Hochschulrechenzentrum (HRZ)
der Philipps-Universität Marburg

Hans-Meerwein-Straße
35032 Marburg
06.10.2003

Auszug aus den

Erläuterungen zum HBFG-Antrag
Linux-Cluster als Compute Server

Arbeitsgruppen aus den Fachbereichen Mathematik und Informatik, Physik, Chemie sowie Pharmazie der Philipps-Universität haben vereinbart, aus ihren Mitteln ein Linux-Cluster zu beschaffen (d.h. die Landesmittel im Rahmen des HBFG dazu bereitzustellen). Das Cluster soll vom HRZ betrieben werden und für folgende Zwecke zum Einsatz kommen:

- High-Performance-Computing: Anwendungen mit hohem Rechenzeit- und Speicherplatzbedarf in den Naturwissenschaften sowie der Mathematik und Informatik.
- Cluster und Grid Computing: Entwicklung und Evaluation von Methoden, Algorithmen und Werkzeugen in der Informatik.

Das Cluster soll – auf der Basis nachträglicher Finanzierungsbeteiligung - auch von allen anderen Anwendern aus der Philipps-Universität genutzt werden können. Entsprechend soll es in den Verbund hessischer Hochleistungsrechner eingebracht werden.

Nach Analyse der für den geplanten Einsatz am Markt verfügbaren Linux-Cluster liegt dem Antrag folgende Konfiguration zugrunde:

- 128 Knoten mit Dual AMD Opteron CPUs 1,6 GHz, 2 GB RAM und 120 GB Disk
- vernetzt über Gigabit-Ethernet für die Datenkommunikation
- und Fast-Ethernet für Service-Zwecke
- 1 Frontend für das System-Management inkl. 1 TB RAID-Array
- sowie Management-Software, Linux, Compiler und Runtime-Libraries

Der Anstoß zur Beschaffung eines Clusters an der Philipps-Universität Marburg ist von Prof. Freisleben (Fachbereich Mathematik und Informatik, vgl. 3.2) ausgegangen, auf ihn geht auch die Wahl Opteron-basierter Knoten zurück. Die Auswahlentscheidung (Abschnitt 2) ist in seiner Arbeitsgruppe erarbeitet worden, dabei wurden Anforderungen der anderen Arbeitsgruppen und des HRZ berücksichtigt.

Hochleistungsrechnen hat an der Philipps-Universität eine lange Tradition. Hauptnutzer waren in der Vergangenheit die Fachbereiche Chemie und Physik, jetzt sind durch Neuberufungen weitere Anwender hinzugekommen. Im HRZ gab es bisher folgende Compute Server:

- Vektorrechner CONVEX C230 mit 3 CPUs (Dez. 1989 – Aug. 1995)
- Parallelrechner IBM SP mit 35 Power2-Knoten (Dez. 1995 – März 2003, 1995/96 in der TOP500-Liste) und 8 Power3-Knoten (Aug. 1999 – Sept. 2003), integriert in den hess. Hochleistungsrechner-Verbund.

Darüber hinaus gibt es Fachbereichs-Cluster, z.B. ein PC-Cluster der Chemie mit 43 Dual-CPU-Knoten (seit 2001; Eigenbau; Hosting durch das HRZ) und ältere Alpha-Cluster in der Physik.

Schließlich wurde und wird erhebliche Rechenkapazität von Marburger Wissenschaftlern auch außerhalb von Marburg in Anspruch genommen:

- Die Systeme des hess. Hochleistungsrechner-Verbunds (1992 – 2000) waren zentral finanziert, so dass sie aus Marburg ausgiebig genutzt werden konnten (insb. das leistungsfähigste System an der TU Darmstadt).
- Die gegenwärtigen Systeme an der TU Darmstadt und der Uni Frankfurt sind zum größten Teil von den Universitäten selbst zu finanzieren; hier hat sich die Philipps-Universität (2001/2002) beteiligt und wird sich auch weiterhin beteiligen (2003-2005).
- Schließlich werden bundesdeutsche Höchstleistungsrechner genutzt, in Stuttgart von der Chemie bzw. in Jülich von der Physik und Chemie.

