



# **Enterprise Computing System z Hardware**

**Prof. Dr.-Ing. Wilhelm G. Spruth  
Dipl. Inf. Gerald Kreißig**

**WS2016/17**

# **System z Hardware Teil 1**

## **Mainframe Alternativen**

# Mainframe Alternativen

| Hersteller    | Name                  | Microprocessor | Betriebssystem |
|---------------|-----------------------|----------------|----------------|
| Fujitsu / Sun | Sunfire, M9000, M5-32 | Sparc          | Solaris        |
| HP            | Superdome             | Itanium        | HP-UX          |
| IBM           | System p              | PowerPC        | AIX            |

Mehrere Hersteller (Dell, HP, IBM, Unisys, andere) stellen große Konfigurationen mit x86 Microprozessoren und Windows/Linux Betriebssystem her.

Die Firma Unisys produziert in kleinen Stückzahlen Großrechner, mit dem OS2200 Betriebssystem , welches auf die UNIVAC 1100/2200-Serie zurückgeht, sowie Rechner mit dem **Master Control Program (MCP)** Betriebssystem, welches auf die Burroughs-B5000-Produktlinie zurückgeht.

Die Firma Hewlett Packard (HP) vertreibt neben dem hauseigenen **HP-UX** (Unix) mehrere weitere Betriebssysteme, die zum Teil aus der Übernahme mehrerer anderer Computer Firmen stammen:

- **MPE/iX** ist eine HP-eigene Entwicklung, die 2010 eingestellt wurde.
- **Tru64 UNIX** wurde von der Firma Digital Equipment entwickelt und läuft auf "Alpha" Microprozessoren. Es wird nicht weiterentwickelt, und HP stellt die Unterstützung Ende 2012 ein.
- **HP NonStop** stammt von der Firma Tandem Computers und läuft auf Itanium Microprozessoren.
- Das **Virtual Memory System (VMS)** Betriebssystem wurde von der Firma **Digital Equipment Corporation (DEC)** entwickelt und läuft ebenfalls auf Itanium Microprozessoren. Microsoft benutzte VMS als Basis für die Entwicklung von Windows 2000.

# **Hardware für betriebswirtschaftliche Großrechner**

Sun (mit dem Co-Operationspartner Fujitsu), Hewlett Packard sowie IBM sind die drei führenden Hersteller von betriebswirtschaftlichen Großrechnern. Einige weitere Hersteller (z.B. Bull mit dem novascale gcos 9010 System oder Unisys mit dem ClearPath System) spielen eine eher untergeordnete Rolle.

Die meisten Implementierungen von betriebswirtschaftlichen Großrechnern verwenden die gleichen oder ähnlichen Technologien und Bausteine, wie sie auch für Arbeitsplatzrechner oder kleine Server eingesetzt werden. Dies hat den Vorteil, die Entwicklungskosten auf eine größere Stückzahl verteilen zu können. Es hat den Nachteil, dass die verwendeten Komponenten (Commodity Parts) nicht für den Einsatz in einem betriebswirtschaftlichen Großrechner optimiert wurden. Zwei Beispiele sind:

- Zuverlässigkeit und Verfügbarkeit richtet sich nach ökonomischen Kriterien , die für den PC Bereich etabliert werden. Durch zusätzliche Einrichtungen versucht man Verbesserungen zu erreichen, die dann aber ins Geld gehen.
- Ein/Ausgabe Einrichtungen (Input/Output, I/O) sind z.B. für den Anschluss von 5 Plattspeichern optimiert, nicht aber für 5 000 oder 50 000 Plattspeicher

Es werden aber zahlreiche zusätzliche Komponenten benötigt, deren Entwicklung einen erheblichen Aufwand erfordert. Als Folge sind betriebswirtschaftlichen Großrechner alles andere als billig, wie das folgende Beispiel eines führenden Herstellers, der Firma Sun/Oracle, zeigt:

# The new flagship of the industry.

**Get It** From \$1.023.047,00 (US)

» Upgrade now and get over 5x performance gains within the same chassis.



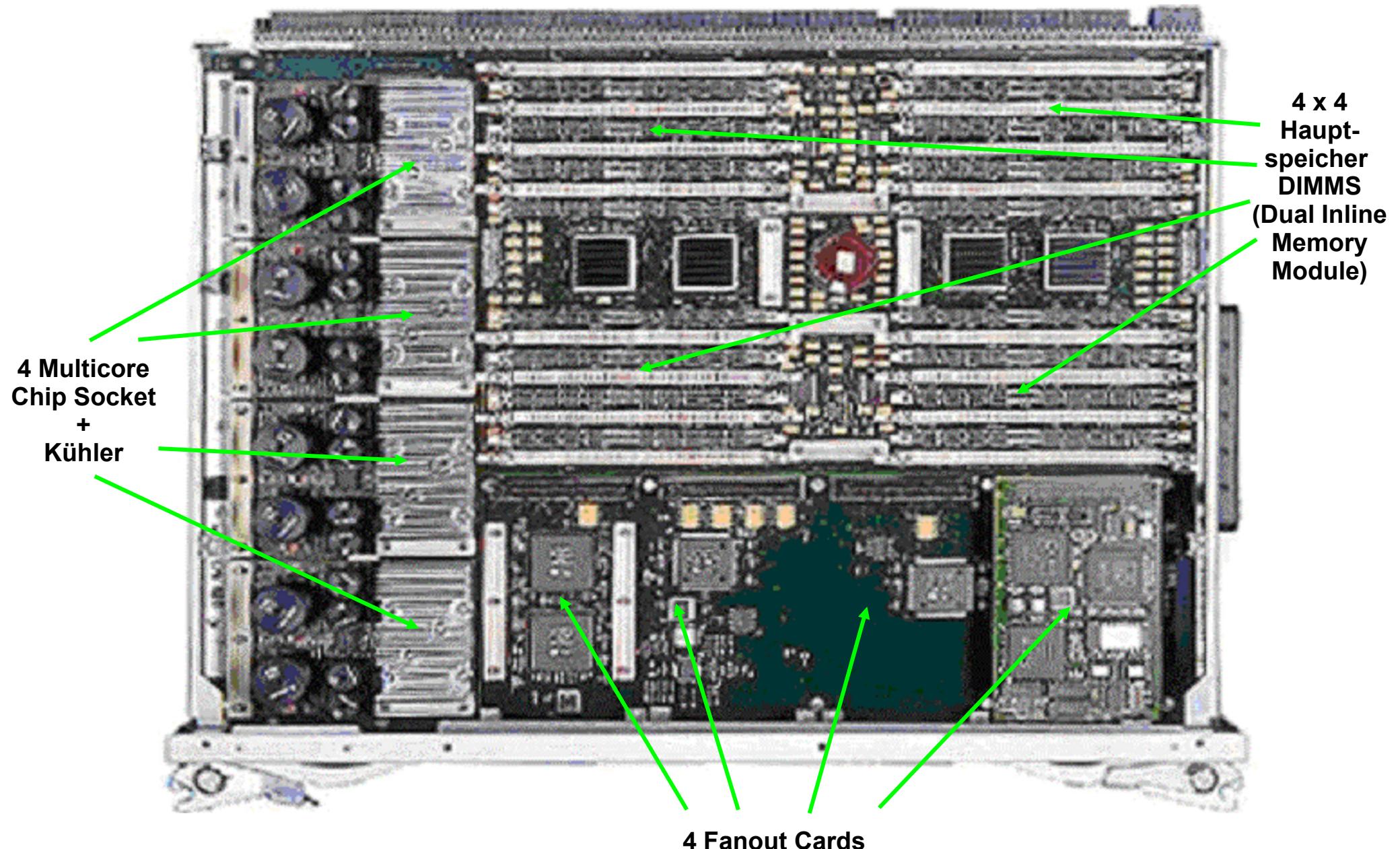
Die Firma Sun verkündet stolz, dass eine Minimalkonfiguration ihres Großrechners für wenig mehr als 1 Million \$ zu haben ist.

Das kleinste Modell der Sun 25k Serie hatte 4 „System Boards“ mit je 4 Dual Core SPARC CPUs, oder insgesamt 16 CPUs. Maximal sind 16 bzw. 18 System Boards möglich. Spätere Modelle benutzten Quad Core CPU Chips. Die E25K System Boards (auch als Prozessor Boards bezeichnet) sind eine evolutionäre Weiterentwicklung der System Boards in der Sun 15k und Sun 10k Serie.

Die Sun System Boards werden auch in anderen Produkten der Firma Sun eingesetzt, z.B. in einer Low End Workstation mit einem einzigen System Board.

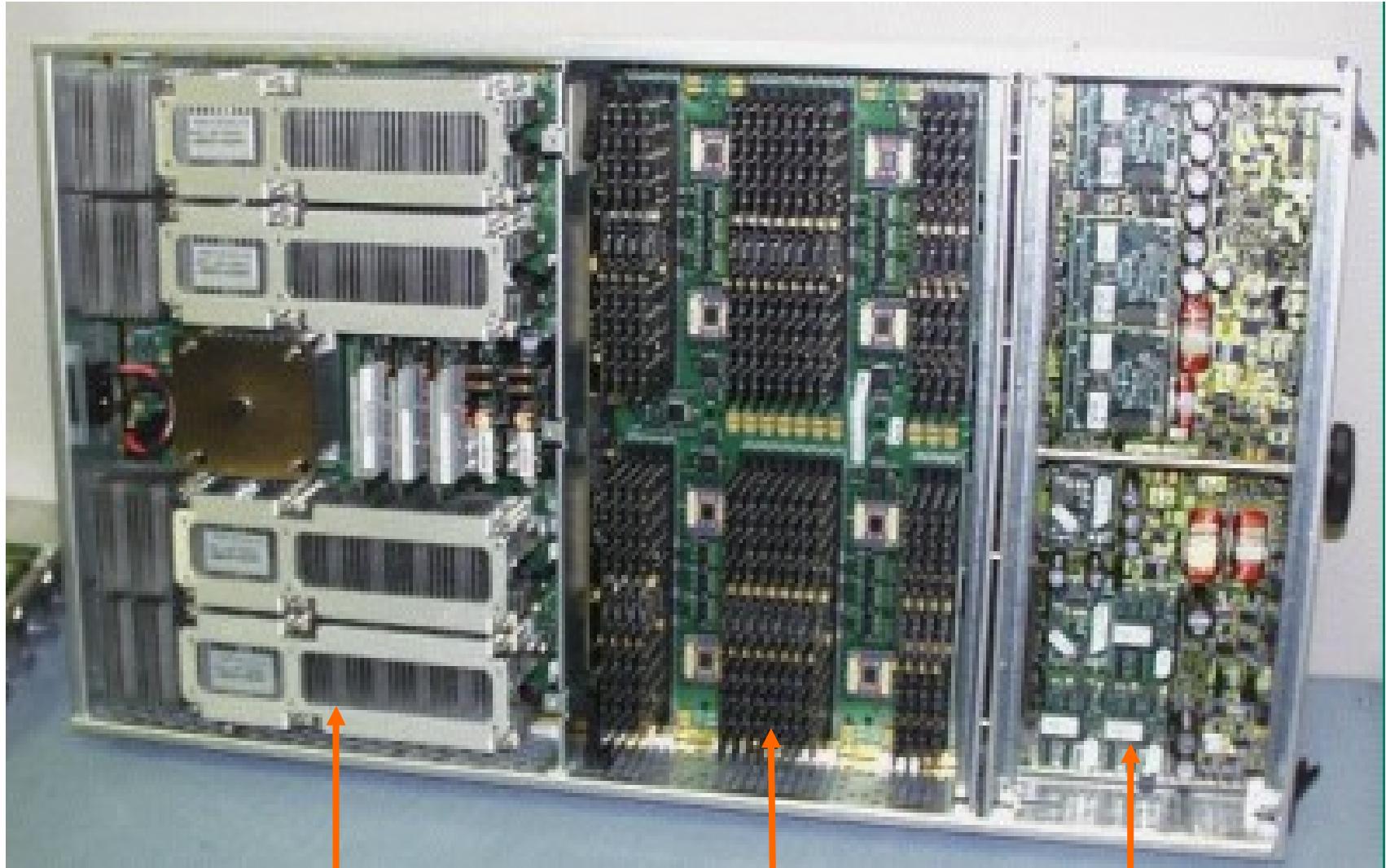
Die Firma Sun hat ihre Produkte kontinuierlich weiterentwickelt, wobei sich vor allem die Anzahl der CPU Cores pro CPU Chip gewachsen ist. Die Folgemodelle wurden als M9000 bezeichnet, und gemeinsam von Sun und von Fujitsu/Siemens vertrieben. Die neueste Version läuft unter dem Namen SPARC M5-32 Server.

Die nächste Abbildung zeigt ein Sun System Board.



## Sun E 15 000 System Board

# Hewlett-Packard Superdome Cell Board

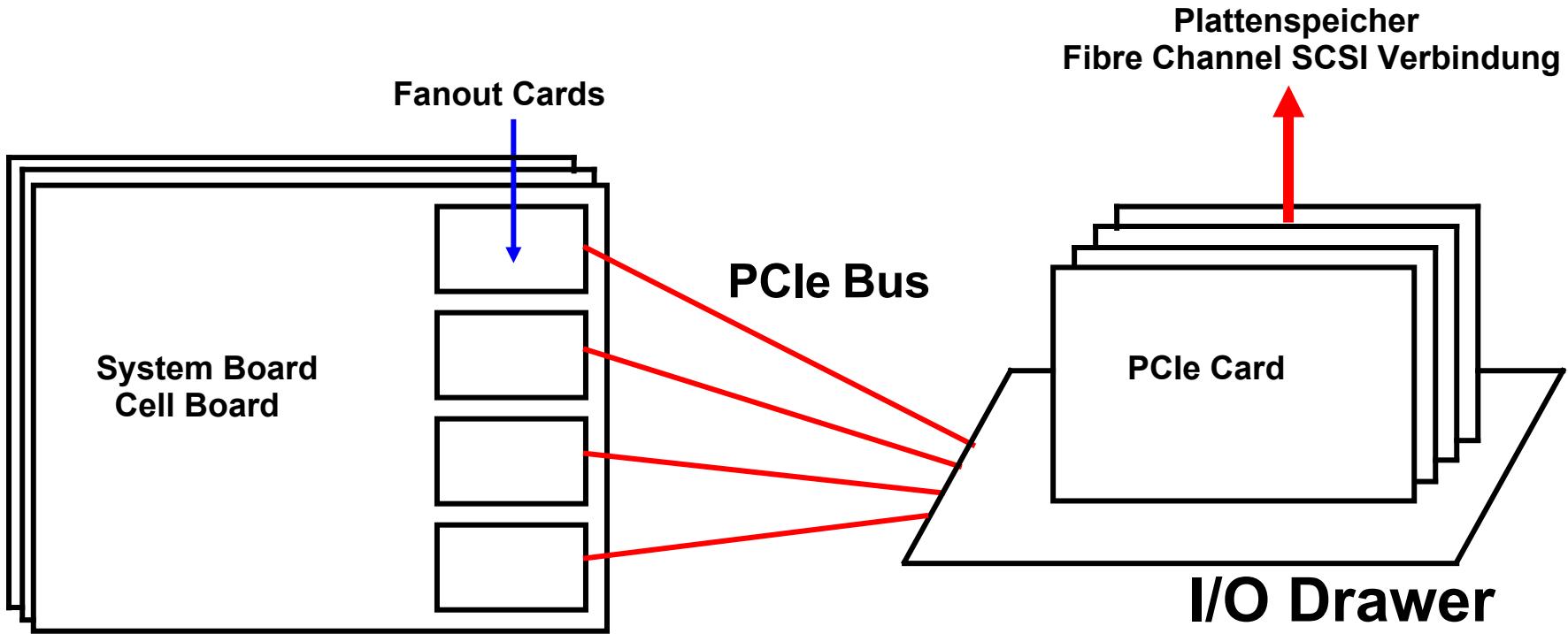


4 Itanium 2  
CPU Chips

Hauptspeicher  
2 x 16 DIMMs

6 Fanout Cards

Das Cell Board des Hewlett Packard Superdome Rechners hat sehr viel Ähnlichkeit mit einem Sun System Board.



