FORSCHUNGSZENTRUM JÜLICH GmbH
Zentralinstitut für Angewandte Mathematik
D-52425 Jülich, Tel. (02461) 61-6402

Interner Bericht

Metacomputing zwischen KFA und GMD
Ein verteilter massiv-paralleler Rechner
im RTB-NRW

Michael Weber, Wolfgang E. Nagel,
Helmut Grund, Roland Völpel**

KFA-ZAM-IB-9523

Oktober 1995
(Stand 26.10.95)

(*) Institut für Wissenschaftliches Rechnen und Algorithmen (SCAI),
GMD - Forschungszentrum Informationstechnik GmbH, D-53757 Sankt Augustin

Dieser Bericht wird erscheinen in:
Mitteilungen - Gesellschaft für Informatik e.V., Parallel-Algorithmen und Rechnerstrukturen

Metacomputing zwischen KFA und GMD Ein verteilter massiv-paralleler Rechner im RTB-NRW

Michael Weber*, Wolfgang E. Nagel*
Zentralinstitut für Angewandte Mathematik (ZAM)
Forschungszentrum Jülich GmbH (KFA)
D-52425 Jülich

Helmut Grund, Roland Völpel
Institut für Wissenschaftliches Rechnen und Algorithmen (SCAI)
GMD - Forschungszentrum Informationstechnik GmbH
D-53757 Sankt Augustin

Zusammenfassung

Hochleistungsrechner, wie Vektor- und Parallelrechner, sind in den Produktionsumgebungen der Forschungseinrichtungen und Universitäten etabliert. Zwischen vielen dieser Institutionen stehen bereits heute schnelle Datenleitungen auf der Basis des *Asynchronous Transfer Mode* (ATM) zur Verfügung. Diese Infrastruktur eröffnet neue Perspektiven der kooperativen Nutzung der regional verteilten Ressourcen.

1 Einleitung

Die Verteilung unterschiedlicher Aufgaben auf die verschiedenen Funktionseinheiten eines Rechners ist eine in der Datenverarbeitung übliche Vorgehensweise. Typischerweise handelt es sich dabei um eine lokale, statische Aufgabenteilung auf Ebene der Hardware. Im Unterschied dazu bezeichnet der Begriff *Metacomputing* die Verteilung einer Anwendung auf zwei oder mehr Rechner, die dynamisch über ein externes Netzwerk gekoppelt werden, und die Verteilung erfolgt typischerweise auf der Software-Ebene. Diesem Konzept liegen unterschiedliche technische und wirtschaftliche Überlegungen zugrunde. Zunächst kann die Leistungsfähigkeit der einzelnen Hardware-Komponenten akkumuliert werden, um so Probleme anzugehen, die auf einem System nicht gelöst werden können. Ebenso kann flexibel auf extreme Anforderungen einzelner Projekte reagiert werden, indem die Ressourcen mehrerer Standorte gebündelt werden.

*E-mail: {m.weber,w.nagel}@kfa-juelich.de

Zwei Konzepte des *Metacomputing* können unterschieden werden, die aber auch in Kombination denkbar sind.

- Ein Weg nutzt die *funktionale Parallelität*, die in vielen Anwendungen steckt, indem das Programm auf die jeweils am besten geeignete Hardware verteilt wird. Dies können Workstations, Vektorrechner oder auch massiv-parallele Rechner sein. Dieses Vorgehen verspricht für die Zukunft hohe Gewinne in der Performance, erfordert jedoch gegenwärtig noch erhebliche Eingriffe in die existierenden Programme sowie die Definition einheitlicher Schnittstellen und Konzepte.

- Der zweite, naheliegendere Weg ist die *Verteilung bereits per Message-Passing parallelisierter Programme* auf Rechner mit verteiltem Speicher. Ist eine Anwendung für eine derartige Architektur parallelisiert, bedeutet es logisch keinen Unterschied, ob der Nachrichtenaustausch zwischen den Rechenknoten innerhalb eines Systems erfolgt oder aber über eine externe Verbindung mit einer CPU in einem zweiten Rechner.

