Kurzer Überblick zur Einbindung eines

Scalable Parallel Processing (SPP) Systems

in das Konzept des Meta Computing

CONVEX Computer GmbH
Büro Dresden
Dr. Kisperth

Dresden 28.5.93

I. CONVEX Konzept für High Performance Computing (HPC)

I. 1. Meta Computing

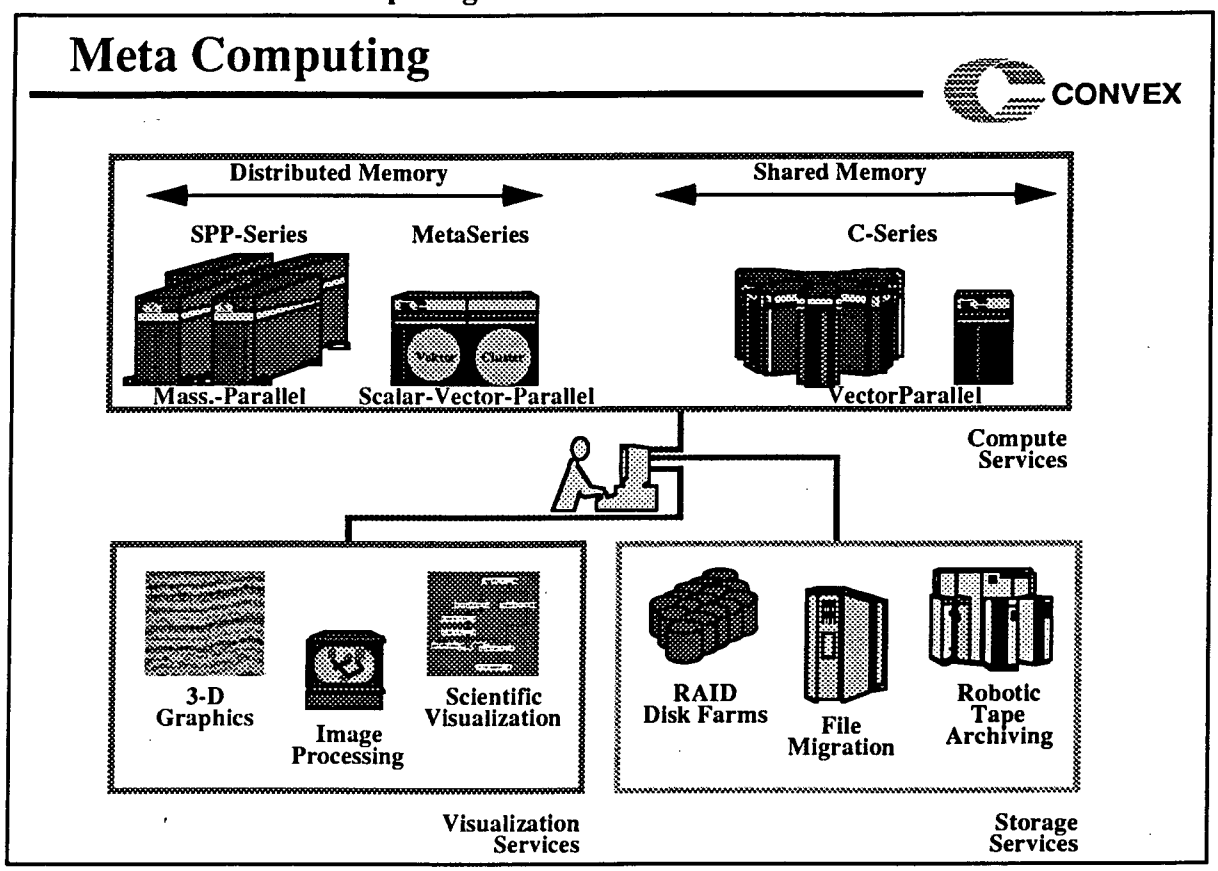


Abb.: 1

In Anlehnung an Larry Smarr, Direktor des NCSA, arbeitet CONVEX als führender Anbieter von HPC-Systemen auf eine Realisierung des Meta Computing Konzepts gem. Abb. 1 hin. Meta Computing stellt dem Benutzer in möglichst transparenter Weise folgende Dienste zur Verfügung:

- **Compute Services**
  - Skalar Architektur (CONVEX MetaSerie: C-Serie + PA-RISC Cluster)
  - Vektor Architektur (CONVEX C-Serie)
  - Parallel Architektur (CONVEX SPP-Serie)
- **Storage Services**
  - Disk Farms auf RAID-Basis (CONVEX VVM)
  - Roboter-Archive (VHS- oder D2-Roboter)
  - Filemigration (UniTree, Fileserv, CSM)
  - Remote Backup (RSB, User Access)
- **Visualization Services**
  - Hochleistungs-Workstations (HP 700-Serie)
  - Hochleistungs-LAN (FDDI, HiPPI, UltraNet)
  - Visual. Tools (AVS, CONVEX EVT - Engineering Visual. Tool)

I. 2. Parallelrechner-Serie: CONVEX SPP

Als Erweiterung der C-Serie, einer Familie von Multiprozessor Vektor-Systemen, entwickelt CONVEX eine Familie hochgradig skalierbarer Parallelrechner: **die CONVEX SPP-Serie.** Die erste Familie der SPP-Serie wird im ersten Quartal 1994 ausgeliefert werden und trägt die Bezeichnung **CONVEX SPP-1.**

