UNIVERSITÄT LEIPZIG

Mainframe Internet Integration

Prof. Dr. Martin Bogdan
Prof. Dr.-Ing. Wilhelm G. Spruth

SS2013

Sysplex Teil 1

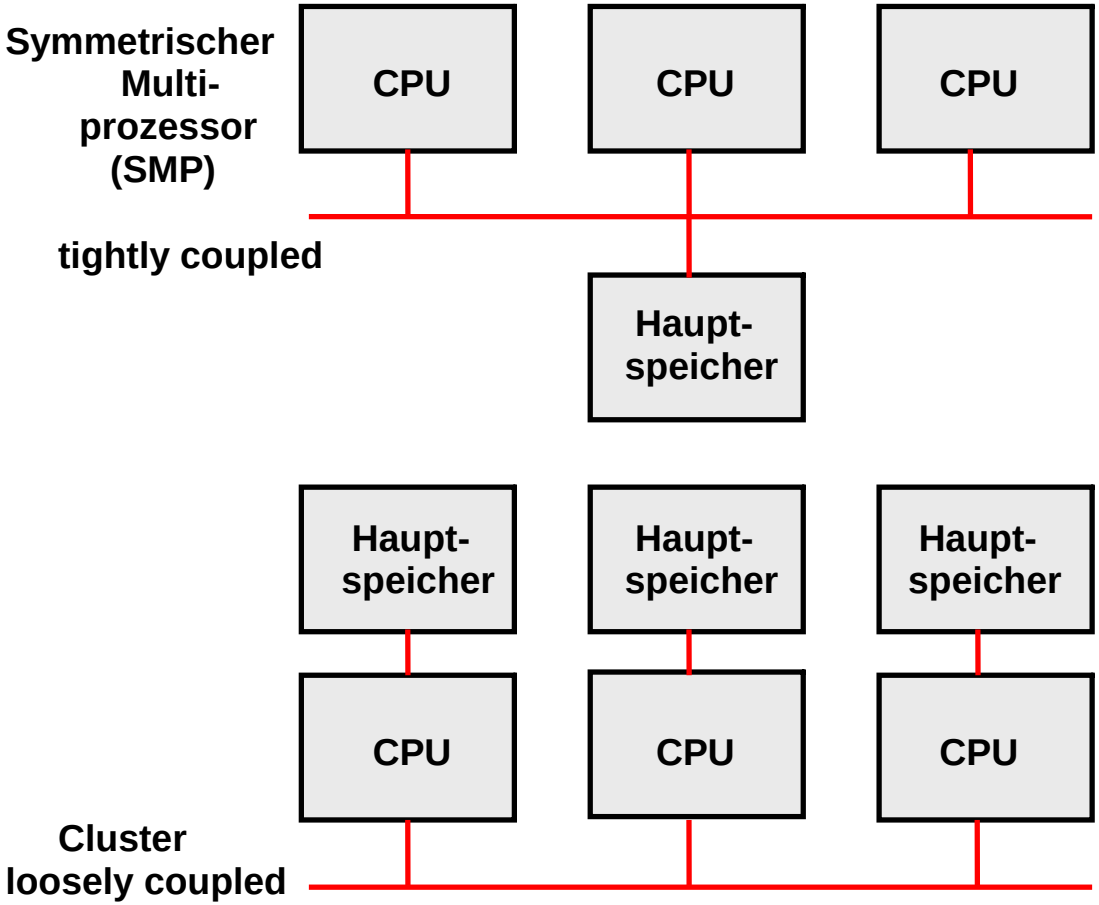
Mehrfachrechner

Mehrfachrechner

Eine einzelne CPU hat eine nur begrenzte Rechenleistung. Die Forderung nach mehr Rechenleistung führte dazu, mehrere Prozessoren zu koppeln.

In Systemen mit mehr als einem Prozessor teilt das Betriebssystem die einzelnen Prozesse oder Threads auf die verschiedenen Prozessoren auf und sorgt damit für eine höhere Leistungsfähigkeit.

Mainframe Systeme erfordern meistens die Leistung von mehr als einer CPU.



Taxonomie von Mehrfachrechnern

Bei Mehrfachrechnern unterscheiden wir 2 Grundtypen:

- **Ein Symmetrischer Multiprozessor (SMP) wird auch als eng gekoppelter (tightly coupled) Mehrfachrechner bezeichnet. Er besteht aus mehreren CPUs (über 100 bei einem zEC12 Mainframe), die alle auf einen gemeinsamen Hauptspeicher zugreifen. In dem Hauptspeicher befindet sich eine einzige Instanz des Betriebssystems.**
- **Ein Cluster (loosely coupled Mehrfachrechner) besteht aus mehreren Rechnern, von denen jeder seinen eigenen Hauptspeicher und seine eigene Instanz eines Betriebssystems hat.**

Crossbar Switch

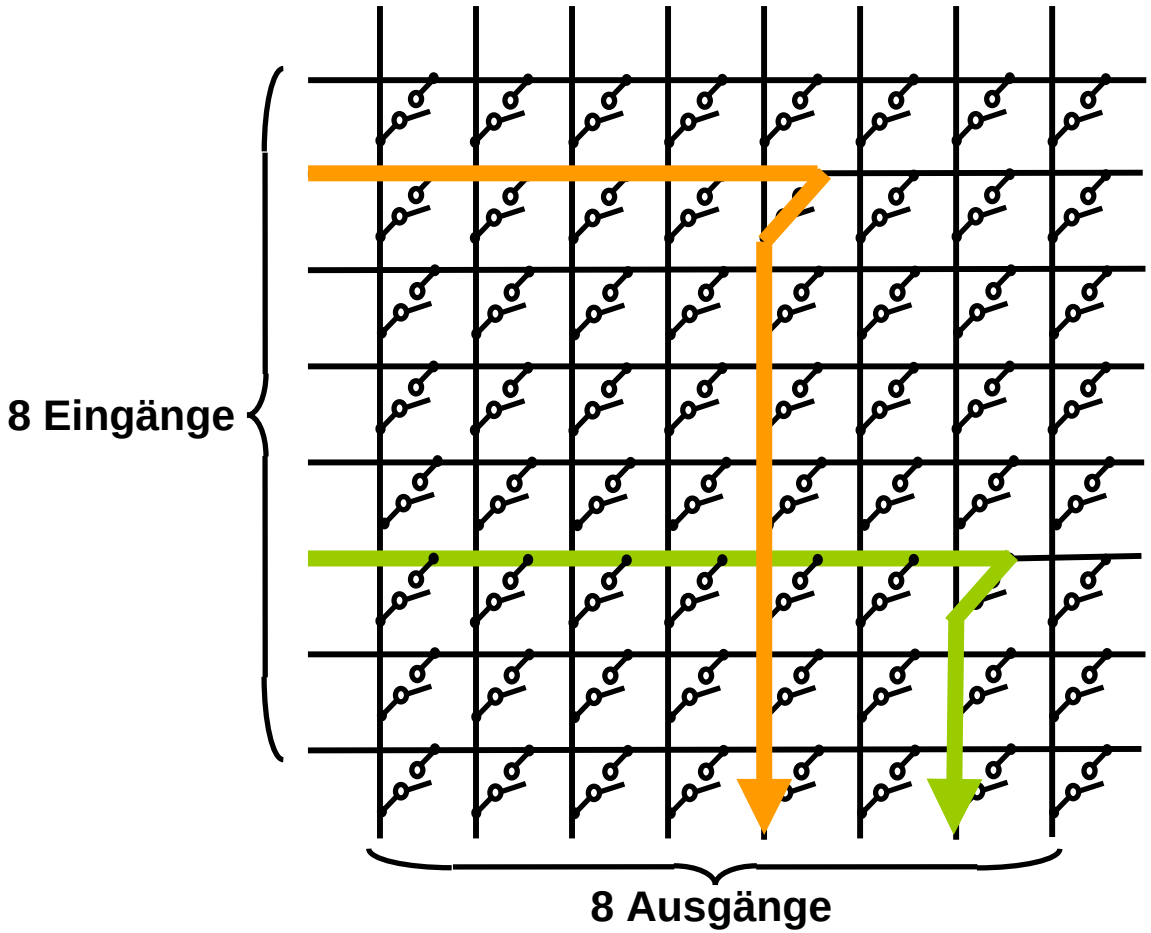
Die CPUs eines Parallelrechners sind über ein Verbindungsnetzwerk miteinander verbunden. Dies kann ein leistungsfähiger Bus sein, z.B. der PCI Bus. Ein Bus hat aber nur eine begrenzte Datenrate. Deshalb setzen viele Implementierungen statt dessen einen Kreuzschienenverteiler (Crossbar Switch, Crossbar Matrix Switch) als Verbindungsnetzwerk ein, der die gleichzeitige Verbindung mehrerer Eingänge mit mehreren Ausgängen ermöglicht.

