**Neuer massiv-
paralleler Rechner**

Neuer massiv-
paralleler Rechner

Heinz W. Pöhlmann/Peter Haas

Mit dem neuen Supercomputer Hitachi SR2201 stellt das Rechenzentrum seinen Nutzern einen weiteren interessanten Parallelrechner mit Zukunftsperspektive zur Verfügung.

Im Rahmen einer Kooperationsvereinbarung stellt die Firma COMPAREX Informationssysteme GmbH, Mannheim, dem Rechenzentrum der Universität Stuttgart für zunächst neun Monate einen massiv-parallel arbeitenden Supercomputer modernster Technologie zur Verfügung. Der Parallelrechner vom Typ Hitachi SR2201/H32 mit 32 Hochleistungsrechenknoten, acht weiteren IO-Knoten, 10 GByte verteiltem Hauptspeicher und einer Peak-Performance von insgesamt 9,6 GFlop/s wurde vor wenigen Tagen am Rechenzentrum in Betrieb genommen und zeichnet sich insbesondere durch schnelle RISC-Prozessoren, jeweils großen lokalen Hauptspeicher und eine hohe Kommunikationsbandbreite im Verbindungsnetzwerk der einzelnen Knoten aus. Außerhalb Japans ist es weltweit die zweite Installation dieses Rechnertyps. Zusammen mit der Universität wird COMPAREX den Rechner evaluieren und die Markteinführung in Europa im technisch-wissenschaftlichen oder ggf. auch im kommerziellen Bereich vorbereiten.

Mit der Baureihe SR2201 stellt die Firma Hitachi, Tokyo, einen interessanten massiv-parallelen Rechnertyp vor, der die Vorzüge der RISC-Technologie mit den Eigenschaften der klassischen Vektorrechner verbindet. Dabei kommen ausgeklügelte Speicherhierarchien auf Mehrlagen-Keramikmodultechnik, dreidimensionale, parallele Crossbars und Magnetplattensysteme in Matrixanordnung zum Einsatz.

Die akademische Nutzung des Rechnersystems ist während der Dauer der Kooperationsvereinbarung kostenlos. Entsprechende Benutzeranträge finden Sie auf dem ftp-Server des RUS unter: `/pub/rus/betrieb/formulare`

In diesem Artikel wird das System mit seinen Mö glichkeiten und Stärken nur kurz beschrieben. Nachfolgende Ausgaben der BI. werden sich gezielt mit den neuartigen Eigenschaften dieses Rechnertyps beschäftigen. Die SR2201/H32 wird vom Personal des Höchstleistungsrechenzentrums Stuttgart (HLRS) betrieben.

Weitere Informationen über das System finden Sie in Kürze im WWW unter:
http://www.rus.uni-stuttgart.de

Technischer Überblick

Die SR2201 wird von Hitachi/COMPAREX als massiv-paralleler Supercomputer propagiert. Es handelt sich um ein System mit hardwaremäßig verteiltem, aber logisch globalem Hauptspeicher, das für die Lösung großer Probleme im Wissenschafts- und Ingenieursbereich in Frage kommt.

Die SR2201 verwendet Hitachis neuen Hochleistungs-RISC-Chip und skaliert über einen sehr weiten Knotenbereich von 8 bis 2048 Prozessoren. Das Verbindungsnetzwerk ist als dreidimensionaler, paralleler Crossbar mit einer Bandbreite von 300 Mbyte/s je Prozessor ausgelegt. Eine Pseudo-Vektorfunktion erlaubt den programmierten By-pass der beiden Prozessor-Caches mittels nicht blockierender Load/Store-Befehle im Rahmen der vollen, lokalen Speicherbandbreite von 1,2 Gbyte/s. Als Betriebssystem wird ein Microkernel UNIX, basierend auf dem Industriestandard Mach 3.0, verwendet.

