

Grid Computing

Anwendungen

Mirosława Utzka
November '04

Universität Duisburg - Essen
Institut für Informatik und Interaktive Systeme

1. Inhaltsverzeichnis:

1. INHALTSVERZEICHNIS:	2
2. EINLEITUNG	3
2.1 MOTIVATION	3
2.2 WENN KOSTEN EINE ROLLE SPIELEN	4
3. ANWENDUNGEN FÜR GRID COMPUTING	5
3.1 DISTRIBUTED SUPERCOMPUTING	5
3.2 HIGH THROUGHPUT COMPUTING	5
3.3 ON-DEMAND-COMPUTING	6
3.4 DATA INTENSIVE COMPUTING.....	6
3.5 COLLABORATIVE COMPUTING.....	6
4. BEISPIELE FÜR GRIDS	7
4.1 GLOBUS TOOLKIT.....	7
4.2 SETI@HOME CLASSIC	7
4.3 IBM ZETAGRID	7
4.4 BOINC.....	7
5. DIE BOINC GRID PLATTFORM	8
5.1 STRUKTUR EINES BOINC PROJEKTES	8
5.1.1 Die Serverseite.....	10
5.1.2 Die Clientseite	10
5.2 BOINC PROJEKTE.....	11
5.2.1 Predictor@home	11
5.2.2 ClimatePrediction.net.....	11
5.2.2 AstroPulse	12
5.2.3 SETI@home II.....	12
5.2.4 Folding@home	13
5.2.5 LHC@home	13
6. SETI@HOME	13
6.1 DER ABLAUF	14
6.1.1 Datengewinnung - Arecibo Radio Observatory.....	14
6.1.2 Welchen Daten erhält man?	15
6.1.3 Wonach wird gesucht?.....	15
6.1.4 Datenintegrität und Störungsbeseitigung	16
6.1.5 Warum ist Persistenz so wichtig?	16
7. AUSBLICK	17
DIE NÄCHSTE GENERATION DES INTERNETS IST DAS GRID.....	17
7. QUELLEN	18

2. Einleitung

Die nächste Stufe des World Wide Web heißt Grid. Doch was ist das eigentlich?

Die Idee des Grids ist es eine zusätzliche allgemein verfügbare Ressource zu schaffen, ähnlich dem Strom. Demnach soll es später möglich sein, seine Rechenleistung oder aber auch die Speicherkapazität wie Strom über eine Steckdose zu beziehen. Dann ist es egal, wo die Ressource herkommt oder wie sie an die richtige Stelle gelangt, es ist nur noch notwendig eine einheitliche Schnittstelle zu besitzen. Am Ende des Monats wird dann zum Beispiel eine Rechnung über die gerechneten CPU-Cycles bzw. über die verbrauchte Speicher Kapazität ausgestellt. Zurzeit ist die Entwicklung im Grid Bereich aber soweit, wie anfangs des 20. Jahrhunderts bezüglich des Stroms. Jeder, der Strom benötigte, musste einen eigenen Stromgenerator besitzen. Mit der Entwicklung des Stromnetzes, auch Power-Grid genannt, wurde ein großer Fortschritt erzielt. Jeder konnte seinen Strom auch ohne eigenen Generator beziehen und sollte man überschüssige Leistung haben, so konnte man diese an andere verkaufen. Aktuell besitzt jeder, der für eine Berechnung einen Computer braucht, einen eigenen Rechner oder Großrechner abhängig von der Aufgabenstellung. Man hofft aber, dass das Grid einen ähnlichen Fortschritt mit seinen Ressourcen erzielen wird, wie damals das Stromnetz mit dem Strom.

2.1 Motivation

Woraus setzt sich aber so ein Grid zusammen?

Zu einem Grid gehören die Computer, die die Berechnungen durchführen sollen, sowie die Software zum Betreiben dieser Computer - so genannte Middleware. Des weiteren Tools und die Software für unsere Anwendungen, das Netzwerk sowie die Endgeräte zu Hause und unterwegs.

Die Software auf diesen Endgeräten nennen man auch Portals, weil sie uns wie ein "Tor" den Zugang zu dieser Internet-Infrastruktur und zu unseren Anwendungen eröffnet. In den nächsten Kapiteln werde ich die verschiedenen Anwendungsbereich von Grid Computing vorstellen und einige Beispiele zu diesen nennen. Danach möchte ich eine Anwendung genauer vorstellen die es ermöglicht gridfähige Projekte einem breiten User Publikum zugänglich zu machen, es handelt sich dabei um die BOINC Plattform. Dabei werde ich auch eins dieser Projekte genauer vorstellen.