Damit verbundenen sind technische und organisatorische Probleme, die an Hand einer Pilotanwendung aufgezeigt und teilweise gelöst werden sollen. Dies ist das Ziel in dem gemeinsamen Projekt 'Verteilter massiv-paralleler Rechner' der Forschungszentrum Jülich GmbH (KFA) und der GMD - Forschungszentrum Informationstechnik GmbH. Als Teilprojekt Z2 wird das Projekt im Rahmen des *Regionalen Testbed* (RTB) NRW vom DFN-Verein seit dem 1. September 1994 gefördert. Innerhalb der zweijährigen Laufzeit soll eine Prototyp-Implementierung einer verteilten Anwendung auf den beiden Parallelrechnern Intel Paragon XP/S10 in der KFA und IBM SP2 in der GMD realisiert werden, die im Rahmen des RTB-NRW durch eine ATM-Datenleitung mit zunächst 34 Mbit/s verbunden sind. In dem folgenden Abschnitt wird zunächst die vorhandene Infrastruktur vorgestellt. Anschließend werden das verwendete Programm und der verteilte massiv-parallele Rechner beschrieben. Nach dem Projektablaufplan und der Vorstellung der projektbegleitenden Arbeiten soll ein Ausblick auf die weitere Arbeit und Entwicklung gegeben werden.

2 Infrastruktur

Das ZAM der KFA und die GMD betreiben seit mehreren Jahren Parallelrechner und Hochgeschwindigkeitsnetze in Produktionsumgebungen und verfügen somit über die entsprechenden Erfahrungen. Für das Projekt wird im ZAM eine Intel Paragon XP/S10 [1] und in der GMD eine IBM SP2 eingesetzt. Tabelle 1 stellt die wesentlichen technischen Daten der Systeme einander gegenüber. Obwohl also beide Systeme die gleiche Architektur aufweisen, zeigen ihre technischen Daten sehr unterschiedliche Stärken und Schwächen, die bei der Konzeption des Projektes berücksichtigt werden sollten.

Als Teilnehmer am RTB-NRW stehen zwischen Jülich und St. Augustin seit März 1995 schnelle Datenleitungen zur Verfügung. Weitere Teilnehmer an diesem Testbed sind die Standorte Aachen, Bonn und Köln. Die Verbindung beruht auf dem *Asynchronous Transfer*

	XP/S10	SP2
Anzahl CPUs	140	37
Hauptspeicher pro CPU [MBytes]	32	128
Rechenleistung pro CPU [Mflops]	75	250
Interne Bandbreite [MBytes/s]	90	34
Topologie	2D-Mesh	Crossbar
Netzwerk-Interfaces	Ethernet, FDDI/HiPPI	Ethernet, HiPPI, ATM

Tabelle 1: Technische Daten der Intel Paragon XP/S10 und der IBM SP2

Mode (ATM) [2]. ATM ist eine neue, noch in der Standardisierung befindliche Technik, die nötige Hardware ist daher noch nicht für alle Systeme verfügbar. Neben hohen Bandbreiten bietet die ATM-Technologie eine Reihe von Merkmalen, die sie in besonderer Weise für *Metacomputing-Anwendungen* interessant macht. Die Möglichkeit, eine 'Quality of Service' auch für Netzverbindungen zu garantieren, ohne hierfür teure Standleitungen schalten zu müssen, hebt die externe Verbindung qualitativ auf das Niveau der internen Verbindungsnetzwerke der Parallelrechner. Inwieweit eine ATM-Verbindung quantitativ den gestellten Anforderungen genügt, soll in diesem Projekt geklärt werden. Gegenwärtig ist im Testbed eine Bandbreite von 34 Mbit/s realisiert. Für die Zukunft existieren Optionen auf 155 und 622 Mbit/s. Unabhängig von den RTB-Projekten werden in der GMD und im ZAM lokale ATM-Netze betrieben und aufgebaut, die eine direkte ATM-Verbindung der Parallelrechner gestatten sollen. Diese Infrastruktur eröffnet neue Perspektiven der kooperativen Nutzung der Systeme.

