Auszug
aus dem

Research Report

Hochleistungsrechner in Hessen

2010/2011

TECHNISCHE UNIVERSITÄT DARMSTADT

GOETHE UNIVERSITÄT

JUSTUS-LIEBIG-UNIVERSITÄT GIESSEN

UNIKASSEL VERSITÄT

Philipps Universität Marburg

Inhalt

Mathematik und Informatik

Biologie, Medizin, Neurowissenschaften

Geowissenschaften

Chemie und Materialwissenschaften

Physik

Impressum

Herausgeber:
Hochschulrechenzentrum der Technischen Universität Darmstadt
im Auftrag des Hessischen Beirats für Hochleistungsrechnen

Redaktion:
Nicole Voß und Dr. Andreas Schönfeld
Hochschulrechenzentrum der Technischen Universität Darmstadt

Texte und Bilder
Wissenschaftlerinnen und Wissenschaftler, die in den Jahren 2010 und 2011
an den Hochleistungsrechnern in Hessen gearbeitet haben.

Gestaltung, Reinzeichnung, Produktion:
Pia Lauck, Dipl.-Designerin – www.desktop-design.de

Vorwort zum Jahresbericht 2010-2011

Liebe Leserinnen, liebe Leser,

die Bedeutung computergestützter Untersuchungen hat in allen Bereichen der Natur- und Ingenieurwissenschaften weiter zugenommen und ist heute ein unverzichtbares Element wissenschaftlichen Arbeitens in der Forschung und auch in der Ausbildung. Die hier von den Arbeitsgruppen an den hessischen Hochschulen zusammengetragenen Forschungsaktivitäten aus den Jahren 2010 bis 2011 belegen dies in eindrucksvoller Weise.

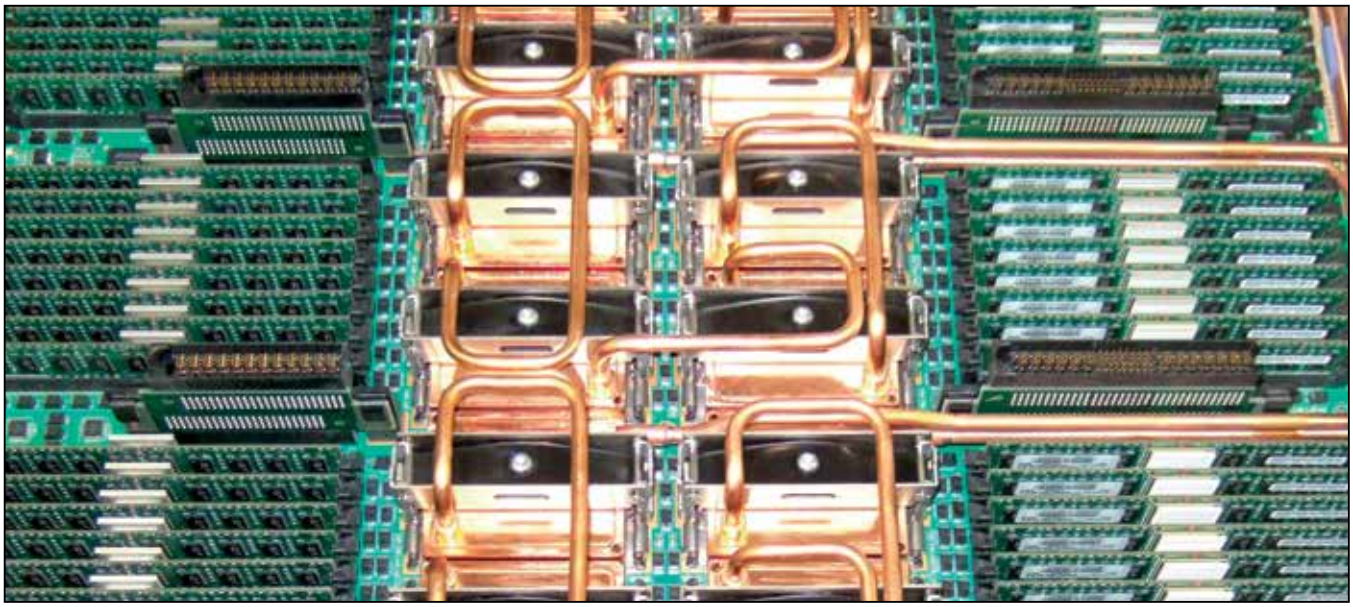
Das Spektrum der Beiträge reicht von subatomaren Stoßprozessen, bei denen Computer nicht nur für die Datenanalyse sondern auch für die Simulation und Interpretation der Messungen herangezogen werden, über die Materialwissenschaften und die Chemie, wo neuartige Substanzen erst im Computer simuliert werden, bevor die teurere, experimentelle Synthese angestrebt wird, bis hin zu den notorisch schwierigen Untersuchungen von Verbrennungen und Turbulenzen.

Es sind die vielen Variablen, die oft nichtlinearen Wechselwirkungen und die breit verteilten Zeitskalen, die einer analytischen Lösung im Wege stehen und die numerische Simulation erfordern. Diesen Anforderungen ist mit schnelleren CPUs alleine nicht beizukommen, es sind auch entsprechend angepasste Algorithmen zu entwickeln. Beides zusammen hat dazu geführt, dass nun in vielen Bereichen eine sehr gute Übereinstimmung zwischen Simulationsergebnissen und experimentellen Beobachtungen erzielt wird. Dieser Erfolg führt zwangsläufig dazu, dass die Anforderungen an Zahl und Komplexität von Untersuchungen in silico weiter zunehmen, nicht nur in der grundlagenorientierten Wissenschaft, sondern auch in den anwendungsnahen Forschungsbereichen und der Industrie. Die Wissenschaftler dürfen sich also auf neue Fragen und neue Herausforderungen freuen.

Die hier dargestellten Forschungsergebnisse dokumentieren zunächst die vielfältigen, international sichtbaren Forschungsleistungen der beteiligten Gruppen. Sie zeigen weiter, welche Möglichkeiten das Hochleistungsrechnen in Hessen bietet und weisen auf die Hürden hin, zu deren Überwindung neue Hard- und Software erforderlich sein wird. Sie sind aber auch ein Ausweis der hohen Qualität der in den Projekten tätigen Mitarbeiterinnen und Mitarbeiter, die über ihre Forschungsbeiträge bestens für eine immer stärker von Computersimulationen dominierte Arbeitswelt vorbereitet werden.

