UNIVERSITÄT PADERBORN
Die Universität der Informationsgesellschaft

Annual Report 2005

PADERBORN
CENTER FOR
PARALLEL
COMPUTING

University of Paderborn
Paderborn Center for Parallel Computing
Fürstenallee 11, D-33102 Paderborn

www.upb.de/pc2

Table of Contents

1 Preface

Das Paderborn Center for Parallel Computing (PC²) ist eine zentrale wissenschaftliche Einrichtung der Universität Paderborn und arbeitet als Forschungs- und Dienstleistungszentrum im Bereich des parallelen und verteilten Rechnens. Die wesentlichen Aufgaben des PC² liegen in der Entwicklung und Bereitstellung innovativer paralleler Rechnersysteme für die Hochschulen des Landes Nordrhein-Westfalen sowie in der Entwicklung von Methoden und Verfahren zur effizienten Nutzung paralleler und verteilter Systeme.

Die wichtigste Veränderung im Jahr 2005 betrifft die Erneuerung der Rechenressourcen. Nach sechs Jahren Einsatz wurde das Clustersystem hpcLine durch das innovative, leistungsfähige Clustersystem ARMINIUS ersetzt, das über integrierte Visualisierungsfunktionalitäten verfügt und damit eine umfangreiche, hochqualitative und echtzeitnahe grafische Darstellung komplexer Vorgänge erlaubt. Diese Eigenschaften sind für viele Anwendungen aus den Ingenieur-, Wirtschafts- und Naturwissenschaften von Interesse, so dass sowohl die bestehenden interdisziplinären Forschungsaktivitäten fortgeführt als auch neue Forschungsfelder erschlossen werden können. Der ARMINIUS Cluster belegte zeitweise den Platz 205 auf der Liste der 500 weltweit schnellsten Rechner und steht allen Nutzern des PC² zur Verfügung. An der Einweihungsfeier am 21. Juni 2005 nahmen herausragende Persönlichkeiten aus Wissenschaft, Wirtschaft und Politik teil und begrüßten den nächsten Ausbauschritt des PC². Dr. Horst Simon – Direktor des Lawrence Berkeley National Laboratory (LBNL) und des National Energy Research Scientific Computing Centers (NERSC) – hielt den Hauptvortrag im wissenschaftlichen Kolloquium und zeigte die Zukunftsperspektiven des Hochleistungsrechnens auf. Eine Vielzahl von wissenschaftlichen Anwendungen aus Natur- und Ingenieurwissenschaften sowie Demonstrationen der im PC² durchgeführten Projekte rundeten die Veranstaltung ab.

Der Systemzugang wird durch stetige Weiterentwicklung der Managementsoftware sowie durch Integration moderner Grid Technologien weiter vereinfacht. Zusätzlich werden individuelle Angebote für unsere Benutzer erstellt, die Zugang zu rechenintensiven Anwendungen für ihre Forschungs- und Projektarbeit benötigen. In einem Kooperationsmodell werden Forschungsanwendungen installiert, gepflegt und deren Benutzer durch das PC² Team betreut. Die aktuelle Auslastung der Maschine von über 90% zeigt den Erfolg dieser erweiterten Dienstleistungsstrategie und betont die Kopplung von leistungsfähiger Hard- und Software mit kompetenter Beratung und Betreuung als einen wesentlichen Baustein für erfolgreiche Forschung in vielen Arbeitsgruppen der Universität Paderborn und darüber hinaus. Die aktuell betreuten

Systeme, Anwendungen, die Ansprechpartner im PC² sowie die Vorgehensweise zur Registrierung neuer Benutzer finden sich in diesem Jahresbericht (näheres ab Seite 47). Schließlich wollen wir auf die regelmäßigen Veranstaltungen „Tag der offenen Tür“ hinweisen, in denen aktuelle Themen des Hochleistungsrechnens vorgestellt werden und eine Führung durch das PC² angeboten wird.

Über den mit Nachdruck verfolgten Dienstleistungscharakter des PC² als Forschungsinstitut mit einem attraktiven Serviceangebot, wurden von der Kerngruppe wichtige und interessante Forschungsbeiträge präsentiert. Die Forschungsfelder parallele numerische Simulation, innovative Hardwarekonzepte, Systemsoftware für Parallelverarbeitung und Grid Computing wurden durch zahlreiche, international anerkannte Forschungsaktivitäten weiter ausgebaut. Ferner wirkte das PC² an einer Vielzahl von nationalen und internationalen Forschungsanträgen mit, so dass mittlerweile elf drittmittelfinanzierte Projekte durchgeführt werden. Im Jahr 2005 kamen die DFG-geförderten Projekte MOVES (*Multi-Objective Intrinsic Evolution of Embedded Systems*) und ReconOS (*Operating System for Dynamically Reconfigurable Hardware*) sowie das BMBF Projekt D-Grid hinzu. Die Firma Intel unterstützt die Entwicklung von Mehrkernprozessoren mit einem der begehrten Intel Research Awards. Ferner wurde das EU Projekt AssessGrid über Risikobewertung und Management in Grid Umgebungen unter der Leitung des PC² genehmigt. Im Rahmen der Zielvereinbarung „VisSim – Kompetenzzentrum für Visualisierung und Simulation“ wurde ein interdisziplinäres Projekt mit den Fakultäten Maschinenbau und EIM gestartet, bei der die effiziente Kopplung von Simulation und Visualisierung am Beispiel von virtuellen Nachtfahrten demonstriert wird. Die Beschreibungen dieser Projekte können dem vorliegenden Jahresbericht entnommen werden.

Das PC² ist weiterhin aktiv an der Etablierung einer europaweiten und nationalen Grid-Infrastruktur beteiligt und wirkt an der Gestaltung dieser Zukunftstechnologie für die Universität Paderborn erfolgreich mit. Ferner ist das neue Clustersystem im Rechnerverbund NRW integriert. Damit wird sowohl der Zugang externer Nutzer auf die Paderborner Ressourcen als auch der Zugang für Forscher der Universität Paderborn auf andere Rechner in NRW deutlich vereinfacht.

In Sachen Personalwesen freuen wir uns sehr, dass der PC² Vorstand durch die Berufung von Prof. Dr. Dellnitz (Mathematik) und Prof. Dr. Platzner (Informatik) ergänzt werden konnte. Frau Dr. Kerstin Wielage verließ das PC² nach erfolgreichem Promotionsabschluss.

Zum Schluss dieses Vorwortes möchten wir alle Interessenten herzlich einladen, sich über das PC² zu informieren, bei uns zu rechnen und aktiv die Zukunft des PC² mitzugestalten.

Paderborn, 27. April 2006

Prof. Dr. Burkhard Monien, Vorsitzender des Vorstandes
Prof. Dr. Odej Kao, Geschäftsführender Leiter

Preface

The Paderborn Center for Parallel Computing (PC²) is one of the central scientific institutes at the University of Paderborn. It serves as a research and service center for parallel and distributed computing. The main tasks of PC² are the development and operation of innovative parallel compute systems for the universities in the federal state of North Rhine-Westphalia (NRW). Moreover, PC² scientists develop methods for the efficient use and management of parallel and distributed systems.

Selecting from various events in the year 2005, we would like to highlight the modernisation of our computing resources. After six years of continuous operation, the hpcLine system was replaced by a new high-performance cluster system named ARMINIUS, which integrates compute and visualisation components and allows a high-quality and near real-time graphic presentation of complex processes. Stereoscopic equipment provides a three-dimensional viewing experience. These properties are of particular importance for manifold applications and the research in natural sciences and engineering. Thus, the existing interdisciplinary activities will be intensified and new research fields will be addressed. The ARMINIUS cluster was ranked on position 205 of the 500 most powerful computers world-wide and is available for all users of PC². In June 2005 many prominent guests from academia, society and the commercial sector celebrated with us the inauguration of the ARMINIUS cluster and the further expansion of PC². Dr. Horst Simon – Director of the Berkeley Labs and the National Energy Research Scientific Computing Centre – gave the key note talk in the scientific colloquium describing the future challenges of high-performance computing. Further talks on research in scientific computing and its applications in natural sciences and engineering as well as a number of demonstrations of projects developed in PC² completed this very successful event.

The access to the compute resources in PC² is continuously improved by steady development of the administration software as well as by integration of modern Grid technologies. Additionally, we offer tailor-made solutions to users with resource requirements for their compute-intensive research and project work. In the scope of a cooperation model, we install and maintain research and standard applications and provide continuous support to the users of these applications. The current utilization of more than 90% underlines the success of our extended service and support strategy. The tight link between high-performance hardware and applications on the one hand and the competent support on the other hand represents an important building block for successful research in many groups of the University of Paderborn and beyond. The list of supported systems, applications, contact data of the support teams as well as the registration for new users are explained in this report (please see page 47). Finally, open days at PC² are taking place every month. In the scope of these events, we present our new services, resources and research results.

Beside the extensive service activities, the core research group of the PC² presented a number of important scientific, internationally recognised contributions in the research fields of parallel numerical simulation, innovative hardware architectures, system software for high-performance computing and Grid Computing. The research and project work has been intensified by compiling and coordinating various national and international funding proposals. Meanwhile eleven funded projects are established and running. In 2005, the DFG-supported projects MOVES (*Multi-Objective Intrinsic Evolution of Embedded Systems*) and ReconOS (*Operating System for Dynamically Reconfigurable Hardware*) as well as the BMBF-funded project D-Grid (*German Grid Infrastructure*) have started. Intel supports the development of heterogeneous multi cores for high-performance computing with one of the prestigious Intel research awards. The EU-funded and PC²-coordinated project AssessGrid related to risk assessment and management in Grid environments has been successfully negotiated. Finally, the VisSim project has been started jointly with our interdisciplinary partners and aims at the creation of a competence centre for distributed visualisation and simulation. The pilot application demonstrates an efficient coupling of simulation and visualisation at the example of a virtual night drive. The associated project descriptions are contained in this report.

Furthermore, PC² is actively supporting the on-going establishment of Grid Computing platforms in Europe as well as in Germany and provides the University of Paderborn with a prominent role in the design of this future technology. Moreover, the new cluster is included in the resource union “RV-NRW”. With this step the access for external users to the resources in Paderborn and the access for researchers from the University of Paderborn to other systems in NRW will be simplified significantly.

In the area of human resources, we are very happy to announce that the PC² board of directors was successfully extended by Prof. Dr. Dellnitz (Mathematics) and Prof. Dr. Platzner (Computer Science). Dr. Kerstin Wielage successfully defended her Ph.D. work.

Finally, as this foreword comes to an end, we would like to welcome all those who are interested in obtaining more information as well as those who want to use the facilities for computations and to participate in carving the future of the PC².

Paderborn, 27. April 2006

Prof. Dr. Burkhard Monien, Chairman of the Board
Prof. Dr. Odej Kao, Managing Director

2 Inside PC²

2.1 Board

The PC² is headed by an interdisciplinary board comprising professors from various working groups.

2.2 Members of the Board

Prof. Dr. Burkhard Monien (Chairman)
Faculty of Electrical Engineering, Computer Science and Mathematics

Prof. Dr. Odej Kao (Managing Director)
Faculty of Electrical Engineering, Computer Science and Mathematics

Prof. Dr. Wilhelm Dangelmaier
Faculty of Business Administration and Economics

Prof. Dr. Michael Dellnitz
Faculty of Electrical Engineering, Computer Science and Mathematics

Prof. Dr. Thomas Frauenheim
Faculty of Science

Prof. Dr. Hans-Ulrich Heiß
Institute for Telecommunication Systems, Technical University Berlin

Prof. Dr. Joachim Lückel
Faculty of Mechanical Engineering

Prof. Dr. Marco Platzner
Faculty of Electrical Engineering, Computer Science and Mathematics

Prof. Dr. Franz Josef Rammig
Faculty of Electrical Engineering, Computer Science and Mathematics

Prof. Dr. Ulrich Rückert
Faculty of Electrical Engineering, Computer Science and Mathematics

Prof. Dr. Otto Rosenberg
Faculty of Business Administration and Economics

Prof. Dr. Hans-Joachim Warnecke
Faculty of Science

Dr. Jens Simon
Paderborn Center for Parallel Computing
Assistant researchers´ representative

Dipl.-Inform. Sabina Rips
Faculty of Electrical Engineering, Computer Science and Mathematics
Assistant researchers´representative

Dipl.-Inform. Axel Keller
Paderborn Center for Parallel Computing
Non researchers´ representative

Christian Biermann
Student representative

2.3 PC² Staff

The following people were assigned to the PC² for the period of time covered by this report.

Dipl.-Inform. Dominic Battré (since December 2005)
Dipl.-Inform. Bernard Bauer
Dr. Stephan Blazy
Dipl.-Inform. Felix Heine
Dipl.-Inform. Matthias Hovestadt
Diana-Mercedes Hunecke (Trainee)
Ulrich Jahnke (Trainee until February 2005)
Dipl.-Inform. Paul Kaufmann (Since September 2005)
Dipl.-Inform. Axel Keller
Michaela Kemper (Secretary)
Dipl.-Ing. Andreas Krawinkel
Dipl.-Inform. Stefan Lietsch (since February 2005)
Dipl.-Inform. Oliver Marquardt
Holger Nitsche

Dr. Jens Simon
Dipl.-Inform. Kerstin Voß (since April 2005)
Dr. Kerstin Wielage (until March 2005)

Additional support was provided by student members who were engaged part time (9.5 h/week and 19 h/week) in tasks which included programming, user support, system administration, etc.

Henrich Blöbaum	Thomas Heinen	Simon Richter
Robert Breitrück	André Höing	Sebastian Ritter
Martin Eikermann	Diana Kleine	Tobias Schumacher
Hesham Elmorsy	Matthias Köhne	Christian Todtenbier
Dominic Eschweiler	Christoph Konersmann	Jan Henrik Wiesner
Christian Fromme	Ralf Kruse	Veit Wittenberg
Alexander Gretencord	Jens Lischka	
Björn Hagemeier	Peter Quiel	

In the year 2005 the PC² employed two trainees to learn the trade of a “computer specialist” (Fachinformatiker) in the field of system integration. This trade is one of the new IT-professions, which was installed in 1997. With the source required to employ trainees provided by the North Rhine-Westphalia government, the PC² was able to oversee this priority assignment.

3 Services

3.1 Operated Parallel Computing Systems

The focus of the Paderborn Center for Parallel Computing is to bring prototypical parallel high performance computing (HPC) infrastructures to productivity. Hence the PC² is not a computing center in the traditional way providing off−the−shelf hardware and software. However, on demand we install and support standard software packages on our systems.

In 2005 the PC² operated seven high performance computing systems and one parallel file system. Five of the HPC systems were dedicated to specific projects and/or working groups. Two HPC systems were available for all users.

3.1.1 Publicly Available Systems

ARMINIUS Cluster

Since 1998, PC² operates the 192 processor hpcLine system (PSC2) from Fujitsu-Siemens Computers. At that time, this system was one of the fastest computer systems of the world (number 351 in the Top500 list). After six years of operation, the PSC2 system was no more attractive for HPC users. Actual requirements of the PC²´s users were: state of the art overall compute performance, fat compute nodes with at least two processors and large size of main memory, high end visualization capabilities, and everything tightly connected with a high performance interconnect. An additional essential requirement from the operating side is that the system has to fit in the computing center in size, electrical power, and cooling.

With the financial support of the state North Rhine-Westphalia and Federal Republic of Germany, PC² was able to accomplish the procurement of a new high performance computer system. In co-operation with Fujitsu-Siemens Computers we designed the ARMINIUS cluster system consisting of 200 compute and 8 visualization nodes. The new system is again the first system of a new hpcLine generation of Fujitsu-Siemens Computers.

Hardware	***Description***
200 dual processor Intel Xeon	3.2 GHz, 1 MByte L2-cache 4 GByte DDR2 main memory 4x InfiniBand PCI-e HCA
8 dual processor AMD Opteron	2.4 GHz, 1 MB L2-cache 12 GByte DDR main memory 4x InfiniBand PCI-e HCA nVidia Quadro FX 4500G PCI-e graphics card
216 port InfiniBand switchFabric	central switch fabric with 18 switch modules each with 12 ports
5 TByte FC-RAID subsystem	All nodes are connected with InfiniBand via an IB/FC switch to the storage subsystem
7 TByte parallel file system	Accessible from all nodes
2 login nodes	Nodes are used for compiling and starting user applications
Stereoscopic rear projection	1.80m x 2.40m screen 2 D-ILA projectors 3D tracking system

Table 1: Hardware specification of the ARMINIUS Cluster

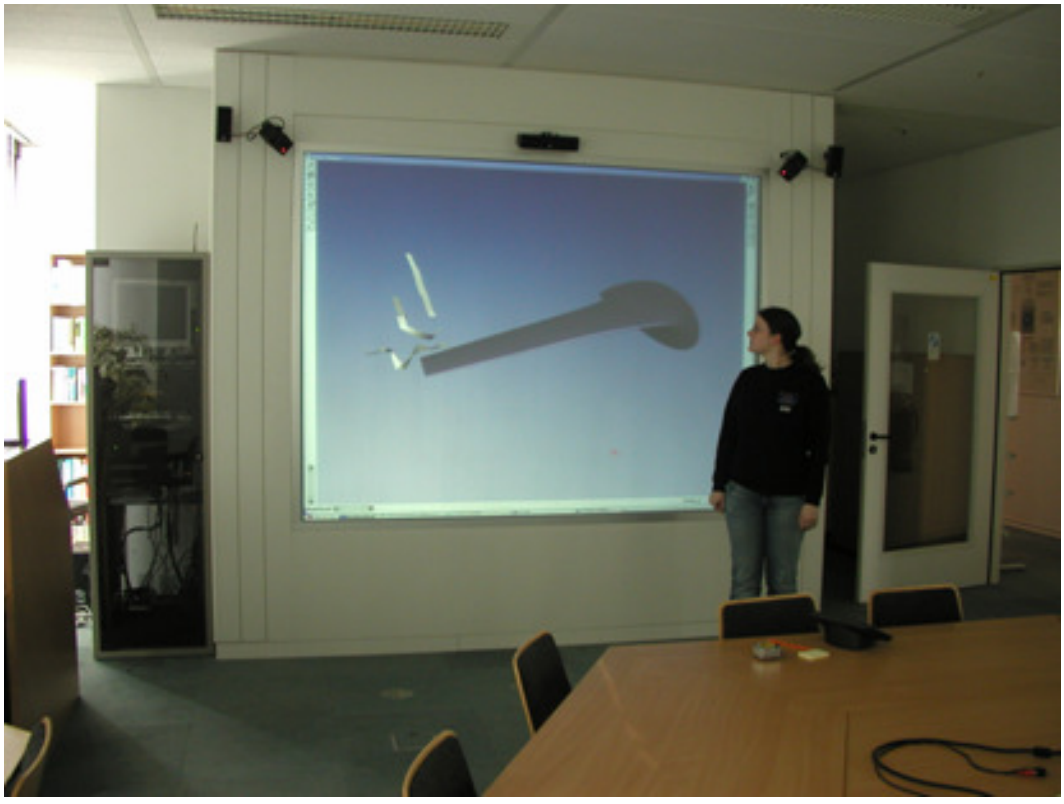

Fig. 1 The stereoscopic rear projection of the ARMINIUS cluster

The ARMINIUS cluster with 208 nodes has a peak performance of 2.6 TFlop/s. This compute performance needs about 70 kWatt electrical power which leads to a nearly equal amount of thermal energy. Our computing center is not able to get that much energy out of the room with the installed air conditioning system. Therefore a special fluid based cooling system was used inside the system. All processors of the compute nodes have special heat sinks which are connected via a heat exchanger to the cooling system of the building. This technique is able to move 50 to 60 percent of the thermal energy directly out of the building; the rest is cooled with the air conditioner. Special cabinets and housings for the nodes are used to adopt the processor cooling technique. We are using housings for the compute nodes which allow maximum flexibility in upgrading the system with additional I/O cards, and/or acceleration cards with GPUs, FPGAs, or cell processors.

We provide all standard system software for cluster systems. A Linux operating system with its software development tools is installed. Additionally, some MPI message passing libaries, thereunder three MPI versions optimized for InfiniBand are available. Also a software layer for fast IP over InfiniBand can be used for standard networking software packages. Scientific libraries for numerical applications are available and the compiler suite optimized for the Intel Xeon processor can be used. The software environment of the ARMINIUS cluster is shown in the following table:

Software	***Description***
RedHat Advanced Server Release 4	Linux operating system 2.6.9 kernel
GNU Tools	e.g. gcc
Intel compiler	C/C++, Fortran
Scali-MPI-Connect	MPI 1 compliant, fail-over from IB to GbE
MPICH-vmi	MPICH 1.2.6 for VMI 2.1 from NCSA
MvAPICH	MPICH 1.2.6 on VAPI from Ohio State University
Intel MKL	Math Kernel Library

Table 2: System Software of the ARMINIUS Cluster

Company	*Components*
Fujitsu-Siemens Computers GmbH	general contractor cluster system
ICT AG	housings, system integration
SilverStorm Technologies	InfinIO 9200 switch fabric 216 ports (max 288 ports)
SilverStorm Technologies Mellanox Inc.	InfiniBand Host Channel adaptor PCI-E IB 4x
nVidia GmbH	graphics cards
Rittal AG	heat exchanger,racks, controlling and management of cooling system
Atotech GmbH	fluid based heat sinks
Intel GmbH	INTEL Xeon EM64T processors HPC software tools
Scali Inc.	MPI Connect
UNITY AG	general contractor visualization equipment
3-Dims GmbH	integration of visualization equipment

Table 3: Companies involved in the development of the ARMINIUS Cluster system.

The official opening with a ceremonial inauguration was at June, 21st 2005. The system has an excellent performance with user applications. It is able to sustain 1.978 TFlop/s out of 2.6 TFlop/s peak performance. Based on the Linpack benchmark for supercomputers, the ARMINIUS cluster is one of the 500 most powerful systems of the world (Rank 205 in the 25th Top-500 list and rank 13th of the german supercomputers). The ARMINIUS system is embedded in two worldwide used Grid computing environments: Globus and UNICORE.

Utilzation

Fig. 2 depicts the utilization of the ARMINIUS Cluster in 2005 (24 hours per day). The average load was 62,72%. Unfortunately, we had a lot of outages in the first 4 months. Table 4 depicts these dates (marked with numbers in Figure 2) were the system was not fully accessible.

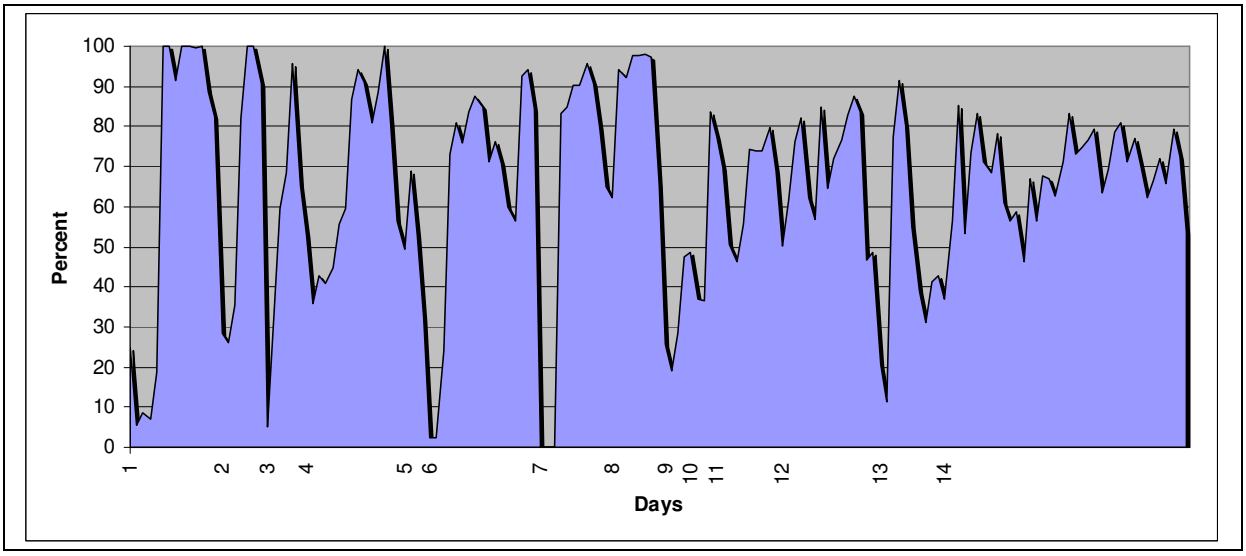

Figure 2: Utilization of theARMINIUS cluster in 2005 (24 hours per day).

Event	Date	Reason
1	22.7	*Maintenance*
2	5.8	*Maintenance*
3	12.8	*Maintenance*
4	18.8	*Maintenance*
5	2.9	*Maintenance*
6	6.9. – 8.9.	*Maintenance*
7	23.9. – 26.9.	*File server crash*
8	4.10.	*File server crash*
9	12.10.	*Failure in the cluster cooling system*
10	16.10. - 17.10.	*Failure in the cluster cooling system*
11	20.10. – 21.10.	*File server crash*
12	30.10 – 31.10	*File server crash*
13	14.11. – 15.11.	*Maintenance*
14	24.11.	*Maintenance*

Table 3 Dates with restricted system access

Paderborn SCI Cluster-2 (PSC2)

Installed	1999
Vendor	Fujitsu-Siemens
Number of nodes / CPUs	96 / 192
Node type	Primergy Server (2x Pentium III, 850 MHz)
Node memory	512 MByte
System memory	48 GByte
Node peak performance	850 MFlop/s
System peak performance	163 GFlop/s
High speed network type	SCI (Scalable Coherent Interface)
High speed network topology	12x8 torus
SCI performance (MPI)	Bandwidth: 84 MByte/s, Latency: 5 µs
Operating system	Linux
Message Passing SW	PVM, MPICH, ScaMPI
Compiler	GNU, Intel, PGI, Lahey-Fujitsu
Debugger	Totalview
Performance Analyzer	Vampir

The PSC2 is a workstation cluster comprising fully functional standard PCs. These nodes are connected by a fast communication network (SCI).

The PSC2 system has been developed in collaboration with the vendor companies (Fujitsu-Siemens and Scali) and has been brought to productivity.
The PSC2 system is embedded in two worldwide used Grid computing environments: Globus and UNICORE.

Utilization
The PSC2 system ran stable with only two major hardware failures. We had to replace one hard disk and one SCI card. The whole system was not available, from January 18th to January 19th (due to a maintenance of the air-condition) and on August 1st (due to a system maintenance). Figure 1 depicts the utilization of the PSC2 in 2005 (24 hours per day). The average load was 22.52%.

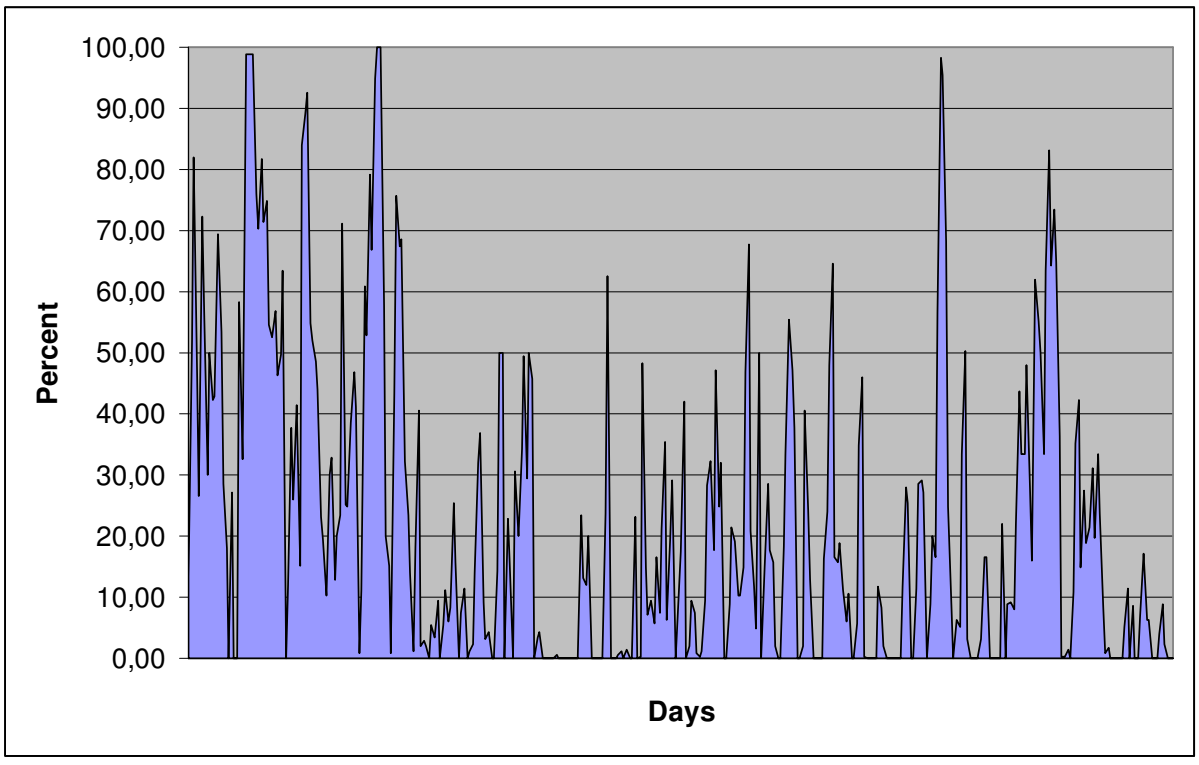

Figure 1 Utilization of the PSC2 cluster in 2005 (24 hours per day).

3.1.2 Dedicated Systems

Itanium2 / InfiniBand cluster (SFB)

Installed	2002
Vendor	Hewlett-Packard
Number of nodes / CPUs	4 / 8
Node type	HP ZX6000 (2x Intel Itanium-2, 1GHz)
Node memory	12 MByte
System memory	48 GByte
Node peak performance	8 GFlop/s
System peak performance	32 GFlop/s
High speed network type	Infiniband and Myrinet
High speed network topology	Switched
Infiniband performance (MPI)	Bandwidth: 446 MByte/s, Latency: 7.79 µs
Myrinet performance (MPI)	Bandwidth: 270 MByte/s, Latency: 11.81 µs
Operating system	Linux
Compiler	GNU, Intel
Message Passing SW	MPICH, ScaMPI

The SFB system is managed by CCS and is dedicated to the members of the “Sonderforschungsbereich 376 Massive Parallelität, Algorithmen, Entwurfsmethoden, Anwendungen”.

Paderborner Linux Cluster Next Generation (PLING)

Installed	2003
Vendor	Hewlett-Packard
Number of nodes / CPUs	32 / 64
Node type	HP RX-2600 (2x Itanium-2 1.3 GHz)
Node memory	4 GByte
System memory	128 GByte
Node peak performance	10.4 GFlop/s
System peak performance	332 GFlop/s
High speed network type	Infiniband
High speed network topology	Switched
Infiniband performance (MPI)	Bandwidth: 751 MByte/s, Latency: 6.51µs
Operating system	Linux
Message Passing SW	MPICH, ScaMPI
Compiler	GNU, Intel

The system is owned by and dedicated to the working group of Prof. Dr. Frauenheim with 25% for public usage.

FPGA Test Cluster

Installed	2003
Vendor	MEGWARE
Number of nodes / CPUs	4 / 8
Node type	2x Intel Xeon, 2.8 GHz
Node Memory	2 GByte
System memory	8 GByte
Node peak performance	5.6 GFlop/s
System peak performance	22 GFlop/s
High speed network type	Myrinet
High speed network topology	Switched
Operating system	Linux

The system is owned by the SFB 376 “Massive Parallelität, Algorithmen, Entwurfsmethoden, Anwendungen” and is dedicated to the working group of Prof. Dr. Monien.

KAO Cluster

Installed	2003
Vendor	Dell
Number of nodes / CPUs	4 / 8
Node type	2x Intel Xeon, 2.4 GHz
Node Memory	1 GByte
System memory	4 GByte
Node peak performance	4.8 GFlop/s
System peak performance	19.2 GFlop/s
High speed network type	Myrinet
High speed network topology	Switched
Disks	365 GByte SCSI-RAID
Operating system	Linux

The system is owned by and dedicated to the working group of Prof. Dr. Kao.

4-way Opteron Cluster

Installed	2004
Vendor	AMD, Fujitsu-Siemens
Number of nodes / CPUs	2 / 8
Node type	4x Opteron, 2.2 GHz
Node Memory	32 GByte
System memory	64 GByte
Node peak performance	4.6 GFlop/s
System peak performance	18.4 GFlop/s
Operating system	Linux

The system is owned by the SFB 376 “Massive Parallelität, Algorithmen, Entwurfsmethoden, Anwendungen” and is dedicated to the working group of Prof. Dr. Monien.

Parallel Virtual File System Cluster

Installed	2001
Vendor	ARBOR Datensysteme GmbH
Number of nodes / CPUs	13
Node type	AMD Athlon XP1600+, 1.4 GHz
Node Memory	1 GByte
System memory	13 GByte
Node disk capacity	65 GByte
System disk capacity	760 GByte
Operating system	Linux

The PVFS[1] is a virtual parallel file system running on clusters of Linux PCs. Virtual in the meaning that the file data is striped over the local file systems of the PCs acting as I/O nodes. Parallel means that data is accessed using multiple streams from the clients to the I/O nodes. We provided a capacity of 380 GByte on 6 I/O-nodes. The PVFS cluster was accessible from the PSC2 cluster, the PLING cluster, the SFB cluster, and their front-ends.

[1] Parallel Virtual File System

3.1.3 System Access

The access to the systems at the PC² is free of charge for all users coming from the academic world e.g. universities or schools. Users from commercial sites are also welcome but may have to pay a fee for using the systems. Please contact the PC² adminstration phone number +49 5251 606303.

Access to systems dedicated to specific user groups (e.g. the PLING cluster) may be denied depending on the requirements of the owner.

To apply for an account for the PC² systems one has to fill in small application forms available on our web server. Refer to http://www.upb.de/pc2/services/access. After handling the application all necessary information will be send via email within one or two days.

The registration information is kept private and will not be disclosed to third parties. It helps us to survey who is using our parallel systems.
To register for a new project one has to provide:

- A project description.
- The name and office address of the research manager and the project leader.
- The name and office address of each project member using the system.
- Additionally needed requirements like disk space or special software packages.

Specialist counseling is available for the following fields:

- Compiler
- Debugging
- Grid Computing
- MPI
- Numerical Applications
- Optimization
- Performance Profiling
- System Access and CCS
- System-Benchmarking and -Evaluation

For detailed information about how to use our systems please also refer to the following URLs.

ARMINIUS-cluster	http://www.upb.de/pc2/services/systems/arminius/
PLING-cluster	http://www.upb.de/pc2/services/systems/pling/
SFB-cluster	http://www.upb.de/pc2/services/systems/ic/
PSC2-cluster	http://www.upb.de/pc2/services/systems/psc/

Please report your problems to:
pc2-gurus@upb.de or call our service number +49 5251 606303.