Mit derartigen Switches können fast beliebige Datenraten erreicht werden. In der folgenden Abbildung ist als Beispiel ein Switch mit 8 Eingängen und 8 Ausgängen gezeigt, der gleichzeitig 8 parallele Verbindungen ermöglicht. In jedem Augenblick kann durch entsprechende Steuerung jeder Eingang mit jedem Ausgang verbinden sein.

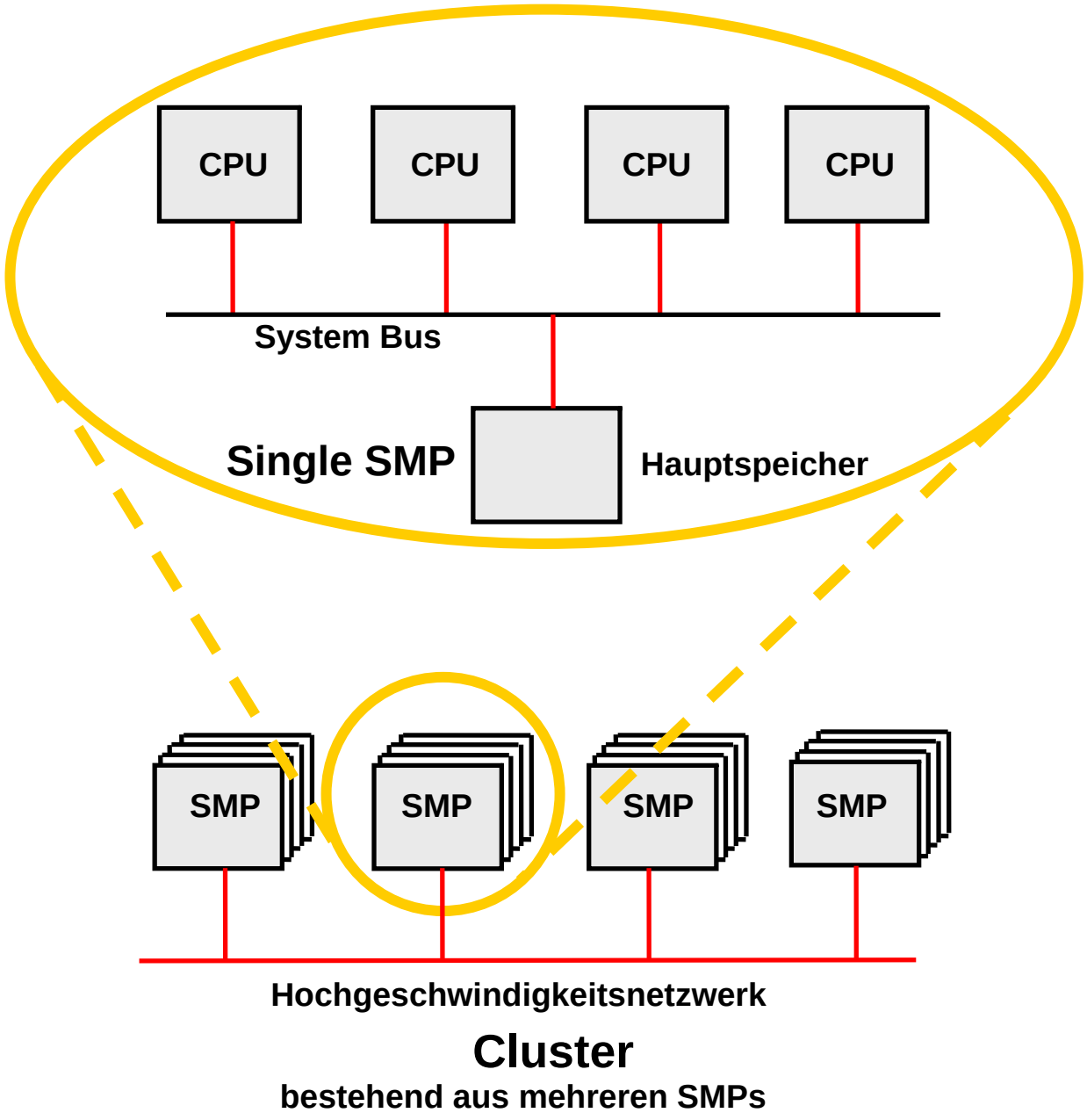
Die Personalisierung der 8 x 8 = 64 Transistorswitche erfolgt in einfachsten Fall durch einen 8 x 8 Bit Speicher, der durch einen eigenen Mikroprocessor angesteuert wird. Durch Änderung des Speicherinhalts können die Verbindungen dynamisch geändert werden.

Zur Erhöhung der Brandbeite lassen sich 8 oder 32 derartige Switche übereinanderstapeln, um einen 8 oder 32 Bit breiten Bus für jeden Eingang/Ausgang zu implementieren.