3 Das parallele Programm

Das Institut für Chemie und Dynamik der Geosphäre (ICG-4) der KFA entwickelt ein Programm (TRACE), das den Wasserfluß in porösen heterogenen Medien berechnet und den Schadstofftransport in Boden und Grundwasser an Hand eines dreidimensionalen Modells simuliert [3]. Dazu werden Ort und Zeit durch Finite-Elemente (FE) und Finite-Differenzen diskretisiert. Die Berechnung realistischer Gebietsgrößen erfordert rund 10^6 FE-Knoten. Neben der Rechenzeit ist daher insbesondere der hohe Speicherplatzbedarf des Programms ein begrenzender Faktor. Die Parallelisierung für den Einsatz auf der Intel Paragon zeigt hier einen Ausweg auf [4]. Auf der Basis des Schwarz'schen Verfahrens wird eine Gebietszerlegung durchgeführt, die es erlaubt, die so entstehenden Gebiete auf die verschiedenen Rechenknoten eines Parallelrechners zu verteilen. Die aktuelle Version des Programms geht von einer gleichmäßigen Lastverteilung aus und teilt den Prozessoren gleich große Gebiete zu. Die Parallelisierung erfolgte nach dem Message-Passing-Pardigma unter Verwendung der Intel NX-Schnittstelle [5]. Zur Erleichterung der Portierung auf andere Schnittstellen wurden die

Intel-Befehle gekapselt und in einem Modul zusammengefaßt. Im Hinblick auf Methoden und Algorithmen wird das Programm von einer Gruppe von Wissenschaftlern permanent fortentwickelt.

4 Der verteilte massiv-parallele Rechner

Abbildung 1 zeigt das Schema des verteilten massiv-parallelen Rechners, wie er zwischen der KFA und der GMD betrieben werden soll. Die Anwendung, das ATM-Netz und die unterschiedliche Hardware wurden bereits vorgestellt.

Bislang konnten sich die Rechnerhersteller weder auf eine einheitliche parallele Programmierschnittstelle einigen, noch unterstützen sie die Möglichkeit, mit anderen Rechnern aus einer Anwendung heraus zu kommunizieren (Remote-Message-Passing, RMP). So muß eine portable Software-Schicht installiert werden, die die folgende Funktionalität bietet.

- *Portabilität:* Da es einen nicht vertretbaren Aufwand bedeutet, zwei verschiedene Programmversionen zu warten, muß auf beiden Systemen eine einheitliche parallele Programmierschnittstelle existieren.
- *Transparenz:* Die Befehlssyntax für das Versenden oder Empfangen von Nachrichten sollte unabhängig vom Ort sein. Das heißt auch, daß die Software in der Lage sein muß, automatisch die verschiedenen Zahlendarstellungen auf den Rechnern umzuwandeln.
- *Effizienz:* Die Schnittstelle soll die Kommunikationsleistung jedes Systems optimal ausnutzen und dies möglichst durch die Einführung von Prozeßgruppen unterstützen.

In dem Projekt sollen diese Ziele durch die Wahl der Bibliothek PARMACS [6] der Firma PALLAS GmbH in Brühl realisiert werden. Mit der Version 6.0 war bereits eine portable Programmierschnittstelle auf der Paragon und der SP2 vorhanden. Mit der Version 7.0 wurde von den Projektpartnern eine aufwärtskompatible Erweiterung in Auftrag gegeben. Damit konnten die Anforderungen der KFA und der GMD bei der Definition und Implementierung berücksichtigt werden. PARMACS 7.0 unterstützt die oben formulierten Anforderungen für das transparente Versenden von Nachrichten zwischen zwei Parallelrechnern. Dabei ist zu berücksichtigen, daß sich die internen und die externen Kommunikationsleistungen (Bandbreite, Latenz) eines *Metacomputers* um rund zwei Größenordnungen unterscheiden.

5 Der Projektablauf

Das Teilprojekt Z2 wurde am 1. September 1994 begonnen und hat eine Laufzeit von zwei Jahren. Während dieser Zeit sind 1.5 Mitarbeiter in der KFA und 0.5 in der GMD mit den Projektarbeiten betraut. Ein Jahr nach dem Start, in der Mitte der Laufzeit des Projektes, sind bereits eine Reihe von Zwischenergebnissen erzielt worden. Eine Übersicht der erreichten und angestrebten Meilensteine zeigt die untere Hälfte der Abbildung 2. Nach