Available Software

Software	Purpose	License	Available on
Abaqus	Finite element analysis	Dedicated	ARMINIUS
Amira	Advanced Visualization, Data Analysis and Geometry Reconstruction	PC²	ARMINIUS
Ansys	3D FEM solvers	Dedicated	ARMINIUS
ATLAS	Automatically Tuned Linear Algebra Software	None	ARMINIUS, SFB
Fluent	Computational fluid dynamics	Dedicated	ARMINIUS, SFB
FFTW	Library for computing the discrete Fourier transform (DFT)	None	ARMINIUS, SFB
Goto Lib	High-performance BLAS Implementation of Kazushige Goto	None	ARMINIUS, SFB
Gromacs	Molecular dynamics	None	ARMINIUS, SFB
MKL	Intel Math Kernel Library	PC²	ARMINIUS, SFB
MOE	Molecular operating environment	Dedicated	ARMINIUS
Matlab	Technical computing	Campus	ARMINIUS
MPICH	MPICH 1.2.6 for Ethernet	None	ARMINIUS, PSC2, SFB
MPICH-vmi	MPICH 1.2.6 for VMI 2.1 from NCSA	None	ARMINIUS, SFB
MvAPICH	MPICH 1.2.6 on VAPI from Ohio State University	None	ARMINIUS
MPIJAVA	MPI for Java	None	ARMINIUS
OpenFoam	Finite element analysis	None	ARMINIUS
Padfem2	Finite element analysis	None	ARMINIUS
PUB	Bulk-Synchronous-Parallel-Model Library	None	ARMINIUS, PSC2
ScaMPI	MPI 1 compliant, fail-over from IB to GbE	PC²	ARMINIUS, PSC2, SFB
StarCD	Finite element analysis	Dedicated	ARMINIUS
Totalview	Parallel Debugger	PC²	PSC2
Vampir	Performance Analysis	PC²	PSC2
VMD	Molecular Visualization	Dedicated	ARMINIUS
VTK	Visualization ToolKit	None	ARMINIUS
xmgrace	Two-dimensional plots of numerical data	None	ARMINIUS

3.2 Teaching

3.2.1 Thesis and Lectures in PC²

Lectures

- Performance-optimale Programmierung (Dr. Jens Simon)
- Webbasierte Informationssysteme (Prof. Dr. Stefan Böttcher, Prof. Dr. Odej Kao)
- Architektur paralleler Rechnersysteme (Dr. Jens Simon)
- Einführung in verteilte Systeme (Prof. Dr. Odej Kao)
- Systemaspekte verteilter Systeme (Prof. Dr. Odej Kao)

Project-Groups

- Peer-2-Peer based Search for Web Services (Prof. Dr. Odej Kao, Felix Heine, Matthias Hovestadt)

Bachelor Thesis

- Bahr, Andreas: Realisierung virtueller Organisationen mit Hilfe des „Virtual Organisation Membership Service“ (VOMS)
- Beisel, Tobias: Analyse und Auswertung von Daten aus UMTS Netzwerken
- Böttcher, Henning: Definition und Verarbeitung von topologischen Informationen innerhalb eines Location Servers
- Dyck, Eugen: Ortsbezogene Druckdienste für PDAs
- Felix, Peter: Eine Public-Key-Infastruktur zur sicheren Chipkartengestützen Authentisierung auf Basis des Microsoft Active Directory
- Freitag, Thomas: Vergleich von Verfahren zur Auffindung von Diensten und Ressourcen
- Fromme, Andre: Ressourcenmanagement mit optimierten Service-Levels
- Gretencord, Alexander: Entwicklung eines Frameworks für ortsbezogene Dienste mittels CORBA Trading Service
- Haertel, Christine: Visualisierung von WS-Resource Properties und WS-Service Groups der WSRF-Spezifikation

- Koch, Philipp: Entwurf eines Replikationskonzeptes mit SQL Server
- Lerch, Nicolas: Server Monitoring und Leistungmessung mit Nagios
- Mense, Holger: Einführung des Network Itrusion Detection Systems „Snort“
- Müller, Stephan: Konzeption und Entwicklung eines alternativen WLAN Zugangsverfahren für die Universität Paderborn
- Paul, Oxana: Untersuchung der Effizienz von relationalen Datenbanken zur Speicherung von XML-basierten Service Level Agreements
- Riemann, Tim: MPI-Kommunikation für qualitätsbasierende verteilte Algorithmen
- Rikanovic, Igor: Klassifizierung von Positionierungstechniken für Location-based Services
- Schumacher, Tobias: Entwurf und Implemtierung einer grafischen Benutzungsschnittstelle für das Resource Management System CCS
- Twiste, Nicola: Konzeption und Implementierung eines verteilten Agentensystems zur Synchronisation von Verzeichnisdiensten am Beispiel Microsoft Active Directory

Master Thesis

- Birkenheuer, Georg: A Framework for Semantic Web Service Development
- Weking, Michael: Erfassung und Visualisierung der Topologie von industriellen Netzwerken
- Finke, Stefan: Einigungsprozesse mit Nachweispflicht auf XML-Basis – Ein Konzept zur zertifikatsbasierten Definition, Aushandlung und Archivierung von Vertragsdokumenten
- Wan, Kan-Sing: Ontology-Based Resource Matching
- Gerdemann, Sebastian: Analyse der Backboneauslastung und Bandbreitenplanung
- Schumacher, Tobias: Untersuchung von Kommunikationsmethoden zwischen FPGA-basierten Systems-on-Chip auf Basis von Message-Passing
- Thygs, Manuela: Swapping Mechanismen für Smart Cards
- Thiele, Kai: Konzept und Entwicklung von kontextbasierten Diensten in Wireless-LAN Umgebungen

PhDs Thesis

Kerstin Wielage:

- Analysis of Non-Newtonian and Two-Phase Flows

PC² Colloquium

- Tobias Schumacher (PC²): Untersuchung von Kommunikationsmethoden zwischen FPGA-basierten Systems-on-Chip auf Basis von Message-Passing
- Stefan Lietsch (PC²): Grundlagen und Einsatzgebiete eines Hochleistungs-Visualisierungs-Clusters
- Andreas Bahr (Studienarbeit am PC²): Realisierung virtueller Organisationen mit Hilfe des "Virtual Organisation Membership Service" (VOMS)
- Dr. Juan José Porta (IBM): Broadband Processor Architecture: power-efficient and cost-effective high-performance processing for a wide range of applications
- Paderborn Center for Parallel Computing: Tag der offenen Tür
- Teilnehmer der Projektgruppe: Abschlussvortrag Projektgruppe PeerThing
- Christine Härtl (Studienarbeit am PC²): Entwicklung eines Clients zum WS-Agreement als Anwendung der WSRF-Spezifikation
- Axel Keller (PC²): PC2-Workshop: CCS: Tipps und Tricks
- Holger Mense (Universität Paderborn): Network Intrusion Detection am Beispiel von Snort (Studienarbeit)
- Nicolas Lerch: Server Monitoring und Leistungsmessung mit Naigos
- Axel Keller, Dr. Jens Simon (PC²): Vorstellung des neuen PC² Cluster-Systems
- Prof. Dr. Maria Specovius (Universität Kassel): Wie man Randwertprobleme auf PC-Grösse zuschneidet
- Robert Breitrück (PC²): Exakte Stromlinien auf unstrukturierten Tetraedernetzen
- Jonas Klevhag, Stefan Möhl (Mitrionics AB): FPGA-Programming
- Steeve Ndong (Diplomarbeit am PC2): Entwicklung einer erweiterbaren Architektur für das Tool ccsMon

3.2.2 PhD at PC²

Dr. Kerstin Wielage
Analysis of Non-Newtonian and Two-Phase Flows

Abstract:

Fluids can be classified according to various criteria. One opportunity is to distinguish Newtonian and non-Newtonian fluids. The difference between both classes of fluids can be observed in a variety of situations. For example, consider two bowls containing two different types of fluids, e.g., water and a polymer solution representing a Newtonian and a non-Newtonian fluid, respectively. Inserting a rotating rod in each bowl, we observe that in case of the Newtonian fluid, a characteristic dip arises near the rotating rod due to centrifugal forces pushing the fluid outwards. In contrast, some non-Newtonian fluids climb up along the rod. This effect is known as the Weissenberg effect.

Particularly, in industrial applications non-Newtonian fluids are of great interest, since many used fluids, e.g., lacquers or polymer solutions show non-Newtonian effects.

In general, real-world experiments of industrial processes can be complex, cost-intensive, and time-consuming. Therefore, numerical simulations are important and can be useful in order to optimize the process with respect to cost or quality aspects. Thus, the investigation of mathematical models governing the flow of non-Newtonian fluids is important regarding both analysis and numerical simulations. In the context of this thesis, the term analysis in the title refers to both the meaning in the mathematical sense and the more informal meaning of “computational analysis”, i.e., the numerical investigation of the behavior of non-Newtonian fluids.

Non-Newtonian fluids are characterized by different features, such as viscosity, elasticity, or memory effects. An important feature of polymeric liquids is the fact that their viscosity changes with the shear rate, so-called generalized Newtonian fluids on which this thesis is focused on. The mathematical model describing the motion of these fluids in the whole space is given by the system

$$\begin{aligned} \frac{\partial}{\partial t}\mathbf{u} + \mathbf{u}\cdot\nabla\mathbf{u} &= \operatorname{div}\mathbf{S} - \nabla p + \mathbf{f} && \text{in } [0,T]\times\mathbf{R}^n \\ \operatorname{div}\mathbf{u} &= 0, && \text{in } [0,T]\times\mathbf{R}^n \\ \mathbf{u}_{|t=0} &= \mathbf{u}_0, && \text{in } \mathbf{R}^n. \end{aligned} \qquad (1)$$

Here, the stress tensor $\mathbf{S}$ is given by $\mathbf{S} = 2\mu\left(\|\mathbf{D}\|^2\right)\mathbf{D}$ with the rate-of-deformation tensor $\mathbf{D} = \frac{1}{2}\left[\nabla\mathbf{u} + (\nabla\mathbf{u})^T\right]$, and the viscosity function μ depending on $\|\mathbf{D}\|^2$ respectively on the shear rate $\dot{\gamma} = \sqrt{2}\|\mathbf{D}\|$, where $\|\mathbf{D}\|$ denotes the Hilbert-Schmidt norm.
The main result of this thesis is the proof of existence of problem (1) in the maximal L_p-regularity class. By means of maximal L_p-regularity, local (in time) strong well-posedness of this model is obtained under certain restrictions concerning the viscosity function. For example, for the viscosity function

$$\mu\left(\|\mathbf{D}\|^2\right) = \mu_0\left(1 + \|\mathbf{D}\|^2\right)^{\frac{m-2}{2}}$$

with $\mu_0>0$ which is often used in the mathematical literature, we obtain local existence in $\mathbf{R}^n, (n>1)$ for $m > \frac{3n-4}{2(n-1)}$, i.e., in the 3D case $m > 5/4$ is sufficient. We emphasize that according to the engineering literature the range of interest concerning m is $m>1$.

In the numerical part, the emphasis is on two-phase flows, since many interesting problems appear in this context. Moreover, we compare experiment and simulation of a binary droplet collision using non-Newtonian fluids, the behavior of which is assumed to be that of generalized Newtonian fluids. Furthermore, assuming the situation of system (1), we show the admissibility of the viscosity function used in the simulation.

3.2.3 Software System "PIRANHA" – "Paderborn Idle Resource Allocation Harness"

Project coordinator	Prof. Dr. Odej Kao, University of Paderborn
Project members	Felix Heine, PC², University of Paderborn
	Matthias Hovestadt, PC², University of Paderborn

General Problem Description

Already in 1965, Gordon Moore predicted that the number of transistors on a processor will double every year [1]. The pace slowed down a bit, reaching the doubling of transistors on a processor every 18 months, but at the bottom line Moore's law is valid until the present day. Since an increased number of transistors comes along with an improved performance of a single processor, system performance should not be an issue. Far from it! In fact, compute power still is a very precious good. With every new generation of processors, computers became a valuable instrument for solving new classes of problems. Nowadays the computation of simulations is widely adopted in research, providing results at an inconceivable level of precision, resulting from the solution of millions of differential equations.

In this light it seems to be paradoxical that more and more compute power worldwide is lying idle and passes unused. The reason for this paradox is the fact that even a standard low-cost PC available in common computer stores offers vast amounts of compute power – particularly compared to high-end compute resources a few years ago. In computer pools and offices within the University of Paderborn, hundreds of computers are installed, waiting to be used by researchers, staff, or students. Hence, these computers normally idle at night or at weekends and holidays. Even if users are logged in and start their work, programs like web browsers, word processors, or development environments only use a fraction of the capacity of modern processors. At the bottom line, the overall amount of wasted computer power is remarkable. It would be greatly appreciated by many research projects if this power could be used for their calculations.

In 2003 a project group named "The do-it-yourself upb.de supercomputer" started at PC². The general goal of this project group was to unleash the idle compute power at the University of Paderborn. For this, a system had to be created which could be used by users and administrators to provide and consume compute power. Non-intrusiveness has been a central aspect in system design, since resource owners will not open their machines for other users if this has negative impact for themselves. Furthermore the system has to hide complexity from the users, such that the virtual supercomputer can be used by normal users, which do not have in-depth knowledge about the system architecture and infrastructure.

Following to an introductory seminar phase, the project group started on evaluating existing tools, creation of a requirement analysis and a general system architecture. The group decided to use “Sun Grid Engine” (SGE) [3] as core technology, since this system complied to most of the basic demands of the project group. During the main phase of the project the core system for harnessing idle compute time has been developed, which builds on top of the SGE for reaching the project group goals. The system has been named “PIRANHA” [2]. Beside basic system tools and server components, it also ships with an intuitive graphical user interface. In a final deployment phase on systems from PC², the project group was able to proof the functionality of the PIRANHA system.
Since PIRANHA turned out to be a valuable tool for harnessing compute power already during the project group phase, PC² is continuing this system even after the ending of the project group. PC² is supporting other working groups in installing the system on their computers. Furthermore PC² provides user and administration support and maintains further development and fulfillment of user requests.

Problem Details and Work Done in the Reporting Period
After the end of the project group in October 2004, the PIRANHA system first had to be completed as shippable software. Beside the preparation of source packages and installable binary packages in RPM format for RedHat Linux systems and in DEB format for Debian Linux systems, this explicitly also included the finalization of software documentation. Before releasing the software to external users, the functionality of RPMs and DEBs has been verified on systems of PC².
In a first phase, the PIRANHA system has been installed on student computer pools of the computer science and mathematics department of the University of Paderborn. Having PIRANHA installed, the operators of these computer pools were able to provide the noticeable compute power to interested users. In fact, the PIRANHA infrastructure has been used for simulations from local users.
As a matter of fact, the application of PIRANHA in real-world environments resulted in numerous new requirements coming from administrators and users. If possible, these requirements have been directly implemented in the PIRANHA software. This work also affected the graphical user interface, making it more intuitive for the user. However, some new requirements were challenging, since they exceeded the requirement profile, which served as the blueprint during system design phase in the project group. Hence, PC² started with a re-design of the PIRANHA system, making the system more flexible and easier to adopt.
An important aspect on re-design affects the SGE, which currently is the core technology behind the PIRANHA system. Despite the fact that SGE is open source and offers large functionality, some interested administrators are fixed to other systems, e.g. the CONDOR system [4]. For also attracting this group of potential

users, the re-design of PIRANHA focuses on abstraction from the underlying resource management system.

Resource Usage at PC²

For system development and testing a realistic installation environment is indispensable. Since the development of PIRANHA is pushed by PC², new beta-versions of the PIRANHA software have been tested on resources within the PC². These resources are partially virtual machines running under “UML” (User Mode Linux) [5], but also real workstation machines running in the offices and in the student pool of PC².

References

[1] G.E. Moore, Cramming More Components Onto Integrated Circuits, Electronics, 38(8), 1965
[2] PIRANHA: http://www.upb.de/pc2/projects/piranha
[3] Sun Microsystems: Sun Grid Engine, http://wwws.sun.com/software/gridware
[4] University of Wisconsin Madison: Condor, http://www.cs.wisc.edu/condor
[5] User Mode Linux, http://user-mode-linux.sorceforge.net

3.2.4 Project Group: PeerThing "Peer to Peer based Search for Web Services"

Project coordinator	Prof. Dr. Odej Kao, PC², University of Paderborn
Project members	Felix Heine, PC², University of Paderborn
	Matthias Hovestadt, PC², University of Paderborn
	Kerstin Voß, PC², University of Paderborn

General Problem Description

Modern Grid environments [1] promise to release the user from the task of finding suitable resources for his jobs. Instead of using hard-coded links to each system, users describe their individual resource requirements. The Grid middleware is in charge of finding matching resources. This is what is called *resource virtualization*.

As Grids and Web Services are on the verge of merging due to the Web Services Resource Framework (WSRF) initiative [2], the problem of finding Grid Resources is a sub-problem of finding Web Services. The *matchmaker* is responsible for finding suitable resources. Currently used matchmakers like Globus MDS [3] or the Condor Class Ads System [4] have some limitations which make them unsuitable for future large scale Grid systems, like the next generation grid as described in the NGG report from the European Commission [5].

The project group "Peer to Peer based Search for Web Services" [6] aims to explore new approaches to this problem by implementing a search engine which is based on a Peer to Peer (P2P) network. While the primary goal is to develop an engine which is suitable for finding compute resources within the university, the system has to be extensible in two ways:

- First, the formalism used to describe the entities must be semantically rich and flexible in order to allow specification of any kind of resources. It should be possible to specify general information about resources, like compatibility relationships. This information should be used during the search process to improve the quality of the results.
- Second, the search engine must be scalable. It has to be able to integrate resources from numerous different providers without noticeable degradation of performance.

The system has to integrate two different kinds of information about a resource: dynamic and static. Static information is concerned with attributes like processor type, overall memory size, or installed operating system. These attributes change infrequent, typically in the range of months or years. A change in such an attribute is due to a hardware or software update or reconfiguration. In contrast, dynamic

attributes are attributes like current processor load which change very frequently. Thus different ways of processing and refreshing these attributes are needed.

Problem Details and Work Done in the Reporting Period

The project group already started in October 2004 with a seminar as reported in the previous annual report. The goal of the seminar was to learn the foundations for the following practical work. During 2005, the members of the project group designed and implemented a P2P based system for resource discovery using the RACER description logic reasoner [7].

The overall system design is shown in Figure 1. The *poolclients* are the workstations which offer their services to the Grid. They are grouped into *pools.* For each pool, a so-called *poolhead* is responsible to represent the pool within the network. The poolheads are connected via a Gnutella-style p2p network. To find suitable resources, static and dynamic data has to be compared. The poolhead collects the *static* resource information from the poolclients and uses the RACER system to store and query this information. The dynamic information is retrieved on the fly from the *poolclients* during the query evaluation. The *userclient* is a graphical user interface used to formulate queries and to send them to the p2p network.

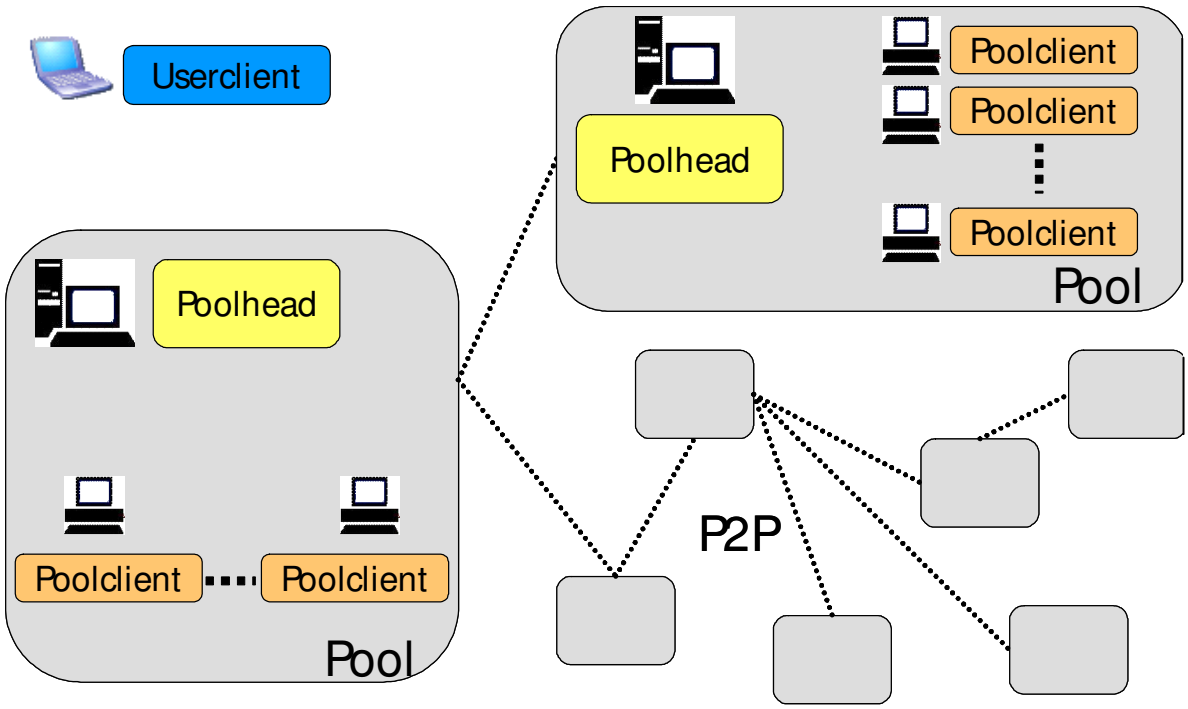

Figure 1: System design

When a user wants to query resources in the system, she starts the userclient (see Figure **2**) on her local machine. The userclient connects to a known poolhead and retrieves an *ontology* from the poolhead which includes a schema-level description of the resources. This ontology is used to drive the user interface. The user builds up a tree of conditions which specifies the features of the desired resources. In each step, the user interface uses the ontology to restrict the further input to useful values. Thus even complicated queries can be formulated easily. Further, the user can specify restrictions on the current values of dynamic attributes.

The integrated use of ontologies is a distinguishing feature of PeerThing. Although there are a number of systems using ontologies in resource discovery, the novel combination of these techniques with a p2p network allows achieving both semantically rich resource discovery and scalability.

When the query is sent to the network, the receiving poolhead both forwards the query to other known poolheads in the network, and evaluates the query locally. As the p2p network continuously optimizes its topology in order to favor local connections, the resources located close to the user are retrieved earlier in the query process. Query results are sent back to the original poolhead directly, bypassing the routing steps during query sending. This locality oriented topology is an important enhancement compared to the original Gnutella implementation, because the user is typically interested in matching resources with good network connections to her local site. Especially for jobs with large input or output data sets, the transfer time can be a critical factor. Thus the PeerThing system has been designed to favor these resources.

The query evaluation within the poolhead starts by evaluating the static parts of the query using the RACER description logic reasoner. The reasoner does not only detect direct matches of the attributes, it is also capable to integrate background knowledge. This means, that e.g. information about compatibility of resources is integrated during the retrieval of matching answers. After the retrieval of matching resources, also the dynamic attributes have to be checked. Therefore, the poolclients are directly contacted to get the latest values of the dynamic attributes. These values are cached for a short time on the poolhead.

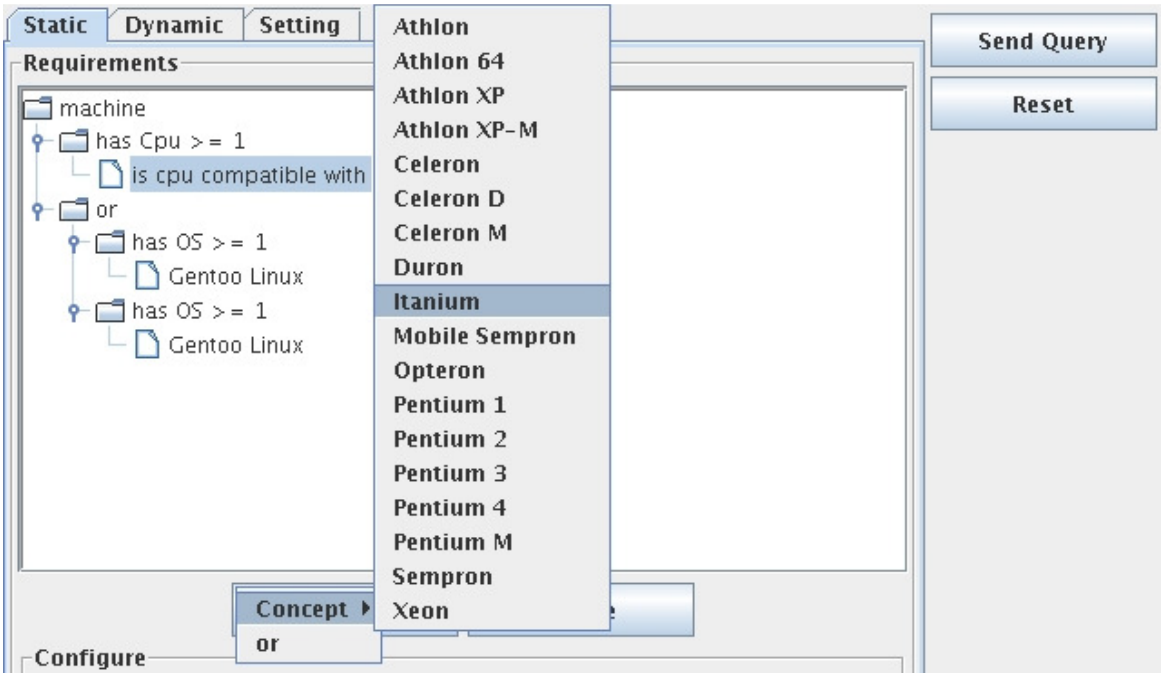

Figure 2: Userclient

The poolclients have integrated various modules using a plugin concept to query different types of attributes. Within the developed prototype, modules to retrieve both static and dynamic resource information from Windows and Linux based workstations have been developed. Due to the flexible plugin concept, various other types of resource information could be integrated.
To achieve fault-tolerance, some of the poolclients within a pool are configured to serve as backup poolheads. They permanently receive current configuration of the p2p network at the working poolhead, so that they can immediately take over in case the poolhead crashed. The poolclients use a distributed election algorithm to determine which of them should take over when a poolhead has failed.

Resource Usage at PC²

The project group used various systems at PC². First, they used development infrastructure including a revision control system, a bug tracking database, and others. Additionally, the developed system was tested both on PC² workstations and other workstations on the university campus.

References

[1] Foster and C. Kesselman (Eds.). The Grid 2: Blueprint for a New Computing Infrastructure. Morgan Kaufmann Publishers Inc. San Francisco, 2004

[2] K. Czajkowski, D. Ferguson, I. Foster, J. Frey, S. Graham, T. Maguire, D. Snelling, and S. Tuecke. From Open Grid Services Infrastructure to WS-Resource Framework: Refactoring & Evolution, 2004

[3] Karl Czajkowski, Steven Fitzgerald, Ian Foster, and Carl Kesselman, *Grid Information Services for Distributed Resource Sharing*, Proceedings of the Tenth IEEE International Symposium on High-Performance Distributed Computing (HPDC-10), 2001

[4] Rajesh Raman, Miron Livny, and Marvin H. Solomon, *Matchmaking: Distributed Resource Management for High Throughput Computing*, HPDC, 1998

[5] H. Bal et al. Next Generation Grids 2: Requirements and Options for European Grids Research 2005-2010 and Beyond, Expert Group Report, 2004

[6] Homepage of the Project Group: http://www.upb.de/StaffWeb/maho/pg2004

[7] Volker Haarslev and Ralf Möller, *RACER User's Guide and Reference Manual Version 1.7.7*, http://www.sts.tu-harburg.de/~r.f.moeller/racer/racer-manual-1-7-7.pdf, 2003

3.3. Collaborations

3.3.1 Ressourcenverbund – Nordrhein-Westfalen (RV-NRW)

Project coordinator:	Dr. Jens Simon, PC², Paderborn University

General Description

The *Ressourcenverbund – Nordrhein-Westfalen* (RV-NRW) is a network of university computer centers of the state North Rhine-Westphalia which provides a network of excellence and cooperative resource-usage of high performance compute systems [1]. Targets of this network are:

- Outsourcing of work besides the main focus of each computer center.
- Providing access to short and expensive resources.

Active member organisations of the RV-NRW are:

- RWTH Aachen
- University Köln
- University Paderborn
- University Münster
- University Siegen
- University Dortmund
- University Duisburg-Essen
- Ruhr-University Bochum
- Open University Hagen

In generally, all systems and services of the *Ressourcenverbund* are available for all scientists of RV-NRW members. The use of the resources is free of charge for this community.

Details and Work Done in the Reporting Period

The resources and services of the RV-NRW will be constantly increased. The list of services includes:

- Access to HPC systems and software
- Consulting HPC users
- Computational Grid
- Certificate Registration Authority
- Access to database
- Backup and archiving

HPC systems and application software

Several high-performance computer systems are available for the users of the RV-NRW. The *Rechen- und Kommunikationszentrum* of the RWTH Aachen provides a Sun Sparc processor based cluster system consisting of 24 SMP-Servers, each with up to 144 processors. The *Paderborn Center for Parallel Computing* of the University Paderborn provides a cluster system with 200 nodes, each node with two Intel Xeon processors. The *Zentrum für Informationsverarbeitung* of the University Münster operates a cluster system with 120 Intel Xeon single processor nodes. The *Zentrum für Angewandte Informatik* of the University Köln provides a 256 nodes cluster system with AMD Opteron dual processor nodes. Finally the *Rechenzentrum* of the Ruhr-University Bochum operates an out-dated 28 processor Hewlett-Packard Superdome SMP system.
Interested scientists apply for access to the RV-NRW compute resource at their local compute center.

Consulting HPC users

The RV-NRW provides a primary point of contact for users for all resources provided within the network. Expert advice will be provided by the appropriate compute center staff responsible for the requested resources. Additionally, courses and material concerning high performance computing are offered to increase the skills and qualifications of the users.

Computational Grid

RWTH Aachen provides a computational grid infrastructure for compute intensive jobs. This grid environment is based on the high throughput computing resource management software Condor and the bioinformatics application BLAST (Basic Local Alignment Search Tool). A web interface is available for easy and comfortable use.
Paderborn Center for Parallel Computing provides an application grid infrastructure for the FLUENT application, a computational fluid dynamics code typically used by chemists. The infrastructure is based on the UNICORE grid middleware [2]. A powerful graphical user interface for the user's local PC is available, which offers a single entry point to all computing resources available to the user.

Certificate Registration Authority

The Open University Hagen provides a Public Key Infrastructure (PKI) for an automatically issue of X.509v3 certificates. Members of the RV-NRW are free to use the dedicated certificate-server.

The University Paderborn, PC² is a registration authority for Grid certificates. In Germany two Certificate Authorities (CAs), Karlsruhe GridKa and Hamburg DFN, are established for grid services. The standard DFN certificates, used for the encryption of e-mails, can not be used for grid services.

Database

The University Dortmund operates a Beilstein Reaction and Material database.

Backup and Archiving

The *Rechen- und Kommunikationszentrum* of the RWTH Aachen provides a Tivoli Storage Management (TSM) archiving and backup system.

Resource Usage at PC²

PC² provides up to 30 percent of the compute resources of the *Arminius* cluster to users of universities of North Rhine-Westphalia.

Further information about the network of excellence is available on the web-pages of the Ressourcenverbund-NRW [1]

References

[1] Ressourcenverbund Nordrhein-Westfalen (in German), http://www.rv-nrw.de

[2] Unicore-based Application-Grid for Fluent Users, PC² Annual Report, 2004

4 Research

4.1 Research Areas at PC²

Since 1990 the PC² is operating massive parallel systems and clusters. The methods and experiences derived will be integrated in new projects. In addition, the PC² evaluates the latest technologies and develops them to implementation stage. The projects which content these activities are listed in the following table.

Selected Projects	*Contact*	*Email*
Administrated Systems at the PC²	Andreas Krawinkel	krawi@upb.de
PC² Technical Administration	Axel Keller, Holger Nitsche	kel@upb.de hn@upb.de

Research on Visualization Cluster

Nowadays, High Performance Computing (HPC) and graphic clusters are implemented successfully as individual solutions. The PC² develops an integrated cluster system which provides the services of HPC and visualization within a system for various applications. The integration of several services guarantees the manageability of resource requirements of complex applications in the future.

The PC² has long term experiences with cluster systems equipped with acceleration cards based on FPGAs (Field-Programmable Gate Arrays) and scalable applications exploiting the performance of standard CPUs und FPGAs. We have designed and installed cluster systems with multiple-CPUs and multiple-FPGAs tidily connected with high-speed interconnection networks.

The PC² offers possibilities for mutual research and project work in this field as well as the access to appropriate and innovative high performance computer systems. Research projects and their persons in charge are presented in the following table.

Selected Projects	*Contact*	*Email*
Development of a Compute and Visualization Cluster	Dr. Jens Simon	simon@upb.de
Advanced Cluster Computing with Hardware Acceleration	Dr. Jens Simon Dr. Ulf Lorenz	simon@upb.de flulo@upb.de

Finite Elements Simulation

One of the PC²'s core competencies is the simulation of physical and chemical processes, which are computed using mathematical methods like finite elements and finite volumes. A focal point is the development of efficient parallel algorithms and data structures for the simulation of highly dynamic processes in structural mechanics. Another focal point is the development of mathematical models for complex flow problems, particularly for material transport and multiple phase flow.
Long years of experience gathered from a multitude of research projects in the field of massive parallel applications have been integrated into the modular "padfem²" development and simulation environment.
The projects in the research of Finite Elements Simulations are:

Selected Projects	***Contact***	***Email***
Mixing in Micro-Reactors with Direct Numerical Simulation	Dr. Stephan Blazy	blazy@upb.de
padfem² - An Efficient, Comfortable Framework for Massivly Parallel FEM-Applications	Oliver Marquardt	marquardt@upb.de

PC² Benchmarking Center

The PC² benchmarking center is specialized in investigating the performance of high-speed networks and parallel computer systems. Typically, these are based on cluster technology. Functional parts or complete systems are evaluated with the help of so-called low-, system-, and application-level benchmarks. Derived from this evaluation new system architectures will be developed.
In addition, the PC² offers assistance in finding a high performance and cost efficient solution for parallel computers for already existing application programs and those which are under development.
The following table presents an overview of the projects in this research area.

Selected Projects	***Contact***	***Email***
System Evaluation, Benchmarking – and Operation of Experimental Cluster Systems	Dr. Jens Simon	simon@upb.de

Resource Management and Grid Computing

Like the World Wide Web the Grid will revolutionize the world of computers by offering on demand worldwide access to computing power.

The PC² has been working in this sector for years and has established itself as a competence center in the field of resource management. Current research focuses on the problem of how to guarantee the use of the resources with Service Level Agreements (SLA). This field is a vital assumption for future commercial use of grid environments.

The various projects focused on resource management and Grid Computing are listed in the following table.

Selected Projects	***Contact***	***Email***
Computing Center Software (CCS)	Axel Keller	kel@upb.de
Scheduling in HPC Reource Management Systems: Queuing vs. Planning	Axel Keller, Matthias Hovestadt	kel@upb.de maho@upb.de
The Virtual Resource Manager Architecture for SLA-aware Resource Management	Matthias Hovestadt	maho@upb.de
Risk Management and Resource Management	Kerstin Voß	kerstinv@upb.de
Semantic Grid Resource Discovery	Felix Heine	fh@upb.de

4.2 Parallel Architectures

4.2.1 System Evaluation, Benchmarking and Operation of Experimental Cluster System

Project coordinator	Dr. Jens Simon, PC², University of Paderborn
Project members	Axel Keller, PC², University of Paderborn
	Andreas Krawinkel, PC², University of Paderborn
	Holger Nitsche, PC², Universtiy of Paderborn
Work partly supported by	Fujitsu-Siemens Computers, Intel

General Problem Description

In the year 2005, PC² has done lots of evaluation of new computation and communication technologies for cluster systems. InfiniBand has become a widely used cluster interconnect and PCI-Express and 64-bit processors are becoming this year a standard for large scale systems.

Problem Details and Work Done in the Reporting Period

The PC² Benchmarking Center has done lots of system evaluation in this year. Most of the work was directly related to the acquisition of the new ARMINIUS cluster system. At first we present some results from the new PCI-Express InfiniBand communication adaptors. Then some performance figures of the ARMINIUS cluster are shown.

Co-operations

The PC² benchmarking center is also doing system evaluation and benchmarking for external companies and organizations. The PC² has a long term co-operation with Fujitsu-Siemens Computers where PC² acts as a competence center for high performance computing.

Company	Funding	Amount
Fujitsu-Siemens Computers	Cooperation “Competence Center PC²”	€30.000 per year
Fijitsu-Siemens Computers	Benchmarking of new hpcLine system	€30.000

Table 4: Financial support for our project activities.

We presented our results on the International Supercomputer Conferences (ISC2004 and ISC2005). Also workshops and user meetings have been visited.

Industrial co-operations enabled us to get early access to newest technology and some joint projects results in sponsoring of hardware and software for the evaluation of high performance systems (Table 5). Mellanox and InfiniCon provided their high-end InfiniBand components - PCI-e based host channel adaptors and first versions of large scale switch fabrics. The company iWill, a motherboard manufacturer from Taiwan, gave us some systems with onboard InfiniBand for lone. Intel provided us the very first white-boards of the new introduced 64-bit Xeon processor line. With these systems PC² was able to make early experiences with the new I/O standard PCI-e. Fujitsu-Siemens, Hewlett-Packard, and iWill provided multiprocessor workstations with different types of 64-bit processors – Intel Xeon, AMD Opteron, and Intel Itanium. High-end graphic cards based on PCI-e were provided by nVidia. Test equipment for processor cooling was provided by Rittal and Atotech.

Company	Components	Duration	Value
Intel	Two Lindenhurst Dual Xeon EM64T DDR2 PCIe-8x	since May 04	€12.000
Mellanox	PCI-Express InfiniBand HCAs 24 port InfiniBand switch	since May 04	€10.000

Table 5: Support of equipment from involved companies (without tax).

References

[1] Primeur weekly, The Paderborn hpcLine cluster: a marriage of Intel and AMD processors, http://www.hoise.com/primeur/05/articles/weekly/AE-PR-08-05-19.html

[2] Primeur weekly, How do you build a supercomputer? Together! http://www.hoise.com/primeur/05/articles/weekly/AE-PR-08-05-17.html

[3] Jens Simon, Low level InfiniBand performance, http://www.upb.de/StaffWeb/jens/Projekte/Benchmarks/Interconnects/infiniband.htm

[4] Jens Simon, MPI performance of interconnects, http://www.upb.de/StaffWeb/jens/Projekte/Benchmarks/Interconnects/mpi_pmb.htm

[5] Top500 List, http://www.top500.org/

4.2.2. Evaluation of Microsoft Windows Compute Cluster Server

Project coordinator	Dr. Jens Simon, PC², University of Paderborn
Project members	Holger Nitsche, University of Paderborn
	Michael Flachsel, University of Paderborn
Work supported by	Microsoft Cooperation, Silverstorm Technologies

General Problem Description

Microsoft Cooperation extends its Windows Server 2003 operating system with some additional tools to fulfill the demands of cluster computing environments [1]. The idea of Windows Compute Cluster Server 2003 is to bring the supercomputing power of high-performance computing to the personal and workgroup level. PC² evaluates the currant versions of the Windows CCS 2003 in a compute center environment.

Problem Details and Work Done in the Reporting Period

Windows CCS is made up of several tools layered on the standard Windows Server 2003. The clustering tools are using existing Microsoft technology like Remote Installation Services (RIS), Microsoft Management Console with a cluster management console plug-in, and Active Directory for network-wide authentication.

The Windows CCS architecture consists of several required and optional components. The optimal deployment can be configured depending on the applications being run on the cluster system. For deployment and management of the cluster a dedicated computer system is needed.

The head node provides user interfaces for administration as well as management services for all compute nodes. Management services include job scheduling and resource management. Automatic compute node deployment can be done by Remote Installation Services (RIS). Internet Connection Sharing (ICS) and Network Address Translation (NAT) can be used to connect private compute nodes with public networks. The head node can also provide the Dynamic Host Configuration Protocol (DHCP) and Domain Name System (DNS) services. For authorization and authentication each node of the cluster must be a member of a special Active Directory domain. This can be an intra or inter cluster AD domain.

The compute nodes provide the computational resources of the cluster. Heterogeneous configurations of compute nodes are supported as long as each node fulfills the hardware and software requirements.

The job scheduler is running on the head node and coordinates the execution of jobs on the compute nodes. The scheduler manages the job queues and all resource allocations. The execution of jobs is done by the local Node Manager Service.
The Message Passing Interface (MPI) is integrated in the Windows CCS to provide an environment for the execution of parallel applications. MPICH-2 from the Argonne National Laboratory is used which utilizes any Ethernet connection as well as the high-performance communication networks InfiniBand and MyriNet. Winsock Direct driver is used as the network protocol. Therefore in principle all networks with Winsock Direct support can be used as cluster interconnect.
Multiple network interfaces are supported. Especially a private, high-speed interconnect dedicated for message passing purposes can be configured. The current Windows CCS supports up to 5 different network topologies. A typical cluster network topology has two network interfaces in the head node and one network interface in the compute nodes. Another typical network topology comes with three network interfaces in the head node and two network interfaces in each compute node which supports more tightly coupled parallel applications. In this case the two network interfaces of a node are used for the separated private networks, one network for all standard network traffic and one high-speed interconnect dedicated for MPI traffic.
The Microsoft cluster software comes up with some hardware and software requirements. The computer architecture of the compute nodes must be a 64-bit Intel Xeon or AMD Athlon/Opteron with at least 0.5 GByte main memory and at least 4 GByte of hard disk. The operating system must be a version based on Windows Server 2003 x64. The head node must have the same architecture with a different hard disk layout. The remote administration and job scheduling is done by default on the head node but a system with 32-bit processor and Windows XP can be used for these administrative components, too.
The Windows Compute Cluster Server 2003 has its benefit in environments where standard Windows applications have to be supported. Now, high performance computing becomes a standard tool for all different kind of non academic applications. HPC goes mainstream also implies a shift of user requirements. The HPC environment must become compatible to the standard environment of the user's PCs, laptops, and PDAs.
In this project, the PC² evaluated the available components of the Windows CCS. Under investigation are the user management, data storage, and job scheduling. These functionalities implemented by Microsoft had to smoothly integrated in the current computing center environment. The interoperability of Linux and Windows services is of high importance for an efficient and reliable operation of the heterogeneous HPC systems.
Also the aspect of high performance had to considered. We have to prove that a Windows CCS system provides similar performance as a Linux cluster system.