## Sun Fire, Superdome Konfiguration

Die System- bzw. Cell Boards haben keinen Platz für den Anschluss von I/O Geräten, besonders Plottenspeichern, von denen evtl. hunderte oder mehr angeschlossen werden müssen. Statt dessen existieren Fanout Cards, die mittels DMA auf den Hauptspeicher zugreifen können.

Die Fanout Cards sind über Kabel mit PCIe Card Steckplätzen auf einem „I/O Cage“ Board verbunden. Am häufigsten sind PCIe Adapter Karten für Plottenspeicher Anschlüsse mittels Fibre Channel SCSI Kabeln anzutreffen. Auf diese Art ist es möglich, hunderte oder mehr Plottenspeicher an einen SUN oder HP Großrechner anzuschließen.

Hewlett Packard (HP) bezeichnet sein Prozessor Board als „Cell Board“. Ansonsten hat es sehr viel Ähnlichkeit mit dem System Board von Sun.

Spezifisch verfügen beide Processor Board Typen über Anschlüsse für 4 bzw. 6 Fanout Adapter Cards, die als Daughter Cards auf das Processor Board aufgesteckt werden. Typischerweise verwendet man für diese Slots PCIe Adapter Cards. Eine derartige PCIe Adapter Card verbindet das Prozessor Board über ein PCIe Kabel mit einem PCIe Board, welches eine Reihe von PCIe Karten-Slots verfügt.

System z verwendet an Stelle von Prozessor Boards sog. „Books“. Das Äquivalent zur Fanout Adapter Card wird als HCA (Host Channel Adapter) Card bezeichnet.

Alle Hersteller außer IBM setzen aus der PC-Welt abgeleitete Technologien für ihre Processor Boards ein. Besonders verwenden die Prozessor Boards die gleiche Printed Circuit Board (PCB) Technologie, die auch für die Mainboards der PCs benutzt wird.

Printed Circuit Boards bestehen aus einem elektrisch isolierenden Trägermaterial (Basismaterial), auf dem Kupferschichten aufgebracht sind. Die Schichtstärke beträgt typischerweise 35 µm. Das Basismaterial besteht meistens aus mit Epoxidharz getränkte Glasfasermatten. Die Bauelemente werden in der Regel auf Lötflächen (Pads) aufgegelistet. Multilayer Boards können aus bis zu 48 Schichten bestehen. PC Mainboards haben in der Regel unter 10 Schichten.

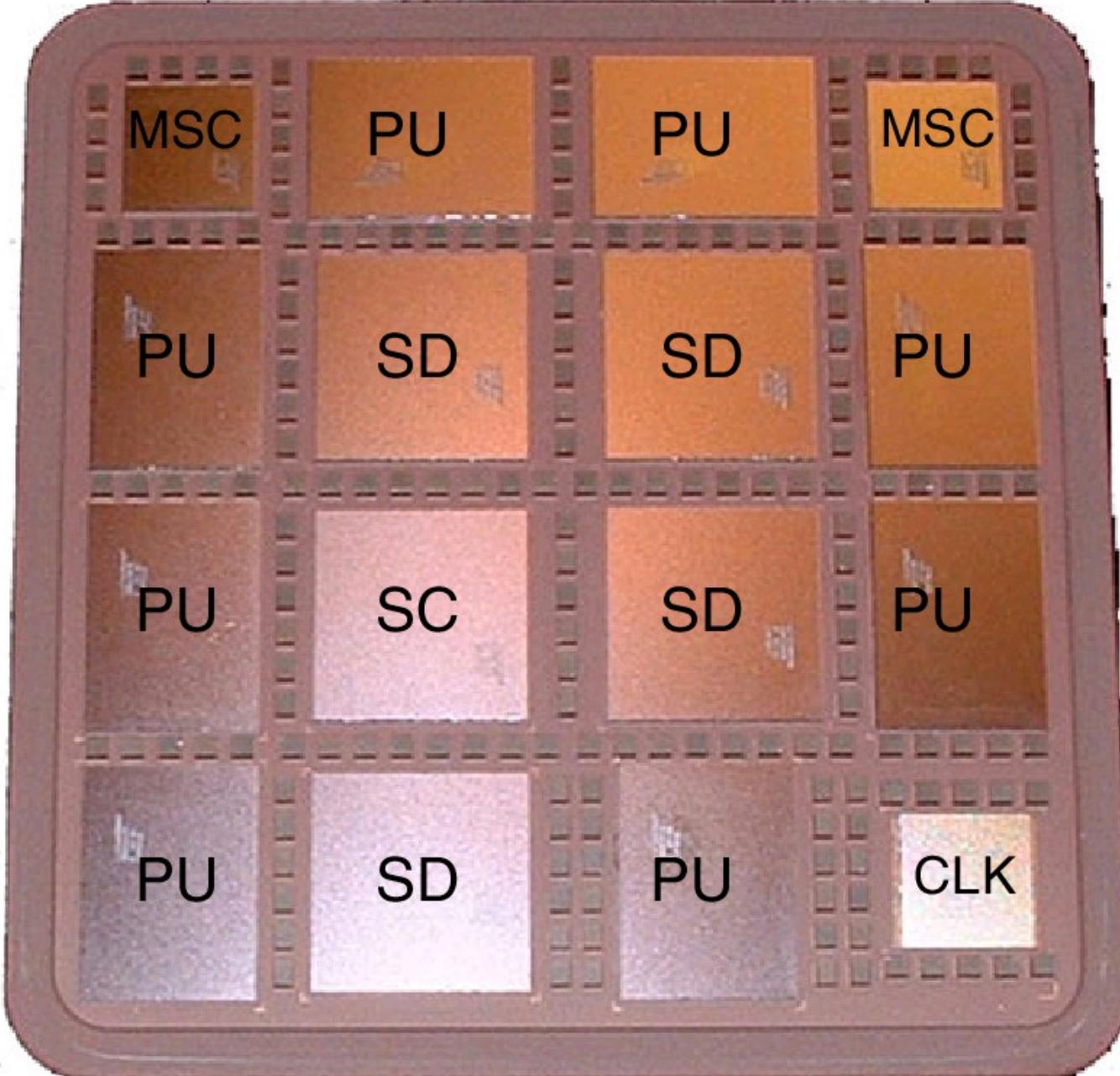
# **System z Hardware Teil 2**

## **Microprocessor Technologie**

## Multi-Chip Module eines z9 Rechners



Im Gegensatz dazu verwenden die System z Mainframes einen fundamental unterschiedlichen Ansatz. An Stelle eines Printed Circuit Boards wird ein „Multichip Module“ (MCM) eingesetzt.



Das z9 Multi Chip Module (MCM) benutzt eine Multilayer Ceramic (MLC) Technologie. Auf dem Module befinden sich:

- 8 dual Core CPU Chips (labeled PU), insgesamt 16 CPU Cores,
- 4 Level 2 (L2) Cache Chips labeled SD,
- 1 L2 Cache Controller Chip labeled SC,
- 2 Hauptspeicher Controller Chips labeled MSC,
- ein Clock Chip (CLK).

In der obigen Abbildung ist ein MCM (Multi Chip Modul), das Kernstück eines z9-Rechners gezeigt. Das MCM besteht aus einem 95 x 95 mm großem Multilagen-Glas-Keramik-Träger mit 102 Verdrahtungslagen. Auf dem Glas-Keramik-Modul sind 16 Chips aufgelötet. Die MCM-Technologie benutzt die Multilayer Ceramic (MLC) Technologie. MLC ermöglicht im Vergleich zur Printed Circuit Board Technologie besonders günstige Signallaufzeiten zwischen den Chips. Der Grund für die günstigeren Signallaufzeiten ist:

1. kleinere Abstände zwischen den Chips
2. günstigere Dielektrizitätskonstante von MLC im Vergleich zur Printed Circuit Technologie

Der Nachteil ist ein schwierigerer Produktionsprozess, den derzeitig in der Computerindustrie nur IBM einsetzt. Mehrere Unternehmen stellen MLC Substrate für Microwave Anwendungen her, z.B.

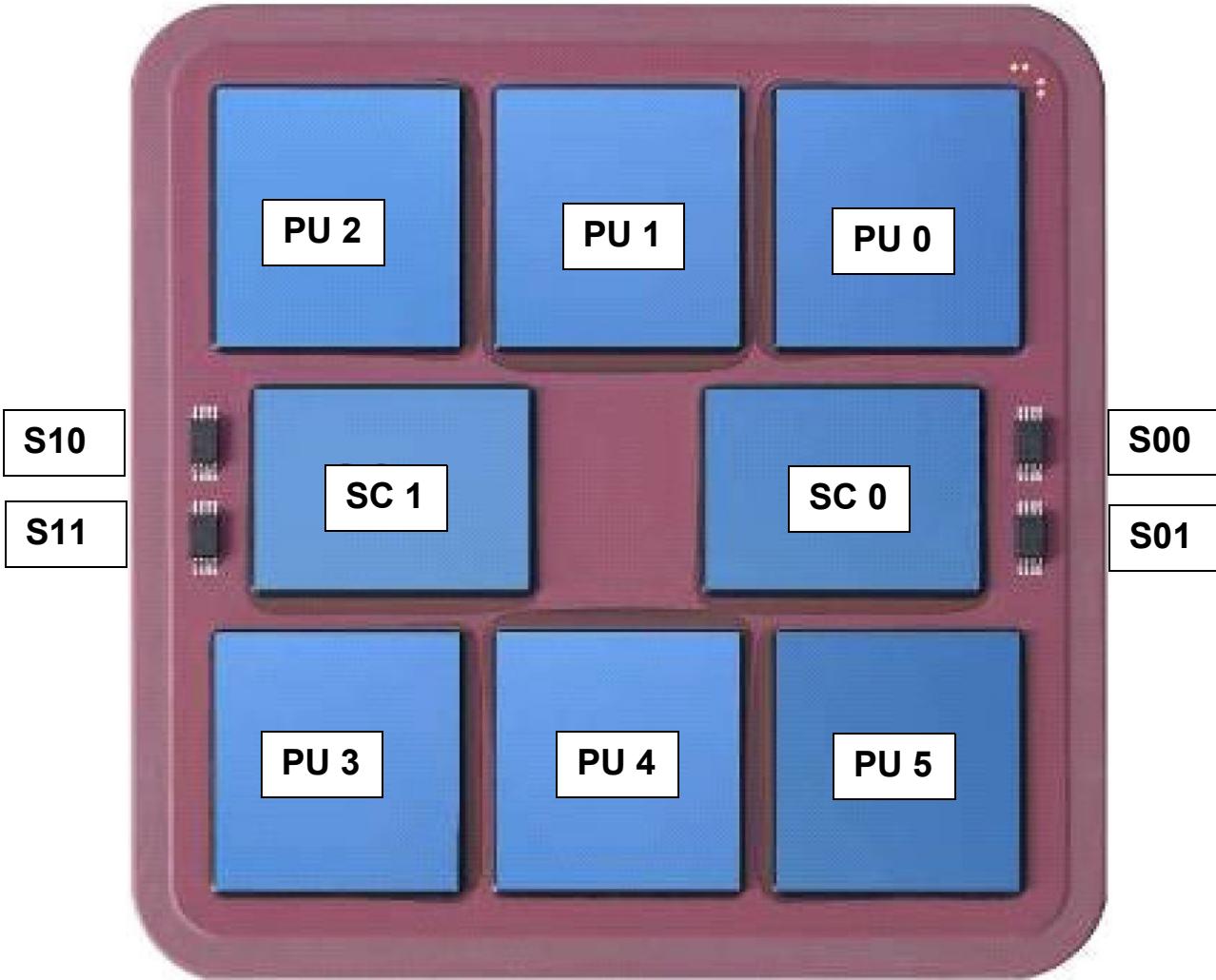
<http://www.ltcc.de/downloads/rd/pub/10-doc-plus-engl-2001.pdf>

IBM bringt verbesserte Mainframe Modelle etwa im 2 ½ Jahres Rhythmus heraus. Die vier letzten Modelle sind:

|              |                              |
|--------------|------------------------------|
| Modell z9    | vertrieben seit Juli 2005    |
| Modell z10   | vertrieben seit Februar 2008 |
| Modell z196  | vertrieben seit Juli 2010    |
| Modell zEC12 | vertrieben seit August 2012  |

Die neuen Modelle beinhalten in der Regel zahlreiche Verbesserungen, vor allem auch in der Halbleiter Technologie. Die MCM Technologie ist dagegen stabil und ändert sich nur wenig. Wir sehen deshalb auf dem MCM für das Modell zEC12 nur weniger, dafür aber größere Chips, wobei sich an den Abmessungen des Modules kaum etwas ändert.

