example. This means that the scanner can safely be operated with a repetition rate of 3 seconds – a typical parameter for many experiments performed in Jülich. While the time needed for data aquisition and analysis are well balanced in our current system, advanced MR imaging techniques which are under development [9] will produce data rates that are an order of magnitute beyond what is feasible today. Analysing this data in realtime will be a challenging task for a supercomputer again.

A 3–D visualization in a virtual reality environment can support the interpretation of the results or can be used for demonstration or educational purposes, but requires high-end visualization hardware. Therefore the functional data are transferred to the 12-processor SGI Onyx 2 in Sankt Augustin as the calculation proceeds. Here it is merged with a high resolution (256x256x128 voxels) image of the subject's head. Such images are usually produced before the actual measurement begins. These data are visualized on a Responsive Workbench, where they can be rotated, zoomed, and sliced in realtime using the VR-software AVOCADO developed by the GMD. Currently, the Workbench in use is located at the GMD. Within this project it is planned to extend AVOCADO such that also remote display systems can be used. Then the data will be displayed on a Workbench that was recently installed in Jülich. This will be implemented as soon as 622 Mbit/s ATM interfaces for the

SGI Onyx 2 are available. They are neccessary since an enormeous amount of data has to be transferred for this application: the workbench has two projection planes, each of them displays stereo images of 1024x768 true color (24 Bit) pixels. This means that less than 8 frames/second can be transferred over a 622 Mbit/s ATM network using classical IP.

Figure 4 shows a prototype, that has been implemented using the AVS (Advanced Visualization System) software running on a workstation. While (on a high–end graphical workstation) the update of the functional data takes about the same amount of time as the display on the 2–D GUI, this setup is too slow for interactive manipulations as described above. It should be noted that similar work is also performed at the Pittsburgh Supercomputing Center [4].

5. Extensions of the Testbed

The testbed is currently extended by connecting new sites to the original link between Jülich and Sankt Augustin and by defining new applications that use those extensions. A dark fibre that links the national German Aerospace Research Center (DLR) and the University of Cologne to the GMD has just been set up. This line is used for projects that range from distributed traffic simulation and visualization to distributed virtual TV-production (in cooperation between GMD, DLR, Academy of Media Arts in Cologne, and echtzeit GmbH). The latter relies on the results of the multimedia project. A new 622 Mbit/s ATM-link between the University of Bonn and the GMD will be the basis for metacomputing projects that deal with multiscale molecular dynamics and lithospheric fluids.

6. Conclusions

This contribution gave an overview over the metacomputing activities in the German Gigabit Testbed West. The underlying 2.4 Gbit/s SDH and ATM technology for the wide area backbone seems to be mature, a neccessary condition for the upgrade of the German scientific network that is planned for the year 2000. In contrast to that, the networking capabilities of the supercomputers that are attached to the testbed have to be improved. The concept of a HiPPI/ATM gateway seems to be viable. A couple of applications that deal with various aspects of metacomputing are using the infrastructure of the testbed. The realtime fMRI project that is described in some detail here uses a quite complex configuration. Up to 5 computers and a MRI–scanner have to cooperate simultaneously. When — as in our case — the bandwidth of the connecting network is sufficiently large this can be accomplished with good performance and stability. Nevertheless, the problem of simultaneous resource allocation in a distributed environment will become more apparent when the application is used for clinical research.

7. Acknowledgements

Many of the activities that are reported in this contribution are not the work of the authors but of several persons in the institutions that participate in the Gigabit Testbed West project. The authors wish to thank D. Conrads, D. Gembris, T. Graf, R. Niederberger, M. Sczimarowski, H. Vereecken, and H. Zilken from the Research Centre Jülich, U. Eisenblätter, H. Grund, F. Hommes, W. Joppich, M. Göbel, M. Kaul, E. Pless, R. Völpel, K. Wolf, P. Wunderling, and L. Zier at the GMD, W. Hiller and T. Störtkuhl at the AWI, V. Gülzow at the DKRZ and J. Henrichs and K. Solchenbach at Pallas GmbH, to mention but a few. We also wish to thank the BMBF for partially funding the Gigabit Testbed West and the DFN for its support.

References

[1] T. Eickermann, H. Grund, and J. Henrichs. Performance issues of distributed MPI applications in a German gigabit testbed. *accepted for Euro PVM/MPI, Barcelona*, September 1999.

[2] D. Erwin. The unicore architecture and project plan. In *Workshop on Seamless Computing*, pages 16–17, ECMWF, Reading, UK, 1997.

[3] I. Foster and C. Kesselman. The globus project: A status report. *Proc. IPPS/SPDP '98 Heterogeneous Computing Workshop*, pages 4–18, 1998.

[4] N. Goddard, G. Hood, J. Cohen, W. Eddy, C. Genovese, D. Noll, and L. Nystrom. Online analysis of functional mri datasets on parallel platforms. *Journal of Supercomputing*, 11:295–318, 1997.

[5] H. Grund, F. Hommes, R. Niederberger, and E.Pless. High speed supercomputer communications in broadband networks. *TERENA–NORDUnet Networking Conference, Sweden*, June 1999.

[6] Message Passing Interface Forum. *MPI-2: Extensions to the Message-Passing Interface*. University of Tennessee, http://www.mcs.anl.gov/mpi/index.html, 1997.

[7] W. Nagel, A. Arnold, M. Weber, H. Hoppe, and K. Solchenbach. Vampir: Visualization and analysis of mpi resources. *Supercomputing*, 63 XII no. 1:69–80, 1996.

[8] S. Ogawa, T. Lee, A. Kay, and D. Tank. Brain magnetic resonance imaging with contrast depending on blood oxygenation. *Proc. Natl. Acad. Sci. USA*, 87:9868–9872, 1990.

[9] S. Posse, S. Wiese, D. Gembris, K. Mathiak, C. Kessler, M.-L. Grosse-Ruyken, B. Elghawaghi, T. Richards, S. Dager, and V. G. Kiselev. Enhancement of BOLD–contrast sensitivity by single–shot multi–echo functional MR imaging. *Magnetic Resonance in Medicine*, in press.