Bzgl. der zentralen Server im HRZ sind gegenwärtig zwei weitere HBFG-Anträge unterwegs (vgl. 1.3), so dass mit dem vorliegenden Antrag das gesamte Spektrum abgedeckt ist:

- Der HBFG-Antrag „Ersatz des zentralen Backup/Archive-Systems“ vom 25.11.2002 ist am 31.07.2003 bewilligt worden. Im Rahmen dieser Maßnahme sollen noch in 2003 ein RISC-Server für TSM, ein Plattensystem mit 2 TB und eine Bandbibliothek mit über 100 TB in LTO2-Technologie beschafft werden.

- Der HBFG-Antrag „Ersatz der zentralen Server“ vom 31.07.2003 umfasst die Beschaffung von 5 RISC-Servern für Mail- und File-Services sowie 20 intel-Servern für andere zentrale Services (während 30 IBM/Sun/intel-Server noch weiter betrieben werden können). Gleichzeitig ist der Einstieg in die SAN-Technologie mit Plattensystemen im Umfang von 8 TB für allgemeine Zwecke vorgesehen, in die auch das Plattensystem des Backup/Archive-Systems integriert werden soll.

Beim ersten Antrag steht RISC für IBM, beim zweiten für Sun; z.Zt. werden vom HRZ sowohl IBM-Server unter AIX als auch Sun-Server unter Solaris betrieben. Wenn die Ausschreibungen zeigen, dass die Reduzierung auf nur einen Hersteller wirtschaftlich zu vertreten ist, soll diese Konsolidierung erfolgen. Linux als Betriebssystem auf den intel-Servern wird in jedem Fall weiter unterstützt, z.Zt. auf der Basis von Debian GNU/Linux.

1. Zu den Geräten

1.1 Spezifikation der beantragten Geräte

Anz	Gerätebezeichnung, Leistungsangaben Leistungen	Einzelpreis in €	Gesamtpreis in €
128	Megware **Clusternode** Dual Opteron Dual AMD Opteron 242 1,6 GHz Mainboard Tyan K8S S2880 2 GB DDR RAM reg. ECC PC2700 DDR333 Festplatte Hitachi 180GXP 120 GB EIDE 7200 rpm 2 x Gigabit-Ethernet 1000TX onboard	2.600,00	332.800,00
1	Megware Cluster **Frontend** Dual Opteron Dual AMD Opteron 242 1,6 GHz Mainboard Tyan K8S S2880 4 GB DDR RAM reg. ECC PC2700 DDR333 Dual Channel U320 SCSI onboard Externes RAID System TripleStor 7 x 180 GB Hitachi Festplatten, 1,08 TB netto bei RAID 5 2 x Gigabit-Ethernet 1000TX onboard Tastatur, Maus, 17" TFT Display, DVD ROM	8.490,00	8.490,00
1	**Interprozessnetzwerk**: Gigabit-Ethernet Switch HP Procurve 9315m 15 Slot Chassis Management-Modul 8 x 16 Port 1000TX Gigabit (= 128 Ports)	114.670,00	114.670,00
1	**Servicenetzwerk**: Fast-Ethernet Switch HP Procurve 4108gl 8 Slot Chassis 6 x 24 Port 10/100 Mbit/s Ethernet Module (=144 Ports) 1 x 6 Port 1000TX Gigabit Modul	8.100,00	8.100,00
4	Megware SlashTwo **Schrank**, BxHxT 850x2200x800 für maximal 40 Nodes in 20 SlashTwo Gehäusen sowie 19" Ebene für Switches	2.930,00	11.720,00
11	**Stromverteilung, Überwachung, Steuerung** Megware ClustSafe 12 Port Switch intelligenter Mehrfachschalter zur Cluster-Konfiguration	960,00	10.560,00
1	**Management-Software** ClustWare Software für Remote Management der Nodes im Cluster	5.980,00	5.980,00
1	Portland **Compiler** 64 bit für Opteron Cluster Development Kit bis 256 CPUs 2 User AMD64 Compiler Fortran, C, C++	7.340,00	7.340,00
1	Betriebssystem Debian GNU/**Linux** Parallele Bibliothek MPICH, Jobscheduling OpenPBS	0,00	0,00
1	Vorinstallation für 7-Tage-Dauertest und **Installation** vor Ort	9.800,00	9.800,00
1	3 Jahre **Garantie** auf das gesamte System, 2 Spare Nodes	21.280,00	21.280,00
1	3 Jahre **Wartung** und Support für das gesamte System	7.500,00	7.500,00
	Summe netto		538.240,00
	Summe inkl. 16 % MwSt		**624.358,40**

