HLRB II

Der neue Höchstleistungsrechner in Bayern: SGI Altix 4700

EIN ENORMER LEISTUNGSSCHUB FÜR DIE WISSENSCHAFT.

SGI

Abb. 1: Die neue SGI Altix 4700.

VON REINHOLD BADER UND MATTHIAS BREHM

Nach einer Betriebsdauer von sechseinhalb Jahren wird der im LRZ-Gebäude in der Innenstadt Münchens betriebene Höchstleistungsrechner in Bayern (HLRB I), eine Hitachi SR8000-F1, Ende Juni 2006 außer Betrieb genommen. Er wird durch den erheblich leistungsfähigeren HLRB II ersetzt. Dabei handelt es sich in der ersten Ausbaustufe um ein Altix 4700-System der Firma SGI mit 4096 Intel Itanium2-Prozessoren. Dieser Rechner wird im Laufe des Junis im obersten Stockwerk des LRZ-Neubaus in Garching installiert.

Die Leistungsdaten des neuen Systems sind bereits in der ersten Ausbaustufe imposant: Mit einer Spitzenrechenleistung von mehr als 26 Billionen Rechenoperationen *(Teraflops)* pro Sekunde wird Forschern aus ganz Deutschland am LRZ wieder ein Rechensystem mit europaweit konkurrenzfähiger Rechenleistung zur Verfügung stehen. Stellt man sich die Rechenoperationen als Nägel vor, die in einem Abstand von 1.5 mm einzuschlagen sind, so müsste man innerhalb einer Sekunde eine Strecke abarbeiten, die den Äquator 975mal umrundet! Aber auch die Größe des Hauptspeichers ist gigantisch: mehr als 17 Terabytes (das sind 17000 Gigabytes) werden sehr umfangreiche und neuartige Simulationen ermöglichen.

Die Vorzüge des neuen Rechners

Die besonderen Vorzüge des neuen Systems bestehen nicht nur in der obengenannten Spitzenrechenleistung, sondern auch in einer breiten Palette von Eigenschaften, deren Zusammenspiel eine sehr effiziente Nutzung des neuen Rechners ermöglicht. Die wichtigsten dieser Eigenschaften seien im Folgenden aufgezählt:

1. Das System ist in 16 Einheiten *(Partitionen)* mit jeweils 256 Prozessoren unterteilt, wobei jeder Partition ein logisch einheitlich ansprechbarer Hauptspeicher von 1 Terabyte zur Verfügung steht; kein anderes System in Europa weist derzeit diese spezielle Fähigkeit auf. In geeigneter Weise parallelisierte Programme können mehrere Partitionen gleichzeitig benutzen. Im Laufe des Betriebs und insbesondere mit der Installation der zweiten Ausbaustufe wird die Größe dieser Partitionen weiter wachsen.

2. Das System weist eine hohe aggregierte Bandbreite zum

Hauptspeicher auf, weil jedem Prozessor ein eigener Speicherkanal zur Verfügung steht. Damit sind datenintensive Simulationen sehr effizient durchführbar. Weil darüber hinaus jedem Prozessor ein 6 Megabytes großer schneller Cache-Speicher zur Verfügung steht, lassen sich manche Anwendungen sogar überproportional zur Zahl der verwendeten Prozessoren beschleunigen.

3. Der für die Ablage und Weiterverarbeitung von Daten verfügbare Hintergrundspeicher ist bezüglich Quantität und Qualität besonders performant ausgelegt worden: Es stehen für große Datensätze 300 Terabytes an Plattenplatz zur Verfügung (dies entspricht dem Inhalt von etwa 100 Milliarden voll beschriebenen DIN A4-Seiten). Die Daten können mit einer aggregierten Bandbreite von 20 Gigabytes/s gelesen oder geschrieben werden. Damit kann theoretisch der Hauptspeicherinhalt des Gesamtsystems innerhalb einer Viertelstunde auf die Platten herausgeschrieben werden. Da viele wissenschaftliche Programme in regelmäßigen Abständen Daten herausschreiben oder einlesen, wird hierdurch ein mitunter deutlicher Flaschenhals, der auf anderen Systemen zu leer stehenden Prozessoren führt, behoben.