Especially the performance of WinSock Direct over high performance communication networks must be investigated.
PC² operates a small cluster with Windows Compute Cluster Server 2003. Now, we are porting applications to this environment. First results will be available early 2006.

Co-operations
PC² takes part in the Microsoft test program for the Windows CCS. Since November 2005 a beta 2 version of Windows CCS is available. The general product availability is scheduled by Microsoft for the first half of 2006. Besides standard Gigabit-Ethernet networks, the high-speed interconnect InfiniBand from Silverstorm Technologies will be used [2]. Silverstorm provides the InfiniBand host channel adaptor cards and the necessary network driver for Windows Server 2003.

References
[1] http://www.microsoft.com/windowsserver2003/ccs/default.mspx
[2] http://silverstorm.com/

4.2.3 Development of Reconfigurable Cluster Systems with Field-Programmable-Gate-Arrays

Project coordinator	Dr. Jens Simon, PC², University of Paderborn
Project members	Tobias Schumacher, PC², University of Paderborn
Work partly supported by	AlphaData Inc.

General Problem Description

Scientific work today has a great demand for computing power. In the past, people tried to satisfy this demand by designing fast but also extremely expensive special-purpose hardware. Today, standard workstations have become very powerful and therefore often can replace these machines. In areas where more power is needed those workstations are connected by a fast communication network to a compute cluster which makes supercomputing much more price efficient than special hardware. On the other hand, the gain of compute power results in a big power consumption and therefore difficulties in cooling. Current research tries to address these problems by using reconfigurable hardware, for example FPGAs. These components provide the opportunity to optimize algorithms in hardware while being nearly as flexible and price efficient as commodity hardware. By using standard CPUs embedded in those FPGAs, these can work much more autonomously without depending on the aid of a host workstation.

Problem Details and Work Done in the Reporting Period

This project addresses the integration of reconfigurable components in cluster computing. One of the possible configurations is using the reconfigurable parts as co-processors while the commodity hardware still is responsible for controlling the distributed tasks. This is the configuration used in the chess monster Hydra[1]. Another approach is building complete Systems-on-Chip based on the PPC 405 CPUs available in the Vortex-II Pro FPGAs. Such Systems-on-Chip can directly run programs written in languages like C and use special purpose custom cores for acceleration of such applications. The used FPGA-boards provide the following features:

- Vortex-II Pro 2VP70-FPGA with two embedded PPC 405 CPUs
- 128MB DDR-SDRAM
- 8MB DDR-SSRAM
- 32MB FPGA-configuration flash
- 32MB PPC boot flash
- Front I/O equipped with 10/100Mbit-Ethernet
- 64Bit PCI-X bridge

Our main work concentrated on communication between two system-on-chip, realized on FPGA-Boards hosted in two different workstations. We created a simple MPI implementation[2] that transmits packages over the PCI bridge to the host system, from where they are forwarded to the destination host. This implementation does not need a full featured operating system but can run under control of the much simpler Xilkernel. This kernel supports multiple processes and threads, but does not need access to a file system for storing configuration data and other files. Using special MPI communicators, PPC-to-PPC communication as well as PPC-to-Host communication can be achieved by this implementation.

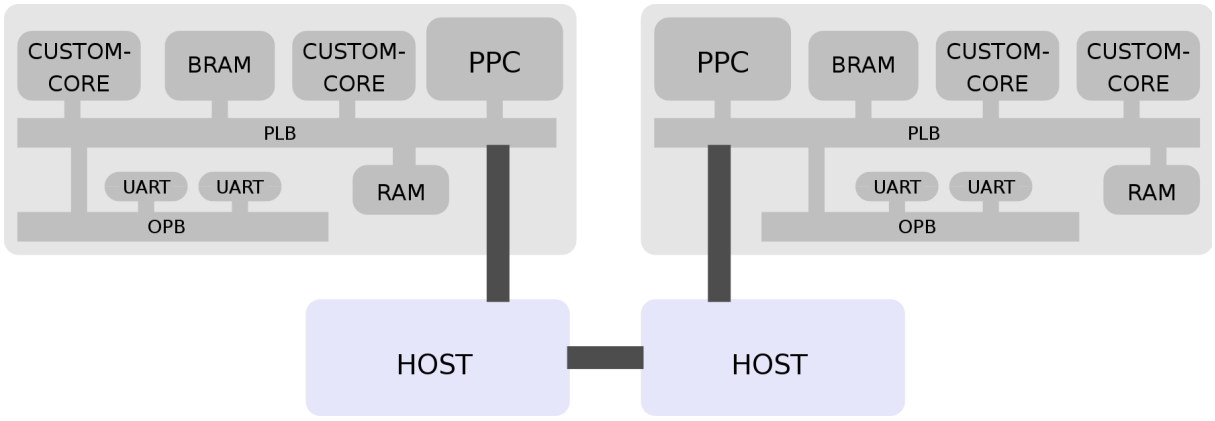

Figure 1: Communication between two PPC-Systems using the PCI-Bridge

We also put some effort in getting Linux running on these system-on-chip. Due to the very limited amount of persistent storage available, the root file system was mounted using NFS. Therefore there was a need to establish a TCP/IP connection between the file server and the PPC systems. The obvious way for this is using the 10/100Mbit-Ethernet connected to the Front I/O of the FPGA boards. Since these devices only provide a disappointing performance, we also created a linux network driver that transmit packets to the host using the PCI bridge. Accordingly, on the host side a network driver exists for communication. Using the PCI-Bridge for TCP/IP communication does not only provide a better performance than the use of the

Ethernet devices. We are now also able to replace the Ethernet module by other hight performance networks, for example RapidIO, while still being able to mount the root file systems using NFS.

Co-operations

This work was supported by AlphaData Inc. and Beam Ltd. AlphaData provided us with the FPGA-Boards including the Ethernet front Panel modules. They also delivered several supporting software like a software development kit for these boards. Part of this software was a basic configuration for the FPGAs implementing a PPC-based system using the devices located on the FPGA boards. This was the base system we customized for our needs in this project.
Additionally, Beam provided us with a release of the Montavista Linux Distribution ported to these systems and with some additional supporting tools. Based on this distribution we implemented the kernel drivers for the TCP communication utilizing the PCI bridge and all the further experiments with Linux on the FPGA boards.

References

[1] The Chess Monster Hydra: http://www.hydrachess.com
[2] Tobias Schumacher, Diplomarbeit “Untersuchung von Kommunikations-methoden zwischen FPGA-basierten Systems-on-Chip auf Basis von Message-Passing”, 2005

4.2.4 Operating Systems for Dynamically Reconfigurable Hardware: From Programming to Execution Models (ReconOS)

Project coordinator	Prof. Dr. Marco Platzner, PC², University of Paderborn
Project members	Enno Lübbers, PC², University of Paderborn

General Problem Description

The project ReconOS aims at the investigation and development of a programming and execution model for dynamically reconfigurable hardware devices. These devices, such as Field-Programmable Gate Arrays (FPGAs), are being used more and more in embedded systems but also as accelerators in general purpose computing. Being a rather new technology, dynamically reconfigurable systems are not sufficiently supported by current design methodologies and tools. Especially the unique feature of partial reconfiguration is still difficult to exploit except for hand-crafted designs.

Problem Details

We plan to add a layer of abstraction to current design methodologies for reconfigurable systems by developing an operating system that manages the (partially reconfigurable) system resources. The operating system will raise the design productivity and flexibility to a level currently achieved only for processor-based systems.

The key points of our current research are the programming model, the execution model, and the automated generation of the runtime system:

- The programming model provides the main abstractions for the application design including the definition of hardware tasks and services for task synchronization, communication, and scheduling.
- The execution model defines the runtime system which enables multitasking in dynamically reconfigurable hardware as well as across hardware/software boundaries [1],[2].
- The automated generation of the runtime system facilitates the adaptation of the operation system to different target architectures. By customization, the runtime system's memory and logic footprints are minimized which is of utmost importance for cost-sensitive embedded systems.

The ReconOS project is funded by the DFG as part of the Priority Programme 1148 “Rekonfigurierbare Rechensysteme”.

References

[1] H. Walder and M. Platzner: A Runtime Environment for Reconfigurable Operating Systems. In Proceedings 14th Int. Conf. on Field Programmable Logic and Applications (FPL), Belgium 2004, Springer

[2] H. Walder and M. Platzner: Reconfigurable Hardware Operating System: From Design Concepts to Realizations. In proceedings of 3rd Int. Conf. on Engineering of Reconfigurable Systems and Algorithms, Nevada, USA, June 2003, CSREA Press

4.2.5 Paderborn BSP Library on InfiniBand

Project coordinator:	Dr. Jens Simon, PC², University of Paderborn
Project members:	Olaf Bonorden, University of Paderborn
	Felix Schulte, University of Paderborn

General Problem Description

The PUB–Library (Paderborn University BSP–Library) is based on a small communication layer with a well defined Application Programming Interface. This API can be used to write network dependent device interfaces. The PUB-Library supports different communication networks and protocols. With the MPI version of the communication layer almost all communication networks are usable. To get most performance out of the PUB-Library, an implementation on the native network protocol will be efficient. The new ARMINIUS cluster of the PC² has an InfiniBand high-speed network for message passing which supports Remote-Direct-Memory-Access (RDMA) between interconnected compute nodes. In this project, a communication layer with RDMA support was developed.

Problem Details and Work Done in the Reporting Period

The PUB–Library is a C–library of communication routines. These routines allow implemente algorithms which are designed for the BSP model (bulk synchronous parallel model). The BSP model divides algorithms into several parts called supersteps. In each superstep a processor can work on local data and send messages. At the end of the superstep a barrier synchronization takes place and all processors receive the messages which were sent in the previous superstep. Further information concerning the BSP model can be received from [1].

PUB offers functions for both message passing and remote memory access. Furthermore, some collective communication operations like broadcast and parallel prefix are also provided. To be more flexible, PUB allows creating independent BSP objects each representing a virtual BSP computer. The PUB–Library is available for several parallel platforms. To port the PUB-Library to a new network protocol, only the small communication layer of the library must be adapted.

The *ARMINIUS* cluster has an InfiniBand interconnect as a high-speed message passing network. The Direct Access Programming Library (DAPL) developed by the DAT-Collaborative is available on InfiniBand. DAPL has a standard API for RDMA operations provided by several networks like IB (InfiniBand), VIA (Virtual Interface

Architecture), and MyriNet (Myricom Inc.). The standard DAPL-API can be used for Remote-Direct-Memory-Access operations also between computer systems with distributed memory.
The PUB-Library already supported shared memory, but this implementation is based on a atomic synchronization operation. However this synchronization operation is not supported in the current DAPL standard. Therefore a special communication protocol and message buffering scheme was developed.
The performance benefit of the PUB-Library with DAPL communication layer over the implementation with the MPI communication layer depends on the compute load of the BSP-processes and the size of messages transmitted over the network.
A communication bounded benchmark leads for large message sizes (> 1 MByte) to a per node communication bandwidth of 460 MByte/s with DAPL and 224 MByte/s with MPI. With small message sizes (< 10 kByte) the performance benefit from DAPL over MPI of factor 2 turns to a disadvantage of a factor up to 10. Small messages suffer from the high overhead for initializing a RDMA operation.
The DAPL implementation can be further optimized by using RDMA for large message sizes and programmed I/O (PIO) for small message sizes. This method is also used in almost all fast MPI implementations.

References

[1] Olaf Bonorden, Ben H. H. Jurlink, Ingo von Otte, and Ingo Rieping, The Paderborn University BSP (PUB) library; Parallel Computing, 29(2), February 2003

[2] Felix Schulte, Implementierung und Evaluation eines Message–Passing-Layers unter Nutzung von Remote-DMA-Funktionen, Studienarbeit, Universität Paderborn, 2005

[3] DAT-Collaborative Website, http://www.datcollaborative.org

4.2.6 Multi-Objective Intrinsic Evolution of Embedded Systems

Project coordinator:	Prof. Dr. Marco Platzner, PC², University of Paderborn
Project members:	Paul Kaufmann, PC², University of Paderborn
Work partly supported by:	German Science Foundation (DFG) – SPP 1183

General Problem Description

This project aims at the investigation of intrinsically evolvable embedded systems. Simulated evolution provides embedded systems with a means to react properly to unforeseen changes in the environment and the system state.

Problem Details and Work Done in the Reporting Period

Emerging reconfigurable hardware architectures combined with biologically inspired methods revealed a new approach of treating problems like robustness, fault-tolerance, self-adaptation and self-optimization in embedded systems. Applying evolutionary algorithms on reconfigurable hardware allow designing systems that are able to reorganize the structure of the problem solving function online. Hence, function recovery as a reaction for changes in the system state is inherent in this method. Further, the evolution of the problem solving function can be targeted at multiple objectives that might be conflicting, as for example speed and power consumption. A trade-off has to be found than that is good enough for solving all needs in the present situation.

To examine evolutionary algorithms on reconfigurable hardware we use a formal model [2] to simulate the process. Core algorithms and a model-invariant framework have been written and a graphical user interface is under development.

The goals of the project are:

- Develop new models and algorithms for intrinsic evolution.
- Investigate multi-objective optimization techniques.
- Examine approach-specific problems like scalability.
- Develop the basic technology for self-reconfiguration of hardware and software functions.

References

[1] M. Platzner: Multi-Objective Evolution of Embedded Systems (MOVES)

[2] J. Miller and P. Thompson: Cartesian Genetic Programming

4.2.7 Medical Image Reconstruction

Project coordinator:	Prof. Dr. Odej Kao, PC², University of Paderborn
Project members:	Stefan Lietsch, PC², University of Paderborn
Work partly supported by:	Heart and Diabetes Center North Rhine-Westphalia (HDZ NRW), Bad Oeynhausen

General Problem Description

Modern medicine more and more tends to do non-invasive examination before going into surgery. One method in this field is the Computer Tomography (CT) i.e. the Positron Emission Tomography (PET). The PET is used to examine processes of metabolism by recording the dispersion of a radioactive tracer injected into the circulation of the patient. The disintegration of the tracer is recognized by the detectors of a PET device. This is the only information that is known and it must be reconstructed to have a real image of the examined area of the body. Obviously this is a very challenging task which also demands a lot of computational power.

In cooperation with the HDZ NRW [5] the PC² wants to find new ways to accelerate this reconstruction process and to enhance its quality. Therefore different approaches are taken and new techniques are researched. The main idea is to parallelize existing algorithms to reach the goal. But also new ways like the utilization of GPUs (Graphics Processor Unit) [3] or new architectures like the Cell-Processor [4] are taken into account. In further steps of the project there are plans to integrate this reconstruction mechanism into visualization software which enables medics to view the data interactively in 3D. This may help them while making life critical decisions.

Problem Details and Work Done in the Reporting Period

Basics of the Positron Emission Tomography

To tackle the problems in Medical Image Reconstruction one first has to understand the principles of Positron Emission Tomography [1]. Figure 1 shows the schematic setup of a PET device. There is a ring of detectors which can be rotated around and moved along the patient. The rotation enables more angles of measurement and thereby a higher accuracy. The movability is used to take “pictures”, so called slices, of multiple, consecutive parts of the body. Those slices can be put together to a 3 dimensional dataset afterwards.

There are two different methods of acquiring the data measured by the detectors. The first is the so called *List Mode* where every single event is written into a list

consecutively. This method can be used to stream process the data but needs a lot of memory and special algorithms to extract the time differences between the single events. The second mode is the *Sinogram Mode* where all events are stored in a data model called Sinogram. A Sinogram is a 2 dimensional array as depicted in Figure 2. Every single point (Bin) in the Sinogram is equivalent to a possible coincidence line and is determined by the distance r from the center and the angle ϑ (as shown in Figure 1). Those Sinograms represent all events detected in a fixed timeframe and are used to reconstruct the corresponding image / slice.

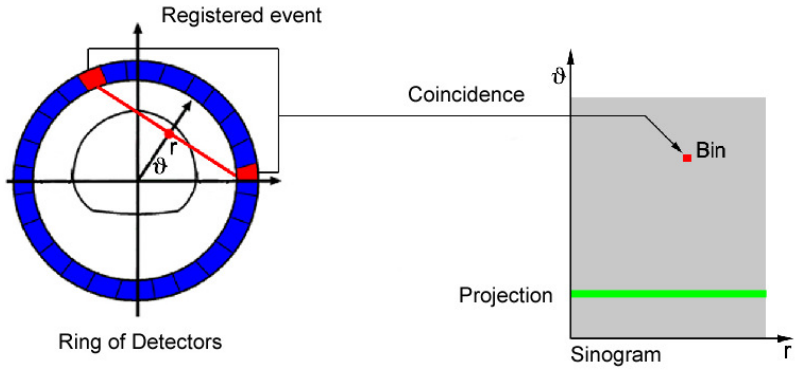

Figure 1: Detector ring of a PET device

Analytic Reconstruction Algorithms

The best-known algorithm in image reconstruction is the Inverse Radon Transformation and its various modifications. All of them have in common to use line integrals to calculate the original image from the measured projections. To correct distortions and reduce noise in the reconstructed images filters are applied to the projection. Those so called Filtered Back Projection (FBP) algorithms work very well and fast with X-Ray based CT since they deliver a very high statistic robustness of measured data. That is there is not much disturbance or distraction in the measured events. PET in contrast delivers significantly lower statistical robust results. Therefore FBP algorithms generate faulty images with a lot of noise and interfering artifacts. This is why a second group of algorithms have evolved.

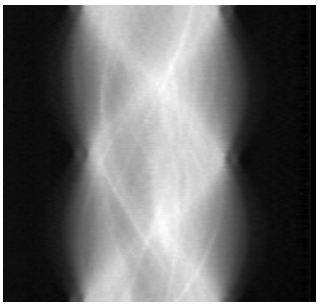

Figure 2: Sinograms (schematic and from real data)

Algebraic Reconstruction Methods

Algebraic or iterative algorithms better adapt to the problems of the PET [2]. There are a large variety of algorithms, but each starts with an assumed image, computes projections from the image, compares the original projection data and updates the image based upon the difference between the calculated and the actual projections. The major advantages of the iterative approach include insensitivity to noise and capability of reconstructing an optimal image in the case of incomplete data. The method has been applied in emission tomography modalities like Single Photon Emitting CT (SPECT) and PET, where there is significant attenuation along ray paths and noise statistics are relatively poor. For more information on CT and the used algotithms see for example [9].

Speeding up the Reconstruction

Although conceptually this iterative approach is much simpler than FBP, for medical applications it has traditionally lacked the speed of implementation and accuracy. This is due to the slow convergence of the algorithm and high computational demands. This is the matter that the PC² in cooperation with the HDZ NRW and other partners wants to improve. Since iterative methods are the only ones that can be practically applied in PET the HDZ and many other hospitals depend on results from those algorithms to make life critical decisions. The main problem of iterative algorithms as mentioned above is the vast demand of computational power. However in the last few years even conventional desktop computers became more powerful than computers available in the time when most of those algorithms where developed. Additionally clustering and new hardware designs as the Cell Broadband Engine ease the utilization of parallel and distributed computing to speed up computation.

To start off several ways of speeding up existing methods were examined. The most obvious one is to distribute the slices to different computers in a cluster and put the results together afterwards. This can easily be done and does not require big programming effort although a lot of optimization in the existing implementations can

be done for the existing systems in PC². The second approach is to go deeper into the algorithms. Since no conventional hospital can afford to operate big cluster systems with the entire maintenance and administration tasks and costs there must be a way to optimize the algorithms and the systems they are running on to fit into the special environment of hospitals. Our approach is to transfer the conclusions we gained from distributing the algorithms on big cluster systems to new specialized hardware like Cell-Processors, FPGAs and GPUs. Since these architectures outperform universal CPUs in optimized tasks by far we claim that one single machine including a Cell processor and optionally an FPGA or GPU can deliver the performance of a small cluster system for the task of image reconstruction. This system could be easily established in hospitals and wouldn't be more expensive to maintain or operate than a common workstation.

Resource Usage

In the first stage of this project we mainly use the resources of the ARMINIUS cluster to study the distributability of the given algorithms. At the same time we start evaluating the possibilities of the new technologies mentioned above. GPUs and FPGAs are already at hand in different systems and preparations are done to integrate them into the ARMINIUS environment. Since the Cell Broadband Engine isn't available yet we had to limit to a software simulation to conduct preliminary tests. As soon as systems become available we plan to implement the software on the real hardware and integrate the other components to a standalone machine. All test data and support for technical questions concerning the PET is provided by the HDZ NRW and its technical employees.

References

[1] C. Kak and Malcolm Slaney, Principles of Computerized Tomographic Imaging, Society of Industrial and Applied Mathematics, 2001

[2] H.Fricke, Ein beschleunigtes iteratives Bildrekonstruktionsverfahren für die Positronen-Emissions-Tomographie, Dissertation Hannover Medical School, 1999

[3] Homepage General-Purpose Computation Using Graphics Hardware: http://www.gpgpu.org/

[4] Homepage IBM Cell Broadband Engine Research: http://www.research.ibm.com/cell/

[5] Hompage of the Heart and Diabetes Center North Rhine-Westphalia: www.hdz-nrw.de/

[6] Herman, G.T. et al Image Reconstruction from Projections-Implemtation and Applications. Topics in Applied Physocs Volume 32. Springer-Verlag 1979

4.3 Tools, Environments, and Interfaces

4.3.1 HPC4U – Highly Predictable Clusters for Internet Grids

Project coordinator	Prof. Dr. Odej Kao, PC², University of Paderborn
Project members	Felix Heine, PC², University of Paderborn
	Matthias Hovestadt, PC², University of Paderborn
	Axel Keller, PC², University of Paderborn

General Problem Description

During the last years, Grid computing became an established instrument for academic users worldwide. Research has led to Grid middleware systems like “UNICORE” or the “Globus Toolkit”. Now the focus turns on attracting commercial users for using Grid infrastructures. However, this new user community has new challenging requirements on the Grid. In this context, already in 2002 the European Commission convened a group of experts. Their task was to identify demands of future Grid systems, which should be commercially successful. By this, they clarified properties and capabilities missing in existing Grid infrastructures.

The work of this group resulted in the idea of the Next Generation Grid (NGG) [2]. Among the core requirements on such an NGG are issues on reliability, transparency, and assurance of Quality of Service (QoS) parameters. In fact, a commercial user will not use Grid infrastructures for computing his deadline bounded or business critical jobs, if the Grid is only able to follow the best-effort approach. In contrast, the commercial user demands for contractually fixed service quality levels, on which the user is able to rely on.

In this context, a Service Level Agreement (SLA) is a powerful instrument, since it allows the description of all expectations and obligations in the business relationship between service provider and the (commercial) service customer [1]. This way, the customer is able to unambiguously specify the requirement profile of his job. In turn, the provider is able to define the service to be provided. Hence, the SLA is only accepted and enforced, if both parties agree on its content.

Numerous research projects already focus on the integration of SLA negotiation and management procedures within the Grid middleware. However, solely focusing on Grid middleware is not sufficient. Since Grid middleware is mediating between incoming user jobs to available local resource management systems (RMS), which

are providing their resources to Grid infrastructures, also these RMS have to be aware of the requirements of the job defined in the SLA. Moreover, RMS have to realize the agreed service quality level in their local domain [4]. However, current RMS are operating on best-effort approach only, not allowing Grid middleware to substantially guarantee on SLA aspects.

A major research focus at PC² is on resource management systems providing an increased level of quality of service as a software-only solution. At this, application transparency is crucial. Arbitrary applications should benefit without recompilation and linkage against special libraries from an increased level of Quality of Service, like fault tolerant job execution. This is of particular interest for commercial Grid environments, since source code of commercial applications is normally not available.

Within the EU-funded project “Highly Predictable Cluster for Internet-Grids” (HPC4U) [3] the PC² is working on an SLA-aware resource management system, utilizing the mechanisms of process-, storage- and network-subsystem for realizing application-transparent fault tolerance. The HPC4U project started in June 2004 and will end in May 2007. The partners are IBM France (FR), Fujitsu Systems Europe Ltd (UK), Seanodes SA (FR), Dolphin Interconnect Solutions AS (NO), Scali AS (NO), Paderborn Center for Parallel Computing, CETIC (BE), and National Supercomputer Centre (SE).

The goal of the HPC4U project (Highly Predictable Cluster for Internet Grids) is to provide an application-transparent and software-only solution of a reliable RMS. It will allow the Grid to negotiate on Service Level Agreements, and it will also feature mechanisms like process and storage checkpointing to realize fault tolerance and to assure the adherence with given SLAs. The HPC4U solution will act as an active Grid component, using available Grid resources for further improving its level of fault Tolerance.

Problem Details and Work Done in the Reporting Period

Resource outages (e. g. power failure of a compute node) usually cause a crash of the job running on these resources. Without checkpointing mechanisms, such a job has to be restarted from the very beginning, so that all computational results are lost. Especially for long running jobs, this is a major drawback. Deadline guarantees for such jobs are at most best effort. If a weather service uses Grid resources for computing the weekend weather forecast, the computed result would be useless if it is finished on Monday due to resource outages.

With system level checkpointing mechanisms available, the resource management system is able to create images of a running process in regular intervals. In case of resource outages, the resource management system can query for compatible and suitable spare resources to resume the job. If such resources were found, the checkpoint dataset (e. g. the process image) is transferred to the new compute node. There the job can restart from the last checkpoint. The impact on the time of completion of the job is minimal, since it is only delayed for the time required for checkpointing, dataset transfer, and restart. This leverages the realization of fault tolerance and the provision of deadline guarantees. The project HPC4U will realize an application-transparent checkpointing, so that arbitrary applications can benefit from this service. However, it is not sufficient to focus on process checkpointing only. The storage subsystem will provide checkpointing mechanisms for the storage partition of a running job. If the resource management system initiates a checkpoint of a running process, it simultaneously starts the checkpointing of the data partition. If the job needs to be restarted due to a resource failure at a later time, both process and storage are restored. This assures the consistency of the running process with its saved data.

If an application is running on multiple nodes in parallel, each node may send messages to the other nodes of the applications (e. g. information exchange or synchronization). If such an application is checkpointed, a node might currently be sending a message to another node. At time of checkpoint, this message might have already left the sender, but not yet reached the recipient. Such a packet is called in-transit. Consistency is a major demand to the checkpointing mechanism. Hence, the network subsystem must handle these in-transit packets. At checkpoint time, the checkpointing subsystem will first freeze the application, so that its state is stable and does not change. Now the network subsystem is invoked to check all stacks for network packets. Also the network itself is checked for currently transmitted packets. All these packets are saved in a network dataset file.

After all goals, requirements, interfaces, and module interconnections have been specified within the first workpackage of HPC4U, the focus of the second workpackage was on realizing fault tolerance for non-parallel applications. The PC² has been involved in all tasks of this workpackage, which were the integration of the fault tolerance building blocks into the RMS, the enhancement of system monitoring capabilities of the RMS, as well as verification and validation. Moreover, PC² actively worked on creating all five deliverables of this workpackage, two of them as responsible partner.

As result of this second workpackage a first version of the envisaged cluster middleware has been realized. This system integrates the software components of all partners and orchestrates their functionality for providing fault tolerance for non-

parallel applications to the user. By this, the project consortium reached an important milestone. The functionality of this software system has been presented during the first annual EC-review meeting.

Resource Usage at PC²

For the runtime of the HPC4U project, a cluster system has been installed as reference platform for the HPC4U project. This cluster consists of five Dell PowerEdge 2650 nodes. Each of these nodes is equipped with two Intel Xeon 2.4 Ghz processors, 1 GB of RAM, 73 GB harddisk and a Myrinet M3F-PCIXD-2 card. Furthermore a Dolphin Interconnect SCI card (AS PSB66) has been integrated into these nodes, since the SCI interconnect is used as standard in HPC4U.
One of these five nodes will be used as a frontend node for the cluster system, where users will be able to submit their jobs. This frontend node will also represent the interface for Grid middleware, as HPC4U will allow Grid users to negotiate on specific service levels. All other nodes of the cluster are used as dedicated compute nodes.

References

[1] Sahai et al. Specifying and Monitoring Guarantees in Commercial Grids through SLA. Technical Report HPL-2002-324, Internet Systems and Storage Laboratory, HP Laboratories Palo Alto, November 2002.

[2] H. Bal et al. Next Generation Grids 2: Requirements and Options for European Grids Research 2005-2010 and Beyond. ftp://ftp.cordis.lu/pub/ist/docs/ngg2_eg_final.pdf, 2004.

[3] Highly Predictable Cluster for Internet-Grids (HPC4U), EU-funded project IST-511531. http://www.hpc4u.org.

[4] L.-O. Burchard et al. The Virtual Resource Manager: An Architecture for SLA-aware Resource Management. In *4th Intl. IEEE/ACM Intl. Symposium on Cluster Computing and the Grid (CCGrid) Chicago, USA*, 2004.

4.3.2 AssessGrid – Advanced Risk Assessment and Management for Trustable Grids

Project coordinator	Prof. Dr. Odej Kao, PC², University of Paderborn
Project members	Matthias Hovestadt, PC², University of Paderborn
	Kerstin Voß, PC², University of Paderborn

General Problem Description

Using Grids for outsourcing is an attractive opportunity to save money for commercial users. However, the commercial Grid utilization has not been yet established. Commercial users require an individual amount of Quality of Service (QoS) since they need guarantees for their own business plans. Contemporary resource providers, however, only work on best effort. To establish the Grid commercialisation, the customer's required QoS has to be guaranteed from resource providers. Therefore, the quality characteristics are defined in contracts between customers and providers, the Service Level Agreements (SLAs).

In order to guarantee SLAs, Grid actors in all Grid layers (the Grid fabric, the Grid middleware, and the Grid service) have to meet negotiated SLAs. The resource provider in the Grid fabric has to prepossess the SLA by offering an adequate job execution. Grid middleware brokers have to select a trustable offer from a resource provider, and the end-user has to estimate the trustiness of trust the offered SLA from brokers or providers. Since SLAs are mandatory for the commercialisation, integrating them into all Grid layers is the topic of current research projects.

However, a gap exists between SLA as a concept and as an accepted tool for the commercial Grid utilization. Providers are cautious about agreeing on SLAs since these contain business risks for them: resources can fail, experts may be unavailable, resources can be overbooked etc. Such events are critical for prepossessing SLAs. This may result in SLA violations and paying penalties.

Hence, providers need risk assessment methods as decision support for accepting/rejecting SLAs, price/penalty negotiations, for initiating fault-tolerance actions, as well as for capacity and service planning. Customers need the risk estimation and aggregated confidence information for provider selection and fault-tolerance/penalty negotiations. Furthermore, they are informed about the existing risk of failure and can estimate their individual consequences.

The project AssessGrid will successively expand all Grid layers with risk assessment and management functionalities. Beyond several layer-oriented loosely coupled modules, a vertically integrated solution of all developed functionalities will be

provided. The several modules will simplify the integration of risk management within different Grid middleware configurations and operating systems.

The Paderborn Center for Parallel Computing, particularly Prof. Dr. Odej Kao, will be the AssessGrid project coordinator. Therewith he will have the position of the legal correspondent between the European Commission and the consortium. The consortium is a well balanced partnership between research (PC² - University of Paderborn [Germany]; IAMSR - Abo Akademi University [Finland]; University of Leeds [United Kingdom]), an SME (CETIC [Belgium]), and even global players (ATOS Origin [Spain]; Wincor Nixdorf [Germany]).

AssessGrid has been proposed as a project in call 5 of the FP6 Specific Programme for “integrating and strengthening the European Research Area” (SP1). It was evaluated with excellent 28 of 30 points. Therewith we achieved the second position of all 26 classical instruments in the strategic objectives 2.5.4 “Advanced Grid Technologies, Systems, and Services” (it is even tied with the first ranked proposal). In consequence a high probability of project funding from the EC is given. The negotiations between the consortium and the EC will presumably take place in the first months of 2006. If funded, the project is planned to start in April or May 2006. It will run 33 month.

Problem Details and Work Done in the Reporting Period

The Paderborn Center for Parallel Computing developed the idea of the project in May 2005. Up to the completed proposal, the PC² worked hard to join a qualified consortium as well as to write an excellent proposal.

At the end of May we participated on the Grid Technology Days in Brussels. This conference is organised by the EC in order to publish information about the current FP call. Since many researchers and experts use this offer from the EC, it also is an opportunity for meeting Grid researchers and experts. On the one hand, we received important information about general requirements for funding. On the other hand we presented our project idea in a short talk in order to get in contact with potential partners. Following to this presentation, we got a positive feedback to our addressed topic from research and industry. Therewith we received the first confirmation that the project idea is good and relevant in current research. In spite of these new contacts, partners with a special expertise in Grid brokerage, risk assessment and management, Grid Economics, as well as an end-user had to be found. In Brussels, we also met CETIC, one of our project partners in HPC4U [1], who also showed their

interest in our new project and became the first member in our AssessGrid consortium.
In the next planning phase, we specified the project idea, addressed topics in detail, and planned the workflow. Further hints about the general requirements of funding were given us at a workshop of the Paderborn University (“Erfolgreiche Antragstellung in EU-Forschungsprogrammen"). In parallel to the project specification and developing the proposal’s concept, we searched for contemplable researchers and institutes with special expertise. On the base of published papers (for example [2] from the University of Leeds), internet presentations, and even personal contacts, we could complete our consortium. The expert for Grid Brokerage is the University of Leeds and for Grid Economics is ATOS Origin.
Risk management and assessment is well known for financial risk estimations. The integration into the Grid is a totally new application area. We could enthuse IAMSR from Abo Akadmi University for this new application area so that they join the consortium. In the same way, we attracted Wincor Nixdorf for the project goals. Their business role at the moment is not a typical Grid provider. However, since their customers outsource business essential functionalities to them, risk management technologies are important for Wincor Nixdorf’s business.

In the next step it was important that all project partners got the same idea of the planned work, workflow, and details of AssessGrid. Accordingly, we organised an one-day preparatory meeting in London at the end of July. After brainstorming and discussions, we fixed the focused objectives, project responsibilities, and workflow. From August until September we wrote in cooperation with the other partners the proposal with extraordinary diligence. The PC² had a special role in that progress since we coordinated the work, wrote main parts of the proposal, and reviewed everything in detail.

References

[1] Highly Predictable Cluster for Internet-Grids (HPC4U), EU-funded project IST-511531. http://www.hpc4u.org.
[2] SLA Management in a Service Oriented Architecture, K. Djemame, M. Haji, and J. Padgett, In Proceedings of ICCSA’2005, Singapore, May 2005, Lecture Notes in Computer Science 3483, pp.1282-1291

4.3.3 D-Grid: German Grid Initiative

Project coordinator	Prof. Dr. Odej Kao, PC², University of Paderborn
Project members	Dominic Battré, PC², University of Paderborn
	Bernard Bauer, PC², University of Paderborn
	Felix Heine, PC², University of Paderborn
	Matthias Hovestadt, PC², University of Paderborn
	Holger Nitsche, PC², University of Paderborn

General Problem Description

The D-Grid (www.d-grid.de) is a German Grid computing initiative started in September 2005. It is supported by the BMBF (Bundesministerium für Bildung und Forschung) and encompasses over 100 institutions all over Germany. The initiative consists of several Grid application projects from scientific communities like High Energy Physics, Life Sciences, or Climate Research. The technological basis for the project is provided by the so-called *integration project*. This project will develop and package a middleware for the community projects and provide a Grid generic infrastructure consisting of high performance compute resources, storage space, and other specialized devices.

The PC² is part of the D-Grid initiative in the context of the integration project and has several tasks. First of all, PC² is responsible to develop and run the D-Grid portal in close cooperation with FZK (Forschungszentrum Karlsruhe). Especially important within this portal is a monitoring section where users receive status information about the D-Grid resources.

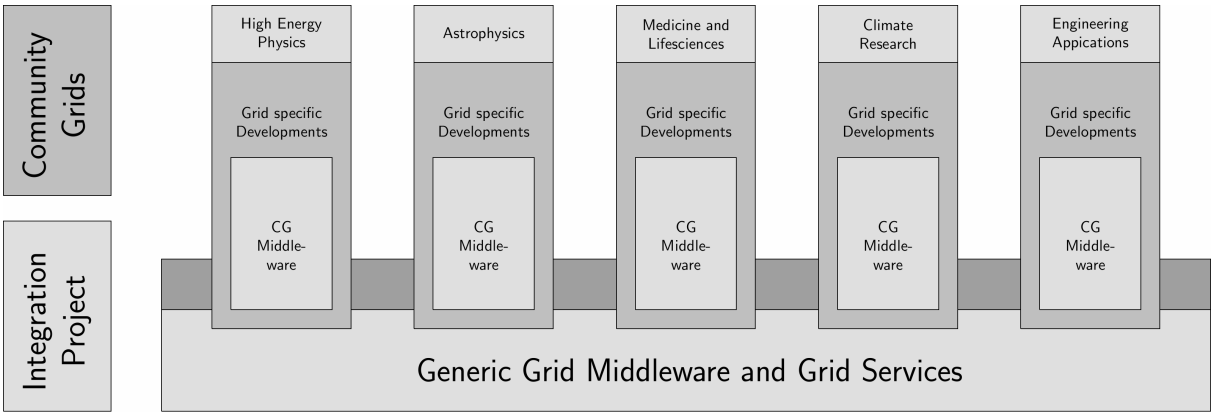

Figure 1: Projects of the D-Grid Initiative 1

Furthermore, PC² integrates its ARMINIUS cluster with the D-Grid using commonly used middleware packages like Globus [2] and UNICORE [3],[4]. Thus, each user of D-Grid may use parts of the ARMINIUS cluster in a transparent way running its local Grid client.
An important part of the security of Grid infrastructures are digital certificates. Each user of D-Grid is required to have a certificate which is used during authentication. The certificates used in D-Grid follow the European policy for Grid certificates, which permits only one tier of Certificate Authorities (CAs) in each nation. Thus, in Germany there are only two authorized CAs. To ease the application and the issuing process for new certificates, the PC² has established and runs a local Registration Authority (RA) for grid certificates. This allows users to apply for certificates locally on site.

