bmb+f - Förderschwerpunkt

Astrophysik

Großgeräte der physikalischen Grundlagenforschung

Schlussbericht vom 9.5.2004 zum Thema:

Entwicklung und Betrieb eines hochparallelen PC-Clusters am Astrophysikalischen Institut Potsdam für numerische Simulationen zur Galaxienentstehung

Zuwendungsempfänger:	Astrophysikalisches Institut Potsdam
Projektleitung:	Prof. Dr. M. Steinmetz
Förderkennzeichen:	05EA2BA1/8
Förderzeitraum:	01.05.02 – 31.12.03
Zuwendung:	767000 EUR
E-Mail:	msteinmetz@aip.de
Projektträger:	Projektträger DESY

Genutzte Großgeräte:

Angaben zum Projekt:

Veröffentlichungen:	11
Konferenzbeiträge:	3
Diplomarbeiten:	0
Dissertationen:	0
Habilitationen:	0
Patente:	0

Das diesem Bericht zugrundeliegende Vorhaben wurde mit Mitteln des Bundesministeriums für Bildung und Forschung gefördert. Die Verantwortung für den Inhalt dieser Veröffentlichung liegt beim Autor.

Schlussbericht-Kurzfassung

Zuwendungsempfänger: ***Astrophysikalisches Institut Potsdam***

Projektleitung: ***Prof. Dr. Matthias Steinmetz***

Entwicklung und Betrieb eines hochparallelen PC-Clusters am Astrophysikalischen Institut Potsdam für numerische Simulationen zur Galaxienentstehung

Motivation

Die neuste Generation von Großteleskopen am Boden und im Weltraum erlaubt uns einen atemberaubend detaillierten Blick auf die Entwicklung des Universums. Diese Daten haben in ihrer Vielfalt und ihrem Detailreichtum den Rahmen klassischer Modelle der theoretischen Astrophysik verlassen. Numerische Simulationen höchster Auflösung sind in der Lage, dieses Bindeglied zwischen Theorie und Beobachtung zu liefern. Mit der zunehmenden Verbreitung von PC-Clustern steht nun dafür eine effektive, kostengünstige und leistungsfähige Alternative zum Einsatz traditioneller Supercomputer zur Verfügung. Aufbau, Optimierung und Test einer solchen Konfiguration waren primäres Ziel des Vorhabens.

Durchführung

Im Rahmen des Projekts wurden am AIP zwei hochparallele PC-Cluster installiert. Der erste aus 72 Intel-Pentium-4 in 36 Dual-Knoten bestehende Testcluster (*OCTOPUS*) wurde im Herbst 2002 installiert und diente primär zum Sammeln von Erfahrungen, um die Hauptinvestition für das Nutzungsprofil des AIP optimal durchführen zu können. Der ursprünglich mit 100Mbit/s Netzwerktechnik ausgestattete Cluster konnte durch Nachrüstung auf 1Gbit/s in seinem Durchsatzvermögen merklich gestärkt werden. Es zeigt sich jedoch, dass die technische Grenze von maximal 4GB pro Knoten zunehmend eine unüberwindbares Problem für die größten Simulationen darstellt. Auf Grund dieser Erfahrung wurde bei der Installation (Winter 2003/2004) des Hauptclusters SANSSOUCI, auf die neue 64-bit-Architektur von AMD-Opteron gesetzt.

Erste Ergebnisse

- Im Rahmen des Projekts wurden zwei PC-Cluster installiert, für hochparallele Anwendungen optimiert und dem Routinebetrieb am AIP übergeben.
- Auf dem Testcluster *OCTOPUS* wurden bereits mehrere wissenschaftliche Simulationsrechnungen durchgeführt, insbesondere zur Entstehung von Galaxien und Galaxienhaufen sowie zur Ausbildung von Magnetfeldern in der Milchstraße. Diese Arbeiten führten bereits zu mehreren Publikationen in referierten Journalen.
- Auf dem Hauptrechner *SANSSOUCI* wurde die bislang weltweit größte gas-dynamische Simulation zur Strukturbildung im Universum gestartet. Gesamtrechen-bedarf wird mehrere Monate auf 256 CPUs sein.
- Mit dem Hauptcluster *SANSSOUCI* wurde mit dem Linpack-Benchmark eine Performance von 623.3 Gflop/s erreicht. Ein entsprechender Antrag für die Aufnahme in die Liste der 500 schnellsten Computer der Welt wurde eingereicht. Damit ist *SANSSOUCI* unseres Wissens nach der schnellste Institutsrechner in Deutschland und der schnellste Rechner an einem astrophysikalischen Institut weltweit.

