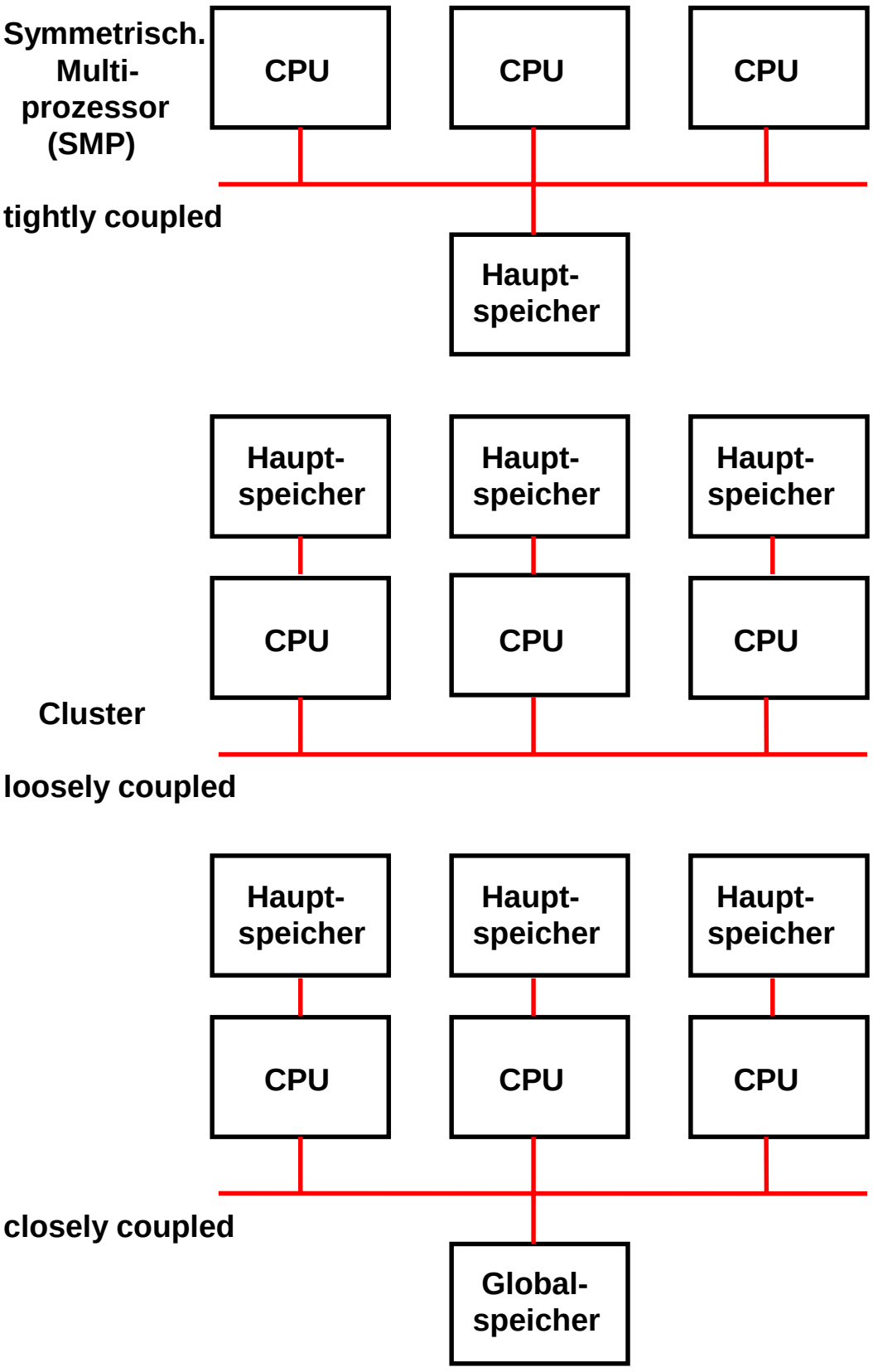
Einführung in z/OS und OS/390

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

WS 2006/2007

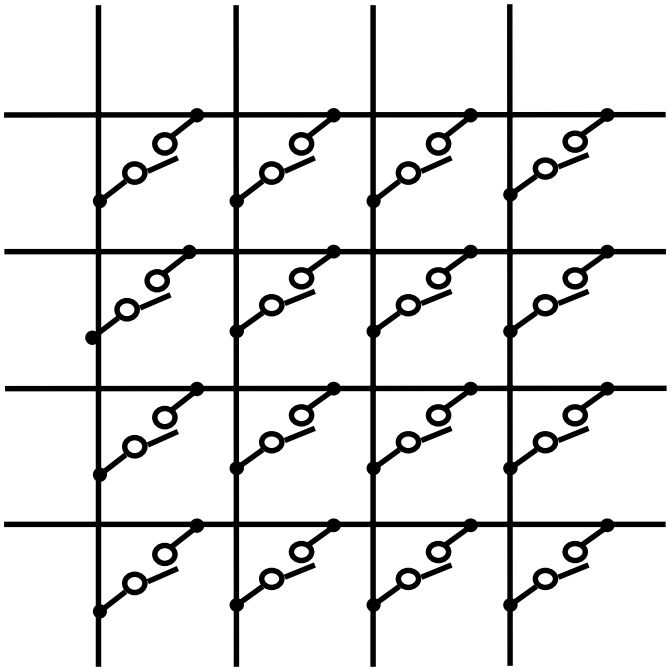
Teil 5

Parallelrechner



Taxonomie von MIMD Parallelrechnern

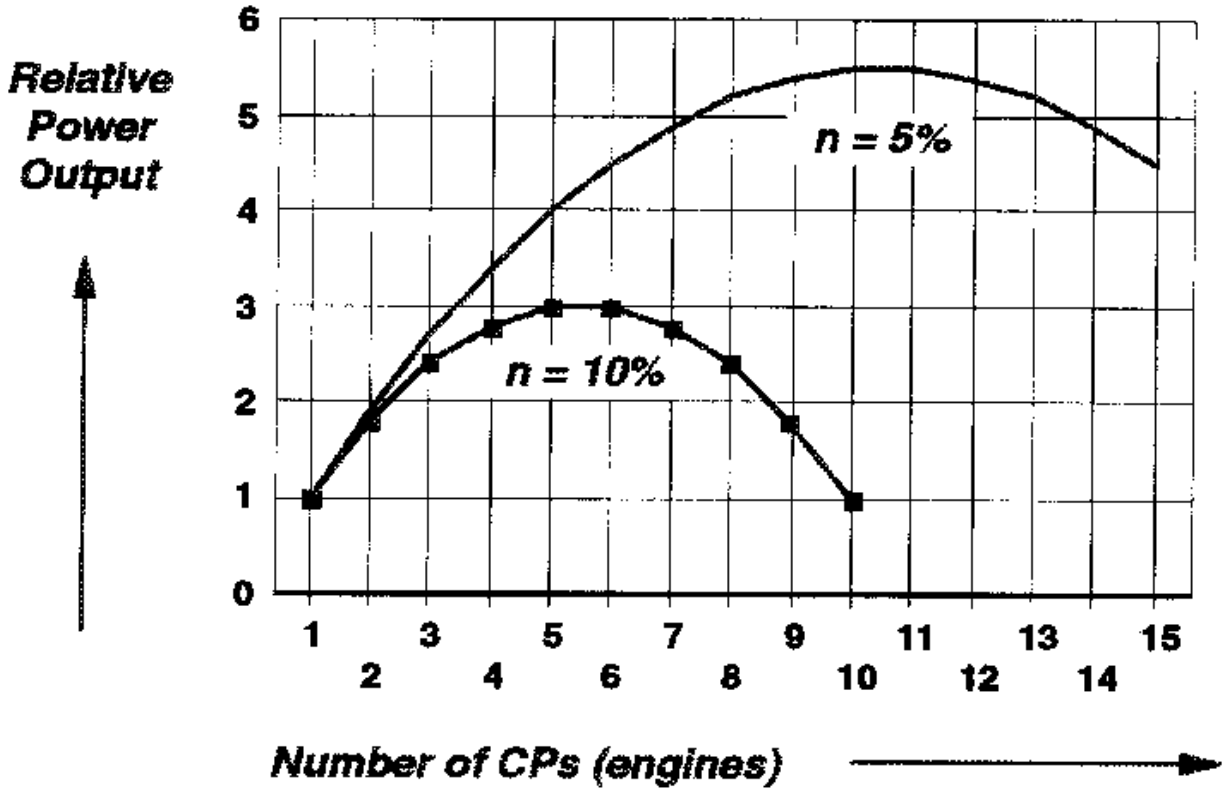
vier Eingänge



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.

Angenommen m = Anzahl CPUs. Der Leistungsabfall pro CPU ist

$$\text{Verlust pro CPU} = n(m-1)$$

Bei einem z9 Rechner mit z/OS ist n << 2 %. Es ist sinnvoll, einen SMP mit bis zu 32 Prozessoren einzusetzen.

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

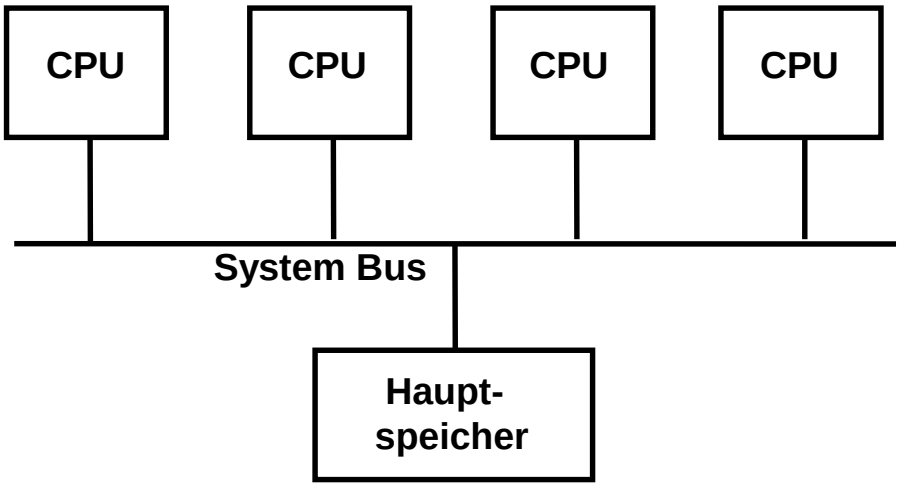
Berücksichtigt

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

Benchmark besteht aus S/390 Maschinenbefehlen, daher

nicht portierbar auf Rechner anderer Architektur

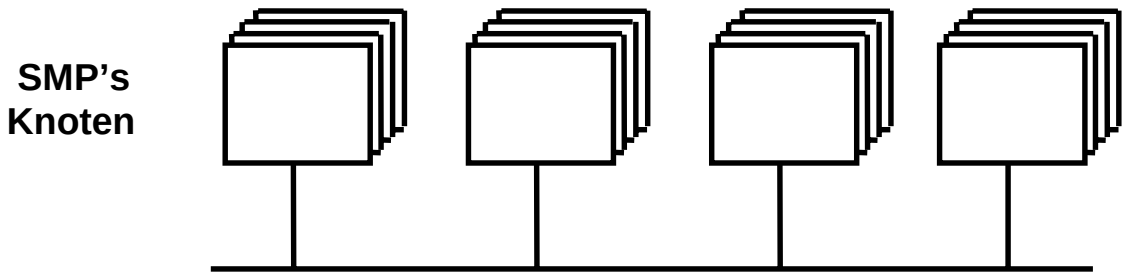
- **(Problem der *„S/390 äquivalenten MIPS*)**



SMP, Prozessor Knoten

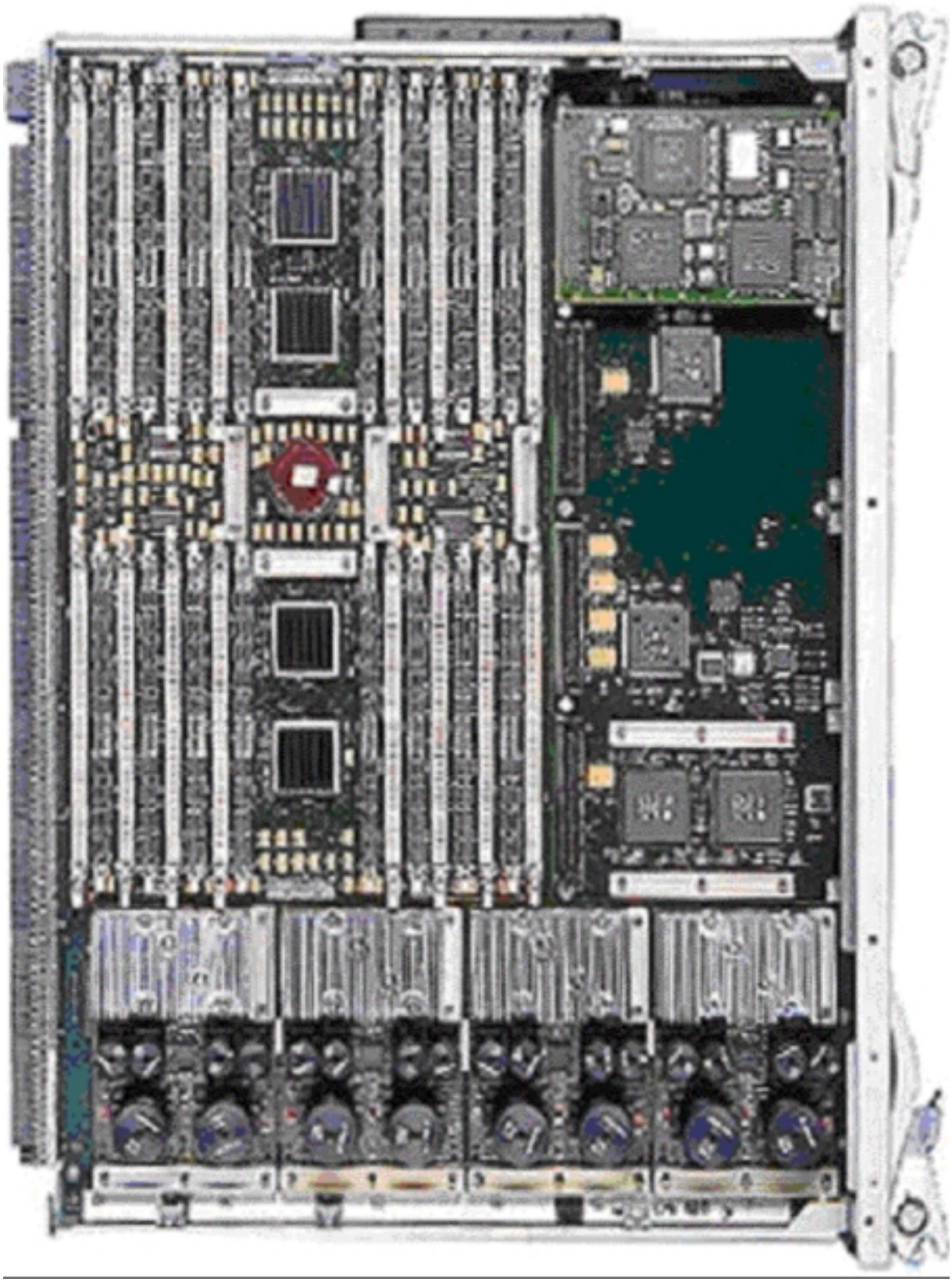
Ein SMP (Symmetric Multiprocessor, Prozessor Knoten, Node) besteht aus mehreren CPU´s, die auf einen gemeinsamen Hauptspeicher zugreifen

