SYSTEM

In HWW installiert: CRAY T3E

intel Paragon MPP: System wurde verdoppelt

In HWW installiert: CRAY T3E

Alfred Geiger

Das massiv-parallele System Cray T3E mit insgesamt 512 Rechenknoten wird derzeit bei der Höchstleistungsrechner für Wissenschaft und Wirtschaft Betriebs GmbH (HWW) in Untertürkheim installiert. Mit einer Peak-Performance von 307 GFLOP/s und einem Hauptspeicher von 65 GB ist diese Maschine nun das Flaggschiff der bei HWW installierten Rechner.

Die Installation der Cray T3E begann bereits im Oktober diesen Jahres mit einem 128 Knoten-System. In der ersten Dezemberwoche erfolgte der Ausbau auf 256 Knoten. Anfang Januar wird das System auf die vollen 512 Knoten aufgerüstet werden, wobei das System die ersten beiden Wochen dann nur als 2*256 Knoten-Maschine benutzt werden kann. Die volle Leistung für die Einzelapplikation dürfte ab Anfang Februar 1997 zur Verfügung stehen.

Im Rahmen des HWW-Konzeptes soll die T3E die Rolle einer Maschine für Anwendungen mit höchsten Leistungsanforderungen übernehmen, bei denen auch ein höherer Programmieraufwand zu rechtfertigen ist.

In diesem Artikel soll das System nur kurz beschrieben und seine Möglichkeiten und Stärken erläutert werden. In den folgenden Ausgaben der BI. wird dann gezielter auf die speziellen Features des Systems eingegangen. Sie können aber bereits jetzt eine Reihe von Informationen über den Umgang mit dem System im WWW finden:
http://www.hlrs.de

Besonders möchten wir die technische Beschreibung der Maschine von Wilfried Oed (Cray Research) ans Herz legen, die Sie über die Einleitungsseite zur T3E erreichen können. Auch sämtliche Handbücher

zur T3E hängen bereits im WWW.

Technischer Überblick

Ähnlich wie beim Vorgängermodell Cray T3D handelt es sich bei der T3E um ein Sy-stem mit hardwaremässig verteiltem, aber logisch globalem Hauptspeicher. Anders als das Vorgängermodell benötigt die T3E kein Front-End System mehr. Für den Benutzer sieht die Maschine deshalb sehr ähnlich aus wie die intel Paragon.

Architektur

Wie bei allen Systemen mit physikalisch verteiltem Hauptspeicher ist der wichtigste Teil der Architektur das Verbindungsnetzwerk zwischen den Knoten.

Beim Design der Cray T3E wurde besonderer Wert darauf gelegt, das Verbindungsnetzwerk so zu gestalten, daß es prinzipiell keine Skalierungslimits kennt (angeboten werden Systeme bis 2048 Knoten), aber insbesondere daß die Kommunikationsparameter so hervorragend sind, daß die Verwendung der vollen Rechenleistung für eine Einzelapplikation mglich wird. Im Gegensatz zu der im URS-Bereich installierten IBM SP/2 handelt es sich hier also nicht um eine durchsatz- sondern um eine leistungsoptimierte Maschine.

Wie bereits bei der Cray T3D handelt es sich beim Verbindungsnetzwerk um einen dreidimensionalen Torus. Im Gegensatz zur T3D hat bei der T3E aber jeder Knoten seinen eigenen Router-Zugang. Alle Links sind bidirektional ausgeführt und laufen mit einer Geschwindigkeit von 500 MB/s. Dies ist der derzeit beste auf dem Markt erhältliche Wert.

Nutzer der Cray T3D haben bisher unter dem Problem einer sehr inflexiblen Partitionierung gelitten, da nur Partitionen mit einer Zweierpotenz von Prozessoren allokiert werden konnten, die noch dazu zusammenhängend angeordnet sein mußten. Bei der T3E kann deutlich flexibler allokiert werden. Die Knotenzahl der Partition kann beliebig sein, allerdings muß die Form der Partition immer noch konvex sein. Ein Zusammensammeln beliebig über die gesamte Maschine verstreuter Knoten für eine Partition, wie bei der intel Paragon, ist bei der T3E leider nicht mö glich. Dies kann für den Benutzer zu dem leidigen Effekt führen, daß zwar die von ihm gewünschte Anzahl von Prozessoren in der Maschine noch frei ist, er diese jedoch nicht erhalten kann, weil sie nicht in einem zusammenhängenden Gebiet liegen. Spätere Releases des Betriebssystems werden jedoch die Mööglichkeit zum Verschieben von Partitionen enthalten. Um dennoch der Gefahr vorzubeugen, daß sich die Maschine im Laufe der Zeit in immer kleinere Partitionen zerlegt, werden wir das System so fahren, daß Jobs, die eine hohe Zahl von Knoten benutzen können, bevorzugt werden. Ein Übungs-, Entwicklungs- und Debug-Betrieb auf der Maschine wird somit nur sehr eingeschränkt mö glich sein. Aus diesem Grund wird für solche Zwecke die intel Paragon erhalten (vgl. Artikel ab Seite 8 in dieser Ausgabe).