Schlussbericht

Zuwendungsempfänger: ***Astrophysikalisches Institut Potsdam***

Projektleitung: ***Prof. Dr. Matthias Steinmetz***

Entwicklung und Betrieb eines hochparallelen PC-Clusters am Astrophysikalischen Institut Potsdam für numerische Simulationen zur Galaxienentstehung

Motivation

Die neuste Generation von Großteleskopen am Boden und im Weltraum erlaubt uns einen atemberaubend detaillierten Blick auf die Entwicklung des Universums und der Galaxien. Diese Daten haben in ihrer Vielfalt und ihrem Detailreichtum den Rahmen klassischer Modelle der theoretischen Astrophysik verlassen. Numerische Simulationen höchster Auflösung sind zwar im Prinzip in der Lage, dieses Bindeglied zwischen Theorie und Beobachtung zu liefern, denoch wurde die dazu notwendige Rechenleistung erst jüngst von Supercomputern erreicht. Mit der zunehmenden Verbreitung von PC-Clustern steht nun eine effektive, kostengünstige und leistungsfähige Alternative zum Einsatz traditioneller Supercomputer zur Verfügung. Aufbau, Optimierung und Test eines solchen PC-Clusters für den Einsatz in der computergestützten Astrophysik war primäres Ziel des Vorhabens.

Durchführung

Das Vorhaben sieht die Installation des Clusters in zwei Phasen vor. In einer ersten Phase sollte im Sommer 2002 ein Testcluster installiert werden, um erste Erfahrungen mit dieser Technologie zu sammeln. Dem sollte im 1. Halbjahr 2003 der Hauptcluster folgen.

Der Testcluster (*OCTOPUS*) wurde deutschlandweit ausgeschrieben, es bewarben sich 8 Anbieter mit 13 verschiedenen Vorschlägen. Der Zuschlag erging an die in Hamburg ansässige Firma Delta-Computer, die eine Lösung mit 36 Knoten vorschlug (für technische Details siehe Tabelle 1). Der Cluster wurde Anfang September 2002 installiert, und konnte schon kurz später für erste Simulationsrechnungen verwendet werden. So konnten schon auf der Sitzung des wissenschaftlichen Beirats des AIP Ende Oktober 2002 die Ergebnisse erster Simulationsrechnungen präsentiert werden. Jedoch zeigt sich die auf dem Cluster installierte Softwareumgebung „score“ als fehleranfällig und wenig flexibel. Zudem konnte zwar in individuellen Tests ein leichter Performancegewinn festgestellt werden, in praktischen Anwendungen aber verschlechterte „score“ die Performance. Da Anfang Januar auf Grund eines Totalausfalls (Fehlschaltung in der Stromversorgung) eines RAID-Systems das ganze Betriebssystem neu zu installieren war, beschlossen wir, auf eine Standard-Linux (redhat-8.1)-Clusterumgebung zu wechseln. Danach arbeitete der Testcluster nicht nur problemlos, sondern auch mit einer höheren Performance. Weitere Tests mit einigen 1Gbit Netzwerkkarten zeigten, dass die Performance von hochparallelen Anwendungen mit dieser schnelleren, aber nur moderat kostenintensiven

Netzwerktechnologie merklich verbessert werden konnte. Der Performancegewinn lag im Bereich von 20-30%, so dass der Testcluster dann ganz auf Gbit-Netzwerk umgestellt wurde. Seit dieser Umstellung läuft der Testcluster problemlos mit nur minimalen administrativen Ausfällen. Die Auslastung liegt durchschnittlich oberhalb 75% (inklusive Leerzeiten durch suboptimales Loadbalancing), und es gibt praktisch keine Ausfallzeiten.

Abbildung 1: Links: der 72-CPU-Testcluster *OCTOPUS*; rechts: der 270-CPU-Hauptcluster *SANSSOUCI*

Im Frühjahr 2003 wurde mit den Vorbereitungen für die Installation des Hauptclusters begonnen. Dazu musste zunächst der ursprünglich für eine Röntgentestanlage vorgesehene Raum im Keller des AIP-Schwarzschildhauses als Rechnerraum umgebaut und klimatisiert werden. Auch die unterbrechungsfreie Stromversorgung musste für die zukünftig erwarteten Lasten des Rechners und der Klimaanlage ausgebaut werden. Da der Rechenraum genügend Standfläche bietet und ein hoher Luftdurchsatz sich technisch einfach bewerkstelligen ließ, war eine klassische Kühlung die kostengünstigste Lösung. Wassergekühlte Schränke, wie sie z.B. am Forschungszentrum Karlsruhe zum Einsatz kamen, waren nicht notwendig. Die Arbeiten wurden im April 2004 abgeschlossen (siehe unten unter Zeitplan), und bietet nun Raum und Luftdurchsatz um 2-3 Cluster der Größe von SANSSOUCI zu bedienen. Jedoch war der Rechenraum schon im Dezember 2003 bedingt benutzbar (Kühlung über Außenluft).

