Entwicklung integrierter IT-Infrastrukturen für die Simulation komplexer geophysikalischer Prozesse

Inaugural-Dissertation
zur Erlangung des Doktorgrades
der Fakultät für Geowissenschaften der
Ludwig-Maximilians-Universität München

vorgelegt von

Jens Oeser

am 17. September 2008

1. Gutachter: Prof. Dr. Hans-Peter Bunge

2. Gutachter: Prof. Dr. Heiner Igel

Tag der mündlichen Prüfung: 19.03.2009

Inhaltsverzeichnis

Verzeichnis der verwendeten Symbole, Einheiten und Abkürzungen

Die folgende Übersicht enthält eine Zusammenstellung der in dieser Arbeit mehrfach unter der gleichen Bedeutung verwendeten Symbole, Einheiten und Abkürzungen. Die Bedeutung für die hier nicht aufgeführten Symbole, Einheiten und Abkürzungen können dem entsprechenden Kontext entnommen werden.

Anstelle des Kommas für die Trennung von ganz- und bruchzahligem Anteil wird in dieser Arbeit in Anlehnung an die Rechnerausgabe der Dezimalpunkt benutzt.

BIOS	Basic Input Output System
BOOTP	Bootstrap-Protocol
CPU	Central Processing Unit (Prozessor)
DHCP	Dynamic Host Configuration Protocol
DNS	Domain Name System
ECC	Error Correction Code
FAI	Fully Automatic Installation
FSC	Fujitsu Siemens Computers
HDD	Hard Disc Drive (Festplatte)
HP	Hewlett-Packard
HPC	High Performance Computing (Hochleistungsrechnen)
IP	Internet Protocol
LDAP	Lightweight Directory Access Protocol
LMU	Ludwig-Maximilians-Universität
LRZ	Leibniz-Rechenzentrum der Bayerischen Akademie der Wissenschaften
MAC	Media Access Control
MWN	Münchner Wissenschaftsnetz
NFS	Network File System
PXE	Preboot eXecution Environment
RAM	Random Access Memory (Arbeitsspeicher)

SCSI	Small Computer System Interface
SMP	Symmetric Multiprocessing
TUM	Technische Universität München

Flop	Floating Point Operations
Flops	Flop pro Sekunde
GFlops	GigaFlop pro Sekunde
PFlops	PetaFlop pro Sekunde
TFlops	TeraFlop pro Sekunde
EFlops	ExaFlop pro Sekunde
kB	KiloByte
MB	MegaByte
GB	GigaByte
TB	TeraByte

P	Anzahl an Prozessoren
t	Laufzeit
$\bar{t}$	Mittelwert der Laufzeit
$t(P)$	Laufzeit für P Prozessoren
$\bar{t}(P)$	Mittelwert der Laufzeit für P Prozessoren
$S_{par}(P)$	paralleler Speedup für P Prozessoren
Ω	dreidimensionales Gebiet
$\partial\Omega$	Rand des Gebietes Ω
$\overline{\Omega}$	abgeschlossenes Gebiet ($\overline{\Omega} = \partial\Omega \cup \Omega$)
Γ_1	Randstück mit Randbedingungen 1. Art
Γ_2	Randstück mit Randbedingungen 2. Art
Γ_3	Randstück mit Randbedingungen 3. Art
$C(\overline{\Omega})$	Raum der in $\overline{\Omega}$ stetigen Funktionen
$C(\Omega)$	Raum der auf Ω stetigen Funktionen
$C^m(\overline{\Omega})$	Raum der auf $\overline{\Omega}$ *m*-mal stetig differenzierbaren Funktionen
$C^m(\Omega)$	Raum der in Ω *m*-mal stetig differenzierbaren Funktionen
$L_2(\Omega)$	Raum der über Ω quadratisch integrierbaren Funktionen
$H^m(\Omega)$	Raum der L_2-Funktionen, deren verallgemeinerte Ableitungen bis zur Ordnung *m* existieren und ebenfalls Element des Raumes $L_2(\Omega)$ sind

V	Raum der Test- und Ansatzfunktionen
V_0	Raum der Funktionen aus V, welche auf dem Randstück Γ_1 gleich Null sind
V_{g1}	Menge der Funktionen aus V, welche die Randbedingungen 1. Art erfüllen
V_h	endlichdimensionaler Teilraum von V
$a(.,.)$	Bilinearform
$l(.)$	Linearform
$\vec{v}$	Vektor
ϕ_i	Finite Elemente Ansatzfunktion
$\phi_\alpha^{(r)}$	Formfunktionen über dem Element $T^{(r)}$
$\mathrm{grad}\, v$	Gradient von v
$\mathrm{div}\, \vec{v}$	Divergenz von $\vec{v}$
$\mathrm{rot}\, \vec{v}$	Rotation von $\vec{v}$
A_h	Steifigkeitsmatrix mit den Komponenten A_{ij}
$\vec{f}_h$	Lastvektor mit den Komponenten f_i
$A^{(i)}$	Elementsteifigkeitsmatrix für das Element $T^{(i)}$
$\vec{f}^{(i)}$	Elementlastvektor für das Element $T^{(i)}$
$\mathcal{T}_h$	Vernetzung (Triangularisierung) des Gebietes Ω
$T^{(i)}$	Element i der Vernetzung
$\hat{T}$	Referenzelement
R_h	Anzahl der Elemente in der Vernetzung $\mathcal{T}_h$
P_h	Anzahl der Knoten in der Vernetzung $\mathcal{T}_h$
$\hat{N}$	Anzahl der Knoten pro Element
h	Diskretisierungsparameter

1 Einleitung

Über die letzten Jahrzehnte hat sich die zur Verfügung stehende Rechenleistung vervielfacht. Dies begründet sich zum einen durch die Weiterentwicklung der Rechnerhardware und zum anderen durch die Fortschritte in den Betriebssystemen und Anwendungen. Der Leistungszuwachs, welcher durch die Hardware begründet ist, folgt dabei bisher immer der von Moore (1965) festgestellten Gesetzmäßigkeit. Demnach verdoppelt sich die Anzahl an Transistoren in den Prozessoren alle achtzehn Monate und mit der Anzahl an Transistoren steigt auch die Rechenleistung exponentiell an. Im Juni 2008 stellte erstmals ein Rechnersystem mehr als 1 PFlops an Rechenleistung bereit. Bei diesem Wachstum kann davon ausgegangen werden, dass die Grenze von 1 EFlops in weniger als zehn Jahren überschritten wird. Demnach ist bereits heute für jede Wissenschaftsdisziplin eine große Rechenkapazität verfügbar.

In den Geowissenschaften und speziell in der Geophysik wurde schon sehr früh damit begonnen diese Rechnerleistung zu nutzen. Einerseits begründet durch die Tatsache, dass viele der Phänomene auf Grund der langen inhärenten Zeiträume im Laborexperiment nur schwer nachvollziehbar sind. Zusätzlich können durch die Verlagerung der Experimente in ein virtuelles Rechenlabor Kosten und Aufwand gesenkt und die Untersuchungen beliebig oft wiederholt werden. Andererseits ist es erst durch diese Rechenlabore und die entwickelten Simulationsanwendungen möglich, die komplexen Prozesse in der Erde zu untersuchen und zu verstehen. Die dabei zu lösenden diskreten Systeme mit heute bis zu 10^{10} Freiheitsgraden können nur auf Hochleistungsrechnern bearbeitet werden. Dies bedeutet konkret, dass erstmalig Skaleninteraktionen geophysikalischer Prozesse über drei Größenordnungen hinweg modelliert werden können. Beispielsweise sind Berechnungen tektonischer Störungszonen (circa 10 km) im Zusammenspiel mit globalen Fließprozessen des Erdmantels (circa 10000 km) oder die Simulation komplexer Erdbebenszenarien (seismische Wellenlänge circa 1 km) in geologischen Beckenstrukturen der Dimension 500 km durchführbar. Im Zentrum steht dabei das quantitative Zusammenführen von Modellen und Beobachtungen. Dazu gewinnen numerisch extrem aufwendige Verfahren der Datenassimilation rasch an Bedeutung. Viele geophysikalische Probleme sind von ihrer Natur her inverse Probleme, deshalb werden Optimierungsverfahren zur Bestimmung komplexer Modellzustände aus den gegebenen Beobachtungen verwendet. Die durchzuführenden Simulationen für die gesamte Erde (Mantelkonvektion und Wellenausbreitung) erfordern auf Grund

der Modellgröße (10^9 und mehr Gitterpunkte) und der vielen Zeitschritte lang anhaltende Rechnungen. Darüber hinaus werden diese Modellrechnungen nicht nur einmal durchgeführt sondern viele Male, bedingt durch die großen zu untersuchenden Parameterräume. Diese Simulationen können also in den Bereich des „Capacity Computing“ eingeordnet werden. Die regionalen und überregionalen Rechenzentren verfügen oft nicht über die notwendigen Rechenkapazitäten, um diese lang andauernden Rechnungen durchführen zu können. Zudem müssen die verfügbaren Rechenzeiten nach Möglichkeit gleichmäßig auf alle Nutzer verteilt werden. Daher liegt es nahe, Rechenanlagen in den Forschungsinstituten wie beispielsweise einem Lehrstuhl einzurichten. Diese können dann gezielt auf die speziellen Anforderungen der Simulationsanwendungen ausgerichtet werden und erreichen somit potenziell hohe Effizienzen.

Das Ziel der vorliegenden Arbeit ist es, am Lehrstuhl für Geophysik der Ludwig-Maximilians-Universität eine IT-Infrastruktur zu entwickeln, mit der die verschiedenen komplexen geophysikalischen Prozesse in der Erde simuliert werden können. Es stellt sich natürlich die Frage wie solche Strukturen aussehen sollten. Aus den Forschungsschwerpunkten ist es möglich ein Anforderungsprofil für die Infrastrukturen abzuleiten. Darüber hinaus werden auch die verfügbaren Techniken Einfluss nehmen. Es gilt also zu jedem Zeitpunkt, möglichst optimal die Hardware auf die Anwendungsprogramme abzubilden. Dies kann durch eine flexible und zugleich gut in das Lehrstuhlnetzwerk integrierte Simulationsinfrastruktur erzielt werden. Heute werden für das Hochleistungsrechnen bevorzugt Rechencluster mit zumeist standardisierter PC-Technik eingesetzt. Damit ist ein sehr gutes Preis-Leistungs-Verhältnis erreichbar. Neben der Rechenleistung ist auch Massenspeicher notwendig. Zum einen müssen die Simulationsergebnisse langfristig gespeichert und zum anderen auch die Observationsdaten lang vorgehalten werden. Diese treffen am Standort des Geophysikalischen Observatoriums in Fürstenfeldbruck ein und sollten wenn möglich auch für die wissenschaftlichen Mitarbeiter in München verfügbar sein. Die gespeicherten Simulations- und Messdaten bedürfen zusätzlich einer geeigneten Weiterverarbeitung und visuellen Darstellung. Die dabei erzeugten Abbildungen sollten sowohl für Veröffentlichungen als auch für Gäste ansprechend sein. Die Zielstellung der Arbeit ist es, das ungehinderte Zusammenspiel zwischen Simulation, Observation, Datenbearbeitung und Visualisierung für die Mitarbeiter am Lehrstuhl für Geophysik zu ermöglichen.

Kapitel 2 der vorliegenden Arbeit beschäftigt sich mit der Simulationsinfrastruktur am Lehrstuhl für Geophysik der Ludwig-Maximilians-Universität. Im Abschnitt 2.1 wird konkret der Hochleistungsrechner, beginnend mit der Konzeption über die Umsetzung und die Benchmarkrechungen bis zur aktuellen Auslastung, diskutiert. Daran schließt sich die Darstellung des Massenspeichersystems mit der Konzeption, Umsetzung und Auslastung an. Abschnitt 2.4 stellt den SMP-Rechner in den Mittelpunkt. Dabei werden ebenso die Konzeption, Umsetzung und Benchmarkrechnungen angesprochen. Im abschließenden Abschnitt 2.5 des Kapitels wird die Konzeption und Umsetzung des Visualisierungssystem vorgestellt.

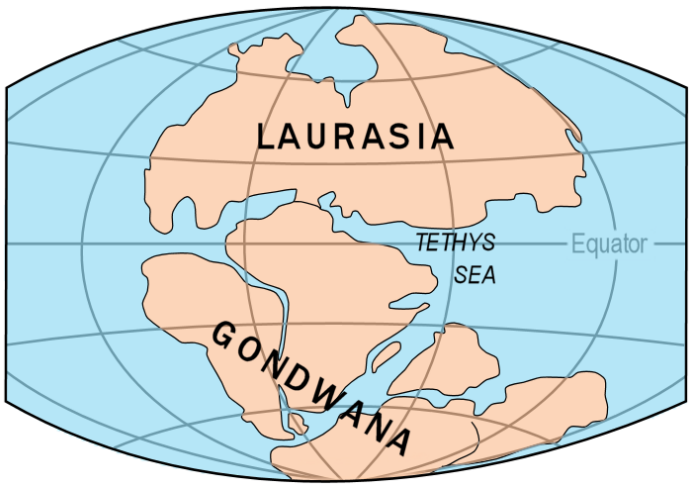
Die Simulationsanwendungen werden im Kapitel 3 besprochen. Ausgehend von einem allgemeinen Überblick zu den aktuellen rechnergestützten Forschungsschwerpunkten in der Geophysik des Departments für Geo- und Umweltwissenschaften der Ludwig-Maximilians-Universität wird in den sich anschließenden Abschnitten 3.2 bis 3.10 eine Übersicht der Simulationsanwendungen für den Rechencluster gegeben.

Im Kapitel 4 werden die Ergebnisse der vorangegangenen Abschnitte in komprimierter Form vorgestellt und Schlussfolgerungen sowie Ausblicke präsentiert.

Die vorliegende Arbeit führte zu zwei Erstautorveröffentlichungen (Oeser et al., 2006, 2009) (siehe Abschnitt 2.2 und 3.1).

2 Simulationsinfrastruktur in der Geophysik

Dem Lehrstuhl für Geophysik am Department für Geo- und Umweltwissenschaften der Ludwig-Maximilians-Universität (LMU) München war es Ende 2005 möglich, aus Mitteln des HBFG-Programms den ersten eigenen, rein für die Simulation von geowissenschaftlichen Problemstellungen konzipierten, Hochleistungsrechner (HPC-Rechner) zu beschaffen. Dabei sollte bereits die Wahl des Namenskürzels TETHYS (Tectonic High Performance Simulator) die klare Ausrichtung auf die Geowissenschaften verdeutlichen, da Tethys auch ein Ozean im Erdzeitalter des Mesozoikums war (siehe Abbildung 2.1(a)). Der Rechencluster wird seit mehr als zwei Jahren für die Simulation der Erdmantelkonvektion, Plattenbewegung, Ausbreitung seis-



(a) Pangaea mit dem Urozean Tethys

(b) Blick auf den HPC-Rechner

Abbildung 2.1: TETHYS: a) ein Ozean im Osten des Superkontinents Pangaea im Erdzeitalter des Mesozoikum und b) Blick in den Rechencluster des Lehrstuhls für Geophysik der Ludwig-Maximilians-Universität

mischer Wellen, Erdmanteltomographie, von Bruchprozessen in der Erdkruste und der Schadwirkung seismischer Ereignisse genutzt, um nur sechs der zahlreichen Anwendungsbeispiele zu nennen. Die Integration der Simulationsinfrastruktur in das neu entworfene und strukturierte Netzwerk der Geophysik wird durch die wissenschaftliche Neuausrichtung des Lehrstuhls auf die Simulation komplexer Prozesse im Erdkörper begleitet. Neben dem Linux-Rechencluster stehen noch ein Massenspeichersystem, ein 3D-Visualisierungslabor sowie ein SMP-Rechner mit 128 GB Arbeitsspeicher zur Verfügung. Ausgestattet mit diesen Komponenten, ist es den Wissenschaftlern am Lehrstuhl möglich, geowissenschaftliche Modellrechnungen durchzuführen, die gewonnenen Datensätze für längere Zeit abzuspeichern und interaktiv darzustellen und zu analysieren.

Die folgenden Abschnitte sollen im Einzelnen die verschiedenen Komponenten der Simulationsinfrastruktur am Lehrstuhl für Geophysik erläutern. Zu Beginn wird das Hauptaugenmerk auf den Hochleistungsrechner gelegt, um danach näher auf das Massenspeichersystem, das SMP-System und die 3D-Visualisierung einzugehen.

2.1 TETHYS

Die Planungen für das HPC-System TETHYS begannen bereits während der Antragsstellung im Rahmen des HBFG-Verfahrens[1] der Deutschen Forschungsgemeinschaft (DFG) im Jahre 2004. Die Schwerpunkte in diesem frühzeitigen Stadium des Projekts lagen dabei auf der Festlegung einer sinnvollen Größe des Rechenclusters. In diesem Zusammenhang wurden erstmals alle möglichen Anwendungen zusammengetragen und, mit Blick auf die typischen Modellgrößen, hinsichtlich ihrer Anforderungen an die Hardware untersucht. Auf Grundlage dieser Überlegungen (siehe Tabelle 2.1) konnte eine Mindestgröße für die Prozessoranzahl und Arbeitsspeichergröße anvisiert werden. Dem HBFG-Antrag lag ein System zu Grunde, welches mindestens aus 128 Rechenknoten (jeder Knoten verfügt über 1 Prozessor und 2 GB Arbeitsspeicher) besteht und in der Summe aller verbauten Festplatten über 6 TB Speicherplatz vorhält. Diese Anforderungen spiegeln im Wesentlichen den Bedarf der Hauptanwendung für den geplanten Rechencluster TERRA (Bunge et al., 1997) wider. Die beständige Weiterentwicklung im Bereich der Computerhardware, -architektur und der kontinuierliche Zuwachs an Anwendungsprogrammen für den Rechencluster bedingt, dass bereits ein Jahr später, im Sommer 2005, diese Überlegungen erneut geprüft und überarbeitet werden mussten. Die folgenden Abschnitte werden die Planungsphase im Vorfeld der Beschaffung, die Inbetriebnahme und die damit verbundene Überprüfung der Leistungsfähigkeit des Systems verdeutlichen.

Anwendung	Art der Simulation	RAM	CPUs
TERRA	Mantelkonvektion (Finite Elemente)	128 GB	64 oder 128
bm3d	Erdbebenquellsimulation (Finite Differenzen)	120 GB	64 oder 128
SPECFEM3D	Wellenausbreitung (Spektrale Elemente)	150 GB	75 oder 150
YAC	Wellenausbreitung (Finite Differenzen)	100 GB	64 oder 128

Tabelle 2.1: Bedarfszusammenstellung (Größe des Arbeitsspeichers (RAM) und Anzahl der Prozessoren (CPUs)) für die wichtigsten Anwendungen des Rechenclusters zum Zeitpunkt der Antragsstellung im HBFG-Programm der DFG

2.1.1 Konzeption

Von Anbeginn erfordert der Aufbau eines Hochleistungsrechners eine genau durchgeführte Planung. Diese sollte unter Einbeziehung aller möglichen Faktoren in einem Anforderungs-

[1] Hochschulbauförderungsgesetz (zum 31.12.2006 ausgelaufen und durch das Großgeräteprogramm der DFG ersetzt)

profil münden. Mit dieser möglichst exakten Darstellung der Gegebenheiten und des Bedarfs lassen sich Probleme und Fehler in der Ausschreibung und der späteren Inbetriebnahme vermeiden. Am Lehrstuhl für Geophysik gab es bis zu diesem Zeitpunkt keinen zentralisierten Betrieb der IT-Infrastruktur. Daher war es ein besonderes Anliegen, die sich bereits in der Umsetzung befindliche zentralisierte EDV-Struktur, einzubeziehen und nach Möglichkeit auch vollständig zu integrieren. Damit kann eine effektive Nutzung der vorhandenen Ressourcen erreicht und zusätzliche Kosten vermieden werden. Das zu erstellende Profil muss eine Lösung/Vorgehensweise für die folgenden Punkte bieten:

Raumkonzept: Bis 2004 war im Gebäude kein zentraler Serverraum vorhanden. Der benötigte Raum sollte durch das Bauamt der Universität so eingerichtet werden, dass ein reibungsloser IT-Betrieb möglich ist. In dem erstellten Raumkonzept finden die Statik, Elektro- und Netzwerkverkabelung sowie Klimatisierung Berücksichtigung.

Anwendungskonzept: In der Analyse des Rechenclustereinsatzes müssen aus den Anwendungen heraus die Fragen nach den Hardwareanforderungen geklärt werden. Neben der Anzahl an Rechenknoten und deren Ausstattung ist die Netzwerkverkabelung von besonderem Interesse. Aus den zur Verfügung stehenden Netzwerkarten wie GBit-Ethernet, Infiniband und Myrinet muss eine sinnvolle Wahl getroffen werden.

Betriebssystemkonzept: Das zu installierende Betriebssystem trägt wesentlich zur Leistungsfähigkeit des Gesamtsystems bei. Man sollte aber nicht nur die erzielte Rechenleistung in die eigenen Überlegungen einbeziehen, sondern auch der Wartungsaufwand wird vorrangig vom installierten Betriebssystem beeinflusst. Darüber hinaus sind mögliche Kosten durch zusätzlich notwendige Softwarelizenzen unerwünscht.

Managementkonzept: In den Bereich Management fällt nicht nur die Wartung des Betriebssystems, sondern auch die Verwaltung und Überwachung der Rechenknoten, der Schutz der Simulationsergebnisse und die Kontrolle des Serverraums.

Servicekonzept: Von allen Herstellern sind für ihre Hardware und den darauf laufenden Betriebssysteminstallationen entsprechende Dienstleistungs- und Supportverträge verfügbar. In wie weit dies, in Anbetracht der Gesamtnutzungsdauer und der daraus entstehenden Kosten, mit Blick auf die mögliche Eigenleistungen sinnvoll ist, muss vorab geklärt werden.

Der Weg zum endgültigen Anforderungsprofil für den Rechencluster soll im Folgenden unter Berücksichtigung aller genannten Punkte diskutiert werden. Das entstandene Gesamtkonzept bildete die Grundlage für die Ausschreibung innerhalb der Beschaffungsphase Ende 2005.

Raumkonzept

Einen Raum in einem Gebäude zu finden, das vor über 30 Jahren mit vorwiegend Laborräumen konzipiert wurde, gestaltete sich schwieriger als gedacht. Nach Auskunft des Bauamts der Universität besteht bei keinem der Räume im Gebäude ein Statikproblem, da alle die auftretenden Lasten aufnehmen können. Unter einem 19 Zoll Serverschrank können erhebliche punktuelle Belastungen entstehen, weil ein voll bestückter Rackschrank (siehe Abbildung 2.1(b)) bis zu 1000 kg Gesamtgewicht besitzt.

Die Grundfläche und Höhe des Raums sind von besonderer Bedeutung, da man bei einer Raumklimatisierung, über im Zimmer installierte Kühlaggregate, davon ausgehen kann, dass es nach der empirischen Formel[2] 2.1 möglich ist, das 1.5-fache der Grundfläche als Wärme in kW abzuführen.

$$\text{Wärme } [kW] = 1.5 * \text{Grundfläche } [m^2] \tag{2.1}$$

Eine grobe Abschätzung der benötigten Kälteleistung innerhalb des HBFG-Antrags ergab einen ungefähren Bedarf von 50 kW. Demnach muss der Raum über eine Grundfläche von etwa 33 m^2 verfügen. Ein Zimmer dieser Fläche konnte schließlich durch die Umgestaltung des bisher zum Teil als Labor genutzten Raums C429 gefunden werden, welcher sich auf der vierten Etage in der Theresienstraße 41 befindet. Die Abbildung 2.2 verdeutlicht die Dimension und geplante Anordnung der verschiedenen Systeme im Raum C429. Aus der gegebenen Grundflä-

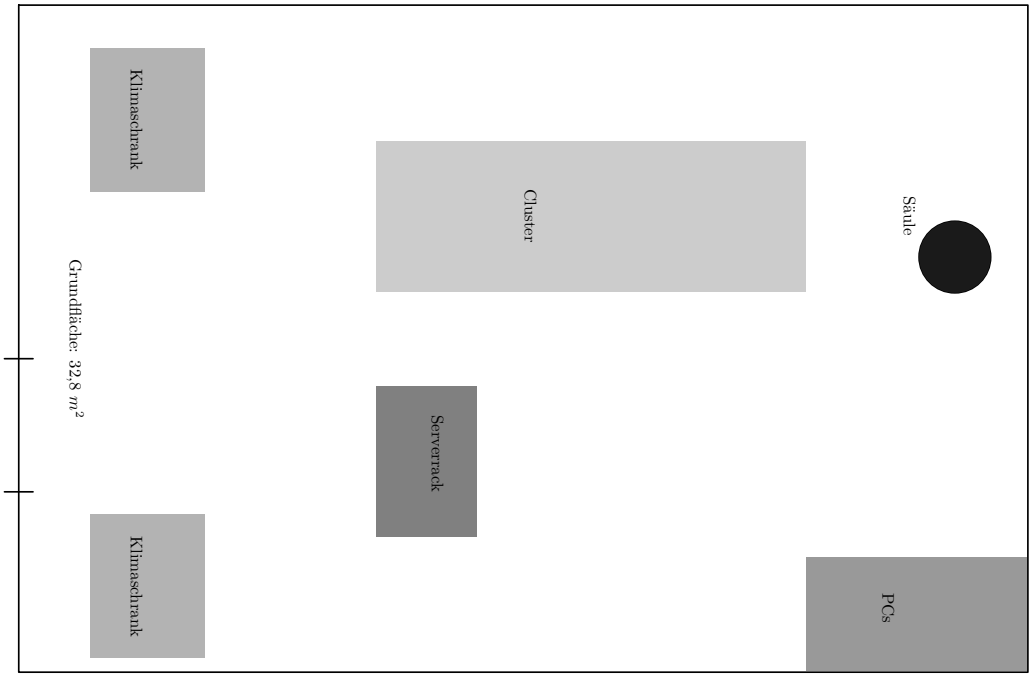


Abbildung 2.2: Schematische Darstellung des geplanten Serverraums C429

[2]persönliche Kommunikation mit D. Baier, Bauamt der Ludwig-Maximilians-Universität

che und einer Raumhöhe von 3.45 m ergibt sich für den Serverraum ein gut nutzbares Raumvolumen von 113 m^3. Damit steht ein genügend großes Luftvolumen zur Verfügung, um bei eventuellen Problemen an der Klimaanlage, den Raum als Puffer zu nutzen.

Mit der Klimatisierung ist auch einer der kritischen Punkte angesprochen. Es wurden mit dem Universitätsbauamt umfangreiche Gespräche über die Ausstattung des Raums mit einer Klimaanlage geführt. In den Besprechungen konnten die verschiedenen Möglichkeiten einen Raum entsprechend zu kühlen diskutiert werden. Neben Luft als Mittel zur Abfuhr der Wärme, besteht auch die Möglichkeit eine Wasserkühlung zu verwenden. Bei dieser Art der Klimatisierung sind die für diesen Zweck günstigeren Eigenschaften von Wasser wichtig. Einerseits beträgt die Wärmekapazität c von Wasser das 4.2-fache von Luft und andererseits hat Wasser fast die 1000-fache Dichte ϱ (siehe Tabelle 2.2). Nimmt man nach Gleichung 2.2 an, dass die aufzunehmende Wärmeänderung δQ und die Temperaturänderung dT für Luft und Wasser gleich sein sollen, so kann mit der Wärmekapazität c, der Dichte ϱ sowie dem Raumvolumen V berechnet werden, wie viel Masse oder Volumen an Wasser benötigt wird, um die gleiche Wärmemenge abzuführen.

Medium	Wärmekapazität $c\ [\frac{J}{kg \cdot K}]$	Dichte $\varrho\ [\frac{kg}{m^3}]$
Luft	1000	1.2
Wasser	4200	1000

Tabelle 2.2: Vergleich der gemittelten Wärmekapazitäten und Dichten für Wasser und Luft

$$\delta Q = c \cdot m \cdot dT \tag{2.2}$$

Wird die Gleichung 2.2 entsprechend der getroffenen Annahme aufgestellt und unter Verwendung von $m = \varrho \cdot V$ umgeformt, so erhält man die in Gleichung 2.3 dargestellte Form.

$$\begin{aligned} \delta Q_{Luft} &= \delta Q_{Wasser} \\ c_{Luft} \cdot m_{Luft} \cdot dT &= c_{Wasser} \cdot m_{Wasser} \cdot dT \\ c_{Luft} \cdot \varrho_{Luft} \cdot V_{Luft} &= c_{Wasser} \cdot \varrho_{Wasser} \cdot V_{Wasser} \\ V_{Wasser} &= \frac{c_{Luft} \cdot \varrho_{Luft} \cdot V_{Luft}}{c_{Wasser} \cdot \varrho_{Wasser}} \end{aligned} \tag{2.3}$$

Werden in diese Gleichung die Werte aus Tabelle 2.2 sowie das Raumvolumen von 113 m^3 eingesetzt, so erhält man das benötigte Wasservolumen $V_{Wasser} = 0.0323\ m^3$ um die gleiche Wärme abzuführen. Dies entspricht dem 3445-stel des notwendigen Luftvolumens. Damit könnte eine Wasserkühlung wesentlich effizienter die entstehende Abwärme aufnehmen und nach außen abführen. Vorausgesetzt die gesamte Anlage wird entsprechend konzipiert, ist es ebenfalls möglich bei einem Ausfall der Wasserkühlung über genügend Puffer zu verfügen, um

ein gefahrloses Abschalten der Rechner zu gewährleisten. Nach Auskunft verschiedener Firmen sind derartige Systeme auch ausgereift. Da allerdings diese Art der Kühlung bis dato an der Universität nicht eingesetzt wurde und auch Bedenken bezüglich der Dichtheit und Störunempfindlichkeit bestanden, wurde von einer solchen Lösung durch das Universitätsbauamt abgesehen. Eine weitere Möglichkeit stellte der Einbau eines doppelten Bodens dar, der aber aus Kostengründen vom Bauamt abgelehnt wurde. Am Schluss hat sich das Universitätsbauamt, welches auch für die Wartung zuständig ist, für die kostengünstigste Lösung entschieden. Es wurden im Serverraum zwei Klimaschränke der Firma Stulz mit einer nominellen Kälteleistung von jeweils 26.5 kW eingebaut. Sie erzeugen eine Luftwalze entlang der Längsseiten des Raums. Diese Art der Kühlung ist nicht optimal, da durch Verwirbelungen leicht heiße Zonen entstehen können. In Folge dessen werden unter Umständen einzelne System unzureichend mit Kühlluft versorgt. Auf diesen Punkt musste beim Aufbau des Rechenclusters besonders geachtet werden.

Neben der Klimatisierung des Raums wurden genügend Netzwerkanschlüsse und ein entsprechend dimensionierter Stromanschluss benötigt. Das Bauamt der Universität setzte den Plan entsprechend des Bedarfs um. Die Stromversorgung für den Rechenclusters konnte über vierzehn 3-Phasen-16-A-CEE-Steckdosen realisiert werden, wobei jeweils zwei im Sicherungskasten zusammengeklemmt sind. Jede Phase ist dabei mit 16 A abgesichert. Für den Anschluss des Rechenclusters an das Netzwerk der Geophysik wurden zwei GBit-fähige Netzwerkdosen im Raum verlegt.

Neben dem Aufbau des Systems in einem Raum der Geophysik hätte auch die Möglichkeit bestanden, den Cluster am Leibniz-Rechenzentrum der Bayerischen Akademie der Wissenschaften (LRZ) aufzustellen. Ein entsprechend ausgearbeitetes Konzept des LRZ sah allerdings viel zu hohe Kosten vor (rund 500 € pro Rechner und Jahr). Nach dem Umzug des LRZ in die neuen Räumlichkeiten am Forschungscampus in Garching wäre darüber hinaus, ein nicht unbedeutender Zeitaufwand durch die weiteren Wege bei Wartungsarbeiten entstanden. Viel wichtiger ist aber die Tatsache, dass die angestrebte vollständige Integration in das Lehrstuhlnetzwerk der Geophysik mit einem Standort im LRZ in der angestrebten Art nicht möglich wäre.

Anwendungskonzept

Die primäre Ausrichtung auf wenige Hauptapplikationen für einen Rechencluster bedingt eine genaue Analyse dieser Programme. Mit dem daraus erarbeiteten Anwendungskonzept kann die Hardware gezielt auf die Applikationen abgestimmt werden und es sind potenziell hohe Effizienzen erreichbar. TERRA (Bunge et al., 1997) dient als wichtigste Anwendung für den Cluster der Simulation von Mantelkonvektionsströmungen. Die zu rechnenden Probleme bewegen sich

in der Größenordnung von etwa 100 Millionen Gitterpunkten, was einem Arbeitsspeicherbedarf von circa 100 GB entspricht. Aus der verwendeten Gebietszerlegung für die Parallelisierung ergibt sich der Bedarf an Prozessoren P als $P = 2^n$ mit $n \in \mathbb{N}$. Die Überprüfung weiterer Applikationen für den Rechencluster im Hinblick auf deren Hardwareanforderungen ergab einen vergleichbaren Bedarf. Tabelle 2.3 zeigt eine Übersicht der weiteren Simulationsprogramme für den zu beschaffenden Rechencluster. Neben der Angabe welche geophysikalische Problemstellung mit Hilfe des Programms bearbeitet wird und welcher numerische Ansatz Verwendung findet (siehe Kapitel 3) , werden der Arbeitsspeicherbedarf und die benötigte Anzahl an Prozessoren für die mittelgroßen, dafür aber ständig zu rechnenden Probleme aufgeführt.

Nahezu alle aufgezählten Anwendungen können beliebig auf die eingesetzte Hardware skaliert werden. Mit dem Rechencluster sollen allerdings keine „one-of-a-kind“ Simulationen durchgeführt werden, sondern systematische Parameterstudien ermöglicht werden. Dies dient zur effizienten Vorbereitung größerer Simulationen auf zentralisierten HPC-Systemen und der Rechenbedarf für diese Untersuchungen liegt vorwiegend in einem mittleren Anforderungsbereich, bezüglich Arbeitsspeicher und Prozessoranzahl, aber mit langen Rechenzeiten. Man spricht dabei auch von „Capacity Computing“ im Gegensatz zum „Capability Computing“ wie beispielsweise auf dem Bundeshöchstleistungsrechner am LRZ.

Anwendung	Art der Simulation	RAM	CPUs
TERRA	Mantelkonvektion (Finite Elemente)	120 GB	128
bm3d	Erdbebenquellsimulation (Finite Differenzen)	120 GB	64 oder 128
SeisSol	Wellenausbreitung (Finite Elemente)	128 GB	64 oder 128
GeoELSE	Wellenausbreitung (Spektrale Elemente)	64 GB	128
ses3d sec3d	Wellenausbreitung (Spektrale Elemente)	60 GB	120
SPECFEM3D	Wellenausbreitung (Spektrale Elemente)	150 GB	75 oder 150
YAC	Wellenausbreitung (Finite Differenzen)	100 GB	64

Tabelle 2.3: Bedarfszusammenstellung (Größe des Arbeitsspeichers (RAM) und Anzahl der Prozessoren (CPUs)) für die wesentlichen Anwendungen des Rechenclusters zum Zeitpunkt der Ausschreibung (kurze Beschreibungen der Simulationsanwendungen sind in den Kapiteln 3.3 bis 3.10 zu finden)

Die Bedarfsangaben in Tabelle 2.3 belegen, dass ein Rechencluster mit 128 Prozessoren nahezu ideal ist. Für den Arbeitsspeicher gestaltet sich dies etwas schwieriger. Es sollte ein Kompromiss für die verschiedenen Anwendungen gefunden werden. Mit 128 GB RAM wäre für 6 von 7 Anwendungen der typische Bedarf gedeckt. Zusätzlich sind noch circa 100 MB Arbeitsspeicher pro Rechenknoten für das Betriebssystem einzuplanen. Jede dieser Applikationen benötigt

darüber hinaus noch temporären Speicherplatz um Zwischenergebnisse abzulegen. Dazu können die lokalen Festplatten oder ein paralleles Dateisystem (siehe Abschnitt 2.1.2) verwendet werden. Für TERRA werden ungefähr 6 TB Festplattenplatz benötigt. Bei den anderen Applikationen bewegt sich der Bedarf in der Größenordnung von 1 TB.

Neben den in Tabelle 2.3 aufgeführten Simulationsanwendungen werden immer wieder vereinzelt kurzzeitige Projekte, zum Beispiel von Gastwissenschaftlern oder Mitarbeitern anderer geowissenschaftlicher Bereiche der LMU, zur Ausführung auf dem HPC-System kommen. Für diese Projekte kann aber, bedingt durch deren Kurzlebigkeit, keine Bedarfsplanung erstellt werden, sodass sie nicht in das Anwendungskonzept einfließen.

Nachdem die Prozessoranzahl und die Arbeitsspeichergröße des geplanten Rechenclusters festgelegt werden konnten, musste eine Entscheidung zu Gunsten eines Prozessortyps getroffen werden. Der HPC-Rechner sollte der Philosophie der ersten Beowulf Rechner folgend aus Standard PC-Komponenten bestehen. Damit ist es möglich ein gutes Preis-Leistungs-Verhältnis zu erreichen. Es war eine Wahl zwischen den etablierten 32-bit und den neuen 64-bit Prozessoren notwendig. Um diese Entscheidung zu vereinfachen sollten kleine Benchmarkrechnungen mit zwei typischen Anwendungen des Rechenclusters durchgeführt werden.

TERRA Für die Hauptanwendung TERRA des HPC-Systems wurde dazu eine Konfiguration erstellt, in welcher mit einer geringen Auflösung für den Erdkörper die Konvektion des zähflüssigen Materials im Erdmantel simuliert wird. Das dabei verwendete Gitter, zur Abstraktion des Erdmantels, verfügt über 1.4 Millionen Gitterpunkte und belegt damit rund 1 GB Arbeitsspeicher. Bei dieser Gittergröße, charakterisiert durch den Wert 64 des Steuerparameters MT, wird aller 125 km ein Gitterpunkt auf der Erdoberfläche erzeugt. Für jeden Gitterpunkt und Zeitschritt sind circa 8000 Gleitpunktzahloperationen („Floating Point Operations“ oder kurz Flop) notwendig. Für die gegebene Anzahl an Gitterpunkten und den im Test zu rechnenden 100 Zeitschritten ergeben sich 1.12×10^{12} Flop. Durch die Unterstützung der HPC-Gruppe von Fujitsu-Siemens-Computers (FSC) konnten mit dieser TERRA-Konfiguration Laufzeituntersuchungen zum einen auf einem System mit zwei Intel[3] Itanium Prozessoren (1.6 GHz Taktfrequenz) und zum anderem auf einem Rechner mit zwei Intel Xeon Prozessoren (3.6 GHz Taktfrequenz) durchgeführt werden. Für alle Tests mit der Anwendung TERRA auf FSC Rechnern wurde von deren HPC-Gruppe der Intel Compiler in Version 8.1 und ein aktuelles MPICH[4] eingesetzt. In Tabelle 2.4 sind die Mittelwerte der Laufzeiten $\bar{t}$ für ein und zwei simultan gestartete Instanzen von TERRA aufgeführt. In beiden Tests zeigte sich, dass entgegen der Erwartung die Laufzeit für zwei Instanzen nicht der Laufzeit einer Instanz entspricht. Dies könnte im Verhal-

[3] `http://www.intel.com`

[4] freie Implementierung des MPI Standards Version 1 (Gropp et al., 1996)

	Itanium		Xeon	
Programminstanzen	1	2	1	2
$\bar{t}\ [s]$	728	890	701	1044

Tabelle 2.4: TERRA: Benchmark auf FSC Rechner mit entweder Intel Itanium oder Intel Xeon Prozessoren – Laufzeiten für ein oder zwei simultan gestartete Programminstanzen von TERRA

ten des Betriebssystems begründet sein, die beiden Prozesse den zwei im System installierten CPUs häufig neu zuzuordnen. Damit muss der Cache jeder CPU beständig aktualisiert werden, was wertvolle Berechnungszyklen ungenutzt lässt. Ein weiterer Grund könnte auch in der Systemarchitektur der Intel Prozessoren zu finden sein, bei der beide CPUs über die Northbridge auf den Arbeitsspeicher zugreifen müssen (siehe Abschnitt A.2). Deren Bandbreite ist allerdings beschränkt, was zu verringerten Datentransferraten zwischen Arbeitsspeicher und CPU führt und damit das Programm verlangsamt. Die Ergebnisse aus Tabelle 2.5 zeigen den Mittelwert der Laufzeit für P Prozessoren $\bar{t}(P)$ und den daraus berechneten parallelen Speedup $S_{par}(P)$ (siehe Abschnitt A.1) für ein und vier Prozessoren auf einem System mit Intel Xeon CPUs (3.6 GHz Taktfrequenz) entweder in einer Single oder Dual Prozessorausführung. Mit vier MPI-Prozessen skaliert TERRA besser auf einem Single Prozessor System und erreicht dabei einen Speedup von 3.3.

P	1	4 (Dual-Xeon)	4 (Single-Xeon)
$\bar{t}(P)\ [s]$	701	337	212
$S_{par}(P)$	1	2.1	3.3

Tabelle 2.5: TERRA: Benchmark auf FSC Rechner mit Intel Xeon Prozessoren – Laufzeiten und Speedup im Vergleich für einen Rechner mit zwei Dual-Xeons und zwei Rechnern mit Single-Xeons

Die Laufzeiten mit mehreren Programminstanzen für AMD[5] Opteron Prozessoren werden in Tabelle 2.6 aufgeführt. Wie zu erkennen ist, liegen die Laufzeiten ebenfalls für zwei oder vier Programminstanzen auf einem System leicht über der Laufzeit für nur eine Programminstanz. Die AMD Opteron Systeme verfügten über ausreichend RAM, der auch jedem Prozessor zugeordnet war. Damit können die höheren Laufzeiten nicht durch die längeren Speicherzugriffszeiten erklärt werden, es könnte sich also wieder um den gleichen Effekt wie bei den Intel

[5] `http://www.amd.com`

	Opteron 252		Opteron 275	
Programminstanzen	1	2	1	4
$\overline{t}\ [s]$	778	870	842	1106

Tabelle 2.6: TERRA: Benchmark auf FSC Rechner mit AMD Opteron 252 und 275 – Laufzeiten für ein, zwei oder vier simultan gestartete Programminstanzen von TERRA

Prozessoren handeln (Verschieben der Prozesse zwischen den CPUs durch das Betriebssystem). Dieses Verhalten belegt die Notwendigkeit im Betriebssystemkern die einzelnen Prozesse möglichst an einen Prozessor zu binden.

Wenn über MPI vier parallele, am selben Problem arbeitende, Prozesse gestartet werden, so verringert sich die Laufzeit für den AMD Opteron 275 von 842 s für einen MPI-Prozess auf 354 s für vier MPI-Prozesse auf einem Rechner ausgestattet mit zwei solcher AMD Opteron CPUs. Zum Vergleich wurde auf zwei Systemen mit jeweils zwei AMD Opteron 250 Prozessoren ebenfalls vier MPI-Prozesse gestartet, was eine Laufzeit von 325 s ergab. Die um 28 s geringere Laufzeit kann durch die 200 MHz höher getakteten Prozessoren erklärt werden.

AMD verwendet für ihre Prozessoren ein Nummernschema aus dem nicht so ohne weiteres die genauen Spezifikationen abgelesen werden können. Deshalb sind in Tabelle 2.7 diese Werte für die relevanten CPUs zusammengetragen. Alle Prozessoren der 200 Serie sind für den Betrieb in Rechnern mit 2 CPU Sockeln ausgelegt, wohingegen die 800 Serie für 8 CPU Sockel pro Rechner entwickelt werden.

AMD Opteron Prozessor	275	250	252	850
Anzahl der Rechenkerne	dual core	single core	single core	single core
Taktfrequenz	2.2 GHz	2.4 GHz	2.6 GHz	2.4 GHz
L1 Cache (Daten/Instruktionen)	64/64 kB	64/64 kB	64/64 kB	64/64 kB
L2 Cache (Daten + Instruktionen)	1 MB	1 MB	1 MB	1 MB

Tabelle 2.7: Spezifikation der in den Benchmarks verwendeten AMD Opteron Prozessoren

In einem weiteren Test sollte der Einfluss schneller Netzwerke wie Infiniband auf die Laufzeit von TERRA untersucht werden. Zu diesem Zweck war es möglich, auf dem Infiniband-Cluster[6] des Lehrstuhls für Rechnertechnik und Rechnerorganisation / Parallelrechnerarchitektur (LRR) der Fakultät für Informatik an der Technischen Universität München (TUM) einige Testläufe durchzuführen. Bei den in ihrem HPC-System verwendeten CPUs handelt es sich um AMD

[6] `http://www.lrr.in.tum.de/Par/arch/infiniband/`

Opteron 850 Prozessoren (siehe Aufstellung in Tabelle 2.7). In jedem Rechenknoten ist neben einem GBit-Ethernet Netzwerkanschluss auch ein Mellanox Infiniband 4x Netzwerkadapter vorhanden. Probleme bei der Übersetzung einer eigenen MPI-Version mit einem der beiden vorhandenen Compiler (Intel Compiler Version 8.1 und PGI Compiler) verhinderten Testläufe für beide Netzwerke mit identischer Compiler und MPI-Version. So wurde für das Infiniband Netzwerk MVAPICH[7] als MPI-Implementierung mit dem PGI Compiler benutzt und für das GBit-Ethernet Netzwerk kam MPICH mit dem Intel Compiler zum Einsatz. In Tabelle 2.8 sind die Ergebnisse des Benchmarks aufgeführt. Die Laufzeiten für die Tests mit dem GBit-Ethernet Netzwerk und dem Intel Compiler sind zum Teil deutlich geringer als die gemessenen Zeiten bei Infiniband Netzwerk und PGI Compiler. Dem gegenüber ist der aus den Laufzeiten berechnete parallele Speedup für das Infiniband Netzwerk mit PGI Compiler deutlich besser. Betrachtet man nur den Einfluss der Netzwerke auf die Laufzeit mit zunehmender Anzahl an Prozessoren, so scheint das schnellere Infiniband Netzwerk Vorteile zu besitzen. Bei dieser Untersuchung wird aber der Compilereinfluss auf die Laufzeiten deutlich. Für den Einzelprozessorlauf ist der Test mit dem Intel Compiler zweieinhalbmal schneller als der mit dem PGI Compiler und zeigt damit die unbedingte Notwendigkeit auf, gut optimierende Compiler zu verwenden.

CPU	Opteron 850			Opteron 850		
Netzwerk	GBit-Ethernet			Infiniband		
Compiler	Intel			PGI		
P	1	4	16	1	4	16
$\bar{t}(P)\ [s]$	890	297	97	2236	581	158
$S_{par}(P)$	1	2.9	9.3	1	3.9	14.4

Tabelle 2.8: TERRA: Benchmark auf dem Rechencluster des LRR an der TUM mit AMD Opteron 850 Prozessoren sowie GBit-Ethernet Netzwerk und Intel Compiler Kombination oder Infiniband Netzwerk und PGI Compiler Kombination – Laufzeiten und Speedup für ein, vier und sechzehn MPI-Prozesse

Um den Einfluss schneller Netzwerke auf die Laufzeit von TERRA noch genauer zu betrachten, war es in einer weiteren Untersuchung möglich, auf dem sich im Probebetrieb befindenden HPC-System der Universität Paderborn[8] verschiedene Testläufe durchzuführen. Jeder Rechenknoten des Clusters ist mit zwei Intel Xeon 3.6 *GHz* Prozessoren ausgestattet. Daneben sind in jedem Knoten eine GBit-Ethernet und eine Mellanox Infiniband 4x Netzwerkschnittstelle verbaut. Für die Tests standen vier Rechner des Clusters zur Verfügung. Während der Rechnun-

[7] `http://mvapich.cse.ohio-state.edu/overview/mvapich/`
[8] `http://wwwcs.uni-paderborn.de/pc2`

gen mit vier MPI-Prozessen wurde nur eine CPU pro System verwendet. Bei den Testläufen mit acht MPI-Prozessen waren beide CPUs pro Knoten im Einsatz und für die Intra-Knoten-Kommunikation wurde dabei das „Shared Memory“ System genutzt. Als Softwareprodukte waren der Intel Compiler in Version 8.1 und ScaMPI in Version 3.3.2 von Scali verfügbar. Damit stand für den Vergleich der Netzwerke eine identische Ausgangsbasis zur Verfügung. Tabelle 2.9 führt die Laufzeiten für beide Netzwerkarten auf. Es ist zu erkennen, dass TERRA bis zu vier MPI-Prozesse gut skaliert, aber dann für acht MPI-Prozesse deutlich abfällt. Dieses Verhalten könnte in der Architektur der Intel Rechner begründet sein, bei der beide Prozessoren gemeinsam über die Northbridge auf den Arbeitsspeicher zugreifen und sich damit die zur Verfügung stehende Bandbreite teilen müssen. Um dies ausschließen zu können, wäre ein Test mit acht Rechenknoten und acht MPI-Prozessen notwendig gewesen. Leider stand keine entsprechende Hardware zum Testzeitpunkt zur Verfügung, da der Test nur innerhalb eines sehr kurzen Zeitfensters während der Inbetriebnahme möglich war. Die geringen Laufzeitvorteile von 14 % bei vier MPI-Prozessen und 12.5 % bei acht lassen Zweifel aufkommen, ob diese sehr preisintensive Netzwerktechnik für den geplanten Rechencluster sinnvoll ist. Vorangegangene Laufzeitanalysen auf den verschiedensten Hochleistungsrechnern weltweit zeigten, dass die für das Netzwerk verwendete Technik unkritisch ist für TERRA. Ein Standard GBit-Ethernet Netzwerk scheint demnach vollkommen ausreichend.