Knoten

Auf dem Knoten befinden sich neben dem Prozessor vom Typ DEC 21164 und dem Hauptspeicher von je 128 MB noch weitere interessante Features, die insbesondere für den schnellen Memory-Transfer zwischen den Knoten wichtig sind. Auf diese soll in späteren Artikeln eingegangen werden.

Der Knoten wird mit 300 Mhz getaktet. Aufgrund der beiden Floating-Point Pipes des DEC 21164 ergibt dies eine Peak-Performance pro Knoten von 600 MFLOP/s. Der Datenbus auf dem Knoten hat dabei eine Bandbreite von 1.2 GB/s.

Der Cache des 21164 ist, wie bereits beim in der T3D eingesetzten Vorgänger 21064, bezogen auf die Leistung recht klein: 8kB L1 und 96kB L2 Cache. Um einen teuren zusätzlichen Off-Chip Cache zu

vermeiden hat Cray ein Stream-Buffer-Konzept eingeführt, das das schnelle Laden von Vektoren erlaubt. Sobald damit erste Erfahrungen vorliegen, werden wir berichten.

I/O

Neben dem Verbindungsnetzwerk für den Nachrichtenaustausch der Anwendung, hat jeder Knoten noch Zugang zu einem I/O-Verbindungsnetzwerk. Dies ist in doppelt ausgelegter und bidirektionaler SCI-Technologie (Scalable Coherent Interface, ein IEEE-Standard) ausgeführt, dem von Cray sogenannten Gigaring. Jeder Gigaring hat eine bi- direktionale Bandbreite von 600 MB/s. Die bei der HWW installierte Maschine wird im Endausbau 13 solcher Ringe haben, im Augenblick allerdings nur zwei.

An die Gigaringe sind I/O-Knoten angeschlossen, die den Übergang auf Disk-Systeme (HWW-Maschine: ca. 500 GB) und Netzwerke (HiPPI, ATM, FDDI) realisieren.

Über die Anzahl der Gigaringe ist auch die I/O-Leistung der Maschine skalierbar.

Betriebssystem

Das Betriebssystem, UNICOS/MK, ist ein microkernel-basiertes UNIX-Derivat. Die Arbeitsweise und Funktionalität ist sehr ähnlich dem OSF/1-Betriebssystem der intel Paragon. Im Gegensatz zu diesem basiert es jedoch nicht auf dem MACH-Microkernel der Carnegie-Mallon University, sondern auf dem von Chorus-Systems in Grenoble entwickelten CHORUS-Microkernel. Dieser hat aufgrund seiner Echtzeitfähigkeiten und seines konsistenteren Designs klare Vorteile gegenüber dem MACH-Kernel.

Auf jedem Knoten der T3E läuft ein Mikrokernel, die UNIX-Server hingegen nur auf wenigen Systemknoten, auch hier also eine große Ähnlichkeit zur Paragon.

Programmiermodelle und Compiler

Es stehen Compiler für FORTRAN 90, C und C++ zur Verfügung. Diese Compiler sind, wie auf Distributed-Memory-Maschinen üblich, Knotencompiler.

Als parallele Programmiermodelle stehen Message-Passing (MPI, PVM, shmlib) und HPF (High Performance Fortran) bereits heute zur Verfügung. Das Cray-proprietäre und von der T3D bekannte Programmiermodell CRAFT wird zu einem späteren Zeitpunkt als Erweiterung von HPF zur Verfügung stehen.