Prof. Dr. Bruno Eckhardt
Vorsitzender des Hessischen Beirates für das Hochleistungsrechnen

Der Hessische Hochleistungsrechner an der Technischen Universität Darmstadt



Systemaufbau

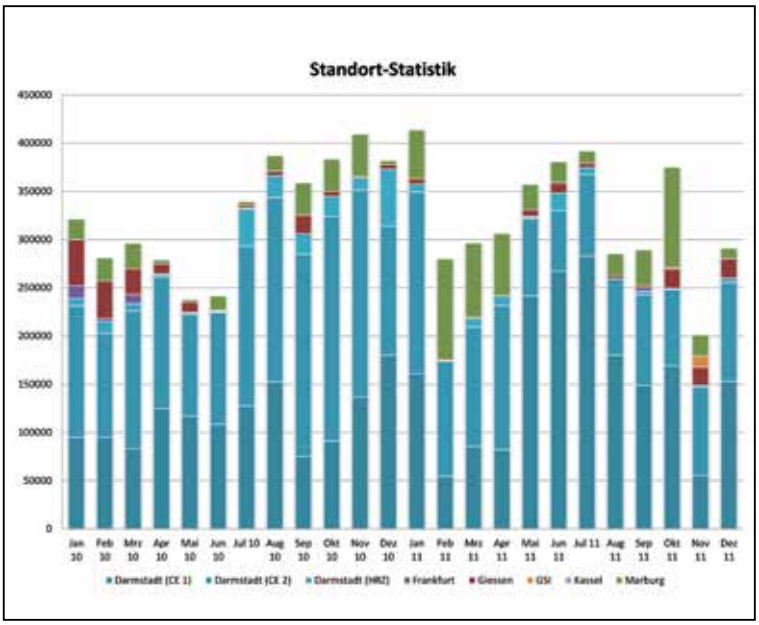
Der Hessische Hochleistungsrechner (HHLR) besteht aus einem Cluster von 18 Rechenknoten mit jeweils 32 Power6-Prozessoren. Die ersten Systeme wurden bereits im November 2008 installiert. Damit gehört der Rechner heute zu den Oldtimern unter den Hochleistungsrechnern. Trotz der nach heutigen Standards eher beschaulichen Gesamtrechenleistung von 11 TFLOP/s (Peak), kann er sich bei kleinen und mittleren Problemstellungen (bis 64 parallele Prozesse) aber noch gut mit modernen Rechnern messen. Die hohe Taktfrequenz von 4,7 GHz und der schnelle Arbeitsspeicher machen es möglich.

Der Rechner wird – nach einigen Verzögerungen im Bau – in der zweiten Jahreshälfte 2012 abgelöst. Anders als in der Vergangenheit werden die vielfältigen Anforderungen dann nicht mehr von einer einzelnen Systemarchitektur abgedeckt. Stattdessen wird es für die wesentlichen Bereiche spezialisierte Systeme geben. Arbeitsspeicherintensive Anwendungen werden zum Beispiel anders behandelt als klassische massiv-parallele Rechenanwendungen.

Auslastung des HHLR

Die mittlere Auslastung des HHLR lag in den Jahren 2010 und 2011 bei etwa 77%. Auffällig ist dabei, dass die Nutzung im Vergleich zu früheren Jahren stark schwankt. So gibt es Monate in den keine 70% der theoretisch verfügbaren Rechenzeit abgerufen wurden, während in anderen Monaten der Rechner mit einer Auslastung von merklich über 90% de facto voll war.

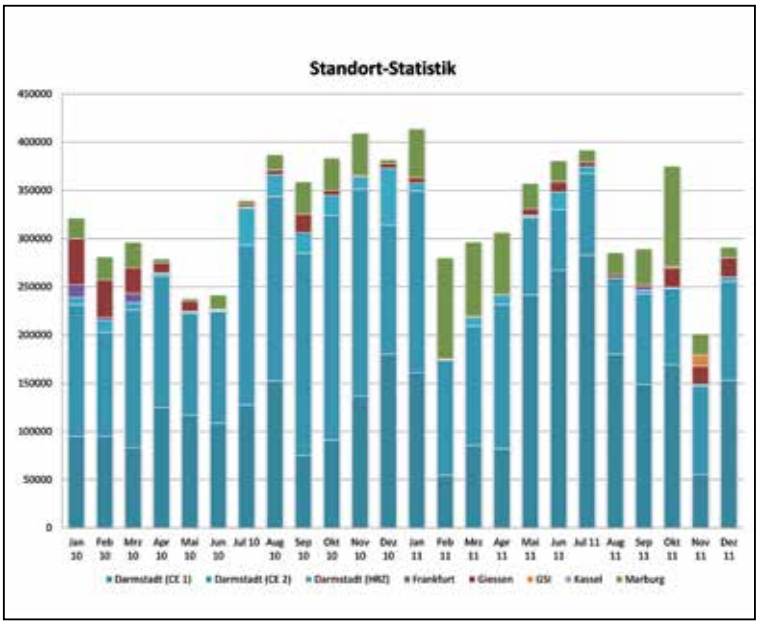
Die Verteilung der Rechenzeit auf die wissenschaftlichen Einrichtungen hat sich im Vergleich zu den Vorjahren erneut verschoben. Der Anteil der TU Darmstadt ist von 71% im Jahr 2009 auf 86% in den Jahren 2010/11 gestiegen. Die Universität Marburg nutzte 2011 14% und die Universität Gießen 3% der Rechenkapazität. Neu in den Nutzerkreis gekommen ist das GSI Helmholzzentrum für Schwerionenforschung. In der Gesamtnutzung des Rechners fällt sie aber mit weniger als 1% (wie auch die Universitäten Kassel und Frankfurt) nicht ins Gewicht.



Parallelität der Jobs

Betrachtet man die Entwicklung der Jobgrößen (Parallelität) so fällt auf, dass zusehends mehr Rechenzeit von Jobs mit mehr als 128 Prozessen (grün) genutzt wird. Trotz dieser Entwicklung hin zu größeren Problemstellungen wird nach wie vor die Hälfte der Rechenzeit von Jobs verwendet, die innerhalb eines Knotens (32 Prozessoren) bearbeitet werden können.