4. Für die Benutzerverzeichnisse mit Programmquellen, Konfigurationsdateien usw. stehen weitere 40 Terabytes an extrem ausfallsicher ausgelegtem Plattenplatz zur Verfügung, auf den ein Zugriff auch von außerhalb des Systems möglich ist. Dieser Plattenbereich zeichnet sich durch hohe Transaktionsraten aus, so dass die effiziente Verarbeitung einer Vielzahl von kleinen Dateien gewährleistet ist.

5. Da das System aus Itanium2 Standard-Prozessoren der Firma Intel aufgebaut ist und als Betriebssystem das inzwischen weit verbreitete Linux verwendet wird, steht ein großes Spektrum an Standard-Softwarepaketen zur Verfügung, die ohne großen Portierungsaufwand auf dem System eingesetzt werden können. Für die von Forschern selbst erstellten Programme steht eine vollständige Entwicklungsumgebung zur Verfügung, die einen fast nahtlosen Übergang von Arbeitsplatzsystemen oder von Clustern auf den neuen Höchstleistungsrechner ermöglicht.

	SGI Altix 4700	Hitachi SR8000
Prozessoren	4096	1344
Spitzenleistung	26,2 Teraflops/s	2,0 Teraflop/s
Hauptspeicher	17,2 Terabytes	1,3 Terabytes
Speicherbandbreite	34,8 Terabytes/s	5,4 Terabytes/s
Plattenplatz	340 Terabytes	10 Terabytes
Latenz des Interconnects	1-6 Mikrosekunden	14 Mikrosekunden

Tabelle 1: Vergleich der Leistungen des alten und des neuen Höchstleistungsrechners.

Die folgende Tabelle gibt eine Übersicht über die wesentlichen Leistungszahlen des neuen Systems im Vergleich zum Vorgänger.
In der hier beschriebenen Konfiguration wird die Altix 4700 bis etwa Mitte 2007 betrieben; danach werden in einer zweiten Ausbaustufe alle Prozessoren durch ein Nachfolgemodell mit jeweils zwei Rechenkernen (statt einem) auf einem Prozessor-Knoten ersetzt.

Außerdem wird zusätzlicher Hauptspeicher und Plattenplatz installiert, sodass in der zweiten Ausbaustufe ein nahezu doppelt so leistungsfähiges System verfügbar sein wird; die vertraglich zugesicherte, durch Benchmark-Programme definierte Anwendungsleistung des Systems wird sich von 7 auf 13 Teraflops/s erhöhen.

Systemarchitektur

Die Systemarchitektur ist eine verteilte Shared-Memory-Architektur, das heißt: der gemeinsame Hauptspeicher ist über die Systemknoten verteilt. Memory-Controller auf den Systemknoten sorgen für den cache-kohärenten Zugriff aller Prozessoren auf diesen gemeinsamen Hauptspeicher. Je nachdem, ob ein Speicherzugriff auf physisch lokale oder auf einem anderen Systemknoten befindliche Daten erfolgt, ergeben sich jedoch unterschiedliche Zugriffszeiten und Bandbreiten. Daher wird die Systemarchitektur auch als *cache-coherent non-uniform memory access* (ccNUMA) bezeichnet. Die effiziente Nutzung eines derart ausgelegten Speichersystems stellt den Programmierer durchaus noch vor Herausforderungen, bietet aber auch große Flexibilität der Nutzung.

Systemknoten

Einzelne Systemknoten der Altix 4700 sind entweder mit Prozessoren ausgestattet oder es handelt sich um Ein/Ausgabe-Knoten. Alle Knotentypen sind in Form von Blades, einer flachen Bauform von Platinen mit gemeinsamer Strom- und Lüftungsversorgung, realisiert.. Diese Blades werden mittels des SGI-Numalink4-Interconnects zu einem Shared-Memory-System zusammengeschaltet. Ein Compute-Blade besteht aus einem Intel Itanium2-Prozessorchip und einem Memory-Controller, der den Prozessor mit dem physisch lokalen Hauptspeicher verbindet sowie zwei Numalink-Kanäle zur Anbindung an den Interconnect bereitstellt (Abb. 2).