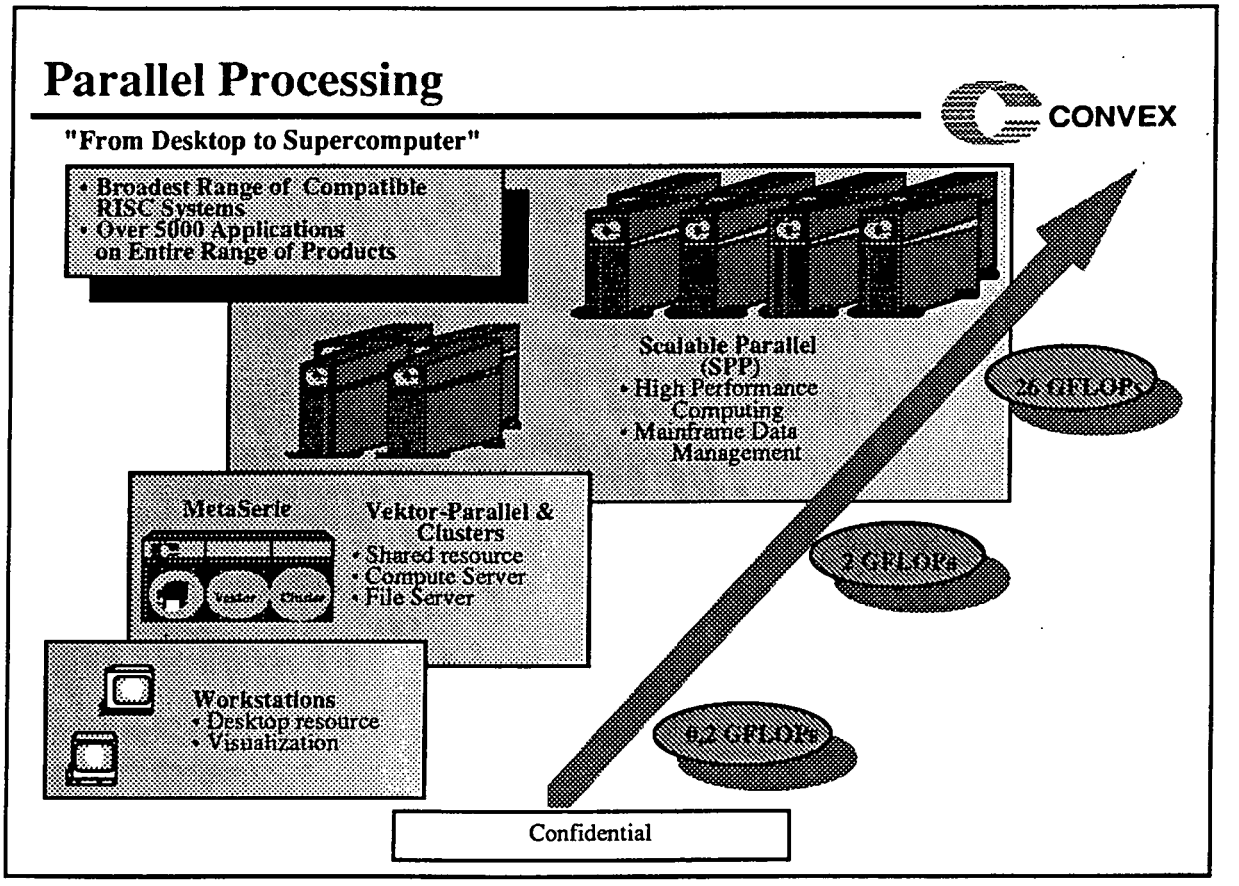


Abb.: 2

Die SPP-Systeme basieren auf dem HP PA-RISC Prozessor. CONVEX und HP haben über die reine Lieferung eines Prozessors im vergangenen Jahr eine strategische Kooperation geschlossen.

Ziel dieser Kooperation ist es, eine kompatible Produktpalette von HP Workstations (0,2 GFLOPs) über die CONVEX MetaSerie (2 GFLOPs) bis zu Hi-End CONVEX SPP-1 Systemen (26 GFLOPs) anzubieten.

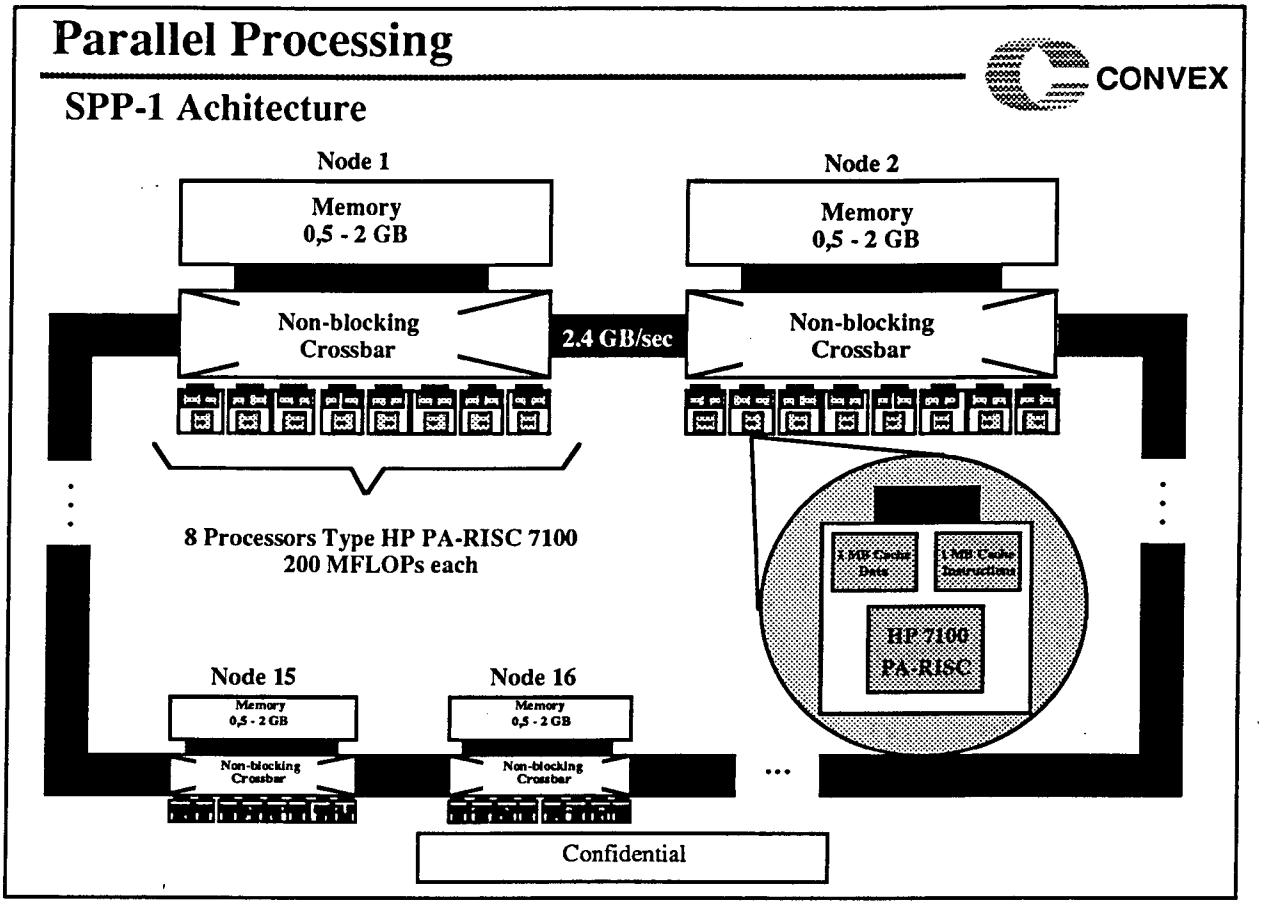


Abb.: 3

Die **CONVEX SPP-1 Architektur** zeichnet sich durch folgende Leistungsmerkmale aus:

- **MIMD**

  Die MIMD-Architektur hat gegenüber der (veralteten) SIMD-Architektur folgende Vorteile:
  - Höhere Einzelprogramm-Leistung
  - Besserer Durchsatz

- **Verwendung von Industriestandard Prozessoren** (HP PA-RISC 7100)

  Nur Prozessor-Chips welche in großen Stückzahlen bei Hochleistungs-workstations eingesetzt werden, gewährleisten eine zukünftige Leistungs-steigerung von 50 - 100 % alle 2 Jahre.
  Proprietäre Chips fallen demgegenüber in der Leistungssteigerung zurück, da die immensen Entwicklungskosten nicht aufgebracht werden können (siehe Abb. 3: KSR, Ncube, Parsytec).

- **Anwenderfreundliches Programmiermodell**

  CONVEX SPP-1 Systeme sind als Produktions-Systeme gedacht. Die Programmierung muß daher einfach sein, damit schnell reale Produktionsanwendungen verfügbar sind.