Dies ist auf die Architektur des HHLR zurückzuführen, da Jobs dieser Größe besonders effizient bearbeitet werden.



Die Hochleistungsrechner (HLR) am Center for Scientific Computing (CSC) der Goethe-Universität

Der Hochleistungsrechner LOEWE-CSC
In Frankfurt läuft seit November 2010 der LOEWE-CSC, einer der Energie effizientesten Großcomputer Europas. Seine Rechenleistung von 299 TFlop/s macht ihn zum derzeit drittschnellsten Supercomputer Deutschlands. Mit 740 MFlop/s pro Watt verbraucht der LOEWE-CSC nur etwa ein Viertel der Energie vergleichbar schneller Computer, zu Investitionskosten, die mit knapp fünf Millionen Euro bei etwa einem Drittel liegen.

Der Frankfurter Rechner ist eine Eigenentwicklung der Goethe-Universität, des Center for Scientific Computing (CSC), des Frankfurt Institute for Advanced Studies (FIAS) und des Helmholtz International Center for FAIR (HIC for FAIR). Das System umfasst 832 Rechenknoten mit 20.928 AMD-Magny-Cours-Kernen sowie 778 GPU (AMD Radeon HD 5870), 56 TB Hauptspeicher und über 2.5PB Festplattenspeicher. Die Rechenknoten sind über ein latenzarmes QDR-Infiniband (40 Gb/s) vernetzt. Der Anteil der Kühlung am Stromverbrauch des Rechners beträgt unter maximaler Last nur 7%. Die Architektur des LOEWE-CSC ist dem sehr heterogenen Anforderungsprofil unterschiedlicher Forschungsprojekte angepasst, denn am LOEWE-CSC arbeiten unter anderem Wissenschaftler aus den Bereichen Physik der elementaren und komplexen Materie, Quantenchemie, Lebenswissenschaften und Klimaforschung.

Den hervorragenden PUE-Koeffizienten von 1,07 verdankt der Rechner einer mit der Firma Knürr gemeinsam entwickelten Wärmetauscher-Technologie: Die Gehäuselüfter blasen die heiße Luft durch den Wärmetauscher in der hinteren Tür. Dadurch brauchen die Racks keine zusätzlichen Ventilatoren. Die Wasserkühlung wurde mit der Firma Infraserv am Standort des Rechners, dem Industriepark Höchst, realisiert. Über zwei Kühltürme wird Flusswasser verdampft und über einen Wärmetauscher das Wasser des inneren Kreislaufs gekühlt.

LOEWE-CSC wurde im Jahr 2011 mit dem GreenIT Best Practice Award in der Kategorie 'Visionäre Gesamtkonzepte' ausgezeichnet. In diesem Jahr gehört LOEWE-CSC zu den Preisträgern im Wettbewerb „365 Orte im Land der Ideen".

Hochleistungsrechner FUCHS und GPU-Cluster Scout
Der Hochleistungsrechner FUCHS ist seit April 2010 in Betrieb. Er wurde aus Mitteln der DFG und des HMWK finanziert. Das System besteht aus 4920 AMD-Istanbul-Kernen mit einer Spitzen-Rechenleistung von 43 TFlop/s, 14 TB Hauptspeicher und 537 TB Massenspeicher. Die Rechenknoten, die bis zu 24 CPU-Kerne und 128 GB Hauptspeicher bieten, sind mit InfiniBand vernetzt.

Der Scout, ein gpGPU-Cluster, enthält neun Recheneinheiten, wobei jede Einheit aus zwei CPU- und drei GPU-Knoten aufgebaut ist. Die CPU-Systeme bestehen aus zwei QuadCore Xeon CPUs mit 16 Gbyte Hauptspeicher. Die GPU-Knoten sind Tesla S1070 Systeme von Nvidia. Jeder GPU-Knoten leistet

4 TFlop/s single precision (sp), bzw. 345 GFlop/s double precision (dp), so dass das Gesamtsystem eine Spitzenleistung von 108 TFlop/s sp, bzw. 9.3 TFlop/s dp erreicht. Ziel des Scout ist es, allen interessierten Arbeitsgruppen Erfahrungen im Einsatz von gpGPU-Systemen zu ermöglichen.