Problem details and work done

RA

As stated in the introduction, a registration authority is an important entity improving the usability of Grid systems. Without a local RA, Grid users from Paderborn would have to travel either to the DFN office in Hamburg or to the GridKA center in Karlsruhe for authentication as these are the places where the German Grid CAs are located. Thus, PC² established a local RA. The RA has customized Web-Pages where users can apply for a personal certificate. From these pages, a form is generated and printed, which has to be signed by the user. Users then have to visit PC² personally, carrying an id-card or passport, in order to verify their identiy. Employees of PC² will sign their request and forward it to the mentioned CA. The certificate is then issued to the user via e-mail.
The procedure is described in detail on the PC² web pages, together with all relevant links, and addresses to get support.

UNICORE

UNICORE is a vertically integrated Grid middleware. It has a three layer architecture shown in Figure 1. Users define and submit their jobs locally using the graphical UNICORE client. The client provides various plug-ins to enter application-specific parameters during the job definition in a convenient way. The job gets submitted to the UNICORE gateway, which is a dedicated server residing within the demilitarized zone (DMZ) of the firewall. The gateway serves as a dispatcher forwarding the job to the target system. The various systems running in the UNICORE framework are virtualized behind a so-called network job supervisor (NJS), which gets the responsibility for the jobs running on the system. The incarnation database (IDB) describes both the hardware aspects and the installed software of underlying system. The user database (UUDB) maps the grid certificates to local usernames and defines who is permitted to use the system. Finally, the target system interface translates

high-level UNICORE commands into commands of the local resource management system, which is the computing center software (CCS) in our case.

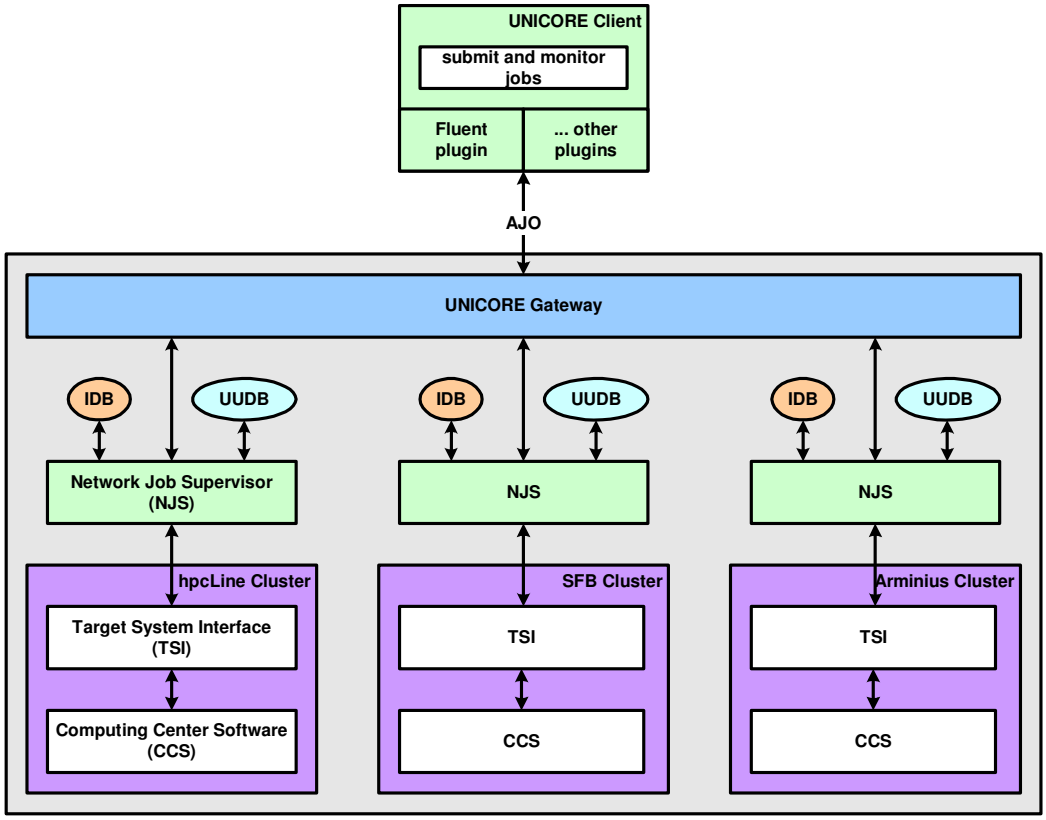

Figure 2: UNICORE at PC²

At PC², we have currently 3 clusters running in our UNICORE infrastructure. The hpc*L*ine system PSC2 [5] and the Itanium-based SFB Cluster [6] used to be integrated in our UNICORE infrastructure in the past already and in the scope of the D-Grid initiative, we have now also integrated the new ARMINIUS cluster [7] as well. In 2006, various UNICORE systems from the D-Grid will be integrated using a central gateway, thus forming a powerful Grid system.

Globus Toolkit

The work on Globus Toolkit [8] started in 1998 at Argonne National Laboratory and University of Chicago. The goal was to establish an open platform for Grid computing, following open standards, e.g. defined in the Global Grid Forum (GGF) [9]. Meanwhile version 4 of the Globus Toolkit is available, which is based on open standard Grid services. Since Globus Toolkit has rapidly evolved and is used in a multitude of projects, it has become the “de-facto standard” in Grid computing. The

Globus Toolkit is in use at the PC² since 2001, allowing users to access compute resources at PC² by submitting compute jobs via the Globus toolkit infrastructure. For the hpc*L*ine system PSC2 this Grid interface has been operational and frequently used since 2002. This way, users have been able to easily access resources at the PC² using standard Globus mechanisms - without knowing about the usage of the resource management system CCS, which is used for operating cluster systems at the PC².

For the D-Grid project, the Globus Toolkit installation at the PC² has been updated to the most recent version 4.0.1. Since this new version has major differences in the interface between the Globus Toolkit and local resource management systems, the already existing interface between Globus and CCS had to be updated. Moreover, also CCS had to be enhanced to support this new interface. Having the new interface available, users from D-Grid are now able to access resources from the ARMINIUS cluster using their local Globus Toolkit 4 system.

Portal

The D-Grid web portal [10] serves the purposes of external presentation and internal collaboration. The Typo3 system, a content management system used for the web sites of many enterprise level companies, has been installed to allow members of the D-Grid initiative and its community projects an easy access to their individual web appearances while preserving a consistent layout throughout the whole web site. Apart from an online rich text editor, Typo3 provides necessary authentication functionalities, automatic menu generation, and support for bilingual content. Within the scope of the portal project, we will implement a Typo3 plug-in to support a certificate based single-sign-on solution throughout various web sites of the D-Grid Initiative and the community projects. This sign-on grants access to internal areas which provide online collaboration tools, such as document sharing facilities, access to an event calendar, and detailed information about resources such as mailing lists, the trouble ticket system, and a resource status map mentioned before. The PC² is in charge of creating and maintaining a central point of information and supporting other members of the D-Grid initiative at questions related to the common web appearance.

D-Grid homepage at http://www.d-grid.de 1

Resource Usage

The ARMINIUS cluster will be part of the new Grid infrastructure D-Grid. For this to work, also a large number of other servers at PC² are used, like the gateway server for UNICORE.

References

[1] Foster and C. Kesselman (Eds.). The Grid 2: Blueprint for a New Computing Infrastructure. Morgan Kaufmann Publishers Inc. San Francisco, 2004
[2] The Globus Alliance, http://www.globus.org
[3] UNICORE Forum e.V., http://www.unicore.org
[4] UNICORE at SourceForge, http://unicore.sourceforge.net
[5] The hpc*L*ine at PC². http://www.upb.de/pc2/services/systems/psc
[6] The SFB Cluster at PC². http://www.upb.de/pc2/services/systems/ic
[7] The ARMINIUS Cluster at PC²: http://www.upb.de/pc2/services/systems/arminius
[8] The Globus Toolkit, http://www.globus.org
[9] The Global Grid Forum, http://www.ggf.org
[10] The D-Grid web portal, http://www.d-grid.de

4.3.4 Computing Center Software (CCS)

Project coordinator	Prof. Dr. Odej Kao, PC², University of Paderborn
Project members	Axel Keller, PC², University of Paderborn
	Sebastian Ritter, PC², University of Paderborn

General Problem Description

The availability of commodity high performance components for workstations and networks made it possible to build up large, PC based compute clusters at modest costs. Large clusters with hundreds of processors operated as multi purpose systems demand for multi user management and efficient administration tools.

A resource management system is a portal to the underlying computing resources. It allows users and administrators to access and manage various resources like processors, memory, networks, or permanent storage.

Workstation clusters are often not only used for high-throughput computing in time-sharing mode (i.e. multiple jobs are sharing the same resources at the same time) but also for running complex parallel jobs in space-sharing mode (only one job is using the resources). This poses several difficulties to the resource management system, which must be able to reserve computing resources for exclusive use and also to determine an optimal process mapping for a given system topology.

The Computing Center Software is such a resource management system. It provides a homogeneous access interface to a pool of different High Performance Computer (HPC) systems, while for system administrators it provides a means for describing, organizing, and managing HPC systems that are operated in a computing center. Hence the name “Computing Center Software”, CCS for short.

CCS itself is a distributed software system possibly consisting of hundreds of entities. Its functional units have been kept modular to allow adaptation to new environments. A CCS island is responsible for one underlying system. It consists of six components. Each part contains several modules and/or daemons which may run asynchronously on different hosts to improve the performance.

Figure 1 depicts the architecture of a CCS island.

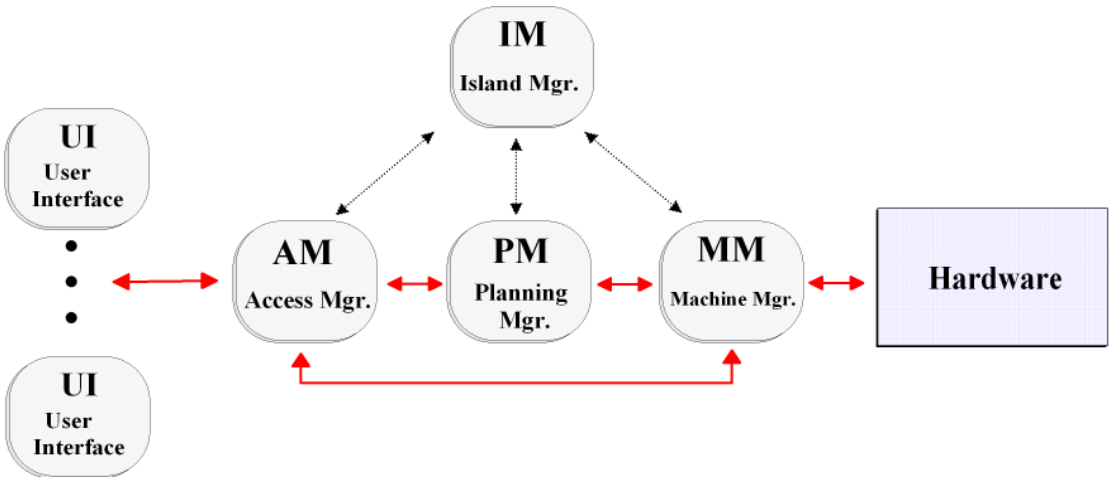

Figure 1: Architecture of a CCS island

1) The *User Interface (UI)* offers X-window or ASCII access to the machine.
2) The *Access Manager (AM)* manages the user interfaces and is responsible for authentication, authorization, and accounting.
3) The Planning *Manager (PM)* schedules the user requests onto the machine.
4) The Machine Manager (MM) provides an interface to the machine specific features like partitioning, job controlling, etc.
5) The *Island Manager (IM)* provides name services and watchdog functions to keep the island in a stable condition.
6) The *Operator Shell (OS)* is the X-window based interface for system administrators to control CCS, e.g. by connecting to the other components.

Problem Details and Work Done in the Reporting Period

Unlike to well-known resource management systems like Condor [5], LSF [6], PBS [7], or SGE [8] CCS is targeted for the support of space-sharing parallel computers. Its resource description facility qualifies CCS to compute an efficient mapping of partitions onto the nodes. Another important difference to the mentioned resource management systems is the way how CCS schedules incoming requests. The criterion for the differentiation of resource management systems concerning the aspect of scheduling is the planned time frame.

Queuing systems try to utilize currently free resources with waiting resource requests. Future resource planning for all waiting requests is not done. Hence, waiting resource requests have no proposed start time.

Due to their design, queuing systems provide no information that answers questions like “Is tomorrow's load high or low?” or “When will my request be started?”. Hence, advanced reservations are troublesome to implement which in turn makes it difficult to participate in a multi-site grid workflow run.

Planning systems in contrast plan for the present and future. Planned start times are assigned to all requests and a complete schedule about the future resource usage is computed and made available to the users.
Hence, planning systems are well suited to participate in grid environments and multi-site application runs. There are no queues in planning systems. Every incoming request is planned immediately. However, planning systems also have drawbacks. The cost of scheduling is higher than in queuing systems.

Figure 2 shows an example of a schedule planned by CCS. The X-axis represents the time and the Y-axis the used number of nodes.

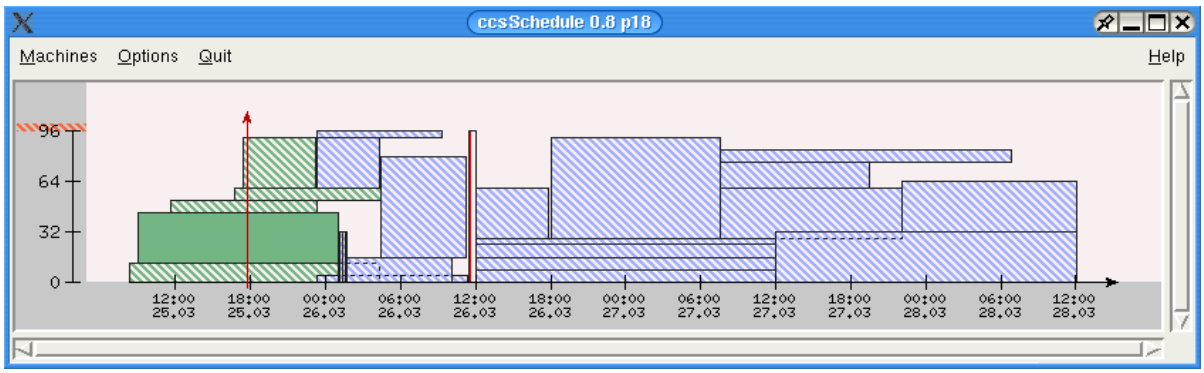

Figure 2: A CCS Schedule

Based on this scheduling paradigm CCS provides the following features:

- *Information about Planned Start Times:* The CCS user interface shows the estimated start time of interactive requests directly after the submitted request has been planned. This output will be updated whenever the schedule changes.
- *Advanced Reservations:* CCS can be used to reserve system resources for a given time. Once CCS has accepted a request, the user is guaranteed access to the requested resources. During the reserved time frame a user can start an arbitrary number of interactive or batch jobs
- *Deadline Scheduling:* Batch jobs can be submitted with a deadline notification. Once a job has been accepted, CCS guarantees the job to be completed at (or before) the specified time.

- *Limit Based Scheduling:* In CCS authorization is project based. One has to specify a project at submit time. CCS knows two different time slots: weekdays and weekend. In each slot CCS distinguishes between day and night. All policies consider the project specific node limits (given in percent of the number of available nodes of the machine). This means that the scheduler will sum up the already used resources of a project in a given time slot. If the time dependent limit is reached, the request in question is planned to a later or earlier slot (depending on the request type: interactive, reservation, deadline etc.).
- *System Wide Node Limit:* The administrator may establish a system wide node limit. It consists of a time slot [start, stop], a number of nodes (N), a threshold (T), and duration (D). The limit defines that during the interval [start, stop] N nodes are kept free for requests which consists of less than T nodes and have duration not longer than D. This ensures that small partitions are not blocked by large ones during the given interval.
- *Admin Reservations:* The administrator may reserve parts or the whole system for a given period of time for one or more projects. Only the specified projects are then able to allocate and release an arbitrary number of requests during this interval on the reserved nodes. Requests of other projects are planned to an earlier or later time. An admin reservation overrides a project limit and the current system wide node limit. This enables the administrator to establish virtual machines with restricted access for a given period of time and a restricted set of users.
- *Duration Change at Runtime:* It is possible to manually change the duration of already running or waiting requests.

Resource Usage at the PC²

CCS is used for the management of the following parallel systems operated at PC²:

1. The ARMINIUS cluster: The system consists of 400 Intel-Xeon processors. The ARMINIUS is accessible via the Globus Grid software toolkit and UNICORE.
2. The PSC2 cluster: The system consists of 192 Pentium-III processors. The PSC2 is accessible via the Globus Grid software toolkit and UNICORE.
3. The SFB cluster: This system consists of 8 Itanium-2 CPUs. The 4 nodes are connected via Ethernet, Myrinet, and Infiniband. The SFB is accessible via UNICORE.

References

[1] Felix Heine, Matthias Hovestadt, Odej Kao, Axel Keller: Provision of Fault Tolerance with Grid-enabled and SLA-aware Resource Management Systems Proceedings of the Parallel Computing Conference 2005, Malaga, Spain

[2] Lars-Olof Burchard, Felix Heine, Matthias Hovestadt, Odej Kao, Axel Keller, Barry Linnert: A Quality-of-Service Architecture for Future Grid Computing Applications Proceedings of the 13th International Workshop on Parallel and Distributed Real-Time Systems, April 2004 (WPDRTS 2005)

[3] Felix Heine, Matthias Hovestadt, Odej Kao, Axel Keller: SLA-aware Job Migration in Grid Environments L. Grandinetti (Edt.): Grid Computing: New Frontiers of High Performance Computing, pp. 185-201, Elsevier, 2005

[4] Lars-Olof Burchard, Felix Heine, Hans-Ulrich Heiss, Matthias Hovestadt, Odej Kao, Axel Keller, Barry Linnert, Jörg Schneider: The Virtual Resource Manager: Local Autonomy versus QoS Guarantees for Grid Applications Future Generation Grids, Springer, 2005

[5] Condor. http://www.cs.wisc.edu/condor

[6] LSF - Load Sharing Facility. http://www.platform.com/products/wm/LSF/index.asp

[7] PBS - Portable Batch System. http://www.openpbs.org

[8] SGE - Sun Grid Engine. http://www.sun.com/software/gridware

[9] PSC2-, PLING-, SFB- Cluster. http://www.upb.de/services/systems

4.3.5 DELIS: Large-Scale P2P Data Management

Project coordinator	Prof. Dr. Friedhelm Meyer auf der Heide, HNI, University of Paderborn Prof. Dr. Burkhard Monien, PC², University of Paderborn Prof. Dr. Odej Kao, PC², University of Paderborn
Project members	Felix Heine, PC², University of Paderborn Giovanni Cortese, Telekom Italia Learning Services Fabrizio Davide, Telekom Italia Learning Services Federico Morabito, Telekom Italia Learning Services

General Problem Description

Large scale data management is a challenging task. Through modern information technologies, more and more data and information is available online. However, it is often difficult to find the relevant information, and to efficiently combine the pieces found in various data sources. The problems lie both in the sheer amount of data and in syntactic and semantic heterogeneities.

Even within single organizations, sometimes a large number of different data sources is available. The need to process this data collection as a whole and to draw conclusions from the aggregated information residing in the sources has lead to the field of data warehousing. The typical approach is to copy every relevant piece of information into a large, centrally managed data warehouse (DWH), which is then used for evaluating queries, e.g. in so called decision support systems [1].

However, in many cases the relevant data sources span multiple organizations. In the collaboration of multiple companies, or when mining the information contained in the World Wide Web, it is necessary to combine multiple heterogeneous information sources which are decentrally controlled. The Semantic Web initiative [2] aims at this goal.

To face these challenges, a system needs to be capable of integrating *a huge number* of information sources which are *syntactically and semantically heterogeneous*, *decentrally managed*, and which may be *highly dynamic*. Two important aspects have to be regarded. First, the underlying infrastructure must be scalable and flexible. Second, the system must handle the heterogeneities of the data sources.

With respect to *infrastructure*, P2P systems [3] have gained much attention of the research community in recent years. They provide a good basis for large-scale data management systems. Compared to traditional approaches, P2P systems offer good scalability features combined with decentralized control and flexibility.

In the context of the EU project DELIS we are working on a P2P data management system which fulfills the above mentioned demands. We utilize a structured P2P network called *Pastry* [4] and use the W3C standard *Resource Description Framework* (RDF) [5] for knowledge description. Within the reporting period, we focused on query evaluation.

Problem details and work done

In general, we assume to have a large P2P network. Each node in the network has some local knowledge stored using RDF. Knowledge in RDF is described as a set of *triples*, which can be regarded as short sentences of the form *subject, predicate, object*. A set of triples forms a labeled graph.

The nodes also have local schema knowledge stored as RDF Schema [6] triples. The schema knowledge does not need to be the same for every node. In fact, we are convinced that it is impossible to ensure synchronization of schema knowledge in large world-wide distributed environments or to restrict the schema to a single common standard. Moreover, it is desirable to allow each node to add locally needed schema information on the fly. If new entities need to be described, new classifications may become necessary. Waiting for a new version of some standard schema does not solve this problem.

However, we assume that there is an ontology which serves as a common schema, at least for some subsets of the nodes. This ontology will be the basis which can be extended locally. Additional schema knowledge may be stored to allow translation from one ontology to the other. Without such common understanding, no interoperability would be possible.

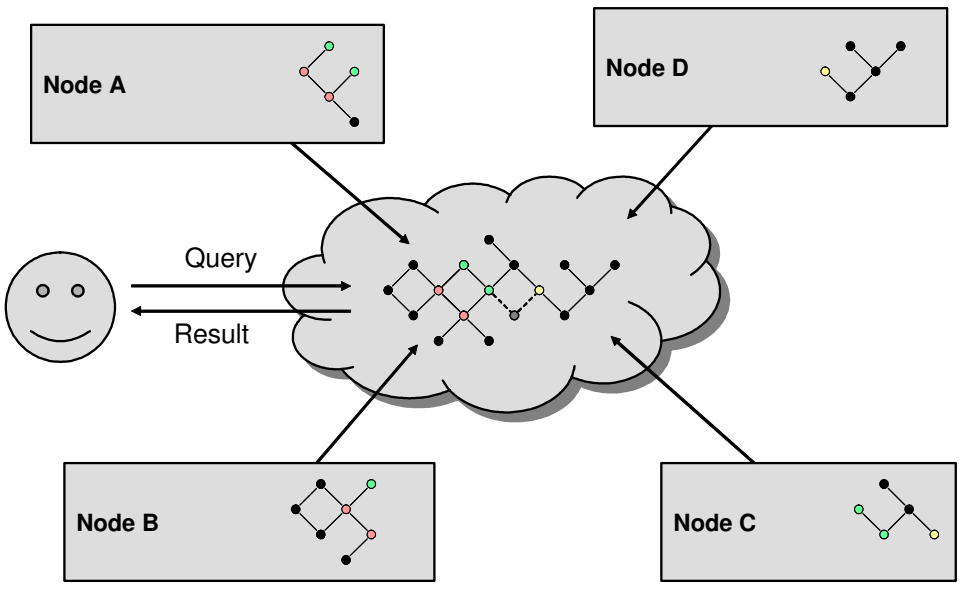

Figure 1: Virtual pool of knowledge

Our desired result is to put all this knowledge from all the nodes virtually in one pool, apply RDFS entailment rules to this pool and evaluate queries with respect to the union of the knowledge, (Figure 84). This approach is very beneficial, as overlaps in the schema knowledge are used to build bridges between different schemas used by different nodes.

A query is formulated as a pattern consisting of multiple triples where parts of the labels are replaced by variables. Thus query answering is essentially sub-graph matching [7]. However, in our scenario, the model graph is distributed over the nodes of the P2P network. Thus query processing is a two-fold process. In the first phase, a sub-graph is fetched from the network which has to be large enough to contain all query results, but should be as small as possible to save network transmission time. In the second phase, the matching sub-graphs for the query are searched locally within the retrieved sub-graph.

In order to be able to retrieve parts of the model graph in a deterministic way, we pre-distribute the triples over the nodes of the network. We use a *distributed hash table* [8], which means that the network splits up the key range of the hash table and assigns responsibility for each sub-range to a well-defined node in the network. Typical implementations of distributed hash tables like the Pastry network need O (*log n*) routing steps to forward a message with a given key to the responsible network node. In our case, the keys are the labels in the RDF graphs. They are hashed and mapped to the nodes of the network.

RDF Schema reasoning is done by evaluating rules. An example of such a rule is

If X is a sub-class of Y,
and A is an instance of X,
then A is also an instance of Y

It states that instances are propagated towards more general concepts in a class hierarchy. As the preconditions of these rules always share a common variable, there will also be a single node in the network which stores all triples of the precondition. Thus the rules can be evaluated locally. However, the resulting triple has to be distributed to the network to the responsible nodes.

The retrieval of the candidates is done triple by triple. To retrieve candidates for a triple of the query graph, at least one element (subject, predicate, or object) has to be known. Thus in each step, a triple is selected which can be used to retrieve candidates. The candidates are used to determine possible values for variables, which are in turn used in the next steps to retrieve candidates for the other triples.

We have developed and compared different versions of the algorithm. They differ in the way how the next triple for candidate retrieval is determined. A basic version simply chooses an arbitrary triple from the set of viable triples. More advanced versions employ look-ahead strategies to select triples with small candidate sets. We have additionally used Bloom-filters [9] to reduce the number of transferred triples.

The final evaluation uses a backtracking algorithm which iterates over all possible combinations of candidates. Clashes in the variable bindings are detected as early as possible to cut parts of the search tree.

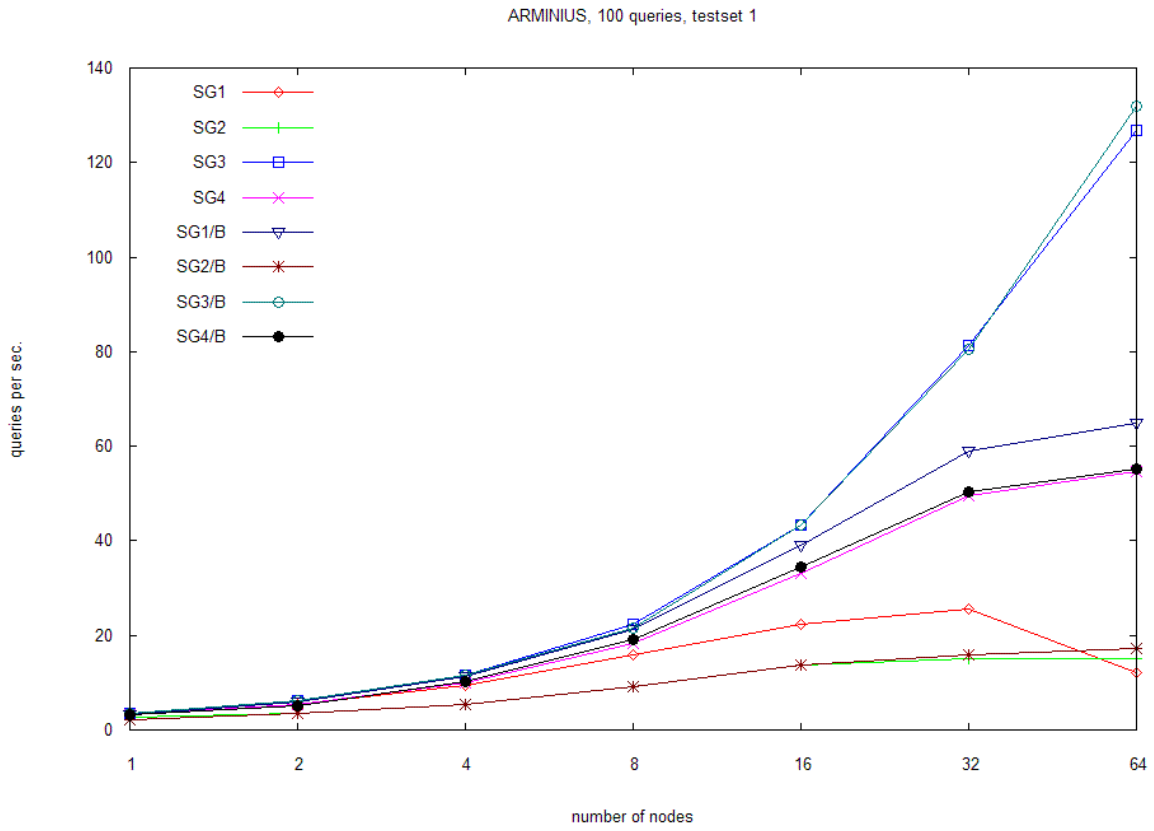

Figure 2: Evaluation on ARMINIUS

The evaluation of our system has been performed on the ARMINIUS cluster of PC². We have generated an artificial model graph together with a set of 100 test queries. We have started the network using various node counts and have measured the overall throughput of the system, by starting a query client on each node which looped over the set of test queries. The result is depicted in Figure 2. The version named SG3 performs best, which uses a sophisticated look-ahead strategy. However, a further refinement of this strategy in version SG4 is performing worse as too much network traffic is generated during the look-ahead. Furthermore, we can see that in most of the versions, the use of Bloom-filters does not significantly increase the performance. However, in version SG1 the version with bloom-filters (SG1/B) performs much better, as the candidate sets are much larger in this version. Thus we conjecture that with larger model graphs Bloom-filters will get more valuable.

We have published the results of our work in two papers; the first [10] appeared already in 2005, while the second will appear in 2006 [11]. Details about the system can be found in these papers.

Resource Usage

P2P networks are difficult to evaluate. As they target large deployed systems with possibly thousands of nodes, it is normally not possible to test the system in the desired scale. Thus simulations of the network are a good way to gain insight into a system's properties. We ran various simulations using the ARMINIUS cluster for the evaluation of our system. Additionally, we have deployed the system on the cluster to measure some properties on the real system. Although the scale is smaller than the target scale we could already see interesting effects like load-balancing influences.

References

[1] William H. Inmon, *Building the Data Warehouse,* John Wiley & Sons, 2005

[2] Tim Berners-Lee, James Hendler, and Ora Lassila, *The Semantic Web*, Scientific American, May 2001

[3] Ralf Steinmetz and Klaus Wehrle, *Peer-to-Peer Systems and Applications*, Springer, LNCS 3485, 2005

[4] Antony Rowstron and Peter Druschel, *Pastry: Scalable, decentralized object location and routing for large-scale peer-to-peer systems*, Proc. of the 18th IFIP/ACM International Conference on Distributed Systems Platforms, 2001

[5] Frank Manola and Eric Miller, *RDF Primer*, http://www.w3.org/TR/rdf-primer, 2004

[6] Dan Brickley and Ramanathan V. Guha, *RDF Vocabulary Description Language 1.0: RDF Schema*, http://www.w3.org/TR/rdf-schema, 2004

[7] Julian R. Ullmann, *An Algorithm for Subgraph Isomorphism*, Journal of the ACM, 23:1, pages 31-42, 1976

[8] John Kubiatowicz, Extracting Guarantees from Chaos, Communications of the ACM, 46:2, pages 33-38, 2003

[9] Burton H. Bloom, Space/Time Trade-offs in Hash Coding with Allowable Errors., Communications of the ACM, 13:7, pages 422-426, 1970

[10] Felix Heine, Matthias Hovestadt, Odej Kao. *Processing Complex RDF Queries over P2P Networks.* Workshop on Information Retrieval in Peer-to-Peer-Networks P2PIR 2005, November 4, 2005

[11] Felix Heine. *Scalable P2P based RDF Querying*. In First International Conference on Scalable Information Systems (INFOSCALE06), to appear, 2006.

4.4 Numerical Algorithms and Applications

4.4.1 Computational modeling of Rare Earth doped GaN

Project coordinator	Prof. Dr. Thomas Frauenheim, PC², University of Paderborn
Project members	Simone Sanna, University of Paderborn
Work supported by	RENiBEI Research Training Network

General Problem Description

In the last decade the attention of the scientific community has been attracted by Rare Earth (RE) doped semiconductors, new promising materials which can be used for a variety of applications including optoelectronic and spintronic. In particular RE-doped semiconductors emit almost monochromatic visible light, whose color depends from the particular RE used as doping substance. Europium (Eu), Erbium (Er) and Thulium (Tm) doped semiconductors have already been successfully exploited as primary color source (Red-Green-Blue) in the realization of phosphor displays. Responsible for the emission spectra of RE doped samples are the sharp optical transitions taking place in the *4f*-orbitals. As the *4f*-electrons are shielded by outer orbitals, the emission spectra of the RE-doped samples are quite unaffected by the semiconductor host: it can be then freely chosen in order to get the best results in terms of light emission and device performance. Gallium nitride (GaN) is a wide gap semiconductor which has proven to be a particularly valid host for the RE as it is transparent to visible light and because the emission is not quenched at room temperature as in the case of other hosts. In Figure1 the cromaticity graph of RE-doped GaN is reported.

The points at the angles of the triangle are those that got doping GaN with Er, Eu and Tm: all the colors within the triangle are potentially reachable by doping GaN in a proper way. The interest around RE-doped GaN grew further last year, as it was doped with Europium (Eu) and Gadolinium (Gd) and used for the realization of spintronic devices. The key objective of this work is to understand the fundamental mechanism ruling the formation of RE defects in GaN and determining the properties of the material.

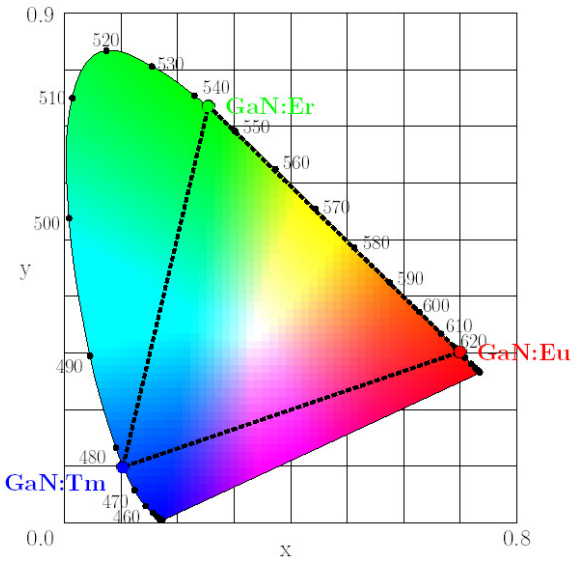

Figure 1: Cromaticity graph of RE-doped GaN

In order to reach this goal we systematically perform computer simulations of the microscopic structure and properties of the defects involved in the luminescence, using the efficient computational scheme DFTB developed at the University of Paderborn and internationally awarded by the scientific community.

Problem details and work done

Despite the increasing interest around RE-doped materials and their commercial applications the theoretical comprehension of these systems lies behind the experimental knowledge. This is mainly due to the fact that modeling the microscopic structure and properties of the defects involved in the luminescence presents a substantial challenge to current theoretical methods like the supercell-approach in the framework of the density functional theory (DFT). While it is possible to investigate issues of defect stability using pseudo-potential based approaches, which avoid the problems of modeling strongly-correlated *f*-electron systems, this cannot address luminescence from these centers.

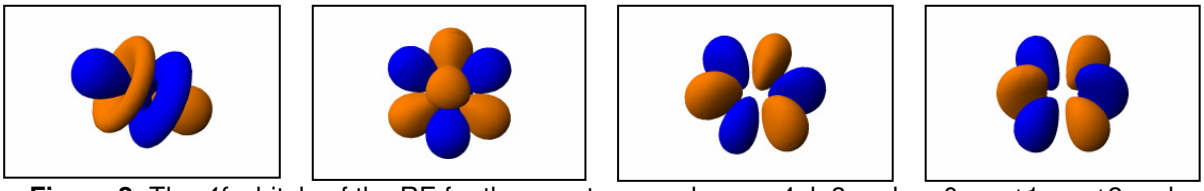

Figure 2: The *4f* orbitals of the RE for the quantum numbers n=4, l=3 and m=0, m=±1, m=±2 and m=±3.

Explicitly treating *4f*-electrons is beyond the reach of the usual mean-field methods normally employed in the DFT and other computational methods which are able to address the problem in a rigorous way like the GW approximation are computationally very demanding. Furthermore, as the RE ions are “heavy atoms” (atomic numbers ranging from 58 to 71) relativistic effects have to be taken in account in the models for the simulations. In attempt to improve the theoretical

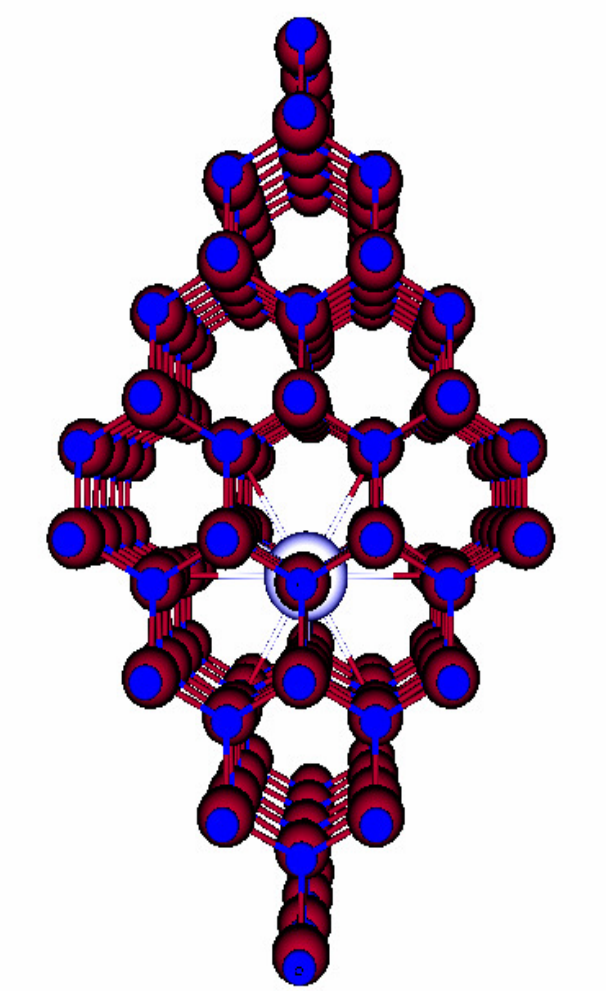

Figure 3: A wurtzite GaN supercell containing an Erbium substitutional. Supercells like this one are used to model different defect configurations and calculate their structural and electronic properties.

description of these systems we present results using density-functional based calculations on the properties of selected lanthanides in wurtzite GaN. Both substitutional defects and complexes with native defects are considered. We account for strong-correlation of the *4f*-shell using a variation of the LDA+U method, and discuss the role of spin-orbit coupling in defect properties.