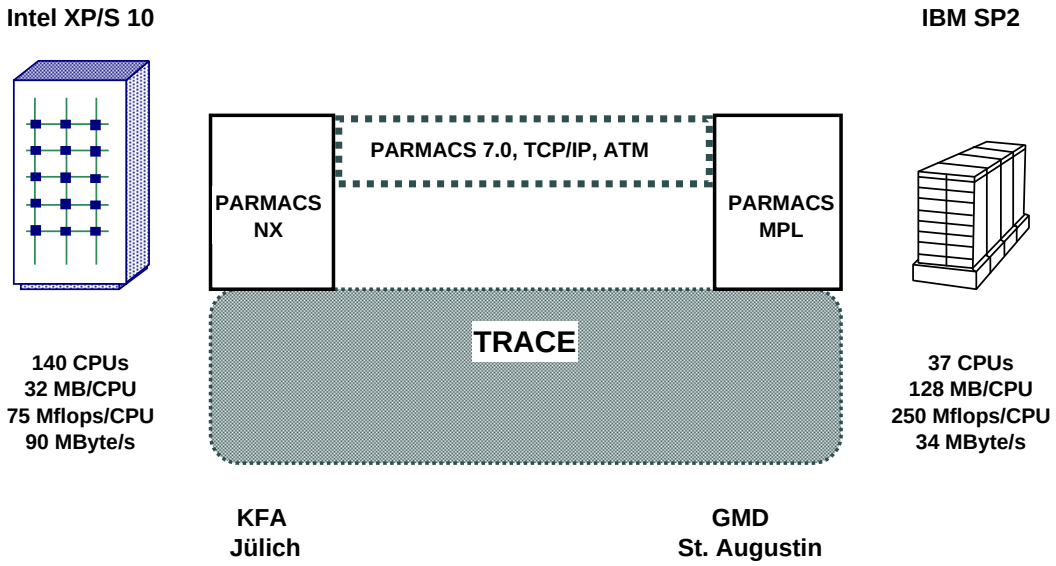

Abbildung 1: Der verteilte massiv-parallele Rechner – Schema

der Fertigstellung einer parallelisierten Arbeitsversion des Programms TRACE auf der Intel Paragon wurde eine mit den PARMACS 6.0 instrumentierte Version erstellt und im Juni 1995 auf die Version 7.0 angepaßt. Tabelle 2 vergleicht die Laufzeiten des noch nicht verteilten Programms für die unterschiedlichen Bibliotheken an Hand einiger Testläufe auf je 4 CPUs der Paragon und der SP2 in der KFA. Der Vergleich zwischen SP2 und Paragon

Bibliothek	1 Zeitschritt[Sek.]	2 Zeitschritte[Sek.]	3 Zeitschritte[Sek.]
Intel Paragon			
NX	100	162	226
PARMACS 6.0	100	163	226
PARMACS 7.0	118	178	244
IBM SP2			
PARMACS 6.0	86	137	182

Tabelle 2: Laufzeiten des Programms TRACE auf 4 CPUs der Intel Paragon und der IBM SP2 für verschiedene Bibliotheken und für eine unterschiedene Zahl von Zeitschritten

zeigt nur einen geringen Geschwindigkeitsgewinn von rund 20%, der jedoch auf der SP2 in der GMD deutlicher ausfallen sollte, da die CPUs der IBM in Jülich über TCP/IP kommunizieren. Obwohl die reine Kommunikationsleistung der PARMACS auf dem Intel-Rechner niedriger ist als die der NX-Bibliothek, zeigt das Programm unter der Version 6.0 innerhalb der Toleranzen die gleichen Laufzeiten wie das originale Programm. Erst unter der Version