CPU	Xeon 3.6 GHz			Xeon 3.6 GHz		
Netzwerk	GBit-Ethernet			Infiniband		
P	1	4	8	1	4	8
$\bar{t}(P)$ $[s]$	1006	374	244	1006	311	214
$S_{par}(P)$	1	2.7	4.1	1	3.2	4.7

Tabelle 2.9: TERRA: Benchmark auf dem Rechencluster der Universität Paderborn mit Intel Xeon 3.6 GHz Prozessoren sowie GBit-Ethernet und Infiniband Netzwerk – Laufzeiten und Speedup für ein, vier und acht MPI-Prozesse

YAC Neben den Benchmarkrechnungen mit TERRA sollten noch Laufzeituntersuchungen mit dem Programm YAC (Ewald, 2006; Ewald et al., 2006) durchgeführt werden. Für dieses Finite Differenzen Simulationsprogramm zur Ausbreitung elastischer Wellen wurde ein Modell als Konfigurationsgrundlage gewählt, welches etwa 3.6 Millionen Gitterpunkte beinhaltet, was in diesem Fall einem Arbeitsspeicherbedarf von ungefähr 855 MB entspricht. In der ersten Untersuchung konnten wieder Testläufe auf den Intel Systemen der HPC-Gruppe von FSC durchgeführt werden. Die Ergebnisse dieser Untersuchung sind in Tabelle 2.10 zusammenge-

	Itanium		Xeon	
Programminstanzen	1	2	1	2
$\bar{t}\ [s]$	1512	2335	1485	2812

Tabelle 2.10: YAC: Benchmark auf FSC Rechner mit entweder Intel Itanium oder Intel Xeon Prozessoren – Laufzeiten für ein oder zwei simultan gestartete Programminstanzen von TERRA

tragen. Wie bereits die Laufzeiten für die identischen Untersuchungen mit TERRA belegen (siehe Tabelle 2.4), steigen diese auf beiden Prozessortypen beim gleichzeitigen Start zweier Programminstanzen an, jedoch bei YAC signifikanter. Dieser deutlichere Anstieg dürfte in der weniger starken Optimierung des Programms begründet liegen. Die limitierte Bandbreite zum Arbeitsspeicher und das Zuordnen der Prozesse zu den CPUs kommt bei diesem Test wieder zum Tragen und begründet die ansteigenden Laufzeiten. Wenn für YAC auf dem Intel Itanium System zwei MPI-Prozesse gestartet werden, so verringert sich die Laufzeit von 1512 s auf 1109 s. Dies entspricht einem parallelen Speedup $S_{par}(P)$ von 1.4. Die gemessenen Zeiten für die Skalierungsuntersuchungen auf dem Intel Xeon System sind in Tabelle 2.11 aufgelistet. Darin ist zu erkennen, dass auf diesen Rechnern YAC etwas besser mit der Anzahl an CPUs skaliert.

P	1	2	4
$\bar{t}(P)\ [s]$	1485	888	442
$S_{par}(P)$	1	1.6	3.4

Tabelle 2.11: YAC: Benchmark auf FSC Rechner mit Intel Xeon Prozessoren – Laufzeiten und Speedup für ein, zwei und vier MPI-Prozesse

Die Laufzeiten für die Tests von YAC auf den AMD Opteron Systemen von FSC werden in Tabelle 2.12 aufgelistet. Auf Grund von Problemen beim Übersetzen des Quelltextes wurde im Gegensatz zu den vorherigen Benchmarkrechnungen der PGI Compiler verwendet. Wie bereits die Tests für TERRA auf den gleichen Systemen gezeigt haben, steigen die Laufzeiten bei mehreren Programminstanzen auf einem System an. Werden über MPI vier parallele Prozesse gestartet, so verringert sich die Laufzeit für den AMD Opteron 275 von 2306 s bei einem MPI-Prozess auf 805 s für vier MPI-Prozesse auf einem Rechner ausgestattet mit zwei solcher AMD Opteron Prozessoren. Dies entspricht einem parallelen Speedup $S_{par}(P)$ von 2.9. Weitere Untersuchungen mit dem Simulationsprogramm YAC auf den HPC-Rechnern des Lehrstuhls Rechnertechnik und Rechnerorganisation / Parallelrechnerarchitektur der TUM und der Uni-

versität Paderborn konnten auf Grund der kurzen Zeitfenster, die für die Tests zur Verfügung standen, leider nicht durchgeführt werden.

	Opteron 252		Opteron 275	
Programminstanzen	1	2	1	4
$\bar{t}\,[s]$	2145	2255	2306	2984

Tabelle 2.12: YAC: Benchmark auf FSC Rechner mit AMD Opteron 252 und 275 – Laufzeiten für ein, zwei oder vier simultan gestartete Programminstanzen von YAC

Wahl der Prozessoren und Netzwerkart Benchmarks und Performanceanalysen zum Zeitpunkt der Beschaffung bestätigten die hohe Rechenleistung der AMD Opteron CPUs im Vergleich zu Intel Xeon Prozessoren. Eine Recherche im Internet Portal der „Standard Performance Evaluation Corporation“[9] (SPEC) liefert für Rechner von FSC die in Tabelle 2.13 aufgeführten Punktzahlen des CPU2000 Benchmarks. Die angegebenen Punktzahlen wurden bei den entsprechenden Tests mit starker Compiler-Optimierung ermittelt. Im CINT2000 Teil sind zwölf rechenintensive Tests aus dem Bereich der Ganzzahlarithmetrik zusammengefasst, wohingegen im CFP2000 Teil vierzehn rechenintensive Tests aus dem Bereich der Gleitpunktzahlen verwendet werden. Da diese Tests zum einem sehr stark von der Optimierung durch den verwendeten Compiler abhängen und zum anderen nicht die Leistung in realen Anwendungen widerspiegeln, ist es schwierig allein aus diesen Werten eine Entscheidung für einen Prozessor- oder Rechnertyp zu fällen. Die in den Tests verwendete Itanium CPU mit 1.6 *GHz* Taktfrequenz erreicht in einem Rechner der Firma HP bei dem CINT2000 Benchmarkteil im Vergleich eine Punktzahl von 1408 und beim CFP2000 Teil 2553 Punkte. Dies belegt die Aussage, dass die Prozessoren der AMD Opteron Familie in einigen Bereichen den viel teureren CPUs der Intel Itanium Reihe gleichwertig oder überlegen sind. Damit können die Itanium CPUs aus der Liste der möglichen Prozessoren für den Rechencluster genommen werden. Zum Zeitpunkt der Beschaffung war es aus Aspekten der Investitionssicherheit nicht sinnvoll reine 32-bit Prozessoren zu beschaffen. Die 64-bit Prozessoren der AMD64 oder EM64T Familie wiesen in den durchgeführten Untersuchungen ihre sehr hohe Leistungsfähigkeit nach. Die Laufzeiten für die AMD Opteron und Intel Xeon CPUs lagen in den durchgeführten Benchmarks in einem vergleichbaren Bereich, deshalb sollte nach einem Blick in die Preistabellen eine Entscheidung gefällt werden. Im Preis-Leistungs-Verhältnis lagen zum Beschaffungszeitpunkt eindeutig die Single Core CPUs der AMD Opteron Familie vorn. Nach den Ergebnissen der Benchmarks sprach

[9] `http://www.spec.org`

System	CINT2000	CFP2000
FSC Primergy Rx200 S2 (Intel Xeon 3.60 *GHz*)	1712	1866
FSC Primergy Rx220 (AMD Opteron 250)	1688	2076
FSC Primergy Rx220 (AMD Opteron 252)	1809	2182

Tabelle 2.13: SPEC: Punkte des CPU2000 Benchmark für die in den Test verwendeten FSC Rechner

auch nichts gegen diese Prozessoren, sodass als Grundlage für die anstehende öffentliche Ausschreibung CPUs der AMD Opteron Serie verwendet wurden.

Die Erfahrungen mit Simulationen auf verschiedenen Hochleistungsrechnern und die unterschiedlichen Untersuchungen zum Netzwerk zeigen, dass eine GBit-Ethernet Netzwerkverkabelung vollkommen ausreichend ist. Die sehr viel teureren Netzwerke vom Typ Infiniband oder Myrinet zeigten keine wesentlichen Vorteile. Dies bedeutet aber nicht, dass für ein GBit-Ethernet Netzwerk nicht auch hohe Kosten entstehen könnten. Dies tritt dann ein, wenn Switche zum Einsatz kommen müssen, die über 100 und mehr Anschlüsse verfügen und von jedem zu jedem anderen Anschluss die volle Transferrate bereitstehen soll (man spricht dabei auch von „full duplex mode“). Um derartige kostenintensive Lösungen umgehen zu können, muss nach alternativen Ansätzen gesucht werden. Ein kaskadiertes Netzwerk wäre eine Option. In Abbildung 2.3 ist ein solcher kaskadierter Aufbau dargestellt. Um einen zentral angeordneten „Core-Switch“, der über mindestens fünf 10-GBit-Anschlüsse verfügt, werden vier „Node-Switches“ mit 10-GBit angeschlossen. Jeder dieser „Node-Switches“ stellt dann mindestens sechzehn 1-GBit-Anschlüsse für die einzelnen Rechenknoten zur Verfügung. Diese angestrebte Lösung ist um den Faktor 2-3 kostengünstiger als ein einzelner großer zentraler Switch.

Nachdem die wesentlichen Hardwaredesignpunkte des Rechenclusters erarbeitet wurden, kann aus dem erstellten Anwendungskonzept, welches die Hardwareanforderung enthält, die Konfiguration für den HPC-Rechner zusammengestellt werden. Diese diente als Grundlage für die öffentliche Ausschreibung Ende 2005.

Betriebssystemkonzept

Mit der Neuausrichtung des Lehrstuhls für Geophysik auf die Simulation komplexer geowissenschaftlicher Prozesse wurde im Jahre 2004 begonnen, die IT-Infrastruktur an die modernen Gegebenheiten und Techniken anzupassen und das komplette Netzwerk mit einer neuen Struktur zu versehen. Diese zentralisierte Infrastruktur bietet viele Vorteile in Bezug auf die Benutzer-, Rechner- und Softwareverwaltung. Die angestrebte vollständige Integration des HPC-Rechners

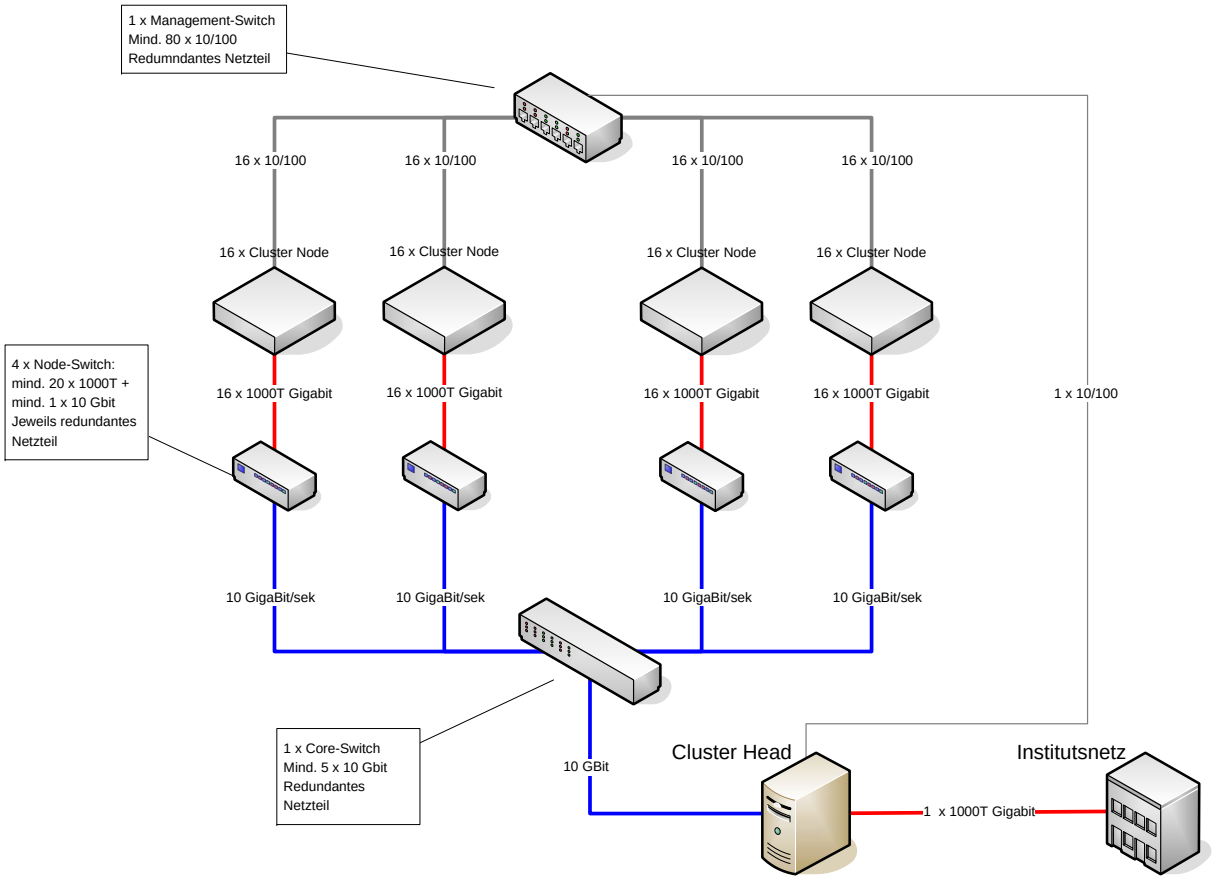


Abbildung 2.3: Schematische Darstellung der geforderten Netzwerkkonfiguration in der öffentlichen Ausschreibung

in die aufgebaute moderne und flexible Infrastruktur war der zentrale Punkt bei den Überlegungen zum Betriebssystemkonzept. Nach Möglichkeit sollten alle sich bietenden Synergieeffekte genutzt werden, was natürlich am besten durch die Verwendung der Kernkomponenten des Lehrstuhlnetzwerks erreicht werden kann. Es handelt sich dabei im Wesentlichen um folgende Punkte:

- Nutzerauthentifizierung an einem der zentralen LDAP Server,
- Bereitstellung der Heimatverzeichnisse jedes Nutzers durch den zentralen NFS Server,
- Einsatz der am Lehrstuhl benutzten Linux Distribution Debian GNU/Linux[10],
- zentrale Softwareverwaltung und Installation der Rechenknoten,
- Schutz des Netzwerks durch die vorhandenen Sicherheitslösungen,
- Nutzung der zentralen Datensicherungslösung am LRZ.

[10] `http://www.debian.org`

Die Verwaltung der Benutzer und deren Zugriffsrechte erfolgt über die zentralen Authentifizierungsserver auf Basis von LDAP. Die Verwendung dieser Datenbestände ermöglicht es, effektiv einer doppelten Datenhaltung und -pflege in mehreren Quellen entgegenzutreten. Bei dem gleichzeitigen Einsatz der zentralen Benutzerverzeichnisse im gesamten Rechencluster ergeben sich für die Anwender Vorteile in der Art, dass es nicht notwendig ist Konfigurationen und Datenbestände auf zwei Systemen vorzuhalten. Dies ermöglicht gleichzeitig eine bessere, weil Festplattenplatz und damit Ressourcen schonende Nutzung der zentralen Datensicherungssysteme am LRZ. Die Verwendung der am Lehrstuhl eingesetzten Linux Distribution Debian GNU/Linux vereinfacht wesentlich das Zurechtfinden der Nutzer im System, da sie bereits auf einem identischen System arbeiten. Für Systemverwalter bietet Debian darüber hinaus sehr viele Vorteile, wobei die herausragende Softwareverwaltung besonders hervorzuheben ist. Über diese können die in nahezu unerschöpflicher Anzahl vorhandenen Softwarepakete aus den Distributionspaketquellen installiert werden. Diese Pakete werden durch das Debian GNU/Linux Projekt ständig mit Softwareaktualisierungen versorgt. Die Vielzahl an fertigen Softwarepaketen ermöglicht es, den Großteil der benötigten Software ohne großen Aufwand zu installieren. Falls für eine Software kein fertiges Paket existiert, so kann dies auf sehr einfache Weise selbst erstellt werden. Das entstandene Softwarepaket kann anschließend in die eigenen Paketquellen eingebunden werden und steht somit für die Installation auf den einzelnen Rechnern zur Verfügung. Mit den vorhandenen Paketquellen kann durch die Verwendung einer automatischen Installationsroutine der gesamte Hochleistungsrechner innerhalb kürzester Zeit (neu) aufgesetzt werden.

Für die Umsetzung dieses Teilkonzeptes wird ein hohes Maß an Verständnis des vorhandenen Lehrstuhlnetzwerks benötigt. Da dieses nicht innerhalb kurzer Zeit erreicht werden kann, wollte der Lehrstuhl die Installation vorrangig selbst ausführen. Bei der Einrichtung von besonderen Komponenten, wie zum Beispiel des Netzwerks, sollte aber auf die technische Unterstützung des HPC-Anbieters zurückgegriffen werden. Durch die eigenständige Umsetzung dieses Teilkonzeptes wurde eine erhebliche Kostenreduktion erhofft.

Managementkonzept

Die Umsetzung des Betriebssystemkonzeptes sollte wie bereits erläutert eigenständig erfolgen. Daran knüpft nahtlos die Notwendigkeit an, das Management des Betriebssystems in eine zentrale Komponente einzugliedern. Darin müssen Routinen hinterlegt werden, die auf bestimmte Ereignisse mit vordefinierten Aktionen reagieren. Dies muss vollständig automatisiert erfolgen und keines Eingriffs eines Systemverwalters bedürfen. Zu solchen Ereignissen zählt beispielsweise der Ausfall der Klimaanlage. Im zentralen Management muss dafür eine Aktion

vorhanden sein, die zum Beispiel den Rechencluster geordnet abschaltet, damit ein Schutz des Gesamtsystems möglich ist. Die ausgelösten Aktionen müssen schon angelaufen sein, bevor das eigentlich verursachende Problem schon viel weit reichendere Folgen verursacht hat.

Um die Schutzmechanismen und das allgemein notwendige Management umsetzen zu können, bedarf es gewisser Hardwarevoraussetzungen. Sie sollten im Vorfeld der Beschaffung untersucht und in der Ausschreibung entsprechend gefordert werden. Welche am Markt verfügbare Hardware kann die geforderte Aufgabe kostengünstig erfüllen? Neben der Überwachung der Systemhardware unterhalb des Betriebssystemlevels (Lüfter und Spannungen) muss eine einfache Systemdiagnose der Rechenknoten möglich sein. In den letzten Jahren hat sich dazu eine betriebssystemunabhängige Lösung etabliert, die als „Intelligent Platform Management Interface“ (IPMI)[11] bekannt ist. Für diese offene Spezifikation, wie auf notwendige Werte und Systemfunktionen über eine Managementschnittstelle zugegriffen werden kann, existiert von jedem Hardwarehersteller eine in Hardware umgesetzte Lösung, die meist über einen Netzwerk- oder serielle Schnittstellenanschluss verfügt. Die Hardware kann dann über herstellereigene Softwareprodukte angesprochen werden. Da es sich aber um eine offene Spezifikation handelt, sind verschiedene freie Softwareprojekte entstanden, die eine herstellerunabhängige Lösung umzusetzen versuchen. Dies schien für das HPC-Projekt genau richtig zu sein, da es eine zukunftssichere Lösung darstellt. Dementsprechend wurde in der Ausschreibung eine IPMI konforme Schnittstelle gefordert. Damit über das Netzwerk auf diese Schnittstelle zugegriffen werden kann ohne das Datennetzwerk zusätzlich zu belasten, sollte ein Managementnetzwerk in den Rechencluster eingeplant werden. In Abbildung 2.3 ist ein solches bereits eingezeichnet. Über dieses einfache Ethernet Netzwerk kann zusätzlich das Betriebssystem verwaltet werden.

Im Management des eigentlichen Serverraums ist die Überwachung der Temperatur neben der Detektion von Wassereinbrüchen und der Entstehung von Feuer wichtig. Eindringendes Wasser oder ein Brandalarm sollte sinnvoller über entsprechende Sensoren an die zentrale Haustechnik gemeldet werden, da meist für solche Ereignisse Aktionen notwendig sind, die nicht vom Systemverwalter ausgeführt werden können. Ein Anstieg der Temperatur deutet meist auf einen Defekt in der Klimaanlage hin. Ab einer gewissen Temperatur im Serverraum leidet nachweislich die eingesetzte Hardware in der Art, dass die zu erwartende Lebensdauer deutlich herabgesetzt wird. Bei einem Totalausfall der Klimatisierung ist schnellstens die Hardware vor Defekten zu schützen, was am besten durch ein kontrolliertes Abschalten der Rechner erfolgen kann. Über im Raum angebrachte Temperatursensoren wird eine solche Überwachung realisiert. Eine einfache Möglichkeit bietet die Integration der Messsensoren in bereits vorhandene oder ohnehin zu beschaffende Geräte. Bei vielen Modellen zur unterbrechungsfreien Stromversorgung (USV) sind solche Managementfunktionalitäten durch entsprechende Erweiterungskarten mög-

[11] `http://www.intel.com/design/servers/ipmi`

lich. Da für den Verwaltungsrechner, auf dem auch ein Teil der Simulationsergebnisse abgelegt werden soll, eine solche Stromversorgung notwendig ist, um die Resultate zu schützen, sollte eine entsprechende Erweiterung in der Ausschreibung gefordert werden. Der eigentliche Schutz des Verwaltungsrechners vor Defekten an der Hardware kann durch den Einsatz eines komplett redundant ausgelegten Systems erfolgen.

Servicekonzept

Die Deutsche Forschungsgemeinschaft (DFG) spricht in ihren Richtlinien zur Beschaffung von Geräten aus Mitteln der bewilligten DFG-Anträge eine Empfehlung bezüglich der zu wählenden Garantie-/Serviceleistungen aus. Es wird für Computerhardware vorgeschlagen, eine Garantie von fünf Jahren zu wählen. Bei Hochleistungsrechnern kann nach Auskunft der zuständigen zentralen EDV-Beschaffung an der LMU auch von den vorgeschlagenen fünf Jahren nach unten abgewichen werden. Diese Möglichkeit wird dadurch begründet, dass HPC-Rechner häufig bereits vor Ablauf der fünf Jahre ersetzt werden müssen/sollen. Dabei wären die zusätzlich erworbenen Dienstleistungsjahre unnötig ausgegebene Mittel.

Ein Blick in die Statistik der Hardwaredefekte verschiedener Hersteller lässt erkennen, dass ein Großteil der vorkommenden Defekte und Probleme zu Beginn des Nutzungszeitraums oder erst nach vier bis fünf Jahren auftritt.

In der Neuausrichtung des Lehrstuhls für Geophysik hat sich der neue Lehrstuhlinhaber Prof. Bunge das Ziel gesetzt, die Versorgung mit entsprechender Rechenleistung nachhaltig zu gestalten. Dies bedeutet für den HPC-Rechner eine kontinuierliche Auffrischung, was durch einen Austausch des Gesamtsystems aller vier bis fünf Jahre möglich ist.

Die Preislisten der verschiedenen Hersteller für HPC-Rechner zeigen einen teilweise übermäßig starken Anstieg der Preise für Garantieleistungen ab dem vierten Jahr. Um diesen Sachverhalt zu verdeutlichen sind in Tabelle 2.14 entsprechende Preisangaben für zwei Hersteller aufgeführt. Als Beispiel sind eine Primergy Rx220 der Firma FSC und ein ProLiant DL160 von HP angegeben. Für beide Geräte sind die Listenkaufpreise der Garantieleistungen für die entsprechende Anzahl an Jahren in der Tabelle aufgeschlüsselt. Bei FSC wird ein garantierter Vor-Ort-Service innerhalb von 48 Stunden angeboten. Wohingegen HP den gleichen Vor-Ort-Service innerhalb von 24 Stunden anbietet.

Zieht man alle dargelegten Punkte in die Überlegung für das angestrebte Servicekonzept mit ein, so scheint das Risiko (selbst zu tragenden Kosten bei auftretenden Defekten) mit einer dreijährigen Garantieleistung überschaubar. Deshalb wurde entschieden, für alle Hardwarekomponenten des Rechenclusters eine dreijährige Garantie mit Vor-Ort-Service zu wählen. Einige Anbieter

	FSC Primergy Rx200	HP ProLiant DL160
3 Jahre	144 €	246 €
4 Jahre	339 €	365 €
5 Jahre	540 €	477 €

Tabelle 2.14: Kosten für die Garantieleistungen nach Jahren aufgeschlüsselt für zwei ausgesuchte Geräte zweier Hersteller

im Netzwerkbereich arbeiten darüber hinaus mit Garantieleistungen die zum Teil deutlich über dem avisierten Zeitraum liegen.

Da die Wartung der Betriebssysteminstallation eigenständig am Lehrstuhl durchgeführt werden soll, kann auf eine entsprechende Dienstleistung durch den Anbieter verzichtet werden. Um bei eventuell auftretenden Problemen aber dennoch gerüstet zu sein, wurde entschieden, drei Arbeitstage eines Technikers in der Ausschreibung zu fordern.

Anforderungsprofil

Im Folgenden soll noch einmal in zusammengefasster Form das Anforderungsprofil für den zu beschaffenden Rechencluster dargestellt werden. Dabei werden die wichtigsten Punkte aus den vorher diskutierten Konzepten aufgegriffen. Das so erstellte Anforderungsprofil diente der zentralen EDV-Beschaffung der Ludwig-Maximilians-Universität als Grundlage für die öffentliche Ausschreibung des HPC-Systems Ende 2005.

Die Einrichtung des zentralen Serverraums für den Lehrstuhl für Geophysik wurde nach den erstellten Vorgaben durch das zentrale Bauamt der Ludwig-Maximilians-Universität im Herbst 2004 durchgeführt. Eine Beschreibung der vorhandenen Klimatisierung und der installierten Strom- und Netzwerkanschlüsse war im Ausschreibungstext enthalten.

Von den mindestens 64 geforderten Rechenknoten (Cluster-Nodes) sollten alle über eine absolut identische Ausstattung verfügen. Der gesamte Ausschreibungstext wurde so weit wie möglich herstellerneutral formuliert. Die in Tabelle 2.15 aufgeführten Merkmale sollten den Anbietern des HPC-Systems als Konfigurationsvorschlag dienen. Eine Abweichung von dieser Beispielkonfiguration sollte begründet werden.

Neben den in der Zusammenstellung aufgeführten Merkmalen, wurden noch allgemeine Hinweise an die Anbieter herausgegeben. Es sollte bei den angebotenen CPUs auf ein ausgewogenes Verhältnis zwischen realer Anwenderleistung und dem Preis des Prozessors geachtet werden und im Zweifel auf ein besseres Preis-Leistungs-Verhältnis gesetzt werden. Bei den CPUs ist weiterhin darauf zu achten, dass eine entsprechend dimensionierte Kühlung vorhanden und die

Prozessor:	– 2 gleiche 64-bit Prozessoren pro Knoten (AMD Opteron 246 oder vergleichbarer Intel Prozessor)
Arbeitsspeicher:	– 2 *GB* (ECC) Speicherausbau (4 *GB* optional) – die aggregierte nominelle Speicherbandbreite eines Knoten soll der zweifachen Bandbreite einer CPU entsprechen
Netzwerk:	– zweimal 1-GBit-Ethernet Netzwerkanschluss onboard (RJ45)
Festplatte:	– mindestens 160 *GB* SATA-Festplatte (hot-plug)
Aufbau:	– 19 Zoll Rackmount Konfiguration (bevorzugt eine Höheneinheit pro Knoten) – Ausstattung mit einem Baseboard Management Controller und IPMI-Kompatibilität – werkzeuglose Montage von Erweiterungskomponenten

Tabelle 2.15: Konfigurationsvorschlag für einen Rechenknoten des geplanten HPC-Systems

Ableitung der entstehenden Wärme sichergestellt ist. Für alle Lüfter sind verpflichtend kugelgelagerte Modelle zu verwenden. Das Stromversorgungsmodul muss ausreichende Leistungsreserven bieten. Der Gesamtaufbau eines Rechenknotens muss so gestaltet sein, dass ein Betrieb unter Volllast stabil und performant möglich ist. Insbesondere hat die Verlegung der Kabel sauber und den Luftstrom nicht behindernd zu erfolgen. Da eine kleine Abweichung im Aufbau eines Rechenknotens größere Probleme bei der Inbetriebnahme nach sich ziehen kann, wurde auch auf identische Mainboards mit einer einheitlichen BIOS-Version geachtet.

Der zentrale Verwaltungsrechner (Cluster-Head) des HPC-Systems soll den Zugriff der Benutzer auf den Rechencluster regeln und verwaltet alle Rechenknoten. Demnach muss dieses System entsprechend den erhöhten Anforderungen aufgebaut werden (Hochverfügbarkeit). Der in Tabelle 2.16 aufgeführte Konfigurationsvorschlag wurde in der Ausschreibung verankert. Auch beim Cluster-Head muss eine den Spezifikationen entsprechende Betriebstemperatur für alle Komponenten und besonders für die Festplatten gewährleistet werden.

In öffentlichen Ausschreibungen werden üblicherweise noch gewisse Unbedenklichkeitsnachweise gefordert. Für den Verwaltungsrechner und die Rechenknoten sind dies a) der Nachweis der Konformitätsverfahren/CE-Zertifizierung 89/36/EWG (EMV); 72/23 EEC (LVD) nach EU-Richtlinien – Nachweis durch akkreditiertes Prüflabor und b) der Nachweis der elektromagnetischen Verträglichkeit gemäß EN55022 class A, EN55024, EN61000-3-2/-3 sowie c) der Nachweis der Produktsicherheit gemäß EN60950.

Die in Abbildung 2.3 dargestellte kaskadierte Vernetzung mittels eines Standard Ethernet Netzwerks wurde auf Grundlage der vorangegangenen Tests erstellt. Im Ausschreibungstext musste

Prozessor:	– ein Standardprozessor (der CPU Typ kann von dem im Rechenknoten verwendeten abweichen) – im Dual-CPU-Mainboard
Arbeitsspeicher:	– 2 GB (ECC) Speicherausbau (4 GB optional)
Netzwerk:	– sechsmal 1-GBit-Ethernet Netzwerkanschluss davon 2 onboard (RJ45) – einmal 10-GBit-Ethernet Netzwerkanschluss
Festplatte:	– zweimal 73 GB SCSI-Festplatte (hot-plug) im RAID-1 Modus
Aufbau:	– 19 Zoll Rackmount Konfiguration – redundante Stromversorgungsmodule und kugelgelagerte Lüfter (beides hot-plug) – redundante gesicherte Stromversorgung (mit eingebauter Überwachung der Raumtemperatur und Raumfeuchte, per SNMP abfragbar) – zweimal 64-bit PCI-X Steckplatz frei (Anschluss eines vorhandenen RAID-Subsystems mit 2 TB Nettokapazität) – Remote Management Board – Ausstattung mit einem Baseboard Management Controller und IPMI-Kompatibilität – werkzeuglose Montage von Erweiterungskomponenten

Tabelle 2.16: Konfigurationsvorschlag für den Verwaltungsrechner des geplanten HPC-Systems

für die Umsetzung dieses Konzeptes nach verschiedenen Netzwerkgeräten gefragt werden. Neben der notwendigen Verkabelung sollte als zentrale Komponente der so genannte Cluster Core Switch (CSS) eingesetzt werden. Von diesem CSS sollten die vier Cluster Node Switches (CNS) mit einer 10-GBit-Verbindung angebunden werden. Jeder der 16 pro CNS geplanten Rechenknoten sollte über 1-GBit mit dem CNS verbunden werden. Für das eigenständige Management Netzwerk ist ein dedizierter Cluster Management Switch (CMS) vorgesehen. Die genauen technischen Anforderungen an die einzelnen Netzwerkgeräte können den folgenden Aufstellungen entnommen werden.

Konfigurationsvorschlag für die vier Cluster Node Switches (CNS):

- mindestens zwanzig 1-GBit-Ethernet Netzwerkanschlüsse (RJ45) (Verbindung zu den einzelnen Rechenknoten)
- mindestens einmal 10-GBit-Ethernet Netzwerkanschluss (Verbindung zum CCS)
- Full Wire Speed/Non-Blocking (Nachweis durch Datenblatt)

- mindestens 60 GBit/s Switching Kapazität
- voll managebar
- redundante gesicherte Stromversorgung
- Jumbo Frame Support

Konfigurationsvorschlag für den Cluster Core Switch (CCS):

- mindestens fünf 10-GBit-Ethernet Netzwerkanschlüsse (viermal CNS und einmal Verwaltungsrechner)
- wünschenswert ist eine mögliche Erweiterung auf sechs oder acht 10-GBit-Ethernet Netzwerkanschlüsse
- Full Wire Speed/Non-Blocking (Nachweis durch Datenblatt)
- voll managebar
- redundante gesicherte Stromversorgung
- Jumbo Frame Support

Konfigurationsvorschlag für den Cluster Management Switch (CMS):

- mindestens 80 Ethernet Netzwerkanschlüsse und ein Ethernet Netzwerkanschluss für den Verwaltungsrechner (RJ45)
- Full Wire Speed / Non-Blocking (Nachweis durch Datenblatt)
- voll managebar
- redundante gesicherte Stromversorgung

Der in der öffentlichen Ausschreibung geforderte Aufbau des Rechencluster schließt natürlich alle dafür notwendigen Teile ein. Diese sind gesondert im Angebot zu vermerken (19 Zoll Rackschränke, Verkabelung für die Stromversorgung und die Datenvernetzung, für Montage notwendige Kleinteile).

Als Service- und Garantieleistungen wurde im Ausschreibungstext zum einen eine Ersatzteilgarantie für alle Hardwarekomponenten von fünf Jahren gefordert und zum anderen ein drei Jahre andauernder Vor-Ort-Service mit einer Reaktionszeit von 48 Stunden. Um diese Dienstleistungsangebote erfüllen zu können, müssen die Anbieter typischerweise zertifizierte Service Partner sein.

2.1.2 Umsetzung

Nachdem im vorangegangenen Abschnitt die Konzeptionsphase zum Hochleistungsrechner diskutiert wurde, soll im Folgenden näher auf die Details während der Umsetzungsphase eingegangen werden. Im Mittelpunkt steht dabei das Resultat der öffentlichen Ausschreibung des Rechenclusters. Nachfolgend wird also dargestellt, was auf Grundlage der Angebote beschafft werden konnte und wie die Inbetriebnahme organisiert wurde. Bei der geführten Diskussion werden auch die aufgetretenen Probleme und Schwierigkeiten genannt.

Die Zahl an eingegangenen Angeboten als Reaktion auf die öffentliche Ausschreibung im Oktober 2005 war ernüchternd. Von den zwei abgegebenen Geboten erfüllte nur eines die Vorgaben des finanziellen Rahmens der zur Verfügung stehenden Mittel aus dem HBFG-Antrag. Nach Auskunft der zentralen EDV-Beschaffung besteht die Möglichkeit die Ausschreibung aufzuheben, wenn weniger als drei Gebote abgegeben wurden. Demnach wäre es möglich ohne eine Bindung an vorgegebene Richtlinien, die ausgeschriebenen Produkte eigenständig zu beschaffen. Da aber das eine in den finanziellen Rahmen passende Angebot sehr gut ausgearbeitet war, schien dieser Schritt nicht notwendig und wurde auch gemeinschaftlich abgelehnt. So konnte durch den Auftraggeber (zentrale EDV-Beschaffung der LMU) in Absprache mit dem Betreiber (Lehrstuhl für Geophysik am Department für Geo- und Umweltwissenschaften) der Auftrag an den Auftragnehmer (Firma Microstaxx GmbH mit Sitz in München) erteilt werden.

Aufbau

Die Lieferung der gemäß dem Angebot bestellten Komponenten des Rechenclusters erfolgte durch den Auftragnehmer Mitte Dezember 2005. Die Firma Microstaxx GmbH führte danach wie gefordert den Hardwareaufbau des HPC-Systems durch. Dies beinhaltete die Aufstellung der 19 Zoll Rackschränke, den Einbau der Rechenknoten in die Schränke, die Integration der Netzwerkkomponenten sowie die Strom- und Netzwerkverkabelung. In Abbildung 2.4 ist in ausgesuchten Bildern dieser Aufbauprozess dargestellt. Nach Abschluss dieser Arbeiten führte der Auftragnehmer für alle Komponenten eine umfassende Funktionsprüfung durch. Im Anschluss daran konnte die Konfiguration der Netzwerkkomponenten (CCS, CNS, CMS) eingespielt werden. Damit war im Januar 2006 der erste Schritt der Inbetriebnahme des Clusters abgeschlossen. Um alle weiteren Schritte erfolgreich durchführen zu können, musste zunächst der erste Teil des neuen, zentralen Massenspeichersystems in Betrieb genommen werden (siehe Abschnitt 2.3). Dies erforderte eine kurze Unterbrechung der Arbeiten am Rechencluster, weil das neue Massenspeichersystem als zentraler Fileserver für alle Benutzerverzeichnisse (auch für den HPC-Rechner) eingesetzt werden sollte. Nach der erfolgreichen Integration des Mas-

Abbildung 2.4: Aufbau der Hardware des HPC-Systems durch die Firma Microstaxx GmbH im Dezember 2005

senspeichersystems in die zentrale IT-Infrastruktur der Geophysik, konnte Ende Februar 2006 mit der Inbetriebnahme des HPC-Systems fortgefahren werden. Bevor die dabei durchgeführten Arbeiten näher zu diskutieren sind, soll die bestellte Hardware des Rechenclusters aufgelistet werden. Alle Bestandteile des Hochleistungsrechners entsprechen den Mindestanforderungen aus der öffentlichen Ausschreibung (siehe die Beispielkonfigurationen in Abschnitt 2.1.1). Die stetig sinkenden Preise für Computerhardware erlaubten es, bei der Auftragsvergabe die um 400 MHz schneller getakteten CPUs vom Typ AMD Opteron 250 als Prozessoren für die Rechenknoten (Cluster-Node) zu wählen. Die Konfiguration eines solchen Knotens wird in Tabelle 2.17 aufgeführt. Insgesamt konnten mit den Geldern des HBFG-Antrags 64 Rechenknoten für den Hochleistungsrechner beschafft werden. Die 128 darin verbauten AMD Prozessoren erreichen theoretisch eine Peak Performance von 624 GFlops. Wie viel davon als reale Rechenleistung für die Anwendungsprogramme des HPC-Systems zur Verfügung steht, wird in Kapitel 2.1.3 diskutiert.

Modell:	FSC Primergy Rx220
Prozessoranzahl:	zwei
Prozessortyp:	AMD Opteron 250 (64 bit, single core)
Taktfrequenz:	2.4 GHz
L1 Cache:	64/64 kB (Daten/Instruktionen)
L2 Cache:	1 MB (Daten + Instruktionen)
Arbeitsspeicher:	2 GB (DDR1)
Festplatte:	160 GB (SATA)
Netzwerkanschluss:	zweimal 1-GBit-Ethernet (onboard)

Tabelle 2.17: Hardwarespezifikation für die Rechenknoten des HPC-Systems

Die Beispielkonfiguration aus Abschnitt 2.1.1 für den Verwaltungsrechner (Cluster-Head) sah ein hochverfügbares Systems vor. Zusätzlich sollte der Rechner auch über mehrere freie Erweiterungssteckplätze verfügen. Dies konnte vom Auftragnehmer nicht mit einem System von FSC auf Basis der AMD Prozessoren verwirklicht werden, da von diesem Hersteller die entsprechenden Optionen nicht für die Primergy Rx220 vorgesehen sind. Der Rechnertyp wurde nur für das Hochleistungsrechnen entworfen, weshalb keine Möglichkeiten zur Redundanzbildung im System vorgesehen sind. Darum ist der Verwaltungsrechner von TETHYS ein System auf Basis von Intel Xeon CPUs. Dieser Primergy Tx300S2 von FSC bietet neben der redundanten Auslegung aller Systemkomponenten (Festplatte, Speicher, CPU, Stromversorgung) auch noch die Möglichkeit, in mehrere freie Steckplätze entsprechende Erweiterungskarten zu integrieren. Tabelle 2.18 legt die Konfiguration dieses Servers dar. In einen der freien Steckplätze konn-

te der SCSI-Controller zum Anschluss des vorhandenen RAID-Subsystems eingebaut werden. Damit wurde den Nutzern die Möglichkeit gegeben, die mitunter recht großen Simulationsdatenmengen an einem zentralen Ort abzulegen. Dieses SCSI-Subsystem verfügt über etwa 1.8 TB Speicherkapazität und kann nicht erweitert werden. Allerdings werden die zu speichernden Simulationsergebnisse über die Nutzungsdauer des Systems hinweg beständig anwachsen, womit auch der Bedarf an Speicherkapazität zunehmen wird. In Abschnitt 2.3 wird daher das weitaus flexiblere und auch größere zentrale Massenspeichersystem ausführlich besprochen.

Modell:	FSC Primergy Tx300S2
Prozessoranzahl:	zwei
Prozessortyp:	Intel Xeon (64 bit)
Taktfrequenz:	3.0 GHz
L1 Cache:	16 kB (Daten) + 16 kμOps (Instruktionen)
L2 Cache:	2 MB (Daten + Instruktionen)
Arbeitsspeicher:	2 GB (DDR2)
Festplatte:	zweimal 73 GB (SCSI) im RAID-1 Modus
Netzwerkanschluss:	zweimal 1-GBit-Ethernet (onboard)
	viermal 1-GBit-Ethernet
	einmal 10-GBit-Ethernet

Tabelle 2.18: Hardwarespezifikation für den Verwaltungsrechner des HPC-Systems

Der Struktur aus Abbildung 2.3 folgend, ist der detaillierte Aufbau des eingerichteten Netzwerks in Abbildung 2.5 dargestellt. Für das Managementnetzwerk des Rechenclusters wurde vom Auftragnehmer ein HP ProCurve Switch 4148 GL verbaut. Dieser stellt sechsundneunzig 100 MBit Ethernet Anschlüsse zur Verfügung. Alle Rechenknoten sowie der Verwaltungsrechner sind über ein eigenes CAT 5 Netzwerkkabel mit dem CMS verbunden. Zur Kommunikation in diesem Netzwerk wird ein separierter IP Adressbereich verwendet. Womit das Managementnetzwerk vollständig vom eigentlichen Datennetzwerk getrennt ist. Es wurde durch die Firma Microstaxx GmbH ebenfalls mit Switches der Firma HP realisiert, wobei als CCS ein HP ProCurve 6400 CL Verwendung findet. Dieser stellt sechs 10-GBit-Ethernet Anschlüsse vom Typ CX4 bereit, über welche die Verbindung zu den CNS verwirklicht wurde. Zusätzlich ist der Verwaltungsrechner über einen Anschluss vom Typ SR Optical mit 10-GBit-Ethernet angebunden. Als CNS kommt jeweils ein HP ProCurve 3400 CL-24 zum Einsatz. Jeder dieser vier Switches stellt vierundzwanzig 1-GBit-Ethernet Anschlüsse sowie zwei 10-GBit-Ethernet Anschlüsse vom Typ CX4 bereit. Wie bereits erwähnt wurde eine Verbindung zwischen den CCS und den CNS über die CX4 Anschlüsse hergestellt. Alle Rechenknoten wurden über ein Netz-

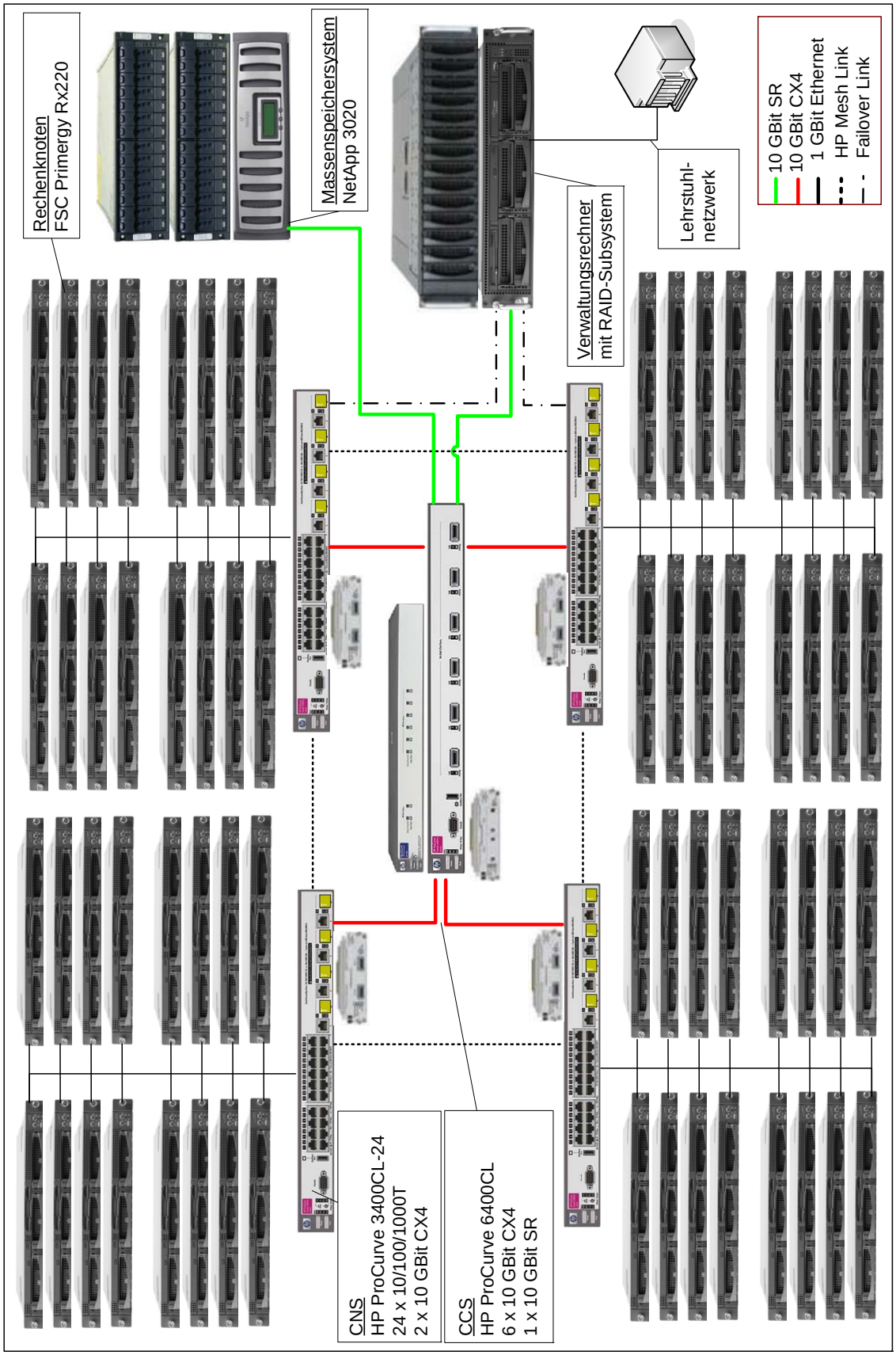


Abbildung 2.5: Darstellung der Netzwerkkonfiguration des HPC-Systems

werkkabel vom Typ CAT 6 mit einem 1-GBit-Ethernet Anschluss an einen CNS verbunden. Somit weist das Datennetzwerk eine baumartige Struktur auf. Alle verbauten Netzwerkkomponenten verfügen über eine redundante Stromversorgung und der Hersteller gibt eine lebenslange Garantie (lebenslang bedeutet in diesem Fall 30 Jahre).

Das gesamte Equipment des HPC-Systems wurde vom Auftragnehmer in drei 19 Zoll Rackschränke verbaut. In einem dieser sind auch die beiden für die Hochverfügbarkeit und Redundanz benötigten USV von APC untergebracht. Die Firma APC[12] bietet für verschiedene ihrer USV Modelle Managementkarten an, mit denen unter anderem die Temperatur gemessen werden kann. Über diese Sensoren erfolgt die Überwachung der Raumtemperatur (und damit der Klimaanlage) im Serverraum.

Der Rechencluster TETHYS (64 Rechenknoten; 1 Verwaltungsrechner; Datennetzwerk; Managementnetzwerk; Montagematerial) wurde mit 230 000 € aus Mitteln des HBFG-Antrags (Kennziffer: 132/943–1) und mit 35 000 € aus Mitteln von Prof. Bunge finanziert. Im Herbst 2006 konnte das HPC-System mit weiteren Rechenknoten ausgebaut werden. Dr. Martin Käser warb erfolgreich, im Rahmen des Emmy Noether Programms der DFG, Mittel für eine Nachwuchsforschergruppe ein. Weitere 10 Rechenknoten konnten damit aus diesem Programm (Kennziffer: KA 2281/2–1) beschafft werden. Zusätzlich wurde aus Lehrstuhlmitteln ein weiterer Rechenknoten erworben und die Firma FSC erklärte sich auf Grund des unzureichend ausgeführten Supports und den daraus entstandenen Problemen bereit, weitere 5 Rechenknoten zu finanzieren (siehe dazu Seite 39). Damit verfügt der HPC-Rechner über 80 Rechenknoten, die zusammen eine theoretische Peak Performance von 768 GFlops erreichen.

Inbetriebnahme

Wie bereits dargestellt, führte die Firma Microstaxx GmbH den gesamten Hardwareaufbau durch. Somit waren von Seiten des Lehrstuhls für Geophysik an der Ludwig-Maximilians-Universität keine weiteren Arbeiten für den Aufbau der Hardware notwendig. Die Konfiguration der Switches wurde ebenfalls durch einen Techniker des Auftragnehmers in Absprache mit dem Betreiber vorgenommen. Nach Möglichkeit sollten die geplanten Netzwerke logisch getrennt werden, wobei eine physikalische Trennung bereits durch die Verwendung dedizierter Netzwerkgeräte erreicht wurde. Diese logische Netzseparierung kann durch die Nutzung verschiedener IP-Adressbereiche ermöglicht werden. Während der Simulationsrechnungen werden die zu übertragenen Daten über den privaten IP-Adressbereich 172.16.10.0/24 verschickt. Im Managementnetzwerk läuft der Datentransfer davon vollkommen getrennt im 172.16.1.0/24 Netz. Ein weiterer privater IP-Adressbereich 172.16.2.0/24 erschien sinnvoll, um die Kommunikation mit den IPMI-Schnittstellen der Rechenknoten zu ermöglichen.

[12] `http://www.apc.com/`

Betriebssystem Nach Abschluss der Aufbauarbeiten und dem vorher notwendigen Einrichten des neuen zentralen Massenspeichersystems durch den Lehrstuhl für Geophysik (siehe Abschnitt 2.3) konnte mit der Installation des Betriebssystems auf den HPC-Rechnern begonnen werden. Dabei war es wichtig, zuerst den Verwaltungsrechner einzurichten, da dieser für alle weiteren Schritte unabdingbar war. Die Installation eines Debian GNU/Linux Betriebssystems soll nicht Gegenstand dieser Arbeit sein, sondern vielmehr die benötigten Programme und Dienste mit deren Hilfe der Betrieb des Rechenclusters effektiv gestaltet werden kann. Als Basis für die Installation diente die Linux Distribution Debian[13] GNU/Linux Sarge (AMD64). Unter AMD64 versteht man den von der Firma AMD für ihre Prozessoren eingeführten speziellen 64-bit Modus (siehe Abschnitt A.3). Die darin enthaltenen Spezifikationen sollten bei der Zusammenstellung einer Linux Distribution beachtet werden, da nur dadurch die optimale Leistung aus einem solchen System gewonnen werden kann. Dies alles wird vom Debian Projekt gewährleistet.

Der Verwaltungsrechner von TETHYS ist die zentrale Komponente des gesamten HPC-Systems. Er stellt alle wichtigen und für den reibungslosen Betrieb notwendigen Dienste zur Verfügung. Deshalb ist es von besonderer Relevanz, dass dieser Rechner einen gewissen Grad an Hochverfügbarkeit erreicht. Bei den angebotenen Diensten des HPC-Rechners handelt es sich um die im Folgenden aufgeführten.

LDAP: Die Umstrukturierung des gesamten Lehrstuhlnetzwerks erforderte auch die Umstellung der Benutzerverwaltung auf eine zentralisierte Struktur. Das eingesetzte System auf Basis der freien OpenLDAP[14] Software verwendet das LDAP Protokoll um Benutzerinformationen in einer baumartigen Struktur abzulegen und die einzelnen Nutzer anhand ihres hinterlegten Passworts zu authentifizieren. Auf dem Verwaltungsrechner wurde dem Master-Slave-Prinzip von OpenLDAP folgend eine Slave-Instanz eingerichtet, die ihren Datenbestand vom LDAP-Master des Lehrstuhlnetzwerks bezieht. Damit kann im Netzwerk des HPC-Rechners unabhängig vom Geophysiknetzwerk die Nutzerverwaltung abgewickelt werden.