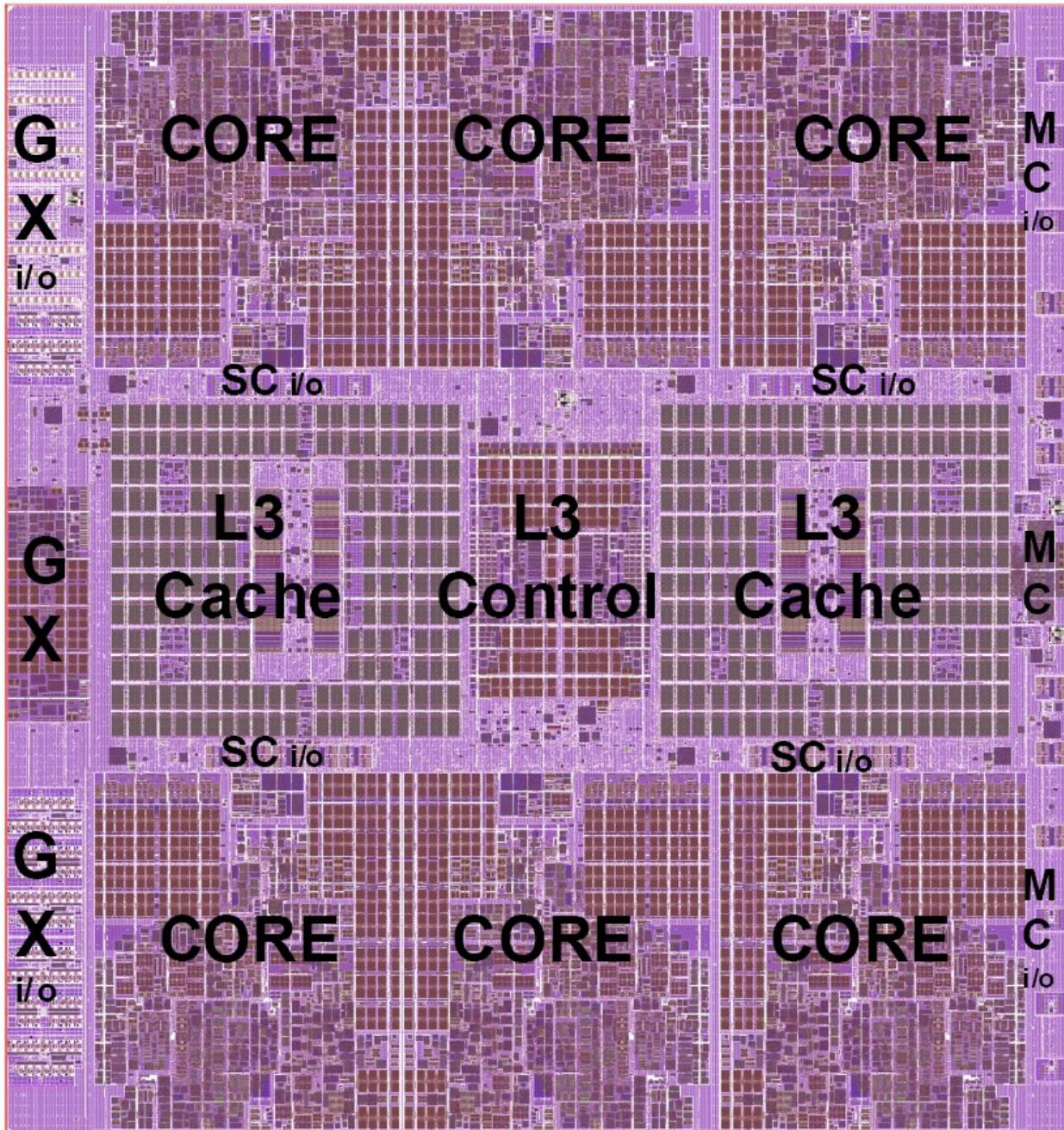
# **zEC12 MCM**



**Auf dem zEC12 Multi Chip Module (MCM) befinden sich:**

- **6 Hex core CPU Chips (labeled PU), insgesamt 36 CPU Cores,**
- **2 L4 Cache Chips labeled SC,**
- **4 EEPROM chips labeled S00, S01, S10 und S11. Sie dienen der Personalisierung (characterization) jedes einzelnen MCMs.**

Das MCM des Modells zEC12 besteht aus einem 96 x 96 mm großem Multilagen-Glas-Keramik-Träger mit 103 Verdrahtungslagen.

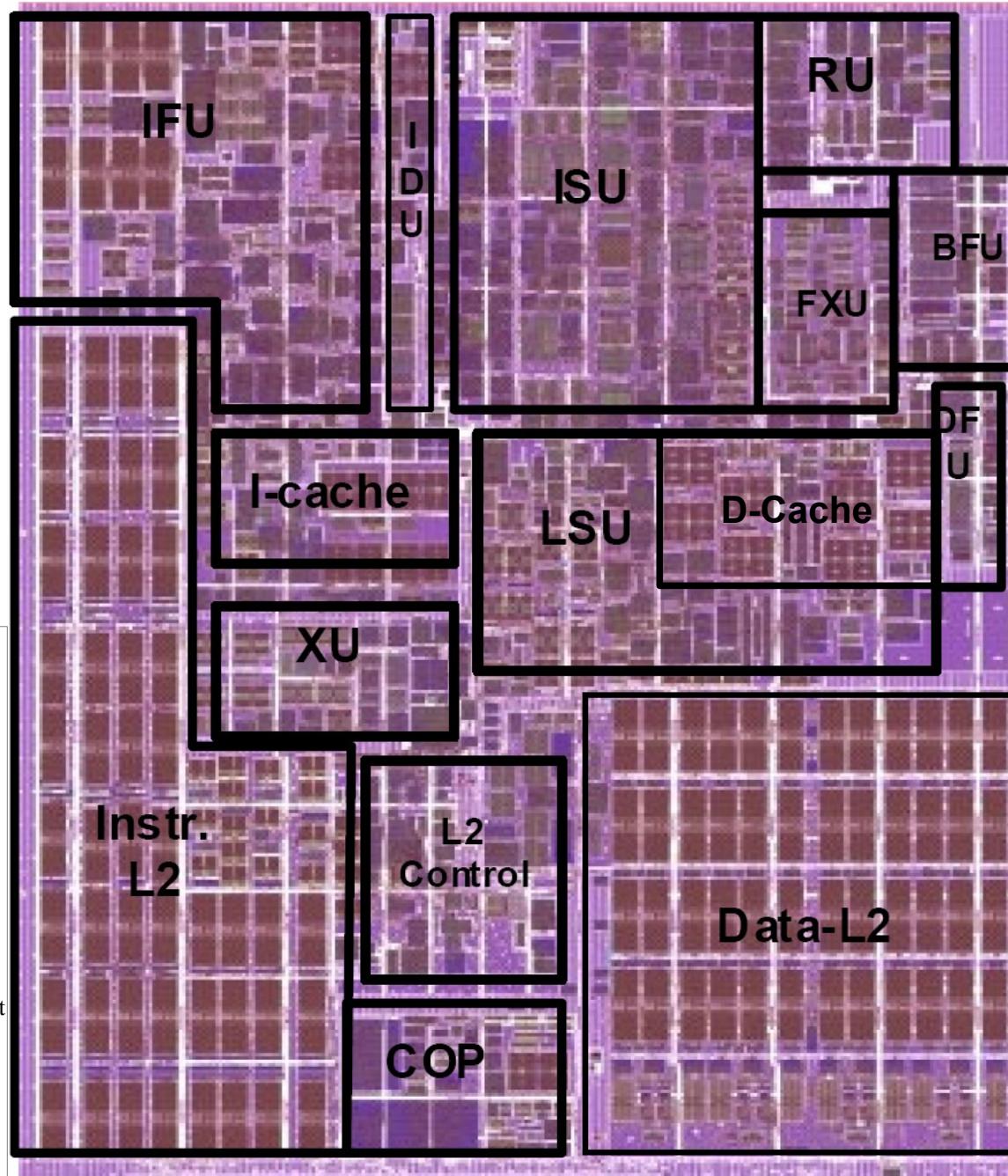


## **zEC12 CPU Chip Layout**

Auf jedem zEC12 CPU Chip befinden sich 6 identische CPU Cores. Die Chips sind mit 5,5 GHz getaktet.

Weiterhin befindet sich auf dem zEC12 CPU Chip ein L3-Cache mit 48 Mbyte, bestehend aus 2 Compartments. Der L3-Cache ist ein Store-in Cache Speicher, der von allen Cores des CPU Chips gemeinsam genutzt wird. Die L3 Control Einheit verfügt über einen integrierten On-Chip-Kohärenz-Manager und einen Crossbar-Schalter. Beide L3 Compartments können gleichzeitig bis zu 160 GByte/s Bandbreite an jeden Core übertragen. Der L3-Cache verbindet die sechs Kernen, GX I/O-Busse und Hauptspeicher-Controller (MCs) mit getrennten Storage Controller (SC)-Chips. Die Hauptspeicher-Controller (MC) Funktion steuert weiterhin den Zugriff auf den Hauptspeicher. Der GX I/O-Bus steuert die Schnittstelle zu den Host Channel Adapter (HCA), welche die Verbindung zu getrennten I/O Adapters herstellen.

Insgesamt steuert das zEC12 CPU Chip den Datenverkehr zwischen den CPU Cores, I/O sowie dem L4-Cache auf dem SC-Chips.



**IFU:**  
Instruction Fetching Unit  
**IDU:**  
Instruction Decode Unit  
**ISU:**  
Instruction Sequence Unit  
**RU:**  
Recovery Unit  
**BFU:**  
Binary Floating-point Unit  
**FXU:**  
Fixed-point Unit  
**DFU:**  
Decimal Floating-point Unit  
**LSU:**  
Load-Store Unit  
**XU:**  
Translation Unit  
**COP:**  
dedicated Co-Processor

## Layout eines CPU Cores

Jeder der sechs Cores hat einen eigenen L1-Cache mit 64 KByte für Befehle und 96 KByte für Daten. In jedem Core befindet sich ein privater L2-Cache mit 1 MByte für Befehle und 1 MByte für Daten.

Jedes CPU Chip hat ca. 2,75 Milliarden ( $10^9$ ) Transistoren. Die Abmessungen sind 23,5 x 21,8 mm. Es implementiert 6 CPU Cores und eine 4-stufige Cache Hierarchie, die aus L1, L2, L3 und L4 Caches besteht.

Im Vergleich dazu hatte die 1988 erschienene CMOS Implementierung eines S/370 Prozessors 200 000 Transistoren.

Der Energieverbrauch beträgt etwa 300 Watt pro zEC12 CPU Chip.

## **Sechs oder acht Cores pro Chip ?**

Ein Blick auf das Layout eines Cores des zEC12 CPU Chips zeigt, dass vier L2 Caches mit 4 MByte Speicherkapazität in etwa den gleichen Platz in Anspruch nehmen wie der gemeinsam genutzte L3 Cache mit insgesamt 24 MByte Speicherkapazität.

Weiterhin zeigt das zEC12 Chip Layout, dass die 6 CPU Cores (ohne L2 cache) deutlich weniger als 50 % der Chip Fläche in Anspruch nehmen. Es wäre denkbar gewesen, auf dem CPU Chip 8 Cores unterzubringen.

Bei der Firma IBM entwickelt die gleiche Mannschaft neue PowerPC und System z CPU Chips. Nicht überraschend weisen beide CPU Core Implementierungen viele Gemeinsamkeiten auf. Die gemeinsam mit dem zEC12 Chip entwickelte Version des PowerPC Microprozessors wird als Power7 bezeichnet.

Während man sich beim Power7 für ein 8 Core Chip entschieden hat, ist das System z Team bei 6 Cores geblieben, um dafür eine sehr komplexe Cache Hierarchie mit maximalen Cache Größen unterzubringen.

Wir werden in der Zukunft öfters Diskussionen erleben, ob mit wachsender Integrationsdichte es besser ist, die Anzahl der Cores zu vergrößern, oder mehr Platz für Cache Speicher zur Verfügung zu stellen.

## **Sun/Oracle T5 und M5 Chips**

Eine ähnliche Entscheidung hat man bei der Firma Sun/Oracle getroffen

Das neue (2013) T5 Sparc Chip der Firma Sun/Oracle hat 16 CPU Cores und 8 MByte L3 Cache. Eine Variante, das M5 Sparc Chip, benutzt identische CPU Cores, hat aber nur 6 an Stelle von 16 Cores. Der freiwerdende Platz wurde benutzt, um den L3 Cache von 8 MByte auf 32 MByte zu vergrößern.

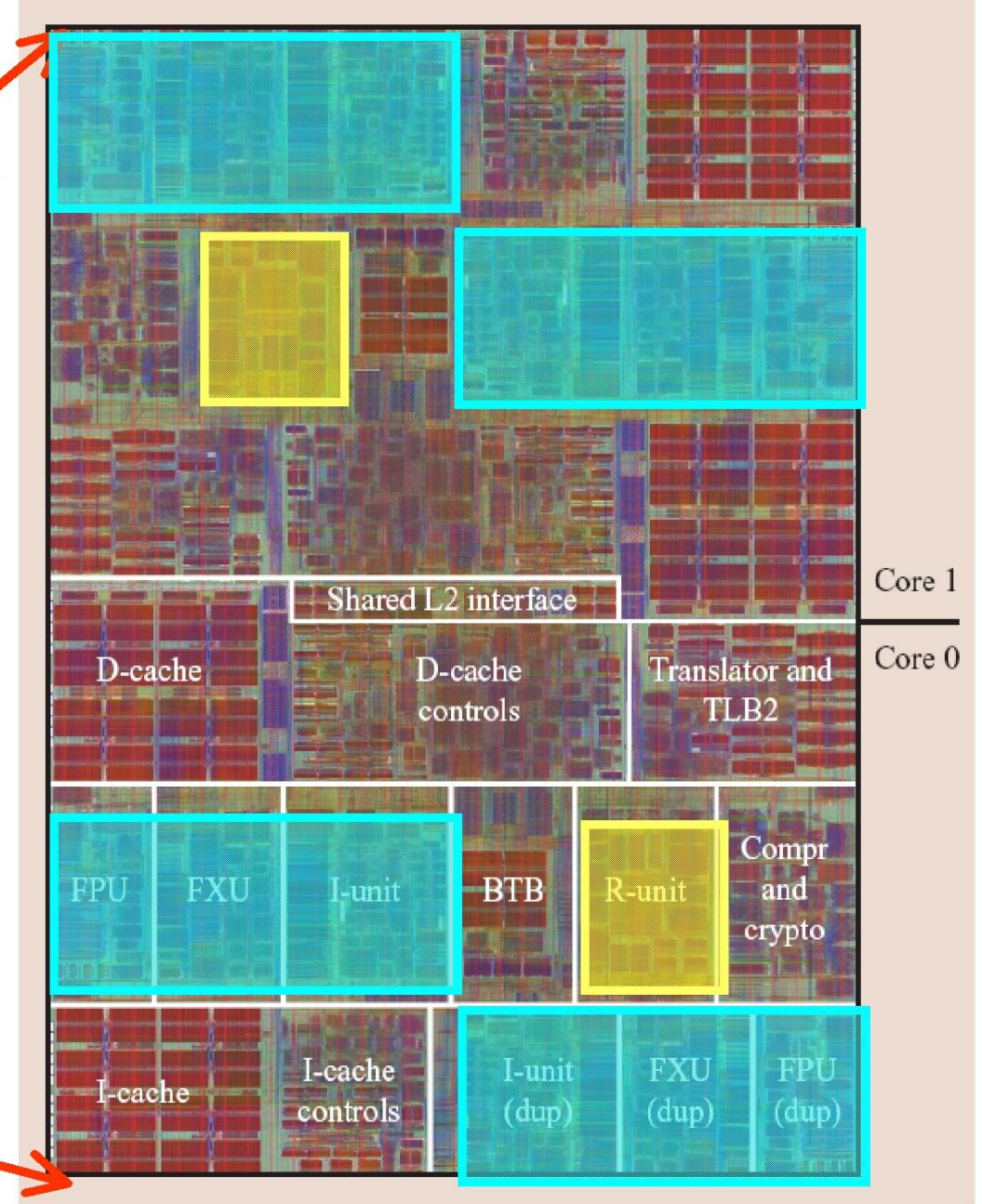
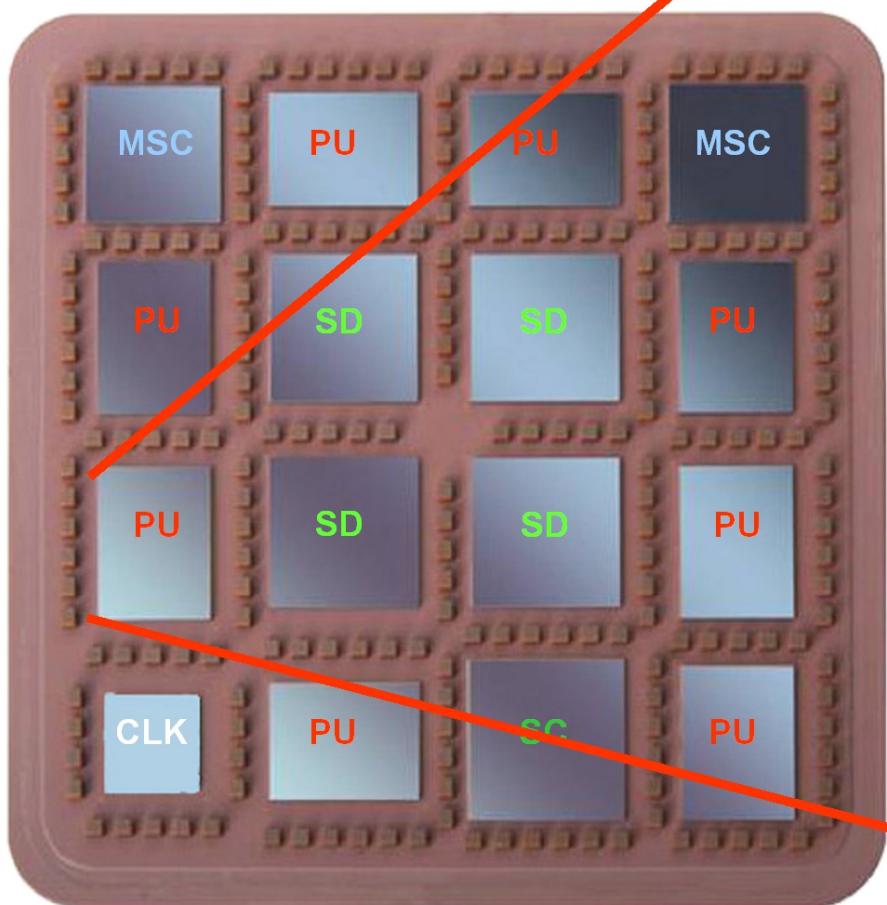
Das T5 Chip ist für High CPU Performance Server vorgesehen, während das M5 Chip für betriebswirtschaftliche Großrechner vorgesehen ist, die evtl. mit Mainframes konkurrieren sollen.