Die Intel Itanium2-CPUs sind mit 1,6 GHz getaktet und haben zwei Multiply-Add-Einheiten. Damit ergibt sich pro Prozessor eine Spitzenleistung von 6,4 Gigaflops/s (6,4 Milliarden Gleitkomma-Operationen pro Sekunde). Jede CPU ist darüber hinaus mit 256 Kilobytes Level 2 Cache und 6 Megabytes Level 3 Cache ausgestattet; im Unterschied zum „normalen" Hauptspeicher laufen diese Caches mit der vollen Systemfrequenz, sodass solche Anwendungen, die ausreichend oft Daten aus dem Cache wiederverwenden können, eine sehr hohe Anwendungsleistung erzielen. Auf der am LRZ installierten Altix 4700 sind die Speicherbänke im Normalfall mit 4 Gigabytes pro Blade bestückt; auf der zur interaktiven Nutzung vorgesehenen Partition stehen sogar 8 Gigabytes pro Blade zur Verfügung. In Phase 2 wird neben dem Austausch der Itanium2-CPU durch einen Doppelkern-Prozessor auch zusätzlicher Hauptspeicher in jeden Knoten eingebaut, um so auch letzteren in etwa zu verdoppeln.

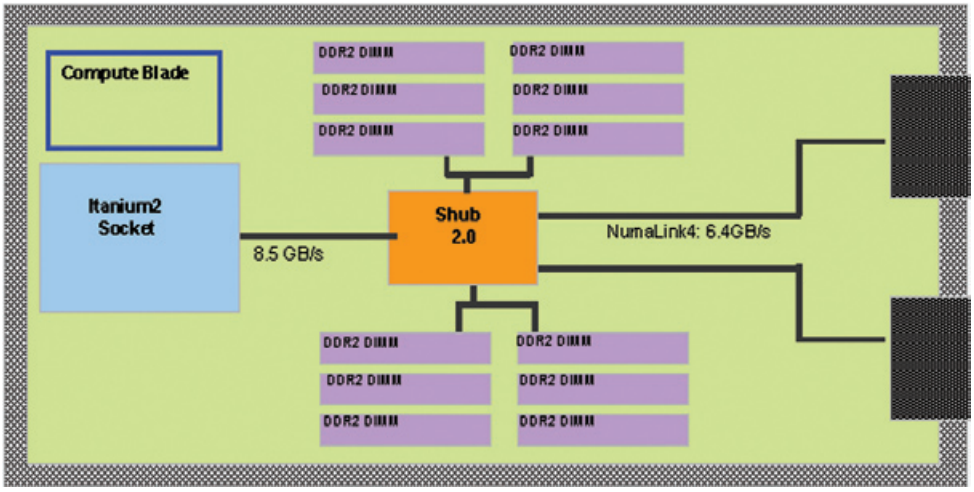

Abb. 2: Schematische Darstellung eines Systemknotens.

Ein Ein/Ausgabe-Knoten besteht aus einer Cache-Kohärenzschnittstelle (TIO-Chip) und einem ASIC, der die gängigen IO-Schnittstellen, wie zum Beispiel PCI-X oder PCI-Express, bereitstellt. Die Kohärenzschnittstelle erlaubt es, Daten cache-kohärent direkt von der IO-Schnittstelle (z.B. PCI-X-Karte) über den Numalink4-Interconnect in den verteilten Hauptspeicher auf den Prozessorknoten zu transportieren.

