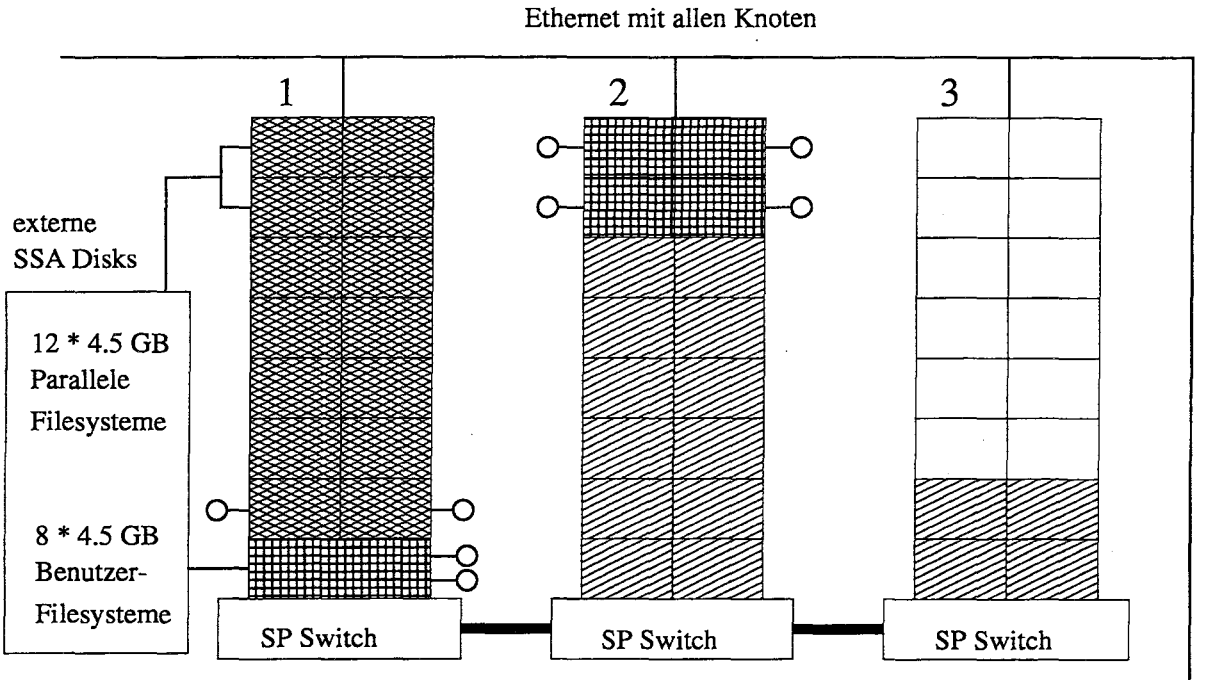
HRZ Uni Marburg

Parallelrechner IBM SP



Frame 1

4 Thin Nodes 2 à 256 MB Mem, 4.5 GB Disk
10 Thin Nodes 2 à 256 MB Mem, 4.5 GB Disk, L2 Cache
1 Wide Node mit 512 MB Mem, 4 * 4.5 GB Disk

Frame 2

4 Thin Nodes 2 à 512 MB Mem, 2 * 4.5 GB Disk, L2 Cache
12 Thin Nodes 2 à 128 MB Mem, 2.2 GB Disk

Frame 3

4 Thin Nodes 2 à 128 MB Mem, 2.2 GB Disk

Stand: 31.12.97

5.4 Parallelrechner IBM SP

Der Parallelrechner IBM RS/6000 SP ist ein Distributed Memory System; bevorzugtes Programmiermodell für parallele Anwendungen ist Message Passing auf der Basis von Bibliotheken wie MPI oder PVM; serielle Nutzungen sind ebenfalls möglich.

Der Parallelrechner ist Anfang Oktober 1995 geliefert und anschließend im Hochschulrechenzentrum installiert worden; die offizielle Freigabe für Benutzer erfolgte am 15.12.95, die Einweihung am 15.05.96. In 1996 ist der Parallelrechner mit einem neuen Switch ausgestattet worden; deshalb war in 1995 auf die Beschaffung des alten HPS+ Switches für das gesamte System verzichtet worden (vgl. vorangegangene Jahresberichte). Für 1997 war ursprünglich ein Ausbau geplant, der aber nicht realisiert werden konnte.

Dokumentationen zur Hardware- und Software-Ausstattung, zum Benutzer-Zugang und zum Betrieb, etc. wurden von Anfang an im WWW bereitgestellt (vgl. http://www.uni-marburg.de/hrz/sp/welcome.html). Darüber hinaus werden laufend Workshops für Benutzer angeboten. Das Accounting ist im März 1996 angelaufen.