2.2 Wenn Kosten eine Rolle spielen

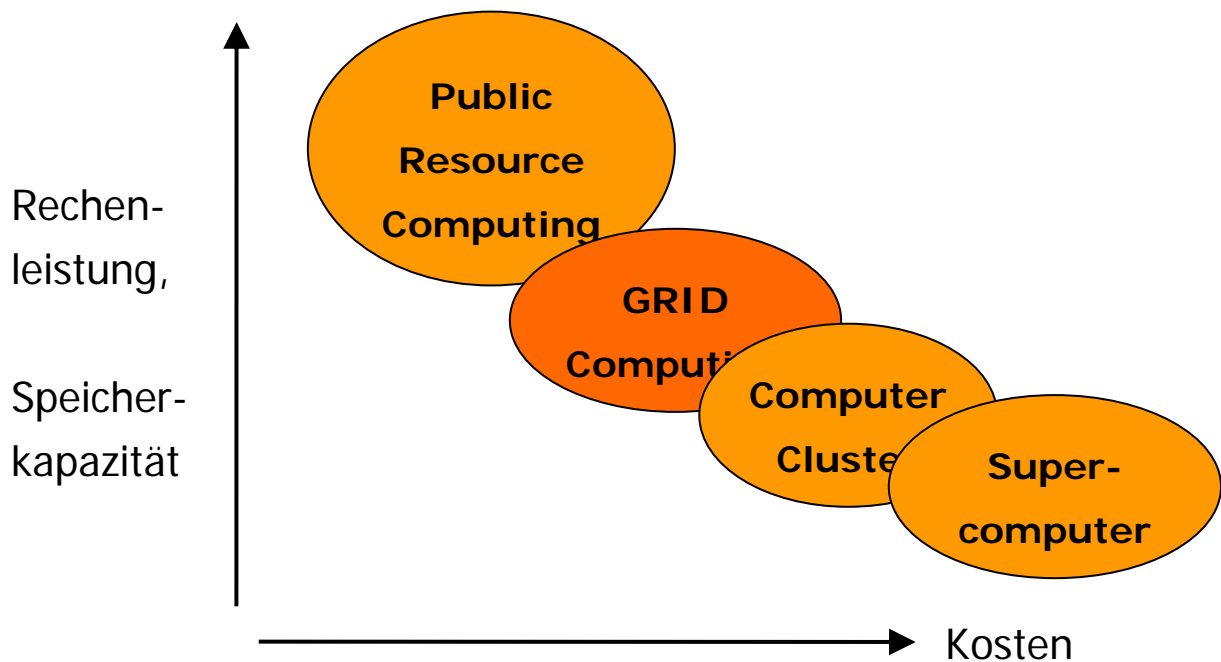


Abb. 1 Preis- Leistungsverhältnis

Anhand der Abb. 1 kann man sich die Beziehung zwischen Rechenleistung/ Speicherkapazität und den Kosten verdeutlichen. So sehen wir dass, Public Ressource Computing im Vergleich zu den anderen Möglichkeiten wie Grid Computing, Computer Cluster und Super Computer das Beste Preis- Leistungsverhältnis hat. Dennoch gibt es diverse Befürchtungen bei Public Ressource Computing die diese Alternative ausschließen. So kann es beispielsweise passieren, dass ein wachsender Teil der privaten CPU Kapazität künftig nicht mehr frei nutzbar sein wird. Das könnte geschehen wenn z.B. Multimedia Geräte als Black Boxes für Video, Games, E-Mail gehandhabt werden oder wenn es Nutzungsbeschränkungen durch Gesetze und Hardwarekontrolle (Software Patente, usw.) geben wird. Eine andere Sorge ist, dass externe User Viren oder Trojaner einschleusen könnten oder gar Informationen stehlen würden.

Trotz des guten Preis- Leistungsverhältnisses des Public Ressource Computings sprechen die Nachteile doch für die sicherere Alternative das Grid Computing.

3. Anwendungen für Grid Computing

3.1 Distributed Supercomputing

Distributed Supercomputing steht für den Zusammenschluss verschiedener Supercomputer, diese können geographisch weit voneinander entfernt sein, zu einem einzelnen, virtuellen Supercomputer. Die Ziele, die dabei verfolgt werden, sind zum einen die Nutzung von gemeinsamen Ressourcen, um größere Probleme lösen zu können und natürlich auch die daraus resultierende Senkung der Hardwarekosten. „Distributed Supercomputing“ beschränkt sich im Vergleich zu „High Throughput Computing“ (s.u.) lediglich auf einen Verbund von Supercomputern, die über Hochgeschwindigkeitsverbindungen gekoppelt sind.

Hierbei steht ganz besonders die Frage im Vordergrund, wie viel Rechenoperationen pro Sekunde erreicht werden können. Dies hat mit der Art der Anwendungen zu tun, die für diese Art der parallelen Verarbeitung besonders geeignet sind.

Als Beispiel möchte ich an dieser Stelle das von CERN ca. 2007 startend LHC Projekt (Large Hadron Collider) nennen. Hier werden riesige Datenmengen, ca. 15 Petabyte pro Jahr, verarbeitet werden müssen. Um dies Datenflut in einer angemessenen Zeit bewältigen zu können wird ein großes Hochgeschwindigkeitsnetzwerk gebaut welches die Weltweit verteilten Super Computer verbinden soll, um diese dann mit den angefallenen Daten zu versorgen.

3.2 High Throughput Computing

Einen anderen Ansatz verfolgt das High Throughput Computing (HTC) Hier ist der Durchsatz der möglichen Berechnungen von besonderer Bedeutung. So gibt es Anwendungen, meist in Wissenschaft und Forschung, bei denen die zu lösenden Probleme von ihrer Komplexität selbst für Supercomputer unmöglich in einer angemessenen Zeit zu lösen sind.

Deshalb ist hier die Frage wie viele FLOPS (Floating Points Per Second) das Rechnersystem pro Sekunde erreicht, nicht so entscheidend sondern vielmehr, wie viele FLOPS pro Monat oder Jahr erreicht werden können, um überhaupt jemals zu einer Lösung des Problems zu gelangen.

Beim High Throughput Computing ist jede auch noch so kleine Rechenleistung willkommen. Deshalb sind bei vielen Projekten, wie z.B. bei den sehr bekannten SETI@home (Search for ExtraTerrestrial Intelligence) auch Privatrechner über das Internet mit dem Rechenzentrum verbunden und stellen somit auch ihre Rechenkapazität zur Verfügung. Privatrechner sind natürlich nicht primär für die Lösung der HTC-Aufgaben vorgesehen, sondern stellen lediglich freie bzw. unbenutzte Ressourcen auf freiwilliger Basis bereit.

Neben der Akzeptanz der Eigentümer stellt dies auch besondere Anforderungen an die Aufteilung der Aufgaben und Zusammenführung der Teillösungen. Hierbei sind besonders Probleme geeignet, die sich gut in voneinander unabhängige Teilprobleme aufteilen lassen.

3.3 On-Demand-Computing

On-Demand-Computing bedeutet übersetzt "Rechenleistung auf Bedarf". Der Begriff des On-Demand-Computing, nicht zuletzt in Verbindung mit Grid-Computing, wurde durch die Assoziation mit dem elektrischen Stromnetz (Power Grid) geprägt.

Die Motivation ist, dass Rechenkapazität in gleicher Weise universell und transparent zur Verfügung gestellt wird, ähnlich wie die elektrische Energie. Das heißt, es handelt sich im Kern um eine Dienstleistung die zum Beispiel von Unternehmen wie IBM oder HP angeboten wird. Diese Dienstleistung können dann Unternehmen nutzen, die durch veränderte Umstände sehr kurzfristig einen stark erhöhten Bedarf an Rechen- oder Speicherkapazität haben.