Architektur

Wie bei allen Systemen mit physikalisch verteiltem Speicher kommt dem Verbindungsnetzwerk zwischen den einzelnen Knoten eine besondere Bedeutung zu. Die größeren Modelle der SR2201 (>128 Prozessoren) verwenden einen dreidimensionalen Cross-bar Switch, um individuelle Verbindungen zwischen einzelnen Prozessoren durchzuschalten. Drei Leitungsbündel je Prozessor genügen für den Anschluß an das dreidimensionale Verbindungsnetzwerk. Die innere Bandbreite des bidirektionalen 3D-Cross-bars beträgt n x 300 Mbyte/s, wenn n die Anzahl der angeschlossenen Prozessoren bezeichnet [1].

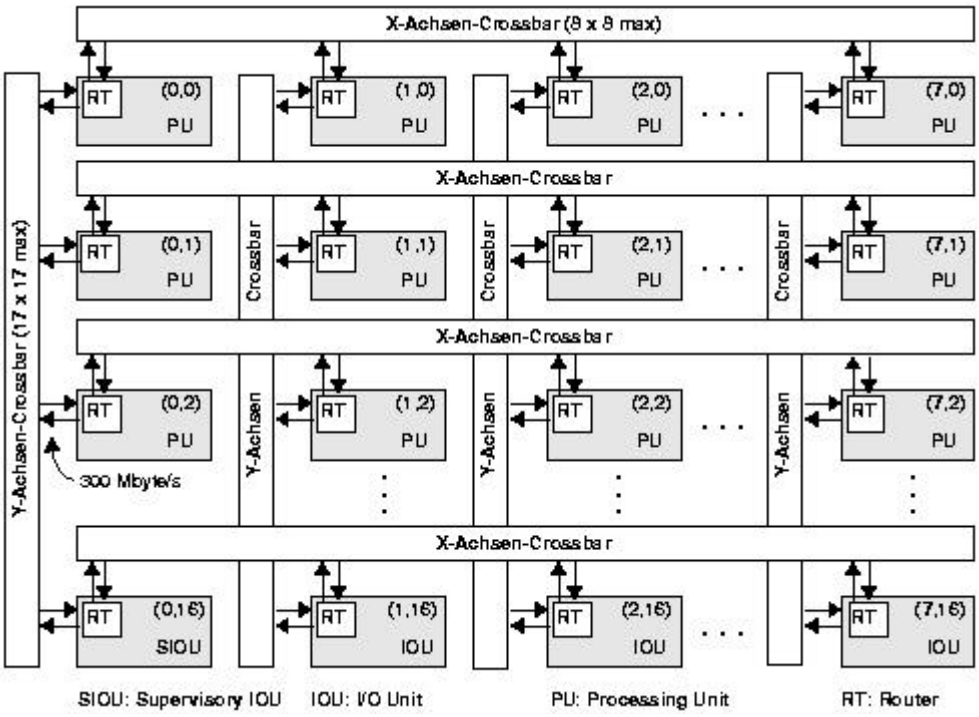

Abb. 1: Grundkonfiguration einer SR2201 bis maximal 128 Prozessoren

Abbildung 1 zeigt die Grundkonfiguration einer mittleren SR2201, bei der ein 2D- Crossbar genügt, um die vorhandenen Prozessoren miteinander zu verknüpfen. Diese Darstellung entspricht dem derzeitigen Ausbaugrad der Anlage am RUS.

Broadcast und Barrier-Synchronisation innerhalb einer Prozessorgruppe sowie direkter Zugriff auf den Speicher eines entfernten Knotens erledigt das Crossbar-Netz in Hardware. Die Latenzzeit des Remote DMA beträgt typisch 4 µs und stellt damit die Grundlage für eine effiziente Implementierung der Message Passing-Mechanismen her.

Die Knotenzahl einer Partition kann beliebig sein. Weiterhin müssen die Knoten einer Partition nicht notwendigerweise in einem zusammenhängenden Gebiet liegen.