Aufbau des Interconnect

Der NumaLink4-Interconnect der Altix 4700 verbindet die einzelnen Blades miteinander. Er unterscheidet sich von handelsüblichen Netzwerk-Technologien dadurch, dass der Zustand des gesamten Hauptspeichers zu jeder Zeit für alle Prozessoren sichtbar ist. Darüber hinaus ist die Latenz beim Zugriff auf Daten auf anderen Blades gering. Der Interconnect besteht aus 8-Port-Routern, 8-Port-Metaroutern und Kabelverbindungen der Nodeboards mit den Routern sowie der Router mit den Metaroutern. Jede Kabelverbindung leistet 6,4 Gigabytes/s (3,2 Gigabytes/s je Richtung non-blocking). Die Router und Metarouter sind als non-blocking Crossbar-Einheiten realisiert und verfügen über acht NumaLink Ports (acht Eingänge und acht Ausgänge). Der Grundbaustein einer Partition ist ein Blade-Chassis, das über zehn Blade-Schächte verfügt, in die Prozessor- oder Ein/Ausgabe-Blades eingebracht werden können. Die NumaLink-Ports werden dabei mit der Backplane verbunden, die die Verbindung dieser zehn Blade-Slots untereinander und zu anderen Blades hin bereitstellt. Abb. 3 zeigt die Topologie einer Backplane eines Blade-Chassis.

Die 8-Port-Router verbinden jeweils vier Prozessor-Blades und (maximal) einen I/O-Blade untereinander und über eine zweistufige Hierarchie von Meta-Routern (hier nicht gezeigt) mit dem System-Interconnect. Diese zweistufige Hierarchie ist doppelt ausgelegt, sodass man die Topologie des Interconnects auch als „dual plane fat tree" bezeichnet. Auf ihrer Basis beruht die Kommunikation innerhalb einer 256-Prozessor-Partition. Für die Kommunikation zwischen unterschiedlichen Partitionen stehen ebenfalls NumaLink4-Verbindungen zur Verfügung. Diese sind jedoch nur als sog. „Mesh-Topologie" ausgelegt und weisen damit weniger Übertragungsbandbreite auf.

Software-Umgebung

Die SGI Altix 4700 wird unter einem Standard-Linux-Betriebssystem betrieben. Als Linux-Distribution kommt Novells SUSE Linux Enterprise Server (SLES 10) zum Einsatz. Für diese Distribution bietet SGI zusätzliche Software für den Einsatz auf großen Systemen in Rechenzentren an: Neben den HPC-Numa-Tools, dem Message Passing Toolkit und der Scientific Subroutine Library (SCSL) sind das z. B. Storage Management Software wie das XFS-Filesystem mit XVM Volume Manager, sowie deren Cluster-Erweiterungen CXFS und CXVM, Accounting-Pakete und der Performance Co-Pilot zur Systemüberwachung. Benutzern stehen zur ausgefeilten Performance-Messung an eigenen Programmen die SGI Tools histx, profile.pl sowie Speedshop zur Verfügung.

Compiler und Tools

Für die Generierung von optimalem Code aus Fortran-, C- oder C++-Quellen kommen die Compiler-Produkte der Firma Intel zum Einsatz; diese unterstützen die entsprechenden Sprach-Standards (Fortran 2003, C99 und ANSI C++) und sind in der Lage, die besonderen Eigenschaften des Itanium2-Prozessors hinsichtlich der hochgradig parallelen Ausführung von Instruktionen auszunutzen. Darüber hinaus wird auch die OpenMP-basierte parallele Programmierung mit Threads durch die Intel-Compiler konform zum OpenMP-Standard 2.5 unterstützt. Die C/C++-Compiler sind kompatibel zu den mit dem Betriebssystem mitkommenden C- und C++-Compilern der GNU Compiler Collection. Darüber hinaus ist auch Sprachmischung zwischen Fortran und C/C++ möglich.

Als Alternative zur Verwendung von mathematischen Funktionen der Linearen Algebra (BLAS, LAPACK und Fourier-Transformationen) in der SCSL ist es auch möglich, die Intel Math Kernel Library (MKL) zu verwenden. Diese stellt darüber hinaus auch schnelle Vektor-Versionen mathematischer Funktionen sowie Löser für dünn besetzte Matrizen (PARDISO) zur Verfügung. Die Integrated Performance Primitives-Bibliothek (IPP) stellt Codecs für die Audio, Film- und Bildverarbeitung, aber auch kryptographische Funktionen bereit. Zur Analyse des Laufzeit- und Kommunikationsverhaltens MPI-paralleler Programme dienen die Intel Tracing Tools (Vampir); für die Performance-Analyse serieller Programme auf der Basis der Itanium Hardware Performance Counter stellt Intel mit VTune ein Tool mit graphischem Interface bereit, das auf dem Login-Knoten des LRZ-Systems verfügbar sein wird. Zur Fehlersuche und Fehlerbehebung in Programmen stehen Debugger von Intel und Etnus (Totalview) zur Verfügung.