Der Parallelrechner kann von allen hessischen Hochschulen genutzt werden; aufgrund der Finanzierung kann die Universität Marburg 60 % der Rechenzeit für ihre Nutzer beanspruchen. Es können sowohl fertige parallele Anwendungen eingesetzt als auch neue entwickelt werden; schließlich können einzelne Knoten des Parallelrechners auch seriell genutzt werden.

In der TOP 500 Supercomputer-Liste (von Dongarra, Meuer und Strohmaier) belegte der Parallelrechner in Marburg im November 1995 Rang 172, im Juni 1996 Rang 216 und im November 1996 Rang 272; im Juni 1997 war er bereits ausgeschieden.

5.4.1 Hardware-Ausstattung

Die IBM RS/6000 SP (wie Scalable POWERparallel System) besteht aus Knoten, die über einen Switch verbunden sind. Die Knoten entsprechen RS/6000 Workstations mit POWER2 Prozessoren; der Switch verbindet jeden Knoten mit jedem anderen Knoten. Der Übergang zum neuen Switch erfolgte Anfang Juli 1996; seitdem ist die Ausstattung unverändert geblieben.

IBM RS/6000 SP

Anzahl Frames	3	
Anzahl Knoten/Frame	max. 16	
Anzahl Knoten insgesamt	35	
Arbeitsspeicher insgesamt	8.2	GByte
Plattenspeicher intern insgesamt	152.2	GByte
Plattenspeicher extern insgesamt	90.0	GByte
Peak Floating Point Performance insgesamt	9.3	GFLOP/s

Ausstattung der Knoten

16 Thin Nodes 2 à 128 MB Arbeitsspeicher und 2.2 GB Plattenspeicher
14 Thin Nodes 2 à 256 MB Arbeitsspeicher und 4.5 GB Plattenspeicher , davon 10 mit 2 MB Level 2 Cache
4 Thin Nodes 2 à 512 MB Arbeitsspeicher, 2*4.5 GB Plattenspeicher und 2 MB Level 2 Cache
1 Wide Node mit 512 MB Arbeitsspeicher und 4*4.5 GB Plattenspeicher

Spezifikation der Knoten

Wide Nodes und Thin Nodes 2 gehören beide zur superskalaren POWER2-Architektur mit 1 Instruction, 2 Floating Point und 2 Integer Units, bis zu 6 Instructions/Cycle, 66.7 MHz und 266.7 MFLOP/s Peak Performance. Unterschiede sind:

	Wide Node	Thin Node 2
Daten Cache	256 KB	128 KB
Level 2 Cache	-	opt. 2 MB
Prozessor-Cache Bus	256 Bit	256 Bit
Arbeitsspeicher	64 MB - 2 GB	64 - 512 MB
Cache-Arbeitsspeicher Bus	256 Bit	128 Bit

Switch

Any-to-any Multi-Stage-Switch: 8 Bit parallel, bidirektional, max. 150 MB/s je Richtung (meßbar: max. 90 MB/s).

Die **Netzanbindung** des Parallelrechners erfolgt durch den Netzanschluß einzelner Knoten:

- Der Wide Node und 6 Thin Nodes 2 sind in ein FDDI-LAN integriert, welches an das FDDI-Backbone des UMRnet angeschlossen ist.
- Der Wide Node ist darüber hinaus in ein eigenständiges FDDI-LAN mit File-, Backup- und Archive-Server integriert.
- Alle Knoten sind mit den Control Workstations zu einem Ethernet-LAN verbunden (inkl. Anschluß an das FDDI-Backbone).

5.4.2 Software-Angebot

Die Software für die IBM RS/6000 SP wird auf allen Knoten bereitgestellt (meistens lokal, zum Teil via NFS vom Wide Node).

Systemsoftware

AIX	4.1.5	AIX Version 4 SPO
PSSP	V2.1	Parallel System Support Programs System Administration, Monitoring und Data Repository, Resource Manager und Switch Support
PE	V2.1	Parallel Environment inkl. MPL und MPI (s.u.) Parallel Operating Environment (POE) Parallel Debugger (PDBX) Parallel Profiling (prof, gprof) Visualization and Performance Monitoring Tool (VT)
RVSD	V1.1	Recoverable Virtual Shared Disk
LL	V1.3	LoadLeveler (inkl. NQS Interface)
PIOFS	V1.2.	Parallel I/O File System
MPE	1.0.11	Multiprocessing Environment Library

