Internet Anwendungen unter OS/390

Dr. rer. nat. Paul Herrmannn
Prof. Dr.rer.nat. Udo Kebschull
Prof. Dr.-Ing. Wilhelm G. Spruth

WS 2004/2005

Teil 2

zSeries Ein-/Ausgabe

Ein/Ausgabe Performance

Das Leistungsverhalten in großen kommerziellen C/S Systemen wird in der Regel weniger durch die CPU Geschwindigkeit und mehr durch dieLeistungsfähigkeit der Speicherverwaltung und des E/A Systems bestimmt.

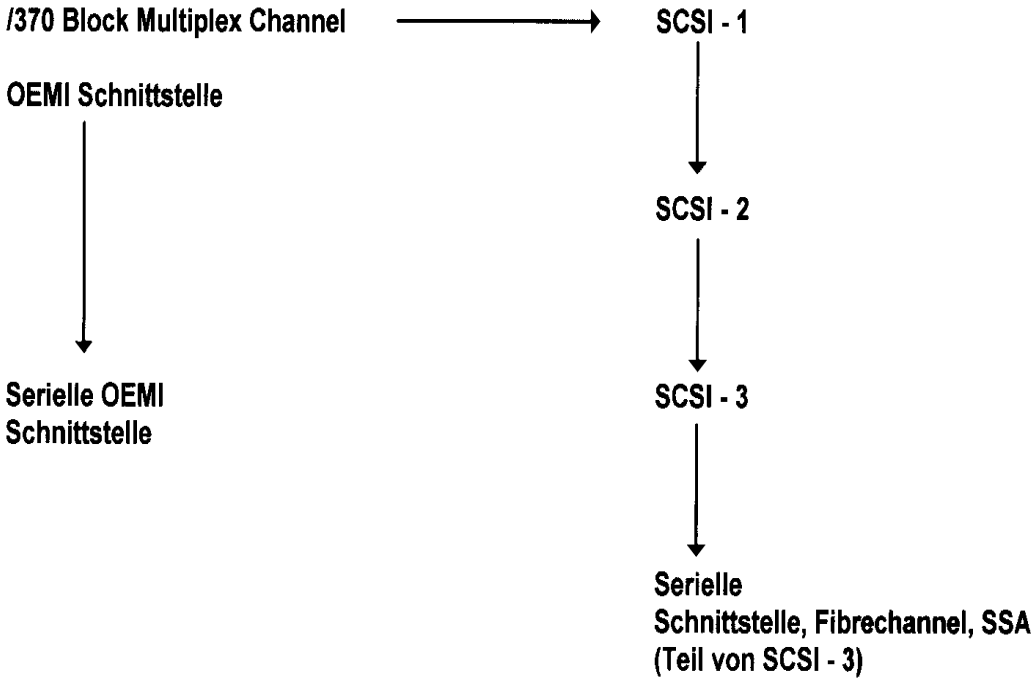
Es ist allerdings sehr schwierig das E/A Leistungsverhalten zu charakterisieren.

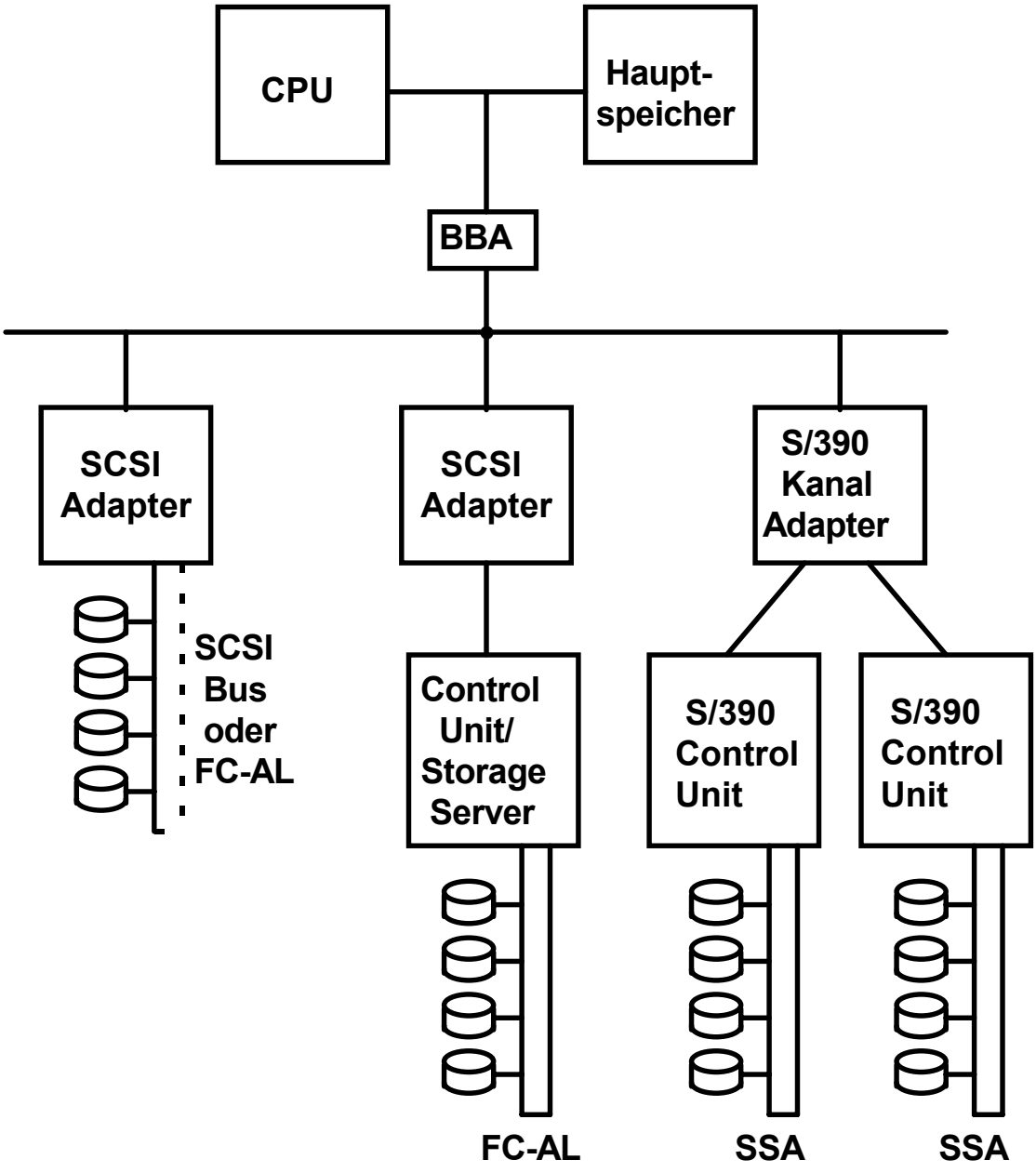
Eine Meßgröße ist die geamte maximale E/A Datenrate. Eine Angabe hierüber enthält das Februar 1996 Heft der Zeitschrift „Manufacturing Systems“. Hiernach kann das S/390 E/A Subsystem 1,000 bis 20,000+ MByte/Minute übertragen. Sehr große UNIX Systeme können 2 bis 100 Mbyte/Minute übertragen.

Ähnliche Ziffern gibt Price Waterhouse als Begründung für die Implementierung ihres Geneva ViewBuilder Produktes unter S/390 an.

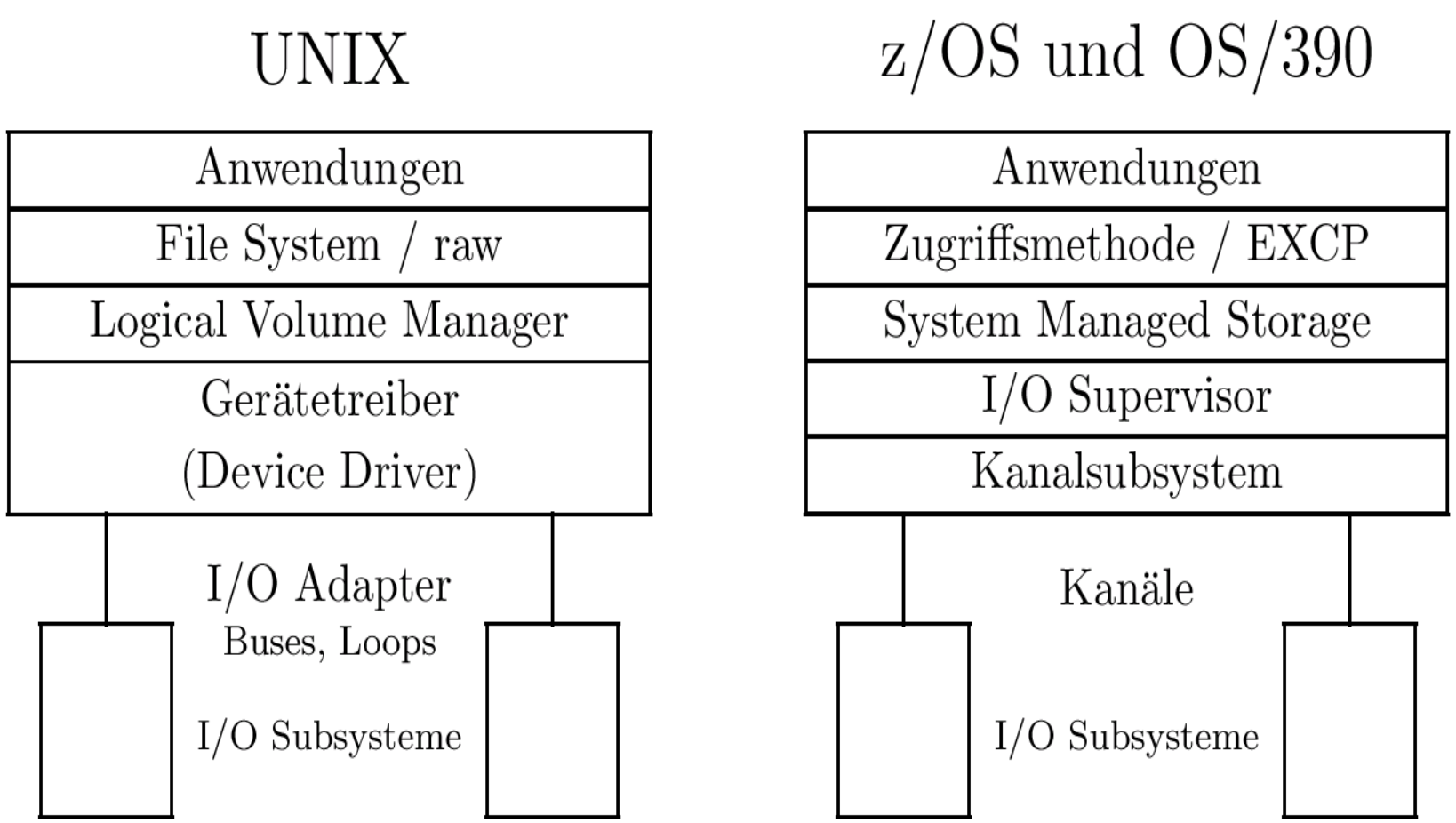
R. K. Roth, E.L. Denna: "Making good on a Promise". Manufacturing Systems (Chilton Publications), vol. 14, no.2, Feb. 1996, p.42-53.

Historische Entwicklung des Peripherie-Busses





Plattenspeicher Anschlußalternativen

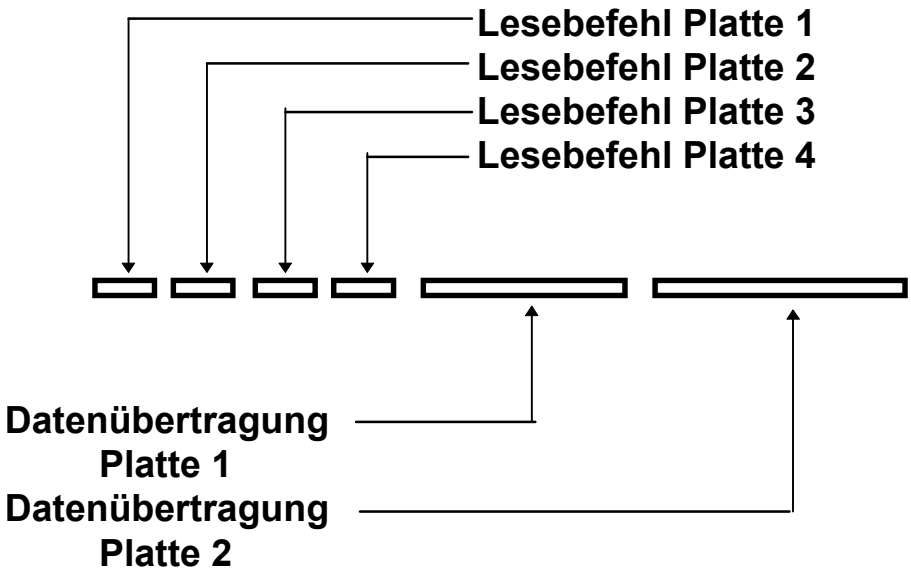

UNIX
Anwendungen
File System / raw
Logical Volume Manager
Gerätetreiber
(Device Driver)
I/O Adapter
Buses, Loops
I/O Subsysteme
z/OS und OS/390
Anwendungen
Zugriffsmethode / EXCP
System Managed Storage
I/O Supervisor
Kanalsubsystem
Kanäle
I/O Subsysteme

Aufgaben der Steuereinheit

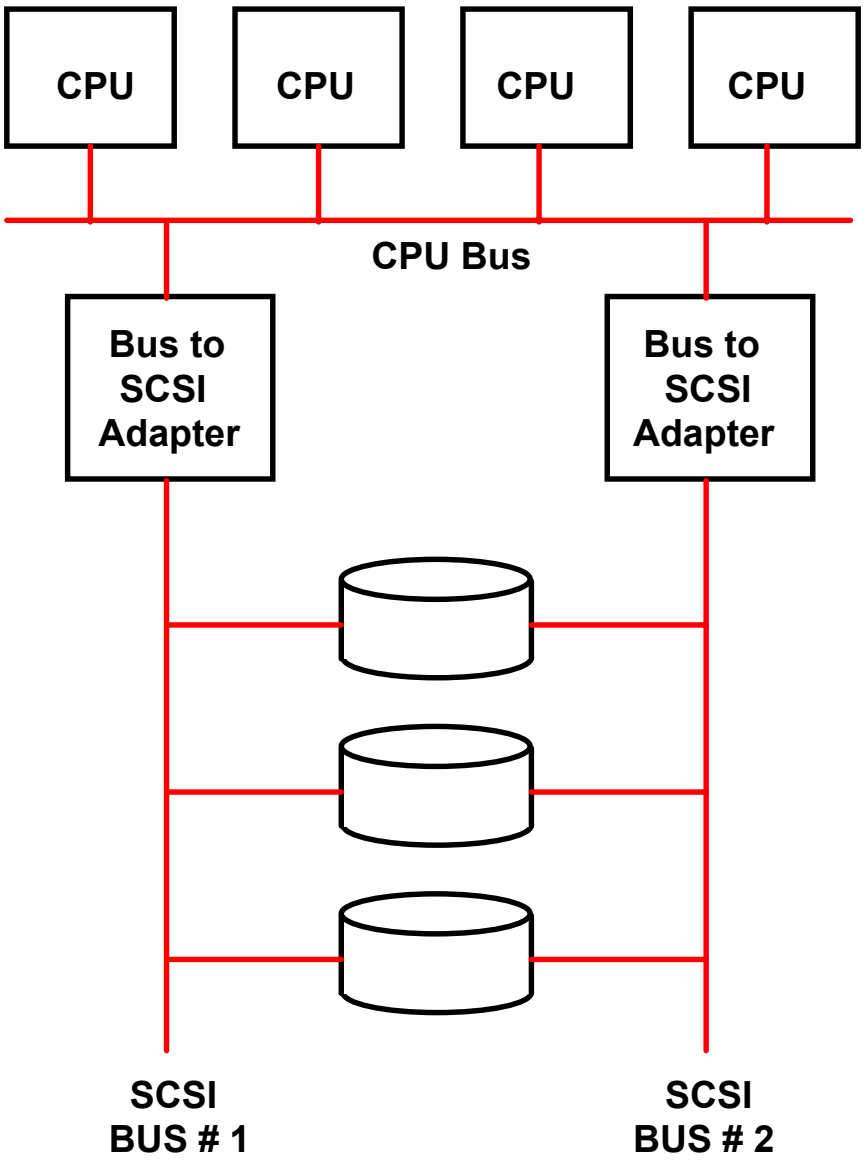
- **E/A - Kommandos (CCW) ausführen, z.B.**
  SEEK
  SEARCH
  READ
  WRITE
- **Command Chanining**
- **Fehlerkorrektur (permanente Fehler sind normal)**
- **E/A - Befehlswiederholung**
- **Statusinformation sammeln und an Zentraleinheit weitergeben**
- **Unterbrechungssignale erzeugen und an**
- **Zentraleinheit weitergeben (CEDE)**
- **Eine von mehreren Festplatten selektieren**
- **Cache - Non Volatile Cache**
- **RAID**

Aufgaben der Festplatten-Elektronik (Teil der Festplatte)

- **Umsetzen der magnetischen Lese / Schreibsignale in Folgen von Bits (R / W Channel)**
- **Spuranfangssignal**
- **Steuerung des Zugriffsmechanismus**
- **Lese / Schreibkopf selektieren (Plattenoberfläche)**
- **Fehler Erkennung (Syndrom Checking, Syndrom = 5 - 6 Bytes)**
- **Status setzen**



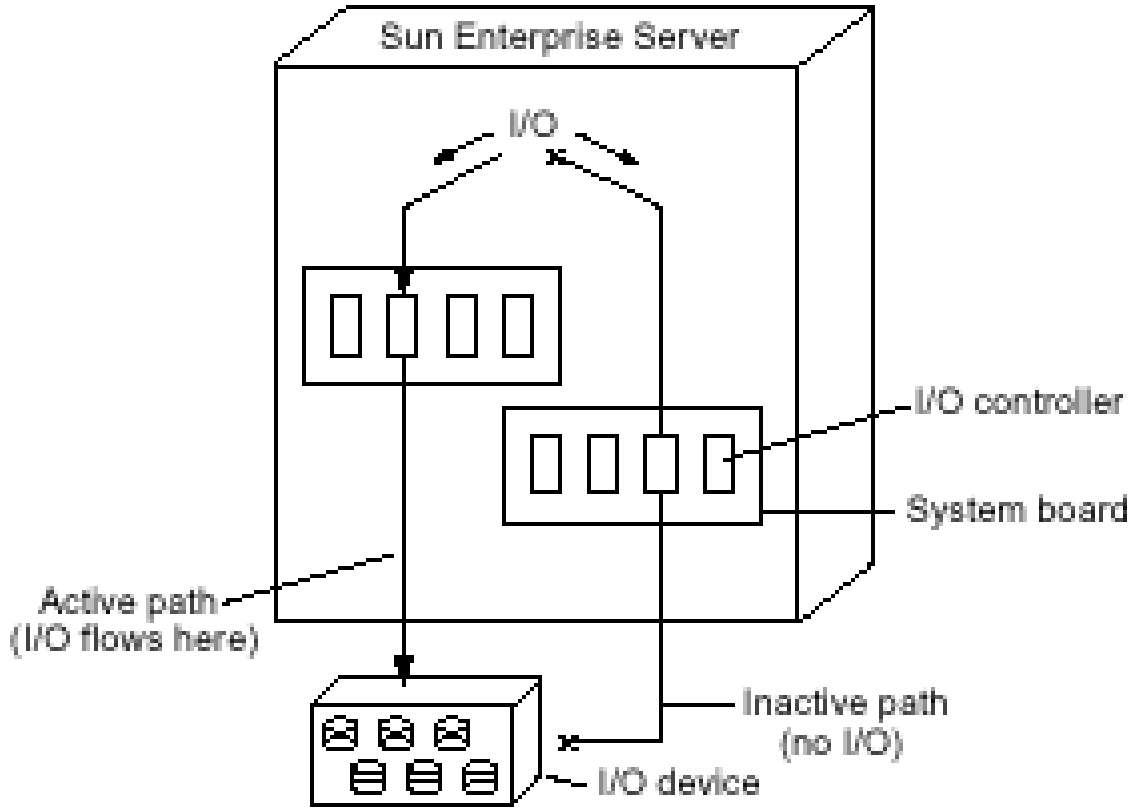
Verzahnte SCSI Plattenzugriffe



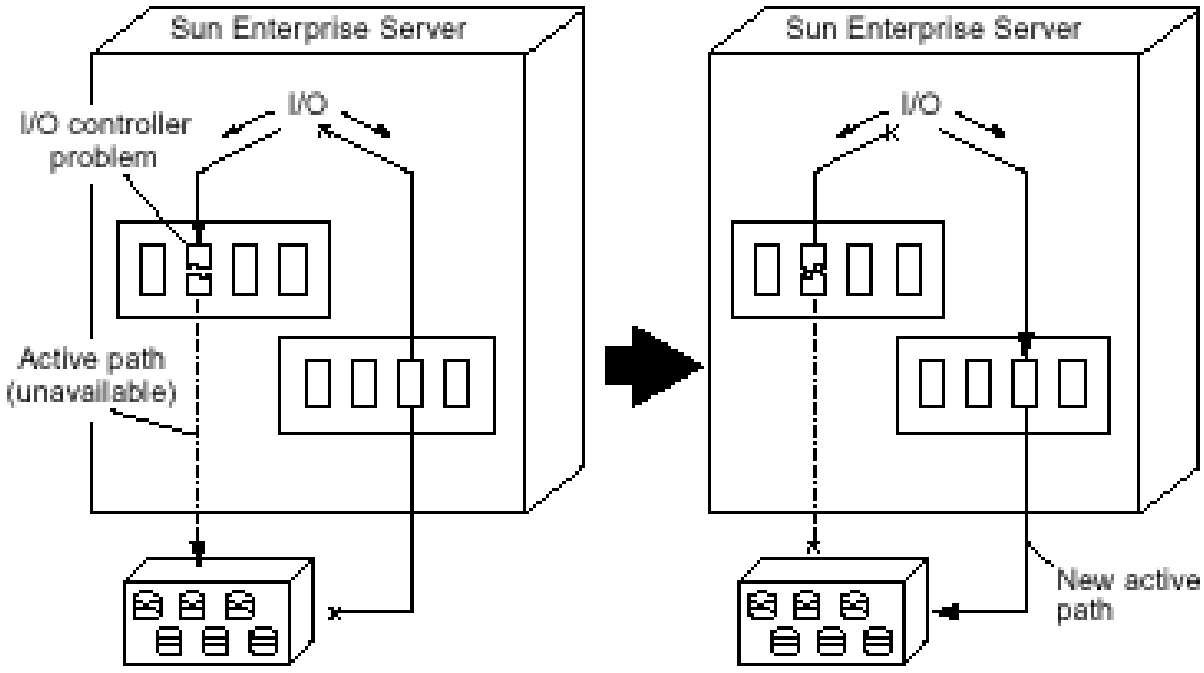
Typische Unix E/A Ansteuerung

Zweiter SCSI BUS in der Regel nur für für Failover, nicht als dynamischer alternativer Übertragungsweg.

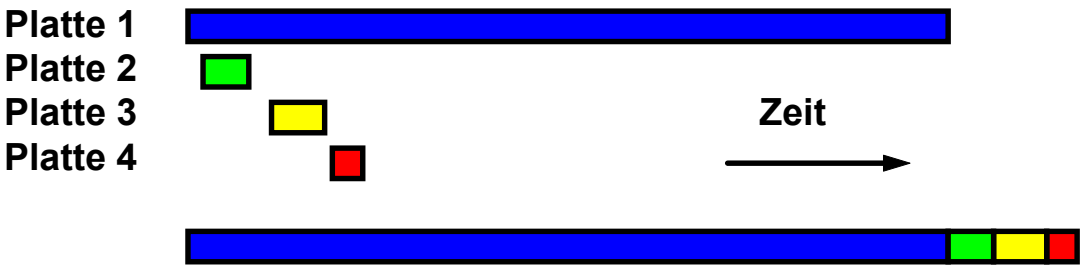
Gute Datenrate, aber: Auf dem Übertragungsweg Mischung von großen und kleinen Datenblöcken. Große Datenblöcke behindern Übertragung von kleinen Datenblöcken.



Ein Plattenspeicher String wird über zwei getrennte elektrische Verbindungen (alternate paths) und zwei SCSI Adapter (I/O Controller) mit dem System verbunden.



Im Fehlerfall wird die Verbindung umgeschaltet.



Unix SCSI Datentransfer

Hohe SCSI Datenraten werden bei E/A Operationen erreicht, die große Blöcke zwischen Platte und Hauptspeicher transportieren. Gleichzeitige E/As für unterschiedliche Platten, die an den gleichen SCSI Kanal angeschlossen sind, verursachen Kanal- Contention, welche die E/A Leistung drastisch verringern kann.

Problematisch bei großen Systemen, auf denen zahlreiche Anwendungen mit unterschiedlichen E/A Anforderungen laufen.

Die Übertragung großer Datenblöcke blockiert die Übertragung zahlreicher kleinen Datenblöcke von anderen Plattenspeichern des gleichen Strings. SCSI E/A basierte Datenbank Systeme benötigen große E/A Puffer Pools um das Contention Problem zu lösen.

Wegen der Schwierigkeit, gemischte E/A Belastungen zu bewältigen, wird in großen Unix Installationen häufig 1 physikalischer Rechner/Anwendung dediziert.

Dies ist ein Grund, warum große Unix Installationen häufig ein Hardware System pro Anwendung dedizieren.

S/390 Utilization vs. UNIX

- Usable Capacity

Utilization	UNIX peak	S/390 Peak	UNIX Average	S390 Average	Usable Capacity Multiplier
CPU utilization at sub-second response time	50-60%	100%	20-30%	65-75%	3X
Disk capacity utilization			20-30%	60-75%	3X

Steve MacKay, Chief Technical Officer, Sun Microsystems, Investors Daily in March, 1999 :

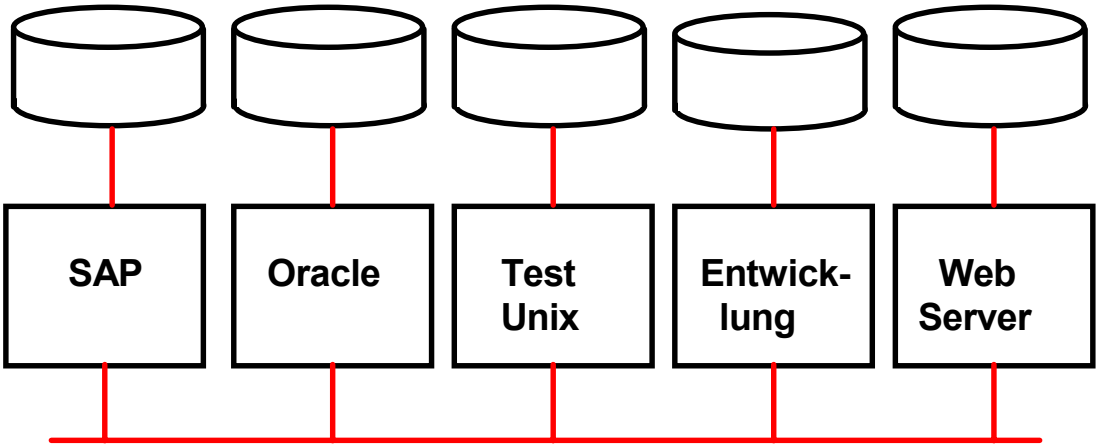
" Peak performance is an area that Sun is working on, and one of the advantages of mainframe over UNIX is that mainframes are capable of running at 85-95% of capacity. UNIX servers run at 20%-30% of peak load".

Steve MacKay is responsible for architecture & technology for the computer systems business at Sun Microsystems.

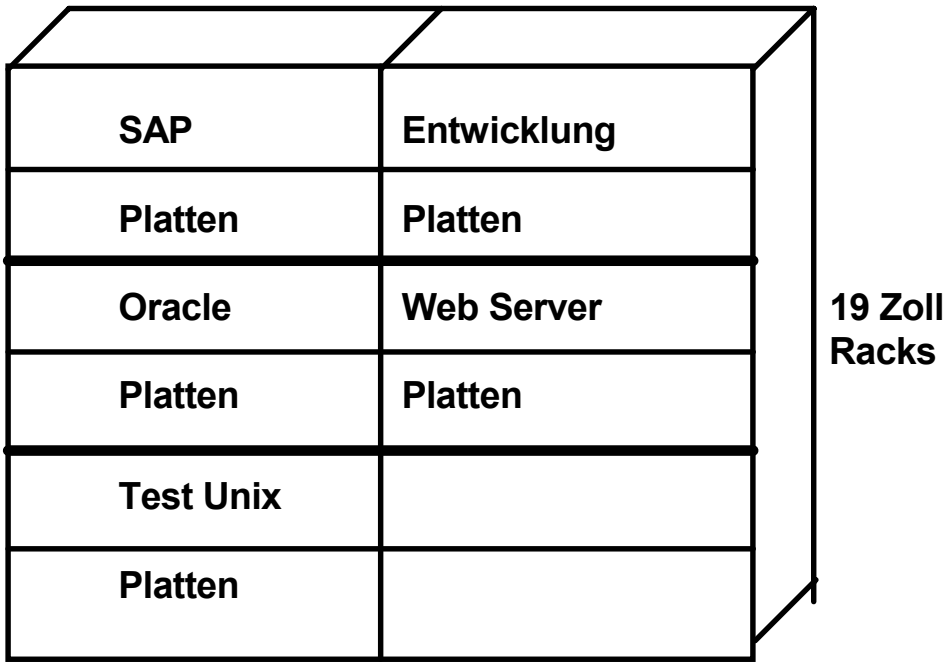
Comparing Typical Disk Requirements
(700 GB database)

	Capacity required by UNIX at 25% Utilization	Capacity Required by S/390 at 70% Utilization	Usable Capacity Multiplier
1 DB Instance	2.8 TB	1 TB	2.8 X
2 DB Instances	5.6 TB	2 TB	2.8 X

Fehler! Textmarke nicht definiert.

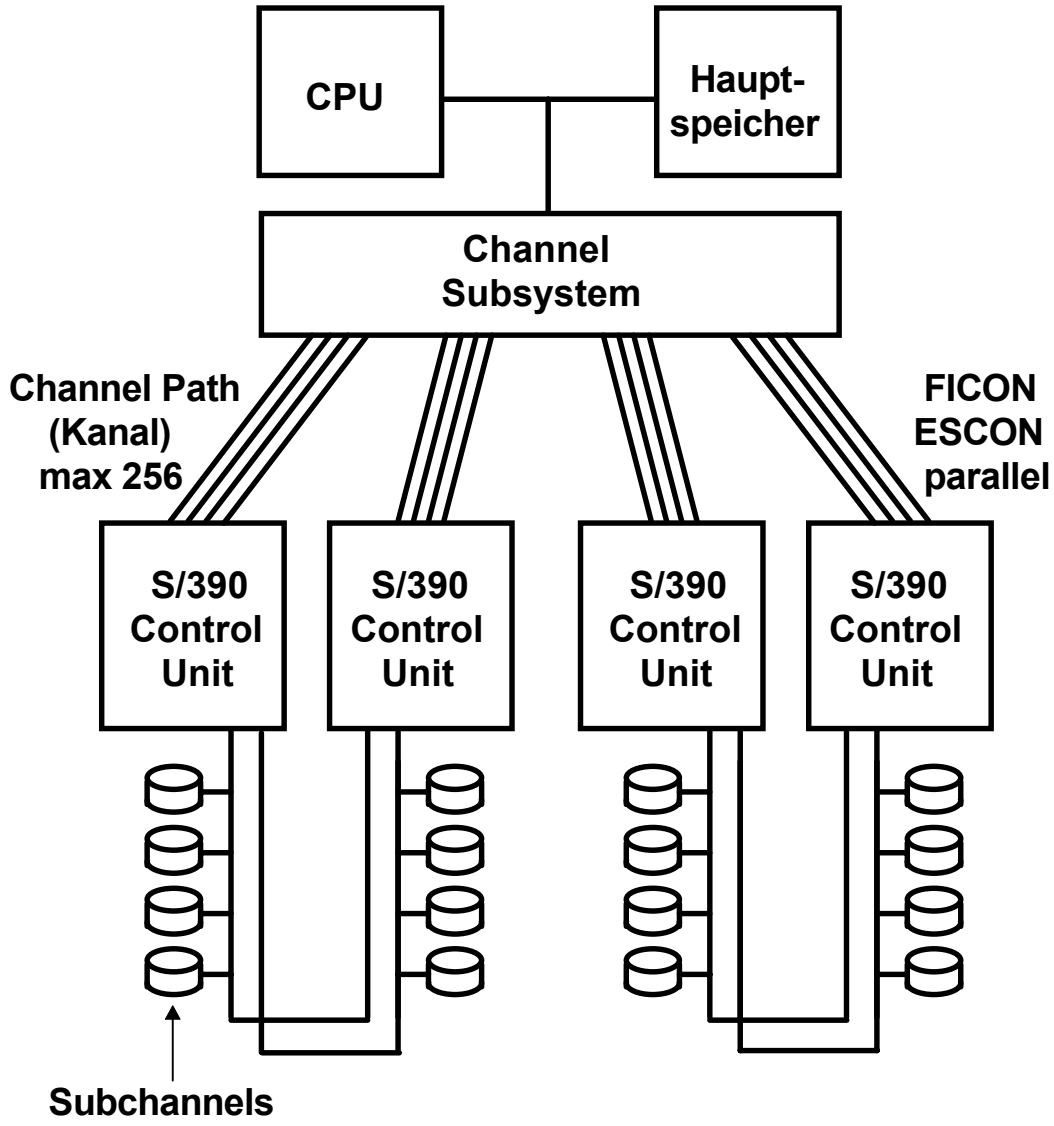


Problem: Administration, Wartung



Multiple Unix Konfiguration

Altenative: LPAR



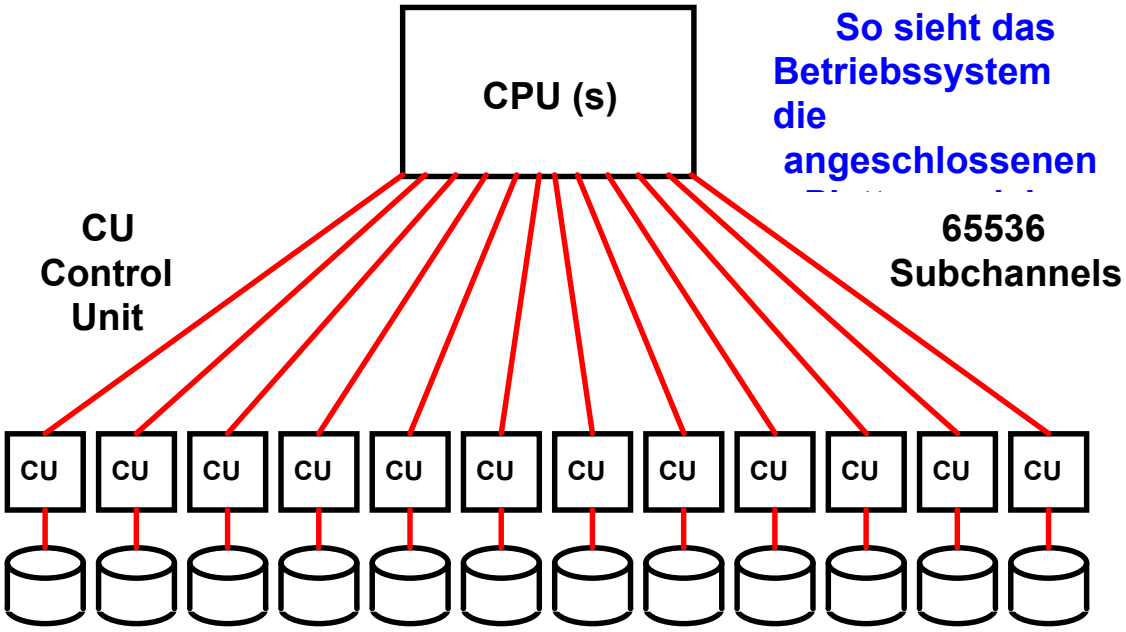
zSeries und S/390 Plattenspeicher Anschluß

Das Channel Subsystem wird durch mehrere Prozessoren (als System Assist Prozessoren, SAP, bezeichnet) und entsprechenden Code verwirklicht. Die SAPs greifen parallel zu den CPUs auf den Hauptspeicher zu und entlasten diese von Ein-/Ausgabe Aufgaben.

Vereinfachte E/A Konfiguration aus Sicht des z/OS Betriebssystems

Plattenspeicher werden bei allen Großrechnern über eine komplexe Konfiguration von (SCSI oder Ficon) Kanälen und Steuereinheiten mit der (den) CPU(s) verbunden.

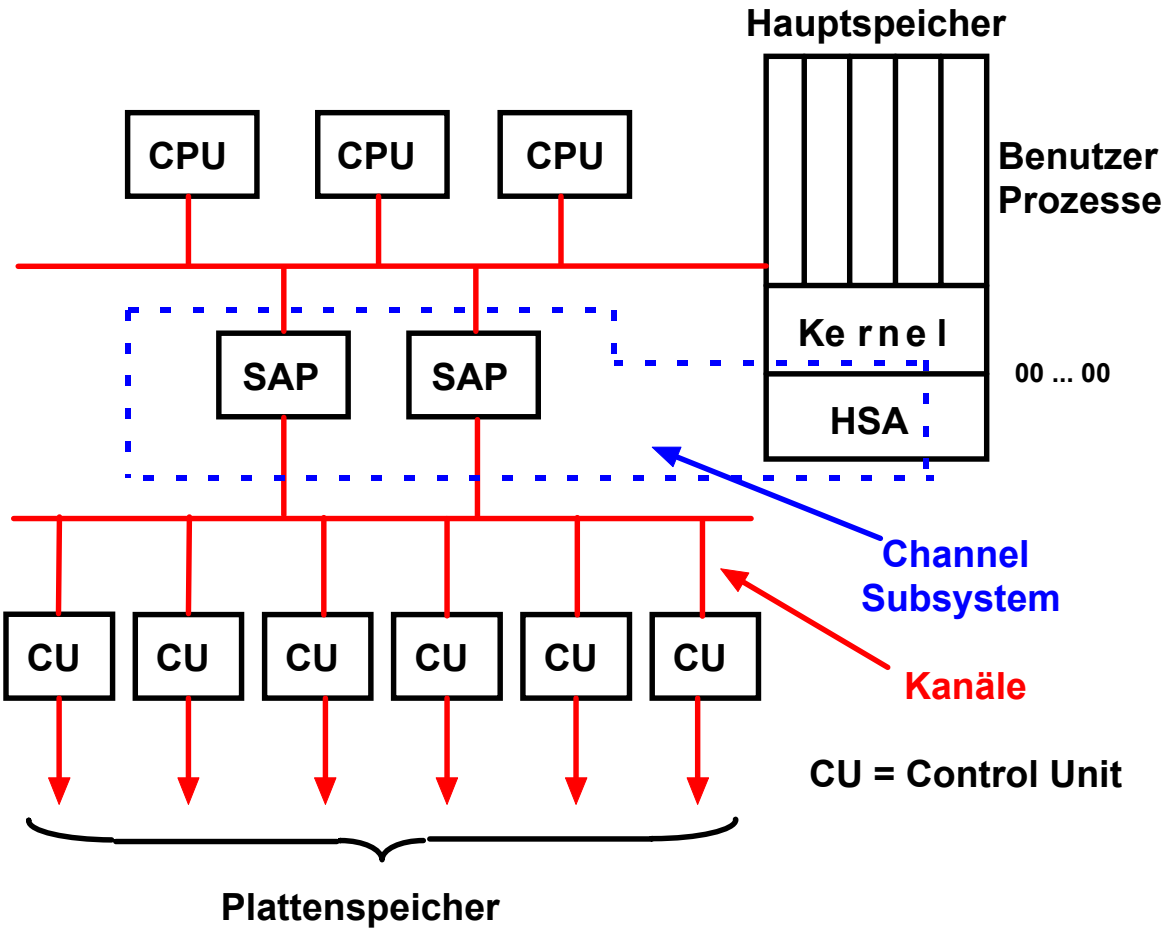
zSeries und S/390 Rechner arbeiten mit einer vereinfachten und standardisierten Sicht der angeschlossenen E/A Struktur (virtuelles E/A Subsystem). Die E/A Ansteuerung des Betriebssystem Kernels kennt die Einzelheiten der E/A Konfiguration nicht.



Jeder Plattenspeicher wird über eine 16 Bit (0 .. 65 535) Subschannel ID angesprochen

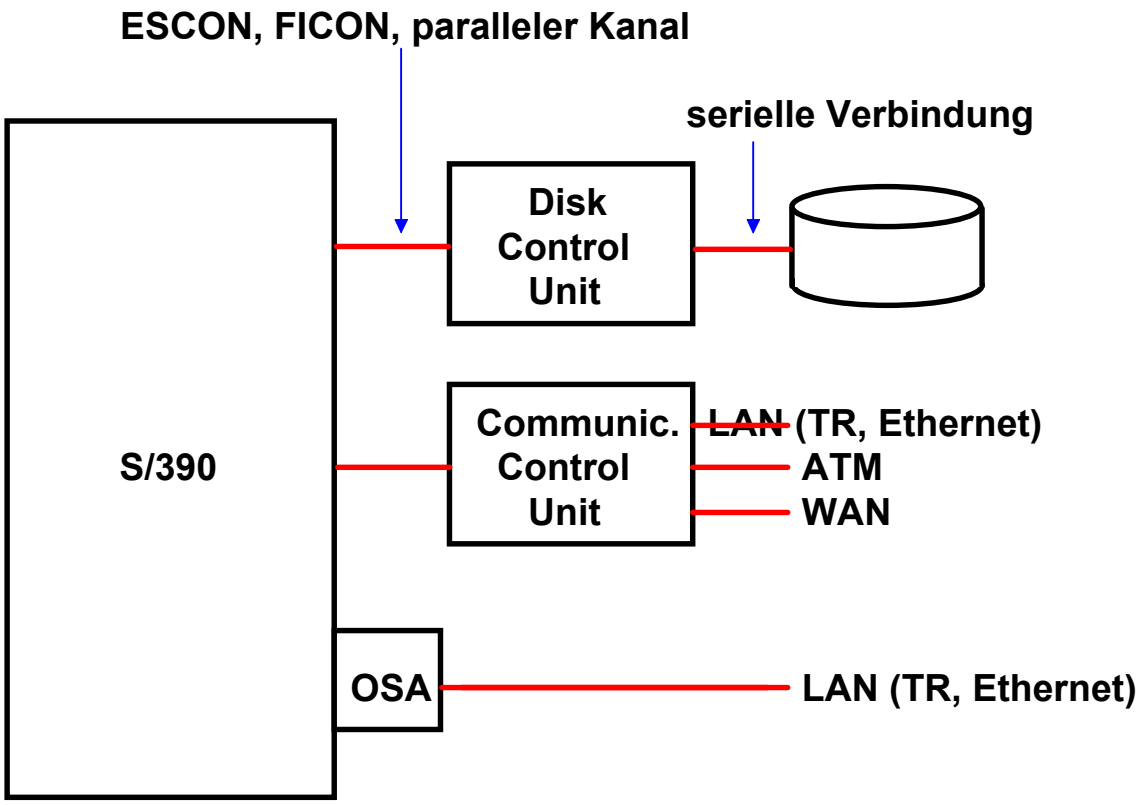
Ein *Channel Subsysten* optimiert die Plattenspeicher Ansteuerung,

zSeries Ein/Ausgabe Anschluss



Die *HSA* (Hardware System Area) ist ein Teil des Hauptspeichers. Sie liegt außerhalb des Adressenraums, auf den die CPUs zugreifen können. Das *Channel Subsystem* besteht aus SAP Prozessoren und Code in der HSA. Es bildet das virtuelle E/A Subsystem, mit dem der Betriebssystem Kernel glaubt zu arbeiten, auf die reale E/A Struktur ab.

Unabhängig von System- und Benutzercode sind damit umfang-reiche Optimierungen der Plattenspeicherzugriffe möglich.



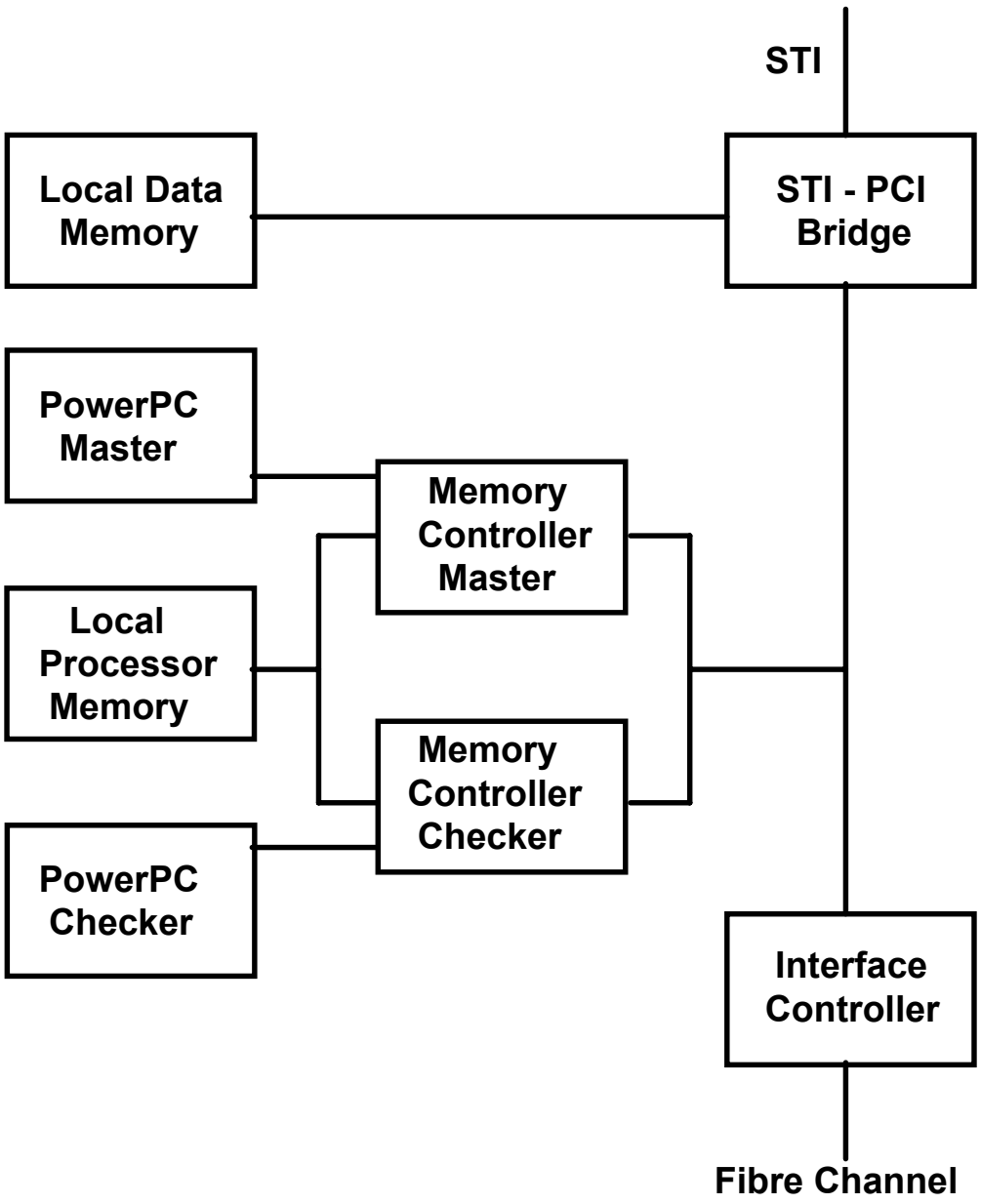
S/390 E/A Konfiguration

E/A Geräte werden grundsätzlich über Steuereinheiten (Control Units) angeschlossen. Steuereinheiten sind meistens in getrennten Boxen untergebracht, und über Glasfaser (ESCON, FICON) an den S/390 Rechner angeschlossen.

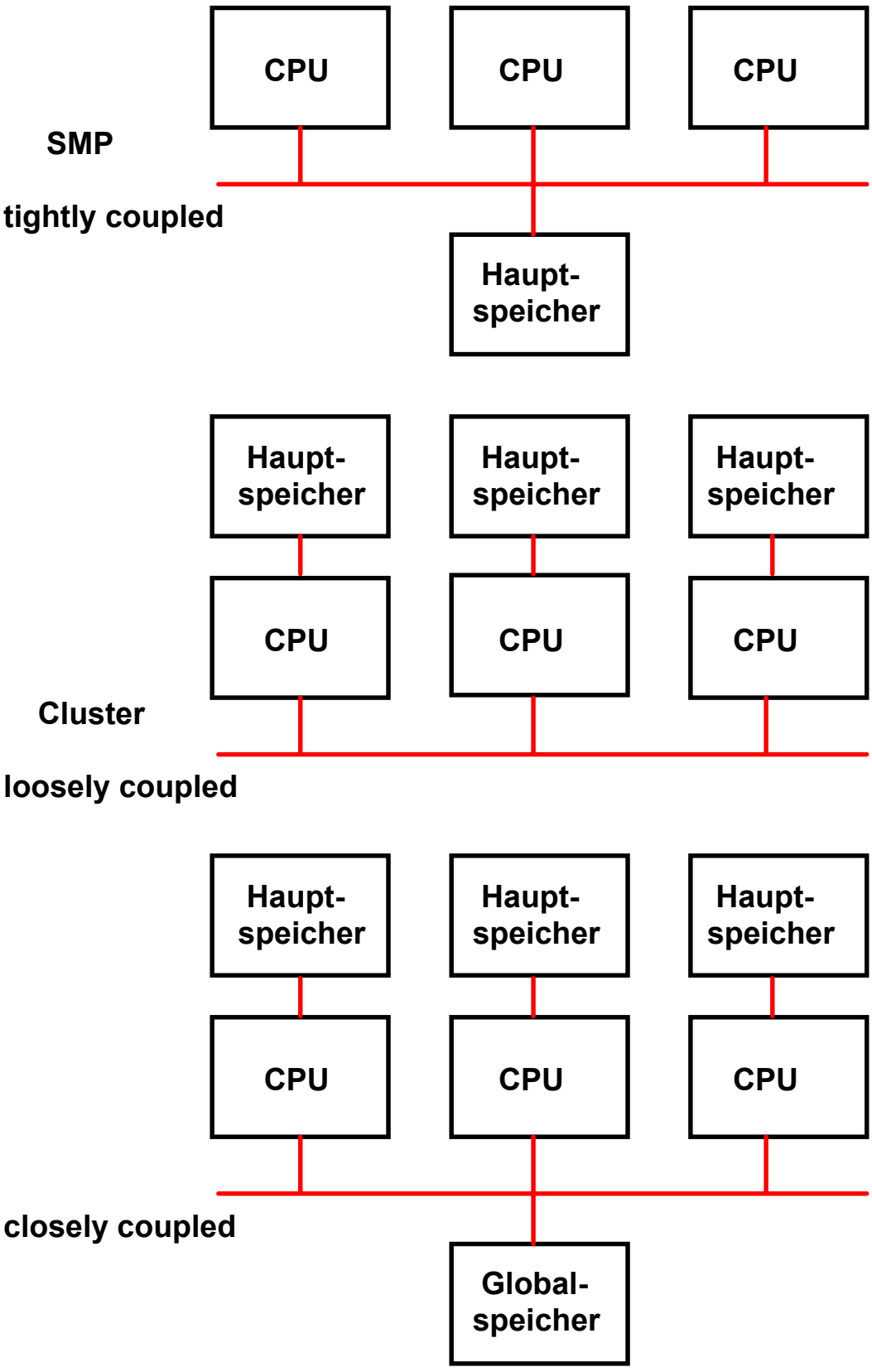
Es existieren viele unterschiedliche Typen von Steuereinheiten. Die wichtigsten schließen externe Speicher (Platten, Magnetbänder Archivspeicher) und Kommunikationsleitungen an.

Es existieren Steuereinheiten für viele weiteren Gerätetypen. Beispiele sind Belegleser für Schecks oder Druckstraßen für die Erstellung von Rentenbescheiden.

Einige Steuereinheiten können in den S/390 Rechner integriert werden. Das wichtigste Beispiel ist der OSA Adapter für den Anschluß von LAN´s.

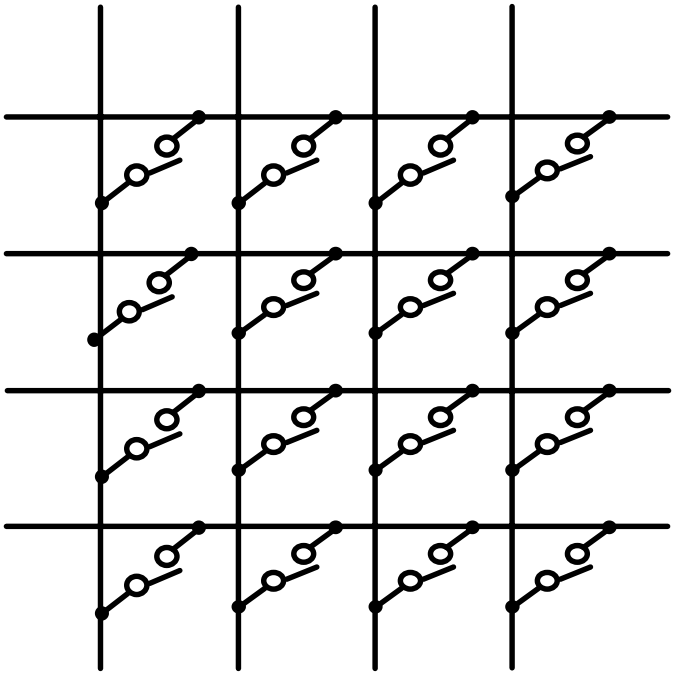


zSeries Fibre Channel Kanal basierend auf der Common I/O Card



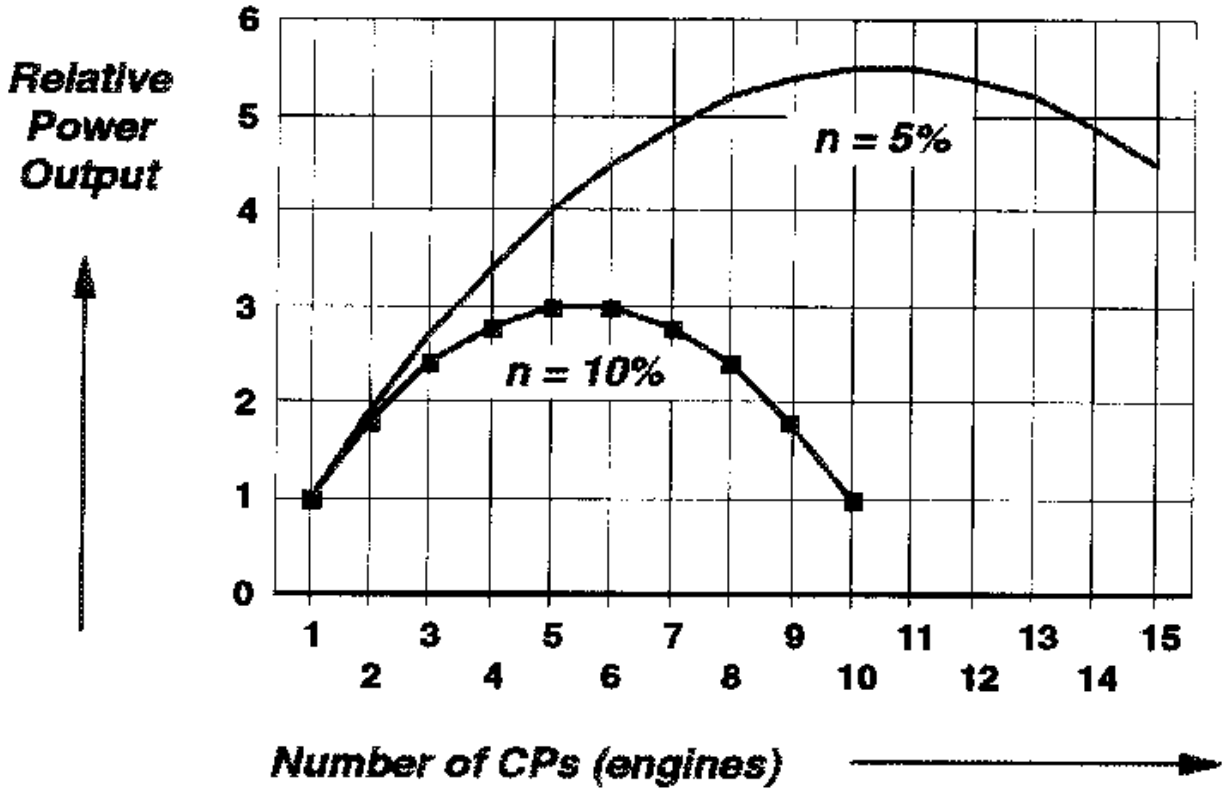
Taxonomie von MIMD Parallelrechnern

vier
Eingänge



vier Ausgänge

4 x 4 Crossbar Matrix Switch



Leistungsverhalten eines Symmetric Multiprocessors (SMP)

Angenommen, ein Zweifach Prozessor leistet das Zweifache minus n % eines Einfach Prozessors. Für n = 10 % ist es kaum sinnvoll, mehr als 4 Prozessoren einzusetzen. Für n = 5 % sind es 8 Prozessoren,

Bei einem z900 Rechner ist n = 2 %. Es ist sinnvoll, einen SMP mit 16 Prozessoren einzusetzen.

Die Gründe für den Leistungsabfall sind Zugriffskonflikte bei der Hardware und Zugriffskonflikte auf Komponenten des Überwachers. Die Letzteren überwiegen.

(S/390) MIPS
Million Instructions Per Second

Performance Benchmark für S/390 Rechner

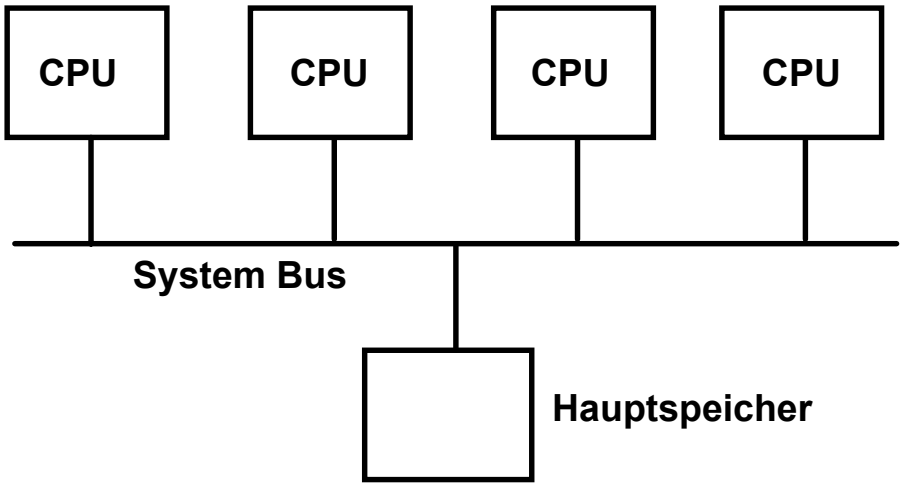
Ausführungszeit für eine Mischung von Maschinenbefehlen

Reine CPU Leistung, keine Ein/Ausgabe

Proprietärer IBM Standard
- **verfügbar seit 1965, ständig erweitert und angepaßt**
- **an realistischen Anwendungsprofilen orientiert**
- **anwendbar für Rechner unterschiedlicher Hersteller**

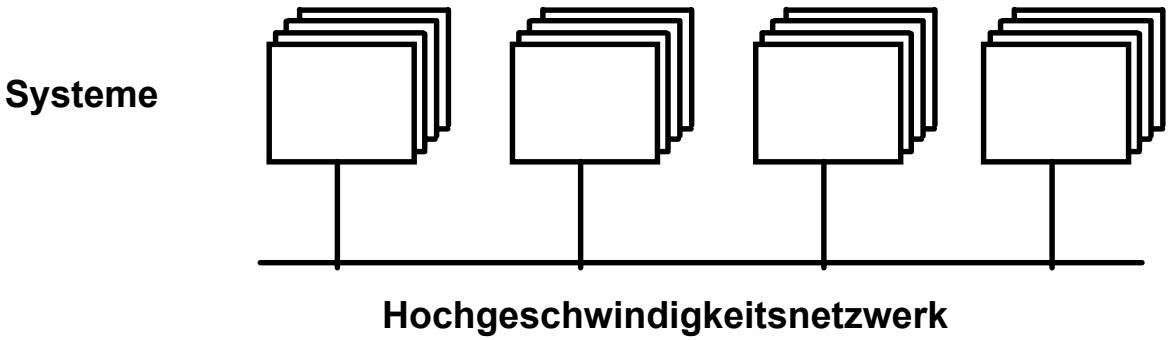
Berücksichtigt
- **Häufigkeit der einzelnen Maschinenbefehle**
- **Cache- und Hauptspeicherzugriffszeiten**
- **Bus Latency/Contention**
- **Cache Misses und Cache-Line reload**
- **Memory Refresh**

Cluster

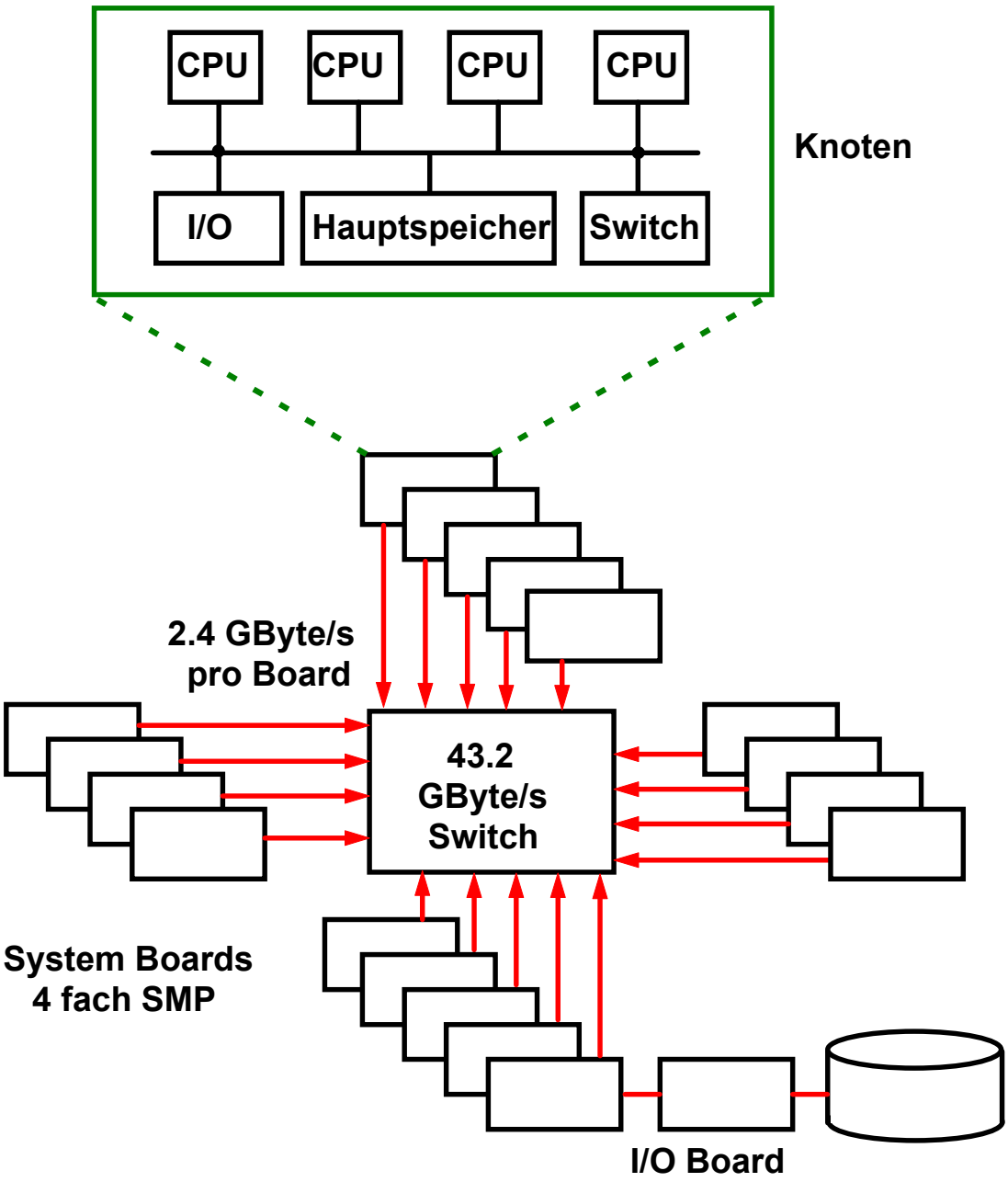


Ein System (Prozessor, Node) besteht aus mehreren CPU´s, die auf einen gemeinsamen Hauptspeicher zugreifen (SMP., Symmetric Multiprocessor).

Im Basisfall nur eine Kopie (Instanz) des Betriebssystems im gemeinsam genutzten Hauptspeicher



Bei einem Cluster werden mehrere Systeme (von denen jedes aus mehreren CPU´s besteht), über ein Hochgeschwindigkeitsnetzwerk miteinander verbunden. Dieses Netzwerk kann ein leistungsfähiger Bus sein, wird aber häufig als Crossbarswitch implementiert.

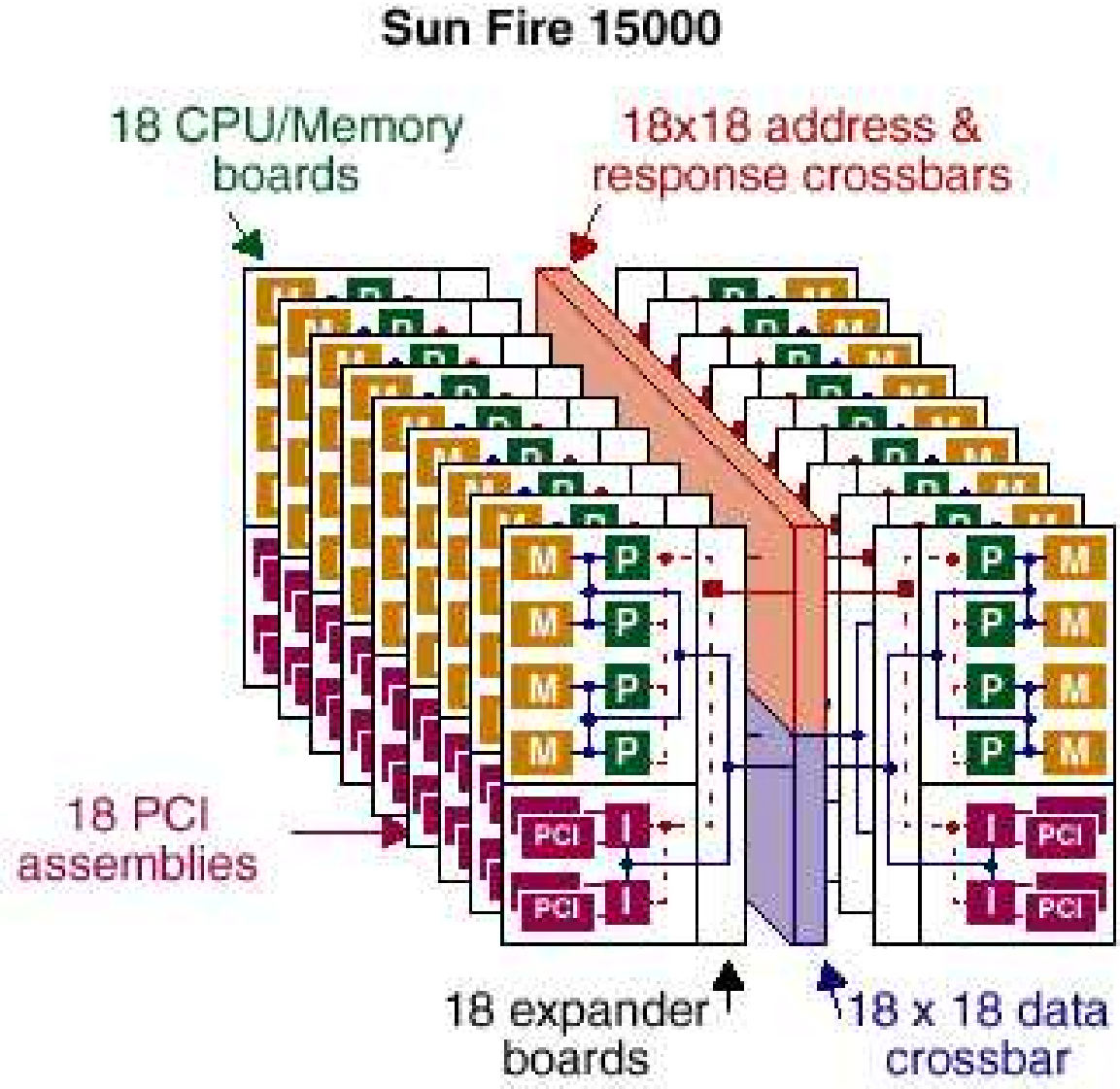


Sun E15K

72 CPU´s

18 System Boards, je 4 CPU/System Board

I/O Controller auf jedem System Board

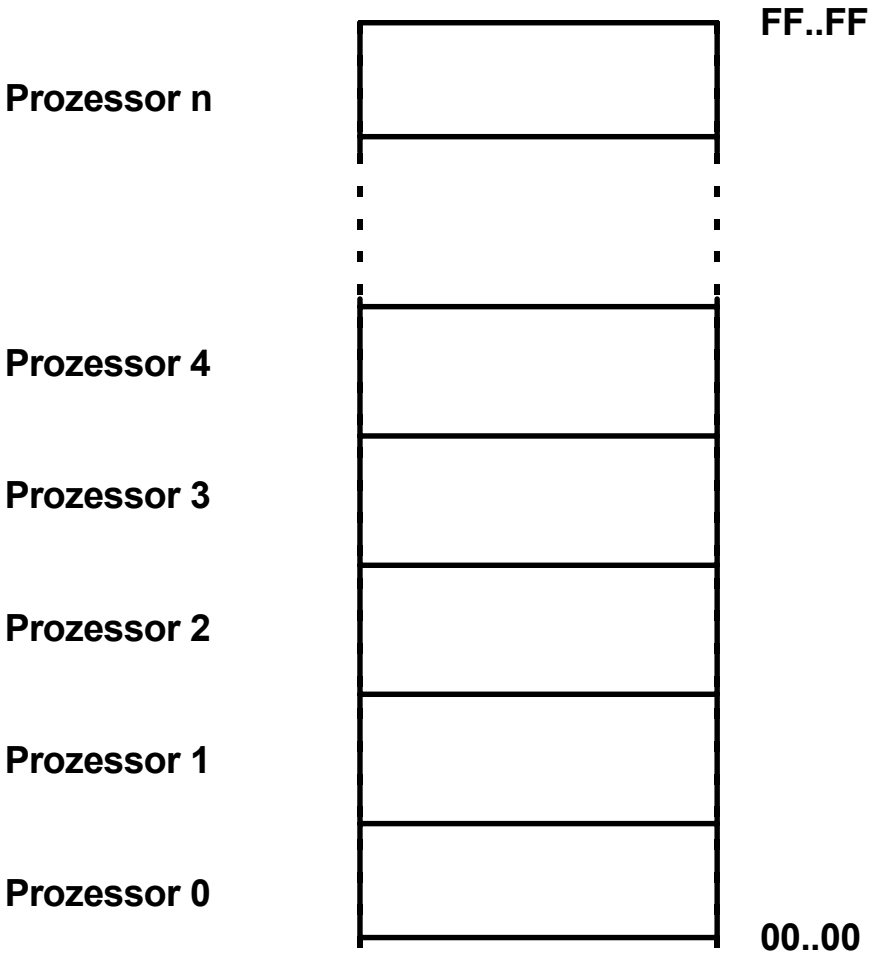


Sun Fire 15000 System hat max 576 Gbyte Hauptspeicher, max 18 CPU/Memory Boards, max 18 Domains, max 18 I/O Boards, max 72 PCI Slots für 72 PCI Karten.

„Board Set“ besteht aus Slot 0 Board, Slot 1 Board und Expander Board. Letzteres nimmt die beiden anderen Boards auf.

Slot 0 Board ist entweder CPU/Memory Board (System Board, 18 max) oder System Controller Board (1 oder 2 max, nicht gezeigt).

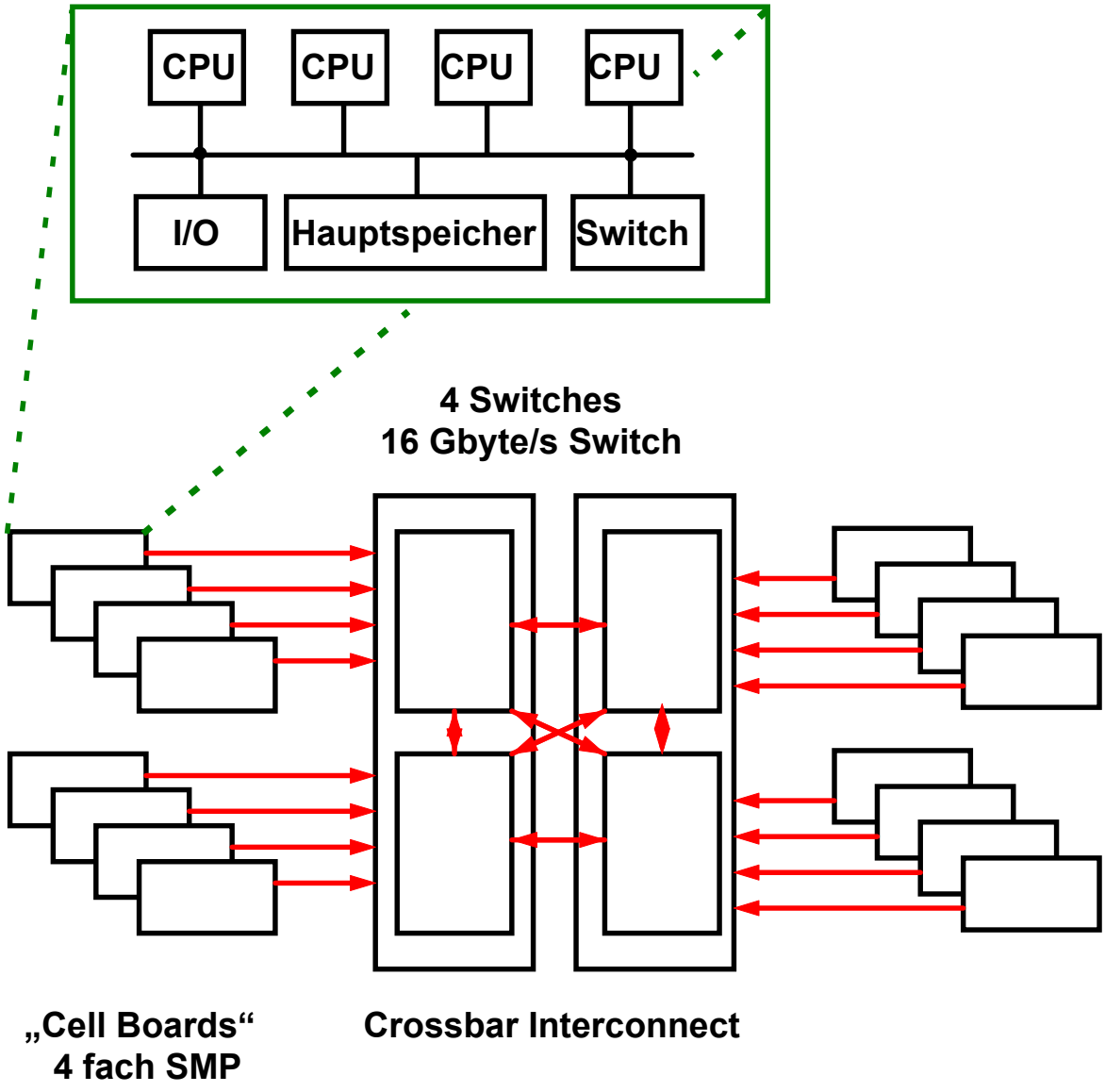
CPU/Memory Board hat 4 Sparc III , 1,2 Ghz CPUs, 8 DIMMS/CPU, 8GByte/CPU, 32GByte total. Hauptspeicher Zugriffszeit 180 ns für Hauptspeicher auf gleichem Board, 333 - 440 ns für Hauptspeicher auf anderem Board.



Non-uniform Memory Architecture
NUMA

Die Knoten eines Clusters haben jeweils einen eigenen lokalen Hauptspeicher.

Alle Hauptspeicher der Knoten bilden einen gemeinsamen realen Adressenraum. Jeder Knoten bildet automatisch einen Ausschnitt dieses Adressenraums auf die absoluten Adressen seines lokalen Hauptspeichers ab.



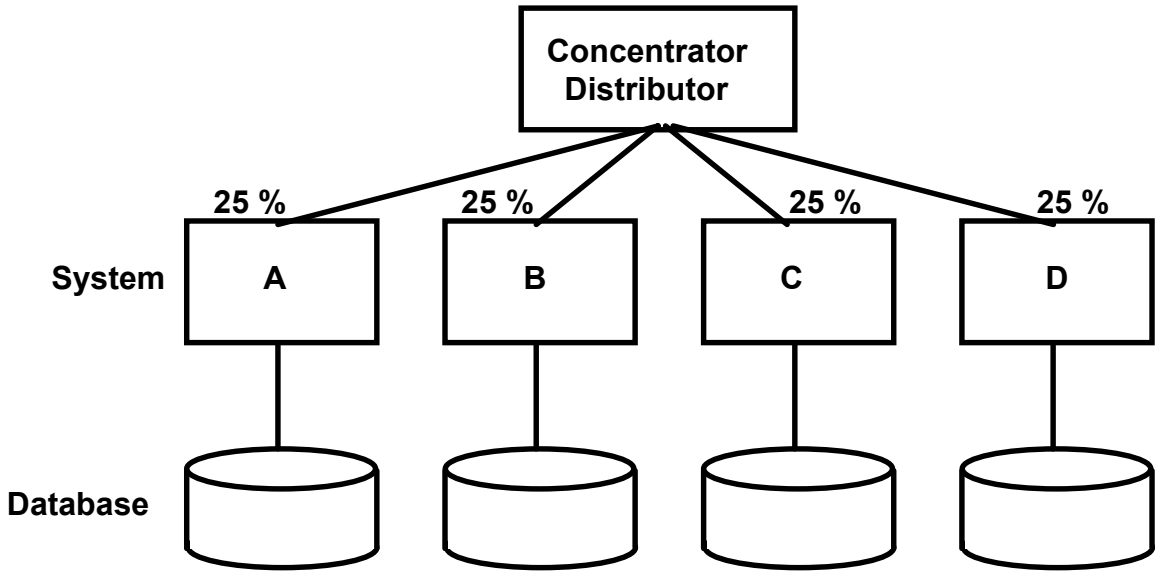
HP Superdome Cluster

64 CPU´s

16 Knoten (Cell Boards), je 4 CPU/Knoten

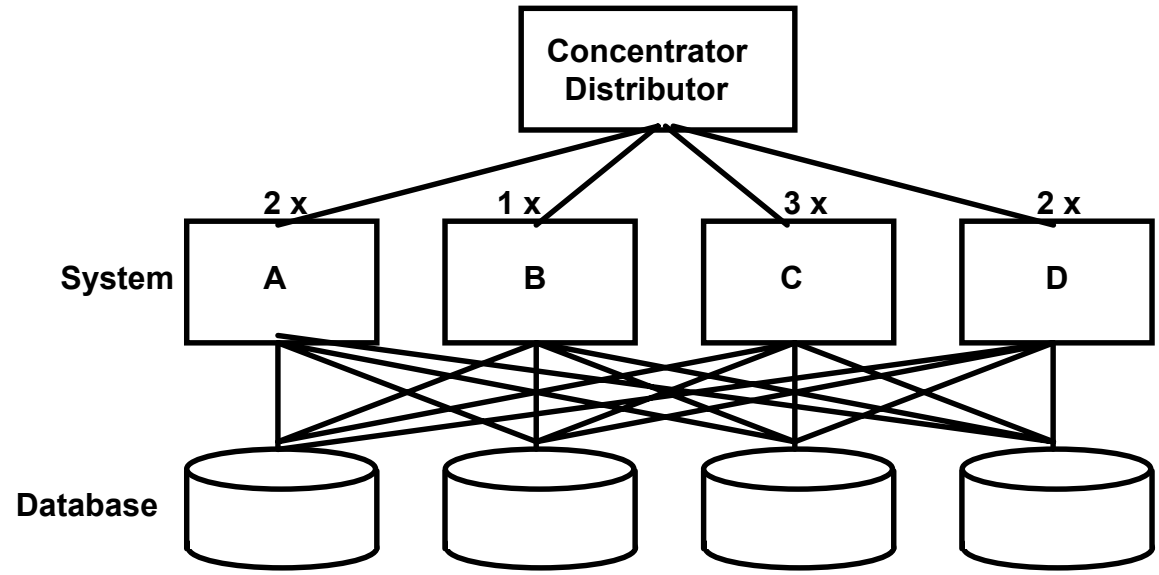
I/O Anschluß auf jedem Cell Board

http://www.serverworldmagazine.com/webpapers/2001/05_hpsuperdome.shtml



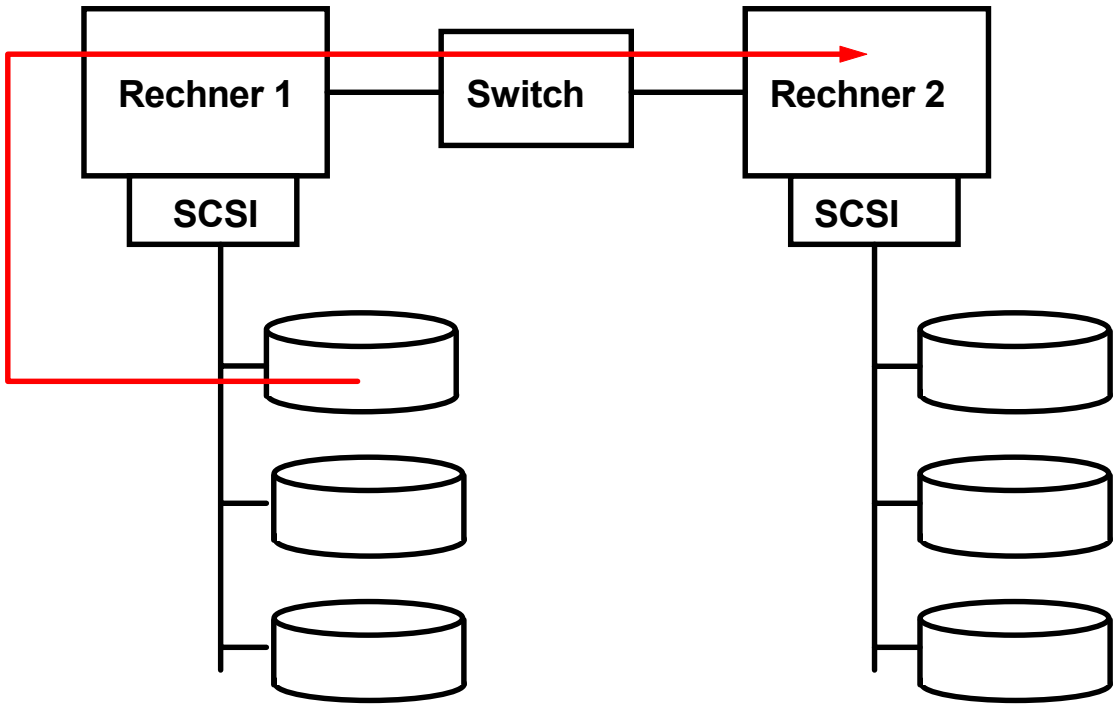
Shared nothing (partitioned data)

Jeder Rechner greift auf seine eigenen Daten zu. Die Arbeitslast wird den einzelnen Rechnern statisch zugeordnet.



Shared data (shared disk)

Jeder Rechner greift auf alle Daten zu. Dynamische Zuordnung der Arbeitslast.

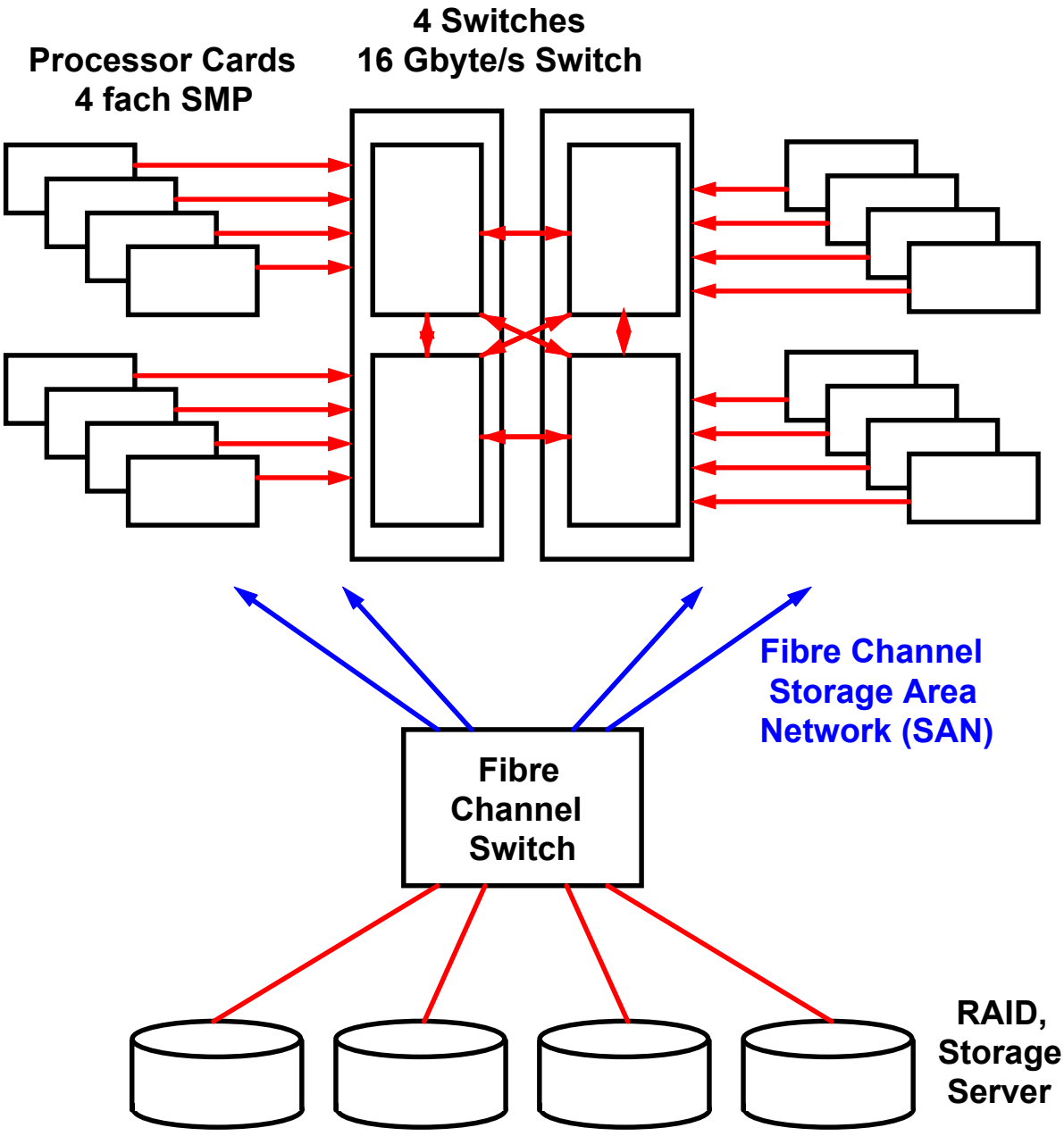


Shared Disk Emulation

Rechner 2 bittet Rechner 1, die gewünschten Daten zu übertragen

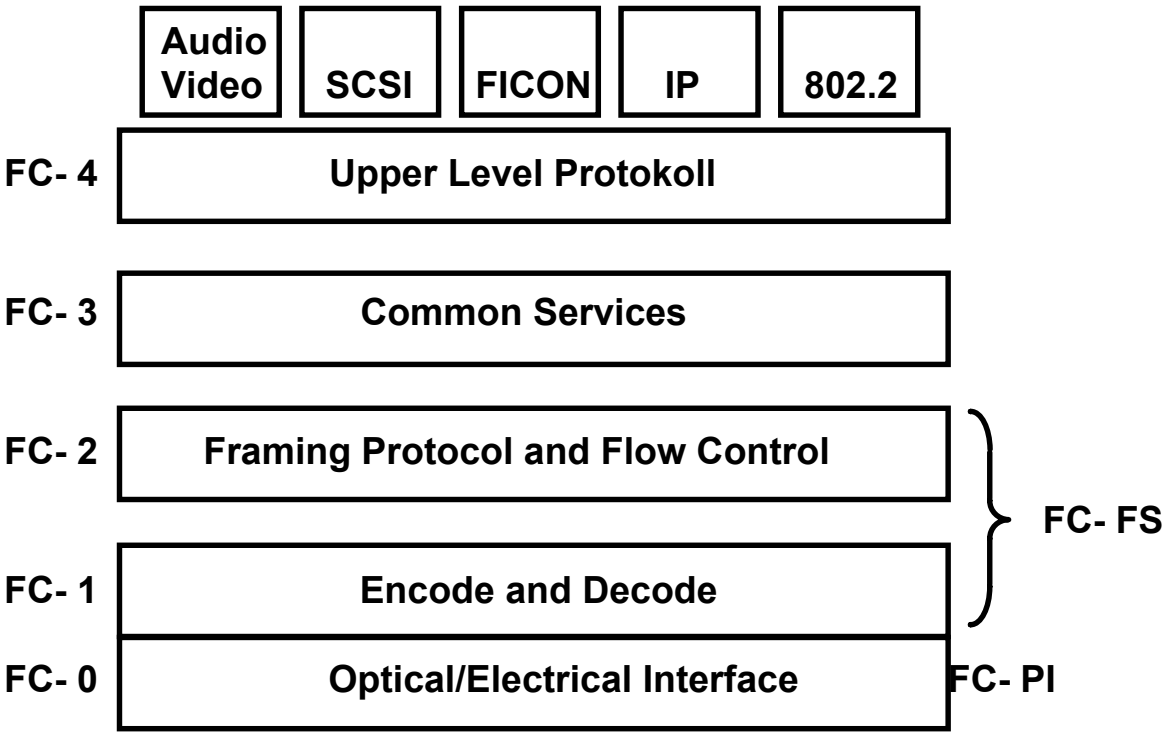
HP Superdome Cluster

64 CPU´s
16 Knoten, je 4 CPU/Knoten
I/O Controller auf jeder Karte

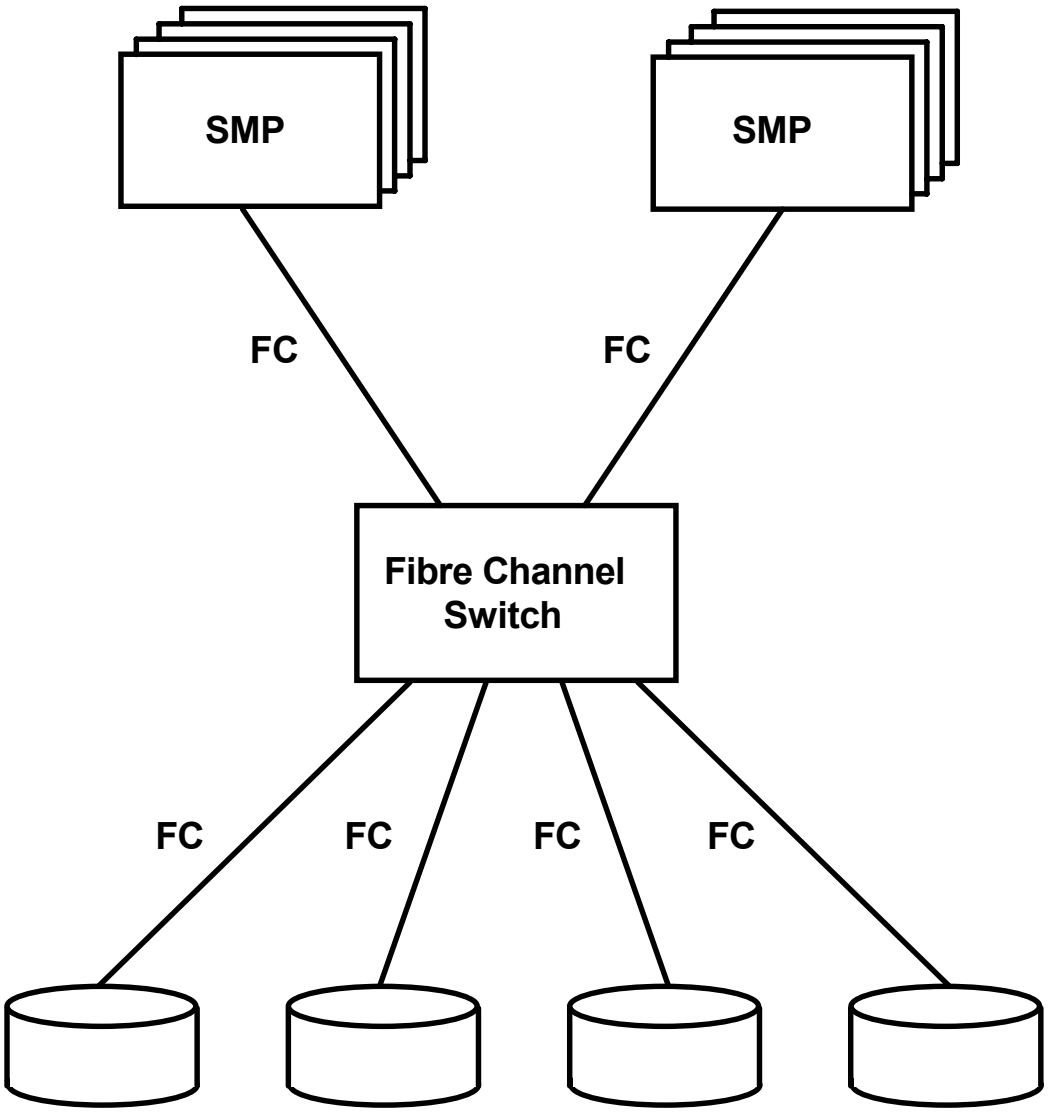


Unterschiedliche Festplattenanschlüsse

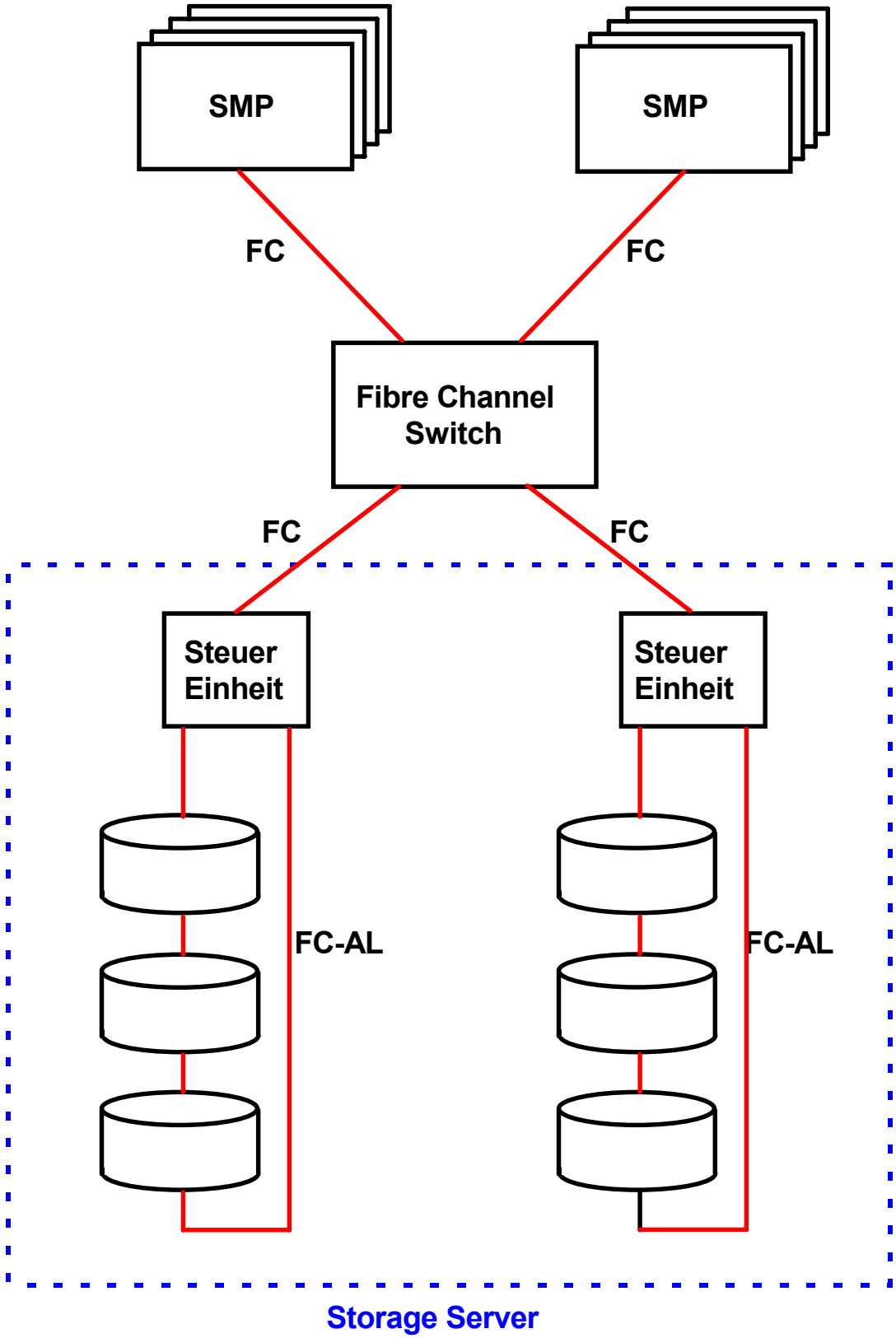
- ATA (IDE)
- SCSI (parallel SCSI)
- Fibre Channel (Serial SCSI)
- SSA (Serial Storage Architecture)



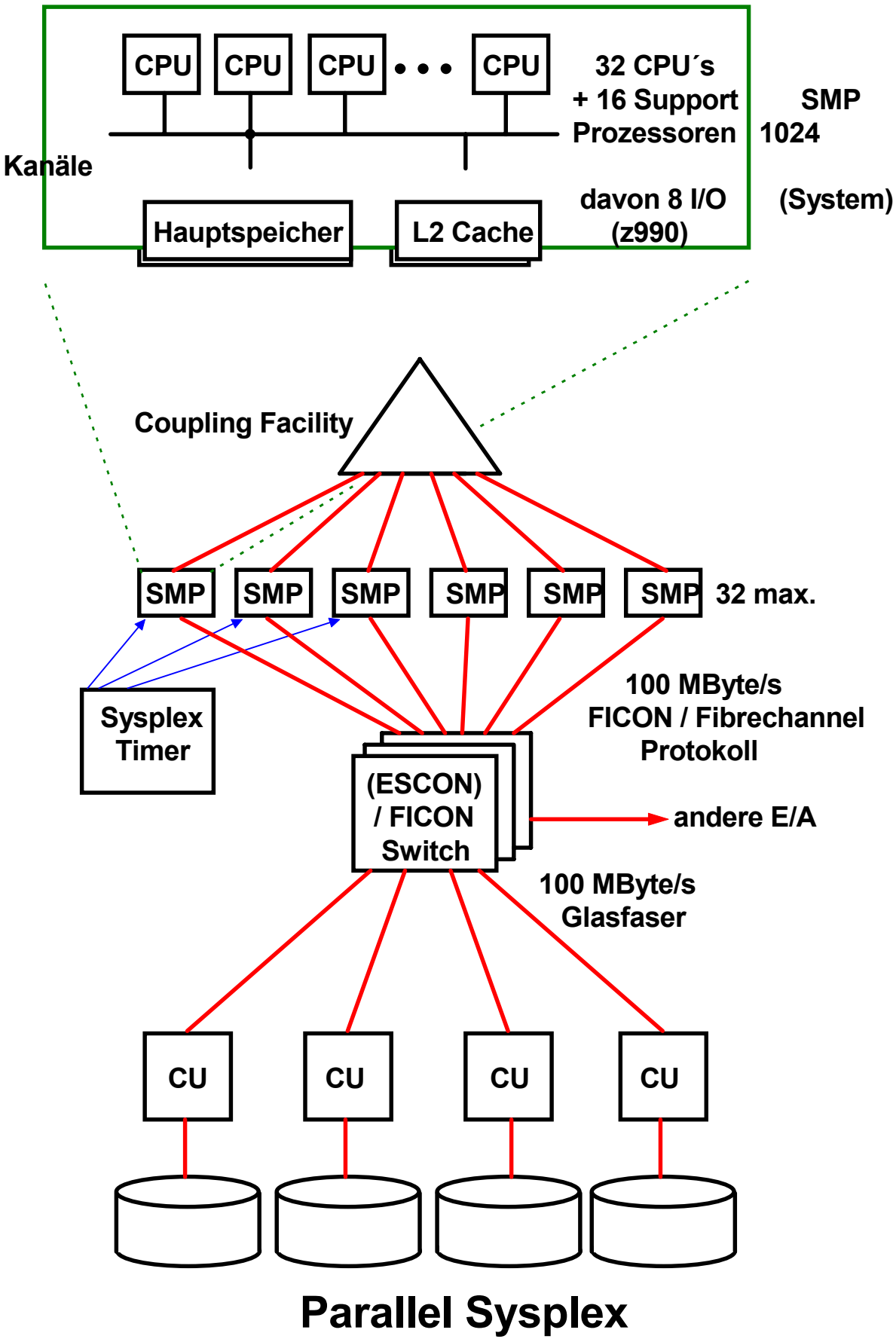
Fibre Channel Standard Architektur



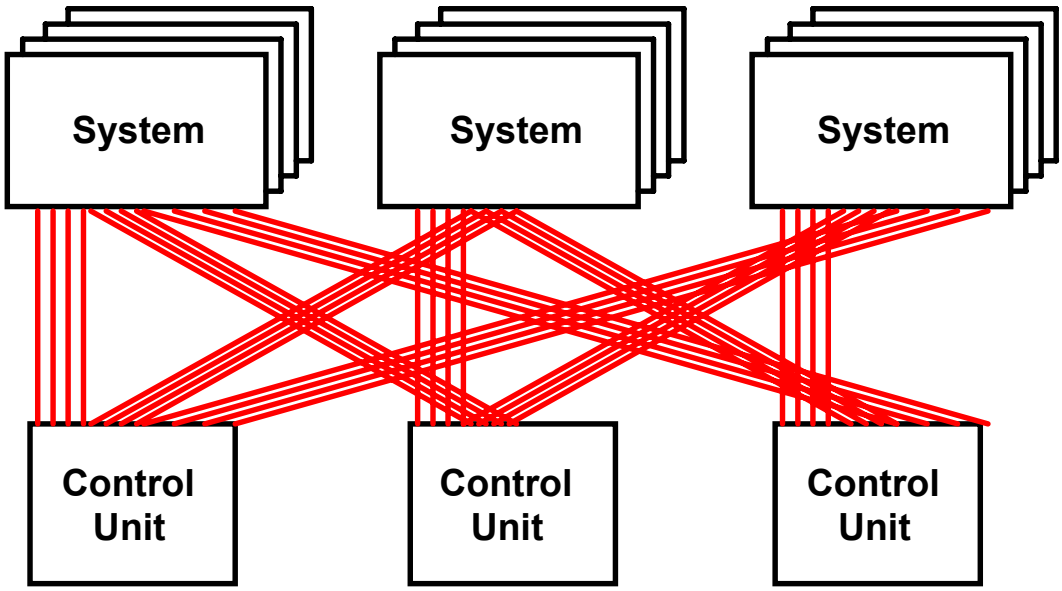
Einfache Fibre Channel Konfiguration



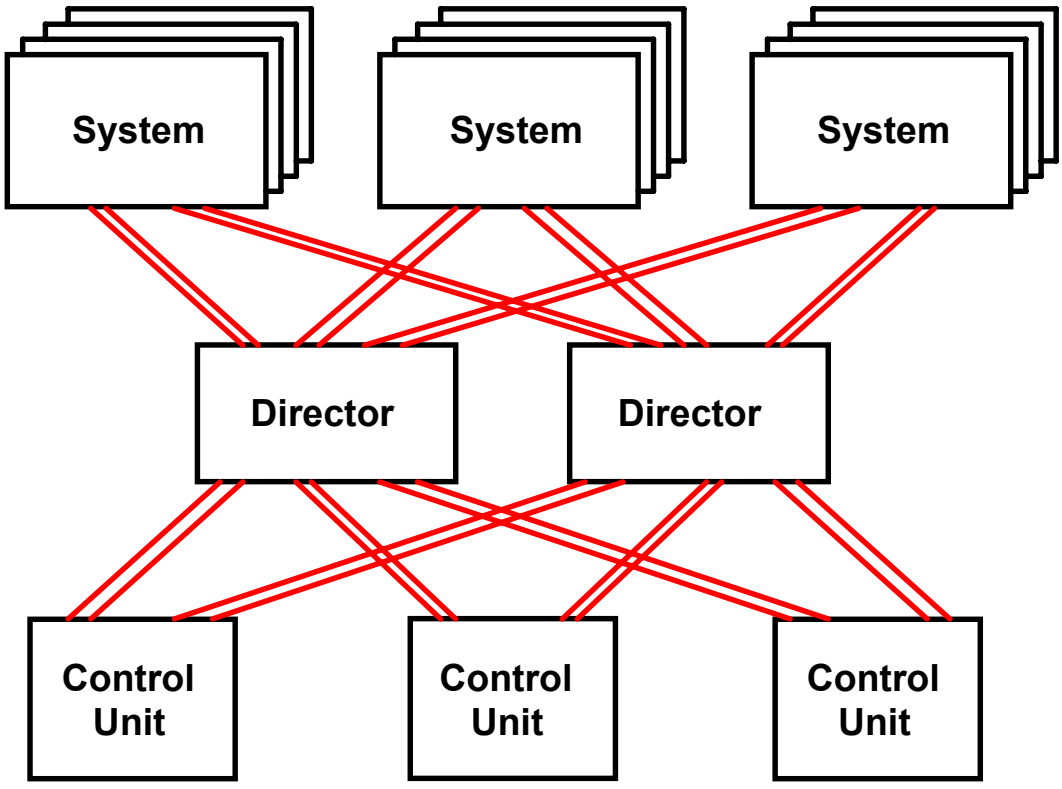
RAID, Cache Funktionalität



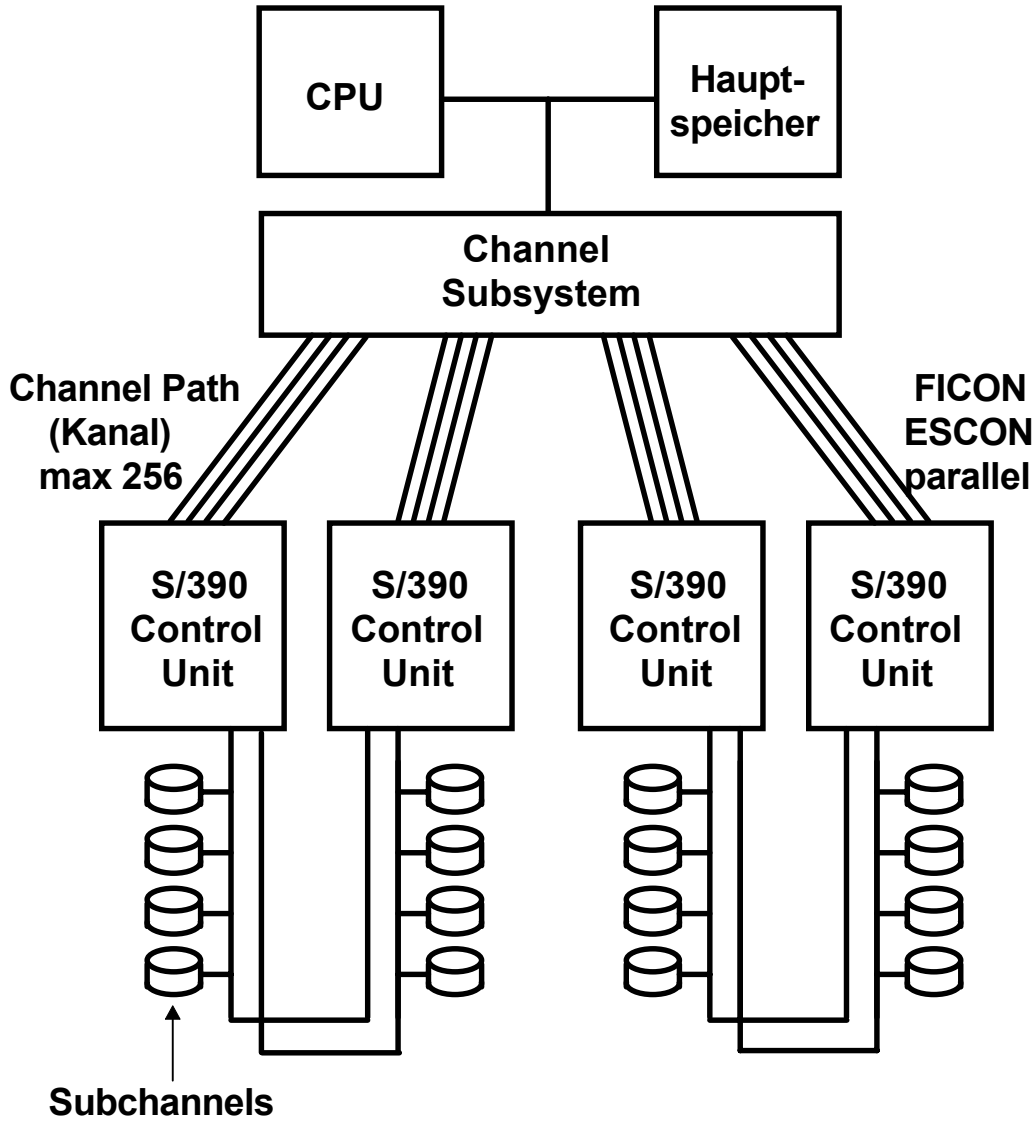
Parallel Sysplex



Parallel Channel Configuration

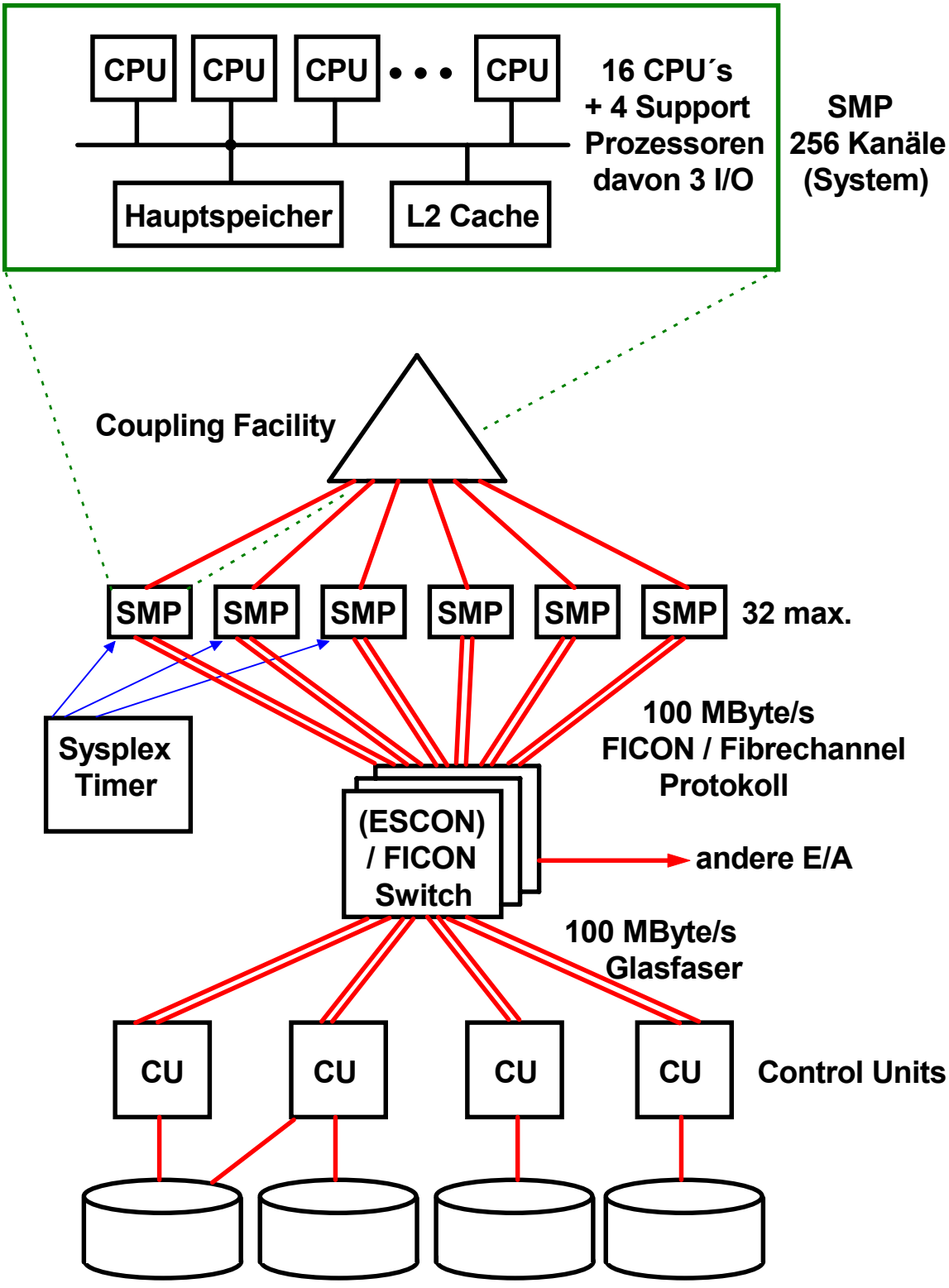


FICON (ESCON) Channel Configuration



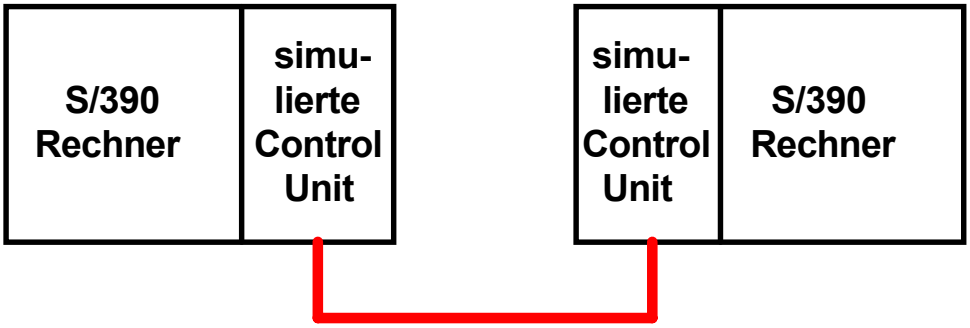
zSeries und S/390 Plattenspeicher Anschluß

Das Channel Subsystem wird durch mehrere Prozessoren (als System Assist Prozessoren, SAP, bezeichnet) und entsprechenden Code verwirklicht. Die SAPs greifen parallel zu den CPUs auf den Hauptspeicher zu und entlasten diese von Ein-/Ausgabe Aufgaben.



Parallel Sysplex

CTC Verbindung
(Channel- to Channel)



Channel- to Channel Verbindung

Cross-System Coupling Facility (XCF)

Die Cross-System Coupling Facility (XCF) verwendet das CTC Protokoll. Sie stellt die Coupling Services bereit, mit denen OS/390 Systeme innerhalb eines Sysplex miteinander kommunizieren.

Sysplex Konfigurationsdaten

Jedes System hat bis zu 256 Kanäle
Jeder ESCON Switch hat bis zu 256 Ports
Bis zu 8 Pfade pro Control Unit

Eine große Installation hat (1999)

- **100- 200 TByte Plattenspeicherplatz installiert (Deutsche Telekom 300 TByte)**
- **15 - 20 ESCON Switche**
- **200 Fiber Optik Anschlüsse pro Switch**
- **8 -10 Systeme**
- **8 - 10 CPU´s / System, 100 CPU´s gesamt**

ESCON Kabel: 17 Mbyte/s, FICON Kabel: 100 MByte/s

Parallel Sysplex Cluster Technology

Mehrfache S/390 Systeme verhalten sich so, als wären sie ein einziges System (Single System Image).

Parallel Sysplex Cluster Technology Komponenten:

- **Prozessoren mit Parallel Sysplex Fähigkeiten**
- **Coupling Facility**
- **Coupling Facility Control Code (CFCC)**
- **Glasfaser Hochgeschwindigkeitsverbindungen**
- **ESCON oder FICON Switch**
- **Sysplex Timer**
- **Gemeinsam genutzte Platten (Shared DASD)**
- **System Software**
- **Subsystem Software**

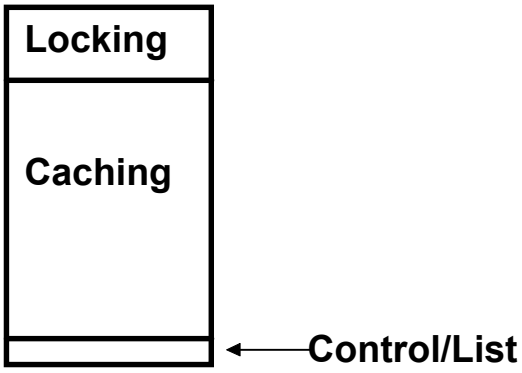
Die Coupling Facility ermöglicht Data Sharing einschließlich Datenintegrität zwischen mehrfachen S/390 Servern

Coupling Facility

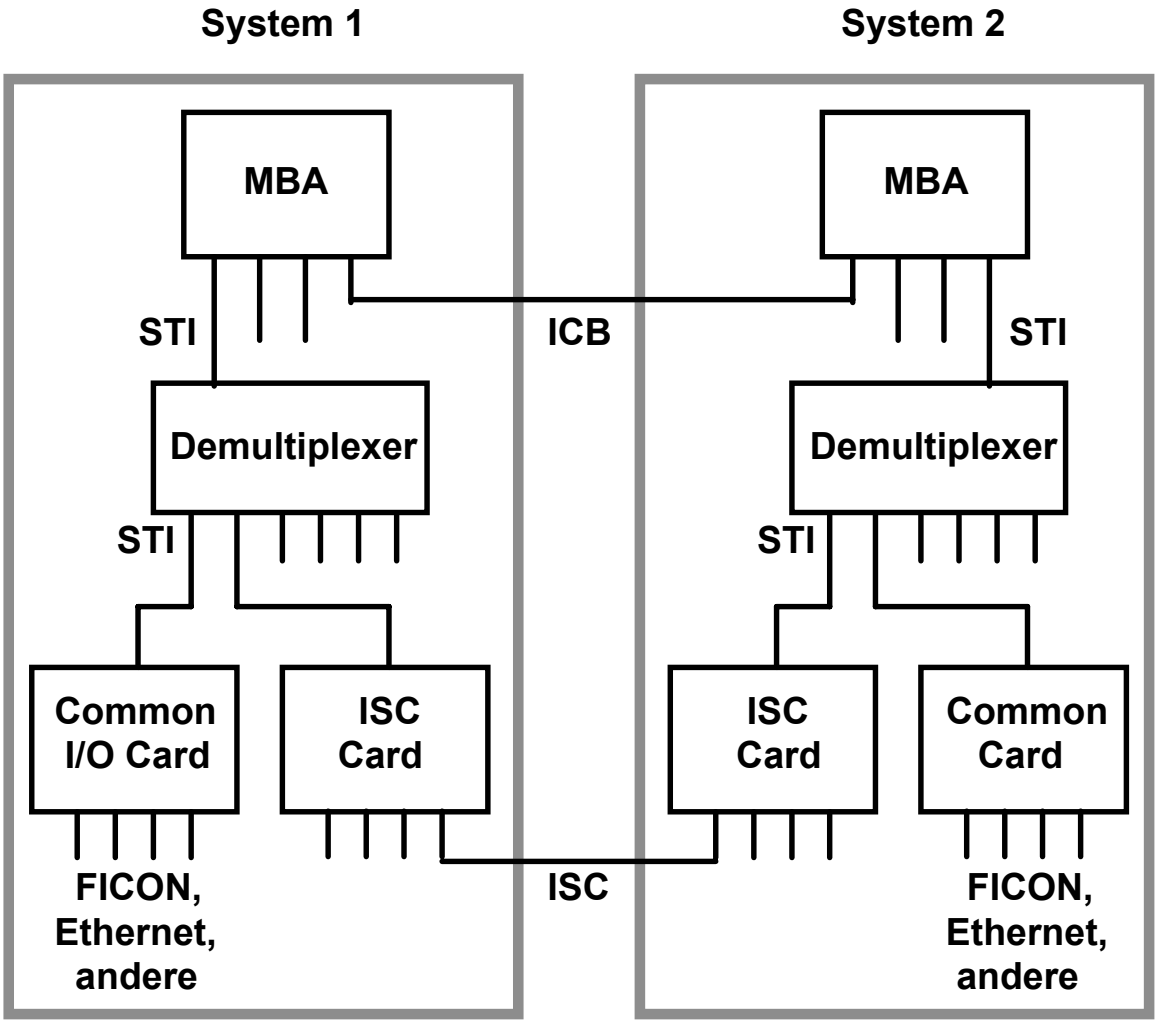
Die Coupling Facility ist in Wirklichkeit ein meiterer S/390 Rechner mit spezieller Software. Ihre Aufgaben sind:

- **Locking**
- **Caching**
- **Control/List Structure Management**

Der größte Teil des Coupling Facility Hauptspeichers wird für das caching von Plattenspeicherdaten eingesetzt.



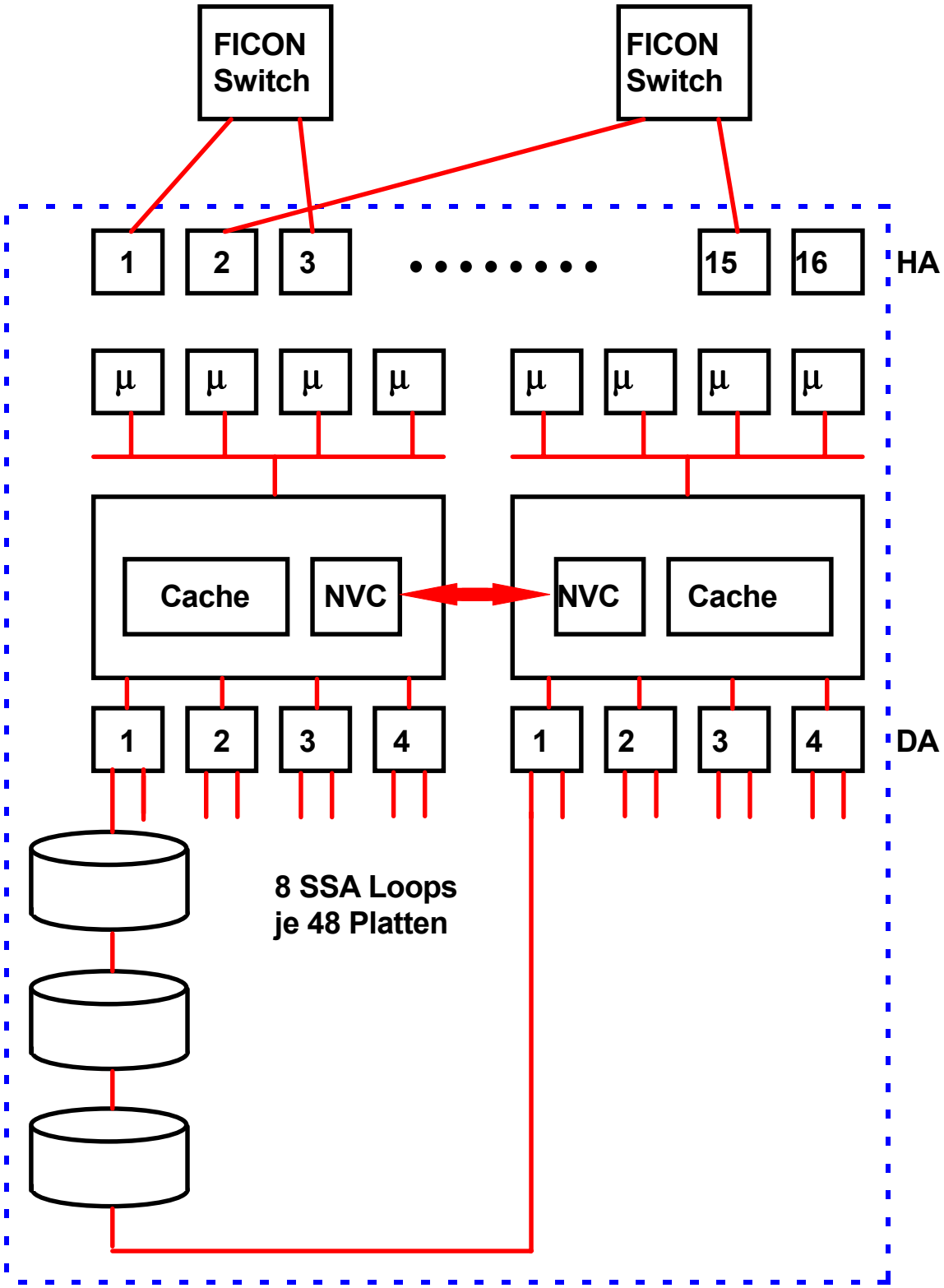
Die Coupling Facility ist über Flasfaser Verbindungen mit einem optimierten Protokoll mit den Rechnern des Sysplex verbunden.



System Area Network

Von jedem I/O Port (MBA Chip) gehen 4 full duplex STI Busse zu 4 Demultiplexoren. Jeder Demultiplexor hat 6 STI full duplexBus Ausgänge. Jeder dieser Ausgänge geht zu einer I/O Card, z.B. einer Common I/O Card oder ISC Card. Jede dieser Karten hat 4 Ausgänge. Von den maximal 96 Ausgängen sind maximal 84 nutzbar für I/O Cards (z.B. FICON, Gigabit Ethernet und andere). Eine spezielle I/O Card ist die ISC Card, die es gestattet, zwei zSeries Server über eine bis zu 20 km lange Glasfaserverbindung zu koppeln.

Alternativ können zwei zSeries Server über den (elektrischen) ICB Bus gekoppelt werden; die maximale Entfernung beträgt hierbei 10 Meter.



Shark Enterprise Storage Server

NVC = Non Volatile Cache (Batterie Back Up)
HA = Host Adapter, DA = Device Adapter

Shark Enterprise Storage Server (ESS)

16 FICON oder FC-SCSI Host Adapter, 1 Link / Adapter

2 Cluster Prozessoren, je 4 x SMP

2 x 8 Gbyte Cache, Teil davon als NVC (non-volatile Cache)

2 x 4 Device Adaptoren, je 320 Mbyte/s, 1 280 Mbyte/s insgesamt

8 SSA Loops, je 160 MByte/s

48 Platten/Loop aufgeteilt in 6 Gruppen zu je 8 Platten

1 RAID Einheit je Gruppe, 6+P+S

48 Platten/Loop, 384 Platten insgesamt

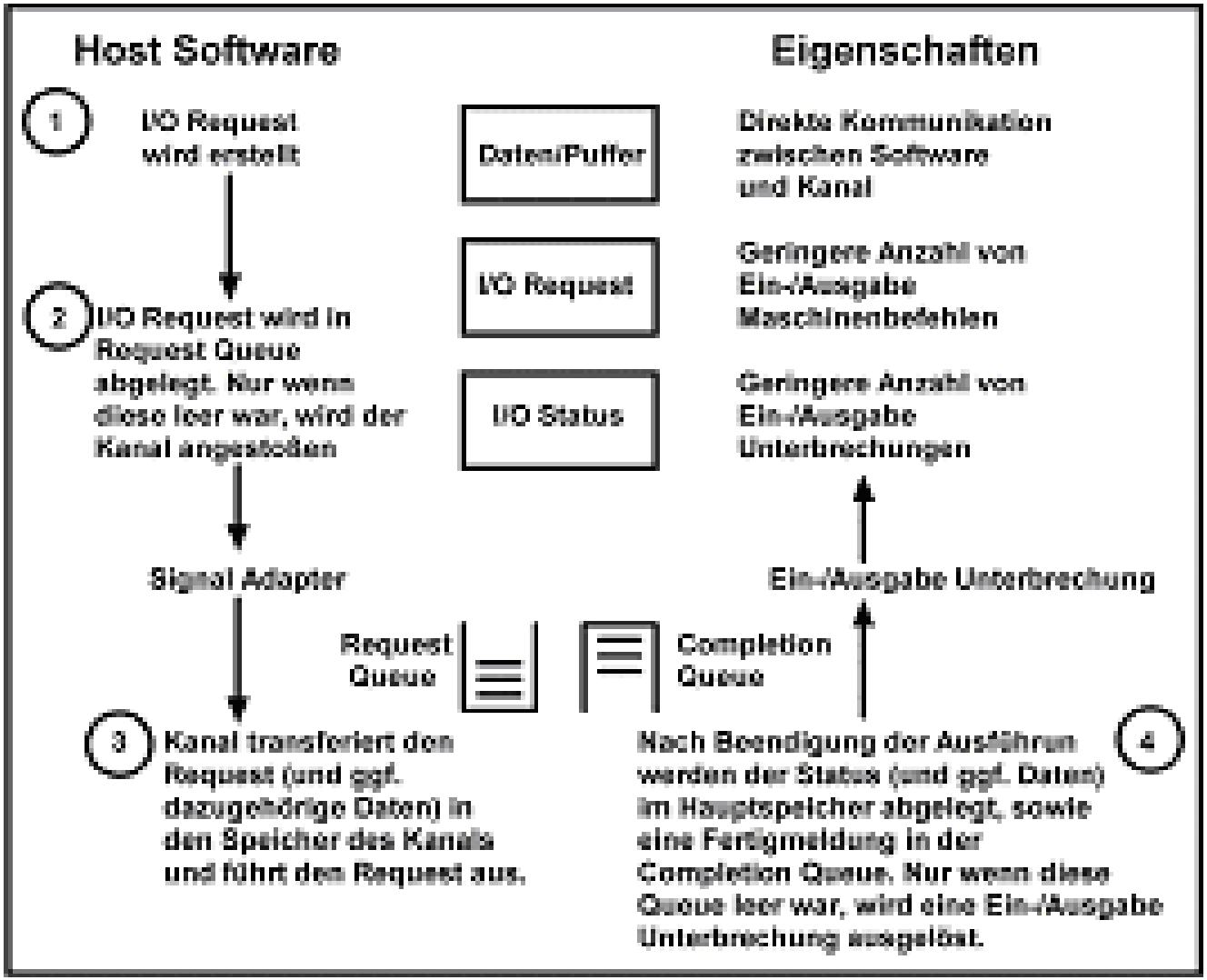
9 oder 18 oder 36 GByte/Platte (kleinere Platten sind schneller)

bis zu 11,2 Tbyte / ESS (2001)

Alternative Datenpfade für jede Übertragung. Alle Komponenten sind doppelt vorhanden. Cache Daten sind gespiegelt. Versagt eine Komponente, gehen keine Daten verloren.

Der Non-Volatile-Cache wird für die Zwischenspeicherung von Schreiboperationen benutzt. Die Idee ist: Wenn Daten einmal im ESS angekommen sind, gelten sie als sicher.

Enterprise Storage Server werden von vielen Firmen angeboten, meistens sowohl mit FC-SCSI als auch mit FICON Anschluß-möglichkeiten: EMC, Hitachi/Sun, MaxData, andere.


Host Software
Eigenschaften
1
I/O Request wird erstellt
Daten/Puffer
Direkte Kommunikation zwischen Software und Kanal
I/O Request
Geringere Anzahl von Ein-/Ausgabe Maschinenbefehlen
2
I/O Request wird in Request Queue abgelegt. Nur wenn diese leer war, wird der Kanal angestoßen
I/O Status
Geringere Anzahl von Ein-/Ausgabe Unterbrechungen
Signal Adapter
Ein-/Ausgabe Unterbrechung
Request Queue
Completion Queue
3
Kanal transferiert den Request (und ggf. dazugehörige Daten) in den Speicher des Kanals und führt den Request aus.
Nach Beendigung der Ausführun werden der Status (und ggf. Daten) im Hauptspeicher abgelegt, sowie eine Fertigmeldung in der Completion Queue. Nur wenn diese Queue leer war, wird eine Ein-/Ausgabe Unterbrechung ausgelöst.
4