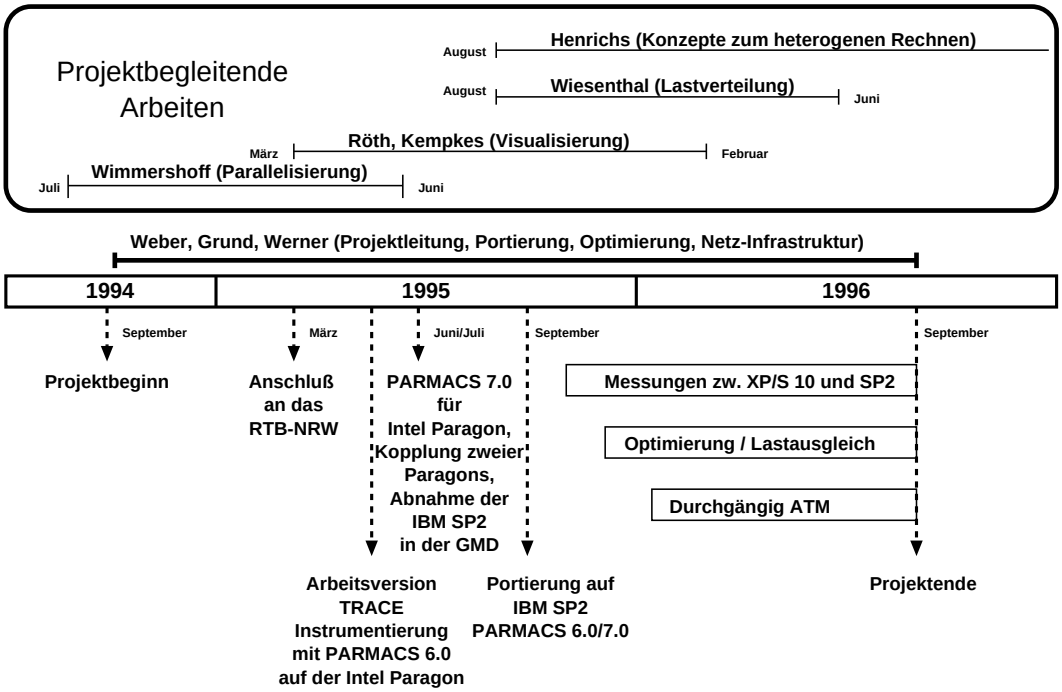

Abbildung 2: Projektüberblick des Teilprojektes Z2

7.0 beobachet man Einbußen in der Größenordnung von 10%. Einen genaueren Vergleich der beiden PARMACS-Versionen mittels des Benchmark-Programms *comms1* aus der GENESIS Test-Suite [7] zeigt Tabelle 3. Die geforderte Effizienz soll in dem folgenden Release erzielt werden. Trotzdem waren erste erfolgreiche Tests zwischen zwei Paragon-Rechnern im ZAM

Bibliothek	Latenz [μs]	Bandbreite [MByte/s]
PARMACS 6.0	49	66
PARMACS 7.0	126	36

Tabelle 3: Vergleich der Latenz und Bandbreite der Version 6.0 und 7.0 der PARMACS, mittels des Programms *comms1* aus der GENESIS-Suite

möglich, auf die das Programm verteilt wurde. Die Kopplung dieser Rechner per Ethernet ließ jedoch nur Tests der Funktionalität zu.

Damit sind die Arbeiten für die restliche Projektzeit vorgegeben. Nach der Auslieferung der PARMACS 7.1 im Oktober 1995 müssen Messungen zwischen der Intel Paragon in Jülich und der IBM SP2 in St. Augustin vorgenommen werden. Diese werden begleitet von lokalen und globalen Optimierungen an dem Programm und dessen Kommunikationsstruktur. Ein zentraler Punkt wird dabei der Lastausgleich zwischen den unterschiedlich leistungsstarken

Rechnern sein. Schon jetzt zeichnet sich ab, daß diese Arbeiten nur dann erfolgreich sein können, wenn es gelingen wird, zwischen den Systemen eine durchgängige ATM-Verbindung zu etablieren.

6 Projektbegleitende Arbeiten

In der oberen Hälfte der Abbildung 2 sind diejenigen Arbeiten dargestellt, die im ZAM begleitend zum Projekt durchgeführt wurden und werden. Insbesondere für die kommenden Meilensteine stellen diese Arbeiten die nötigen Hilfsmittel zur Verfügung. Das Kommunikationsverhalten und die Interpretation der Ergebnisse sind derart komplex, daß graphische Hilfsmittel hierfür unabdingbar sind. Die Fehlersuche, Optimierung und auch die Lastverteilung werden durch die im ZAM entwickelte X Window basierte Visualisierungsumgebung PARvis unterstützt. PARvis gestattet es beispielsweise, einzelne Nachrichten, wie in Abbil-