Die Ausschreibung für den Hauptcluster (*SANSSOUCI*) erfolgte EU-weit im Juli 2003, und es bewarben sich 9 Firmen mit 20 Angeboten, teilweise auf 32-bit Intel-Xeon, teilweise auf 64-bit AMD-Opteron basierend. Vier Anbieter wurden in die nähere Auswahl genommen und die angebotenen Architekturen intensiv auf ihre Leistungsfähigkeit insbesondere für AIP-übliche Anwendungen überprüft. Dabei erwies sich die neuen 64-bit-Opteron-Prozessoren als die bessere Alternative, insbesondere für Anwendungen mit großem Speicherbedarf. Aufgrund des schnellen Datenbus, der zudem individuell für jede CPU zur Verfügung steht (bei den Intel-Xeon-Prozessoren teilen sich 2 CPUs einen Bus), zeigte der Opteron eine deutlich

höhere Performance trotz nominell niedrigerer Peak-Performance. Die in Österreich/England beheimatete Firma Quant-X/Compusys unterbreitete das mit Abstand beste Angebot (für Details, siehe Tabelle 1) und erhielt Ende Oktober den Zuschlag.

Anfang Dezember wurde der Hauptcluster dann in Potsdam installiert. Die Testphase und Feinabstimmung nahm deutlich mehr Zeit in Anspruch als beim Testcluster *OCTOPUS*, was im wesentlichen auf noch unausgereifte Software für die neuen Prozessoren zurückzuführen ist. Insbesondere die Temperaturkontrolle, die das automatische Herunterfahren des Rechners bei Überhitzungsgefahr garantiert, war anfänglich nur wenig verlässlich. Die Arbeiten wurden aber erfolgreich abgeschlossen und der *SANSSOUCI* feierlich von der Ministerin für Wissenschaft, Forschung und Kultur, Frau Professor Wanka, am 15. Januar 2004 eingeweiht.

	Octopus (Testcluster)	***SANSSOUCI*** (Hauptcluster)
CPU	76×Intel Pentium4 2.2GHz in 36 Knoten 2 Frontend	270×AMD-Opteron 1.8 GHz in 130 Knoten, 3 I/O-Rechnern und 1 Visualisierungsrechner
Hauptspeicher	36×3 Gbyte 2×4 Gbyte	101×4 Gbyte 32×6 Gbyte 1x16 GByte
Lokaler Plattenplatz	36×80 Gbyte	134×160 GByte
Globaler Plattenplatz	2 RAID mit je 1.7 Tbyte	4 RAID mit je 3 Gbyte
Netzwerk	100 Mbit 1 Gbit	1Gbit
Peak-Performance	281.4 Gflop/s (@64 CPU)	921.6 Gflop/s (@256 CPU)
Linpack-Performance	~150 Gflop/s (@64 CPU)	623.3 Gflop/s (@256 CPU)

Tabelle 1: Technische Eigenschaften der beiden PC-Cluster

Pünktlich zum Abgabetermin am 15.4.2004 wurde auf 256 Prozessoren von *SANSSOUCI* mit dem Linpack-Benchmark eine Performance von 623.3 Gflop/s erreicht. Dies übertrifft die Leistung des AMD-eigenen Clusters derselben Größe, aber mit schnellerem Netzwerk (10Gbit-Myrinet) um mehr als 2Gflop! Ein entsprechender Antrag für die Aufnahme in die Liste der 500 schnellsten Computer der Welt wurde eingereicht. Damit ist *SANSSOUCI* unseres Wissens nach der schnellste Institutsrechner in Deutschland und der schnellste Rechner an einem astrophysikalischen Institut weltweit. Durch die Kopplung der beiden Cluster erwarten wir, bis zum Herbst etwa 750-800 Gflop/s zu erreichen, und somit auch für die nächste top-500 Liste gut gewappnet zu sein.

Seit dem Abschluss der Benchmark-Rechnungen Ende April steht nun SANSSOUCI den Mitarbeitern des AIP für anspruchsvolle hochparallele Simulationen zur Verfügung.