Dadurch, dass man keine teure Hardware kaufen muss, sondern diese im Bedarfsfall vergleichsweise günstig mieten kann ist On-Demand-Computing eine gute und preisgünstige Alternative.

3.4 Data Intensive Computing

Data Intensive Computing wird vorzugsweise in großen digitalen Büchereien und Datenbanken verwendet, man konzentriert sich dabei auf die Synthetisierung von neuen Informationen. Die Anwendungen dabei haben das Ziel „Die Nadel im Heuhaufen“ zu finden. Man verwendet hierbei auch die Ansätze aus der KI, welche in dem sog. Data Mining Verfahren verwendet werden.

Einer der Einsatzgebiete für Data Intensive Computing ist die Hochenergie Physik. Hier müssen große Datenmengen gespeichert werden, möglicherweise mehrere Petabyte pro Jahr, und nach „besonderen“ Vorkommen untersucht werden.

Weiteres Einsatzgebiete wäre die Wettervorhersage in modernen meteorologischen Systemen oder in der medizinischen Forschung.

3.5 Collaborative Computing

Die Mensch-zu-Mensch-Interaktion steht im Vordergrund bei Collaborative Computing. Dabei können Arbeitsgruppen in Forschung und Wirtschaft gemeinsam an Problemen arbeiten. Teures Equipment kann durch „Collaborative Computing“ geteilt und Kosten gesenkt werden. Echtzeit-Interaktion und virtuelle Realität sind der Hauptaugenmerk bei diesen Anwendungen. Ein Beispiel für Collaborative Computing ist das VRGeo Projekt. Hier werden geologische Informationen visualisiert diese können dann von verschiedenen Teams Weltweit in Echtzeit bearbeitet werden.

4. Beispiele für Grids

Es existiert eine sehr große Anzahl verschiedener Anwendungen im Bereich des Grid Computings, deshalb möchte ich an dieser Stelle nur einige nennen, die zu den bekanntesten auf diesem Gebiet gehören.

4.1 Globus Toolkit

Globus ist ein Ansatz für eine Grid Software Infrastruktur. Mit der Hilfe von Globus soll das Grid möglichst effizient und einfach gesteuert werden. Es ist eine Sammlung von Diensten, deren einzelne Teile nach und nach in ein System integriert werden können. Globus Toolkit findet Einsatz in verschiedenen Grid Projekten wie z.B. dem European Datagrid oder dem NASA Power Grid. Unterstützt wird Globus von vielen Firmen und Institutionen wie IBM und Microsoft.

4.2 SETI@home Classic

Das an der Universität von Kalifornien, Berkeley, seit Mai 1999 durchgeführte ursprüngliche SETI@home Projekt, ist der größte und leistungsfähigste Rechnerverbund der jemals existiert hat.

Seti@home fasste erstmals die, ansonsten ungenutzte, Leistung von vielen hunderttausend Computern über das Internet zusammen und forschte nach „besonderen“ Signalen aus dem Weltall.

4.3 IBM ZetaGRID

Das ZetaGRID ist ebenfalls eine Grid-Software, die unbenutzte CPU-Kapazität von einzelnen Computern für verteiltes Rechnen über das Internet nutzt. Das Projekt welches den Beweis der Riemannschen Hypothese mit dieser Software anstrebt, hat bis jetzt über 11 000 Computer weltweit vernetzt.

4.4 BOINC

Es ist eine noch recht neue Plattform für das Grid Computing, die aber aufgrund ihrer Herkunft ein großes Potenzial bietet. Sie ist aus dem Seti@home Classic Projekt entstanden und da dieses mit über 5 Millionen User das größte seiner Art ist hofft man, dass BOINC (Berkeley Open Infrastructure for Network Computing) einen ähnlichen Erfolg haben wird. Da diese Plattform noch so neu ist möchte ich sie im nächsten Kapitel genau vorstellen.

5. Die BOINC Grid Plattform

Die positiven Erfahrungen mit dem seit Mai 1999 laufenden Projekt SETI@home Classic, hat zu dieser Weiterentwicklung geführt. Die positiven Erfahrungen sollten weiter genutzt werden, dabei aber die Nachteile vermieden werden.

Ein Nachteil der Software war die Verknüpfung zwischen dem Projekt und der Infrastruktur, die bei SETI@home bisher nur für einen Zweck - die Suche nach Signalen aus dem Weltraum - genutzt werden konnte.

Im Juni 2004 wurde mit der Ablösung des bisherigen Clients begonnen.

Die neue Infrastruktur wurde zweistufig aufgebaut. Die erste Stufe übernimmt die Technik für Verwaltung und Verteilung von Daten und Programmen, die Teilnehmerverwaltung und die Statistiken usw.

Auf dieser Plattform können wissenschaftliche Projekte wie SETI@home oder Astropulse aus Berkeley, aber auch beliebige andere Projekte, wie z.B.

Predictor@home oder ClimatePrediction.net aufsetzen. Es können mehrere Projekte gleichzeitig verwaltet werden, wobei z.B. auch auf Client-Seite die vorhandene Kapazität den Einzelprojekten prozentual zugeordnet werden kann.

5.1 Struktur eines BOINC Projektes

Da die einzelnen Projekte getrennt voneinander verwaltet werden können, muss jeder Auftraggeber für ein Projekt eine eigene Serverinfrastruktur bereitstellen, wobei redundante Server zur Lastverteilung und Erhöhung der Verfügbarkeit möglich sind. Projektteilnehmer müssen auf ihrem heimischen PC das BOINC Clientprogramm installieren und können sich dann für Projekte anmelden. Zu jedem Projekt erhält ein Teilnehmer eine eigene Account ID.

Die Teilnehmer können generelle Einstellungen an der Client-Software vornehmen z.B. Dinge wie die Pufferung von Work Units, die Nutzung von Festplattenplatz und Netzwerkbandbreite, sowie die Aufteilung der Ressourcen unter den Projekten. Die Accountinformationen werden jeweils auf dem Server gespeichert, bei dem der Teilnehmer eine Account ID hat. Dadurch können innerhalb des BOINC Netzwerkes Konfigurationen zwischen den Projekten ausgetauscht werden, so dass allgemeine Einstellungen zu den Projekten an jedem Server möglich sind.