Das M5 Chip wird in dem derzeitigen Spitzenprodukt, dem Sparc M5-32 Server eingesetzt, Nachfolger des bisherigen M9000 Servers. Ein M5-32 Server hat 32 M5 CPU Chips, oder bis zu 192 CPU Cores.

Einzelheiten unter:

<http://www.oracle.com/technetwork/server-storage/sun-sparc-enterprise/documentation/o13-024-m5-32-architecture-1920556.pdf?ssSourceSiteId=ocomen>

# Zusätzliche Eigenschaften des System z CPU Chips



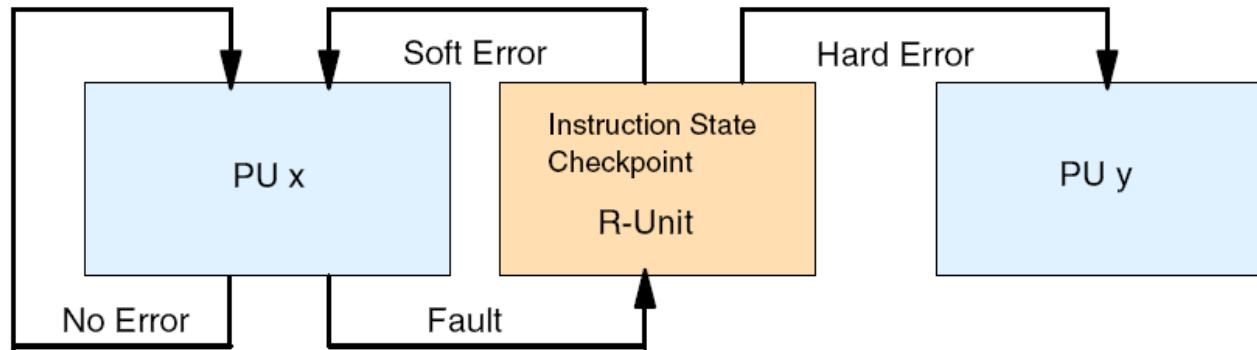
## Zusätzliche Eigenschaften des System z CPU Chips

Im Vergleich zum Power PC, x86, Sparc oder Itanium Chip wird bei System z ein sehr viel höherer Aufwand für Sicherheit und Verfügbarkeit getrieben. Dies sei am Beispiel des oben dargestellten z9 CPU Chips erläutert:

Es handelt sich um ein Dual Core Chip. Die obere und die untere Hälfte stellen je einen Core da. In der Mitte befindet sich die Schnittstelle zum L2 Cache (getrennte Chips auf dem gleichen MCM), die von beiden Cores gemeinsam genutzt wird.

Die wichtigsten Elemente in jedem Core sind die Instruction Unit (I-unit), Fixed Point Execution Unit (FXU) und Floating Point Execution Unit (FPU). Jedes Core enthält alle drei Units in zweifacher Ausführung. Maschinenbefehle werden unabhängig und (nahezu, aber nicht exakt parallel) auf beiden Kopien der I-, FXU- und FPU Units ausgeführt. Mittels einer Compare Funktion wird verifiziert, dass beide Kopien das gleiche Ergebnis erzeugen. Wenn nicht, greifen automatische Fehlerbehebungsmaßnahmen ein, z.B. eine Maschinenbefehlwiederholung (instruction retry). Dies geschieht unbemerkt vom Betriebssystem oder Benutzerprogramm.

# Recovery Unit



Von besonderem Interesse ist in diesem Zusammenhang die „Recovery Unit“ (RU). Wenn während der Ausführung eines Maschinenbefehls ein Fehler entdeckt wird, versucht die Instruction Unit des CPU Cores den Befehl ein zweites Mal auszuführen, in der Hoffnung, dass diesmal kein Fehler auftritt. Wenn eine Maschinenbefehlwiederholung nicht erfolgreich ist (z.B. ein permanenter Fehler existiert), wird ein Relocation Process gestartet. Dieser bewirkt, dass die Prozessausführung auf einem anderen CPU Core fortgesetzt wird. Das ist möglich, weil in jedem Augenblick der vollständigen Architekturstatus der CPU in ihrer R-Unit zwischenspeichert wird, wobei diese Zwischenspeicherung wiederum über Hamming Fehlerkorrekturcodes abgesichert ist.

# Cryptography Accelerator

Der z9 Rechner hat eine Compression und Crypto Unit, beim zEC12 als Coprocessor bezeichnet. Die Crypto Unit wird u.a. eingesetzt, um die Verschlüsselung und Entschlüsselung von SSL (Secure Socket Layer) Nachrichten zu beschleunigen.

Die folgende Abbildung demonstriert den Nutzen der Crypto Unit. Angenommen ist eine 6-Prozessor Einheit, die eine theoretische Leistung von 600% verglichen mit einem einzelnen Prozessor erbringen kann.

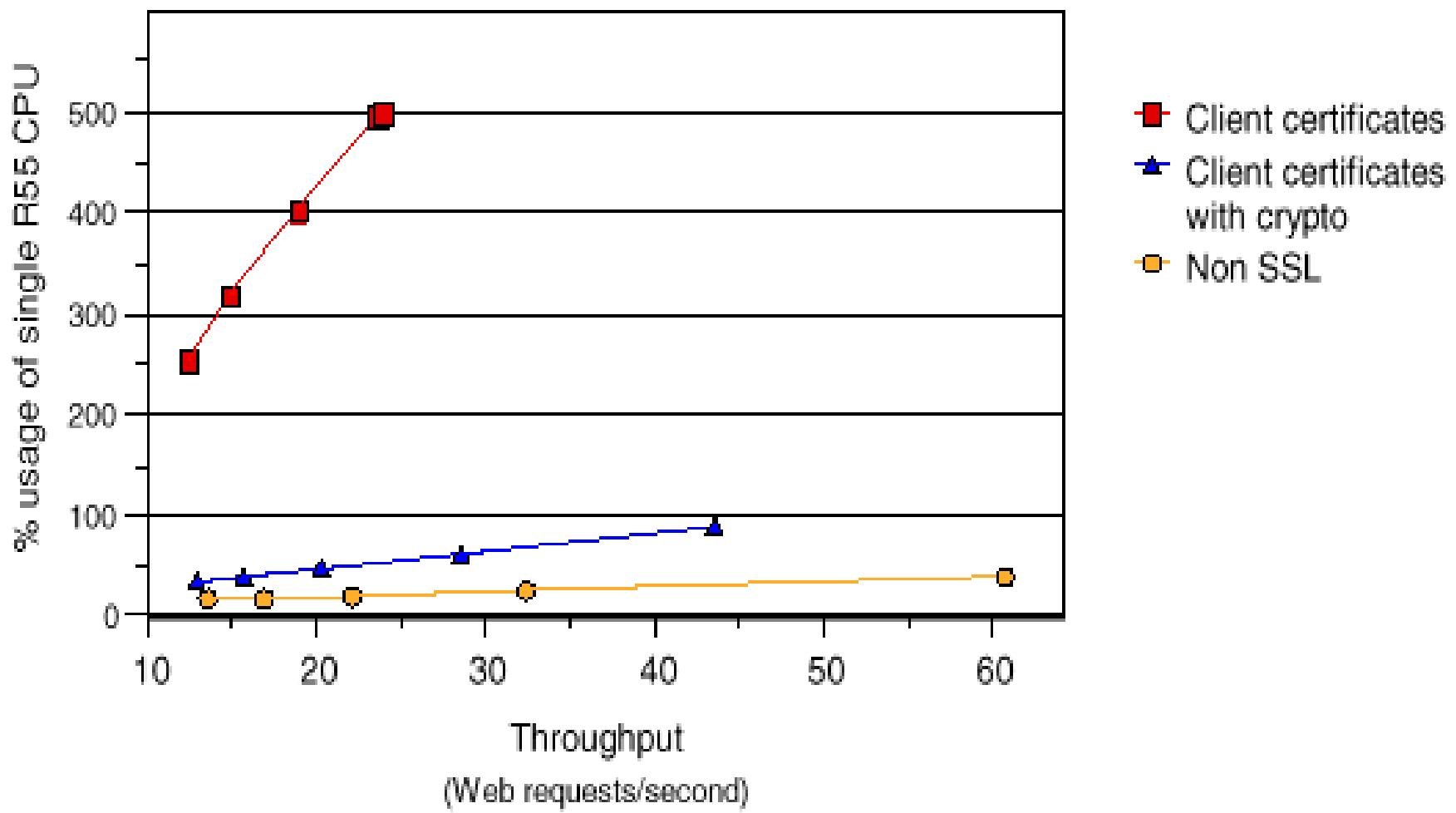
Die gelbe Kurve stellt die CPU Auslastung für eine bestimmte Art von Web Requests pro Sekunde dar. Ohne SSL beträgt die Auslastung bei 60 Transaktionen/s weniger als 50 % der Leistung einer einzigen CPU.

Beim Einsatz von SSL, aber ohne Crypto Unit (rote Kurve) steigt die CPU Auslastung schon bei 25 Transaktionen/s auf 500 % (fünf CPUs).

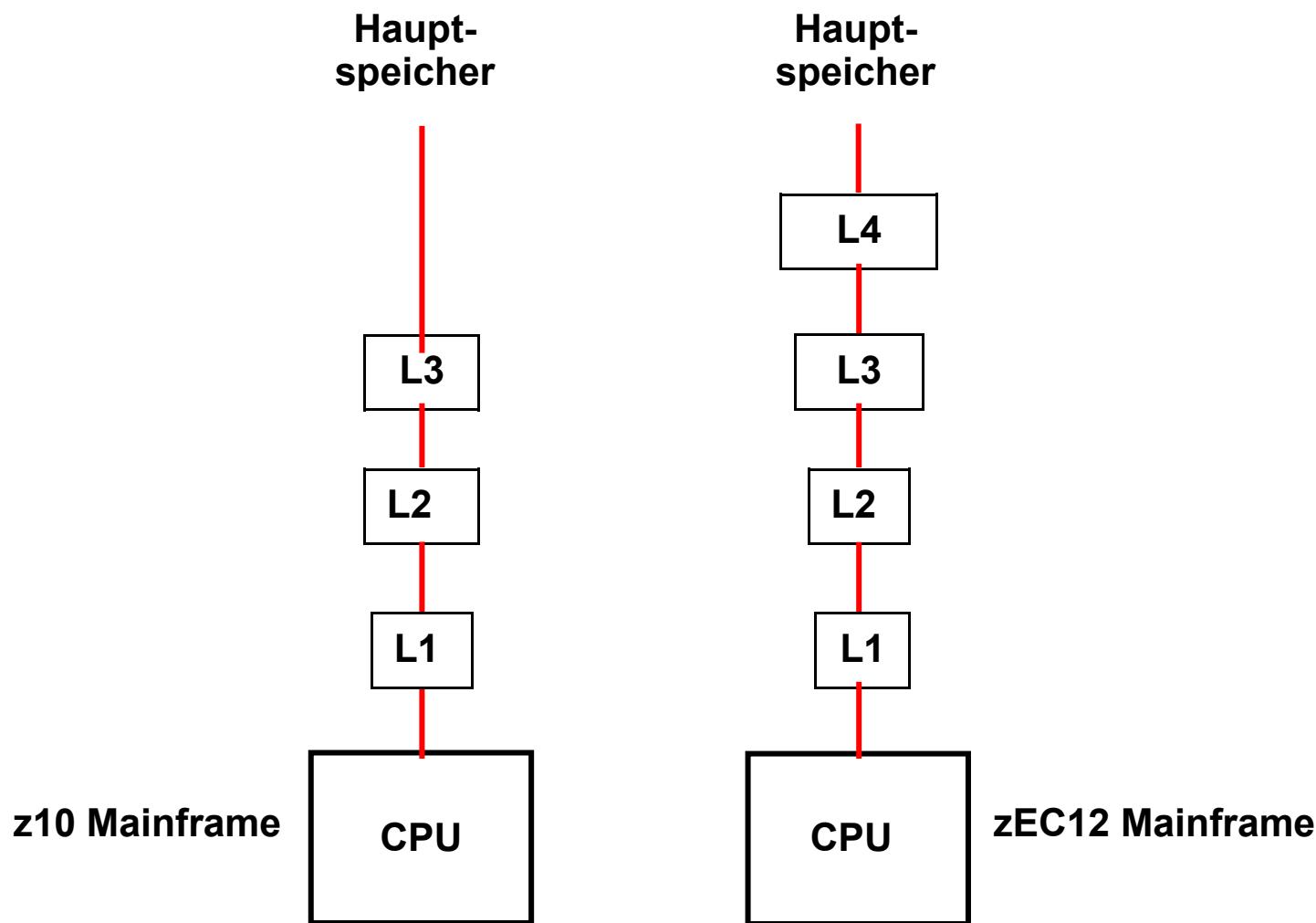
Wird die Crypto Unit für die Verschlüsselung eingesetzt, entsteht die blaue Kurve. Die CPU Auslastung ist zwar deutlich höher als ohne SSL, aber deutlich besser als SSL ohne Crypto Unit.