Aus diesem Grund wurde folgendes Programmier-Modell gewählt:

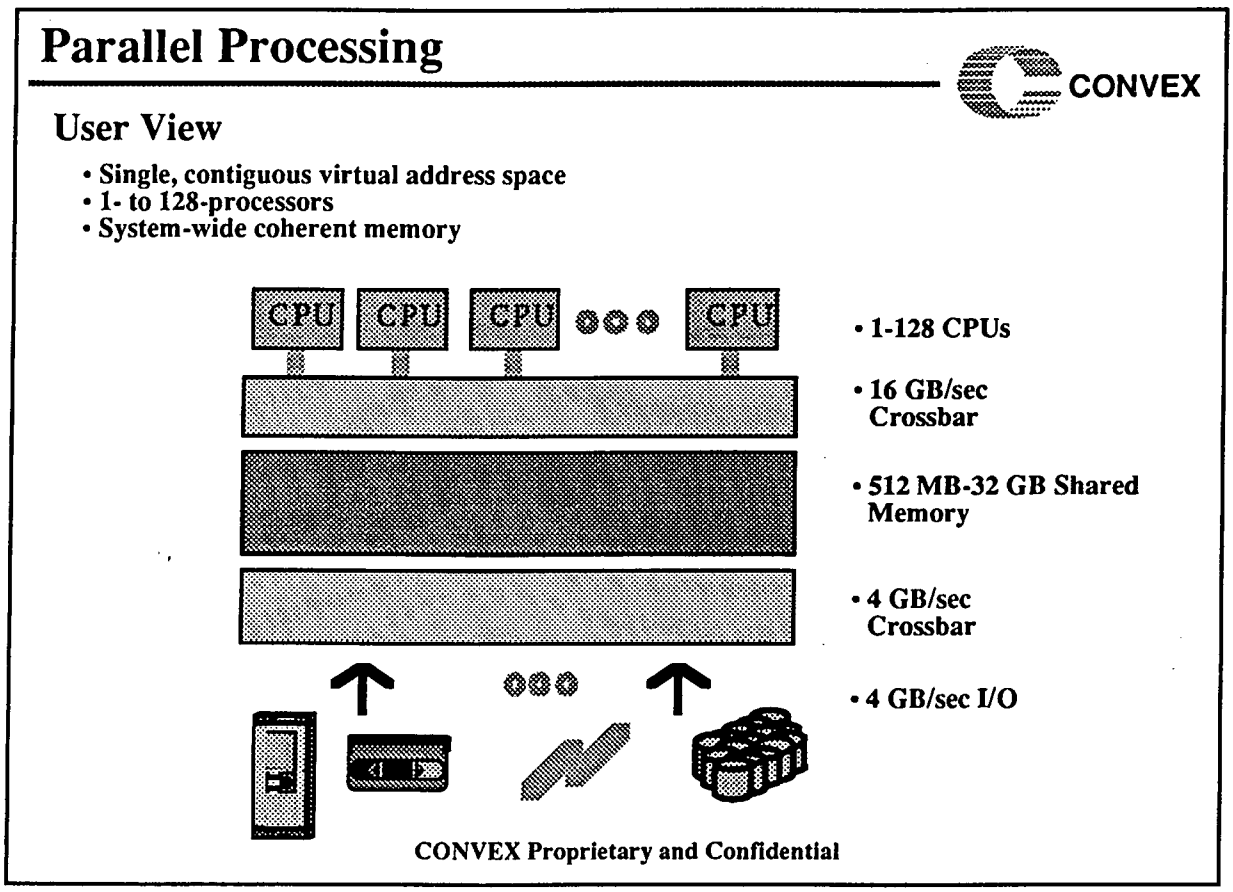


Abb.: 4

Der Benutzer (Programmierer) sieht eine Shared Memory Maschine mit 1 - 128 Prozessoren, sehr großem Hauptspeicher und großer Memory- und I/O-Bandbreite.

- **Globally Shared Distributed Virtual Memory (GSDVM)**
  - Evolution aus der CONVEX C-Serie (Shared Memory)
  - 64-Bit Adressraum
  - Speicherhierarchie (autom.): Cache, local, global (d.h. remote)
  - Skalierbare Architektur: 1 - 8 Knoten à 8 Prozessoren
  - Autom. Parallelisierung durch Compiler (FORTRAN, C, C++)