DNS: Im Rechencluster werden private IP-Adressbereiche verwendet. Für diese können im zentralen DNS System des Münchner Wissenschaftsnetzes (MWN) keine Namen eingetragen werden, womit auch eine Namensauflösung nicht möglich wäre. Dies ist aber für die einfache Nutzung des Rechners durch die Anwender notwendig. Damit musste auf dem Verwaltungsrechner ein DNS Dienst eingerichtet werden, der für die Namensauflösung in der Domain `cluster.geophysik.uni-muenchen.de` verantwortlich ist.

[13] `http://www.debian.org`
[14] `http://www.openldap.org`

DHCP: Für den HPC-Rechner ist eine automatische Installation vorgesehen (siehe den weiter unten aufgeführten Punkt FAI). Innerhalb des Installationsprozesses müssen die einzelnen Rechenknoten eine IP-Adresse zugewiesen bekommen. Dies kann über den DHCP Dienst realisiert werden, in dem für jede MAC-Adresse genau eine IP-Adresse hinterlegt und dem Knoten während des Startvorgangs zugewiesen wird.

BOOTP und PXE: Jeder Rechenknoten ist in der Lage durch die eingebauten PXE-fähigen Netzwerkkarten über das Netzwerk ein Betriebssystem zu starten. Dazu werden ihm durch den BOOTP Dienst die notwendigen Dateien (Betriebssystemkern) für den Start zur Verfügung gestellt. Damit kann die automatische Installation mit FAI (siehe weiter unten) ohne die Verwendung sonstiger Medien durchgeführt werden, was eine enorme Erleichterung bei 80 Rechenknoten darstellt.

NFS: Der BOOTP Server stellt den Rechenknoten nur den für den Start notwendigen Betriebssystemkern und einige Parameter bereit. Sobald der Kern die Initialisierung der Hardware abgeschlossen hat, muss über das Netzwerk ein Dateisystem mit allen weiteren Daten/Informationen eingebunden werden. Dieses Netzwerkdateisystem kann über das NFS-Protokoll (siehe Seite 133) jedem Knoten zur Verfügung stehen. Die Freigaben enthalten ebenfalls alle für die Installation benötigten Dateien. Der damit eingerichtete zentralisierte Aufbau ermöglicht eine erhebliche Reduktion des sonst notwendigen Administrationsaufwands.

FAI: Für eine automatische Installation können verschiedene Systeme/Softwareprodukte verwendet werden. Eine freie Implementierung für Debian GNU/Linux basierte Rechner bietet das FAI-Projekt[15] an. Dazu werden die bereits genannten und erklärten Dienste benötigt. Das Konzept hinter FAI basiert auf Klassen. Jedem Rechner können anhand verschiedenster Merkmale zu Beginn der Installation eine oder mehrere Klassen zugewiesen werden. Diese beeinflussen danach den weiteren Verlauf der Installation, indem zum Beispiel eine andere Festplattenaufteilung oder Softwareauswahl verwendet wird. Mit diesem Konzept lassen sich in kurzen Zeiträumen komplexe Installationen verwirklichen. Da alle Rechenknoten identisch eingerichtet sein sollen, ist die erstellte FAI-Konfiguration einfach strukturiert. Diese Art der zentralisierten Installation vieler Rechner wurde später, auf Grund der guten Erfahrungen, für das Lehrstuhlnetzwerk der Geophysik übernommen. Beispielsweise ist es möglich, innerhalb von weniger als zwei Stunden den gesamten Rechencluster neu zu installieren. Bei sorgfältiger Planung kann der Wechsel zu einer neueren Version von Debian GNU/Linux in zwei bis drei Tagen durchgeführt werden.

[15] `http://www.informatik.uni-koeln.de/fai`

Welche Softwareprodukte auf einem HPC-Rechner benötigt werden, hängt im Wesentlichen von den darauf laufenden Anwendungen ab. Alle in Tabelle 2.3 aufgeführten Anwendungen sind Eigenentwicklungen durch Wissenschaftler am Lehrstuhl für Geophysik oder in Kooperation mit anderen geowissenschaftlichen Einrichtungen entwickelte Software. Es kommen keine kommerziellen Simulationsprogramme zum Einsatz. Dies bedeutet für die zu installierende Software, dass vorwiegend Entwicklungssoftware benötigt wird. Es sind vor allem hochoptimierende Compiler unabdingbar, da sie durch die angewendeten Optimierungen in der Lage sind, die Laufzeiten der Simulationen erheblich zu verkürzen. Weiterhin wird Software benötigt mit deren Hilfe ein paralleler Programmablauf möglich ist. Diesbezüglich existieren verschiedene Standards, von denen das „Message Passing Interface“ (MPI) die größte Verbreitung hat und auch von allen Anwendungen genutzt wird. Für diese beiden Bedarfsfälle sollte nach Möglichkeit Software verwendet werden, die keine oder nur geringe Kosten verursacht. Als Umsetzung des MPI-Standards Version 1 bietet sich die freie Software MPICH[16] (Gropp et al., 1996; Gropp & Lusk, 1996) an. Sie kann mit der installierten GNU Compiler Collection (GCC[17]) für das Debian System übersetzt werden. Auf Grund der guten Benchmarkergebnisse für TERRA und da AMD keine eigenen Compiler bereitstellt, wurden als optimierende Compiler die von Intel entwickelten ausgewählt. Ein positiver Nebeneffekt ist die Tatsache, dass für die Intel Compiler eine kostengünstige Lizenzierung über das LRZ möglich ist. Mit der Version 9.1 der Intel C/C++ und Fortran Compiler wurde ein weiteres Mal MPICH übersetzt und im System als Debian Paket installiert. Den MPI-Standard in Version 2 setzt OpenMPI[18] (Gabriel et al., 2004; Graham et al., 2006) um und wurde ebenso als Debian Paket installiert, womit zwei verschiedene MPI-Implementierungen zur Verfügung stehen. Die Fehlererkennung in parallelen Programmen auf dem HPC-Rechner wird durch den installierten Totalview Debugger stark erleichtert.

Normalerweise werden auf Rechenclustern auch Jobverwaltungssysteme eingerichtet. Diese sorgen für eine optimale Auslastung des Gesamtsystems, in dem sie, die vom Nutzer in die verschiedenen Warteschlangen eingereihten Rechenaufträge, möglichst optimal auf die verfügbaren Rechenknoten verteilen. In einer Nutzerbesprechung nahmen jedoch die Anwender von einem solchen System Abstand. Stattdessen wurde eine unkompliziertere Art etabliert. Zu diesem Zweck enthalten zwei Dateien disjunkte Mengen von Rechenknoten. Damit kann der Nutzer die Knotenauswahl für MPI steuern und die parallelisierte Simulationsanwendung auf dem HPC-System verteilen. Dies ermöglicht eine Verwendung des HPC-Rechners ähnlich zu einem Arbeitsplatzrechner, hat aber den Nachteil einer nicht optimalen Auslastung. Jeder Nutzer muss vor dem Starten des Rechenauftrags auf dem Überwachungssystem (siehe `cacti` im nächsten

[16] `http://www-unix.mcs.anl.gov/mpi/mpich1`
[17] `http://gcc.gnu.org/`
[18] `http://www.open-mpi.org`

Absatz) prüfen, ob die gewählten Rechner frei sind. Für einen zukünftigen Rechencluster muss allerdings eine Jobverwaltung eingerichtet werden, da die Vorteile eines solchen die wenigen Nachteile überwiegen.

Das Management der Rechenknoten wird über freie Software verwirklicht. Dazu werden zum einen Standardprogramme verwendet und zum anderen spezialisierte Software genutzt. Für das Einspielen von Softwareaktualisierungen in das Debian System kann über `ssh` („Secure Shell“) auf den Rechenknoten das entsprechende von Debian gelieferte Programm (`apt-get` oder `aptitude`) gestartet werden. Der Verwaltungsrechner dient dabei als lokal im Cluster vorhandener Spiegel der Debian Pakete. Die Hardwareüberwachung der Rechenknoten wird zweigeteilt durchgeführt. Über die IPMI-Schnittstelle jedes Rechenknotens können alle Parameter mit Hilfe des Programms `ipmitool` ausgelesen werden. Dazu zählen zum Beispiel das „System Event Log“, die Spannungen und Lüfterdrehzahlen. Das Betriebssystem selbst ist über `hdparm` in der Lage die eingebaute Festplatte und deren SMART-Werte[19] zu überwachen. Sobald sich in den ständig gelesenen Werten der Festplatte Änderungen ergeben, wird der Systemverwalter entsprechend vordefinierter Regeln benachrichtigt. Die Überwachung der Auslastung gewisser Systemkomponenten wird mittels `cacti`[20] durchgeführt (siehe Abschnitt 2.1.4 und Abbildung 2.8). Dabei werden über das SNMP-Protokoll[21] in regelmäßigen Abständen die verschiedenen Rechenknoten nach vorher festzulegenden Parametern befragt. Anschließend werden diese in einer RRD-Datenbankdatei[22] gespeichert. Der Vorteil dieser Datenbank in einer Datei ist, dass mit zunehmendem Alter der Daten eine Mittelung und Ausdünnung erfolgt, sodass eine konstante Dateigröße erreicht wird. Die Überwachungslösung kann über eine Intranetseite angesprochen werden, womit sich jeder Anwender einen Überblick über die gegenwärtige Auslastung des Rechenclusters verschaffen kann und entsprechend den Anmerkungen im vorherigen Absatz die Rechenaufträge auf die Rechenknoten verteilen.

Die IPMI-Schnittstelle wird zusätzlich noch für das Ein- und Ausschalten der Rechenknoten verwendet. Dies ist zum Beispiel bei einer Neuinstallation erforderlich, wenn darüber hinaus die Art des Systemstarts geändert wird (Start von lokalen Medien oder über PXE). Das geordnete Herunterfahren der Rechner ist auch bei einem Klimaanlagendefekt unabdingbar. Über die in den USV verbauten Managementkarten wird der Verwaltungsrechner beim Überschreiten einer Temperaturschwelle durch eine SNMP-Nachricht informiert. Dieser startet sofort ein `SHELL` Programm, welches durch IPMI jeden Rechenknoten herunterfährt. Im Detail bedeutet das, dass dem Betriebssystem ein ACPI-Ereignis durch die IPMI-Schnittstelle übermittelt wird, worauf das Betriebssystem mit dem geordneten Herunterfahren beginnt. Sobald auf allen

[19] Self-Monitoring, Analysis and Reporting Technology – kurz SMART
[20] `http://www.cacti.net`
[21] Simple Network Management Protocol
[22] Round Robin Database

Rechenknoten das Ausschalten initiiert wurde, verschickt das `SHELL` Programm eine E-Mail an den Systemverwalter. Die Firma netplace Telematic GmbH in München stellt für ihre Kunden ein E-Mail zu SMS[23] System bereit. Damit wird der Systemverwalter zusätzlich per SMS benachrichtigt. Bei diesem radikalen Vorgehen steht der Schutz der Hardware an erster Stelle in dem Sinne, dass auf die laufenden Anwendungen keine Rücksicht genommen werden kann. Es musste in den mehr als 2 Jahren des Rechnerbetriebs bereits mehrfach angewendet werden, da häufiger Probleme mit der Klimaanlage auftreten. Die Unzuverlässigkeit der Klimatisierung ist bisher das größte Problem bei der Nutzung des Hochleistungsrechners TETHYS.

Probleme mit der FSC Primergy Rx220 Kleinere Schwierigkeiten mit den Rechenknoten vom Typ FSC Primergy Rx220 konnten bereits während der Inbetriebnahme des Rechenclusters festgestellt werden. Es handelte sich dabei um eine fehlerhafte Festplatte, Probleme an zwei Mainboards und mehrfach Defekte an Arbeitsspeichermodulen. Diese wurden zeitnah durch den technischen Support des Auftragnehmers behoben. Somit konnte der Betrieb des Rechenclusters fortgesetzt werden. Bei der Vielzahl an neuen Rechnern treten immer wieder vereinzelt Schäden auf, die aber im Allgemeinen unbedenklich sind, da das Verhalten zu Beginn der Nutzungsdauer realistischerweise erwartet werden muss.

Im Anschluss an die Betriebssysteminstallation auf allen Rechenknoten wurden vom Betreiber des HPC-Rechners verschiedene Testrechnungen (siehe Abschnitt 2.1.3) durchgeführt. Darin sollte im Wesentlichen die volle Funktionstüchtigkeit geprüft und der Nachweis der aus den Benchmarks im Vorfeld zu erwartenden Leistungsfähigkeit erbracht werden. Bei diesen Untersuchungen im Februar/März 2006 traten vermehrt Probleme mit dem Arbeitsspeicher auf. Am Anfang reagierte der Auftragnehmer, in Absprache mit dem Hersteller, mit dem Austausch der betroffenen Module. Laut „System Event Log“ (lesbar über die IPMI-Schnittstelle) handelte es sich zu Beginn immer um korrigierbare ECC-Speicherfehler[24]. Im weiteren Verlauf veränderte sich das Fehlerbild. Es traten gehäuft nicht korrigierbare ECC-Speicherfehler auf. Die Reaktion des Rx220 BIOS auf diesen Fehler ergibt einen sofortigen Reset des Rechners, der für die laufenden Rechnungen wie ein Systemabsturz aussieht. Auf Grund dieses dramatischen Zustands musste verstärkt an einer Lösung gearbeitet werden. Daher wurde die Problematik vom Auftragnehmer beim technischen Support von FSC eskaliert und mit zusätzlichem Personal an der Lösung gearbeitet. Doch erschien der technische Support nicht in der Lage das Problem einzugrenzen. Verschiedene Sitzungen mit dem Auftragnehmer und FSC sowie dem

[23] „Short Message Service“

[24] ECC ist ein Fehlerkorrekturverfahren, welches der Erkennung und Korrektur von Fehlern bei der Speicherung und Übertragung von Daten dient. Dazu werden vor der Speicherung und Übertragung den Daten zusätzliche Bits hinzugefügt, welche dann zur Bestimmung von Fehlern genutzt werden.

Lehrstuhl für Geophysik führten zu keiner Lösung. Zu diesem Zeitpunkt waren nach Ansicht des Auftraggebers und Betreibers keine weiteren Verzögerungen annehmbar, weshalb in einem Schreiben an den Vorstand von FSC rechtliche Schritte durch die LMU angekündigt wurden, falls die Probleme nicht umgehend beseitigt werden sollten. Gleichzeitig musste auf das unannehmbare Auftreten des technischen Supports hingewiesen werden. Innerhalb weniger Stunden führte dieser Schritt des Lehrstuhls für Geophysik zur intensiveren Fehlersuche durch FSC. Der technische Support arbeitete verstärkt an einer Lösung und konnte Anfang August 2006 eine solche vorstellen. Diese bestand in einer geänderten Initialisierung des Rechners noch bevor das eigentliche BIOS gestartet wird. Die Wirkung der als BIOS Aktualisierung verpackten Lösung konnte in einem sechswöchigen Test nachgewiesen werden. Im Einvernehmen aller Seiten wurde danach das Problem als behoben angesehen.

Die fast 6 Monate andauernde Lösungssuche stellte eine nicht unerhebliche Belastung für den kontinuierlich notwendigen Wissenschaftsbetrieb dar. Bedingt durch die ständigen Ausfälle der Rx220 kam es teilweise zu einer Einstellung der Nutzung des HPC-Rechners durch die Anwender. Damit wurden wissenschaftliche Arbeiten und notwendige Veröffentlichungen behindert. Aus diesem Grund wurde vom Betreiber und Auftraggeber nach einer Aufwandsentschädigung durch FSC gefragt. Das Resultat der sich anschließenden Verhandlungen mit FSC waren die fünf zusätzlich zur Verfügung gestellten Primergy Rx220, eine Primergy Rx200S3 gedacht als Server für das Lehrstuhlnetzwerk und zwei energieeffiziente Arbeitsplatzrechner.

Paralleles Dateisystem Nachdem die Probleme mit dem Arbeitsspeicher der Rechenknoten behoben waren, konnte über die Nutzung der noch freien Festplattenbereiche eines jeden Knotens nachgedacht werden. Diese Festplattenbereiche (auch als Partitionen bezeichnet) sollten den Nutzern des HPC-Systems zur Verfügung gestellt werden. Dazu können verschiedene Ansätze verfolgt werden. Ein solcher ist das Einbinden der Partitionen als ein lokales temporäres Dateisystem. In diesem ist es jedem Anwender möglich, seine Daten während oder auch nach der Simulation abzulegen. Es ändert sich aber an der bestehenden Fragmentierung der jeweils pro Knoten verfügbaren 120 GB Kapazität nichts. Vom Benutzer können keine Dateien erzeugt werden, die über den lokal verfügbaren Festplattenplatz hinausgehen. Zusätzlich ist es für jeden Nutzer nur über selbst zu erstellende Programme möglich, die herausgeschriebenen Simulationsergebnisse von den betroffenen Knoten einzusammeln. Da dieses Vorgehen umständlich für die Anwender ist, sollte ein anderer Ansatz genutzt werden. Bei diesem wird aus den einzelnen Partitionen jedes beteiligten Rechners ein großes paralleles Dateisystem generiert. Damit wird die Fragmentierung des Speicherplatzes aufgehoben, wodurch gleichzeitig jeder Nutzer die Möglichkeit erhält, größere Dateien zu erstellen. Durch den einheitlichen Zugriff, der auch von jedem Knoten aus gleich ist, entfällt die Notwendigkeit Programme zu schreiben die ein-

zig der Dateneinsammlung dienen. Den genannten Vorteilen stehen auch Nachteile gegenüber. Das Management eines solch großen Dateisystems ist keine triviale Aufgabe und erfordert eine exakte Planung. Tritt an einem Knoten ein Festplattendefekt auf, so ist häufig mit einem totalen Verlust der Daten zu rechnen, da diese parallelen Dateisysteme keine Möglichkeit der Redundanzbildung bieten. Auch ist das Dateisystem nicht nutzbar, wenn ein Rechenknoten nicht verfügbar ist. Dieser Sachverhalt muss den Benutzern des HPC-Rechners verdeutlicht werden, ansonsten bestünde die Möglichkeit, dass wichtige Daten auf dem parallelen Dateisystem abgelegt werden, wo sie stark gefährdet wären.

Mit der fortschreitenden Entwicklung der Rechencluster in den letzten Jahren, entstanden zunehmend mehr Implementierungen derartiger paralleler Dateisysteme. Alle diese Softwareprodukte haben das gemeinsame Anliegen, eine möglichst schnelle und effiziente parallele Ein- und Ausgabe (Input/Output oder auch I/O) zu entwickeln. Die über die Jahre entstandene Kluft zwischen der Rechen- und Netzwerkleistung der HPC-Rechner auf der einen Seite und der I/O-Leistung auf der anderen, kann über ein solches Dateisystem verkleinert werden. Die unterschiedlichen Herangehensweisen bei der Implementierung und die verfolgten Entwicklungsziele der einzelnen verfügbaren parallelen Dateisysteme legen bereits die Vermutung nahe, dass nicht jede Anwendung mit jedem Dateisystem die optimale Leistung erzielen kann. Deshalb sollte bereits im Vorfeld eine Sichtung/Auswahl erfolgen. Im wissenschaftlichen Umfeld werden oftmals freie Softwareprodukte bevorzugt, da diese bei dem vorherrschenden Kostendruck meist eine sehr gute Lösung darstellen.

Die verteilten Dateisysteme werden häufig auch als Cluster-Dateisysteme bezeichnet. Um eine Wahl treffen zu können, war es notwendig, zuerst einen Überblick über die verfügbaren parallelen Dateisysteme zu bekommen. Zu diesem Zweck wurden für das GNU/Linux Betriebssystem mögliche Kandidaten in der folgenden Liste zusammengetragen. Es besteht dabei nicht der Anspruch einer vollständigen Aufzählung, es sollen stattdessen die verbreitetsten Implementierungen aufgelistet werden.

OCFS2: Das „Oracle Cluster File System“[25] bietet ab Version 2 die Möglichkeit, neben Oracle Datenbanken auch normale Dateien abzuspeichern. Damit scheint das POSIX[26] kompatible Dateisystem als Cluster-Dateisystem geeignet. Allerdings konnten zu diesem Zeitpunkt in einer Internetrecherche keine entsprechenden Verwendungsinformationen gefunden werden, was an einer entsprechenden Nutzung Zweifel aufkommen lässt.

[25] `http://oss.oracle.com/projects/ocfs2/`

[26] Unter Portable Operating System Interface (POSIX) versteht man ein gemeinsam von der IEEE und der Open Group für Unix entwickeltes standardisiertes Application Programming Interface (API), welches zwischen den Applikationen und dem Betriebssystem die Schnittstelle darstellt.

GPFS: Von IBM wird das „General Parallel File System“[27] angeboten. Es wurde von Anfang an als Cluster-Dateisystem entwickelt und verfügt über eine breite Nutzerbasis sowie genügend Referenzinstallationen. Die Nutzung dieses Dateisystems hätte den Erwerb von entsprechenden Lizenzen erfordert, was in Anbetracht des ausgeschöpften finanziellen Rahmens nicht möglich war.

GFS: Das „Global File System“[28] wird vom Linux Distributor Red Hat entwickelt und vermarktet. Dennoch fallen nur beim Abschluss entsprechender Serviceverträge weitere Kosten bei der Nutzung an. Vom Lehrstuhl für Theoretische Physik der Ludwig-Maximilians-Universität wurde im selben Zeitraum ein vergleichbarer Rechencluster beschafft. Auf diesem installierte der Lehrstuhl für Physik GFS und konnte auch nach langen Testreihen und Konfigurationsänderungen keine zufrieden stellende Leistung erreichen.

Lustre: Das „Lustre“[29] Dateisystem wurde im Jahre 2006 noch von einer eigenständigen gleichnamigen Firma entwickelt und vermarktet. Entsprechende Lizenzkosten fielen nur bei abgeschlossenen Serviceverträgen an. Die frei verfügbare Version des Dateisystems ist in ihrer Funktionalität ein wenig eingeschränkt gegenüber der kommerziell vertriebenen Variante. Durch den in der Zwischenzeit erfolgten Kauf der Firma Lustre durch Sun Microsystems hat sich dies aber zum Positiven für die Anwender entwickelt. Das LRZ führte für die von ihnen betriebenen HPC-System eigenständige Untersuchungen zur Leistungsfähigkeit der verteilten Dateisysteme durch. Dabei erreicht das Lustre Dateisystem nach Auskunft des LRZ sehr gute Werte und wird mittlerweile standardmäßig eingesetzt.

PVFS: Das „Parallel Virtual File System“[30] (Carns et al., 2000) wurde bereits in den vorherigen Rechenclustern von Prof. Bunge verwendet. Die dadurch gewonnenen Erfahrungen waren durchweg positiv. Das LRZ wies darauf hin, dass bei kleinen Dateien laut seiner Tests eine schlechtere Skalierung zu erwarten ist, aber PVFS im Vergleich zu Lustre einfacher zu konfigurieren ist.

Alle hier aufgeführten Dateisysteme bieten vorbereitete Softwarequellpakete für GNU/Linux an. Der Installationsaufwand unterscheidet sich aber teilweise erheblich. Für alle Produkte muss entweder der Betriebssystemkern angepasst oder ein Kernelmodul für diesen übersetzt werden. Darüber hinaus sind noch die notwendigen Programme und Bibliotheken für den Zugriff auf das

[27] `http://sourceforge.net/projects/gpfs/`
[28] `http://sources.redhat.com/cluster/gfs/`
[29] `http://www.lustre.org`
[30] `http://www.pvfs.org`

verteilte Dateisystem zu installieren. Die Anpassung an die eigenen Bedingungen wird bei allen in den entsprechend zu verteilenden Konfigurationsdateien durchgeführt. Bereits fertige, für die einfache Installation vorbereitete, binäre Softwarepakete existieren für Debian GNU/Linux Sarge von keinem der Produkte. Für alle ist demnach mehr oder weniger Aufwand notwendig, wobei allerdings für PVFS ältere Versuche Debian Pakete zu erstellen im Internet verfügbar sind. Die dazugehörigen Quellen können somit als Basis für individuelle Anpassungen dienen. Die bisher aufgeführten Überlegungen, die Erkenntnisse aus Woitaszek et al. (2005) und die begrenzt zur Verfügung stehende Zeit führten zum Entschluss, in der ersten Ausbaustufe von TETHYS, PVFS den Vorzug zu geben. Bei einer anstehenden Aktualisierung der Debian GNU/Linux Distribution sollten aber entsprechende Tests mit den Anwendungen unter reellen Bedingungen zumindest auf Lustre und PVFS Dateisystemen durchgeführt werden.

Wie alle genannten verteilten Dateisysteme verfügt PVFS über eine standardisierte Dateisystemschnittstelle, die über ein Kernelmodul realisiert wird. Mit dieser ist es den Anwendern möglich, die normalen Dateisystembefehle für den Zugriff auf die Daten zu benutzen (wie zum Beispiel `ls`, `cp`, `mv` usw.). Die zu speichernden Dateien werden in entsprechende Dateisegmente aufgeteilt und an die zur Verfügung stehenden PVFS-Datenserver verschickt. Diese legen die Informationen auf dem in der Konfigurationsdatei festgelegten Speicherplatz ab, welcher bei TETHYS der noch freien Partition entspricht. Der Festplattenbereich kann mit einem Dateisystemformat der eigenen Wahl eingerichtet werden, da PVFS die Daten in einer eigenen Datei darauf ablegt. Die Metainformationen (Benutzer- und Gruppenzugehörigkeit oder Dateirechte und Zeitstempel) werden von einem PVFS-Metadatenserver verwaltet, der durch einen einzelnen Rechner oder einen Rechnerverbund repräsentiert werden kann. Die aktuelle PVFS-Konfiguration für TETHYS enthält für die verwendete Version 1.5.1 neun Metadatenserver und 81 Datenserver (dies beinhaltet die Gesamtanzahl an Rechenknoten plus den Verwaltungsrechner). Da für die Anzahl der Metadatenserver keine Richtlinie in der PVFS-Dokumentation gefunden werden konnte, wurden jeweils zwei Rechner pro Cluster Node Switch zusammen mit dem Verwaltungsrechner als Metadatenserver eingerichtet. Damit erreicht das für die Ablage von temporären Dateien eingerichtete PVFS-Dateisystem eine Größe von 9.7 TB.

2.1.3 Benchmarks

Nach der Inbetriebnahme von TETHYS galt es, die Leistungsfähigkeit des HPC-Rechners zu prüfen. Es sollte zum einen geklärt werden, ob die in den Benchmarks erreichten Laufzeiten (siehe Kapitel 2.1.1) mit dem beschafften Rechencluster bestätigt werden können. Zum anderen galt es, den Einfluss der Netzwerktopologie auf die Performance genauer zu betrachten. Des Weiteren sollte für die Hauptanwendung TERRA des Rechners die Skalierung genauer untersucht werden.

Die bei den im Vorfeld der Beschaffung durchgeführten Benchmarks verwendeten Anwendungen TERRA und YAC wurden erneut für die Laufzeituntersuchungen mit der identischen Konfiguration eingesetzt (siehe dazu die Seiten 13 und 17). Allerdings kam bei der Übersetzung der Programme nicht mehr der Intel Compiler in Version 8.1 zum Einsatz, sondern die neuere Version 9.0. Standardmäßig wird in den `makefiles` für TERRA und YAC nur die Compiler Option `-O3` eingesetzt. Für die MPI-Unterstützung fand auch auf den Knoten des TETHYS Rechenclusters MPICH in Version 1.2.5 Verwendung. Damit war zumindest der Unterschied im Setup der Laufzeituntersuchungen zu TERRA nur marginal. Es wäre zu erwarten, dass es zu einer Reduktion der Laufzeiten kommt, da durch die neuere Compiler Version bessere Optimierungen angewendet werden können.

TERRA

Für die Laufzeitmessungen mit TERRA auf den AMD Opteronsystemen von FSC und des Lehrstuhls für Rechnertechnik und Rechnerorganisation / Parallelrechnerarchitektur der TUM kamen Einzel-, Vier- und Sechzehn-Prozessorkonfigurationen zum Einsatz. Es galt für diese, die Laufzeiten auf TETHYS für jeweils 100 Zeitschritte der Mantelkonvektionssimulation mit der Gittergröße MT=64 zu bestimmen. Bei dieser Simulationsrechnung sind insgesamt 7 Millionen Freiheitsgrade vorhanden. Die ermittelten Zeiten sind in Tabelle 2.19 zusammengestellt. Für verschiedene Kombinationen aus Prozessor-, Knoten- und Switchanzahl wird darin auch der aus den angegebenen Laufzeitmittelwerten errechnete parallele Speedup aufgeführt. Mittels unterschiedlicher Kombinationen sollte ein Gefühl für den Einfluss der Zweiprozessorknoten und der Netzwerktopologie (siehe Abbildung 2.5) auf TERRA entwickelt werden. Beispielsweise ist es möglich, acht MPI-Prozesse auf vier Rechenknoten an einem Switch, auf vier Rechenknoten an vier Switches, auf acht Rechenknoten an einem Switch oder auf acht Rechenknoten an vier Switches zu verteilen. Wenn zwei MPI-Prozesse auf einem Rechenknoten laufen, so treten sie unter Umständen in Konkurrenz um Rechnerressourcen. Werden vier statt ein Switch verwendet, so verändert sich der Nachrichtenweg zwischen den Rechenknoten und die Über-

tragungsrate zwischen den vier Cluster Node Switches ist auf 10-GBit beschränkt. Dies gilt es genauer zu betrachten.

Zunächst soll allerdings die Laufzeit für den Einzelprozessorlauf betrachtet werden. Auf den Rechnern von FSC wurden dafür mit dem AMD Opteron 252 $\bar{t} = 777\ s$ und mit dem AMD Opteron 275 $\bar{t} = 842\ s$ ermittelt. Beide Laufzeiten liegen unter der für TETHYS mit 851 s. Im Vergleich zum AMD Opteron 275 bedeutet dies 1 Prozent mehr Laufzeit, trotz 200 MHz höher getakteter CPU. Zieht man die 890 s auf dem AMD Opteron 850 (gleiche Taktfrequenz wie die CPUs in TETHYS) des LRR der TUM in die Betrachtung ein, so wurden unter Umständen von FSC nicht weiter spezifizierte Compileroptionen aktiviert. Eine genauere Untersuchung dieses Sachverhaltes bedürfte der Kenntnis dieser Optionen. Die HPC-Gruppe von FSC führte auch einen Test mit den in TETHYS verbauten AMD Opteron 250 CPUs durch. Bei vier MPI-Prozessen verteilt auf zwei Rechnern konnte eine Laufzeit von 325 s ermittelt werden. Dieses Testsetup entspricht der in Tabelle 2.19 aufgeführten Kombination von vier Prozessoren, zwei Rechenknoten und einem Switch. Die Laufzeit dafür beträgt 313 s und liegt somit zwischen der für die FSC Rechner und den 297 s auf dem HPC-System des LRR an der TUM. Die 97 s Laufzeit auf 16 AMD Opteron 850 Prozessoren des LRR Rechenclusters sind mit den 92 s auf TETHYS vergleichbar (acht Rechenknoten und vier Switches). Betrachtet man die Laufzeiten aus Tabelle 2.19 genauer, so fällt auf, dass mit mehr eingesetzten Switches keine Ver-

Prozessoranzahl P	Knotenanzahl	Switchanzahl	$\bar{t}\ [s]$	$S_{par}(P)$
1	1	1	851	1
4	2	1	313	2.7
	2	2	321	2.7
	4	1	227	3.8
	4	4	227	3.8
8	4	1	142	6.0
	4	4	141	6.0
	8	1	104	8.2
	8	4	104	8.2
16	8	1	93	9.2
	8	4	92	9.3
	16	1	60	14.1
	16	4	63	13.6

Tabelle 2.19: TERRA: Benchmark auf TETHYS für die Gittergröße MT=64 und 100 Zeitschritte – untersucht wird der Einfluss der Netzwerktopologie und der Anzahl an Prozessen pro Rechenknoten

besserung oder Verschlechterung der Laufzeit einhergeht. Die Verdoppelung der Rechenknoten bei gleicher Prozessoranzahl bringt hingegen eine deutliche Leistungssteigerung. Es wäre daher notwendig, diesen Sachverhalt mit geeigneten Analysewerkzeugen genauer zu betrachten. Mögliche Erklärungen wären, dass die Datenkommunikation innerhalb eines Rechenknotens mit dem eingesetzten MPICH nicht optimal verläuft oder die Rechenknoten in Konkurrenz um Rechnerressourcen stehen. Mit Blick auf kommende Mehrkernprozessoren sollte daher über eine mögliche Umstellung bzw. Optimierung in TERRA nachgedacht werden.

YAC

Entgegen den Benchmarks mit dem Programm YAC auf den Rechnern der HPC-Gruppe von FSC konnte der PGI Compiler zum Übersetzen des Quelltextes nicht verwendet werden. Damit sind diese Ergebnisse nicht direkt mit denen auf TETHYS unter Nutzung des Intel Compilers vergleichbar. Das verwendete Modell für die Testläufe entspricht dem bereits auf Seite 17 eingeführten und basiert auf etwa 3.6 Millionen Punkten im finite Differenzengitter. Die Laufzeiten für die Ein-, Zwei- und Vierprozessoruntersuchungen sind in Tabelle 2.20 zusammengestellt. Auch bei YAC wurde die Anzahl an Rechenknoten und Switches variiert, um eine Vorstellung vom Programmverhalten zu entwickeln.

Die Laufzeit von 3461 s auf TETHYS liegt deutlich über den 2306 s des Einzelprozessorlaufs bei FSC und zeigt damit eindrucksvoll die Optimierungsmöglichkeiten auf. Vergleicht man allerdings die Laufzeiten für vier MPI-Prozesse, so sind die 805 s auf den AMD Opteron 275 Rechnern von FSC mit den 727 s auf TETHYS vergleichbar (zwei Rechenknoten und ein Switch), da die AMD Opteron 275 CPUs um 200 MHz langsamer getaktet sind. Die Werte für

Prozessoranzahl P	Knotenanzahl	Switchanzahl	$\bar{t}\,[s]$	$S_{par}(P)$
1	1	1	3461	1
2	1	1	1244	2.8
	2	1	1113	3.1
	2	2	1108	3.1
4	2	1	727	4.8
	2	2	714	4.8
	4	1	582	5.9
	4	4	584	5.9

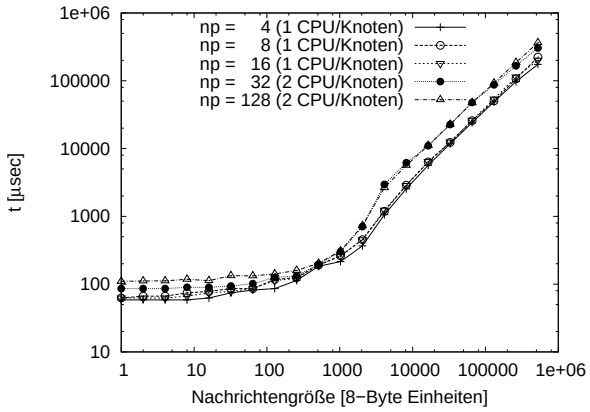
Tabelle 2.20: YAC: Benchmark auf TETHYS – untersucht wird der Einfluss der Netzwerktopologie und der Anzahl an Prozessen pro Rechenknoten

den Speedup zeigen eine überlineare Skalierung, welche sich aber wieder relativiert, könnte die Einzelprozessorlaufzeit gesenkt werden. Dies bedürfte allerdings einer genaueren Analyse des Programms. Wie bereits für TERRA zu sehen, zeigt sich auch bei dieser Untersuchung, dass kein Einfluss der Netzwerktopologie auf die Laufzeiten zu verzeichnen ist. Eine Verringerung der Laufzeit kann aber, ebenso wie für TERRA, durch den Einsatz von gleich vielen Rechenknoten wie MPI-Prozessen erreicht werden. Das finite Differenzen Programm zur Simulation der seismischen Wellenausbreitung ist demnach auch eine Anwendung für mögliche Optimierungen bei Mehrkernsystemen.

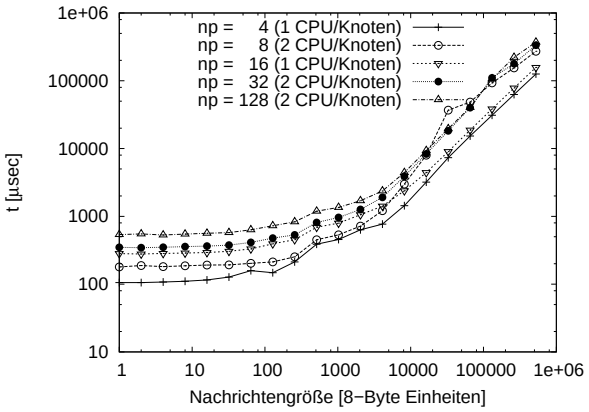
Intel MPI-Benchmark (IMB)

Der Kommunikationsdurchsatz für das eingesetzte GBit-Ethernet Datennetzwerk kann mit Hilfe des Intel MPI-Benchmark (IMB)[31] untersucht werden. Die Voraussetzung für die Nutzung dieses Benchmarks ist eine vorhandene MPI-Implementierung in Version 1 und ein C Compiler. Auf TETHYS wurden dafür MPICH in Version 1.2.5 und der Intel Compiler in Version 9.0 verwendet. Die bei der „Domain Decomposition“ mit parallelen Mehrgitterlösern auftretenden Kommunikationsschemata können exemplarisch mit den im IMB enthaltenen „Exchange“ und „Allreduce“ Tests nachgestellt werden. Im „Exchange“ Test wird angenommen, dass die Prozesse eine periodische Prozesskette bilden. Jeder Prozess tauscht dabei Daten mit den beiden Nachbarn in der Kette aus. Dies ist vergleichbar mit der Aktualisierung von inneren Randwerten, zum Beispiel nach dem Glättungsschritt beim Mehrgitterverfahren. Der „Allreduce“ Test hingegen steht für gemeinschaftliche Kommunikation aller Prozesse. Die korrespondierende `MPI_ALLREDUCE` Methode sammelt alle Daten von allen Prozessen ein und führt mit diesen eine vorher definierte Operation aus (beispielsweise `MIN`, `MAX` oder `MULT`). Anschließend wird das Ergebnis zurück an alle Prozesse gesendet. Bei der Berechnung von Residuennormen in parallelen iterativen Verfahren müssen entsprechende Methoden verwendet werden. Weiterführende Informationen können der Dokumentation des Benchmarks entnommen werden. Die Abbildungen 2.6(a) und 2.6(b) zeigen die Ergebnisse dieser Benchmarks auf TETHYS, jeweils für unterschiedliche Nachrichtengrößen und zunehmende Anzahl an Prozessoren (np). Die ermittelte Performance spiegelt die zu erwartende Kommunikationsleistung per TCP/IP über ein GBit-Ethernet Netzwerk wider. Es ist auch eine Skalierung mit zunehmender Nachrichtengröße zu beobachten. In den ersten vier Fällen (np=4, 8, 16, 32) werden nur Rechenknoten eines Switches benutzt, wohingegen bei 128 CPUs (np=128) alle vier Cluster-Node-Switches (CNS) Verwendung finden und damit auch der zentrale Cluster Core Switch (siehe Abbildung 2.5). In Ab-

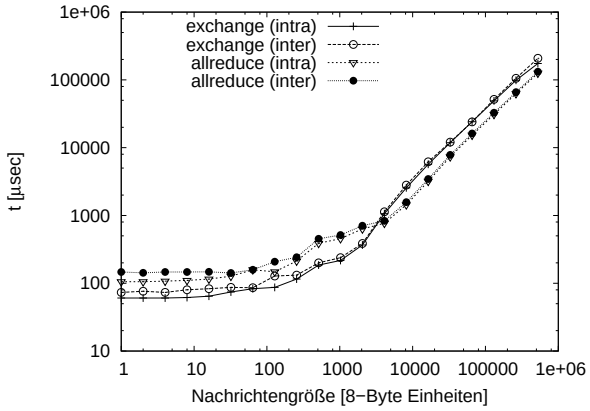
[31] Bezug des Softwarepakets über `http://www3.intel.com/cd/software/products/asmo-na/eng/219848.htm`



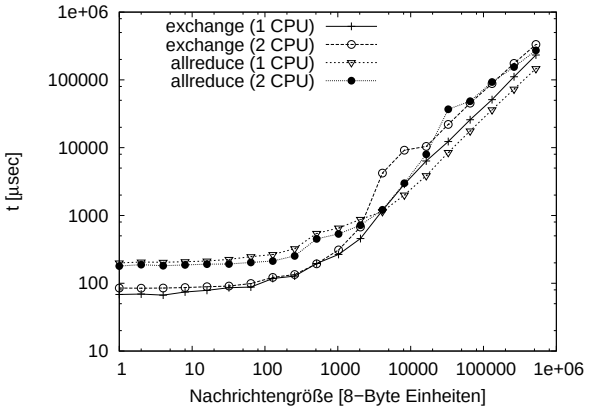
(a) Performance für den „Exchange“ Benchmark



(b) Performance für den „Allreduce“ Benchmark



(c) Vergleich zwischen intra und inter Cluster Node Switch Kommunikation



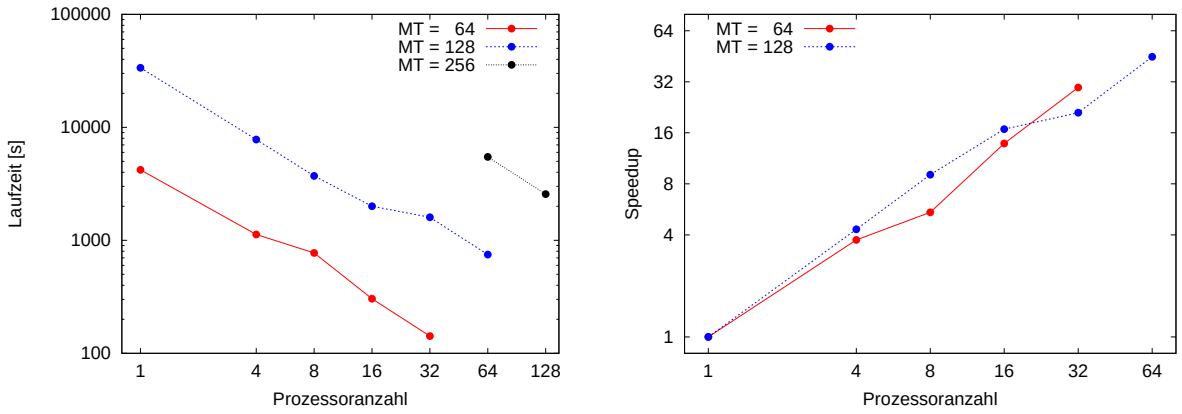
(d) Einfluss der Zweiprozessorknoten auf die Kommunikationsperformance

Abbildung 2.6: Ergebnisse des IMB Benchmark auf dem HPC-Rechner TETHYS

bildung 2.6(c) wird der Vergleich zwischen einer Konfiguration, bei der vier Prozesse entweder auf vier Rechenknoten an einem CNS (intra) oder auf allen vier CNS (inter) laufen, dargestellt. Die gleichwertige Skalierung für beide Konfigurationen bestätigt, dass die hierarchische Netzwerktopologie fehlerfrei funktioniert und keine offensichtlichen Engpässe zeigt. Der Effekt bei der Nutzung der Rechenknoten als Zweiprozessormaschinen wird in Abbildung 2.6(d) gezeigt. In diesem Test werden zum einen acht CPUs auf acht Rechenknoten und zum anderen acht Prozessoren auf vier Rechenknoten verwendet. Aus den Graphen können geringe Unterschiede insbesondere beim „Exchange“ Benchmark mit verschiedenen Nachrichtengrößen abgelesen werden. Dieser Effekt könnte durch den Einsatz einer für SMP optimierten Kommunikationsschicht in MPICH behoben werden.

Skalierungstest für TERRA

Im folgenden Test soll die Performance von TERRA auf dem Rechencluster TETHYS im Mittelpunkt stehen. Dazu wird mit drei verschiedenen Diskretisierungen (MT=64, 128 und 256) die Laufzeit für unterschiedliche Prozessoranzahlen bestimmt. Aus den angegebenen Gitterdiskretisierungen ergeben sich Problemgrößen mit 1.4, 10 und 85 Millionen Gitterpunkten, was einem Gitterpunkt aller 100, 50 und 25 km auf der Erdoberfläche entspricht. Die Abbildung 2.7(a) stellt die Laufzeiten für diese Simulationen mit jeweils 500 Zeitschritten für variierende Prozessoranzahlen graphisch dar, wobei die Zeit für den Einzelprozessorlauf für MT=128 auf Grund des begrenzten Arbeitsspeicher extrapoliert wurde. Aus den dargestellten Graphen kann eine annähernd linear abfallende Laufzeit abgelesen werden. Dieser lineare Zusammenhang zwischen der Anzahl an Prozessoren und der Laufzeit spiegelt sich auch in Abbildung 2.7(b) wider, wobei der parallele Speedup für die Gittergrößen MT=64 und 128 angegeben wird. Für einen Geowissenschaftler ist der so genannte Scaleup viel interessanter als der dargestellte Speedup. In diesem Zusammenhang sei erwähnt, dass für MT=128 mit 16 Prozessoren eine Laufzeit von 2002 s ermittelt wurde und für MT=256 mit 128 CPUs eine Zeit von 2564 s, wobei in beiden Fällen eine vergleichbare Rechenlast pro Prozessor entsteht (siehe dazu auch Abbildung 2.7(a)). Basierend auf den notwendigen 8000 Flop pro Gitterpunkt und Zeitschritt ergibt sich für die Problemgröße MT=256 mit 128 CPUs eine System Performance von 140 GFlops. Dies entspricht circa 22 % der theoretisch erreichbaren 624 GFlops mit den 128 in TETHYS verbauten AMD Opteron 250 Prozessoren. Durch Optimierungen für die AMD Opteron Architektur kann dieser Wert noch erhöht werden.



(a) Laufzeiten für variierende MT

(b) Speedup auf TETHYS

Abbildung 2.7: TERRA: Laufzeiten für verschiedene Problemgrößen (MT) in Abhängigkeit zur Anzahl an Prozessoren und der daraus berechnete parallele Speedup

2.1.4 Auslastung

Betreibt man einen Rechencluster der Größe von TETHYS, so muss man sich stets die Frage nach der Nutzung stellen. Man sollte keine subjektiven Einschätzungen zur Auslastung des Systems geben, sondern nach Möglichkeit realistische Zahlen zur Nutzung heranziehen. Dies wäre bei der Verwendung einer Auftragsverwaltungssoftware durch deren eingebaute Statistiken sicherlich einfach durchführbar. Für TETHYS wurde entschieden keine solche Software einzusetzen, womit auch keine Nutzungsstatistiken verfügbar sind. An Stelle dessen kann aber die für das HPC-System eingesetzte Überwachungssoftware `cacti` verwendet werden (siehe dazu Seite 35 ff.). Mit Hilfe eines Hintergrundprozesses fragt die Software `cacti` alle fünf Minuten systemrelevante Daten von jeden Rechenknoten ab und speichert sie anschließend in einer RRD-Datenbankdatei. Das besondere an dieser Datei ist deren konstante Größe. Je älter die Daten in der RRD-Datei werden, desto stärker werden sie gemittelt. Dies ist für Nutzungsstatistiken ausreichend, da meist nur die aktuellsten Daten mit einer hohen zeitlichen Auflösung benötigt werden. Auftretende Trends über längere Zeiträume gehen in der Mittelung nicht unter, da ihre Periode meist größer als die Zeiträume für die Mittelwertbildung ist.

In Abbildung 2.8 sind exemplarisch die Auslastung der CPUs, die Zahl aktiver Prozesse und die Auslastung des Datennetzwerks für einen Rechenknoten dargestellt. Der ausgewählte Knoten A02 unterscheidet sich in der Nutzung unwesentlich von anderen. Der dargestellte Zeitraum vom 01.07.2006 bis zum 31.08.2008 umfasst die durch `cacti` verfügbare Zeitspanne. Betrachtet man die einzelnen Darstellungen genauer, so fallen immer wieder Bereiche mit keinen oder sehr wenigen Daten auf, beispielsweise Anfang Juli 2006, Ende September 2006, Anfang April 2007 und den gesamten Mai bis Anfang Juni 2008. Diese leeren Abschnitte entstanden durch die notwendige Abschaltung des Rechenclusters auf Grund von Problemen mit der Klimaanlage. Die äußerst lange Abschaltungsphase Mitte des Jahres 2008 wurde durch einen Defekt der Regelelektronik in einem der Klimaschränke verursacht. Mit einem Klimaschrank können nur die zentralen Server des Lehrstuhls sicher betrieben werden. Die für die Reparatur beauftragte Firma war durch Lieferengpässe des Herstellers nicht in der Lage, den Defekt frühzeitiger zu beheben. In den anderen Phasen konnten die Probleme eher behoben werden. Damit zeigt sich, dass nicht der Betrieb des Rechencluster die größte Herausforderung darstellt, sondern die kontinuierliche Klimatisierung des Raums.