Crossbar Switch Silizium Chips mit je 256 Ein/Ausgängen lassen sich in der heutigen Silzium Technik kostengünstig herstellen.



Beispiel: 8 x 8 Crossbar Switch



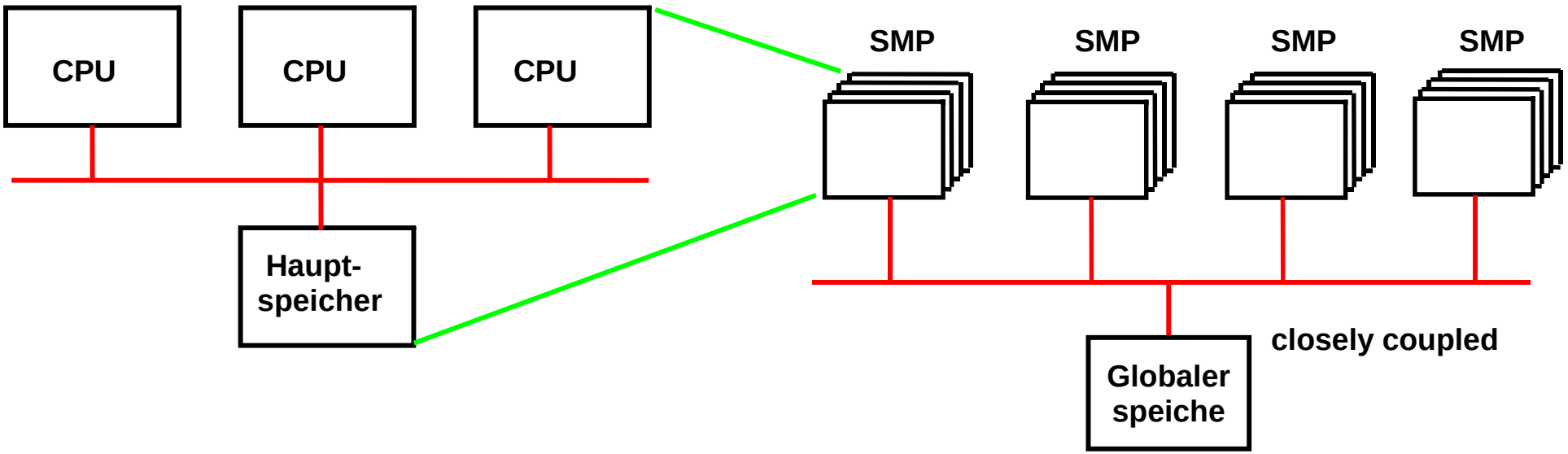
Häufig (heute fast immer) sind die Elemente eines Clusters nicht einzelne CPUs, sondern SMPs, die als Knoten (und in der IBM Terminologie als „Systeme“) bezeichnet werden.

Jeder Knoten ist ein SMP. Er besteht aus mehreren CPUs, die auf einen gemeinsamen Hauptspeicher mit einer einzigen Instanz des Betriebssystems zugreifen.

Die Knoten sind über ein Hochgeschwindigkeitsnetzwerk miteinander verbunden, das in der Regel als Crossbar Switch implementiert wird.

Die Großrechner von HP, IBM und Sun haben diese Struktur.

Ein SMP wird als eng gekoppelter, ein Cluster als loose gekoppelter Multiprozessor bezeichnet.



Ein Cluster, bei dem die Knoten aus einzelnen CPUs oder aus SMPs bestehen, kann durch einen globalen Speicher erweitert werden. Diese Konfiguration wird als nahe gekoppelt (closely coupled) bezeichnet. Der Mainframe Sysplex ist eine derartige Konfiguration.

Unix Großrechner

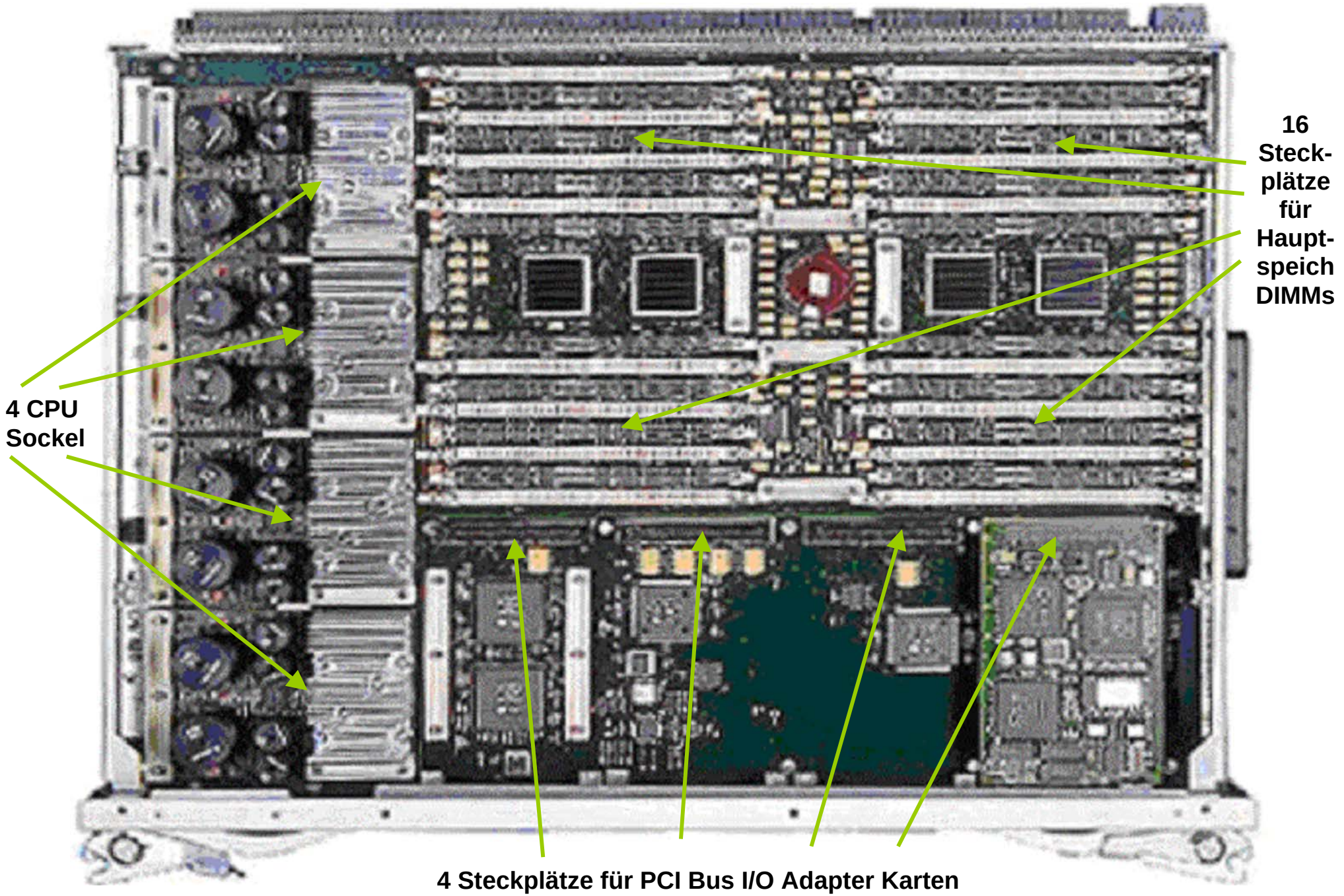
Die folgenden Abbildungen zeigen Beispiele von Unix basierten Großrechnern der Firmen Sun und Hewlett Packard. Das erste Beispiel betrifft den E 25 000 Rechners der Firma Sun.