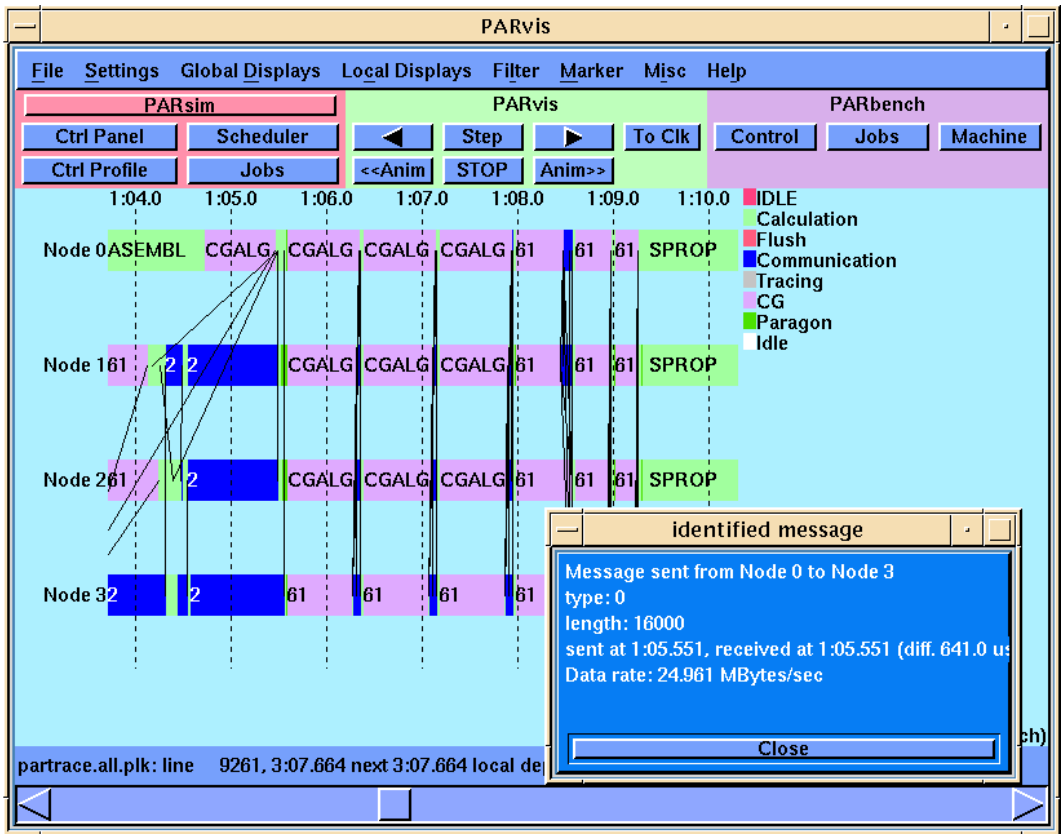

Abbildung 3: Beispiel einer PARvis-Darstellung des Programms TRACE auf 4 CPUs

dung 3 von Knoten 0 an Knoten 3, auszuwählen und die Nachrichtenlänge, -Kennung und Bandbreite zu messen und darzustellen. Eine ausführliche Darstellung der umfangreichen

Möglichkeiten von PARvis findet man in [8]. Zur Interpretation und Visualisierung der Gebietszerlegung und der numerischen Ergebnisse entsteht eine Umgebung unter der Software AVS [9]. Die Abbildungen 3 und 4 präsentieren zwei Darstellungen der beiden Tools. Weitere