Rechenbetrieb

Der größte Teil des Altix 4700-Superclusters wird in der Regel über das Warteschlangensystem PBSPro der Firma Altair zugänglich sein, für interaktive Arbeiten, Entwickeln und Testen von Programmen sowie kleinere Produktionsläufe steht eine der sechzehn Partitionen teilweise zur Verfügung. Die Maximallaufzeit großer paralleler Programme wird im Normalfall auf etwa zwei Tage begrenzt sein; der Programmierer wird daher selber dafür sorgen, in regelmäßigen Abständen die für den Neustart seines Programms notwendigen Daten auf den Hintergrundspeicher zu schreiben. In Ausnahmefällen kann jedoch für einzelne Nutzer auch eine längere Programmlaufzeit zugelassen werden. Für solche Programme, die über Computational Steering-Mechanismen zur Laufzeit interaktiv kontrolliert werden sollen, kann man im Warteschlangensystem einen ausgewählten Startzeitpunkt sicherstellen, zu dem der Nutzer von seinem Arbeitsplatz aus das Programm bedienen kann.

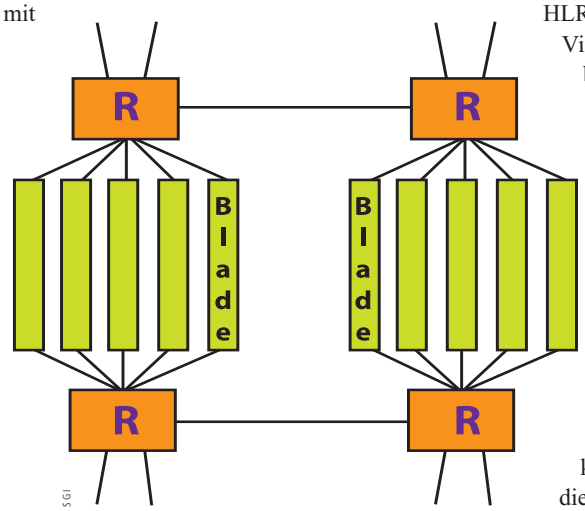

Abb. 3:
Jeder der hier gezeichneten 8-Port-Router (R) verbindet jeweils vier Prozessoren-Blades mit maximal einem I/O-Blade.

Nahtloser Übergang

Das LRZ hat sich von Anfang an bemüht, für die Anwender einen nahtlosen Übergang auf das neue System zu ermöglichen. Seit Juli 2005 betreibt es deshalb eine SGI Altix 3700 mit 64 Prozessoren als Migrationssystem. Dort ist im Wesentlichen die Softwareumgebung installiert, wie sie auch auf dem HLRB II verfügbar sein wird. Viele Nutzer haben dadurch bereits ihre Programme an das neue System anpassen können.Auch ein Wachstumspfad für künftige Programmentwicklungen ist vorhanden: Programme, die auf dem Linux-Cluster des LRZ und insbesondere auf der darin integrierten 128-Prozessor Altix 3700 Bx2 entwickelt wurden, können, wenn sie mit diesen Ressourcen nicht mehr auskommen, auf den neuen Höchstleistungsrechner gebracht werden. Zum ersten Mal in der Geschichte des LRZ steht damit eine Leistungspyramide mit einer weitgehend einheitlichen Programmier- und Prozessorumgebung zur Verfügung, beginnend beim Linux-PC auf dem Schreibtisch, über das Linux-Cluster und die 128-Prozessor Altix auf Landesebene, bis hinauf zum bundesweit genutzten neuen Höchstleistungsrechner in Bayern.

Dr. Reinhold Bader ist Mitarbeiter der Gruppe Hochleistungsrechnen am Leibniz-Rechenzentrum, Dr. Matthias Brehm leitet diese Gruppe.