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

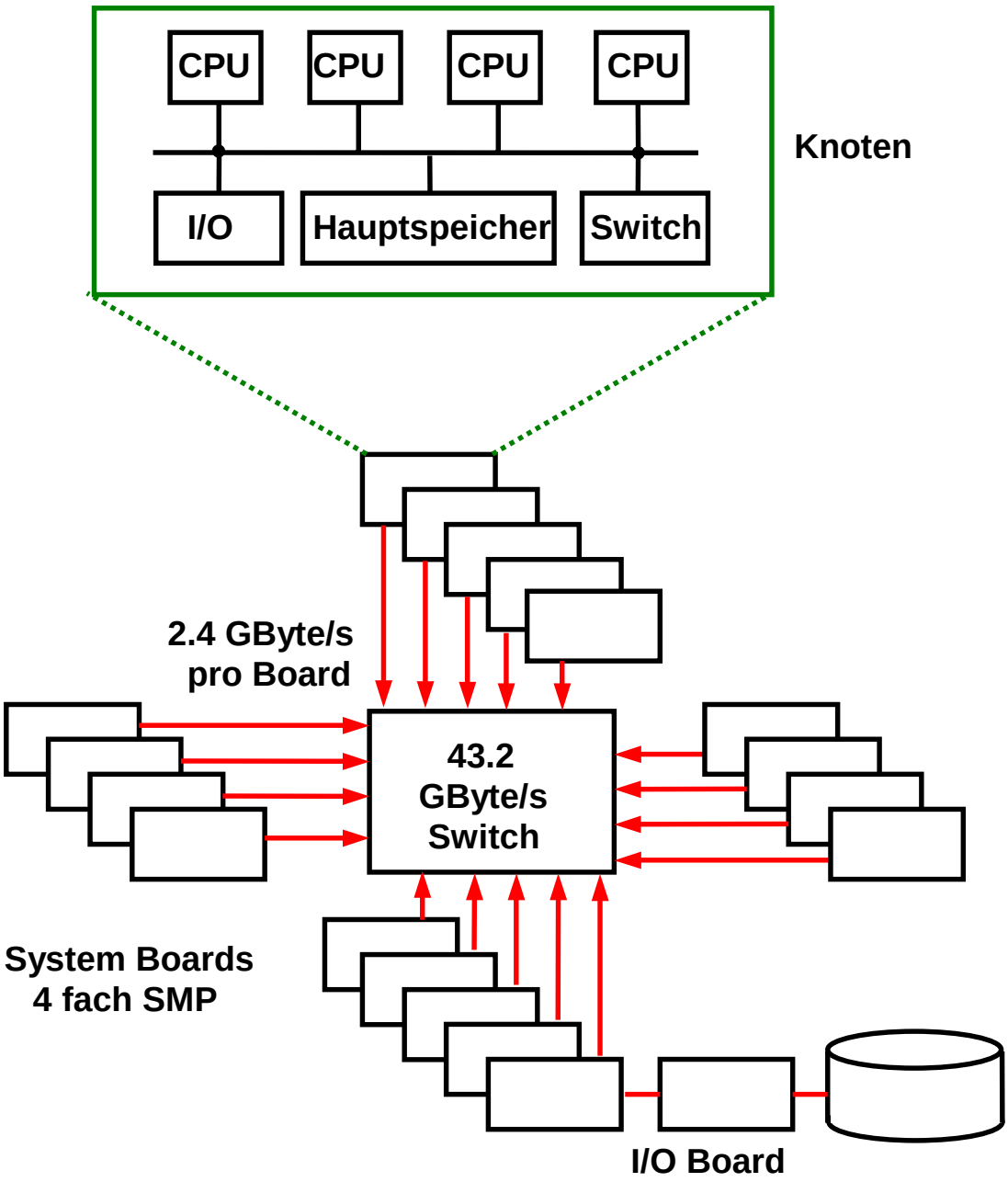


Cluster

Bei einem Cluster werden mehrere SMP's (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 E 10 000 System Board

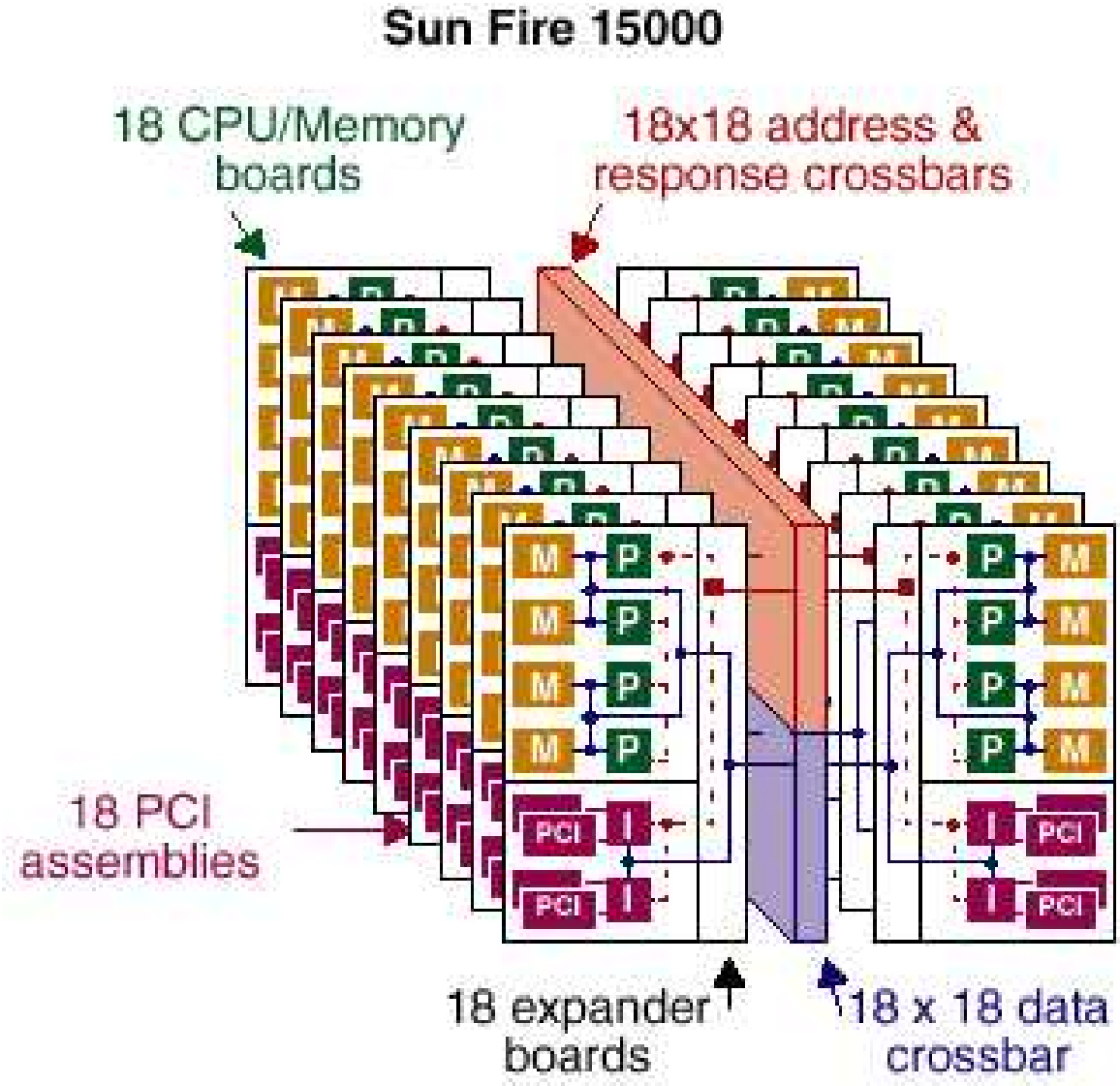


Sun E15K

72 CPU´s

18 System Boards, je 4 CPU/System Board

I/O Controller auf jedem System Board

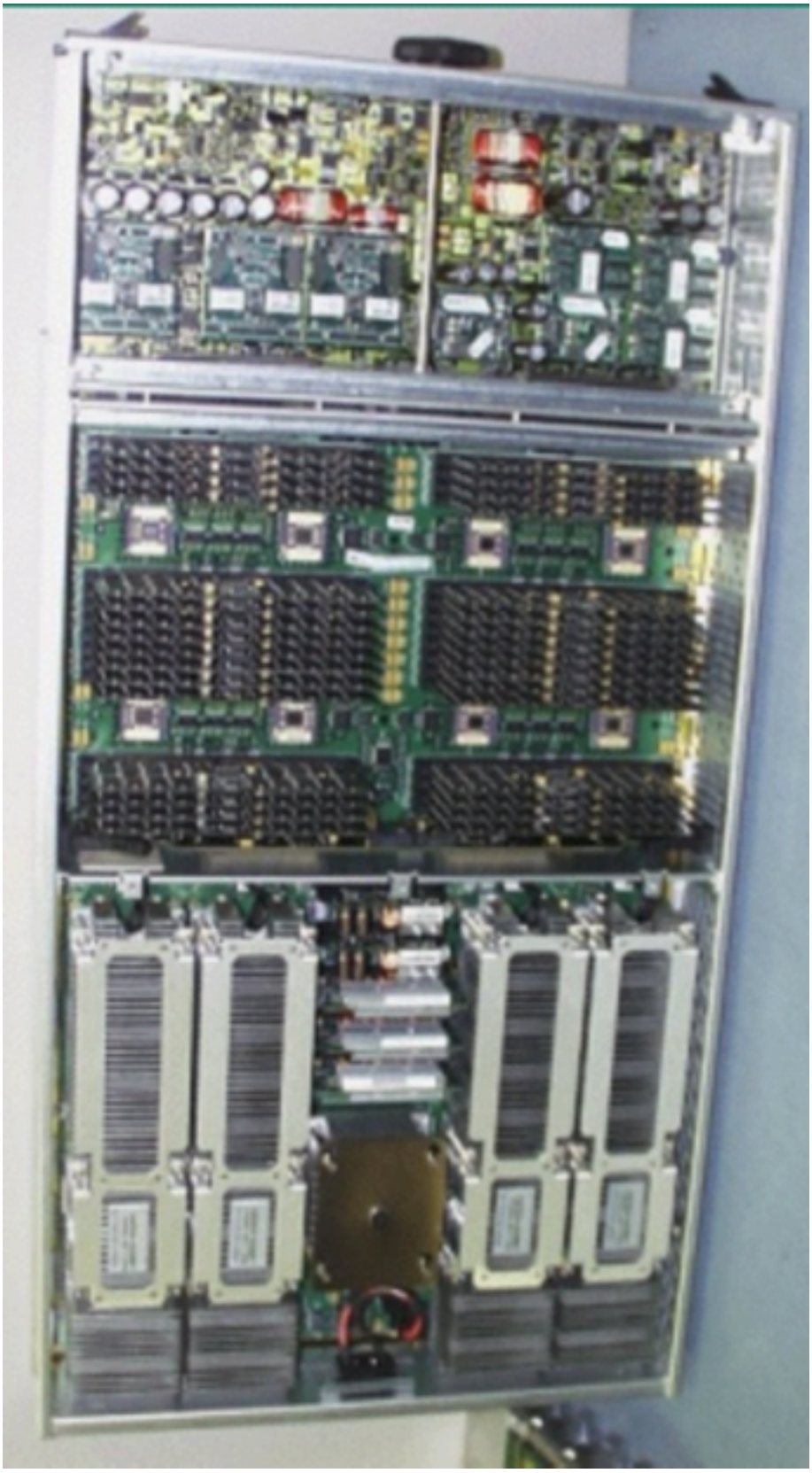


Sun Fire 15000 System hat max 576 Gbyte Hauptspeicher, max 18 CPU/Memory Boards, max 18 Domains, max 18 I/O Boards, max 72 PCI Slots für 72 PCI Karten.

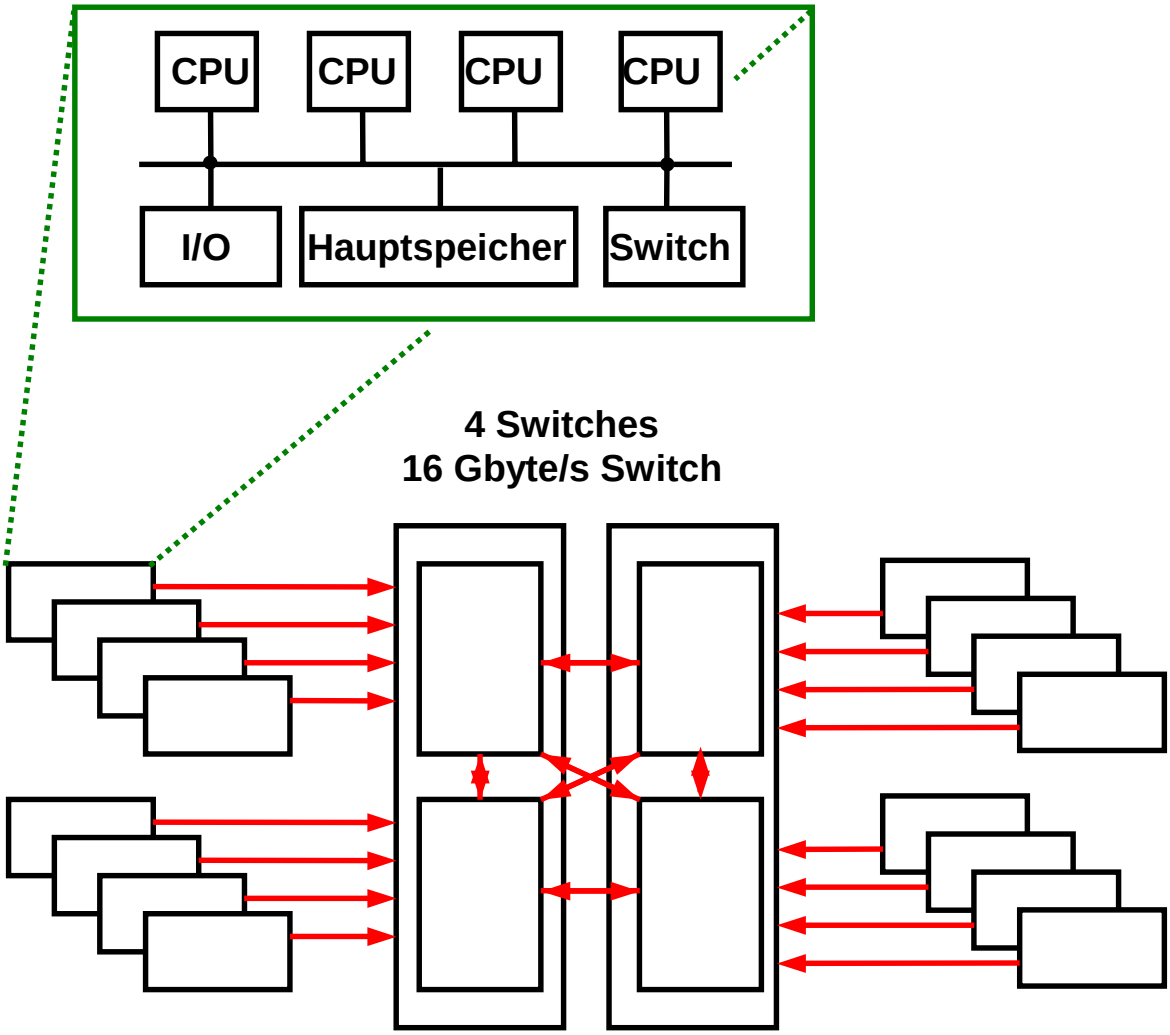
„Board Set“ besteht aus Slot 0 Board, Slot 1 Board und Expander Board. Letzteres nimmt die beiden anderen Boards auf.

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.



HP Superdome Cell Board



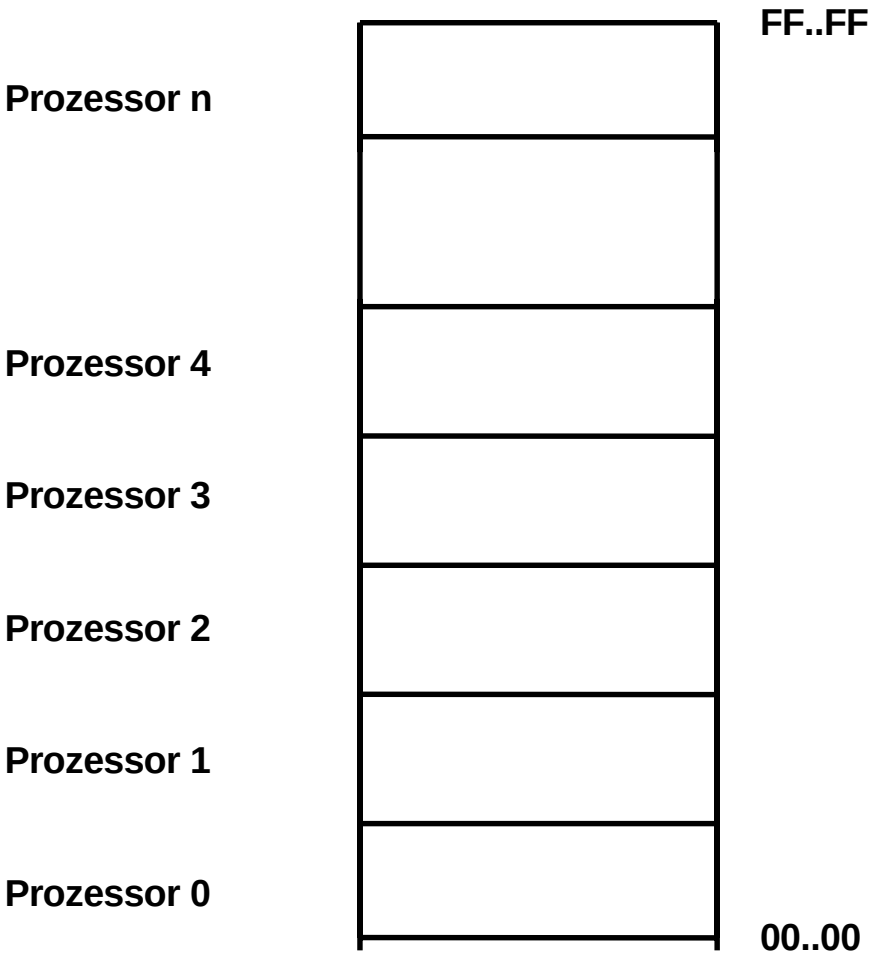
HP Superdome Cluster

64 CPU´s

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

I/O Anschluß auf jedem Cell Board

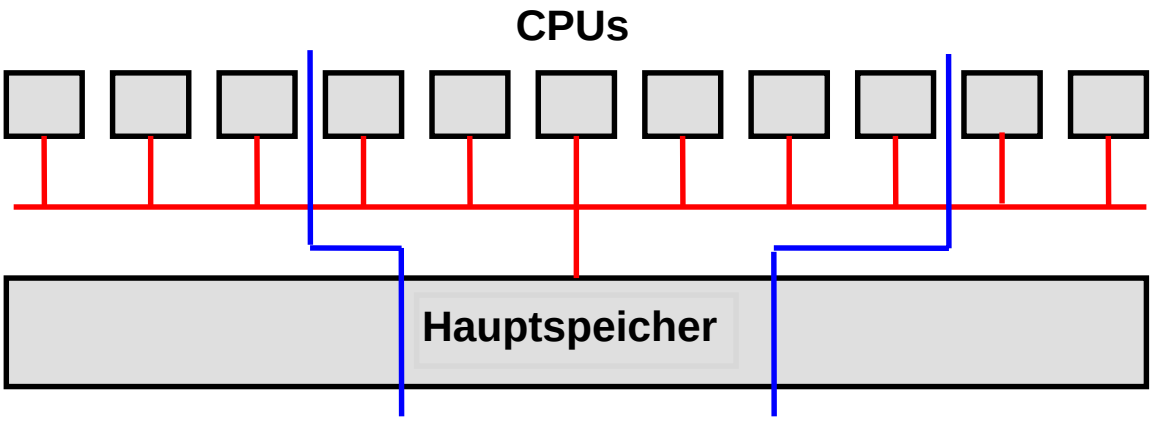
http://www.serverworldmagazine.com/webpapers/2001/05_hpsuperdome.shtml



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.



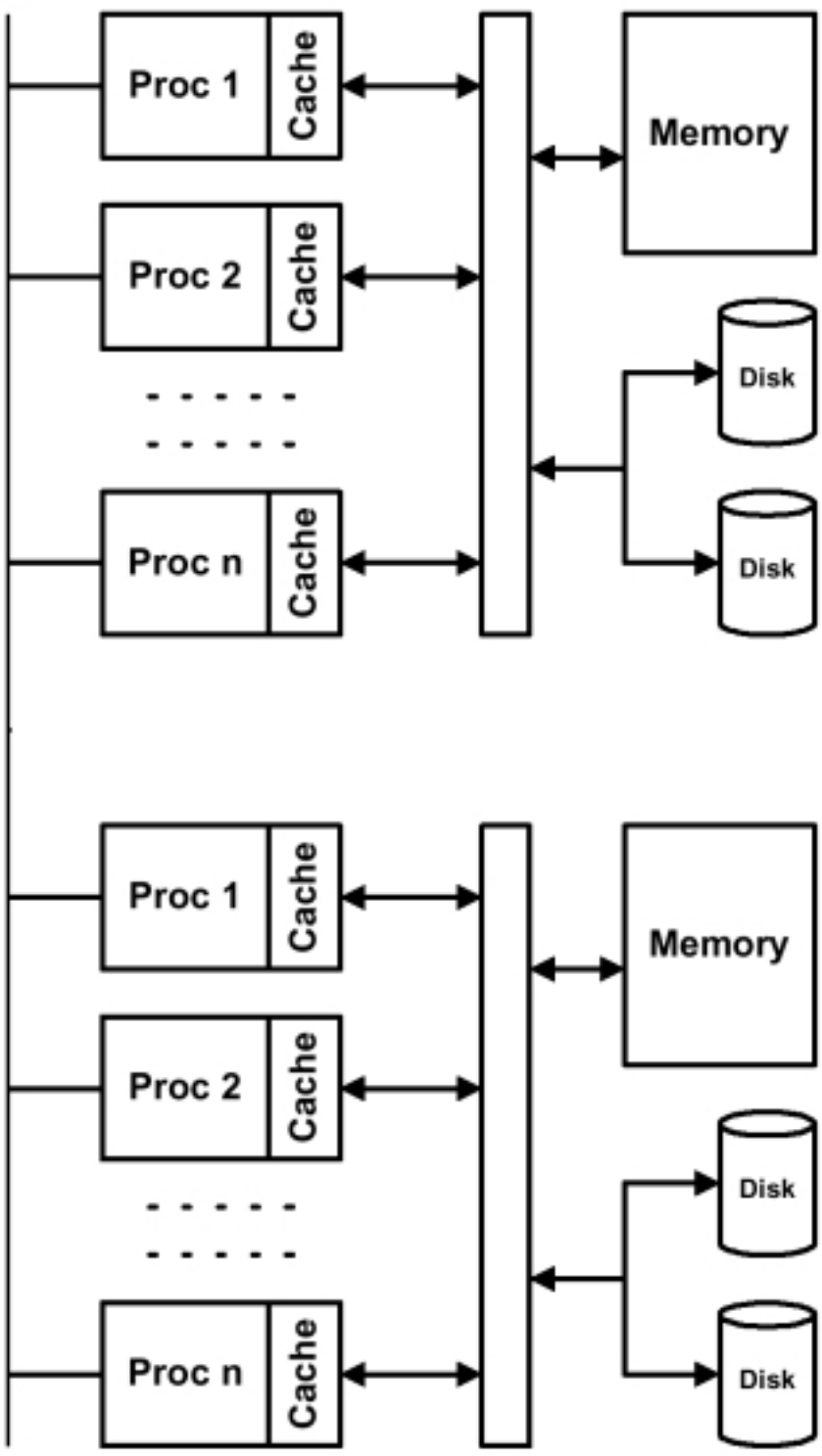
Aufteilung eines Großrechners in mehrere SMPs

z/OS unterstützt symmetrische Multiprozessoren (SMP) mit bis zu 24 CPUs. Bei Unix, Linux und Windows Betriebs-systemen liegt die Grenze für Transaktions- und Datenbankanwendungen eher bei 12 CPUs.

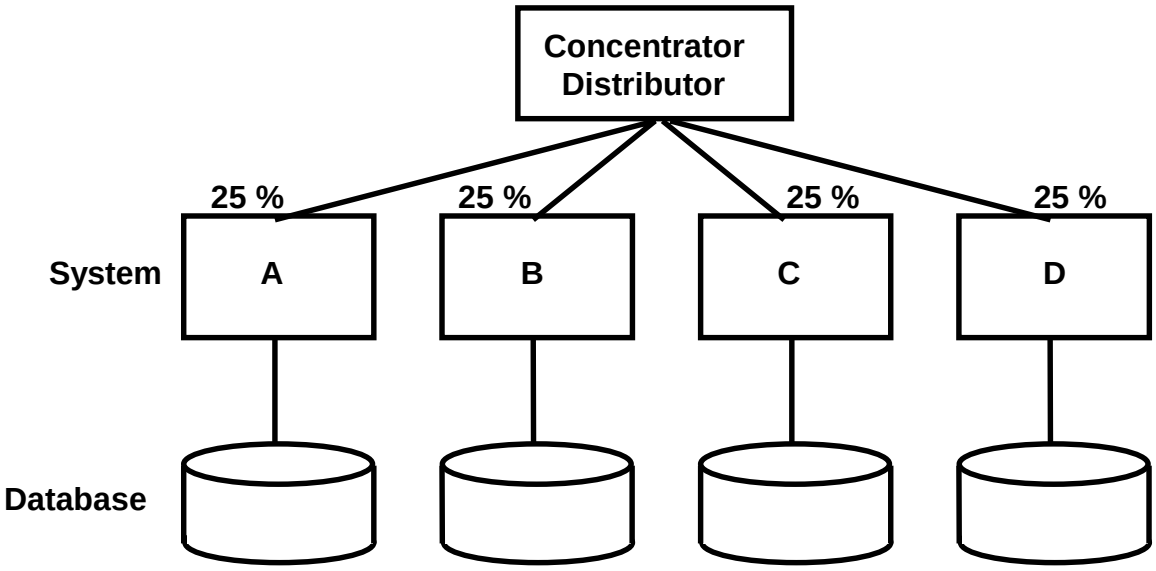
Moderne Großrechner (Systeme) verfügen über wesentlich mehr CPUs. Sie werden deshalb in mehrere SMPs aufgeteilt, die über einen zentralen Switch miteinander kommunizieren.

Der Systemadministrator kann den gesamten Hauptspeicher in unterschiedlichen Größen auf die einzelnen Hauptspeicher aufteilen.

Bei den Sunfire und HP Superdome Rechnern ist die Granularität der SMPs jeweils 4, 8 oder 12 CPUs. zSeries und z/OS erlauben eine beliebig kleine Granularität

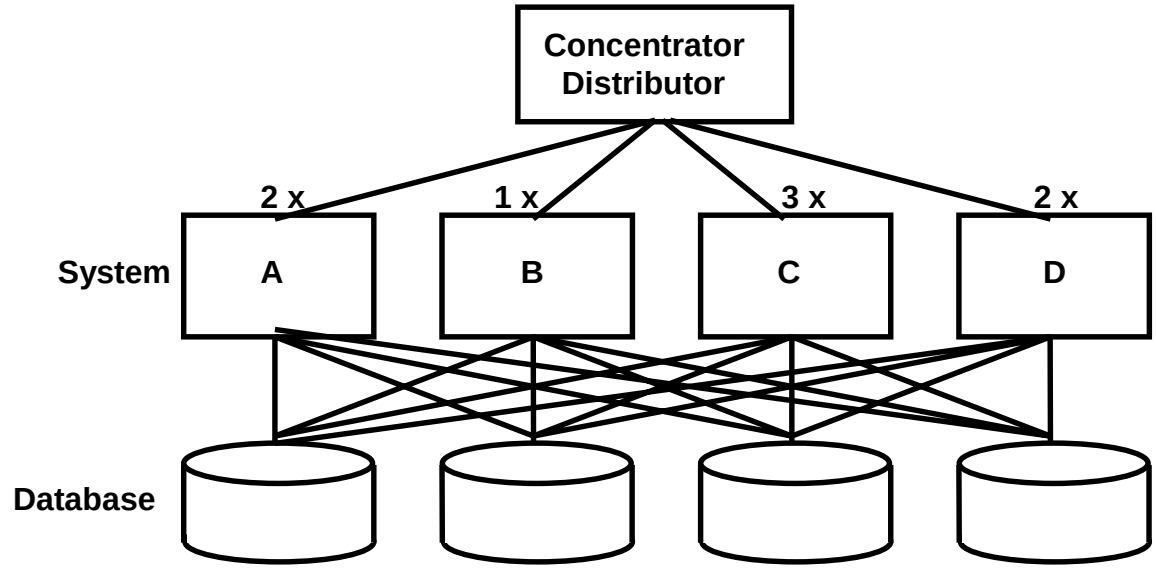


Shared Nothing Modell



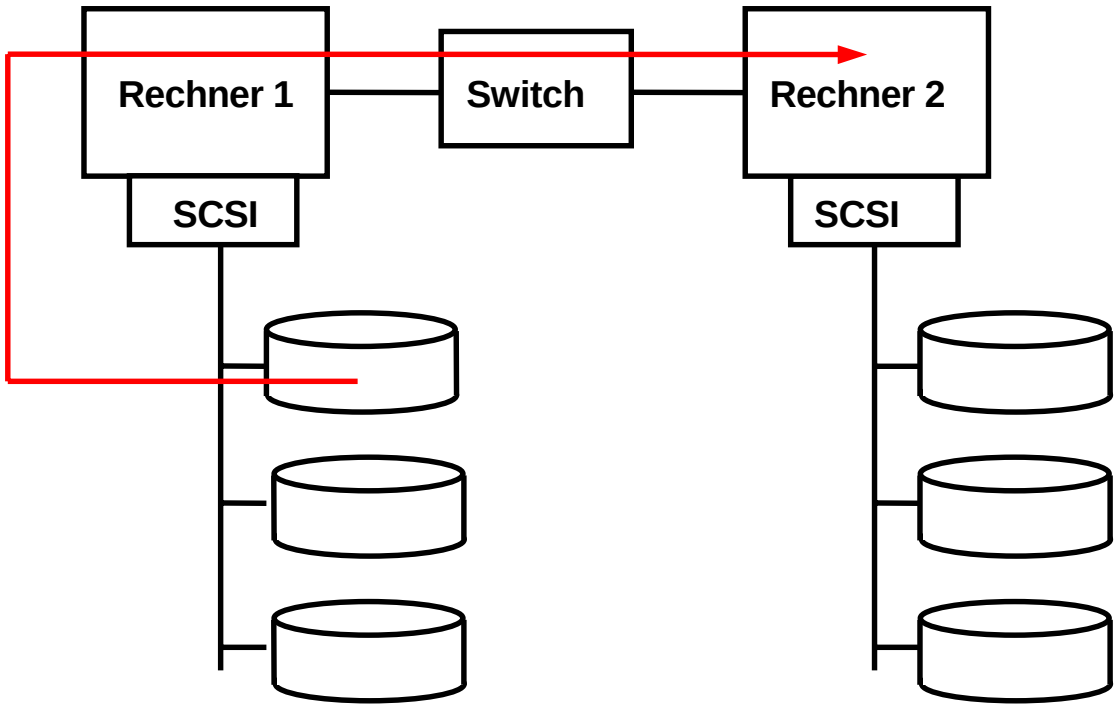
Shared nothing (partitioned data)

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



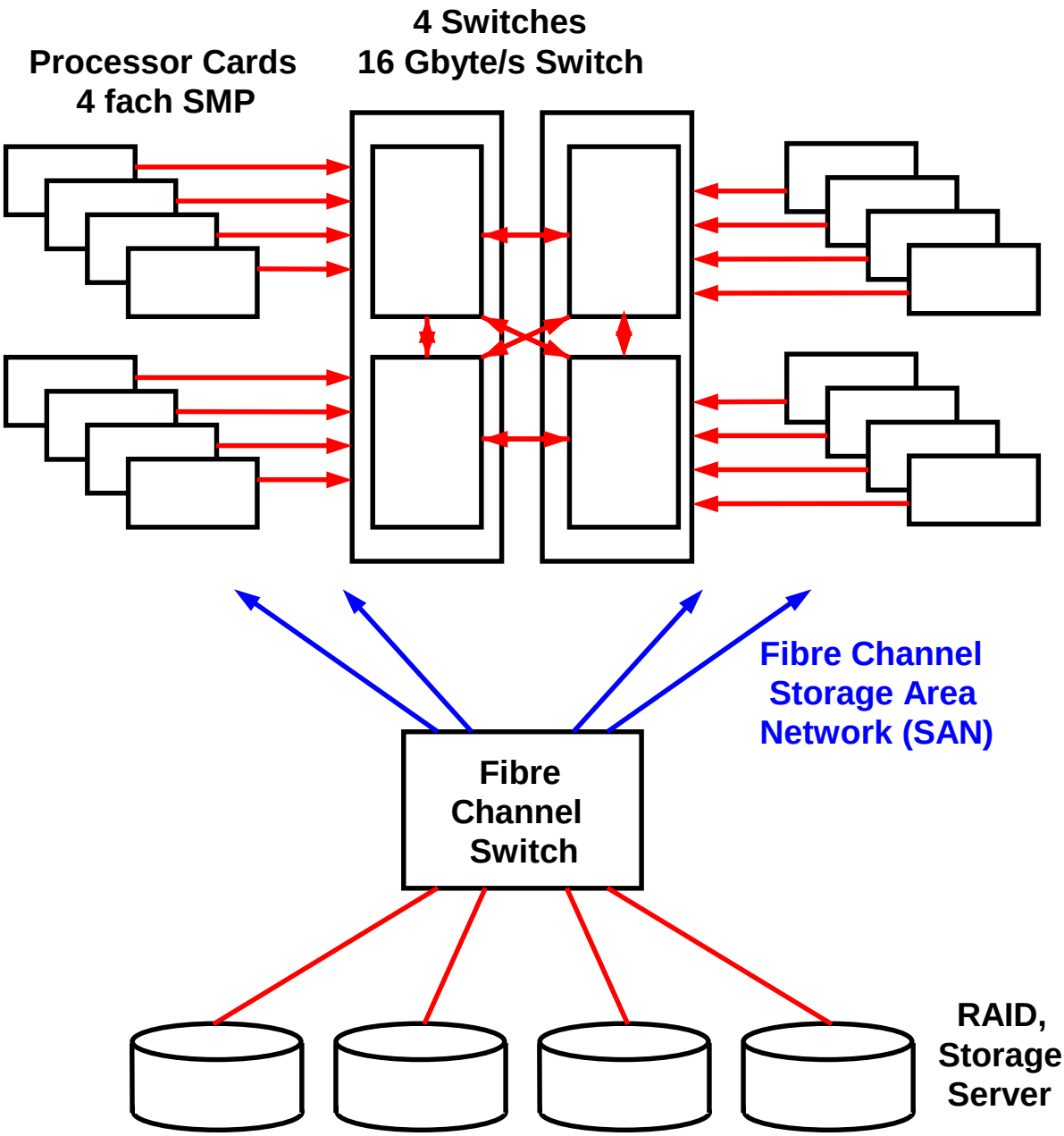
Shared data (shared disk)

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



Shared Disk Emulation

Rechner 2 bittet Rechner 1, die gewünschten Daten zu übertragen

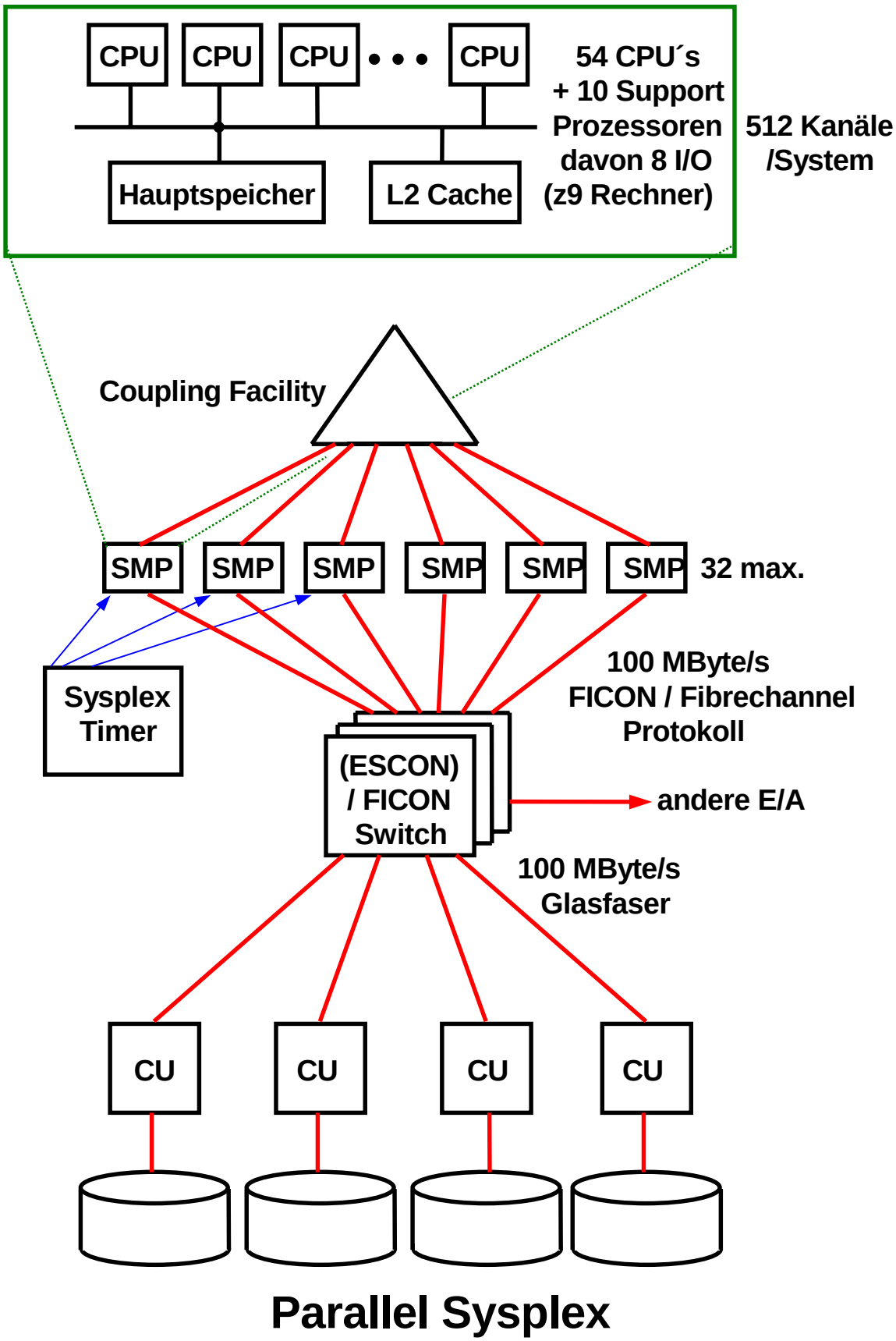


HP Superdome Cluster

64 CPU´s

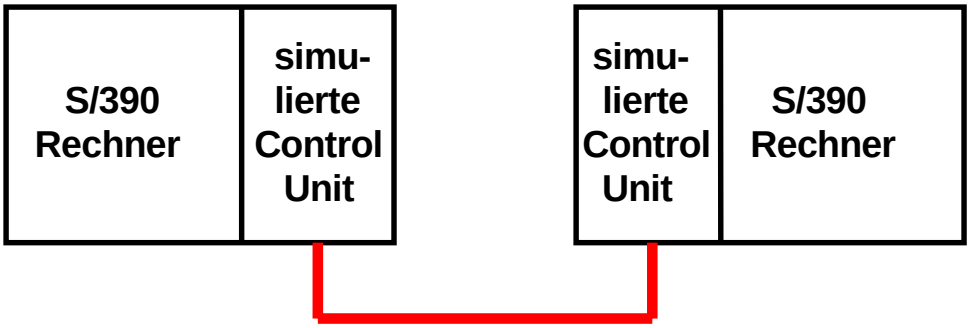
16 Knoten, je 4 CPU/Knoten

I/O Controller auf jeder Karte



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 4 Channel Subsystems
Jedes Channel Subsystem hat bis zu 256 Kanäle
Jeder FICON 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/FICON 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 z/OS oder 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 z/OS oder S/390 Servern

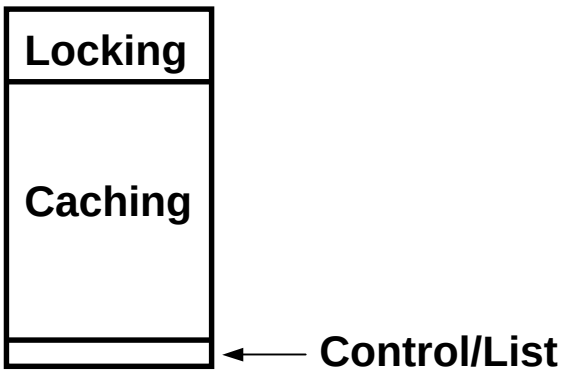
Der Sysplex Zeitgeber (Timer) stellt allen z/OS und OS/390 Instanzen eine gemeinsame Zeitbasis zur Verfügung. Dies ermöglicht korrekte Zeitstempel und Ablaufsequenzen bei Datenbank Änderungen. Dies ist besonders bei Datenbank-Recovery Operationen wichtig.

Coupling Facility

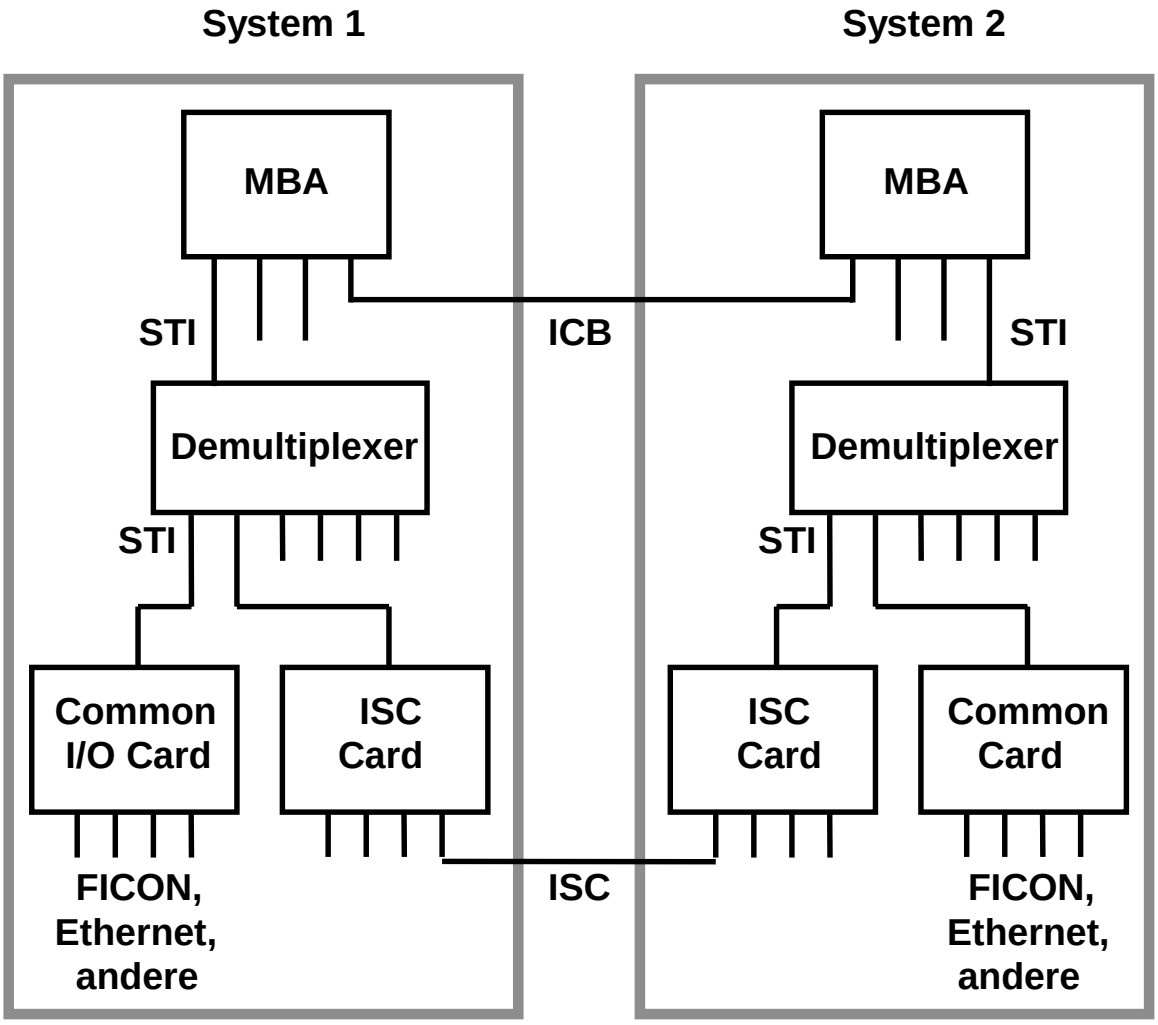
Die Coupling (CF) Facility ist in Wirklichkeit ein weiterer zSeries Rechner mit spezieller Software. Die Aufgaben der CF sind:

- **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 und spezieller Hardware Unterstützung mit den Systemen 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.

Plattenspeicher
Ein/Ausgabe Konfiguration

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.

Unterschiedliche Festplattenanschlüsse

ATA (IDE) und Serial ATA

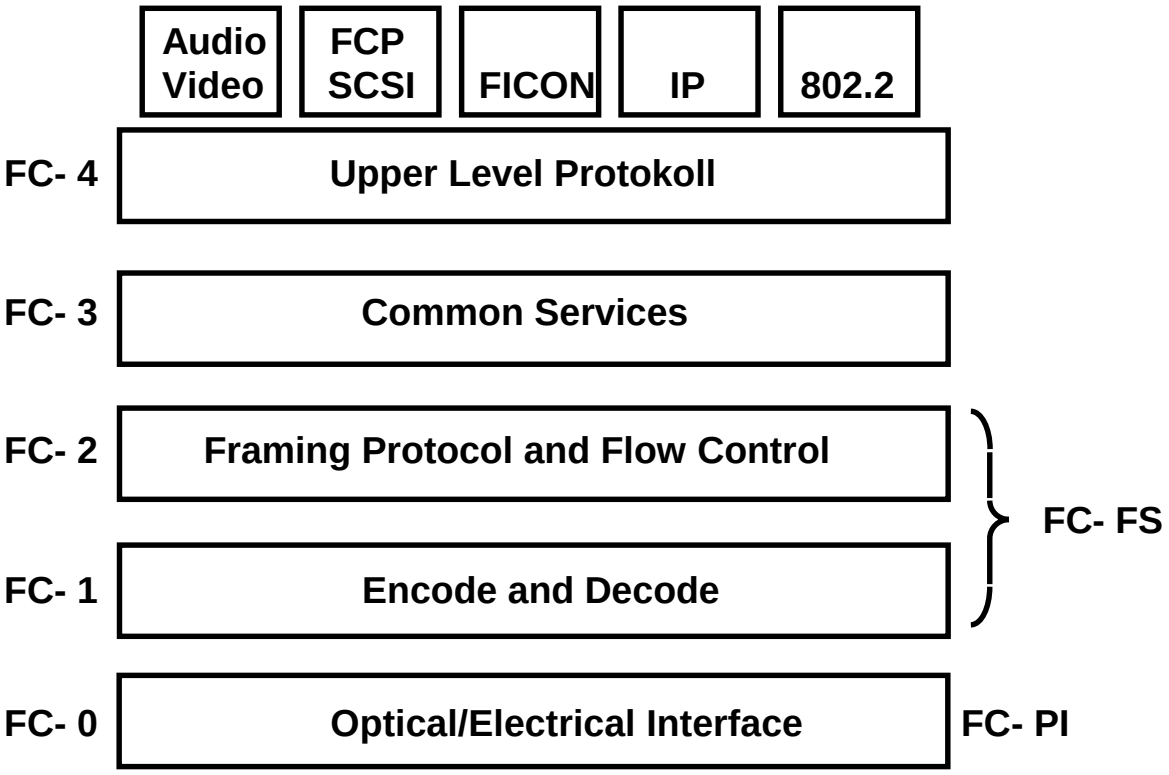
SCSI (parallel SCSI) und SAS (Serial Attached SCSI)

SCSI Fibre Channel Protocol (FCP)

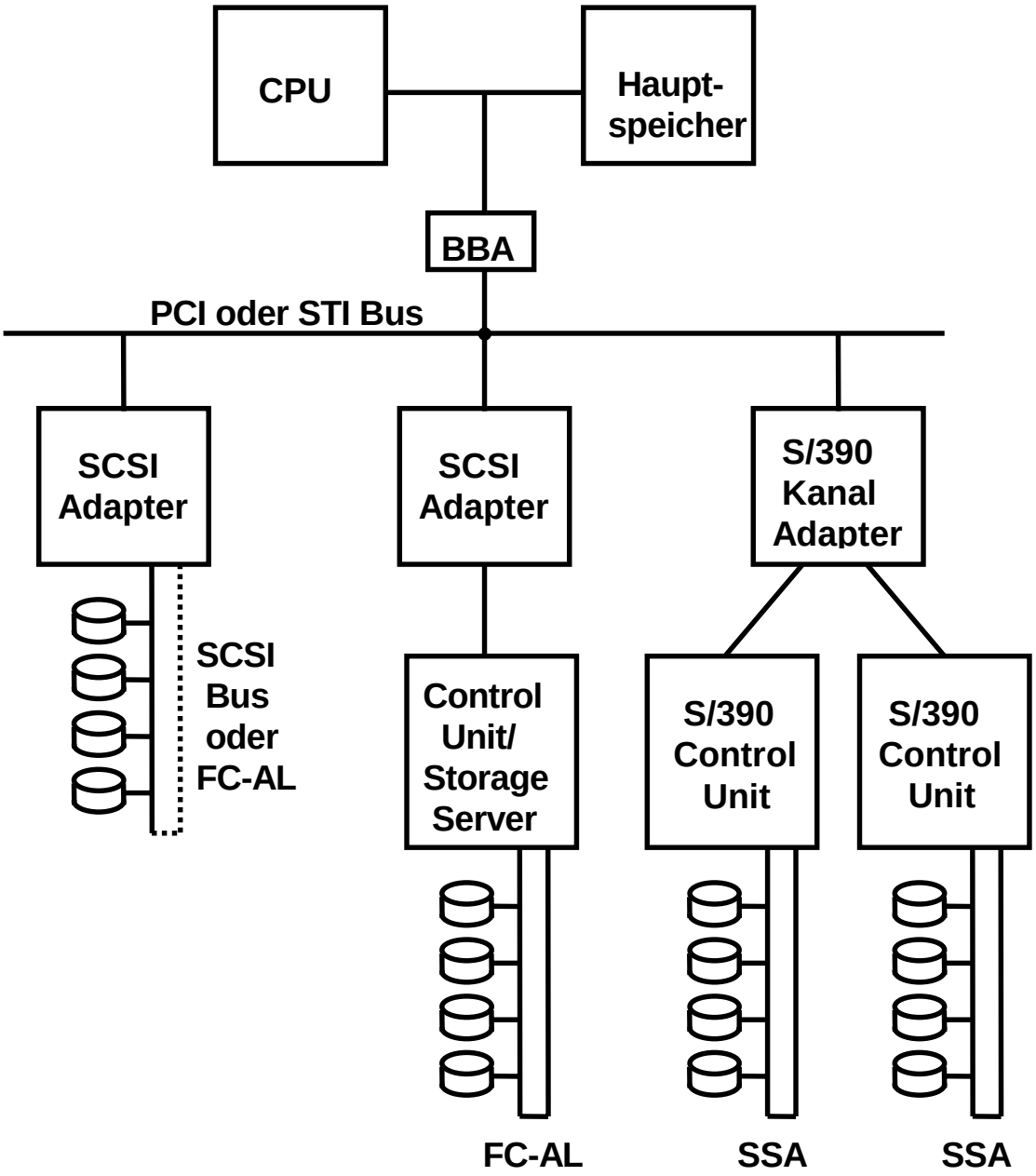
SCSI Fibre Channel Arbitrated Loop (FC-AL)

SSA (Serial Storage Architecture)

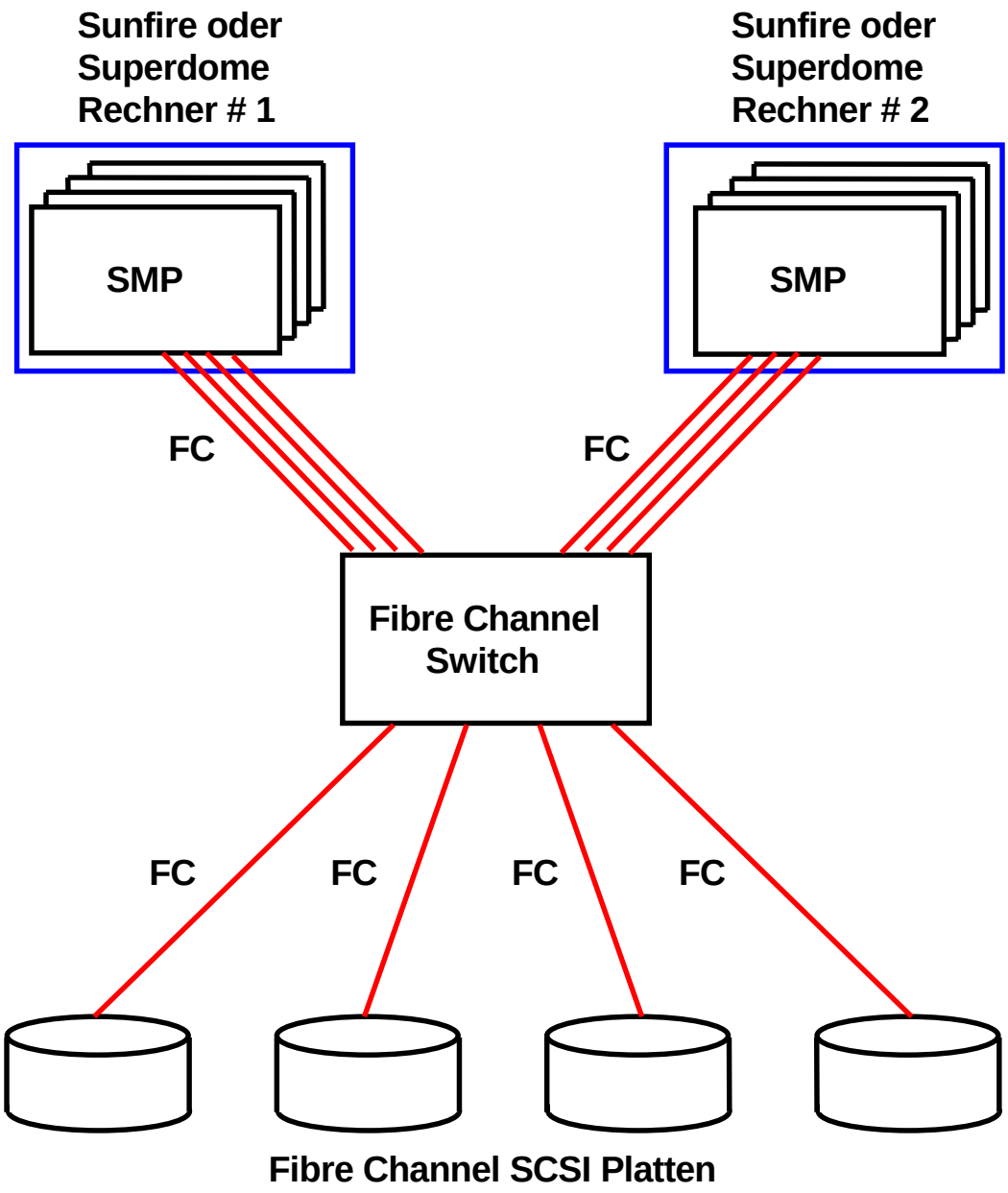
Beim Anschluss einer größeren Anzahl von Plattenspeichern an einen Server ist der Fibre Channel das dominierende Protokoll.



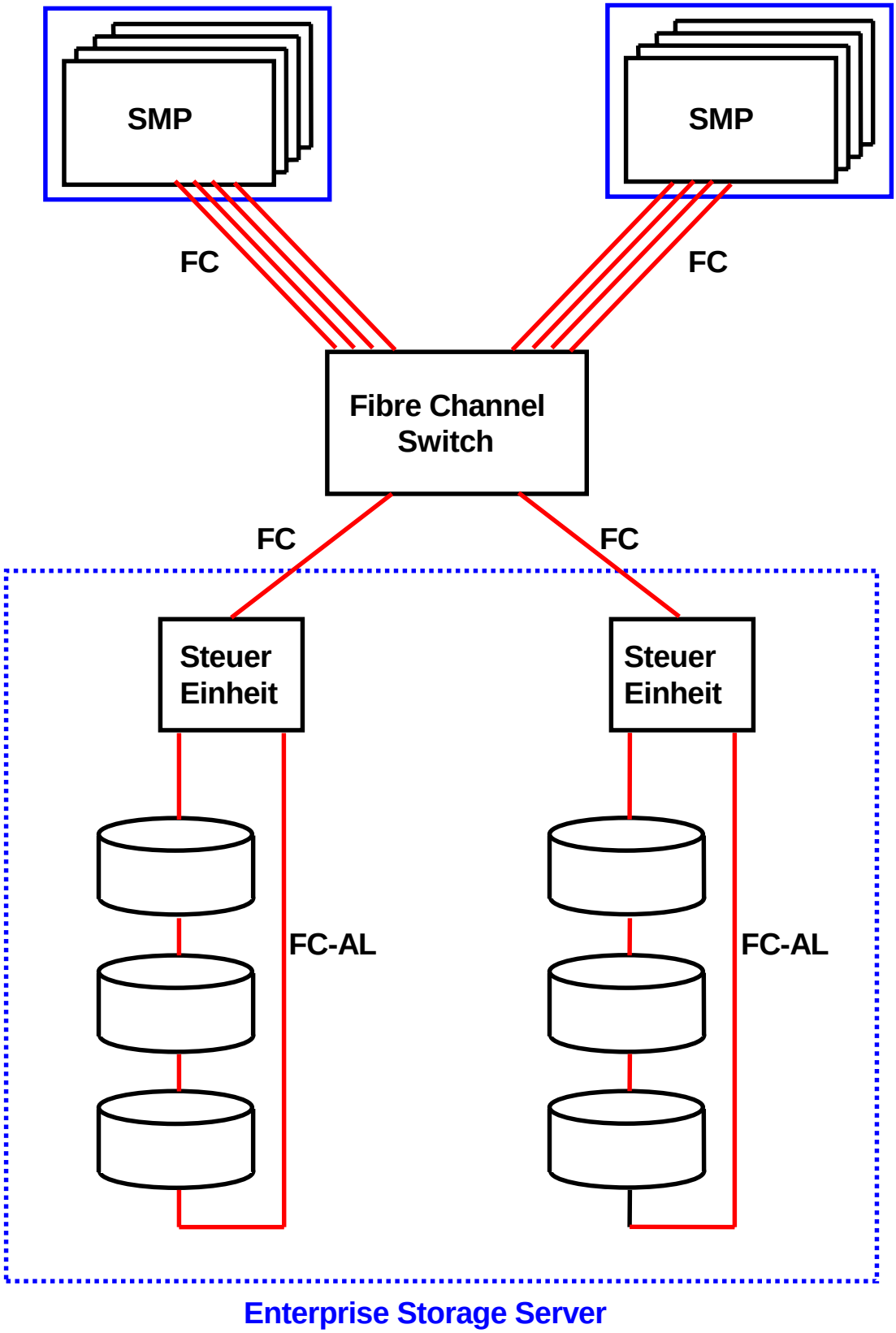
Fibre Channel Standard Architektur



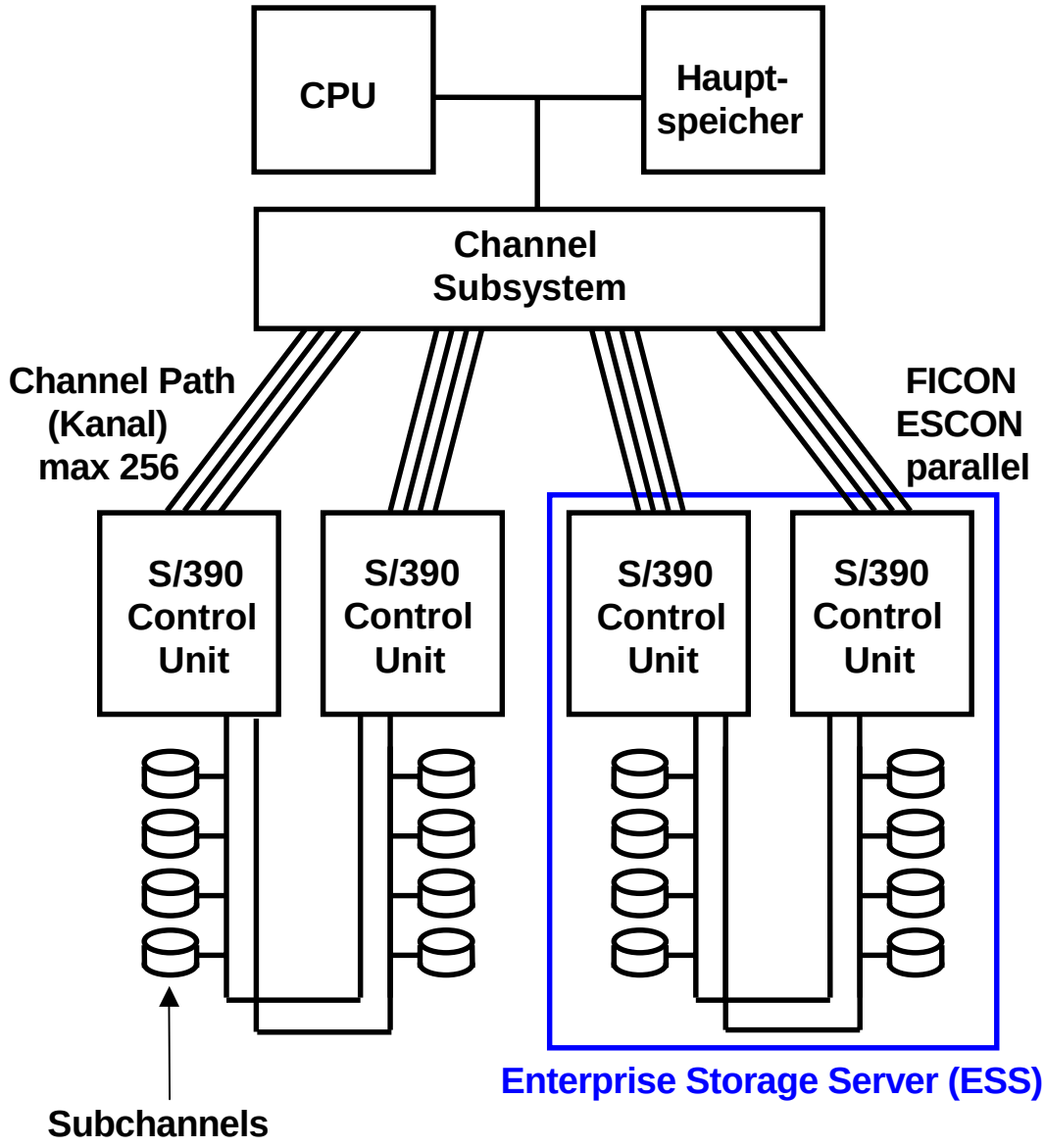
Plattenspeicher Anschlußalternativen



Einfache Fibre Channel Konfiguration



RAID, Cache Funktionalität



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.

Enterprise Storage Server

Im Wesentlichen besteht jeder Enterprise Storage Server aus vier Teilen:

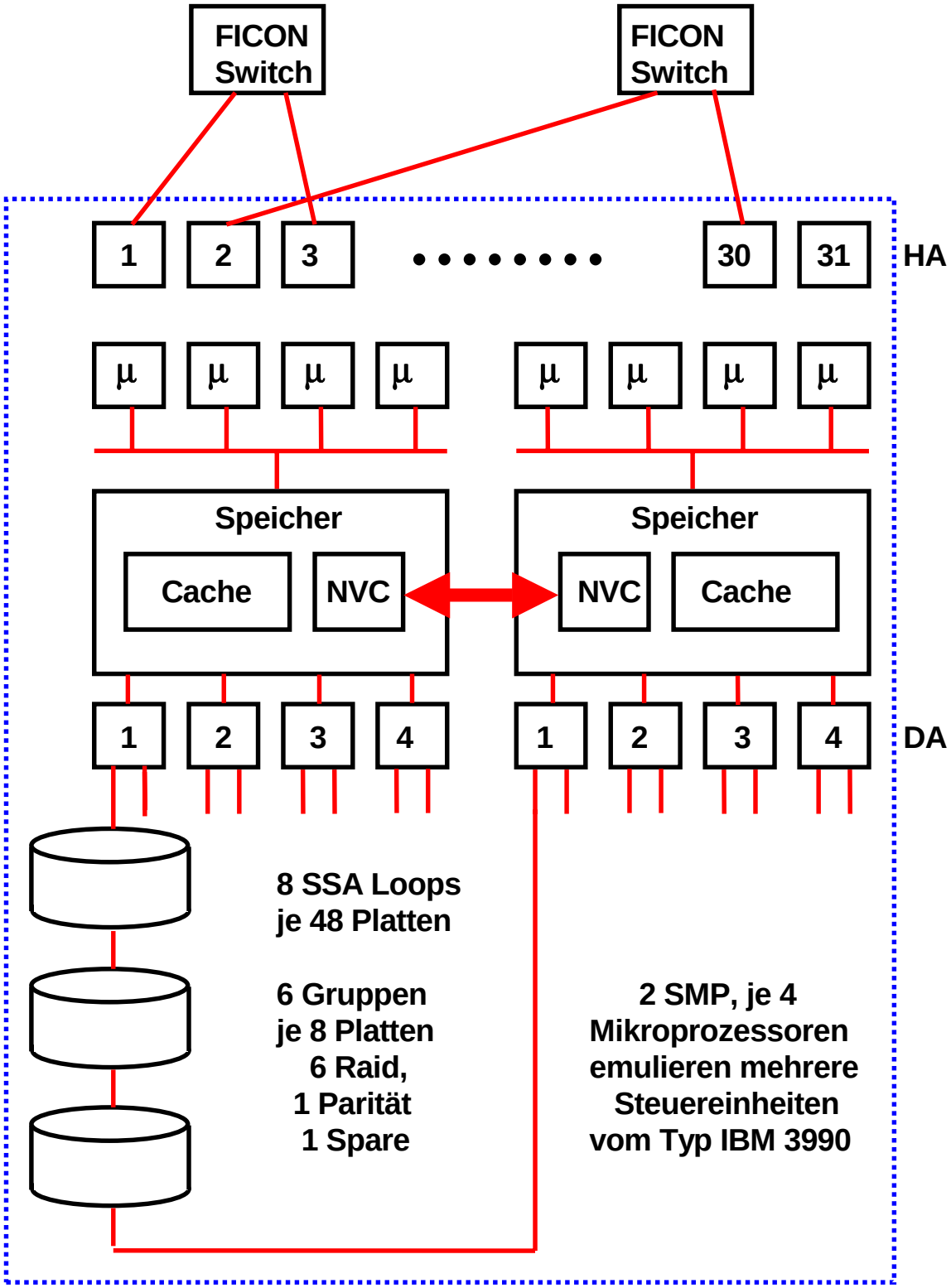
1. **Front End, welches die Schnittstelle zu den Rechnern darstellt. Bei S/390 und System z sind dies ESCON- oder FICON-Kanäle; bei UNIX-Systemen serielle SCSI/Fiberchannel Kanäle.**

2. **Cache, welcher aus zwei Teilen besteht. Dem Cache für Daten, die gelesen werden sollen, und dem Cache für Daten, die geschrieben werden sollen. Letzterer heisst Non-Volatile Storage und bezeichnet damit Cache, der extra gegen Stromausfälle gepuffert ist.**

3. **Back-End-Kanäle, welche bei den meisten heutigen Storage Processoren SCSI-Kanäle sind.**

4. **Back Store. Dieser verwendet SCSI-Platten als Bausteine, die häufig als RAID konfiguriert werden. Jede der Platten verfügt noch einmal, ähnlich wie PC-Platten, über einen eigenen vergleichsweise kleinen Cache.**

Heutige Enterprise Storage Server besitzen sehr grosse Caches bis zu 256 GByte. Der Non-Volatile Storage kann deutlich kleiner sein, da er nur zum vorübergehenden Zwischenspeichern der Schreibzugriffe benötigt wird. Diese werden dann asynchron auf den Back Store geschrieben, so dass die Anwendung davon nichts bemerkt. Die bedeutensten Hersteller von Storage Processors sind die Firmen EMC, IBM, Hitachi und StorageTek.



DS 8300 Enterprise Storage Server

NVC = Non Volatile Cache (Batterie Back Up)

HA = Host Adapter, DA = Device Adapter, µ = Mikroprozessor

Enterprise Storage Server (ESS)

Beispiel IBM DS8300, Modell 9A2, Stand 4Q 2005

Bis zu 32 FICON oder SCSI-FCP Host Adapter, 4 FICON oder FCP Ports / Adapter

- **2 Cluster Prozessoren, je 4 x SMP**
- **2 x 128 Gbyte Cache, Teil davon als NVC (non-volatile Cache)**
- **2 x 4 oder 2 x 8 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 5 Einheit je Gruppe, 6+P+S**
- **48 Platten/Loop, 384 Platten insgesamt**
- **73 oder 146 oder 300 GByte/Platte**
- **bis zu 115,2 TByte / ESS (IBM DS8300, Modell 9A2)**
- **100 000 I/O Operations / s .**

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 SCSI-FCP als auch mit FICON Anschluß-möglichkeiten: EMC, Hitachi/Sun, MaxData, andere.

IBM Model 800
Enterprise Storage Server (Shark)

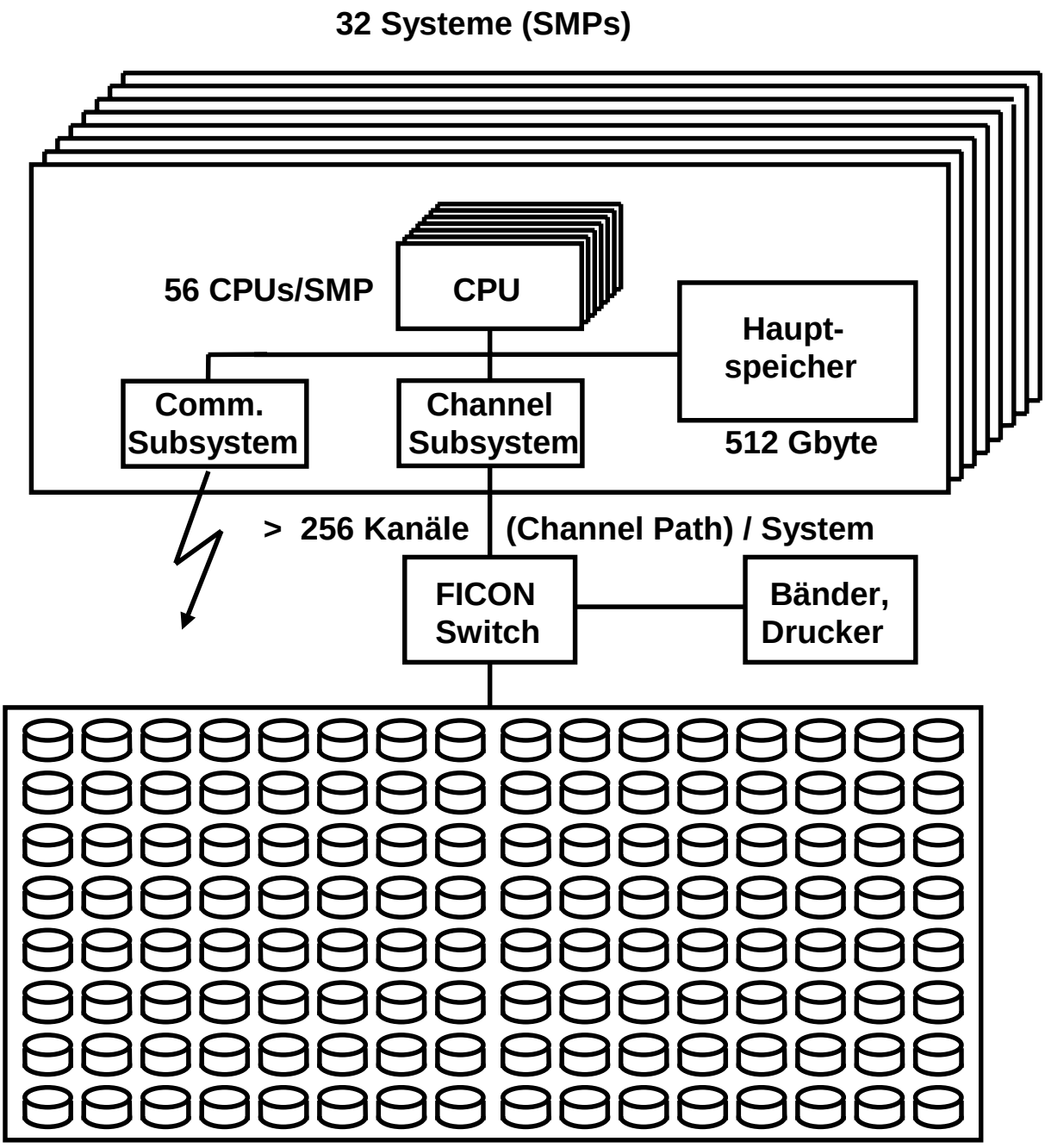
16 Host Adapter (HA zur Verbindung mit dem Host, die entweder je 2 ESCON- oder 2 SCSI- Kanäle oder je 1 FICON bzw. Fibre Channel nach aussen darstellen können.

Intern besteht der Storage Processor aus je 2 Clustern, die jeweils aus den PowerPC RISC-Prozessoren, dem Cache Storage und dem Non-Volatile Storage bestehen. Zum Back Store besitzt jedes Cluster 4 Device Adapter. Die Adapter arbeiten immer paarweise, und die Plattenstränge (Disk Arrays) oder *Ranks* sind über eine SSA Loop mit den Device Adaptern verbunden.

SSA bedeutet Serial Storage Architecture und stellt eine serielle Kreisverbindung (Loop) für SCSI-Platten dar. Es existieren 2 Lese- und 2 Schreibverbindungen, von denen jede mit 40 MB/s arbeitet, was eine Gesamtkapazität von 160 MB/s ergibt.

Die Ranks können als RAID Ranks oder als einfache Platten konfiguriert werden. Das Letztere bezeichnet man auch als Just a Bunch of Disk (JBOD). In der S/390 Welt hat die RAID-Konfiguration eine grössere Bedeutung, da sie die Daten gegen Geräteverluste am besten absichert. Dabei wird häufig eine RAID-5 Konfiguration bestehend aus 8 Platten verwendet. Der erste Rank besteht aus 6 Platten für die Daten, einer Platte für Parity und einer Spare-Platte. Jeder weitere Rank besitzt 7 Datenplatten und eine für Parity.

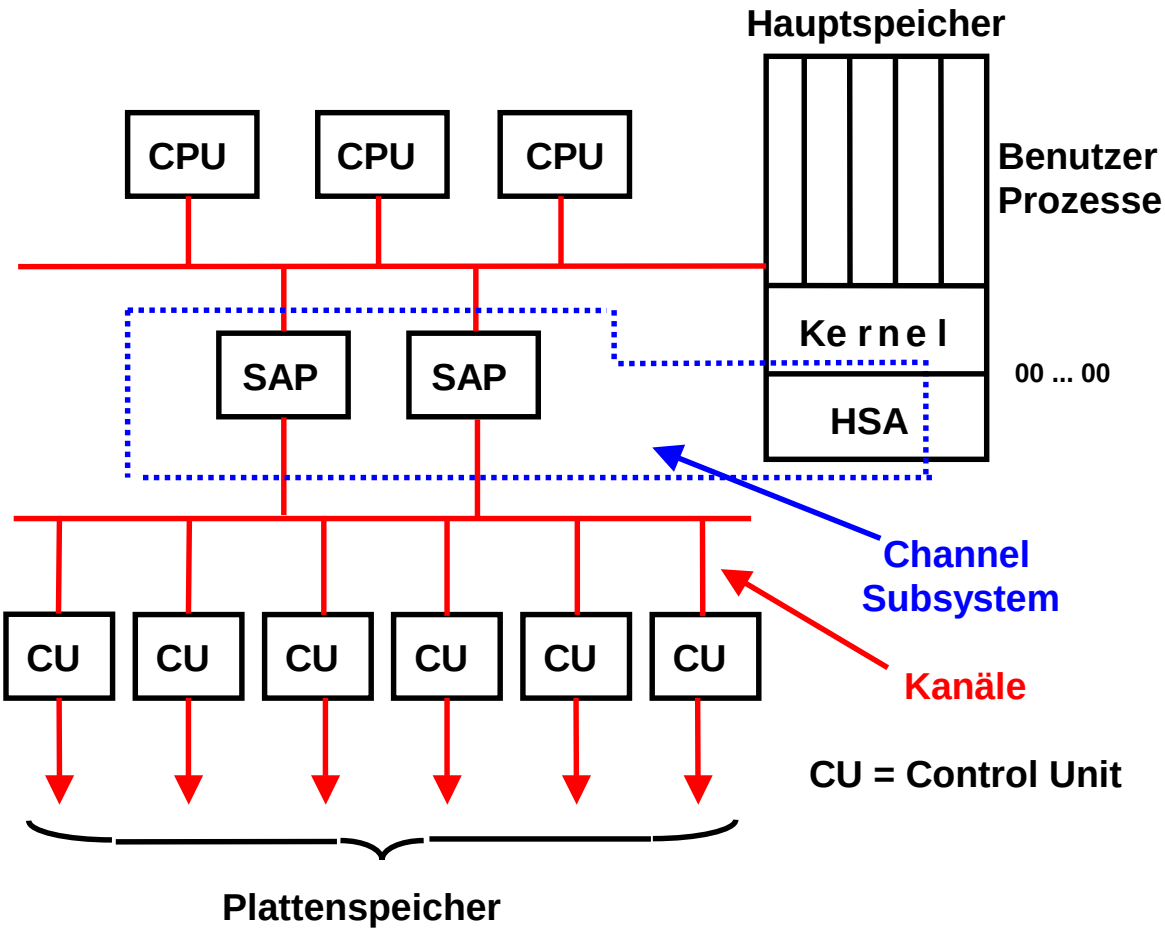
Der Enterprise Storage Server unterstützt eine Vielzahl von Funktionen, die die Datensicherung unterstützen sowie das parallele Schreiben und Lesen von Daten von demselben Logical Volume.



256 000 Plattenspeicher (Devices)
10 - 1000 TeraByte Plattenspeicherplatz

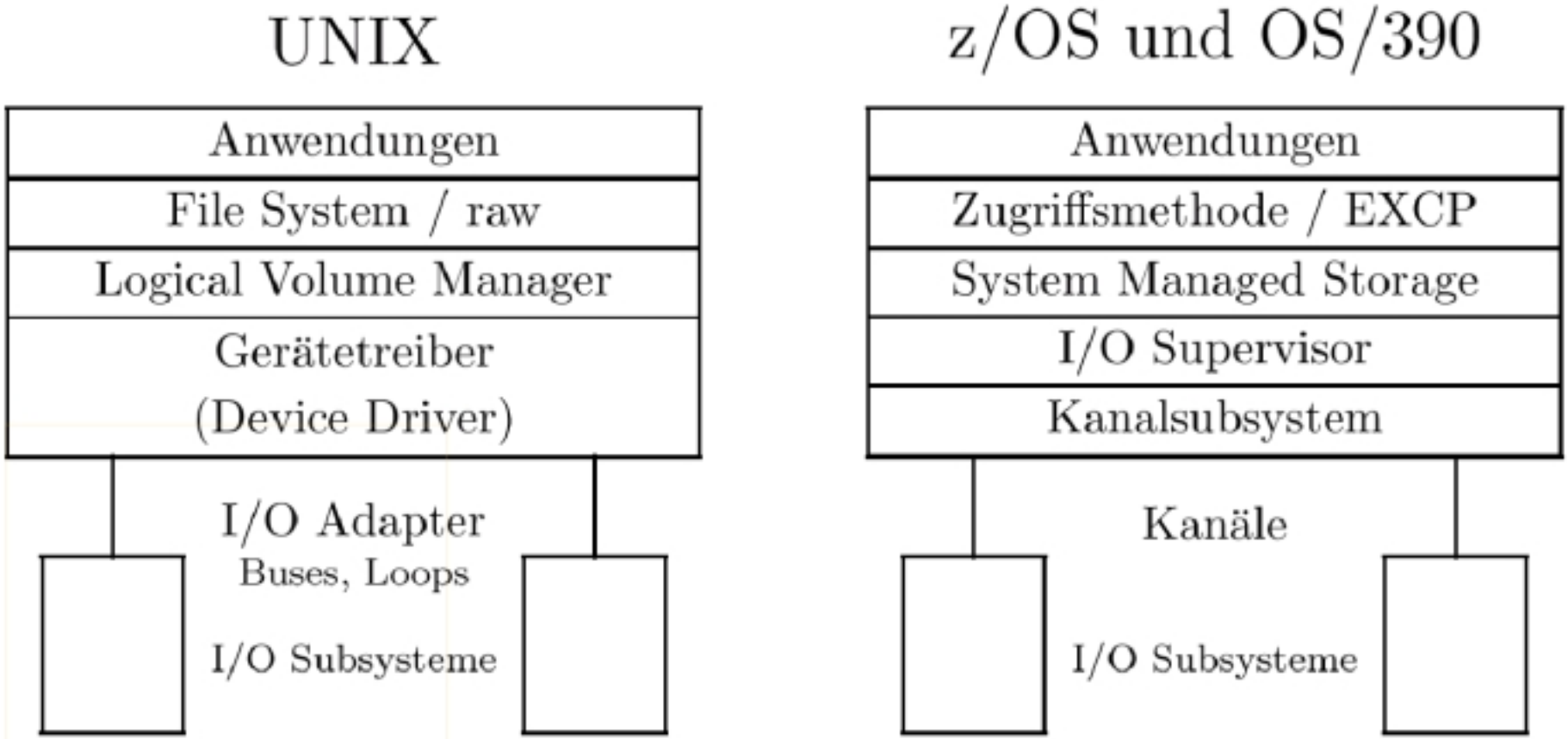
z Series (S/390) Großsystemkonfiguration

zSeries Ein/Ausgabe Anschluss

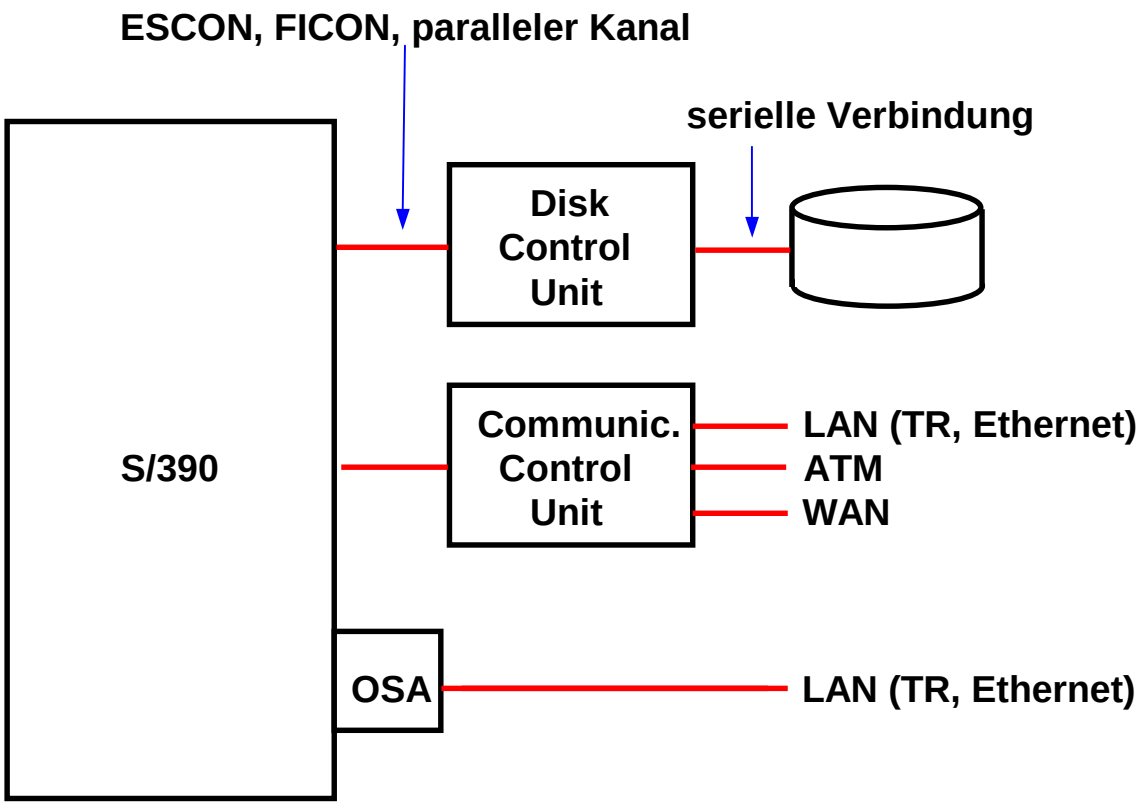


Die *HSA* (Hardware System Area) ist ein Teil des Hauptspeichers. Sie liegt außerhalb des Adressenraums, auf den die CPUs zugreifen können. Das *Channel Subsystem* besteht aus SAP Prozessoren und Code in der HSA. Es bildet das virtuelle E/A Subsystem, mit dem der Betriebssystem Kernel glaubt zu arbeiten,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

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



Unterschied in der z/OS und Unix Ein/Ausgabe Architektur



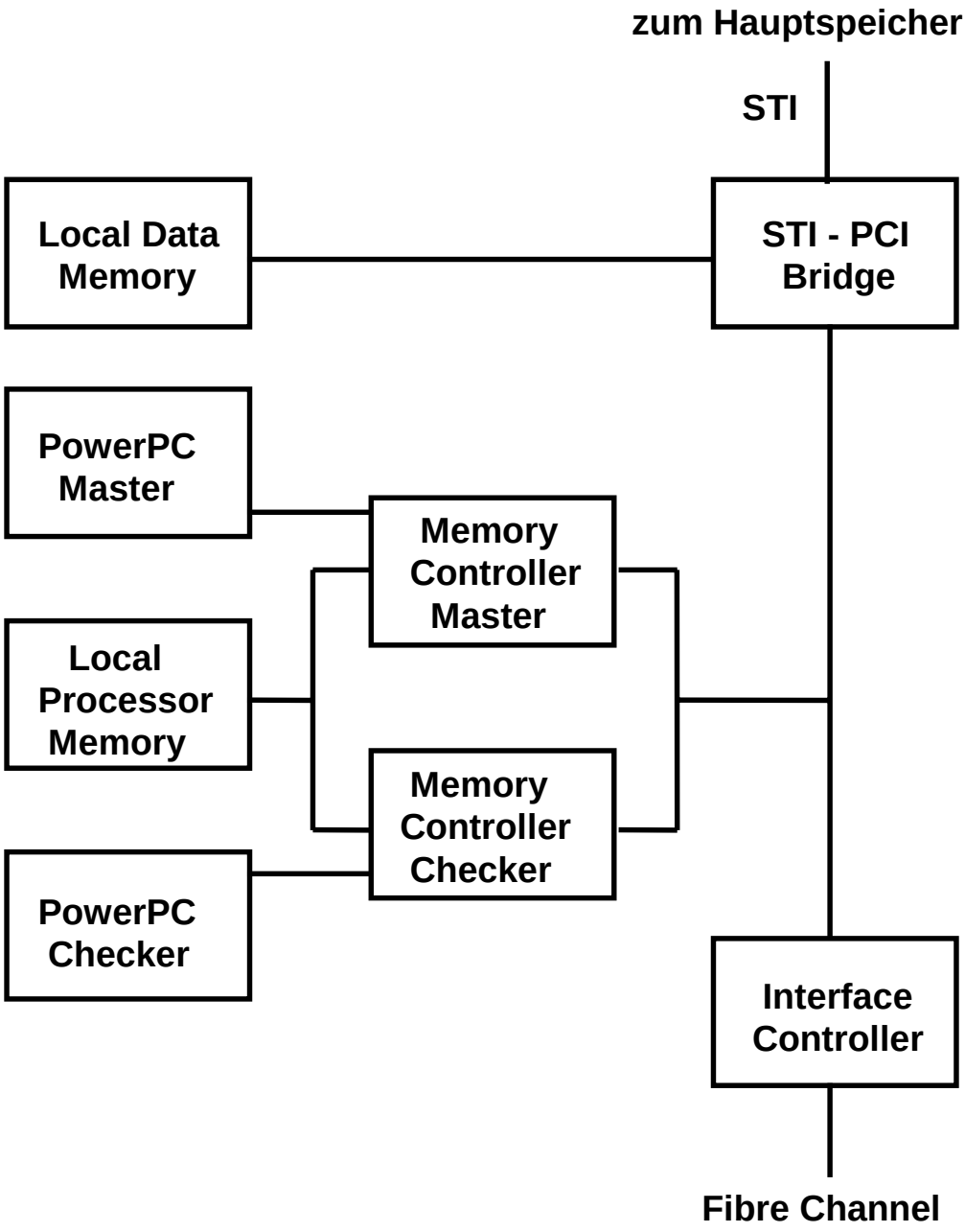
S/390 E/A Konfiguration

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.

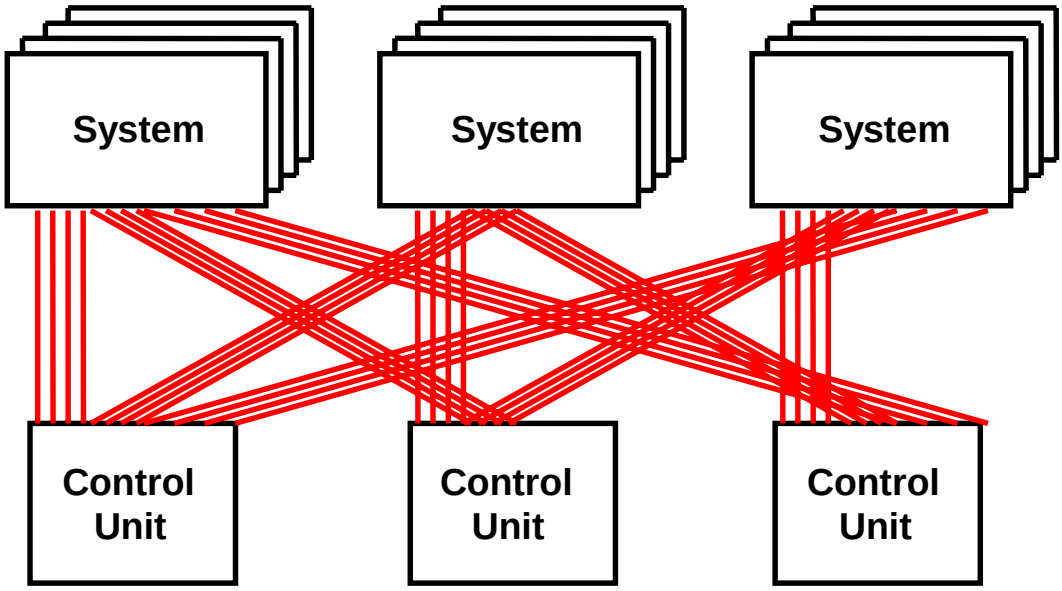
Es existieren viele unterschiedliche Typen von Steuereinheiten. Die wichtigsten schließen externe Speicher (Platten, Magnetbänder Archivspeicher) und Kommunikationsleitungen an.

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

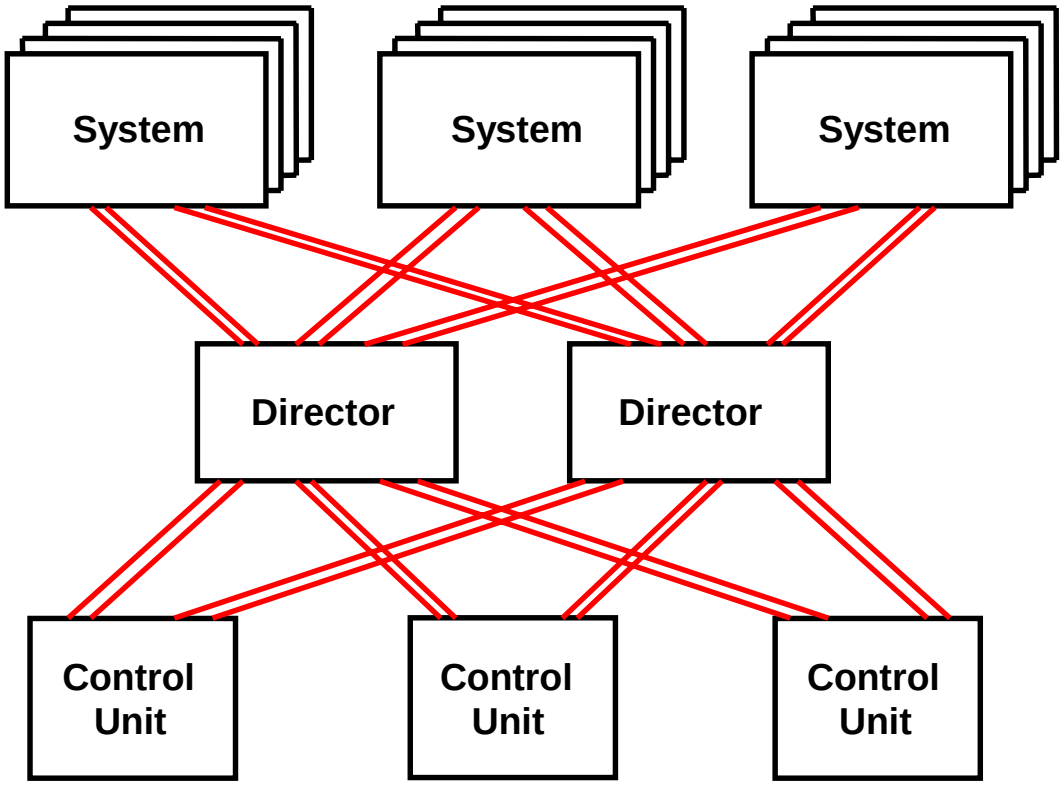
Einige Steuereinheiten können in den S/390 Rechner integriert werden. Das wichtigste Beispiel ist der OSA Adapter für den Anschluß von LAN´s.



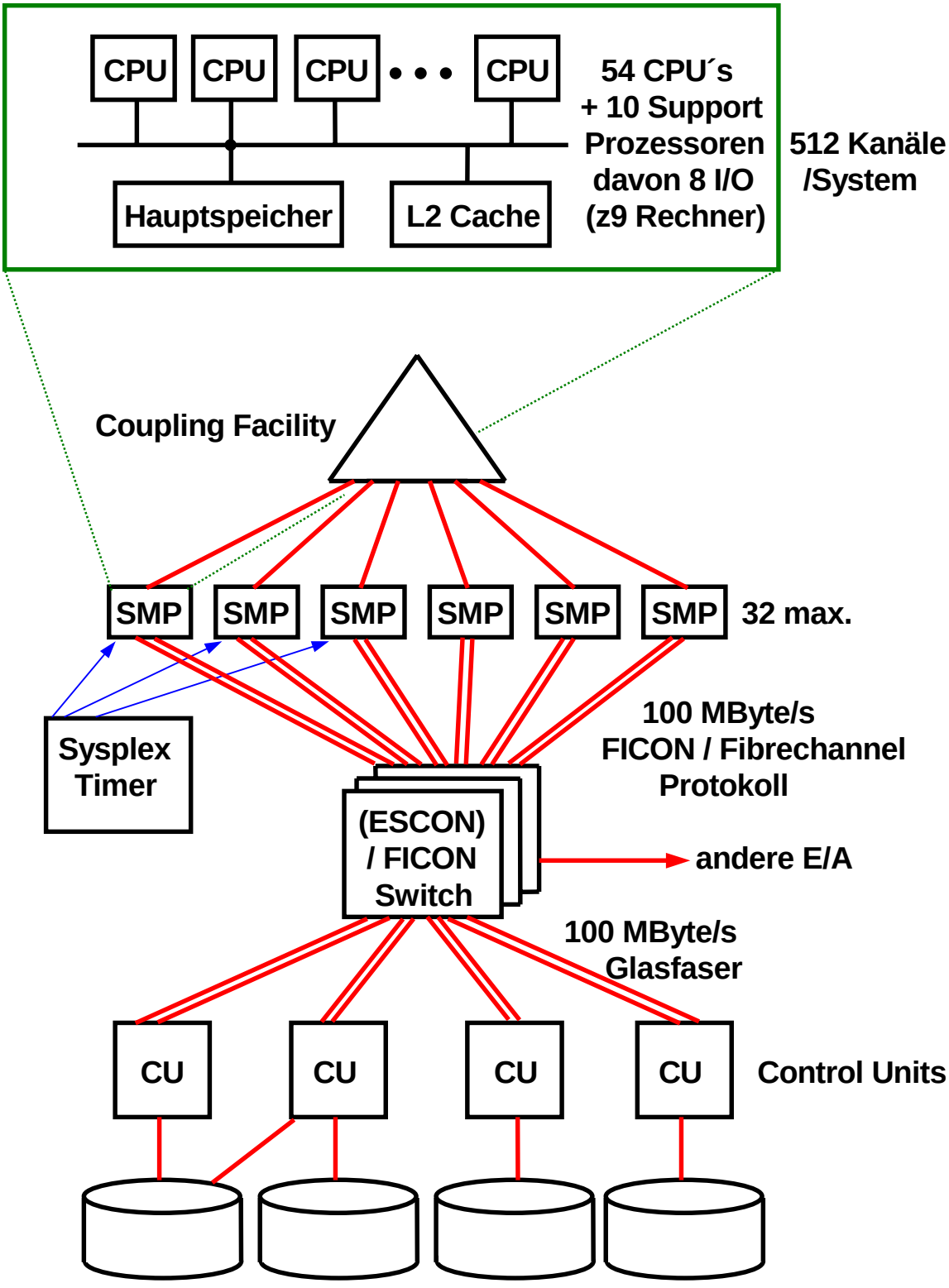
zSeries Fibre Channel Kanal basierend auf der Common I/O Card



Parallel Channel Configuration



FICON (ESCON) Channel Configuration

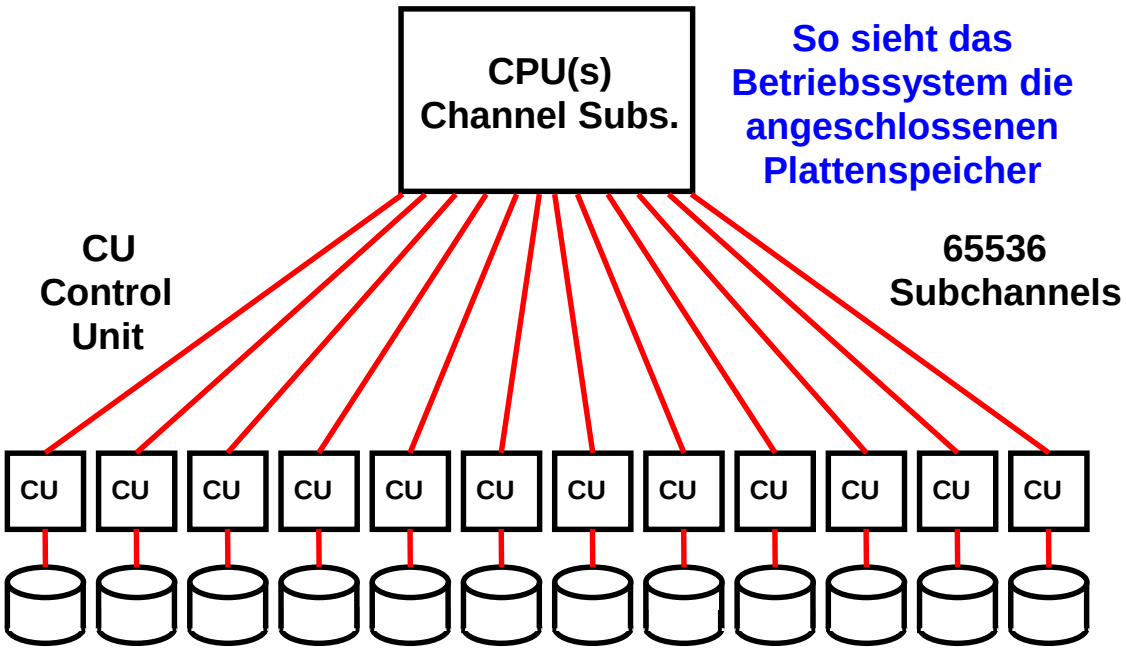


Parallel Sysplex

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.
65 536 Subchannels (E/A Geräte) pro Channel Subsystem.
Ein z9 Rechner kann über bis zu 4 Channel Subsystems verfügen.