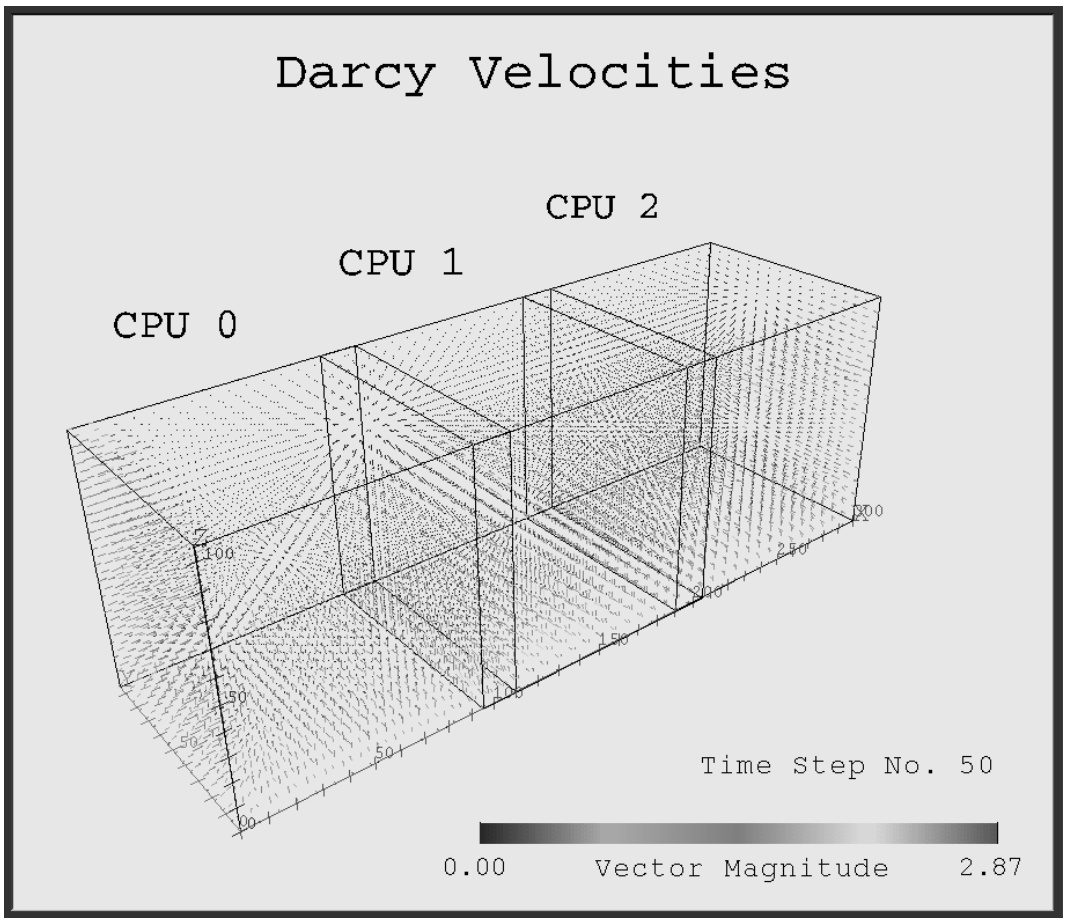

Abbildung 4: Darstellung der Gebietszerlegung für 3 Gebiete

Arbeiten (zum Beispiel von Frau Wiesenthal) sollen an der Anwendung TRACE neue Strategien zur Lastverteilung auf die heterogene Hardware – CPUs und Verbindungsnetzwerke– entwickeln. Die Erzeugung unterschiedlich großer Gebiete deutet sich hier als Lösungsweg an. Langfristig werden neue Programmierkonzepte zum heterogenen Rechnen entworfen und realisiert werden.

7 Zusammenfassung und Ausblick

Im Rahmen des Teilprojektes Z2 'Verteilter massiv-paralleler Rechner' sind zur Mitte der Laufzeit die Grundlagen, Konzepte und Hilfsmittel erarbeitet worden, die die Voraussetzungen dafür liefern, daß die Anwendung TRACE prototypisch auf die zwei Rechner verteilt

werden kann. Die Erfahrungen, die dabei bislang gesammelt wurden, zeigen, daß der entscheidende Punkt in der zweiten Hälfte des Projektes die Anbindung der beiden Rechner Intel Paragon und IBM SP2 an die Hochleistungsdatenleitung sein wird. Die drei Niveaus der Kommunikationsleistungen, rund 90, 4 und 34 MByte/s für Paragon, ATM und SP2 stellen die technische Herausforderung dar. Abbildung 5 veranschaulicht den Status und die mögliche Entwicklung der Kommunikationsstrecke, indem die internen und externen Bandbreiten aufgetragen sind. Das Ziel muß es sein, die Schnittstellen Rechner/Datenleitung durchgängig

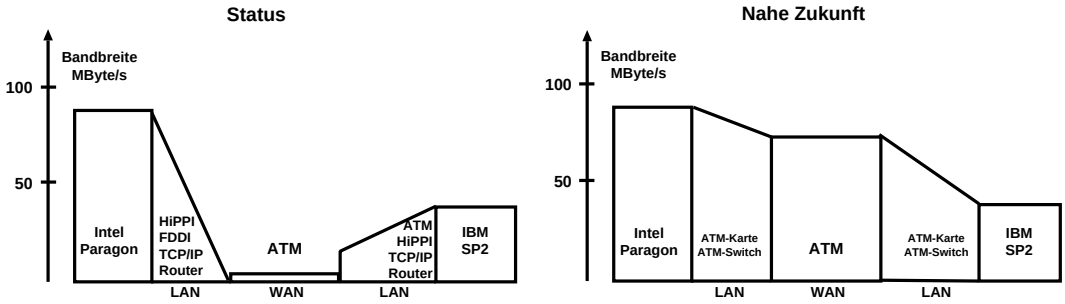

Abbildung 5: Status und mögliche Entwicklung der Kommunikationsstrecke zwischen Paragon und SP2