Die einzelnen Projekte sind unabhängig voneinander und haben ihre eigenen Anwendungsprogramme, Datenbanken und Server. Dadurch sind sie nicht vom Status anderer Projekte betroffen. Jedes Projekt wird über eine sogenannte Master URL identifiziert die auf ein XHTML Dokument verweist, welches das Projekt beschreibt.

Man kann das BOINC Projekt technisch in eine Server- und eine Client-Seite aufteilen. Die einzelnen Komponenten lassen sich dann wiederum entweder dem BOINC Framework zuzuordnen, oder dem wissenschaftlichen Projekt selbst.

Anhand der folgenden Grafik möchte ich die Funktionsweise des BOINC Frameworks erläutern.

Hierbei sind die projektspezifische Komponenten rot und BOINC Bestandteile blau dargestellt.

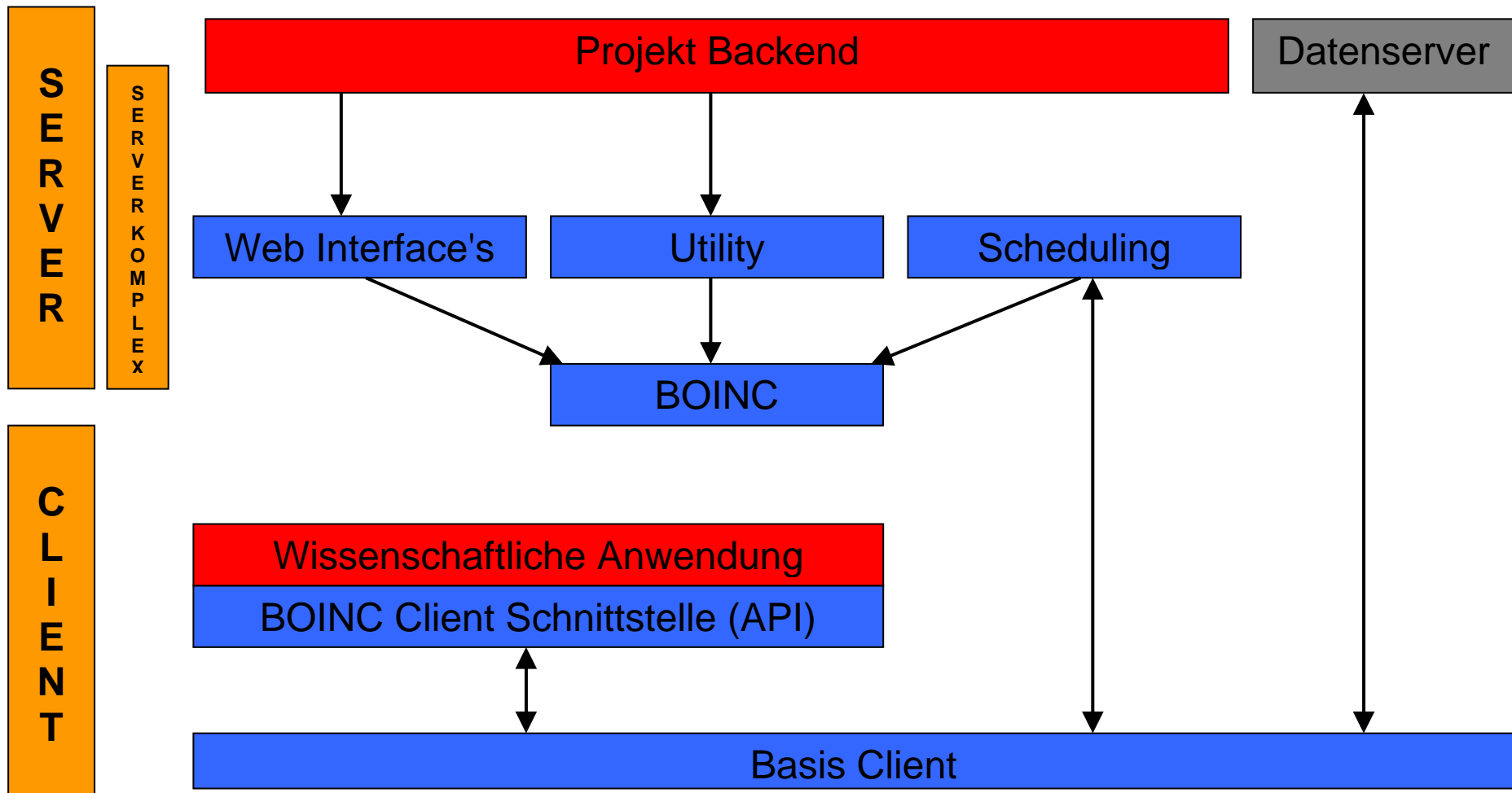


Abb. 2 BOINC Struktur

5.1.2 Die Serverseite

Projekt Backend

Das Projekt Backend stellt die Anwendungen und Arbeitspakete (Work Units) bereit und prüft die Korrektheit der Ergebnisse und verwaltet die berechneten Ergebnisse.

Daten Server

Die Verteilung der Arbeitspakete erfolgt über Daten Server, die nicht zur Projektinfrastruktur selber gehören müssen. Diese können auch von anderen Organisationen, an anderen Standorten betrieben werden und damit Bandbreitenengpässe vermeiden.

BOINC Server Komplex

Scheduling Server

Ein oder mehrere Scheduling Server sorgen für die Kommunikation mit den Clients.

BOINC Datenbank

Die BOINC Datenbank speichert als relationale MySQL Datenbank, die Informationen über Arbeit, Ergebnisse und Teilnehmer.

Web Interface's

Die Web Interface's bilden die -mit Internetbrowser bedienbare- Benutzeroberfläche, getrennt für Teilnehmer und Entwickler

Utility Programme

Die Utility Programme (Dienstprogramme) sind die Schnittstelle vom Server Komplex zum Projektbackend.

5.1.2 Die Clientseite

Basis Client

Der Basis Client ist das Basisprogramm auf dem Rechner des Teilnehmers. Er kommuniziert mit dem Scheduling Server, holt die Work Units ab, liefert die Ergebnisse zurück und puffert einen Arbeits- und Ergebnisvorrat in einem lokalen Cache.