Mit shmlib stellt Cray bereits, im Vorgriff auf den Standard MPI-2, eine Schnittstelle für einseitiges Message-Passing (auch Remote-Copy genannt) zur Verfügung. Was bei der Benutzung der einzelnen Programmiermodelle zu beachten ist, erfahren Sie auf immer aktuellem Stand in unseren WWW-Seiten: http://www.hlrs.de

Zugang

Die Maschine ist im Netz unter dem Namen hwwt3e.hww.de erreichbar. Sie ist für alle Kunden der HWW zugänglich und wird im akademischen Bereich durch RUS/HLRS bundesweit angeboten. Der Zugriff erfolgt über die an der HWW üblichen Mechanismen Secure Shell (ssh) und Secure Copy (scp). Der Batch-Zugang wird im Laufe der nächsten Wochen etabliert werden.

Erste Leistungsmessungen

Zunächst wurden Messungen der Kommunikationsleistung zwischen den Knoten durchgeführt. Da wir

in Stuttgart eines der ersten Systeme haben und an Hard- und Software noch ständig Upgrades durchgeführt werden, können die hier angegebenen Werte aber zum Zeitpunkt der Drucklegung dieses Artikels bereits (im positiven Sinne) überholt sein.

Die Messung der Bandbreite zwischen zwei beliebigen Knoten mit Remote-Memory Copy führte auf Werte im Bereich von 300 MB/s und bei MPI je nach MPI-Call auf Werte zwischen 70 und 300 MB/s. Im Vergleich zu der Peak-Performance von 500 MB/s ist das zwar nicht berauschend, dennoch handelt es sich um die besten jemals in einem solchen System erzielten Werte.

Die Latenzzeiten halten sich für Remote-Memory Copy im Bereich um 1 Mikrosekunde und für MPI im Bereich um 10 Mikrosekunden. Auch das Bestzeiten.

Messungen auf dem Einzelknoten brachten die für DEC-Prozessoren typischen 10-20 % der Peak-Performance. Ähnlich wie beim Verbindungsnetzwerk muß man auch hier sagen, daß die Werte absolut gesehen sehr gut liegen, nicht jedoch im Vergleich zur Peak-Performance. Von Prozessoren wie intel's Pentium Pro und IBM's Power 2+ ist man hier eine höhere Ausbeute der Peak-Performance gewohnt.

Von Seiten des HLRS wurden bisher drei parallele Codes auf die Maschine gebracht, alle aus dem Bereich der numerischen Strö mungsmechanik. Es wurde die Skalierbarkeit dieser Codes über die gesamte zum Testzeitpunkt zur Verfügung stehende Knotenzahl (128) getestet. Bei zwei aus dem Bereich der Universität Stuttgart stam-menden Codes (Ein Verbrennungscode des IVD und ein Code zur Berechnung reagierender Strö mungen beim Wiedereintritt von Raumfahrzeugen des IRS (URANUS)) ist dies auch in hervorragender Weise gelungen. Beispielsweise konnte beim URANUS Code auf 118 Knoten auf Anhieb ein Speedup von ca. 80 erzielt werden, verbunden mit einer Absolutleistung, die bisher auf keiner Maschine erreicht werden konnte. Dieser Wert ist bemerkenswert, da bei gleichbleibender Gesamtproblemgröße die Granularität auf einer so großen Knotenzahl schon extrem fein wird.

Bei Codes mit hoher I/O-Last, wie dem kommerziellen CFD-Code STAR-CD hielten sich die Erfolge hingegen bisher in Grenzen, da, wie bereits oben geschildert, von 13 bestellten Gigaringen erst zwei installiert sind, damit also die I/O-Leistung erst einen Bruchteil des im Endausbau erwarteten Wertes erreicht.

Dr. Alfred Geiger, NA-5719
E-Mail: geiger@hlrs.de
Web: http://www.hlrs.de

---

intel Paragon MPP:
System wurde verdoppelt

Shannon Miller/Dirk Sihling/Alfred Geiger

Die intel Paragon war das erste massiv-parallele System, das an der Universität Stuttgart als für alle Institute allgemein zugänglicher Rechner 1992 beschafft wur-de. In der Zwischenzeit wurde eine Vielzahl wichtiger Codes aus verschiedenen Anwendungsgebieten auf dieses System gebracht, wobei sich die Paragon als ide-ale Plattform für die Entwicklung portabler Software erwies. Ohne den Einsatz der Paragon als Pathfinder System wäre ein Schritt in die Spitzenklasse heutiger Pa-rallelrechner, wie er jetzt mit der an anderer Stelle in dieser Ausgabe vorgestellten Cray T3E vollzogen wurde, niemals möglich gewesen.