HESSISCHER BEIRAT FÜR HOCHLEISTUNGSRECHNEN

KONZEPT ZUM HOCHLEISTUNGSRECHNEN IN HESSEN FÜR DIE JAHRE 2003 – 2005

Dezember 2002

1 Einleitung

Die Bedeutung des Wissenschaftlichen Rechnens als einer Schlüsseltechnologie für nahezu alle Bereiche von Wissenschaft und Technik ist heutzutage unumstritten. Trotz der enormen Fortschritte im Bereich der Rechnertechnologie ist hierbei in vielen Anwendungsfällen der Einsatz von Hochleistungsrechnern unerlässlich, nicht zuletzt auch aufgrund der stetig wachsenden Komplexität der Problemstellungen. Das Hochleistungsrechnen trägt entscheidend zur Reduktion von Entwicklungszeiten neuer Methoden und Technologien bei und es können Probleme angegangen werden, die aufgrund der hohen Anforderungen an die Rechenleistung anderweitig einer numerischen Simulation nicht zugänglich sind.

Der Verfügbarkeit einer adäquaten Hochleistungsrechnerkapazität kommt damit für den Wissenschaftsstandort Hessen eine überaus wichtige Bedeutung zu, da dies die Grundlage für eine national und international konkurrenzfähige Forschung im Bereich des Wissenschaftlichen Rechnens darstellt. Mit der Inbetriebnahme des neuen hessischen Hochleistungsrechners Anfang 2002 an der TU Darmstadt wurde in einem ersten Schritt eine wichtige Voraussetzung für eine Grundversorgung an Hochleistungsrechnerkapazität geschaffen. Der künftig stark steigende Bedarf macht einen kontinuierlichen weiteren Ausbau dieser Kapazität erforderlich, um diese Grundversorgung auch in Zukunft sicher zu stellen.

Mit dem vorliegenden Konzept macht der *Hessische Beirat für Hochleistungsrechnen* einen Vorschlag für die konkrete Umsetzung des weiteren Ausbaus der Hochleistungsrechnerkapazität in Hessen für die Jahre 2003–2005, der der Bedarfssituation und den organisatorischen und finanziellen Rahmenbedingungen Rechnung trägt.

Der Hochleistungsrechenbedarf soll weiterhin durch zentrale Rechenkapazitäten auf Landesebene gedeckt werden. Aufgrund unterschiedlicher Anforderungen der Nutzer besteht hierbei Bedarf für zwei unterschiedliche Architekturen:

- ein System für Anwendungen mit feingranularer Parallelität mit vergleichsweise hohen Anforderungen an die Kommunikationsleistung (Hauptnutzer: aus Natur- und Ingenieurwissenschaften)
- ein System für Anwendungen mit grobgranularer Parallelität mit vergleichsweise geringen Anforderungen an die Kommunikationsleistung (Hauptnutzer: aus Naturwissenschaften, Informatik und Wirtschaftswissenschaften)

Dieser Bedarf soll durch zwei entsprechende zentrale Systeme unterschiedlicher Architektur abgedeckt werden:

- Das SMP-Cluster an der TU Darmstadt unter Verantwortung des *Darmstädter Zentrums für Wissenschaftliches Rechnen (DZWR)*, das Anfang 2004 um eine 2. Ausbaustufe erweitert werden soll.

- Ein MPP-Cluster an der Universität Frankfurt unter Verantwortung des Frankfurter *Center for Scientific Computing (CSC)*, das in zwei Ausbaustufen Ende 2002 und 2003 neu beschafft werden soll.

Beide Rechnersysteme sind für Nutzer aller hessischen Universitäten zugänglich. Die Rechenkontingente auf beiden Systemen werden jeweils nach der finanziellen Beteiligung verteilt, wobei Landesmittel zu gleichen Teilen auf alle umgelegt werden (siehe Abschnitt Finanzierung).

Das Konzept, welches nachfolgend im Detail erläutert ist, wird einvernehmlich von allen hessischen Universitäten unterstützt. Die Eckpunkte sollen im Rahmen der Zielvereinbarungen des HMWK mit den Universitäten festgeschrieben werden.