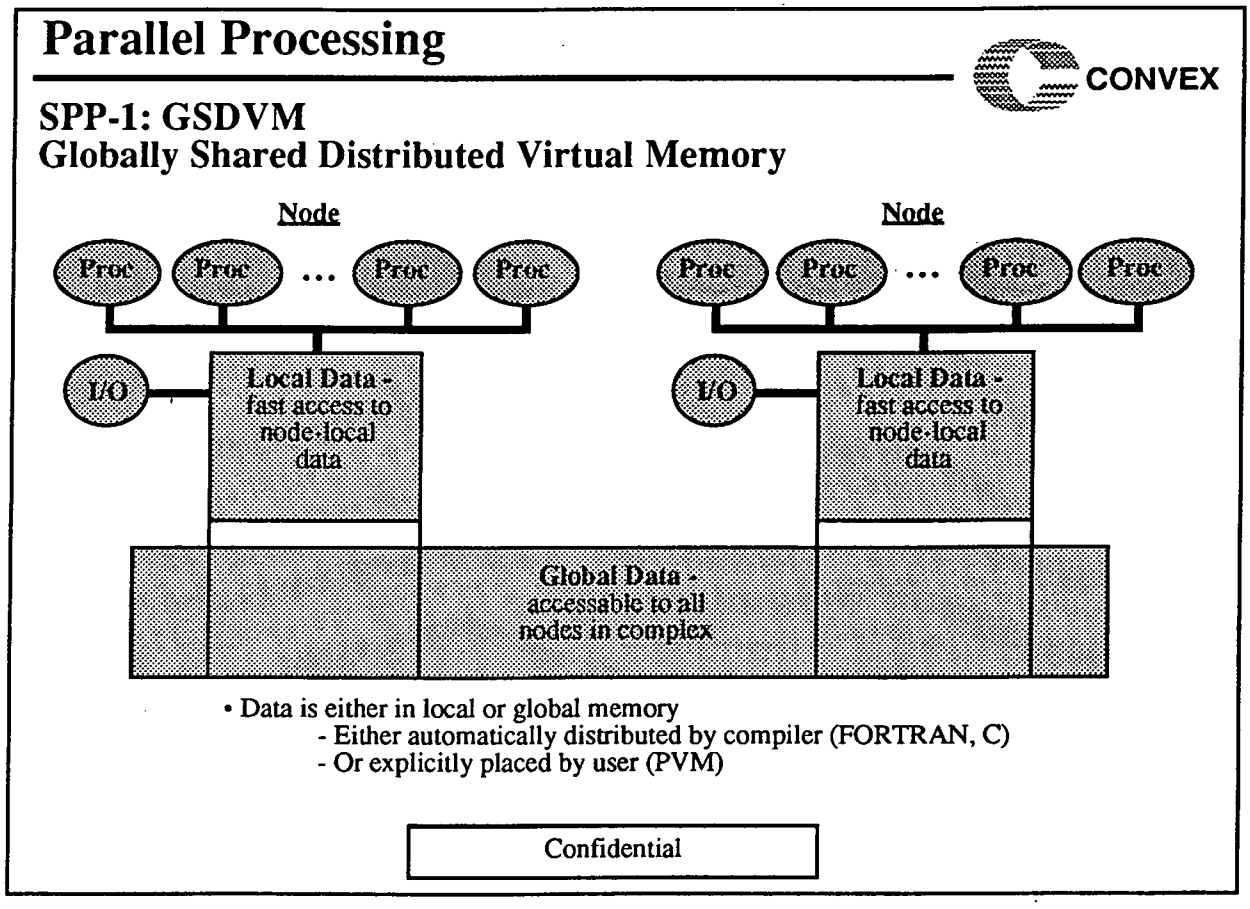


Abb.: 5

- **Hohe Leistung auch für schlecht parallelisierende Anwendung**

Robustheit des Systems gegenüber unterschiedlicher Anwendung ist eine Hauptforderung, will man dem Anspruch eines Produktionssystems für den techn./wiss. Einsatz gerecht werden.

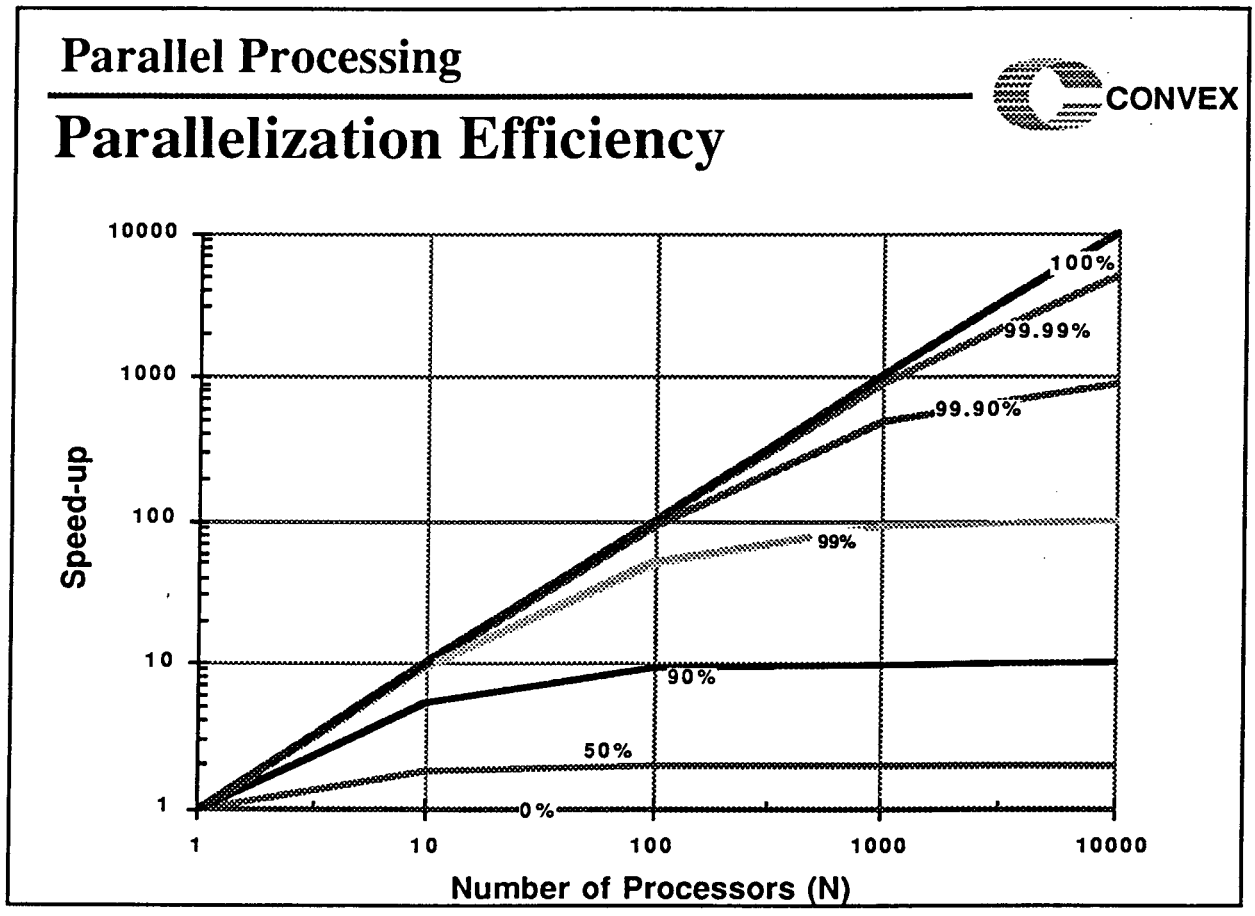


Abb.: 6

Aus diesem Grund wird im CONVEX SPP-1 System der derzeit leistungsstärkste Prozessor, der HP PA-RISC 7100, eingesetzt.