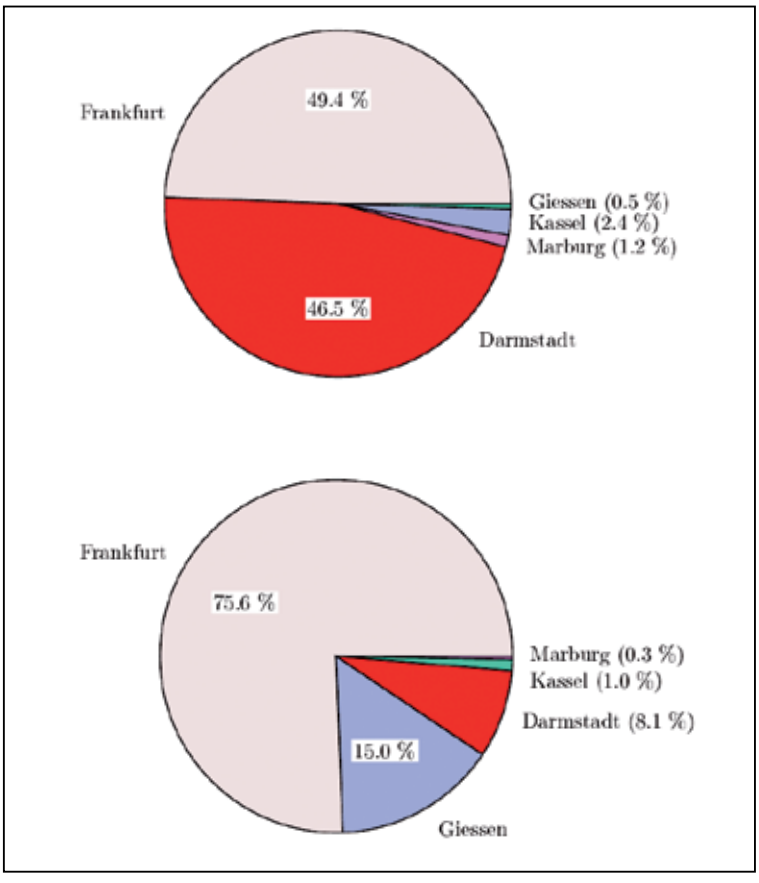
Auslastung der Hochleistungsrechner

Mehr als 650 Forscher aus über 100 Arbeitsgruppen der Naturwissenschaften, der Mathematik und der Informatik nutzen die Hochleistungsrechner des CSC. Die Rechner versorgen einerseits die Frankfurter Wissenschaftler mit Rechenkapazität, stehen über den hessischen HLR-Verbund aber auch den anderen Hochschulen des Landes zur Verfügung. Beide Rechner sind im Jahresmittel zu 85% ausgelastet. Entsprechend des sehr heterogenen Nutzungsprofils naturwissenschaftlicher Anwendungen, verteilt sich die abgegebene Rechenzeit zu % auf Ein-Kern-Jobs und zu 85% auf parallele Anwendungen.

Der Studiengang Computational Science an der Goethe-Universität

Das CSC organisiert den Master-Studiengang Computational Science. Der Studiengang wird vollständig in englischer Sprache angeboten und erlaubt somit auch internationalen Studierenden den Zugang ohne Sprachbarriere.

Ziel des Studiengangs ist die Vermittlung der theoretischen Grundlagen sowie praktischer Fähigkeiten im wissenschaftlichen Rechnen. Voraussetzung für die Zulassung ist ein Bachelor in Informatik, Geowissenschaften, Mathematik, Meteorologie, Neurowissenschaften, Chemie oder Physik. Aufbauend auf der Basis mathematischer und methodischer Grundlagen des jeweiligen Bachelor-Studiums, bietet der Master-Studiengang eine Einführung in das moderne wissenschaftliche Rechnen, sowohl allgemein als auch im Hinblick auf das jeweilige wissenschaftliche Arbeitsgebiet, in dem sich die Studierenden spezialisieren.
http://www.physik.uni-frankfurt.de/mpcs/index.html



Nutzung der HLR LOEWE-CSC (links) und FUCHS (rechts) nach Standorten im Jahr 2011

Anpassung des LINPACK-Benchmarks für den LOEWE-CSC

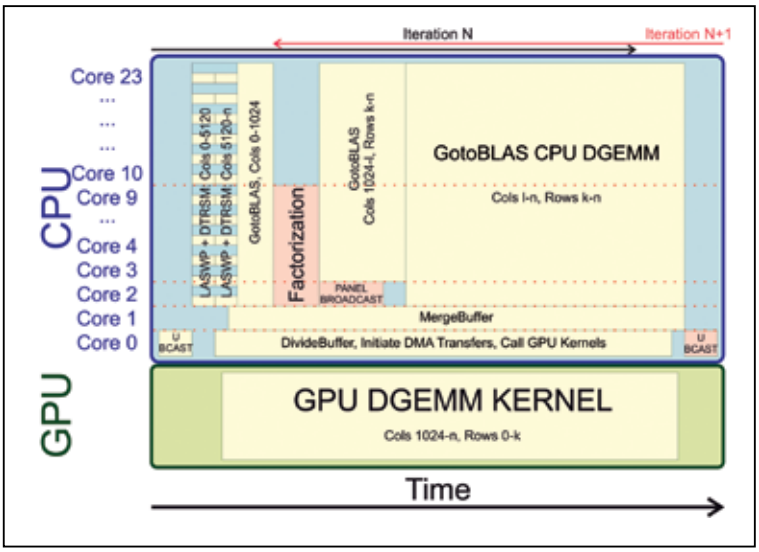
Der LOEWE-CSC-Rechencluster ist ein heterogenes System, dessen Rechenleistung sowohl von gewöhnlichen Prozessoren (CPUs) als auch von Grafikkarten (GPUs) bereitgestellt wird. Zur Messung der Spitzenrechenleistung eines Clusters wird für gewöhnlich der Linpack-Benchmark eingesetzt. Linpack misst die erreichte Rechenleistung anhand der Anzahl von Fließkomma-Operationen, die der Cluster während des Lösens eines dicht besetzten Gleichungssystems durchzuführen vermag.

Die vorhandenen Implementierungen des Linpack-Benchmarks sind auf herkömmliche Cluster ausgelegt und können die von den Grafikkarten zusätzlich bereitgestellte Rechenleistung nicht oder nur eingeschränkt nutzen. Aus diesem Grund wurde für den LOEWE-CSC eine neue Variante des Benchmarks entwickelt.

Den größten Teil der Laufzeit des Linpack-Benchmarks nimmt eine Subroutine namens DGEMM ein, die eine Matrixmultiplikation durchführt. Matrixmultiplikation führt abwechselnd Multiplikationen und Additionen durch und ist damit wie geschaffen für Grafikkarten.

Die Linpack-Version für den LOEWE-CSC ist wie folgt designt: Die Matrixmultiplikation wird zum größten Teil auf die GPU ausgelagert. Die CPU übernimmt alle übrigen Subroutinen. Insofern noch freie CPU-Ressourcen zur Verfügung stehen, übernimmt die CPU Teile der Matrixmultiplikation, sodass eine vollständige Auslastung beider Prozessoren gewährleistet ist. Hierfür wird ein dynamischer Scheduler nach dem Work-Stealing-Prinzip eingesetzt.

Der ursprüngliche Linpack-Benchmark arbeitet seriell und kann auf einem Rechenknoten nicht gleichzeitig die Matrixmultiplikation und andere Berechnungen durchführen, da diese jeweils das Ergebnis des anderen benötigen. Eine genaue Analyse der Abhängigkeiten zeigt jedoch, dass nur ein gewisser Teil der Matrixmultiplikation benötigt wird, um die nachfolgenden Schritte durchzuführen. Indem die Matrix geschickt zerlegt wird, kann dieser Teil zuerst berechnet werden, sodass danach beide Teilaufgaben unabhängig sind und parallelisiert werden können.