Aus der Darstellung für die Auslastung der CPUs (siehe Abbildung 2.8) können als Nutzungsmittelwert durch die Anwender (User) 28.5 Prozent und für das System 16.5 Prozent abgelesen werden. Dies ergibt in Summe 45 Prozent Auslastung des Systems. Der Wert ist nicht zufrieden stellend und muss verbessert werden. Während der Erstellung dieser Graphen konnte allerdings auch ein noch nicht verstandener Effekt beobachtet werden, bei dem die Darstellung nicht die reale Auslastung durch die Nutzer (User) zeigte, da das gleichzeitig gestartete Sy-

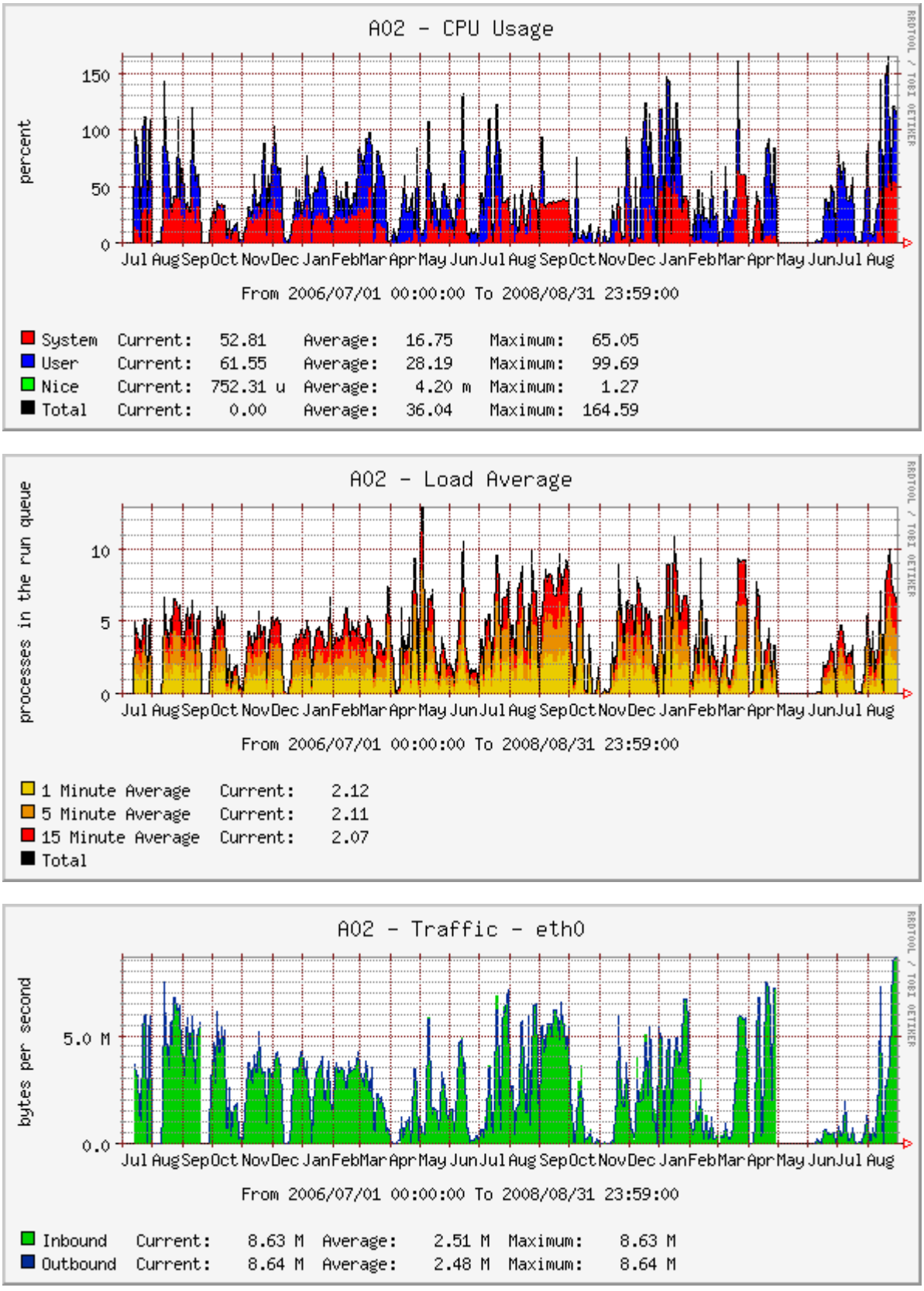


Abbildung 2.8: Graphen für die Auslastung der CPUs, der Zahl aktiver Prozesse und der Auslastung des Datennetzwerks für den Rechenknoten A02

stemprogramm `top` eine nahezu volle Nutzung des Rechners aufwies. Dieser Eindruck wird auch durch die mittlere Darstellung in Abbildung 2.8 (Anzahl aktiver Prozesse des Systems) bestätigt. Der Mittelwert gibt knapp zwei aktive Prozesse an. Wenn man davon ausgeht, dass ein Systemprozess vorhanden ist, so läuft mindestens noch ein Anwendungsprozess, also eine Simulation. Dies spricht für eine bessere Auslastung der Rechenknoten. Die Graphen für die Auslastung des Datennetzwerks bestätigen mit dem Mittelwert von etwa 2.5 MB/s die bereits gewonnene Erkenntnis, dass die Netzwerktopologie für den Rechencluster keinen Engpass darstellt. Mit dem Maximum von 7.5 MB/s bleibt die Übertragungsrate im Datennetzwerk deutlich unter den maximal möglichen 125 MB/s. Es kann also zusammenfassend gesagt werden, dass der Rechencluster sehr gut akzeptiert und genutzt wird. Dies verdeutlicht auch den Bedarf an Rechenzeit am Lehrstuhl für Geophysik.

2.2 Cluster Design in the Earth Sciences: TETHYS

Computational modeling is a powerful tool in the Earth Sciences. In the solid Earth important simulation areas include seismic wave propagation, rupture and fault dynamics in the lithosphere, creep in the mantle, and magneto-hydrodynamic flow linked to magnetic field generation in the core. These problems rank among the most demanding calculations computational physicists can perform today. They exceed the limitations of the largest high-performance computing systems by a factor of ten to one hundred measured both in terms of the demands on capacity and capability of systems. Off-the-shelf high-performance Linux clusters are useful to ease the limitations in capacity computing by exploiting price advantages in mass produced PC hardware. Here we review our experience of building a 128 processor AMD Opteron Gigabit Ethernet Linux cluster. The machine is operated at the scientific department level, targeted directly at large-scale geophysical and tectonic modeling and is funded by the German Ministry of Education and Science and the Free State of Bavaria. We observe an aggregate system performance of 140 GFLOPs out of a theoretical 624 GFLOPs peak.

This section was published as:

J. OESER, H.-P. BUNGE AND M. MOHR (2006), **Cluster Design in the Earth Sciences: TETHYS**, in *High Performance Computing and Communications – Second International Conference, HPCC 2006, Munich, Germany*.

2.2.1 Introduction

The Earth's interior is complex, consisting of three distinct regions nested one inside the other. Starting from the outside, there is first the brittle crust, then the solid mantle, and finally the (mostly) liquid core. As a result of convective and other forcings, all three regions are in motion, albeit on different time scales. On the longest time scale solid state convection (creep) overturns the solid mantle once in about every 100-200 million years (Bunge et al., 1998). This overturn is the primary means by which our planet rids itself of primordial and radioactive heat. It gives rise to large-scale geologic activity such as plate tectonics and continental drift. Reflecting this importance geophysicists have performed computer-based studies of mantle convection for decades, initially with simple 2D approaches (Jarvis & McKenzie, 1980), and recently in fully 3D spherical models (Glatzmaier, 1988; Tackley et al., 1993; Bunge et al., 1996; Zhong et al., 2000).

On a shorter time scale of perhaps 1 to 1000 years convection of the liquid iron core generates Earth's magnetic field. The field is sustained by a complicated dynamo process that probably operated throughout much of Earth's history. Only recently have geophysicists been able to study dynamo action in sophisticated magneto-hydrodynamic models of the core (Glatzmaier & Roberts, 1995; Kuang & Bloxham, 1997). On an even shorter time scale of hours to seconds both the core and the mantle are traversed by seismic sound waves, and seismologists are now turning to computer models to study seismic wave propagation through our planet (Igel & Weber, 1995; Komatitsch & Tromp, 1999b).

Geophysical modeling has benefited greatly from the advent of large-scale parallel computers. Focusing on the Earth's mantle, computing resources are now sufficient to simulate convection at full vigor in 3D spherical geometry. One current frontier lies in applying a range of data-assimilation techniques to study the geologic history of deep Earth flow (Bunge et al., 2002; McNamara & Zhong, 2005). Data-assimilation is, of course, a familiar tool in numerical weather prediction. A powerful approach to data-assimilation exists through variational methods, see e.g. Courtier & Talagrand (1987); Wunsch (1996). In mantle convection this technique requires one to solve a numerical adjoint code and the forward model iteratively to find solutions to a non-linear inverse problem (Bunge et al., 2003). Unfortunately adjoint modeling comes at a heavy computational price. Weeks to months of dedicated integration time are needed to solve these problems even on some of the most powerful parallel machines currently in use at national compute centers. Such resources are often out of reach for individual research groups and academic departments.

While parallel computing has moved into the mainstream, it is not economically feasible on a department budget to invest in dedicated commercial high-performance parallel machines sufficient to handle, for example, forward and adjoint mantle circulation models with earth-like Rayleigh numbers on the order of 10^9. The spatial resolution in such a simulation needs to be at least 20 km throughout the mantle, involving about 100,000,000 numerical grid points. However, it appears practical to address such problems at the academic department level by building cost-efficient *departmental* supercomputers.

In this paper we report on our approach on building such a system for capacity computing in the Earth sciences. In order to give an impression of the requirements and challenges of cluster computing in this field we consider as an example application the above mentioned problem of mantle convection. We start in Chap. 2.2.2 with a brief overview of the governing equations before describing our solution algorithm and its parallelisation in Chap. 2.2.3. We then report how these requirements influenced the design of our Tethys cluster in Chap. 2.2.4 and provide some first benchmarks.

2.2.2 A model for mantle convection

Generally speaking mantle convection is a flow process driven by temperature variations. As such it can be described by the time-dependent compressible Navier-Stokes equations that constitute a system of partial differential equations (PDEs) describing the conservation of mass and momentum in combination with an equation for energy conservation, see e.g. Landau & Lifschitz (1987). However, several simplifications to this model are possible. The first one is to assume a quasi-static flow field, i.e. the time-derivative is dropped from the momentum equations. One then exploits the small magnitude of the flow velocities to also drop here the non-linear convection terms. Finally one employs the so called Boussinesq approximation, see again e.g. Landau & Lifschitz (1987), whose main assumption is that density variations may be neglected except for the buoyancy terms. Thus, one arrives at a generalised Stokes problem coupled to a time-dependent energy equation. Using standard notation in which the divergence of a matrix field A is understood as a vector whose k-th entry is the (scalar) divergence of the k-th column of A and $(\operatorname{grad}\vec{u})_{ij} = \partial u_j/\partial x_i$ this system can be written as:

$$\operatorname{div}\vec{u} = 0 \tag{2.4}$$

$$\operatorname{div}\left[\nu\left(\operatorname{grad}\vec{u} + (\operatorname{grad}\vec{u})^T\right)\right] - \operatorname{grad} p + \varrho\alpha\left(T - T_0\right)\vec{g} = 0 \tag{2.5}$$

$$\varrho\, c_p\left(\frac{\partial T}{\partial t} + \vec{u}\cdot\operatorname{grad} T\right) - \operatorname{div}\left(\kappa \operatorname{grad} T\right) - \varrho H = 0 \tag{2.6}$$

Here the dependent variables are velocity $\vec{u}$, pressure p and temperature T, α is the coefficient of thermal expansion, c_p the specific heat at constant pressure, H the rate of internal heat production per unit volume, $\vec{g}$ the gravitational acceleration, κ the thermal diffusivity, ϱ the density and ν the kinematic viscosity of the fluid. Note that we have also dropped the inertial and coriolis term from the momentum equation, because of their small relative amplitude. Their absence distinguishes the slow creeping flow of the mantle (where viscous forces dominate) from other perhaps more familiar geophysical flows such as ocean circulation or the magneto-hydrodynamic flow problem of the core.

2.2.3 TERRA: Algorithm and Parallel Issues

In order to solve the coupled non-linear PDE system of mantle convection given in Sect. 2.2.2 we use a Finite Element technique. Our numerical modeling code TERRA is a parallel version of the vectorised model introduced in Baumgardner (1985). The code is widely used for mantle convection studies. Supported by NASA's High Performance Computing and Communication *(HPCC)* initiative TERRA demonstrated sustained parallel performance better than 100 GFLOPs (10^{11} floating point operations per second) in 1998 on a 1024 Processor Cray T3E-1200.

In TERRA the mantle is discretised using a computational grid adopted from the atmospheric community. The mesh is based on the regular icosahedron (Williamson, 1968) and displays the remarkable property of providing an almost uniform triangulation of the spherical surface. The icosahedral grid allows one to avoid the *pole-problem* of conventional latitude-longitude grids, where narrow wedge-shaped computational cells near the pole impose a severe Courant limitation on the time step. Starting from the icosahedron a Finite Element surface mesh is built recursively by splitting nodal distances in half and inserting new nodes at the midpoints, see e.g. Baumgardner & Frederickson (1985) for a detailed description. A graphical representation of the first three steps is given in Fig. 2.9. Note that each time we repeat this process lateral resolution in the domain is doubled. We can thus achieve arbitrary degrees of mesh refinement. The surface grid is then expanded into 3D by radial replication down to the inner surface of the mantle. In this fashion one obtains prismatic elements with spherical triangles on top and bottom. The icosahedral discretisation also yields a convenient data structure for our code. Combining pairs of the original twenty icosahedral triangles to form ten diamonds on the sphere, we obtain ten logically rectangular blocks. Topologically the spherical mesh thus appears as one single logically rectangular problem domain.

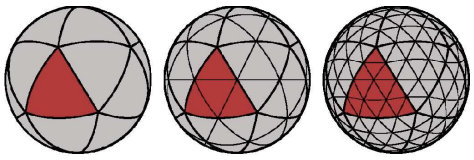


Figure 2.9: Three successive mesh-refinements of the icosahedral grid.

After converting the problem into its weak formulation and choosing piece-wise linear Finite Element ansatz functions the resulting system of equations is solved using the following approach. The temperature T is integrated in time with an explicit second order approach, i.e. the modified Euler method. Thus, for each time step the non-differential part of (2.6) must be evaluated twice. This requires knowledge of $\vec{u}$ at the respective simulation time. The latter is determined by solving the Stokes problem (2.4), (2.5) for the velocity field $\vec{u}$ and the pressure p via a Schur complement approach, see e.g. Benzi et al. (2005). This is accomplished by an inner-outer iteration pair, where the outer conjugate gradient (CG) iteration is used to compute the pressure and in each CG step a multigrid method is employed to determine the new search direction. The velocity is computed together with p in the CG method. For details of the algorithm see e.g. Verfürth (1984); Yang & Baumgardner (2000).

We, thus, see that at the core of the mantle convection code TERRA lies the problem of efficiently solving a discrete linear system stemming from an elliptic boundary value problem. This task arises not only in mantle convection simulations, but also e.g. in geo-potential problems and (quasi-)static viscoelastic analysis. In fact it can be found at the heart of numerous applications ranging from computational fluid dynamics over chip layout to bioelectric field simulations.

It is well-known that multigrid methods, see e.g. Trottenberg et al. (2001), are among the most efficient methods for solving large elliptic systems. Their key advantage is the possibility to reach an optimal linear scaling of computational expense relative to the number of unknowns of the problem. This makes these methods at least competitive in cost with the fast transform schemes (FFT) available for spectral codes. We note that high computational efficiency is essential for global mantle flow problems that involve millions of grid points. The nested structure of the icosahedral grid lends itself naturally to multigrid, as each grid is a subset of the next finer grid level.

Contrary to standard FFT approaches for PDEs both the CG and MG method can be parallelised in a straightforward manner using a domain decomposition (or more precisely a grid partitioning approach) in our case, see e.g. Trottenberg et al. (2001); Hülsemann et al. (2005). This is because the FEM approach leads only to a spatially restricted coupling of unknowns at nodes, which is similar to the local stencils of a Finite Difference setting. Splitting the grid into

sub-domains and associating each sub-domain with one parallel process the main effort of parallelisation lies in the co-ordination of information exchange along the sub-domain boundaries. In a distributed memory setting this can be achieved by explicit message passing based on the MPI standard[32].

In TERRA grid partitioning focuses on the ten diamonds that compose the icosahedral grid. Each associated 3D block is partitioned by powers of four. If four processes are used they work on the blocks associated with the upper, left, right and lower quarter of each diamond. Hence each process works on one quarter of the global problem domain. This approach can be repeated further, separately for the upper and the lower hemisphere, to achieve domain decompositions for any power of two.

Due to the above mentioned locality most of the work within each sub-domain can be performed independently of the others, with communication limited mostly to the exchange of boundary data among nearest neighbours. Depending on the volume-to-surface ratio of the sub-domains this local communication property assures high parallel efficiency. However, without going into too much detail, we must mention two aspects here. The first one is that due to the grid hierarchy employed in the MG method the volume-to-surface ratio naturally degrades on coarser grid levels. TERRA employs the standard approach of coarse grid agglomeration to counter this effect, we again point to Hülsemann et al. (2005) for this and alternative solutions. The second remark concerns the fact that TERRA uses a 2D partitioning of the grid. From a pure volume-to-surface point of view this is of course less favourable than a 3D splitting, where the grid is also partitioned in the radial direction. However, 2D splitting also has two distinct advantages. First, TERRA uses line-smoothing in the radial direction in order to ensure good MG convergence rates in the presence of strong viscosity variations. The choice of 2D decomposition avoids cutting these lines by sub-domain interfaces, which is a considerable performance and implementation advantage for the smoother. Second, 2D splitting results in a significant reduction of the number of messages that must be passed between processes compared to the 3D case. Using e.g. a partitioning with 64 sub-domains we need approximately 2,560 (2D) versus 3.520 (3D) messages for one boundary update after a single smoothing step. This reduction of about 27% may be considerable in a system with (relatively) high latency interconnects, cf. the design of Tethys in Sect. 2.2.4 in this respect.

[32] A PVM implementation of TERRA is also still available.

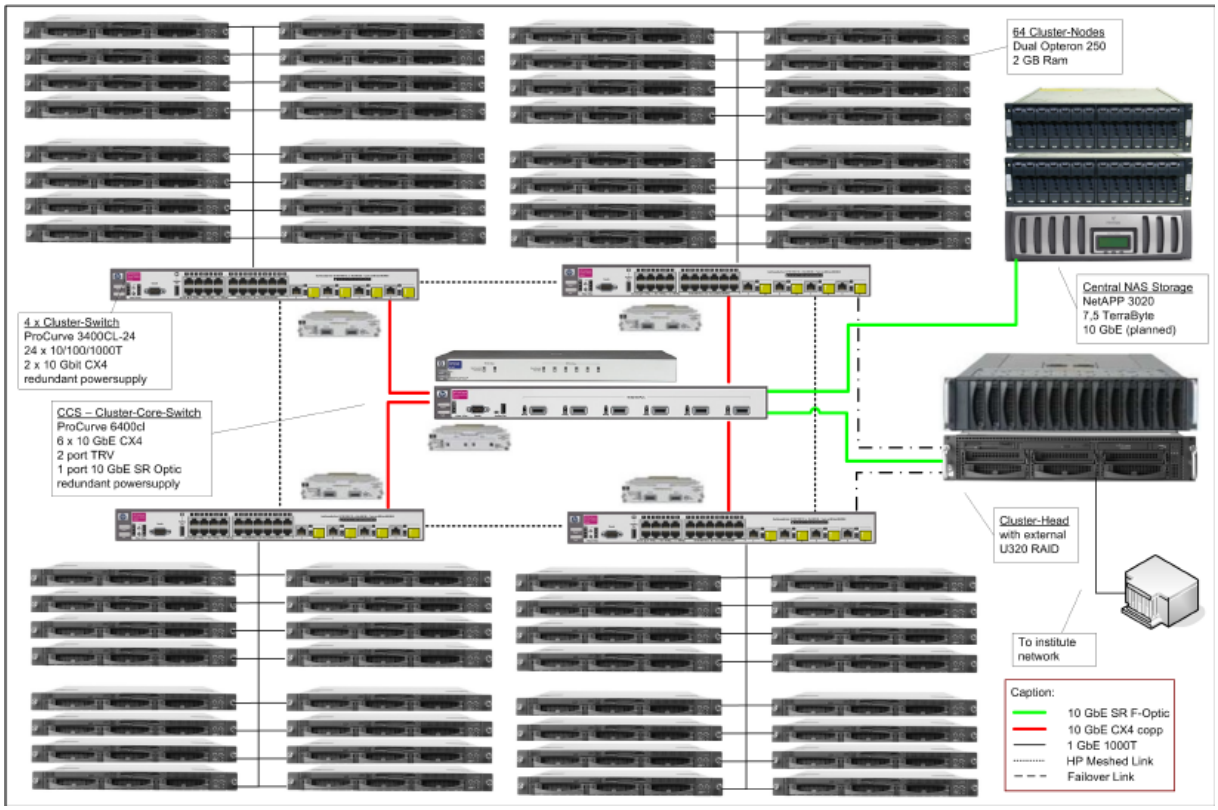


Figure 2.10: Schematic representation of the setup of Tethys and its interconnect topology.

2.2.4 Munich Earth Modeling Cluster Tethys

Design of Tethys

Global mantle flow studies rely on modern parallel computers. Taking mantle convection simulations with 100 million finite element grid points (to approximate the dynamic regime of vigorous mantle convection) as representative, about 70 Gigabytes (GB) of main memory and a sustained system performance of about 100 GFLOPs are needed to complete at least one convective overturn per hour. These requirements are now well in reach of dedicated PC-clusters. Such systems are called *Beowulf*, after the initial *Beowulf* PC-cluster project at the Goddard Space Flight Center [`http://www.beowulf.org`].

Our current cluster, named *Tethys* for Tectonic High Performance Simulator, includes 64 AMD Opteron 250 (64 bit, single core) 2.4 GHz dual-processors, each equipped with 2 GB RAM and 160 GB of disc-space. A dual-processor configuration improves the price performance ratio of our system and allows for some flexibility for those codes that require very large per-processor memory. Table 2.21 summarises the Tethys configuration, Figure 2.10 gives an overview of the setup of the cluster and its network topology. The 64 compute nodes of Tethys are arranged in four bricks consisting of 16 nodes each. The nodes within one such brick com-

Table 2.21: Hardware specification of the 64 compute nodes.

no. of processors	two per node
type of processors	AMD Opteron 250 (64 bit, single core)
clock speed	2.4 GHz
L1 cache	64/64KB (data/instruction)
L2 cache	1MB (data + instruction)
local memory	2 GB RAM (DDR1)
local storage	160 GB
network interface	1000T Ethernet (2 ports)

municate via a single 1000T cluster node switch. For inter-brick communication the four cluster node switches are attached to a central cluster core switch.

Some Benchmarks

We start our exposition by considering the communication performance of the Tethys cluster. In order to evaluate the latter we employ the Intel MPI benchmarks (IMB) suite[33]. For the sake of brevity we restrict ourselves to the *Exchange* and *Allreduce* benchmarks, since they represent exemplary communication patterns in domain decomposition approaches for parallel multigrid. In the *Exchange* test processes are assumed to form a periodic process chain. Each process exchanges data with both of its neighbours in the chain. This resembles the update of internal boundary values e.g. after one MG smoothing step. The *Allreduce* test, on the other hand, represents a collective communication involving all processes. The corresponding MPI_ALLREDUCE method gathers data from all processes performing a global reduction operation, e.g. a summation, on the fly and returns the result to all processes. This situation typically occurs in the computation of the residual norm in parallel iterative methods. Further details of the benchmarks can be found in the user's guide.

Figures 2.11 and 2.12 give timing results for both tests for different message sizes and increasing number of processes (np). Performance is in line with expectations for TCP/IP based communication over GBit-Ethernet and there is a linear scaling with increasing message size. Note that the first four cases (np = 4, 8, 16, 32) involve only nodes belonging to a single brick. Thus, communication involves a single cluster node switch only. The case having 128 CPUs (np = 128), instead, involves all four bricks and hence the entire cluster network including

[33] We employ IMB version 2.3 in combination with MPICH 1.2.5.3 an the Intel Compiler suite version 9.0 for these tests.

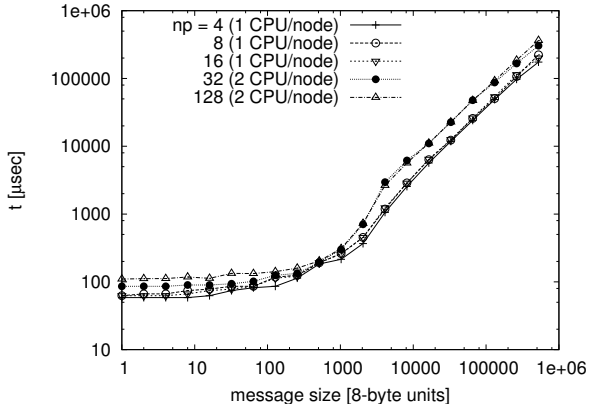


Figure 2.11: Performance for *Exchange* benchmark.

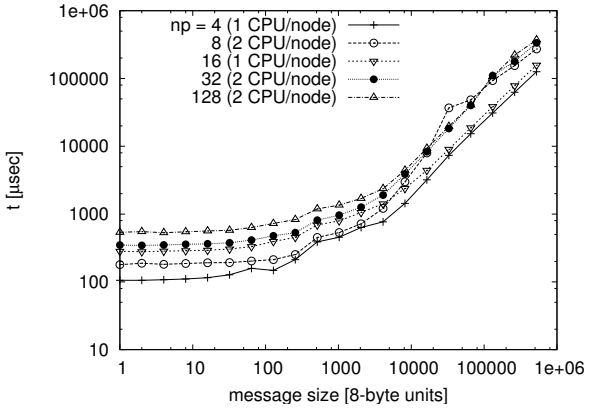


Figure 2.12: Performance for *Allreduce* benchmark.

the cluster core switch, see Fig. 2.10. We also show the direct comparison of a setting with 4 processes running either on 4 nodes within one brick or running on four nodes each in a separate brick in Fig. 2.13. We verify from the similar scaling of the two configurations that the hierarchical system of our network topology works quite well, and does not present any serious bottle-neck to global communication.

Finally we measure in Fig. 2.14 the effect of using our cluster nodes in dual-processor mode. In the first setting we use 8 CPUs on 8 nodes and in the second one 8 CPUs on 4 nodes. We note small differences especially in the *Exchange* performance for certain message lengths. However, this might be eliminated by employing an SMP communication layer for intra-node communication within MPICH.

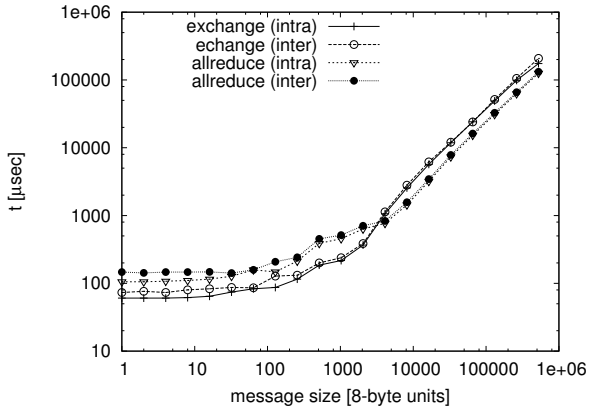


Figure 2.13: Comparison between intra- and inter-brick communication.

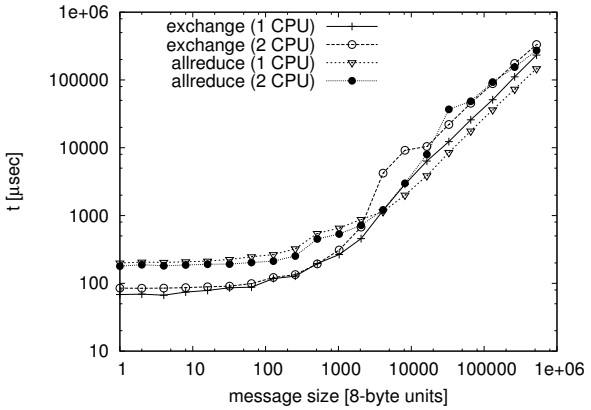


Figure 2.14: Influence of double-CPU nodes on communication performance.

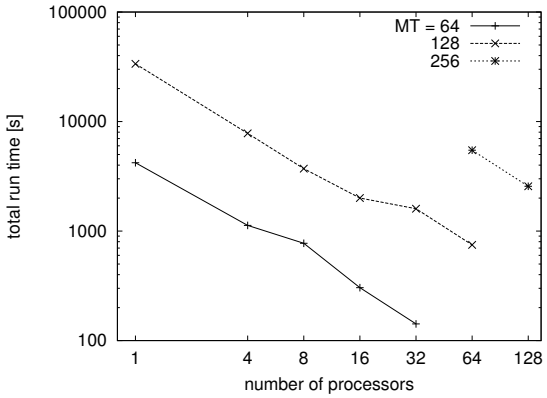


Figure 2.15: Run times for different problem sizes and increasing number of processes.

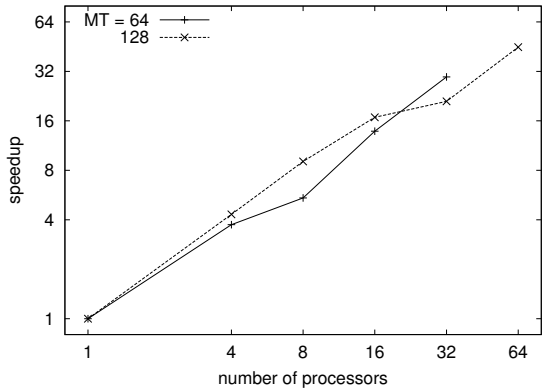


Figure 2.16: Speedup of the TERRA code on the Tethys cluster.

We now turn to the performance of our mantle convection code TERRA. For our initial test we considered three different discretisations marked by mt=64, 128, and 256. These result in a problem size of about 1, 10 and 85 million grid points and reflect a surface resolution of 100, 50 and 25 km. Figure 2.15 reports the wall clock times[34] of a single simulation run consisting of 500 time-steps for different numbers of processes np. We observe a nice, approximately linear decay with increasing np. This is directly reflected in Fig. 2.16 where the speedup for mt=64 and 128 is plotted. From the Earth scientist's point of view the scaleup of the application is of course more interesting than the mere speedup. In this respect we point out that we obtain a run-time of 2002 s (for the mt=128 case on 16 processes) and 2564 s (for the mt=256 case on 128 processes), which both lead to the same workload per process, see again Fig. 2.15. Based on an approximate requirement of 8000 operations per node and time-step the latter example results in an aggregate system performance of 140 GFLOPs out of a theoretical 624 GFLOPs peak. These are initial results and an improved scale-up might be obtained by tuning the code to the Opteron architecture.

[34] The value for mt=128 on a single process is extrapolated (due to memory limitations).

2.2.5 Conclusions & Outlook

We have built a large-scale geophysical modeling cluster at Munich University's (LMU) Geosciences department. The machine serves as a departmental supercomputer to perform a range of geosciences simulations, including compute intensive variational data-assimilation calculations for global mantle convection studies. We observe parallel efficiencies of better than 80% and an aggregate system performance of 200 GFLOPs. We conclude that cost-efficient Beowulf clusters should take an increasing role in performing large-scale capacity-oriented geosciences simulations.

2.2.6 Acknowledgement

Tethys was funded by the German Ministry of Education and Research (BMBF) and the Free State of Bavaria by means of the HBFG program. We would also like to thank Microstaxx GmbH and the High-Performance Group of Fujitsu-Siemens Computers for their support.

2.3 Massenspeichersystem

Die vorgestellten Rechnersysteme in den Abschnitten 2.1, 2.4 und 2.5 werden innerhalb ihrer Nutzungsdauer am Lehrstuhl für Geophysik dazu verwendet, Simulationsrechnungen durchzuführen, die erhaltenen Simulationsergebnisse auszuwerten und schließlich die daraus gewonnenen Erkenntnisse darzustellen. Jede Einzelrechnung generiert Datenmengen im Bereich von mehreren GigaByte. Damit entstehen in den nächsten Jahren der Nutzung Datensätze die mehrere TeraByte Speicherplatz belegen werden. Die enormen Datenmengen werden noch vergrößert durch den Einsatz der am LRZ vorhandenen Rechnersysteme für „one of a kind“ Simulationen. Diese Simulationsergebnisse müssen mittelfristig auch für die Weiterverarbeitung gespeichert werden. Zusätzlich entstehen in dem, dem Lehrstuhl für Geophysik angegliederten, „Geophysikalischen Observatorium Fürstenfeldbruck“ kontinuierlich Datensätze der Größe von ungefähr 100 GB pro Monat. Auch wenn man davon ausgehen kann, dass nicht alle Simulationsergebnisse dauerhaft vorrätig zu halten sind, so ergeben sich allein aus den Observatoriumsdaten 1.2 TB benötigter Speicherplatz pro Jahr. Eine Abschätzung der erforderlichen Kapazität zur Aufnahme der hier bereits erwähnten Datenmengen ergibt einen jährlichen Speicherplatzbedarf am Lehrstuhl für Geophysik der Ludwig-Maximilians-Universität von circa 5 TB.

2.3.1 Konzeption

Im Folgenden sollen nun die Möglichkeiten diskutiert werden, ein derartiges Projekt zu verwirklichen. Die wesentlichen Merkmale der zu beschaffenden Datenspeicherinfrastruktur sind in der sich nun anschließenden Aufzählung zusammengetragen, wobei in allen Bereichen die Effektivität an erster Stelle steht.

- Das Speichersystem muss flexibel mit den steigenden Anforderungen erweiterbar sein.
- Mit den wachsenden Datenmengen muss das System gut skalieren.
- In den nächsten Jahren muss ein Ausbau auf rund 50 TB Nutzkapazität möglich sein.
- Eine zentralisierte Struktur für das Gesamtsystem ist unabdingbar.
- Der Datenzugang muss unabhängig vom vorhandenen Betriebssystem möglich sein.

Eine Analyse der im Jahre 2004 vorhandenen IT-Infrastruktur in der Geophysik zeigte die Notwendigkeit eines vollständigen Neubeginns auf. Dieser radikale Schnitt mit der bisherigen Speicherinfrastruktur schien in Anbetracht der Neuausrichtung des Lehrstuhls zur Simulation komplexer Prozesse im Erdkörper notwendig. Bis dahin wurde eine dezentrale Speicherung der

Daten auf den Arbeitsplatzrechnern der Anwender bevorzugt. Die zu diesem Zeitpunkt erhältlichen PCs verfügten nur über eine begrenzte Festplattenkapazität in der Größe von maximal 300 bis 400 GB. Damit war zwangsläufig jeder Nutzer mit großen Datensätzen gezwungen, die eigenen Dateien über mehrere Rechner zu verteilen. Dies führt früher oder später zu Leistungs- und Kapazitätsgrenzen, die dazu zwingen weitere Dateisysteme/Festplatten hinzuzufügen. Da die Rechner unabhängig voneinander arbeiten, müssen sie auch getrennt verwaltet werden. Außerdem verbringen die Anwender viel Zeit damit, einzelne Dateien zu lokalisieren. Zusätzlich erzeugen die auf vielen PCs verstreuten Datensätze bei der Weiterverarbeitung einen nicht unerheblichen Mehraufwand, da sie vor der Bearbeitung zusammengetragen werden müssen. Eine Minderung der Probleme können in diesem Fall RAID-Subsysteme[35] bieten, da sie aus mehreren Festplatten durch eine entsprechende Verschaltung größere Speicherbereiche generieren. Dabei erhöht sich nicht nur der zusammenhängende Speicherplatz, sondern durch intelligente Auswahl des RAID-Levels auch die Datensicherheit (RAID-1, RAID-5, RAID-6). In solchen Subsystemen (meist über SCSI-Schnittstellenkarten mit dem Rechner verbunden) kann nur eine begrenzte Anzahl an Festplatten eingebaut werden, deshalb ist eine spätere Kapazitätserweiterung kaum möglich. Abhängig vom Betriebssystem des Rechners, welcher das RAID-System angeschlossen hat, kann der Zugriff auf die Dateien für alle vorhandenen Betriebssysteme eingerichtet oder nicht eingerichtet werden. Diese bis dato in der Geophysik verwendete Infrastruktur scheint demnach nicht als Basis für die anzustrebende Neustrukturierung geeignet.

Eine funktionale Einheit aus Hard- und Software für ein solches Speichersystem ist aus diesem Grund angebracht. Anbieter für solch eine Speicherinfrastruktur konnten in einer Internetrecherche und in einer Besprechung mit der zentralen EDV-Beschaffung der Ludwig-Maximilians-Universität gefunden werden. Nach Möglichkeit sollen Produkte bevorzugt werden, die eine freie Software Basis besitzen, damit Kosten eingespart werden können. Allerdings verwenden die entsprechenden Anbieter Rechner mit angeschlossenen RAID-Subsystemen, wodurch die geforderte Erweiterbarkeit nicht ohne größeren Aufwand realisierbar ist. Zusätzlich konnten für viele dieser Systeme keine Verfügbarkeitswerte gefunden werden. Damit scheint der geplante Einsatz als zentrale Speicherinfrastruktur für derartige Systeme fraglich aber nicht ausgeschlossen. Neben diesen häufig kleineren Unternehmen existieren noch Firmen, die auf Herstellung und Vertrieb großer Speicherinfrastrukturen spezialisiert sind. Die von diesen Unternehmen vertriebenen kommerziellen Produkte werden von ihnen häufig als Massenspeichersysteme umworben, wobei eine Einteilung in NAS und SAN unternommen wird. Diese Unterteilung liegt in der Art Massenspeicheranbindung an das lokale Netzwerk begründet. Unter „Network Attached Storage“ (NAS) versteht man ein System, welches im Unterschied zu „Direct Attached Storage“ (DAS) ein eigenständiges System mit integriertem Betriebssystem darstellt. Solch ein

[35] RAID steht für „redundant array of independent disks“ (`http://de.wikipedia.org/wiki/RAID`)

einfach zu verwaltender Dateiserver kann bereits durch eine handelsübliche Festplatte, versehen mit einer Netzwerkschnittstelle und einem minimalistischen Betriebssystem, realisiert werden. Je mehr Festplatten in einem derartigen System eingebaut werden, desto höher ist auch der Aufwand bei der Hardwareintegration. Davon unberührt bleibt aber der einfache Zugriff auf die Daten über das Netzwerk mit standardisierten Protokollen wie NFS und CIFS (siehe dazu Seite 133). Bei diesen Protokollen werden die entsprechend benötigten Dateien über das Netzwerk durch den Client vom Server abgerufen. Vom „Storage Area Network“ (SAN) spricht man, wenn über ein extra dafür eingerichtetes Netzwerk die Festplattensysteme direkt mit den Rechnern verbunden sind. Es ist dabei unerheblich wie viele Rechner oder Festplattensysteme über das SAN miteinander kommunizieren. Es müssen lediglich genügend Netzwerkanschlüsse zur Verfügung stehen. Damit stellt das SAN eine Erweiterung des DAS dar, bei der die punktuelle Verbindung zwischen Speicher und Server durch ein Netzwerk hergestellt wird. Derzeit wird hauptsächlich das „Fibre Channel“ Protokoll für die Netzwerkkommunikation eingesetzt. Diese Technik erfordert jedoch die Nutzung spezieller (kostenintensiver) Hardware. Dies ist aber mit Blick auf die erreichbaren Transferleistungen sinnvoll. Der Zugriff auf die Daten erfolgt beim SAN im Gegensatz zum NAS blockbasiert.

Es stellt sich die Frage, welche dieser beiden Varianten eines Massenspeichersystems die geeignetere für die Geophysik ist. Die Planung für den Massenspeicher sah vor, dass neben den Simulations- und Observationsdaten auch die Benutzerverzeichnisse auf das System verschoben werden sollten. Die Vielzahl an Arbeitsplatzrechnern am Lehrstuhl macht es fast unmöglich, ein SAN Netzwerk aufzubauen, da die damit verbundenen Kosten enorm wären. Für wenige Server wäre eine solche Lösung aber im Hinblick auf die Virtualisierung attraktiv. Eine Speicherinfrastruktur, die beide Arten ermöglichen kann, wäre zu bevorzugen.

Speicherkonzept

Im nächsten Abschnitt werden nun konkrete Anforderungen an die Massenspeicherinfrastruktur zusammengetragen. Dazu ist eine genauere Analyse der verschiedenen in der Geophysik vertretenen Arbeitsschwerpunkte notwendig. In der Magnetik- und Geodynamikarbeitsgruppe liegt das Hauptaugenmerk der Forschung auf der Modellierung der Dynamik des Erdinneren und der Lithosphäre, während sich die Seismologiearbeitsgruppe auf die Ausbreitung seismischer Wellen und der Modellierung komplizierter Erdbebenszenarien konzentriert. Damit ergibt sich für jede Arbeitsgruppe durch deren unterschiedliche Forschungsthemen ein anders gewichteter Bedarf an Speicher. Es soll nun versucht werden, diesen für jede Arbeitsgruppe getrennt aufzuzeigen.

Geodynamik Die Durchführung von iterativen Inversionsmethoden auf Basis des adjungierten Verfahrens erzeugt in der Geodynamik den Hauptbedarf an Speicherkapazität. Bei diesen Inversionsverfahren werden in TERRA die Druck-, Geschwindigkeits- und Temperaturverteilungen einer Vorwärtssimulation abgespeichert. Die gespeicherten Datensätze werden danach bei der Inversion wieder verwendet. Einer der vielen zeitlichen Momentaufnahmen bei der globalen Mantelkonvektionssimulation belegt rund 5 GB Speicher. Bei 10000 Zeitschritten einer Simulation entstehen Simulationsdaten der Größenordnung 50 TB. Durch spezielle Verfahren kann der Bedarf deutlich verringert werden, allerdings zum Preis erhöhter Rechenzeiten. Von 5 bis 10 TB benötigtem Speicherplatz kann trotzdem ausgegangen werden. Die Bereitstellung tomographischer und tektonischer Modelle oder von INSAR und GPS Daten für die verschiedenen Simulationsrechnungen erfordert weitere 5 TB. Bei der Auswertung der Modellierungen und Inversionen entstehen zusätzlich 5 bis 10 TB auf Grund der aufwendigen Nachbearbeitung und Visualisierung. Letztendlich führt dies zu einer notwendigen Speicherkapazität von 15 bis 25 TB für die Geodynamik.

Seismologie Wie in der Geodynamik liegt auch in der Seismologie der Hauptbedarf an Speicherkapazität in den Simulations- und Inversionsrechnungen. Die notwendigen hohen numerischen Auflösungen und die Speicherung langer Zeitreihen von Wellenfeldern (ungefähr 300 GB pro Zeitreihe) erfordern Massenspeicherkapazitäten der Größenordnung 20 bis 25 TB, wobei die Speicherung von zeitlichen Momentaufnahmen und Auswertung der Daten in dieser Angabe bereits inbegriffen sind. Somit ist der Bedarf aus der Seismologiearbeitsgruppe vergleichbar mit dem aus der Geodynamik.

Magnetik Die Generierung komplexer dreidimensionaler tektonischer Modelle mit Hilfe paläomagnetischer Daten stellt den Hauptbedarf an Speicherkapazität dar. Die erwähnten Messdaten und die daraus erstellten Modelle belegen bis zu 5 TB Massenspeicher.

Geophysikalisches Observatorium Fürstenfeldbruck Die Seismologiearbeitsgruppe befasst sich neben der Simulation seismischer Prozesse aktiv mit der Datenerhebung und automatischen Auswertung von seismologischen Netzwerken am geophysikalischen Observatorium. Dazu werden die deutschen Regionalnetzstationen FUR (Fürstenfeldbruck) und WET (Wettzell) zur Beobachtung der weltweiten Erdbebenaktivität betrieben. Die 21 Stationen des regionalen seismologischen Netzwerks (Bayernnetz) werden zur Überwachung und wissenschaftlichen Auswertung der Erdbebenherde Bayerns und angrenzender Gebiete genutzt. Darüber hinaus besteht eine enge Verknüpfung mit der Vulkanseismologie sowie der Beobachtung von Rotationsbewegungen bei Erdbeben. Alle diese Bereiche sind äußerst datenintensiv. Es fällt

ein kontinuierliches Datenvolumen von knapp unter 5 GB pro Station und Monat an. Ein neuer und besonders wichtiger Forschungszweig in der Seismologie zielt von der ereignisgestützten Datenerhebung und Auswertung auf die kontinuierliche Aufzeichnung, Auswertung und Archivierung. Dies wird motiviert durch neuere Arbeiten, in denen gezeigt wird, wie wichtig die bisher als Rauschen angesehenen Aufzeichnungen zwischen Ereignissen sind, um zusätzliche Untergrundparameter zu bestimmen. Dies bedingt eine weitere Steigerung des dauerhaft zu speichernden Datenvolumens in der Art, dass Daten für 3 bis 5 Jahre online vorgehalten werden sollten und erst danach archiviert werden. Pro Jahr ergibt sich für die mehr als 20 Stationen des Erdbebennetzes über 2 TB Speicherbedarf. Will man die Daten der letzten 5 Jahre mit hoher Verfügbarkeit bereitstellen, so müssen insgesamt 10 TB Speicherplatz auf dem Massenspeichersystem vorhanden sein.

Die erdmagnetische Warte des geophysikalischen Observatoriums (FUR) ist hinsichtlich ihrer hohen Datenqualität bekannt. Die Langzeitbeobachtungen des Erdmagnetfeldes werden seit über 100 Jahre durchgeführt. Die auf Papier durchgeführten Magnetfeldaufzeichnungen zu Beginn des Stationsbetriebs werden zur Zeit digitalisiert und im Internet zur Verfügung gestellt. Für diese Aufgabe werden ungefähr 500 GB Speicherplatz benötigt. Die Größe der aktuellen Datensätze wird in Zukunft noch stärker als bisher ansteigen, da das internationale Netzwerk der geomagnetischen Observatorien (INTERMAGNET[36]) eine Erhöhung der Messrate von einer Messung pro Minute auf eine Messung pro Sekunde vorschreibt. Der notwendige Speicherplatzbedarf wird in diesem Zusammenhang um den Faktor 200 ansteigen (der Faktor 200 resultiert aus der Anwendung numerischer Filter, sodass magnetische Messungen im Bereich von 10 bis 50 Hz durchgeführt werden müssen). Zusammengefasst ergibt sich für die Magnetikgruppe am Observatorium ein Speicherplatzbedarf in der Größenordnung von 1 TB.

Lehrstuhlnetzwerk Das Massenspeichersystem soll primär dazu verwendet werden, die bereits aufgeführten Datenmengen aus den verschiedenen Forschungsschwerpunkten aufzunehmen. Die Planungen für diese Infrastruktur sahen eine Kapazität von etwa 50 TB vor. Die bisher dezentral organisierten Dienste im Lehrstuhlnetzwerk sollten mit der Neustrukturierung zentralisiert werden. Dazu ist die Speicherung der privaten Benutzerverzeichnisse auf einem zentralen Massenspeichersystem notwendig. Für dieses Datenvolumen unter 2 TB ist der schnelle und betriebssystemunabhängige Zugriff auf die eigenen Daten eines jeden Anwenders genauso wichtig wie die Sicherheit der privaten Daten. Um eine hohe Effizienz erzielen zu können, sollte das neue Massenspeichersystem diese Aufgabe mit übernehmen.

[36] `http://www.intermagnet.org/`

Aufteilung der Speicherinfrastruktur Die Aufteilung des Lehrstuhls für Geophysik in zwei Standorte und die zu erwartenden hohen Kosten ließen eine nicht einfache Realisierung erwarten, zumal der Observatoriumsbetrieb nach Möglichkeit nahezu unterbrechungsfrei laufen sollte. Daher war eine detaillierte zeitliche Planung des Speicherplatzbedarfs notwendig. Es wurde eine Staffelung bei der Beschaffung angedacht. Am wichtigsten war die Migration der verstreut abgelegten Benutzerverzeichnisse auf ein zentrales System sowie die Verfügbarkeit von geringen Speicherkapazitäten für die ersten Simulationsergebnisse. Zur Realisierung dieses Zieles erschienen 4 bis 5 TB nutzbare Kapazität ausreichend.

In Folge dieses ersten Schrittes war für circa zwei Jahre später der weitere Ausbau geplant. Dabei sollte der Speicherplatz für die Simulationsrechnungen bis zu einer Größe von 12 bis 15 TB erweitert werden. Es musste auch eine Lösung für die kontinuierlichen Messdaten des Observatoriums angedacht werden. Während der Standort München über eine direkte Hochgeschwindigkeitsanbindung an das MWN und damit in das Deutsche Wissenschaftsnetz (WiN) verfügt, hat der Standort Fürstenfeldbruck nur eine 1 MBit-Verbindung[37]. Da einerseits die zu speichernden Simulationsdaten im Standort Innenstadt anfallen und die zu speichernden Observationsdaten am Standort Fürstenfeldbruck zusammenlaufen, andererseits der Standort Fürstenfeldbruck nur über eine verhältnismäßig langsame Anbindung verfügt, wurde im Konzept eine Aufspaltung des Massenspeichersystems in der folgenden Form vorgesehen.

Der weitaus größere Teil des Massenspeichersystems sollte in München installiert werden, um die dort durch Berechnungen auf dem vorhandenen HPC-Rechner anfallenden Simulationsdaten, sowie vorselektierte Ergebnisse von Simulationsläufen auf den Rechenclustern des LRZ zur Weiterverarbeitung zu speichern. Die Installation des zweiten Teils des Massenspeichersystems war für das geophysikalische Observatorium Fürstenfeldbruck vorgesehen, wo die Speicherung der von den Messstationen einlaufenden Daten erfolgt. Diese Daten stehen den vor Ort arbeitenden Wissenschaftlern mit hoher Geschwindigkeit für die Auswertung im Rahmen ihrer Forschungen zur Verfügung. Dies bedeutet für das Observatorium einen Speicherplatzbedarf von 10 bis 12 TB. Zusätzlich wurde angedacht, die Observationsdaten zentral in München als Datenspiegel vorzuhalten. Durch die aufzubauende Synchronisation der Systeme in München und Fürstenfeldbruck, wird die Datensicherheit signifikant erhöht. Dies stellt einen wichtigen Aspekt dar, da verlorene Messdaten nicht erneut erhoben werden können. So können die Daten vom System in München Forschern weltweit zur Verfügung gestellt werden. Dank der besseren Netzwerkanbindung kann dies mit wesentlich höherer Geschwindigkeit erfolgen. Die größte Schwierigkeit dieser Lösung besteht darin, eine effiziente Synchronisation der Daten zwischen den Standorten zu schaffen, da die Netzwerke mit unterschiedlicher Geschwindigkeit an dass Internet angeschlossen sind.

[37] Vom LRZ war 2007 ein Ausbau auf 20 MBit geplant, was 2008 auch erfolgte.

Nach einer Besprechung mit Mitarbeitern des Leibniz-Rechenzentrums und der zentralen EDV-Beschaffung wurde ein wichtiger Ansatz zur Datensynchronisation verfolgt. Vom LRZ wird für interne Dienste ein System auf Basis der NetApp Technologie[38] eingesetzt. Eine Komponente dieses kommerziellen Systems ist die effiziente Synchronisation von Datenbereichen, bei der nur geänderte Blöcke des Dateisystems transferiert werden. Die am LRZ vorhandenen NetApp Systeme liefen zu diesem Zeitpunkt ohne Probleme und mit einer sehr hohen Verfügbarkeit und Geschwindigkeit[39]. Bei der öffentlichen Ausschreibung sollte nach Möglichkeit eine vergleichbare Synchronisationstechnik für das Massenspeichersystem gefordert werden.

Für einen möglichen weiteren Ausbau der Massenspeicherinfrastruktur in der Geophysik, sollte die Option bestehen, auf bis zu 30 TB Kapazität zu erweitern. Diese Aufstockung könnte in einem sich anschließenden dritten Schritt erfolgen.

Anforderungsprofil

Der angedachte zweigeteilte Prozess zur Einrichtung einer Massenspeicherinfrastruktur am Lehrstuhl für Geophysik wurde für einen Zeitraum von etwa zwei Jahren geplant. Dementsprechend war es notwendig, zwei getrennte Anträge im HBFG-Programm der DFG (später das Großgeräteprogramm der DFG) zu stellen. Die getrennte Antragsstellung bedingt auch den Erwerb der Infrastruktur über zwei öffentliche Ausschreibungen.