Die hier wiedergegebenen Messdaten verwenden einen sehr einfache Web Request, der selbst nur wenig CPU Auslastung bewirkt. Bei komplexeren Web Requests ist der Unterschied weniger dramatisch.

**SSL handshakes with client certificates**  
**CWS direct connection**  
**Throughput vs. CPU usage**

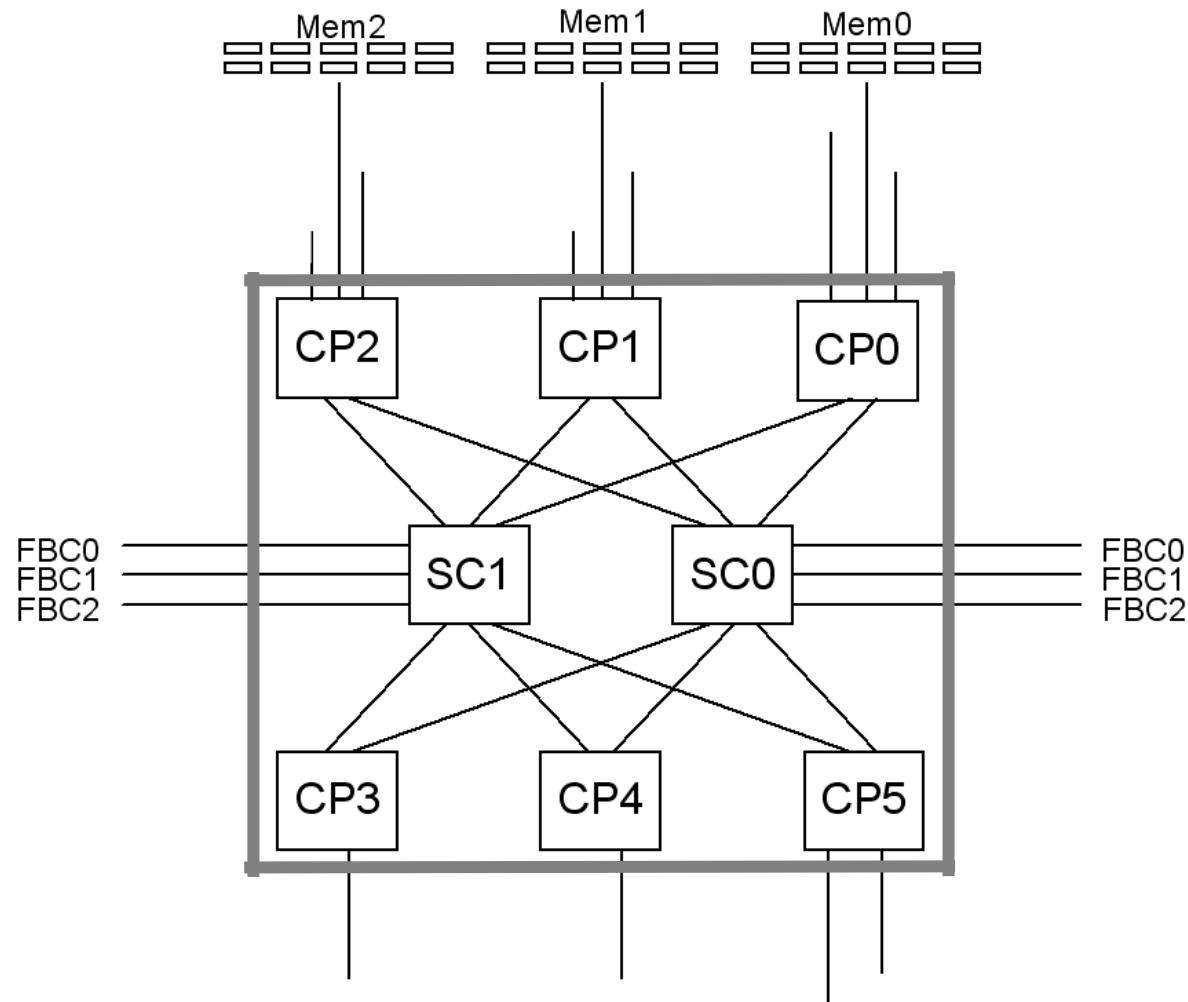


# Mainframe Cache Hierarchien



Heutige Mainframes haben eine drei- oder vier-stufige Cache Hierarchie.

# z12 Multichip Module



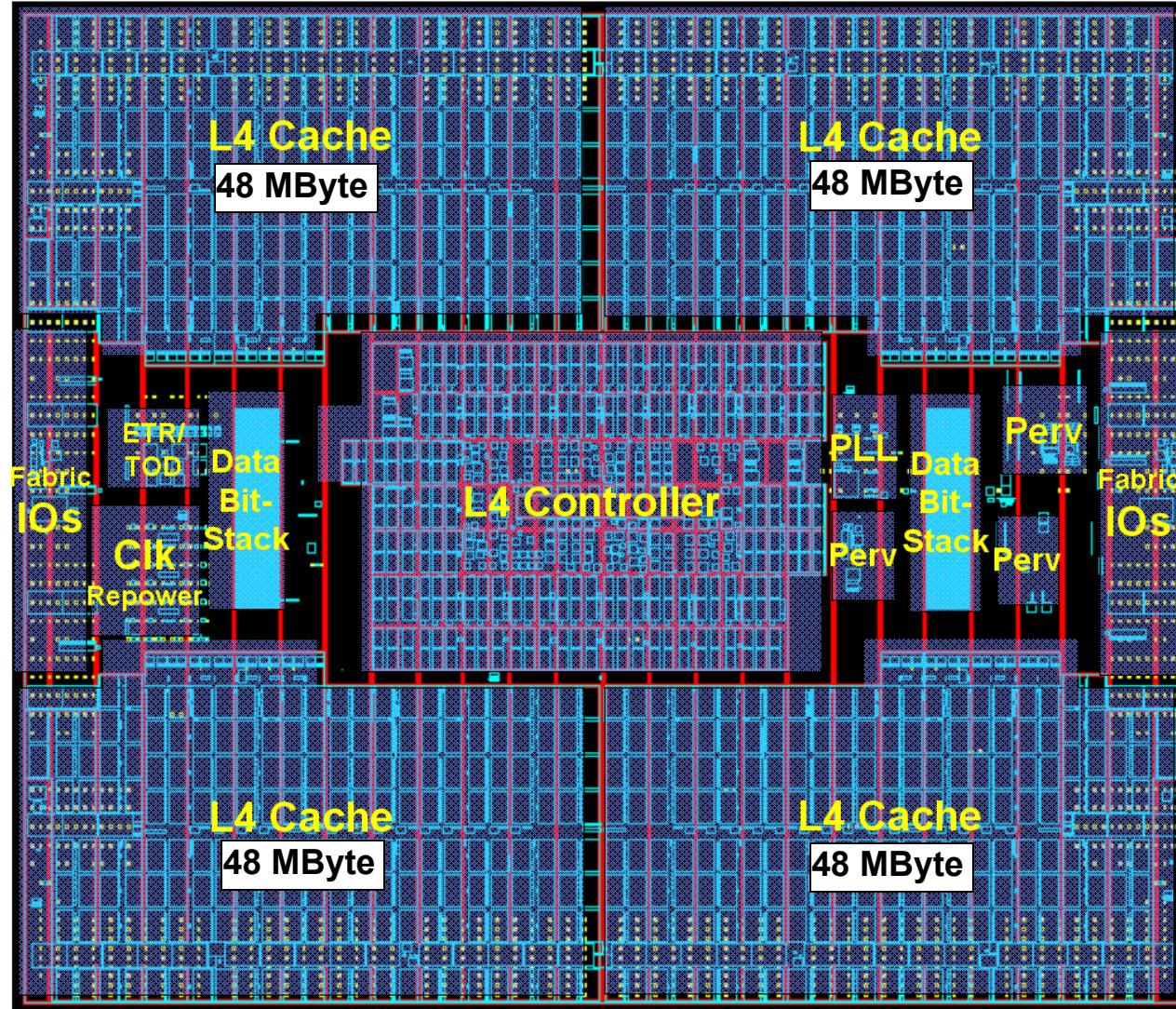
Auf dem zEC12 Multichip Module befinden sich sechs CPU Chips (CP0 ..CP5) sowie zwei L4 Cache Chips SC0, SC1.

Jedes CPU Chip ist mit beiden L4 Cache Chips direkt verbunden  
Außerdem enthalten 3 der 6 CPU Chips eine Memory Management Unit und sind direkt mit den Hauptspeicher DIMMs verbunden, von denen bei Bedarf eine Cache Line nachgeladen werden kann.

Ein z12 System kann vier als „Books“ bezeichnete Baugruppen enthalten. Jedes Book beinhaltet ein Multichip Module.

Die L4 Cache Chips aller Books sind miteinander über „Fabric Book Connectivity (FBC)“ Leitungen verbunden und bilden einen gemeinsam genutzten Cache.

## L4 Cache Chip



Neben den sechs CPU Chips befinden sich auf dem zEC12 MultiChip Module (MCM) noch zwei der hier gezeigten L4 Cache (SC) Chips. Jedes Chip hat Abmessungen von 28.4 x 23.9 mm, enthält 3.3 Milliarden ( $10^9$ ) Transistoren und 2,1 Milliarde dynamische Speicherzellen (eDRAM). Neben dem Cache Controller speichert es 192 MByte.

# eDRAM

Ein Static Random Access Memory (SRAM) benötigt zwischen 4 – 10 Transistoren in einer Flip-Flop Schaltung für jedes gespeicherte Bit. Ein Dynamic Random Access Memory (DRAM) benötigt 1 Transistor und einen kleinen Kondensator für jedes gespeicherte Bit. Ein DRAM Speicher kann auf einer gegebenen Chip Fläche wesentlich mehr Bits unterbringen als ein SRAM Speicher, ist dafür aber wesentlich langsamer (z.B. 100 ns versus 1 ns Zugriffszeit).

Die Hauptspeicher nahezu aller Rechner verwenden fast immer DRAMS, häufig in der Form von SIMM oder DIMM Steckkarten. Cache Speicher werden fast immer in SRAM Technologie implementiert.

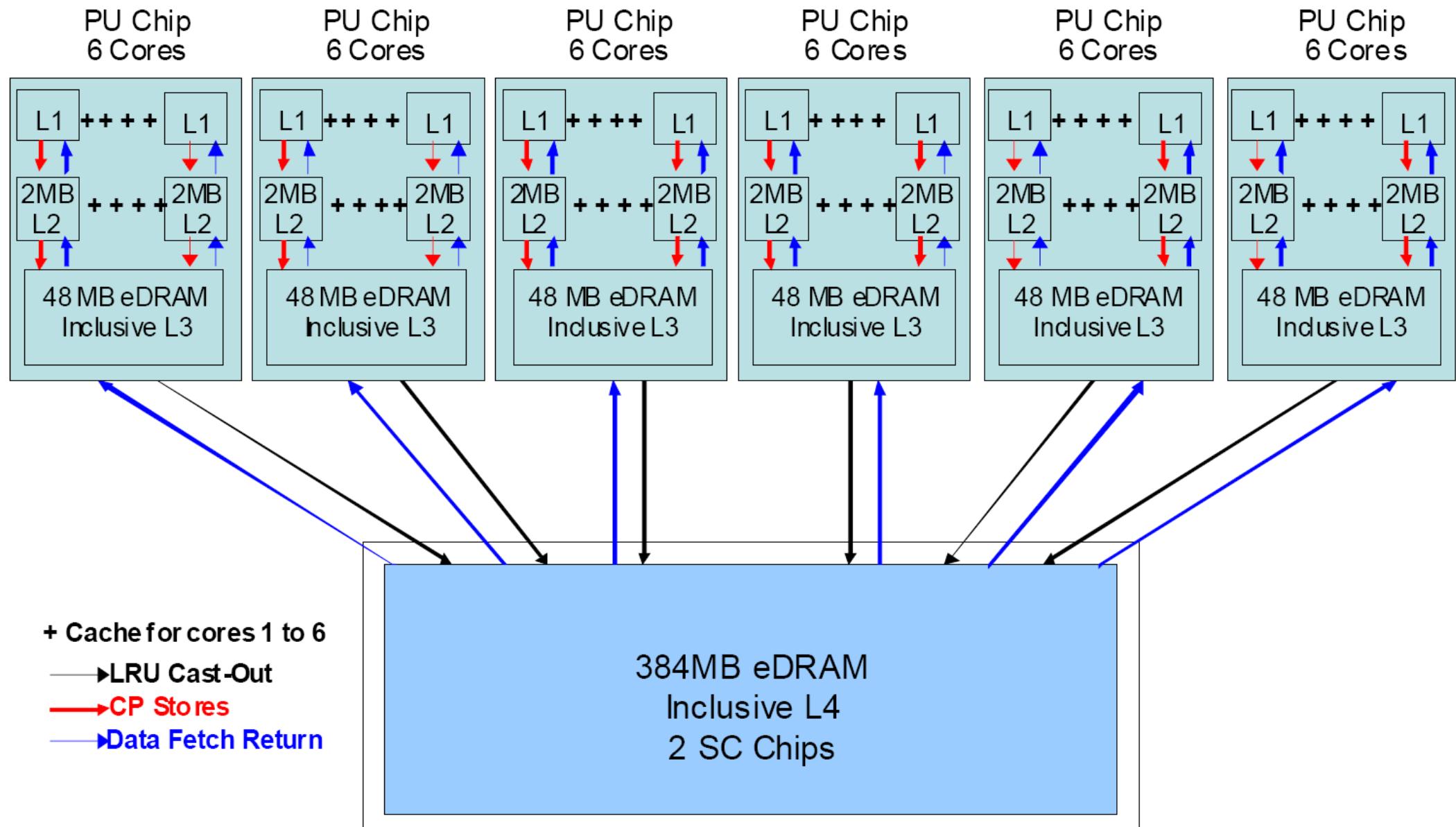
Die L3 und L4 Caches eines zEC12 Rechners werden in einer neuartigen eDRAM (embedded DRAM) Technologie implementiert. eDRAM ist nahezu so schnell wie SRAM, benötigt aber viel weniger Platz.

Ein **embedded DRAM (eDRAM)** ist ein auf DRAM basierender eingebetteter Speicher. Das bedeutet, das eDRAM ist auf dem gleichen Chip wie der Microprozessor (die CPU) integriert (eingebettet). Im Gegensatz zu externen DRAM-Speichermodulen wird er häufig wie ein transistorbasiertes SRAM als Cache genutzt.

Das Einbetten des Speichers ermöglicht gegenüber externen Speichermodulen die Nutzung größerer Busse und höhere Arbeitsgeschwindigkeiten. Durch den geringeren Platzbedarf ermöglicht eDRAM verglichen mit SRAM eine höhere Datendichte; somit kann bei gleicher Chip-Größe potentiell mehr Speicher genutzt werden. Jedoch machen die Unterschiede im Herstellungsprozess zwischen DRAM und Transistorlogik die Integration auf einem Chip kompliziert, das heißt, es sind in der Regel mehr Prozessschritte notwendig, was die Kosten erhöht.

eDRAM wird nicht nur als L3- und L4-Cache-Speicher im zEC12 Rechner, sondern auch in vielen Spielekonsolen genutzt, z.B. Xbox 360 von Microsoft, Wii von Nintendo und PlayStation 3 von Sony.