Jeder Client Rechner bekommt beim ersten Anmelden an einem BOINC Server eine eindeutige ID übertragen. Unter dieser ID werden alle Verbindungen zwischen Client und BOINC Server nummeriert und in der BOINC Datenbank und auf dem lokalen Rechner gespeichert.

Der Basis Client ermittelt und verwaltet auch Informationen wie Anzahl und Typ der CPU's, Plattenplatz, Arbeitsspeicher u.a. Diese werden in der BOINC Datenbank gespeichert und dienen zur Entscheidung, ob einem Client eine bestimmte Arbeitsaufgabe übertragen wird.

Ein Basis Client kann mehrere Wissenschaftliche Anwendungen unterstützen und verwalten.

Client API

Über die BOINC API und Grafik API kommunizieren der Basis Client und die eigentliche Wissenschaftliche Anwendung miteinander. Der Basis Client liefert eine grafische Fortschrittanzeige für jedes Projekt. Die Projektanwendung kann zusätzlich aufwändigere Grafikausgaben bis hin zu Bildschirmschonern erzeugen. Für die Grafikausgabe wird OpenGL, für die Kommunikation XML verwendet.

Wissenschaftliche Anwendung

Eine Wissenschaftliche Anwendung repräsentiert eine spezielle verteilte Berechnung und besteht aus einem Programm (möglicherweise in verschiedenen Versionen für unterschiedliche Plattformen), Work Units und Ergebnissen. Ein Projekt kann mehrere Anwendungen umfassen. Diese Anwendungen erledigen die eigentliche wissenschaftliche Arbeit. Bei Seti@home ist es zum Beispiel die „Schnelle Fourier-Transformation“ (FFT)

Wissenschaftliche Anwendungen müssen nicht als Open Source entwickelt werden.

5.2 BOINC Projekte

5.2.1 Predictor@home

Das Predictor@home Projekt am Scripps Instiut in Kalifornien/USA erprobt neue Algorithmen und Methoden zur Vorhersage von Proteinstrukturen. Mit einem anderen Ansatz, aber auf einem ähnlichen Gebiet tätig wie das Folding@home Projekt, betreibt das Team um Professor Charles L. Brooks III Grundlagenforschung auf dem Gebiet der Molekularbiologie und Bioinformatik. Ziel beider Projekte ist es sich ergänzend das Wissen über Proteine, ihre Funktionen und ihre Struktur zu vergrößern und damit die Möglichkeiten zu Bekämpfung von Krankheiten zu verbessern, die mit Proteinen in Verbindung stehen.

Das Predictor@home Projekt findet im Umfeld des CASP (Critical Assessment of Techniques for Protein Structure Prediction) Experimentes statt. Das ist ein, im 2-Jahres-Zyklus stattfindender, internationaler Forscherwettbewerb, der versucht den Fortschritt in der Strukturvorhersage zu ermitteln. Es werden bereits gelöste, aber noch unveröffentlichte Proteinstrukturen verwendet und deren ursprüngliche Sequenzen als Aufgabe ins Internet gestellt. Die eingereichten Vorhersagen der teilnehmenden Wissenschaftler werden von einer unabhängigen Jury bewertet und die Ergebnisse gegen Ende des ersten Jahres auf eine Konferenz den Teilnehmern bekannt gegeben und im folgenden Jahres veröffentlicht.

5.2.2 ClimatePrediction.net

Das climateprediction.net Projekt der University of Oxford, der Rutherford Appleton Laboratorien und der Open University will langfristige Vorhersagen über die klimatischen Entwicklungen im 21ten Jahrhundert erarbeiten

Im wesentlichen lösen Klimamodelle die selben Gleichungen wie Wettervorhersagen, plus vieler zusätzlicher, um langsame Veränderungen in der Entwicklung des Ozeans und des See-Eises vorherzusagen.

Da climateprediction.net Vorhersagen über 50 bis 100 Jahre machen will, statt über ein paar Tage oder Woche, können sie nicht versuchen so detailliert zu arbeiten wie

die Wettervorhersagen. Mehr als eine Handvoll Langzeitsimulationen wäre selbst für den größten Supercomputer zuviel.

Statt dessen hat man den Ansatz gewählt, ein state-of-the-art Klimamodell sehr gründlich zu testen, in dem ein großes Ensemble von Modellen simultan auf tausenden von PC's läuft. Jedes Modell in einem Ensemble wird eine geringfügig abweichende Physik haben um die Unsicherheit in klimatischen Schlüsselprozessen darzustellen.

Jedes Modell wird auf einem PC am Stück durchgerechnet, also nicht in Arbeitspakete zerlegt. Deshalb ist die Systemanforderungen für dieses Projekt für BOINC ein Novum. Das ClimatePrediction.net Projekt benötigt rund 60 MB Arbeitsspeicher und 600 MB freien Plattenplatz für ein Klimamodell. Auf einem Pentium III mit 1.400 MHz können durchaus mehr als 900 Stunden CPU-Zeit für die Berechnung eines Modells erforderlich sein.

Damit erreicht dieses Distributet Computing Projekt sicherlich die Grenze des Möglichen, doch löst die rasante Weiterentwicklung der Hardware vermutlich dieses Problem von allein.

5.2.2 AstroPulse

Astropulse nutzt die bereits vorhanden SETI@home I Daten des Radioteleskopes in Arecibo nach einer anderen (nur eine Mikrosekunde anhaltenden) Sorte von Radio Signal Impulsen durchsuchen.

Dieser Signal Typ unterscheidet sich von denen, die mit SETI@home I gesucht werden. Es handelt sich um sehr schnelle Impulse. Es werden die vollen 2,5 MHz Bandbreite benötigt, während SETI@home I dieses Frequenzband in 256 einzelne 10kHz Sub-Bänder unterteilt hat.

Streuereffekte bei diesen Impulsen, wenn sie das interstellare Medium (das dünne Gas, das den Raum zwischen den Sternen in unserer Galaxie füllt) durchqueren, muss man mit aufwendigen Algorithmen korrigieren, was sehr rechenintensiv ist. Also eine typische Aufgabe für Distributed Computing.

5.2.3 SETI@home II

Das Southern Hemisphere Search Projekt (SETI@home II) soll, das an der Universität von Kalifornien, Berkeley seit Mai 1999 durchgeführte, SETI@home Projekt um die Observation des südlichen Sternenhimmels erweitern.