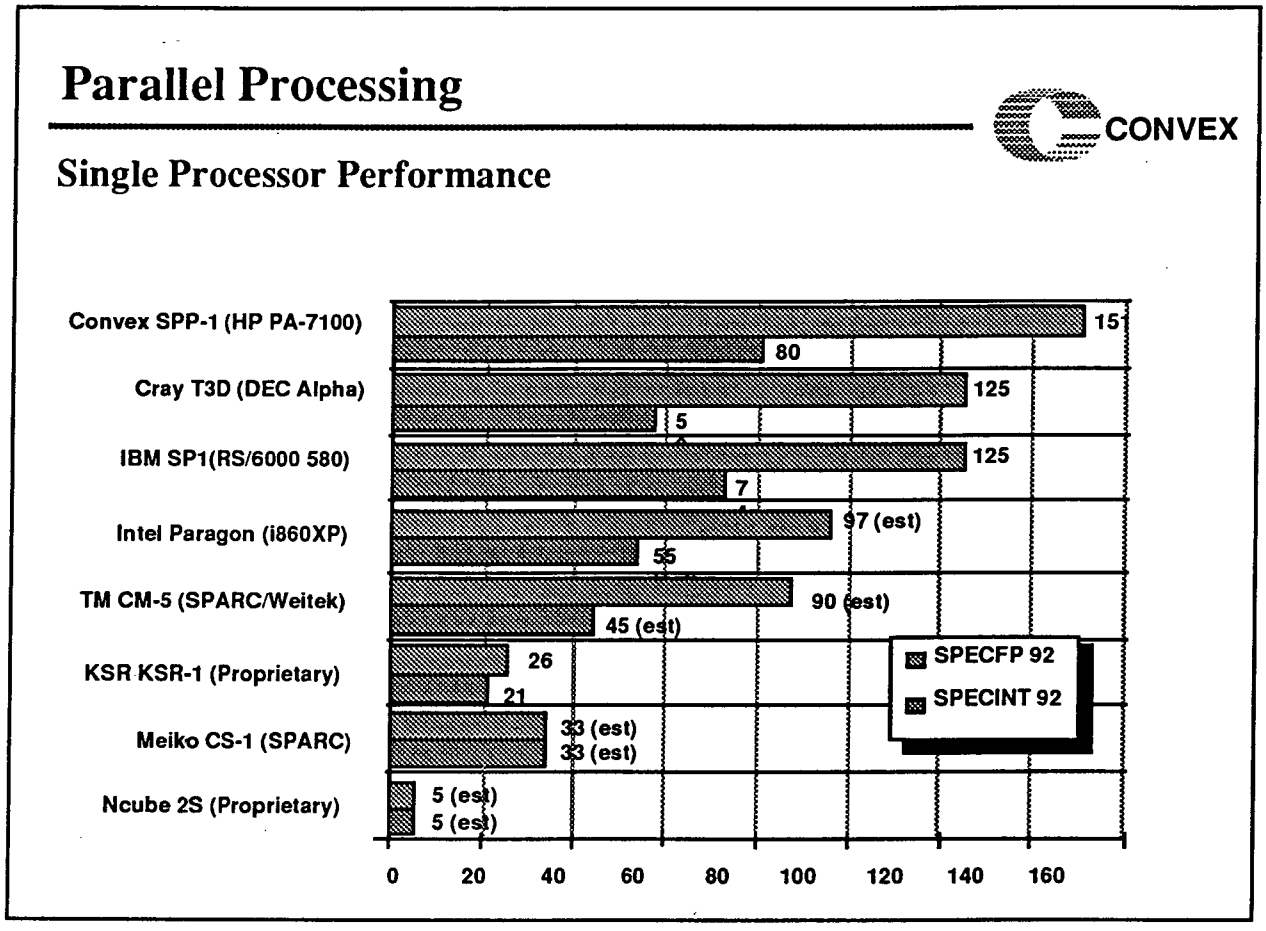


Abb.: 7

Dieser Prozessor ist, bei einer Taktfrequenz von ca. 100 MHz und einer Peak-Performance von 200 MFLOPs den in bisher verfügbaren MPP-Systemen, wie INTEL Paragon, TMC CM-5 oder gar KSR KSR-1, eingesetzten Prozessoren, deutlich überlegen.

- **Skalierbarkeit SPP-1**

Das CONVEX SPP-1 System ist wie folgt skalierbar:

Parallel Processing

CONVEX

SPP-1 Skalierbarkeit

Anzahl Prozessoren	:	**8 - 128**
Anzahl Knoten (à 8 Prozessoren)	:	**1 - 16**
Peak-Leistung pro Prozessor	:	**0,2 GFLOPs**
Peak-Leistung pro Knoten (8 Proz.)	:	**1,6 GFLOPs**
Peak-Leistung im Vollausbau	:	**25,6 GFLOPs**
Max. Memory pro Knoten	:	**2 GB**
Max. Memory imVollausbau	:	**32 GB**
Interconnect Bandbreite (Ring)	:	**2.4 GB/s**
Interconnect Latency (Ring)	:	**< 3 µs**