The work, realized in the year 2005, can be divided in two parts, the first being the generation of a parameter set for our computational scheme and the second being the proper simulation of defect and defect complexes. In the DFTB approach a set of parameters including both atomic properties and interatomic “hopping” and repulsive potentials is needed for each type of atom pair considered. We chose to parameterize the rare earths Pr, Eu, Er and Tm because of their optical properties and Eu and Gd because of their magnetic properties. We also developed parameters for the simulation of the host GaN and its “classic” dopants, H, C and O.

It is in fact still unknown whether the RE-related emission is due to the interaction of the RE with the host atoms (Ga and N) or, at least partially, with its impurities (i.e. Er-O complexes). The parameter were tested against experimental results (where present) and against other more sophisticated, but also far more computationally demanding, theoretical approaches, finding in all the cases a very good agreement. After the testing and validation of the parameters, different defects and defect complexes are simulated with supercells reproducing the crystal symmetry. A thorough systematic study of the RE-defects in GaN (including formation energy, electronic and structural properties etc.) is needed in order to find the defect responsible for the observed emission.

Each supercell is fully relaxed in order to find stable defect configurations. The number of possible defect types, geometric configurations, charge states and spin configurations makes this systematic work very demanding and only possible with a very efficient computational scheme (DFTB) running on a modern, high-developed computational architecture like the ARMINIUS cluster. The next step of this study will be the extension of the simulation to other very promising III-V semiconductor nitride-systems like AlN and InN.

Resource Usage

The calculations were executed at the ARMINIUS cluster at the PC². The FORTRAN90 code of DFTB was locally compiled with the Intel compiler. Geometry optimization, electronic structure calculation and other simulations were run as serial processes running on one processor at time. Calculations were executed on a weekly basis, each calculation lasting from a minimum of a couple of hours to a maximum of one week.

References

[1] A. J. Steckl, J. Heikenfeld, D. S. Lee and M. Gartner, Mat. Science and Eing., B81, 97 (2001).
[2] G. M. Dalpian and Su-Huai Wie, Phys. Rev. B 72, 115201 (2005).
[3] Th. Frauenheim, G. Seifert *et. al*, J. Phys. Cond. Matter 14, 3015 (2002).
[4] J. S. Filhol, R. Jones, M. J. Shaw, P. R. Briddon, Appl. Phys. Letts. 84, 2841 (2004).
[5] B. Hourahine, S. Sanna, et al. Phisica B, *in press* (2006)

4.4.2 Computational studies on epoxy adhesion at the surface of native γ-Al_2O_3

Project coordinator	Prof. Dr. Thomas Frauenheim, PC², University of Paderborn
Project members	Jan M. Knaup, University of Paderborn
Work supported by	DFG SPP 1155

General Problem Description

During the past two decades aluminum has constantly gained importance in technical applications, with its use spreading from aerospace to automotive applications and now covering nearly every area of industrial and consumer appliances. At the same time adhesive technology has achieved advances which enable adhesive bonding of metal parts to nowadays supplement or even replace traditional metal joining technologies such as welding, bolting or riveting. Correlated to this is the development of metal-resin or metal-resin-fiber compound materials, which have recently advanced to the point of technical application. The demands for lighter and at the same time stronger materials, in order to build more fuel-efficient, i.e. lighter, vehicles and aircrafts without sacrificing structural strength leads to a strong interest in the development and improvement of fiber reinforced aluminum-polymer hybrid materials and adhesive bonding technology for aluminum. For both of these technologies, the aluminum-polymer adhesion is of crucial importance.

Since, outside high vacuum environments, aluminum instantly develops a surface oxide layer, the problem of organic adhesion on aluminum translates into the problem of organic adhesion on native aluminum oxide. The improvement of adhesion technology requires an understanding of the underlying chemical processes of the bonding of organic adhesives to the alumina surface. In this study we aim at finding a suitable methodology for gaining insight into the initial bonding of adhesive molecules as well as the bonding competition between different organic species at the native Al_2O_3 -surface. To gain a better understanding of these adhesion phenomena, we work to model the binding behaviour of a simple epoxy adhesive system at the surface of alumunim oxide.

Problem details and work done in the Reporting Period

As a first step towards understanding the adhesion of organic molecules, we investigate the reaction products and –paths of the initial adsorption reactions of the components of our model adhesive at the hydroxylated surface of γ-Al_2O_3. These components are diglycidylesterbisphenol-A (DGEBA) as the resin, diethyltriamine (DETA) as the hardener and 3-aminopropylmethoxysilane (sold as: Dynasilan AMEO, here referred to as AMEO) as an adhesion promoter component. We calculate the structures and energies of the isolated organic molecules and the adsorbed molecules on the surface model using the *self-consistent charge density-functional based tight binding* (SCC_DFTB) method, to obtain the reaction energies of the adsorption reactions. Figure 1 shows the calculated products structures of the AMEO adsorption reaction at two different surface sites. Our surface model contains 374 atoms in a 2D-periodic supercell, together with the organic compounds, this leads to about 400 atoms for the image geometries.

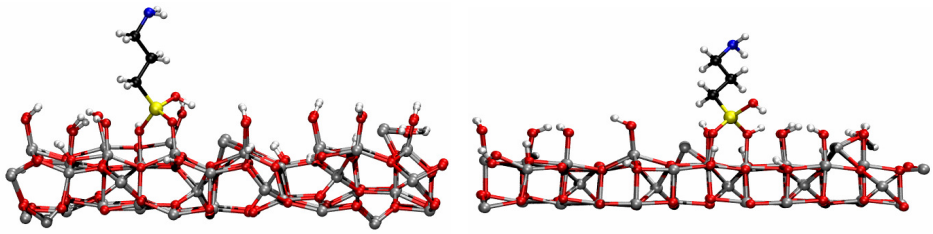

Figure 1: Structures of AMEO adsorbed at two different sites of the alumina surface model. (Al: grey, O: red, C: black, N: blue, Si: yellow, the surface model is truncated at the bottom).

We then employ the nudged elastic band (NEB) method to calculate the total energies of a chain of intermediate structures, called images, between the educt and product geometries. These images sample the minimum energy path of the given reaction and allow to asses the reaction barriers and kinetics. Figure 2 shows the MEP-s obtained for the adsorption of AMEO. Figures 2 (b) and (c) show the two most important transition states along the adsorption path. We sample between 7 and 11 images along each reaction path.

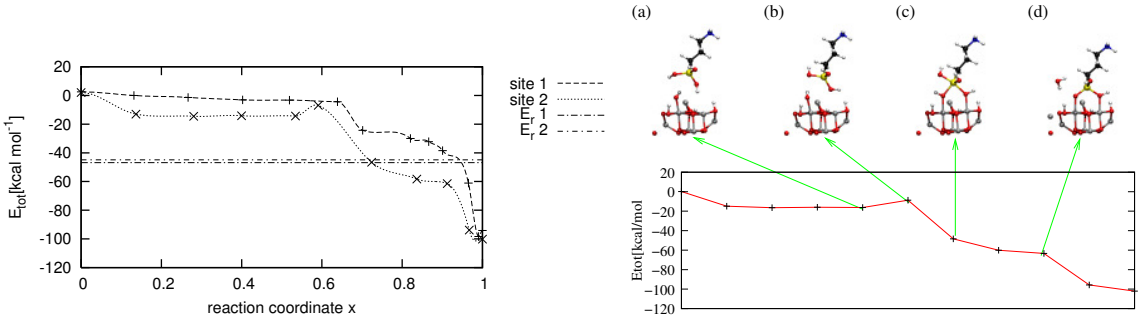

Figure 2: Interpolates minimum energy paths of the AMEO adsorption (left) and selected image geometries along the adsorption path of AMEO at site 2 (right) (Al: grey, O: red, C: black, N: blue, Si: yellow, +-symbols show calculated image total energies, the red line serves to guide the eye.)

From this data we aim to construct a model of the competition between the organic compounds for the adsorption sites available at the surface.
Additional analyses of the reaction paths allow to asses the ranges of electronic and mechanical disturbance of the surface, induced by the adsorption reactions. Figure 3 shows the root of mean square displacements (RMSD) and maximum charge differences of the atoms during the adsorption of AMEO at site 2.

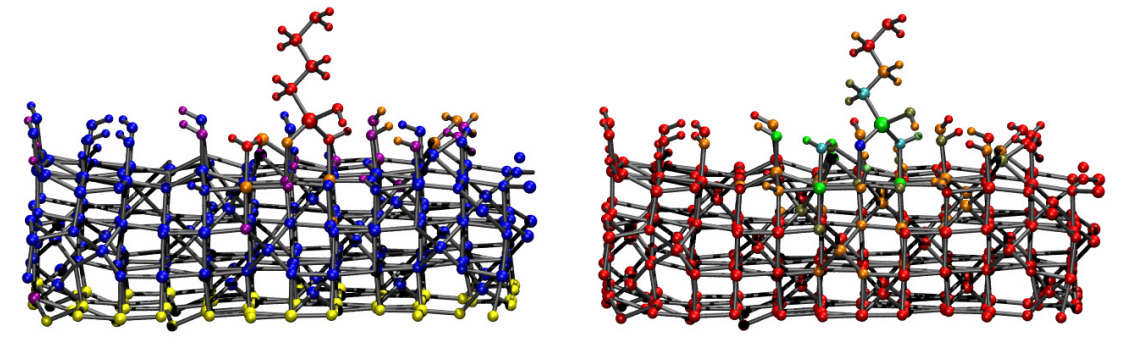

Figure 3: (left) Movement of the atoms during the adsorption of AMEO atsite 2, color coded by RMSD over the entire reaction path: RMSD < 0.05 : blue, 0.05 < RMSD < 0.1 : purple, 0.1 < RMSD < 0.3 : orange, RMSD > 0.3 : red. Fixed atoms are yellow (right) Mulliken charge difference \Delta C=C_{max}-C_{min} of each atom for the reaction of AMEO at site 2 in units of electrons. ΔC=0.0: red, ΔC=0.2: green, ΔC=1.0: blue, colors are continuously blended.

These results will be used to help us in developing a multi-scale coupling method for modeling reactions at organic/inorganic hybrid interfaces, by providing an insight into the influence ranges across the model.

Resource Usage

We use the ARMINIUS cluster at the PC² in two different modes: As a compute cluster for our serial quantum mechanical code: we employ to calculate the structures and total energies of our reaction educts and products, and as a parallel cluster for our MPI parallelized code to search for the reaction paths. For the latter, we use a Python wrapper application to distribute single runs of our serial code across several nodes via MPI. The serial program runs take about 15 to 20 minutes and a number of calls equal to the number of images along the path are completely independent of each other. For our project, this allows us to parallelize up to 11 total energy calculations with negligible communications requirements and therefore with nearly perfect scaling behavior. Because of the low communications overhead of our application, we use the Ethernet interconnect of the cluster.

We use the ARMINIUS Cluster on a daily basis, with up to 35 nodes and 70 processors in parallel. Since summer 2005 we have used an estimated 50000 total CPU-hours.

References

[1] P.V. Straznicky, J.F. Laliberté, C. Poon, A. Fahr, Polymer Composites 21, 558 (2004).

[2] K. Wefers and C. Misra, Tech. Rep. Oxides abd Hydroxides of Aluminum, Alcoa Laboratories (1987).

[3] F. Holleman and E. Wiberg, Lehrbuch der Anorganischen Chemie (de Gruyter, Berlin, 1995), pp. 1081-1083.

[4] Th. Frauenheim, G. Seifert, M. Elstner, Z. Hajnal, G. Jungnickel, D. Porezag, S. Suhai and R. Scholz, phys. stat. sol. (b) 217, 41 (2000)

[5] H. Jónsson, G. Mills and K.W. Jacobsen, Classical and Quantum Dynamics in Condensed Phase Simulations (World Scientific, Singapore, 1998), chap. Nudged elastic band method fir finding minimum energy paths of transitions, pp. 387-405

[6] J. M. Knaup, C. Köhler, Th. Frauenheim, M. Amkreutz, P. Schiffels, B. Schneider, O.-D. Hennemenn, in preparation (2006)

4.4.3 PAL/CSS Online Freestyle Chess Tournament Participation

Project coordinator	Prof. Dr.-Ing. Dietmar P.F. Möller, Department of Computer Science, University of Hamburg
Project members	Kai Himstedt, Department of Computer Science, University of Hamburg

NOTE: The work was done by externals in cooperation with members of the PC² core group (by a short and very friendly contact to Dr. Rainer Feldmann and Bernard Bauer) and by using the available hardware at the PC² which was administrated and configured for our special requirements with great commitment by Holger Nitsche.

General Problem Description

Chess matches are an appropriate measure to determine relative playing strengths and to evaluate parallel algorithms of chess programs. Taking computer chess as an example, a new idea, named *optimistic pondering*, is presented in Himstedt [1]: how to use a distributed environment for asynchronous parallel game-tree search. In sequential chess programs, searching ahead with the expected opponent's move is the de facto standard to avoid the processor being idle while the opponent is on move and thinking [2], [3]. The basic principle of *optimistic pondering* is to start searching ahead not only at the beginning of the opponent's thinking time but already during the own thinking time with another processor. This idea makes it possible (naturally with consequences concerning the efficiency) to use more than two processors in an analogous way which in the progression of the game leads to a kind of pipeline principle (see Himstedt [1] for some efficiency considerations). The abstract concept of a meta-chess engine was chosen to embed the idea into a prototypical environment. A special advantage of this concept is the possibility to use existing concrete chess programs (so called chess engines) as "workers" (e.g. CRAFTY [4], DEEP SHREDDER [5] etc.) without the necessity of any engine modification.

To test the implementation of the meta-chess engine under real match conditions and to get a first impression about the relative playing strengths of the system the participation in the PAL/CSS Online Freestyle Chess Tournament [6] was chosen. Freestyle chess goes back to an idea of Prof. Ingo Althöfer and is characterized by the fact that the use of any external help (e.g. computer assistance or especially chess programs) is allowed during the matches. Those were of course very suitable conditions to participate with the meta-chess engine using DEEP SHREDDER for the

worker engines. About 50 participants from 20 different countries started in the qualification part of the tournament. Among the participants were even some International Masters (IMs) and International Grandmasters (GMs). The well known HYDRA [7] chess system, which is closely connected to the PC² and supported by the PAL group in Abu Dhabi in the United Arab Emirates (one of the initiators of the tournament) also participated.

Problem Details and Work Done in the Reporting Period

The meta chess engine concept is composed of two major components to achieve an utmost independence from a concrete chess engine. One of these components is a proxy chess engine which behaves to the graphical user interfaces (GUIs) (e.g. XBoard and WinBoard [8], Arena [9] or the ChessBase GUI [10]) like a normal chess engine by being controllable with an engine protocol (xboard/WinBoard [11]). The proxy engine performs no tree search itself, but has some kind of a master role to control (completely transparent for the GUI) the *optimistic pondering* with distributed worker clients. The worker clients form the second major component of the meta-chess engine. The communication is technically based on the freely available message passing interface MPI in the Mpich2 implementation variant [12], [13]. A worker client is a more abstract component based on a small C++ framework to handle different engine protocols (xBoard/WinBoard [11] and UCI [14] up to now) and performs no tree search either. For the real tree search each worker client controls an associated unmodified concrete chess engine. That means of course that these concrete base engines may themselves run on multiprocessor systems.

Very often it is tried with the help of typical test suites to evaluate the performance of parallel algorithms for chess programs (e.g. in [15], [16], [17], [18]). Such suites are based on chess positions and used to test e.g. how fast a mate (decoupled from a game) is found or how many “correct solutions” can be found and in which – hopefully short – time. The existence of an opponent (even necessary for *optimistic pondering*) as some kind of real time condition is ignored by the test suites. The PAL/CSS Online Freestyle Chess Tournament, on the contrary, formed implicitly real match conditions, of course. Generally, a match condition is mainly defined by the rules for the time control (e.g. how many moves have to be played in which time), and a major consequence of time control is that each side has after all usually 50% of the total time for a game available to think on the opponent’s time. An obvious goal is to utilize the total time for the own tree search.

During the tournament, held from May 28 to June 19, 2005, the PC² and especially the SFB 376 provided us with additional computing power of a Quad-Opteron-System (2.2 GHz, 32 GB main memory, 2 GB of which were used by DEEP SHREDDER

for the transposition table) for our distributed cluster environment. The cluster environment consisted until then mainly of four Dual-Xeon-Systems (2.67 GHz, 2 GB main memory) provided by the Regional Computer Centre at Hamburg University [19]. With this environment the 7th place of 48 in the qualifying part and the 15th place of 45 in the main tournament was reached under the user name VoidChessICC.

Himstedt [1] concludes that *optimistic pondering* does not need high communication bandwidths due to its asynchronous character. The asynchronous behavior is characterized by the quasi event-driven handling of the rather infrequently received new information from the real chess engines (i.e. principal variations) containing the currently expected best opponent's move. This makes it particularly suited for distributed environments, also because there is no demand for a shared memory support. The participation in the Freestyle Chess Tournament with the strongest computer of the cluster placed in Paderborn and the other computers placed in Hamburg, linked via an ordinary 1 GBit/s TCP/IP ethernet connection, confirmed this in practice.

Optimistic methods have often proved to be valuable for computer science in the past, if for these methods it can be assumed that the optimistic presumption is normally met (e.g. for optimistic locking in data base systems (Kung and Robinson [20]) and even to recompile modified sources in parallel with an optimistic "Make" approach (Bubenik and Zwaenepoel [21], cf. Himstedt [22])). Due to the high ponder hit probabilities (i.e. the opponent usually plays the expected best move from the program's point of view) the optimistic presumption should often be met for the optimistic pondering method. The results from the tournament are so promising that additional experiments are done at the moment to get more performance results and with more detail. Furthermore, adding workstations to the cluster is planned to analyze the behavior of *optimistic pondering* with a greater number of processors.

References

[1] K. Himstedt: An Optimistic Pondering Approach for Asynchronous Distributed Game-Tree Search. In ICGA Journal. Vol. 28 No. 2 (2005):77-90.

[2] R. Hyatt: Using Time Wisely. In ICCA Journal. Vol. 7 No. 1 (1984):4-9.

[3] Althöfer, C. Donninger, U. Lorenz, V. Rottmann: On Timing, Permanent Brain and Human Intervention. In H. J. van den Herik, I. S. Herschberg, J. W. H. M. Uiterwijk (Eds). Advances in Computer Chess 7. University of Limburg. Maastricht (1994):285-296.

[4] CRAFTY (Hyatt's home page): http://www.cis.uab.edu/info/faculty/hyatt/hyatt.html

[5] DEEP SHREDDER: http://www.shredderchess.com/shredderdeep.html
[6] PAL/CSS Online Freestyle Chess Tournament (held from May 28 to June 19, 2005 on www.playchess.com): http://www.computerschach.de
[7] HYDRA: http://www.hydrachess.com
[8] XBoard and WinBoard graphical user interfaces for chess: http://www.tim-mann.org/xboard.html
[9] Arena, graphical user interface (GUI) for chess engines: http://www.playwitharena.com/
[10] ChessBase: http://www.chessbase.com
[11] xboard/WinBoard Chess Engine Communication Protocol: http://www.tim-mann.org/xboard/engine-intf.html
[12] Message Passing Interface (MPI) Forum Home Page: http://www.mpi-forum.org/
[13] Mpich2 Home Page: http://www-unix.mcs.anl.gov/mpi/mpich2/index.htm
[14] Universal Chess Interface (UCI) Protocol: http://www.chessbase.com/download/index.asp?cat=UCI%2DEngines
[15] M. Newborn: Unsynchronized Iteratively Deepening Parallel Alpha-Beta Search. In IEEE Transactions on Pattern Analysis and Machine Intelligence. Vol. 10 No. 5 (1988):687-694.
[16] B. C. Kuszmaul: The Startech Massively-Parallel Chess Program. In ICCA Journal. Vol. 18 No. 1 (1995):3-19.
[17] R. D. Blumofe, C. F. Joerg, B. C. Kuszmaul, C. E. Leiserson, K. H. Randall, Y. Zhou: Cilk: An Efficient Multithreaded Runtime System. In Journal of Parallel and Distributed Computing. Vol. 37 No. 1 (1996):55-69.
[18] M. G. Brockington: Asynchronous Parallel Game-Tree Search. PhD Thesis. University of Alberta. Edmonton 1998.
[19] Regional Computer Centre, University of Hamburg Home Page: http://www.rrz.uni-hamburg.de/RRZ/e.index.html
[20] H. T. Kung, J. T. Robinson: On Optimistic Methods for Concurrency Control. In ACM Transactions on Database Systems. Vol. 6 No. 2 (1981):213-226.
[21] R. Bubenik, W. Zwaenepoel: Optimistic Make. In IEEE Transactions on Computers. Vol. 41 No. 2 (1992):207-217.
[22] K. Himstedt: Verfahren zur Vermeidung redundanter Übersetzungen in modularen Softwaresystemen. Diplomarbeit im Fach Informatik. Universität Hamburg. Hamburg 1993.

4.4.4 Color Tuning in Rhodopsins

Project coordinator	Prof. Dr. Thomas Frauenheim, PC², University of Paderborn
Project members	Michael Hoffmann, University of Paderborn Dr. Peter König, University of Madison Marius Wanko, University of Paderborn
Work supported by	DFG FG Retinal Action

General Problem Description

Rhodopsins are the most prominent family among the photoreceptor proteins [12]. The family comprises photosensitive reseptor proteins in animal eyes including human rod and cone visual pigments as well as archaeal-type rhodopsins found in halophilic archaea. The latter ones work as phototaxis receptors (sensory rhodopsin), light-driven proton (bacteriorhodopsin) or chloride (halorrhodopsin) pumps.

All rhodopsins share some structural features [13]: The crucial light-absorbing chromophore is retinal (vitamin-A aldehyde), a conjugated polyene chain. It is bound covalently via a protonated Schiff.

Interestingly, the protein environment drastically modulates the absorption maximum of the chromophore from a value of about 440 nm in organic solvents to values ranging from 425 to 560 nm in light sensitive cone pigments responsible for color vision. The mechanism by which protein environments regulate the absoption maximum of the chromophore (spectral tuning) is therefore of fundamental importance for understanding the process of color vision and is investigated in this project. The project is the part of the Forschergruppe 490 “Molecular Mechanisms of Retinal Protein Action” supported by the Deutsche Forschergemeinschaft (http://phys.upb.de/retinal/).

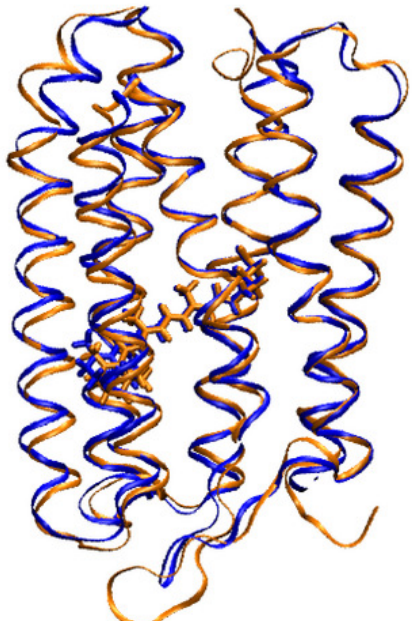

Abbildung 1: QM/MM optimized structures of bR (blue) and sRII (red)

Problem details and work done

Sensory rhodopsin (sRII) and bacteriorhodopsin (bR) are two prominent rhodopsins. Despite theier high structural similarity (Figure 1), the absorption maximum of sensory rhodopsin II (sRII) is strongly blue-shifted by about 70 nm relative to that of bacteriorhodopsin (bR).

Previous theoretical investigations came to contradictory conclusions regarding the cause for the spectral shift [1,2]. Additionally, there are disagreement with experimental findings.

Because of these open issues, we revisited the problem of color regulation in bR vs. sRII with focus on reproducing and interpreting the available experimental data [8]. After verification of the used methods [9,10], we applied QM/MM calculations to determine the absorption spectra of these two retinal proteins. In a first step the effect of mutations on the absorptions energies was compared with experimental values. Our results are in very good agreement with the experimental findings. In a second step, the focus was shifted to the investigation of the mechanisms of the spectral tuning. For this purpose, the contribution of different protein side chains to the spectral shift has been investigated by successively switching off the electrostatic interaction of various residues with the chromophore.

These studies confirm the results from the mutation experiments by showing that no single residue is responsible solely fort he shift of the absorption maximum. The effect seems to be a collective phenomenon to which several distinctive effects contribute, e.g., the counter ions as well as a few prominent point mutations [8].

In a last step QM/MM moleculary dynamics simulations have been performed to sample the conformational space. Based on the obtained trajectories, vertical absorption energies werde calculated along the trajectorieswhich in turn were used to obtain the absoption spectra and eventually the spectral shift. The results reproduce the shift of the maximum between bR and sRII (Figure 2).

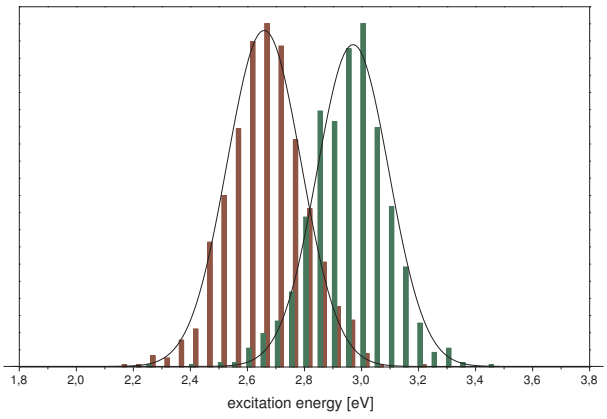

Figure 2: Histogramme der Anregungsenergien für bR (rot) und SRII (grün)

Resource Usage

For the calculations, the programm suite CHARMM [3] was used in combination with SCC-DFTB [4] in a QM/MM framework [9,11]. The excitation energies were calculated with the semi-empirical OM2 method [5,6] in a multi-reference approach [7] for the ground and excited state. Parts of the calculations were conducted using the ARMINIUS Cluster at the PC².

References

[1] Hayashi, S., Tajkhorshid, E., Pebay-Peyroula, E. Royant, A., Landau, E.M., Navarro, J., Schulten, K. J. Phys. Chem. B 105, 10124 (2001).
[2] Ren, L., Martin, C.H., Wise K.J., Gillespie, N.B., Luecke, H., Lanyi, J.K., Spudich, J.L., Birge, R.R. Biochemistry 40, 13906 (2001).
[3] König, P.H., Hoffmann, M., Frauenheim, Th., Cui, Q., eingereicht
[4] Wanko, M., Hoffmann, M., Strodel, P., Koslowski, A., Thiel, W., Neese, F., Frauenheim, Th., Elstner, M., J. Phys. Chem. B im Druck
[5] Hoffmann et. al. in Vorbereitung
[6] B. Brooks, R. Bruccoleri, B. Olafson, D. States, S. Swaminathan, M. Karplus, J. Comp. Chem. 4, 187 (1983)
[7] Elstner, M., Porezag, D., Jungnickel, G., Elsner, J., Haugk, M., Fraunheim, Th., Suhai, S., Seifert, G., Phys. Rev. B 58, 7260 (1998)
[8] Cui, Q., Elstner, M., Kaxiras, E., Frauenheim, Th., Karplus, M., J. Phys. Chem. B 105, 569 (2001)
[9] Weber, W.: Dissertation Universität Zürich (1996)
[10] Weber, W., Thiel, W.: Theor. Chem. Acc. 103, 495 (2000)
[11] Koslowski, A., Beck, M.E., Thiel, W.: J. Comput. Chem, 24, 714 (2003)
[12] Van der Horst, M.A. and Hellingwerf, K.J. Acc. Chem. REs. 37, 13 (2004)
[13] Spudich, J.L. and Yang, C., Jung, K. and Spudich, E.N.: Annu. Rev. Cell Dev. Biol. 16, 365 (2000)

4.5 Models and Simulation

4.5.1 Parallel Tetrahedral Refinement Strategy Using Locally Based Object-Namespaces

Project coordinator	Prof. Dr. Burkhard Monien, PC², University of Paderborn
	Dr. Stephan Blazy, PC², University of Paderborn
Project members	Oliver Marquardt, PC², University of Paderborn
Work supported by	German Science Foundation (DFG) – SFB 376

General Problem Description

The numerical simulation of industrial applications is a very important appliance in the field of engineering. In chemical or physical applications simulations can help to get a better understanding of basic or complex processes and local phenomena. Studying these processes and phenomena with high precision requieres the usage of extreme computation power [2, 4]. Parallel computer resources, like cluster systems of vector nodes or PC-based nodes, make it possible to investigate such computation-intensive problems.

To obtain meaningful solutions from simulation models the approximation involves a large number of unknowns for the discretization of these processes and phenomena. The number of unknowns is typically limited by the amount of memory per compute-node. Thus, it is important to distribute positions of unknowns to areas in the discretized simulation domain where they are needed to achieve high solution precision with minimal error influence. The process of increasing or decreasing the number of unknowns in locally bounded areas of the simulation domain is called *adaptation process* and requires a modification of the discretization.

Efficient parallel computation depends on uniformly partitioning the discretization and distributing the sub-problems to each compute-node of the parallel system. During a simulation, increasing or decreasing the number of unknowns in some partitions through an adaptation process creates an imbalance, hence, some compute-nodes need more or less time to solve their sub-problems than other compute-nodes. To avoid such imbalance, because it increases the global computation run time, *workload balancing* strategies [7, 8] are applied to the distributed problem. These strategies decide which unknowns move to other partitions in order to achieve in a

uniformly distributed set of unknowns. Typically, a workload balancing phase is performed after each adaptation phase.
Numerical simulation environments identify a discretization point in a simulation domain by a globally unique number. If the adaptation process creates during the simulation new discretization points inside a partition, there exist a problem assigning a globally unique number to these points. A solution to this problem is to perform a global communication step to determine such a unique number or to map a fixed number interval to each partition from which the adaptation process can receive new identifiers. Both solutions are sub-optimal, because global communication steps are time-consuming and fixed number intervals can be too small if the adaptation process creates large numbers of unknowns. Vice versa, coarsening the discretization, i.e. reducing the number of unknowns, produces similar problems.

Our massively parallel numerical simulation environment *padfem²* [1] focuses on a completely different identifying scheme. All discretization points are identified by numbers which are locally defined and unique only in the partition they are assigned to. Moreover, all objects, the simulation domain is constructed from, like edges, faces and volumes, take their identifier numbers from these locally defined, object-specific namespaces. Because of this algorithm model, where a global identification of objects is not possible anymore, modifications to the simulation domain by the adaptation process require a new strategy.

Problem details and work done

The numerical simulation environment padfem² deals with discretizations of three dimensional domains using tetrahedral meshes. There exist several methods to adapt tetrahedral meshes, for instances bisection methods using the longest edge or irregular refinement with closure tetrahedra [3, 5, 6]. In padfem² a modified parallel irregular refinement method with four possible closure tetrahedra is used.

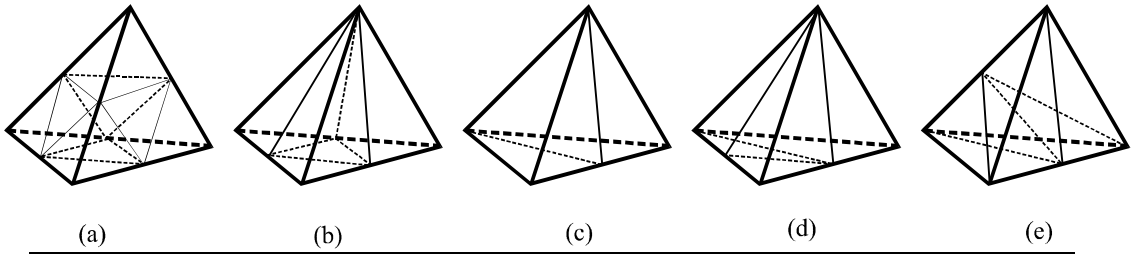

Figure 1: Regular refinement (a) and closure (b-e) patterns.

If the error estimator of the problem instance marks a tetrahedron to be refined, the tetrahedron has to be divided regularly into eight sub-tetrahedra, thus, introducing six new descretization points on the edges of the coarse tetrahedron (cf. Fig. 1a). Unfortunately, this refinement pattern creates so-called hanging objects, i.e. vertices, edges and faces which need special mathematical treatment. To avoid such hanging objects closure tetrahedra are applied to the mesh to eliminate them. There are 2^6-2 = 62 possible closure tetrahedra for each tetrahedron's refined edge combination. Because of symmetry and easy handling reasons, this number can be reduced to four closures (cf. Fig. 1b-e).

The regular refinement pattern is the only pattern which creates new discretization points. Applying it to the mesh can trigger a domino effect of re-applying this pattern to other tetrahedra not marked by the error estimator. The reason for this can be found in the refinement rule using the regular refinement pattern if three refined edges of a tetrahedron do not reside completely on a face (cf. Fig. 1b) or the tetrahedron has more than three refined edges. This is the price for the reduced number of possible closure tetrahedra. Therefore, the complete refinement phase consists of three steps: firstly, applying the regular refinement pattern to marked tetrahedra, secondly, identifing and resolving the domino wave through the mesh by applying also the regular refinement pattern, and finally, eliminating all hanging objects by applying the closure patterns. In this three-step algorithm step one and three can be done in parallel without communication between neighboring partitions, because these steps are independent and modify the mesh only locally. Before, during and after the second step, information about refined edges on partition boundaries have to be exchanged between neighboring partitions to maintain consistency of the distributed mesh. During the reporting period we designed, implemented and tested an iterative two-step algorithm for the partition border adjustment inside the domino wave resolving phase using locally based object-namespaces.

The key feature of locally based object-namespaces is an efficient mapping of partition border object identifiers shared between neighboring partitions. Each partition must know the identifier number of a shared partition border object from its neighbors; faces are shared exactly between two partitions, vertices and edges at least between two partitions. Our simulation environment padfem² uses a complex hash table mechanism per partition neighbor for this mapping. These tables have to be updated every time an object on the partition border is modified, i.e. new vertices are introduced when edges are refined or faces are divided. These modification information are collected and exchanged with neighbors at specific communication points in the algorithm.

The Two-Step Adjustment Algorithm

The two-step adjustment algorithm is executed in every iteration of the domino wave resolving to maintain mesh consistency and conformity. For the algorithm partition border edges are classified into two types. Type 1 edges are construction edges from the coarse tetrahedron and are divided into two sub-edges by the regular refinement pattern. Note, that there could be border edges which are not refined and, therefore, will be untouched by the algorithm. Type 2 edges are newly created edges on the old coarse tetrahedron's faces. Their construction vertices are from the newly inserted discretization points from type 1 edges.

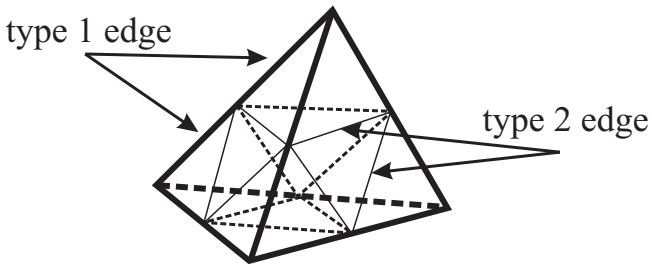

Figure 2: Two types of tetrahedron edges.

With this classification, the two-step adjustment algorithm performs as follows. In the setup phase, firstly, we determine for all partition border edges of all neighbors, if an edge is of type 1 or type 2, not refined edges will be ignored. The first step of the adjustment algorithm consists of collecting all local identifiers of edges of type 1, sending them to the neighboring partitions, receiving their type 1 edge identifiers from there and performing a local splitting of the associated local partition border edges from the remote edges. The mapping is done by lookup operations in the identifier hash tables of the neighbors. After that an identifier information exchange about newly created sub-edges of type 1 edges is needed. Hence, after the first step the algorithm propagates on the partition borders the splitting of coarse tetrahedra by the regular refinement rule. It is important, that partitions have to exchange the newly created identifiers between all neighboring partitions, because inconsistency might occur with edges on boundaries shared by more than two partitions and only one partition initiates the refinement of the edge. The other partitions must exchange their local identifiers also among each other. Step 1 includes two communication operations, i.e. exchanging which edges should be refined and exchanging newly created identifiers.

The second step performs the same exchange for type 2 edges. The previously determined type 2 edges from the setup phase are collected and sent to the neighboring partitions. After that the partition receives from its neighbors their type 2 edge identifiers. Until now, everything is fine, but there might be cases where an

adaptation front crosses a partition border for the first time and tetrahedra on the opposite border are not marked for refinement, which means type 2 edges are not created on that side and there is no feasible mapping of local identifiers to remote identifiers and vice versa. If such a situation occurs, the irresolvable type 2 edges have to be saved temporarily for the next adjustment operation and the associated remote tetrahedron is marked for refinement with the regular pattern. For this reason the two steps can not be combined to reduce the number of communication operations. After processing the second step, the adjustment algorithm finishes, and the resolving of the domino wave continues. All partitions determine tetrahedra with three or more refined edges, locally apply the regular refinement pattern and perform the next partition border adjustment. The number of iterations for the domino wave resolving is mainly affected by the problem-specific and time-dependent error estimator and the initial partitioning of the simulation domain.

A simple test case for the correctness of the parallel refinement algorithm is the high resolution determination of a surface of a geometric shape. Figure 3 shows a cross-section through a unit cube containing a sphere. The error estimator is designed to resolve the surface of this sphere with high precision. In the figure adaptation step 5 is depicted. It can be clearly seen that on the partition border (red shape) a high density of refined tetrahedra faces lies.