Darüber hinaus ermöglicht die Aufspaltung der Matrix in mehrere Teile, eine Pipeline für den GPU-Arbeitsfluss einzurichten. Für jeden Matrixteil müssen nacheinander ein Vorbereitungsschritt auf der CPU, der Transfer auf die GPU, die eigentliche Matrixmultiplikation, der Rücktransfer und eine Nachbearbeitung durchgeführt werden. Die Pipeline ermöglicht es, dass diese Schritte für verschiedene Matrixteile überlappen. Dies stellt sicher, dass zu jeder Zeit die GPU durch Matrixmultiplikation voll ausgelastet ist. Die Abbildung vermittelt einen Überblick über das komplexe Scheduling.

Die Matrixmultiplikation auf der Grafikkarte erreicht mit 494 GFlop/s ca. 90% der spezifizierten Maximalleistung. Berücksichtigt man noch den Transfer sowie Vor- und Nachbearbeitung bleiben noch 465 GFlop/s für den Anwender nutzbare Leistung übrig. Übernimmt die CPU einen entsprechenden Teil der Matrix, steigert sich die Gesamtleistung der Matrixmultiplikation auf 624 GFlop/s. Der Linpack-Benchmark, der zum größten Teil aus der Matrixmultiplikation, aber auch aus anderen Subroutinen besteht, wobei letztere nicht gleichermaßen effizient implementiert werden können, kommt noch auf 563 GFlop/s. Nutzt man mehrere Rechenknoten gleichzeitig, reduziert sich dies wegen Verlusten durch die Latenz der Datenübertragung auf 525 GFlop/s pro teilnehmendem Knoten. Der LOEWE-Cluster insgesamt kommt in der Messung auf 299 TFlop/s.

Das Linux-Cluster im IT-Servicezentrum der Universität Kassel

Das Linux-Cluster besteht zurzeit aus 160 Maschinen mit insgesamt 1200 Prozessorkernen. Jede Maschine hat mindestens zwei AMD-Opteron-Prozessoren und 8 GB Hauptspeicher. 62 dieser Systeme sind mit einer Infiniband-Vernetzung ausgestattet. Das Cluster läuft unter dem Betriebssystem CENTOS 5.4 mit PBS-Torque-Resource-Manager und Maui-Cluster-Scheduler.

Zehn Maschinen stehen für die interaktive Nutzung zur Verfügung. Der Zugriff auf Plattenspeicher erfolgt über GPFS-Dateisysteme. Von den 1200 Prozessoren wurden 336 Prozessoren aus Mitteln eines Fachgebietes im Fachbereich Naturwissenschaften beschafft und stehen deshalb nur den Mitarbeitern dieses Fachgebietes zur Verfügung. Im Jahr 2009 wurden als letzte Ergänzungsbeschaffung 40 Doppelprozessorsysteme mit je zwei AMD-Sechskernprozessoren und 64 bzw. 128 GB Hauptspeicher in Betrieb genommen.

Im laufenden Jahr wird das Linux-Cluster um 3300 Prozessorkerne (AMD-Opteron 6276) erweitert. Die Erweiterung erfolgt über einen Großgeräteantrag des Fachbereichs Mathematik und Naturwissenschaften gemäß Artikel 91b des Grundgesetzes. Die Clustererweiterung soll insbesondere für den Betrieb von massiv-parallelen Anwendungen in der Theoretischen Physik genutzt werden.

Das Hessische Ministerium für Wissenschaft und Kunst fördert die Nutzung des Linux-Clusters für massiv-parallele Anwendungen im Rahmen der Stärkung der Methodenkompetenz im Hessischen Hochleistungsrechnen. Im Zuge der Erweiterung des Linux-Clusters ist auch die Aktualisierung des Betriebssystems und der Anwendungssoftware geplant. Aktuell wird das Cluster auf das Betriebssystem Scientific Linux 6 umgestellt.

Anwendungssoftware

Abaqus-6.10-2 (x86-64), ACML 4.1.0, jrMan, Gausian 03, GROMACS 3.3.3-1, Matlab R2011b, Mathematica 7.0, MD Nastran 2.1 (x86-32), MD Patran 2.1 (x86-32), Meep-0.20.3, MPI: mpich-1.2.7, MPI: LAM/MPI version 7.1.2, MPI: mpich2-1.0.5p4 , MPI: mvapich2-1.0.1, mpb-1.4.2, NWChem 5.1, OpenFOAM-1.5, Pixie-2.2.4, R-2.8.1 (Rmpi mit lme4)

Compiler

gcc-4.1.2, Portland-7.0-5 (C, C++, Fortran), Intel-12.1 (C, C++, Fortran)

Auslastung

In den Jahren 2010 und 2011 wurden insgesamt 14.000.000 CPU-Stunden zur Verfügung gestellt, von denen 6.500.000 Stunden von ca. 50 Nutzern abgerufen wurden. Die größten Nutzergruppen kommen aus den Fachbereichen Physik, Maschinenbau und Elektrotechnik.

Die Hochleistungsrechner der GSI

Die GSI betreibt eine große, weltweit einmalige Beschleunigeranlage für Ionenstrahlen. Forscher aus aller Welt nutzen die Anlage für Experimente in der Grundlagenforschung. Die wohl bekanntesten Resultate sind die Entdeckung von sechs neuen chemischen Elementen und die Entwicklung einer neuartigen Tumortherapie mit Ionenstrahlen. In den nächsten Jahren wird an der GSI das neue internationale Beschleunigerzentrum FAIR (Facility for Antiproton and Ion Research) entstehen.

Die wichtigsten Anforderungen an die Hochleistungsrechner-Systeme der GSI sind:

- Die Analyse der Daten der GSI-Experimente, die vorbereitenden und begleitenden Simulationsrechnungen sowie die Langzeitarchivierung der Daten und Ergebnisse
- Die Beteiligung als Tier-2-Zentrum am weltweiten Grid zur Auswertung des LHC-Experiments ALICE
- Rechnungen zur Vorbereitung der FAIR-Beschleunigeranlage und der FAIR-Experimente
- Unterstützung der für die Schwerionenphysik notwendigen Theorie-Rechnungen

Das System

Das HLR-System der GSI ist für die Analyse von Experimentdaten der Kern- und Hochenergiephysik optimiert, das heißt für die Verarbeitung großer Datenmengen bei höchster I/O-Leistung.