# Cache Struktur eines zEC12 Multichip Modules



Die System z Prozessoren verwenden eine 4-stufige Cache Hierarchie.

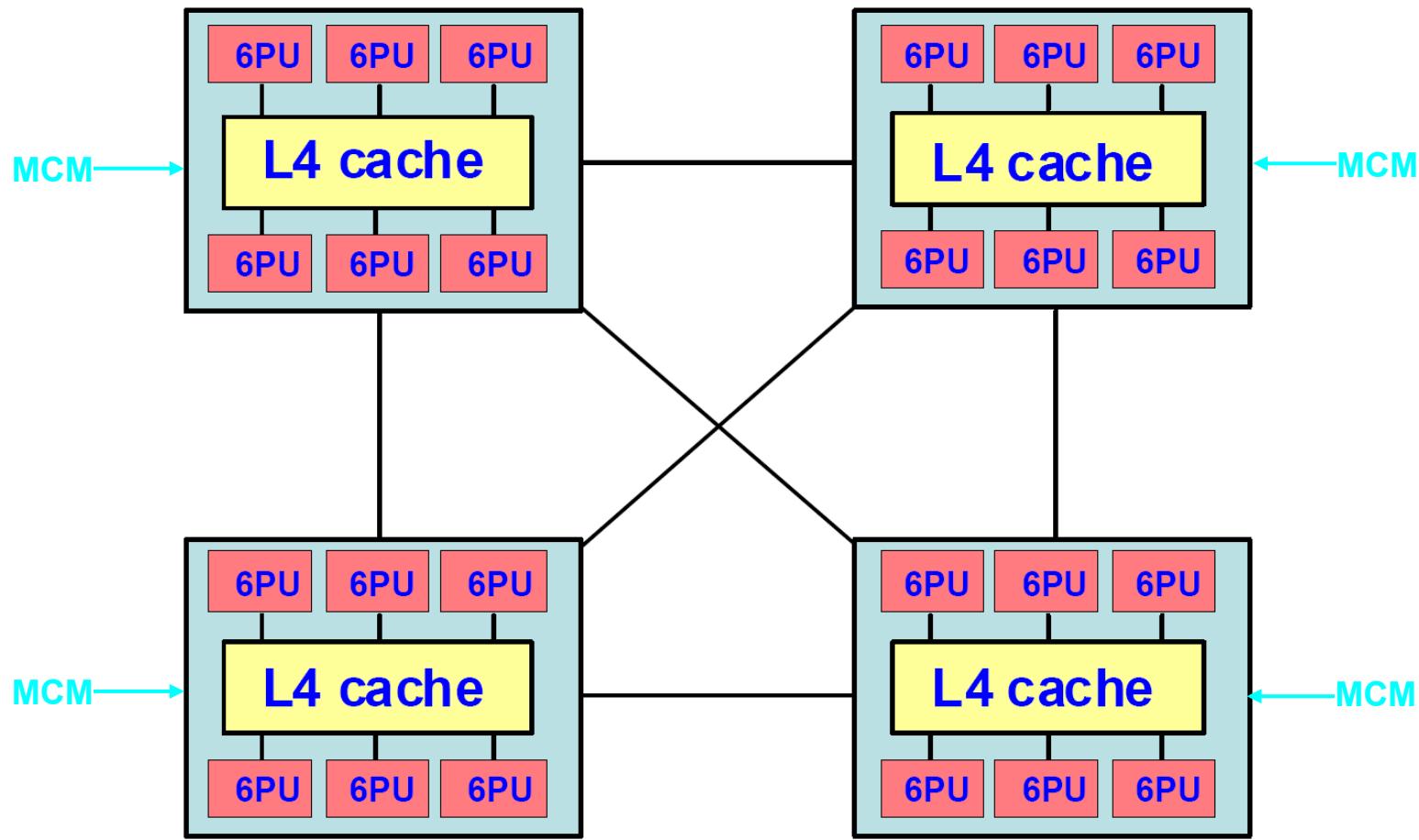
Auf dem CPU Chip befinden sich 6 Cores. Jeder Core hat seine eigenen privaten L1 und L2 Cache. L1 und L2 verwenden unterschiedliche SRAM Technologien und haben unterschiedliche Zugriffszeiten.

Alle 6 Cores des CPU Chips verwenden einen gemeinsam genutzten (shared) L3 Cache.

Die 6 CPU Chips eines Multichip Modules verwenden einen gemeinsam genutzten L4 Cache.

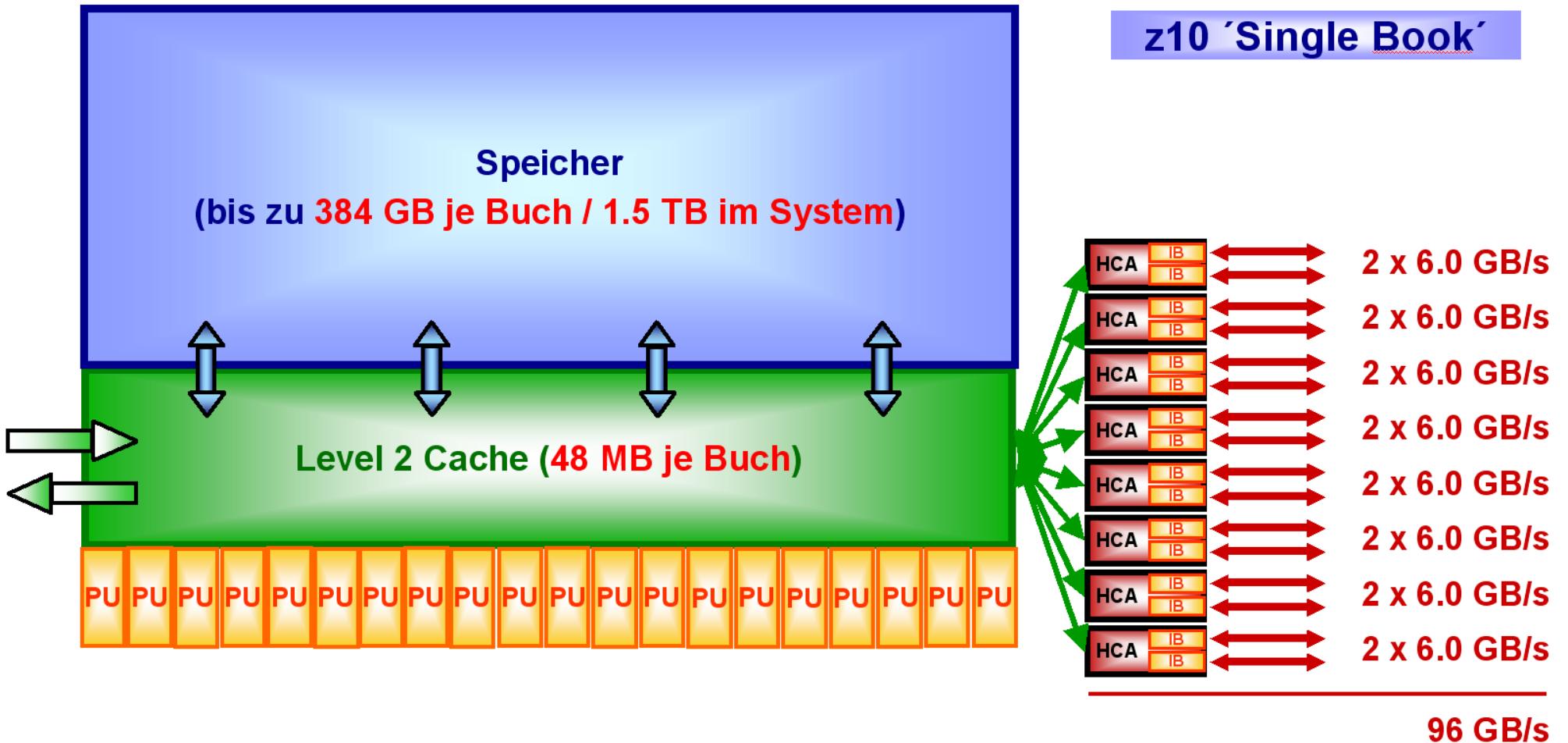
Das MCM ist Bestandteil einer als „Book“ bezeichneten Baugruppe. Ein zEC12 Rechner enthält bis zu 4 derartiger Books und damit 4 MCMs. Die vier L4 Caches der vier MCMs sind miteinander verbunden und bilden einen von allen CPU Cores gemeinsam genutzten NUMA (Non-Uniform Memory Architecture) L4 Cache.

Ein NUMA Speicher ist dadurch gekennzeichnet, dass die Zugriffszeiten zu Teilen des Speichers unterschiedlich sein können.



Die vier L4 Caches (je 2 L4 Cache Chips) der vier z12 Books mit je 36 CPU Cores sind mittels einer Punkt zu Punkt Topologie über „Fabric Book Connectivity (FBC)“ Leitungen direkt miteinander verbunden. Damit ist ein direkter Datenaustausch zwischen den vier L4 Caches möglich.

Alle  $4 \times 36 = 144$  Cores greifen auf die L4 Caches aller Books direkt zu. Die vier L4 Caches implementieren einen NUMA (Non Uniform Memory Architecture) Cache.

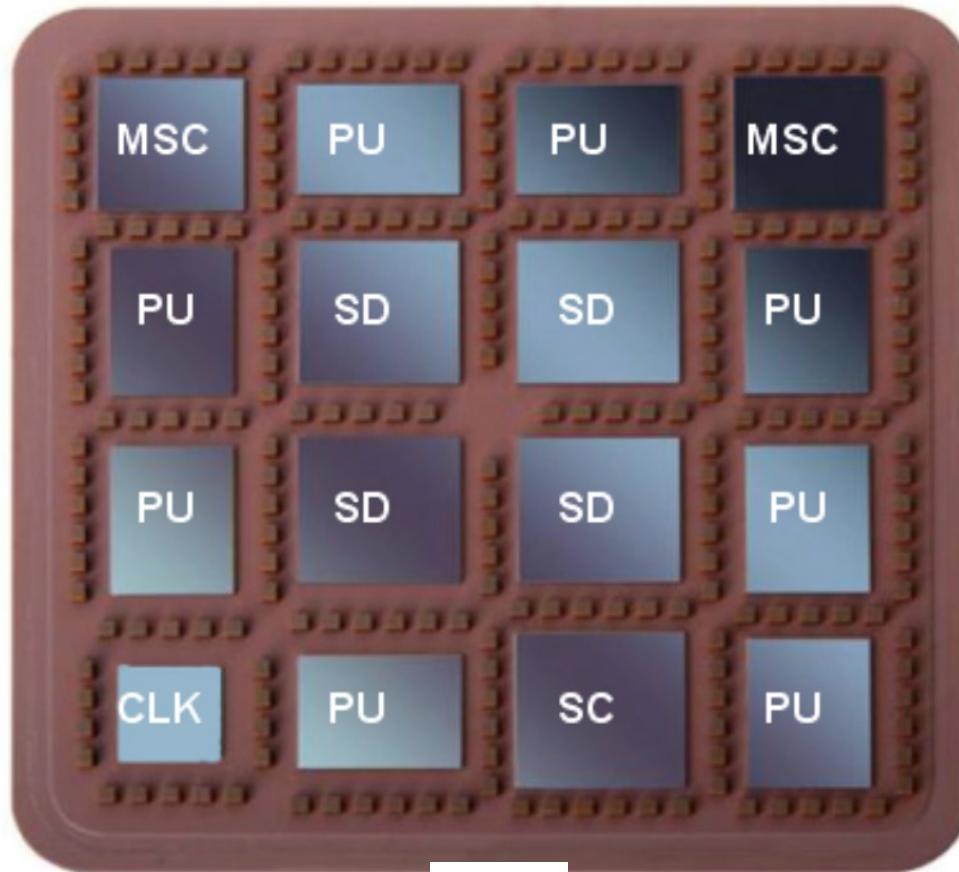


Ein weiteres Mainframe Alleinstellungsmerkmal: In allen nicht-Mainframe Servern (und allen PCs) bewirken die Fanout Adapter Karten einen Datentransfer zwischen Plattspeicher und Hauptspeicher. In den Mainframe Rechnern erfolgt der Datentransfer zwischen Plattspeicher und Cache. Damit sind natürlich wesentlich höhere Datenraten möglich.

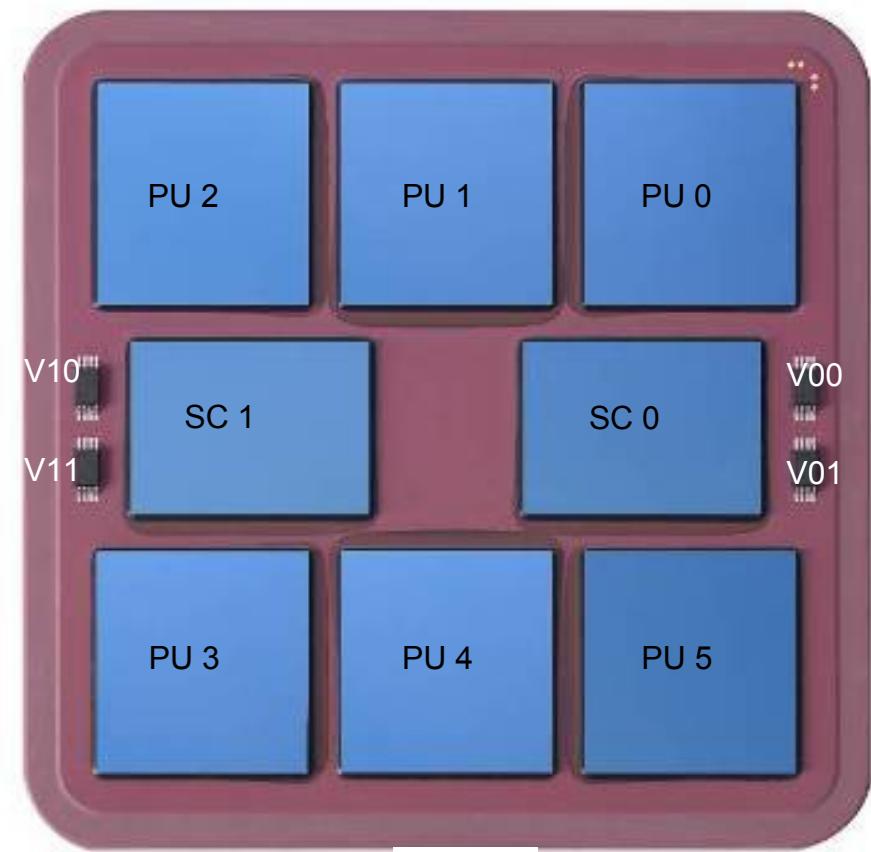
Dies wurde bisher auf Grund von Cache Kohärenzproblemen für unmöglich gehalten. Die eingesetzten Kohärenzalgorithmen wurden bisher auch noch nicht von IBM veröffentlicht.

# **System z Hardware Teil 3**

## **Multichip Module**



**z9**



**z12**

Vergleich der z9 und z12 Multichip Multilayer Ceramic Module . Die Abmessungen und die Anzahl der Verdrahtungsebenen der beiden Module sind praktisch identisch. Die Anzahl der Chips ist jedoch halbiert und die Chips sind größer geworden. Die CPU Chips (PU) haben 2 (z9) bzw. 6 (z12) Cores, und der gemeinsame Cache (SD/SC) wuchs von 40 (z9) auf 384 (z12) MByte.