Mit BOINC Infrastruktur werden auch in diesem Distributed Computing Projekt die von einem Radioteleskop - diesmal Parkes in Australien - aufgezeichneten Daten, in handliche Päckchen zerlegt, an die heimischen Computer in aller Welt verteilt und von diesen dann nach Signalen außerirdischen Lebens durchsucht.

Parkes ist das größte einzelstehende Radioteleskop auf der Südhalbkugel und kann den kompletten südlichen Sternenhimmel überwachen. Es wird bereits seit 1998 von Southern SERENDIP für die SETI Forschung genutzt.

Die meisten S.E.T.I. Projekte haben bisher den nördlichen Himmel abgesucht. Aus diesem Grund und weil die Sternkonstellationen der südlichen Hemisphäre, nach Ansicht vieler Wissenschaftler, bessere Erfolgsaussichten bieten, ist es für das Team in Berkeley eine großartige Möglichkeit, dass SETI Australien bei Southern Hemisphere Search mit ihnen kooperiert.

5.2.4 Folding@home

Folding@home arbeitet mit einem ähnlichen Aufbau wie SETI@home I. Die sonst brachliegende Leistung vieler Computer wird genutzt, um zu simulieren, wie sich bestimmte Proteine falten, das heißt wie sie sich räumlich anordnen.

Dabei geht es um Grundlagenforschung zum besseren Verständnis dieser Vorgänge. Man erhofft sich davon unter anderem Fortschritte bei der Behandlung von Krankheiten wie Alzheimer, Mukoviszidose und Creutzfeldt-Jakob.

Aminosäureketten falten sich in Millionen Einzelschritten zu einem dreidimensionalen funktionsfähigen Protein. Dabei gibt es eine gewaltige Anzahl von Möglichkeiten und damit auch Fehlerquellen. Der Versuch diese Vorgänge zu simulieren ist extrem rechenintensiv und führt häufig nicht zum gewünschten Erfolg.

Den Wissenschaftlern der Pande Group der chemischen Fakultät der Stanford Universität ist es 2002 mit Folding@home gelungen, die etwa 10 Mikrosekunden lange Faltung eines Proteins abzuleiten und dabei signifikante Übereinstimmung mit experimentell ermittelten Daten zu erzielen. Erstmals hat damit ein Distributed Computing Projekt Anerkennung in der wissenschaftlichen Fachwelt gewonnen. Durch seine strukturelle Ähnlichkeit zu SETI@home I bietet sich Folding@home für eine Umsetzung auf BOINC sehr gut an.

5.2.5 LHC@home

Das LHC@home Projekt des CERN Forschungszentrum in der Schweiz simuliert Teilchen bei ihrer Reise durch den 26,7 Kilometer langen Ring des im Bau befindlichen neuen Teilchenbeschleunigers LHC (Large Hadron Collider) um ihre Flugbahn zu studieren. Die Ergebnisse dienen zur Überprüfung der langfristigen Stabilität der Hochenergiepartikel im LHC und liefern Daten für die Justierung der neu eingebauten Magnete.

Die SixTrack-Anwendung des LHC@home Projektes simuliert 60 Teilchen zur Zeit und lässt sie den Ring 100.000 mal umrunden. Das wären weniger als 10 Sekunden in der wirklichen Welt. Es reicht aber um zu erkennen, ob der Teilchenstrahl in einem stabilen Orbit verbleibt oder in die Wand der Vakuumröhre einschlägt. Eine solche Kursabweichung könnte ernsthafte Folgen haben, bis hin zur erzwungenen Abschaltung der Anlage für Reparaturen.

6. Seti@home

Nachdem ich im vorherigen Kapitel grob einige Projekte der BOINC Plattform vorgestellt habe, möchte ich nun spezieller auf die Funktionsweise von Seti@home (BOINC) eingehen um den Ablauf der Datenverarbeitung zu erläutern. Den Ablauf werde ich von der Datengewinnung über Daten Auswertung/Berechnung bis hin zu Datenzusammenführung versuchen anschaulich darzustellen, siehe Abb.3

Kurze Beschreibung des Ablaufs:

Nachdem man den BOINC Client geladen und installiert hat muss man sich bei BOINC registrieren. Über die Daten- und Scheduling- Server bekommt man dann ca. 340 kB große, zu analysierende Datenpakete, die so genannte Work Unit, zugeschickt, welches im Anschluss ausgewertet werden. Man muss nicht, die ganze Zeit online sein, während der Auswertung baut BOINC keine Verbindung zu seinem Server auf.

Immer wenn der Rechner ungenutzte Ressourcen hat läuft die Berechnung im Hintergrund bzw. wird als Bildschirmschoner dargestellt. Die Bildschirmschoner Funktion ist ganz wichtig bei der Motivation der User. So hätte Seti@home Classic nie so viele User motivieren können, wenn es nicht den optisch interessanten Bildschirmschoner und die Rang Listen gehabt hätte.

Ist ein Paket fertig berechnet, meldet sich das Programm, man nimmt wieder kurz Kontakt mit dem BOINC-Server auf und das Programm sendet automatisch das Ergebnis und lädt direkt ein neues Paket zur Verarbeitung.

Dieser Ablauf kann auch ohne Zutun des Users im Hintergrund ablaufen.

6.1 Der Ablauf

6.1.1 Datengewinnung - Arecibo Radio Observatory

Die Daten die man bei Seti@home verwendet, kommen von dem größten Radioteleskop der Welt dem Arecibo Radio Observatory

Der Spiegel ist in den Krater eines inaktiven Vulkans eingelassen und er besteht aus ca. 14.000 perforierten Aluminium-Panelen, von denen jede etwa eine Fläche von 6 Quadratmetern hat. Die Höhe über dem Meeresspiegel beträgt 497 Meter

Die Empfänger des Teleskops sind hochmodern und sehr empfindlich. Sie arbeiten in einem Bad aus flüssigem Helium, um die Temperatur und damit das temperaturbedingte Eigenrauschen der Empfänger möglichst gering zu halten.

