

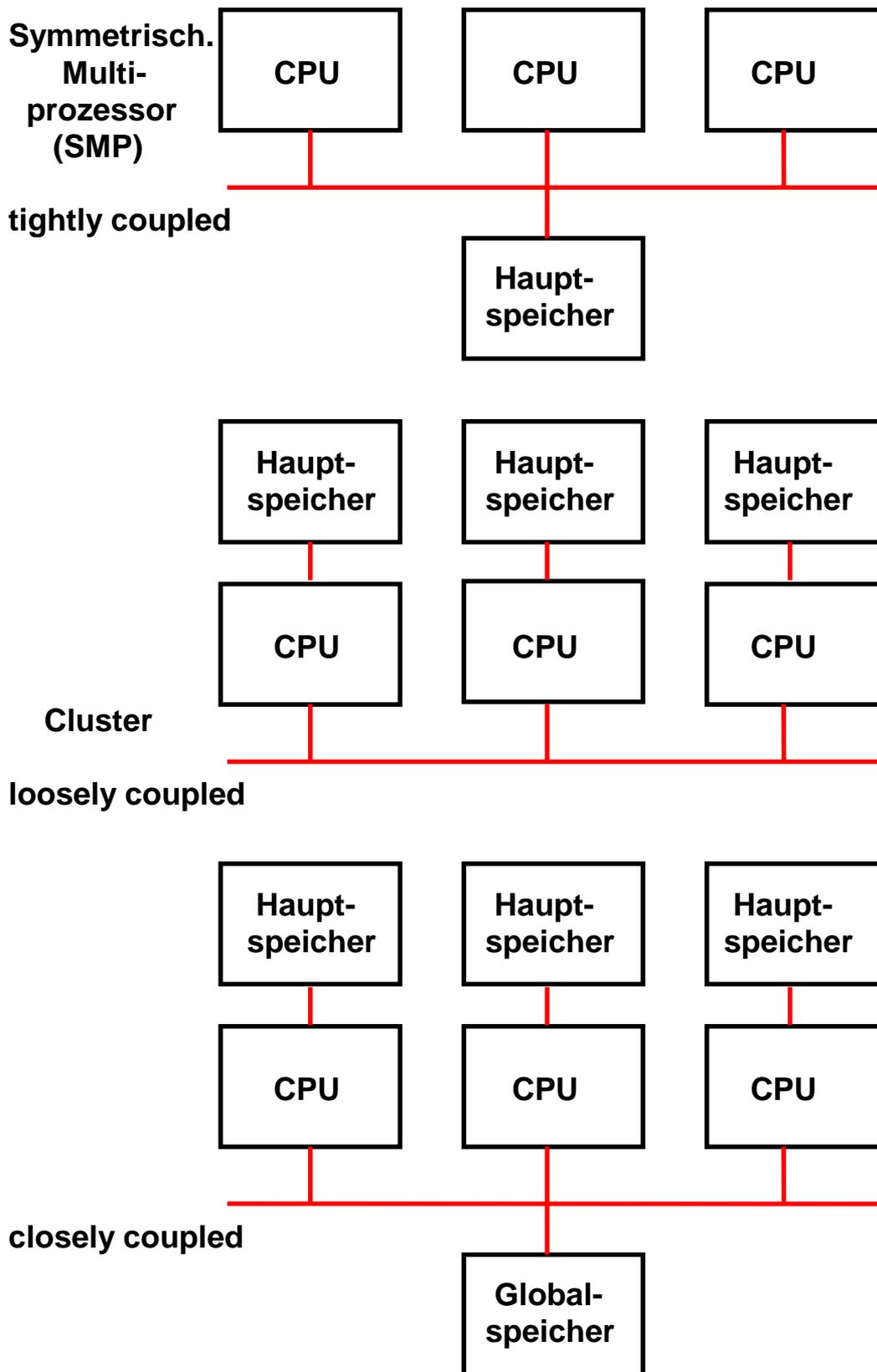
Einführung in z/OS

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

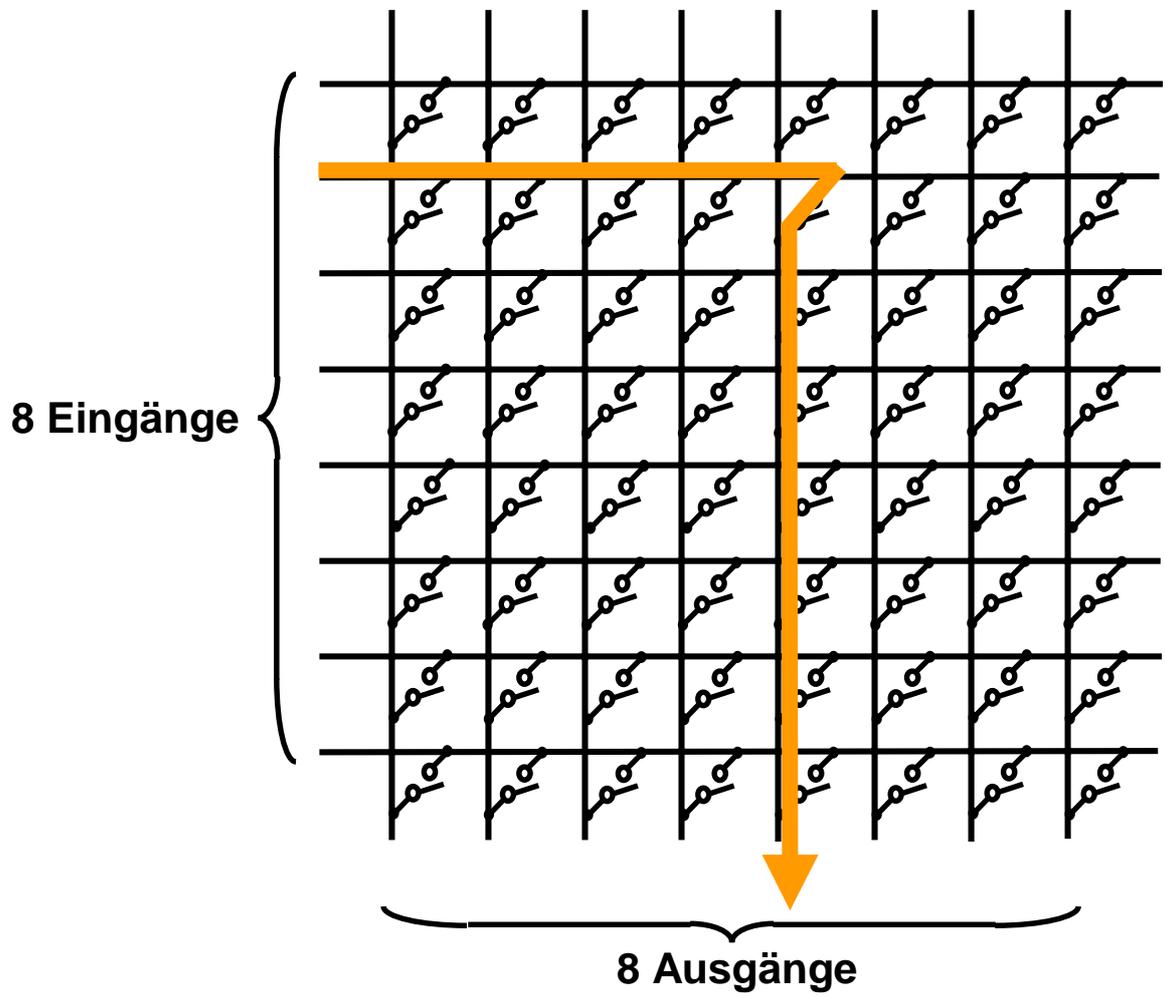
WS 2008/2009

Teil 5

Parallelrechner, I/O

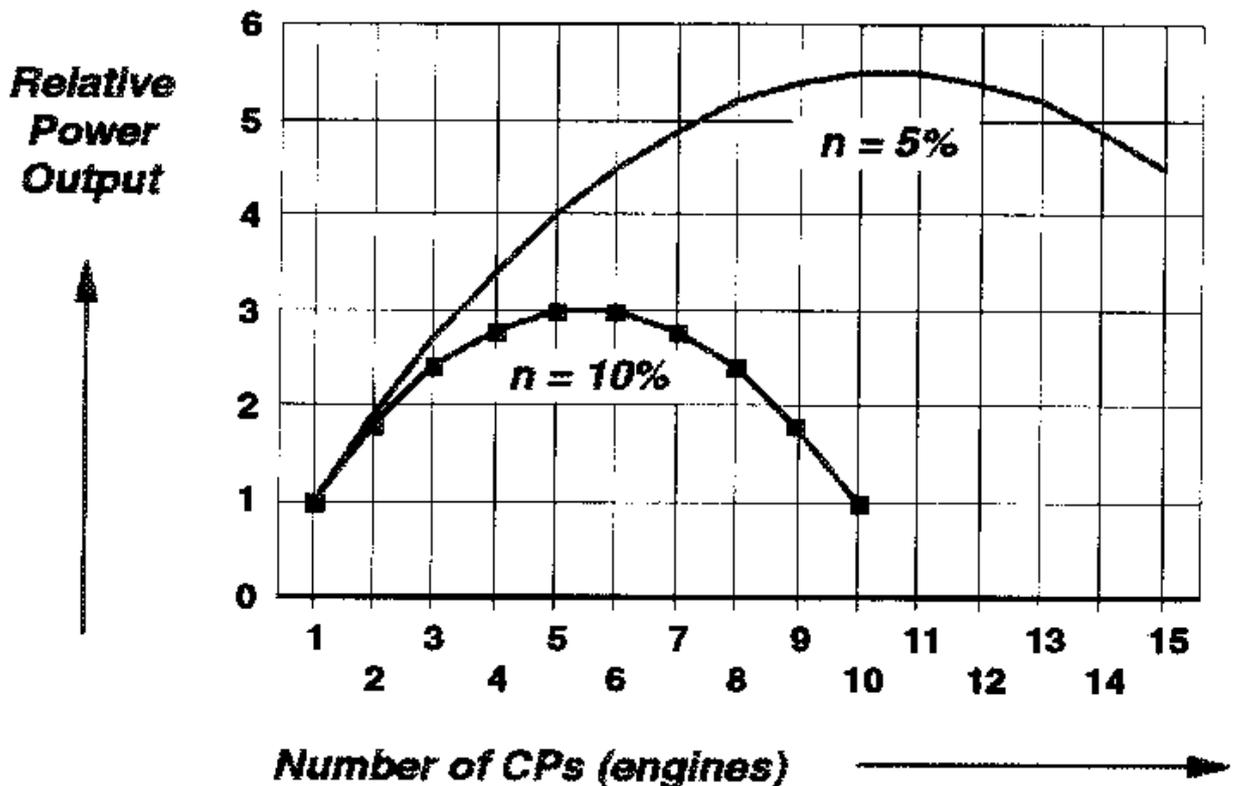


Taxonomie von MIMD Parallelrechnern



8x 8 Crossbar Matrix Switch

Leistungsverhalten eines Symmetric Multiprocessors (SMP)



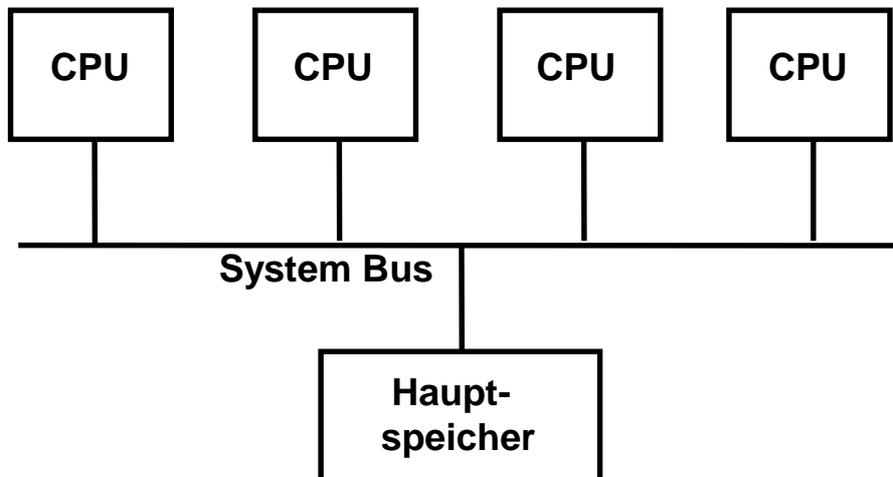
Angenommen, ein Zweifach Prozessor leistet das Zweifache minus $n\%$ eines Einfach Prozessors. Für $n = 10\%$ ist es kaum sinnvoll, mehr als 4 Prozessoren einzusetzen. Für $n = 5\%$ sind es 8 Prozessoren.

Angenommen $m =$ Anzahl CPUs. Der Leistungsabfall pro CPU ist

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

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

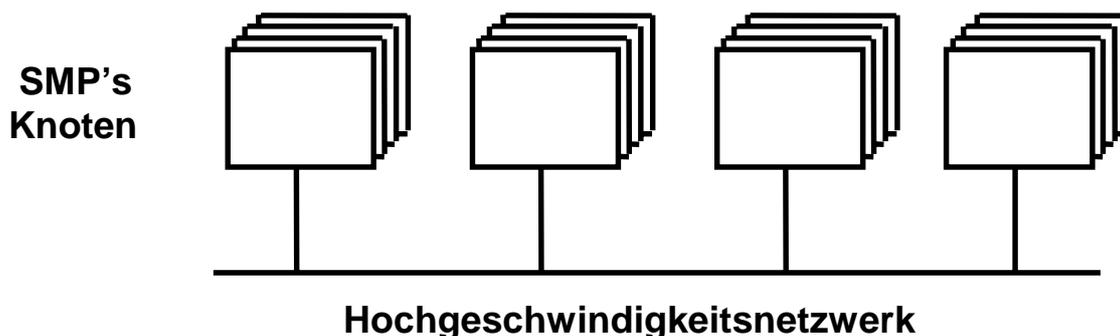
Die Gründe für den Leistungsabfall sind Zugriffskonflikte bei der Hardware und Zugriffskonflikte auf Komponenten des Überwachers. Die Überwacherkonflikte überwiegen.



SMP, Prozessor Knoten

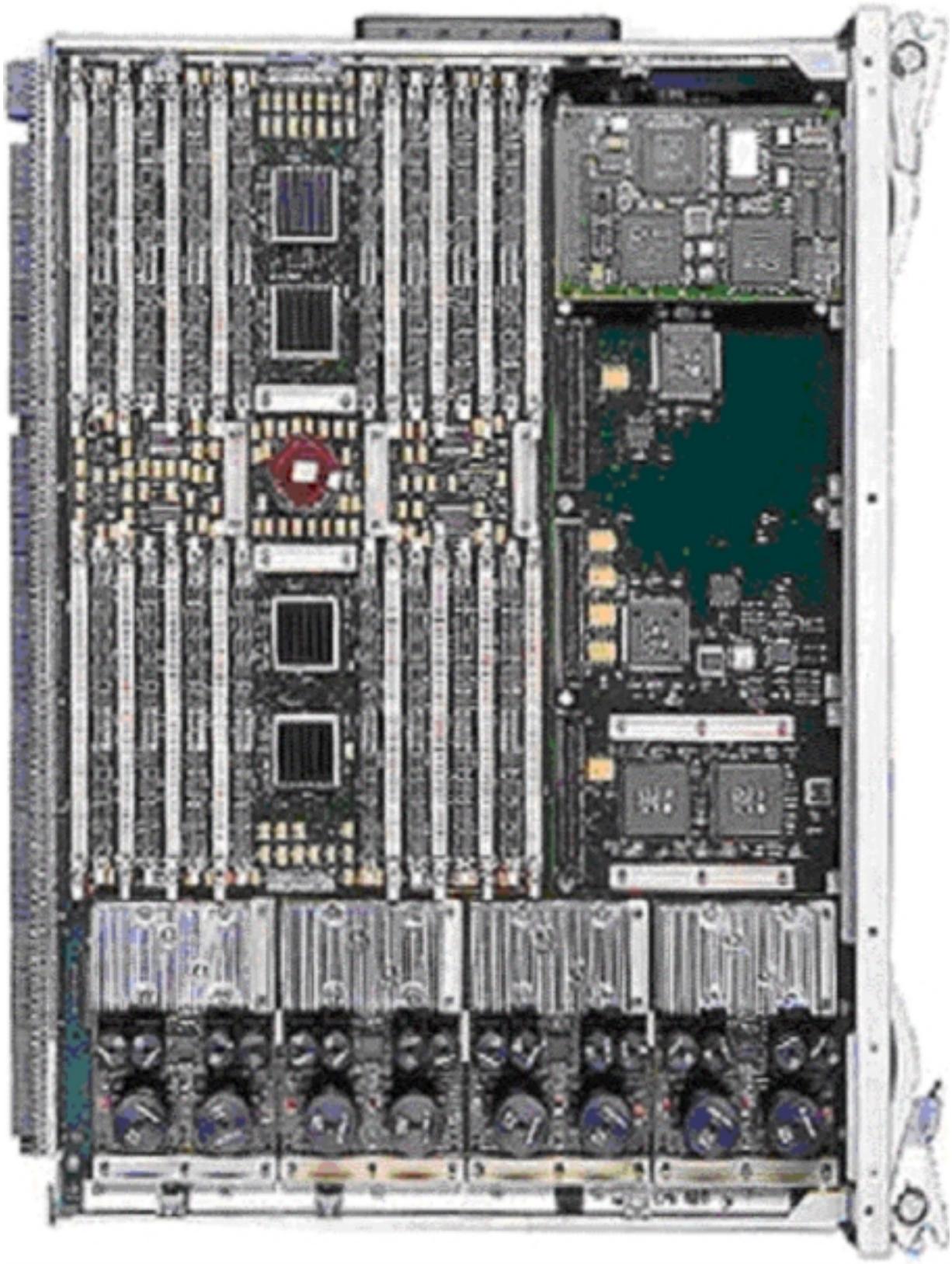
Ein SMP (Symmetric Multiprocessor, Prozessor Knoten, Node) besteht aus mehreren CPU's, die auf einen gemeinsamen Hauptspeicher zugreifen