Knotencharakteristik

Als Knotenprozessor kommt durchweg der sogenannte HARP-1E (Hitachi Advanced RISC Processor 1 Enhance) zum Einsatz [2]. Dabei handelt es sich um einen 32-bit Superscalar RISC, der bis zu vier Operationen je Maschinentakt durchführen kann; davon sind jeweils zwei Festkomma- und zwei Fließkomma-Operationen. Bei einer Taktfrequenz von 150 MHz beträgt die Rechenleistung maximal 300 MFlop/s.

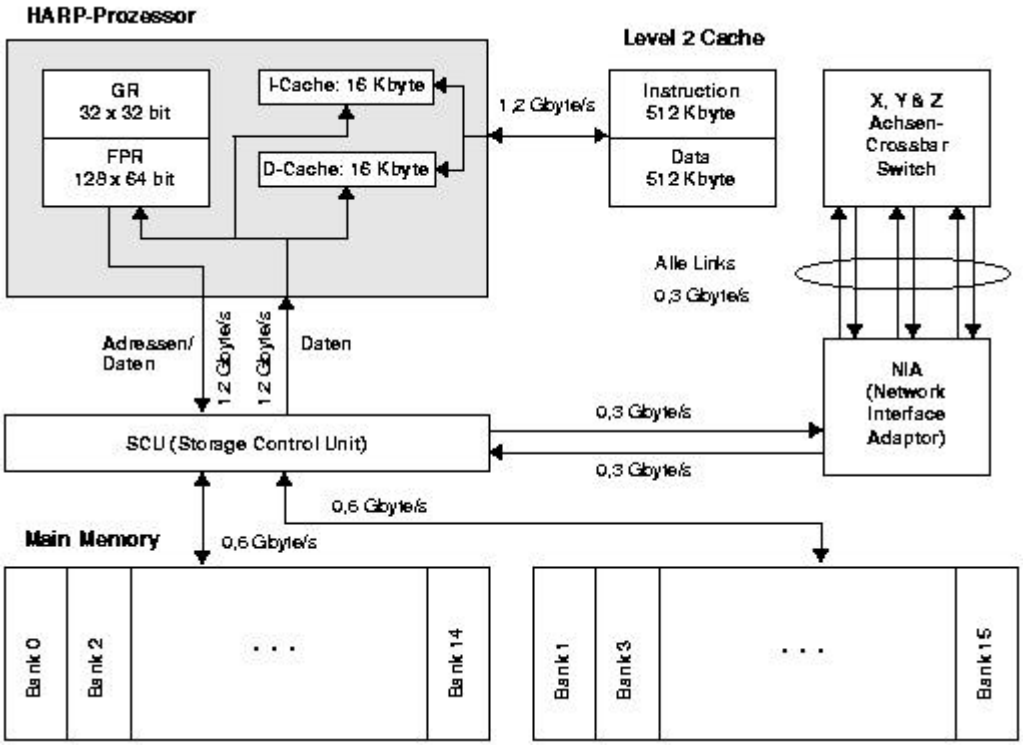

Abb. 2: Speicherorganisation des HARP-1E-Knotens

Obige Abbildung 2 enthält eine schematische Darstellung der Speicherorganisation auf einem Knoten. Bemerkenswert sind hier der extrem große Fließkommaregistersatz (128 x 64 bit), das Second Level Cache (2 x 512 Kbyte) und die lokale Speicherbandbreite von 1,2 Gbyte/s. Die Größe des Hauptspeichers beträgt 256 Mbyte (1 Gbyte max.).

Große numerische Berechnungen, deren Daten nicht in der Cache-Hierarchie Platz finden, sind ein wunder Punkt jeglicher RISC-Architektur. Abbildung 3 zeigt den dramatischen Abfall der Rechenleistung nach dem ersten bzw. zweiten Cache-Überlauf. In diesem Fall ist es oft besser, die Daten vorausschauend aus dem Hauptspeicher direkt in die internen Fließkommaregister zu kopieren.