Die wichtigsten Kennzahlen sind:

- 621 Rechenknoten in Intel64/AMD64-Architektur unter Debian GNU/Linux verteilt auf zwei Rechencluster;
- Cluster-Filesystem Lustre: 140 Fileserver mit ca. 5000 Festplatten und einer Kapazität von 3,5 PB;
- Archiv, basierend auf zwei Automatic Tape Libraries mit einer Kapazität von 3,5 PB;
- Backbone-Ethernet-Switch Brocade BigIron RX-32 mit einer Kapazität von 5,12 Tb/s.

Für das HLR-Cluster betreibt die GSI eine Strategie des kontinuierlichen Ausbaus und der kontinuierlichen Erneuerung. Ein- bis zweimal pro Jahr werden die jeweils am besten geeigneten, aktuell am Markt verfügbaren Gerätegenerationen beschafft. Dadurch wird die schnelle Weiterentwicklung und Leistungssteigerung der IT-Komponenten optimal genutzt.

Bis Ende 2010 entstand ein Rechen-Cluster mit 521 Rechnern unterschiedlicher Leistungsfähigkeit („Harpertown“, „Clovertown“, „Gainestown“), mit insgesamt etwas mehr als 3000 Kernen und 2 bis 4 GB Hauptspeicher pro Kern. Als Betriebssystem werden verschiedene Versionen von Debian Linux eingesetzt. Zur Installation und Pflege des Systems werden FAI und cfengine benutzt. Das Monitoring basiert auf Nagios, Torrus und Monalisa. Die Nutzer nutzen auf dem Cluster ihr GSI-weites Home-Filesystem und submittieren ihre Jobs via LSF.

Im zweiten Quartal 2011 baute die GSI einen zweiten Cluster auf, bestehend aus 100 Systemen

mit insgesamt 1600 Kernen. Dabei wurden neue Software-Technologien erprobt, die es erlauben, das System auf mehrere zehntausend Kerne zu erweitern. Das Cluster-Management wurde auf das Ruby basierte System Chef umgestellt. Das Batchsystem wurde durch SGE ersetzt. Zur Verteilung von Benutzersoftware wird cvmfs genutzt.

Die Systeme der GSI sind darauf optimiert, mit Tausenden von Jobs gleichzeitig große Datenmengen zu lesen. Im Zentrum steht das Cluster-Filesystem Lustre. Der Metadaten-Server (MDS) besteht aus zwei redundanten Systemen mit jeweils 48 Kernen und 128 GB Hauptspeicher, die via drdb synchron gehalten werden. Die Daten befinden sich auf 140 Fileservern mit ca. 5000 SATA-Festplatten und einer Kapazität von 3,5 PB. Das Lustre-System und die Rechenfarmen sind mittels Backbone-Ethernet-Switch Brocade BigIron RX-32 so verbunden, dass sich eine aggregierte Bandbreite von 0,5 Tb/s ergibt.

Zur Archivierung der Experimentdaten wird das bei GSI entwickelte System gStore verwendet. Es basiert zurzeit auf zwei Automatic Tape Libraries (ATL 3584-L23) mit einer Kapazität von 3,5 PB und 16 Data-Mover-Servern mit einem Lese/Schreibcache von insgesamt 200 TB, die in einem Storage Area Network (SAN) miteinander verbunden sind. gStore wurde im Hinblick auf einfache Skalierbarkeit entwickelt, sowohl im Hinblick auf Datenkapazität als auch auf I/O-Bandbreite. Mit der derzeitigen Konfiguration erreicht man Transferraten von 1.2 GB/s zwischen den Data-Mover-Platten und der Tape Library, 5 GB/s zwischen den Data-Movern und Klienten im LAN (Lustre) und bis zu 1 GB/s von der Experiment-Datenaufnahme zu den Data-Movern.

Ausbaupläne

Im Berichtszeitraum war die HLR-Infrastruktur über drei Standorte verteilt:

- Das Hauptrechenzentrum der GSI hat 400 qm Grundfläche und eine Anschluss- und Kühlkapazität von 180 kW. Mit einer traditionellen Luftkühlung und ausgelegt auf die niedrigen Leistungsdichten der Mainframe-Ära dient es als General-Purpose-Rechenzentrum und ist bis an die Grenzen der Kapazität ausgelastet. Es beherbergt etwa die Hälfte der Lustre-Fileserver.
- Für die Rechenknoten wurde ein zweiter Standort im Beschleuniger-Betriebsgebäude mit 200kW Kühlleistung in Betrieb genommen.
- Die zweite Hälfte der Lustre-Fileserver befindet sich in einem alten Container. Dort testet die GSI ein innovatives Kühlkonzept: Die Rechner befinden sich in geschlossenen 19-Zoll-Schränken, mit einem passiven Wärmetauscher in der hinteren Tür. Die Luft von etwa 40-45° C vor dem Wärmetauscher wird auf die Raumtemperatur von etwa 27-28° C gekühlt. Durch die hohe Raumtemperatur ist freie Kühlung möglich. Dies führt zu einem PUE kleiner 1,1 im Jahresmittel. In den über 1,5 Jahren Betrieb gab es keine erhöhten Ausfälle von Servern oder Platten.

Nach dem erfolgreichen Test des neuen Kühlkonzepts wurde ein Rechenzentrum als eine Art zweistöckiges Hochregallager aufgebaut. Pro Stockwerk stehen 48 12-Zoll-Schränke zur Verfügung. Die Infrastruktur mit 2 MVA Strom und 1,2 MW Kühlkapazität (ausbaubar auf 1,8MW) war Ende 2011 fertiggestellt. Im ersten Quartal 2012 wird dort ein Rechencluster mit ca. 10000 Kernen und einem latenzarmen Infiniband Netz, sowie ein Lustre-System mit 1PB in Betrieb genommen. Danach sollen dort alle HLR-Systeme der GSI zentralisiert werden. Gegen Ende 2012 ist geplant ein größeres mit GPUs bestücktes System aufzubauen. Damit wird das HLR-Cluster der GSI auch für Anwendungen des klassischen Hochleistungsrechnens attraktiv. Es soll dann in größerem Umfang für GitterQCD-Rechnungen verwendet werden.