Das ist nötig, da die zu empfangenden Signale sehr, sehr schwach sind und nur diese auch verstärkt werden sollen. Sie arbeiten auf Frequenzen zwischen 50 Megahertz und 10 Gigahertz, d.h. zwischen dem 6m- und dem 3cm-Band

Insgesamt 26 Motoren steuern die Plattform. Innerhalb der Kuppel befindet sich ein gesonderter Raum, in dem ein Radarsender mit einer Sendeleistung von 1 Megawatt betrieben wird. Die ausgesendeten Radarsignale werden an Objekten in unserem Sonnensystem reflektiert, und die Echos werden empfangen und ausgewertet. Dadurch kann man z.B. Informationen über die Oberflächenbeschaffenheit oder die dynamischen Eigenschaften dieser Objekte sammeln.

Das Radioteleskop wurde in Puerto Rico erbaut und kann auf Grund seiner Bauart nur den Himmelsbereich des Weltraums beobachten, welcher durch die natürliche Drehung der Erde über dem Teleskop erscheint, d.h. nur 30 Prozent der nördlichen Hemisphäre

Es nimmt Radiowellen auf, die von direkt über ihm kommen. Das "Blickfeld", d.h. der Ausschnitt, den es am Himmel sieht, ist also sehr klein. Der Durchmesser des Mondes würde am Himmel fünfmal so groß erscheinen wie der Ausschnitt, den das Teleskop sieht.

Das SETI@home Programm bekommt pro Übertragung ein Datenpaket mit einer Aufzeichnungslänge von ca. 107 Sekunden. Innerhalb dieser 107 Sekunden bewegt sich der Fleck, den das Teleskop am Himmel sieht, um etwa seinen sechsfachen Durchmesser in Richtung der Erddrehung. Man bekommt also die Daten aus einem kleinen Streifen am Himmel, das 1/5 des Mondes breit und etwas mehr als 1 Monddurchmesser lang ist.

Die Daten werden auf hochdichten Bändern beim Arecibo Teleskop in Puerto Rico gespeichert, ungefähr ein 35 Gbyte Band pro Tag, dann nach Berkeley übersandt, in kleine Einheiten aufgeteilt und über das Internet an Teilnehmer auf der ganzen Welt zur Analyse übermittelt. Da Arecibo nicht über eine ausreichend hohe Internet-Bandbreite verfügt, müssen die Daten auf herkömmliche Weise nach Berkeley verschickt werden.

6.1.2 Welchen Daten erhält man?

Neben der zeitlichen Auswahl (den 107 Sekunden) wird die Datenmenge, die jeder erhält, noch weiter eingeschränkt. Dadurch dass sehr viele Frequenzen gleichzeitig aufgenommen werden müssen sie erst noch weiter unterteilt werden. Das heißt dass ein solches "Frequenzband" weiter in 256 Teile aufgeteilt wird. So ein 107 Sekunden langes 256stel der Daten erhält nun jeder SETI@home User.

6.1.3 Wonach wird gesucht?

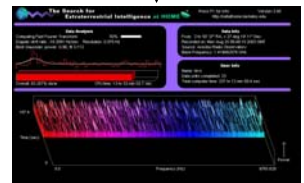
Wenn man irgendwo Signale von intelligenten Lebewesen finden will, muss man sie von solchen unterscheiden, die auch ganz natürlich irgendwo erzeugt werden. Deshalb hört man zum Beispiel einen Frequenzbereich ab, in dem störende, natürliche „Hintergrundgeräusche“ möglichst schwach sind. Dies ist bei Radiofrequenzen um ca. 1.4 GHz der Fall. In diesem Bereich dürfen sich auch keine irdischen Sender aufhalten. Außerdem glaubt man, dass intelligente Lebewesen wahrscheinlich diesen Frequenzbereich zum Versenden ihrer Nachrichten verwenden würden. Erstens ist es hier "relativ leise/dunkel am Himmel", zweitens ist dies in der Nähe einer physikalisch interessanten Frequenz, die vom häufigsten Element des Universums (Wasserstoff) ausgesandt wird. Man hofft also, dass sich jemand dort bemerkbar machen würde, wo jemand anderes (WIR) auch am ehesten hinhören würde. Ob diese Annahme gerechtfertigt ist, muss sich noch herausstellen.

Ablauf

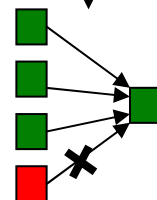
1. Daten Sammeln



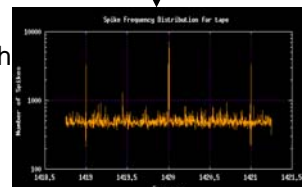
2. Signale (Kandidaten) finden



3. Daten Integrität prüfen



4. Störungen die durch Radiofrequenz-Interferenzen entstehen entfernen



5. Endgültige Signale (Kandidaten) finden – Persistente Signale

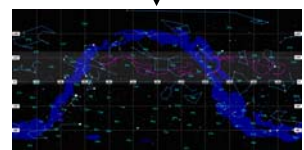


Abb. 3 Ablauf beim Seti@home Projekt

6.1.4 Datenintegrität und Störungsbeseitigung

Nachdem die Daten Pakete auf dem eigenen Rechner ausgewertet wurden, wobei digitale Signale, überwiegend über Fourier Transformationen bei unterschiedlicher Dauer und Chirp-Raten ausgewertet werden, und die Ergebnisse zurück zum BOINC Server übermittelt wurden, muss deren Integrität vor der Weiterverarbeitung geprüft werden.

6.1.5 Warum ist Persistenz so wichtig?

Man erwarten, dass die meisten (aber nicht alle) der aufgezeichneten Radiosignale von der Erde oder von Satelliten stammen bzw. durch Rauschen oder andere natürliche Phänomene verursacht wurden.

Wenn das Radioteleskop zufällig in Richtung Alpha Centauri zeigt und währenddessen Radiosignale empfängt, kann dies z.B. ein Fernseh-Satellit ausstrahlen, der sich gerade über dem Teleskop befindet. Man erkennt aus den Daten nur, dass dieses Signal aufgezeichnet wurde, während die Antenne auf Alpha Centauri ausgerichtet war.