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

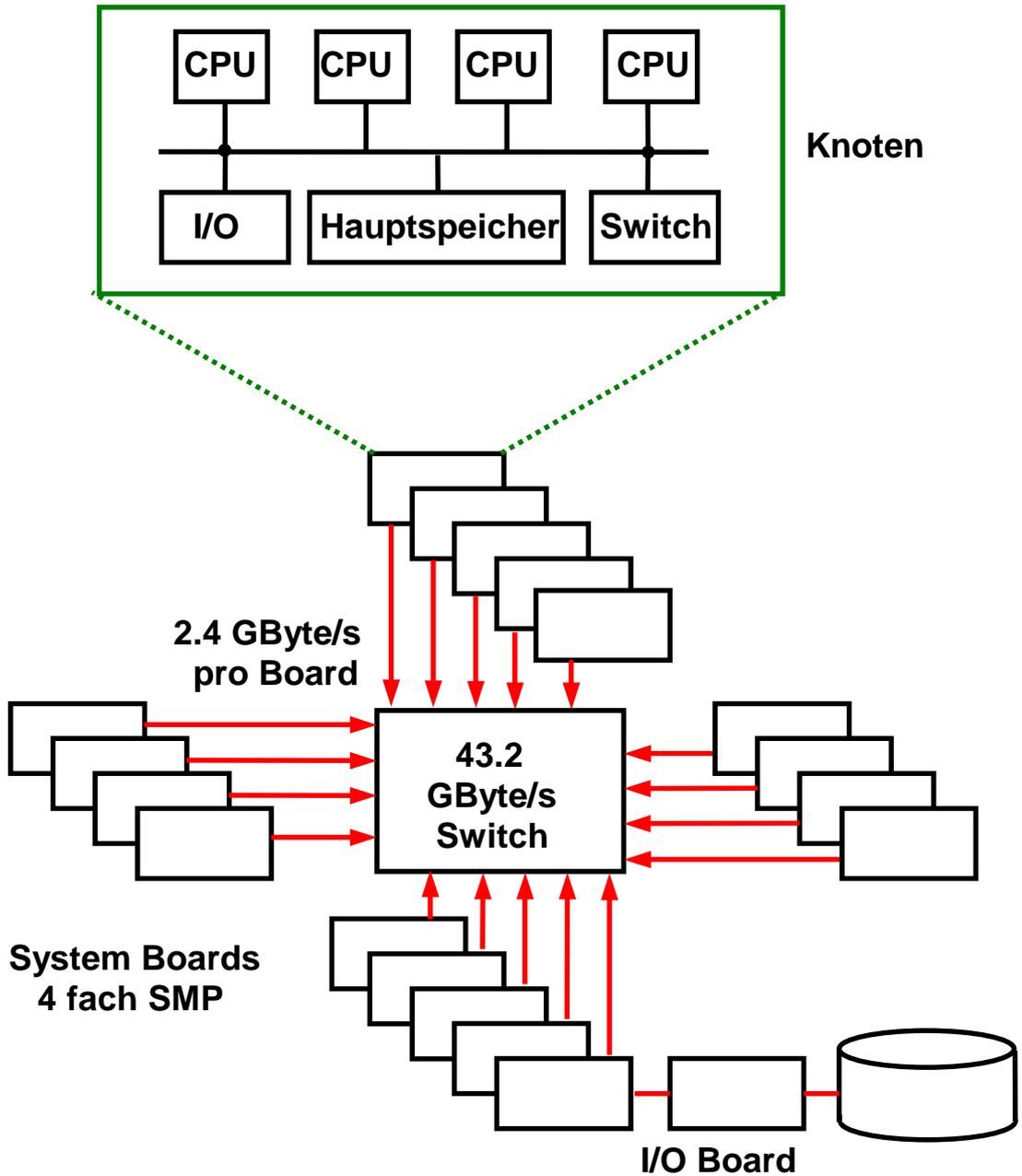


Cluster

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.

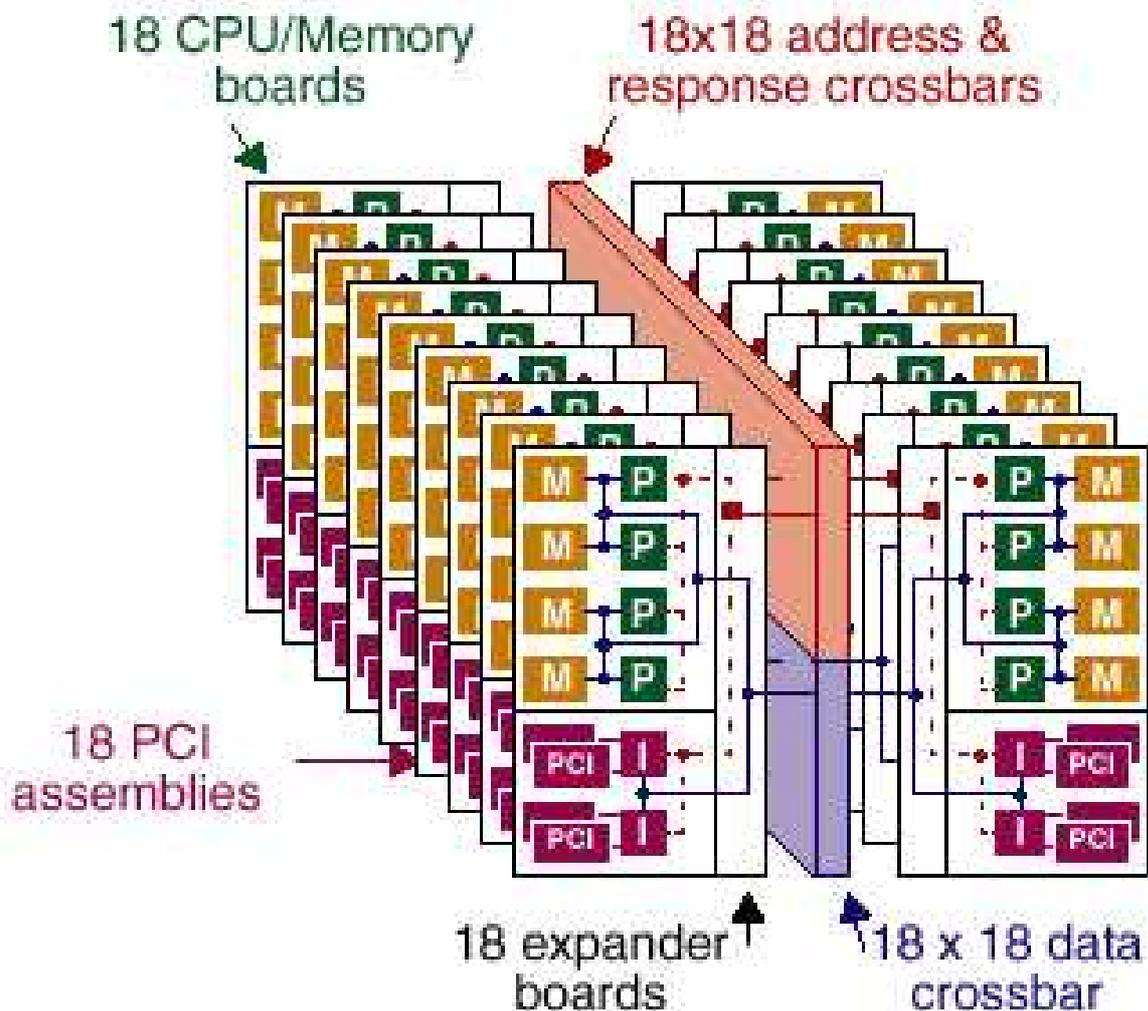


Sun E 10 000 System Board



Sun E15K
 72 CPU's
 18 System Boards, je 4 CPU/System Board
 I/O Controller auf jedem System Board

Sun Fire 15000



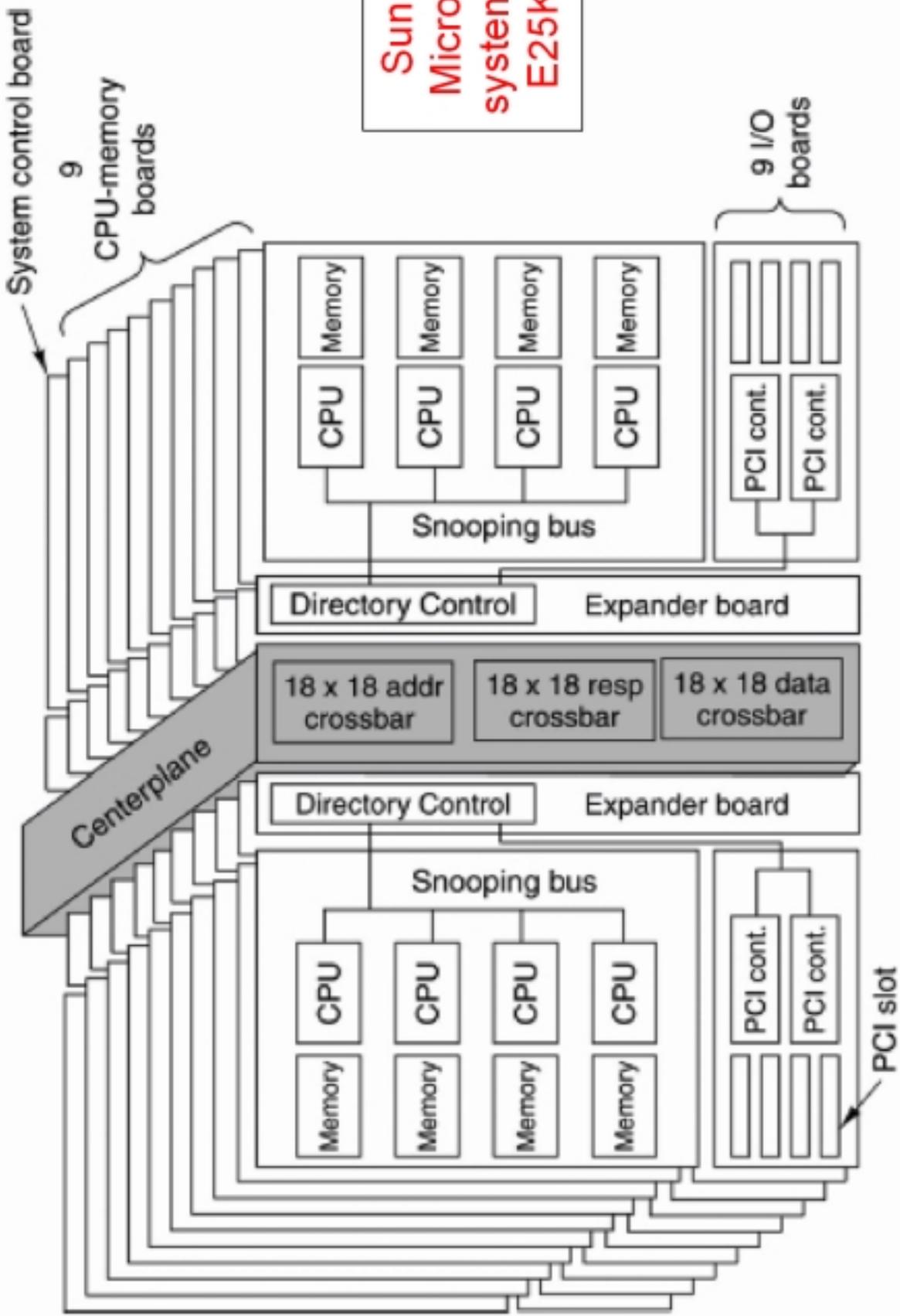
Sun Fire 15000 System hat max 576 Gbyte Hauptspeicher, max 18 CPU/Memory Boards, max 18 Domains, max 18 I/O Boards, max 72 PCI Slots für 72 PCI Karten.

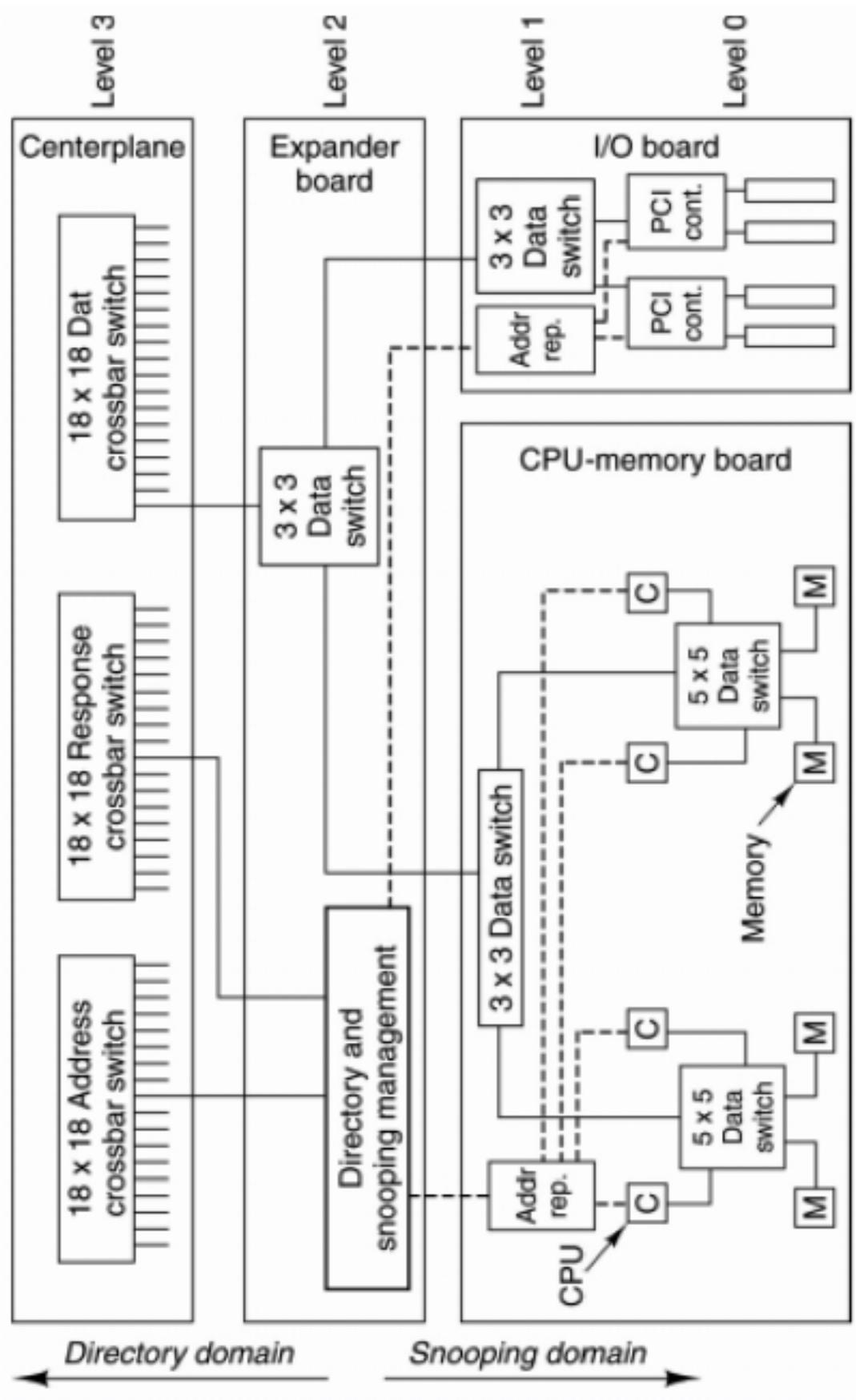
„Board Set“ besteht aus Slot 0 Board, Slot 1 Board und Expander Board. Letzteres nimmt die beiden anderen Boards auf.