2 Bedarfssituation

Der hessische Beirat für Hochleistungsrechnen hat zur Kenntnis genommen, dass an den hessischen Universitäten ein wachsender Bedarf an zentraler Hochleistungsrechenkapazität besteht. Dies gilt insbesondere im Hinblick auf eine national und international konkurrenzfähige Forschung, aber auch im Sinne einer modernen, zukunftsorientierten Ausbildung. Die Möglichkeit der Nutzung zentraler Hochleistungsrechner ist von essentieller Bedeutung für den Erfolg von Sonderforschungsbereichen, Graduiertenkollegs und einer Vielzahl unterschiedlicher Einzelprojekte. Durch die im Rahmen des Generationenwechsels an den hessischen Universitäten anstehenden Neuberufungen ist abzusehen, dass die Nachfrage nach zentraler Rechenleistung in den folgenden Jahren drastisch steigen wird.

Die Dringlichkeit des Bedarfs an zentraler Rechenleistung wird dadurch unterstrichen, dass verschiedene Fachbereiche bzw. einzelne Forschergruppen der hessischen Universitäten Eigenbeteiligungen zur Finanzierung von Hochleistungsrechnern zugesagt haben. Das führte im Jahr 2001 zur Beschaffung des SMP-Clusters an der TU Darmstadt, der seit Beginn des Jahres 2002 den hessischen Universitäten zur Verfügung steht und bereits jetzt überlastet ist.
Parallel dazu entwickelte sich an der Goethe Universität Frankfurt eine Initiative zur Gründung des *Center for Scientific Computing*, in dem sich 20 Arbeitsgruppen aus den Fachbereichen Physik, Chemie, Informatik und Bioinformatik, Geowissenschaften, Mathematik, Wirtschaftswissenschaften, sowie das Hochschulrechenzentrum der Universität zusammengeschlossen haben mit dem Ziel, einen wesentlichen Anteil an Berufungsmitteln zu bündeln und als Eigenbeteiligung zur Finanzierung eines zweiten Hochleistungsrechners am Standort Frankfurt zur Verfügung zu stellen.

Das wissenschaftlich breit gefächerte Interesse an zentraler Rechenleistung bedingt eine Nachfrage nach unterschiedlichen, den Aufgabenstellungen angepassten Rechnerarchitekturen:

- Applikationen, die eine Parallelverarbeitung auf SMP-Systemen unterstützen (z.B. Strukturrechnungen komplexer Quantensysteme).
- Monte-Carlo ähnliche Anwendungen, die von einer hohen Zahl an Einzelprozessoren profitieren und in der Regel ohne nennenswerte Prozessorkommunikation auskommen (z.B. Vielteilchendynamik in Stoßprozessen).

Mit der Installation des HLR Darmstadt wurde die SMP-Architektur als eine Säule des hessischen HLR Konzeptes bereits realisiert. In den Jahren 2002/2003 wird das CSC am Standort Frankfurt ein MPP-Cluster aus 360 leistungsstarken Einzelknoten aufbauen und somit eine zu Darmstadt alternative Architektur unterstützen.

Die Befürwortung zweier unterschiedlicher Architekturen im gemeinsamen hessischen HLR Netz hat neben der wissenschaftlichen Begründung auch eine ökonomische Komponente. Die immer noch rasante technologische Entwicklung auf dem Rechnermarkt lässt nicht voraussehen, welche Architektur in Zukunft am ehesten den Anforderungen an einen wissenschaftlichen Rechenbetrieb genügen wird. Im Sinne der Nachhaltigkeit der Investitionen der öffentlichen Hand ist deshalb ein paralleler Betrieb beider Architekturen von Vorteil.

3 Technische Zielvorstellungen

Anfang 2002 wurde in Darmstadt der neue hessische HLR in Betrieb genommen. Die eingesetzte Architektur ist ein nachrichtengekoppeltes System von drei SMP-Knoten (Regatta) der Firma IBM mit jeweils 32 Power 4 Prozessoren. In der zweiten Ausbaustufe (Anfang 2004) ist die Erweiterung des SMP-Systems geplant. Ziel ist mit der dann aktuellen Prozessortechnologie, eine Verdreifachung der Leistung und des Speicherplatzes der ersten Ausbaustufe zu erreichen.

Die geplante Cluster Struktur in Frankfurt zeichnet sich durch viele leistungsfähige Einzelknoten aus, bestehend aus jeweils zwei Intel basierten Prozessoren mit gemeinsamen Speicher. Durch Verwendung von Standardkomponenten ergibt sich ein sehr günstiges Preis-Leistungs- Verhältnis pro Knoten. Die starke Modularisierung zu preiswerten Einheiten erlaubt einen flexiblen Ausbau des Clusters, abhängig von dem zur Verfügung stehenden finanziellen Rahmen.