auf der ATM-Technik aufzubauen. Dazu werden sowohl im ZAM als auch in der GMD lokale ATM-Netze aufgebaut, die den direkten Zugang ins Testbed erlauben sollen. Auf der Seite der Paragon wird an der Beschaffung einer ATM-Hardware für diesen Rechner gearbeitet, da die existierenden FDDI- und Ethernet-Schnittstellen den Anforderungen nicht genügen. Die SP2 der GMD ist bereits mit einer ATM-Karte ausgerüstet. Hemmend erweist sich die noch nicht abgeschlossene Standardisierung beim ATM, die wie im Falle Intel das Angebot an verfügbarer Hardware einschränkt und die die Verwendung des Protokolls TCP/IP nötig macht. Der daraus resultierende Overhead beeinflußt insbesondere die Latenz so negativ, daß hier ein kritischer Flaschenhals existiert, der algorithmisch kaum aufgelöst werden dürfte. Mittelfristig ist zudem eine Erhöhung der Bandbreite im Testbed auf 155 bzw., wie in Abbildung 5 dargestellt, 622 Mbit/s wünschenswert, um so zumindest in Einzelfällen in die Größenordnung der rechnerinternen Leistungsdaten zu kommen.

Neben den programmiertechnischen Aspekten ist die Frage der Betriebsmittelverwaltung für den beschriebenen *Metacomputer* zu untersuchen. Es muß sichergestellt sein, daß zur vorgesehenen Laufzeit des Programms die notwendigen Ressourcen sowohl auf den beiden Rechnern als auch im Netz zur Verfügung stehen. In der Testphase werden diese Bedingungen auf der Basis expliziter Absprachen erfüllt werden. Es sollen aber Perspektiven und Strategien entwickelt werden, welche auch die automatische Ausführung des verteilten Programms erlauben.

Konzeptionell werden weitere Überlegungen die Verteilung von Programmen auf eine heterogene Hardware unter funktionalen Aspekten betreffen. Dieser Weg verspricht meßbare

Beschleunigungen bei der Ausführung der Programme, die bei der vorgestellten transparenten Verteilung nur schwer erzielt werden dürften. Auch TRACE wird in seinen weiteren Entwicklungsstufen diese Möglichkeit bieten. Mit der Berücksichtigung des Teilchentransportes im Modell kommt eine neue Komponente hinzu, die lose gekoppelt auf einer geeigneten Architektur gerechnet werden kann, während der Wasserfluß sehr gut auf einem Rechner wie der Intel Paragon läuft. Die bislang gesammelten Erfahrungen werden dafür einen grundlegenden Beitrag liefern.

Literatur

[1] R. Berrendorf et al., Intel Paragon XP/S – Architecture, software environment, and performance, Interner Bericht Forschungszentrum Jülich, KFA-ZAM-IB-9409, 1994

[2] D. Conrads, ATM – die Vermittlungs- und Multiplextechnik des Breitband-ISDN, Interner Bericht Forschungszentrum Jülich, KFA-ZAM-IB-9504, 1995

[3] H. Vereecken et al., TRACE: A mathematical model for reactive transport in 3D variably saturated porous media, KFA/ICG-4 Internal Report No. 501494, 1994

[4] R. Wimmershoff, Entwicklung und Implementierung einer dreidimensionalen Partitionierungsstrategie für das Programm TRACE auf einem massiv-parallelen Rechner, Diplomarbeit im Fach Elektrotechnik an der RWTH Aachen, Lehrstuhl für Technische Informatik und Computerwissenschaften, 1995

[5] Intel Supercomputer Systems Division, Paragon user's guide, No. 312489-002, 1993

[6] R. Calkin et al., Portable programming with the PARMACS message-passing library, Parallel Comput. 20, 615-632, 1994

[7] A. Hey, The GENESIS distributed memory benchmarks, Parallel Computing, 17 (10-11), 1275-1283, 1991

[8] W. E. Nagel et al., Performance visualization of parallel programs - The PARvis environment -, Proceedings 1994 Intel Supercomputing Users Group (ISUG) Conference, pp. 24-31, 1994

[9] A. Kempkes, Diplomarbeit im Fach Elektrotechnik an der RWTH Aachen, Lehrstuhl für Technische Informatik und Computerwissenschaften, zur Veröffentlichung vorgesehen