Confidential

Abb.: 8

- **SPP-1Architektur**

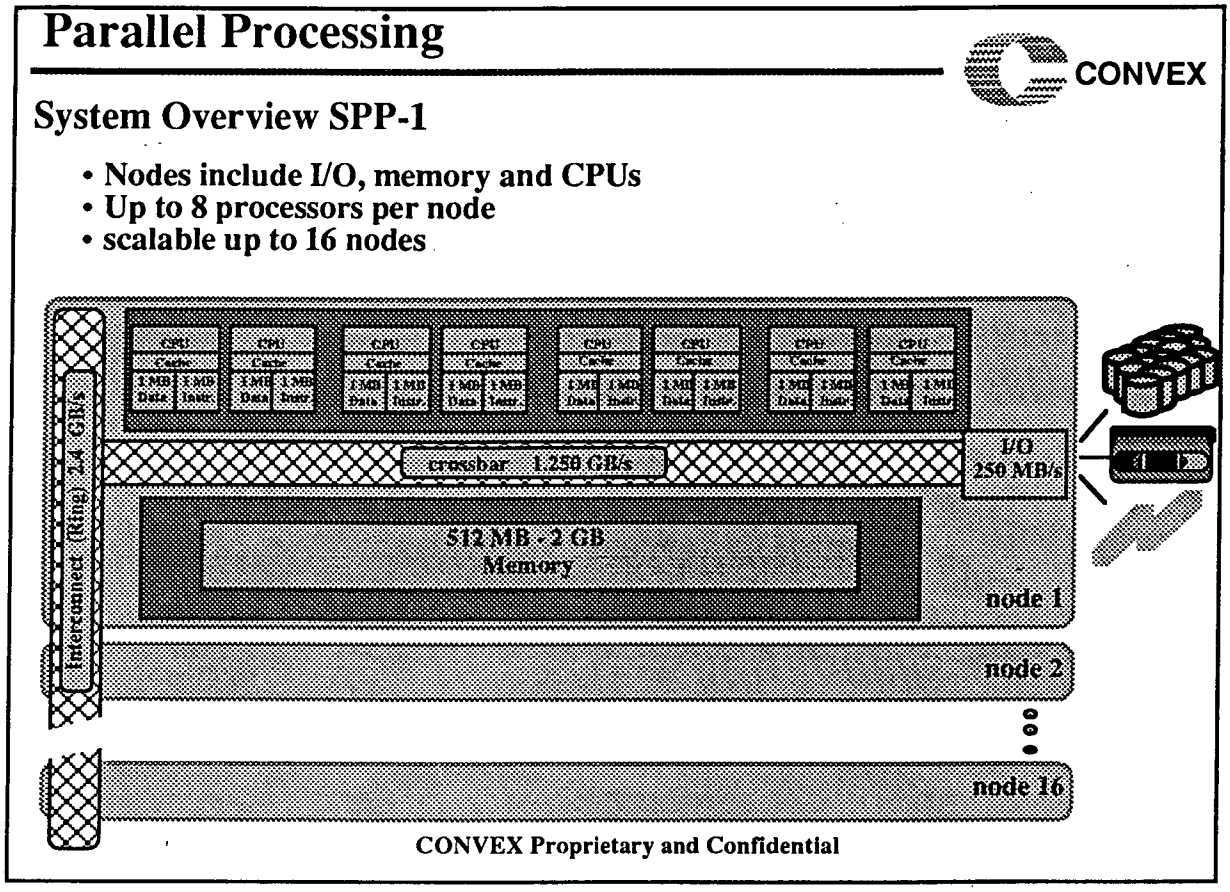


Abb.: 9

Das CONVEX-SPP-1 Parallelrechner-System besteht aus 1 bis 16 Knoten, welche durch einen Ring miteinander verbunden sind. Die Bandbreite des Ringes beträgt 2.4 GB/sec; die Latency < 3 µs. Jeder Knoten besteht aus 8 Prozessoren des Typs HP PA-RISC 7100 mit 200 MFLOPs Peak-Performance für die Verarbeitung von 64-Bit Gleitkommazahlen.
Die 8 Prozessoren innerhalb eines Knotens haben Zugriff auf einen lokalen Hauptspeicher im Sinne einer Shared Memory Maschine (analog der CONVEX C-Serie). Jeder Prozessor hat einen Daten- und Instruktions-Cache von je 1 MB. Das Programm sieht einen globalen Adressraum, der sich über alle Knoten erstreckt. Werden Daten angesprochen, die außerhalb eines Knotens liegen, so erfolgt dieser Zugriff automatisch durch die Hardware. Es wird ferner Demand Paging unterstützt, deshalb sprechen wir von einem Globally Shared Distributed **Virtual** Memory (GSDVM) Programmier-Modell, das durch den Compiler direkt unterstützt wird. Im Gegensatz dazu steht ein Message Passing Modell, bei dem der Programmierer im Quellcode von Hand den Datenaustausch steuern muß (z.B. mittels PVM). Dieses Programmier Modell wird ebenfalls unterstützt.
Das Betriebssystem (OSF/1 AD Mach Kernel) läuft symmetrisch (SMP) auf allen Prozessoren innerhalb eines Knotens.
Das CONVEX SPP-1 System unterstützt unter OSF/1 das HP Application Binary Interface (ABI). Damit sind vom ersten Tag an über 5000 Anwendungen in einer Produktionsumgebung verfügbar. Diese laufen auf jeweils einem Prozessor. Parallele Versionen werden in kurzer Zeit verfügbar sein.

II. 1. CPU-Leistung

Z. B. ein System besteht aus 32 Prozessoren (Variante 1 und 2) und 48 Prozessoren (Variante 3) des Typs HP PA-RISC. Die Einzel-CPUs haben folgende Leistungsdaten:

- 100 MHz Taktfrequenz
- 200 MFlops Peak (64 bit)
- 200 Mips
- IEEE 754 Floating Point Format
- IEEE 754 Floating Point Arithmetik

daraus ergibt sich für das Gesamtsystem:

- 6400 MFlops Peak (Variante 1+2) bzw. 9.600 MFlops (Variante 3)

Folgende Leistungen bei Standardbenchmarks werden erreicht (pro Prozessor):

- Linpack 1000x1000: 107 MFlops
- SpecInt92: 80
- SpecFloat92 151

II. 2. Kommunikation

Jeder Knoten besitzt sein lokales Memory (bis 2 GByte), der Zugriff der CPUs darauf erfolgt über einen non-blocking Crossbar. Die Summe der lokalen Memories ist das globale Memory, das durch ein Ring-Netzwerk verbunden ist (siehe auch Abb.3).
Aus Sicht einer CPU gibt es nur einen (globalen) Adressraum, jedoch entsteht je nach Ort der Daten eine unterschiedliche Zugriffszeit. Dabei sind folgenen Stufen möglich:

(1) Daten im Cache
(2) Daten im lokalen Memory
(3) Daten im Netzwerk-Cache (= lokales Memory)
(4) Daten im Memory eines anderen Knotens

	Latency	Transferrate
(1)	< 0.01 usec	800 M Byte/sec
(2)+(3)	< 0.5 usec	250 M Byte/sec
(4)	< 3 usec	160 M Byte/sec

Für typisch Software, die die Cache-Lines ausnutzt ergibt sich daraus folgende Zugriffszeit, wobei die Cache-Zugriffszeit 1 sei, größere Zahlen bedeuten langsameren Zugriff:

(1)	1
(2)+(3)	5
(4)	15 - 20

Die Cache Größen der CPUs sind 1 MByte Daten und 1 MByte Instruktions Cache, die Cache-Lines sind 32 Byte. Die Größe des Netzwerk-Caches ist konfigurierbar während des Bootens.

Betont sei, daß die Cache Coherency in Hardware durchgeführt wird.

II. 3. Memory-Architektur (Programmier-Modell)

Wie bereits unter I.2 erwähnt hat das System ein "Globally Shared Distributed Virtual Memory". Dadurch ist es möglich (unterstützt von den Compilern), das System durch ein sogenanntes "Data Parallel" Programmier-Modell zu programmieren. Das bedeutet, daß Source-Code nicht explicit in Host und Node Programme aufgesplittet und mit Parallelanweisungen versehen werden muß, sondern der Compiler (eventuell unterstüzt durch Direktiven) automatisch Parallel-Code erzeugt.

Ferner werden pthreads wie auch Message Passing Libraries z.B. PVM unterstützt.

II. 4. Wachstumspfad

Das angebotene System kann bis zu 128 Prozessoren (16 Knoten) ausgebaut werden. Zukünftige Systeme werden bis zu 512 Prozessoren (8x8 Knoten) haben. Die Binärkompatibilität bleibt erhalten (nächste Generation von HP PA-RISC).

II. 5. Compiler und Bibliotheken

Es werden folgende Compiler angeboten, wobei alle Compiler das oben beschriebene Programmiermodell unterstützen, also auto-parallelisierend sind:

- Fortran77, mit Fortran90 Erweiterungen zur Matrixverarbeitung
- Ansi-C
- Convex Interprocedural Optimizer. Dieser, bereits auf der C-Serie vorhandene Compiler, analysiert eine Applikation als Ganzes.
  Dadurch werden nicht nur logische Fehler zwischen einzelnen Routinen gefunden, sondern vor allem Datenabhängigkeiten geprüft und damit automatisches (= nicht durch Direktiven gesteuertes) Parallelisieren ganzer Unterprogrammbäume ermöglicht.

Ein HPF Compiler ist in Planung, Convex ist Mitglied im HPF Kommitee.

Als Bibliothek steht eine hochoptimierte "Vektorlibrary" zur Verfügung. Diese enthält neben einfachen Matrixoperationen folgende Standardlibraries:

- Linpack
- Eispack
- Blas, Blas2 und Blas3
- Minpack

II. 6. Betriebssystem

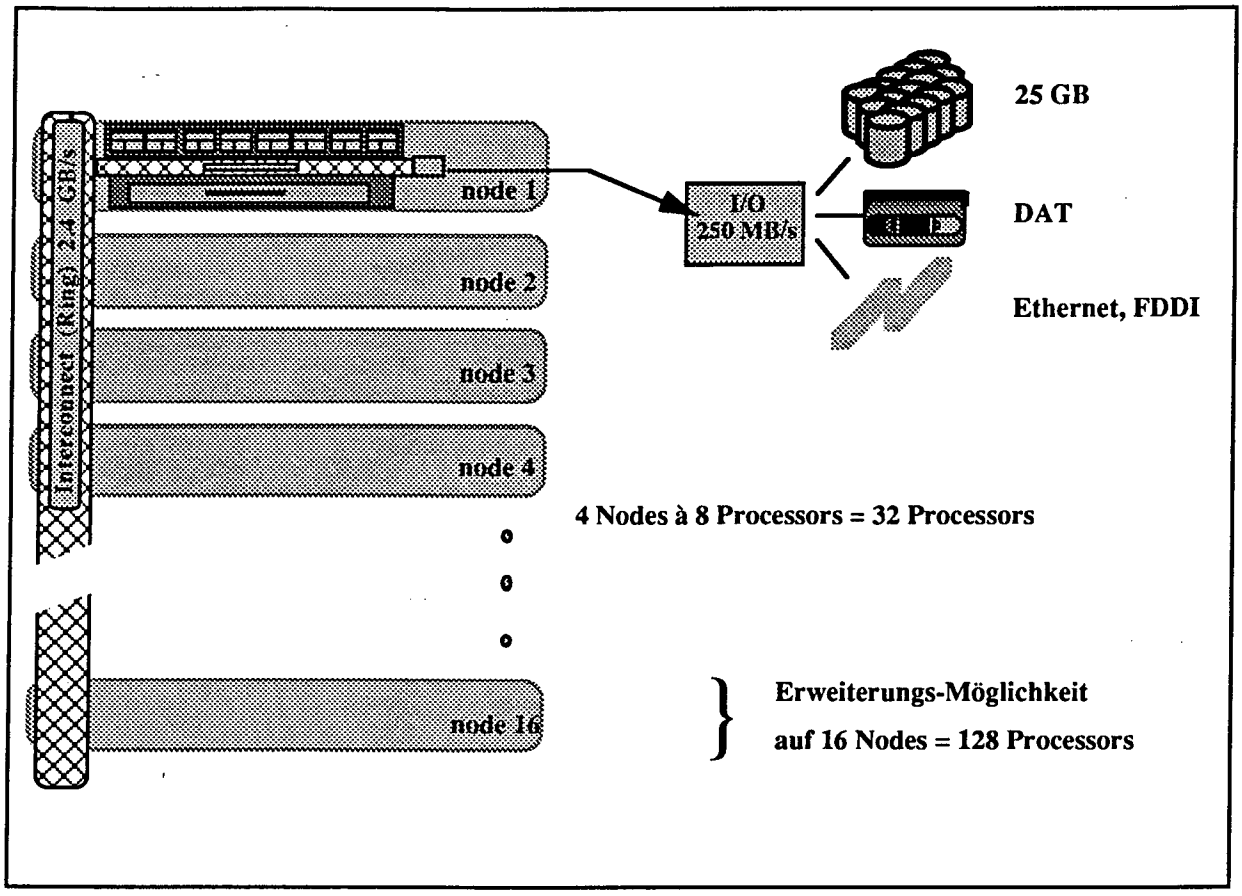


Abb. 10

Das Betriebssystem kann im Sinne des symetrischen Multiprozessings auf jedem SPP-1 Prozessor laufen. Der Kernel ist OSF/1 AD, der auf dem „Mach 3.0 Micro Kernel" beruht.

Am API (Applikations Program Interface) erscheint das System als HP/UX, mit all seinen Standard-Utilities. Durch Unterstützung des HP/UX Application Binary Interface (ABI) sind mehr als 5000 Anwendungsprogramme verfügbar.

Es existieren sowohl das HP/UX als auch das ConvexOS Systemcall Interface.

Erweiterungen des OSF/1 AD beinhalten die Partitionierbarkeit des Systems sowie verbesserte I/O Performance.

II. 7. Partitionierbarkeit

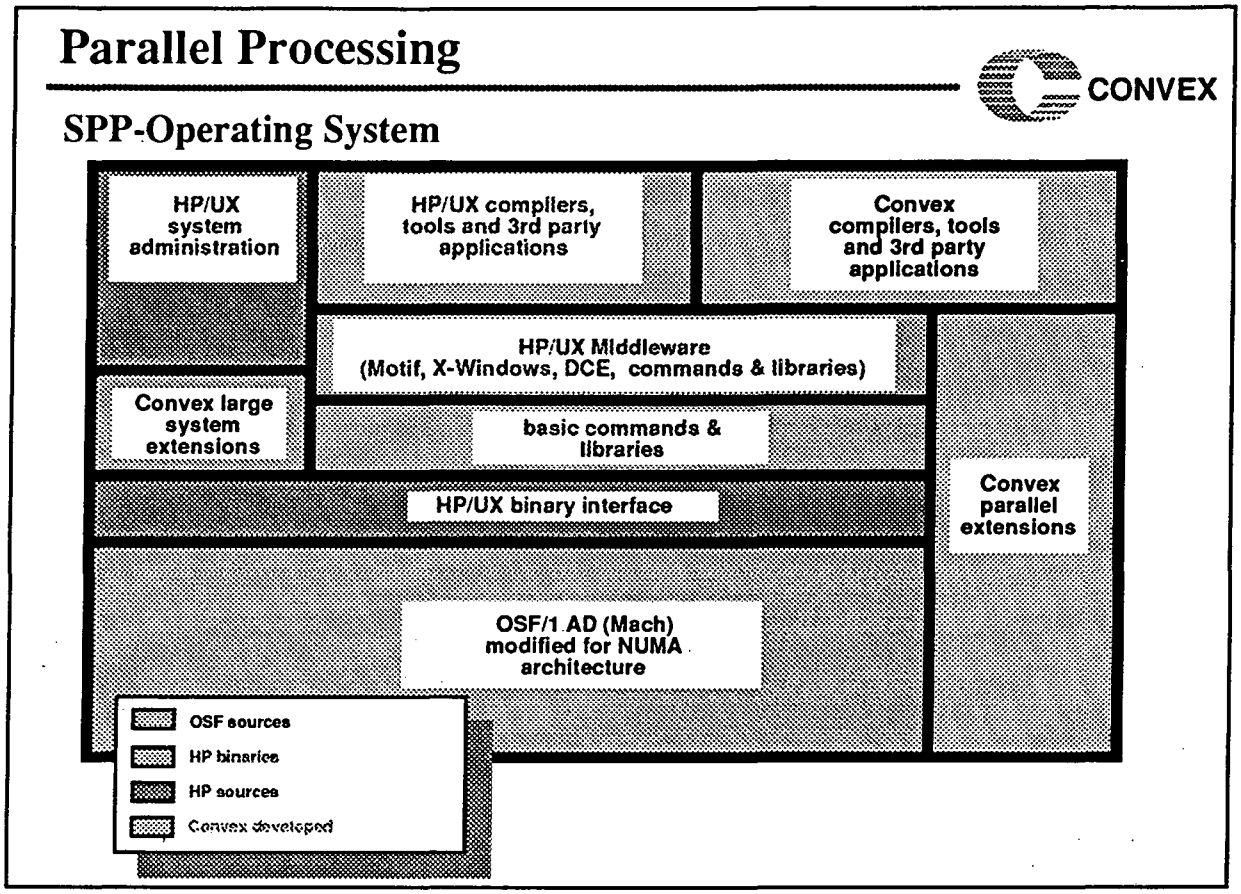


Abb.: 11

Das System kann in beliebig viele Partitionen mit beliebig vielen Prozessoren unterteilt werden. Resource-Beschränkungen stehen im Rahmen der Standard Unix Möglichkeiten sowie dem NQS Batch Systems zur Verfügung. Ein Accounting System ist vorhanden.

II. 8. I/O System

Jeder Knoten hat sein eigenes I/O-System. Die Platten sind SCSI Disken mit Raid 5. Die Geschwindigkeit des SCSI Busses beträgt 20 MByte/sec (fast/wide SCSI).
Die Verteilung auf andere Knoten erfolgt mit 160 MByte/sec (siehe auch II.2 Kommunikation).

II. 9. System Überwachung

Tools dazu sind in Arbeit, Details sind jedoch noch nicht bekannt.