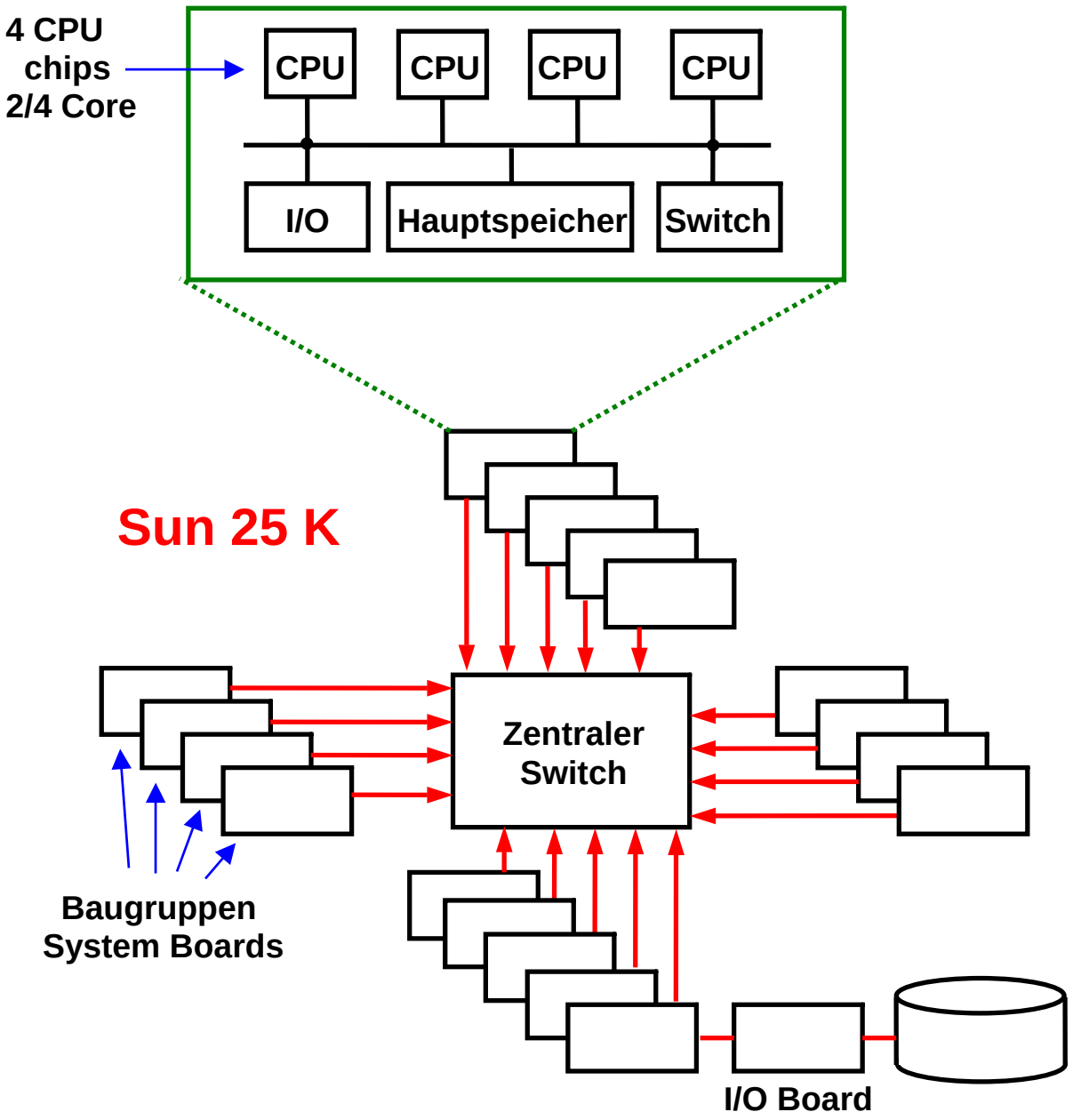
Zentralstück des Rechners ist ein Crossbar Switch, der sich auf einem Motherboard zusammen mit 16 Steckplätzen für sog. „System Boards“ befindet. Die folgende Abbildung zeigt ein derartiges System Board.

Auf dem System Board befinden sich 4 CPU Chip Sockel. Zu sehen sid die Kühlkörper. CPU Chips waren früher Single Core, sind heute (2012) in der Regel aber Dual Core oder Quad Core (oder mehr). Weiterhin befinden sich auf dem System Board 16 Steckplätze für Hauptspeicher DIMMs. Als Dual Inline Memory Module (DIMM) werden Speichermodule für den Arbeitsspeicher von Computern bezeichnet. Im Gegensatz zu Single Inline Memory Modulen (SIMM) führen DIMMs auf den Anschlusskontakten auf der Vorderseite und auf der Rückseite der Leiterplatte unterschiedliche Signale.

Zusätzlich enthält das System Board 4 Steckplätze für sog. Daughter Cards. Es stehen eine Reihe unterschiedlicher Arten von Daughter Cards zur Verfügung, aber die allermeisten Steckplätze werden von I/O Adapter Karten eingenommen, die z.B. mit einem SCSI Kabel eine Verbindung zu Plattenspeichern ermöglichen.