IBM bringt etwa alle 2½ Jahre verbesserte Mainframe Modelle heraus. So erschien das Modell z9 im Juli 2005, Modell z10 im Februar 2008, Modell z196 im Juli 2010 und Modell zEC12 im Juli 2012.

Bei unserem Mainframe Rechner an der Uni Tübingen handelt es sich um ein Modell z9 .



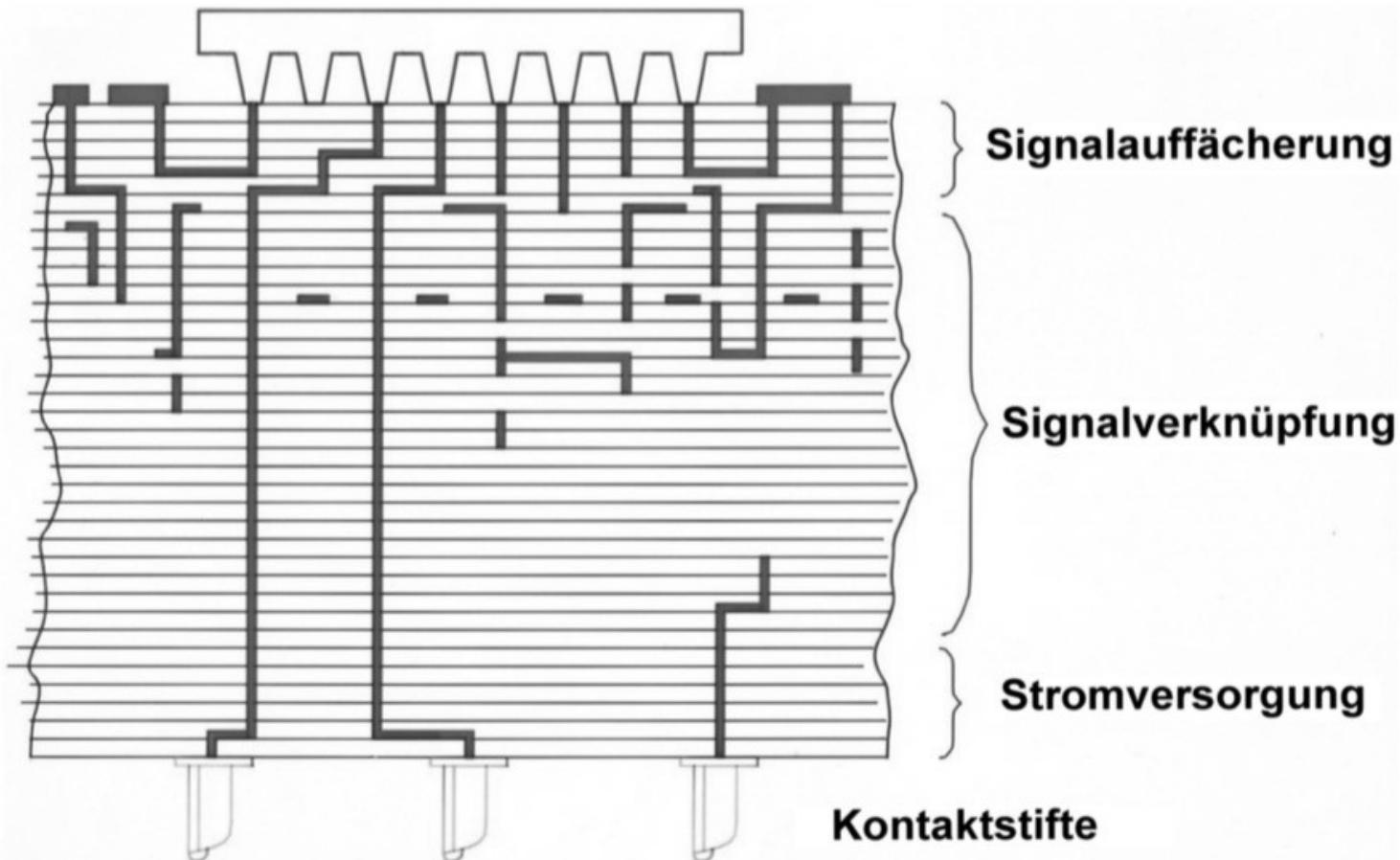
## **z9 Multi-Chip Module**

**Multichip Module (MCM)** benutzen die **Multilayer Ceramic (MLC)** Technologie. Ein MLC Module hat etwa 100 Verdrahtungslagen und ist wenige mm dick. Gezeigt ist das Einpassen eines z9 MCM in eine Halterung.

**Das MCM wird zusammen mit Hauptspeicher DIMMs und zusätzlichen Cards auf ein Printed Circuit Board aufgelötet.**

**Die MLC Module Technology hat sich seit 1980 evolutionär weiter entwickelt: Weniger und dafür größere Chips. Die Anzahl der Transistoren/Chip wuchs um etwa einen Faktor  $10^6$ .**

# Multi Layer Ceramic Technologie



Die circa 100 Verdrahtungsebenen des eines Multilayer Ceramic Technologie (MLC) Multichip Modules ermöglichen effektive Verbindungen der CPU und Cache Chips.

Das zEC12 - Multilayer Ceramic (MLC) Multichip Module benutzt einen Glas-Keramik-Träger mit 103 Glas-Keramik Verdrahtungslagen und 7356 Land Grid Array (LGA) Connections. In dem 96 x 96 mm-Modul sind Leiterbahnen mit einer gesamten Länge von mehr als 500 Meter untergebracht. Innerhalb der verschiedenen Schichten entstehen komplexe Verdrahtungsmuster. Die senkrechten Verbindungen zwischen den Schichten bestehen aus leitenden Bohrungen (VIAs), die wiederum innerhalb einer Schicht in horizontalen Leiterbahnen weitergeführt werden und an einer Bohrung zu einer darunter- oder darüberliegenden Schicht enden usw. Die früher verwendeten Kontaktstifte werden heute durch Kontaktpunkte (Lands) ersetzt.

Das heute verwendete Glas-Keramik-Material tritt an Stelle der früher verwendeten Aluminiumoxid (AL<sub>2</sub>O<sub>3</sub>)-Keramik. Es hat eine um 1/3 geringere Dielektrizitätskonstante und damit eine kürzere Signallaufzeit der Leitungsverbindungen. Eine sehr ähnliche Packungstechnologie hat sich in der Hochfrequenztechnik bewährt.

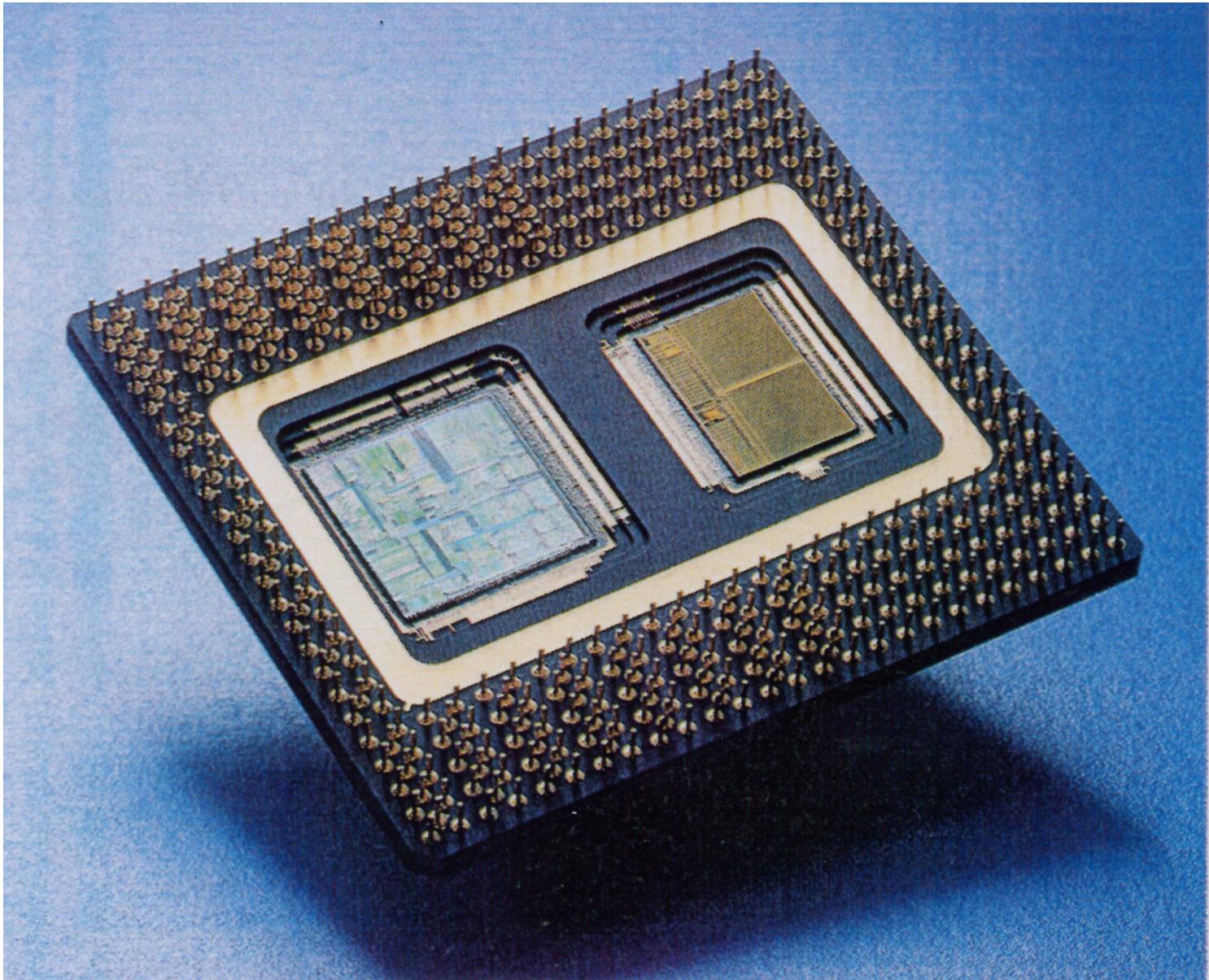
Die Glas-Keramik-Technologie steht im Gegensatz zu den normalerweise in Printed Circuit Boards (PCB) verwendeten organischen Materialien und Wirebond Peripheral Interconnect-Verfahren mit Lead Frame- oder Pin Grid Array (PGA)-Verbindungen. Diese haben schmalbandige induktive Diskontinuitäten (Drähte, Pins, Leads) und lange Netze mit erhöhter Laufzeitverzögerung.

## **Intel Pentium Pro**

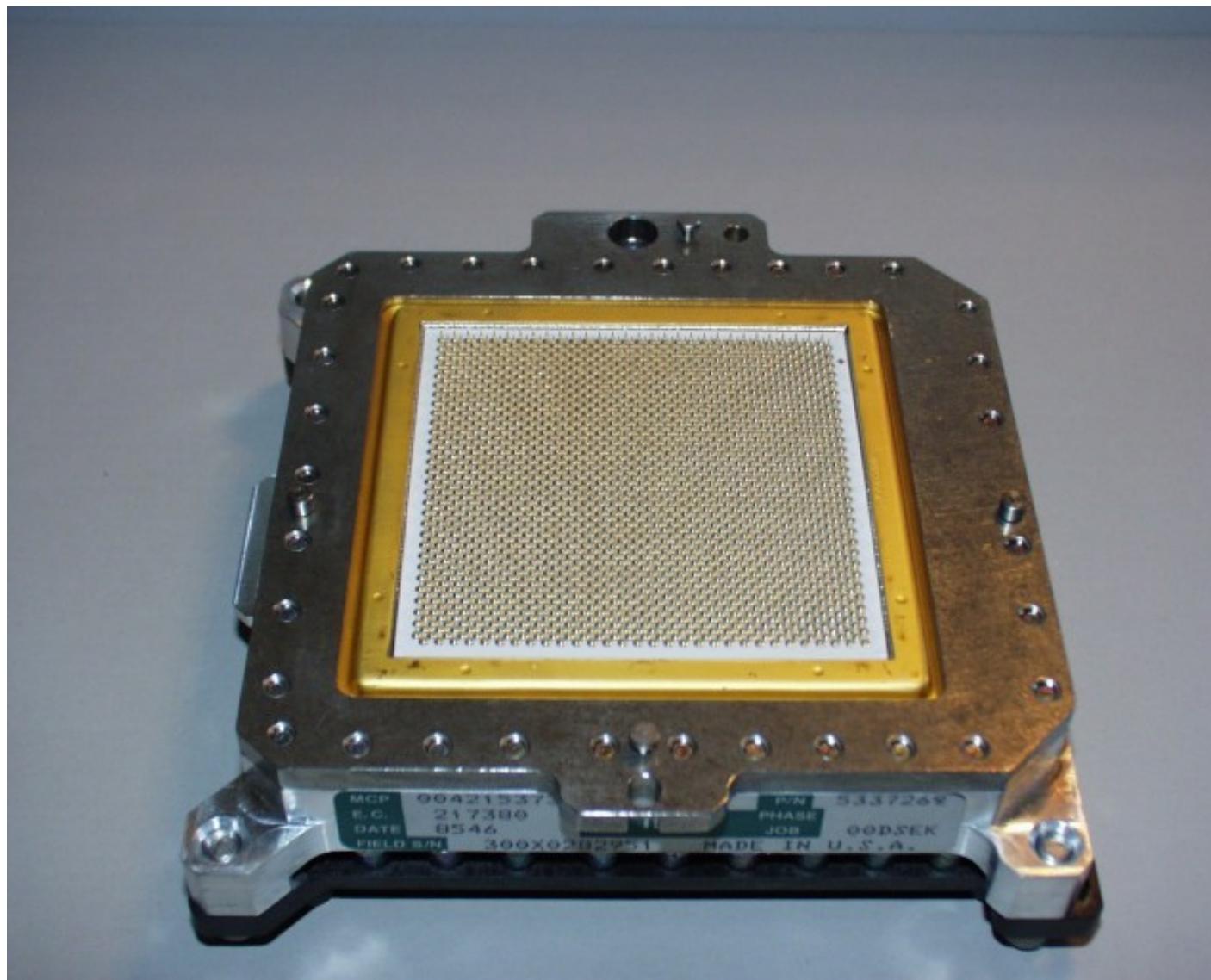
Bei der ursprünglichen Einführung des Pentium hatte die Firma Intel eine ähnliche MCM-Technologie für den Pentium Pro eingesetzt. Der 1995 von Intel herausgebrachte Pentium Pro Microprozessor verwendete ein 387 Pin Multi Layer Ceramic (MLC) Multi Chip Carrier (MCM) Module. Das Modul enthielt zwei Chips, ein CPU chip und ein getrenntes L2 Cache Chip, beide auf dem gleichen MLC Substrate. Die MCM-Technologie ermöglichte besonders günstige Signallaufzeiten zwischen den Chips.

Der Pentium Pro hatte deshalb eine überdurchschnittliche Leistung. Er wurde für Server und für High-End Desktop Processoren eingesetzt und war auch in Supercomputern wie ASCI Red benutzt worden.