II. 10. Parallelisierungswerkzeuge

Die Parallelisierung auf dem System erfolgt über die autoparallelisierenden Compiler. Weiter werden folgende Standard Tools zur Verfügung stehen:

- PVM
- Linda
- MPI (wenn/falls der Standard definiert ist)
- Forge 90
- PARMACS
- andere

II. 11. Debugger/Profiler

Essentiell ist, daß die gelieferten Profiler/Debugger voll den automatisch parallelisierten und optimierten Code (wie bereits jetzt auf der C-Serie) unterstützen.
Es stehen zur Verfügung:

- CXpa
- CXdb
- Performance Monitor (HW)

II. 12. Zukünftige Standards

DCE, DFS, AFS und ATM werden unterstützt.

II. 13. Nutzung des Rechners

Das System kann sowohl interaktiv als auch im Batch betrieben werden, wobei der Zugriff auf Partitionen erfolgt. Der interaktive Betrieb erfolgt mittels „telnet“ oder „rlogin“, als Batchsystem steht NQS zur Verfügung.

II. 14. Netz-Anbindung

Es stehen folgende HW-Medien zur Verfügung:

- Ethernet
- FDDI
- zukünftig HiPPI
- zukünftig Fiberchannel

Als SW-Protokoll wird TCP/IP angeboten.

II. 15. Optimierende Matrix Multiplikation

Die Parallelisierung einer Matrix-Multiplikation wird automatisch vom Compiler durchgeführt. Darüber hinaus stehen das FORTRAN 90 Instrinsic „matmul“ sowie die Vector-Library-Routinen sgemm/dgemm zur Verfügung.

II. 16. Graphikanschluß

Grundlage für das Graphik-System ist X-Windows. Als „konfigurierbare" Graphikapplikation steht AVS zur Verfügung (siehe Datenblatt).

II. 17. Infrastruktur-Anforderungen

Bei dem angebotenem CONVEX SPP-1 Parallelsystem handelt es sich um ein luftgekühltes System. Für den Betrieb reichen die Standard-Einrichtungen eines Rechenzentrums aus.

Leistungsaufnahme:	ca. 2.5 kW pro Knoten, d.h. ca. 10 kW für das angebotene System bzw. 12.5 kW
Raumtemperatur:	21 - 27 ° C
Stellfläche:	ca. 41 x 92 cm pro Knoten, d.h.
zusätzliche Servicefläche:	ca. 1 m rundherum

II. 18. Praktische Erfahrungen

Das angebotene System wird im 4. Quartal 1993 für Benchmarks zur Verfügung stehen. Praktische Erfahrungen liegen für die Programmierung mit PVM auf der CONVEX MetaSerie vor.
Automatische Parallelisierung per Compiler in FORTRAN und C kann heute schon demonstriert werden. Dazu dient die C-Serie mit ihren bis zu 8 Prozessoren und Shared Memory, dies entspricht einem Knoten im SPP-1 System.

II. 19. Wartungskonzept

Betriebsstörungen (HW oder SW) werden an die Techn. Supportzentrale (0130/90 40) gemeldet. Von dort, oder von den Geschäftsstellen, wird eine Ferndiagnose durchgeführt. Falls erforderlich wird ein Kundendiensttechniker die Reparatur vor Ort vornehmen. Es wird eine Responsezeit von max. 8 h bis zum Eintreffen des Technikers eingehalten. Die Wartungszeiten sind MO bis FR von 8°° - 17°° Uhr. Erweiterte Wartungszeiten (24 h, 7 Tage) sowie kürzere Reaktionszeiten sind gegen Aufpreis möglich.

Ersatzteile werden zentral (Frankfurt) bzw. auch dezentral in den Geschäftsstellen (für Hannover: GS Hamburg) gelagert. In allen Geschäftsstellen sind sowohl HW- als auch SW-Supportingenieure verfügbar.

Physical Characteristics

CONVEX

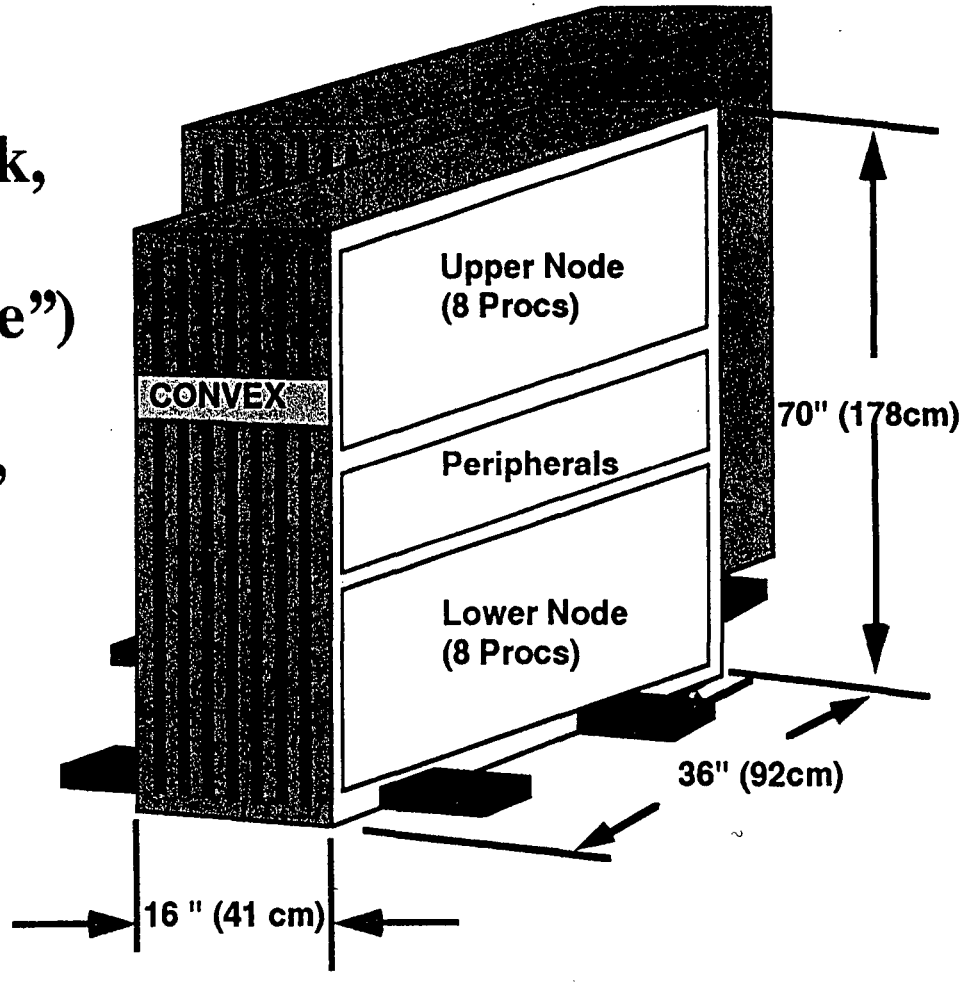
2500 watts/node

Built-in peripherals (disk, tape, and connectivity)

~65dBA ("business office")

Completely air-cooled

Click-together "Lego™" node expansion



- Four-node (32 processor) configuration