512 MByte DIMM (Wikipedia)


16 Steck-plätze für Haupt-speich DIMMs
4 CPU Sockel
4 Steckplätze für PCI Bus I/O Adapter Karten

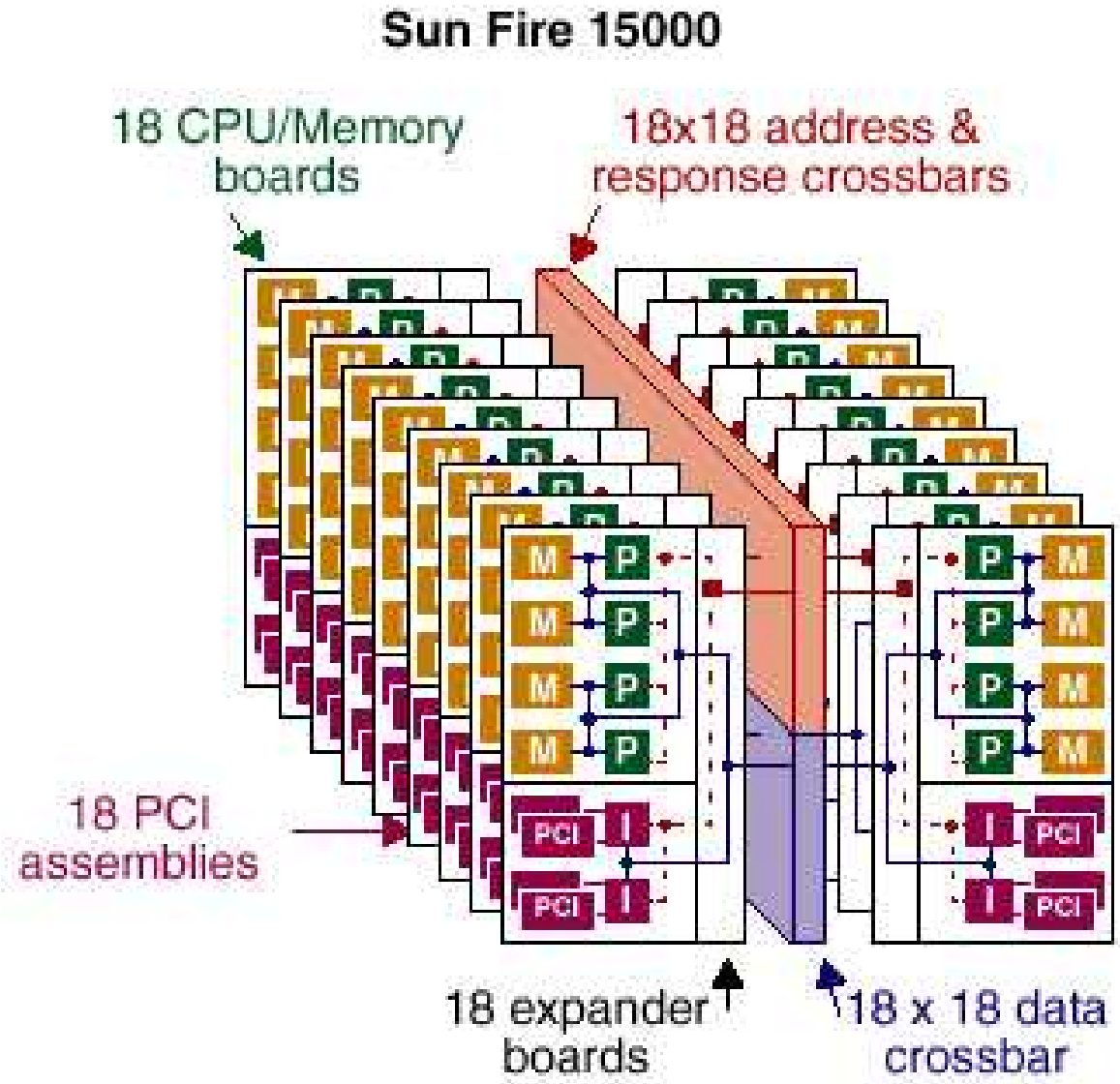


Ein Sun 25K oder HP Superdome Großrechner enthält 16 (oder 18) Printed Circuit Board (PCB) Baugruppen (System Boards), jedes mit 4 CPU Chip Sockeln und bis zu 128 GByte Hauptspeicher, einem Anschluss an einen zentralen Switch sowie I/O Adapter auf jedem System Board.

Die I/O Adapter sind mit PCI Bus Kabeln mit einer Reihe von I/O Cages verbunden, in denen sich Steckkarten für den Anschluss von I/O Geräten, besonders Festplattenspeichern, befinden.

Die CPUs eines jeden System Boards können nicht nur auf den eigenen Hauptspeicher, sondern auch auf den Hauptspeicher eines jeden anderen System Boards zugreifen. Hiermit wird eine „Non-Uniform Memory Architecture“ (NUMA) verwirklicht.

Eine moderne Form eines System Boards wird als „Blade“ bezeichnet, und befindet sich in vielen Low End Produkten der Firma Sun.



Die 16 (oder 18) System Boards befinden sich in Steckplätzen auf einem Crossbar Board. Dieses enthält eine Reihe von Crossbar Chips, welche alle System Boards miteinander verbindet. Dies sind spezifisch ein 16 x 16 Daten Crossbar Chip, ein 16 x 16 Response Crossbar Chip und ein 16 x 16 Adressen Crossbar Chip.

Das Crossbar Board enthält eine Reihe von (elektronischen) Schaltern. Mit diesen lassen sich die System Bards in mehrere Gruppen aufteilen und die Gruppen voneinander isolieren.

Da ein Unix SMP mit 64, 128 oder 256 CPUs bei Transaktions- und Datenbank-anwendungen nicht mehr skaliert, kann der Rechner mit Hilfe der Schalter in mehrere voneinander isolierte SMPs aufgeteilt werden.

Eine solche Aufteilung wird als „Harte Partitionierung“ bezeichnet.

Die Firma Sun hat ihre Produkte kontinuierlich weiterentwickelt ohne dass sich am Konzept in den letzten 10 Jahren viel geändert hat. Die neuesten Modelle werden als SPARC Enterprise M9000 Server oder M-Series Server bezeichnet, und gemeinsam von Sun und von Fujitsu/Siemens vertrieben.

Ein derartiger Rechner ist nicht gerade billig. Eine Minimalkonfiguration hat einen Listenpreis von über 1 Million $. Je nach Ausstattung gehen die Preise von dort steil nach oben. Die Konkurrenzprodukte von Hewlett Packard und IBM sind eher noch teurer.

Hewlett-Packard Superdome Cell Board

4 Itanium 2 CPUs 1,73 GHz

64 Gbyte Hauptspeicher

E/A Bus Anschlüsse

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

Hewlett Packard Superdome

Der Hewlett Packard (HP) Superdome Rechner ist ähnlich aufgebaut wie der Sun Fire 25K bzw. Sun M9000 Rechner. Das System Board wird als Cell Board bezeichnet und hat 4 Sockel für 4 Itanium (Tukwila) CPU Chips mit je 4 CPU Cores. Neben den CPU Chips befinden sich auf dem Cell Board Sockel für Hauptspeicherp DIMMs sowie Steckplätze für PCI Bus I/O Adapter Karten. Wie beim Sun Fire 25K sind die Cell Boards über einen zentralen Cross Bar Matrix Switch miteinander verbunden.