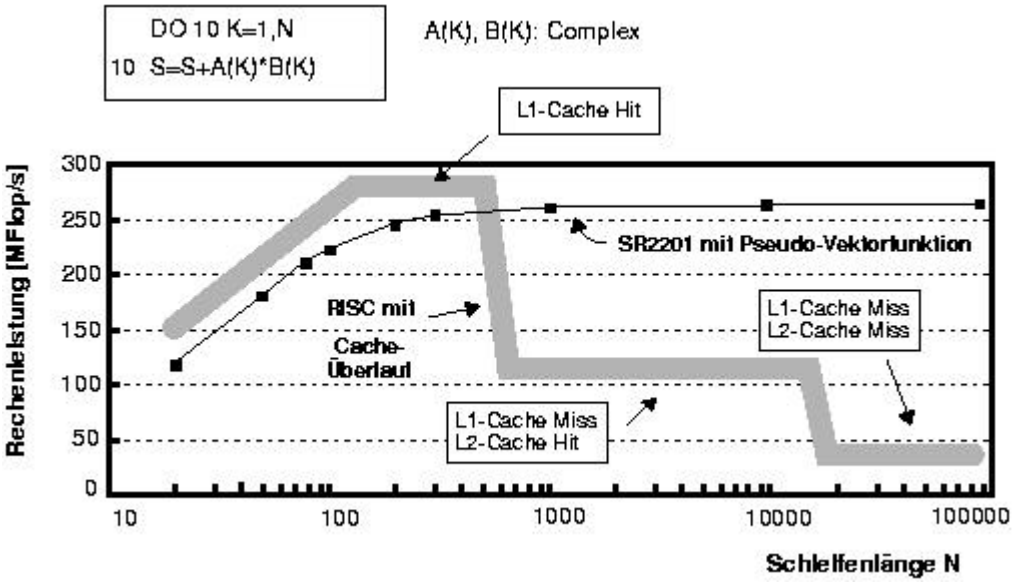

Abb. 3: Pseudo-Vektorverarbeitung beim Skalarprodukt

Die HARP-Architektur enthält als neues Merkmal die sogenannte Pseudo-Vektorfunktion. Darunter versteht man den programmierten I/O von Vektor- und Matrixelementen durch Preload- und Poststore-Befehle, die vom übrigen Programmfluß zeitlich entkoppelt (nicht blockierend) sind. Bei einer typischen Hauptspeicherlatenzzeit von 50 Maschinentakten sind allerdings in der Größenordnung von 128 Fließkommaregistern erforderlich. Von diesen 128 Fließkommaregistern sind jeweils 32 aktiv; sie werden analog zur SPARC-Architektur in zyklisch umlaufenden Registerfenstern adressiert.

Natürlich geht lediglich die Preload-Instruktion mit einem echten Cache Bypass einher. Beim Rückschreiben von Registerinhalten in den Hauptspeicher wird aus Konsistenzgründen immer ein Cache Update durchgeführt.

I/O

Die Hitachi SR2201 stellt ein außerordentlich homogenes System dar, insbesondere was die Implementierung der Ein-/Ausgabe anbelangt. Die Supervisory I/O Unit (SIOU) und die gewöhnlichen I/O Units (IOUs) sind, abgesehen von einer Peripheriebus-Erweiterung, identisch zu den Rechenknoten. Jeder IOU-Knoten treibt vier Magnetplatten an einem gemeinsamen 16-bit SCSI-2 Bussystem. Die Grenzrate dieses Bussystems von 20 Mbyte/s wird im praktischen Betrieb annähernd erreicht.