Die Spezifikation beruht auf der Preisauskunft von Megware (Chemnitz); für die Beschaffung wird eine EU-Ausschreibung erforderlich (vgl. 1.5). Erst dann soll im Rahmen des Finanzvolumens über die endgültige Konfiguration (insb. bzgl. Arbeitspeicher, Netzwerk) entschieden werden.

1.2 Geräte, die durch die Beschaffung ersetzt werden sollen

IBM SP (genauer: RS/6000 Scalable POWERparallel System SP); 3 Frames mit RS/6000 Knoten; ein Spezial-Switch verbindet jeden Knoten mit jedem anderen Knoten:

- 43 Knoten insgesamt; darunter
  16 Power2 Thin Nodes 2 mit jeweils 128 MB Arbeitsspeicher und 2.2 GB Plattenspeicher
  14 Power2 Thin Nodes 2 mit jeweils 256 MB Arbeitsspeicher und 4.5 GB Plattenspeicher, davon 10 mit 2 MB Level 2 Cache
  4 Power2 Thin Nodes 2 mit jeweils 2 MB Level 2 Cache, 512 MB Arbeitsspeicher und 2 * 4.5 GB Plattenspeicher für System, Paging, /tmp
  1 Power2 Wide Node mit 512 MB Arbeitsspeicher und 4 * 4.5 GB Plattenspeicher für Benutzer-Filesysteme
  8 Power3 Thin Nodes mit jeweils 2 Prozessoren, 8 GB Arbeitsspeicher, 2 * 18.2 GB Plattenspeicher und 4 MB Level 2 Cache
- 22.1 GFLOP/s Peak Performance
  (je Power2 Node 266.6 MFLOP/s, je Power3 Node zweimal 800 MFLOP/s)
- 40.2 GB Arbeitsspeicher und
- 443.4 GB interner SCSI Plattenspeicher
- SP Switch: Multi-Stage-Switch (voll vermaschte Schaltung von 4 x 4 Switch-Chips in Stufen), redundante Verbindung von jedem Knoten mit jedem, bidirektional, 8 Bit parallel, max. 150 MB/s je Richtung.
- 36.0 GB (8 * 4.5 GB) externer SSA Plattenspeicher: Benutzer-Filesysteme
- Control Workstation RS/6000 Mod. 7043-140
- Backup Control Workstation RS/6000 Mod. 43P

In der TOP500 Supercomputer-Liste hat die IBM SP im Nov. 1995 Rang 172 belegt, im Juni 1996 Rang 216 und im Nov. 1996 Rang 272; im Juni 1997 war sie bereits ausgeschieden.

Die Power2-Knoten, der Switch und die Peripherie sind in 1995 für ca. 2,5 Mio. DM beschafft worden, die Power3-Knoten in 1999 für ca. 678 TDM. Die erforderlichen Landesmittel sind vollständig aus den Investitionsmitteln des HRZ aufgebracht worden.

Im Vergleich dazu: Die CONVEX C230, der erste Compute-Server (1989-95), hatte in 1989 ca. 3,37 Mio. DM gekostet. Zu den erforderlichen Landesmitteln waren vom Fachbereich Chemie (Prof. Reetz, Leibnizpreisträger 1989) 750 TDM beigesteuert worden.