Auch die Firma Hewlett verwendet eine vereinfachte Form eines Cell Boards (als „Blade“ bezeichnet) in vielen Low End Produkten des Unternehmens. Cell Boards haben gegenüber Blades zahlreiche zusätzliche Funktionen. Beispiele sind:

- **Verbindung über einen zentralen Switch**
- **NUMA Fähigkeit**
- **Partitionsmöglichkeiten (siehe nächstes Thema)**
- **Zusätzliche unterstützende Hardware für das HP-UX Betriebssystem**
- **Verbesserte I/O Einrichtungen, höhere I/O Kapazität**
- **Einrichtungen für eine zentrale Administration**
- **Zusätzliche Verbesserungen für Ausfallsicherheit, Zuverlässigkeit und Verfügbarkeit**
- **.....**

IBM bietet als Alternative zu den Sun M9000 oder HP Superdome Systemen mit den Solaris oder HP-UX Betriebssystemen das ähnliche Produkt „System p“ mit den Betriebssystemen AIX und i5/OS an. Aus Platzgründen wird auf diese Entwicklung nicht näher eingegangen. AIX ist ein besonders funktionsreiches Unix Betriebssystem.

Sysplex

Der Sysplex (SYStem Processing compLEX) ist eine 1990 für Mainframes eingeführte lose Rechnerkopplung von IBM Mainframes. Oft wird diese einfache Form des Clusters auch Base Sysplex genannt.

Der Parallel Sysplex ist eine Weiterentwicklung des Base Sysplex, dessen Einführung 1994 erfolgte. Hierbei handelt es sich um eine nahe Rechnerkopplung (closely coupled). Da der Base Sysplex praktisch ausgestorben ist, bezeichnet ein Sysplex in der Umgangssprache fast immer einen Parallel Sysplex.

Beim Parallel Sysplex übernimmt die Coupling Facility (CF) die Funktion eines globalen Arbeitsspeichers. Innerhalb der CF können drei verschiedene Datenstrukturtypen erzeugt und verwaltet werden. Dies sind Sperr- (Lock), Listen- und Cache-Strukturen, die u.a. genutzt werden um Konkurrenzsituationen beim Zugriff auf geteilte Rescource zu verwalten. Die Hardware der Coupling Facility besteht aus einem regulären Mainframe Rechner, auf dem der „Coupling Facility Control Code“ (CFCC) an Stelle eines Betriebssystems läuft.

Nutzen Subsysteme wie CICS, DB2 oder IMS die durch den Parallel Sysplex bereitgestellten Einrichtungen, so können sie sich gegenüber dem Anwender als eine einzige Anwendung präsentieren. Man spricht dann auch von einem Single System Image (SSI).

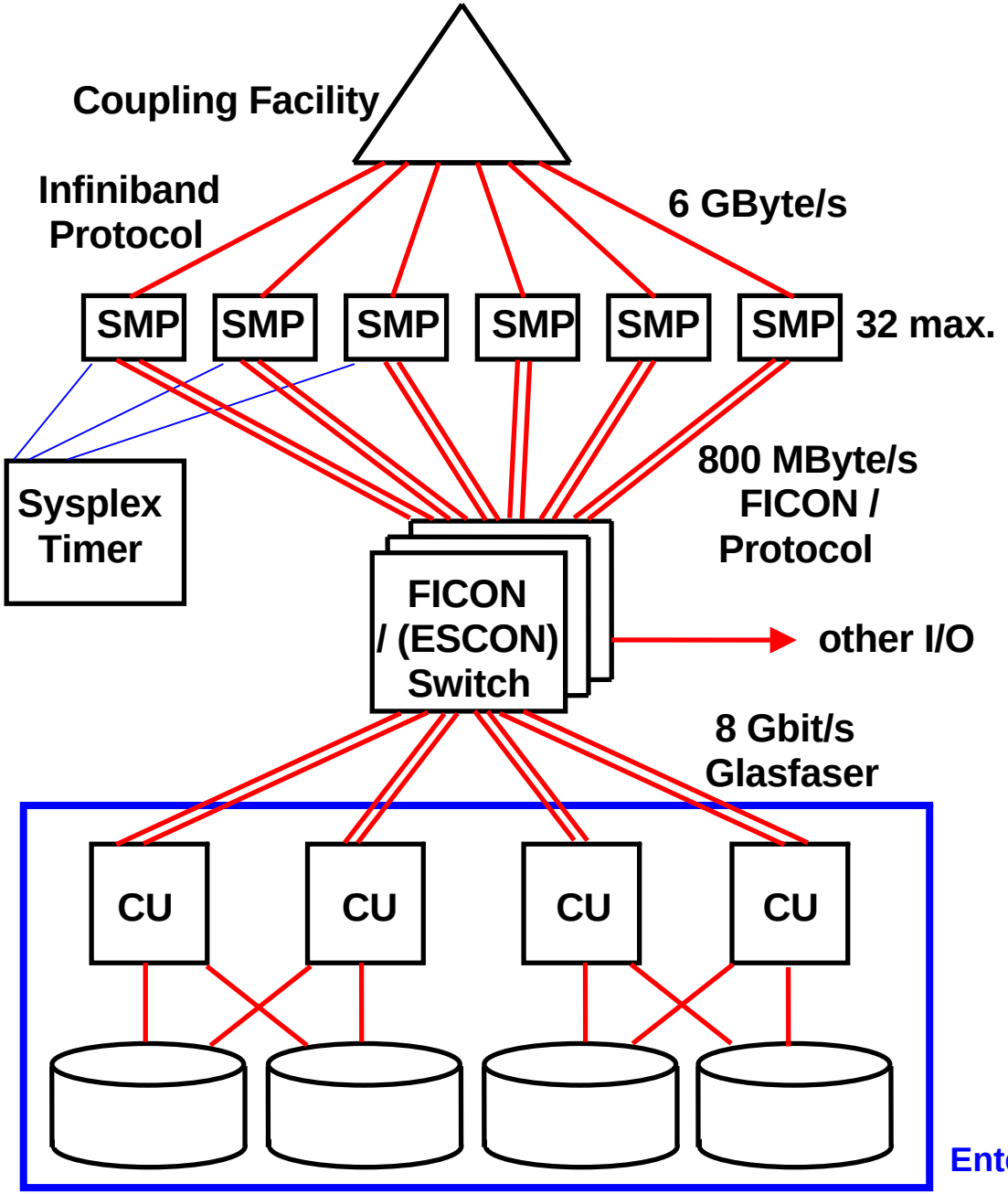
Literatur: Wilhelm G. Spruth, Erhard Rahm, Sysplex-Cluster Technologien für Hochleistungs-Datenbanken. Datenbank-Spektrum, Heft 3, 2002, S. 16-26. http://www-ti.informatik.uni-tuebingen.de/~spruth/Mirror/Sysplex3.pdf

Sysplex mit Coupling Facility

Die folgende Abbildung zeigt den Aufbau der Parallel Sysplex-Architektur. Während die vorgestellten Cluster von Sun und HP eine lose Kopplung verwenden, handelt es sich beim Parallel Sysplex um eine nahe gekoppelt (closely coupled) Cluster-Architektur. Der Sysplex besteht aus den Prozessor-Knoten (fast immer SMPs), gemeinsamen Plattenspeichern (Shared Storage Devices), Netzwerk-Controllern und den Kern-Cluster-Technologie-Komponenten. Letztere umfassen den (die) als „FICON Director“ bezeichneten Switch(es), den Sysplex-Timer und die „Coupling Facility“ (CF). Die Coupling Facility enthält den für die nahe Kopplung charakteristischen globalen Speicher zur Realisierung globaler Kontrollaufgaben und ist von allen Knoten schnell zugreifbar.