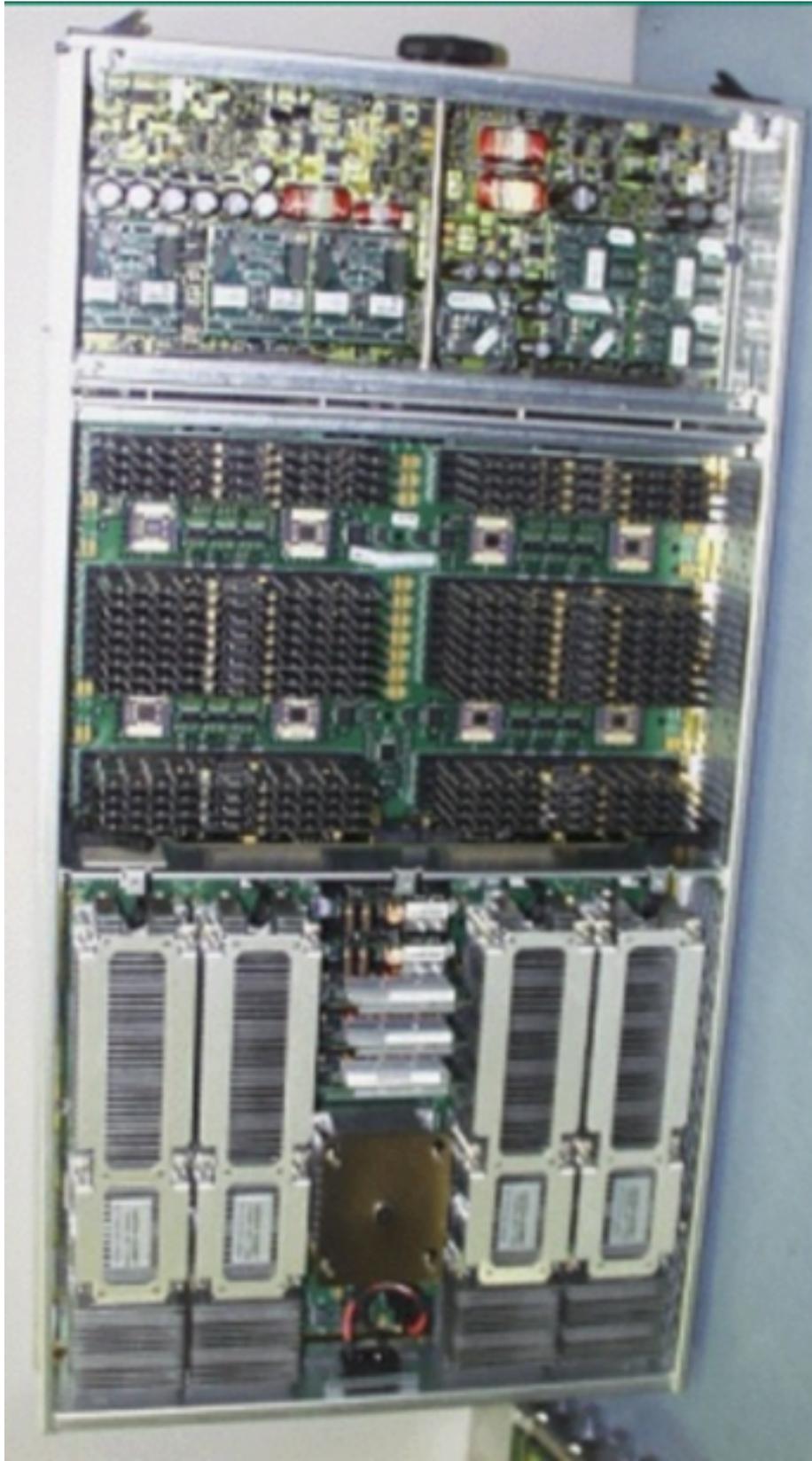
Slot 0 Board ist entweder CPU/Memory Board (System Board, 18 max) oder System Controller Board (1 oder 2 max, nicht gezeigt).

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.

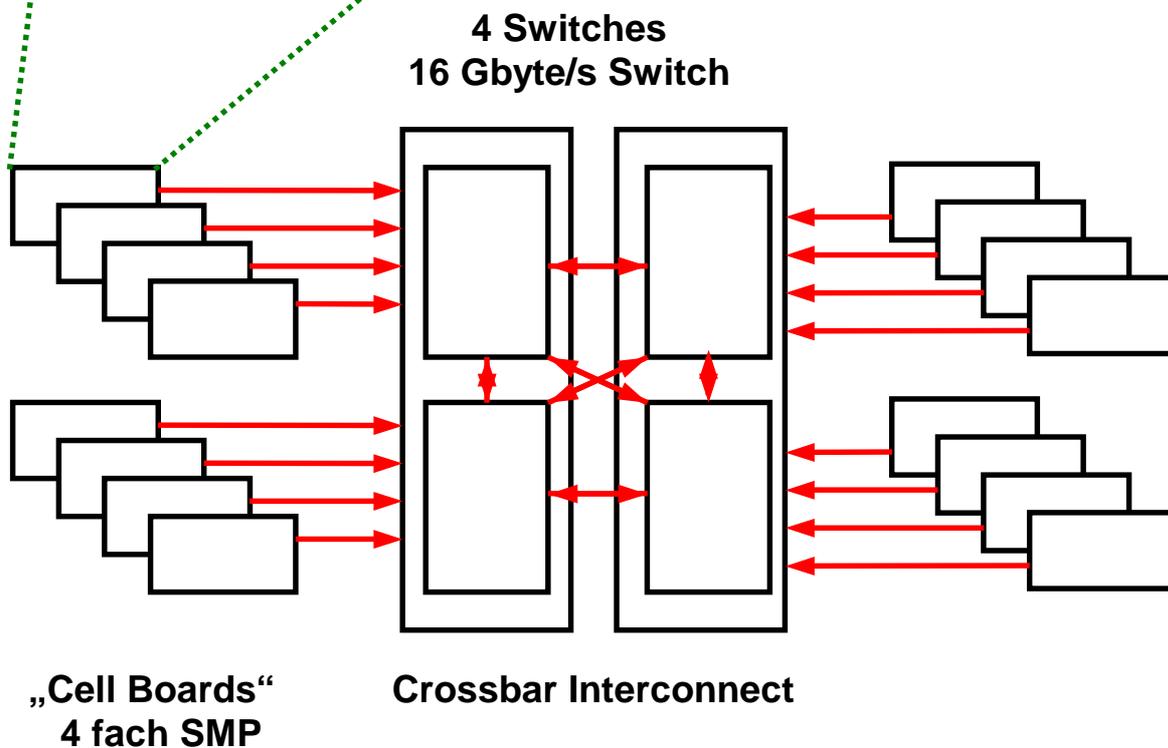
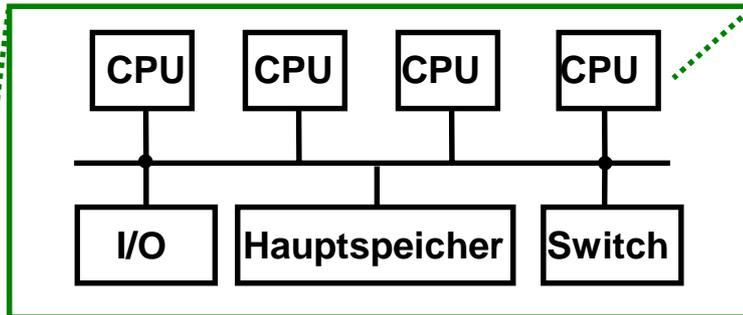
Sun
Micro-
systems
E25K







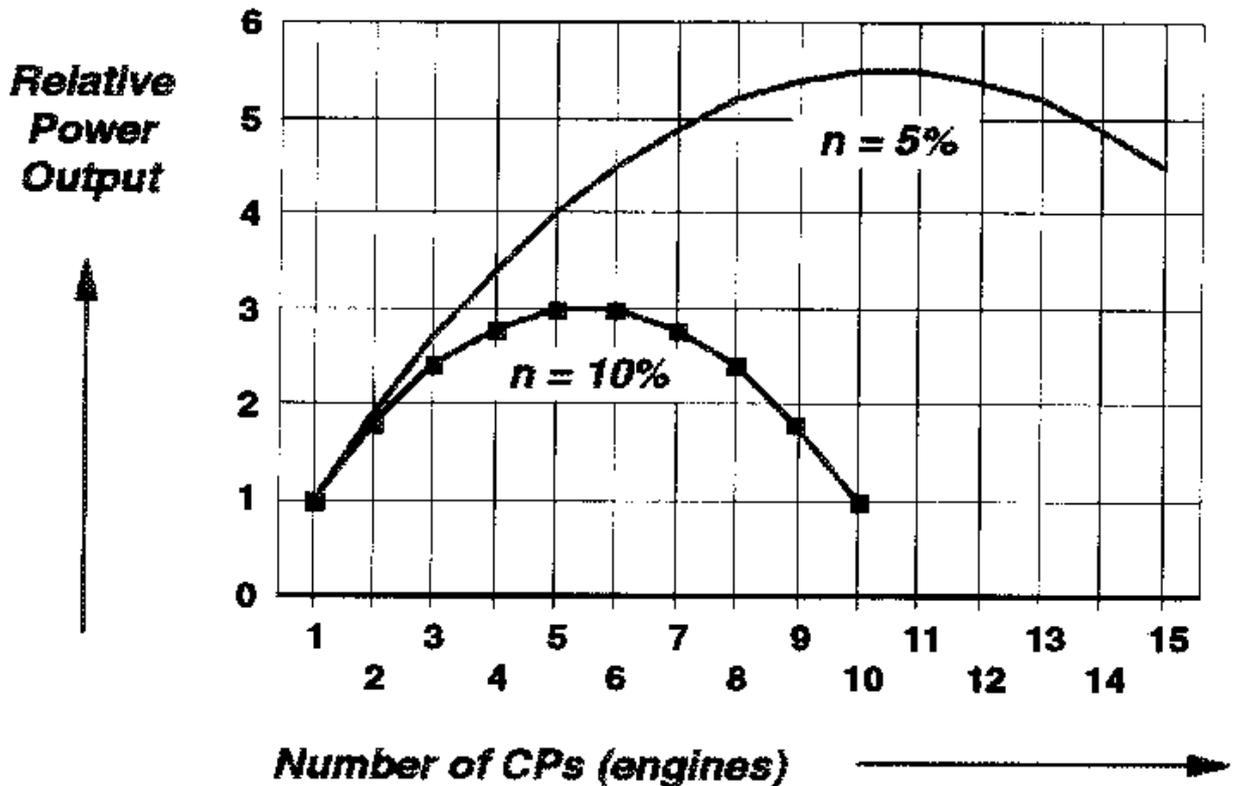
HP Superdome Cell Board



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

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 $n \ll 2\%$. 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

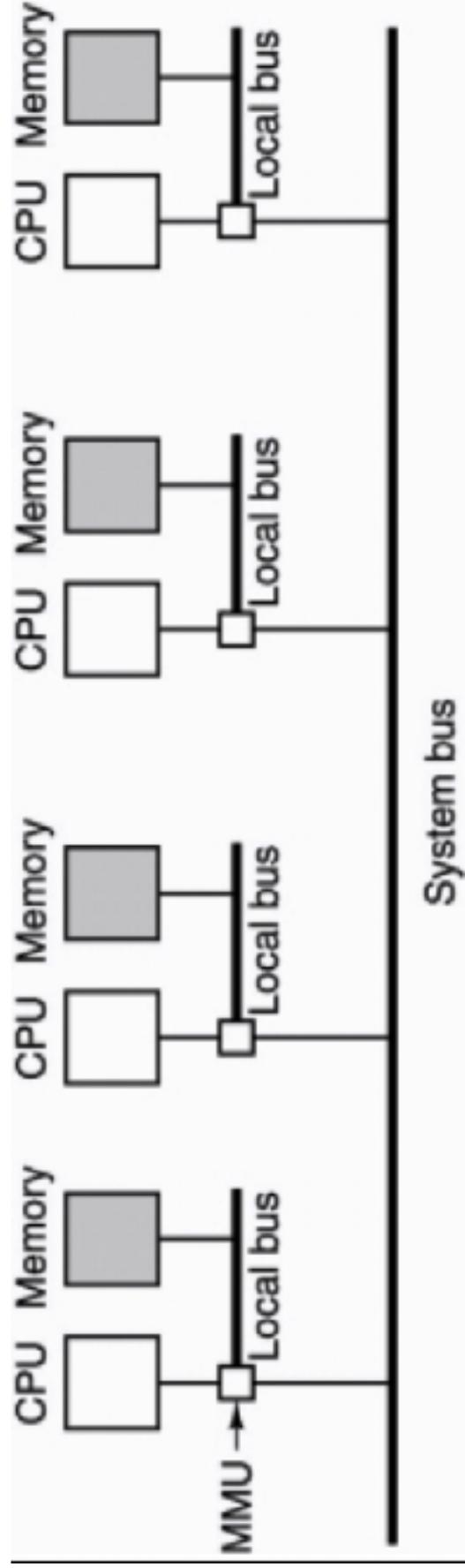
**verfügbar seit 1965, ständig erweitert und angepaßt
an realistischen Anwendungsprofilen orientiert
anwendbar für Rechner unterschiedlicher Hersteller**

Berücksichtigt

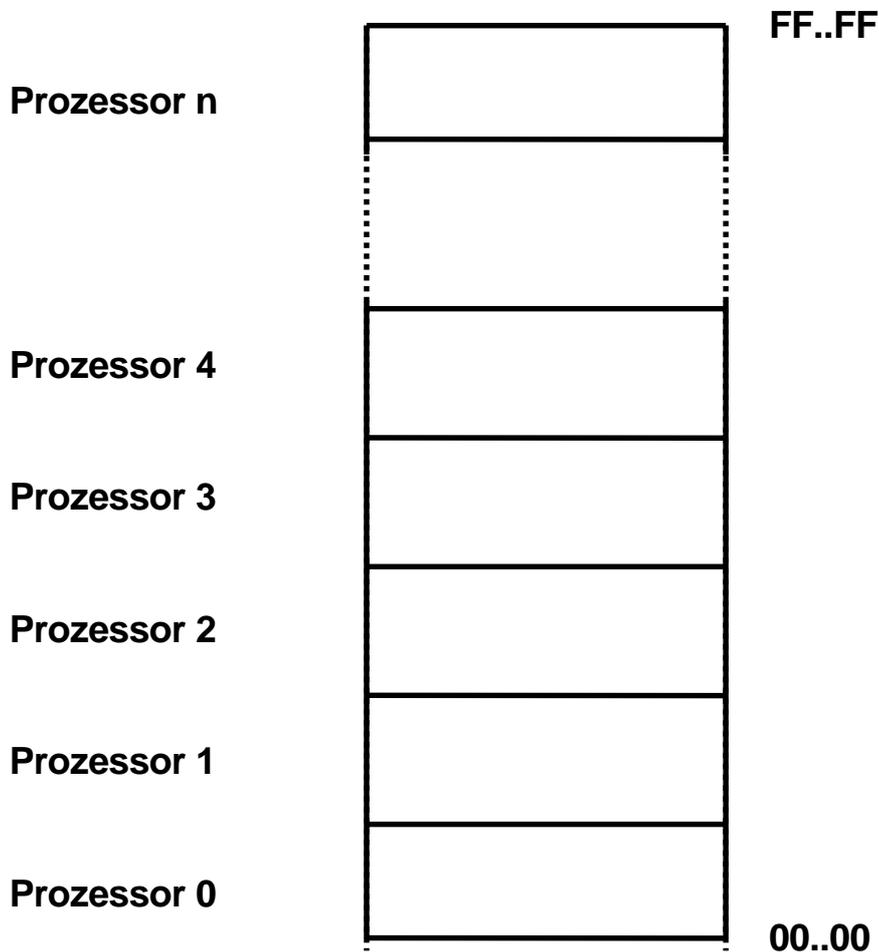
**Häufigkeit der einzelnen Maschinenbefehle
Cache- und Hauptspeichierzugriffszeiten
Bus Latency/Contention
Cache Misses und Cache-Line reload
Memory Refresh**

Benchmark besteht aus S/390 Maschinenbefehlen, daher

**nicht portierbar auf Rechner anderer Architektur
(Problem der „S/390 äquivalenten MIPS)**



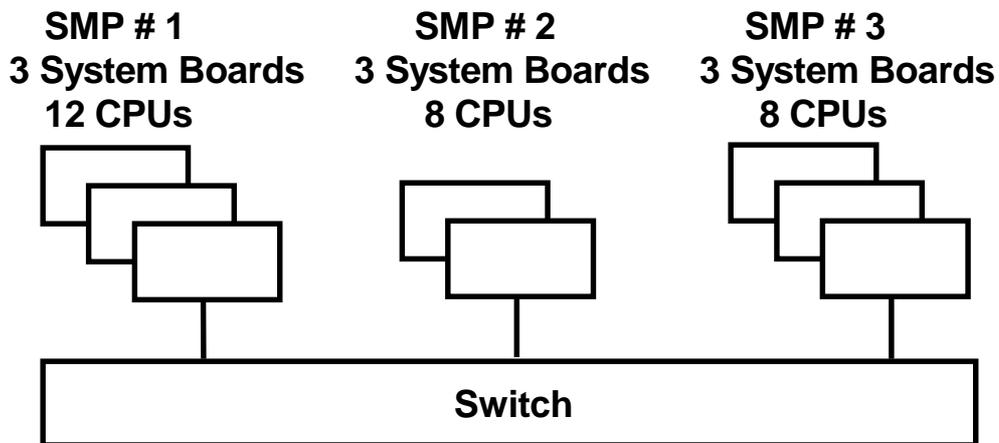
NUMA Rechner mit zwei Ebenen von Bussen



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 Sun Fire oder HP Superdome Servers mit 8 System Boards in mehrere parallel laufende SMPs