3.1 Auswahlkriterien

Wichtige Kriterien für die Auswahl an beiden Standorten sind:

- eine hohe Rechenleistung pro Knoten,
- eine ausgewogene Rechen-, Speicher- und I/O-Kapazität,
- eine angemessene periphere Infrastruktur (Hard- und Software) für einen effizienten Rechenzentrumsbetrieb,
- eine leistungsfähige, skalierbare und hochverfügbare Plattenperipherie,
  - Plattentyp und -kapazität für temporäre Datenhaltung der Nutzerdaten, sehr schneller Zugriff (hoch performant)
  - Plattentyp und -kapazität für mittelfristige Datenhaltung und Analyse der Nutzerdaten, schneller Zugriff, hohe Ausfallsicherheit

  - Plattentyp und -kapazität für langfristige Datenhaltung der Nutzerdaten (Archivierung)

- Datensicherung und Archivierung durch Einbindung in die vorhandene Infrastruktur der jeweiligen Rechenzentren,
- schnelle Netzanbindung der Systeme an das jeweilige Universitätsnetz und das G-WIN (Gigabit Ethernet oder ATM). Darüber hinaus besteht eine schnelle Netzverbindung (z.Z. 100 Mbit/s) zwischen den beiden Standorten TUD und JWGU im Rahmen des Südhessen-Wissenschaftsnetzes,
- Betriebssystem, Filesystem (max. Grösse des Filesystems, max. Grösse eines Files) mit Systemerweiterungen für
  - Accounting
  - Prioritätensteuerung
  - leistungsfähige und erweiterbare Batchverarbeitung
- verfügbare Standard-Software:
  - Entwicklungswerkzeuge
  - Compiler (C, C++, Fortran 95)
  - Bibliotheken (Numerik, MPI, CERN, NAG)
  - Debugging-Tools
  - Performance-/Analyse-Tools
- Fernwartbarkeit des Systems (Remote-Diagnose),
- erforderliche technische Infrastruktur (Leistung: Strom, Klima, Fläche,....),
- Wartungskosten (Hardware, Betriebssystem und Basis-Software),
- Referenz-Installationen,
- Personelle Unterstützung durch den Hersteller bei Installation, Einbindung in die Infrastruktur, Schulung Support-Konzept, Vorortbetreuung.

3.2 Realisierung

3.2.1 Standort Darmstadt

Für den Standort Darmstadt wurde nach ausführlicher Diskussion mit den potentiellen Anwendern ein Cluster aus 3 SMP-Systemen mit je 32 CPUs ausgewählt. Es handelt sich um IBM "pSeries 690 Regatta" mit POWER4 Prozessoren mit 1.3 GHz Taktfrequenz. Jeder Knoten ist mit 64 GB Haupt- speicher ausgestattet. Die Knoten sind zunächst untereinander und mit dem Fileserver mit je 2 Gigabit Ethernet Kanälen "Etherchannel" vernetzt. Der Einsatz einer IBM "Switch2" Verbindung ist noch für 2002 geplant.

Die nominelle Spitzenleistung des Systems liegt bei 0.5 Tflop/s. Im für die TOP500 Liste maßgeblichen LINPACK Benchmark wird eine tatsächliche Rechenleistung von 0.234 Tflop/s erreicht. Die ver- besserte Kopplung der Knoten mit dem "Switch2" läßt eine weitere Erhöhung dieses Wertes erwarten.

Zwei der Knoten stehen komplett für Rechnungen im Batch-Betrieb zur Verfügung. Der dritte Knoten ist logisch in zwei Systeme partitioniert. Die größere mit 28 CPUs und 56 GB Hauptspeicher rechnet ebenfalls im Stapel-Betrieb. 4 CPUs mit 8 GB Hauptspeicher dienen zum inter- aktiven Arbeiten, vor allem zur Programm-Entwicklung und Optimierung.

Der Fileserver besteht aus zwei identischen Knoten mit IBM RS64-Prozessoren in einer hochverfügbaren Konfiguration. Es stehen zunächst 1.5 TB Plattenplatz für den HHLR zur Verfügung. Zusätzlich sind auf jedem Knoten 0.25 TB hochperformante lokale Platten für temporäre Daten vorhanden.