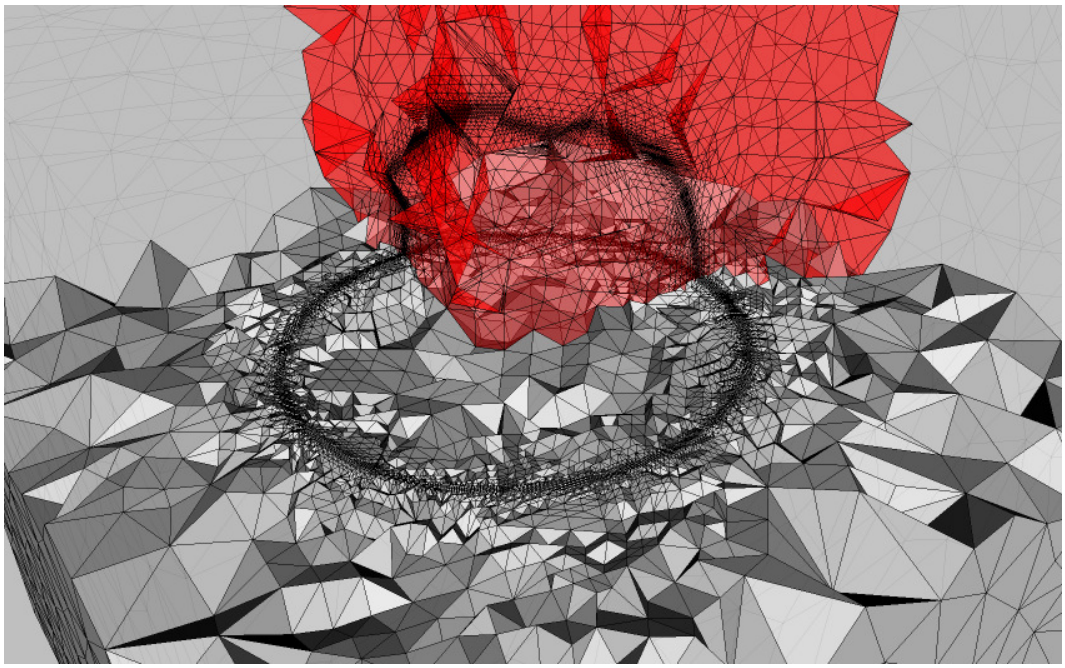

Figure 3: Cross-section (grey) of the simulation domain showing partition border of a partition (red).

Table 1 shows some statistic data for the previously described shape adaptation test case. Five adaptation steps were performed on a unit cube domain with 19837 tetrahedra. The domain was partitioned into 4 equally sized pieces and mapped onto a Quad-Opteron SMP system with 32 GByte memory.

	Time [sec.]	Domino Waves	Total Tetras	Closure Tetras
Step 1	0.469	11	40401	1508
Step 2	2.094	16	133262	8321
Step 3	7.332	14	467203	32839
Step 4	24.800	12	1726931	128143
Step 5	106.120	14	6662375	506091

Table 1: Adaptation statistics for the test case (sphere in unit cube)

Our further investigations focus on effective workload balancing strategies and the influence on numerical algorithms and methods while using locally based object-namespaces.

Resource Usage

The numerical simulation environment padfem² was designed for the efficient usage of massively parallel systems. Thus, it runs on all cluster systems of the PC² and can use all supported interconnection networks. Typically, the problems we solve with padfem² need a large amount of time for computation. Therefore, fast interconnects are preferred, mainly the Infiniband interconnect on the x64-based ARMINIUS cluster system and the SCI interconnect on the x86-based PSC2 using the Scali ScaMPI library on both systems. The problems are partitioned in that way, that they can be mapped via CCS on cluster partitions ranging in size from 2 to 128 compute-nodes. Time-dependent simulations require also a large amount of disk space for the simulation data sets. During the reporting period we used disk space from a few MBytes up to several hundreds of GBytes from the RAID disk array of the ARMINIUS system. For the analysis of these data sets the visualization environment consisting of the 3D immersive desk with passive tracking system and the visualization software Amira installed at the PC² were frequently used.

References

[1] S. Blazy, O. Kao, O. Marquardt: padfem² -- An Efficient, Comfortable Framework for Massivley Parallel FEM-Applications, Proc. of the European PVM/MPI User's Group Meeting (EuroPVM/MPI) 2003

[2] S. Blazy, O. Marquardt: A Characteristic Algorithm for the 3D Navier-Stokes Equation using padfem². Proc. of the 15th IASTED Intl. Conf. on Parallel and Distributed Computing and Systems (PDCS) 2003

[3] S. Blazy, O. Marquardt: Parallel Refinement of Tetrahedral Meshes on Distributed-Memory Machines. Proc. of the 23rd IASTED Intl. Conf. on Parallel and Distributed Computing and Networks (PDCN) 2005

[4] S. Blazy, O. Marquardt: Parallel Finite Element Computations of Three-Dimensional Benchmark Problems. Proc. of the 18th Symposium on Simulationtechniques (ASIM) 2005

[5] Bey, J.: Tetrahedral grid refinement, Computing 55, 1995, pp. 355-378.

[6] Castanos, J. G., Savage, J. E.: Parallel Refinement of Unstructured Meshes, Proc. IASTED Int. Conf. on Parallel and Distributed Computing and Systems, MIT, Boston, USA, 1999.

[7] Walshaw, C., Cross M., Everett M. G.: Parallel Dynamic Graph Partitioning for Adaptive Unstructured Meshes. J. Parallel Distributed Computing, 1997, pp. 102-108.

[8] Karypis, G., Kumar, V.: Parallel Multilevel k-way Partitioning Scheme for Irregular Graphs, Technical Report #36, Minneapolis, MN 55454, May 1996.

4.5.2 Theoretical and numerical Investigation of nonreactive and reactive fluid mixing in a T-shaped micro mixer

Project coordinator	Prof. Dr. Dieter Bothe, University of Paderborn Prof. Dr.-Ing. Hans-Joachim Warnecke, PC², University of Paderborn
Project members	Carsten Stemich, University of Paderborn B.Sc. A. Lojewski, University of Paderborn
Work supported by	Deutsche Forschungsgemeinschaft (DFG)

General Problem Description

The large area to volume ratio of micro devices gives prospect of better yield and selectivity than in conventional tubes, since diffusive fluxes of mass and heat scale with the area, while the rate of changes corresponding to sources and sinks are proportional to the volume. Indeed, theoretical investigations of the scaling behavior [1] support the fact that in relation to the time scale of reaction diffusive transport is faster than in conventional mixers. For that micro devices allow for fast chemical reaction and provide better thermal control. Today a number of reactions are performed in reactors, whose specific length scales are about some 100 µm [2, 3], and it is expected that in the future complete processes will take place in devices which are combined on a single chip (“Lab-/Fab-on-a-chip”). To tab the full potential of this new technology, a fundamental understanding of the transport processes on the relevant time and length scales and their interaction with chemical reactions is required. Furthermore, the mixing of chemical species is of special interest, because it is an essential condition for chemical reactions.

Since well-defined flow conditions can be adjusted preferentially in the laminar regime, this flow condition is of special interest. Typical residence times in micro channels are significantly smaller than in macro systems. Hence, to obtain a suitable mixture for these short residence times, the contact area between higher and lower concentrated regions has to be increased. The existing concentration gradients are transferred to smaller scales, where they are diffusively dissipated.

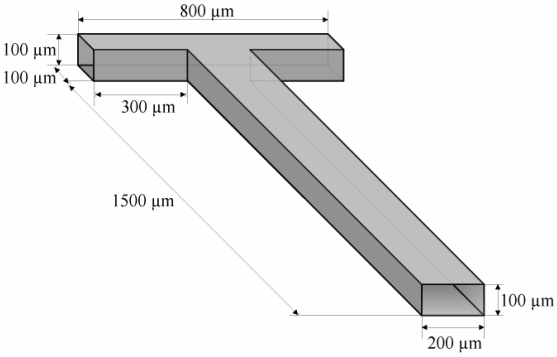

Figure 1: T-shaped micro mixer.

In this project the flow and mixing behavior of a T-shaped micro mixer with rectangular cross sections (figure 1) is investigated, as well as a neutralization reaction of the type $A + B \rightarrow C$. In the case of a quasi-instantaneous reaction, the time-scale of the reaction is much shorter than the time-scales of convection and diffusion. Hence, to avoid numerical difficulties, the simulations are based on the theoretical introduced simplification for diffusion controlled instantaneous reactions [4].

The calculations were carried out with the CFD program FLUENT6.2, which approximately solves the Navier-Stokes equations based on the Finite-Volume method. To evaluate the mixing behavior, the intensity of mixing I_{M} based on Danckwerts intensity of segregation [5] and a complemented characteristic length scale of segregation calculated from the potential for diffusive mixing Φ are used. The latter is closely related to the specific contact area and is efficiently computed employing a formula from geometric measure theory.

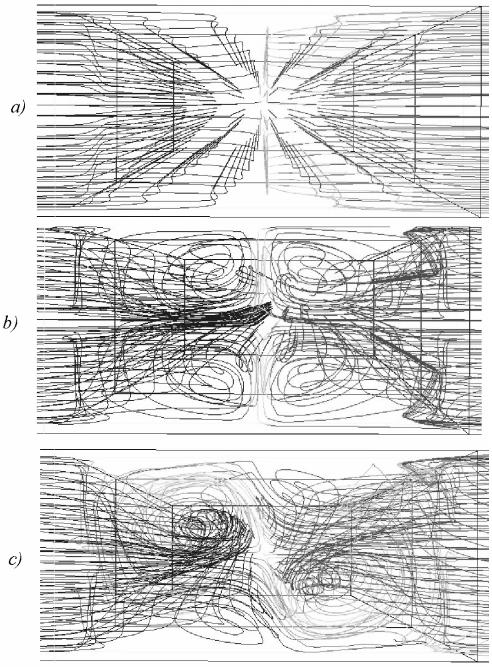

Figure 2: Path lines in the entrance of the mixing channel at average flow velocities of a) 0.1 m/s, b) 0.9 m/s, c) 1.1 m/s.

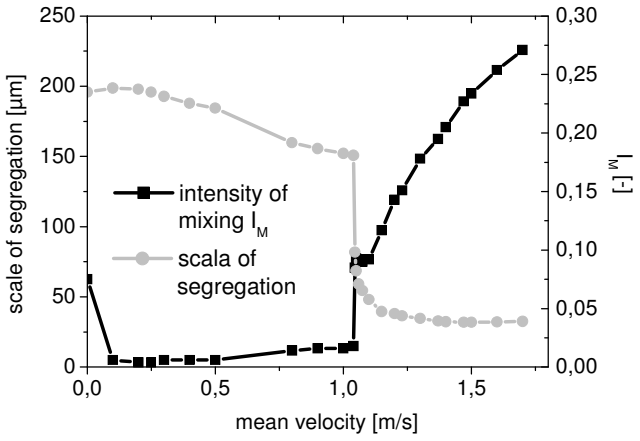

Figure 3: Intensity of mixing and scale of segregation versus mean velocity.

Problem Details and Work Done in the Reporting Period

Depending on the mean velocity, three different flow regimes (figure 2) can be observed for the analyzed Reynolds numbers (Re < 240) which are investigated concerning their mixing characteristics. In all three cases, both inlet streams attain the mixing zone (entrance area of the mixing channel), meet each other and flow into the mixing channel. In the first (laminar) regime, both inlet streams flow side by side through the mixing channel. The resulting species distribution consists of two completely segregated regions. The second case (vortex-regime) causes the same species distribution although a secondary flow type in the form of a double vortex pair is being formed. The third flow regime (engulfment-regime) which occurs at higher Reynolds numbers is more efficient for laminar mixing. A double vortex pair is formed as well, but the symmetry concerning the inlet channels is destroyed. Thereby fluid elements attain the opposing half of the mixing channel. With the twisting of both fluid streams, contact area is additionally produced which causes a decrease of the length scales of segregation and an increase of the mixing intensity.

The intensity of mixing is approximately zero for the first two flow regimes. For very slow mean velocities, diffusion is fast compared with the hydrodynamic residence time and the intensity of mixing is increased. By exceeding a mean velocity of 1.05 m/s the intensity of mixing increases noticeably (figure 3). Simultaneously the length scale of segregation decreases due to the intertwinement of the two vortex pairs from 200 µm, which correspond to the width of the mixing channel, to a value of 40 µm [6].

The hydrodynamics of liquid flow inside micro devices with typical channel widths of 100 µm is adequately described by the incompressible Navier-Stokes equations. For reacting flows, these equations have to be complemented by the species equation with the source/sink term modeling the chemical kinetics. In case of fast chemical reactions the latter term leads to several numerical difficulties, but for quasi-instantaneous reactions the simplification introduced theoretically by Toor [4] can be employed. For this purpose, only the difference of the two species concentrations is considered. The resulting equation has no source term and is of the form

$$\partial_t \phi + \mathbf{u} \cdot \nabla \phi = D(\phi) \Delta \phi \text{ with } \phi = c_A - c_B .$$

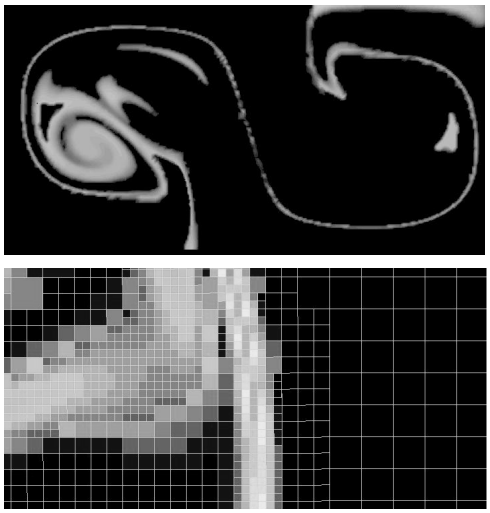

Figure 4: Simulated intensity of the indicator on a cross section at Re = 186 (top). Section with adaptively refined grid (bottom).

The simulations of the neutralization reaction were done on a locally refined grid with up to 18 million cubic grid cells and a resolution of down to 0.6 µm (figure 4 bottom) per cell to resolve the finest gradients which occur due to the high Schmidt-Numbers (300 for HCl and 470 for NaOH). Furthermore, a third order discretization scheme for momentum and species equation is used to avoid numerical diffusion. For the measurements of AK Raebiger (Institute of Environmental Process Engineering, University of Bremen), which are used for validation, a visualization of transport phenomena with a pH-sensitive indicator is being used. Therefore, this indicator is calculated as a further scalar. After obtaining a stationary situation, the intensity of the signal is calculated from the distribution of both scalars (figure 4 top).

Resource Usage at PC²

The numerical investigations of the T-shaped micro mixer with rectangular cross sections have been performed with a commercial CFD-Tool, FLUENT 6.2. The necessary resolution which requires a computational domain with up to 18 million cubic grid cells in the mixing channel can only be reached with high parallelization. Therefore, calculations with locally refined grids were done using the ARMINIUS cluster with up to 11 nodes.

References

[1] D. Bothe, C. Stemich, H.-J. Warnecke: Fluid mixing in a T-shaped micro mixer, *Chem. Eng. Sci*. 2006, in press.

[2] P. Löb, H. Löwe, V. Hessel: Fluorinations, chlorinations and brominations of organic compounds in micro reactors, *J. Fluorine Chem.* 2004, 125, 1677-94.

[3] M.-A. Schneider, T. Maeder, P. Ryser, F. Stoessel: A microreactor-based system for the study of fast exothermic reactions in liquid phase: characterization of the system, *Chem. Eng. J*. 2004, 101, 241-50.

[4] H.L. Toor, S.H. Chiang: Diffusion-controlled Chemical Reaction, *AIChE J.* 1959, 5 (3), 339-44.

[5] P.W. Danckwerts: The definition and measurement of some characteristics of mixtures, *Appl. Sci. Res*. 1952, A3, 279-96.

[6] D. Bothe, C. Stemich, H.-J. Warnecke: Theoretische und experimentelle Untersuchungen der Mischvorgänge in T-förmigen Mikroreaktoren – Teil I: Numerische Simulation und Beurteilung des Strömungsmischens, *Chemie Ingenieur Technik*, 76 (10), 2004, 1480-84.

4.5.3 Numerical Simulation of Fluid flow and heat transfer in Thermoplates

Project coordinator	Prof. Dr.-Ing. J. Mitrovic, University of Paderborn
Project members	Dipl.-Ing. Boban Maletic, University of Paderborn
	Dr. rer. Nat. H. Raach, Thermal Process Engineering and Plant Technology, University of Paderborn

General Problem Description

As heat transfer devices, thermoplates are encountered in several branches of engineering practice, e.g. as condensers or evaporators in thermal process technology and cooling technique. In comparison to shell-and-tube heat exchanges, their installation is relatively simple, and the periphery is drastically reduced.

A thermoplate consists of two metallic sheets, which are spot-welded over the whole surface according to an appropriate pattern, whereas the edges – except for connecting tubes – are continuously seam-welded.

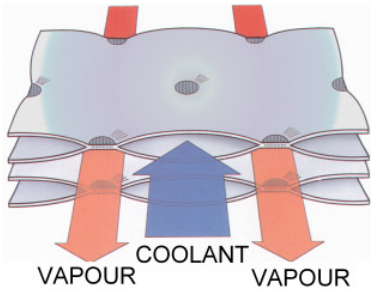

Figure 1: Fluid flow arrangement in thermoplates (DEG-Eng.).

By applying a hydro-form technique, a channel having a complex shape is established between the sheets (see Figure 1). One fluid is conducted through this channel, the other one through the channel created by two neighbouring thermoplates that are assembled in parallel at certain spacing thus making a thermoplate heat exchanger.

The objective of the numerical experiments is to numerically obtain the sets of the geometrical parameters that, in interaction with process parameters, should pave the way for optimal heat transfer of the inside fluid which is assumed to pass the thermoplate as a single phase (coolant in Figure 1). The geometry of the simulated three-dimensional domain is shown in Figure 2. It consists of a strip of a thermoplate channel, the latter being unbounded in spanwise (z) direction. The semi-circles represent the welding spots, which are arranged in a staggered manner

In the simulations, the geometrical parameters, such as the streamwise welding spots pitch, s_L, and the maximal distance between the metallic sheets, δ, have been varied, whereas both the transversal pitch, s_T, and the radius, *R*, of the welding spots were kept constant. To obtain some deeper insights into the flow and temperature fields in the developing region, a relatively short length, *L*, of the strip was chosen first. This length was then gradually increased up to $L = 500\ mm$, and, at low fluid velocity, the fully developed region was included into the simulation domain. The Reynolds number, *Re*, that is, the fluid inlet velocity, u_{IN}, and the distance, δ, have also been varied. The fluid inlet temperature, T_{IN}, and the wall temperature, T_W, were taken to be constant.

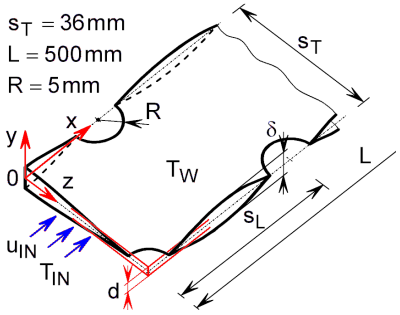

Figure 2: Geometry and dimensions of the three-dimensional simulated domain.

The fluid flow was considered to be laminar, incompressible, at steady-state and three-dimensional. Water of constant physical properties is adopted for the numerical experiments. The velocity and the temperature fields are governed by the equations of continuity, momentum and energy.

In the numerical simulations, the commercial software *StarCD* was employed. The calculations have largely been performed on the PC² system for parallel computing.

Problem details and work done

The velocity field

Figure 3 shows the velocity field in the *x*0*z* symmetry plane close to the channel inlet ($x < 147$ mm) at the smallest and at the largest Reynolds number Re adopted in the numerical experiments. At the Reynolds number of $\mathrm{Re} = 50$, the velocity field is largely smooth and the flow quiet with relatively narrow separation zones behind the welding spots. However, the velocity field is extremely heterogeneous, revealing a strong velocity variation both in transversal (*z*) and axial (*x*) direction, which ranges almost up to the factor of 4 with respect to the fluid inlet velocity. With the strict laminar flow, the fluid is detained in the recirculation zones, and the surface area corresponding to these zones is expected to be less effective for heat transfer.

At the Reynolds number, $\mathrm{Re} = 3800$, fluid separation establishes between the welding spots with reattachment at the contour of the neighbouring, downstream spots, and a comparatively large portion of channel is occupied by recirculation zones. These zones are responsible for local fluid acceleration. In the middle of the channel, there is a meandering fluid core that is bounded by the recirculation zones and that only touches the welding spots.

Heat transfer

The distribution of the heat flux as the vector intensity $|\dot{\mathbf{q}}|$ is illustrated in Figures 4. The heat flux is to decreases in streamwise (x) direction and is becoming smaller in the recirculation zones than in the central part of the strip. The heat flux fluctuates

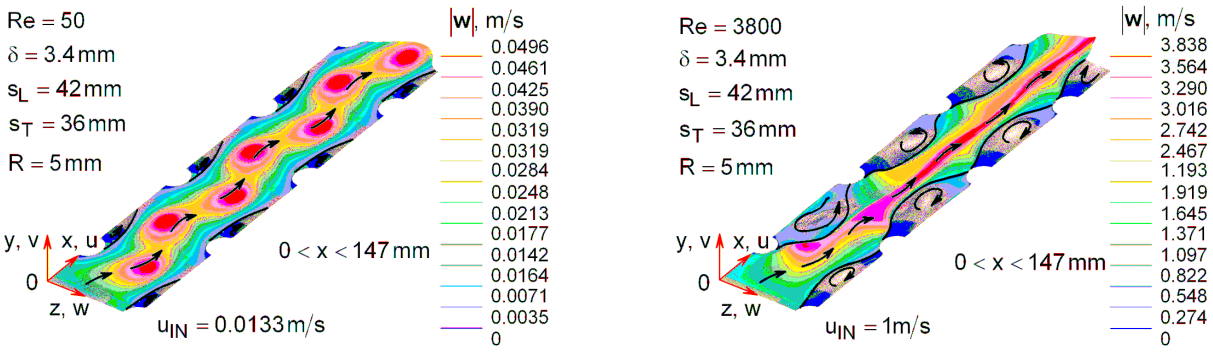

Figure 3: Velocity field in the plane $y = 0$ at $\mathrm{Re} = 50$ and $\mathrm{Re} = 3800$;

significantly not only in spanwise, but also in streamwise direction. The positions of the maxima of $\dot{q}(x)$ nearly coincide with the positions of the welding spots.

Conclusions

Thermoplates are efficient heat transfer devices, which are used in several branches of engineering practice. The complex geometry of the inside channel of such a plate provides the flowing fluid a pronounced three-dimensional character. Our comprehensive and at present not finished investigations aim at obtaining a better understanding of the inside fluid flow and heat transfer, in order to develop some relationships, which could serve as indicators when trying to optimise the channel geometry with respect to the thermo-fluid characteristic of the thermoplate.

Despite the fact that not all of the parameters have been varied, the results obtained are conclusive regarding the heat transfer potential of the thermoplate in comparison to a flat channel with plane walls. Depending on the geometrical and process related parameters, the heat transfer improvement in the thermoplate ranges nearly up to the factor of 4. Further numerical simulations should also provide information about the arrangement and shape of the welding spots and show a way, how to shape the thermoplate to optimise its heat transfer characteristic.

Acknowledgments

We would like to acknowledge the friendly support by the PC² team

4.5.4 Parallel Molecular Dynamics using Gromacs in Alzheimer research

Project coordinator	Prof. Dr. Gregor Fels, University of Paderborn
Project members	Jens Krüger, University of Paderborn

General Problem Description

In modern biochemical and medicinal chemistry research the employment of vast computer resourses is immanent in a broad variety of approaches. In our research work we are investigating processes related to communication between cells that is the molecular basis of how cells talk to each other instance. These processes are involved in, learning, recognition, and memory and are therefore closely related to the Alzheimer's disease. In this respect our research is focused on the so called cholinergic synapse, a communication switch board, at which nerve cells communicate by help of the neurotransmitter acetylcholine. The two most important proteins in this process are the acetylcholine receptor that receives the message by binding the neurotransmitter (Figure 1), and the enzyme acetylcholinesterase which is responsible for terminating the signaling process by cleaving the neurotransmitter acetylcholine at the enzyme's catalytic site (Figure 2). Understanding the functioning of both theses proteins on the molecular basis is a prerequisite for the development of a symptomatic Alzheimer therapy.

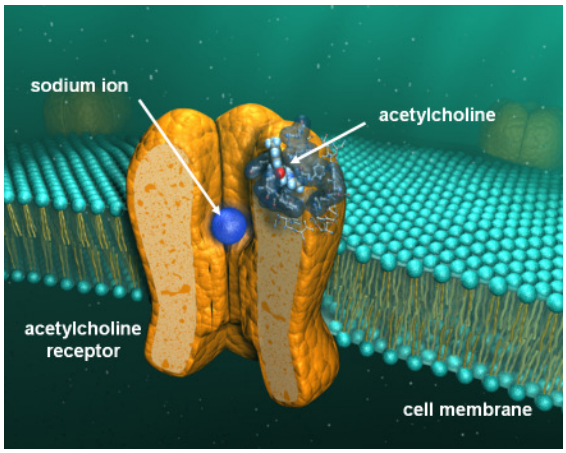

Figure 1: Nerve cell membrane with embedded acetyl-choline receptor

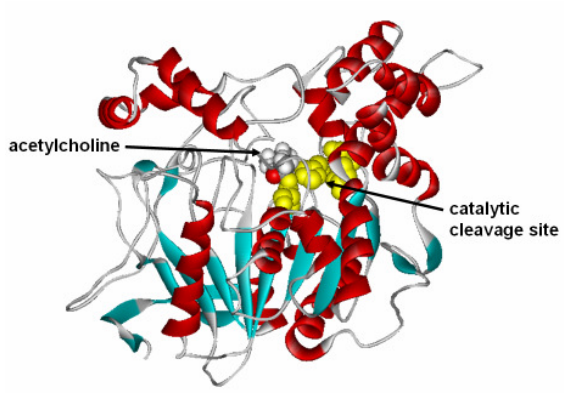

Figure 2: Acetylcholinesterase with acetylcholine at the enzyme cleavage site.

Perhaps even more important is to investigatie the appearance of aggregations of so called β-amyloid peptides at the cholinergic synapse of Alzheimer patients. This peptide is known to form plaques by packing multiple copies itself into fibrils, which then virtually block the synaptic function, and ultimately lead to the death of the

interacting nerve cells. Molecular processes that would allow preventing the plaque formation or perhaps would dissolve the plaques could eventually lead to a causal therapy of the Alzheimer's disease.

All these biological processes are in the focus of our research, in which we, among other methods, use computational chemistry techniques to study the dynamics and the function of the acetylcholine receptor protein, to analyze the binding of inhibitory drugs to the acetylcholinesterase, and to investigate the amyloid aggregation and the dissolution by small drug-like molecules.

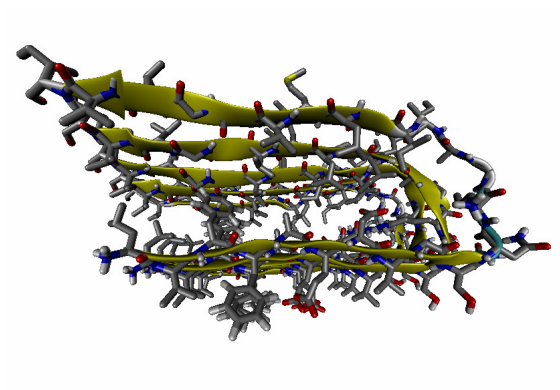

Figure 3: β-Amyloid fibril forming neurotoxic plaques

Many of these investigations only became possible because of the enormous development in computational techniques within the last few years, particularly it was the advert of computer clusters that allow computation of biological processes on a real life time scale. Without the powerful modern computer clusters we would be still bound to investigating static pictures of biological processes rather than the dynamics of the real life.

Problem details and work done

Since more than 30 years, the nicotinic acetylcholine receptor is in the focus of neuronal research. In the absence of crystallographic data, homology modeling has provided a comparatively clear static picture of the receptor structure [1]. In contrast, a dynamic view of this trans-membrane protein on a molecular level has only recently emerged from molecular dynamics calculation on the superb computational resources of the ARMINIUS-Cluster of the PC².

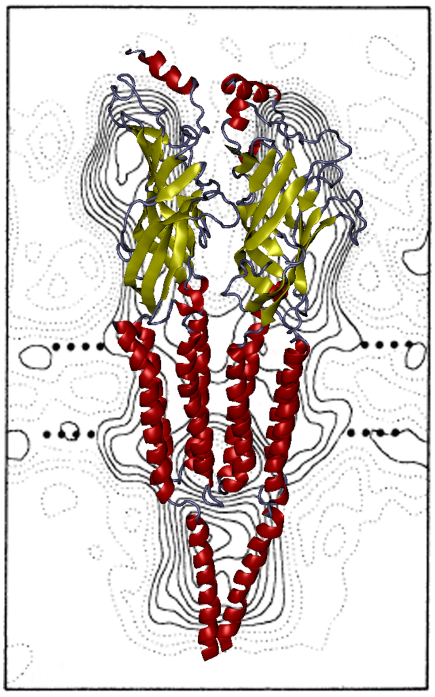

Figure 4: Nicotinic acetylcholine receptor with the extracellular protein chains (above dotted lines) responsible for ligand binding, the trans-membrane region (between dotted lines) that resemble the receptor integrated ion channel, and the intracellular receptor protein at the bottom of the picture.

The system size of over 150.000 atoms, all interacting with each other, had previously made it extremely difficult to handle such big systems. Now, the ARMINIUS-Cluster, allows us to compute multiples of 25 ns simulations of the fully solvated receptor protein embedded in a model membrane (“DPPC bilayer”). From theses calculations we gain knowledge of the asymmetric subunit movements, which are triggered by the binding of the neurotransmitter acetylcholine. These movements are responsible for opening the receptor integrated ion channel. Their computational characterization will ultimately allow us to fully understand the activation of the receptor and the opening of the ion channel by ligand binding. (See Figure 4)

In terms of our investigations of the acetylcholinesterase, we have concentrated on the investigation of the enzyme’s active site, with the potential of developing new drugs for the inhibition of this enzyme. [2, 3]. The binding behavior of a ligand when interacting with the enzyme can numerically be expressed as its affinity to the protein. Again, all the previous investigations of the binding affinities of acetylcholinesterase was restricted to comparatively static proteins which only insufficiently describes the conformational rearrangement that is necessary for the event of ligand binding. This process, known as induced fit, now is available for much more correct computation on computer clusters. We, therefore, have implemented the Linear Interaction Energy method (LIE) for the acetlycholineesterase on the ARMINIUS Cluster, and we were able to dramatically improve the calculation of

binding affinities of known enzyme-ligand systems [4]. As a consequence, we are now able to predict the affinities of new drugs - yet to be synthesized - more precisely and thereby can assist in devolping and screening of new drugs for acetylcholinesterase inhibition. The molecular dynamic method yields far superior results compared to classical rigid protein docking. As show from Figure 5, a predictive index of 0.92 can be reached, with the dynamic LIE-method demonstrating the good conformance of computationally derived data with experimental data [5].

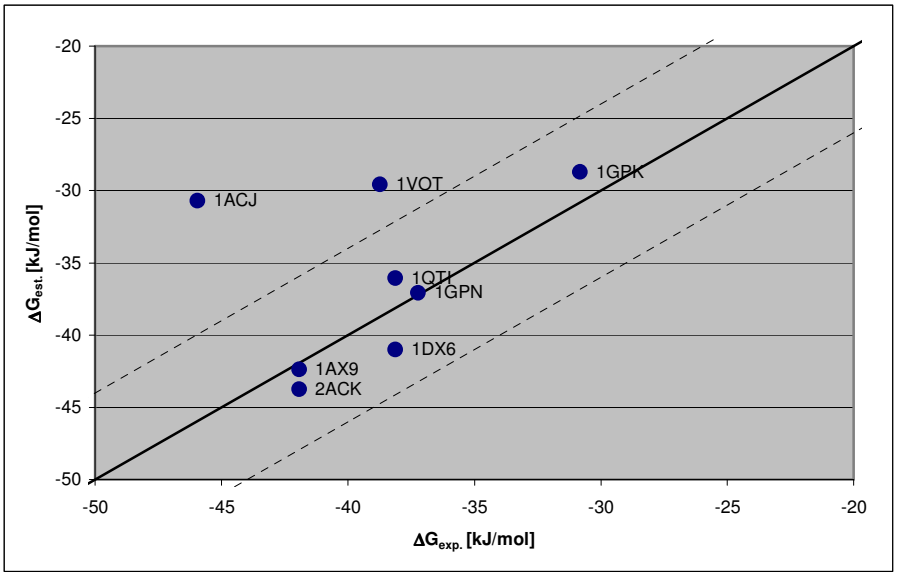

Figure 5: Prediction quality of molecular dynamics derived interaction energies. Points that are closer to the diagonal reflect a better result quality.

Resource Usage

Computational simulations of realistic biological systems are aiming at as much simulation time as possible in a minimum of real CPU-time. In real life, enzyme reactions, ligand recognitions and (at least some) channel gating, for instance, proceed on a submikro- to nanosecond time scale. While protein folding with its microsecond and longer time scales are still not accessible with today's computer power, many other biological processes are now within reach for computations on computer clusters. In this respect, the ARMINIUS-Cluster is not only standing out because of its massive CPU power, but, even more important, provides us with low latency Infiniband interconnects. In the software package GROMACS that we massively employ, we frequently have all-to-all node communications of rather small packages, so we benefit extraordinayly from this feature [6].

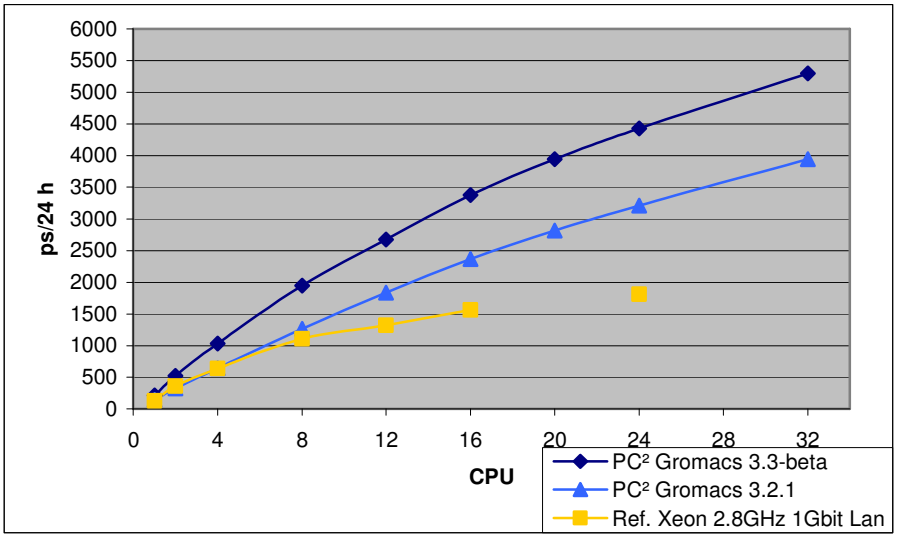

Figure 6: Performance and scaling behavior of a test simulation with different GROMACS versions (blue and light blue) in comparison with a so far leading system (yellow).

If we compare the scaling abilities of Gromacs on the ARMINIUS-Cluster with currently leading clusters (Fig. 6), a 2.5 fold gain in performance can be observed. To achieve maximum performance we usually use 6 to 12 nodes for one to six weeks, always depending on the system size and topology. Due to the allocation limits and recommended application checkpointing, we have to break our simulations in smaller portions and have to merge them at the end of the simulation. This leads to intense hard disk usage and we easily generate data amounts of 10 GB per simulation run.

References

[1] N. Unwin: Refined Structure of the Nicotinic Acetylcholine Receptor at 4 Å Resolution. J. Mol. Biol. (2005) *246*, 967-989.

[2] Pilger, C., Bartolucci, C., Lamba, D., Tropsha, A., and Fels,G. Accurate Prediction of the Bound Conformation of Galanthamine in the Active Site of Torpedo Californica Acetylcholinesterase Using Molecular Docking. J. Mol. Graph. Model., (2001), 19, 288-296

[3] L. Alisaraie, L.A. Haller, and G. Fels: A QXP-based Multi-Step Docking Procedure for Accurate Prediction of Protein-Ligand Complexes. J. Chem. Inf. Comp. Sci. (2006) in press

[4] M. Almlöf, B. Brandsdal and J. Åqvist: Binding Affinity Prediction with Different Force Fields: Examination of the Linear Interaction Energy Method. J. Comp. Chem. (2004) *25*, 1242-1254.

[5] D. Pearlmann and P. Charifson: Are Free Energy Calculations Useful in Practice? A Comparison with Rapid Scoring Functions for the p38 MAP Kinase Protein System. J. Med. Chem. (2001) *44*, 3417-3423.

[6] E. Lindahl, B. Hess and D. van der Spoel: GROMACS 3.0: A package for molecular simulation and trajectory analysis. J. Mol. Mod. (2001) *7*, 306-317

4.5.5 VOF-Simulation of Bubbles Rising in Shear Flow

Project coordinator	Prof. Dr. Dieter Bothe, CCES, RWTH Aachen Prof. Dr.-Ing. Hans-Joachim Warnecke, PC², University of Paderborn
Project members	Martin Schmidtke, University of Paderborn Michael Kröger, University of Paderborn

General Problem Description

Gas liquid flows appear in common contact apparatus in chemical or process engineering such as bubble columns, loop reactors, or aerated stirred vessels. Usually, the flow inside these reactors consists of a continuous liquid phase and a dispersed gas phase. For a valid up-scaling of such multiphase reactors, the relevant physico-chemical processes need to be well understood. As a complement to expensive experiments, numerical simulations offer a unique and cost-efficient chance for a detailed investigation of bubbly flows. Additionally, all relevant quantities such as velocity or pressure are completely accessible within the computational domain.

Numerical simulations of hydrodynamics and mass transfer are carried out with an extended version of the highly parallelized code FS3D (in cooperation with the ITLR Stuttgart), which employs an advanced Volume of Fluid (VOF) method. The high degree of parallelization of the code allows high resolution of the computational domain, such that the relevant length scales inside the liquid phase are resolved during the simulations. In previous works FS3D was used to simulate the free rise of bubbles in homogenous liquid flow without [1] and with mass transfer [2-4]. Recently, we considered the rise behavior of bubbles in shear flows [5].