Message Passing Bibliotheken

PVMe	V2.1	Parallel Virtual Machine
MPI	1.1	Message Passing Interface

Sprachprozessoren

FORTRAN	V3.2.3	XL FORTRAN for AIX V3 & V4
HPF	V1.1.0	XL High Performance Fortran
C++	V3.1.4	C++ for AIX, inkl. C for AIX
PASCAL	V2.1.4	XL Pascal

Mathematische Bibliotheken

OSLp	V1.1.1	Parallel OSL Optimization Subroutine Library
PESSL	V1.2	Parallel ESSL Engineering and Scientific Subroutine Library
NAG	Mark 17	NAG Library (optimiert für POWER2-Architektur)

Anwendungssoftware

Gaussian94	Theoretische Chemie (seriell und parallel)

Tools

FORTRAN	f90convert, f90split, TkfPW,...
GNU-Tools	emacs, tar, gzip, patch, recode, Perl, RCS, bash,...
Graphik	xmgr, gnuplot
Monitor	rs2hpm: Power2 Hardware Performance Monitor
Sonstiges	xftp, ncftp, Tcl/Tk, tcsh,...

5.4.3 Betrieb und Nutzung

Betriebs-Konfiguration: Nach dem Einbau des neuen SP-Switches Anfang Juli 1996 wurde die Aufgabenverteilung zwischen den einzelnen Knoten neu konfiguriert. Zuvor waren 14 Knoten für den parallelen, 5 Knoten für den seriellen und 12 Knoten für den gemischt parallelen/seriellen Batch-Betrieb reserviert (aufgrund der Scheduling-Strategie des LoadLevelers konnte der gemischte Pool jedoch fast ausschließlich nur seriell genutzt werden); jeweils zwei Knoten waren für den interaktiven Zugang bzw. für kurze Test-Jobs reserviert. Seit dem Einbau des neuen Switches sind 16 Knoten für den parallelen und 17 Knoten für den seriellen Batch-Betrieb vorgesehen; jeweils 1 Knoten ist für den interaktiven Zugang bzw. für kurze Test-Jobs reserviert.

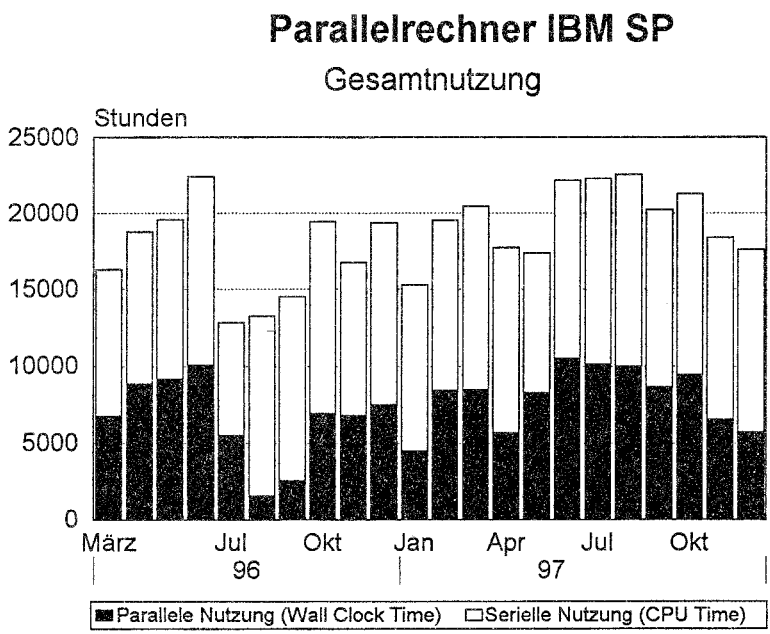
Bei den maximal in 1997 verfügbaren 365 * 24 * 35 = 306 600 Knotenstunden gab es die folgenden Ausfallzeiten:

Hardware-Störungen	2 094	Knotenstunden
Software-Störungen	574	Knotenstunden
sonstige Störungen	1 259	Knotenstunden
Wartung	2 541	Knotenstunden

Die indirekten Ausfallzeiten – verursacht durch den vorzeitigen Abbruch von Batch-Jobs bei Störungen – addieren sich auf ca. 3 000 Knotenstunden. Weitere unvermeidbare Ausfallzeiten ergeben sich durch das kontrollierte Zurückfahren der Batch-Queues vor Wartungsterminen. Die Ausfallzeiten von insgesamt ca. 9 500 Knotenstunden beliefen sich damit auf 3 %.