Erste Ausbaustufe Für die Realisierung der ersten Ausbaustufe konnte im Rahmen des HBFG-Antrags für die Beschaffung eines Rechenclusters (siehe dazu Kapitel 2.1) entsprechende Mittel im Umfang von 110 000 € eingeworben werden (Kennziffer: 132/943-1). In Zusammenarbeit mit der zentralen EDV-Beschaffung der Ludwig-Maximilians-Universität wurde in der öffentlichen Ausschreibung für das HPC-System Ende 2005 auch die zentrale Massenspeicherinfrastruktur ausgeschrieben. Nach den bereits aufgeführten Abschätzungen zum Speicherplatzbedarf forderte die Ausschreibung eine Bruttokapazität von mindestens 7 TB. Bei Nutzung entsprechender RAID-Level und anderer Techniken zur Erhöhung der Datensicherheit entsteht ein Speicherplatzverlust, der bei den 7 TB zu der geplanten Nutz- oder Nettokapazität von 4 bis 5 TB führt. Erwartet wurde ebenfalls, dass der Massenspeicher auf bis zu 50 TB flexibel ausgebaut werden kann. Die Unterstützung der Netzwerkprotokolle NFS und CIFS war ebenso Voraussetzung (siehe dazu Abschnitt A.4). Durch Internetrecherchen, Besprechungen mit verschiedenen Anbietern sowie der zentralen EDV-Beschaffung der Ludwig-Maximilians-Univer-

[38] `http://www.netapp.com`

[39] Bereits seit 2003 werden entsprechende Systeme verwendet, unter anderem für die E-Maildienste und später auch für die HPC-Rechner (`http://www.lrz.de/wir/medien/netapp-2007-01.pdf`).

sität konnten die weiteren Erwartungen zu den technischen Eigenschaften des Projekts ausgearbeitet werden. Dabei war die Berücksichtigung, der für die Zukunft geplanten Synchronisation bereits in dieser Phase besonders wichtig, um nicht später das erworbene System austauschen zu müssen oder andere Zusatzkosten entstehen zu lassen. Dementsprechend enthielt die öffentliche Ausschreibung Ende 2005 die im Folgenden aufgeführten Forderungen nach spezifischen Produktmerkmalen.

Hochverfügbarkeit und Disaster Recovery:

- Verfügbarkeit des Systems von mindestens 99,95 %
- nachweisbare mehrjährige Marktreife durch Referenzen
- Unterstützung redundanter Netzwerkanbindungen
- im System müssen mehrere Versionsstände der Daten online verfügbar sein (zum Beispiel für Datensicherungen, Tests, Konsistenzprüfungen), auf die im Bedarfsfall ohne Leistungseinbußen zurückgegriffen werden kann (Snapshots) – diese müssen ohne Zusatzkosten für die maximale Ausbaustufe des Systems zur Verfügung stehen – eine nachträgliche kostenpflichtige Lizenzierung ist nicht möglich
- Konfigurationsänderungen müssen weitgehend online durchführbar sein
- telefonischer Herstellersupport für das System von täglich 24 Stunden
- RAID-Technologie
  - der Ausfall von zwei Festplatten innerhalb einer RAID-Gruppe muss durch die RAID-Technologie ohne Datenverlust abgesichert sein
  - Hot Spare Festplatten müssen in einem Pool für beliebige RAID-Gruppen zur Verfügung stehen
- Disaster Recovery
  - es muss die Möglichkeit bestehen, die Aktiv/Aktiv-Controllereinheiten auf zwei Standorte aufzuteilen, dabei kann die Entfernung zwischen den Standorten auch mehrere Kilometer betragen
  - optional: synchrone Replizierung (Spiegelfunktionalität), die zwei Kopien der Onlinedaten physisch getrennt aufbewahrt, um diese gegen sämtliche Arten von Hardwareausfällen zu schützen

Skalierbarkeit:

- Veränderung der Speicherbereiche (engl. „Volumes“)
  - die Speicherkapazität der Volumes sollte im laufenden Betrieb linear und innerhalb von Sekunden erweiterbar sein – die neue Kapazität muss sofort und ohne Leistungseinbußen zur Verfügung stehen – eine RAID-Neuberechnung ist unerwünscht
  - die Volumes müssen im laufenden Betrieb vergrößert und verkleinert werden können

  - Schreib-/Leseleistung für kleine und große Volumes muss gleich gut sein
- Produktlinie
  - das Speichersystem muss skalierbar sein (von 0.5 bis 50 TB)
  - idealerweise ist die gesamte Produktlinie mit dem gleichen Betriebssystem bzw. Dateisystem ausgestattet
  - der zeitliche Aufwand bei der Aktualisierung zu einem leistungsfähigeren, skalierbareren Produkt sollte möglichst gering sein (idealerweise unter 15 Minuten) – zeitintensive Datenmigrationen zwischen Alt- und Neusystem sind nicht erwünscht
  - das System muss den gemischten Betrieb von FC und SATA Festplatten parallel verarbeiten können[40]

Backup:

Für das Erstellen von Sicherungskopien (Backup) auf den zentralen Systemen des LRZ ist es notwendig auf feste Versionsstände (Snapshots) des Dateisystems zugreifen zu können. Folgende Funktionalitäten müssen für die Snapshots zur Verfügung stehen:

- das Erstellen und Wiederherstellen von Snapshots darf selbst bei Volumes mit Größen von mehreren TeraByte nur wenige Sekunden dauern und die Performance des Gesamtsystems nicht reduzieren
- Kopiervorgänge und doppelte Datenvorhaltung sind nicht erwünscht
- pro Volumen sollte die Vorhaltung von mindestens 200 Snapshots möglich sein, wobei die zusätzlich benötigte Speicherkapazität lediglich in der Veränderung der Blöcke zwischen den einzelnen Snapshots liegt
- optional: nicht nur das Wiederherstellen von einzelnen Dateien muss möglich sein, sondern auch das eines gesamten Dateisystems – Snaprestore

SAN-/NAS-Dienste:

Zur Konsolidierung sämtlicher Dateiserver unterschiedlicher Betriebssysteme ist es notwendig, dass die Speicherlösung die Netzwerkprotokolle NFS und CIFS unterstützt. Der primäre Einsatzbereich des Massenspeichersystems ist NFS, wozu zwingend die Protokollversionen 2 bis 4 Voraussetzung sind. Eine nachträgliche kostenpflichtige Lizenzierung des CIFS-Protokolls sollte gewährleistet sein. Zur Anbindung von blockorientierten Zugriffen muss eine gleichzeitige Anbindung von FC-SAN und/oder iSCSI[41] neben NFS und CIFS zur Verfügung stehen. Die Administration von SAN und NAS muss dabei gemeinsam erfolgen. Die Größe der virtuellen Festplatten (LUNs) im SAN muss flexibel von wenigen MB bis in den TB Bereich gestaltet werden können. Eine Erweiterung der LUNs muss im laufenden Betrieb möglich sein. NFS-Exporte und CIFS-Freigaben können für jede Partition unabhängig eingerichtet werden.

[40] „Fibre Channel“ (FC) und „Serial Advanced Technology Attachment“ (SATA)

[41] „Internet Small Computer System Interface“ (iSCSI) stellt Speicher über eine Netzwerkverbindung bereit.

Zugriff von Unix-Benutzern auf Windows-Daten und umgekehrt kann realisiert werden, wobei die eingestellten Zugriffsrechte eingehalten werden müssen.

Einfaches Handling:
Das Speichersystem sollte innerhalb kürzester Zeit (unterhalb 1 Stunde) installierbar sein. Der grundlegende Umgang mit dem System sollte in wenigen Stunden erlernbar sein. Das Erstellen von Snapshots, die Wiederherstellung von Dateien oder die Erweiterung von Volumes sollte ohne größeren Aufwand möglich sein.

In der öffentlichen Ausschreibung wurden der Aufbau und die Inbetriebnahme durch einen Systemingenieur des jeweiligen Herstellers sowie die dafür notwendigen Teile gefordert, welche allerdings gesondert im Angebot zu vermerken sind (19 Zoll Rackschrank, Verkabelung für die Stromversorgung und die Datenvernetzung, notwendige Kleinteile für Montage).

Als Service- und Garantieleistungen wurden im Ausschreibungstext zum einen eine Ersatzteilgarantie für alle Hardwarekomponenten von 5 Jahre gefordert und zum anderen einen 3 Jahre andauernden Vor-Ort-Service mit einer Reaktionszeit von 48 Stunden.

Zweite Ausbaustufe Der Antrag im Großgeräteprogramm der DFG für die Erweiterung der Massenspeicherinfrastruktur am Lehrstuhl für Geophysik wurde Ende 2007 bewilligt. Es standen damit 250 000 € für die zweite und auch größere Ausbauphase des Speichersystems zur Verfügung (Geschäftszeichen der DFG: INST 86/1040-1 FUGG).

Das vorhandene Netzwerk des Geophysikalischen Observatoriums Fürstenfeldbruck schien den Anforderungen, die mit der geplanten Erweiterung einhergehen würden, nicht gewachsen. Aus diesem Grunde war es in Kooperation mit dem LRZ, als Betreiber der Netzwerkinfrastruktur möglich, die aktiven Hardwarekomponenten durch neueres und zugleich leistungsfähigeres Material zu ersetzen. Die angestrebte Datensynchronisation zwischen beiden Standorten erfordert eine direkte Kommunikation zwischen den Bestandteilen des Speichersystems. Dies bedingte den Ausbau der vorhandenen Firewall-Infrastruktur auf Basis des „Astaro Security Gateway“[42], um nach Möglichkeit eine hochverfügbare VPN-Verbindung einrichten zu können[43]. Die Erweiterung der Firewallsysteme konnte durch die Integration einer Master-Slave-Lösung von Astaro erreicht werden. Da die Erfahrungen der letzten Jahre mit der Astaro-Lösung überaus positiv waren, wurde an dem Einsatz dieser Technologie festgehalten. Für die Schaffung der Voraussetzungen in Fürstenfeldbruck und München wurden weitere 29 000 € aus dem Großgeräteantrag verwendet.

[42] `http://www.astaro.com`

[43] VPN steht für „Virtual Private Network“ und ist ein Protokoll, um verschiedene Standorte virtuell und verschlüsselt miteinander zu verbinden.

Laut Auskunft der zentralen EDV-Beschaffung der Ludwig-Maximilians-Universität stellte die zweite Ausbaustufe eine Folgebeschaffung der Ersten im Jahre 2005 dar. Deshalb war es nicht notwendig eine öffentliche Ausschreibung durchzuführen. Stattdessen wurden von der Abteilung für EDV-Beschaffung entsprechende Angebote eingeholt und daraufhin die Hard- und Software für den Ausbau des Massenspeichersystems bestellt.

2.3.2 Umsetzung

Die erste Ausbaustufe der Massenspeicherinfrastruktur wurde zusammen mit dem Rechencluster im Oktober 2005 ausgeschrieben. Wie bereits erwähnt passte nur eines der Angebote in den finanziellen Rahmen und die Firma Microstaxx GmbH aus München erhielt den Zuschlag für die Lieferung des Massenspeichersystems. Von dem mittelständigen Unternehmen wurde eine Speicherinfrastruktur auf Basis der bereits angesprochenen NetApp-Technologie entsprechend des Anforderungsprofils in der öffentlichen Ausschreibung angeboten. Das System beinhaltet eine als Filerhead bezeichnete FAS 3020 Grundeinheit, auf der die gesamte Logik (Hard- und Software) sowie die Schnittstellen nach außen enthalten sind. Dazu wurden noch drei Festplatten-Shelfves geliefert. Davon sind zwei der Shelfves mit jeweils vierzehn 144 GB fassenden FC-Festplatten ausgestattet und das verbleibende Shelf verfügt über vierzehn 250 GB fassende SATA-Festplatten. Es wurde zusätzlich für das NFS- und CIFS-Protokoll jeweils eine Lizenz erworben. Die Lieferung des Systems der ersten Ausbauphase erfolgte im Dezember 2005. Den Aufbau der Hardware übernahm wie beim Rechencluster die liefernde Firma Microstaxx GmbH. Das fertig in den 19 Zoll Rackschrank integrierte FAS 3020 System ist in Abbildung 2.17(a) zu sehen.

Bevor die Inbetriebnahme der Speicherinfrastruktur angesprochen wird, sollen für die Nachvollziehbarkeit die Komponenten der zweiten Ausbaustufe angesprochen werden. Die zentrale EDV-Beschaffung der LMU holte Angebote für den weiteren Ausbau des NetApp Systems ein. In kontrovers geführten Verhandlungen mit den Firmen NetApp und Microstaxx konnte eine zufrieden stellende Konfiguration gefunden werden, die zudem auch finanziell in den Rahmen der zur Verfügung stehenden Mittel passte. Die Lieferung des Systems erfolgte Ende Februar 2008 nachdem die Netzwerkvoraussetzungen geschaffen waren. Diese zweite Ausbaustufe umfasst drei weitere Festplatten-Shelfves mit jeweils vierzehn 750 GB fassenden SATA-Festplatten für die Erweiterung der Kapazitäten in München. Für den Standort Fürstenfeldbruck konnten ein FAS 3020 Filerhead und zwei jeweils vierzehn SATA-Festplatten umfassende Festplatten-Shelfves bezogen werden. Neben den Lizenzen für das NFS- und CIFS-Protokoll für beide Standorte war weiterhin eine Lizenzierung für SnapMirror notwendig. Das vom Lehrstuhl für Geophysik aufgebaute System ist in Abbildung 2.17(b) dargestellt.

(a) München (b) Fürstenfeldbruck

Abbildung 2.17: Beide NetApp FAS 3020 Systeme am entsprechenden Standort

Inbetriebnahme

Die Inbetriebnahme aller Systeme des Herstellers NetApp erfolgt durch einen Systemingenieur der Firma. Dabei werden zuerst der Einbau und die richtige Verkabelung der Einzelkomponenten geprüft, bevor das Speichersystem in Betrieb genommen wird. Während des ersten Starts werden die wichtigsten Einstellungen, wie Netzwerkadresse und „root“ Passwort, eingerichtet. Danach wird unter Anleitung des Systemingenieurs die weitere Konfiguration erarbeitet und beispielhaft die verschiedenen Einstellungsmöglichkeiten besprochen. Sollten im Anschluss daran keine Fragen offen sein, so ist die Inbetriebnahme für den NetApp-Mitarbeiter abgeschlossen und es gilt das FAS 3020 an die eigenen Bedingungen anzupassen. Dazu steht der technische Support der Firma jederzeit zur Verfügung.

Die Einarbeitung in die Konfiguration des Massenspeichersystems bedurfte für den ersten FAS 3020 Filer mehr als zwei Wochen bis die Konfiguration genau den Anforderungen entsprach. In Absprache mit dem technischen Support musste in dieser Phase, Anfang Januar 2006, bereits das Betriebssystem aktualisiert werden, um die Netzwerkkommunikation mit dem LDAP-Dienst verschlüsseln zu können. Über das LDAP-Protokoll bezieht das Massenspeichersystem die Nutzerdaten, vorwiegend für das NFS-Protokoll und die Speicherplatzbeschränkungen. Die beiden FC-Shelfves des FAS 3020 bilden für das System eine einzige RAID-Gruppe der Größe 2.7 TB, in welcher zwei Volumes konfiguriert sind. Im ersten ist das Betriebssystem installiert und verfügt über 15 GB Speicherplatz. Die Benutzerverzeichnisse werden im zweiten Volume der Größe 2.65 TB gespeichert (GNU/Linux bindet dieses Volumen nach `/home` ein). Die Nutzung der schnelleren FC-Festplatten ermöglicht für diese beiden Speicherbereiche einen Leis-

tungsvorteil, der gerade für die Benutzerdaten wichtig erscheint. Ein einfacher Test mit dem `/home` Volume sollte die Leistungsfähigkeit bestätigen. Dazu wurde der Speicherbereich über NFS Version 3 in das System eingebunden und das Programm `dd` schrieb kontinuierlich Daten in dieses Dateisystem. Zwischen beiden Systemen besteht eine 1-GBit-Ethernet Netzwerkverbindung. Nach kurzer Zeit erreichte die Datentransferrate vom Rechner zum Massenspeicher die Kapazitätsgrenze der Netzwerkverbindung und blieb bis zum Ende auf diesem hohen Niveau. Die transferierten 125 MB/s konnten durch das FAS 3020 System ohne Probleme auf die Festplatten geschrieben werden. Während dieses Tests zeigte der im „Data ONTAP“ Betriebssystem enthaltene Systemmonitor keine erhöhte Auslastung. Dies legt den Schluss nahe, dass vom Massenspeichersystem noch mehr Leistung erwartet werden kann. Ein Shelf mit SATA-Festplatten wurde im Anschluss als Datenspeicherbereich für die Simulationsergebnisse eingerichtet. Die Migration der Benutzerdaten von dem bis dahin verwendeten System konnte Ende Januar 2006 abgeschlossen werden. Seit diesem Zeitpunkt läuft das System ohne nennenswerte Probleme durchweg stabil und performant. Einzig für Softwareaktualisierungen oder den Einbau der zusätzlichen Shelfves der zweiten Ausbaustufe musste das System neu gestartet werden. Mit der Neustrukturierung des Microsoft Windows Netzwerks am Lehrstuhl verwenden alle Nutzer das identische Benutzerverzeichnis auf GNU/Linux und Microsoft Windows. Dies stellt eine erhebliche Erleichterung für die Benutzer und Systemverwalter dar.

Über den Zeitraum von Januar 2006 bis März 2008 konnte ausreichend Erfahrung mit dem Massenspeichersystem FAS 3020 gesammelt werden. Damit war es möglich, innerhalb eines Arbeitstages nach der Inbetriebnahme durch den NetApp-Ingenieur, das System vollständig in das Netzwerk des Observatoriums einzubinden. Die 750 GB fassenden SATA-Festplatten der beiden Shelfves sind zu einer RAID-Gruppe mit 12 TB Kapazität zusammengefasst. In dieser wurden sieben Speicherbereiche angelegt. Im Ersten der Größe 20 GB ist wieder das Betriebssystem installiert. Das zweite als Benutzerverzeichnis `/home` konfigurierte Volume umfasst 1.5 TB. Die verbliebene Kapazität verteilt sich auf die Datenspeicherbereiche für die gemessenen Daten der Seismologie und Magnetik. Die Auflistung in Tabelle 2.22 gibt einen Überblick über die konfigurierten Datenspeicherbereiche und nennt deren Verwendung. In der zweiten Ausbaustufe wurden dem FAS 3020 System in München drei weitere Shelfves mit jeweils vierzehn 750 GB SATA-Festplatten hinzugefügt. Die bereits vorhandene RAID-Gruppe für die Simulationsdaten vergrößert sich damit, sodass jetzt 10 TB Speicherkapazität zur Verfügung stehen. Außerdem erfolgte die Einrichtung einer neuen RAID-Gruppe, welche die in Tabelle 2.22 aufgeführten Volumes beinhaltet. Diese dienen der Synchronisation der Messdaten vom Geophysikalischen Observatorium Fürstenfeldbruck zum Standort des Lehrstuhls für Geophysik in München. Zum Zwecke der Datenspiegelung wird eine als Snapmirror bezeichnete Beziehung zwischen beiden Massenspeichersystemen hergestellt, wofür die VPN-Verbindung notwendig ist. Danach erfolgt die Synchronisation der Volumes in zeitlich vorgegebenen Intervallen. Bei

Volume	Kapazität	Verwendung
bay200	7.5 TB	Messdaten der festen Stationen des Bayernnetzes
bay_mobil	1.5 TB	Messdaten der mobilen Stationen des Bayernnetzes
bay_event	0.5 TB	Event-Daten des Bayernnetzes
magsphere	0.5 TB	geprüfte Messdaten der Geomagnetikstation (FUR)
lamont	0.25 TB	ungeprüfte Messdaten der Geomagnetikstation (FUR)

Tabelle 2.22: Volumes für die Messdaten des Observatoriums in Fürstenfeldbruck und des Datenspiegels in München

jeder anstehenden Spiegelung wird auf beiden Systemen durch Snapmirror ein Snapshot des Dateisystems angelegt und nur die geänderten Datenblöcke werden übertragen. Dieses effiziente Synchronisationsverfahren stellt einen vollständigen Datenspiegel am Lehrstuhl sicher. Allerdings sind die Datensätze in München nur les- und nicht schreibbar, was jedoch vollkommen ausreichend ist, um allen Forschern weltweit die Messdaten mit hoher Geschwindigkeit für ihre Projekte anzubieten. Durch diese Aufgabenteilung zwischen beiden Standorten wird eine gleichmäßige Auslastung der Ressourcen erreicht. Mit Stand September 2008 läuft die Datenspiegelung seit nunmehr fünf Monaten ohne nennenswerte Probleme.

2.3.3 Auslastung

Eine Überwachung der Last verschiedener Systemkomponenten der beiden FAS 3020 ist gegenwärtig nicht vorhanden. Aus diesem Grund können keine Graphen für beispielsweise den Anstieg der Speichernutzung in verschiedenen Volumes gezeigt werden. Allerdings scheint das zweiteilige Konzept gut zu passen, da nach der zweijährigen Nutzung das Volume für die Simulationsdaten zu 90 Prozent gefüllt war und mit der zweiten Ausbaustufe und der Erweiterung genau zum richtigen Zeitpunkt entsprechende Neukapazitäten verfügbar waren. Bis August 2008 hat sich dieser Speicherbereich bereits wieder zu über 40 Prozent gefüllt, was etwa 4 TB entspricht. Die Benutzerverzeichnisse am Lehrstuhl in München erreichen zum gleichen Zeitpunkt eine Größe von 1.5 TB, es sind also 65 Prozent der Kapazität bereits belegt. Wohingegen am Observatorium gerade einmal 20 Prozent der 1.5 TB belegt werden. Die Messdaten der Seismologie und Magnetik belegen zwischen 10 und 40 Prozent des zur Verfügung stehenden Speicherplatzes. Die belegte Kapazität wird sich durch den kontinuierlichen Datenzuwachs der nächsten Jahre beständig vergrößern.

Würde man aus der aktuellen Speicherplatzbelegung der Simulationsdaten und deren Entwicklung einen Rückschluss auf einen Zeitpunkt für die nächste Ausbaustufe dieses Volumes ziehen, so müsste in 3 Jahren neuer Speicherplatz erworben werden.

2.4 SMP-System

2.4.1 Konzeption

Die gewonnenen Erkenntnisse aus den einzelnen Simulationsanwendungen auf TETHYS ließen Anfang 2008 erste Überlegungen reifen, ein Rechnersystem zu beschaffen, das in seinen Charakteristika so nicht am Lehrstuhl vorhanden war, aber für eine ausgewogene Infrastruktur notwendig erscheint. Da die beiden Anwendungen TERRA und YAC, wie viele der aktuell laufenden Programme auch, in den Testläufen Potential für weitere Verbesserungen auf Mehrkernsystemen zeigten, sollte der angedachte Rechner über entsprechende Hardware verfügen. Im März 2008 bot sich dann die Möglichkeit, aus Mitteln des Großgeräteantrags zum Aufbau einer Massenspeicherinfrastruktur ein solches System zu erwerben.

Unter einem Mehrkernsystem versteht man einen Rechner, bei dem jeder Prozessor mehrere Rechenkerne besitzt. Dies bedeutet, dass im „Die“[44] der gegenwärtig erhältlichen CPUs zwei, vier oder acht solcher Kerne verschmolzen sind. Für die Parallelisierung der Anwendungsprogramme eröffnet diese Technologie neue Möglichkeiten, da der Aufwand, ein parallel arbeitendes Programm zu erstellen, durch die geschickte Verwendung von Softwarebibliotheken und Compiler Optionen reduziert werden kann. In derartige Rechnerhardware kann meist auch eine größere Menge an Arbeitsspeicher integriert werden, womit sich weitere Nutzungsmöglichkeiten für dieses System ergeben. Konkret könnte es sich um Software handeln, die keine eingebaute Parallelisierung hat aber dennoch auf viel RAM zugreifen muss. Als Beispiele für derartige Anwendungsprogramme wären die 3D-Gittergenerierung oder 3D-Datenvisualisierung zu nennen. Damit bestand für die Geophysik die Möglichkeit, mit Hilfe eines Rechners ein breites Anforderungsprofil abzudecken.

2.4.2 Umsetzung

Eine zusammen mit der zentralen EDV-Beschaffung der LMU durchgeführte Recherche ergab, dass sich solche als SMP-Systeme[45] bezeichnete Rechner in einem akzeptablen Preisrahmen bewegen, wenn vier CPUs und 128 GB Arbeitsspeicher ausreichend sind. Diese Ausstattungsmerkmale entsprachen den Vorstellungen zum neuen SMP-Rechner. So wurden von der EDV-Beschaffung entsprechende Angebote eingeholt. Die drei angebotenen Systeme von Sun Microsystems, Fujitsu-Siemens-Computers und HP waren für die Vergleichbarkeit alle identisch

[44] Als „Die“ wird der kleine Block aus Halbleitermaterial bezeichnet, der sich in der Mitte des Prozessors befindet und leicht metallisch leuchtet. Auf diesem Block sind alle integrierten Schaltkreise untergebracht.

[45] SMP kann für „Symmetric Multiprocessing“ und/oder „Shared Memory Processing“ stehen.

ausgestattet (vier Intel Xeon X7350 CPUs mit jeweils vier Rechenkernen, 128 *GB* Arbeitsspeicher und drei 147 *GB* große Festplatten mit RAID-Controller). Keines der angebotenen Systeme trat durch weitere Merkmale besonders hervor, sodass am Ende der Angebotspreis entschied. Der Auftrag wurde an die Firma Schulz Bürozentrum GmbH aus München für ihren offerierten HP ProLiant GL 580 G5 Server erteilt. Lieferung, Einbau und Inbetriebnahme des Rechners erfolgte daraufhin im April 2008. Durch die zentralisierte Installationsinfrastruktur auf FAI-Basis war die eigentliche Inbetriebnahme des SMP-Systems (Aufspielen des Betriebssystems) innerhalb weniger Stunden durchgeführt. Die genauen technischen Spezifikationen des Rechners sind in Tabelle 2.23 zusammengestellt. Abbildung 2.18 zeigt das System eingebaut in den 19 Zoll Rackschrank mit herausgezogenem Mainboard. Auf dem von HP auch als „Computing Board“ bezeichnetem Mainboard sind die vier CPU-Sockel mit den kupfernen Kühlkörpern und die eingebauten 128 *GB* Arbeitsspeicher zu erkennen. Für die Beschaffung des Systems waren 26 000 € aus Mitteln des Antrags im Großgeräteprogramm notwendig.

Modell:	HP ProLiant GL 580 G5
Prozessortyp:	Intel Xeon X7350 QuadCore (EM64T)
Anzahl Prozessoren:	vier
Anzahl Rechenkerne:	vier pro Prozessor / sechzehn insgesamt
Taktfrequenz:	2.93 *GHz*
L1 Cache:	32/32 *kB* (Daten/Instruktionen) pro Rechenkern
L2 Cache:	zweimal 4 *MB* (Daten + Instruktionen) pro Prozessor
Arbeitsspeicher:	128 *GB* (DDR2)
Festplatte:	dreimal 147 *GB* (SAS) im RAID-1 Modus
Netzwerkanschluss:	zweimal 1-GBit-Ethernet (onboard)

Tabelle 2.23: Hardwarespezifikation für das beschaffte SMP-System

Bei der Namenswahl für das SMP-System zeigten die Mitarbeiter der Geophysik Kreativität und entschieden sich für COREDUMP, allerdings verbunden mit der Hoffnung, dass derartige Ereignisse nicht allzu oft auftreten. Als Betriebssystem wird auf dem Rechner Debian GNU/Linux Etch (AMD64[46]) eingesetzt, wie mittlerweile auf allen Arbeitsplatzrechnern am Lehrstuhl. Diese neuere Distribution unterscheidet sich von der auf TETHYS eingesetzten im Wesentlichen durch aktualisierte Programmversionen. Für die Entwicklung von Simulationsoftware sind dabei von besonderem Interesse die neueren Versionen der MPI-Bibliotheken MPICH (1.2.7) und OpenMPI (1.2.6) sowie die aktuelleren GCC Compiler. Neben Etch hielt ebenso die

[46] siehe Abschnitt A.3

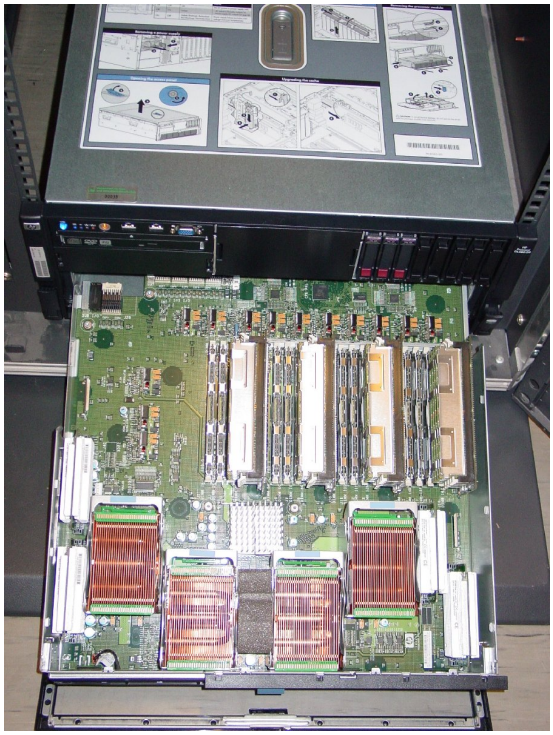


Abbildung 2.18: COREDUMP: HP ProLiant GL 580 G5 Server mit herausgezogenem Mainboard („Computing Board“) – zu sehen sind die vier CPU-Sockel mit den kupfernen Kühlkörpern und die 128 *GB* Arbeitsspeicher

Intel „Cluster Toolkit Compiler Edition“[47] (CTCE) 3.1 Einzug in die Softwarekonfiguration. Mit der Anschaffung dieses Softwarepakets soll dem Umstand Rechnung getragen werden, dass in zukünftigen Projekten verstärkt an der Entwicklung und Optimierung der Simulationsanwendungen geforscht wird, was entsprechende Software voraussetzt. In einigen Projekten werden die verschiedenen Komponenten der CTCE bereits getestet.

2.4.3 Benchmarks

Um eine Bewertung der Leistungsfähigkeit des SMP-Systems zu ermöglichen, wurden mit TERRA einige vergleichende Benchmarkrechnungen durchgeführt. Aus den Tests im Vorfeld der HPC-Beschaffung und zum Nachweis der Leistungsfähigkeit von TETHYS kam wieder die Konfiguration für MT=64 zum Einsatz (siehe Abschnitt 2.1.1 und 2.1.3). Die Laufzeiten der Testrechungen für ein, vier, acht und sechzehn Prozessoren sind in Tabelle 2.24 zusammengetragen. Für die Vergleichbarkeit sind die Laufzeiten und der Speedup für die identische Untersuchung auf TETHYS (siehe Tabelle 2.19) mit in Tabelle 2.24 eingefügt. Vergleicht man die Laufzeiten des Einzelprozessorlaufs von COREDUMP mit denen von TETHYS, so ist diese

[47] `http://www.intel.com/cd/software/products/asmo-na/eng/375500.htm`

auf dem SMP-System für die identische Problemgröße (MT=64) um 33 % geringer und spiegelt damit die Erwartungen an die aktuelleren Prozessoren wider. Darüber hinaus erkennt man anhand der Laufzeiten, dass TERRA auf dem Mehrkernrechner bis hin zu acht eingesetzten Prozessoren/Rechenkernen gut skaliert. Die Laufzeit für P=16 ist aber etwa 12 Prozent höher als die für P=8. Dieses Verhalten kann vielleicht durch die Tatsache erklärt werden, dass alle Rechenkerne in Konkurrenz um Rechnerressourcen stehen. Dies betrifft beispielsweise den Zugriff auf den Arbeitsspeicher oder auch den L2 Cache, den sich zwei der vier Rechenkerne bei den verbauten Intel Xeon Prozessoren teilen müssen. Beim Einsatz aller sechzehn Kerne sind somit keine freien Reserven mehr vorhanden. Diese Beobachtung stimmt mit denen anderer Einrichtungen überein. Demnach ist es nicht immer ratsam, gleich viele Prozesse wie Rechenkerne zu nutzen[48]. Der Vergleich des parallelen Speedup für COREDUMP und TETHYS ergibt für beide einen Anstieg bis P=8. Für sechzehn Prozessoren fällt der Speedup bei COREDUMP aus den oben genannten Gründen wieder.

	COREDUMP		TETHYS	
Prozessoranzahl P	$\overline{t}\ [s]$	$S_{par}(P)$	$\overline{t}\ [s]$	$S_{par}(P)$
1	568	1	851	1
4	182	3.1	227	3.8
8	98	5.8	104	8.2
16	111	5.1	60	14.1

Tabelle 2.24: TERRA: Benchmark für die Problemgröße MT=64 und 100 Zeitschritte – Vergleich zwischen dem SMP-System COREDUMP und dem HPC-Rechner TETHYS

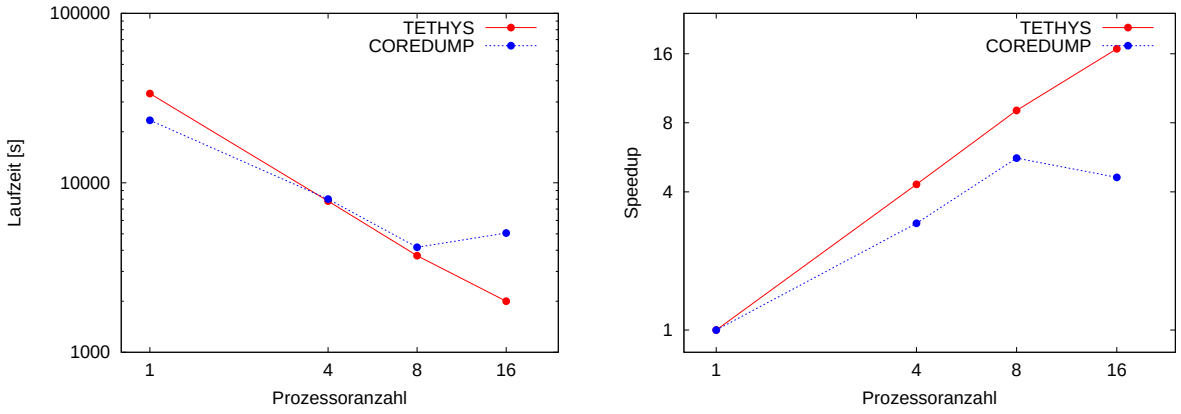
Im weiteren Verlauf sollten mit der Konfiguration aus den Skalierungsuntersuchungen (Gittergröße MT=128, siehe Seite 49) die Laufzeiten sowie der Speedup für Prozessoranzahlen von eins, vier, acht und sechzehn ermittelt werden. Die Graphen in Abbildung 2.19 stellen die Ergebnisse dieser Untersuchung ebenso wie Tabelle 2.25 dar. In beiden sind zum Vergleich die Werte für TETHYS (siehe Abschnitt 2.1.3) eingetragen. Wie bei der Problemgröße MT=64 sinkt die Laufzeit bis zu acht Prozessoren, um dann für sechzehn Prozessoren wieder zu steigen. Dementsprechend verhält sich auch der Speedup, bei dem eine gute Skalierung bis hin zu acht Prozessoren zu verzeichnen ist. Allerdings liegen die Laufzeiten schon bei vier und acht genutzten CPUs leicht über den vergleichbaren Zeiten auf TETHYS. Dieses Verhalten entspricht nicht dem aus dem vorhergehenden Test mit MT=64. Eine Beobachtung der laufenden Prozesse während der Rechnungen bestätigte den Verdacht, dass vom Betriebssystemkern die Prozesse

[48] persönliche Kommunikation mit Dr. M. Mohr, Geophysik

zwischen den Prozessoren verschoben werden. Damit bleiben wertvolle Rechenzeitscheiben[49] ungenutzt, da unter anderem der CPU-Cache erneuert werden muss und auch das Verschieben Zeitscheiben kostet. Es muss demnach unbedingt die Prozessoraffinität erhöht werden. Es besteht die Möglichkeit dies über die MPI-Laufzeitumgebung, die Nutzung von CPU-Sets (Linux-Betriebssystemoption, um für Prozesse die genutzten Prozessoren festzulegen) oder mit Hilfe des `schedtool` Programms zu realisieren. Eine ausführlichere Darstellung der entsprechenden Möglichkeiten im Linux Betriebssystemkern ist in Kunst & Quade (2008) enthalten.

	COREDUMP		TETHYS
Prozessoranzahl P	$\bar{t}\ [s]$	$S_{par}(P)$	$\bar{t}\ [s]$
1	23368	1	33648
4	8009	2.9	7800
8	4162	5.6	3715
16	5051	4.6	2002

Tabelle 2.25: TERRA: Benchmark für die Problemgröße MT=128 und 500 Zeitschritte – Vergleich zwischen dem SMP-System COREDUMP und dem HPC-Rechner TETHYS – Für P=1 musste die Laufzeit auf TETHYS geschätzt werden, da nicht genügend Arbeitsspeicher zur Verfügung stand. Darum wurde auf die Angabe des Speedup verzichtet.



(a) Mittelwert der Laufzeit für MT=128 auf TETHYS und COREDUMP

(b) paralleler Speedup für MT=128 auf TETHYS und COREDUMP

Abbildung 2.19: TERRA: Vergleich der Laufzeiten und des Speedup zwischen TETHYS und COREDUMP für die Problemgröße MT=128 und 500 Zeitschritte

[49] Der Linux Betriebssystemscheduler teilt die zur Verfügung stehende Rechenzeit in Zeitscheiben der Dauer 100 ms ein.

Laut Dokumentation benutzt OpenMPI eine fest eingebaute Variante zur Steuerung der Prozessoraffinität eines MPI-Prozesses. In einer Untersuchung sollte geklärt werden, ob die vermutete Laufzeitreduktion eintritt. Zu diesem Zweck wurde die vorher genutzte Vierprozessorversion von TERRA mit der Gittergröße MT=128 und 500 Zeitschritten eingesetzt. Unmittelbar nach dem Start des Testlaufs mit OpenMPI konnte in der Ausgabe des `top` Befehls beobachtet werden, dass jeder einzelne der vier MPI-Prozesse fest einer CPU zugeordnet war. Dies führte dazu, dass die CPU-Rechenzyklen nicht ungenutzt blieben und somit die Laufzeit von 8009 s auf 6243 s sank. Die erzielte Laufzeitverringerung um 28 % ist beachtlich und zeigt deutlich die vorhandenen Optimierungsmöglichkeiten auf.

Für den zweiten Test bei der Betrachtung des Einflusses der Prozessoraffinität wurde auf dem SMP-System COREDUMP ein CPU-Set erzeugt. In diesem befanden sich jeweils nur einer der vier Rechenkerne. Jeder der vier vorhandenen Prozessoren wurde dem CPU-Set exklusiv zugeordnet. Diese Rechenkerne waren damit nur durch TERRA für die Laufzeituntersuchung nutzbar. Wie bei dem OpenMPI-Test auch, konnte in der `top` Ausgabe beobachtet werden, dass die Prozesse entsprechend der Vorgabe des CPU-Sets nur auf den gewählten Prozessoren ausgeführt wurden. Dementsprechend sank auch die Laufzeit auf 6261 s ab und ist damit mit der des OpenMPI-Testlaufs vergleichbar. Nutzern ohne spezielle Berechtigungen ist es nicht gestattet, auf einem GNU/Linux System solche CPU-Sets anzulegen.

Da der Befehl `schedtool` die identischen Systemaufrufe wie OpenMPI benutzt, wurde auf entsprechende Testläufe verzichtet. Für den Anwender ist mit Sicherheit die Verwendung von OpenMPI der einfachste Weg, die Prozessoraffinität zu erhöhen.

Tabelle 2.26 stellt die Laufzeiten der Untersuchungen zum Einfluss der Prozessoraffinität nochmals zusammen.

	MPICH	OpenMPI	CPU-Set
Prozessoranzahl P	$\overline{t}\ [s]$	$\overline{t}\ [s]$	$\overline{t}\ [s]$
4	8009	6243	6282

Tabelle 2.26: TERRA: Benchmark für die Problemgröße MT=128 und 500 Zeitschritte – Laufzeiten für vier Prozesse auf COREDUMP – Vergleich der beiden Möglichkeiten die Prozessoraffinität (OpenMPI und CPU-Set) festzulegen mit MPICH

In einer abschließenden Untersuchung sollte betrachtet werden, wie die Intel Compiler Option für die automatische Parallelisierung des Programms durch den Compiler die Laufzeit der Einzelprozessorversion beeinflusst. Zur Verwendung kam wieder die TERRA-Konfiguration mit der Gittergröße MT=128 und 500 Zeitschritten. Das Programm wurde dementsprechend mit

der Option `-parallel` neu übersetzt. Schon kurz nach dem Start von TERRA konnte in der Ausgabe von `top` beobachtet werden, wie den einzelnen Rechenkernen die „Threads“[50] von TERRA zugeordnet wurden. Zur Laufzeit des Programms werden per se so viele „Threads“ generiert, wie Rechenkerne/CPUs vorhanden sind. Der Intel Compiler identifizierte demnach im Quelltext von TERRA Programmteile, die automatisch in mehrere „Threads“ aufgeteilt werden können, um einen beschleunigten Programmablauf zu erreichen. Die erhoffte Laufzeitreduzierung trat durch die gemessenen 20163 s auch ein. Verglichen mit den vorher bestimmten 23368 s ohne die Compiler Option ergibt sich ein Laufzeitvorteil von 13.7 %. Neben der Option automatisch so viele „Threads“ wie Rechenkerne durch das Programm erzeugen zu lassen, besteht noch die Möglichkeit über die Umgebungsvariable `OMP_NUM_THREADS` die Anzahl der „Threads“ einzustellen. Bereits bei der Bewertung der Ergebnisse in Tabelle 2.24 konnte festgestellt werden, dass der Einsatz aller Rechenkerne auf COREDUMP nicht sinnvoll erscheint. Aus diesem Grund wurde über die erwähnte Umgebungsvariable die Anzahl der von TERRA zu nutzenden „Threads“ auf acht reduziert. Damit konnte eine Verringerung der Laufzeit auf 19990 s erreicht werden. Im Vergleich zu den 20163 s bei sechzehn „Threads“ ist dies nochmals eine Leistungssteigerung (14.5 % Laufzeitvorteil zu den 23368 s ohne automatische Parallelisierung) und eine Bestätigung dafür, nicht alle Rechenkerne auf dem SMP-System zu verwenden. Die automatische Parallelisierung durch den Compiler ist demzufolge eine sinnvolle Optimierung vor allem für Programme, die nicht in parallelisierter Form vorliegen.

Bei vorhandenem Programmquelltext besteht für diese Anwendungen darüber hinaus die Möglichkeit OpenMP (`http://openmp.org`) einzusetzen. Mit dieser standardisierten Programmierschnittstelle kann der Anwender auf SMP-Systemen eine einfache Parallelisierung erreichen, indem er durch spezielle Direktiven den Compiler anweist, beispielsweise die Abarbeitung von Schleifen in mehrere „Threads“ aufzuteilen. Falls der Compiler die OpenMP Direktiven nicht kennt, so werden sie ignoriert. OpenMP stellt demnach eine einfache Möglichkeit dar, die eigenen Anwendungen zu beschleunigen.

[50] Ist ein Ausführungsstrang des Programms, der in diesem Fall durch den Compiler ermöglicht wird.

2.5 3D-Visualisierung

Die am Lehrstuhl für Geophysik verfügbare Infrastruktur für die Simulation komplexer Prozesse im Erdkörper (siehe Abschnitt 2.1 und 2.4) generiert sehr viele große Datensätze. Diese werden durch die am geophysikalischen Observatorium in Fürstenfeldbruck gewonnenen Messdaten noch zusätzlich erhöht. Mit der Massenspeicherinfrastruktur ist der Lehrstuhl in der Lage, die Datenmengen zuverlässig, hochperformant und über längere Zeiträume hinweg vorrätig zu halten. Sollen diese Daten später zeitgemäß graphisch aufbereitet werden, so stößt man mit den derzeit genutzten Darstellungsmethoden leicht an die Grenzen des Möglichen.

2.5.1 Konzeption

Die aktuell eingesetzten numerischen Verfahren bei der geowissenschaftlichen Modellrechnung bestimmen die Lösung für drei Raumdimensionen. Allerdings sind die meisten vorhandenen Visualisierungsprogramme nur in der Lage, Schnitte durch die Räume oder Volumendarstellungen zu erzeugen. Um die Daten in ihrer Komplexität zu betrachten und zu untersuchen, wird ein hohes Maß an räumlicher Vorstellungskraft vom Betrachter verlangt. Ungeübte verlieren dabei recht schnell den Überblick und bedürfen einer zielgerichteten Führung durch die Darstellungen. Vergleichbare Probleme entstehen auch, wenn Abbildungen erstellt werden, da beispielsweise schon im Voraus bekannt sein muss, welche Wertebereiche darzustellen sind. Häufig ist eine spätere interaktive Anpassung oder Änderung der graphischen Darstellungen nur schwierig zu bewältigen.

In Diskussionen zu den aufgezeigten Problemen entstand die Idee, ein 3D-Visualisierungslabor am Lehrstuhl für Geophysik einzurichten. Mit der Umsetzung der Pläne zu solch einem Labor wurde Ende 2004 begonnen. Für die dreidimensionale Darstellung geowissenschaftlicher Datensätze sollte der Einsatz von standardisierter PC-Technik eine kostengünstige Lösung bieten. Als weitere Anforderungen an das System sind zu nennen:

- effizienter Umgang mit großen Datensätzen (mehrere GB),
- interaktive und intuitive Dialoge für die Änderung visueller Parameter,
- Darstellung mit räumlichen Seheindruck,
- Auslegung für größere Benutzergruppen,
- möglichst ohne Lizenzkosten für weitere Software.

2.5.2 Umsetzung

Durchgeführte Internetrecherchen zeigten, dass für diese Zielstellung eine sogenannte „Geowall“[51] die richtige Lösung darstellt. Um im menschlichen Gehirn einen dreidimensionalen Seheindruck zu erzeugen, müssen die dargestellten Bilder perspektivisch von zwei minimal verschiedenen Blickwinkeln generiert werden. Durch unterschiedliche Bilder für jedes Auge, entweder mit:

- kleinen Bildschirmen,
- Brillen mit Prismen,
- beiden Bildern auf einem Bildschirm und Filterbrillen für den Betrachter,

ist dies umsetzbar. Für die letztgenannte Methode können verschiedene Filtertechniken zum Einsatz kommen. Bei der „Geowall“ werden polarisierende Filter genutzt. Damit sind Abbildungen mit vollem Farbraum in Kombination mit preiswerten passiven Filterbrillen möglich, allerdings zum Preis einer aufwendigeren Bildschirmkonfiguration, die polarisiertes Licht darstellen kann.

Entsprechend den Angaben auf der Projektseite (`http://www.geowall.org`) wurde die Technik in einem vormals als Lager genutzten Raum aufgebaut. Dieser befindet sich mitten im Gebäude und bedarf somit keiner Tageslichtverdunklung. Nach Abschluss der notwendigen Vorbereitungsarbeiten konnte mit dem Aufbau des Projektionssystems begonnen werden.

Die Liste der benötigten Hardware umfasst:

zwei DLP-Projektoren: Obwohl die beiden NEC LT 245 DLP-Projektoren[52] sehr heiß werden, eignen sie sich für eine Montage übereinander (siehe Abbildung 2.20). Sie können die zusätzliche Abwärme des anderen Gerätes auch über einen längeren Nutzungszeitraum hinweg tolerieren. Die DLP-Technik ist notwendig, da nur damit keine weitere Polarisation des Lichts erfolgt.

zwei gleiche Polarisationsfilter mit passenden Polarisationsbrillen: Die linear polarisierenden Filter aus Glas sind, wie in Abbildung 2.20 zu sehen, über eine selbst gestaltete Halterung vor den Projektoren angebracht. Kunststofffilter verändern durch die Wärme ihre Form und eigenen sich daher nicht. Die notwendigen Brillen für die Betrachter müssen auch linear polarisiert sein, um den dreidimensionalen Seheffekt zu erzeugen.

[51] `http://www.geowall.org`

[52] DLP steht für „Digital Light Processing“. Ein Mikrospiegelarray ist der Basisbaustein eines Verfahrens der Bildprojektion, das auch als DLP bezeichnet wird. Dieser Baustein ist in der Lage mittels matrixförmig angeordneter, einzeln beweglicher Mikrospiegeln und einer starken Lichtquelle ein Bild zu projizieren. (siehe `http://de.wikipedia.org/wiki/DLP-Projektor`)

eine silberne Leinwand: Die Leinwand von Vision24 ist auf einem Holzrahmen montiert (siehe Abbildung 2.21). Es wird eine silberne Projektionsfläche benötigt, da die Reflektionen auf vielen Bildschirmen (vor allem mit weißer Fläche) das Licht polarisieren und damit den eigentlich gewünschten dreidimensionalen Effekt zerstören.

einen Arbeitsplatzrechner mit guter Graphikkarte: Der zu Beginn genutzte PC wurde mit Debian/GNU Linux Sarge betrieben, allerdings erweitert um die offiziellen Treiber von NVidia[53] für deren Quadro FX 1400 Graphikkarte (Quad-Buffered). Mit diesen Treibern kann der Aktiv-Stereo-Modus zusammen mit dem Clone-Modus (beide Projektoren zeigen das identische Bild aus leicht verschiedenen Blickwickeln) aktiviert werden. Die Ausstattung des Rechners mit 2 *GB* Arbeitsspeicher und Intel Pentium 4 CPU mit 3 *GHz* entsprach der Üblichen am Lehrstuhl für Geophysik. Die NEC Projektoren limitierten die Bildschirmauflösung auf 1024 mal 768 Pixel.

Für die gesamte Technik wurden weniger als 5 000 € aus Lehrstuhlmitteln investiert. Dass die eingesetzte Hardware passend ausgewählt wurde, bestätigten die ersten Testbeispiele durch die erreichte hohe Darstellungsqualität und Geschwindigkeit.