Zeitplan

Der im Antrag dargelegte Zeitplan sah die Installation des Testclusters im Sommer 2002 und die des Hauptclusters im 1. Halbjahr 2003 vor. Der Zeitplan für den Testcluster konnte eingehalten werden. Die Installation des Hauptclusters verzögerte sich auf Ende 2003. Die ursprünglich anvisierte Lösung, den Cluster bei der IT-Firma *Panmedium* in Potsdam unterzubringen, erschien nach Verhandlungen wirtschaftlich wenig attraktiv, so dass entschieden wurde, auch den Hauptcluster am AIP unterzubringen. Die Mittel für die Klimatisierung des neuen Rechnerraums am AIP mussten trotz Kürzungen des AIP Haushalts durch Bund und Land Brandenburg intern erwirtschaftet werden (ein im Frühjahr 2003 gestellter Aufstockungsantrag an das BMBF wurde auf Grund der Aussichtslosigkeit Ende 2003 zurückgezogen). Durch diese Verschiebung ergab sich zudem die Möglichkeit, auch die zukunftsträchtigeren 64-bit Prozessoren der AMD-Opteron-Familie in die Auswahlentscheidung mit einzubeziehen.

Weitere Ergebnisse

- In Zusammenarbeit mit der Universität Potsdam (Lehrstuhl Bettina Schnor) wurde im Rahmen einer Semesterarbeit von Sven Friedrich der ausfallsichere Betrieb eines RAID-Systems mit zwei Frontend-Servern erfolgreich getestet. Solche Software-Lösungen sind von großer Bedeutung für zahlreiche wirtschaftliche Anwendungen (z.B. Banken, Flugsicherung etc.).
- Auf dem Testcluster *OCTOPUS* wurden bereits mehrere wissenschaftliche Simulationsrechnungen durchgeführt, insbesondere zur Entstehung von Galaxien und Galaxienhaufen sowie zur Ausbildung von Magnetfeldern in der Milchstraße. Diese Arbeiten führten bereits zu mehreren Publikationen in referierten Journalen (siehe Publikationsliste)
- Auf dem Hauptrechner *SANSSOUCI* wurde die bislang weltweit größte gas-dynamische Simulation zur Strukturbildung im Universum gestartet. Gesamtrechenbedarf wird mehrere Monate auf 256 CPUs sein.
- Auf dem Testcluster wurde in Zusammenarbeit von Dr. Stefan Gottlöber (AIP) mit Prof. Kravtsov (University of Chicago) und Prof. Klypin (New Mexico State University) wesentliche Teile des „adaptive mesh refinement“ N-Body/Hydrocode entwickelt. Wesentlicher Aspekt war dabei die schnelle Verfügbarkeit des Clusters in Phasen des „debuggings“. Große Produktionsläufe mit diesem Code werden am Höchstleistungsrechenzentrum Jülich und dem Supercomputer des *Lawrance Livermore National Laboratory* durchgeführt.

Zusammenarbeit mit Industriepartnern

In der Vorbereitung und Durchführung des Vorhabens ergaben sich eine Reihe von Zusammenarbeiten mit industriellen Partnern der IT-Branche.

- In einem „Letter of intent“ mit der Panmedium-Stiftung wurde eine Kollaboration zum Clusterrechnen gestartet. Neben dem Erfahrungsaustausch und der Möglichkeit, im Problemfall auf einen erfahrenen Partner zurückgreifen zu können, stand auch die Überlegung im Vordergrund, für den Hauptcluster die klimatisierten Rechnerräume der Panmedium-Stiftung in den ehemaligen Roten Kasernen in Potsdam zu nutzen. Das Verhandlungsergebnis war dann aber wirtschaftlich wenig attraktiv (im Vergleich zur Anmietung bei Panmedium

amortisieren sich die Investitionskosten für die Klimatisierung bereits nach wenigen Jahren), die Zusammenarbeit erfolgt seitdem primär auf informeller Ebene.
- In einem Kooperationsvertrag mit der in München ansässigen Firma Partec werden derzeit für parallele Rechnungen auf Opteron-Architekturen optimierte Netzwerkprotokolle entwickelt und getestet.
- Da der Hauptcluster die derzeit größte Opteron-Installation von AMD in Europa ist, wurde während der Feinabstimmung des Hauptclusters direkt mit Experten von Quant-X/Compusys, von AMD und von MSI (Hersteller des Motherboard) zusammengearbeitet.