Nutzung

Die Hauptnutzer sind die GSI-Experimente HADES und FOPI, das ALICE-Experiment am CERN, die FAIR-Experimente CBM, PANDA und NUSTAR, die Theorie-Gruppe der GSI, die Beschleunigertheorie, und die Abteilung Sicherheit und Strahlenschutz. Die Verteilung der Ressourcen wird monatlich abgestimmt und kann kurzfristig den jeweiligen Bedürfnissen angepasst werden. Mit der Ausnahme der beiden Theoriegruppen, die auch parallele MPI Jobs benötigen, wird das System durch serielle Batchjobs im High-Throuput-Betrieb genutzt. Die wichtigsten Programmiersprachen sind Fortran und C++.

Für das ALICE Experiment ist die GSI als Tier-2-System in das internationale Gridsystem für die LHC-Experimente integriert. Dafür stehen etwa 1000 Jobslots und 400TB Plattenplatz zur Verfügung und werden kontinuierlich genutzt. Die deutschen ALICE-Gruppen können zusätzlich weitere 1000 Jobslots für das innovative System der Analyse-Trains verwenden. Hierbei kombinieren die Physiker ihre Analyse-Schritte als verschiedene Wagen eines Zuges. So müssen die Daten nur einmal gelesen werden. Zusammen mit der riesigen I/O-Bandbreite des Lustre-Systems führt dies dazu, dass mehrmals pro Woche Hunderte von Terabytes mit neuen Analyseschritten ausgewertet werden können. Dies führt zu einem wichtigen Vorsprung der deutschen Gruppen bei der Auswertung von ALICE.

Die Rechencluster der Philipps-Universität Marburg

Abbildung:
Der Cluster MaRC2 nach Aufbau und Installation. Für den Betrieb von MaRC2 wurde die Klimatisierung im HPC-Serverraum renoviert. Eine Kaltgang-Einhausung soll zudem künftig die Luftführung optimieren.

Der alte Linux-Cluster MaRC

Die Systemarchitektur des HPC-Clusters MaRC (Marburger Rechenclusters) wurde bereits im vorangegangenen Jahresbericht beschrieben (vgl. [1]). Das folgende Diagramm zeigt die fortgeschriebene Nutzungsstatistik bis zum Januar 2012.

Die Zeitspanne der wirtschaftlichen Nutzung von MaRC neigt sich dem Ende zu. Nach Ablauf der Garantiephase wurden Hardware-Ausfälle meist nicht mehr kompensiert. Stattdessen wurden aus funktionsfähigen Einzelteilen defekter Compute Nodes lauffähige Systeme zusammengestellt, soweit der Aufwand vertretbar war. Mehrere Stromausfälle in den Jahren 20 und 2011 führten nicht nur zum Verlust von Rechenzeit (sichtbar in der Nutzungsstatistik), sondern verursachten auch Defekte bei RAM-Modulen in den Compute Nodes der zweiten Ausbaustufe und damit weitere Ausfallzeiten.

MaRC2 – der neue HPC-Cluster

Im Sommer 20 stellte Professor Bruno Eckhardt einen DFG-Antrag für Forschungsgroßgeräte, an dem zehn Marburger Forschungsgruppen und das HRZ beteiligt waren. Dieser Antrag wurde im Februar 2011 bewilligt. Durch die großzügige Förderung der DFG und des HMWK konnte so das Nachfolgesystem MaRC2 im Rahmen einer EU-Ausschreibung beschafft und im Februar 2012 installiert werden. Der Testbetrieb für die ersten Nutzer startete im März 2012.

Der Cluster MaRC2 basiert auf der AMD-Interlagos-Architektur. Jeder der 88 Compute Nodes verfügt über vier 2.3 GHz CPUs mit jeweils 16 Cores (=8 Module). Damit sind pro Compute Node insgesamt 64 CPU Cores verfügbar, die auf 256 GB RAM zugreifen können. Dies ermöglicht eine hohe Parallelisierung im Rahmen des Shared-Memory-Programmiermodells, zum Beispiel via OpenMP. Der hohe Speicherausbau war für viele Anwendungen von zentraler Bedeutung.

Alle Compute Nodes sind über ein QDR Infiniband-Netzwerk verbunden.

Einen deutlichen Ausbau gegenüber dem Vorgänger MaRC hat auch das I/O-Subsystem erfahren, und zwar sowohl bzgl. Performance als auch bzgl. der Kapazität. Die I/O-Leistung des Fileservers von MaRC hatte sich in der letzten Betriebsphase für viele Anwendungen als Flaschenhals erwiesen. MaRC2 verfügt über zwei clusterweite File-Services: Das Home-Filesystem ist per NFS angebunden und zielt auf eine hohe Verfügbarkeit. Die Scratch-Fileserver stellen ein performantes paralleles Filesystem (FhGFS) bereit. Die nutzbare Kapazität beträgt für beide File-Services jeweils über 30 TB.

Auf den Einsatz von GPUs wurde verzichtet, da die Unterstützung der Benutzer bei der Nutzung einen zu hohen Betreuungsaufwand erfordert hätte, der derzeit nicht geleistet werden kann. Für Anwendungen, die vom Einsatz von GPUs in hohem Maße profitieren, steht im Rahmen des hessischen HLR-Verbunds mit dem CSC Cluster in Frankfurt eine prominente Alternative bereit.

Besonderes Gewicht wurde bei der Ausschreibung von MaRC2 auf die Energieeffizienz des Systems gelegt: Die HPL-Effizienz von MaRC2 liegt bei 563 MFlops/Watt.