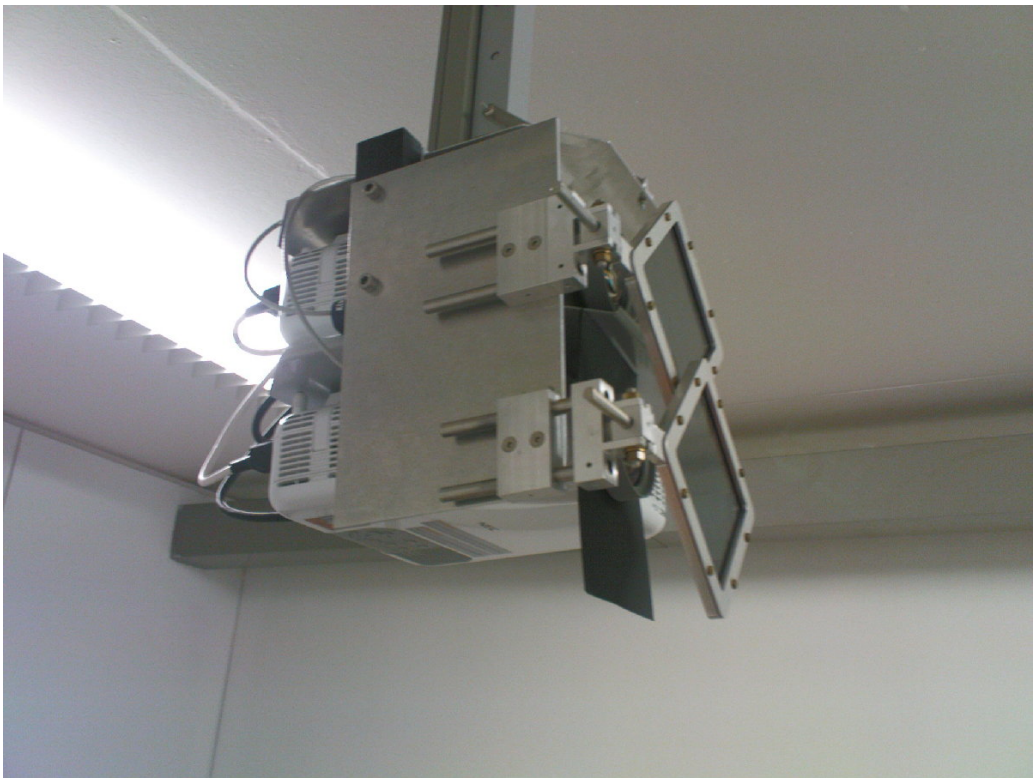


Abbildung 2.20: Anordnung der DLP-Beamer in einer selbstgebauten Halterung zusammen mit den davor angebrachten Polarisationsfiltern

[53] `http://www.nvidia.com`

Abbildung 2.21: 3D-Visualisierungslabor in Benutzung – zu sehen ist ein Schnitt durch die Erde überlagert durch die Küstenlinien

In seiner Diplomarbeit konnte Christoph Moder auf eindrucksvolle Weise zeigen, welche neuen Möglichkeiten sich durch die 3D-Visualisierung für die Geowissenschaften eröffnen (Moder, 2006). So konnte von ihm das Programm `paraview`[54] für den Einsatz auf einem „Geowall“-System entsprechend modifiziert werden. Es wird der Aktiv-Stereo-Modus benötigt. Zusätzlich war es durch eine angepasste Konfiguration beim Übersetzen des Quelltextes von `paraview` möglich, die vorhandene MPI-Unterstützung für die Zusammenarbeit mit dem HPC-System TETHYS bei Visualisierungsaufgaben zu verwenden (Moder, 2006; Moder et al., 2007). Bedingt durch einen Hardwaredefekt am Rechner wurde nach mehr als zweijähriger intensiver Verwendung des 3D-Visualisierungslabors[55] entschieden, ein leistungsfähigeres System zu beschaffen. Dies erschien in Anbetracht des häufig während Visualisierungsversuchen auftretenden Arbeitsspeichermangels sinnvoll. Der neue, hochwertige Arbeitsplatzrechner (FSC Celsius M 420) verfügt über eine Intel Core2 6600 CPU mit zwei Rechenkernen und 2.4 *GHz* Takt-

[54] `http://www.paraview.org`

[55] Für den Raum hat sich in Anlehnung an die „Star Trek“-Serien der Name „Holodeck“ etabliert.

frequenz. Neben den 8 *GB* Arbeitsspeicher ist die für die 3D-Darstellung notwendige NVidia Quadro FX 4600 Graphikkarte im Rechner verbaut. Ebenso wie auf COREDUMP ist die Linux-Distribution Debian/GNU Linux Etch (AMD64) installiert, erweitert um die offiziellen Treiber von NVidia. Die erhofften Leistungssteigerungen mit dem neuen System konnten im letzten halben Jahr nachgewiesen werden.

Im mittlerweile mehr als dreijährigen Nutzungszeitraum hat vor allem Christoph Moder viel Erfahrung mit der 3D-Visualisierung gesammelt. Viele der Abbildungen, die heute in den Veröffentlichungen der Geophysik verwendet werden, sind im Visualisierungslabor entstanden. Doch können diese nicht den 3D-Effekt reproduzieren, der während der Betrachtung der Visualisierungsdaten im „Holodeck“ entsteht. Aus diesem Grund soll beispielhaft nur eine der mit `paraview` erstellten Abbildungen aufgeführt werden, welche freundlicherweise von Christoph Moder zur Verfügung gestellt wurde (siehe Abbildung 2.22). Diese Darstellung zeigt das Ergebnis einer Modellrechnung mit TERRA. Die berechneten Temperaturvariationen bei der Mantelkonvektion sind als Querschnitt durch den Erdmantel unterhalb des Mittelatlantischen Rückens zu sehen. Zur Orientierung sind in dem Bild die Kontinente und Plattengrenzen eingezeichnet. Von einem Temperaturmittelwert ausgehend, werden die Temperaturvariationen im Erdmantel als Isoflächen bei -600 und $+400$ Kelvin gezeigt. Es ist deutlich das aufsteigende heiße Material am Mittelatlantischen Rücken zu sehen.

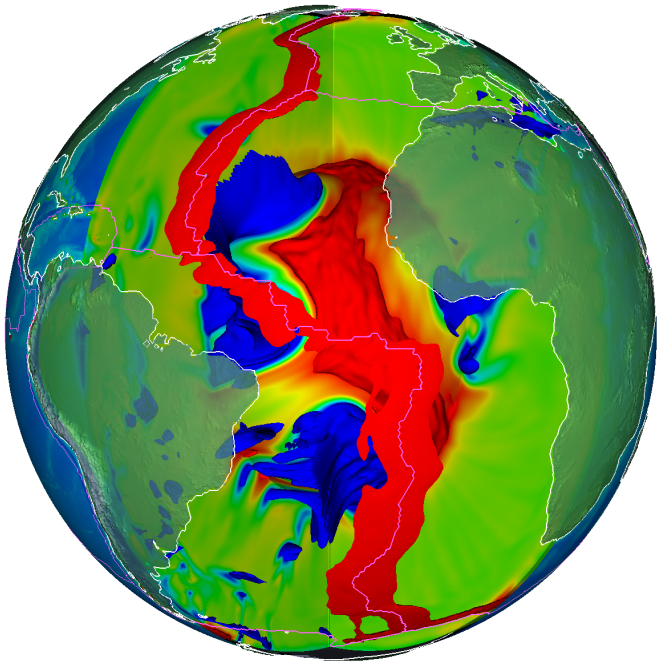


Abbildung 2.22: Temperaturvariation im Erdmantel unter dem Mittelatlantischen Rücken – die Isoflächen der Temperaturvariation sind um einen Mittelwert bei -600 und $+400$ Kelvin gewählt

3 Simulationsanwendungen in der Geophysik

Die folgenden Abschnitte geben einen Überblick zu den verfügbaren parallelisierten Simulationsprogrammen am Lehrstuhl für Geophysik der Ludwig-Maximilians-Universität. Die Mehrzahl der Anwendungen modelliert die Wellenausbreitung für den gesamten Erdkörper beziehungsweise nur für Teilbereiche der Erde. Die eingesetzten numerischen Verfahren reichen von der Methode der finiten Differenzen über die Methode der finiten Elemente bis hin zur Spektrale-Elemente-Methode (SEM). Diese große Bandbreite begründet sich durch die Vor- und Nachteile der eingesetzten Verfahren und den zu bearbeitenden Problemstellungen in der Geophysik. Meist werden die damit gewonnenen Simulationsergebnisse für weiterführende Studien benötigt. Einen Schwerpunkt bildet die Modellierung und Inversion der Erdmantelkonvektion mittels der Finiten-Elemente-Methode (FEM). In dieser Anwendung fließen viele der gewonnen Erkenntnisse zur Struktur des Erdkörpers, der Plattenbewegung und dem chemisch-physikalischen Verhalten des Materials im Erdmantel ein. Bei der Modellierung der Bruchprozesse in der Erde stehen die Quellmechanismen für Erdbeben und die Ausbreitung der entstehenden seismischen Wellen im Vordergrund. Das dabei eingesetzte numerische Verfahren basiert auf der Finiten-Differenzen-Methode (FDM). Die Schadwirkung auf Bauwerke durch seismische Wellen in Gebieten mit komplizierter Bodenstruktur ist ebenso eine Anwendung.

Im Folgenden soll mit einem Überblick zu aktuellen geophysikalischen Modellierungen der globalen Plattentektonik, Mantelkonvektion und seismischen Wellenausbreitung begonnen werden. Das Überschreiten von bisher existierenden Grenzen bei der Simulation von geophysikalischen Phänomenen ist nur durch den kontinuierlichen Anstieg an verfügbarer Rechenleistung auf den verschiedensten HPC-Systemen möglich geworden und wird am Anfang im Mittelpunkt stehen. Anschließend werden in einem kurzen Überblick die mathematischen Grundlagen angesprochen. Darauf folgt die Diskussion der geophysikalischen Simulationsanwendungen für den Rechencluster TETHYS, wobei das Hauptaugenmerk nicht auf dem verwendeten mathematischen Ansatz liegt, sondern vielmehr das Einsatzgebiet eines jeden Programms verdeutlicht werden soll.

3.1 Frontiers in Computational Geophysics: simulations of mantle circulation, plate tectonics and seismic wave propagation

We review recent progress in geophysical modeling of global plate tectonic, mantle convection and seismic wave propagation problems, paying particular attention to novel adjoint methods for the efficient inversion of seismic and tectonic data. We observe that the continuing growth in high performance and cluster computing promises the crossing of long standing barriers in the simulation of first-order geophysical phenomena.

This section was submitted as:
J. OESER, H.-P. BUNGE, M. MOHR AND H. IGEL (2009), **Frontiers in Computational Geophysics: simulations of mantle circulation, plate tectonics and seismic wave propagation**, in *100 Volumes NNFM and 40 Years Numerical Fluid Mechanics*, accepted in Juli 2008.

3.1.1 Introduction

Geophysics differs from other scientific disciplines in its focus on processes one can neither repeat nor control. Examples include the nucleation of an earthquake as brittle failure along faults, or the dynamic processes of ductile (creeping) flow in the earth's interior which give rise to plate tectonics and the endogenic (internally driven) geologic activity of our planet. The inherent experimental limitations and the indirect nature of our observations explain in part why there is such a remarkable impact and success of high-performance computing (HPC) in this field. And indeed many a geophysical observable are only interpretable through the use of sophisticated modelling tools. Another reason for the prominence of HPC lies in the recent crossing of long standing thresholds in capacity and capability computing. This allows us today to implement models having in excess of 1 billion grid points. The development makes it feasible for the first time to overcome in three dimensional (3D) models the great disparity of length scales which characterises important geophysical phenomena: an earthquake rupturing a fault segment over a distance of some 100 km while emanating seismic energy throughout the planet (10,000 km), or the peculiar nature of plate tectonics with deformation concentrated along plate boundaries of 10-100 km width separated by plates of dimension 1,000-10,000 km serve as example. Before we address challenges and recent successes in global geophysical modelling, let us take a brief look at the gross structure and inherent dynamic time scales of our planet.

The earth's interior is complex, consisting of three distinct regions. Starting from the outside there is first the cold lithosphere, which is dominated by brittle behaviour. It then follows the solid mantle, which deforms slowly over geologic time by a mechanism known as ductile creep. Finally near the earth's centre there is the (mostly) liquid core. As a result of convective and other forcings, all three regions are in motion, albeit on different time scales. On the longest time scale solid state convection (creep) overturns the mantle once in about every 100-200 million years (Bunge et al., 1998). This overturn is the primary means by which our planet rids itself of primordial and radioactive heat. Tectonic processes operate on shorter time scales, up to a few million years or so. They include rapid variations in plate motions, which are revealed by the recent arrival in the earth sciences of highly accurate space geodesy techniques, such as the global positioning system GPS (Dixon, 1991). On still shorter time scales of perhaps 1-1000 years convection of the liquid iron core generates the earth's magnetic field through a complicated dynamo process that probably operated throughout much of earth's history (Hollerbach, 1996). Only recently have geophysicists been able to study dynamo action in sophisticated magneto-hydrodynamic models of the core. We will not concern ourselves with these models and refer to the recent review by Glatzmaier (2002). On a time scale of hours to seconds both the core and the mantle are traversed by seismic sound waves, and seismologists

are now turning to computer models to study seismic wave propagation through our planet (Igel, 2002).

3.1.2 Mantle Flow and Circulation Modelling

The mantle comprises approximately 70 $\%$ of the earth's volume and convects with surprising vigour. Its thermal Rayleigh number, estimated at 10^6 to 10^8 (Davies, 1999) exceeds the critical value at which convection begins by a factor of 10^3 to 10^5, yielding flow velocities of 1-10 cm/year and an upper thermal boundary layer (known as plates) of thickness of 50-100 km depth. The advent of powerful computers allows us to resolve the flow in realistic 3D spherical geometry, and a number of high-resolution, parallelised mantle convection models are now available. The models have provided crucial insight into key parameters governing the behaviour of global mantle flow, such as the effects of mantle phase transitions, a depth-wise increase in viscosity and the partitioning of internal (radioactive) and external (core derived) heating (Tackley et al., 1993; Bunge et al., 1996; Zhong et al., 2000; Stemmer et al., 2006).

Mantle convection can mathematically be modelled by a coupled system of three equations, see e.g. Oeser et al. (2006); Stemmer et al. (2006), describing the conservation of mass, momentum and energy. These differ from the standard Navier-Stokes system of convection driven fluid dynamics in that respect that due to the high Prandtl number (on the order of 10^{24}) inertial terms in the momentum equations can be dropped. This reflects the creeping nature of the flow. Note also that for similar reasons Coriolis and centrifugal forces may safely be neglected.

Mass conservation is a constraint on the velocity field of the Stokes problem, and the coupled system of mass and momentum conservation, after discretisation by standard techniques like finite-element and finite-volume approaches, give rise to a saddle-point problem, which one solves for a velocity field satisfying the divergence-free condition. Most mantle convection codes adopt Uzawa-type algorithms for this purpose, see e.g. Benzi et al. (2005), often employing conjugate gradients for the outer and multigrid for the inner iteration. Multigrid employs a hierarchy of nested computational grids, so that near-and far-field components of the momentum balance are effectively solved at once. We show the nested structure of the icosahedral grid adopted in the Terra code (Bunge et al., 1997) as an example in Fig. 3.1.

Similarly one often treats the energy equation through mixed finite volume, finite difference methods for the advected and conducted heat flux. The Péclet number of the mantle is large (in the range of 10-100), that is heat transport in the mantle is controlled primarily by advection outside of thermal boundary layers. This makes finite volume methods, which are conservative and easily adapted to unstructured meshes, an effective solution approach.

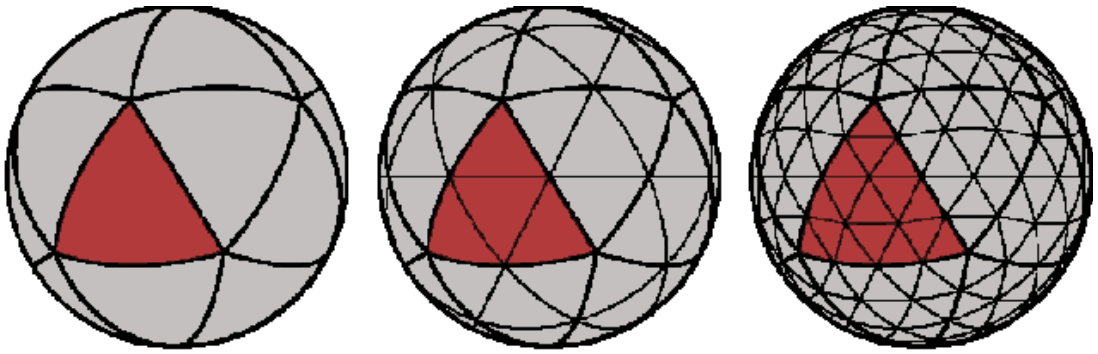


Figure 3.1: Three successive mesh-refinements of the icosahedral grid

It is common to use the term *circulation* to describe the motion of the mantle, in analogy to the general circulation of the oceans and atmosphere. A number of mantle circulation models (MCMs) have been constructed recently (Bunge et al., 1998; Bunge & Grand, 2000; McNamara & Zhong, 2005), and a representative MCM at high numerical resolution (about 100 million grid points) is shown in Fig. 3.2. MCMs differ from traditional convection models in that they include geologic information on the history of subduction (Bunge et al., 1998). This allows them to make explicit predictions on the large-scale thermal structure of the mantle, which is an essential component if one wants to assess the force balance of plate motion.

In general MCMs compare well with tomographic mantle models (Sigloch et al., 2008), which constrain earth structure from independent seismic observations. MCMs suffer, however, in a fundamental way from lack of initial condition information. The difficulty becomes more challenging the further back in time one wants to model the evolution of mantle buoyancy forces, say over the past 10-100 million years. Lack of initial condition information is a problem shared with circulation models of the ocean and the atmosphere.

To overcome the initial condition problem one must formulate a large scale fluid dynamic inverse problem. Essentially one seeks optimal initial conditions that minimise, in a weighted least squares sense, the difference between what a mantle convection model predicts as mantle heterogeneity structure and the heterogeneity one actually infers from, say tomography. This class of problems is known in different contexts as e.g. *history matching* or *variational data-assimilation*, meaning that model parameters are inferred from a variational principle through the minimisation of a cost function F. The necessary condition for a minimum of F, that the variation $\nabla F = 0$, leads to the usual mantle convection equations coupled to a corresponding set of so-called *adjoint equations*.

The adjoint equations, which have been derived recently (Bunge et al., 2003; Ismail-Zadeh et al., 2004), together with large-scale simulations showing that flow can be inferred back in time for at least 100 million years, are nearly identical to the forward model except for forcing terms.

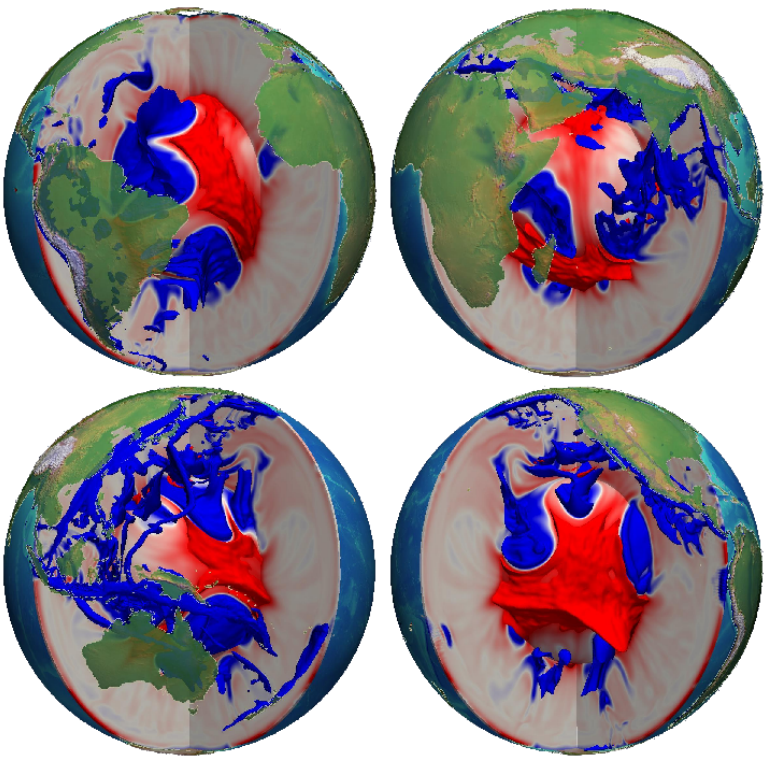


Figure 3.2: 3D representation of temperature variations in a high resolution Mantle Circulation Model (MCM), see text. Shown are four cross sectional views from 35 (upper left), 125 (lower right), 215 (lower left) and 305 (upper right) degrees longitude. Continents with colour-coded topography and plate boundaries (cyan lines) are overlain for geographic reference. Iso-surfaces of temperature are taken to be at -600 and +400 Kelvin. The +400 iso-surface was clipped in the uppermost 500 kilometres in order to allow views into the mantle underneath the mid-ocean-ridge systems which span large parts of the oceanic upper mantle. The colour-scale is saturated at -400 and 400 Kelvin. About 100 million numerical grid points are used, providing a grid point spacing of at most 20 km throughout the mantle, sufficient to resolve the convective vigour of global mantle flow.

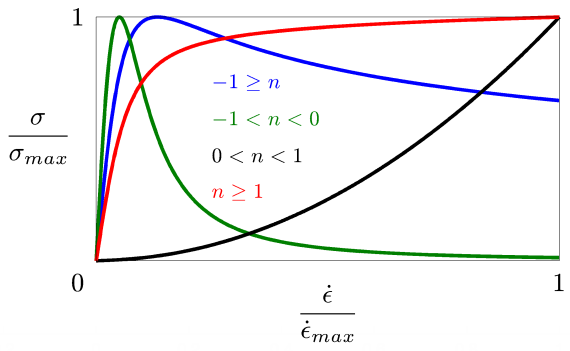
Unfortunately adjoint modelling of global mantle flow at realistic convective vigour comes at a heavy computational price. Weeks to months of dedicated integration time are needed to solve this class of problems even on some of the most powerful parallel machines currently in use. Such resources, however, are coming within reach of topical PC-clusters dedicated to capacity computing (Oeser et al., 2006).

3.1.3 Plate Tectonics and Boundary Forces

A long persistent challenge in geophysics is the computational treatment of plate tectonics, because it is difficult to simulate shear failure along plate boundaries. One strategy, developed more than 30 years ago, models known plate structures and their influence on mantle flow by specifying regions that move in a plate-like manner (Davies, 1989; Ricard & Vigny, 1989; Gable et al., 1991). An alternative approach adopts highly non-linear (non-Newtonian) viscous creep, strain-rate weakening rheologies, and viscoplastic yielding (Zhong et al., 1998; Tackley, 2000; Richards et al., 2001).

In Moresi & Solomatov (1998) explored the effects of strongly temperature-dependent viscosity combined with a plastic yield stress: the former causes the cold upper boundary layer (lithosphere) to be strong, while the latter allows the boundary layer to fail locally in regions of high stress. The success of this *ductile* approach to plate tectonics, measured through a so-called *plateness*, is evident when one applies exotic rheologies with an extreme form of strain softening. One such rheology, where both viscosity and stress decrease with increased strain rate, is known as *self-lubrication*, see Bercovici (1995). We summarise its essence in Fig. 3.3. Unfortunately, self-lubrication requires the use of power-law exponents ranging between -1 and 0 (see Fig. 3.3). These values do not agree with laboratory experiments of ductile deformation performed on olivine, which find n in the range 2 to 5, see e.g. Kirby (1983).

The challenge to develop plate-like behaviour in convection models reflects the difficulty to account for brittle failure and reactivation of pre-existing faults in the uppermost cold region of the lithosphere. The high strength in the upper part of the lithosphere expresses the resistance of rocks at low temperature to break, or slide past each other when already faulted. Experimental results indicate a simple linear relationship to parameterise this behaviour, where shear stress is proportional to the effective normal pressure through a friction coefficient. Geodynamicists have introduced weak zones at the surface of mantle convection models in an attempt to account for brittle failure in the lithosphere (Davies, 1989). The logical development of this approach is the inclusion of discontinuities directly into the computational grid and the representation of faults through contact-element interfaces. This has been done, for example, in the modelling work of Zhong & Gurnis (1995) and the global neo-tectonic model of Kong & Bird (1995).



$$\begin{aligned}\sigma(\dot{\epsilon}) &= \nu(\dot{\epsilon}, T, z) \cdot \dot{\epsilon} \\ &= A \cdot (\gamma + \dot{\epsilon}^2)^{\frac{1-n}{2n}} \cdot e^{\frac{B+Cz}{T}} \cdot \dot{\epsilon}\end{aligned}$$

$$\sigma(\dot{\epsilon}) \propto \dot{\epsilon}^{\frac{1}{n}} \quad \text{for} \quad \dot{\epsilon} \to \infty$$

$$\sigma(\dot{\epsilon}) \propto \dot{\epsilon} \quad \text{for} \quad \dot{\epsilon} \to 0$$

Figure 3.3: Generalised power law rheology, where stress σ is proportional to strain rate ϵ through a viscosity ν that depends on temperature T, strain rate and depth where A, B, C, γ and n parameterise the dependence. Note that so-called *self-lubrication* rheology arises only for a narrow and unphysical band of power law exponents ranging from -1 to 0, which is not observed for geologic materials.

Today the neo-tectonic models have reached a high level of maturity allowing them to account for surface topography, regional variations of lithosphere density and thickness according to either Pratt or Airy isostatic compensation, thermal regime of the lithosphere - based on heat flow measurements and crustal radioactive decay - and for realistic plate configurations (Richardson & Coblentz, 1994; Bird, 1998). The models typically use finite-element formulations to solve the equations of mass and momentum conservation in the Stokes limit that we have seen before, and compute the instantaneous force balance and associated plate velocities. The use of finite elements makes it feasible to implement empirical, depth-dependent rheologies of the lithosphere to account for ductile as well as brittle deformation. We show the computational grid from the global lithosphere model of Kong & Bird (1995) in Fig. 3.4.

A first-order result in plate tectonic modelling is the recent explanation of the plate motion change off-shore of South America (see Fig. 3.5). For the Nazca/South America plate margin a variety of data indicate a significant decline (by some 30%) in convergence velocity over the past 10 Myrs. The ability to consider past as well as present plate motions provides important constraints for our understanding of the plate tectonic force balance, because changes in plate motion are necessarily driven by changes in one or more driving or resisting forces. By explicitly coupling MCMs, which provide estimates on the mantle buoyancy field, to neo-tectonic models Iaffaldano et al. (2006, 2007), shows that the recent topographic growth of the Andes is a key factor controlling the long-term evolution of plate motion in this region.

Further supporting evidence for the dominant effect of Andean topography on plate boundary forcing along the Nazca/South America margin comes from gravity and stress field measurements (Song & Simons, 2003; Heidbach et al., 2007; Zoback, 1992). Heidbach, Iaffaldano and Bunge (Heidbach et al., 2008) show that these independent observables can also be reproduced from the coupled models.

3.1.4 Seismic Wave Propagation

In seismology there has been a gap between observations and theory for several decades in that the quality and quantity of observations far exceeds the traditional methods of seismic modelling. Although the existing tomographic images of the mantle have greatly contributed to our understanding of the planet's dynamics, the inversions of seismic observables usually involve substantially simplified forward models, namely ray theory and finite normal mode summations. Ray theory is only applicable to the arrival times of high frequency waves, therefore significantly reducing the amount of exploitable information. Conversely normal mode approximations rely on smoothly varying structure and long period waveforms, resulting in a limitation of resolution.

The fact that today's computational power is sufficient to accurately solve the wave equation in realistic earth models (Igel & Weber, 1995; Komatitsch et al., 2000) is prompting new

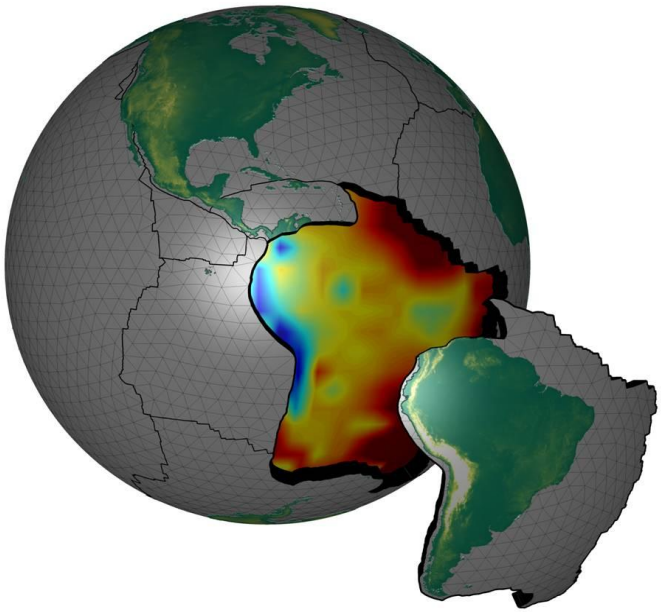


Figure 3.4: Grid of global neo-tectonic SHELLS model coupled to a global mantle circulation model (see text); colours represent temperatures (red=hot, blue=cold) at a depth of 200 km below the surface.

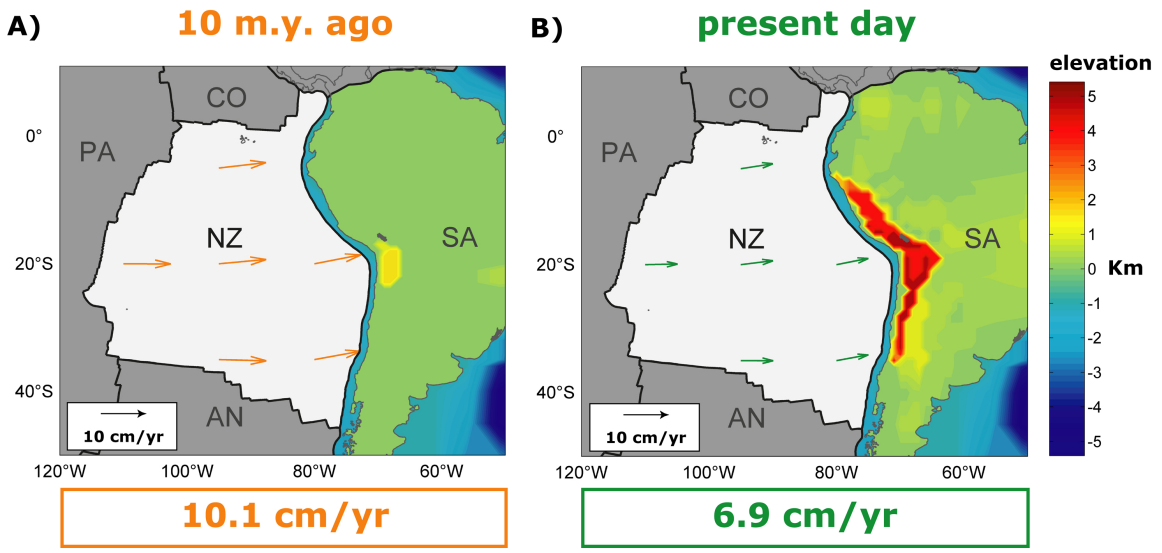


Figure 3.5: Computed velocity for Nazca plate (NZ) relative to South America plate (SA) from global plate motion simulations. Topography of South America plate from numerical simulations is shown together with colourbar. Plate boundaries are in black, coastline in gray. a) For 10 Myrs ago we compute a convergence rate of 10.1 cm/yr with topography of South America plate inferred from geological indicators. b) For present day we compute a convergence rate of 6.9 cm/yr with topography of South America plate from the ETOPO5 database. 4 km of topography lifted up over the last 10 Myrs can account for the slow down of Nazca plate.

efforts to replace the approximate ray-theory and normal mode forward models by the exact forward model of full seismic wave propagation, and to invert for seismic waveforms with shorter periods. The expectation is that the resulting increase of exploitable information will translate into an increase of resolution especially in regions poorly sampled by seismic rays.

In analogy to the efforts of using adjoint theory in geodynamics, adjoint methods are explored in seismology. The approach allows us to compute the derivative with respect to the parameters by combining the synthetic forward wavefield and an adjoint wavefield governed by a set of adjoint equations and adjoint subsidiary conditions. This concept was introduced by Tarantola (1984, 1988) into the field of seismology. Recently, the adjoint method was used in the context of finite-frequency traveltime kernels (Tromp et al., 2005) and regional seismic models (Fichtner et al., 2006a,b). It is expected that these models will yield great improvements in the imaging of earth structures.

3.2 Partielle Differentialgleichungen in der Geophysik

Partielle Differentialgleichungen (PDGLen) bilden die Basis vieler mathematischer Modelle der Physik, Chemie, Biologie und natürlich auch der Geophysik. Sie dienen der Beschreibung entsprechender Vorgänge in der Natur. Die Lösung einer PDGL ist häufig der Schwerpunkt in geophysikalischen Anwendungen und wird vorwiegend mit Hilfe numerischer Verfahren durchgeführt. Die nun folgenden Seiten sollen einen kurzen Überblick zu partiellen Differentialgleichungen und deren Lösung mittels numerischer Methoden geben. Eine ausführliche Einführung in partielle Differentialgleichungen der Naturwissenschaften ist beispielsweise in Morton & Mayers (2005) enthalten. Alle in den Abschnitten 3.3 bis 3.10 aufgeführten geophysikalischen Simulationsanwendungen setzen numerische Verfahren ein, um die in ihren Problemstellungen auftretenden PDGL zu lösen.

3.2.1 Partielle Differentialgleichungen

„Eine Differentialgleichung ist eine Gleichung zwischen gesuchten Funktionen einer oder mehrerer Veränderlicher, diesen unabhängigen Veränderlichen und den Ableitungen der gesuchten Funktionen nach den unabhängigen Veränderlichen. Hängen die gesuchten Funktionen nur von einer unabhängigen Veränderlichen ab, so spricht man von einer gewöhnlichen Differentialgleichung; hängen die gesuchten Funktionen von mehr als einer unabhängigen Veränderlichen ab, so heißt die Differentialgleichung partielle Differentialgleichung.“ (Bronstein & Semendjajew, 1991). Allgemein kann man folgende Aufgabenstellung formulieren:

Gesucht ist eine Funktion $u : \Omega \rightarrow \mathbb{R}$ mit $x := (x_1, x_2, \ldots, x_d)^T \in \Omega \subset \mathbb{R}^d$, sodass

$$F\left(x, u, \frac{\partial u}{\partial x_1}, \ldots, \frac{\partial u}{\partial x_d}, \frac{\partial^2 u}{\partial x_1^2}, \frac{\partial^2 u}{\partial x_1 \partial x_2}, \ldots, \frac{\partial^p u}{\partial x_d^p}\right) = 0 \,, \tag{3.1}$$

wobei $d > 1$ die Dimension des Problemgebiets Ω und p die Ordnung der durch die Funktion F beschriebenen Differentialgleichung ist.

Zumeist bezeichnet $d \in \{2, 3, 4\}$ die Raum- und Zeitdimension ($d = 4$ ist ein dreidimensionales Raum-Zeit-Problem). Beschreibt eine der unabhängigen Variablen x_d die Zeit, so spricht man von einem instationären Problem, andernfalls von einem stationären Problem. Nimmt man eine grobe Einordnung vor, so bezeichnen Gleichgewichtsprobleme stationäre Zustände, Ausbreitungsvorgänge instationäre Zustände während charakteristische Systemzustände durch Eigenwertprobleme beschrieben werden. Beschränkt man sich auf die Betrachtung von linearen partiellen Differentialgleichungen 2. Ordnung ($p = 2$) und auf ein zweidimensionales Gebiet

($d = 2$), so lässt sich die partielle Differentialgleichung (3.1) formulieren als:

$$A\frac{\partial^2 u}{\partial x_1^2} + 2B\frac{\partial^2 u}{\partial x_1 \partial x_2} + C\frac{\partial^2 u}{\partial x_2^2} + D\frac{\partial u}{\partial x_1} + E\frac{\partial u}{\partial x_2} + Fu = G \qquad \forall\,(x_1, x_2) \in \Omega\,, \tag{3.2}$$

mit den Koeffizientenfunktionen A, B, C, D, E, F und G. Diese lineare partielle Differentialgleichung (3.2) bezeichnet man als

$$\text{elliptisch, falls} \quad AC - B^2 > 0 \quad \forall\,(x_1, x_2) \in \Omega, \tag{3.3}$$

$$\text{parabolisch, falls} \quad AC - B^2 = 0 \quad \forall\,(x_1, x_2) \in \Omega, \tag{3.4}$$

$$\text{hyperbolisch, falls} \quad AC - B^2 < 0 \quad \forall\,(x_1, x_2) \in \Omega. \tag{3.5}$$

Diese Typeneinteilung ist notwendig für die Lösung der PDGL und die dabei korrekt zu stellenden Zusatzbedingungen. Als Beispiel für eine elliptische partielle Differentialgleichung ist die Poissongleichung zu nennen, für eine parabolische die Wärmeleitungsgleichung und für eine hyperbolische die Wellengleichung. Im Allgemeinen besitzt eine PDGL mehrere Lösungen und um eine bestimmte davon zu erhalten, sind gewisse Zusatzbedingungen nötig. Man legt daher Randbedingungen (RB) und/oder Anfangsbedingungen (AB) fest. Randbedingungen sind Vorgaben an die Funktion auf dem Rand $\partial\Omega$ des Gebietes Ω. Demgegenüber legen Anfangsbedingungen fest, welche Werte die Lösung an bestimmten Stellen zu einem Anfangszeitpunkt (typischerweise $t = 0$) haben soll. Die genannten Bedingungen sind notwendige, aber nicht hinreichende Voraussetzungen für die Eindeutigkeit von Lösungen. Die am häufigsten auftretenden Randbedingungen sind:

$$\text{Dirichletsche RB} \qquad u(x_1, x_2) := \varphi(x_1, x_2) \qquad \text{auf } \Gamma_1 \subset \partial\Omega, \tag{3.6}$$

$$\text{v. Neumannsche RB} \qquad \frac{\partial u(x_1, x_2)}{\partial n} := \gamma(x_1, x_2) \qquad \text{auf } \Gamma_2 \subset \partial\Omega, \tag{3.7}$$

$$\text{Cauchysche RB} \qquad \frac{\partial u(x_1, x_2)}{\partial n} + \alpha u(x_1, x_2) := \beta(x_1, x_2) \qquad \text{auf } \Gamma_3 \subset \partial\Omega. \tag{3.8}$$

Hierbei sind φ, γ, β vorgegebene Funktionen und $\partial u(x_1,x_2)/\partial n$ ist die Ableitung auf dem Rand in Richtung der äußeren Normalen. Dirichletsche Randbedingungen werden auch als Randbedingungen 1. Art, v. Neumannsche als Randbedingungen 2. Art und Cauchysche als Randbedingungen 3. Art bezeichnet.

Zur Vereinfachung der folgenden Gleichungen kann der Laplaceoperator Δ wie folgt definiert werden:

$$\Delta u := \frac{\partial^2 u}{\partial x_1^2} + \frac{\partial^2 u}{\partial x_2^2}\,. \tag{3.9}$$

Typischerweise treten elliptische partielle Differentialgleichungen im Zusammenhang mit stationären (zeitunabhängigen) Problemen auf. Ein Merkmal von elliptischen Gleichungen ist, dass sie oftmals einen Zustand minimaler Energie beschreiben. Wie bereits erwähnt sind die Laplace- und Poissongleichung die bekanntesten Beispiele. Durch diese Gleichungen werden etwa die (stationäre) Temperaturverteilung in einem Körper oder die elektrostatische Ladungsverteilung in einem Körper beschrieben. Die Poissongleichung mit Dirichletschen Randbedingungen lautet wie folgt:

$$-\Delta u = f \quad \text{in } \Omega \tag{3.10}$$

$$u = g \quad \text{auf } \partial\Omega\,. \tag{3.11}$$

Gleichung (3.10) definiert physikalisch eine Potentialgleichung, wobei die Lösung u das Potentialfeld und f als die „Quelle“ meist eine Massendichte, elektrische Ladungsdichte oder eine Potentialströmung beschreibt. Durch g wird der Wert von u auf dem Rand $\partial\Omega$ vorgegeben. Ist in Gleichung (3.10) $f = 0$, so spricht man von der Laplacegleichung. Da es sich um ein zeitunabhängiges Problem handelt, werden für die Berechnung der Lösung nur Randbedingungen benötigt.

Durch parabolische partielle Differentialgleichungen werden vergleichbare Phänomene wie bei elliptischen PDGL beschrieben, aber im instationären (zeitabhängigen) Fall. Das bekannteste Beispiel einer solchen Gleichung ist die Wärmeleitungsgleichung, die beispielswiese das Abkühlen und Aufheizen eines Körpers oder Diffusionsprozesse veranschaulicht. Parabolische partielle Differentialgleichung benötigen Rand- und Anfangsbedingungen, um die Lösung zu berechnen. Wenn κ als räumlich variable Wärmeleitfähigkeit oder Diffusionskoeffizient bezeichnet wird, so kann die Wärmeleitungsgleichung wie folgt dargestellt werden:

$$\frac{\partial^2 u}{\partial x^2} = \frac{1}{\kappa}\frac{\partial u}{\partial t} \qquad \text{in } \Omega\,,\ \forall\, t > 0 \tag{3.12}$$

$$u(x,0) = u_0(x) \qquad \text{in } \Omega \tag{3.13}$$

$$u(x,t) = g(x) \qquad x \in \partial\Omega\,,\ \forall\, t > 0. \tag{3.14}$$

Mit $g(x)$ sind die Randwerte bestimmt.

Die typische hyperbolische Modellgleichung ist die Wellengleichung, welche allgemein Wellen und deren Ausbreitung in Medien beschreibt. Verglichen zu elliptischen und parabolischen Gleichungen werden die Lösungen der hyperbolischen PDGL nur sehr schwach oder gar nicht gedämpft, wodurch sich die Wellen über weite Distanzen ausbreiten können. Wie bei parabolischen partiellen Differentialgleichungen sind Anfangs- und Randbedingungen notwendig, um die Lösung zu berechnen. Man benötigt zwei Anfangswerte, den Funktionswert und die zeitli-

che Ableitung davon. Eine hyperbolische Wellengleichung lautet wie folgt:

$$\frac{\partial^2 u}{\partial x^2} = \frac{1}{c^2}\frac{\partial^2 u}{\partial t^2} \qquad \text{in } \Omega,\ \forall\, t > 0 \tag{3.15}$$

$$u(x,0) = u_0(x) \qquad \text{in } \Omega \tag{3.16}$$

$$\frac{\partial u}{\partial t}(x,0) = u_1(x) \qquad \text{in } \Omega \tag{3.17}$$

$$u(x,t) = g(x) \qquad x \in \partial\Omega,\ \forall\, t > 0. \tag{3.18}$$

In der Geophysik sind kompliziertere Probleme als die bisher aufgeführten Modellgleichungen anzutreffen. Diese sind vielfach von mehreren Parametern abhängig, ortsvariabel und weisen eine komplexe Modellgeometrie auf, somit kann die Lösung nicht analytisch berechnet werden. Daher kommen numerische Verfahren für das Lösen der partiellen Differentialgleichungen zum Einsatz. Beispiele für derartige Verfahren sind die im Folgenden kurz beschriebenen Methoden der finiten Differenzen und finiten Elemente.

3.2.2 Finite-Differenzen-Methode

Da die Finite-Differenzen-Methode (FDM) relativ einfach zu implementieren ist, wurde sie bereits sehr früh in den Simulationsanwendungen der Geophysik eingesetzt. Bei der FDM wird zuerst das Gebiet Ω in eine finite (endliche) Anzahl von Gitterpunkten zerlegt. Abbildung 3.6 stellt eine solche Unterteilung für den eindimensionalen Fall dar.

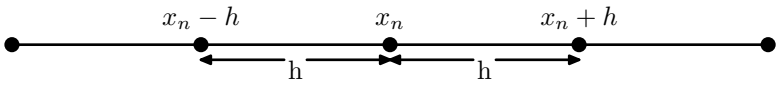


Abbildung 3.6: Eindimensionales Finite-Differenzen-Gitter – Gitterlinie mit den Gitterpunkten und dem Diskretisierungsabstand h zwischen den einzelnen Gitterpunkten

Für zwei- oder dreidimensionale Gebiete bilden die Kreuzungspunkte der senkrecht aufeinander stehenden Gitterlinien die Diskretisierungspunkte. Im nächsten Schritt werden die partiellen Ableitungen der partiellen Differentialgleichungen durch Differenzenausdrücke approximiert.

Es soll nun beispielhaft die Funktion $f : [0,1] \to \mathbb{R}$ betrachtet werden. Die erste Ableitung der Funktion am Punkt $x_n \in (0,1)$ ist definiert als:

$$\frac{d}{dx} f(x_n) = \lim_{h \to \infty} \frac{f(x_n + h) - f(x_n)}{h}\ . \tag{3.19}$$

Für ein Gitter auf $[0,1]$ (vergleichbar zu dem in Abbildung 3.6) mit einem kleinen Gitter- oder Diskretisierungsabstand h kann die Ableitung näherungsweise ausgedrückt werden durch:

$$\frac{d}{dx} f(x_n) \approx \frac{f(x_n + h) - f(x_n)}{h}\ , \tag{3.20}$$

wobei der Term auf der rechten Seite als finite Differenz oder Differenzenquotient bezeichnet wird. Der Approximationsfehler kann durch die Verwendung eines zentralen Differenzenquotienten der Form:

$$\frac{d}{dx}f(x_n) \approx \frac{f(x_n+h)-f(x_n-h)}{2h} \tag{3.21}$$

verringert werden. Diese Approximation wird für jeden Gitterpunkt durchgeführt. Damit kann die partielle Differentialgleichung in ein System von Differenzengleichungen überführt werden. Die Lösung dieses entstandenen Gleichungssystems kann mit unterschiedlichen Algorithmen (direkte oder iterative Löser) erfolgen. Bei komplexen Geometrien, so wie sie in der Natur anzutreffen sind, ist die Bestimmung der Differenzenquotienten schwieriger. Dies ist ein Nachteil der Methode. Ausführliche Beschreibungen der Finiten-Differenzen-Methode sind beispielsweise in Morton & Mayers (2005) und unter `http://en.wikipedia.org/wiki/Finite_difference` zu finden. Eine detaillierte Darstellung der FDM für die seismische Wellenausbreitung ist in Moczo et al. (2004) enthalten.

3.2.3 Finite-Elemente-Methode

Ein weiteres Verfahren zur Lösung partieller Differentialgleichungen ist die Finite-Elemente-Methode (FEM). Eine kurze Einführung aus Sicht der Ingenieurwissenschaften enthält Jung & Langer (2001). In Braess (1997) sind ebenfalls ausführliche Darstellungen zu finden. Im ersten Schritt der FEM wird das endliche Gebiet Ω in Teilgebiete zerlegt, den sogenannten finiten Elementen $T^{(i)}$. Typische geometrische Objekte, die als finite Elemente genutzt werden, sind Dreiecke und Vierecke für zwei Raumrichtungen und Tetraeder, Hexaeder und Prismen für dreidimensionale Gebiete. Als Beispiel für eine solche Zerlegung ist in Abbildung 3.10 auf Seite 113 ein Gitter bestehend aus Hexaedern und im Vergleich auf Seite 116 in Abbildung 3.12 ein Tetraedergitter dargestellt. Eine solche Zerlegung des Gebietes Ω

$$\mathcal{T}_h = \{T^{(i)} : i = 1, 2, \ldots, R_h\} \tag{3.22}$$

wird auch Triangulierung oder Vernetzung genannt. Unter der charakteristischen Größe h wird der maximale Elementdurchmesser aller Elemente verstanden:

$$h = \max_{i=1,2,\ldots,R_h} h^{(i)}\,, \tag{3.23}$$

wobei der Elementdurchmesser $h^{(i)}$ für ein Element $T^{(i)}$ als maximaler Abstand zweier beliebiger Punkte in $\overline{T}^{(i)}$ festgelegt ist. Die Triangulierung $\mathcal{T}_h$ soll folgende Eigenschaften besitzen:

- $\overline{\Omega} = \bigcup_{i=1}^{R_h} \overline{T}^{(i)}$ beziehungsweise $\overline{\Omega}_h = \bigcup_{r=1}^{R_h} \overline{T}^{(r)} \longrightarrow \overline{\Omega}$ für $h \to 0$, d. h. die Vereinigung aller Elemente $T^{(i)}$ soll das Gebiet Ω exakt überdecken oder bei kleiner werdendem h immer besser approximieren,

- für alle $i, i' \in \{1, 2, \ldots, R_h\}$ mit $i \neq i'$ ist

$$\overline{T}^{(i)} \cap \overline{T}^{(i')} = \begin{cases} \emptyset & \text{oder} \\ \text{eine gemeinsame Ecke} & \text{oder} \\ \text{eine gemeinsame Kante} & \text{oder} \\ \text{eine gemeinsame Fläche (in 3D).} \end{cases} \tag{3.24}$$

Das mit diesen finiten Elementen erstellte Gitter ist in der Lage, auch schwierige Modellgeometrien abzubilden und somit eine realitätsnahe Lösung der partiellen Differentialgleichungen zu ermöglichen. Zur Beschreibung einer Vernetzung werden alle Knoten P_j $(j = 1, 2, \ldots, P_h)$ und Elemente $T^{(i)}$ $(i = 1, 2, \ldots, R_h)$ global nummeriert. Zusätzlich werden in jedem Element $T^{(i)}$ die Knoten $P_\alpha^{(i)}$ lokal von $\alpha = 1, \ldots, \hat{N}$ durchnummeriert. $\hat{N}$ bezeichnet die Anzahl der Knoten pro Element.

Für die Gleichstromgeoelektrik als Teildisziplin der Geophysik soll nun die schwache Formulierung oder Variationsformulierung aufgestellt werden, um diesen Schritt der FEM zu veranschaulichen. Die elliptische partielle Differentialgleichung des elektrischen Potentials $\varphi(\vec{r})$ für eine beliebige Verteilung der Leitfähigkeit $\sigma(\vec{r})$ im Halbraum lautet:

$$\operatorname{div}\left(\sigma(\vec{r}) \operatorname{grad} \varphi(\vec{r})\right) = -I\delta(\vec{r} - \vec{r}_s) \,, \tag{3.25}$$

wobei I den elektrischen Strom, δ die Dirac-Funktion und $\vec{r}_s$ die Quellposition bezeichnet. Nach Multiplikation mit Testfunktionen ψ, Integration über Ω, Anwendung der GREENschen Formel und unter Einbeziehung der Randbedingungen:

$$\varphi = g_1 \text{ auf } \Gamma_1 \,, \quad \frac{\partial \varphi}{\partial \vec{n}} = g_2 \text{ auf } \Gamma_2 \,, \quad \frac{\partial \varphi}{\partial \vec{n}} + \alpha\varphi = \alpha\varphi_0 \text{ auf } \Gamma_3$$

erhält man die allgemeine Variationsformulierung der gleichstromgeoelektrischen Randwertaufgabe:

Gesucht ist $\varphi \in V_{g_1}$, sodass

$$a(\varphi, \psi) = l(\psi) \quad \forall \psi \in V_0 \tag{3.26}$$

gilt mit

$$\begin{aligned} a(\varphi, \psi) &= \int_\Omega (\operatorname{grad} \psi)^T \, \sigma \operatorname{grad} \varphi \, dV + \int_{\Gamma_3} \sigma \, \alpha\varphi \, \psi \, ds \,, \\ l(\psi) &= \int_\Omega I\delta(\vec{r} - \vec{r}_s) \, \psi \, dV + \int_{\Gamma_2} \sigma \, g_2 \, \psi \, ds + \int_{\Gamma_3} \sigma \, \alpha\varphi_0 \, \psi \, ds \,, \\ V_{g_1} &= \{\varphi \in H^1(\Omega) : \varphi = g_1 \quad \text{auf } \Gamma_1\} \,, \\ V_0 &= \{\psi \in H^1(\Omega) : \psi = 0 \quad \text{auf } \Gamma_1\} \,, \\ \sigma &\in L^2(\Omega), \sigma \in L^2(\Gamma_3) \,. \end{aligned}$$

Wie aus der Aufgabe (3.26) hervorgeht, nehmen die Randbedingungen in unterschiedlicher Weise Einfluss auf die Variationsformulierung. Auf die Definition der zulässigen Funktionen V_{g_1} und die Menge der Testfunktionen V_0 hat die Randbedingung 1. Art Einfluss. Dahingegen führen die Randbedingungen 2. und 3. Art zu weiteren Termen in der symmetrischen, positiv definiten Bilinearform $a(.,.)$ und der Linearform $l(.)$.