Abbildung 4 zeigt die aktuelle I/O-Konfiguration am RUS. Ein Großteil der dargestellten Magnetplatten ist zu einem parallelen Filesystem zusammengezogen worden. Parallele Filesysteme mit frei konfigurierbarem File- und Blockstriping sind möglich. Je nach verwendetem Pfadnamen können in ein- und demselben Filesystem Dateien einmal auf File- und das andere mal auf Blockebene verschränkt werden [3]. Die Gesamtkapazität der lokalen Magnetplatten beträgt 67 Gbyte.

Die Netzanbindung der SR2201 erfolgt über insgesamt drei Ethernet- und eine HIPPI-Schnittstelle. Die 10-Mbit/s Ethernetschnittstellen ermöglichen den Benutzerzugang. Das High Performance Parallel Interface (HIPPI, 800 Mbit/s) ist im wesentlichen für den File Service reserviert.

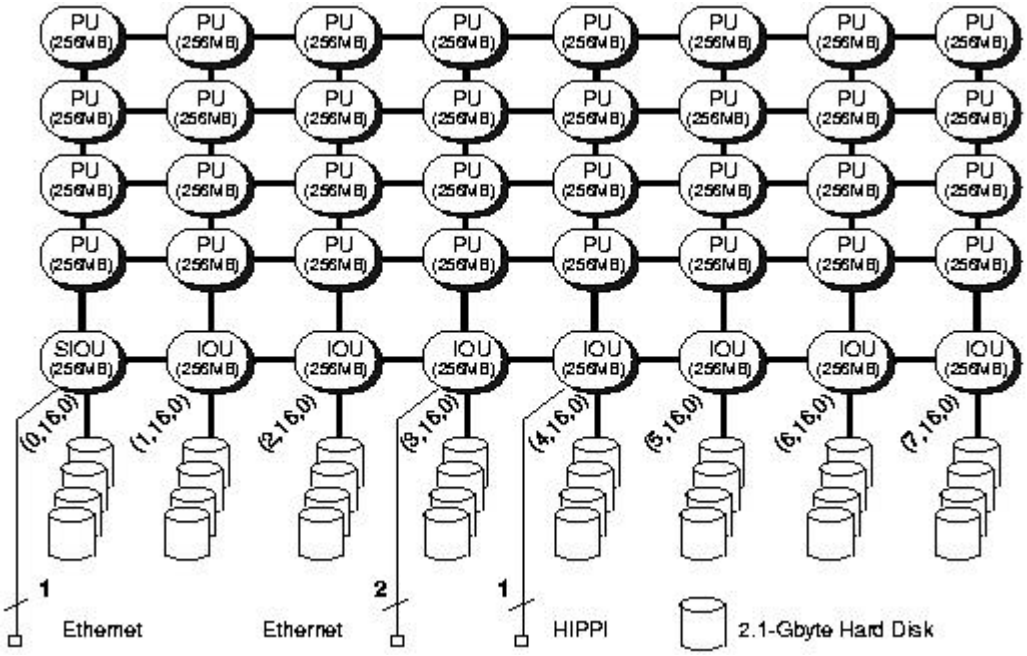

Abb. 4: I/O-Konfiguration der SR2201 am RUS

Betriebssystem

Das UNIX-Betriebssystem der SR2201, Hi-UX/MPP, basiert auf dem Mach 3.0 Microkernel der Carnegie-Mellon University. Eine Kopie davon läuft auf jedem Rechenknoten. Für den Betriebssystem-Code und die zugehörigen Datenstrukturen werden etwa 13 Mbyte Hauptspeicher benötigt.

Hi-UX ist mit vielen Industrie- und de facto-Standards kompatibel, wie z.B. mit:

- OSF und BSD 4.3-Systemschnittstellen
- POSIX ISO/IEC DIS 9945-1
- IEEE 1003.2 Shell Utilities
- X/Windows und Motif Graphical Interface Standards
- ANSI/IEEE Std 754-1985, Fließkommadarstellung

Im Gegensatz zu den Rechenknoten wird auf den (Supervisory) I/O Units ein OSF/1-Betriebssystem verwendet. Der UNIX-Server steht nur auf den I/O-Knoten zur Verfügung.