Zu den CPUs kommen noch weitere Ein-/Ausgabe-Prozessoren (System Assist-Prozessoren, SAPs). Heutige Installationen haben bis zu 200 CPUs. Die Knoten müssen nicht homogen sein, d.h. es können unterschiedliche Mainframe Modelle eingesetzt werden.

Zu den Sysplex-Komponenten werden spezifische Maschinenbefehle sowie Betriebssystemdienste zur Verfügung gestellt, mit denen eine effiziente Durchführung der Cluster-Aufgaben für alle Knoten erreicht wird, insbesondere zur Kommunikation, Ein/Ausgabe, sowie für globale Steuerungsaufgaben wie Synchronisation, Kohärenzkontrolle und Lastbalancierung. Diese Dienste werden in allgemeiner Form realisiert und von unterschiedlichen Software-Subsystemen, insbesondere Web Application Server, Datenbanksystemen und Transaktionsservern für den Cluster-Einsatz verwendet. IBM hat dazu die wichtigsten Subsysteme an das Arbeiten innerhalb eines Sysplex angepasst. Eine Anpassung der Anwendungen an den Sysplex ist in der Regel nicht erforderlich und i.a. auch nicht möglich.



Sysplex mit Coupling Facility

Ein Sysplex besteht aus bis zu 32 System z Rechnern (Knoten), die über FICON Glasfasern und FICON Switche mit Plattenspeichern in der Form von einem oder mehreren Enterprise Storage Servern verbunden sind. Ein Enterprise Storage Server emuliert mehrere Control Units (CU).

Weiterhin sind 1 (oder in der Regel 2) Coupling Facilities vorhanden.

Eine Coupling Facility ist ein regulärer System z Rechner, auf dem „Coupling Facility Code“ an Stelle eines Betriebssystems läuft.

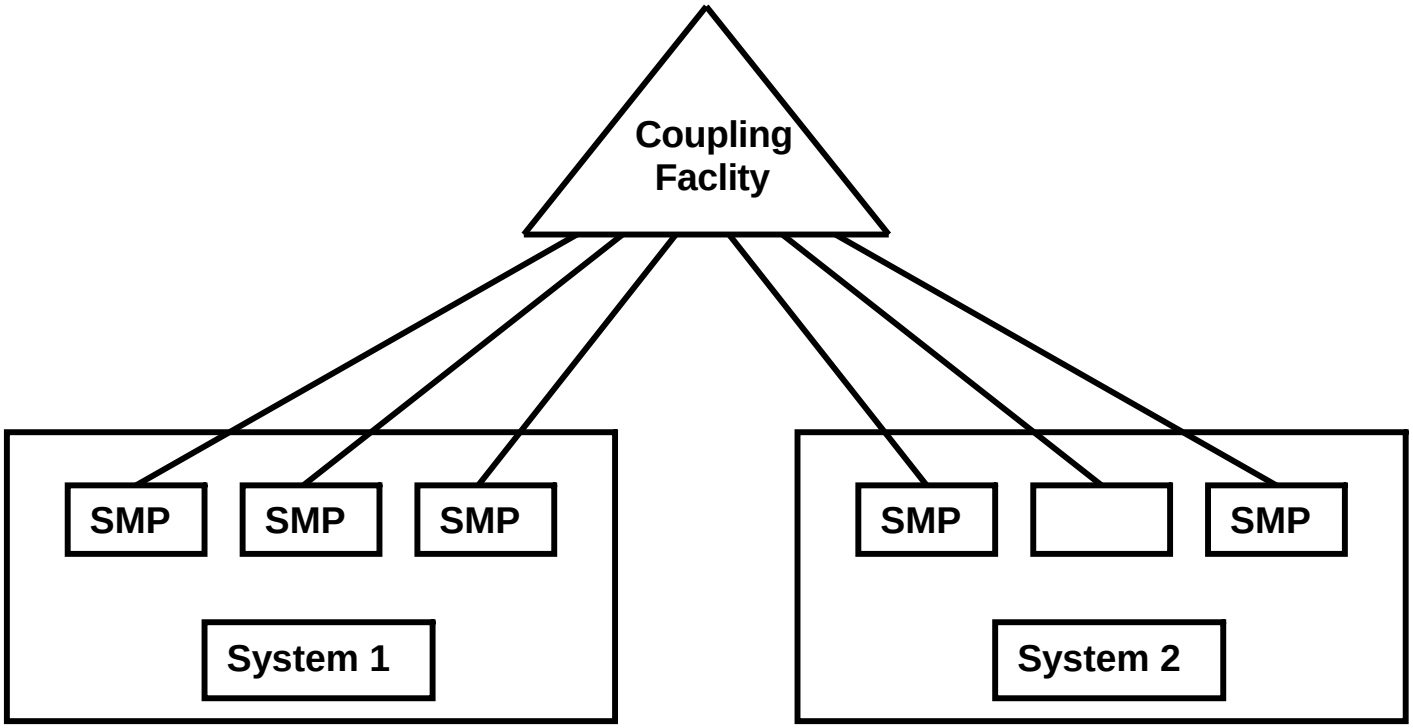
Weiterhin sind 1 oder 2 Zeitgeber (Sysplex Timer) vorhanden. Bei den heutigen Rechnern ist die Zeitgeberfunktion in die CPU Chips integriert.

Jeder Knoten kann mit bis zu 256 „Kanälen“ mit der Außenwelt verbunden werden, mit bis zu 800 Mbyte/s pro Verbindung. Die Verbindung der Systeme sowohl untereinander als auch zu den Festplattenspeichern erfolgt über ein oder mehrere FICON (FIbre CONnection)-Switches. Da jedes System 256 Ein-/Ausgabe-Kanäle anbinden kann, hat diese Verbindungsstruktur die Aufgabe, bei 32 Systemen maximal 256 * 32 = 2^{13} Kanäle zu verwalten. Damit ist jeder Knoten in der Lage, auf alle Plattenspeicher und alle anderen Knoten des Sysplex direkt zuzugreifen (Shared Disk-Modell). Die Steuereinheiten der Plattenspeicher (Control Units) implementieren eine Storage Server- und SAN-Funktionalität.