Die schwache Formulierung sucht die Lösung in einem unendlichdimensionalen Funktionenraum (im Beispiel V_0). Da dieser aber zu groß und komplex ist, wird beim GALERKIN-Verfahren der Raum V durch einen endlichdimensionalen Raum

$$V_h \subset V\,, \quad \dim V_h = n < \infty \tag{3.27}$$

approximiert. Damit lautet die diskrete schwache Formulierung:

Gesucht ist $\varphi_h \in V_h$, sodass

$$a(\varphi_h, \psi_h) = l(\psi_h) \qquad \forall \psi_h \in V_h \tag{3.28}$$

gilt.

Da die Dimension des Raumes V_h kleiner als unendlich ist, gibt es einen endlichen Satz linear unabhängiger Funktionen $\phi_i \in V_h \quad (i = 1, \ldots, n)$, die eine Basis von V_h bilden. Diese Funktionen werden auch als Ansatzfunktionen bezeichnet. Damit gibt es für jedes $v \in V_h$ eine Darstellung der Form

$$v = \sum_{i=1}^{n} v_i \phi_i\,. \tag{3.29}$$

Durch Einsetzen der Darstellung (3.29) in (3.28) und Verwendung der ϕ_i als Testfunktion ψ_h ergibt sich ein lineares Gleichungssystem für die Koeffizienten $\varphi_{h,i}$ in der Darstellung von φ_h über der Basis $\{\phi_i\}$.

Die grundlegende Idee ist es, als Ansatz- oder Testfunktionen ϕ_i solche zu verwenden, die einen kleinen Träger besitzen. Man wählt üblicherweise global stetige, stückweise polynomiale Funktionen, die nur über sehr wenigen Elementen von Null verschieden sind. Die Wahl von Ansatzfunktionen mit lokalem Träger führt zu einer schwach besetzten Steifigkeitsmatrix A_h. Die Ansatz- und Testfunktionen ϕ_i werden lokal über den finiten Elementen $T^{(i)}$, die den Knoten P_j enthalten, durch Formfunktionen $\phi_\alpha^{(i)}$ definiert. Diese werden durch Abbildung von Formfunktionen erhalten, die auf Referenzelementen festgelegt sind. Damit kann für jedes finite Element $T^{(i)}$ die Elementsteifigkeitsmatrix $A_h^{(i)}$ zusammengestellt werden.

Das lineare Gleichungssystem wird elementweise aufgebaut. Dabei wird für jedes Element $T^{(i)} \in \mathcal{T}_h$ sein Beitrag zur Steifigkeitsmatrix A_h und zum Lastvektor $\vec{f}_h$ berechnet. Für ein Element $T^{(i)}$ mit $\hat{N}$ Knoten erhält man die sogenannte Elementsteifigkeitsmatrix $A_h^{(i)}$ der Größe

$\hat{N}\times\hat{N}$. Diese wird durch Aufaddieren an den entsprechenden Knoten in die globale Steifigkeitsmatrix A_h eingebaut. Auf gleiche Weise wird der Lastvektor $\vec{f_h}$ aus den einzelnen Elementlastvektoren $\vec{f}_h^{(i)}$ assembliert.

Die während des Aufbaus der Elementsteifigkeitsmatrix und des Elementlastvektors zu bestimmenden Integrale können nicht immer analytisch berechnet werden. Deshalb nutzt man numerische Integrationsformeln zur näherungsweisen Berechnung. Dabei werden die Integrale nicht über dem Element $T^{(i)}$, sondern über dem Referenzelement $\hat{T}$ berechnet. Die verwendeten Quadraturformeln haben folgende Gestalt

$$\int_{\hat{T}} w(\xi)d\xi \approx \sum_{r=1}^{p} \alpha_r w(\xi_r)\,, \tag{3.30}$$

wobei $w(\xi_r)$ der Integrand, α_r die Quadraturgewichte und ξ_r die Quadraturstützstellen sind. Bei Newton-Cotes-Quadraturformeln werden die Stützstellen vorgegeben und die Gewichte so gewählt, dass ein möglichst hoher Exaktheitsgrad erreicht wird. Dem gegenüber wird bei Gauß-Quadraturformeln neben den Gewichten auch die Lage der Stützstellen verändert, um eine Maximierung des Exaktheitsgrads zu erreichen. Dies führt oft zu weniger Stützstellen bei gleichem Exaktheitsgrad. Wenn auf den Elementen als Basisfunktionen Polynome der Ordnung p verwendet werden, so müssen zur genauen Auswertung die Quadraturformeln Polynome vom Grad $2p-2$ exakt integrieren.

Das Resultat der Finite-Elemente-Diskretisierung ist ein lineares Gleichungssystem $A_h\vec{\varphi}_h = \vec{f_h}$ ($\vec{\varphi}_h = (\varphi_{h,1}\dots\varphi_{h,n})^T$), welches im Allgemeinen sehr groß ist. Dabei wächst die Dimension mit kleiner werdendem Diskretisierungsparameter h an. Durch die Wahl von Ansatz- und Testfunktionen mit lokalem Träger ist die Systemmatrix schwach besetzt, d.h. in jeder Zeile der Matrix A_h sind nur wenige Einträge von Null verschieden. Bei einer geeigneten Knotennummerierung hat die Systemmatrix eine Bandstruktur. Die aufgezählten Eigenschaften können geschickt für die effiziente Lösung des linearen Gleichungssystems mit direkten oder iterativen Lösern ausgenutzt werden. Zu den direkten Lösungsverfahren zählt der Gauß-Algorithmus und zur Klasse der iterativen Gleichungslöser gehören unter anderem die Mehrgitter- und vorkonditionierten Krylov-Unterraumverfahren (Barrett et al., 1994; Saad, 2003; Duff et al., 1989). Die gleichen Methoden können auch für die Gleichungssysteme der Finiten-Differenzen-Methode eingesetzt werden, welche ähnliche Eigenschaften besitzen.

Eine ausführlichere Darstellung der Finite-Elemente-Methode am Beispiel der Gleichstromgeoelektrik ist in Oeser (2004) enthalten. Weiterführende Literatur zur FEM kann unter `http://en.wikipedia.org/wiki/Finite_element_method` aufgerufen werden.

Die spektrale Elemente Methode und das diskontinuierliche Galerkin Verfahren als spezielle Varianten der FEM finden ebenfalls häufig Anwendung in der Geophysik. Für eine Erklärung

dieser Verfahren sei auf die entsprechende Fachliteratur (Komatitsch & Vilotte, 1998; Komatitsch & Tromp, 1999a,b; Komatitsch et al., 2000; Käser & Dumbser, 2006; Dumbser & Käser, 2006; Käser et al., 2007) verwiesen.

3.3 Mantelkonvektion – TERRA

Die Strömungsprozesse im Inneren der Erde (gemeint ist der innere und äußere Erdkern sowie der silikatische Erdmantel) dienen der sekulären Kühlung des Erdkörpers. Im flüssigen, äußeren Erdkern erzeugen diese Strömungen über das Prinzip des Dynamo das irdische Magnetfeld und im festen Erdmantel (plastische Verformung in geologischen Zeiträumen) wirken sie als Antriebsmechanismen für die Plattentektonik. Die Simulationsanwendung TERRA dient der Modellierung dieser Strömungsprozesse im Erdmantel und mit Hilfe der Inversion können die dynamischen Fließprozesse im Erdmantel quantitativ in die Vergangenheit rekonstruiert werden. Damit besteht die Möglichkeit, die großen tektonischen Veränderungen seit der Kreidezeit (vor rund 100 Millionen Jahren) besser zu verstehen. Das parallelisierte Programm TERRA verwendet für die Modellierung und Inversion dieser Prozesse ein strukturiertes Gitter aufgebaut aus Ikosahedern. Als numerisches Verfahren wird die Methode der finiten Elemente verwendet und zur Berechnung der Lösung wird ein Mehrgitterverfahren eingesetzt. Mehr zu den genutzten Verfahren und den geowissenschaftlichen Ergebnissen sind in Bunge et al. (1996, 1997), Bunge & Grand (2000), Bunge & Davies (2001), Bunge et al. (2003) und Kennett & Bunge (2008) enthalten. Kontinuierliche Parameterstudien dienen bei der Mantelkonvektionssimulation zum Finden einer optimalen Lösung. Dazu sind lang anhaltende Rechnungen auf dem Rechencluster TETHYS notwendig. Die dabei in der Summe erforderlichen 120 GB Arbeitsspeicher verteilen sich auf 64 Rechenknoten mit jeweils einem beziehungsweise zwei genutzten Prozessoren. Diese fest vorgegebene CPU-Anzahl begründet sich durch die bei der Modellrechnung verwendete Gebietszerlegung. Diese ist in Abbildung 3.7 dargestellt. Die Zahl der notwendigen Prozessoren P berechnet sich durch $P = 2^n$ mit $n \in \mathbb{N}$. Die vier Einzeldarstellungen in Abbildung 3.8 geben die Temperaturvariationen im Erdmantel wieder. Die vier Querschnitte mit den Blick-

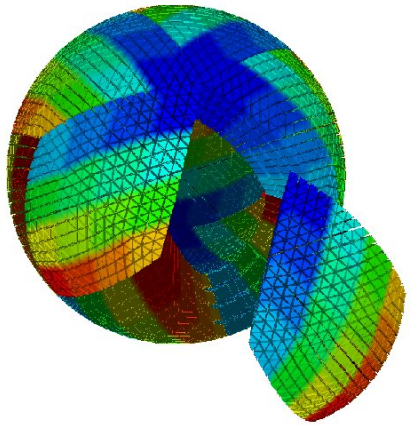


Abbildung 3.7: Die Farbkodierung kennzeichnet die Gebietszerlegung für das strukturierte Gitter. Bereiche gleicher Farbgebung werden einem Prozessor zugeordnet. Die dargestellte Konfiguration entspricht einer Parallelisierung für 16 CPUs.

winkeln aus 35 (oben links), 125 (unten rechts), 215 (unten links) und 305 (oben rechts) Grad geographische Länge wurden mit einem hochauflösenden Mantelzirkulationsmodell berechnet. Die abgebildeten Kontinente enthalten farblich kodiert ihre Topographie und sind als geographische Referenz gedacht. Ausgehend von einer angenommenen mittleren Temperatur wird die Variation der Temperatur im Erdmantel als Isofläche bei -600 und $+400$ Kelvin gezeigt. Die Isofläche bei $+400$ Kelvin wurde in den äußeren 500 km des Erdkörpers herausgeschnitten, um einen ungehinderten Blick auf den Erdmantel unterhalb der mittelozeanischen Rücken zu ermöglichen. Die Abbildungen wurden von Prof. Hans-Peter Bunge zur Verfügung gestellt und basieren auf einer Modellrechnung mit einem Gitter der Größe 100 Millionen Punkten. Dies entspricht einem Gitterpunktabstand von 20 km im Mantel der Erde.

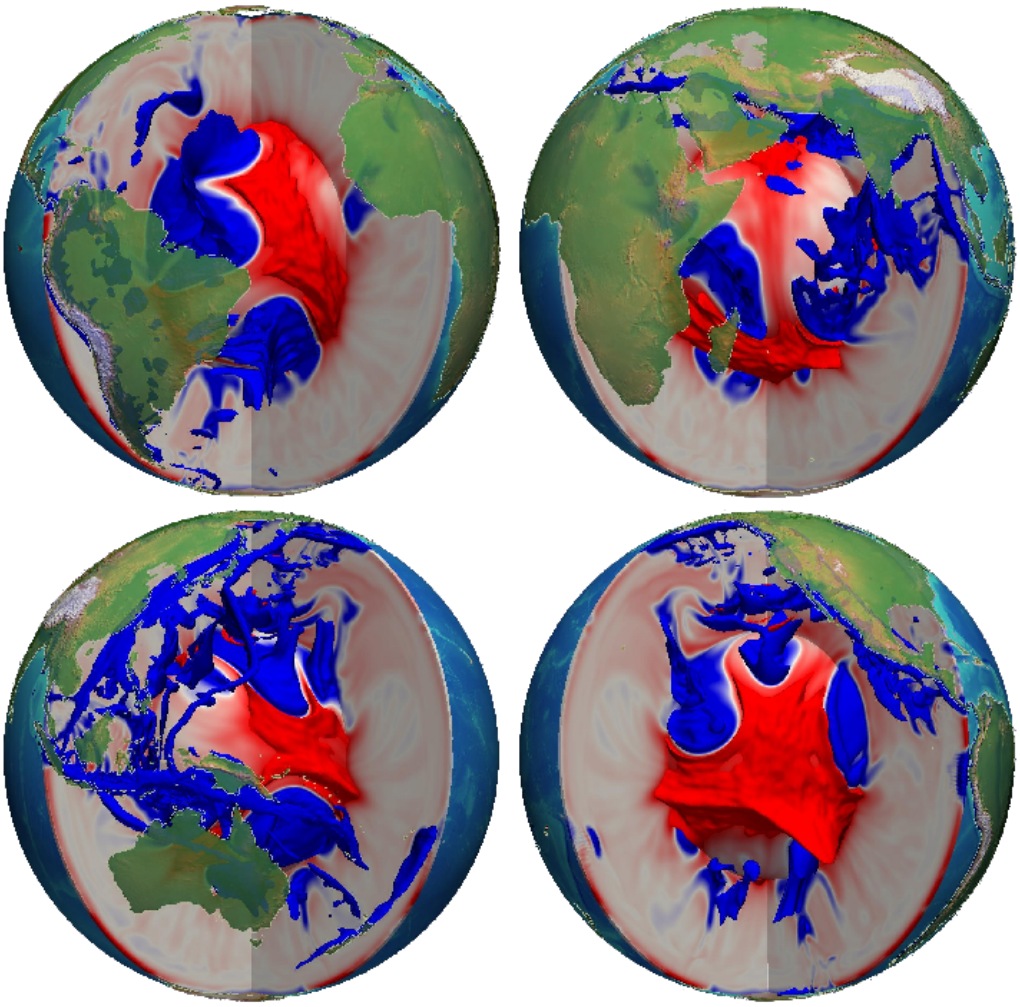


Abbildung 3.8: Temperaturverteilung im Erdmantel dargestellt aus verschiedenen Blickwinkeln. Die Isoflächen der Temperaturvariationen um einen Mittelwert wurden bei -600 und $+400$ Kelvin gewählt. Die Farbskala erreicht bei -400 und $+400$ Kelvin ihre volle Sättigung.

3.4 Erdbebenszenarien – GeoELSE

Die Simulationsanwendung `GeoELSE` wird am CRS4[1] („Center for Advanced, Research and Studies in Sardinia“) und an der „Politecnico di Milano“ DIS[2] („Department of Structural Engineering“) entwickelt. Das Einsatzgebiet des Programms ist die Modellierung der Wellenausbreitung in zwei- und dreidimensionalen Gebieten. Das eingesetzte numerische Verfahren basiert auf der Methode der spektralen Elemente mit einem unstrukturierten Hexaedergitter. Mehr zur Numerik und `GeoELSE` selbst kann Faccioli et al. (1997), Komatitsch & Vilotte (1998), Stupazzini (2004), Zambelli (2006), Scandella (2004) und Igel et al. (2008) entnommen werden. Die Parallelisierung der Simulationsanwendung basiert auf einer Gebietszerlegung mit Hilfe der Software METIS[3], woraus sich aber keine spezielle Vorgabe für die Prozessoranzahl ergibt. Stattdessen kann das Programm flexibel auf die vorhandene Hardware abgestimmt werden. Bei großen Simulationsrechnungen auf TETHYS werden bis zu 128 CPUs und 64 *GB* Arbeitsspeicher genutzt. Demgegenüber sind für die Standardberechnungen 32 Prozessoren und 10 *GB* RAM ausreichend. Diese werden sehr häufig auf dem Rechencluster durchgeführt.

Abbildung 3.9 stellt die mit `GeoELSE` berechnete Wellenausbreitung im Tal von Grenoble dar. Diese Simulation ist Teil eines Genauigkeitsbenchmarks für Modellierungsprogramme zur Ausbreitung seismischer Wellen, mit dem Ziel anhand von realen Messdaten die verschiedenen Simulationsanwendungen vergleichen zu können. Eine genauere Beschreibung ist in Bard et al. (2006a), Bard et al. (2006b) und Stupazzini et al. (2008) zu finden oder unter `http://esg2006.obs.ujf-grenoble.fr`.

Abbildung 3.9: Grenoble Benchmark: Ausbreitung einer elastischen Welle im Tal von Grenoble – dargestellt sind die Wellenfronten zu drei Zeitpunkten (Quellmodell: S1, Hypozentrum: H1, linear viskoelastischen Bodenverhalten)

[1] `http://www.crs4.it`
[2] `http://www.stru.polimi.it/EN`
[3] `http://glaros.dtc.umn.edu/gkhome/views/metis`

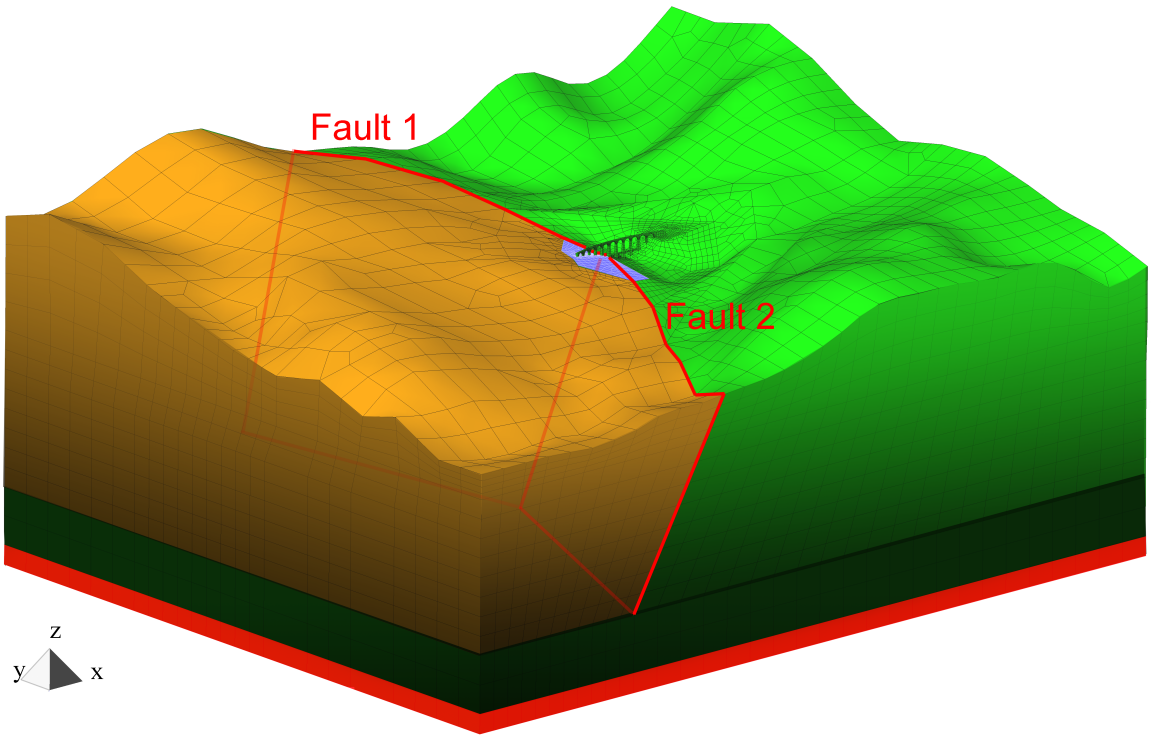


Abbildung 3.10: Hexaedergitter für die „Acquasanta“ Eisenbahnbrücke in Norditalien und das umgebende Bodenmodell

In Abbildung 3.10 wird das verwendete Hexaedergitter für eine weitere Modellrechnung mit `GeoELSE` gezeigt. Die darin dargestellte „Acquasanta“ Eisenbahnbrücke in Norditalien stellt ein typisches Bauwerk dar, welches durch Erdbeben in seiner Struktur gefährdet ist. Weiterhin sind zwei Verwerfungen („Fault 1“ und „Fault 2“) genau in der Zone zu sehen, in der auch die zwei unterschiedlichen Gesteine (kalkhaltiger Schiefer – braune Farbgebung und Serpentin – grüne Farbgebung) aufeinander treffen. Mit dem blauen Farbton wird der Bereich gekennzeichnet, in der Erosionsablagerungen vorhanden sind. In mehreren Simulationen mit dem Modellierungsprogramm wurde der Einfluss der seismischen Welle auf die Bewegung der Brücke untersucht. Abbildung 3.11 zeigt exemplarisch zu zwei Zeitpunkten ($T = 2.0\ s$ und $T = 4.0\ s$) den Betrag des Verschiebungsvektors („Disp“) und die Verformung der Brücke. Bemerkenswert in diesem Zusammenhang ist die Veränderung der Brückenbewegung zwischen beiden Zeiten, bei $T = 2.0\ s$ in x-Richtung und bei $T = 4.0\ s$ verstärkt in y-Richtung. Mehr zu den durchgeführten Untersuchungen ist in Igel et al. (2008) zu finden. Alle Abbildung zu `GeoELSE` wurden freundlicherweise von Marco Stupazzini zur Verfügung gestellt.

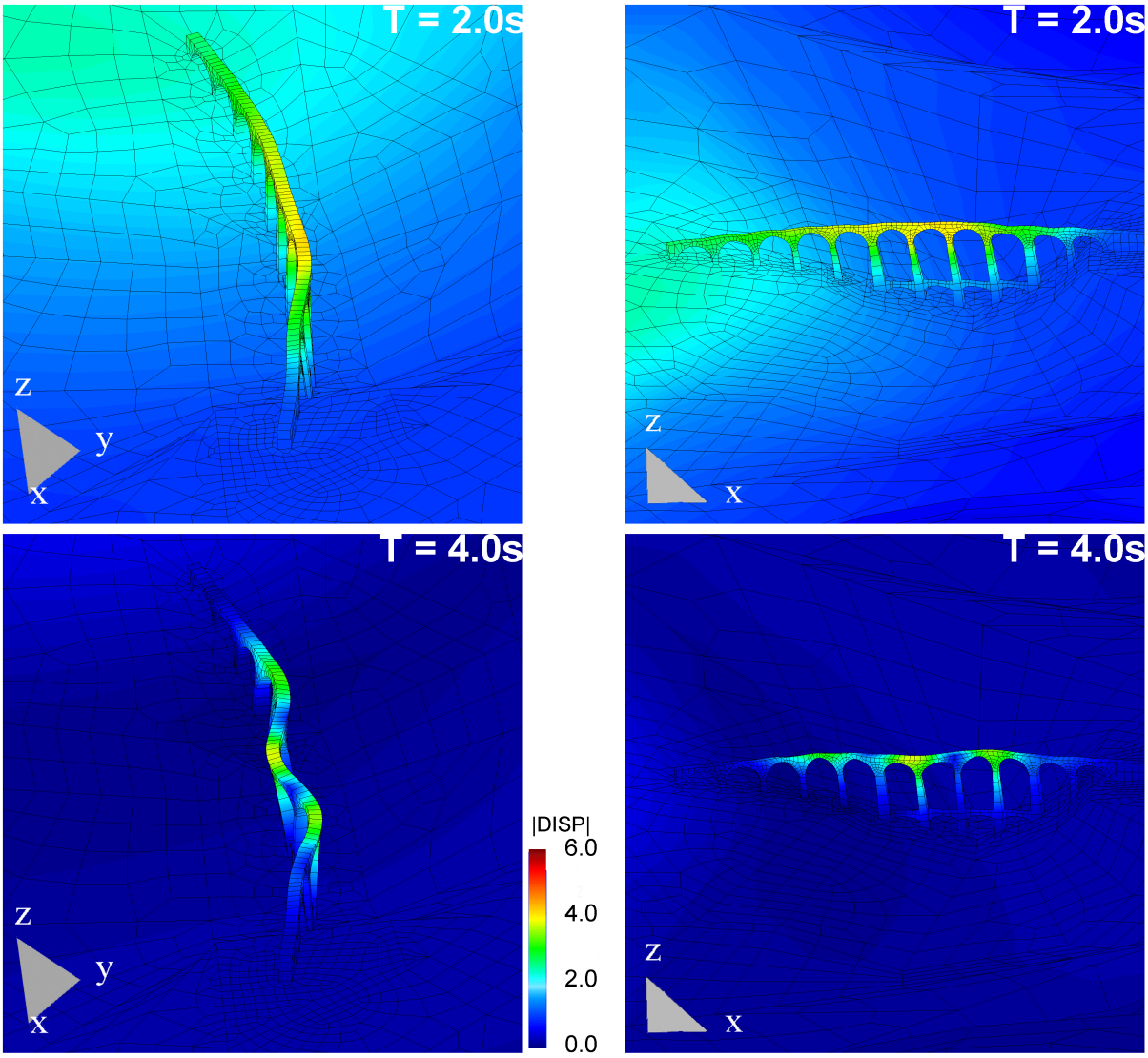


Abbildung 3.11: Betrag des Verschiebungsvektors („Disp“) und Verformung der „Acquasanta“ Eisenbahnbrücke in Norditalien

3.5 Wellenausbreitung – SeisSol

Das Verständnis und die Modellierung seismischer Phänomene ist heute von unschätzbarem Wert in der Erdbeben- und Explorationsseismologie. Die Simulationsanwendung `SeisSol` ist ein Programm, um die Entstehung und Ausbreitung von seismischen Wellen in Gebieten mit komplexer Geometrie und Parametern zu modellieren. Die Entwicklung von `SeisSol` wird gegenwärtig durch Martin Käser innerhalb seines Projekts im Emmy Noether Programm der DFG am Lehrstuhl für Geophysik der Ludwig-Maximilians-Universität durchgeführt. Die eingesetzten numerischen Verfahren beruhen auf der diskontinuierlichen Galerkin (DG) Finite-Elemente-Methode und der Zeitintegration unter Nutzung der „Arbitrary high order DERivatives“ (ADER) für die Ansatzfunktionen. Allgemein wird vom ADER-DG Verfahren gesprochen. Eine detaillierte Beschreibung der eingesetzten numerischen Verfahren und deren Anwendung für die Simulation der Wellenausbreitung ist in Käser & Dumbser (2006), Dumbser & Käser (2006), Käser et al. (2007), de la Puente et al. (2007) und Dumbser et al. (2007) zu finden. Die Modellierung komplexer Geometrien erfordert die Verwendung eines unstrukturierten Tetraedergitters. Die zu Beginn genutzte Gebietszerlegung mit Hilfe der Software METIS[4] wurde durch „Space Filling Curves“ (SFC) ersetzt (Felder et al., 2006). Dies ermöglicht eine sehr flexible und überaus effiziente Verwendung des Rechenclusters TETHYS und der HPC-Systeme am LRZ. Generell richtet sich der Bedarf an Prozessoren und Arbeitsspeicher nach den vorhandenen Ressourcen, je mehr vorhanden ist, desto aufwendigere Simulationsmodelle können berechnet werden.

In Abbildung 3.12 ist das unstrukturierte Tetraedergitter für den Vulkan Merapi dargestellt. Es wurde von Martin Käser zur Verfügung gestellt und zeigt neben der Raumdiskretisierung auch die Gebietszerlegung für die Parallelisierung.

[4] `http://glaros.dtc.umn.edu/gkhome/views/metis`

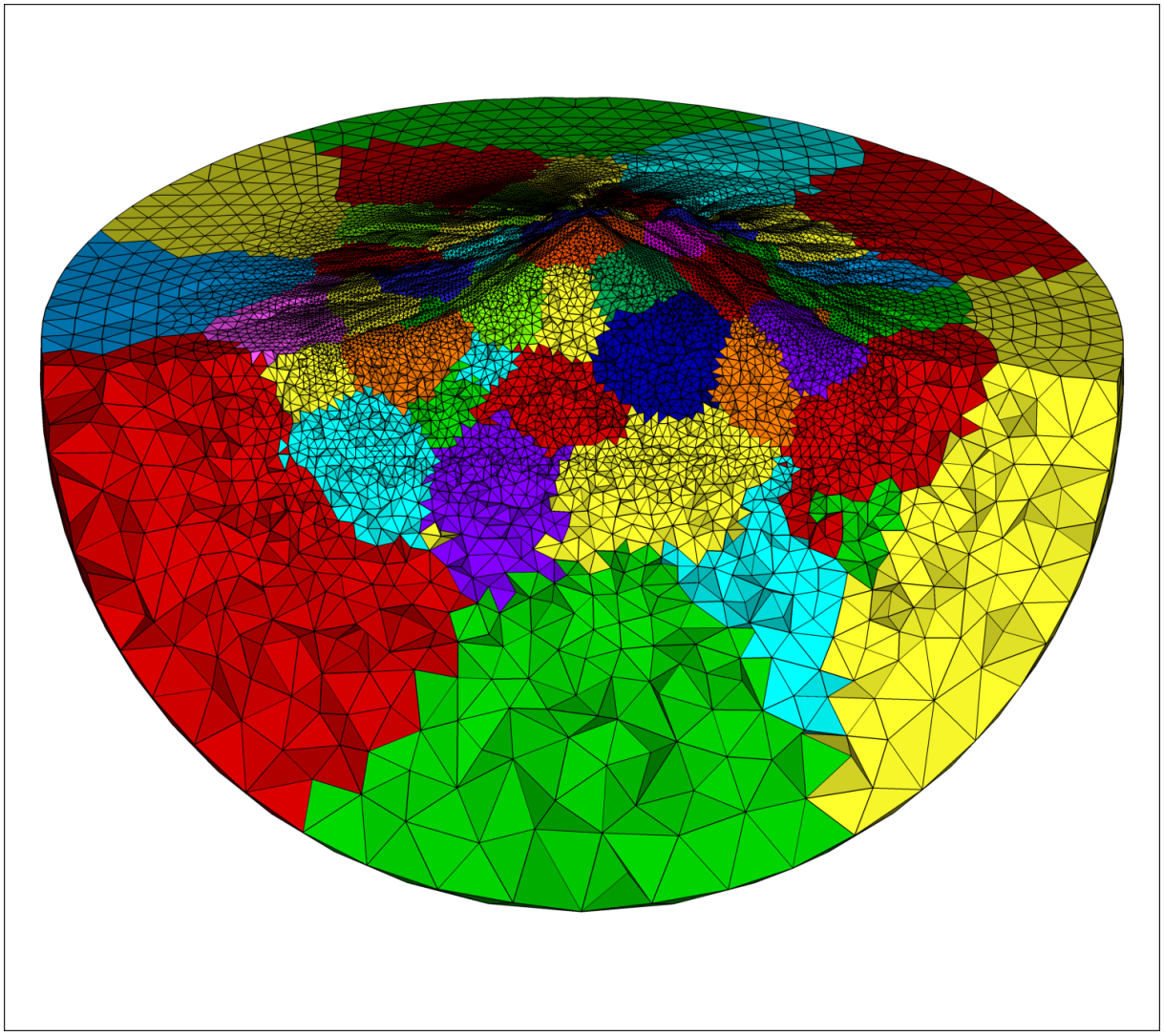


Abbildung 3.12: Schnitt durch das Tetraedergitter für den Vulkan Merapi mit der Gebietszerlegung durch METIS – jede Farbe entspricht einem Teilgebiet, welches auch einem Prozessor zugeordnet ist

3.6 Visualisierung – ParaView

Die Visualisierungssoftware `paraview`[5] verwendet das „Visualization Toolkit“[6] für die Verarbeitung und Darstellung der Visualisierungsdatensätze sowie die Qt-Bibliotheken[7] für die graphische Benutzerschnittstelle des Programms. Damit basiert die „Open Source“ Software auf weltweit eingesetzten Softwareprodukten und ermöglicht eine große Plattformunabhängigkeit (es existieren für nahezu alle Betriebssysteme vorgefertigte binäre Softwarepakete). Im Jahr 2000 startete das Softwareprojekt und stellt mittlerweile eine flexible Infrastruktur zur Visualisierung bereit, mit der kleine wie auch große Datensätze parallel oder nicht parallel verarbeitet werden können.

Im Abschnitt 2.5 wurde bereits der Einsatz von `paraview` im 3D-Visualisierungslabor des Lehrstuhls für Geophysik angesprochen. Die beiden Arbeiten von Moder (2006) und Moder et al. (2007) legten den Grundstock für den heute weit verbreiteten Einsatz der Anwendung in der Geophysik im Allgemeinen und in Kombination mit dem HPC-Rechner TETHYS im Speziellen. Dabei wird nicht primär die Rechenleistung des Clusters benötigt, sondern der Arbeitsspeicher eines jeden Rechenknotens, da die Informationen zum Koordinatengitter und den damit verbundenen Datensätzen gespeichert werden müssen. Die Anzahl an Knoten im Rechencluster ermöglicht die effiziente Bearbeitung von großen Datenmengen bei der Darstellung. Für die parallele Datenverarbeitung bietet `paraview` die eingebaute Client-Server-Architektur. Die drei Bestandteile des Client-Server-Modells sind der Datenserver (Speicherung der Daten und Ausführung von Filteroperationen), der Renderserver (Berechnung der Bilder aus den Oberflächen) und der Client (Darstellung der Bilder und Bereitstellung des Benutzerschnittstelle). Für die Datenserver können sehr gut die Rechenknoten verwendet werden, da sie über genügend Arbeitsspeicher und Rechenleistung verfügen. Als Renderserver werden wenige Rechner benötigt, die aber mit einer guten Graphikkarte ausgestattet sein sollten, da ansonsten die Bildberechnung per Software etwas mehr Zeit benötigt. Der eigentliche Client läuft auf dem Rechner, welcher mit der 3D-Darstellungshardware ausgestattet ist. Startet man `paraview` in einer Einzelprozessorversion, so laufen alle drei Komponenten auf dem einen Rechner. Bei der parallelen Nutzung werden auf allen beteiligten Rechnern die notwendigen Server über MPI gestartet und kommunizieren von diesem Zeitpunkt an über MPI miteinander. Detaillierter werden die verschiedenen Komponenten und ihr Zusammenspiel in Moder (2006) beschrieben.

Die Abbildung 3.13 zeigt eine mit der Anwendung `paraview` generierte Darstellung des Bergs Hochstaufen und der im Jahre 2002 unter- oder innerhalb des Bergmassivs durch Starkre-

[5] `http://www.paraview.org`
[6] `http://www.vtk.org`
[7] `http://www.trolltech.com/products/qt`

gen induzierten Erdbeben (Hainzl et al., 2006; Kraft et al., 2006). Die Graphik wurde freundlicherweise von Christoph Moder zur Verfügung gestellt und ist neben Abbildung 2.22 ein weiteres Beispiel für die dreidimensionalen Darstellungsmöglichkeiten in `paraview`.

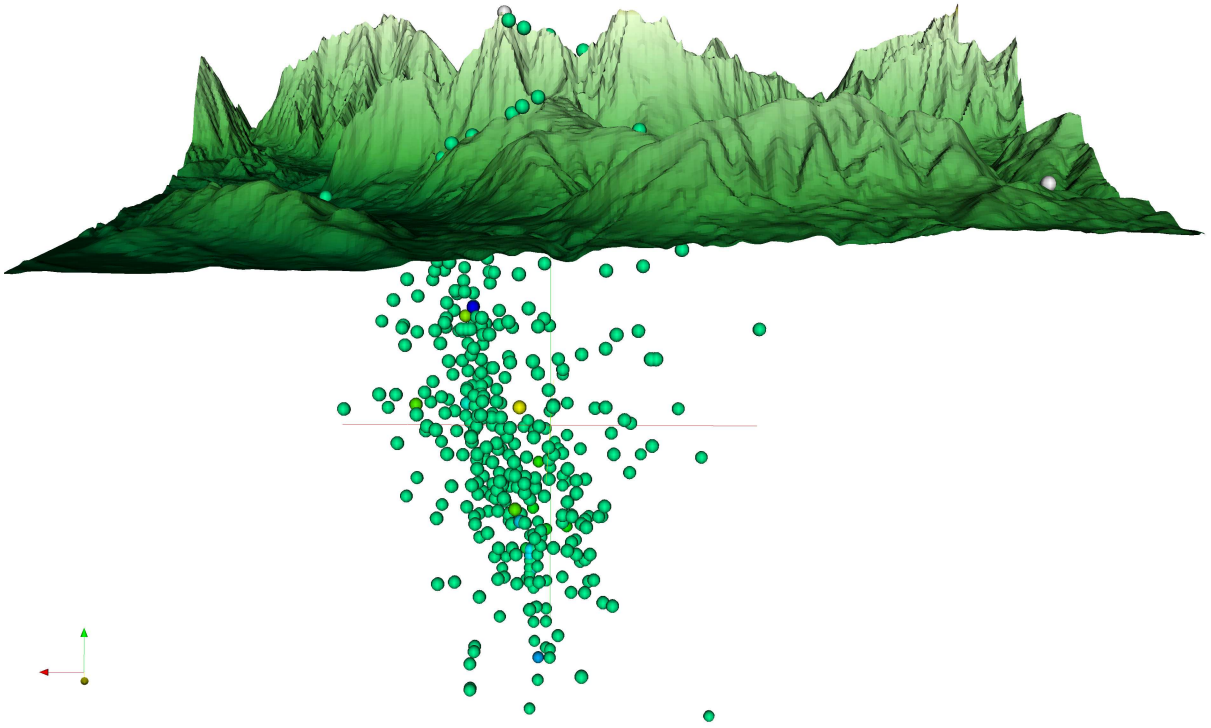


Abbildung 3.13: Blick von Norden auf den Berg Hochstaufen mit den Erdbeben des Jahres 2002 (farbige Kugeln) und den Seismometerstationen (weiße Kugeln), dreifach überhöhte Darstellung

3.7 Bruchausbreitung – bm3d

Die von Gilbert Brietzke in seiner Promotion am Lehrstuhl für Geophysik der Ludwig-Maximilians-Universität entwickelte Anwendung `bm3d` wird für die Modellierung von Erdbebenquellmechanismen als spontane Bruchausbreitung und die Simulation von Erdbebenszenarien sowie von „Strong Ground Motion“ eingesetzt. Die physikalischen Modelle basieren dabei zum einen auf der dynamischen Bruchausbreitung entlang sich reibender Flächen und zum anderen auf der Wellenausbreitung in heterogenen, linearelastischen, isotropen und anisotropen Medien mit einer freien Oberfläche. Das verwendete numerische Verfahren beruht auf der Methode der finiten Differenzen mit einem „Staggered Grid“ zweiter oder vierter Ordnung. Die Randbedingung auf der Verwerfung ist sowohl mit Hilfe des „Traction-At-Split-Node“-Verfahren als auch mittels „Stress-Glut“-Methode implementiert. Der Einsatz von „Perfectly Matched Layers“ (PML) verhindert das Entstehen künstlicher Reflexionen der seismischen Wellen an den Modellrändern. Mehr zu den numerischen Verfahren und Simulationsergebnissen können Brietzke et al. (2007, 2008) entnommen werden. Bei der Parallelisierung des Programms unter Verwendung von MPI wird keine spezielle Gebietszerlegung verwendet, deshalb können beliebige Prozessoranzahlen eingesetzt werden. Die notwendige Arbeitsspeichergröße hängt von der zu bearbeitenden Problemstellung ab. Je nach Größe des Modells werden in Summe zwischen 10 und 120 GB benötigt. Dieser Bedarf kann mit dem Rechencluster TETHYS und dem SMP-Rechner COREDUMP gut abgedeckt werden. Für die Simulationsrechnungen wird auch die am LRZ vorhandene Recheninfrastruktur genutzt.

Gilbert Brietzke stellte freundlicherweise die Abbildung 3.14 zur Verfügung, in welcher ein Simulationsergebnis von `bm3d` zu sehen ist. Es wird die spontane Bruchausbreitung und der Spannungsabfall an einer „Strike-Slip“-Verwerfung zu drei aufeinanderfolgenden Zeitpunkten (2 s, 7.5 s und 13 s) nach der Bruchkeimbildung gezeigt. Die Scherspannung („Shear Stress“) wird farblich kodiert auf der senkrechtstehenden Grenzfläche dargestellt. Die daraus resultierende Verschiebung („Displacement“) an der Oberfläche ist als farbige Fläche gezeichnet, wobei die Gitterverzerrung stark überhöht ist. Amplitude und Richtung des Geschwindigkeitsfelds an der Erdoberfläche werden durch die Verwendung von Kegel repräsentiert.

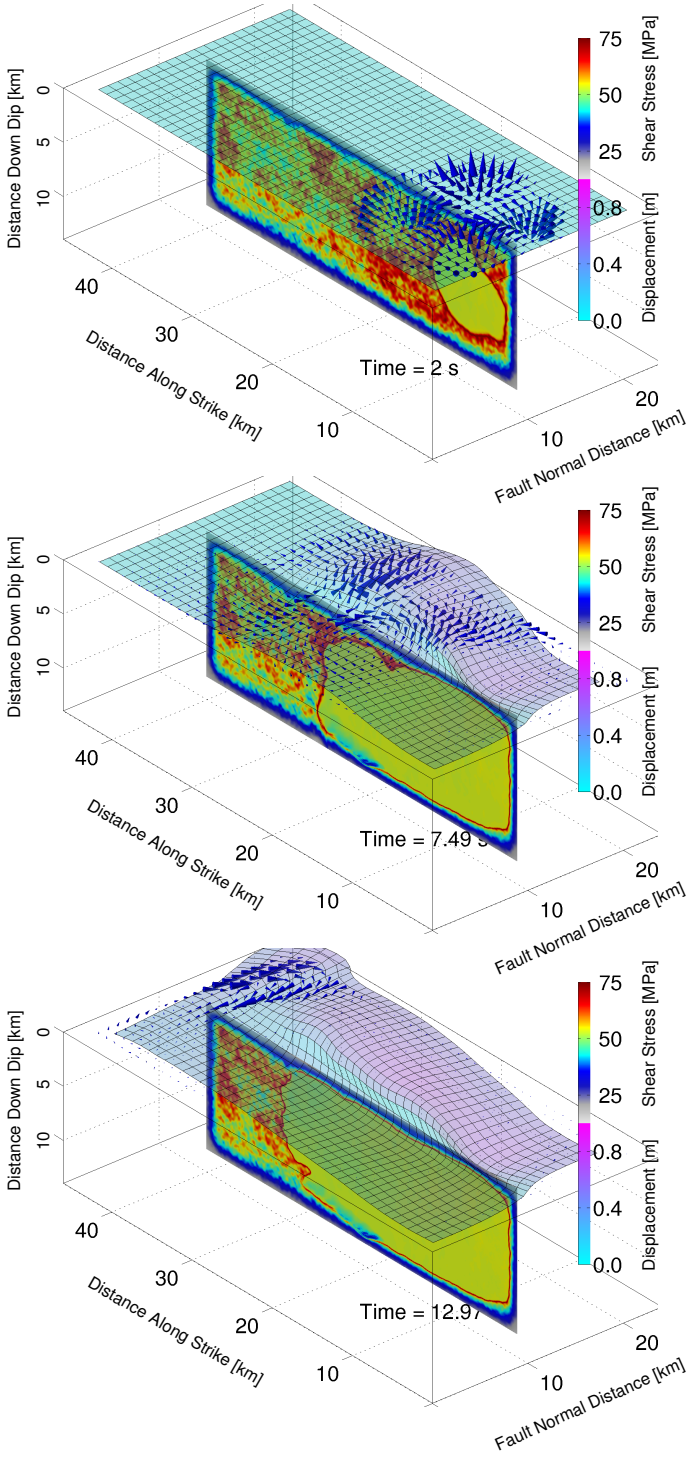


Abbildung 3.14: Bruchausbreitung und Spannungsabfall an einer „Strike-Slip“-Verwerfung zu drei Zeitpunkten – die Scherspannung („Shear Stress“) wird farblich kodiert auf der senkrechtstehenden Grenzfläche und die daraus resultierende Verschiebung („Displacement“) an der Oberfläche dargestellt – Amplitude und Richtung des Geschwindigkeitsfelds an der Erdoberfläche werden durch Kegel repräsentiert

3.8 Wellenform-Tomographie – sec3d/ses3d

Die beiden Anwendungen `sec3d` und `ses3d` dienen der Simulation der elastischen Wellenausbreitung in kartesischen und sphärischen Koordinaten, der seismischen Wellenform-Tomographie sowie der Modellierung und Inversion von finiten Erdbebenquellen (Fichtner et al., 2006a,b; Fichtner & Igel, 2008; Fichtner et al., 2008). Andreas Fichtner entwickelt im Rahmen seiner Promotion am Lehrstuhl für Geophysik der Ludwig-Maximilians-Universität die parallelisierten Programme mit dem breiten Anwendungsspektrum. Als numerische Verfahren werden die spektrale Elemente Diskretisierung im Raum und die Methode der finiten Differenzen zweiter Ordnung für die Zeitdiskretisierung eingesetzt. Die Parallelisierung des Programms erlaubt eine sehr flexible Verwendung des HPC-Systems TETHYS. Der Rechencluster wird vorwiegend zur Weiterentwicklung der Anwendungen und zur Durchführung systematischer Parameterstudien eingesetzt. Für derartige Untersuchung werden mehr als 100 Prozessoren benötigt und die Rechenzeit beträgt bis zu zehn Stunden. Für größere Modellrechnungen wären allerdings mehr als 200 CPUs notwendig. Dieser Bedarf kann gegenwärtig mit TETHYS nicht abgedeckt werden. Im Vergleich zur Prozessoranzahl ist die Größe des Arbeitsspeichers unkritisch. Die vorhandenen 1 GB pro CPU sind ausreichend dimensioniert. Während der Simulationsrechnungen ist ein temporärer Festplattenspeicherplatz von mehr als 1 TB notwendig.

Von Andreas Fichtner wurde freundlicherweise die Abbildung 3.15 zur Verfügung gestellt. Darin ist das vorläufige Resultat einer Wellenforminversion zu sehen, wofür das Programm `ses3d` die synthetischen Seismogramme berechnete.

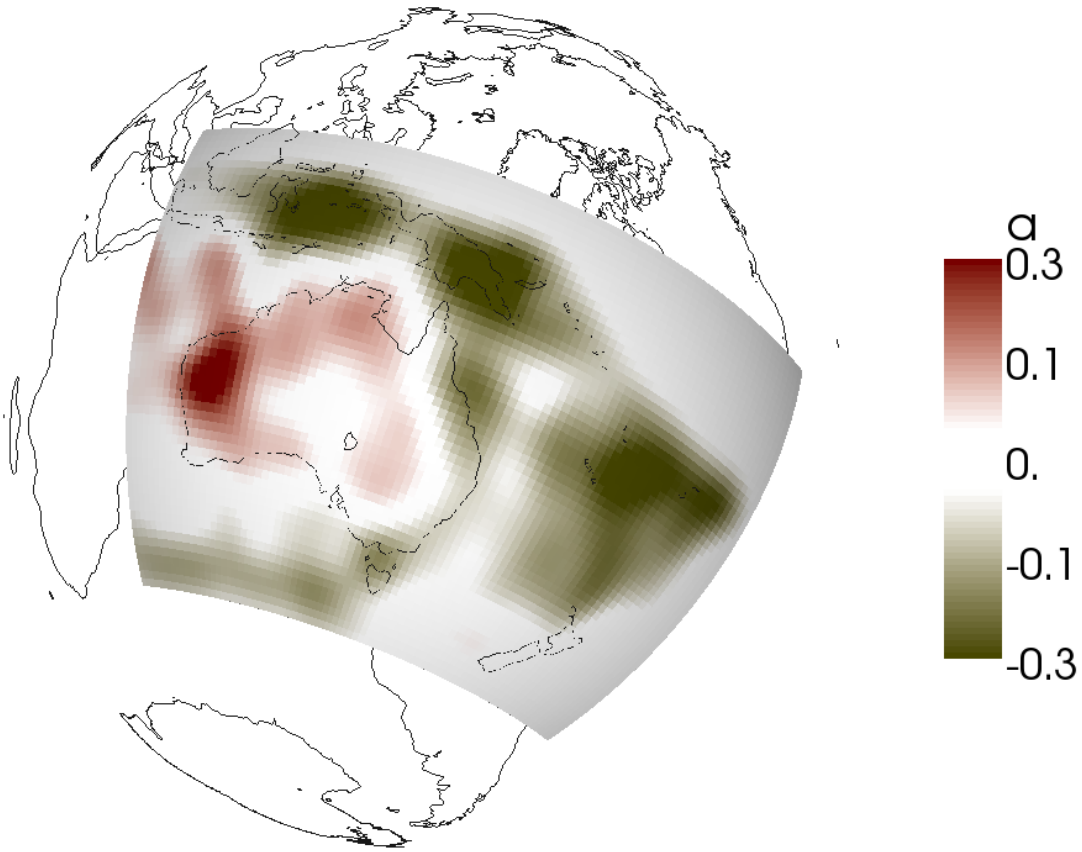


Abbildung 3.15: Wellenforminversion unter Australien – absolute Variation der S-Wellengeschwindigkeit a $[km/s]$ in 100 km Tiefe

3.9 Wellenausbreitung – SPECFEM3D

Die Entwicklung der Simulationsanwendung `SPECFEM3D` wird vom „Computational Infrastructure for Geodynamics“[8] (CIG) in Caltech koordiniert. In seiner Forschungsarbeit am Lehrstuhl für Geophysik der Ludwig-Maximilians-Universität setzt Bernhard Schuberth das Programm zur Modellierung der globalen Ausbreitung elastischer Wellen ein. Das vollständig parallelisierte Programm benutzt ein Hexaedergitter („cubed sphere“– siehe Abbildung 3.16) um den Erdkörper mit seiner inneren Struktur darzustellen. In den Algorithmen werden nach Möglichkeit alle in der Erde auftretenden und für die Simulation relevanten Effekte berücksichtigt. Dies sind beispielsweise die laterale Variation der seismischen Geschwindigkeit und Dichte, die Elliptizität des Erdkörpers, die Topographie und Bathymetrie, die Masse der Ozeane, die Rotation der Erde sowie die Gravitation. Die verwendeten numerischen Verfahren basieren auf der Methode der spektralen Elemente. Aus der Gebietszerlegung ergibt sich die Notwendigkeit eine fest vorgegebenen Prozessoranzahl P zu nutzen. Diese berechnet sich durch $P = 6 \cdot s^2$, wobei s der Anzahl an Gitterteilbereichen in einer Richtung entspricht. Das Hexaedergitter des „cubed sphere“ wird in sechs Teilbereiche zerlegt, wobei jeder aus r Elementen in einer Raumrichtung besteht. Damit kann s dargestellt werden als $s = r/(8 \cdot n)$, mit $n \in \mathbb{N}$ und $r \in 16 \cdot \mathbb{N}$. Mehr zu den genutzten Methoden und deren Anwendung in der Geophysik ist in Komatitsch & Vilotte (1998), Komatitsch & Tromp (1999a, 2002a,b) und Komatitsch et al. (2002) zu finden.