Weitere Aussichten

Die weiteren Erfolgsaussichten des Projekts sind exzellent, das erworbene Know-how kann in vielen Bereichen der computergestützten Astrophysik angewendet werden.
Insbesondere demonstrierten die auf dem PC-Clustern entwickelten Programme auch hervorragende Performance auf hochparallelen Supercomputern, und machen so die PC-Cluster als Entwicklungsplattform am AIP unverzichtbar. Als nächster großer Schritt wird nun die Kopplung heterogener und verteilter Cluster über den GRID angegangen. Dies wird zunächst durch die Kopplung der beiden im Rahmen dieses Projekts erworbenen Cluster am AIP geschehen, später werden dann verschiedene PC-Cluster am AIP und am Albert-Einstein-Institut in Golm gekoppelt. Auch haben erste Planungen für eine Kopplung zwischen Ressourcen in Deutschland (Potsdam, Garching), der Schweiz und England bereits stattgefunden.

Schlussfolgerungen

Unsere Erfahrungen mit PC-Clustern als Rechenmaschinen für astrophysikalische Simulationen lässt sich wie folgt zusammenfassen:

- PC-Cluster bieten eine leistungsfähige, gut beherrschbare und kostengünstige Alternative zum traditionellen Einsatz von Supercomputern an Rechenzentren.
- Ebenso sind sie auf Grund der hohen Verfügbarkeit und Flexibilität ein extrem leistungsfähiges Instrument zur Programm- und Algorithmenentwicklung.
- PC-Cluster sind, sofern sie nur von einer überschaubaren Anzahl von Nutzern beansprucht werden auch mit geringem Personalaufwand (im Produktionsbetrieb ca. 0.5 FTE) gut zu beherrschen.
- Bewährt hat sich insbesondere der Einsatz von Standard-Software und Hardware. Spezielle Cluster-Software bringt oft nur geringen Performance-gewinn bei merklichem administrativen Aufwand. Nicht selten stellte sich Performanceverlust statt Performancegewinn ein.
- Hauptproblem im Betrieb eines PC-Clusters sind die Abwärme (Klimatisierung) sowie die Betriebskosten für Strom und Kühlung, die über den typische Betriebszeitraum (5 Jahre) Kosten in der Größenordnung der Anschaffungskosten erzeugen.

Schlussbericht - Veröffentlichungen

Zuwendungsempfänger: ***Astrophysikalisches Institut Potsdam***

Projektleitung: ***Prof. Dr. Matthias Steinmetz***

Entwicklung und Betrieb eines hochparallelen PC-Clusters am Astrophysikalischen Institut Potsdam für numerische Simulationen zur Galaxienentstehung

Y.Ascasibar, G.Yepes, V.Müller, S.Gottlöber
The radial structure of galaxy groups and clusters
MNRAS **346** (2003) 731

J.Bailin, M.Steinmetz
Coupling between satellite dwarfs and the Milky Way warp,
Proceedings *Satellites and Tidal Streams*, La Palma, Spain, May 26-30 2003,
in press (astro-ph/0310199)

N.Dziourkevitch, D.Elstner, G.Ruediger
Interstellar turbulence driven by the magnetorotational instability
Astronomy and Astrophsics (Letters) (2004) submitted

N.Dziourkevitch, D.Elstner, G.Ruediger
3D global simulations of galactic magnetic fields and gas flows
Ap&SS **284** (2003) 757

P.Egorov, G.Ruediger, U.Ziegler
Vorticity and helicity of the solar supergranulation flow-field
Astronomy and Astrophysics (2004) submitted

S.Gottlöber, E. Lokas, A.A. Klypin, Y. Hoffman
The structure of voids
MNRAS **344** (2003) 715-724 (astro-ph/0305393)

M.Hoeft, J.P. Mücket, S. Gottlöber
Velocity dispersion profile in dark matter halos
Astrophys. J., **602** (2004) 162 (astro-ph/0311083)

V.Müller, C.Maulbetsch
Superclusters and Voids in the Sloan DSS
Proceedings IAU Coll. 195 *Outside of Galaxy Clusters: intense life in the suburbs*, ed. A.Diaferio 2004

Ruediger G., Shalybkov D.
Linear instability of magnetic Taylor-Couette flow with Hall effect
Phys. Rev. E **69** (2004) 01630

U. Ziegler
An ADI-based adaptive mesh Poisson solver for the MHD code NIRVANA
Comp. Phys. Commun. **157** (2004) 207

U. Ziegler
A central-costrained transport scheme for ideal magnetohydrodynamics
J. Comput. Phys. **196** (2004) 393