Die von IBM entwickelte FICON-I/O-Architektur basiert auf dem Kanal-Subsystem (channel subsystem) der Mainframe I/O-Architektur. Dieses integriert System Assist Prozessoren (SAPs), Kanäle sowie die „Staging Hardware“. Die SAPs führen die Kommunikation zwischen den Knoten und den Kanälen durch. Diese führen zur Datenübertragung ein Kanalprogramm aus, wobei über die Steuereinheiten (Control Units) auf die Plattenspeicher zugegriffen wird. Die Staging Hardware stellt die Kommunikationspfade zwischen den I/O-Prozessoren, den Kanälen und dem Rest des Systems zur Verfügung.

In der FICON-Architektur bildet der FICON-Switch (FICON Director) die Kerneinheit. Er implementiert eine Switched Point-to-Point-Topologie für System z I/O-Kanäle und Steuereinheiten. Ein FICON-Switch kann bis zu 60 Kanäle und Steuereinheiten dynamisch und nicht-blockierend (Crossbar Switch) über seine Ports miteinander verschalten. Normalerweise werden eine Reihe von FICON-Switches parallel genutzt. Entfernungen von bis zu 3 km für optische Übertragungen sind möglich. Die zulässigen Entfernungen erhöhen sich beim Einsatz einer sogenannten Extended Distance Laser Link-Einrichtung auf 20, 40 oder 60 km. Dies ist für Unternehmen wichtig, die aus Katastrophenschutz Gründen zwei getrennte Rechenzentren betreiben.

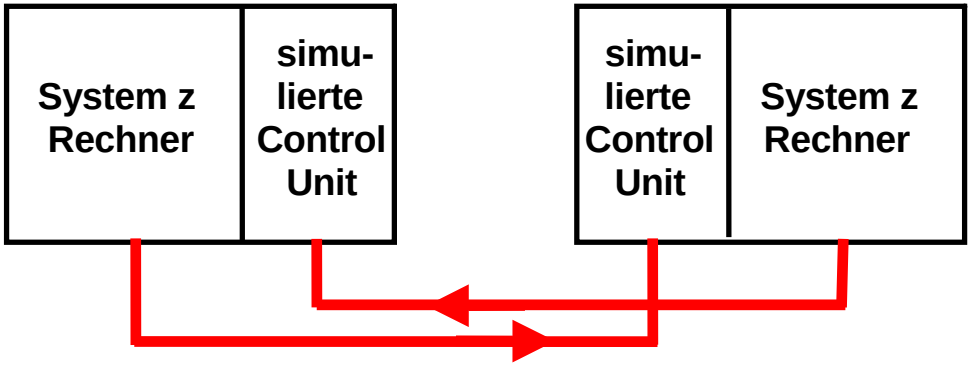
Die Kommunikation zwischen den Knoten und den Plattenspeichern erfolgt über das FICON-Protokoll. Zur Kommunikation der Knoten untereinander wird das „Channel-to-Channel“ (CTC)–Protokoll eingesetzt, normalerweise in einer Full-Duplex-Anordnung. Hierbei betrachtet jeder sendende Knoten den Empfänger als eine Ein-/Ausgabe-Einheit. Der Empfänger simuliert für diesen Zweck einen eigenen Typ einer Mainframe-Steuereinheit. Das z/OS-Betriebssystem stellt einen Basisdienst, die „Cross System Coupling Facility“ (XCF) zur Verfügung, um über eine CTC-Verbindung eine Kommunikation mit einer anderen z/OS-Instanz innerhalb eines Sysplex zu bewerkstelligen. Dieser Dienst wird wiederum von Subsystemen wie den Datenbanksystemen DB2 und IMS für den Nachrichtenaustausch eingesetzt.



In einem Sysplex werden bis zu 32 Knoten unterstützt. Jeder Knoten stellt eine einzelne CPU, oder in den allermeisten Fällen einen SMP, dar und enthält maximal 101 CPUs, insgesamt also maximal 3232 CPUs. Dieser Wert kann in der Praxis nicht erreicht werden, weil SMPs in der Regel nicht bis zu 101 CPUs skalieren. Eine große Sysplex Installation besteht deshalb in der Regel aus mehreren physischen Rechnern (Systeme in der IBM Terminologie), von denen jeder mehrere SMPs (Knoten) enthält. Aus logischer Sicht spielt es dabei keine Rolle, ob zwei SMPs im gleichen oder in getrennten physischen Rechnern untergebracht sind.

Jeder Knoten ist über eine eigene Glasfaserverbindung (Coupling Link) mit der Coupling Facility verbunden.

Zur Erhöhung der Reliability und Availability werden in der Praxis fast immer 2 Coupling Facilities eingesetzt, von denen die zweite CF als Backup für die erste CF dient,



Channel- to Channel Verbindung

Cross-System Coupling Facility (XCF)

Die Cross-System Coupling Facility (XCF) ist eine Komponente des z/OS Kernels.

Sie verwendet das CTC (Channel To Channel) Protokoll und stellt die Coupling Services bereit, mit denen z/OS Systeme innerhalb eines Sysplex miteinander kommunizieren.

Für eine CTC Verbindung stellt ein System z oder S/390 Rechner (als Slave bezeichnet) eine emulierte Control Unit zur Verfügung. An diese ist ein anderer Rechner (Master) über eine normale FICON Glasfaserverbindung angeschlossen. Der Master kommuniziert mit dem Slave wie mit einer normalen I/O Einheit.

Die CTC Verbindung ist unidirectional. Aus Symmetriegründen werden CTC Verbindungn normalerweise paarweise eingerichtet.

Mittels XCF können die Rechner eines Sysplex untereinander kommunizieren.

Parallel Sysplex Cluster Technology

Zu den Parallel Sysplex Cluster Technology Komponenten gehören:

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

Die Coupling Facility ermöglicht Data Sharing einschließlich Datenintegrität zwischen mehrfachen z/OS Servern

Literatur

Wilhelm G. Spruth, Erhard Rahm:
Sysplex-Cluster Technologien für Hochleistungs-Datenbanken.
Datenbank-Spektrum, Heft 3, 2002, S. 16-26.

download: http://www-ti.informatik.uni-tuebingen.de/~spruth/publish.html

Sonderheft des IBM Journal of Research and Development, Vol. 36, No.4, July 1992 zum Thema Sysplex Hardware:

Sonderheft des IBM System Journal, Vol. 36, No.2, April 1997 mit folgenden Aufsätzen:

J. M. Nick, B. B. Moore, J.-Y. Chung, and N. S. Bowen: S/390 cluster technology: Parallel Sysplex, p. 172

N. S. Bowen, D. A. Elko, J. F. Isenberg, and G. W. Wang: A locking facility for parallel systems, p. 202

G. M. King, D. M. Dias, and P. S. Yu: Cluster architectures and S/390 Parallel Sysplex scalability, p. 221

J. W. Josten, C. Mohan, I. Narang, and J. Z. Teng: DB2's use of the coupling facility for data sharing , p. 327

T. Banks, K. E. Davies, and C. Moxey: The evolution of CICS/ESA in the sysplex environment, p. 352

J. P. Strickland: VSAM record-level data sharing, p. 361