Problem Details and Work Done in the Reported Period

In typical gas-liquid flows bubbles rise in non-homogenous flow fields. For example, in bubble columns, the liquid flows downwards near the walls, but upwards at the centre of the reactor. Hence the flow field in the ambient liquid exhibits a so-called shear flow. For optimization of bubbly flows in chemical reactors it is fundamental to understand how bubbles behave in a shear flow. To study this question, several experimental investigations have been carried out (e.g. [6,7]). These investigations reveal that small bubbles, which are nearly spherical in shape because of high surface tension forces, are pulled into the direction of faster counter-flow. Larger bubbles, which have a less stable shape, tend to drift into regions of slower counter-

flow during their rise. Until today, it has not been completely understood why large bubbles drift into opposite direction of small bubbles. Since numerical simulations give detailed velocity and pressure fields, they can be used to study this phenomenon. First investigations in this direction by Tomiyama [8] and Tryggvason [9] have shown opposite drift directions already in 2D computations. The advantage of two dimensional simulations is twofold: One point is the requirement of less computational power (so the grid resolution can be increased) and the other is a rather simple visualization of the obtained data.

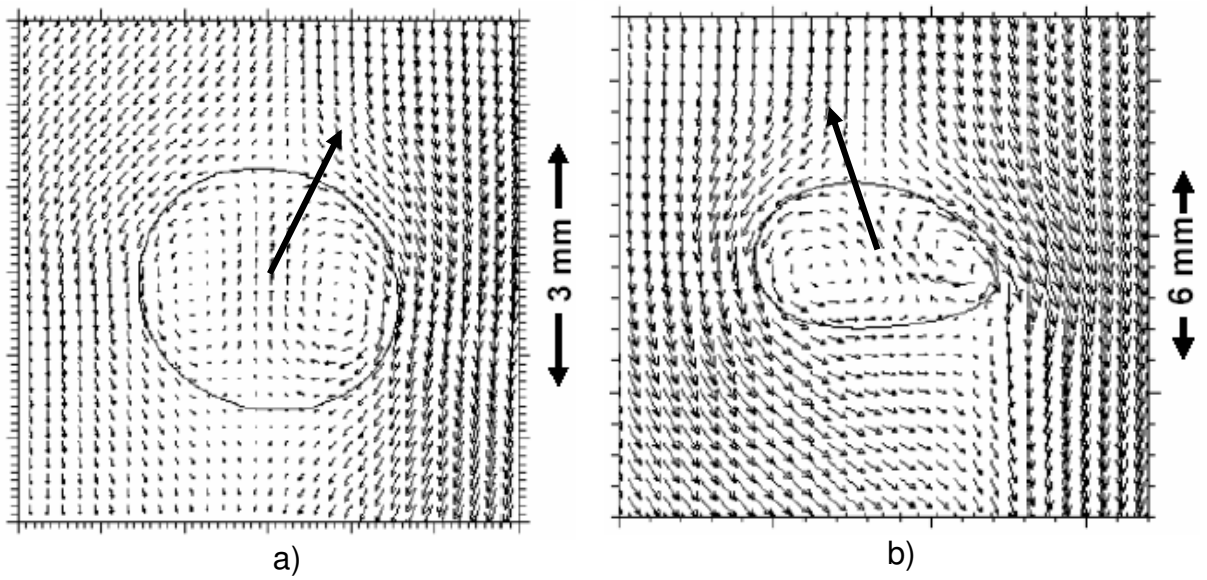

Figure 2: Velocity field of small bubble with equivalent diameter 3 mm (a) and large bubble with equivalent diameter 6 mm (b) in a 2-D simulation. Note the different length scales. Bold arrows indicate the direction of the relative bubble velocities.

Figures 2-4 show the results of 2D simulations for a small and a large air bubble in glycerol, computed with FS3D on the ARMINIUS cluster. Figure 2 shows a vector plot of the velocity fields. The vertical counter-flow increases with the horizontal displacement to the right. In this shear flow, the small bubble drifts to the right, which is the region of faster counter-flow (a). The large bubble drifts into the opposite direction (b). In the small bubble, the eddy on the right side is larger, whereas in the large bubble the left eddy is the bigger one. So, both bubbles drift into the direction of the larger eddy. This has also been observed by Tryggvason, who states that the direction of the lift force depends on the total bubble circulation [9].

Figure 3 shows the dynamic pressure fields around the bubbles. Due to the surface tension, the pressure inside the bubble is much higher. It exceeds the range displayed by the greyscale, so that the bubbles appear as black. Zones of low pressures are pale. They are situated behind the bubbles in an asymmetric manner.

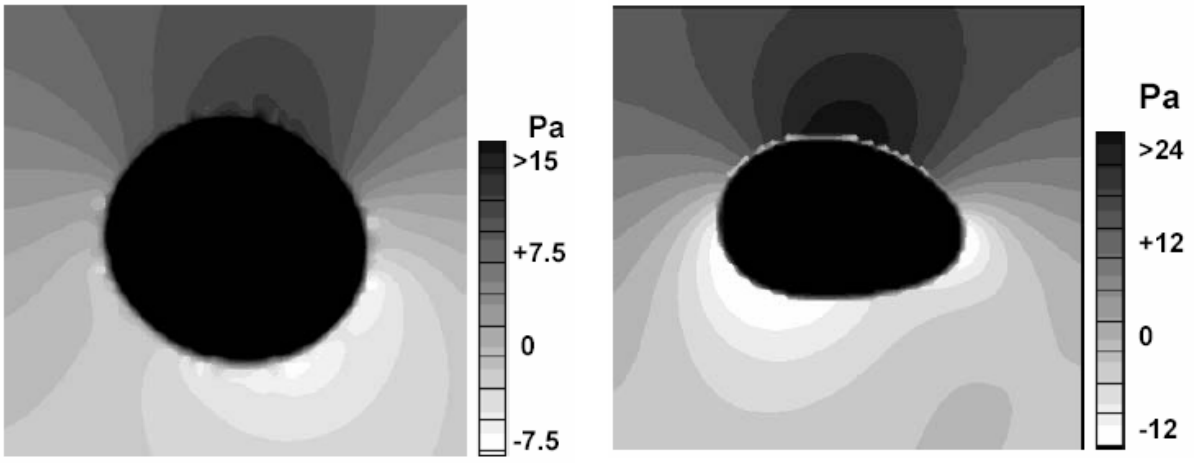

Figure 3: Dynamic pressure in liquid phase around the small bubble (a) and the large bubble (b).

When liquid passes bubbles, fluid elements near the bubble surface are accelerated. This causes low pressure due to the Bernoulli effect on both sides of the bubble. For spherical bubbles, this effect is stronger in the region of faster counter-flow. Large bubbles experience an asymmetric distortion in shear flows, so that the major depression occurs at the opposite side. Both bubbles drift into the direction of the highest depression.

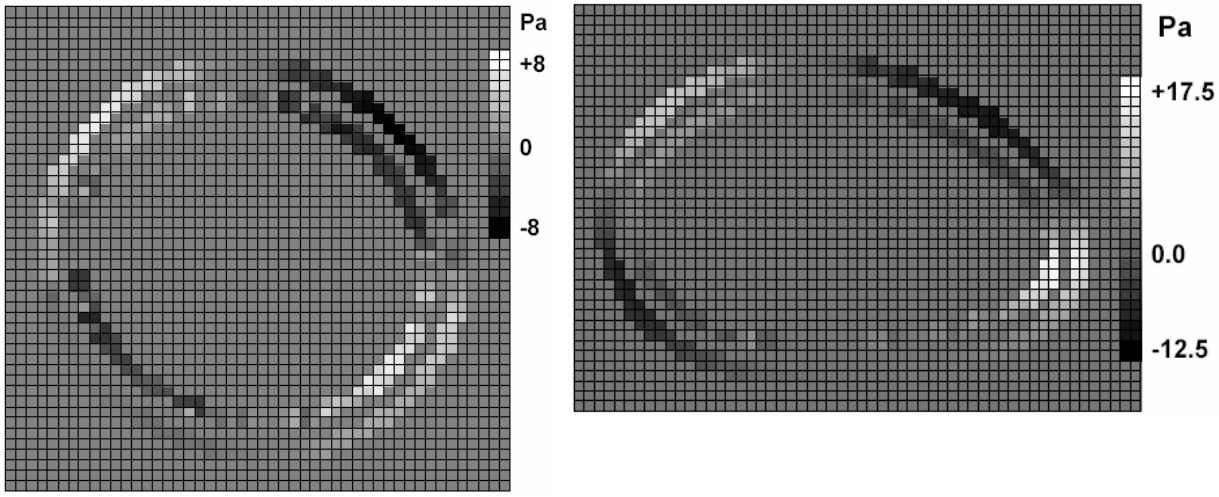

Figure 4: Horizontal component of dynamic pressure on bubble surfaces (outer curves), and horizontal component of shear stress on bubble surface (inner curves).

In addition to pressure, shear stress also has a significant contribution to lift force. The horizontal components of both forces (pressure and shear) on the bubble surface are displayed in Figure 4. Let $\mathbf{e}_y$ be the unit vector in the horizontal direction, then $p_y = p(-\mathbf{n})\cdot\mathbf{e}_y$ is the horizontal component of the dynamic pressure on a surface element with the outer normal vector **n**. In a similar way, the horizontal component of the shear stress is obtained. An integration of the horizontal contribution of the dynamic pressure over the bubble surface yields a force in the direction of the bubble migration in both cases. An integration of the shear stress over the bubble surface gives a force in the direction opposite to the bubble migration. Apparently, the dynamic pressure causes the lateral migration, whereas the shear stress acts as a frictional force against this movement.
Up to now, the contributions of dynamic pressure and shear stress to the lift force have been investigated for spherical bubbles (see Kurose et al., [10]). Our studies extend these considerations to the case of non-spherical bubbles. In the future, we aim to extend our work to bubbles in 3D.

Resource Usage at PC²

The simulations performed, required a very high resolution of the computational domain (1-16 mio. cells). On a single processor, the governing equations can be solved for a maximum number of 250.000 cells. Therefore, parallelization is necessary in order to compute large domains or to provide high resolution, respectively. The number of required CPUs varies from 2 to 64. The employed Volume of Fluid code FS3D (based on MPI) is completely parallelized and runs on the PSC2 and the ARMINIUS cluster.

References

[1] Koebe, M.; Bothe, D.; Prüss, J.; Warnecke, H.-J.: 3D Direct Numerical Simulation of Air Bubbles in Water at High Reynolds-Numbers. In: ASME FEDSM 2002, Montreal, Canada, 2002, pp. 1–8

[2] Bothe, D.; Koebe, M.; Wielage, K.; Warnecke, H.-J.: VOF-Simulations of Mass Transfer from Single Bubbles and Bubble Chains Rising in Aqueous Solutions. In: ASME FEDSM 2003, Honolulu, USA, 2003, pp. 1–7

[3] Bothe, D.; Koebe, M.; Warnecke, H.-J.: VOF-Simulation of the Rise Behavior of Single Air Bubbles with Oxygen Transfer to the Ambient Liquid. In: IBW2 Conference on Transport Phenomena with Moving Boundaries, VDI, 2003, pp. 1–13.

[4] Bothe, D.; Koebe, M.; Wielage, K.; Prüss, J.; Warnecke, H.-J.: Direct Numerical Simulation of Mass Transfer between Rising Gas Bubbles and

Water. In: Sommerfeld, M. (Ed.): Bubbly Flows - Analysis, Modelling and Calculation, Springer, Berlin, Heidelberg, New York, 2003

[5] Schmidtke, M.; Bothe, D.; Warnecke, H.J.: VOF-Simulation of the Rise Behaviour of Single Air Bubbles in Linear Shear Flows, in: Proc. 3rd Int. Berlin Workshop on Transport Phenomena with Moving Boundaries, 2005, pp. 89-100

[6] Tomiyama, A.; Tamaia, H.; Zun, I.; Hosokawaa, S.:Transverse migration of single bubbles in simple shear flows. In: Chemical Engineering Science 57 (2002), pp. 1849-1858

[7] Tomiyama, A.; Sou, A.; Zun, I.; Kanami, N.; Sakaguchi, T.: Effects of Eötvös Number and Dimensionless Liquid Volumetric Flux on Lateral Motion of a Bubble in a Laminar Duct Flow. In: Advances in Multiphase Flow 1995, pp. 3-15

[8] Tomiyama, A.; Zun, I.; Sou, A.; Sakaguchi, T.: Numerical analysis of bubble motion with the VOF method. In: Nuclear Engineering and Design 141 (1993), pp. 69-92

[9] Tryggvason, G.; Ervin, E. A.: The Rise of Bubble in a Vertical Shear Flow In: Journal of Fluids Engineering, Vol 119 (1997), pp. 443-449

[10] Kurose, R., Misumi, R., Komori, S.: Drag and lift forces acting on a spherical bubble in a linear shear flow, Int. J. Multiphase Flow 27, pp.1247-1258 (2001).

4.5.6 Shape Optimizing Load Balancing for Parallel Numerical Simulations

Project coordinator	Prof. Dr. Burkhard Monien, PC², University of Paderborn
Project members	Henning Meyerhenke, University of Paderborn Stefan Schamberger, University of Paderborn
Work supported by	DFG Research Training Group GK-693 DFG Collaborative Research Centre SFB 376

General Problem Description

Finite Element Methods (FEM) are very important in engineering for analyzing physical processes modelled by Partial Differential Equations (PDE). The domain on which the PDEs have to be solved is discretized into a mesh, and the PDEs are transformed into a set of equations defined on the mesh's elements (see e. g. [5]). Due to the sparseness of the discretization matrices these equations are typically solved by iterative methods such as Conjugate Gradient (CG) or multigrid. Since an accurate approximation of the original problem requires a very large amount of elements, this method has become a classical application for parallel computers. The parallelization of numerical simulation algorithms usually follows the Single-Program Multiple-Data paradigm: Each of the P processors executes the same code on a different part of the data. Thus, the mesh has to be split into sub-domains, each being assigned to one processor. To minimize the overall computation time, all processors should roughly contain the same amount of elements. Furthermore, since iterative solution algorithms perform mainly local operations, the parallel algorithm mostly requires communication at the partition boundaries. Hence, these should be as small as possible due to the very expensive communication costs involved.

The described problem can be expressed as a graph partitioning problem or, depending on the application, it may become a repartitioning problem because some areas of the simulation space might be coarsened and/or refined over time due to simulation dynamics unknown beforehand. This can cause an imbalance between the processor loads and therefore delay the simulation. To avoid this, the distribution of elements needs to be rebalanced by interrupting the application and solving the repartitioning problem. To keep the interruption as short as possible, it is necessary to find a new balanced partitioning with small boundaries quickly, with the additional objective not to cause too many elements to change their processor. Migrating elements can be an extremely costly operation since a lot of data has to be sent over communication links and reinserted into complex data structures.

In this work we present a heuristic addressing for the graph partitioning as well as the repartitioning problem. While existing approaches explicitly minimize the edge-cut, our heuristic applies a diffusive process inside a learning framework to determine well-shaped partition boundaries. This yields a low number of boundary vertices and therefore reduces the resulting communication volume. We apply an algebraic multigrid solver to compute the solution of the diffusive process with the advantage that its hierarchy only needs to be constructed once for each learning operation. In addition, we demonstrate that we can use this hierarchy to obtain better solutions with the same number of learning steps.

Problem details and work done

The learning framework used to obtain partitions with few boundary vertices is called *Bubble* [37]. It transfers the idea of the cluster analysis algorithm *k-means* [6] to graphs and starts with an initial, often randomly chosen vertex (seed) per partition. All sub-domains are then grown simultaneously in a breadth-first manner. Colliding parts form a common border and keep on growing along this border. After the whole graph has been covered and all vertices have been assigned to a partition this way, each component computes its new center, which acts as the seed in the next iteration. This is usually repeated until a stable state, where the movement of all seeds is small enough, is reached. Figure 1 illustrates the three main operations.

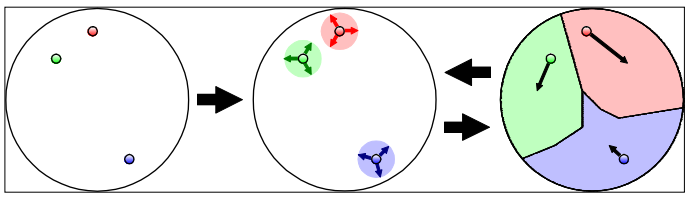

Figure 1: The three operations of the learning Bubble framework: Init: Determination of initial seeds for each partition (left). Grow: Growing around the seeds (middle). Move: Movement of the seeds to the partition centers (right).

These three steps can be implemented in several ways. Many approaches used previously show major disadvantages for our given problem (cf. [6] for a broader discussion). They can be overcome by the growth mechanism FOS/C based on disturbed diffusion [6]. It has the desired properties to send more load to densely connected regions of the graph and to put more load on vertices closer to the seed than on vertices further away. Load values are then interpreted as distances: the higher the load for some given vertex, the shorter its diffusion distance to the seed vertex. This load distribution is computed by solving the linear system $Lw = d$, where L is the Laplacian matrix of the unweighted, connected input graph G, d the drain

vector responsible for the disturbance of the diffusion and w the desired load vector (cf. [6] for more details). Note that d depends on which vertex is the current seed.

The resulting algorithm, which can either be invoked with or without a valid partitioning, is sketched in Figure . Note that it inhibits a large amount of parallelism, particularly when solving the P linear systems (one for each partition) simultaneously. These linear systems all have different drain vectors but the same system matrix. This would not be exploited if we solved them by using a conjugate gradient (CG) solver. Since the algebraic multigrid (AMG) hierarchy built for one linear system if can be reused for all others with the same matrix as well, we can expect superior run-times compared to CG for non-trivial system sizes.

```
Algorithm Bubble-FOS/C(G, π, l, i)
01 in each loop l
02   if π is undefined
03     π = determine-seeds(G)                    /* initial seeds */
04   else
05     parallel for each partition p             /* contraction */
06       solve L w_p = d_p and normalize w_p
07     π(v) = { p : w_p(v) ≥ w_p(u) ∀u ∈ V
               { −1 : otherwise
08   parallel for each partition p /* compute partitioning */
09     solve L w_p = d_p and normalize w_p
10   π(v) = p : w_p(v) ≥ w_q(v) ∀q ∈ {1, ..., P}
11   in each iteration i      /* optional consolidation with ...*/
12     parallel for each partition p
13       solve L w_p = d_p and normalize w_p
14     π(v) = p : w_p(v) ≥ w_q(v) ∀q ∈ {1, ..., P}
15     scale-balance(π)                  /* ... scale balancing */
16   greedy-balance(π)                   /* greedy balancing */
17 return smooth(π)                      /* smoothing */
```

Figure 2: Sketch of the refinement algorithm

If the initial partitioning is undefined or of low quality, many learning steps are required to find a good partitioning (as experienced e. g. in [104]). To avoid this, we adopt the multilevel scheme known from many graph partitioners. Rather than computing an additional hierarchy based on matchings, we use the existing AMG hierarchy. This is possible because each matrix in this hierarchy corresponds to a (possibly edge weighted) graph, and two graphs of consecutive levels have a similar structure. We perform Bubble-FOS/C as a refinement heuristic on each level. This reduces the number of required learning iterations and therefore the runtime considerably.

We have implemented our algorithm in C++ and parallelized the most time consuming parts – solving linear systems and constructing the AMG hierarchy – with POSIX threads. For the latter we use PMIS coarsening [140] to reduce the number of nodes in the next level substantially, so that the total number of created levels remains modest. In cases where PMIS coarsens too much, we neglect its result and apply a slightly modified CLJP coarsening [55] instead. Currently, we use a simple M-matrix interpolation from [141]. The algorithm then follows the multilevel paradigm by starting the computation on the lowest hierarchy level. On each level, the Bubble-FOS/C algorithm is applied and its partitioning result is interpolated to the next level according to the respective prolongation matrix P. All linear systems are solved by Full Multigrid V-cycles, which join the concept of nested iteration with V-cycles [147],

until the desired error tolerance is reached. A standard CG implementation serves as the direct solver on the lowest level inside the V-Cycle.

Experiments and Resource Usage

In this section we present some of our experiments executed on the two identical systems *ying* and *yang* of the PC². These SMP machines facilitate multithread parallelization, which is easier and faster to implement than parallel code for distributed-memory. They have four Opteron processors (2.2 GHz, 1 MB cache) each and run Linux (SMP-Kernel 2.4.21) as OS. The compiler we use is gcc 3.4.1 with level 2 optimization and multithread support. The included test set comprises eight FEM graphs with a modest number of vertices (from 9800 to 36476, for details cf. [102]). We restrict our presentation to 12-partitionings of two-dimensional FEM meshes.

Graph	CG				AMG			
	Time		Cut	Boundary	Time		Cut	Boundary
	1 cpu	4 cpu			1 cpu	4 cpu		
biplane9	61.15 s	26.04 s	774	1136	18.27 s	7.10 s	672	955
crack	15.93 s	5.57 s	1157	1142	3.98 s	1.59 s	1017	1004
crack (dual)	53.12 s	20.96 s	489	949	16.42 s	6.00 s	447	865
grid100x100	8.62 s	2.62 s	684	1000	5.54 s	2.26 s	575	949
stufe10	73.73 s	30.87 s	769	1156	22.83 s	8.73 s	574	725
shock9	138.53 s	54.73 s	1137	1673	40.41 s	15.21 s	961	1480
whitacker	12.20 s	4.62 s	984	970	3.68 s	1.60 s	966	957
whitacker (dual)	47.97 s	18.68 s	488	967	13.66 s	5.43 s	493	973

Table 1: Comparison between the solutions applying the CG solver without a learning hierarchy and the AMG approach with learning hierarchy.

Rather than using the traditional edge-cut metric, we list the total number of boundary vertices as a much more accurate measure of communication costs [3]. At first, we compare our new approach with a modified version of the heuristic from [104]. For the latter, we omit the extra vertex, reduce the number of learning steps and solve the linear systems by a standard CG implementation. Learning is only performed on the original graph, while for the AMG approach, we perform 1+level learning steps and consolidations, respectively. As can be seen from Table 1, the new approach is about three times faster on average than the old one. Moreover, in almost all cases it attains fewer cut-edges and boundary vertices. The current straightforward implementation without the use of neither scientific libraries nor processor bound threads achieves a speedup of about 2.5 on four CPUs.

The comparison with the state-of-the-art sequential libraries Metis and Jostle shows that these are significantly faster than our thread-parallelized algorithm (Table 1).

Yet, this run-time investment is supposed to pay off whenever partitionings of significantly higher quality with respect to the number of boundary vertices are found.

Graph	Metis			Jostle			Bubble-FOS/C		
	Time	Cut	Boundary	Time	Cut	Boundary	Time	Cut	Boundary
biplane9	0.03 s	670	1142	0.15 s	647	1104	7.10 s	672	955
crack	0.02 s	1041	1030	0.06 s	1031	1018	1.59 s	1017	1004
crack (dual)	0.02 s	466	919	0.08 s	450	893	6.00 s	447	865
grid100x100	0.03 s	584	1006	0.09 s	549	992	2.26 s	575	949
stufe10	0.02 s	570	948	0.15 s	546	919	8.73 s	574	725
shock9	0.07 s	1010	1663	0.19 s	909	1665	15.21 s	961	1480
whitacker	0.01 s	1005	992	0.11 s	966	953	1.60 s	966	957
whitacker (dual)	0.01 s	528	1048	0.11 s	515	1027	5.43 s	493	973

Table 2: Comparison of the solutions computed by Metis, Jostle and the shape optimizing approach using the AMG solver and the learning hierarchy.

Table 2 shows that the Bubble-FOS/C heuristic succeeds in most cases to produce comparable edge-cut values and – more importantly – better numbers of boundary vertices. While Jostle obtains fewer boundary vertices for the graph whitacker, our approach delivers the best results in all other displayed cases concerning this metric. The balance values of the partitionings are not shown explicitly because all partitioners stay within the predefined range of 3% imbalance.

We can therefore conclude that the presented heuristic for graph partitioning and load balancing computes high-quality partitionings w.r.t. the number of boundary vertices. Unfortunately, its current implementation requires a long run-time, although replacing former diffusion schemes by FOS/C and the CG solver by the Algebraic Multigrid Method result in a substantial speedup for solving of the linear systems involved. Furthermore, the constructed AMG hierarchy can be applied to improve the learning process in terms of time and quality without the need to compute and store a separate matching hierarchy.

References

[1] R. Diekmann, R. Preis, F. Schlimbach, and C. Walshaw. Shape-optimized mesh partitioning and load balancing for parallel adaptive FEM. J. Parallel Computing, 26:1555–1581, 2000.

[2] G. Fox, R. Williams, and P. Messina. Parallel Computing Works! Morgan Kaufmann, 1994.

[3] B. Hendrickson. Graph partitioning and parallel solvers: Has the emperor no clothes? In Irregular'98, number 1457 in LNCS, pages 218–225, 1998.

[4] V. E. Henson and U. Meier-Yang. BoomerAMG: A parallel algebraic multigrid solver and preconditioner. Appl. Numer. Math., 41(1):155–177, 2002.

[5] G. Karypis and V. Kumar. MeTis: A Software Package for Partitioning Unstrctured Graphs, Partitioning Meshes, [...], Version 4.0, 1998.

[6] J. B. MacQueen. Some methods for classification and analysis of multivariate observations. In Proc. of 5th Berkeley Symposium on Mathematical Statistics and Probability, pages 281–297. Berkeley, University of California Press, 1967.

[7] H. Meyerhenke, B. Monien, and S. Schamberger. Accelerating shape optimizing load balancing for parallel FEM simulations by algebraic multigrid. To appear in Proc. 20th IEEE International Parallel & Distributed Processing Symposium (IPDPS'06).

[8] H. Meyerhenke and S. Schamberger. Balancing parallel adaptive fem computations by solving systems of linear equations. In Proc. Euro-Par 2005 (LNCS 3648), pages 209–219, 2005.

[9] S. Schamberger. On partitioning FEM graphs using diffusion. In HPGC, Intern. Par. and Dist. Processing Symposium, IPDPS'04, page 277 (CD), 2004.

[10] S. Schamberger. A shape optimizing load distribution heuristic for parallel adaptive FEM computations. In Parallel Computing Technologies, PACT'05, number 2763 in LNCS, pages 263–277, 2005.

[11] H. D. Sterck, U. Meier-Yang, and J. Heys. Reducing complexity in parallel algebraic multigrid preconditioners. Technical Report UCRL-JRNL-206780, Lawrence Livermore National Laboratory, 2004.

[12] K. Stüben. An introduction to algebraic multigrid. In U. Trottenberg, C. W. Oosterlee, and A. Schüller, editors, Multigrid, pages 413–532. Academic Press, London, 2000. Appendix A.

[13] U. Trottenberg, C. W. Oosterlee, and A. Schüller. Multigrid. Academic Press, London, 2000.

[14] C. Walshaw. The parallel JOSTLE library user guide: Version 3.0, 2002.

4.5.7 Active Support of the Analysis of Material Flow Simulation in a Virtual Environment

Project coordinator	Prof. Dr. Friedhelm Meyer auf der Heide, HNI, University of Paderborn
Project members	Dr. Matthias Fischer, HNI, University of Paderborn Michael Kortenjan, HNI, University of Paderborn Jens Krokowski, HNI, University of Paderborn
Work supported by	DFG Research Training Group

General Problem Description

The coupling of visualization and simulation of highly detailed virtual scenes imparts the user an intuitive understanding of complex problems. The desire for high detailing demands efficient algorithms to render the scene in real time. We investigate algorithms to disburden the graphics hardware from rendering of hidden parts of the virtual scene. We apply our methods to virtual scenes of material flow simulations.

Problem details and work done

A system is developed, which interconnects simulation and 3D-visualization of production processes (see left fig.). The user can intervene actively in the simulation by actions in the 3D environment, and examine that way, in which effects his changes to the simulation will result. Thereby he is supported by the simulation, which automatically recognizes significant processes. The points, at which these processes take place, are particularly emphasized in the 3D-visualization. Beyond that, paths guiding the user through the virtual factory to these positions are suggested (see [WSC05]). The navigation of the user along such a path requires the representation of the 3D world in real time. Due to 3D models provided by CAD systems being highly detailed, methods to reduce the scene's complexity become necessary. Therefore, we developed a new data structure (Size Equivalent Cluster Trees), making it possible to reduce details of the visible models so far that displaying the entire scene becomes possible in realtime (see [AFG06]). This reduction of the models takes place depending on the position of the user and automatically increases the degree of detail in the environment of the user and at significant points (see middle and right fig.).

Actually, we extend the system such that the material flow simulator shall manage multiple parallel and time synchronous simulations The parallel execution of multiple

simulations should overcome the conflict of simulation run repetition for a good statistical basis and real-time immersive visualization. A side effect will be the reduction of time needed for simulation experiments. The planned simulation tool has the feature of triggered cloning, i.e., the user can decide during runtime to clone a set of simulations after changing parameters to preserve the original system. The simulations will be aggregated by visualization and statistics. The rendering is planned to overlay several simulations using effects like inking and transparency. Simulation data will be aggregated with statistical functions and diagrams. We use the ARMINIUS cluster for the implementation of parallel rendering algorithms and simulations. This work is a co-operation with the research group "Business Computing, esp. CIM" of Prof. Dr.-Ing. habil. Wilhelm Dangelmaier.

References

[1] Fischer, Matthias; Mueck, Bengt; Mahajan, Kiran; Kortenjan, Michael; Laroque, Christoph; Dangelmaier, Wilhelm: Multi-User Support And Motion Planning of Humans And Humans Driven Vehicles In Interactive 3D Material Flow Simulations In: Winter Simulation Conference (WSC´ 05), pp. 1921-1930, 2005.

[2] Kortenjan, Michael; Schomaker, Gunnar: Size Equivalent Cluster Trees - Realtime Rendering of Large Industrial Scenes In: 4th International Conference on Virtual Reality, Computer Graphics, Visualization and Interaction (Afrigraph 2006), pp. 107-116, 2006.

4.6 Parallel Computer Graphics & Multimedia

4.6.1 mLB -- Load Balancing Support in Heterogeneous Environments

Project coordinator	Prof. Dr. Odej Kao, PC², University of Paderborn
Project members	Sabina Rips, PC², University of Paderborn

General Problem Description

The complexity of CFD (Computational Fluid Dynamics) applications requires a high amount of computational power. In case of complex scenarios only parallel applications running on parallel machines enables an efficient simulation. The lack of availability of such systems as well as the desire for simulating much more complex problems can only be overcome by a platform as the Grid.

The heterogeneous characteristics of the Grid in view of the network as well as the processing elements require an appropriate load balancing for an efficient calculation. We developed the tool *mLB* that supports CFD applications concerning load balancing running on heterogeneous and dynamical environments such as the Grid.

Problem Details and Work Done

FEM (Finite Element Method) applications are an often used method for solving CFD problems. The simulation space is mapped to a mesh by partitioning the whole space into small units (cells). After some calculation steps, these cells have to exchange information, each with its neighboring cell. For an efficient calculation, all processors must reach these synchronization points at the same time. In order to achieve this the amount of cells must be suitably distributed over the processors.

During calculation the mesh changes and additional load is generated mostly on just a few of the processors. As this behavior is not predictable, the re-balancing of the load must be done during runtime.

Meta Load Balancer (*mLB*) [3] is used to support such applications. The following two topics are the main goals of *mLB*:

- to optimize the application's overall runtime
- to minimize overhead of load balancing itself

In order to optimize runtime, load must be distributed dependending on the processors' performance. *mLB* determines the available capacity of each processor and recommends an appropriate amount of load to the application's load balancer.

Minimizing the load balancing overhead is achieved by

- using only a part of the processors and
- doing load balancing between nodes with fast connections.

In order to achieve these goals, decisions are based on a hierarchical cluster structure. This structure is set up based on values gathered by a network analysis at runtime. Fast connected machines are represented by clusters at low levels whereas slow connections can be found at high levels (see figure 1).

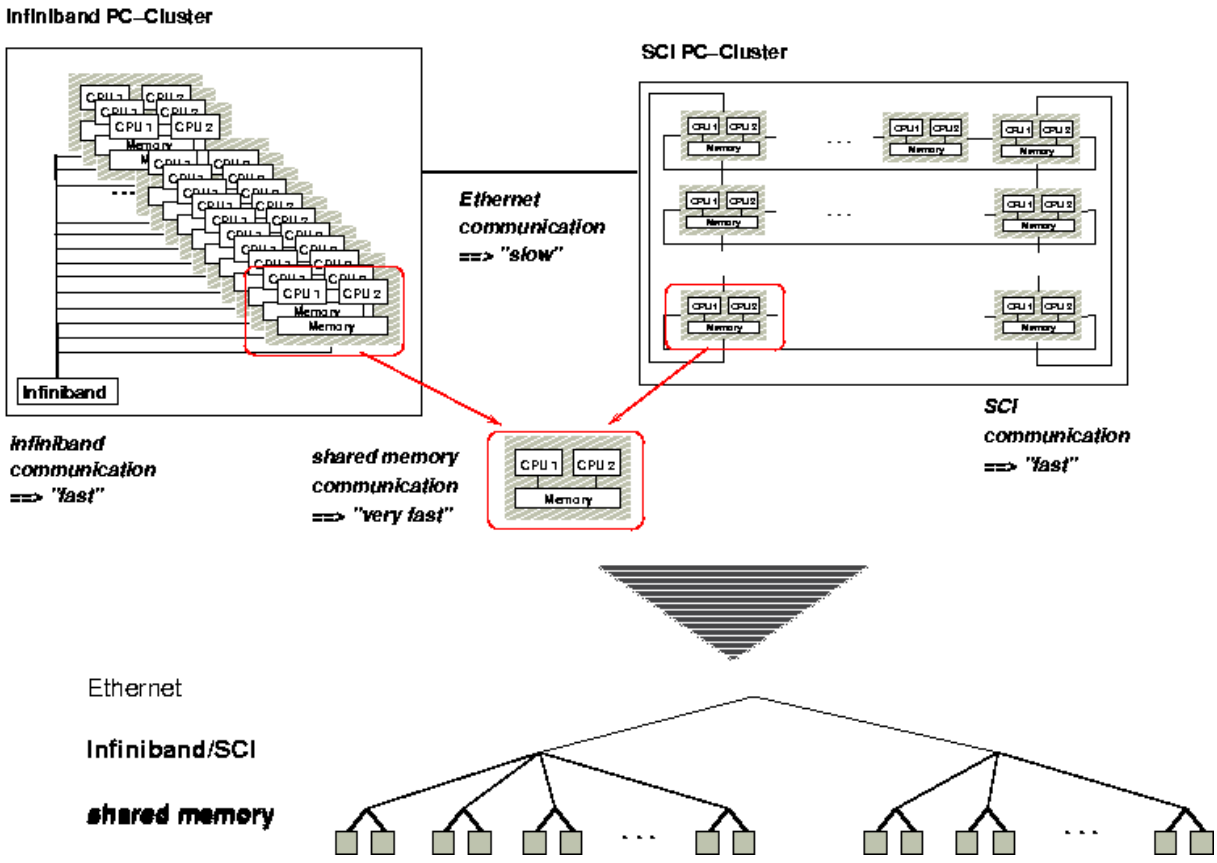

Figure1: Mapping the environment to a hierarchical cluster structure

When load balancing is initiated, it uses this structure to decide, which processors need to participate. If balance can be achieved by just re-balancing the load inside a cluster, only the cluster members are involved in this procedure. All other processors can do other work or, if necessary, do load re-distribution between one another, independently of other clusters. This mechanism avoids the usage of slow connections.

mLB determines *which processors* participate on load balancing and what is the *optimal load* of them. The decision *which part* of the load has to migrate to *which processor* is the task of a load balancing tool. The fact, that *mLB* is designed as a distributed tool without having a global instance, requires the usage of a parallel load balancing tool, such as Jostle [1] or ParMetis[2]. Furthermore, the tools must be able to accept user presettings as the proposed amount of load.

Integration into CFD Codes

mLB considers two areas. The first one is the network monitoring. This is done by using the MPI profiling interface. Whenever a collective MPI call is done, *mLB* measures the networks load. This is fully transparent for the application. The only action, the developer has to do is to link an extra library to his MPI application.
For load balancing support, the load balancing call inside the application has to be exchanged by an *mLB* routine. In order to simplify the integration, *mLB* uses an interface similar to as the load balancer Jostle. When load balancing must be done, *mLB* uses the information gathered by the network monitoring to decide which processors shall participate on load balancing.
Calculation capacities of the processors are measured inside the CFD application. Together with the current load this information is passed to *mLB*. *mLB* processes the optimal load and passes these data to the load balancer started on the involved processors only. After load balancing the application continues as usual.

mLB had been integrated into *padfem2* [4], a parallel FEM code, developed by PC². The aim was to evaluate the correct behavior of *mLB* and to investigate dynamic parameters that influence the efficiency of it.
We studied three main issues:

- Does the hierarchy detection work well?
- Is the load distributed proportionally to the processors' capacities? This means are the synchronization points reached by all processors at the same time?
- Is it feasible for the application to redistribute only parts of the mesh?

Our tests showed that our requirements had been satisfied. The hierarchical structure of the underlying network was detected. The load was distributed in the way that the idle waiting times, occurring when a processor reaches the synchronization points too early, were minimized.
One of our main questions in the reporting period was the last question. Is it really possible to pick only a part of the mesh and redistribute this part as it would be isolated from the rest? The application *padfem2* was able to handle it. After cell migration *padfem2* continued its run as usual.

Resource Usage

For testing, we used the hpcLine. Its different interconnects build an appropriate environment for testing our hierarchy discovery. When using ScaMPI, inter-node communication is done via Infiniband whereas intra-node communication is done via shared memory. The high compute power of the hpcLine enabled the testing of *mLB* with a real CFD application.
We used up to 40 hpcLine compute nodes and started two processes on each. The hierarchy detection worked and under went further optermization. For this test phase, the resources were used daily within a period of about three month.