3.2.2 Standort Frankfurt

Für den Standort Frankfurt ist der Aufbau eines Clusters auf der Basis von PC Standardkomponenten mit Intel kompatibler CPU geplant. Als Betriebssystem soll Linux eingesetzt werden. Jeder Rechenknoten soll mit zwei CPU und 1 GB Hauptspeicher ausgestattet werden. Die Knoten sind untereinander durch zwei getrennte physikalische Netze verbunden. Um sowohl den Anforderungen an MPI-basierten Anwendungen als auch dem hohen Bedarf an skalarer Leistung (bzw. trivial parallelen Anwendungen) zu genügen, werden zwei verschiedene Qualitäten an Vernetzung angeboten. Ein Teil der Knoten wird über ein schnelles Myrinet2000 vernetzt, das wegen seiner geringen Latenzzeit ideal für clusterparallele Anwendungen geeignet ist. Die verbleibenden Knoten werden mit FastEthernet verbunden. Die Anbindung aller Knoten an die zentralen Fileserver erfolgt kaskadiert über FastEthernet und GigabitEthernet.
Der Fileservice soll zunächst über 1 TB verfügen, jedoch gut ausbaufähig sein. Durch zwei unabhängige Fileserver, die sich gegenseitig absichern, wird eine hohe Verfügbarkeit garantiert.
Die Installation des Systems erfolgt in mehreren Phasen:

- Die Erstinstallation in 2002 soll den Fileservice, die Vernetzung und 80 Zweiprozessor-SMP-Knoten umfassen. 32 Knoten werden mit dem leistungsfähigeren Myrinet2000 vernetzt. In dieser Ausbaustufe wird eine theoretische Maximalleistung von 0.5 Tflop/s erreicht. In dem für die Einstufung in die TOP500 Liste maßgeblichen LINPACK Benchmark tragen nur die 32 Myrinet Knoten mit einer Gesamtleistung von 0.11 Tflop/s bei. Das Investitionsvolumen beträgt ca 300.000 Euro.

- In der zweiten Stufe (2003) werden der Fileservice erweitert und hochverfügbar gemacht, sowie weitere Knoten installiert. Abhängig von den verfügbaren Investitionsmitteln ist ein Ausbau um 280 Knoten geplant. Lassen sich alle 360 Knoten idealerweise mit Myrinet2000 vernetzen, erreicht das Cluster eine Gesamtleistung von 1.2 Tflop/s (LINPACK Benchmark).

- Die folgenden Ausbaustufen (2004-2005) beinhalten eine Erweiterung der Knotenzahl und eine Anpassung an die zu erwartende technologische Entwicklung.

4 Betriebs- und Nutzungskonzept

Hinsichtlich der Organisation des Betriebs des Rechners sind institutionell die folgenden Einrichtungen involviert:

- der hessische Beirat für Hochleistungsrechnen,
- das Darmstädter Zentrum für wissenschaftliches Rechnen (DZWR),
- das Frankfurter Center for Scientific Computing (CSC),
- die Rechenzentren der Universitäten.

Dem Darmstädter DZWR und dem Frankfurter CSC kommt die Funktion eines Kompetenzzentrums im Bereich des Hochleistungsrechnens zu. Die notwendige fachübergreifende Kompetenz ist durch die interdisziplinäre Zusammensetzung der beiden Zentren, sowie durch entsprechende Aktivitäten der Mitglieder in Forschung und Lehre gewährleistet. Damit werden beide Zentren einen Beitrag zu den folgenden Aufgaben leisten:

- Entwicklung bzw. Weiterentwicklung von Anwendungssoftware für den Hochleistungsrechner in den verschiedenen Anwendungsbereichen,
- Unterstützung von Nutzern bei der Portierung von Anwendungssoftware.
- Ausbildung von wissenschaftlichem Nachwuchs im Bereich des Wissenschaftlichen Rechnens durch entsprechende Lehrangebote,
- Technologietransfer in die Industrie im Rahmen von Kooperationsprojekten,
- Organisation von regelmäßigen Benutzer-Kolloquien, die den Erfahrungsaustausch zwischen allen Nutzern des Rechners ermöglichen,
- Kontaktpflege und Zusammenarbeit mit anderen im Bereich des Hochleistungsrechnens tätigen Arbeitsgruppen im In- und Ausland (z.B. Workshops, Forschungsprojekte).