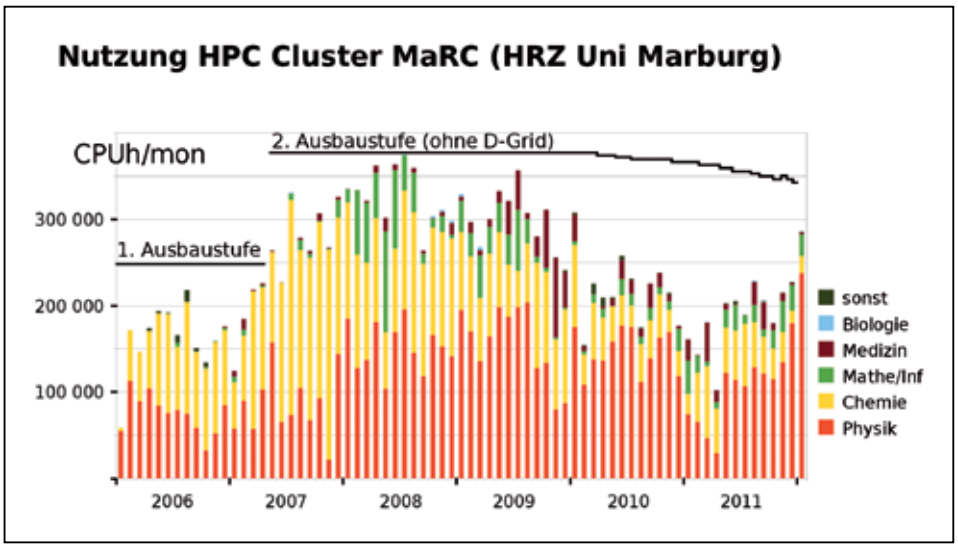


Abbildung:
Nutzung von MaRC durch verschiedene Fachbereiche. Knoten der zweiten Ausbaustufe wurden ihrer höheren Taktfrequenz entsprechend mit dem Faktor 1.2 gewichtet.

Das zentrale Hochleistungsrechen-Cluster der Justus-Liebig-Universität Gießen

Skylla heißt das zentrale Hochleistungsrechen-Cluster der Justus-Liebig-Universität Gießen. Mit einer theoretischen Rechenleistung von 9,3 TFLOPs und insgesamt 2304 GB Hautspeicher unterstützt es die Wissenschaftlerinnen und Wissenschaftler dabei, komplizierte Probleme binnen Stunden oder Tagen zu lösen, für die normale Rechner mehrere Jahre bräuchten.

2009 von der Firma Clustervision geliefert, aufgebaut und konfiguriert, wurde das Cluster im Mai 2011 auf den jetzigen Stand ausgebaut. Skylla setzt sich aktuell aus 86 Rechenknoten mit insgesamt 992 Prozessorkernen, zwei Frontend-Rechnern, einem Lustre-Dateisystem und dem internen Kommunikationssystem zusammen. Diese Komponenten sind in vier Racks mit dazwischen angeordneten Kühleinheiten eingebaut. Durch die Kühleinheiten ist die Klimatisierung des kompletten Systems von der Raumklimatisierung unabhängig.

Die Rechenknoten sind in Twin-Chassis untergebracht, die je zwei Motherboards mit je zwei Vier- oder Sechs-Kern AMD-Prozessoren enthalten. Der Hauptspeicherausbau ist so gewählt, dass jedem Rechenkern mindesten 2 GB RAM zur Verfügung stehen. Für Aufgaben, die einen schnellen Zugriff möglichst vieler Rechenkerne auf einen gemeinsamen Hauptspeicherbereich erfordern, sind sechs Rechenknoten mit je vier Acht-Kern Prozessoren (32 Rechenkerne) und 64 GB Hauptspeicher ausgestattet. Der Zugang der Nutzer zum System erfolgt über die beiden Frontend-Rechner, die außer als Login- auch als File- und Queuing- Server dienen.

Schnelles Dateisystem

Insgesamt stehen 29 TB Plattenspeicherplatz zur Verfügung. Die Homeverzeichnisse (7,5 TB) der Benutzer und ein Scratch-Bereich (3,5 TB) zur Zwischenspeicherung von Daten sind über den Fileserver per NFS ansprechbar. Darüber hinaus steht ein Lustre-Dateisystem mit 18 TB zur Verfügung. Es ist für große Datenmengen vorgesehen, auf die im Verlauf einer Berechnung schnell und möglichst verzögerungsfrei zugegriffen werden muss. Es ist nicht zur permanenten Datenaufbewahrung gedacht und nicht in die Datensicherung einbezogen. Das Lustre-Dateisystem ist durch zwei MDS-Server als aktiv/passiv HA-Cluster für die Metadaten und zwei OSS-Knoten für die Nutzdaten realisiert Die Metadaten liegen auf einem externen RAID-System (RAID10 + Hotspare) mit einer Nettokapazität von ca. 1 TB. Für einen schnellen Zugriff werden im RAID-System 15*146-GB-SAS-Platten mit 15.000 RPM verwendet.

Für die interne Kommunikation zwischen den Clusterknoten stehen drei separate Daten-Netzwerke zur Verfügung: Als Standardnetzwerk dient ein 1 GBit/s-Ethernet. Für die Interprozesskommunikation und den Dateizugriff kann wahlweise ein Infiniband-Netzwerk (Mellanox-Switch, 4x Double Data Rate,

theoretisch 20 GBit/s)genutzt werden. Schließlich wird für den Zugriff auf die IPMI-Ports der Knoten (z.B. zur Hardwareüberwachung) ein 10/100-MBit-Ethernet-Netzwerk eingesetzt.

Software und Compiler

Als Betriebssystem auf den Knoten wird Scientific Linux, das auf Red Hat Enterprise Linux (RHEL) aufbaut, verwendet. Das Queuing-System „Sun Grid Engine (SGE)" dient zur Jobsteuerung. Zur Administration des Clusters wird das proprietäre Produkt „Bright Cluster Manager" von Clustervision genutzt. Für die Interprozesskommunikation sind folgende Software-Produkte und -Bibliotheken installiert: LAMMPI, MPICH und OpenMPI. Als Programmiersprachen stehen Fortran und C zur Verfügung. Dazu sind die folgenden Compiler installiert: GNU Compiler Collection, Intel, Portland PGI Compiler, Open64 Compiler Suite, Oracle Solaris Studio und g95. Die Liste der Anwendungssoftware umfasst die Produkte: Mathematica, Octave, Geant, R (mit verschiedenen Modulen), CMake, gnuplot, Radiance und MrBayes.

Mehr als 100 Benutzer aus acht Arbeitsgruppen teilen sich die zur Verfügung stehende Rechenkapazität, wobei der Hauptanteil auf Arbeitsgruppen aus der Physik und der Chemie entfällt. Der Auslastungsgrad des Clusters liegt bei 60%.