Programmiermodelle und Compiler

Zur Verfügung stehen Compiler für FORTRAN 90, C und C++. Parallel FORTRAN, ein Hitachi-eigener Übersetzer, unterstützt HPF (High-Performance FORTRAN) Version 1. HPF-Programme werden dabei in FORTRAN 90-Quellcode übersetzt.
Die SR2201 unterstützt die folgenden parallelen Programmiermodelle:

- PVM 3.3.10 (Originalversion mit Remote DMA)
- MPI (MPICH Version 1.0.11, alle APIs, Originalversion mit Remote DMA)
- PARMACS
- Express (und PVM) als Teil von ParallelWare

Leistungsfähige Debug- und Analyse-Werkzeuge erleichtern die Programmentwicklung. Eine Reihe numerischer Bibliotheken wurde für die parallele Ausführung auf der SR2201 optimiert, darunter u.a.

MSL2, Matrix/MPP, Matrix/MPP/SSS.

Das RUS bemüht sich um Standardpakete unabhängiger Softwarehersteller.

Weitere Informationen dazu erhalten Sie in aktueller Form auf unseren WWW-Seiten:
http://www.rus.uni-stuttgart.de

Zugang

Die Hitachi SR2201/H32 ist im Netz der Universität Stuttgart unter ihrem Primärnamen `hitachi.rus.uni-stuttgart.de` erreichbar. Sie wird für Nutzer aus dem akademischen Bereich - während der Dauer der o.g. Kooperationsvereinbarung - auf Antrag kostenfrei bereitgestellt.

Der Zugriff erfolgt über die Internet-Dienste Telnet, FTP und gegebenenfalls NFS.
Der Batch-Zugang wird im Laufe der nächsten Wochen etabliert werden.

Verwendete Abkürzungen

ANSI	American National Standards Institute
API	Application Programming Interface
BSD	Berkeley System Distribution
CPU	Central Processing Unit
DMA	Direct Memory Access
DRAM	Dynamic Random Access Memory
FDDI	Fibre Distributed Data Interface
FTP	File Transfer Protocol
HARP	Hitachi Advanced RISC Processor
HIPPI	High Performance Parallel Interface
HLRS	Höchstleistungsrechenzentrum Stuttgart
HPF	High-Performance FORTRAN
IEEE	Institute of Electrical and Electronics Engineers
IP	Internet Protocol
IOU	Input/Output (Processing) Unit
LAN	Local Area Network
MPI	Message Passing Interface
MPP	Massively Parallel Processor
MTU	Maximum Transmission Unit
NFS	Network File System
NIA	Network Interface Adapter
OSF	Open Software Foundation
PVM	Parallel Virtual Machine
PVP	Pseudo Vector Processing
PU	Processing Unit
RFC	Internet Request for Comment
RISC	Reduced Instruction Set Computer
RT	Router
RUS	Rechenzentrum Universität Stuttgart
SCSI	Small Computer Systems Interface
SCU	Storage Control Unit
SIOU	Supervisory IOU
SPARC	Scalable Processor Architecture
TCP	Transmission Control Protocol
WWW	World Wide Web

Literatur

[1] Hitachi Europe, Ltd.: SR2201 Introduction, October 1996

[2] Hitachi Ltd.: Performance Optimization on the Hitachi SR2201, Hitachi, Ltd., Gerneral Purpose Computer Division, 1996

[3] Haas, P., Beisel, Th.: Hitachi SR2201: Disk and Network I/O, RUS Technical Report, March 17, 1997

Dr. Heinz W. Pöhlmann, NA-5992
E-Mail: poehlmann@rus.uni-stuttgart.de

Peter Haas
E-Mail: haas@rus.uni-stuttgart.de