Die Rechenzentren der hessischen Universitäten betreiben den Rechner gemeinsam unter Federführung der Rechenzentren der TU Darmstadt und der JWGU Frankfurt, an denen die Rechner installiert sind. Zu den Aufgaben der Rechenzentren an den beiden Standorten gehören:

- Bereitstellung von Räumlichkeiten und der zugehörigen Infrastruktur,
- Administration und Operating (24-Stunden-Betrieb),
- Betriebssteuerung und Überwachung der Betriebsvorgaben,
- Fehlerverfolgung und -behebung
- Benutzerverwaltung

- Datensicherung

Die Rechner sind für Nutzer aller hessischen Universitäten zugänglich. Die einzelnen Rechenzeitkontingente richten sich vorrangig nach den finanziellen Beteiligungen der Hochschulen, Fachbereiche bzw. Fachgebiete. Dies wird durch eine entsprechende Prioritätenvergabe erreicht.

5 Finanzierung für die Jahre 2003 - 2005

Beide Systeme in Darmstadt und Frankfurt werden von den hessischen Universitäten gemeinsam finanziert, so dass sich jede Universität - gemäß ihrem Nutzungsschwerpunkt - an mindestens einem der beiden Systeme finanziell beteiligt. Sofern alle Universitäten ihren Finanzierungsanteil leisten, beteiligt sich auch das HMWK mit einem Förderbetrag von insgesamt 750.000 Euro für die Jahre 2003 bis 2005 an der Gesamtversorgung. In Anbetracht der unterschiedlichen Bedarfssituationen und der unterschiedlichen finanziellen Rahmenbedingungen an den hessischen Universitäten erscheint dem wissenschaftlichen Beirat für das Hochleistungsrechnen in Hessen dieser Zwang zur gemeinsamen Finanzierung in seiner kurz- und mittelfristigen Wirkung eher behindernd für die Forschung im Bereich des Wissenschaftlichen Rechnens und den darauf aufbauenden Disziplinen. Das Ministerium wird gebeten, diese Vorgabe zu überprüfen und die zugesagte Förderung auch dann zu gewähren, wenn sich nicht alle Universitäten sofort an der Finanzierung beteiligen.

Die Nutzungskontingente, die auf den HMWK-Anteil entfallen, werden auf alle Universitäten umgelegt, somit erhält jede Universität auf beiden Systemen ein Mindestkontingent von rd. 5 %. Sollten sich die Anforderungen der Universitäten im Laufe der Zeit verändern (z.B. durch Neuberufungen), ist eine Erhöhung des jeweiligen Kontingents einer Hochschule durch eine Aufstockung der finanziellen Beteiligung möglich. Die derzeit geplante Verteilung der Nutzungskontingente mit den entsprechenden Finanzierungsbeiträgen der Universitäten ist in der nachfolgenden Tabelle dargestellt. (Alle Beträge sind in 1000 Euro angegeben.)

a) SMP-Cluster Darmstadt (2. Ausbaustufe)

				Nutzungskontingent	
Institution	2003	2004	2005	nominal	real
TU Darmstadt	225	225	225	54.44%	60.4 %
JWGU Frankfurt	0	0	0	0.00%	6.1 %
JLU Gießen	30	30	30	7.26%	13.3 %
PU Marburg	0	50	50	8.06%	14.1 %
U Kassel	0	0	0	0.00%	6.1 %
HMWK	125	125	125	30.24%	0.0 %
Land gesamt	380	430	430		

b) MPP-Cluster Frankfurt

Institution	2002	2003	2004	2005	Nutzungskontingent nominal	real
TU Darmstadt	0	0	0	0	0.00%	4.8 %
JWGU Frankfurt	300	225	225	225	62.50 %	67.3 %
JLU Gießen	0	20	20	20	3.85 %	8.7 %
PU Marburg	0	0	0	0	0.00%	4.8 %
U Kassel	0	50	50	50	9.61 %	14.4 %
HMWK	0	125	125	125	24.04 %	0.0 %
Land gesamt		420	420	420		

Die für die 2. Ausbaustufe in Darmstadt und den Aufbau des Clusters in Frankfurt erforderlichen Bundesmittel werden über die Rahmenplanmaßnahme "Hessischer Hochleistungsrechner" der TUD sowie durch Bundesmittelvorwegabzug für die JWGU bereitgestellt.