Wenn aber dieses Signal aus Richtung Alpha Centauri erneut bei verschiedenen anderen Gelegenheiten aufgezeichnet wird, sinkt die Wahrscheinlichkeit dafür, dass die Ursache der Signale auf der Erde zu finden ist und es wird wahrscheinlicher, dass die Quelle tatsächlich im Raumbereich um Alpha Centauri herum liegt. Daher ist es wichtig, Signale zu finden, die mehrfach aus derselben Himmelsrichtung empfangen wurden

Die Wichtigkeit persistenter Signale aus einer bestimmten Raumregion wird mit der Zahl der gefundenen Übereinstimmungen immer größer.

Leider muss man sagen, dass bis jetzt keine „interessanten/ besonderen“ Signale gefunden wurden. Aber man gibt die Hoffnung nicht auf. Das nächstes Jahr anlaufende Seti@home II Projekt weckt erneut bei vielen Forschern die Erwartung, doch noch entsprechende Signale zu finden.

7. Ausblick

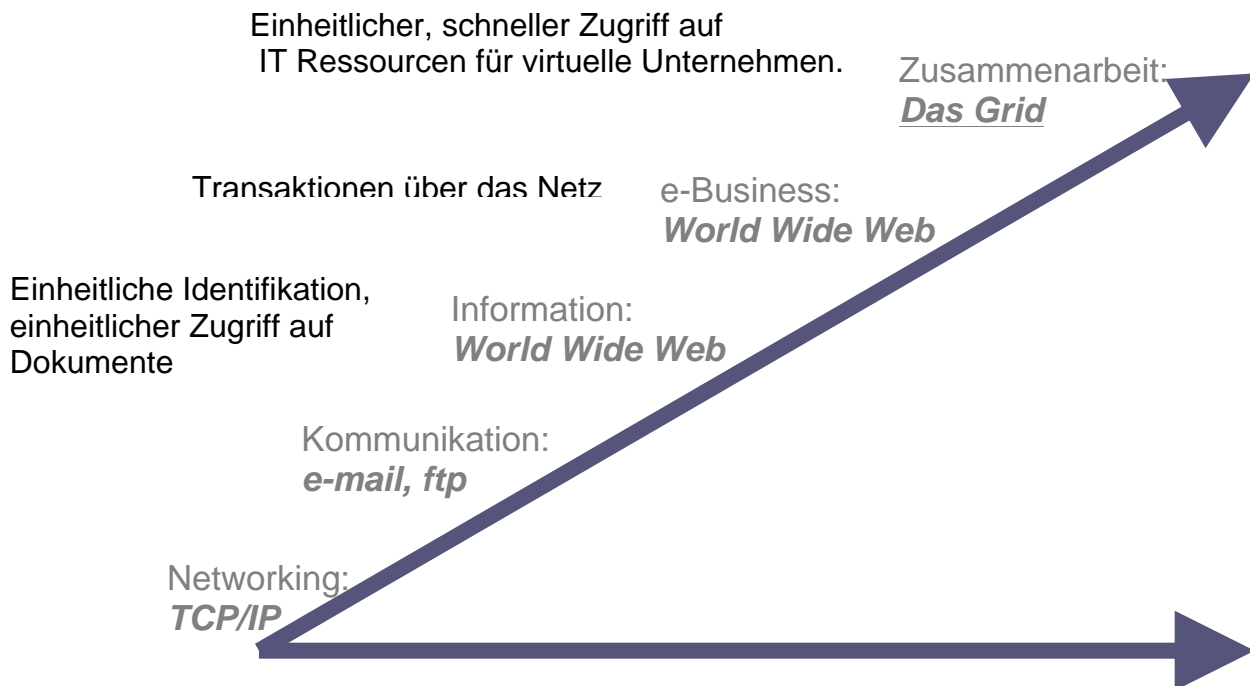


Abb. 4 Netzwerk Entwicklung

Die nächste Generation des Internets ist das Grid

So oder so ähnlich liest man über das Grid Computing in vielen Berichten die sich mit diesem Thema beschäftigen und diesen Eindruck bekommt man auch, wenn man sich die Netzwerkentwicklung in der Abb.4 der letzten Jahre anschaut.

Ob und wann das Grid jemals den Status des WWW erhält, kann man aber schwer vorhersagen.

Andererseits wird das Grid Computing von vielen Firmen massiv unterstützt, so ist es durchaus vorstellbar, dass das Grid irgendwann von uns als so selbstverständlich angesehen wird wie zur Zeit das Internet.

Zu den Firmen die Grid Computing unterstützen gehören zum Beispiel in der Forschung: CERN, NASA, US National Labs usw.

In der IT- Branche sind dies Firmen wie beispielsweise IBM, Microsoft und Sun.

Ebenfalls darf man auch die Regierungen nicht vergessen, welche sehr viel Geld in diesen Forschungsbereich investieren.

Wenn die Grid Technologien mittelfristig in allen Betriebssystemen integriert oder zumindest optional verfügbar sein werden, werden diese also auch eine gute Chance haben, größer und anerkannter zu werden.

Die Grundlagen wurden z.B. durch das Globus Toolkit oder die BOINC Software gelegt. Welcher Grid Software die Zukunft gehören wird, muss sich noch heraus stellen.

Sollte es gelingen die bestehenden Anwendungen „gridfähig“ zu machen, so wie bis jetzt einige wenige technisch- wissenschaftliche Anwendungen, dann steht dem WWG World Wide Grid nichts mehr im Wege. Bis dahin ist es noch ein weiter Weg, der über die Definition der entsprechenden Standards bis hin zu Implementierung in die Betriebssysteme führt.

7. Quellen

Seti@home

<http://setiathome.ssl.berkeley.edu/>

<http://www.setigermany.de>

Genome@home

<http://www.stanford.edu/group/pandegroup/genome/>

BOINC

<http://boinc.berkeley.edu/>

<http://www.boinc.de/>

Grid Computing Recherche auf folgenden Seiten:

<http://www.zdnet.de/>

<http://www.silicon.de/>

<http://www.tecchannel.de/>

<http://www.computerwoche.de/>

<http://www.golem.de>

<http://www.nzz.ch>

<http://www.heise.de/>

<http://www.computerworld.com/>

CERN

<http://www.cern.ch/>

ZetaGrid

<http://www.zetagrid.net/>

Globus Toolkit

<http://www.globus.org/>