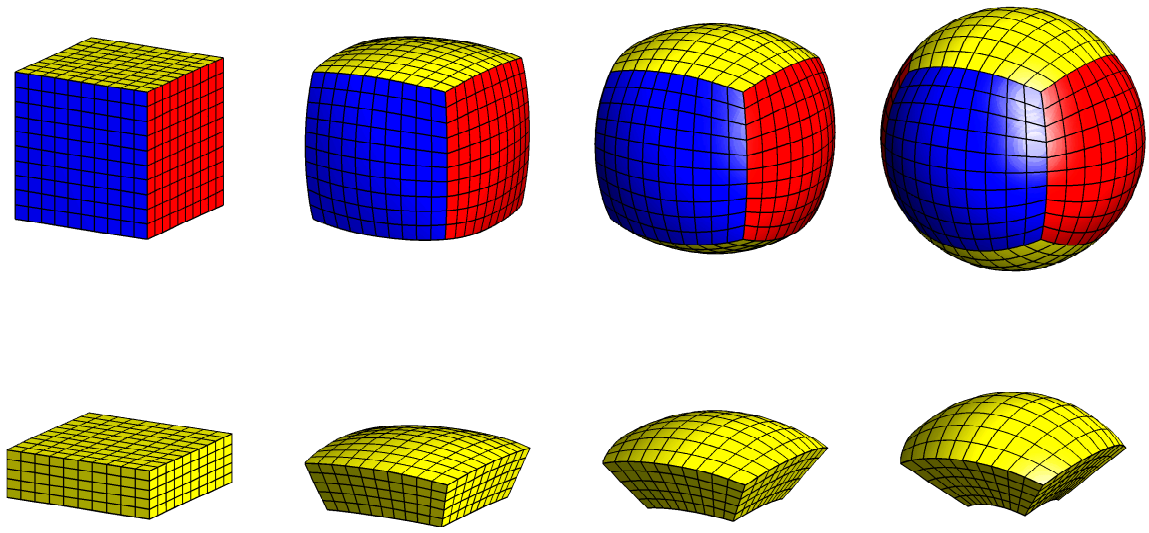


Abbildung 3.16: „cubed sphere“ Gitter – dieses numerische Gitter auf einer Kugeloberfläche wird erzeugt durch die Projektion von sechs gleichmäßigen kartesischen Gittern von der Oberfläche eines Würfels auf die Einheitskugel

[8] `http://www.geodynamics.org/cig`

Bernhard Schuberth stellte freundlicherweise die Simulationsergebnisse in Abbildung 3.17 zur Verfügung. Die einzelnen Bilder zeigen die berechnete Wellenausbreitung für das Sumatra Erdbeben im Dezember 2004 mit einer ausgedehnten Quelle.

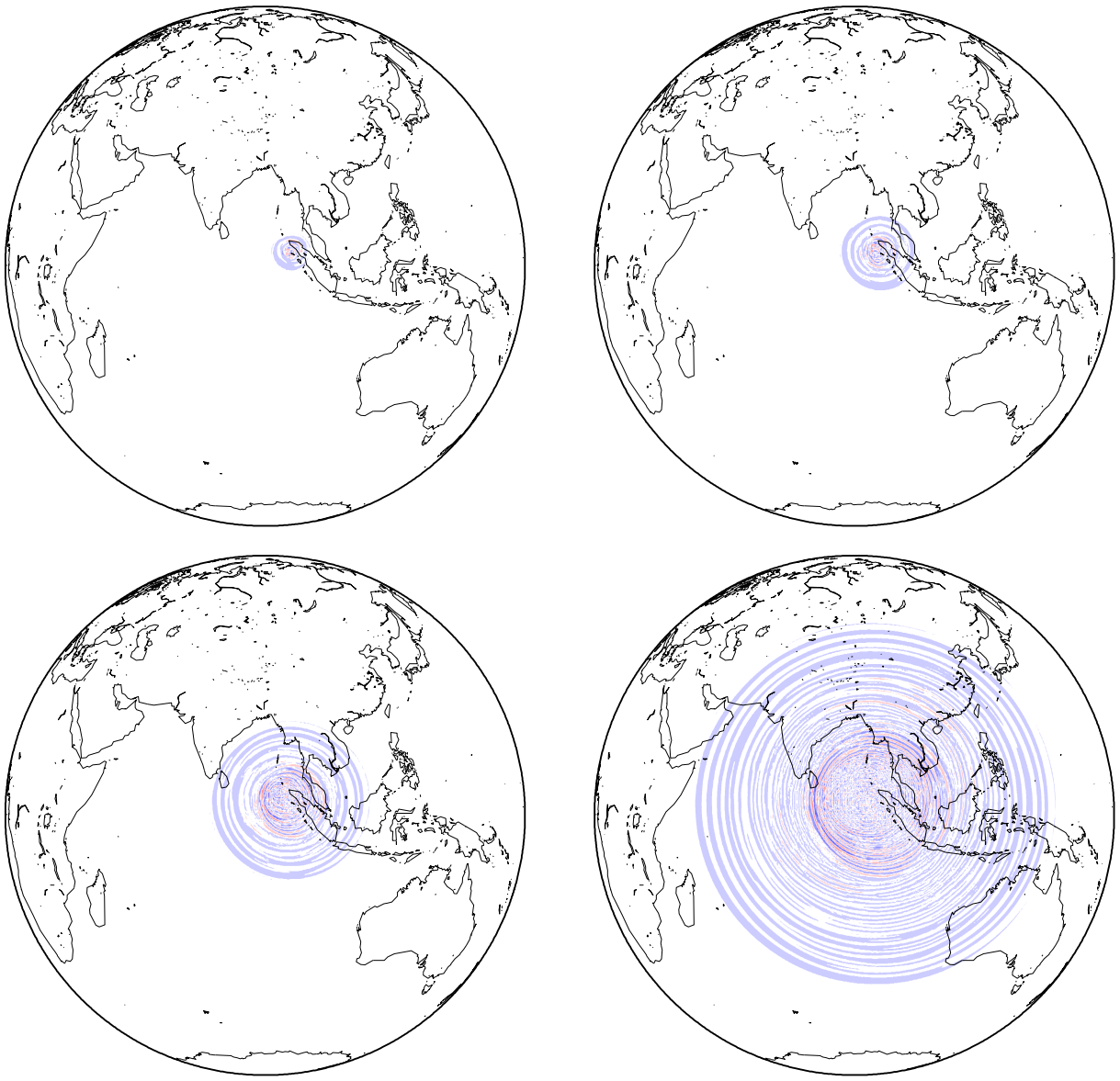


Abbildung 3.17: Ausbreitung der elastischen Welle des Sumatra Erdbebens am 26.12.2004 – dargestellt sind die Wellenfronten zu unterschiedlichen Zeitpunkten nach dem Beben

3.10 Wellenausbreitung – YAC

Die Simulationsanwendung YAC entwickelte Michael Ewald im Rahmen seiner Promotion am Lehrstuhl für Geophysik der Ludwig-Maximilians-Universität. Sie wird zur Modellierung der elastischen Wellenausbreitung in drei Raumdimensionen eingesetzt. Das Ziel seiner Arbeit war die Betrachtung der auftretenden Bodenbewegung während historischer Erdbeben in der niederrheinischen Bucht. Geologische Vorkenntnisse zur Struktur des Gebietes ermöglichten detaillierte Simulationen der Ausbreitung elastischer Wellen für diese Region. Im Anschluss kann daraus die auftretende Bodenbewegung berechnet werden. Damit ist es möglich, die Erdbebengefährdung für die niederrheinischen Bucht zu bestimmen. Aus dieser können die notwendigen Sicherheitsvorkehrungen erarbeitet werden, um im Falle zukünftiger Erdbeben eventuelle Schäden zu vermeiden. Das vollständig parallelisierte Programm verwendet die Finite-Differenzen-Methode mit einem „Staggered Grid“, um die Wellenausbreitung zu berechnen (Ewald, 2006; Ewald et al., 2006). Da keine spezielle Gebietszerlegung angewendet wird, entfällt die Notwendigkeit eine fest vorgegebene Anzahl an Prozessoren einzusetzen. Die in TETHYS verfügbare Arbeitsspeichergröße und Prozessoranzahl deckt den Bedarf der Simulationsanwendung YAC ab. Abbildung 3.18 zeigt ein von Michael Ewald freundlicherweise zur Verfügung gestelltes Simulationsergebnis, welches die Ausbreitung einer elastischen Welle in der niederrheinischen Bucht darstellt. Das Ziel der Modellrechnung war die Verifikation der in YAC implementierten Algorithmen anhand des sehr guten Messdatensatzes für das Alsdorf Erdbeben vom 22.07.2002. Dieses Beben erreichte eine Magnitude von $M_L = 4.9$.

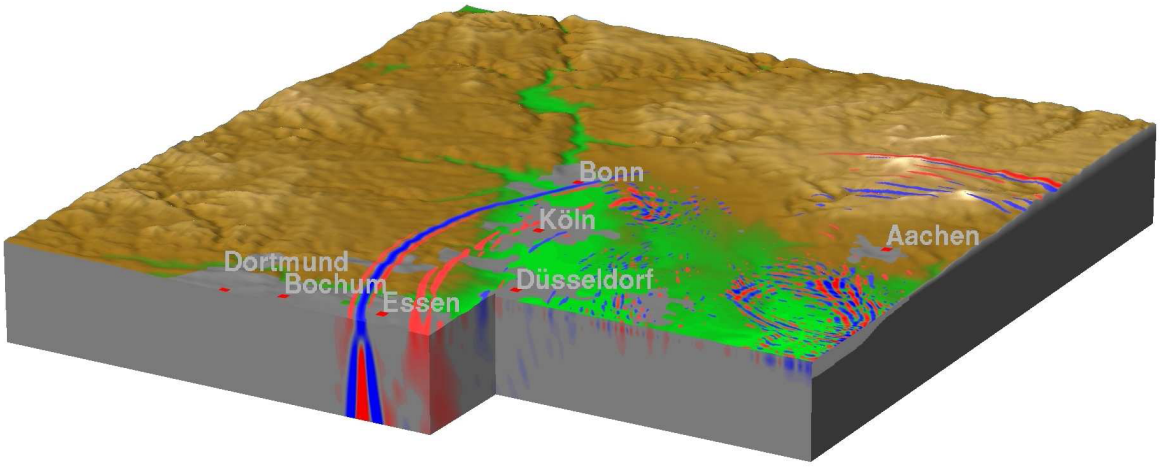


Abbildung 3.18: Ausbreitung einer elastischen Welle in der niederrheinischen Bucht – dargestellt sind die Wellenfronten (rot und blau)

4 Zusammenfassung

Das Ziel der vorliegenden Arbeit war die Entwicklung einer integrierten IT-Infrastruktur für die Simulation komplexer geophysikalischer Prozesse. Mit der Neuausrichtung des Lehrstuhls für Geophysik an der Ludwig-Maximilians-Universität zur Modellierung derartiger Problemstellungen bestand einerseits die Möglichkeit und andererseits die Notwendigkeit, die bis zu diesem Zeitpunkt eingesetzten EDV[1]-Strukturen zu analysieren und zu überarbeiten. Die bei der Prüfung angetroffene Vielfalt an Betriebssystemen und Betriebssystemversionen machte es erforderlich, mit einer vollständig zentralisierten und integrierten IT-Infrastruktur neu zu beginnen. Durch die erfolgreiche Umgestaltung konnte die Basis für die ausstehende Simulationsinfrastruktur geschaffen werden. Das erarbeitete Gesamtkonzept für den Lehrstuhl basiert auf folgenden wesentlichen Merkmalen:

- Absicherung der Netzwerke mit Hilfe einer hochverfügbaren „Astaro“ Firewall Lösung,
- Einführung eines „Single Sign On“ (SSO) mit LDAP als Datenbasis,
- einheitliche Benutzerverzeichnisse auf allen Rechnersystemen,
- bevorzugter Einsatz freier Software,
- Reduktion auf zwei Betriebssysteme (Debian GNU/Linux und Microsoft Windows XP).

In Folge der Umgestaltung der IT-Umgebung in der Geophysik konnte nicht nur der Aufwand für die Systemverwalter verringert werden, sondern es war auch möglich, die Effektivität beim Umgang mit den Rechnern für jeden Anwender zu steigern.

Der nächste Schritt sah den Aufbau der notwendigen Simulationsinfrastruktur und deren Integration in das neu strukturierte Lehrstuhlnetzwerk vor. Der vollständige Prozess der Beschaffung des HPC-Systems TETHYS beinhaltete die Antragsstellung im HBFG-Programm der Deutschen Forschungsgemeinschaft, die Konzeption im Vorfeld der Ausschreibung, die Inbetriebnahme und zum Abschluss die Leistungsbewertung des Rechenclusters. Die genau durchgeführten Planungen in den einzelnen Phasen ermöglichten ein HPC-System, welches:

- bestmöglich auf die geophysikalischen Simulationsanwendungen abgestimmt ist,
- eine hohe Rechenleistung bietet,

[1] Elektronische Datenverarbeitung

- für die Nutzer einfach zu bedienen ist,
- für die Systembetreuer effizient zu verwalten ist und
- eine hohe Stabilität im Rechnerbetrieb aufweist.

Ein wichtiges Anliegen der Neuausrichtung des Lehrstuhls für Geophysik hin zur Simulation komplexer geophysikalischer Prozesse war die kontinuierliche Bereitstellung von Rechenleistung für das in der Modellierung wichtige „Capacity Computing“. Mit dem HPC-Rechner TETHYS konnte dazu der Grundstein gelegt werden. Demgegenüber war und wird für das „Capabilty Computing“ auch in Zukunft die HPC-Infrastruktur des Leibniz-Rechenzentrums unabdingbar sein. Nach der nun gut dreieinhalbjährigen Nutzungsdauer des Rechenclusters TETHYS ist es an der Zeit, die Planung und Antragsstellung für das Folgesystem zu erarbeiten. Die in dem gesamten Zeitraum gesammelten Erfahrungen sind für den zukünftigen HPC-Rechner von unschätzbarem Wert. So hat es sich unter anderem gezeigt, dass die vorhandene Raumklimatisierung nicht den gestellten Anforderungen an das System gewachsen ist. Durch Defekte an der Anlage treten häufig Unterbrechungen im Rechenbetrieb auf. Die knapp bemessenen Etats werden es sicher nicht ermöglichen, die vorhandene Klimaanlage gegen ein stabileres System auszutauschen, weshalb die Wartung und Pflege umso bedeutender ist. Die Arbeitsspeicherprobleme an den Rechenknoten zeigten, dass in zukünftigen Ausschreibungen eine strengere Abnahme erforderlich ist. So könnte nach der Installation des Rechners ein zweimonatiger Probebetrieb erfolgen, in der das System die Stabilität nachweisen muss. Erst nach dessen erfolgreichem Abschluss kann die Abnahme erfolgen. Für das Nachfolgesystem wird die zu beschaffende Hardware wieder den wichtigsten Aspekt darstellen. Es gilt für die wesentlichen Simulationsanwendungen, Benchmarks zu entwickeln und im Vorfeld der Ausschreibung auf verschiedenen Rechnersystemen einzusetzen. Die dabei erhaltenen Erkenntnisse erlauben die Wahl der nach Möglichkeit optimalen Rechnerplattform. Die verfügbaren Mehrkernprozessoren sind darin von besonderem Interesse, wie die Benchmarkrechnungen auf TETHYS und COREDUMP eindrucksvoll belegen.

Zeitgleich mit dem Aufbau der Hochleistungsrechenkapaziät in der Geophysik konnte auch die Basis für das Massenspeichersystem gelegt werden. Mit der ersten Ausbaustufe wurden die, für die Nutzer der Geophysik bereits einheitlich vorliegenden, Benutzerverzeichnisse auf ein performantes Speichersystem verschoben. Das NetApp FAS 3020 System versorgt seit Januar 2006 alle Rechner mit den Benutzerverzeichnissen über die Protokolle NFS und CIFS. Dazu zählen auch die 81 Rechenknoten des HPC-Systems. Zusätzlich konnte ein Datenspeicherbereich der Größe 2 TB eingerichtet werden, der die Zwischenresultate und Simulationsergebnisse vorhält. Für die zweite Ausbaustufe wurde bereits bei der Beschaffung dieses Speichersystem darauf geachtet, dass eine effiziente Synchronisierung von Speicherbereichen zwischen den Standorten der Geophysik in München und Fürstenfeldbruck möglich ist. In dieser Ausbaustufe konnte

zum einen der Speicherbereich für die Simulationsdaten auf 10 TB ausgebaut werden und zum anderen wurde ein zweites NetApp FAS 3020 System am geophysikalischen Observatorium in Fürstenfeldbruck aufgestellt. Dieses weitere Massenspeichersystem stellt wiederum die Benutzerverzeichnisse im Observatorium bereit und ermöglicht das Ablegen der anfallenden Observationsdaten der nächsten fünf bis sechs Jahre in mehreren Speicherbereichen der Gesamtgröße 10 TB. Die abgelegten Messdaten werden über SnapMirror mit Volumes der gleichen Größe in München synchronisiert. Damit sind die Magnetfeld- und Seismologiedaten nicht nur für die Mitarbeiter in Fürstenfeldbruck, sondern auch für alle Wissenschaftler weltweit mit hoher Geschwindigkeit verfügbar. Das entwickelte Massenspeicherkonzept mit zwei NetApp Systemen eignet sich auf Grund der hohen Leistungsfähigkeit und des geringen Wartungsaufwands für die Nutzung in geowissenschaftlichen Forschungseinrichtungen.

Das Visualisierungslabor und SMP-System vervollständigen die IT-Infrastruktur in der Geophysik. Mit dem als HOLODECK bezeichnetem System zur interaktiven, dreidimensionalen Darstellung von Simulationsergebnissen ist es den wissenschaftlichen Mitarbeitern am Lehrstuhl möglich, ihre Simulations- und Messdaten geeignet darzustellen. Die Anpassung der Simulationsanwendungen für die in Zukunft weit verbreiteten Mehrkernsysteme wird durch das SMP-Systems COREDUMP ermöglicht. Zusätzlich ist dieser Rechner auf Grund des großen Arbeitsspeichers und des SMP-Charakters ideal für Anwendungen geeignet, die nicht parallelisiert sind, aber viel RAM benötigen. Zu dieser Kategorie zählen vereinzelte Simulationsanwendungen (auch kommerzielle Produkte wie Matlab) und die Softwareprodukte für das Erstellen von Gittern in FEM Simulationen. Auf COREDUMP konnten bereits wichtige Erkenntnisse für das TETHYS Nachfolgesystem in verschiedenen Benchmarkrechnungen gesammelt werden.

Die Weiterentwicklung der Rechner bedingt natürlich auch eine ständige Anpassung der IT-Infrastruktur am Lehrstuhl für Geophysik. Durch die flexible und zentralisierte Architektur ist die Geophysik aber bestens für die neuen Herausforderungen gerüstet. Dabei sollte die Stabilität immer im Mittelpunkt vor eventuellen Neuheiten stehen, denn nur durch eine stabile Simulationsinfrastruktur waren viele der bisherigen wissenschaftlichen Arbeiten möglich.

A Terminologie

A.1 Speedup

Als Maß für den erreichten Parallelisierungsgrad wird im Allgemeinen der Speedup verwendet. Er charakterisiert den Faktor des Geschwindigkeits- bzw. Laufzeitgewinns. Für eine feste Problemgröße lässt sich der parallele Speedup $S_{par}(P)$ über

$$S_{par}(P) = \frac{t(1)}{t(P)} \tag{A.1}$$

berechnen, wobei P die Anzahl der verwendeten Prozessoren beschreibt. Die Bestimmung der Laufzeit t erfolgt für einen und P Prozessoren mit ein und demselben parallelen Algorithmus. Der erreichbare parallele Speedup $S_{par}(P)$ für P Prozessoren ist im Optimalfall gleich P. Dies ist gleichbedeutend mit einer perfekten Skalierbarkeit des parallelen Algorithmus, d.h. beim Einsatz von P Prozessoren reduziert sich die Laufzeit auf $1/P$ der Laufzeit für einen Prozessor.

Mit steigender Prozessoranzahl ist auch eine Steigerung der gesamten Problemgröße wünschenswert. Damit lässt sich das gegebene Problem aber nicht mehr auf einem Rechner ausführen, was zu Problemen mit der bisherigen Definition des Speedup führt. Es ergibt sich eine Definition des Speedup normiert auf eine feste Laufzeit $S_{skal}(P)$.

$$S_{skal}(P) = \frac{\text{Problemgröße auf } P \text{ Rechnern bei fester Laufzeit } t}{\text{Problemgröße auf 1 Rechner bei fester Laufzeit } t} \tag{A.2}$$

Ziel ist dabei nicht die Reduzierung der Gesamtlaufzeit, sondern eine Erhöhung der Problemgröße bei gleich bleibender Laufzeit. Daher erscheint P=1 im Nenner. Dies wird als skalierbarer Speedup $S_{skal}(P)$ oder Scaleup bezeichnet.

Durch die Definition des parallelen Speedup ist es möglich, verschiedene Implementierungen eines Algorithmus für eine Rechnerarchitektur zu vergleichen. Allerdings sind große Speedup-Werte noch kein Hinweis auf eine gute Parallelisierung, da die Speedup-Zahlen für die Implementierung eines Algorithmus auf verschiedenen Architekturen unterschiedlich sein können. Es muss zusätzlich noch ein Vergleich der Laufzeiten vorgenommen werden. Da es immer einzelne Programmteile gibt, die rein sequentiell abgearbeitet werden müssen, wird sich der erreichte

Speedup nur an P annähern. Im AMDAHLschen Gesetz wird daher der Speedup unter Beachtung des sequentiellen (A_s) und parallelen Anteils (A_p) am Algorithmus wie folgt berechnet:

$$S = \frac{1}{A_s + \frac{A_p}{P}} . \tag{A.3}$$

Bei der Implementierung sollte daher auf einen geringen Anteil an rein sequentiell abzuarbeitenden Programmcode geachtet werden.

A.2 Chipsatz

Mehrere zusammengehörige integrierte Schaltkreise, die auf einer Platine verbaut sind und gemeinsam bestimmte Aufgaben erfüllen, werden im Allgemeinen als Chipsatz bezeichnet. Im Speziellen wird der auf einer PC-Hauptplatine verbaute Chipsatz damit angesprochen, welcher den verbauten Prozessor in seinen Aufgaben unterstützt.

Er setzt sich heute meist aus der North- und Southbridge zusammen, deren Namen sich aus der Lage zum Prozessor ableiten lässt. Dabei liegt die Northbridge näher am Prozessor, wohingegen die Southbridge weiter entfernt angesiedelt ist. Diese beiden Komponenten koordinieren den Datentransfer zwischen den verschiedenen Geräten. In der Northbridge sind meist die schnelleren und auch aufwendigeren Funktionen wie die Schnittstelle zum Arbeitsspeicher integriert, wohingegen die Southbridge langsamere Geräte wie die seriellen Schnittstellen anspricht. Diese Aufteilung hat sich bei den AMD Opteron Prozessoren geändert, da bereits in der CPU die Schnittstelle zum Arbeitsspeicher eingebaut ist, fällt die separierte Northbridge weg und es existiert nur noch ein Chip für alle anderen Aufgaben. In Zukunft ist mit der fortschreitenden Miniaturisierung eine weitere Zunahme, der in den Prozessor integrierten Aufgaben zu erwarten.

A.3 AMD64

Die früher als x86-64 bekannte Systemarchitektur wird seit AMDs Einstieg in den 64-Bit Prozessormarkt auch als AMD64 benannt, da von AMD einige neue Befehle hinzugefügt wurden. Intel stieg kurze Zeit später auch in diesen Markt ein und deren Prozessoren verwenden das Kürzel EM64T, sind aber weitgehend kompatibel zu AMD64.

Mit AMD64 wurde von der Firma AMD ein sanfter Wechsel von 32-Bit hin zu 64-Bit gestartet, da die Prozessoren einen vollwertigen 32-Bit Prozessor darstellen, dessen Register im 64-Bit Modus verbreitet werden, können 32-Bit Anwendungen uneingeschränkt ausgeführt werden.

Der darüber hinaus noch verfügbare 64-Bit Modus ermöglicht vor allem das Ansprechen größerer Speicherbereiche und bringt durch die verbreiterten Register auch Leistungsverbesserungen mit.

Die Breite einer virtuellen Adresse ist bei AMD64 48 Bit, womit 256 TB angesprochen werden können. Da die Prozessoren aber zur Zeit nur 40 Adresspins haben, kann physikalisch nur 1 TB Arbeitsspeicher angesprochen werden. Es lassen sich aber jederzeit mehr Pins durch AMD hinzufügen.

Welche Vor- und Nachteile bieten sich durch die Verwendung dieser Systemarchitektur?

Nachteil – Speicherverbrauch: Da alle Adresswerte nun 64-Bit breit sind, verbraucht ihre Speicherung auch doppelt soviel Platz. Wenn diese Werte zwischen RAM und CPU bewegt werden müssen, so müssen auch doppelt so viele Bytes transferiert werden.

Neutral – Gleitpunktzahloperationen: Da die SSE-Register schon immer 128-Bit breit waren und die SSE-Einheit den größten Teil der Berechnungen für Multimedia und Mathematik durchführt, ist mit keiner Verbesserung zu rechnen.

Vorteil -– Registeranzahl: Die AMD64 Architektur verfügt über die doppelte Zahl an allgemeinen Registern, weshalb Zwischenwerte nicht immer in den Arbeitsspeicher ausgelagert werden müssen und eine Beschleunigung zu erwarten ist.

Vorteil — Adressbreite: Bei der Verarbeitung von Daten im GB-Bereich entfallen aufwendige Zugriffsberechnungen mit 32-Bit Differenzen zu Basisadressen.

Vorteil — Integerarithmetik: Da die Multiplikationen mit ganzen Zahlen größer als 32-Bit erheblich schneller abläuft, profitieren Anwendungen mit Ganzzahlenberechnungen davon.

Allgemein ist der Nutzen stark von den einzelnen Anwendungen abhängig und den darin benutzten Algorithmen.

Eine volle Unterstützung dieser AMD64 Architektur war zuerst in den Betriebssystemen basierend auf GNU/Linux vorhanden. Mittlerweile ist für alle Betriebssysteme eine entsprechend angepasste Version verfügbar.

A.4 NFS- und CIFS-Protokoll

Das „Network File System“ (NFS) Protokoll wurde von Sun Microsystems Inc. entwickelt, um den Zugriff auf Dateien über ein Netzwerk zu ermöglichen. Der Umgang mit den Dateien über

das Netzwerk ist vergleichbar zu lokal auf der Festplatte abgelegten Dateien. Das Protokoll ist nach den Dokumenten RFC 1094, RFC 1813 und RFC 3530 standardisiert. Es verwendet das Netzwerktransportprotokoll TCP/IP und ist seit Version 4 nicht mehr zustandslos.

Das „Common Internet File System“ (CIFS) wurde 1996 von Microsoft eingeführt. Es handelt sich dabei um eine erweiterte Version von SMB („Server Message Block“). Das CIFS-Protokoll bietet neben Datei- und Druckdiensten auch noch weitere für Microsoft Windows spezifische Dienste an.

Literaturverzeichnis

Bard, P.-Y., Chaljub, E., Cornou, C., Cotton, F., & Gueguen, P. (2006a). *Third International Symposium on the Effects of Surface Geology on Seismic Motion*, volume 1. Laboratoire Central des Ponts et Chaussées.

Bard, P.-Y., Chaljub, E., Cornou, C., Cotton, F., & Gueguen, P. (2006b). *Third International Symposium on the Effects of Surface Geology on Seismic Motion*, volume 2. Laboratoire Central des Ponts et Chaussées. in press.

Barrett, R., Berry, M., Chan, T. F., Demmel, J., Donato, J., Dongarra, J. J., Eijkhout, V., Pozo, R., Romine, C., & van der Vorst, H. (1994). *Templates for the Solution of Linear Systems: Building Blocks for Iterative Methods*. Society for Industrial and Applied Mathematics.

Baumgardner, J. R. (1985). Three-Dimensional Treatment of Convective Flow in the Earth's Mantle. *Journal of Statistical Physics*, *39*(5/6), 501–511.

Baumgardner, J. R. & Frederickson, P. O. (1985). Icosahedral discretization of the two-sphere. *SIAM Journal on Numerical Analysis*, *22*(6), 1107–1115.

Benzi, M., Golub, G., & Liessen, J. (2005). Numerical solution of saddle point problems. *Acta Numerica*, *14*, 1–137.

Bercovici, D. (1995). A source-sink model of the generation of plate tectonics from non-newtonian mantle flow. *Journal of Geophysical Research*, *100*, 2013–2030.

Bird, P. (1998). Testing hypotheses on plate-driving mechanisms with global lithosphere models including topography, thermal structure, and faults. *Journal of Geophysical Research*, *103*, 10115–10129.

Braess, D. (1997). *Finite Elemente : Theorie, schnelle Löser und Anwendung in der Elastizitätstheorie*. Springer-Verlag Berlin, Heidelberg, New York.

Brietzke, G. B., Cochard, A., & Igel, H. (2007). Dynamic Rupture Along Bimaterial Interfaces in 3D. *Geophysical Research Letters*, *34*, L11305.

Brietzke, G. B., Cochard, A., & Igel, H. (2008). Importance of Bimaterial Interfaces for Earthquake Dynamics and Strong Ground Motion. *Geophysical Journal International*, –. (submitted).

Bronstein, I. N. & Semendjajew, K. A. (1991). *Taschenbuch der Mathematik*. B. G. Teubner Verlagsgesellschaft, Stuttgart Leipzig.

Bunge, H.-P. & Davies, J. H. (2001). Tomographic images of a mantle circulation model. *Geophysical Research Letters*, *28*(1), 77–80.

Bunge, H.-P. & Grand, S. (2000). Mesozoic plate-motion history below the northeast Pacific Ocean from seismic images of the subducted Farallon slab. *Nature*, *405*, 337 – 340.

Bunge, H.-P., Hagelberg, C. R., & Travis, B. J. (2003). Mantle circulation models with variational data assimilation: inferring past mantle flow and structure from plate motion histories and seismic tomography. *Geophysical Journal International*, *152*(2), 280–301.

Bunge, H.-P., Richards, M., & Baumgardner, J. (1996). The effect of depth dependent viscosity on the planform of mantle convection. *Nature*, *379*, 436–438.

Bunge, H.-P., Richards, M., & Baumgardner, J. (1997). A sensitivity study of 3-D spherical mantle convection at 10exp8 Rayleigh number: Effects of depth dependent viscosity, heating mode and an endothermic phase change. *Journal of Geophysical Research*, *102*, 11991–12007.

Bunge, H.-P., Richards, M., Lithgow-Bertelloni, C., Baumgardner, J., Grand, S., & Romanowicz, B. (1998). Time scales and heterogeneous structure in geodynamic earth models. *Science*, *280*, 91–95.

Bunge, H.-P., Richards, M. A., & Baumgardner, J. R. (2002). Mantle circulation models with sequential data-assimilation: Inferring present-day mantle structure from plate motion histories. *Philosophical Transactions of the Royal Society of London: Series A*, *360*(1800), 2545–2567.

Carns, P. H., Walter, Ross, R. B., & Thakur, R. (2000). PVFS: A Parallel File System for Linux Clusters. In *Proceedings of the 4th Annual Linux Showcase and Conference*, (pp. 317–327)., Atlanta, GA. USENIX Association.

Courtier, P. & Talagrand, O. (1987). Variational assimilation of meterological observations with the adjoint vorticity equation I: Numerical results. *Quarterly Journal of the Royal Meteorological Society*, *113*, 1329–1347.

Davies, G. F. (1989). Mantle convection model with a dynamic plate-topography, heat-flow and gravity anomalies. *Geophysical Journal International*, *98*, 461–464.

Davies, G. F. (1999). *Dynamic earth: plates, plumes, and mantle convection*. Cambridge Atmospheric and Space Science Series. Cambridge University Press.

de la Puente, J., Käser, M., Dumbser, M., & Igel, H. (2007). An Arbitrary High Order Discontinuous Galerkin Method for Elastic Waves on Unstructured Meshes IV: Anisotropy. *Geophysical Journal International*, *169*(3), 1210–1228.

Dixon, T. H. (1991). An introduction to the Global Positioning System and some geological applications. *Reviews of Geophysics*, *29*, 249–276.

Duff, I., Erisman, A., & Reid, J. (1989). *Direct Methods for Sparse Matrices*. Monographs on Numerical Analysis. Oxford University Press.

Dumbser, M. & Käser, M. (2006). An Arbitrary High Order Discontinuous Galerkin Method for Elastic Waves on Unstructured Meshes II: The Three-Dimensional Isotropic Case. *Geophysical Journal International*, *167*(1), 319–336.

Dumbser, M., Käser, M., & Toro, E. (2007). An Arbitrary High Order Discontinuous Galerkin Method for Elastic Waves on Unstructured Meshes V: Local Time Stepping and p-Adaptivity. *Geophysical Journal International*, *171*(2), 695–717.

Ewald, M. (2006). *Numerical Simulations of Earthquake Scenarios in the Lower Rhine Embayment Area: Numerische Simulation von Erdbebenszenarien im Raum der Niederrheinischen Bucht*. PhD thesis, LMU Munich: Faculty of Geosciences.

Ewald, M., Igel, H., Hinzen, K.-G., & Scherbaum, F. (2006). Basin-related effects on ground motion for earthquake scenarios in the Lower Rhine Embayment. *Geophysical Journal International*, *166*, 197–212.

Faccioli, E., Maggio, F., Paolucci, R., & Quarteroni, A. (1997). 2D and 3D elastic wave propagation by a pseudo-spectral domain decomposition method. *Journal of Seismology*, *1*, 237–251.

Felder, M., Käser, M., & Dumbser, M. (2006). Towards Optimising a Novel Seismological Solver Code. Technical Report 2006-04, Leibniz-Rechenzentrum der Bayerischen Akademie der Wissenschaften.

Fichtner, A., Bunge, H.-P., & Igel, H. (2006a). The adjoint method in seismology: I - Theory. *Physics of The Earth and Planetary Interiors*, *157*(1-2), 86–104.

Fichtner, A., Bunge, P., & Igel, H. (2006b). The adjoint method in seismology: II - Applications: traveltimes and sensitivity functionals. *Physics of the Earth and Planetary Interiors*, *157*(1-2), 105–123.

Fichtner, A. & Igel, H. (2008). Efficient numerical surface wave propagation through the optimization of discrete crustal models - a technique based on non-linear dispersion curve matching (DCM). *Geophysical Journal International*, *173*(2), 519–533.

Fichtner, A., Kennett, B. L. N., Igel, H., & Bunge, H.-P. (2008). Theoretical background for continental and global scale full-waveform inversion in the time-frequency domain. *Geophysical Journal International*, –. (submitted).

Gable, C. W., O'Connell, R. J., & Travis, B. J. (1991). Convection in 3 dimensions with surface plates – generation of toroidal flow. *Journal of Geophysical Research*, *96*, 8391–8405.

Gabriel, E., Fagg, G. E., Bosilca, G., Angskun, T., Dongarra, J. J., Squyres, J. M., Sahay, V., Kambadur, P., Barrett, B., Lumsdaine, A., Castain, R. H., Daniel, D. J., Graham, R. L., & Woodall, T. S. (2004). Open MPI: Goals, Concept, and Design of a Next Generation MPI Implementation. In *Proceedings, 11th European PVM/MPI Users' Group Meeting*, (pp. 97–104)., Budapest, Hungary.

Glatzmaier, G. A. (1988). Numerical simulations of mantle convection: Time-dependent, three-dimensional, compressible, spherical shell. *Geophysical and Astrophysical Fluid Dynamics*, *43*, 223–264.

Glatzmaier, G. A. (2002). Geodynamo simulations - how realistic are they? *Annual Review of Earth and Planetary Sciences*, *30*, 237–257.

Glatzmaier, G. A. & Roberts, P. H. (1995). A three-dimensional, self-consistent computer simulation of a geomagnetic field reversal. *Nature*, *377*, 203–209.

Graham, R. L., Shipman, G. M., Barrett, B. W., Castain, R. H., Bosilca, G., & Lumsdaine, A. (2006). Open MPI: A high-performance, heterogeneous MPI. In *Proceedings, Fifth International Workshop on Algorithms, Models and Tools for Parallel Computing on Heterogeneous Networks*, Barcelona, Spain.

Gropp, W., Lusk, E., Doss, N., & Skjellum, A. (1996). A high-performance, portable implementation of the MPI message passing interface standard. *Parallel Computing*, *22*(6), 789–828.

Gropp, W. D. & Lusk, E. (1996). *User's Guide for* `mpich`*, a Portable Implementation of MPI*. Mathematics and Computer Science Division, Argonne National Laboratory. ANL-96/6.

Hainzl, S., Kraft, T., Wassermann, J., & Igel, H. (2006). Evidence for rain-triggered earthquake activity. *Geophysical Research Letters*, *33*.

Heidbach, O., Iaffaldano, G., & Bunge, H.-P. (2008). Topography growth drives stress rotations in the central Andes: 3 Observations and models. *Geophysical Research Letters*, *35*. in press.

Heidbach, O., Reinecker, J., Tingay, M., Müller, B., Sperner, B., Fuchs, K., & Wenzel, F. (2007). Plate boundary forces are not enough: Second- and third-order stress patterns highlighted in the world stress map database. *Tectonics*, *26*.

Hollerbach, R. (1996). On the theory of the geodynamo. *Physics of the Earth and Planetary Interiors*, *98*, 163–185.

Hülsemann, F., Kowarschik, M., Mohr, M., & Rüde, U. (2005). Parallel Geometric Multigrid. In A. M. Bruaset & A. Tveito (Eds.), *Numerical Solution of Partial Differential Equations on Parallel Computers*, number 51 in Lecture Notes in Computational Science and Engineering. Springer. ISBN: 3-540-29076-1.

Iaffaldano, G., Bunge, H.-P., & Buecker, M. (2007). Mountain belt growth inferred from histories of past plate convergence: A new tectonic inverse problem. *Earth and Planetary Science Letters*, *260*, 516–523.

Iaffaldano, G., Bunge, H.-P., & Dixon, T. H. (2006). Feedback between mountain belt growth and plate convergence. *Geology*, *34*, 893–896.

Igel, H. (2002). 3D Seismic Wave Propagation on a Global and Regional Scale: Earthquakes, Fault Zones, Volcanoes. In *High Performance Computing in Science and Engineering*. Heidelberg: Springer Verlag. ISBN 3-540-00474-2.

Igel, H., Käser, M., & Stuppazini, M. (2008). *Encyclopedia of Complexity and System Science*, chapter Simulation of SeismicWave Propagation in Media with Complex Geometries. Springer-Verlag. in press.

Igel, H. & Weber, M. (1995). SH-wave propagation in the whole mantle using high-order finite differences. *Geophysical Research Letters*, *22*(6), 731–734.

Ismail-Zadeh, A., Schubert, G., Tsepelev, I., & Korotkii, A. (2004). Inverse problem of thermal convection: numerical approach and application to mantle plume restoration. *Physics of the Earth and Planetary Interiors*, *145*, 99–114.

Jarvis, G. T. & McKenzie, D. P. (1980). Convection in a compressible fluid with infinite prandtl number. *Journal of Fluid Mechanics*, *96*, 515–583.

Jung, M. & Langer, U. (2001). *Methode der finiten Elemente für Ingenieure*. B. G. Teubner Stuttgart, Leipzig, Wiesbaden.

Käser, M. & Dumbser, M. (2006). An Arbitrary High Order Discontinuous Galerkin Method for Elastic Waves on Unstructured Meshes I: The Two-Dimensional Isotropic Case with External Source Terms. *Geophysical Journal International*, *166*(2), 855–877.

Käser, M., Dumbser, M., de la Puente, J., & Igel, H. (2007). An Arbitrary High Order Discontinuous Galerkin Method for Elastic Waves on Unstructured Meshes III: Viscoelastic Attenuation. *Geophysical Journal International*, *168*(1), 224–242.

Kennett, B. & Bunge, H.-P. (2008). *Geocontinua*. Cambridge University Press.

Kirby, S. H. (1983). Rheology of the lithosphere. *Reviews of Geophysics*, *21*, 1458–1487.

Komatitsch, D., Barnes, C., & Tromp, J. (2000). Simulation of anisotropic wave propagation based upon a spectral element method. *Geophysics*, *65*, 1251–1260.

Komatitsch, D., Ritsema, J., & Tromp, J. (2002). The spectral-element method, Beowulf computing, and global seismology. *Science*, *298*(5599), 1737–1742.

Komatitsch, D. & Tromp, J. (1999a). Introduction to the spectral-element method for 3-D seismic wave propagation. *Geophysical Journal International*, *139*(3), 806–822.

Komatitsch, D. & Tromp, J. (1999b). Introduction to the spectral element method for three-dimensional seismic wave propagation. *Geophysical Journal International*, *139*, 806–822.

Komatitsch, D. & Tromp, J. (2002a). Spectral-element simulations of global seismic wave propagation-I. Validation. *Geophysical Journal International*, *149*(2), 390–412.

Komatitsch, D. & Tromp, J. (2002b). Spectral-element simulations of global seismic wave propagation-II. 3-D models, oceans, rotation, and self-gravitation. *Geophysical Journal International*, *150*(1), 303–318.

Komatitsch, D. & Vilotte, J. P. (1998). The spectral-element method: an efficient tool to simulate the seismic response of 2D and 3D geological structures. *Bull. Seismol. Soc. Am.*, *88*(2), 368–392.

Kong, X. & Bird, P. (1995). Shells: A thin-shell program for modeling neotectonics of regional or global lithosphere with faults. *Journal of Geophysical Research*, *100*, 22129–22132.

Kraft, T., Wassermann, J., Schmedes, E., & Igel, H. (2006). Meteorological triggering of earthquake swarms at Mt. Hochstaufen, SE-Germany. *Tectonophysics*, *424*(3-4), 245–258.

Kuang, W. L. & Bloxham, J. (1997). An earth-like numerical dynamo model. *Nature*, *389*, 371–374.

Kunst, E.-K. & Quade, J. (2008). Kern-Technik: Kernel- und Treiberprogrammierung mit dem Kernel 2.6 – Folge 41. *Linux Magazin*, *09*, 88–91.

Landau, L. & Lifschitz, E. (1987). *Fluid mechanics*. Pergamon Press.

McNamara, A. K. & Zhong, S. (2005). Thermochemical structures beneath Africa and the Pacific Ocean. *Nature*, *437*(7062), 1136.

Moczo, P., Kristek, J., & Halada, L. (2004). *The Finite-Difference Method for Seismologists. An Introduction*. Comenius University, Bratislava.

Moder, C. (2006). Visualization in the Geosciences. Diplomarbeit, LMU, Department of Earth and Environmental Sciences, Geophysics.

Moder, C., Bunge, H.-P., Igel, H., & Schuberth, B. (2007). Visualization in the Geosciences with Paraview and Geowall. In Benger, W., Heinzl, R., Kapferer, W., Schoor, W., Tyagi, M., Venkataraman, S., & Weber, G. (Eds.), *Proceedings of the 4th High-End Visualization Workshop*, (pp. 147–155)., Berlin. Lehmanns Media.

Moore, G. E. (1965). Cramming more components onto integrated circuits. *Electronics*, *38*(8).

Moresi, L. & Solomatov, V. (1998). Mantle convection with a brittle lithosphere: thoughts on the global tectonic styles of the earth and venus. *Geophysical Journal International*, *133*, 669–682.

Morton, K. W. & Mayers, D. F. (2005). *Numerical Solution of Partial Differential Equations, An Introduction* (2 ed.). Cambridge University Press.

Oeser, J. (2004). Parallelisierung geoelektrischer Finite-Elemente-Modellrechnungen auf Linux-Clustern. Diplomarbeit, Technische Universität Bergakademie Freiberg, Fakultät für Geowissenschaften, Geotechnik und Bergbau, Institut für Geophysik.

Oeser, J., Bunge, H.-P., & Mohr, M. (2006). Cluster Design in the Earth Sciences: TETHYS. In Gerndt, M. & Kranzlmüller, D. (Eds.), *High Performance Computing and Communications – Second International Conference, HPCC 2006, Munich, Germany*, volume 4208 of *Lecture Notes in Computer Science*, (pp. 31–40). Springer.

Oeser, J., Bunge, H.-P., Mohr, M., & Igel, H. (2009). Frontiers in Computational Geophysics: simulations of mantle circulation, plate tectonics and seismic wave propagation. In *100 Volumes NNFM and 40 Years Numerical Fluid Mechanics*, volume 100 of *Notes on Numerical Fluid Mechanics and Multidisciplinary Design*. Springer.

Ricard, Y. & Vigny, C. (1989). Mantle dynamics with induced plate tectonics. *Journal of Geophysical Research*, *94*, 17543–17559.

Richards, M. A., Yang, W. S., Baumgardner, J. R., & Bunge, H. P. (2001). Role of a low-viscosity zone in stabilizing plate tectonics: Implications for comparative terrestrial planetology. *Geochem. Geophys. Geosys.*, *2*.

Richardson, R. M. & Coblentz, D. D. (1994). Stress modeling in the andes: Constraints on the south america intraplate stress magnitudes. *Journal of Geophysical Research*, *99*, 22015–22025.

Saad, Y. (2003). *Iterative Methods for Sparse Linear Systems* (2nd ed.). Society for Industrial and Applied Mathematics.

Scandella, L. (2004). *Numerical Evaluation of Transient Ground strains for the seismic response analisys of underground structures*. PhD thesis, Politecnico di Milano (Italy).

Sigloch, K., McQuarrie, N., & Nolet, G. (2008). Two-stage subduction history under north america inferred from finite-frequency tomography. *Nature Geoscience*, –. (in review).

Song, T. R. A. & Simons, M. (2003). Large trench-parallel gravity variations predict seismogenic behavior in subduction zones. *Science*, *301*, 630–633.

Stemmer, K., Harder, H., & Hansen, U. (2006). A new method to simulate convection with strongly temperature- and pressure-dependent viscosity in a spherical shell: Applications to the earth's mantle. *Physics of the Earth and Planetary Interiors*, *157*, 223–249.

Stupazzini, M. (2004). *A spectral element approach for 3D dynamic soil-structure interaction problems*. PhD thesis, Politecnico di Milano (Italy).

Stupazzini, M., Paolucci, R., & Igel, H. (2008). Near-fault earthquake ground motion simulation in the Grenoble Valley by a high-performance spectral element code. *Bull. of the Seism. Soc. of America*, –. (submitted).

Tackley, P. J. (2000). Self-consistent generation of tectonic plates in time-dependent, three-dimensional mantle convection simulations, part 1: Pseudoplastic yielding. *Geochem. Geophys. Geosys.*, *1*.

Tackley, P. J., Stevenson, D. J., Glatzmaier, G. A., & Schubert, G. (1993). Effects of an endothermic phase transition at 670 km depth on a spherical model of convection in Earth's mantle. *Nature*, *361*, 699–704.

Tarantola, A. (1984). 3-dimensional inversion without blocks. *Geophysical Journal of the Royal Astronomical Society*, *76*, 299–306.

Tarantola, A. (1988). Theoretical background for the inversion of seismic waveforms, including elasticity and attenuation. *Pure and Applied Geophysics*, *128*, 365–399.

Tromp, J., Tape, C., & Liu, Q. (2005). Seismic tomography, adjoint methods, time reversal and banana-doughnut kernels. *Geophysical Journal International*, *160*, 195–216.

Trottenberg, U., Oosterlee, C., & Schüller, A. (2001). *Multigrid*. Academic Press. ISBN: 0-12-701070-X.

Verfürth, R. (1984). A Combined Conjugate Gradient-Multigrid Algorithm for the Numerical Solution of the Stokes Problem. *IMA Journal of Numerical Analysis*, *4*, 441–455.

Williamson, D. (1968). Integration of the barotropic vorticity equations on a spherical geodesic grid. *Tellus*, *20*, 642–653.

Woitaszek, M., Cope, J., Oberg, M., & Tufo, H. M. (2005). Shared Parallel Filesystems in Heterogeneous Linux Multi-Cluster Environments. In *Proceedings of the 6th LCI International Conference on Linux Clusters: The HPC Revolution*.

Wunsch, C. (1996). *The Ocean Circulation Inverse Problem*. Cambridge University Press.

Yang, W.-S. & Baumgardner, J. R. (2000). A matrix-dependent transfer multigrid method for strongly variable viscosity infinite Prandtl number thermal convection. *Geophysical and Astrophysical Fluid Dynamics*, *92*(3–4), 151–195.

Zambelli, C. (2006). *Experimental and theoretical analysis of the mechanical behaviour of cohesionless soils under cyclic-dynamic loading*. PhD thesis, Politecnico di Milano (Italy).

Zhong, S., Zuber, M. T., Moresi, L., & Gurnis, M. (2000). Role of temperature-dependent viscosity and surface plates in spherical shell models of mantle convection. *Journal of Geophysical Research*, *105*, 11063–11082.

Zhong, S. J. & Gurnis, M. (1995). Mantle convection with plates and mobile, faulted plate margins. *Science*, *267*, 838–843.

Zhong, S. J., Gurnis, M., & Moresi, L. (1998). The role of faults, nonlinear rheology, and viscosity structure in generating plates from instantaneous mantle flow models. *Journal of Geophysical Research*, *103*, 15255–15268.

Zoback, M. L. (1992). First and second order patterns of stress in the lithosphere: The world stress map project. *Journal of Geophysical Research*, *97*, 11703–11728.

Lebenslauf

Persönliche Daten:

geboren:	04. Oktober 1977 in Oschatz
Staatsangehörigkeit:	deutsch
Familienstand:	ledig

Schulausbildung:

1984 - 1992	Polytechnische Oberschule Dr. Salvador Allende Großweitzschen
1992 - 1994	Martin Luther Gymnasium Hartha Außenstelle Leisnig
1994 - 1996	Martin Luther Gymnasium Hartha *Abschluss:* allgemeine Hochschulreife

Grundwehrdienst:

1996 - 1997	Panzeraufklärungskompanie 370 Weberstedt

Studium:

1997 - 2004	Studium Geophysik an der TU Bergakademie Freiberg *Abschluss:* Diplom Geophysiker (April 2004)
2001 - 2002	Studienarbeit *Thema:* ”Lösung großer linearer Gleichungssysteme in Linux-Clustern mittels MPI und PETSc”
2003 - 2004	Diplomarbeit *Thema:* ”Parallelisierung geoelektrischer Finite-Elemente-Modellrechnungen auf Linux-Clustern”
1998 - 2003	Hilfswissenschaftler am Institut für Geophysik

Berufstätigkeit:

2004 -	Wissenschaftlicher Mitarbeiter am Lehrstuhl für Geophysik