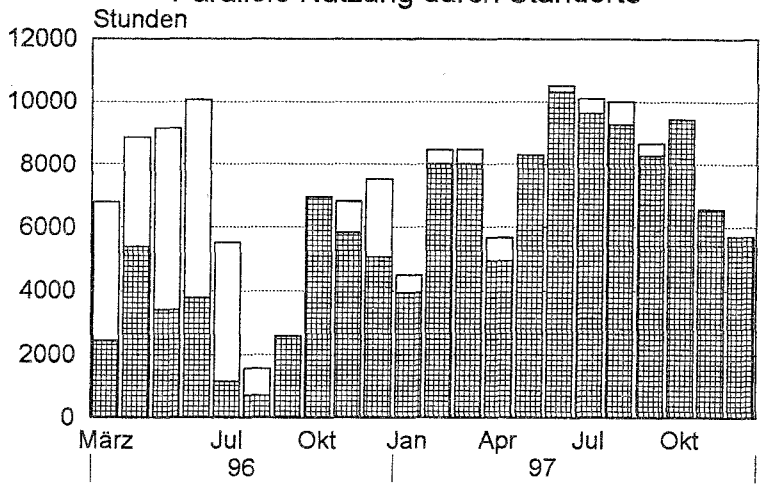
Accounting: Beim parallelen Batch-Betrieb werden die Knoten dediziert an die Jobs vergeben; die Nutzung wird deshalb in Knotenstunden gemessen (Wall Clock Time * Anzahl Knoten). Beim seriellen Batch-Betrieb laufen auf den Knoten maximal 2 Jobs (Time Sharing), so daß die Nutzung in CPU-Stunden gemessen wird.

Die nachfolgenden Diagramme zeigen die Gesamtnutzung der IBM RS/6000 SP, aufgeschlüsselt nach paralleler und serieller Nutzung, sowie die Nutzung durch die einzelnen Hochschul-Standorte bzw. durch die Arbeitsgruppen mit dem höchsten Bedarf an Rechenzeit, verteilt über alle Standorte.

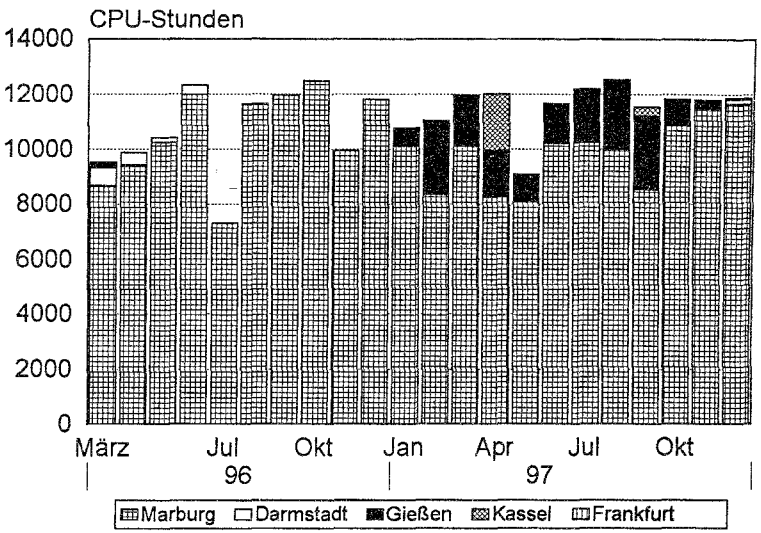


Parallelrechner IBM SP

Parallele Nutzung durch Standorte

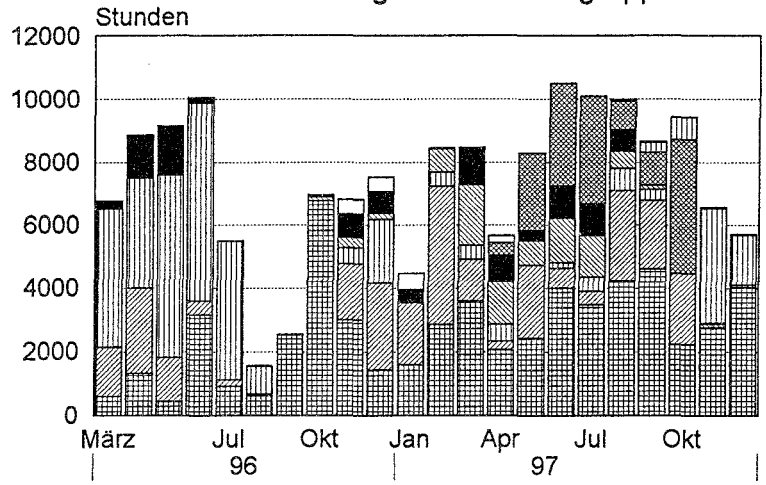


Serielle Nutzung durch Standorte



Parallelrechner IBM SP

Parallele Nutzung durch Arbeitsgruppen



Serielle Nutzung durch Arbeitsgruppen