Harte Partition. 4 CPUs pro System Board

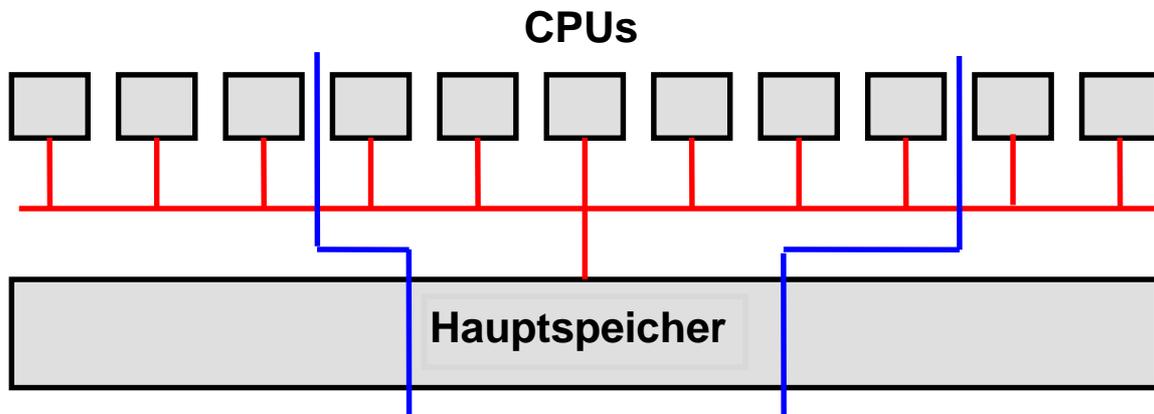
Jeder SMP hat ein eigenes Betriebssystem-

Aufteilung erfolgt statisch

System Administrator kann während des laufenden Betriebes die Zuordnung der System Boards zu den einzelnen SMPs ändern

Für Transaktionsanwendungen realistisch kaum mehr als 3 System Boards (12 CPUs) pro SMP

(Ausnahme: z/OS kann 24 – 32 CPUs in einem SMP betreiben)



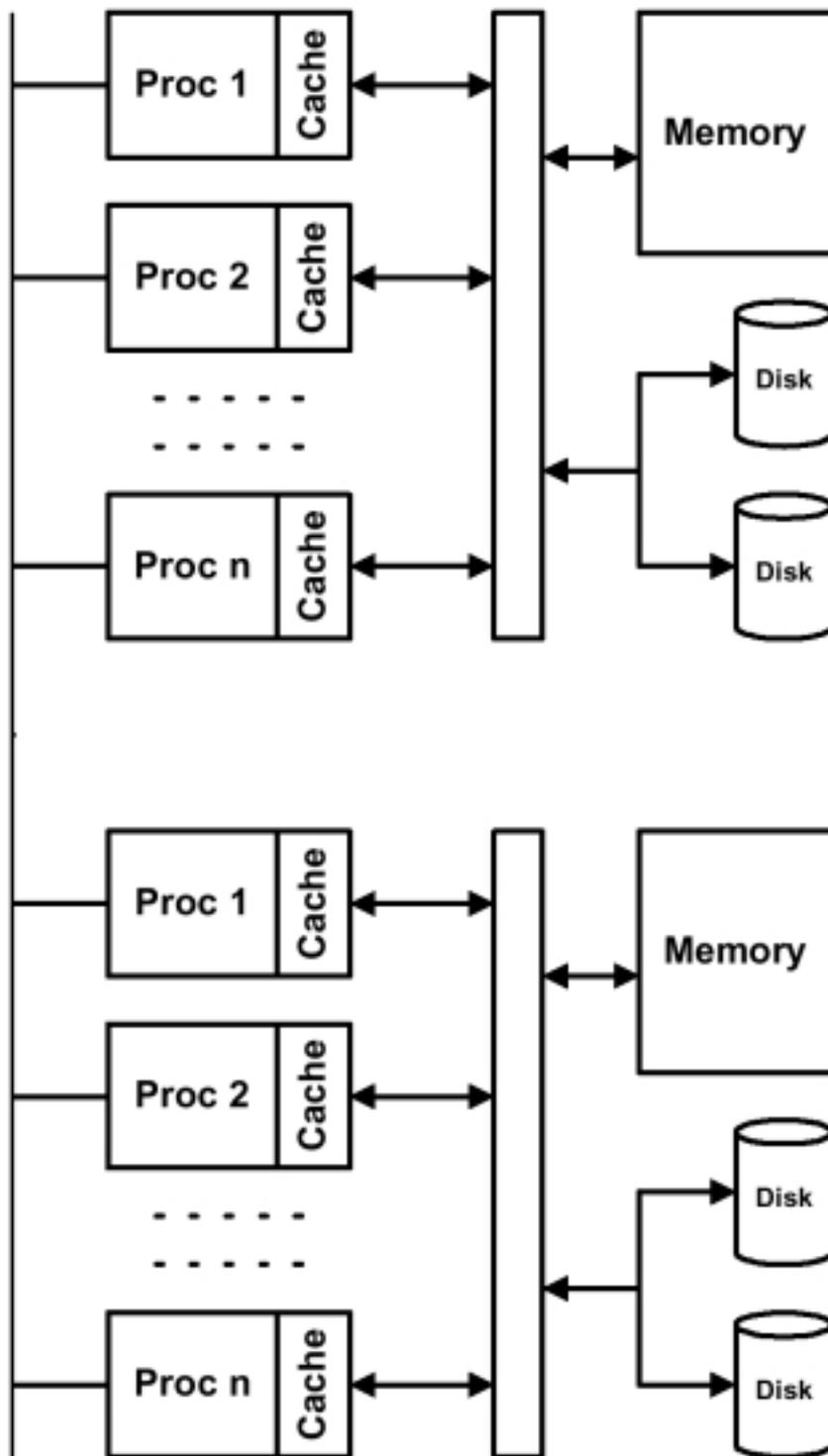
Aufteilung eines Großrechners in mehrere SMPs

z/OS unterstützt symmetrische Multiprozessoren (SMP) mit bis zu 24 – 32 CPUs. Bei Unix, Linux und Windows Betriebssystemen liegt die Grenze für Transaktions- und Datenbank Anwendungen eher bei 12 CPUs.

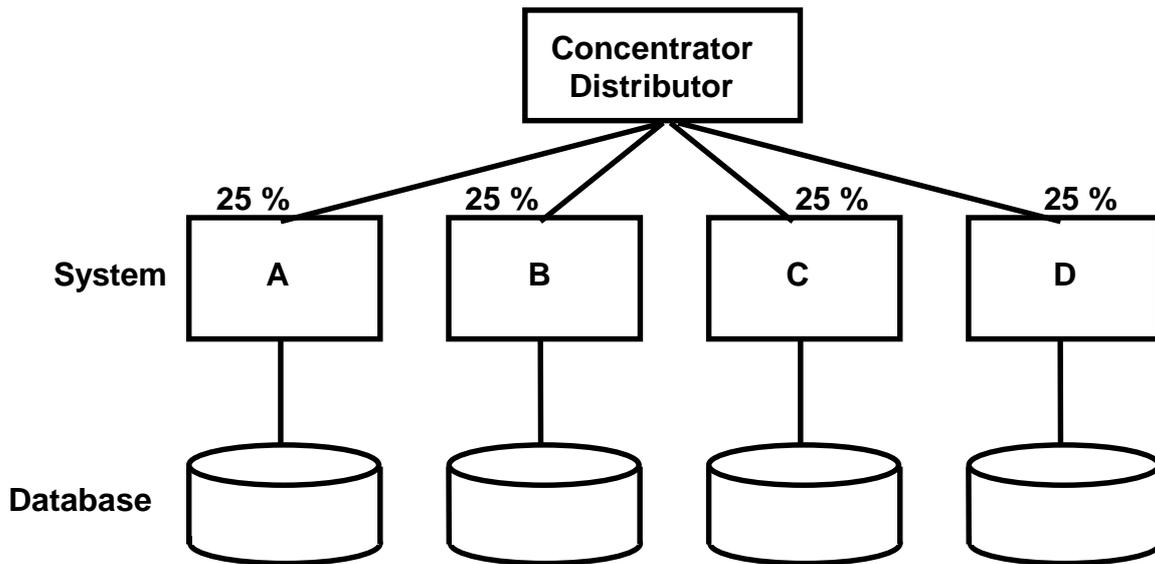
Moderne Großrechner (Systeme) verfügen über wesentlich mehr CPUs. Sie werden deshalb in mehrere SMPs aufgeteilt, die über einen zentralen Switch miteinander kommunizieren.

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

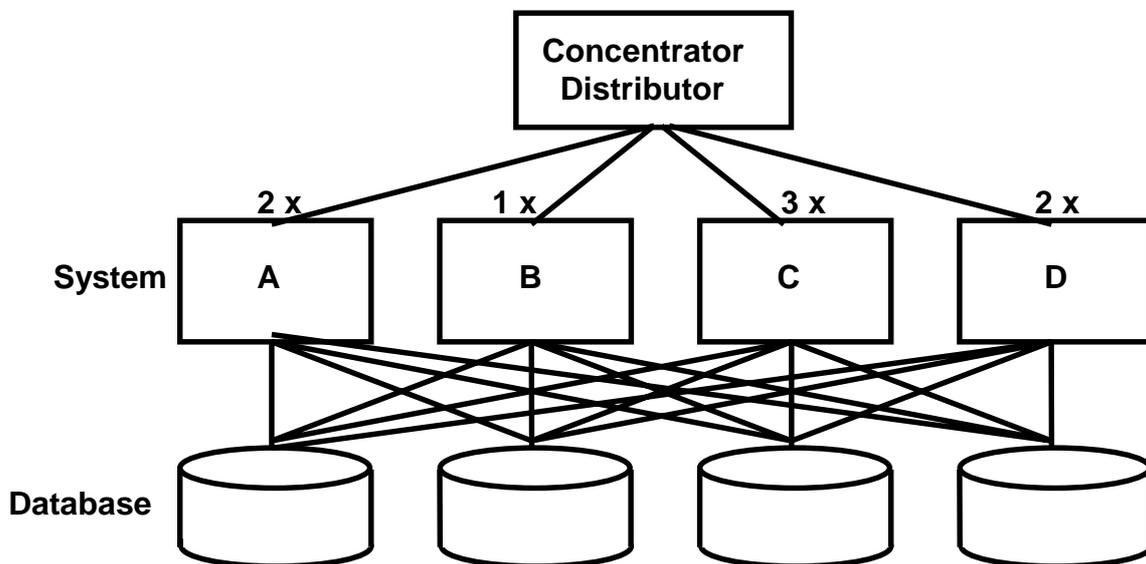


Shared Nothing Modell



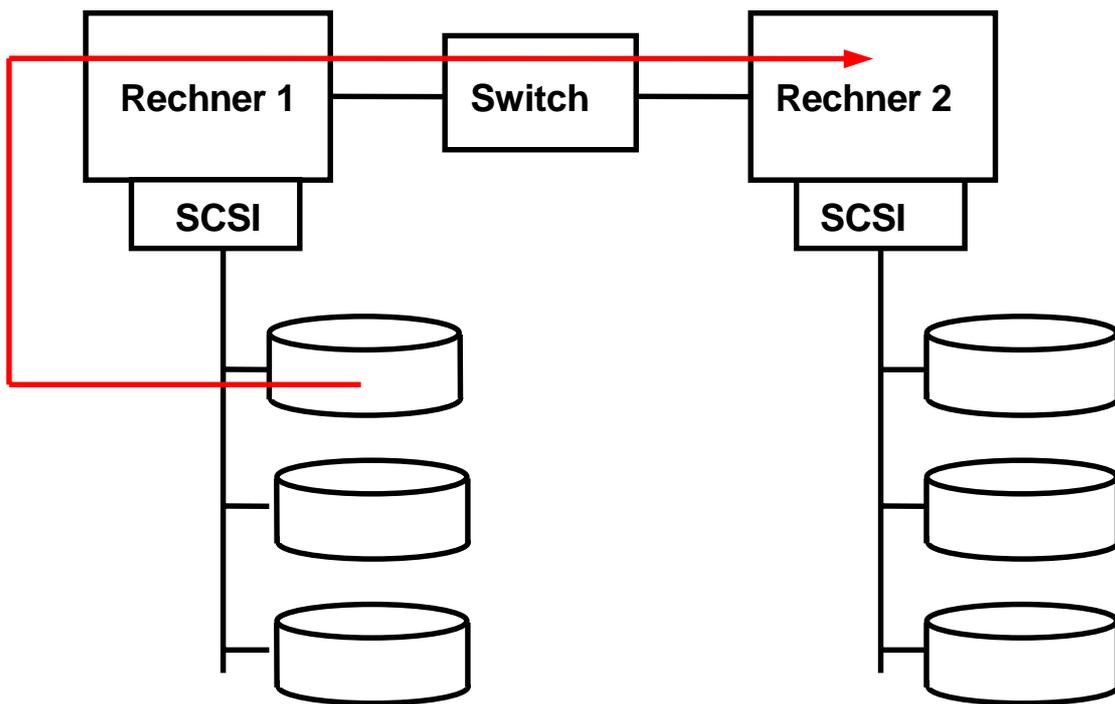
Shared nothing (partitioned data)

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



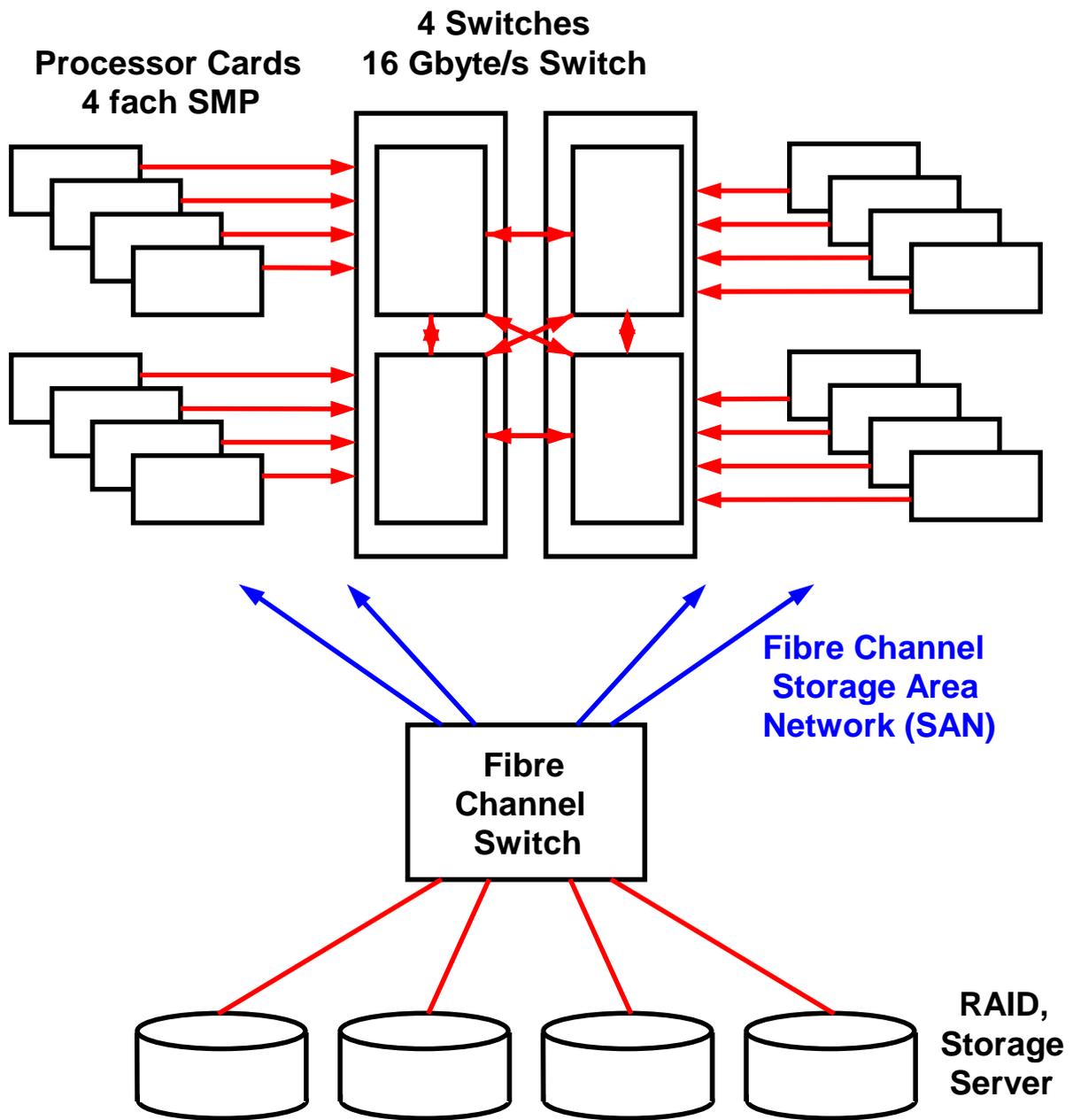
Shared data (shared disk)