References

[1] C. Walshaw, M. Cross, *Multilevel Mesh Partitioning for Heterogeneous Communication Network*. Future Generation Comput. Syst., 17(5): 601 – 623, 2001
[2] Kirk Schloegel, George Karypis, Vipin Kumar, *A Unified Algorithm for Load-balancing Adaptive Scientific Simulations*. Proceedings of Supercomputing 2000
[3] S. Rips, *Load Balancing Support for Grid-enabled Applications*. Proceedings of ParCo '05, September 2005.
[4] S. Blazy, O.Marquardt, *padfem2 – An Efficient, Comfortable Framework for Massively Parallel FEM-Applications*. Proc. Of the European PVM/MPI User's Group Meeting

4.6.2 Target Agreement VisSim

Project coordinator	Prof. Dr. Odej Kao, PC², University of Paderborn Prof. Dr. Jörg Wallaschek, HNI, University of Paderborn
Project members	Cord Bauch, L-LAB, University of Paderborn Jan Bersenbrügge, HNI, University of Paderborn Matthias Fischer, HNI, University of Paderborn Stefan Lietsch, PC², University of Paderborn Henning Zabel, C-LAB, University of Paderborn
Work supported by	Ministry of Innovation, Science, Research and Technology of the State of North Rhine-Westphalia

General Problem Description

In early 2005 the University of Paderborn and the Ministry of Innovation, Science, Research and Technology of North Rhine-Westphalia signed a target agreement [3] to support research projects and enable the university to form new, interdisciplinary facilities. Four different proposals were accepted. One of them is the project *VisSim* which main focus lays on distributed visualization and simulation.

The main competence of the PC², as a part of the *VisSim* group, is in High Performance Computing and Visualization. Therefore we develop concepts how to connect simulation and visualization efficiently and how to gain performance by distributing these tasks. In cooperation with the other members of the project *VisSim* (C-Lab, L-Lab, FG Gausemeier, FG Meyer auf der Heide, FG Rammig, FG Wallaschek,) we based our research on an existing application called Lightdriver. This is a driving simulator originally developed by the company Hella to realistically simulate new headlights which they develop. The second version of the Lightdriver called Virtual Night Drive [1] was developed by Jan Bersenbrügge of the FG Gausemeier. Its main innovation is the utilization of programmable vertex and pixel shaders to improve realism and performance of the application. The next step is to take the application to more complex scenarios like multi-user support and to further increase performance by distributing visualization and simulation. This also brings along the need for remote availability of the service since not every potential user has

the required computational power on his site. Therefore integration in the Grid environment is planned too.
A higher objective of the *VisSim* project is to generalize the results and conclusions gained through the sample application Virtual Night Drive and use them for universal concepts in the field of distributed visualization and simulation on HPC systems. This enables new possibilities regarding interactive simulation of realistic events such as flow simulation, rapid prototyping, or driving/flight simulations.

Problem details and work done

Code-Redesign

Before adding new features to the existing Virtual Night Drive a complete redesign of the source code had to be done. The original applications focus is on the realistic simulation of light. Therefore other components like the dynamic simulation or the interaction interface didn't have high priorities. Additionally the existing application based on Microsoft Windows. Since most of the cluster systems run Linux the source code had also to be ported to this Operating System.
The first step was to separate the visualization part of the Virtual Night Drive from the other parts, namely the dynamic simulation and the input handling. This leads to independent components which are able to run separately. Thereby for example simulations can be exchanged without interfering with other components of the system. Figure 1 shows the components of the system and their dependencies.

Concept for a flexible VisSim Architecture

To couple and synchronize all components and to provide unified interfaces for existing and additional simulations the Commuvit communication server [4] was integrated.

Visualization

The visualization bases on OpenSceneGraph [5] as the original Virtual Night Drive does. It is also extended by new features like switching between cars or having a global overview. Moreover we decided to utilize a system called Chromium to be more flexible in distributing and scaling the visualization. Chromium [2] is a framework to deliver OpenGL streams over networks. It supports IP-based networks but also faster ones like Infiniband i.e. the SDP-over-IB-Protocol. This is very sensible since OpenGL streams can grow very big and latency is a big problem in interactive visualization. Chromium offers different modules to transport and modify the OpenGL stream. The simplest way is to just send the stream to another computer and process it on its GPU. This can be used to route the graphical output to a desired

device. Chromium also offers more complex functionalities like generating stereoscopic images or tile the output for multi-segment displays. The application doesn't have to be changed in any way since Chromium acts as a substitute of the OpenGL library and works with (nearly) every OpenGL aware application.

Dynamic Simulation

To enable a realistic driving simulator a dynamic simulation had to be integrated into the Virtual Night Drive. The physical model of the car is simulated in real-time to have effects like suspension and a lifelike drivability. In the first version of the Virtual Night Drive this simulation was done by proprietary software called Vortex [8]. In the next step this software will be replaced by more powerful or open source systems like Matlab®/Simulink® [7] or Open Dynamics Engine™ [6]. Moreover since not only one but many cars need to be simulated possibilities to distribute the dynamic simulation need to be considered. The dynamic simulation and the visualization are connected through the Commuvit communication server mentioned above. Since the Commuvit offers a unified interface for the Dynamic Simulation all existing and to-be-developed systems can be used without changing the other components of the system (visualization, input etc.)

Input Devices

Simple mouse/keyboard input is not sufficient for a realistic driving simulator. Therefore a support for special input devices like steering wheels and pedals is needed. The Commuvit server offers an interface for the input devices and maps this input to the corresponding dynamics simulation. Depending on the setup of the system many different input devices connected to various computers can be used to control different cars. For the special requirements the L-Lab / Hella has, which is to test the headlights with real test people, realized even an integration of an original Smart car.

Other Simulations

In future revisions of the software other simulations are thinkable. For example the light simulation could be extended by features like adaptive light or driver supporting systems. These additional simulations could be connected through the Commuvit server and be distributed if needed.

Remote Visualization

For remote users that also want to see the visualization output of the cluster and possibly interact with simulations like the Virtual Night Drive a uniform access to the

resources is needed. Since many clients do neither have high graphical performance nor are they connected with very high bandwidth a system that can handle these limitations is needed. To tackle the first problem we decided to base our system on video streams so that the client only needs the ability to play these streams. The second problem can be solved by offering different steps of compression of these streams to be able to serve low bandwidths with lower quality streams and higher bandwidths with higher quality streams. By integrating this service into a Grid environment theoretically users from all over the world can access the Visualization resources of the ARMINIUS cluster in a uniform way.

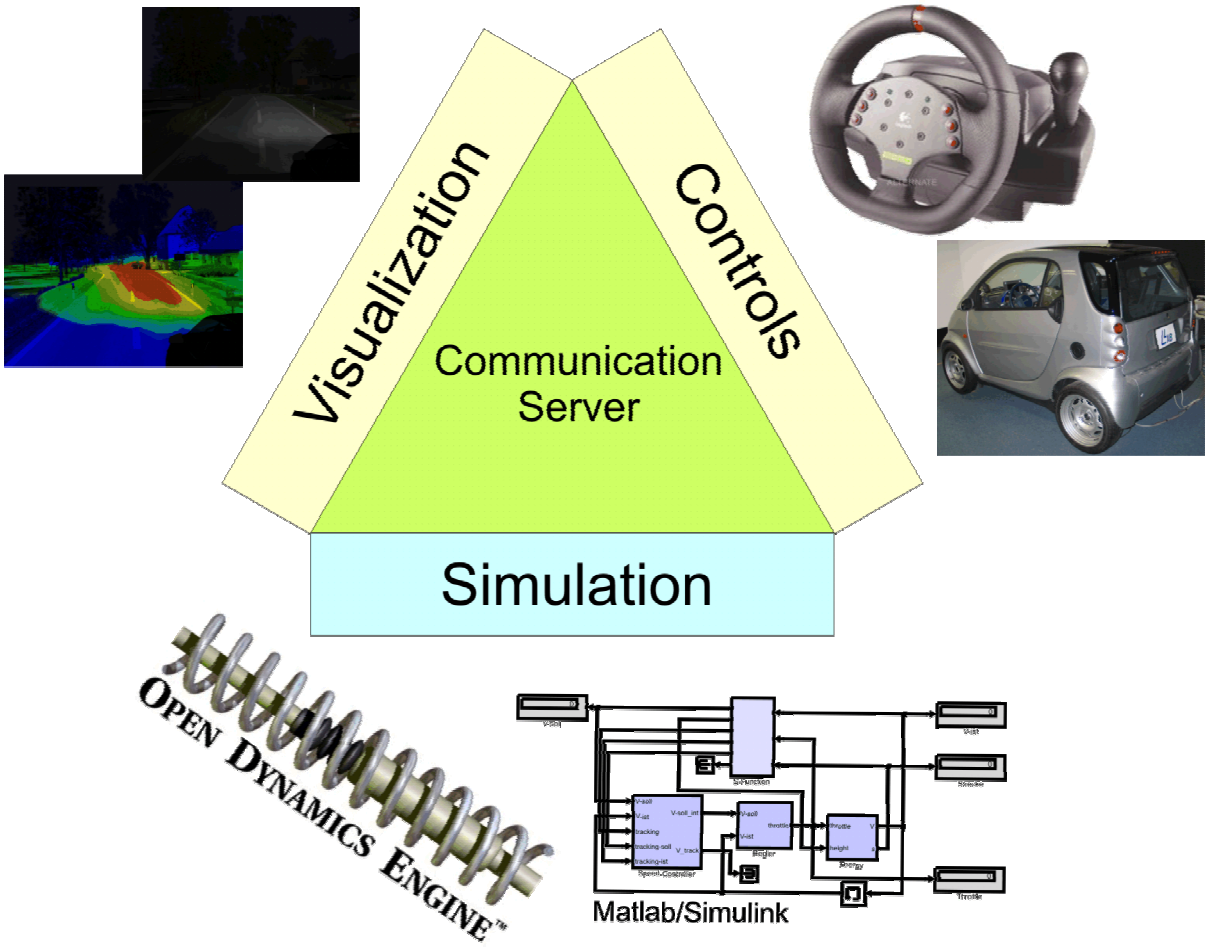

Figure 1 Architecture for distributed and interactive Visualization and Simulation

Outlook

As mentioned in the introduction the goal of this target agreement is to abstract the conclusions gained through the sample application Virtual Night Drive and develop a universal platform for simulation and visualization. This includes redesigning the communication server to provide universal interfaces and to better fit into the existing environment. Another topic is the realization of dynamically distributed visualization

as well as the integration of visualization resources into common Grid environments. Additionally plans for the organizational future of the target agreement need to be developed. This will also include finding new projects and partners from different backgrounds to establish the developed platform. The aim is to pave the way to distributed simulation and visualization for researchers whose primary objectives don't lies on computing but e.g. life sciences, business or medical science.

Resource Usage

The project *VisSim* mainly used the resources of the ARMINIUS Cluster of the PC² and its Visualization capabilities. Simulation and Commuvit ran on the compute nodes sequentially for the time being. In further versions the dynamic simulation will be distributed to support more complex multi-vehicle scenarios. The visualization components of the application massively take advantage of the specialized graphics hardware of ARMINIUS and the connected output devices. The big stereoscopic back projection wall which is driven by two of the visualization nodes in combination with the custom made driving seat offers an environment to realistically test Virtual Night Drive's features interactively. The other 6 visualization nodes are used to drive the multi-segment display to generate high-resolution images. Both settings can be realized by Chromium without special configuration of the application or by setting up the instances of the Virtual Night Drive visualization to synchronize over Commuvit.

References

[1] Gausemeier, Jürgen, Jan Berssenbrügge, Carsten Matysczok, and Klaus Pöhland: Real-time representation of complex lighting data in a Night Drive simulation. In Proceedings of the Immersive Projection Technology and Virtual Environments 2003, Zürich, 2003.

[2] Humphreys, Greg, Mike Houston, Ren Ng, Randall Frank, Sean Ahern, Peter D. Kirchner und James T. Klosowski: Chromium: a stream-processing framework for interactive rendering on clusters.j-TOG, 21(3):693–702, Juli 2002.

[3] Target Agreement between University of Paderborn and the Ministry of Innovation, Science, Research and Technology of the State of NRW (in german) http://www.innovation.nrw.de/Hochschulen_in_NRW/zielvereinbarungen/UniPaderborn_II.pdf

[4] Hompage Commuvit/C-Lab (Cooperation between University of Paderborn and Siemens Business Services GmbH & Co. OHG) http://www.c-lab.de/

[5] Homepage OpenSceneGraph http://www.openscenegraph.org/
[6] Homepage Open Dynamics Engine™ http://www.ode.org/
[7] Homepage Matlab®/Simulink® http://www.mathworks.com/
[8] Homepage Vortex http://www.cm-labs.com/

5 Summary of References (alphabetical order)

[1] A Quality-of-Service Architecture for Future Grid Computing Applications Proceedings of the 13th International Workshop on Parallel and Distributed Real-Time Systems, April 2004 (WPDRTS 2005)

[2] Alisaraie, L.; Haller, L.A. and Fels, G.: A QXP-based Multi-Step Docking Procedure for Accurate Prediction of Protein-Ligand Complexes. J. Chem. Inf. Comp. Sci. (2006) in press

[3] Almlöf, M.; Brandsdal, B. and Åqvist, J.: Binding Affinity Prediction with Different Force Fields: Examination of the Linear Interaction Energy Method. J. Comp. Chem. (2004) 25, 1242-1254.

[4] Althöfer, I.; Donninger, C.; Lorenz, U. and Rottmann, V.: On Timing, Permanent Brain and Human Intervention. In H. J. van den Herik, I. S. Herschberg, J. W. H. M. Uiterwijk (Eds). Advances in Computer Chess 7. University of Limburg. Maastricht (1994):285-296.

[5] Arena, graphical user interface (GUI) for chess engines:

[6] Bal et al, H.: Next Generation Grids 2: Requirements and Options for European Grids Research 2005-2010 and Beyond. ftp://ftp.cordis.lu/pub/ist/docs/ ngg2_eg_final.pdf, 2004.

[7] Berners-Lee, T.; Hendler, J. and Lassila, O.: The Semantic Web, Scientific American, May 2001

[8] Bey, J.: Tetrahedral grid refinement, Computing 55, 1995, pp. 355-378.

[9] Blazy, S.; Kao, O. and Marquardt, O.: padfem² -- An Efficient, Comfortable Framework for Massivley Parallel FEM-Applications, Proc. of the European PVM/MPI User's Group Meeting (EuroPVM/MPI) 2003

[10] Blazy, S. and Marquardt, O.: A Characteristic Algorithm for the 3D Navier-Stokes Equation using padfem². Proc. of the 15th IASTED Intl. Conf. on Parallel and Distributed Computing and Systems (PDCS) 2003

[11] Blazy, S. and Marquardt, O.: Parallel Refinement of Tetrahedral Meshes on Distributed-Memory Machines. Proc. of the 23rd IASTED Intl. Conf. on Parallel and Distributed Computing and Networks (PDCN) 2005

[12] Blazy, S. and Marquardt, O.: Parallel Finite Element Computations of Three-Dimensional Benchmark Problems. Proc. of the 18th Symposium on Simulationtechniques (ASIM) 2005

[13] Blazy, S. and Marquardt, O.: padfem2 – An Efficient, Comfortable Framework for Massively Parallel FEM-Applications. Proc. Of the European PVM/MPI User's Group Meeting

[14] Bloom, B.H.: Space/Time Trade-offs in Hash Coding with Allowable Errors., Communications of the ACM, 13:7, pages 422-426, 1970

[15] Blumofe, R.D.; Joerg, C.F., Kuszmaul, B.C., Leiserson, C.E.; Randall, K.H. and Zhou, Y. Cilk: An Efficient Multithreaded Runtime System. In Journal of Parallel and Distributed Computing. Vol. 37 No. 1 (1996):55-69.

[16] Bonorden, O.; Juurlink, B.H.H.; von Otte, I. and Rieping, I.: The Paderborn University BSP (PUB) library; Parallel Computing, 29(2), February 2003

[17] Bothe, D.; Koebe, M.; Wielage, K. and Warnecke, H.-J.: VOF-Simulations of Mass Transfer from Single Bubbles and Bubble Chains Rising in Aqueous Solutions. In: ASME FEDSM 2003, Honolulu, USA, 2003, pp. 1–7.

[18] Bothe, D.; Koebe, M. and Warnecke, H.-J.: VOF-Simulation of the Rise Behavior of Single Air Bubbles with Oxygen Transfer to the Ambient Liquid. In: IBW2 Conference on Transport Phenomena with Moving Boundaries, VDI, 2003, pp. 1–13

[19] Bothe, D.; Koebe, M.; Wielage, K.; Prüss, J. and Warnecke, H.-J.: Direct Numerical Simulation of Mass Transfer between Rising Gas Bubbles and Water. In: Sommerfeld, M. (Ed.): Bubbly Flows - Analysis, Modelling and Calculation, Springer, Berlin, Heidelberg, New York, 2003

[20] Bothe, D.; Stemich, C. and Warnecke, H.J.: Fluid mixing in a T-shaped micro mixer, Chem. Eng. Sci. 2006, in press.

[21] Bothe, D.; Stemich, C. and Warnecke, H.J.: Theoretische und experimentelle Untersuchungen der Mischvorgänge in T-förmigen Mikroreaktoren – Teil I: Numerische Simulation und Beurteilung des Strömungsmischens, Chemie Ingenieur Technik, 76 (10), 2004, 1480-84.

[22] Brickley, D. Guha, R.V.: RDF Vocabulary Description Language 1.0: RDF Schema, http://www.w3.org/TR/rdf-schema, 2004

[23] Brockington, M.G.: Asynchronous Parallel Game-Tree Search. PhD Thesis. University of Alberta. Edmonton 1998.

[24] Brooks, B.; Bruccoleri, R.; Olafson, B.; States, D.; Swaminathan, S.; Karplus, M. and Comp, J.: Chem. 4, 187 (1983)

[25] Bubenik, R. and Zwaenepoel, W.: Optimistic Make. In IEEE Transactions on Computers. Vol. 41 No. 2 (1992):207-217.

[26] Burchard, L.-O.: Et al. The Virtual Resource Manager: An Architecture for SLA-aware Resource Management. In 4th Intl. IEEE/ACM Intl. Symposium on Cluster Computing and the Grid (CCGrid) Chicago, USA, 2004.

[27] Castanos, J. G., and Savage, J. E.: Parallel Refinement of Unstructured Meshes, Proc. IASTED Int. Conf. on Parallel and Distributed Computing and Systems, MIT, Boston, USA, 1999.

[28] ChessBase: http://www.chessbase.com

[29] Condor. http://www.cs.wisc.edu/condor

[30] Crafty (Hyatt's home page): http://www.cis.uab.edu/info/faculty/hyatt/hyatt.html

[31] Cui, Q.; Elstner, M.; Kaxiras, E.; Frauenheim, Th.; and Karplus, M.: J. Phys. Chem. B 105, 569 (2001)

[32] Czajkowski, K.; Ferguson, D.; Foster, I.; Frey, J.; Graham, S.; Maguire, T.; Snelling, D. and Tuecke, S.: From Open Grid Services Infrastructure to WS-Resource Framework: Refactoring & Evolution, 2004

[33] Czajkowski, K.; Fitzgerald, S.; Foster, I. and Kesselman, C.: Grid Information Services for Distributed Resource Sharing, Proceedings of the Tenth IEEE International Symposium on High-Performance Distributed Computing (HPDC-10), 2001

[34] Dalpian G.M. and Wie, S.H.: Phys. Rev. B 72, 115201 (2005).

[35] Danckwerts, P.W.: The definition and measurement of some characteristics of mixtures, Appl. Sci. Res. 1952, A3, 279-96.

[36] DAT-Collaborative Website, http://www.datcollaborative.org

[37] Deep Shredder: http://www.shredderchess.com/shredderdeep.html

[38] Diekmann, R.; Preis, R.; Schlimbach, F. and Walshaw, C.: Shape-optimized mesh partitioning and load balancing for parallel adaptive FEM. J. Parallel Computing, 26:1555–1581, 2000.

[39] Elstner, M.; Porezag, D.; Jungnickel, G.; Elsner, J.; Haugk, M.; Frauenheim, Th.; Suhai, S. and Seifert, G.: Phys. Rev. B 58, 7260 (1998)

[40] Filhol, J.S.; Jones, R.; Shaw, M.J. and Briddon, P.R.: Appl. Phys. Letts. 84, 2841 (2004).

[41] Foster,I. and Kesselman, C.: (Eds.). The Grid 2: Blueprint for a New Computing Infrastructure. Morgan Kaufmann Publishers Inc. San Francisco, 2004

[42] Fox, G.; Williams, R. and Messina, P.: Parallel Computing Works! Morgan Kaufmann, 1994.

[43] Frauenheim, Th.; Seifert, G.; Elstner, M.; Hajnal, Z.; Jungnickel, G.; Porezag, D.; Suhai, S. and R. Scholz: Phys. stat. sol. (b) 217, 41 (2000)

[44] Frauenheim, Th. And Seifert, G.: Et. al, J. Phys. Cond. Matter 14, 3015 (2002).

[45] Fricke, H.: Ein beschleunigtes iteratives Bildrekonstruktionsverfahren für die Positronen-Emissions-Tomographie, Dissertation Hannover Medical School, 1999

[46] Fox, G.; Williams, R. and P. Messina. Parallel Computing Works! Morgan Kaufmann, 1994.

[47] Future Generation Grids, Springer, 2005

[48] Gausemeier, J.; Berssenbrügge, J.; Matysczok, C. and Pöhland, K.: Real-time representation of complex lighting data in a Night Drive simulation. In Proceedings of the Immersive Projection Technology and Virtual Environments 2003, Zürich, 2003.

[49] Grandinetti, L. (Edt.): Grid Computing: New Frontiers of High Performance Computing, pp. 185-201, Elsevier, 2005

[50] Graphs, Technical Report #36, Minneapolis, MN 55454, May 1996.

[51] Haarslev, V. and Möller, R. RACER User's Guide and Reference Manual Version 1.7.7, http://www.sts.tu-harburg.de/~r.f.moeller/racer/racer-manual-1-7-7.pdf, 2003

[52] Hayashi, S., Tajkhorshid, E.; Pebay-Peyroula, E.; Royant, A.; Landau, E.M.; Navarro, J. and Schulten, K. J.: Phys. Chem. B 105, 10124 (2001).

[53] Heine, F.; Hovestadt, M. and Kao, O.: Processing Complex RDF Queries over P2P Networks. Workshop on Information Retrieval in Peer-to-Peer-Networks P2PIR 2005, November 4, 2005

[54] Heine, F.: Scalable P2P based RDF Querying. In First International Conference on Scalable Information Systems (INFOSCALE06), to appear, 2006.

[55] Hendrickson, B.: Graph partitioning and parallel solvers: Has the emperor no clothes? In Irregular'98, number 1457 in LNCS, pages 218–225, 1998.

[56] Henson, V.E. and Meier-Yang, U.: BoomerAMG: A parallel algebraic multigrid solver and preconditioner. Appl. Numer. Math., 41(1):155–177, 2002.

[57] Highly Predictable Cluster for Internet-Grids (HPC4U), EU-funded project IST-511531. http://www.hpc4u.org.

[58] Himstedt, K.: An Optimistic Pondering Approach for Asynchronous Distributed Game-Tree Search. In ICGA Journal. Vol. 28 No. 2 (2005):77-90.

[59] Himstedt, K.: Verfahren zur Vermeidung redundanter Übersetzungen in modularen Softwaresystemen. Diplomarbeit im Fach Informatik. Universität Hamburg. Hamburg 1993.

[60] Holleman, A.F. and Wiberg, E.: Lehrbuch der Anorganischen Chemie (de Gruyter, Berlin, 1995), pp. 1081-1083.

[61] Hompage Commuvit / C-Lab (Cooperation between University of Paderborn and Siemens Business Services GmbH & Co. OHG) http://www.c-lab.de/

[62] Homepage General-Purpose Computation Using Graphics Hardware: http://www.gpgpu.org/

[63] Homepage IBM Cell Broadband Engine Research: http://www.research.ibm.com/cell/

[64] Hompage of the Heart and Diabetes Center North Rhine-Westphalia: www.hdz-nrw.de/[VND]

[65] Homepage OpenSceneGraph http://www.openscenegraph.org/

[66] Homepage Open Dynamics Engine™ http://www.ode.org/

[67] Homepage Matlab®/Simulink® http://www.mathworks.com/

[68] Homepage of the Project Group: http://www.upb.de/StaffWeb/maho/pg2004

[69] Homepage Vortex http://www.cm-labs.com/

[70] Hourahine, B. and Sanna, S.: Phisica B, in press (2006)

[71] http://www.chessbase.com/download/index.asp?cat=UCI%2DEngines

[72] http://www.hoise.com/primeur/05/articles/weekly/AE-PR-08-05-19.html

[73] http://www.hoise.com/primeur/05/articles/weekly/AE-PR-08-05-17.html

[74] Hydra: http://www.hydrachess.com

[75] http://www.innovation.nrw.de/Hochschulen_in_NRW/zielvereinbarungen/UniPaderborn_II.pdf

[76] http://www.microsoft.com/windowsserver2003/ccs/default.mspx

[77] http://www.mpi-forum.org/

[78] http://www.playwitharena.com/

[79] http://www.rrz.uni-hamburg.de/RRZ/e.index.html

[80] http://silverstorm.com/

[81] http://www.tim-mann.org/xboard.html

[82] http://www.tim-mann.org/xboard/engine-intf.html

[83] Humphreys, G.; Houston, M.; Ng, R.; Randall F.; Ahern, S.; Kirchner, P.D. and Klosowski, J.T.: Chromium A stream-processing framework for interactive rendering on clusters. j-TOG, 21(3):693–702, Juli 2002.

[84] Hyatt, R. Using Time Wisely. In ICCA Journal. Vol. 7 No. 1 (1984):4-9.

[85] Infrastructure. Morgan Kaufmann Publishers Inc. San Francisco, 2004

[86] Inmon, W.H.: Building the Data Warehouse, John Wiley & Sons, 2005

[87] Jens/Projekte/Benchmarks/Interconnects/mpi_pmb.htm

[88] Jens/Projekte/Benchmarks/Interconnects/infiniband.htm

[89] Jónsson, H.; Mills, G. and Jacobsen, K.W.: Classical and Quantum Dynamics in Condensed Phase Simulations (World Scientific, Singapore, 1998), chap. Nudged elastic band method fir finding minimum energy paths of transitions, pp. 387-405

[90] Kak, A.C. and Slaney, M.: Principles of Computerized Tomographic Imaging, Society of Industrial and Applied Mathematics, 2001

[91] Karypis, G. and Kumar, V.: MeTis: A Software Package for Partitioning Unstrctured Graphs, Partitioning Meshes, [...], Version 4.0, 1998.

[92] Karypis, G. and Kumar, V.: Parallel Multilevel k-way Partitioning Scheme for Irregular

[93] Knaup, J.M.; Köhler, C.; Frauenheim, TH.; Amkreutz, M.; Schiffels, P.; Schneider, B. and Hennemenn, O.-D.: In preparation (2006)

[94] Koebe, M.; Bothe, D.; Prüss, J. and Warnecke, H.-J.: 3D Direct Numerical Simulation of Air Bubbles in Water at High Reynolds-Numbers. In: ASME FEDSM 2002, Montreal, Canada, 2002, pp. 1–8.

[95] Kubiatowicz, J.: Extracting Guarantees from Chaos, Communications of the ACM, 46:2, pages 33-38, 2003

[96] Kung, H.T. and Robinson, J.T.: On Optimistic Methods for Concurrency Control. In ACM Transactions on Database Systems. Vol. 6 No. 2 (1981):213-226.

[97] Kurose, R.; Misumi, R. and Komori, S.: Drag and lift forces acting on a spherical bubble in a linear shear flow, Int. J. Multiphase Flow 27, pp.1247-1258 (2001).

[98] Kuszmaul, B.C.: The Startech Massively-Parallel Chess Program. In ICCA Journal. Vol. 18 No. 1 (1995):3-19.

[99] Lindahl, E.; Hess, B. and Van der Spoel, D.: GROMACS 3.0: A package for molecular simulation and trajectory analysis. J. Mol. Mod. (2001) 7, 306-317

[100] Löb, P.; Löwe, H. and Hessel, V.: Fluorinations, chlorinations and brominations of organic compounds in micro reactors, J. Fluorine Chem. 2004, 125, 1677-94.

[101] LSF - Load Sharing Facility. http://www.platform.com/products/wm/LSF/index.asp

[102] MacQueen, J.B.: Some methods for classification and analysis of multivariate observations. In Proc. of 5th Berkeley Symposium on Mathematical Statistics and Probability, pages 281–297. Berkeley, University of California Press, 1967.

[103] Manola, F. and Miller, E.: RDF Primer, http://www.w3.org/TR/rdf-primer, 2004

[104] Meyerhenke, H.; Monien, B. and Schamberger,S.: Accelerating shape optimizing load balancing for parallel FEM simulations by algebraic multigrid. To appear in Proc. 20th IEEE International Parallel & Distributed Processing Symposium (IPDPS'06).

[105] Meyerhenke, H. and Schamberger, S.: Balancing parallel adaptive fem computations by solving systems of linear equations. In Proc. Euro-Par 2005 (LNCS 3648), pages 209–219, 2005.

[106] Message Passing Interface (MPI) Forum Home Page:

[107] Moore, G.E.: Cramming More Components Onto Integrated Circuits, Electronics, 38(8), 1965

[108] Mpich2 Home Page: http://www-unix.mcs.anl.gov/mpi/mpich2/index.htm

[109] Newborn, M.: Unsynchronized Iteratively Deepening Parallel Alpha-Beta Search. In IEEE Transactions on Pattern Analysis and Machine Intelligence. Vol. 10 No. 5 (1988):687-694.

[110] PAL/CSS Online Freestyle Chess Tournament (held from May 28 to June 19, 2005 on www.playchess.com): http://www.computerschach.de

[111] PBS - Portable Batch System. http://www.openpbs.org

[112] Pearlmann, D. and Charifson, P.: Are Free Energy Calculations Useful in Practice? A Comparison with Rapid Scoring Functions for the p38 MAP Kinase Protein System. J. Med. Chem. (2001) 44, 3417-3423.

[113] Pilger, C.; Bartolucci, C.; Lamba, D.; Tropsha, A. and Fels, G.: Accurate Prediction of the Bound Conformation of Galanthamine in the Active Site of *Torpedo Californica* Acetylcholinesterase Using Molecular Docking. J. Mol. Graph. Model., (2001), *19*, 288-296

[114] PIRANHA: http://www.upb.de/pc2/projects/piranha

[115] Platzner, M.: Multi-Objective Evolution of Embedded Systems (MOVES)

[116] Miller, J. and Thompson, P.: Cartesian Genetic Programming

[117] Primeur weekly, The Paderborn hpcLine cluster: a marriage of Intel and AMD processors,

[118] Primeur weekly, How do you build a supercomputer? Together

[119] PSC2-, PLING-, SFB- Cluster. http://www.upb.de/services/systems

[120] Proceedings of the Parallel Computing Conference 2005, Malaga, Spain

[121] Provision of Fault Tolerance with Grid-enabled and SLA-aware Resource Management Systems

[122] Raman, R; Livny, M. and Solomon, M.H.: Matchmaking: Distributed Resource Management for High Throughput Computing, HPDC, 1998

[123] Regional Computer Centre, University of Hamburg Home Page:

[124] Ren, L.; Martin, C.H.; Wise K.J.; Gillespie, N.B.; Luecke, H.; Lanyi, J.K.; Spudich, J.L. and Birge, R.R.: Biochemistry 40, 13906 (2001).

[125] Rips, S.: Load Balancing Support for Grid-enabled Applications. Proceedings of ParCo '05, September 2005

[126] Rowstron, A. and Druschel, P.: Pastry: Scalable, decentralized object location and routing for large-scale peer-to-peer systems, Proc. of the 18th IFIP/ACM International Conference on Distributed Systems Platforms, 2001

[127] Sahai et al, A.: Specifying and Monitoring Guarantees in Commercial Grids through SLA. Technical Report HPL-2002-324, Internet Systems and Storage Laboratory, HP Laboratories Palo Alto, November 2002.

[128] Schamberger, S.: On partitioning FEM graphs using diffusion. In HPGC, Intern. Par. and Dist. Processing Symposium, IPDPS'04, page 277 (CD), 2004.

[129] Schamberger, S.: A shape optimizing load distribution heuristic for parallel adaptive FEM computations. In Parallel Computing Technologies, PACT'05, number 2763 in LNCS, pages 263–277, 2005.

[130] Schloegel, K.; Karypis, G. and Kumar, V.: A Unified Algorithm for Load-balancing Adaptive Scientific Simulations. Proceedings of Supercomputing 2000

[131] Schmidtke, M.; Bothe, D. and Warnecke, H.J.: VOF-Simulation of the Rise Behaviour of Single Air Bubbles in Linear Shear Flows, in: Proc. 3rd Int. Berlin Workshop on Transport Phenomena with Moving Boundaries, 2005, pp. 89-100

[132] Schneider, M.-A., Maeder, T.; Ryser, P. and Stoessel, F.: A microreactor-based system for the study of fast exothermic reactions in liquid phase: characterization of the system, Chem. Eng. J. 2004, 101, 241-50.

[133] Schulte, F.: Implementierung Evaluation eines Message–Passing-Layers unter Nutzung von Remote-DMA-Funktionen, Studienarbeit, Universität Paderborn, 2005

[134] Schumacher, T.: Diplomarbeit “Untersuchung von Kommunikationsmethoden zwischen FPGA-basierten Systems-on-Chip auf Basis von Message-Passing”, 2005

[135] SGE - Sun Grid Engine. http://www.sun.com/software/gridware

[136] SLA-aware Job Migration in Grid Environments

[137] Simon, J.: Low level InfiniBand performance, http://www.upb.de/StaffWeb/

[138] Simon, J.: MPI performance of interconnects, http://www.upb.de/StaffWeb/

[139] Steckl, A.J.; Heikenfeld, J.; Lee, D.S. and Gartner, M.: Mat. Science and Eing., B81, 97 (2001).

[140] Steinmetz, R. and Wehrle, K.: Peer-to-Peer Systems and Applications, Springer, LNCS 3485, 2005

[141] Sterck, H.D.; Meier-Yang, U. and Heys, J.: Reducing complexity in parallel algebraic multigrid preconditioners. Technical Report UCRL-JRNL-206780, Lawrence Livermore National Laboratory, 2004.

[142] Stüben, K.: An introduction to algebraic multigrid. In U. Trottenberg, C. W. Oosterlee, and A. Schüller, editors, Multigrid, pages 413–532. Academic Press, London, 2000. Appendix A.

[143] Straznicky, P.V.; Laliberté, J.F.; Poon, C. and Fahr, A.: Polymer Composites 21, 558 (2004).

[144] Sun Microsystems: Sun Grid Engine, http://wwws.sun.com/software/gridware Ullmann, J.R.: An Algorithm for Subgraph Isomorphism, Journal of the ACM, 23:1, pages 31-42, 1976

[145] Target Agreement between University of Paderborn and the Ministry of Innovation, Science, Research and Technology of the State of NRW (in german)

[146] The ARMINIUS Cluster at PC²:

http://www.upb.de/pc2/services/systems/ARMINIUS

[147] The Chess Monster Hydra: http://www.hydrachess.com

[148] The D-Grid web portal, http://www.d-grid.de

[149] The Globus Alliance, http://www.globus.org

[150] The Global Grid Forum, http://www.ggf.org

[151] The Globus Toolkit, http://www.globus.org

[152] The hpcLine at PC². http://www.upb.de/pc2/services/systems/psc

[153] The SFB Cluster at PC². http://www.upb.de/pc2/services/systems/ic

[154] The Virtual Resource Manager: Local Autonomy versus QoS Guarantees for Grid Applications

[155] Tomiyama, A.; Tamaia, H.; Zun, I. and Hosokawaa, S.:Transverse migration of single bubbles in simple shear flows. In: Chemical Engineering Science 57 (2002), pp. 1849-1858

[156] Tomiyama, A.; Sou, A.; Zun, I.; Kanami, N. and Sakaguchi: T.: Effects of Eötvös Number and Dimensionless Liquid Volumetric Flux on Lateral Motion of a Bubble in a Laminar Duct Flow. In: Advances in Multiphase Flow 1995, pp. 3-15.

[157] Tomiyama, A.; Zun, I.; Sou, A.; and Sakaguchi, T.: Numerical analysis of bubble motion with the VOF method. In Nuclear Engineering and Design 141 (1993), pp. 69-92.

[158] Toor, H.L. and Chiang, S.H.: Diffusion-controlled Chemical Reaction, AIChE J. 1959, 5 (3), 339-44.

[159] Top500 List, http://www.top500.org/

[160] Trottenberg, U.; Oosterlee, C.W. and Schüller, A.: Multigrid. Academic Press, London, 2000.

[161] Tryggvason, G. and Ervin, E. A.: The Rise of Bubble in a Vertical Shear Flow In: Journal of Fluids Engineering, Vol 119 (1997), pp. 443-449

[162] UNICORE Forum e.V., http://www.unicore.org

[163] UNICORE at SourceForge, http://unicore.sourceforge.net

[164] University of Wisconsin Madison: Condor, http://www.cs.wisc.edu/condor

[165] Unwin, N.: Refined Structure of the Nicotinic Acetylcholine Receptor at 4 Å Resolution. J. Mol. Biol. (2005) 246, 967-989.

[166] User Mode Linux, http://user-mode-linux.sorceforge.net

[167] Walder, H. and Platzner, M.: A Runtime Environment for Reconfigurable Operating Systems. In Proceedings 14th Int. Conf. on Field Programmable Logic and Applications (FPL), Belgium 2004, Springer

[168] Walder, H. and Platzner, M.: Reconfigurable Hardware Operating System: From Design Concepts to Realizations. In proceedings of 3rd Int. Conf. on Engineering of Reconfigurable Systems and Algorithms, Nevada, USA, June 2003, CSREA Press

[169] Walshaw, C.; Cross M. and Everett M. G.: Parallel Dynamic Graph Partitioning for Adaptive Unstructured Meshes. J. Parallel Distributed Computing, 1997, pp. 102-108.

[170] Walshaw, C.: The parallel JOSTLE library user guide: Version 3.0, 2002.

[171] Walshaw, C. and Cross, M.: Multilevel Mesh Partitioning for Heterogeneous Communication Network. Future Generation Comput. Syst., 17(5): 601 – 623, 2001

[172] Wanko, M.; Hoffmann, M.; Strodel, P.; Koslowski, A.; Thiel, W.; Neese, F.; Frauenheim, Th. and Elstner, M.: J. Phys. Chem. B im Druck

[173] Wefers, K. and Misra, C.: Tech. Rep. Oxides abd Hydroxides of Aluminum, Alcoa Laboratories (1987).

[174] XBoard and WinBoard graphical user interfaces for chess:

[175] xboard/WinBoard Chess Engine Communication Protocol: