

Internet Anwendungen unter OS/390

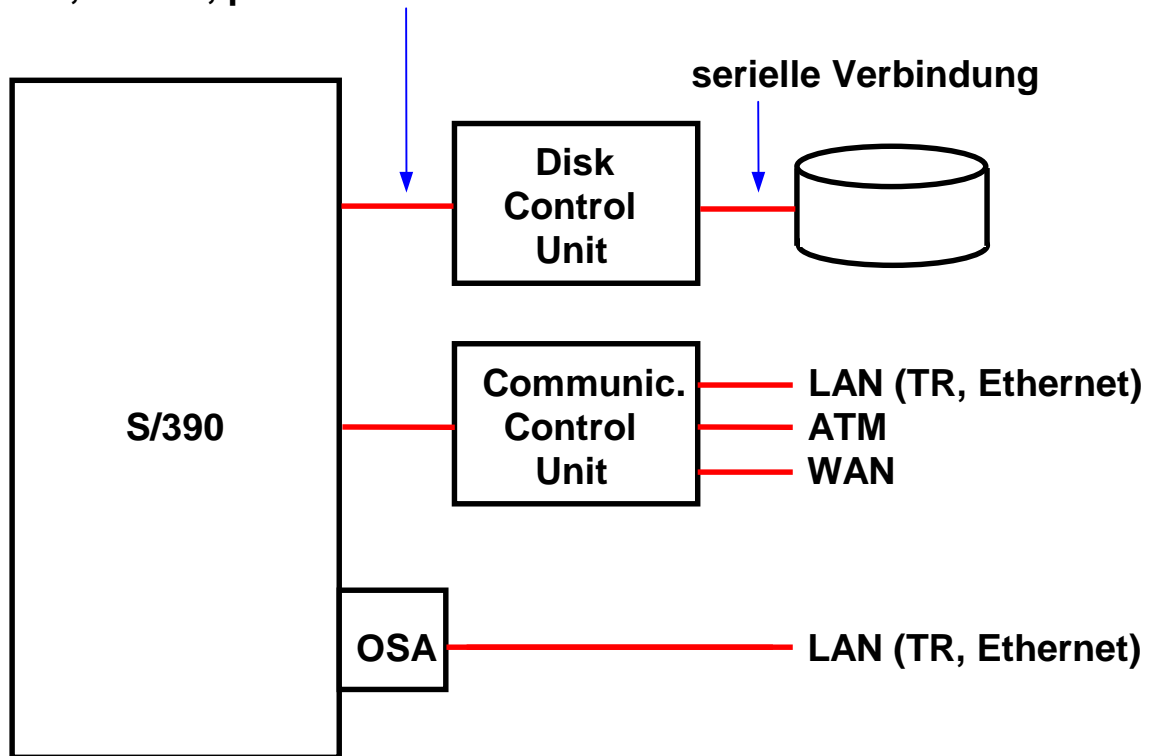
**Dr. rer. nat. Paul Herrmannn
Prof. Dr.rer.nat. Udo Kebschull
Prof. Dr.-Ing. Wilhelm G. Spruth**

WS 2001/2002

Teil 2

ES/390 Betriebssysteme

ESCON, FICON, paralleler Kanal



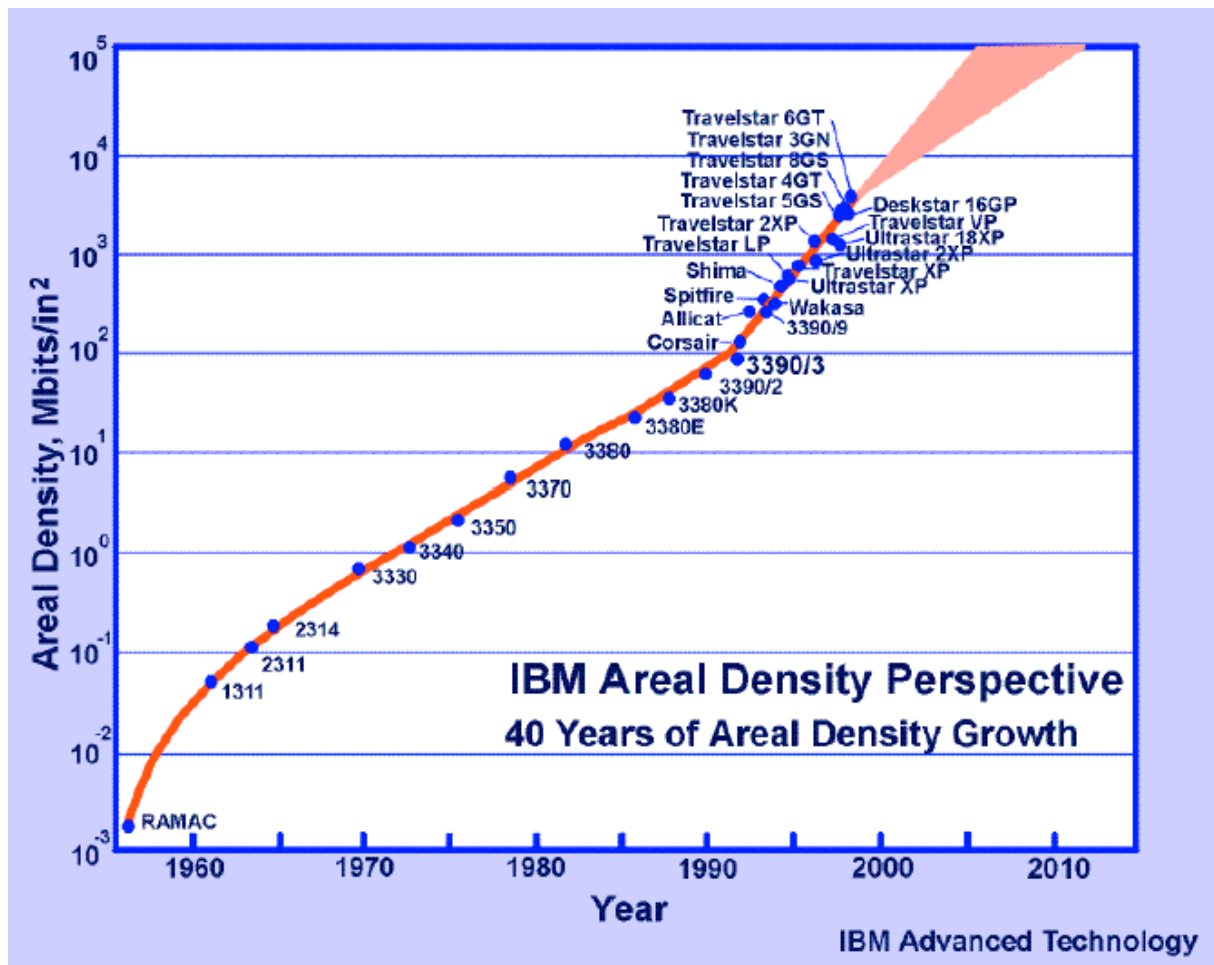
S/390 E/A Konfiguration

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.



3310

Entwicklung der Speicherdichte von IBM Plattenspeichern

Ein/Ausgabe Performance

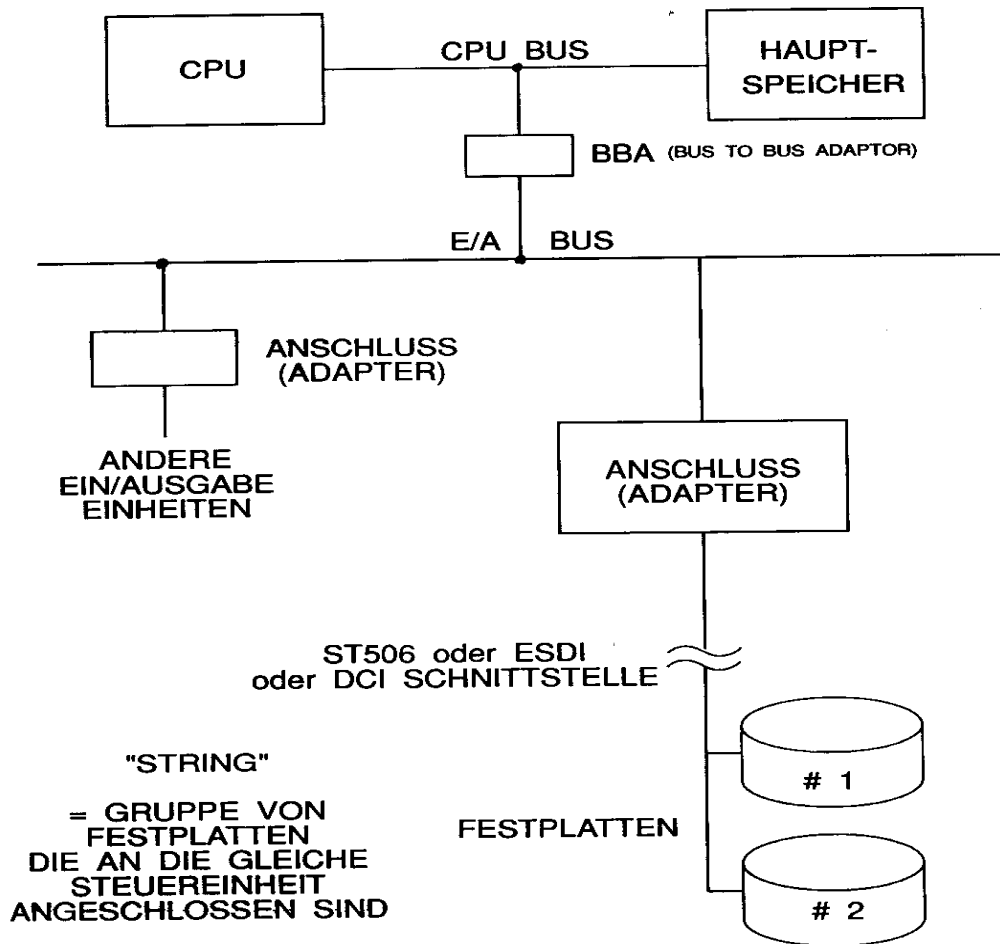
Das Leistungsverhalten in großen kommerziellen C/S Systemen wird in der Regel weniger durch die CPU Geschwindigkeit und mehr durch die Leistungsfähigkeit der Speicherverwaltung und des E/A Systems bestimmt.

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

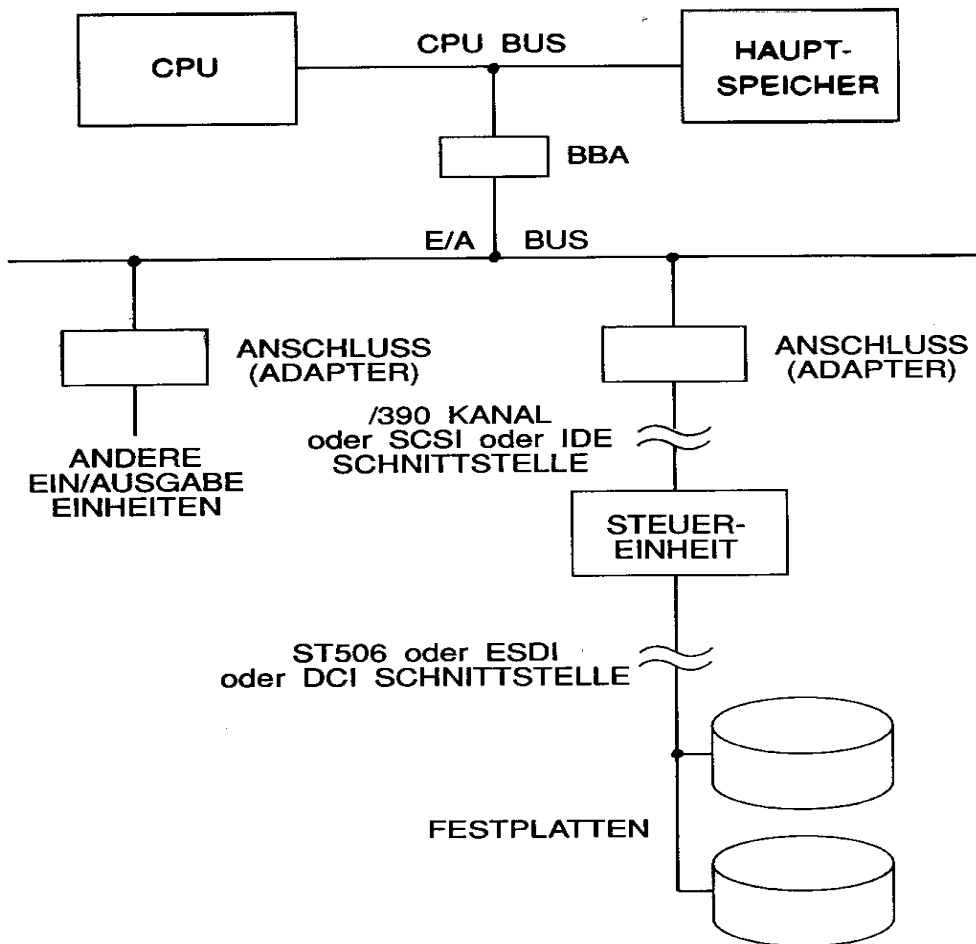
Eine Meßgröße ist die gesamte maximale E/A Datenrate. Eine Angabe hierüber enthält das Februar 1996 Heft der Zeitschrift „Manufacturing Systems“. Hiernach kann das S/390 E/A Subsystem 1,000 bis 20,000+ MByte/Minute übertragen. Sehr große UNIX Systeme können 2 bis 100 Mbyte/Minute übertragen.

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.

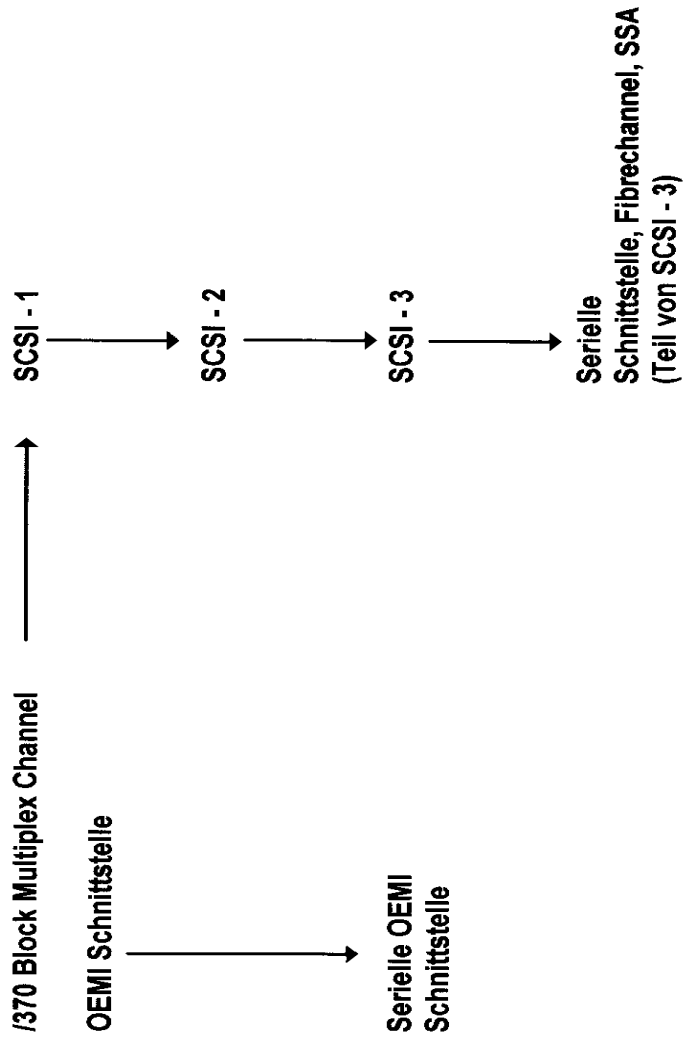


PLATTENSPEICHERANSCHLUSS 1



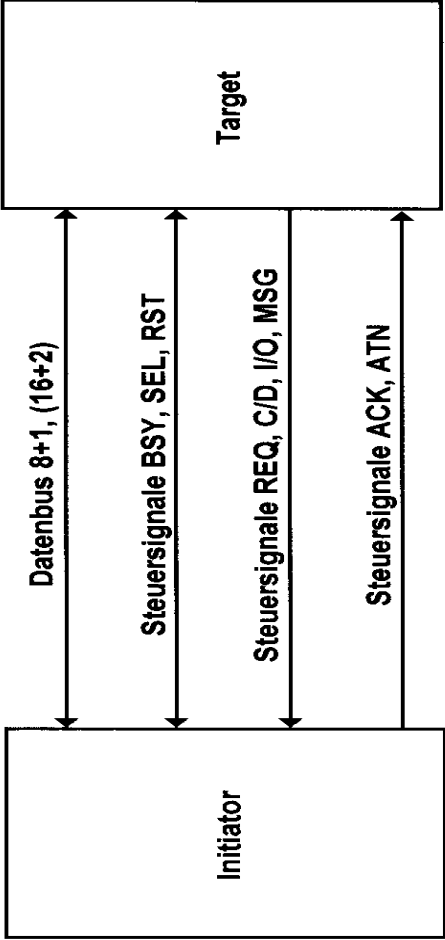
PLATTENSPEICHERANSCHLUSS 2

Historische Entwicklung des Peripherie-Busses

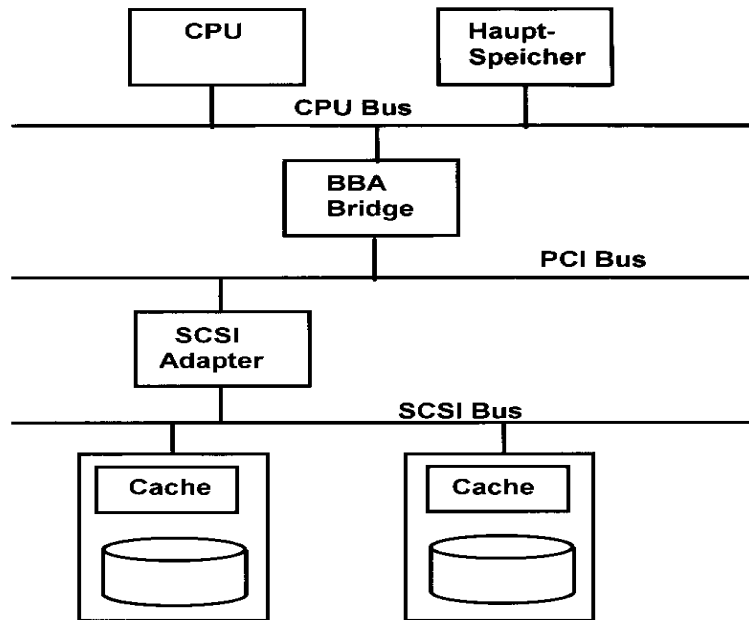


ar1.41D.wwg

wgs c5-97

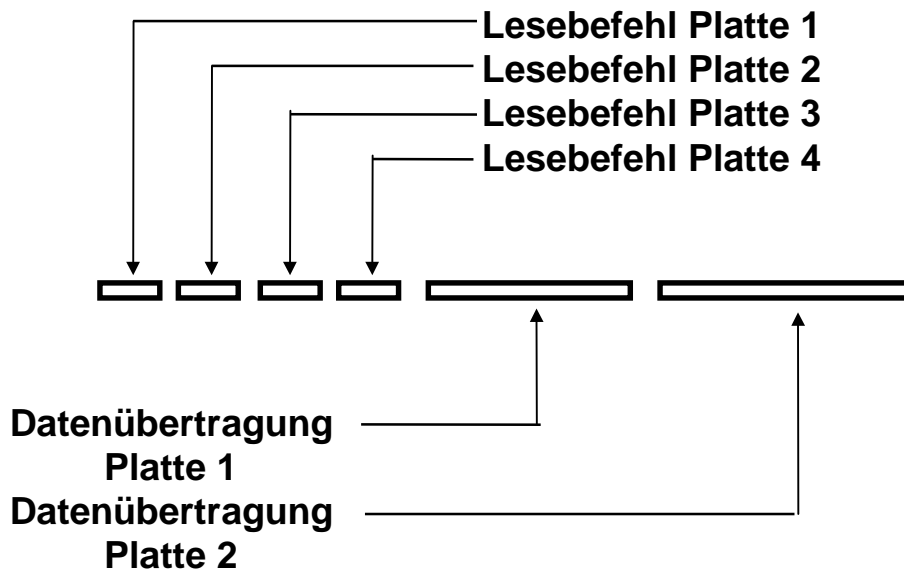


SCSI BUS

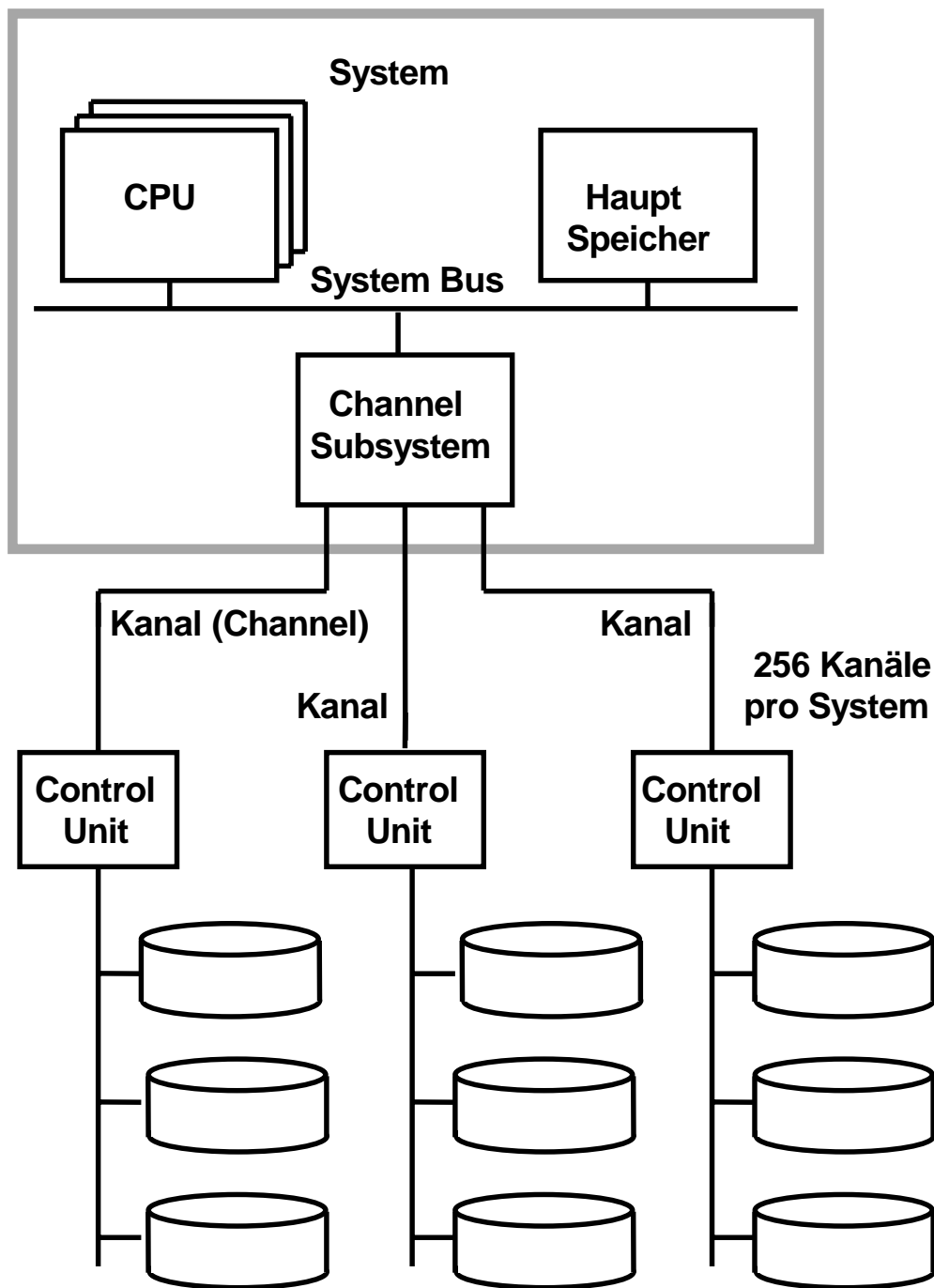


Disconnect/Reconnect beim SCSI Controller

1. SCSI Controller gibt Lese Kommando an Platten Elektronik weiter
2. Platten Elektronik gibt SCSI Bus wieder frei
3. Lesen der Daten in den Cache
4. Übertragung Cache-Hauptspeicher



Verzahnte SCSI Plattenzugriffe



S/390 E/A Konfiguration

Bis zu 65 536 Subchannels pro System. Normalerweise 1 E/A Gerät pro Subchannel. Jedes E/A Gerät ist, unabhängig von seinem physikalischen Anschluß, logisch über einen „Subchannel“ mit dem Channel Subsystem verbunden.

Subchannel und Channel Path

Die Aufgabe der E/A Ansteuerung wird von den drei Baugruppen

- Channel Subsystem
- Control Unit
- E/A Gerät Elektronik

kooperativ übernommen.

Als Channel Path wird die physikalische und logische Verbindung zwischen einem System und einer Control Unit bezeichnet. Während einer E/A Operation kann der Datentransfer über einen Subchannel auf unterschiedliche Channel Path abgebildet werden. Ein Channelpath wird durch eine 8 Bit CHPID gekennzeichnet.

(Die Begriffe „Channel Path“ und „Channel“ (Kanal) sind weitestgehend identisch.)

Jedem angeschlossenen Gerät (z.B. Plattenspeicher) ist ein Subchannel zugeordnet. Jeder aktive Subchannel wird durch einen Speicherblock innerhalb des Hauptspeichers dargestellt und durch eine 16 Bit Adresse gekennzeichnet.

Zwei Channel Path Typen:

- Die (ältere) „Parallel I/O Interface“ überbrückt Distanzen bis zu 130 m. Sie hat Ähnlichkeit mit der parallelen SCSI Interface und wird mit einem parallelen Kupferkabel implementiert.
- Die „Serial I/O Interface“ verwendet Glasfaser, und erlaubt Datenraten von 17 („Escon“) oder 100 („Ficon“) Mbyte/s. Es können Entfernungen bis zu 43 km überbrückt werden.

Ein System kann über mehrere Channel Path mit einer Control Unit verbunden werden, und eine Control Unit kann an mehrere Systeme angeschlossen werden.

Ein E/A Gerät kann mit mehr als einer Control Unit verbunden werden.

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 (CSW)

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)

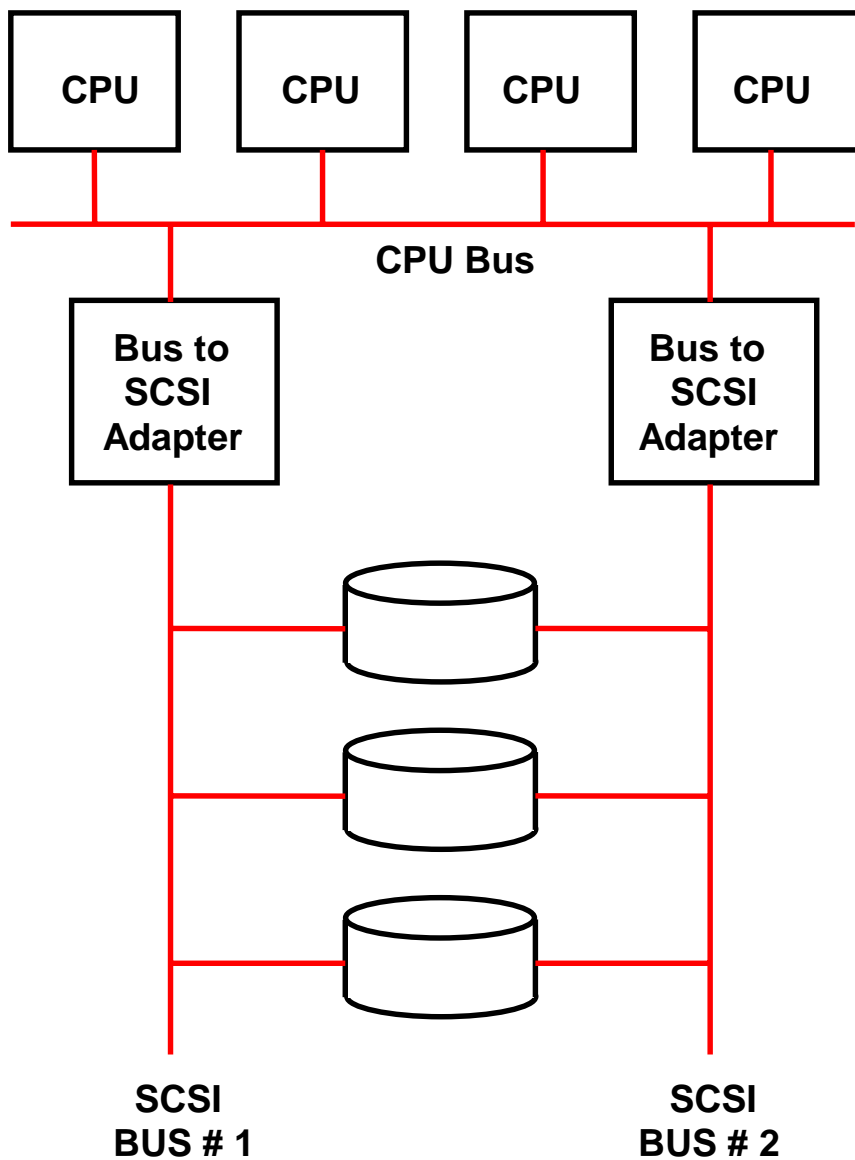
Spuranfangssignal

Steuerung des Zugriffsmechanismus

Lese / Schreibkopf selektieren (Plattenoberfläche)

**Fehler Erkennung
(Syndrom Checking, Syndrom = 5 - 6 Bytes)**

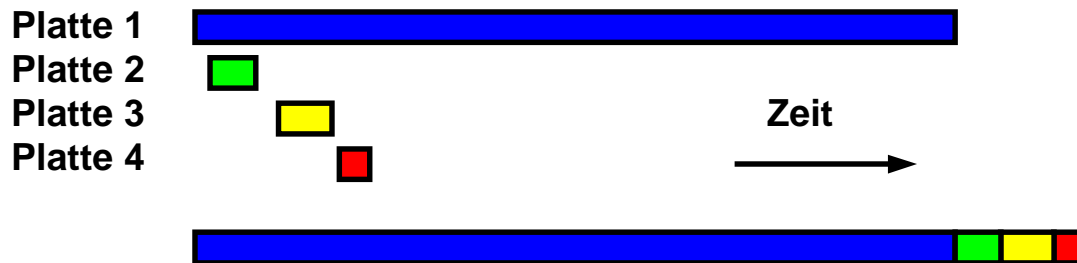
Status setzen



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.



Unix SCSI Datentransfer

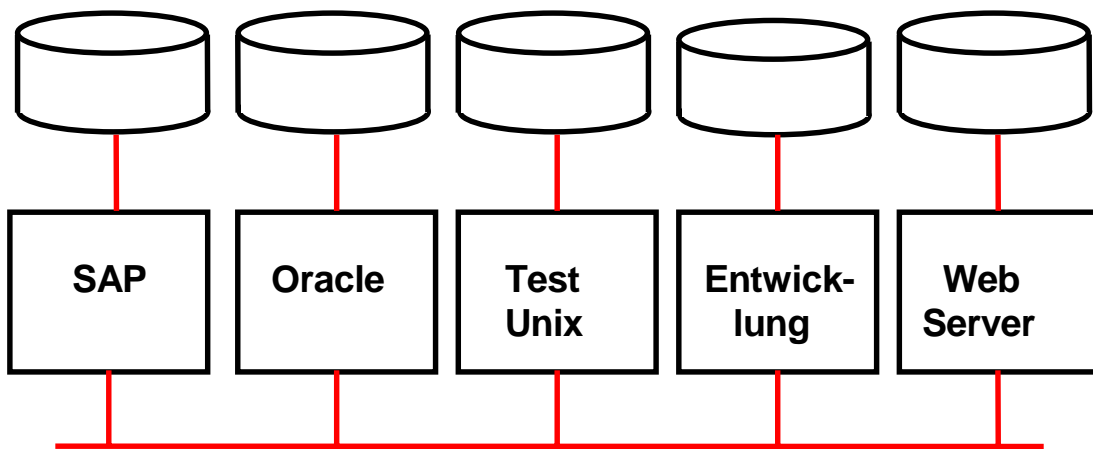
Hohe SCSI Datenraten werden bei E/A Operationen erreicht, die große Blöcke zwischen Platte und Hauptspeicher transportieren. Gleichzeitige E/As für unterschiedliche Platten, die an den gleichen SCSI Kanal angeschlossen sind, verursachen Kanal- Contention, welche die E/A Leistung drastisch verringern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

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

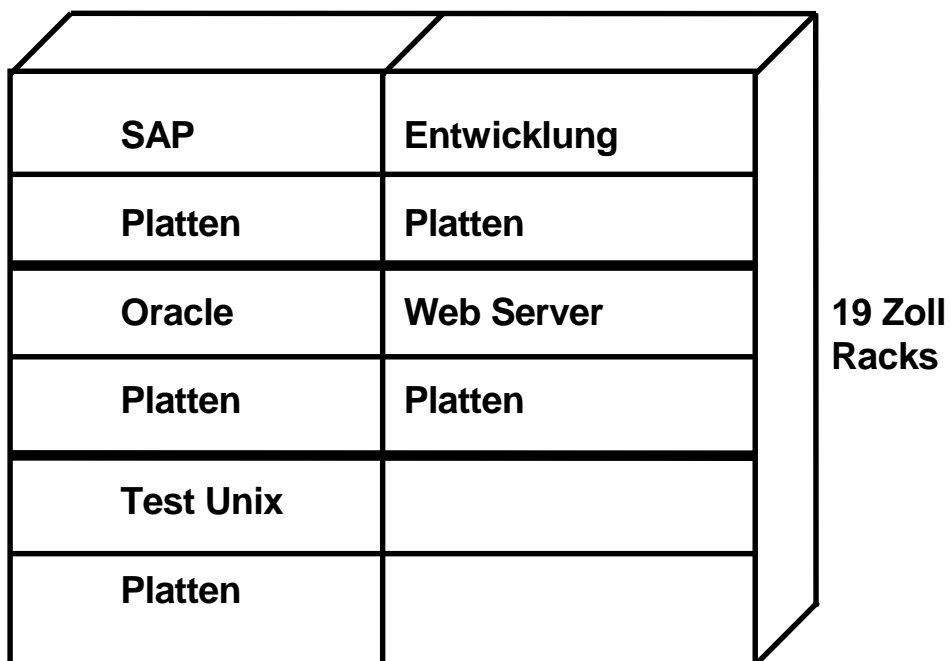
Die Übertragung großer Datenblöcke blockiert die Übertragung zahlreicher kleinen Datenblöcke von anderen Plattenspeichern des gleichen Strings. SCSI E/A basierte Datenbank Systeme benötigen große E/A Puffer Pools um das Contention Problem zu lösen.

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.



Problem: Administration, Wartung



Multiple Unix Konfiguration
 Alternative: LPAR

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 :

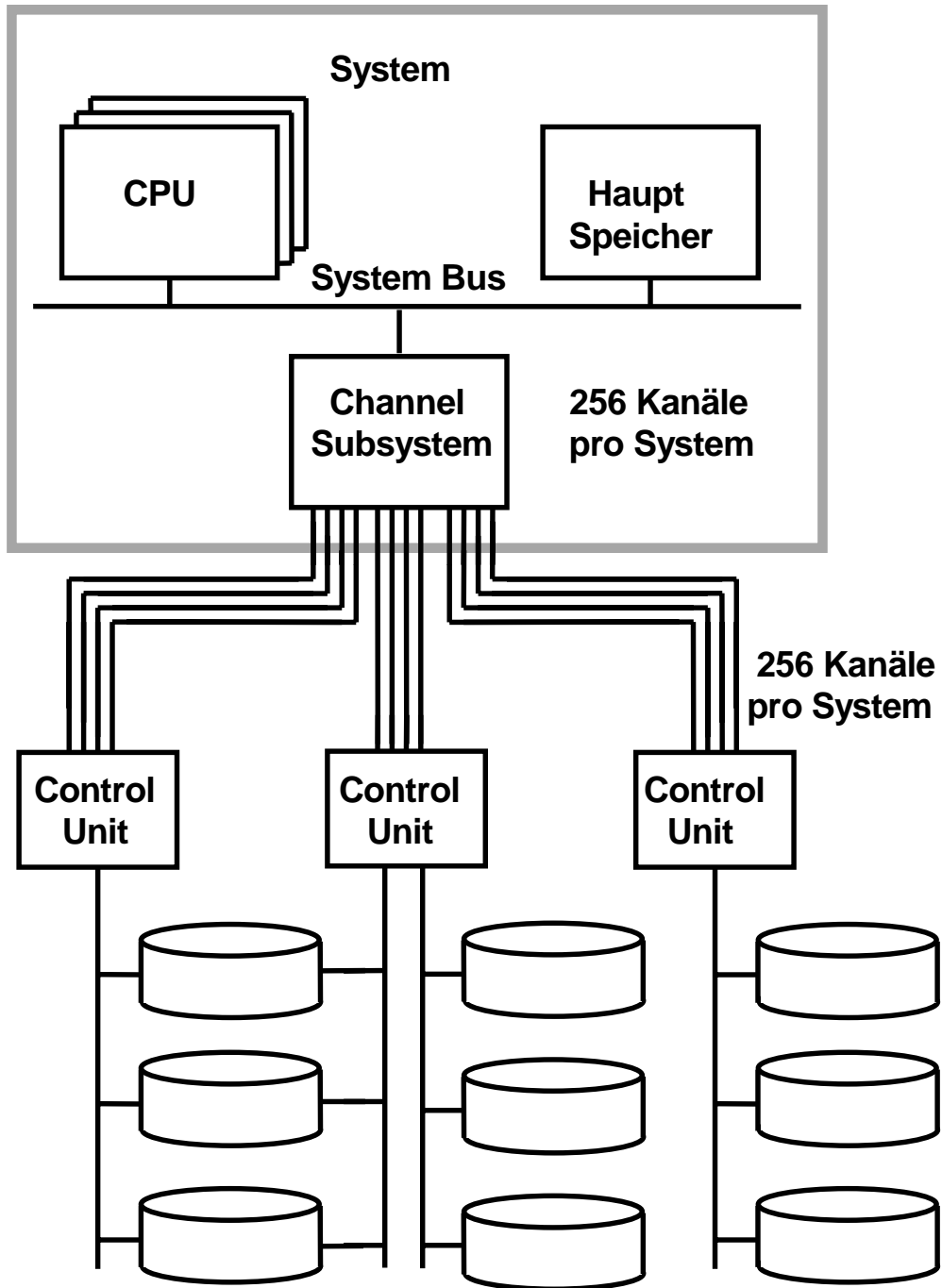
“ Peak performance is an area that Sun is working on, and one of the advantages of mainframe over UNIX is that mainframes are capable of running at 85-95% of capacity. UNIX servers run at 20%-30% of peak load”.

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.

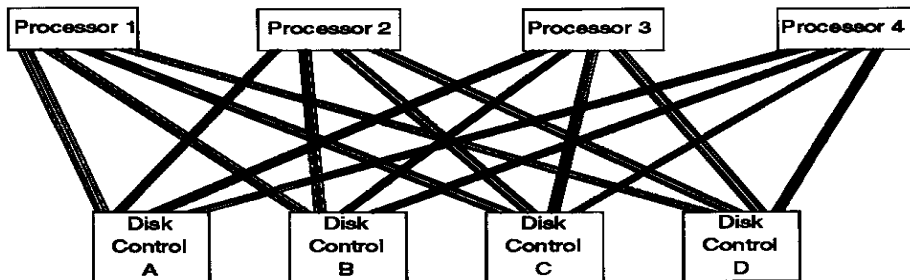


S/390 E/A Konfiguration

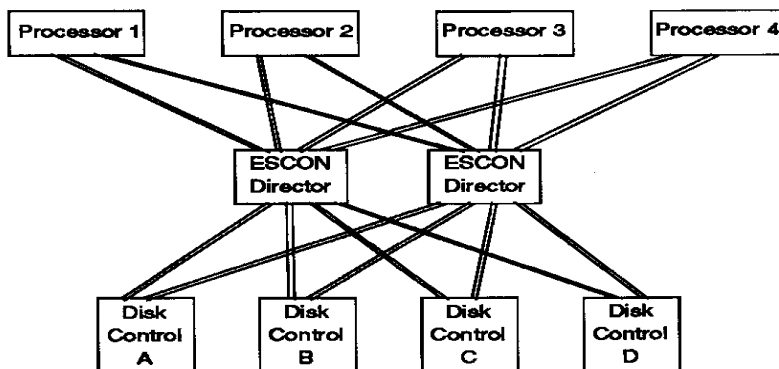
Ein System kann über mehrere (z.B. 4) Kanäle mit einer Control Unit verbunden werden, und ein E/A Gerät kann an mehr als eine Control Unit angeschlossen werden

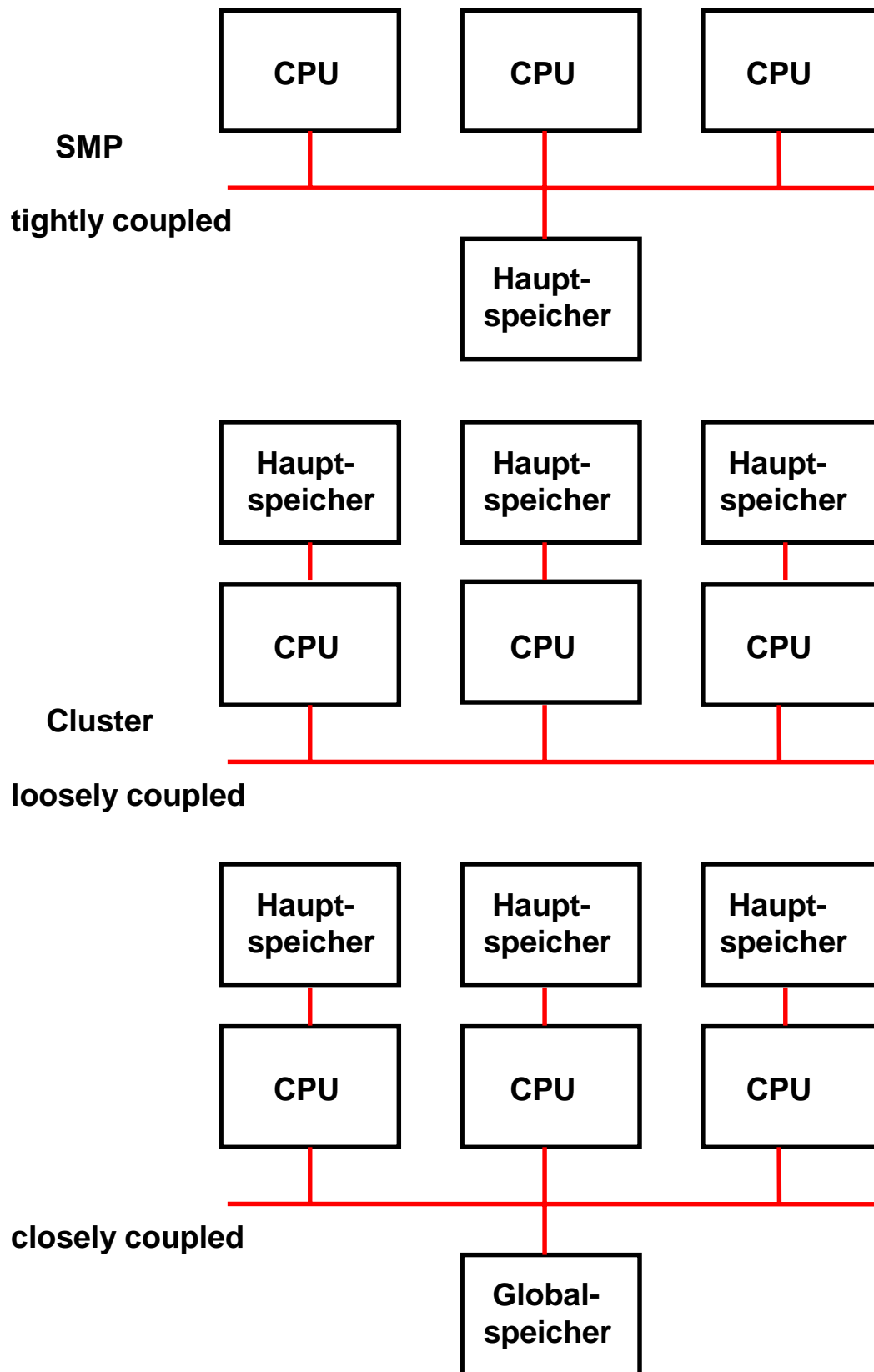
Ein Channel Path ist die logische Verbindung eines E/A Gerätes mit einem Channel Subsystem

Parallel Channel Configuration

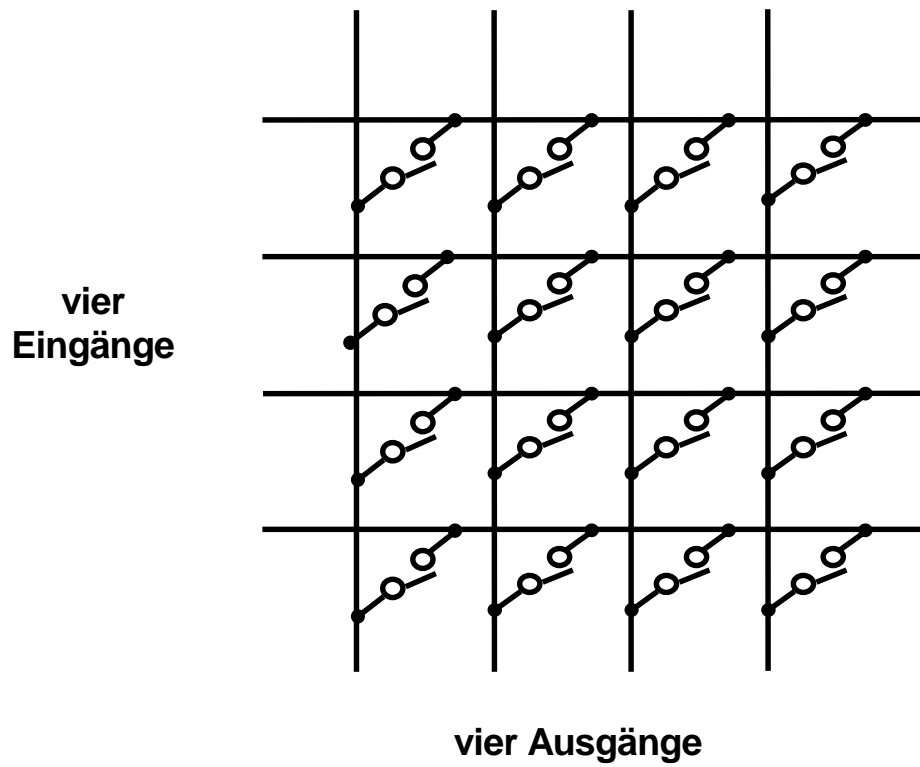


ESCON Channel Configuration

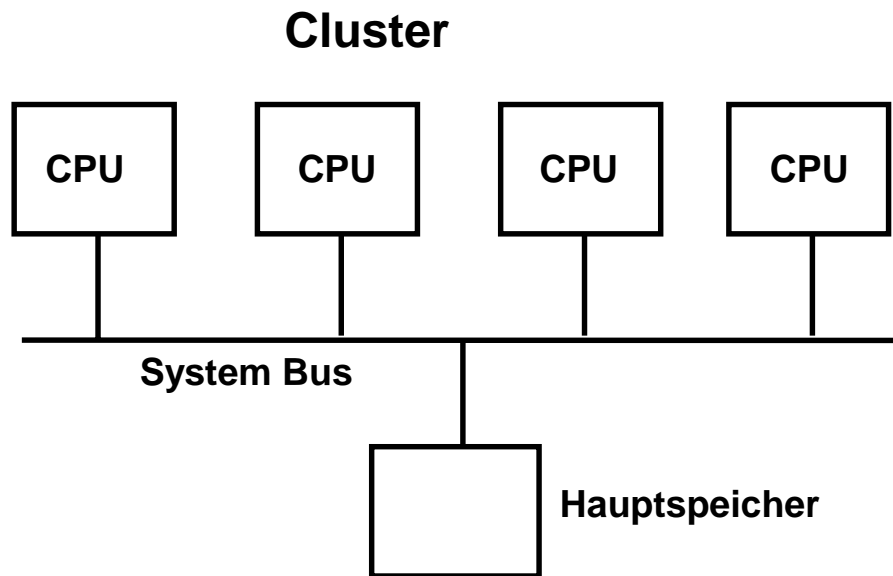




Taxonomie von MIMD Parallelrechnern

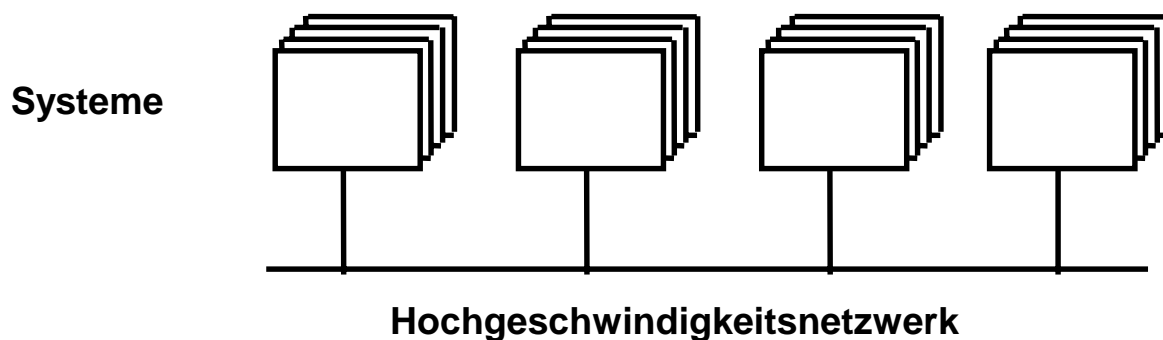


4 x 4 Crossbar Matrix Switch

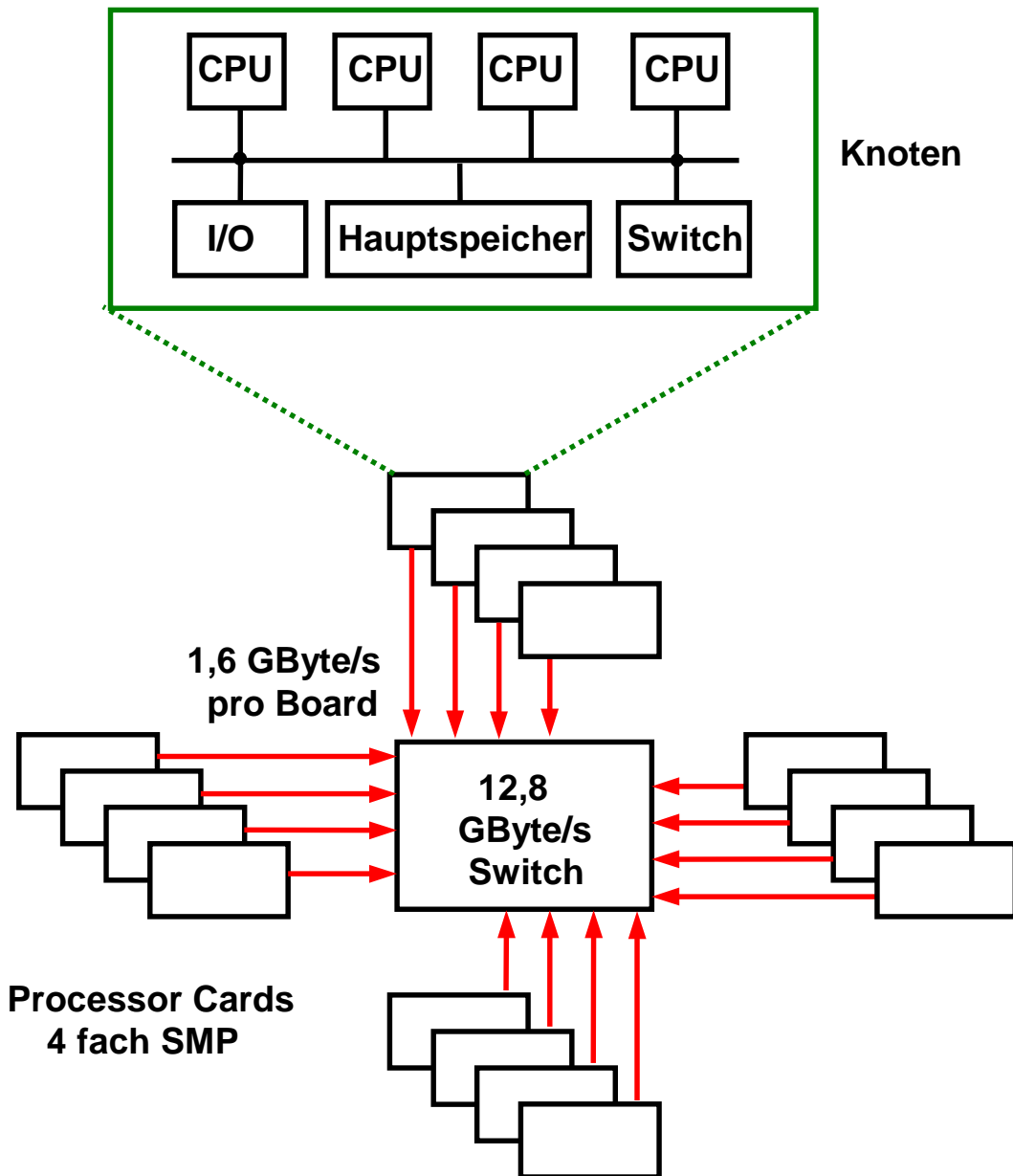


Ein System (Prozessor, Node) besteht aus mehreren CPU's, die auf einen gemeinsamen Hauptspeicher zugreifen (SMP., Symmetric Multiprocessor).

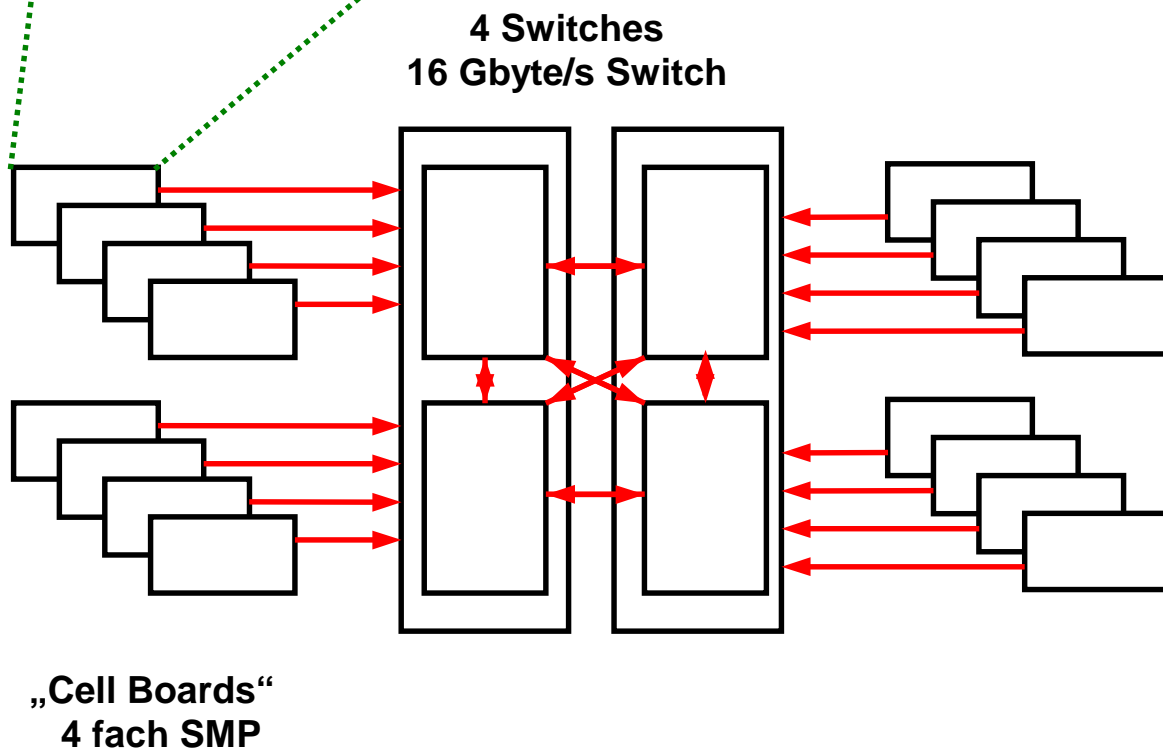
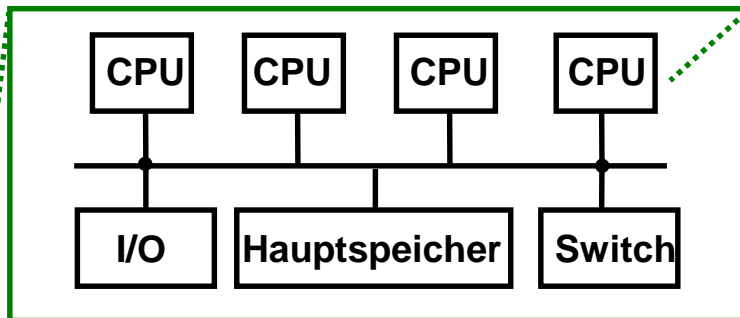
Im Basisfall nur eine Kopie (Instanz) des Betriebssystems im gemeinsam genutzten Hauptspeicher



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.

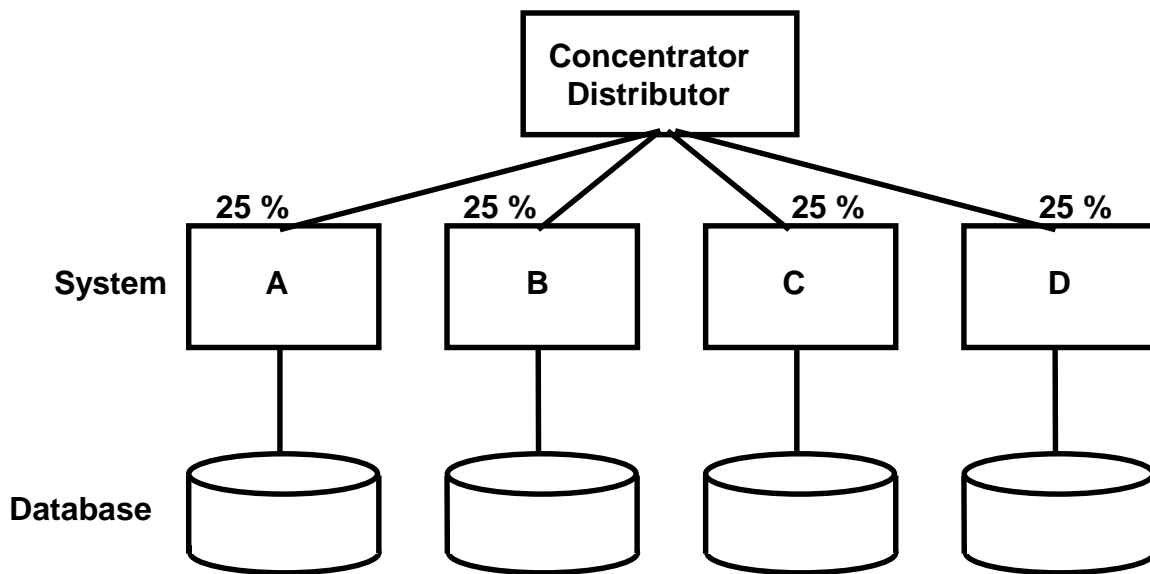


Sun E10000 Cluster
64 CPU's
16 Knoten, je 4 CPU/Knoten
I/O Controller auf jeder Karte



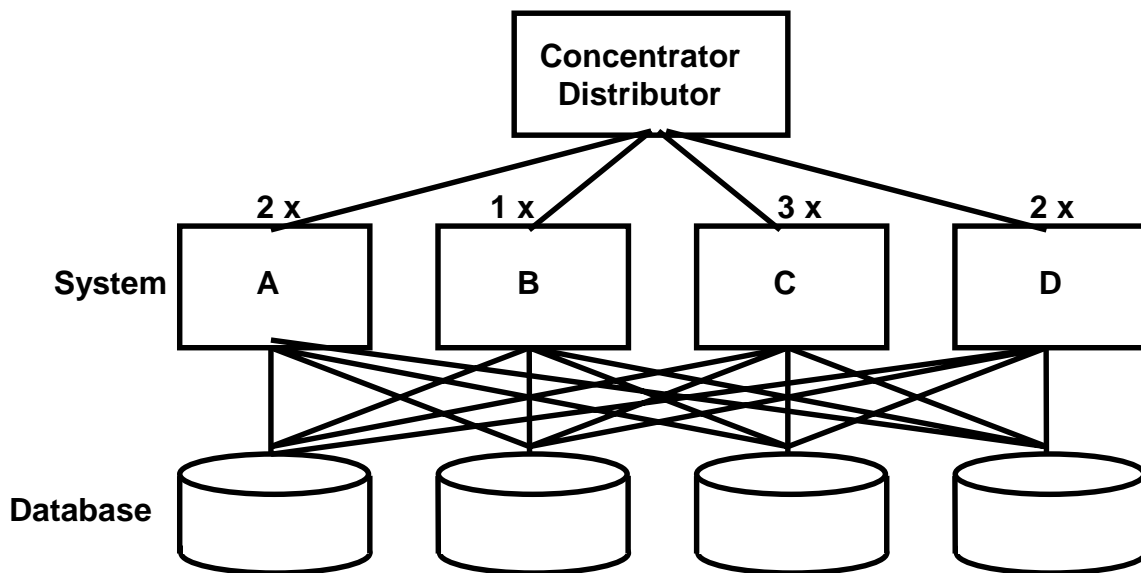
HP Superdome Cluster
64 CPU's
16 Knoten (Cell Boards), je 4 CPU/Knoten
I/O Anschluß auf jedem Cell Board

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



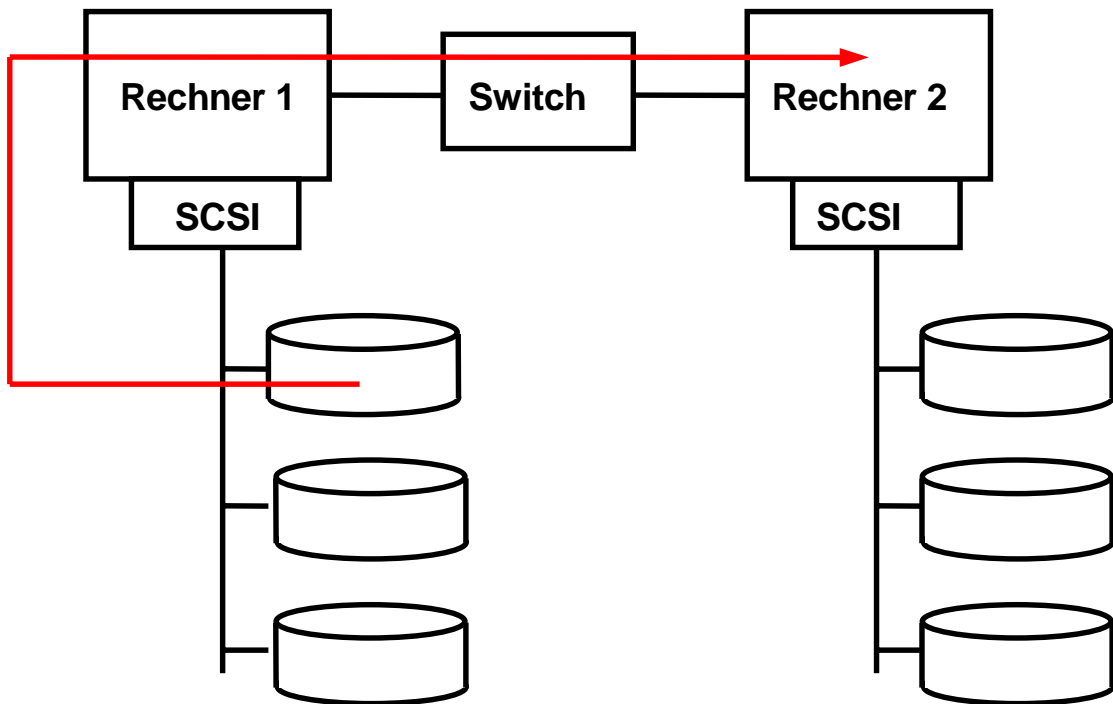
Shared nothing (partitioned data)

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.

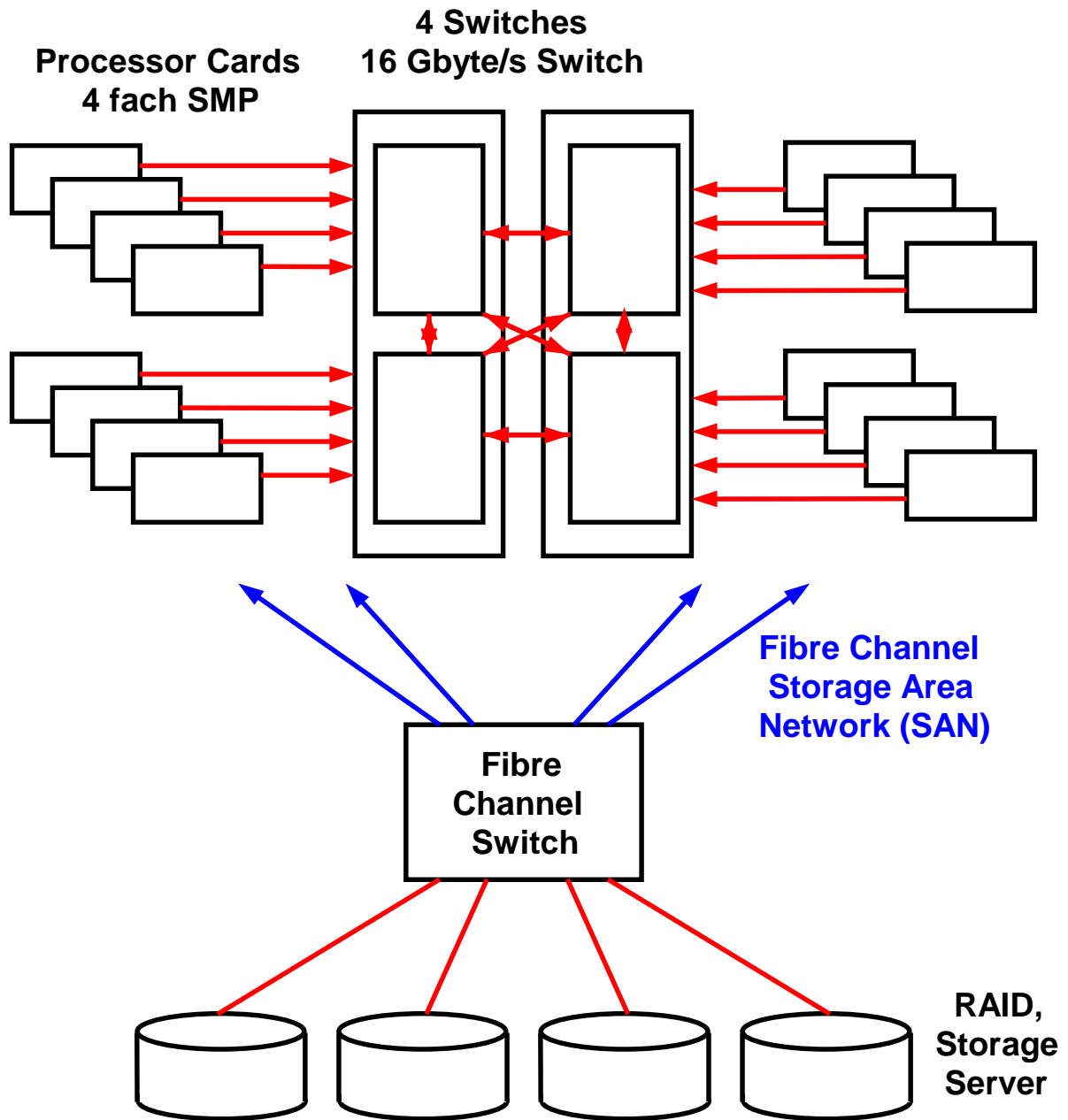


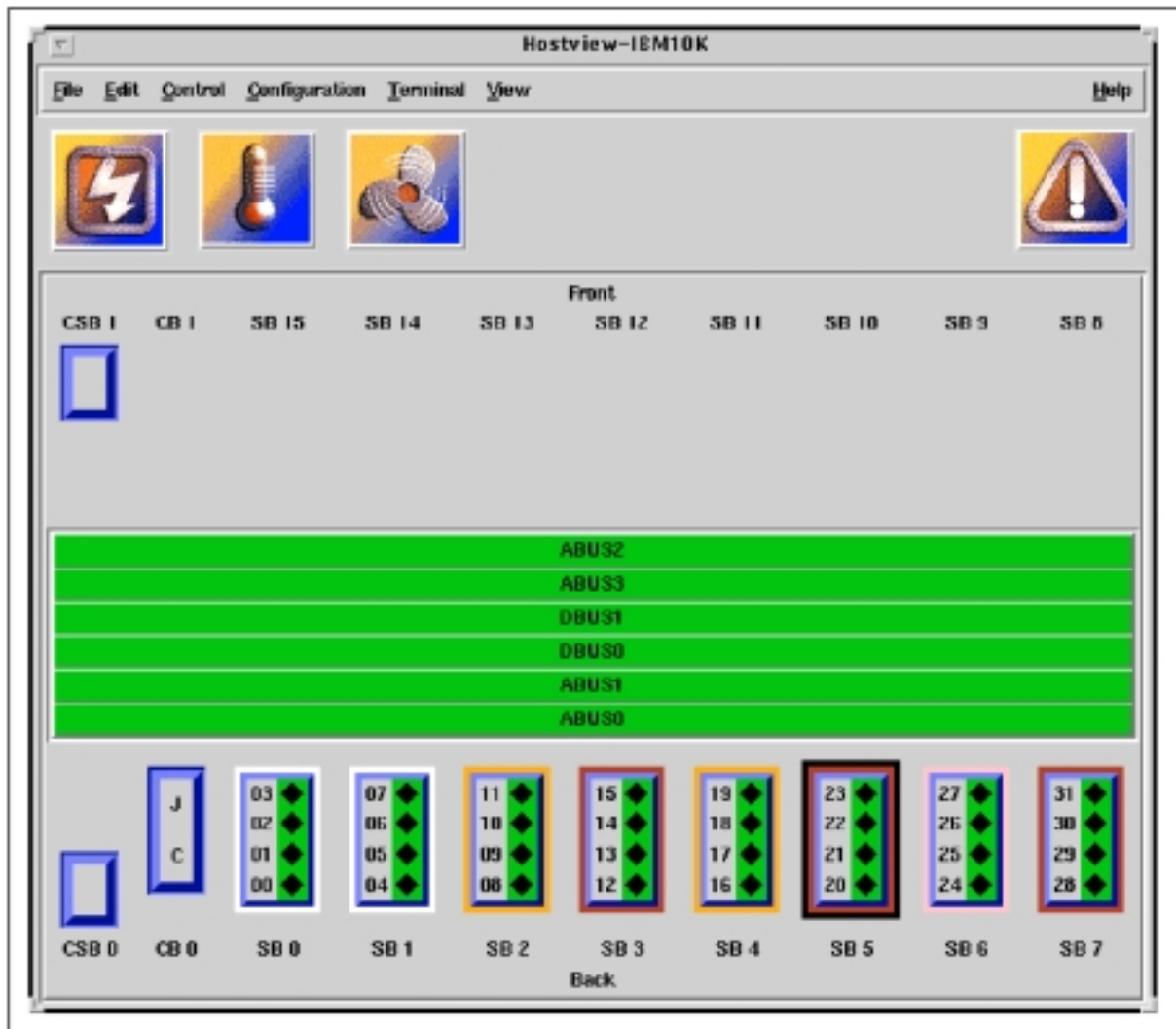
Shared Disk Emulation

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

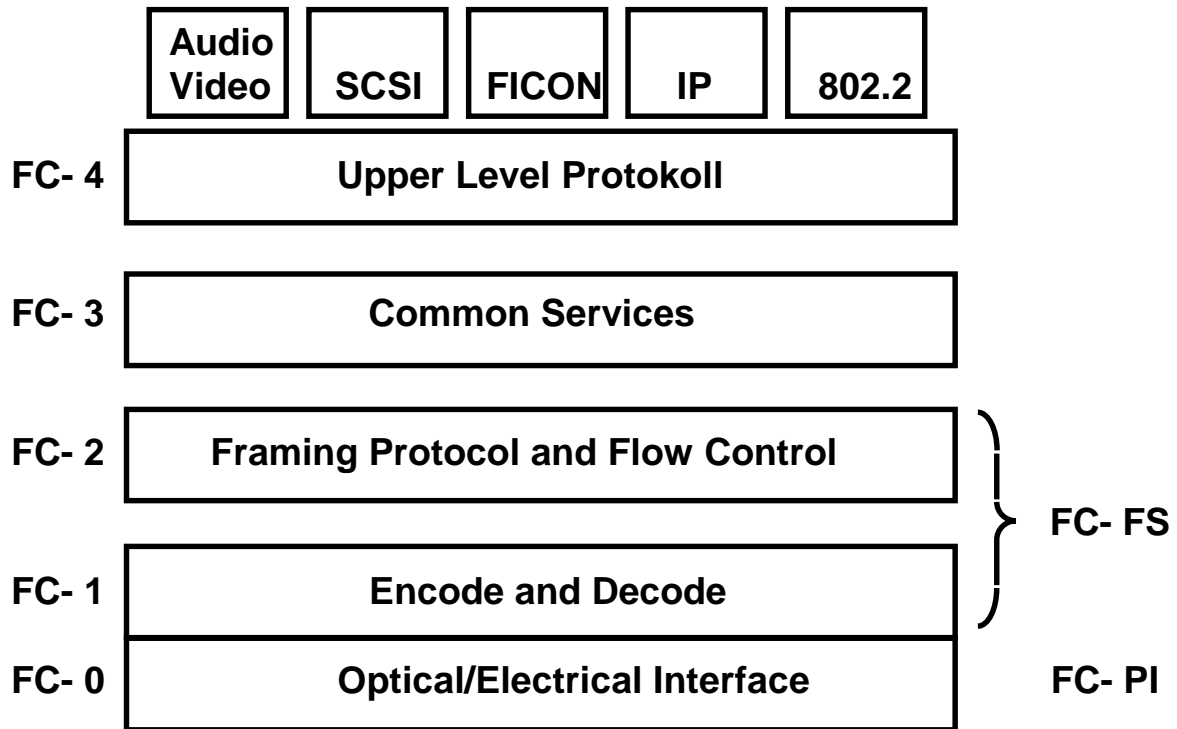




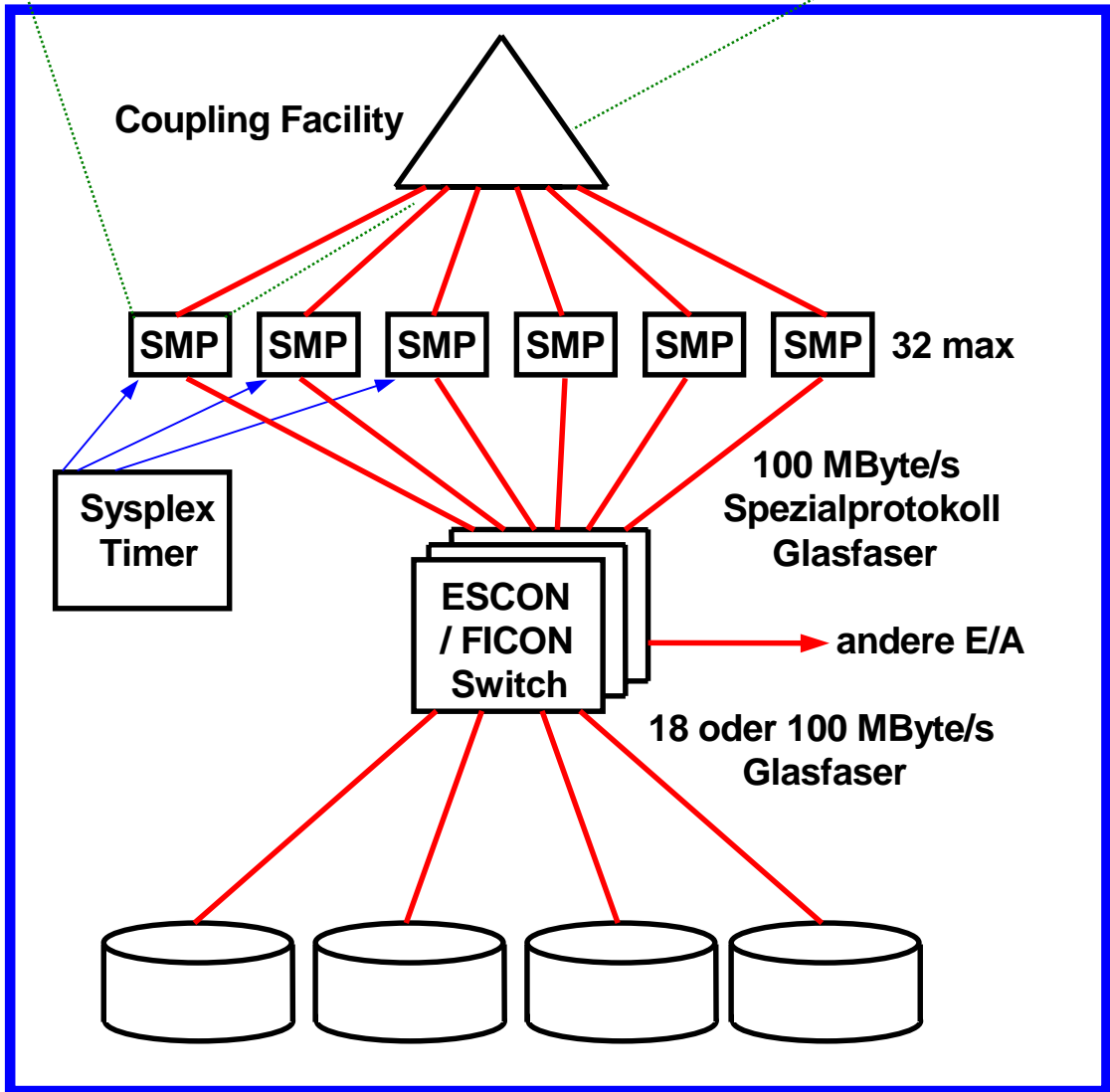
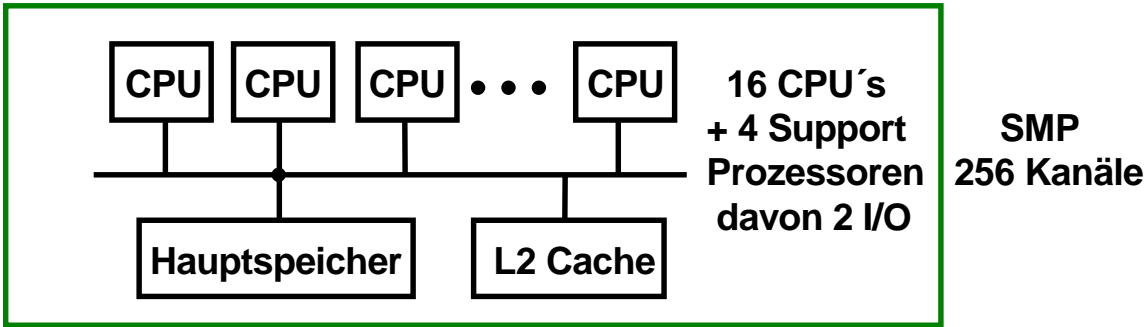
Sun E 10 000 Administrator Konsole

Dargestellt sind 8 Prozessor Boards SB0 .. SB7.

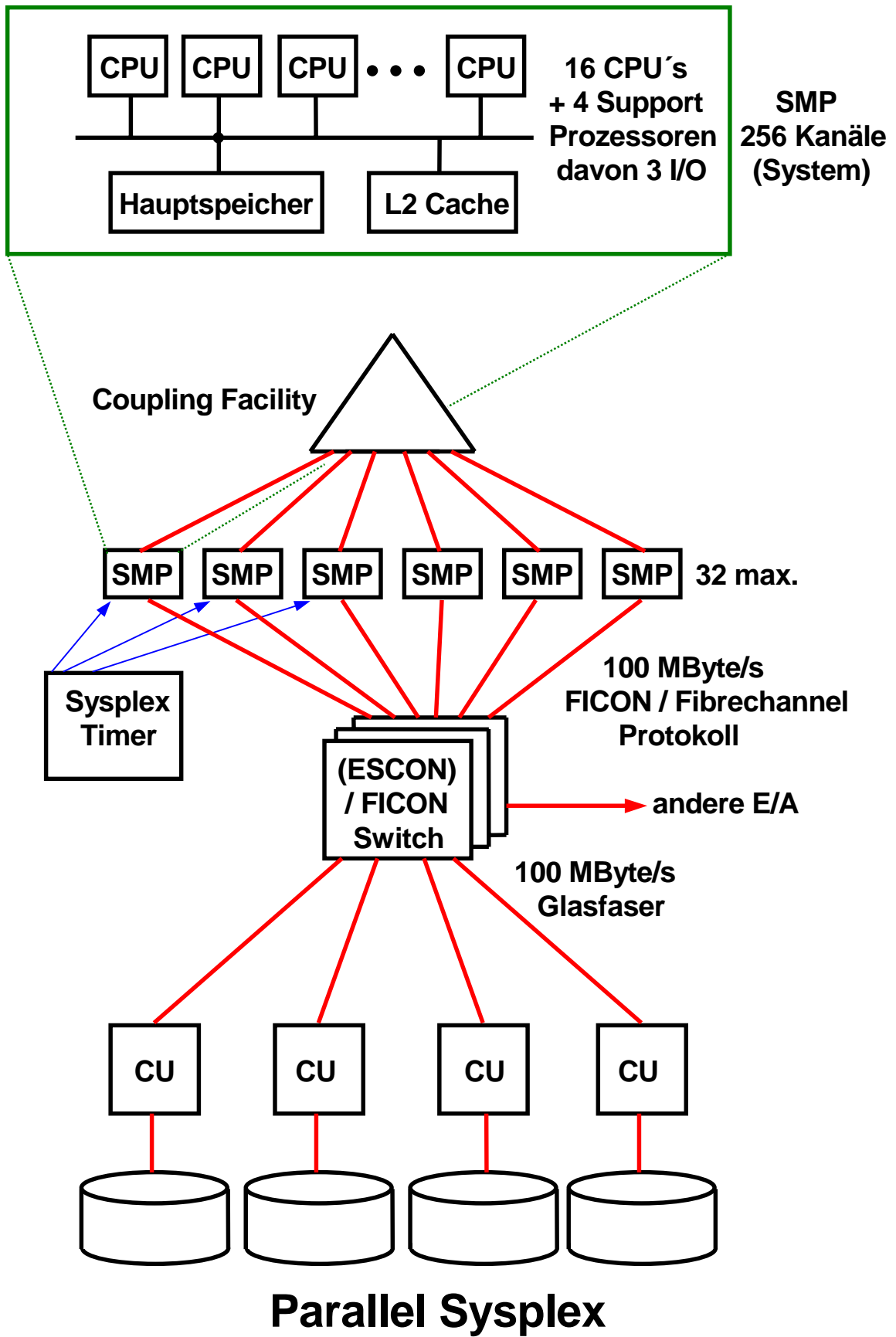
Cluster: SB0, SB1
 SP2, SB4
 SP3, SB7



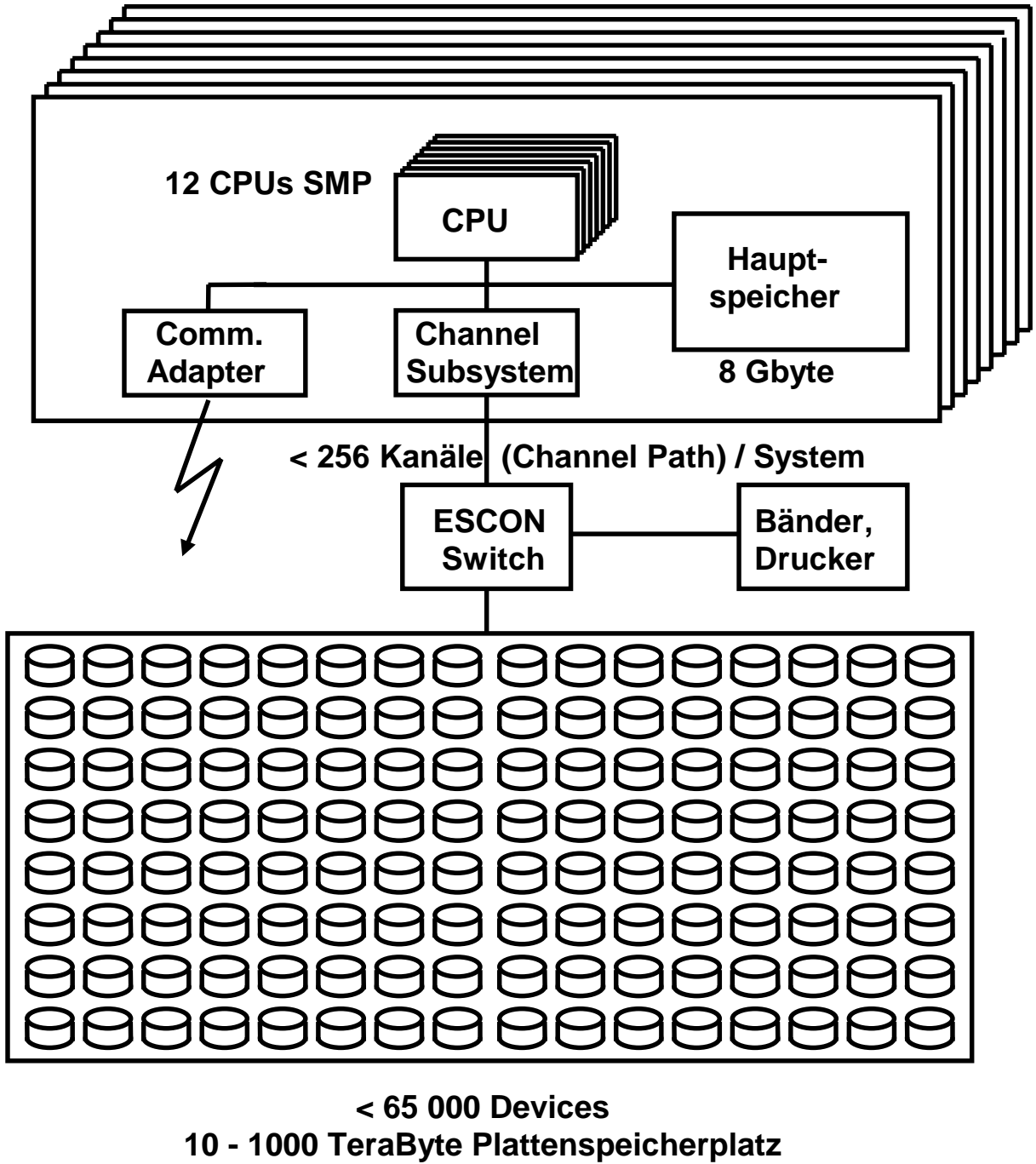
Fibre Channel Standard Architektur



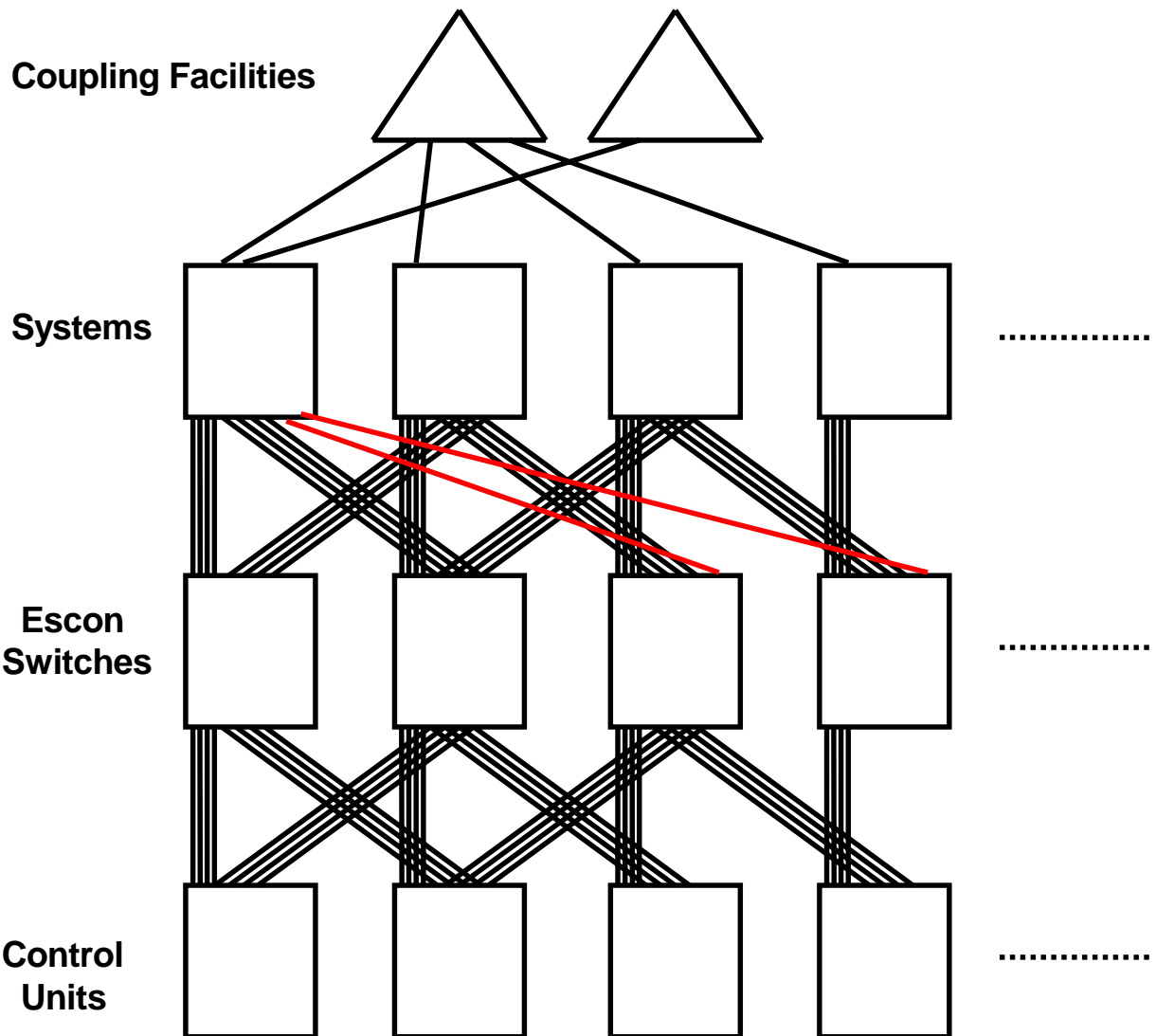
Parallel Sysplex



32 Systeme , 12 CPUs / System



S/390 Großsystemkonfiguration



Jedes System hat bis zu 256 Channels

Jeder Escon Switch hat bis zu 256 Ports

Große Installationen haben (1999)

8 - 10 Systeme

Durchschnitt 8 CPU /System

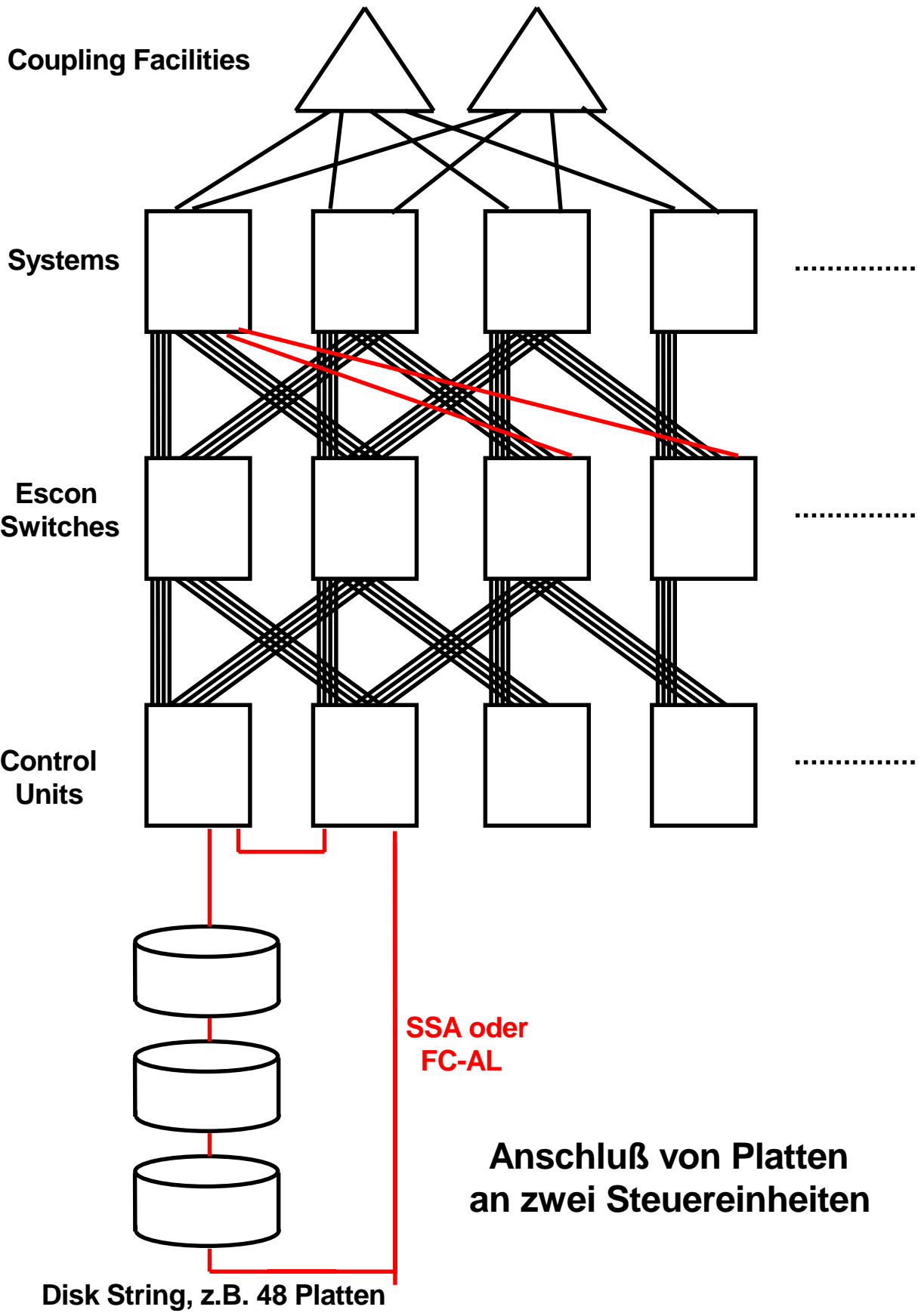
100 - 200 Tbyte Plattenspeicherplatz

(Telekom hatte 1999 300 Tbyte installiert)

15 - 20 Escon Switches

Escon = 17 Mbyte/s = 200 Mbit/s

Ficon = 100 MByte/s



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

es 0416 ww6

wgs 09-99

Parallel Sysplex Cluster Technology

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

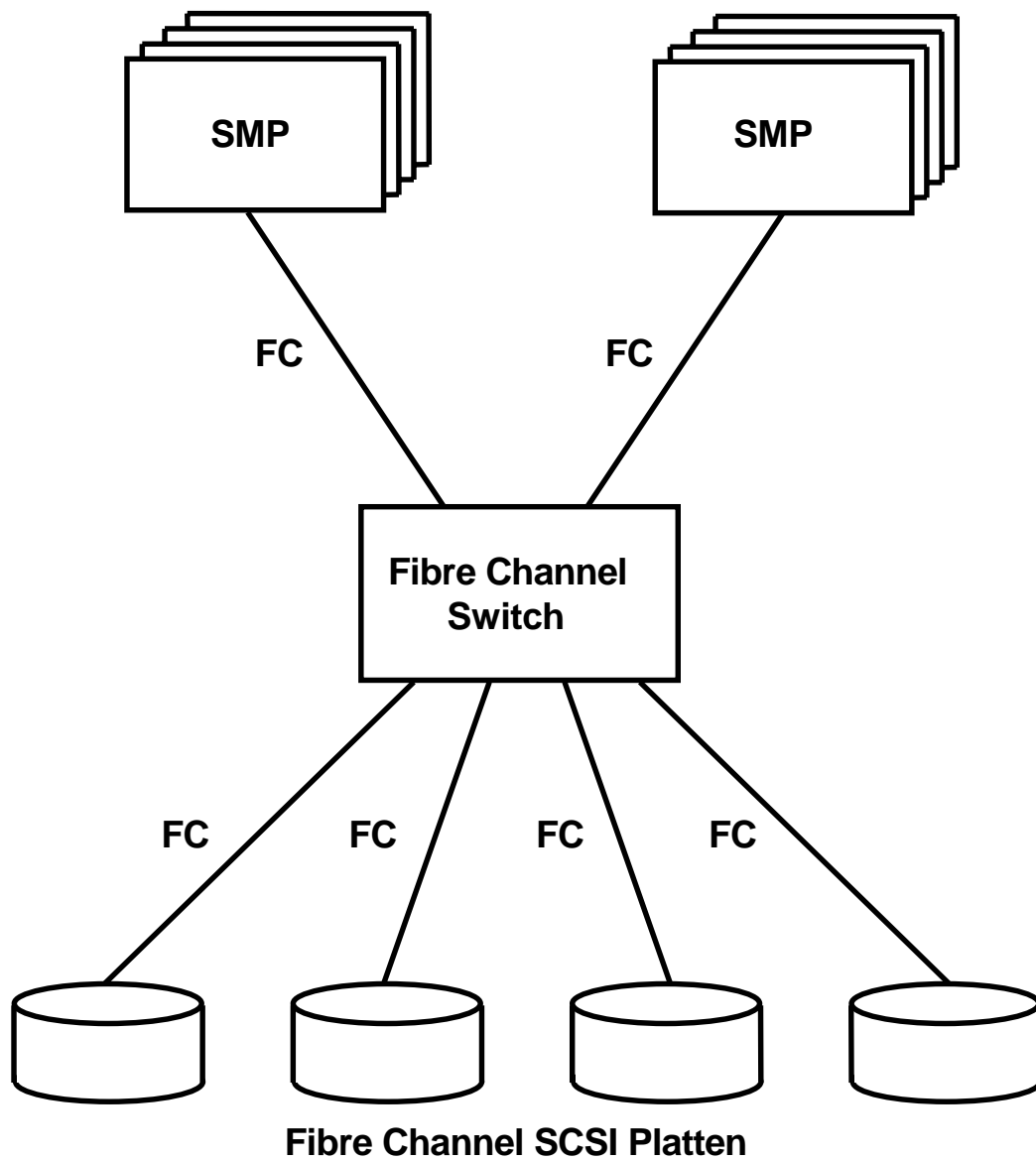
Parallel Sysplex Cluster Technology Komponenten:

- **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

es 0409 ww6

wgs 04-99



Einfache Fibre Channel Konfiguration

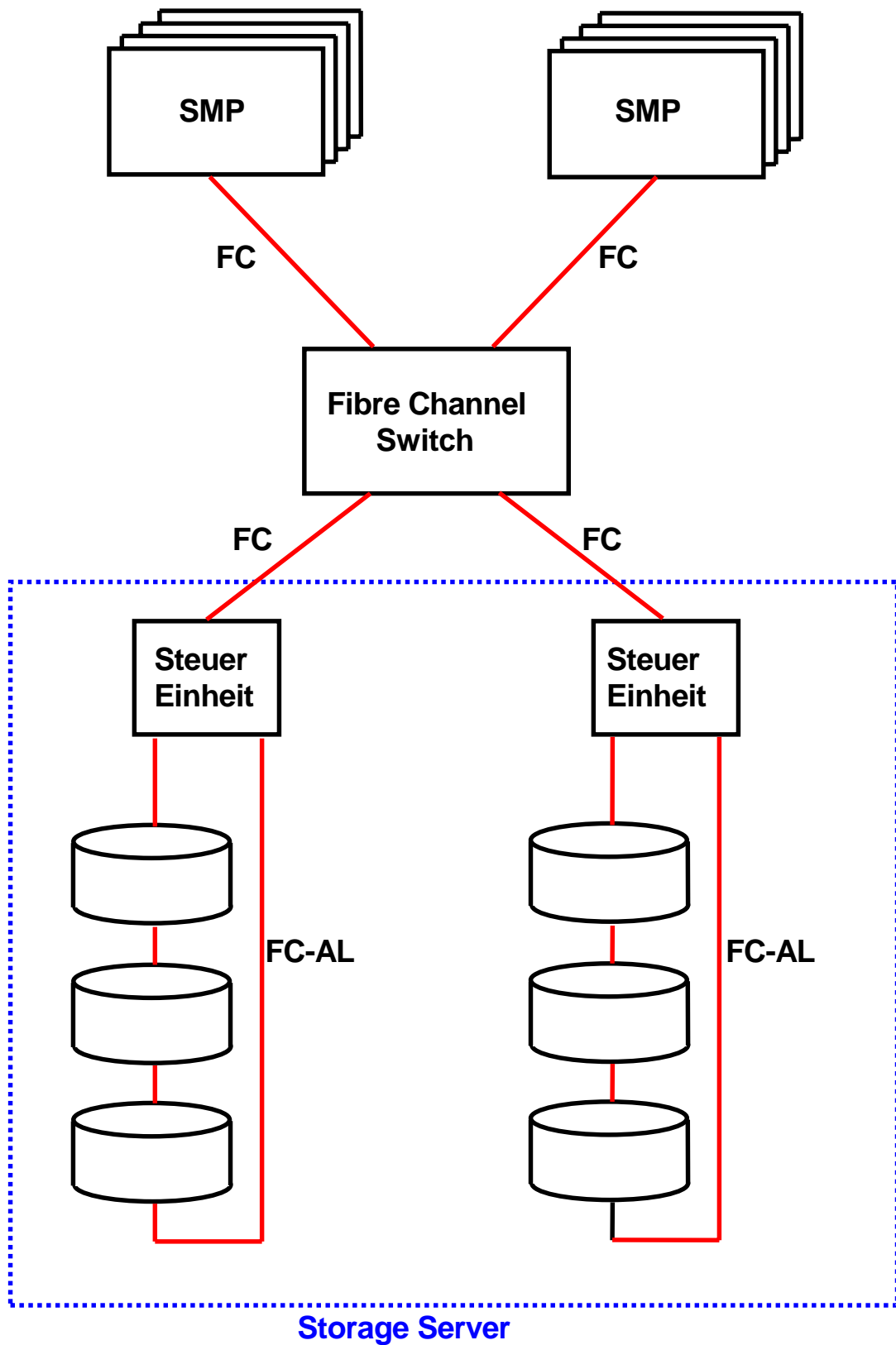
Unterschiedliche Festplattenanschlüsse

ATA (IDE)

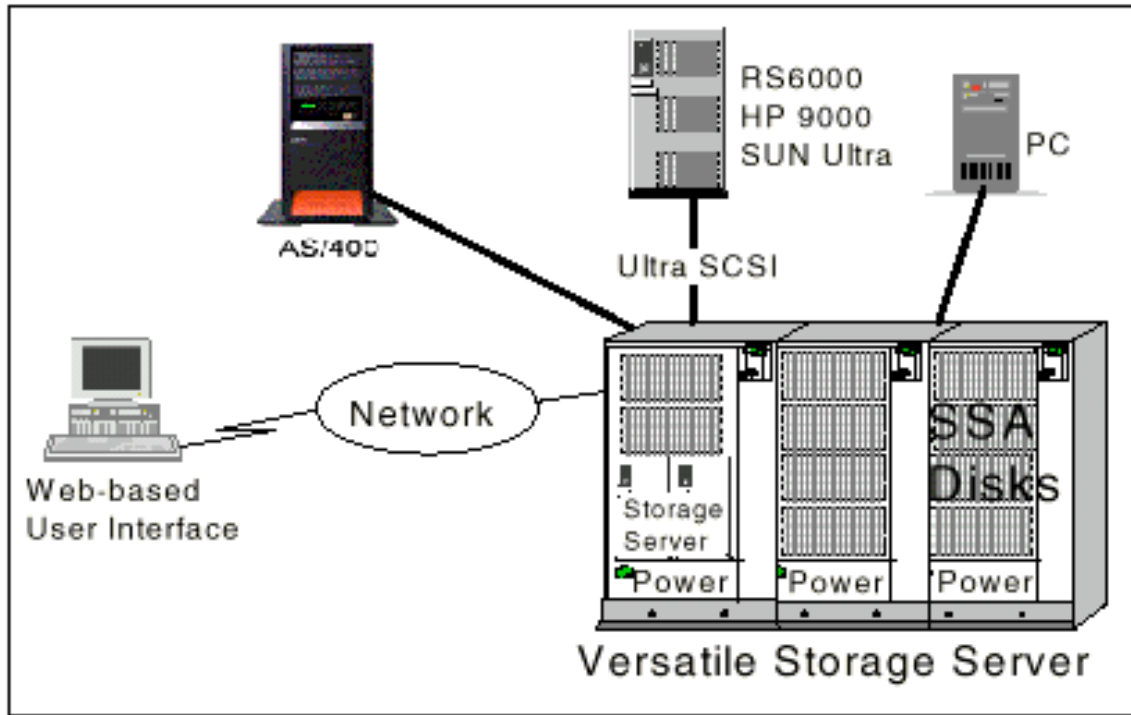
SCSI (parallel SCSI)

Fibre Channel (Serial SCSI)

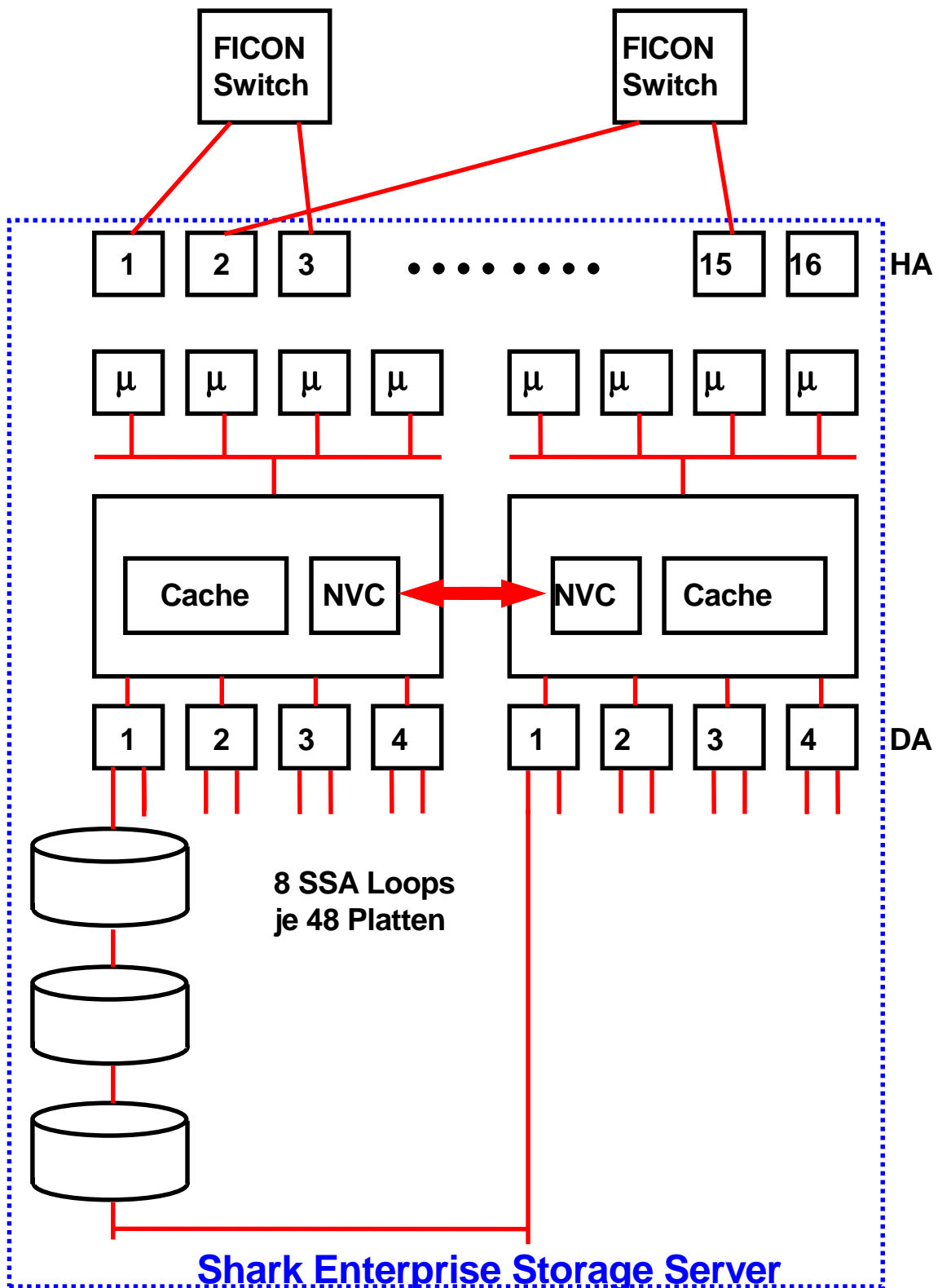
SSA (Serial Storage Architecture)



RAID, Cache Funktionalität



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 Adaptern, 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)

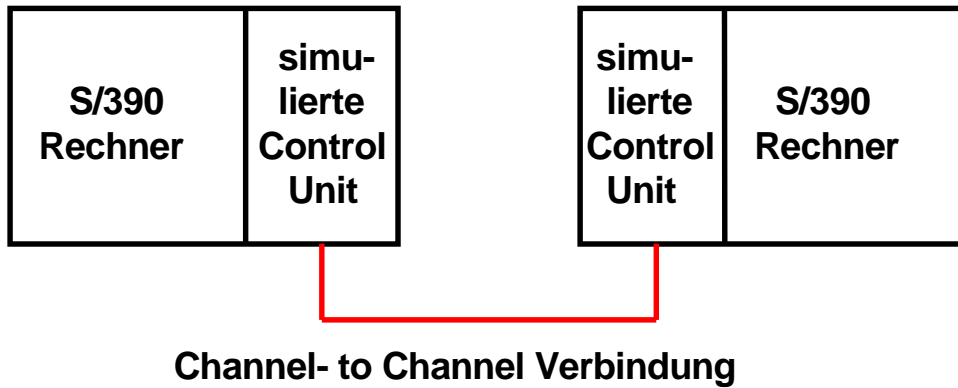
bis zu 11,2 Tbyte / ESS (2001)

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.

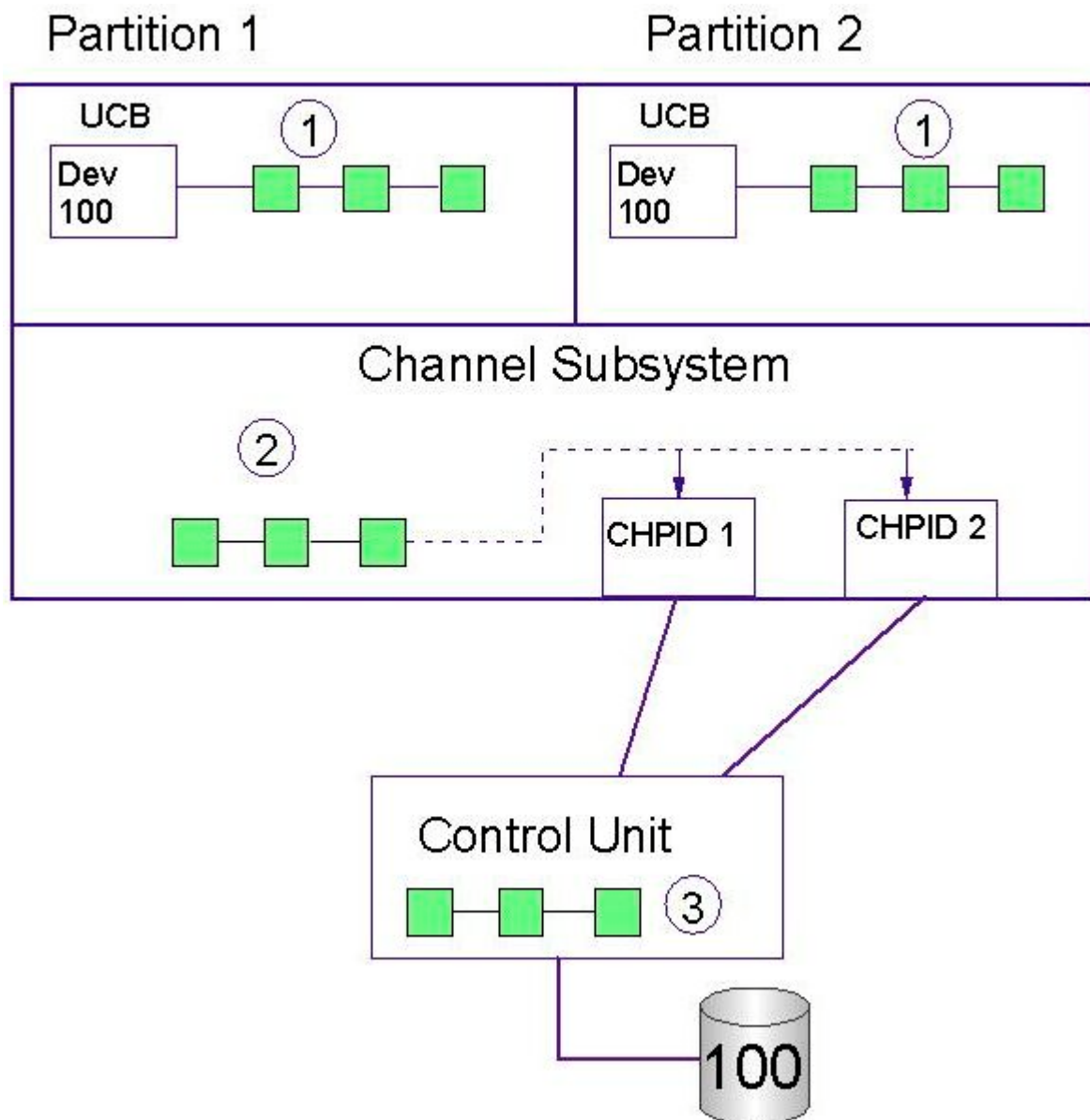
Enterprise Storage Server werden von vielen Firmen angeboten, meistens sowohl mit FC-SCSI als auch mit FICON Anschlußmöglichkeiten: EMC, Hitachi/Sun, MaxData, andere.

CTC Verbindung (Channel- to Channel)



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.



Queuing Points

IOS UCB Queue

In CSS waiting for channel

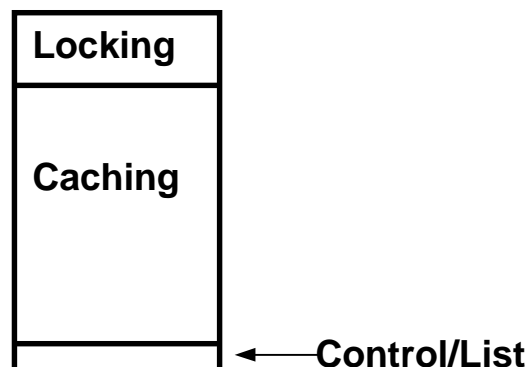
In Shark control unit

Coupling Facility

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

- Locking
- Caching
- Control/List Structure Management

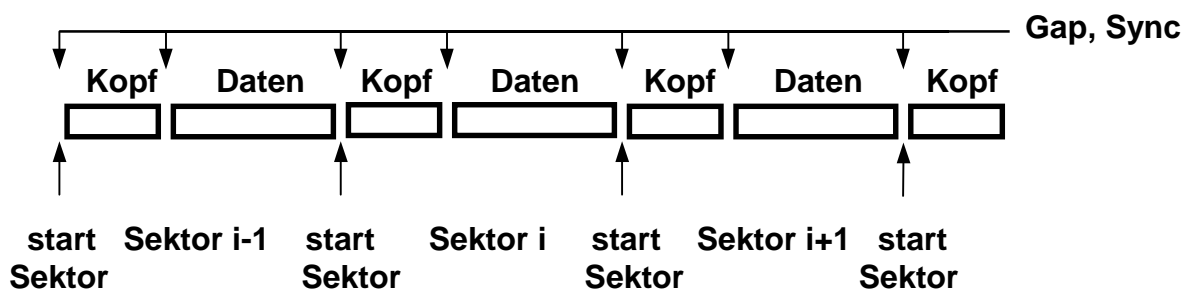
Der größte Teil des Coupling Facility Hauptspeichers wird für das caching von Plattenspeicherdaten eingesetzt.



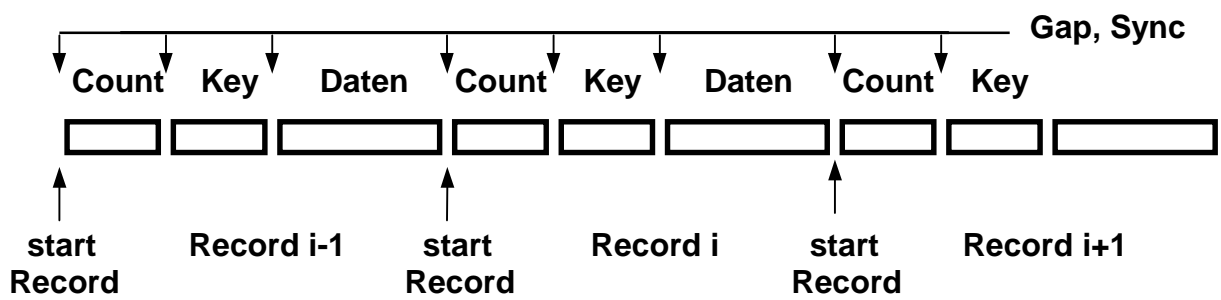
Die Coupling Facility ist über Glasfaser Verbindungen mit einem optimierten Protokoll mit den Rechnern des Sysplex verbunden.

IBM 9037 Sysplex Timer provides a common time reference to all OS/390 images. Enables proper sequencing and time stamping of updates to shared data bases (critical to recovery of shared data).

S/390 Coupling Facility runs Coupling Facility Control Code



Fixed Block Record Format



Count, Key, Data (CKD) Format

CKD (Count, Key, Data) Format

Eine Spur besteht aus physikalischen Records (unterschiedlicher Länge)

Ein Record ist die Dateneinheit, die durch einen Lese/Schreibbefehl zwischen Hauptspeicher und Plattenspeicher bewegt wird

Im einfachsten Fall entspricht ein physikalischer Record einem logischen Record (Datensatz, z.B. einer Struktur in C)

Records können unterschiedlich lang sein. Das Count Feld eines CKD Records gibt an, wie lang das Datenfeld ist.

Late Binding: Alle Dateien werden zum Zeitpunkt der Benutzung spezifiziert

Dateisystem

Ein Dateisystem verbirgt die Eigenschaften der physikalischen Datenträger (Plattenspeicher, CD, evt. Bandlaufwerke) weitestgehend vor dem Anwendungsprogrammierer.

Unix, NT

Dateien sind strukturlose Zeichenketten, welche über Namen identifiziert werden. Hierfür dienen Dateiverzeichnisse, die selbst wie Dateien aussehen und behandelt werden können.

Dateien werden sequentiell gelesen. Ein Direktzugriff wird mit Hilfe von Funktionen programmiert, die gezielt auf ein bestimmtes Zeichen in der Datei vor- oder zurücksetzen.

OS/390

Das Dateisystem (Filesystem) wird durch die Formattierung eines physikalischen Datenträgers definiert. Bei der Formattierung der Festplatte wird die Struktur der Datei festgelegt. Es gibt unterschiedliche Formattierungen für Dateien mit direktem Zugriff (DAM), sequentiellen Zugriff (SAM) oder index-sequentiellen Zugriff (VSAM).

Dateizugriffe benutzen an Stelle eines Dateiverzeichnisse „Kontrollblöcke“, welche die Datenbasis für unterschiedliche Betriebssystemfunktionen bilden.

Dateiverwaltung

Dateien müssen unter OS/ 390 bezüglich Größe und physikalischem Adressenbereich manuell eingerichtet werden.

Unter Linux oder NT muß bezüglich physikalischer Zuordnung nur die Partition (c:, d:, e:) angegeben werden. Wo innerhalb der Partition die Datei zu liegen kommt, bestimmt das Betriebssystem. Die Platzverwaltung erfolgt automatisch. Nachteil der dynamischen Speicherplatzverwaltung:

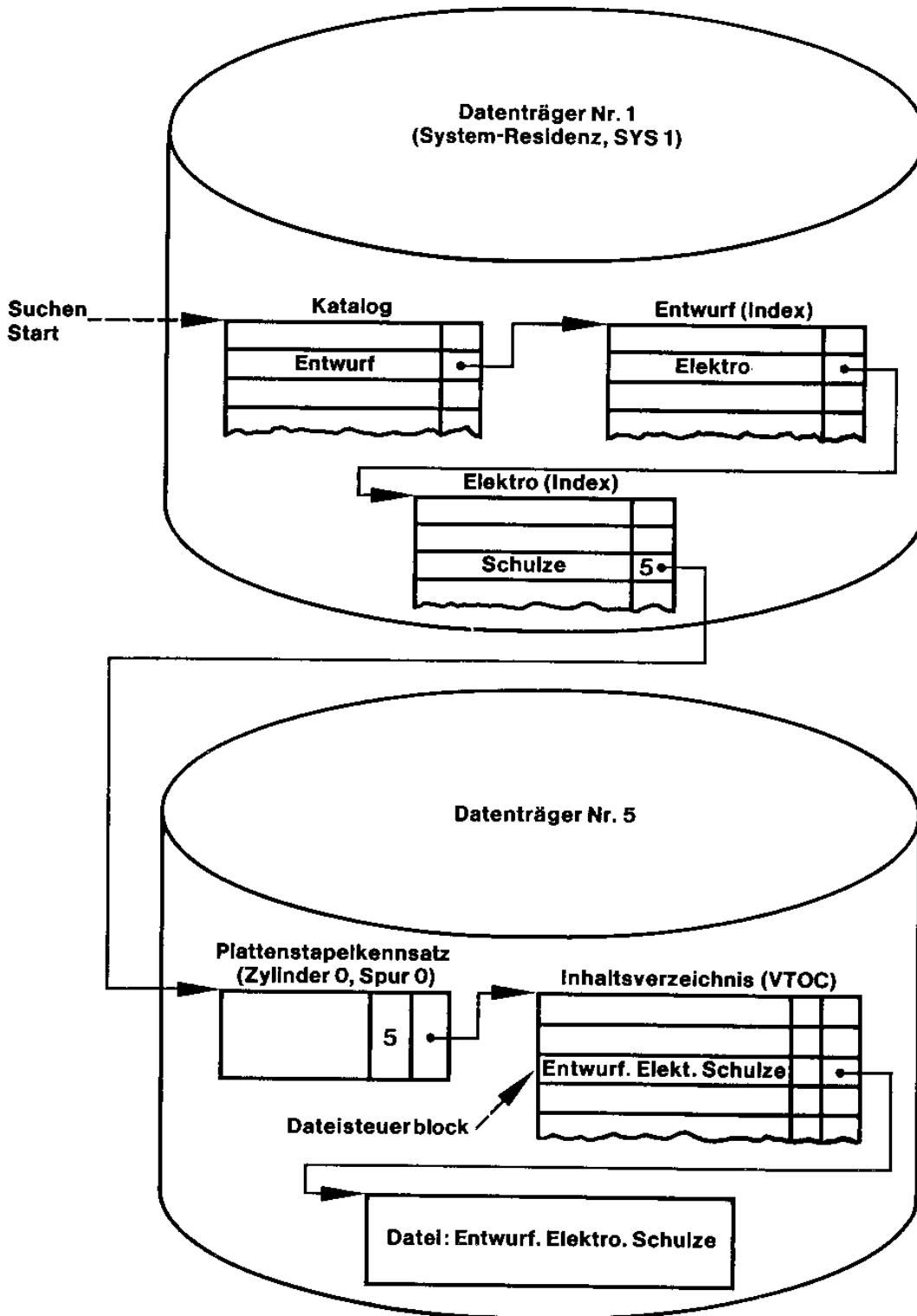
- **Fragmentierung**
- **Nicht-optimale Speicherplatzzuordnung**
(Zugriffseigenschaften des Lese/Schreibkopfes nicht ausgenutzt)

Unter OS/390 ordnet der Systemadministrator manuell allen Dateien einen in der Größe und Anordnung vorgegebenen Platz zu. Der Platz für eine Datei enthält Reserven für ein späteres Wachstum.

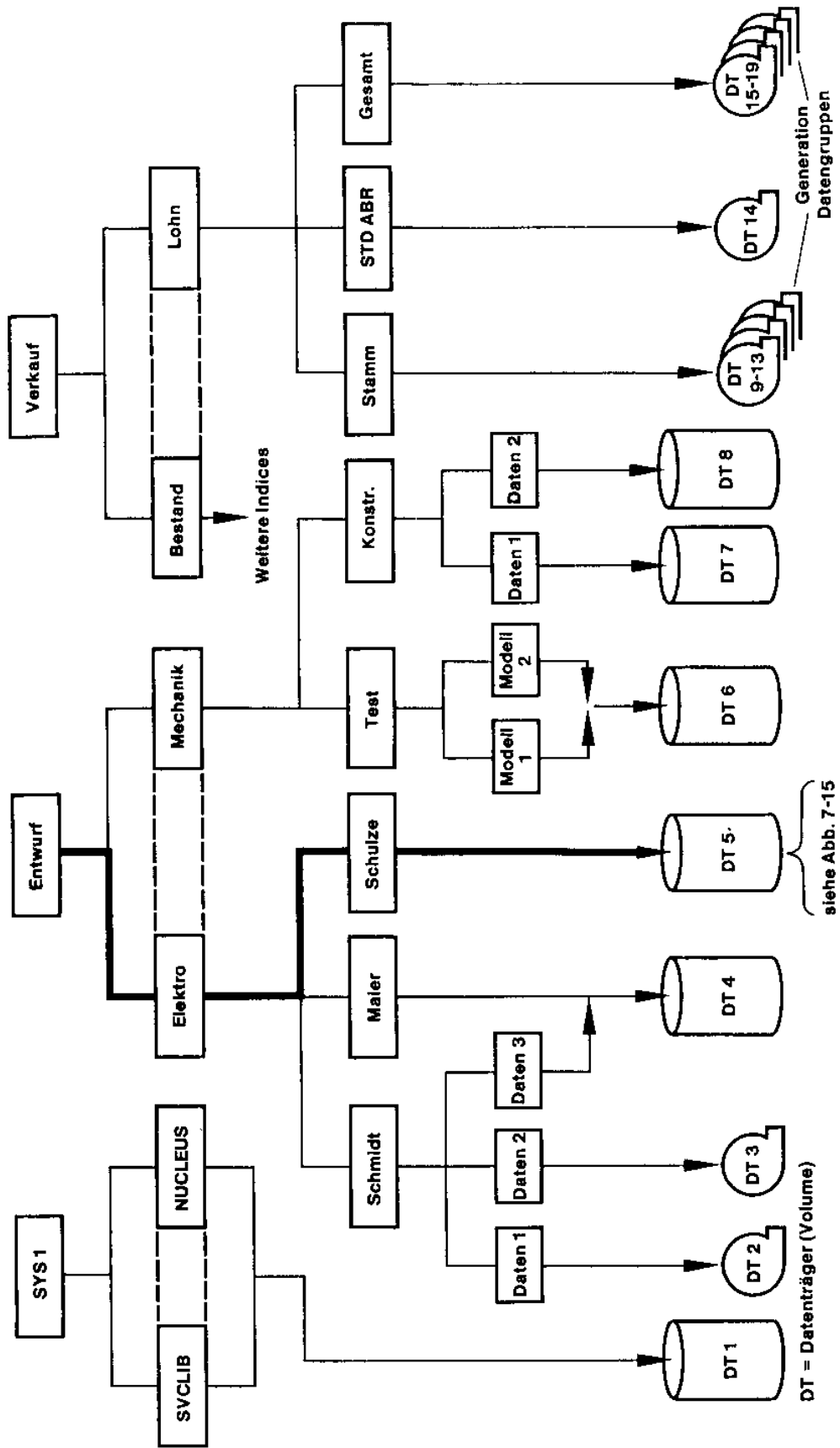
Wird die Reserve erschöpft, so kann als Notbehelf ein Block zusätzlichen Plattenspeicherplatzes in der Form eines „Extends“ angehängt werden. In diesem Fall entstehen ebenfalls Fragmentierungsprobleme, allerdings nicht so schwerwiegend wie bei der automatischen Speicherplatzverwaltung. Dennoch muß der Speicherplatz von Zeit zu Zeit neu zugeordnet werden.

Vorteil des OS/390 Verfahrens: Die manuelle Zuordnung auf der Plattenspeicherfläche erlaubt Optimierung bezüglich Zugriffszeit und Durchsatz. Keine Fragmentierung.

Nachteil: Erhöhter Verwaltungsaufwand.



Anordnung des Katalogs



Katalog-Struktur

Inhaltsverzeichnis

(Directory, Volume Table of Content, VTOC)

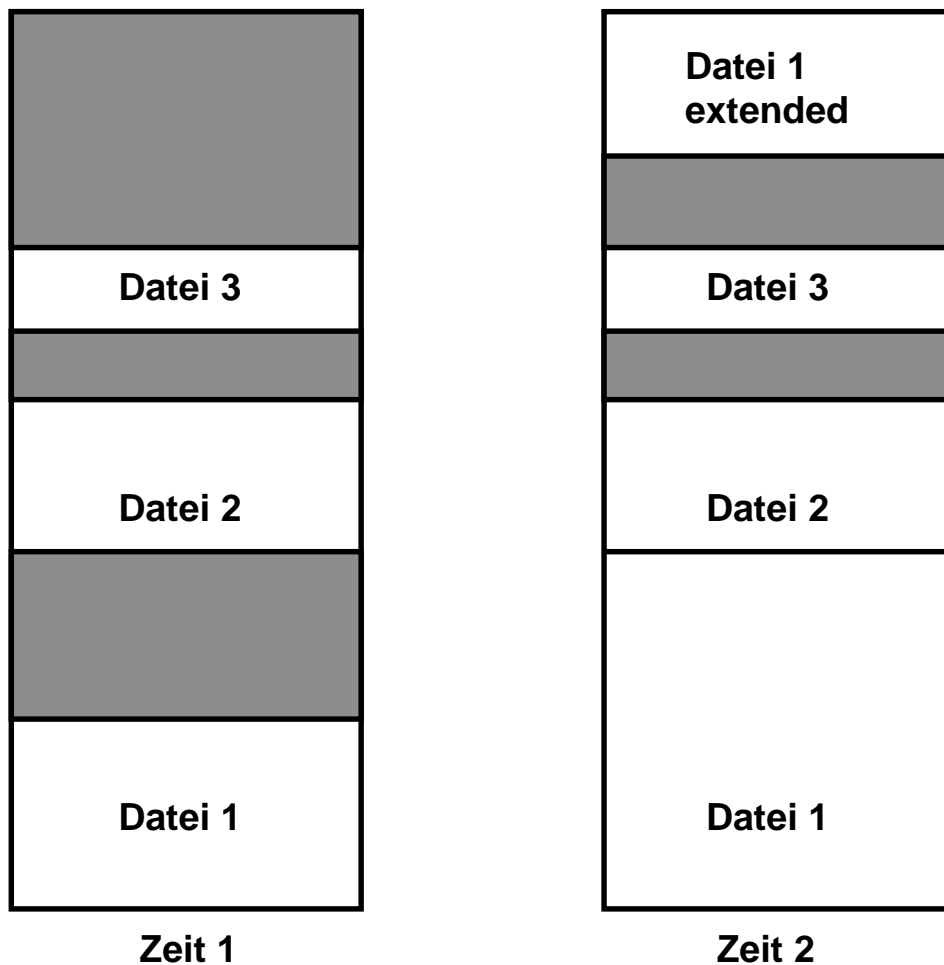
Befindet sich an einer vorgegebenen Adresse auf dem Plattenspeicher.

Enthält eine Serie von Einträgen, je 1 Eintrag (Data Set Control Block) pro Datei auf der Festplatte.

Jeder Eintrag enthält:

**Namen der Datei (z.B. COMMAND.COM in MS-DOS)
Anfangsadresse der Datei auf der Festplatte
Länge der Datei
Datum und Uhrzeit des letzten (Schreib-) Zugriffs
Zusatzinformation, z.B. Datei Attribute**

Bei den verschiedenen Betriebssystemen (selbst des gleichen Herstellers) ist das Prinzip ähnlich, die Detailstruktur jedoch sehr unterschiedlich (andere Anfangsadresse, andersartiger Aufbau).

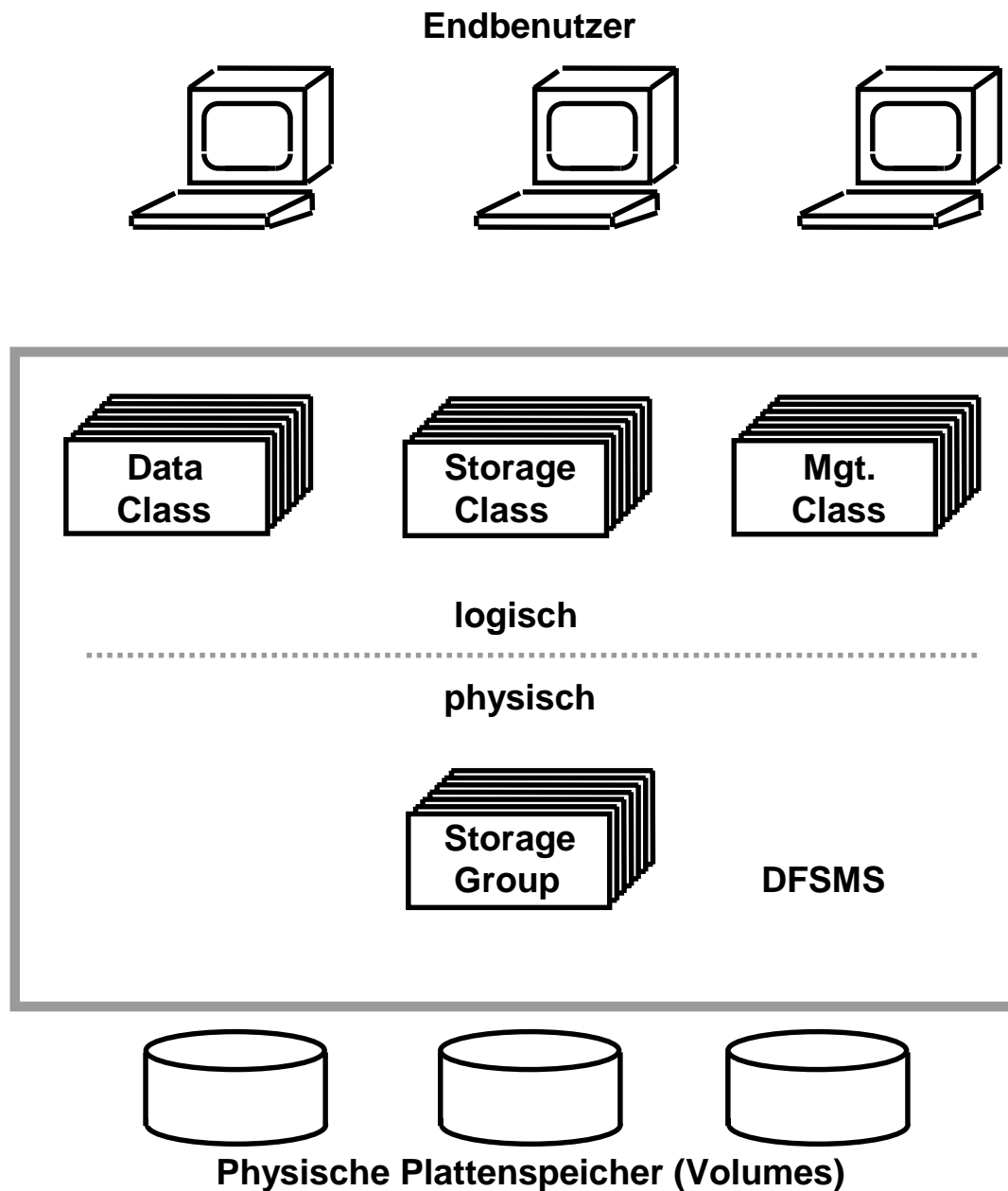


Simple Contiguous

Jeder Datei wird ein fester Platz zugeordnet.

Wächst die Datei über den zugeordneten Platz hinaus, steht ein „Extend“ zur Verfügung.

Extends fragmentieren den Plattenspeicherplatz, jedoch nicht so stark wie bei Unix und Windows File Systemen



Data Facility Storage Management System

Der Benutzer sieht nur die logische Sicht der Daten:

- vorbereitete Datenmodelle in den Datenklassen
- in den Storageklassen festgelegte Serviceanforderungen
- Management Kriterien für die Auslagerung der Daten

DFSMS Klassen

Beim Neuanlegen einer Datei müssen angegeben werden:

Data Class

File System Attribute wie Record Format, Record Länge, Schlüssellänge bei VSAM Dateien

Storage Class

Performance Angaben wie Antwortzeit (ms), Nutzung von Read/Write Caching, Verwendung der Dual Copy Funktion

Management Class

Migration nach x Tagen, Anzahl Backup Versionen, schrittweiser Abbau der Backup Versionen, Löschen der Daten

DFSMS bewirkt die Zuordnung der Datei zu einer Storage Group, einer Gruppe von Plattenspeichern (Volumes)

Lebenszyklus einer Datei

Anlegen der Datei durch den Benutzer

Benutzung (Schreiben und/oder Lesen der Daten)

Sicherungskopien anlegen

Platzverwaltung (freigeben / erweitern / komprimieren)

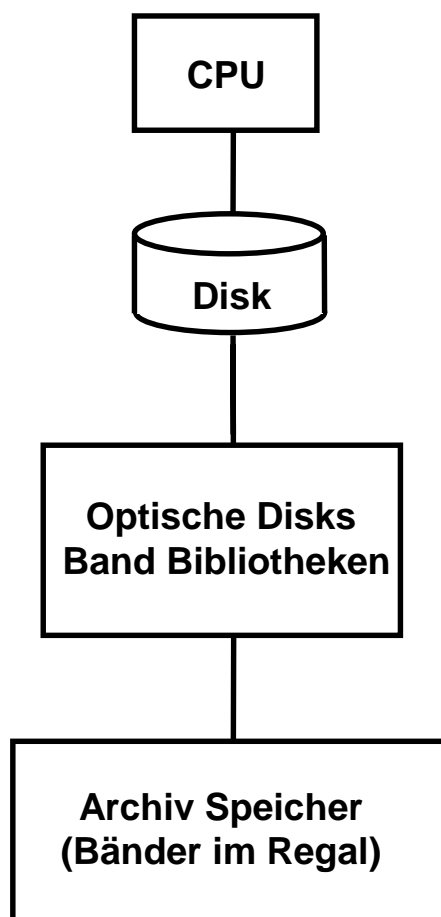
Auslagern von inaktiven Dateien, sowie Wiederbenutzung

Ausmustern von Sicherungskopien

Wiederherstellung von Dateien

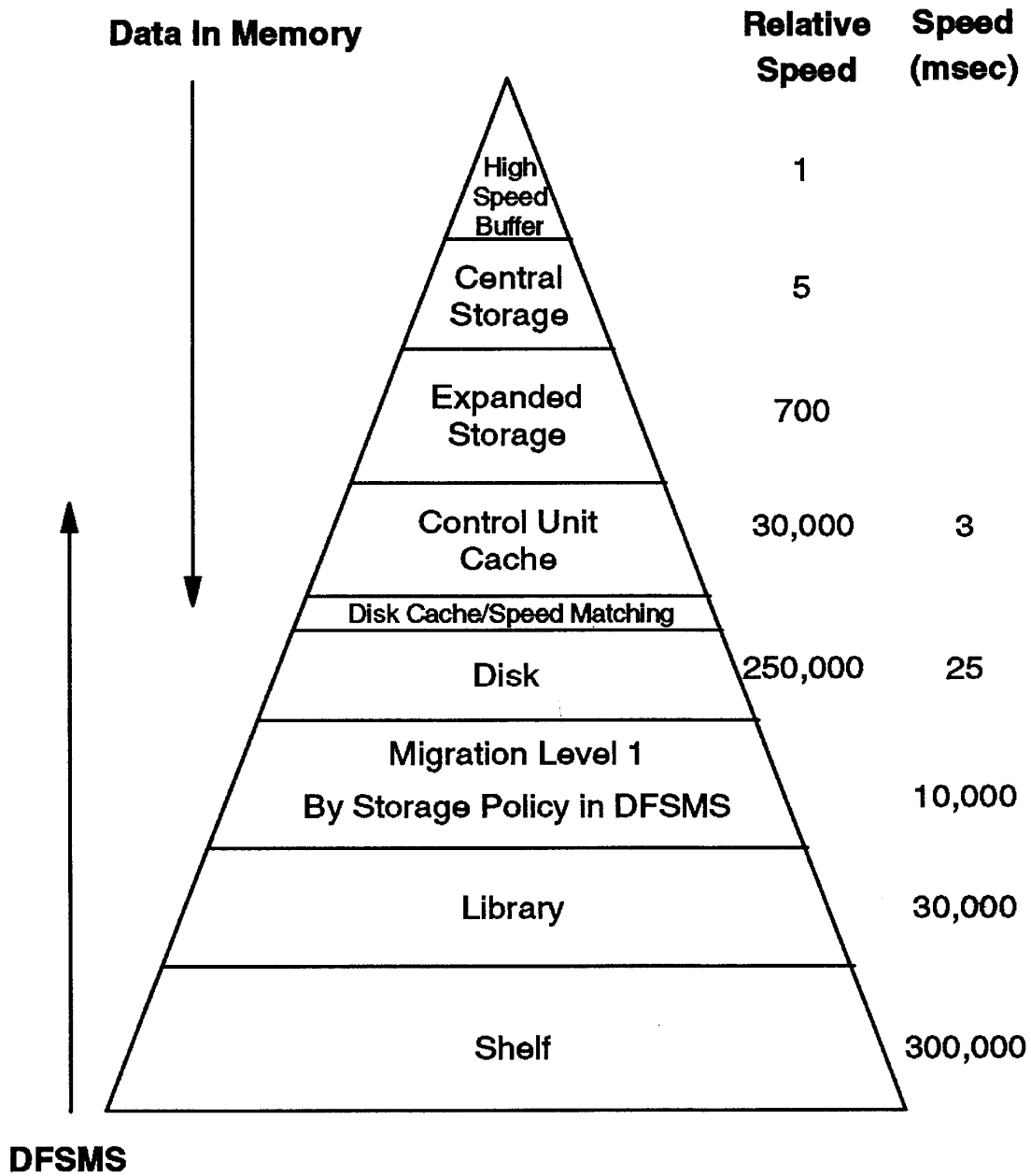
Löschen der Datei

Hierarchical Storage Management



DFSMS (Data Facility Storage Management Subsystem) enthält eine „Hierarchical Storage Management“ (HSM) Komponente. Diese arbeitet mit einem Regelwerk, wonach selten gebrauchte Daten automatisch vom Plattenspeicher auf Archivspeicher migriert werden. Im Bedarfsfall holt DFSMS die Daten automatisch wieder zurück.

Hierarchical Storage Management wird bereits während der Erstellung einer neuen Datei eingesetzt, um Policies für das Leistungsverhalten, Backup, Migration und Cache Eigenschaften (immer, nie cached) festzulegen.



System Managed Storage