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



Shared Disk Emulation

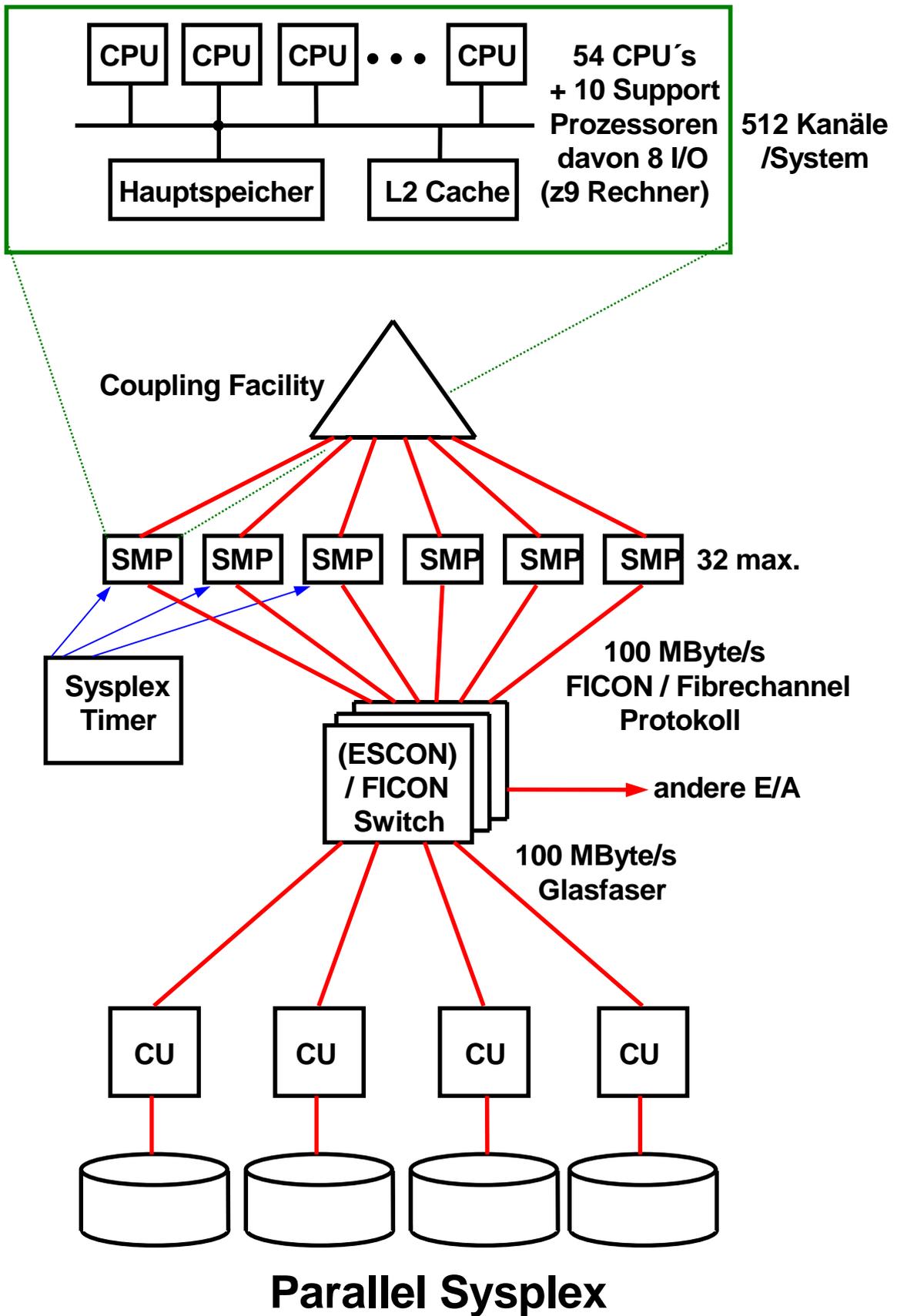
Rechner 2 bittet Rechner 1, die gewünschten Daten zu übertragen



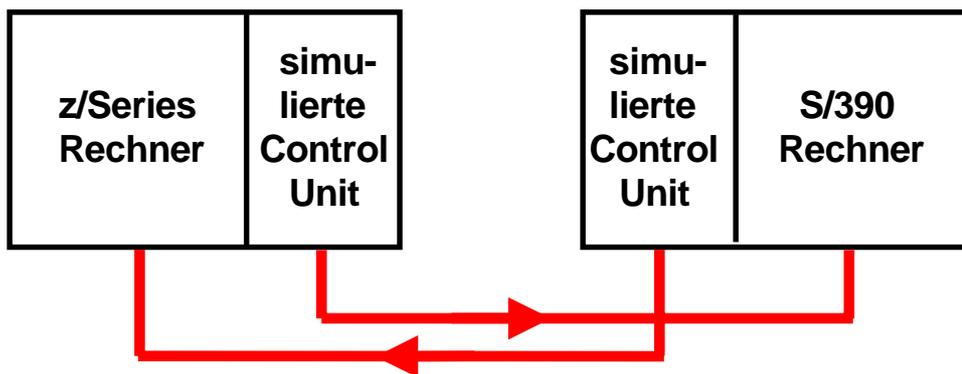
HP Superdome Cluster

64 CPU's

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



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.

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)

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.

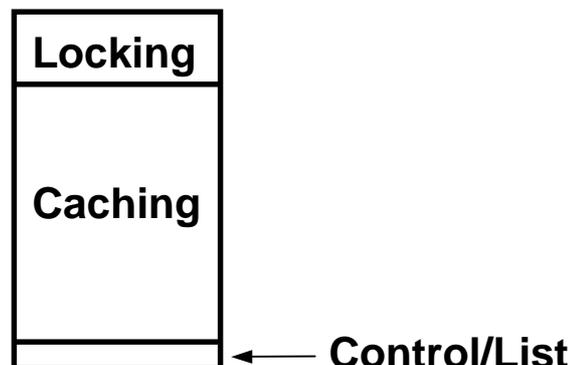
**Jedes System hat bis zu 4 Channel Subsystems
Jedes Channel Subsystem hat bis zu 256 Kanäle
Jeder FICON Switch hat bis zu 256 Ports
Bis zu 8 Pfade pro Control Unit**

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.

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 die Leistungsfähigkeit der Speicherverwaltung und des E/A Systems bestimmt.

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

Trotzdem besteht Einigkeit, dass die E/A Leistung von Mainframes in Bezug auf

- **gleichzeitig aktiver Plattenspeicher und der**
- **gesamten übertragenen Datenrate**

bei Mainframe Rechnern deutlich größer als bei allen anderen plattformen ist.

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), FC-AL Vorläufer

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

Der Unterschied zwischen FCP und FC-AL ist gering. FCP und SAS unterscheiden sich in der Verwendung von Glasfaser bzw. Kupferkabel Anschlüssen. Obwohl FCP und SAS beides (sehr ähnliche) serielle SCSI Protokolle verwenden, wird der Ausdruck „Serial SCSI“ ausschließlich für SAS Platten verwendet.

SAS Platten werden auch als Nearline-Platten bezeichnet. FCP Platten können dank des Glasfaser anschlusses über größere entfernungen angeschlossen werden.

FATA

Fibre Channel ATA

In der Praxis verwenden SCSI Platten bessere mechanische und elektronische Komponenten als ATA Platten. Dies bewirkt

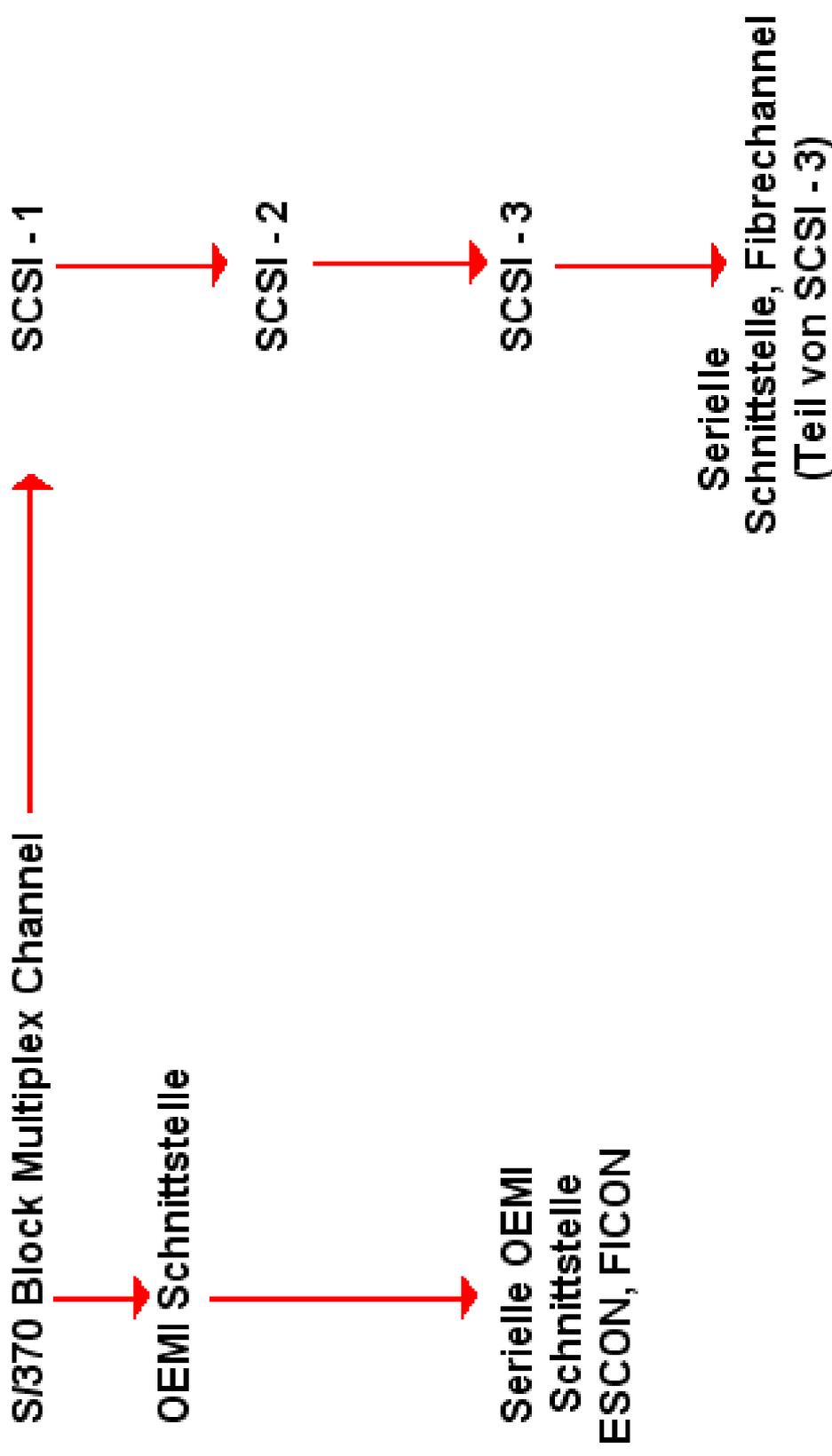
- **schnellere Zugriffszeiten**
- **höhere Zuverlässigkeit**
- **deutlich höhere Kosten**

Aus Zuverlässigkeitsgründen verwenden SCSI Platten selten die neueste Plattenspeichertechnologie. Dies ist einer der Gründe, warum für einen PC Platten mit einer höheren Speicherkapazität erhältlich sind als dies im Mainframe Bereich der Fall ist.

FATA-Laufwerke sind ATA-Plattenlaufwerke für Fibre Channel (FC). Sie arbeiten mit den mechanischen Komponenten von ATA-Festplatten, jedoch mit vorgeschalteter FC-Schnittstelle. In anderen Worten, es sind SATA Platten, bei denen die elektrische serielle ATA Schnittstelle durch eine Fibre Channel Schnittstelle ersetzt wurde. Die FATA-Technologie hat den Vorteil, dass sie in einer Mischung mit anderen FC-Laufwerken betrieben werden können.

FATA Platten werden auch als Nearline-Platten bezeichnet. Sie werden im Großrechnerbereich dann eingesetzt, wenn Zuverlässigkeit und Zugriffszeit weniger wichtig sind, z.B. um Bilddateien (Images) zu archivieren.

Historische Entwicklung des Peripherie-Busses



Fibre Channel (FC)

Fibre Channel ist für serielle, kontinuierliche Hochgeschwindigkeitsübertragung großer Datenmengen konzipiert worden. Die erreichten Datenübertragungsraten liegen heute bei 4 Gbit/s und 8 Gbit/s, was im Vollduplex-Betrieb für Datentransferraten von 800 MB/s ausreicht.

Als Übertragungsmedium findet man gelegentlich Kupferkabel (hauptsächlich innerhalb von Storage-Systemen; überbrückt bis zu 30 m), meistens aber Glasfaserkabel. Letzteres erfolgt meist zur Verbindung von Rechnern mit Storage-Systemen oder aber von Storage-Systemen untereinander. Hierbei werden Entfernungen bis zu 10 km überbrückt. Der Zugriff auf die Festplatten erfolgt blockbasiert.

Es können generell zwei Arten von Fibre-Channel-Implementierungen unterschieden werden, die Switched Fabric, die meist als Fibre Channel, oder kurz FC-SW, bezeichnet wird und die Arbitrated Loop, kurz als FC-AL bekannt.

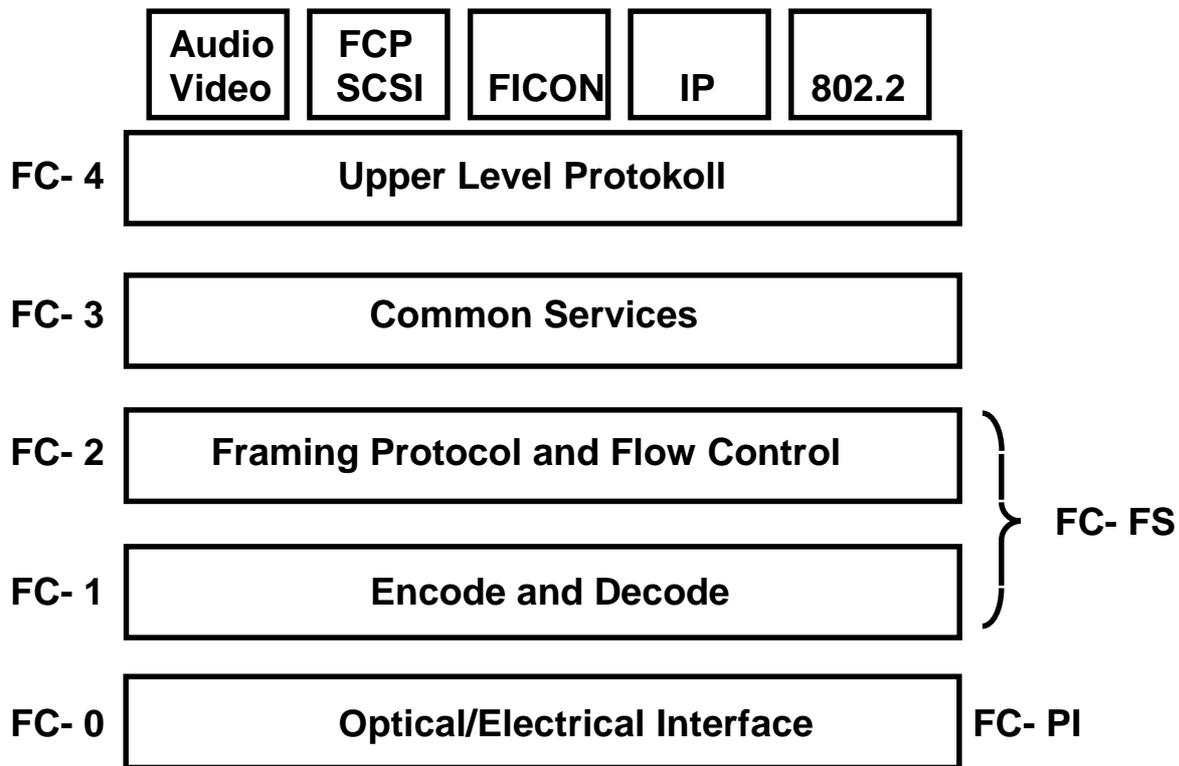
Fibre Channel Protocol (FCP) ist das Schnittstellenprotokoll von SCSI über den Fibre Channel.

Switched Fabric und Arbitrated Loop

Bei der Fibre Channel-Switched Fabric werden Punkt-zu-Punkt-Verbindungen (Point To Point) zwischen den Endgeräten geschaltet, beim Fibre Channel-Arbitrated Loop handelt es sich um einen logischen Bus, bei dem sich alle Endgeräte die gemeinsame Datenübertragungsrate teilen.

Bei der Fibre Channel-Switched Fabric handelt es sich um die leistungsfähigste und ausfallsicherste Implementierung von Fibre Channel. In den meisten Fällen ist Switched Fabric gemeint, wenn nur von Fibre Channel gesprochen wird. Im Zentrum der Switched Fabric steht der Fibre Channel Switch oder der Director. Über dieses Gerät werden alle anderen Geräte miteinander verbunden, so dass es über den Fibre Channel Switch möglich wird, direkte Punkt-zu-Punkt-Verbindungen zwischen je zwei beliebigen angeschlossenen Geräte zu schalten.

FC-AL erlaubt es, bis zu 127 Geräte an einem logischen Bus zu betreiben. Dabei teilen sich alle Geräte die verfügbare Datenübertragungsrate (bis 4 GBit/s). Die Verkabelung kann sternförmig über einen Fibre Channel Hub erfolgen. Es ist auch möglich, die Geräte in einer Schleife (Loop) hintereinander zu schalten (Daisy Chain), da viele Fibre-Channel-Geräte über zwei Ein- bzw. Ausgänge verfügen. Dies ist z.B. beim IBM DSS 8000 Enterprise Storage Server der Fall.

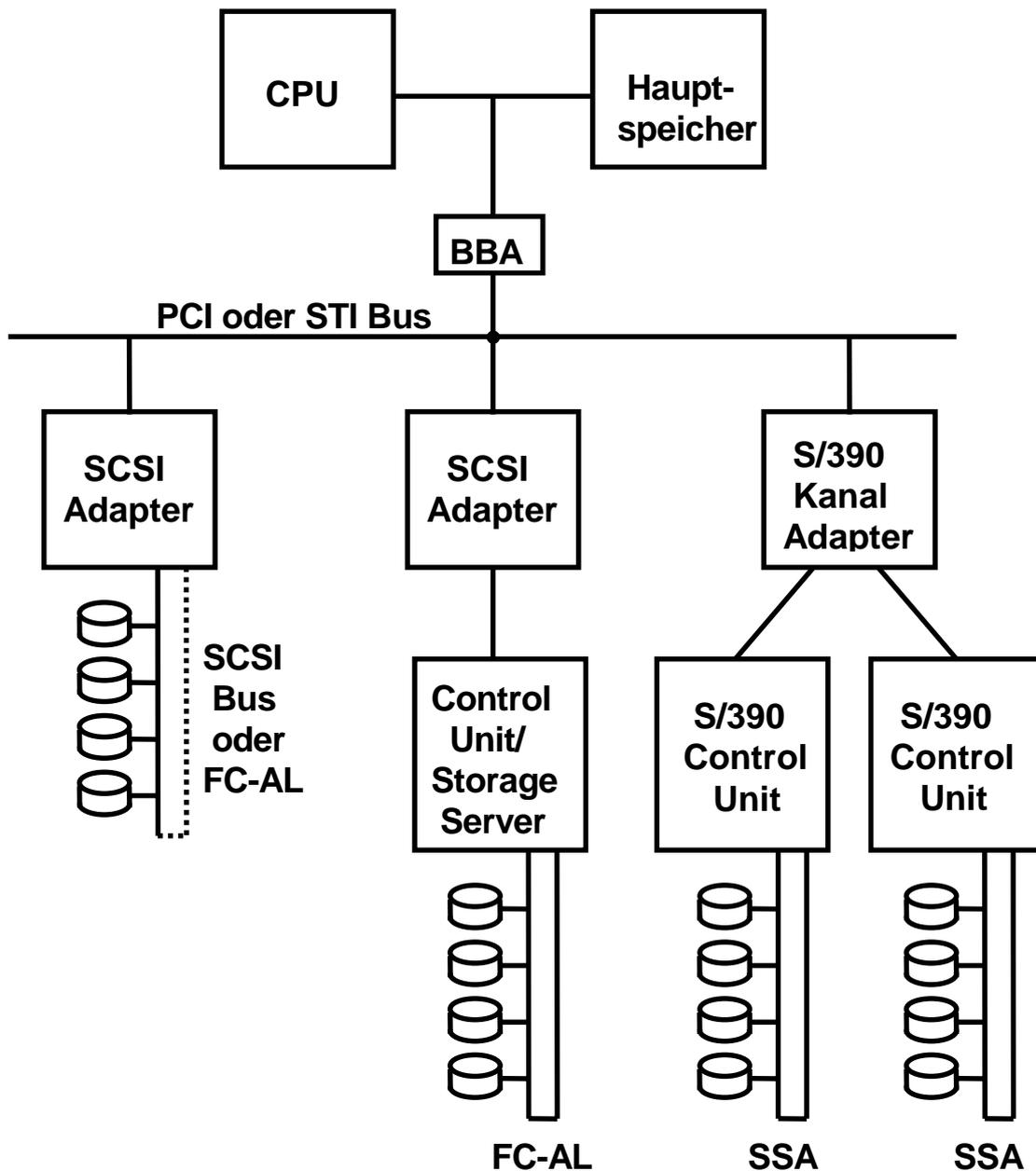


Fibre Channel Standard Architektur

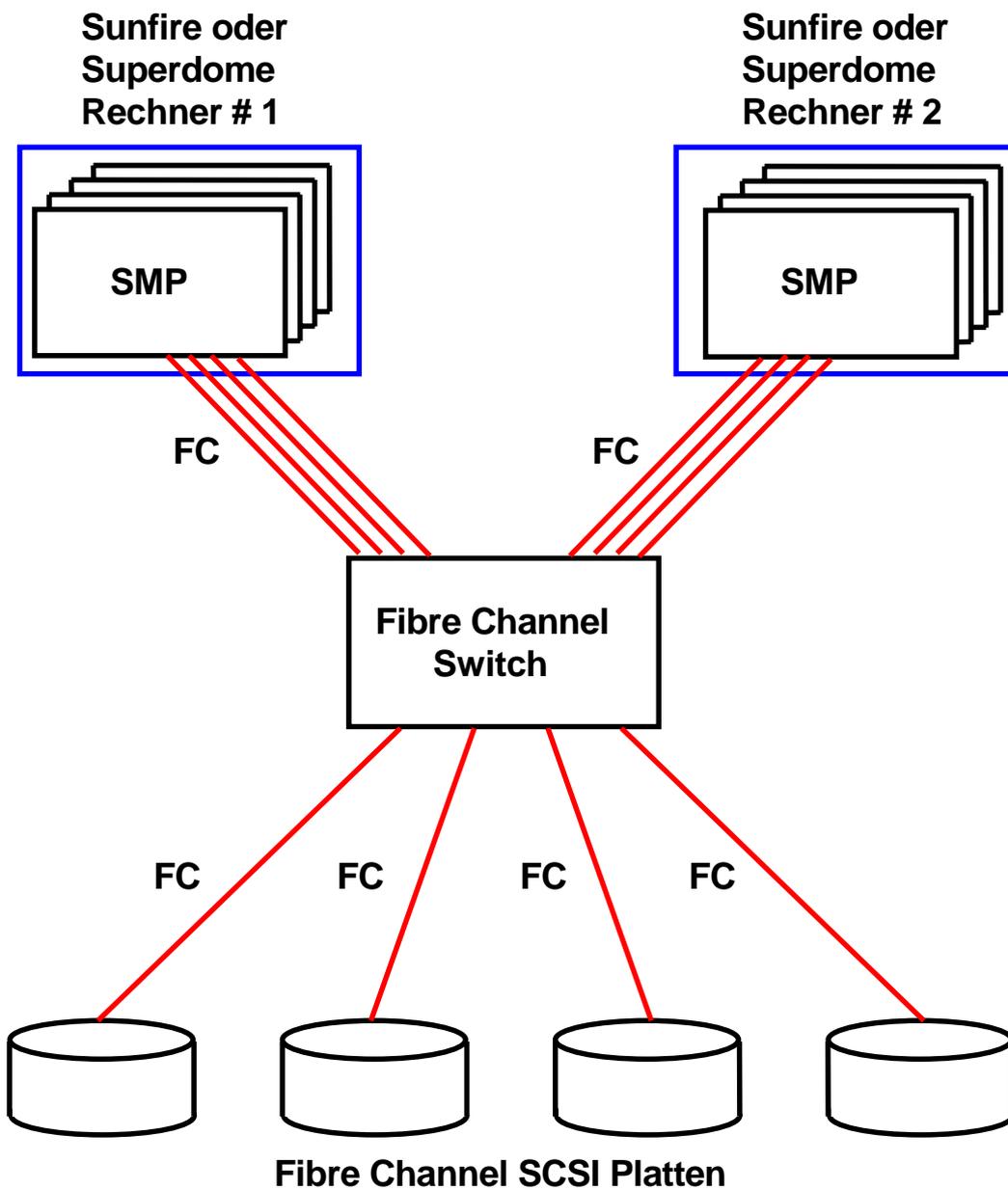
Die Fibre Channel Architektur verwendet ein Schichtenmodell vergleichbar mit (aber unabhängig von) den TCP/IP oder OSI Schichtenmodellen. Die unterste Schicht verwendet in den meisten Fällen optische Kabel. Wichtig ist besonders die oberste Schicht FC – 4.

Hierüber ist es möglich, unterschiedliche Protokolle zu betreiben. FC SCSI ist eine serielle Form des SCSI Protokolls, die über Fibre Channel Verbindungen erfolgt. (Das als „Serial SCSI“ bezeichnete Protokoll benutzt keinen Fibre Channel).

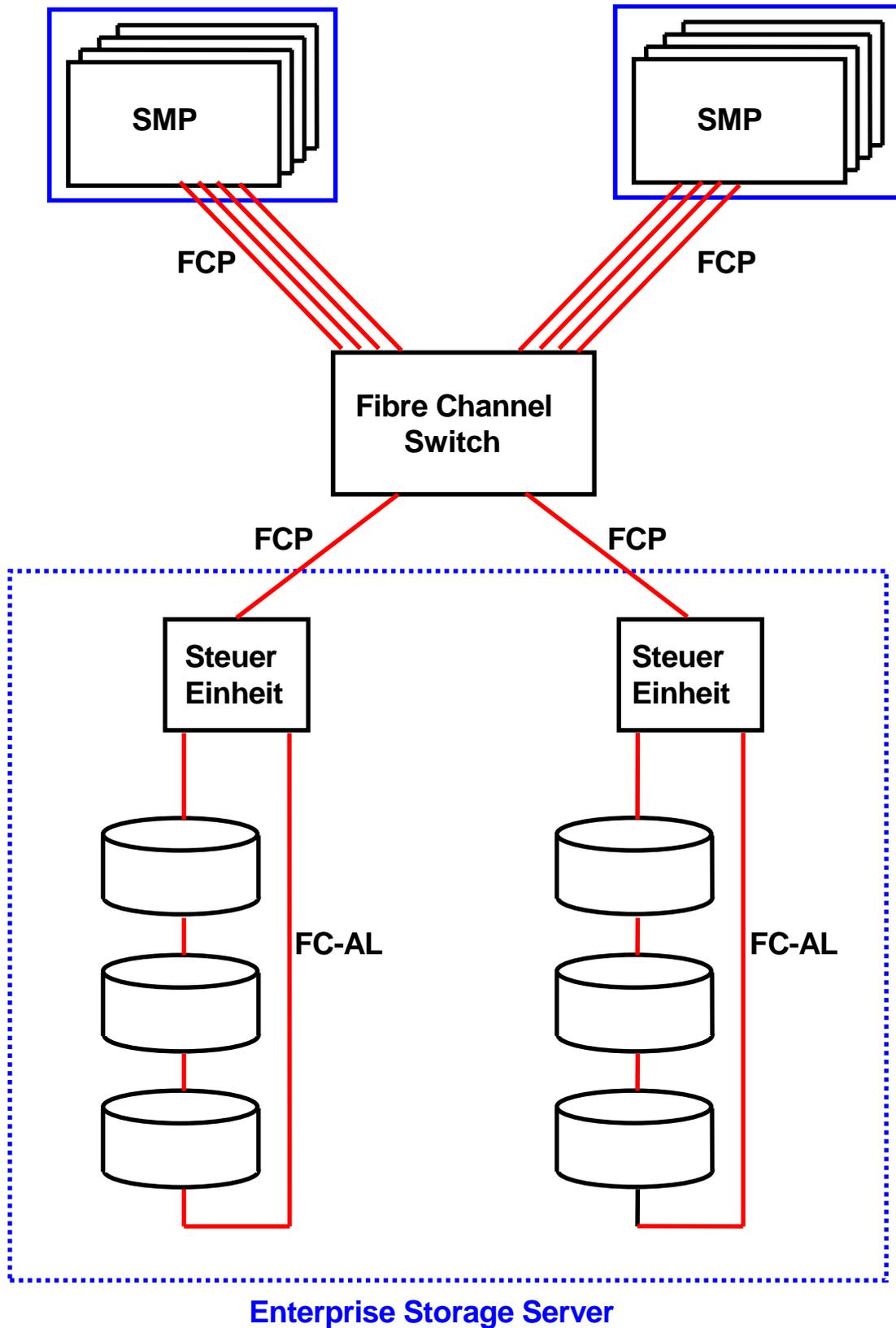
FICON ist das universell von Mainframes eingesetzte Protokoll, um Rechner miteinander und mit I/O Geräten zu verbinden.



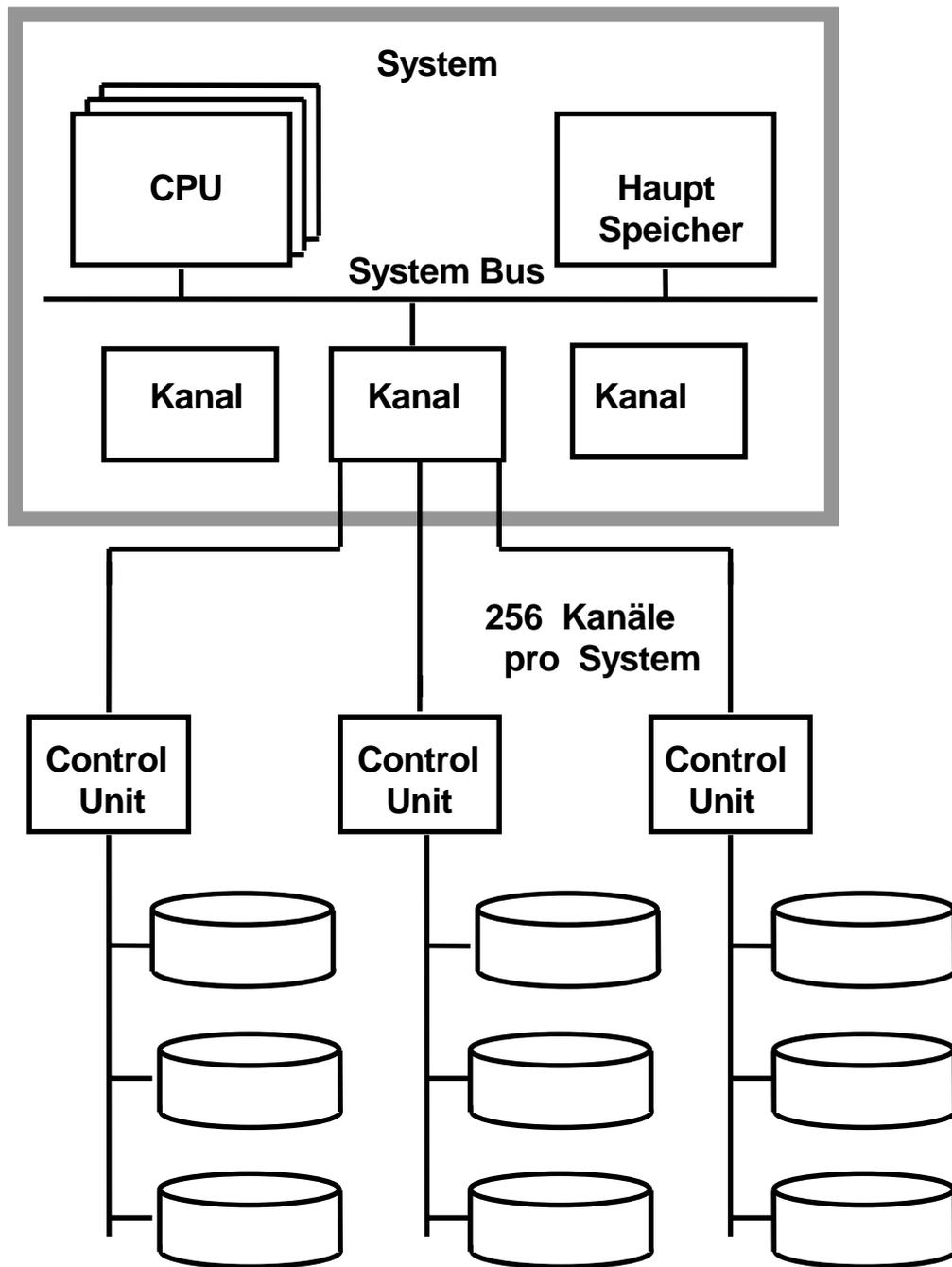
Plattenspeicher Anschlußalternativen



Einfache Fibre Channel Konfiguration

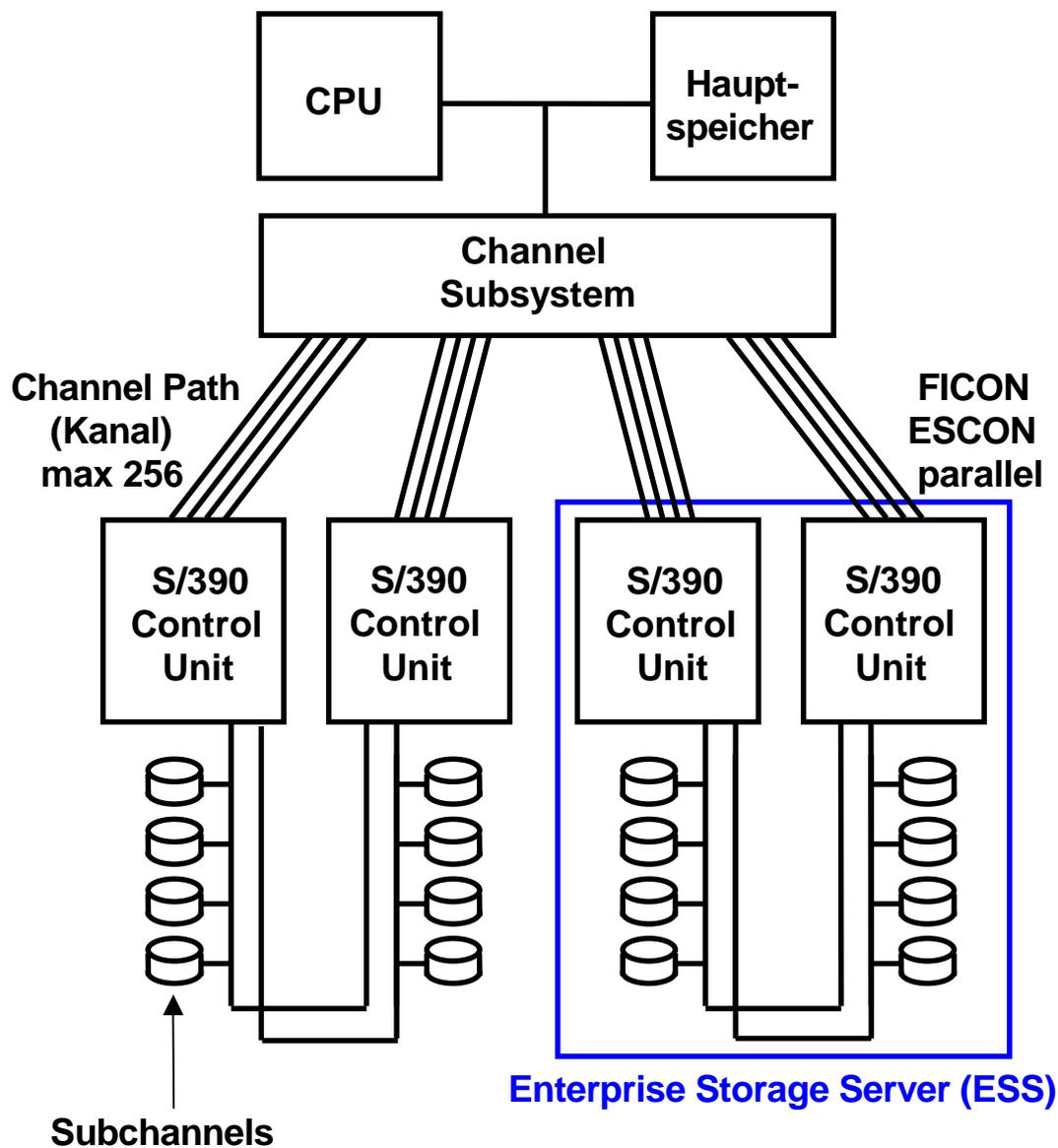


RAID, Cache Funktionalität



S/390 E/A Konfiguration

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



zSeries und S/390 Plattenspeicher Anschluß

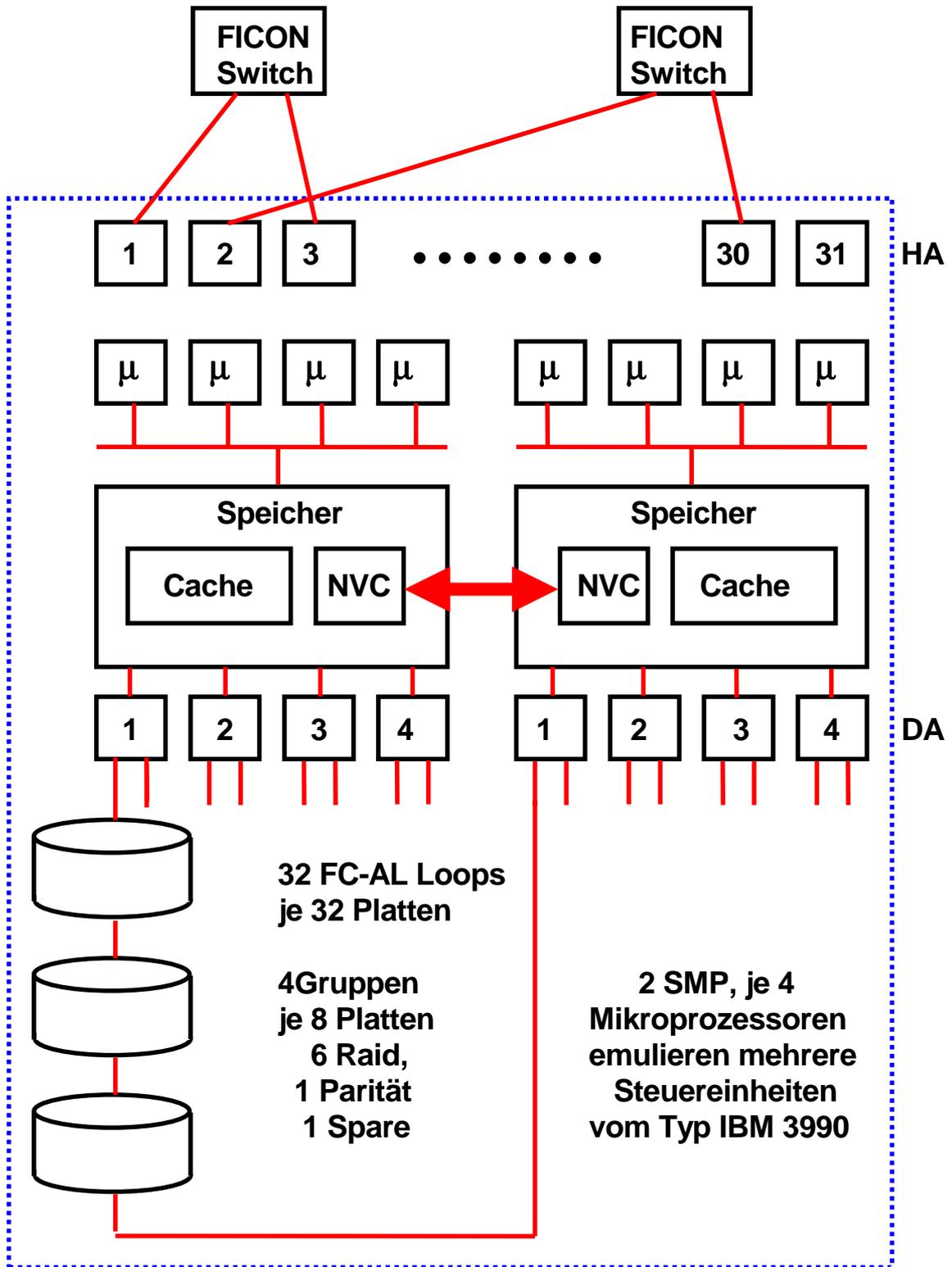
Das Channel Subsystem wird durch mehrere Prozessoren (als System Assist Prozessoren, SAP, bezeichnet) und entsprechenden Code verwirklicht. Die SAPs greifen parallel zu den CPUs auf den Hauptspeicher zu und entlasten diese von Ein-/Ausgabe Aufgaben.

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 FCP 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 FC-AL 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 bedeutendsten 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 2008

2 Cluster Prozessoren, je 4 x SMP

**4 FC/FICON Hostanschlüsse (Ports) pro Host Adapter,
je 4-Gbit/s, 400 MByte/s pro Port (4/5 code)**

**32 Host Adapter, 128 FC/FICON Hostanschlüsse total,
154.000 I/O-Operationen pro Sekunde (IOPs) pro Adapter**

4,9 Millionen IOPs pro DS 8300

256 GByte Cache

**sechzehn FC-AL Geräteadapter (Device Adaptor),
je 4-Port pro Geräteadapter**

64 Ports, 32 FC-AL Loops

1024 Laufwerke (Disk)

je 450 GByte FC-AL, 460 TByte total

500 GByte FATA, 512 TByte total

32 Laufwerke/Loop

RAID 5, 6 oder 10 (Redundant Array of Independent Disks)

Remote-Mirroring

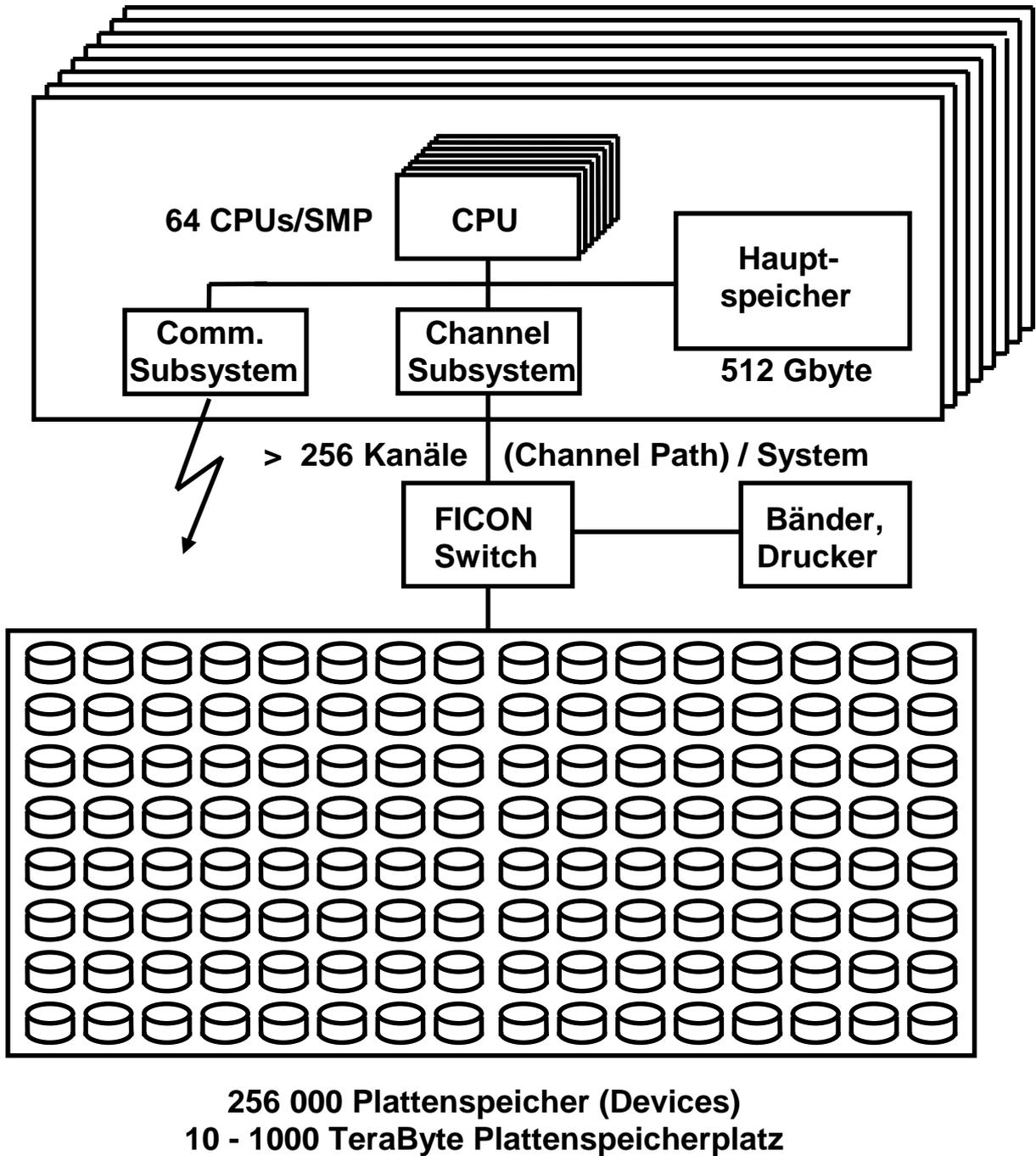
<http://www-03.ibm.com/systems/de/storage/disk/ds8000/index.html>

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.

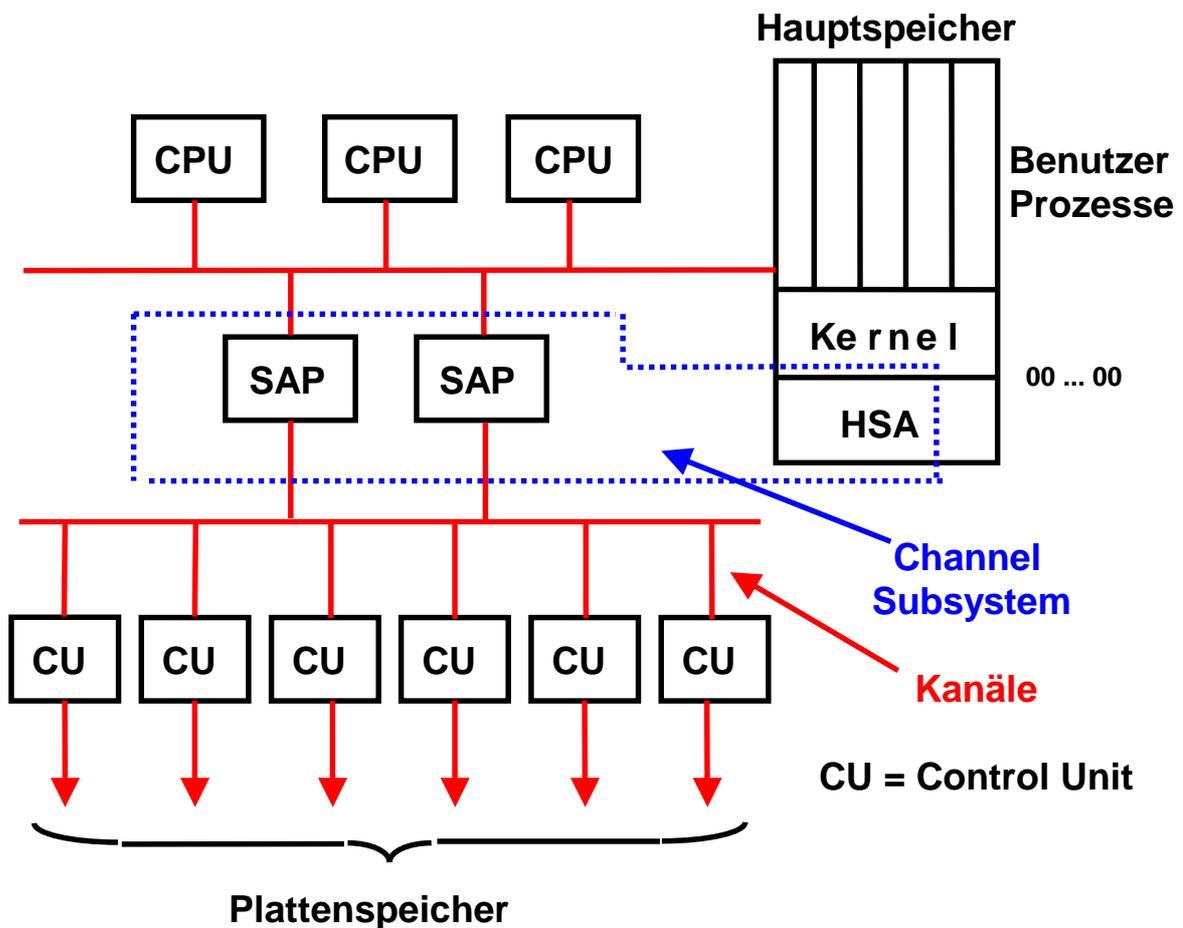
Enterprise Storage Server werden von vielen Firmen angeboten, meistens sowohl mit SCSI-FCP als auch mit FICON Anschlussmöglichkeiten: EMC, Hitachi/Sun, MaxData, andere.

32 Systeme (SMPs)



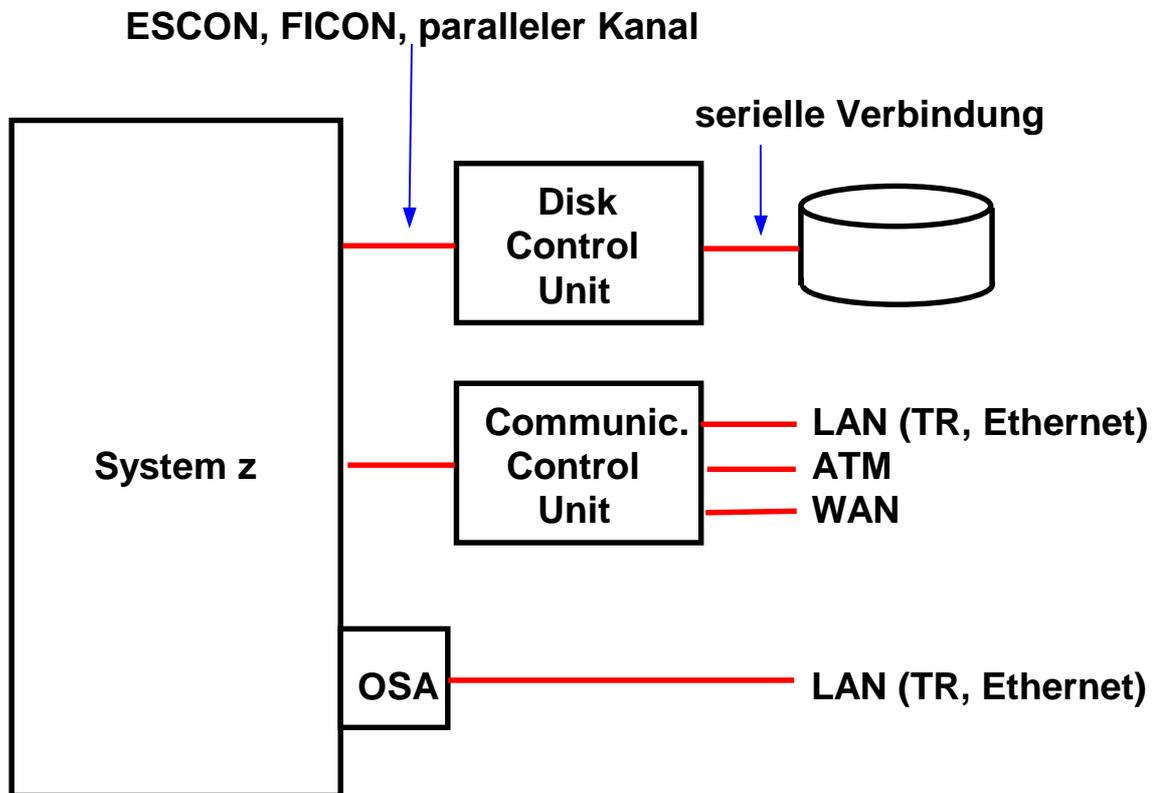
z Series (S/390) Großsystemkonfiguration

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 umfangreiche Optimierungen der Plattenspeicherzugriffe möglich.



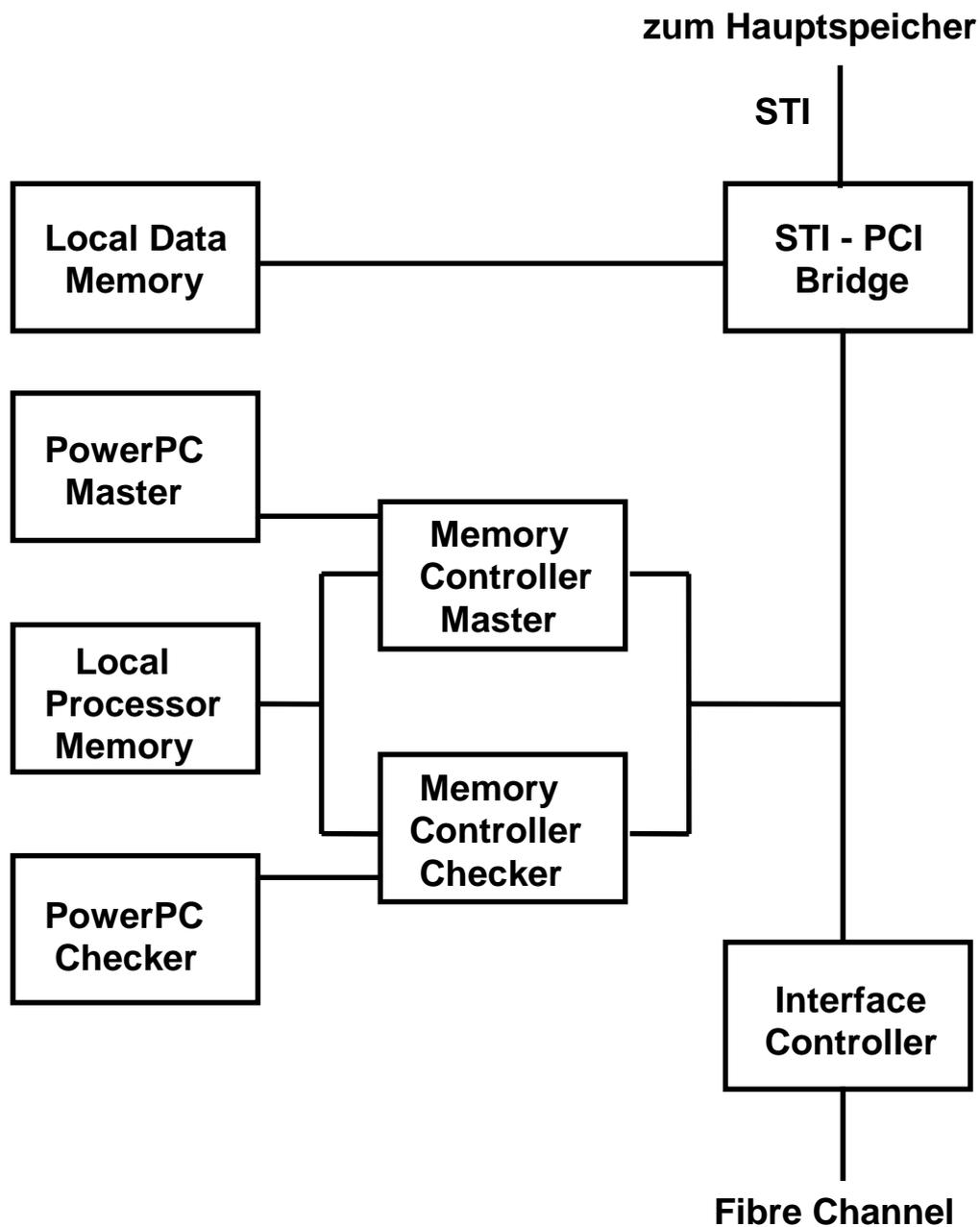
S/390 E/A Konfiguration

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.

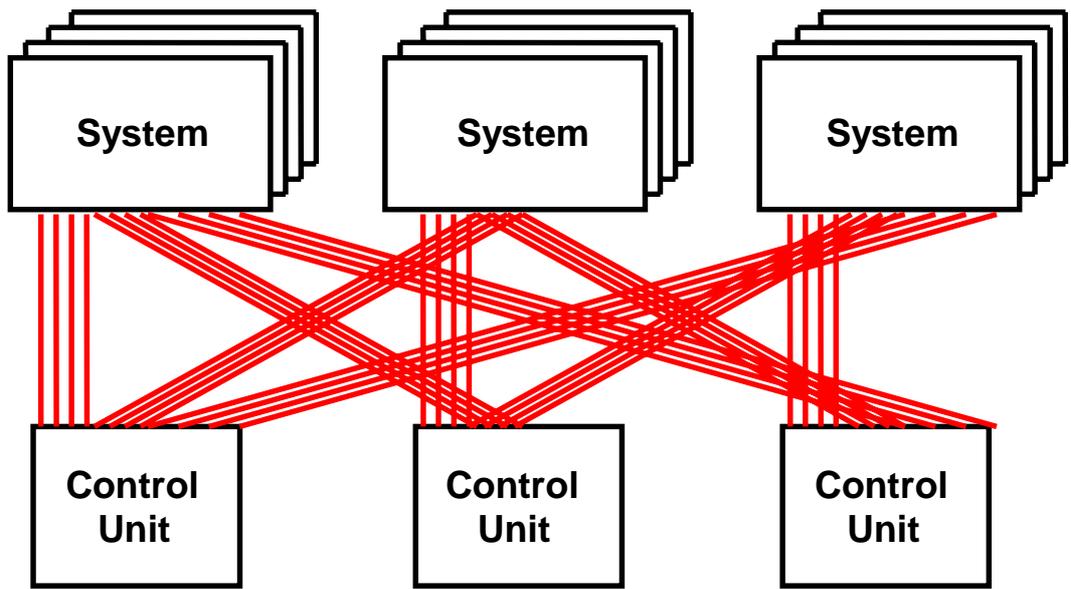
Es existieren viele unterschiedliche Typen von Steuereinheiten. Die wichtigsten schließen externe Speicher (Platten, Magnetbänder Archivspeicher) und Kommunikationsleitungen an.

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

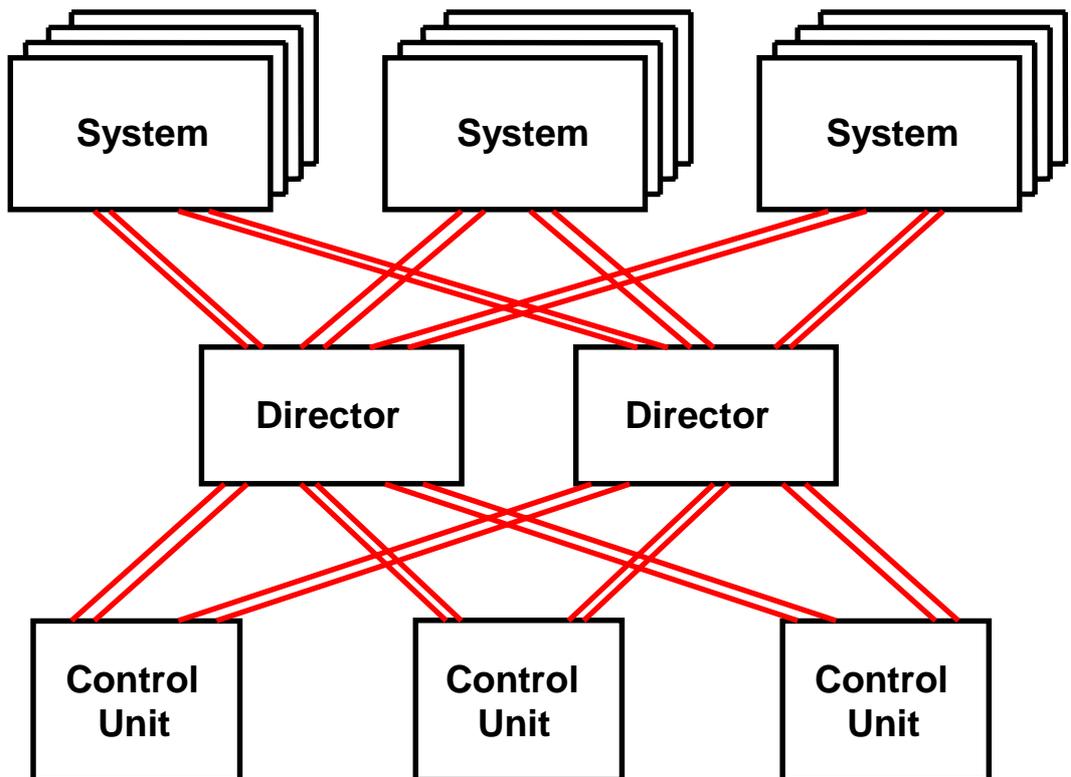
Einige Steuereinheiten können in den S/390 Rechner integriert werden. Das wichtigste Beispiel ist der OSA Adapter für den Anschluß von LAN's.



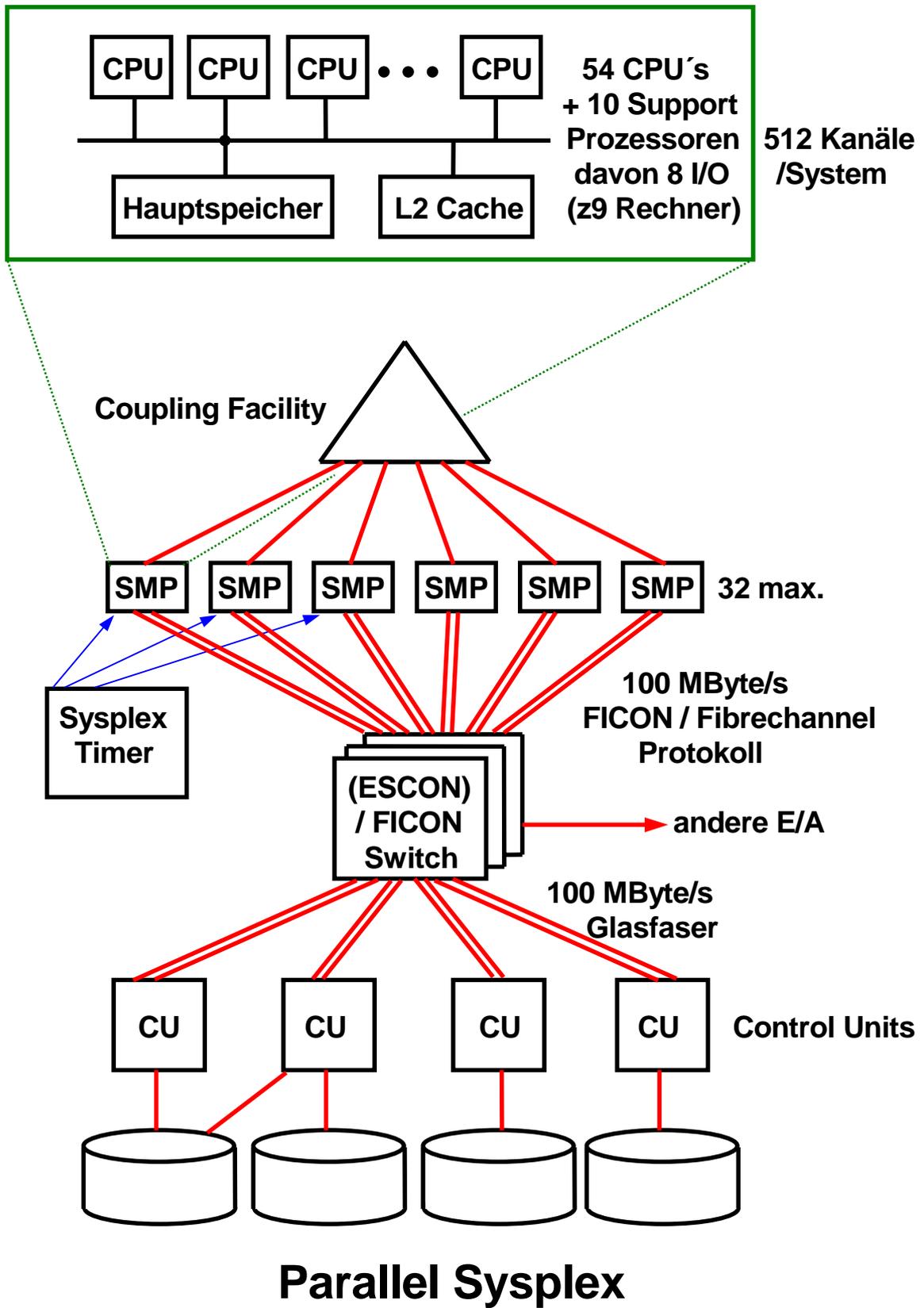
zSeries Fibre Channel Kanal basierend auf der Common I/O Card



Parallel Channel Configuration



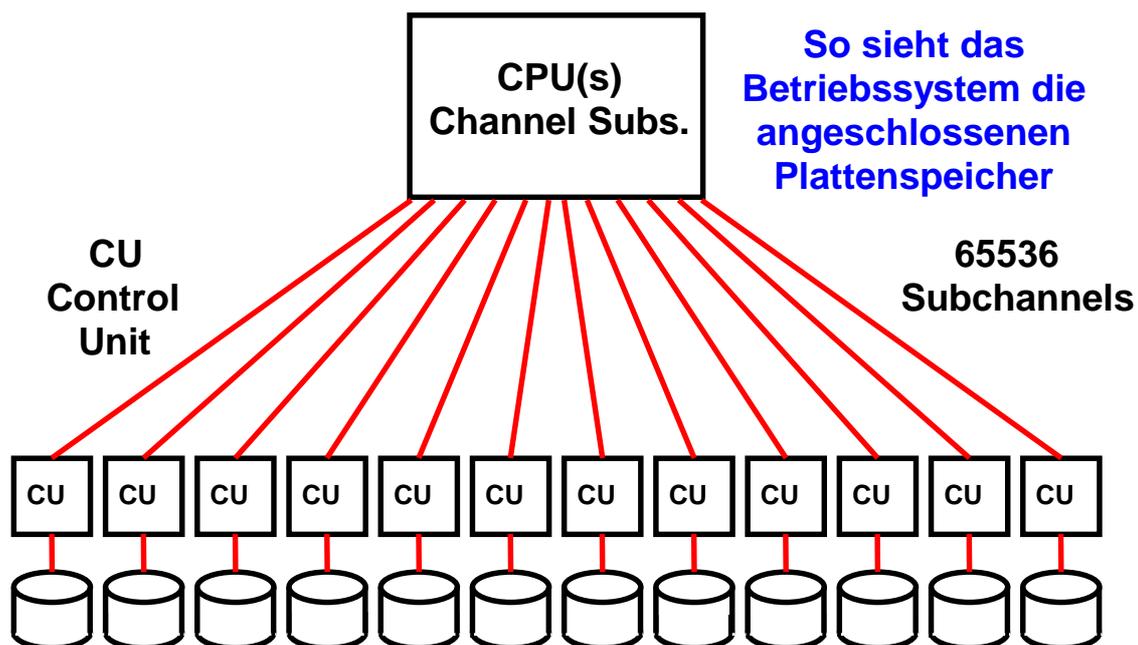
FICON (ESCON) Channel Configuration



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) Subchannel ID angesprochen

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