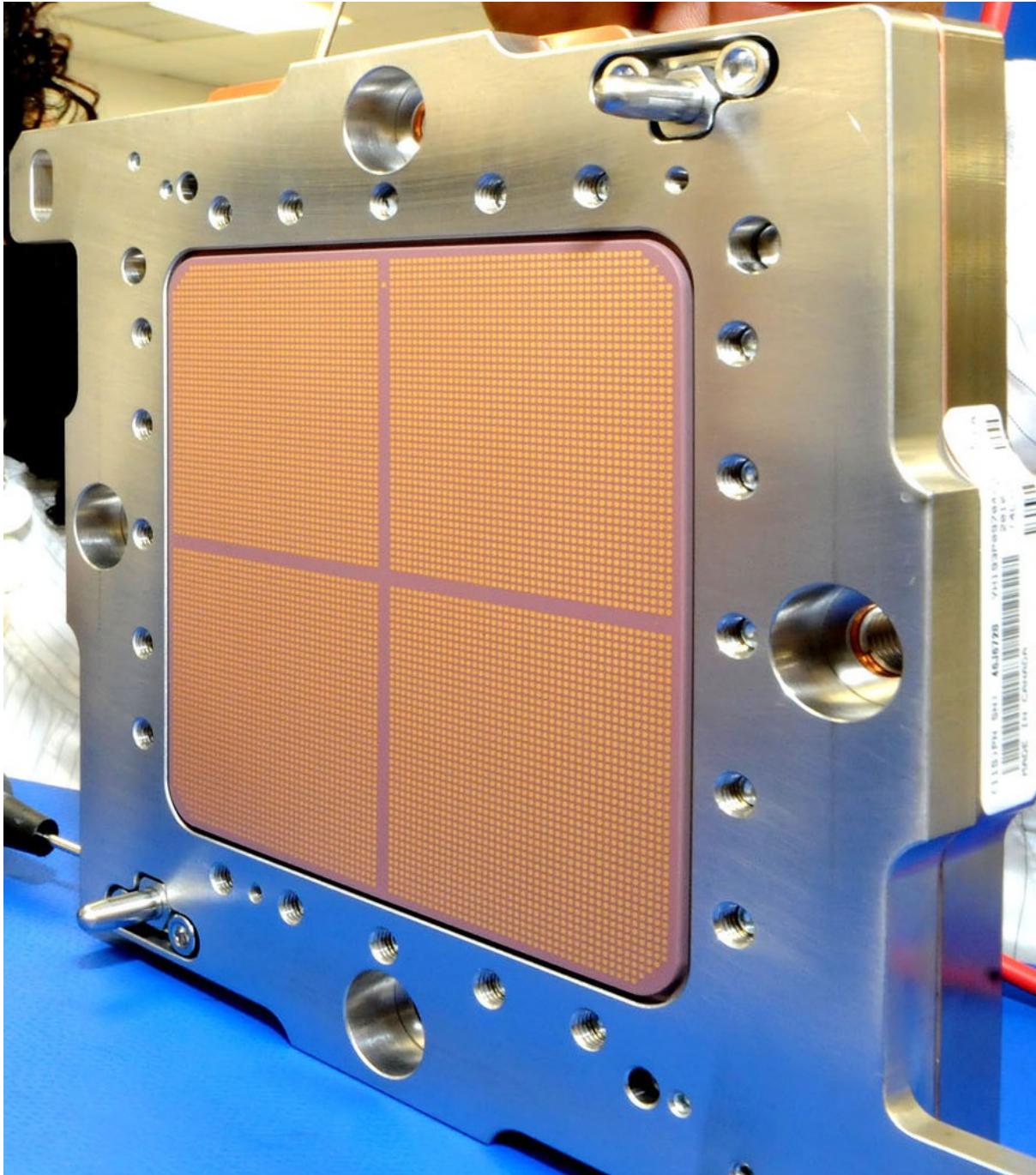
Der Pentium Pro wurde seinerzeit von den Server-Herstellern wegen seiner guten Leistungsdaten sehr geschätzt. Intel hatte aber Schwierigkeiten mit den Produktionskosten und hat deswegen die MCM-Technologie bis jetzt noch nicht wieder verwendet.



Der 1995 von Intel herausgebrachte Pentium Pro Microprozessor verwendete ein 387 Pin Multi Layer Ceramic (MLC) Multi Chip Carrier (MCM) Module und hatte eine überdurchschnittliche Leistung.



Die Rückseite des MLC Modules enthielt früher vergoldete Kontaktstifte, die die Verbindung mit einem Printed Circuit Board aufnehmen, welches gleichzeitig die Hauptspeicher DIMMs und die Host Channel Adapter Cards aufnimmt.



## System z Multichip Module

Im System zEC12 MCM sind die Kontaktstifte durch Kontaktpunkte (Lands) ersetzt.

Es sind 7356 Lands vorhanden.  
Das 96 x 96 mm große MLC  
Modul hat 103  
Verdrahtungslagen.

Es können 1800 Watt an Wärme  
abgeführt (gekühlt) werden.

## Land Grid Array

Ein **Land Grid Array (LGA)** ist ein Verbindungssystem für integrierte Schaltungen (IC, integrated circuit).

Beim LGA-System werden die Anschlüsse des integrierten Schaltkreises auf seiner Unterseite in Form eines schachbrettartigen Feldes (grid array) von Kontaktflächen (lands) ausgeführt. Es ist eng verwandt mit dem **PGA**-System (Pin Grid Array), welches statt der Kontaktflächen die bekannten „Beinchen“ (pins) besitzt, und dem **BGA**-System (Ball Grid Array), welches Lötperlen benutzt.

LGA-Prozessoren werden meistens auf Sockel gesetzt, die federnde Kontakte enthalten, was eine geringere mechanische Beanspruchung der Kontakte zur Folge hat. Andere LGA-ICs werden aber oft auch wie PGA-ICs direkt verlötet. BGA-ICs sind hingegen ausschließlich zum Verlöten gedacht, sie bringen das nötige Lötzinn in Form der Lötperlen gleich mit. Alle drei Varianten sind hauptsächlich für ICs mit Hunderten bis über Tausend Anschlüssen gedacht.

Das Land Grid Array ist im Gegensatz zum Pin Grid Array für höhere Frequenzen geeignet und günstiger zu produzieren.

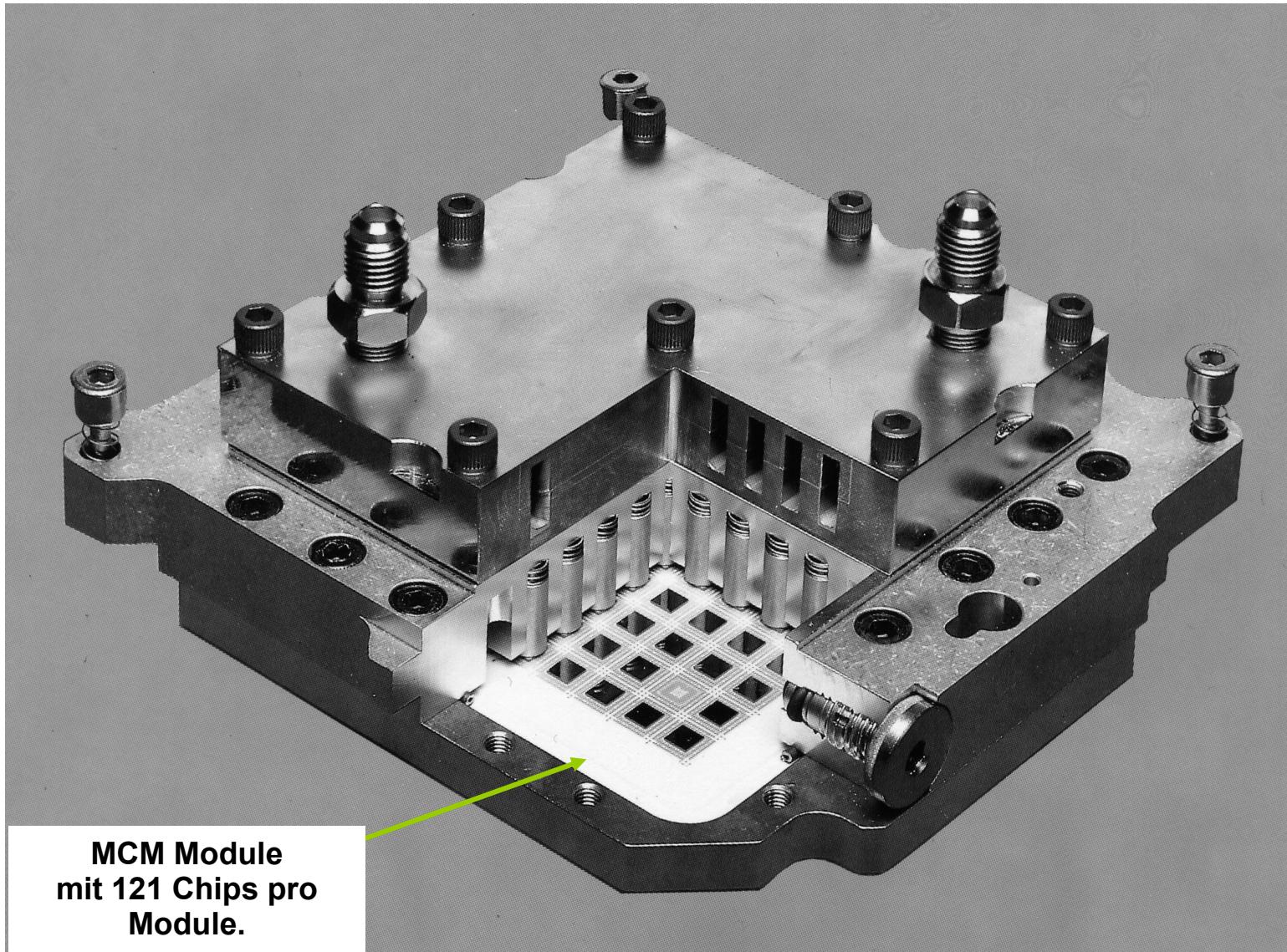
## Thermal Conduction Module

Ein **Thermal Conduction Module (TCM)** ist eine Baugruppe, welche ein Multilayer Ceramic (MLC) Multichip Module (MCM) aufnimmt, und die für die Energieabfuhr (Kühlung) erforderliche Hardware enthält.

Die TCM und MLC Technologie wurde von IBM in den 80er Jahren zur Produktionsreife gebracht und seitdem kontinuierlich weiterentwickelt. An der Grundkonzeption hat sich allerdings erstaunlich wenig geändert.

Die meisten Mainframe Modelle benutzen seitdem diese Technologie. Sie zeichnet sich durch besonders hohe Zuverlässigkeit aus. Die elektrischen Eigenschaften bewirken eine besonders kurze Signallaufzeit zwischen den Chips eines MCM. Ein z196 TCM hat eine Kühlleistung von 1,8 KWatt, etwa soviel Energie wie ein Bügeleisen abstrahlt.

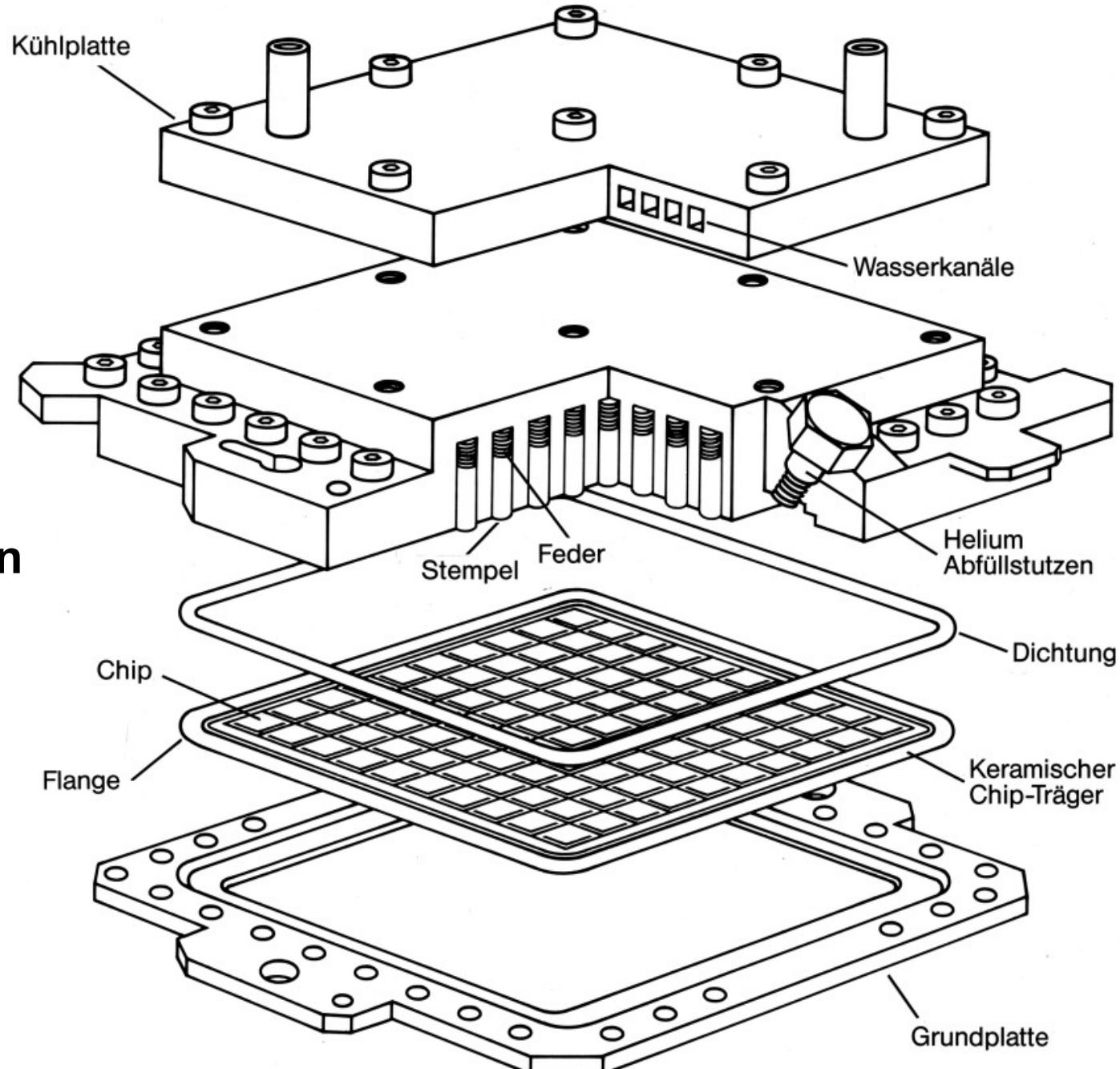
# Thermal Conduction Module (TCM)



**MCM Module  
mit 121 Chips pro  
Module.**

Dargestellt ist ein 1987 gefertigtes Thermal Conduction Module (TCM), mit 121 Chips auf dem MCM, und 704 circuits/chip. Mehrere dieser Module bildeten die CPU einer S/370 Modell 3081 Mainframe CPU.

# Thermal Conduction Module



Eine MCM-Baugruppe enthält neben dem keramischen Chip-Träger die Mechanik für die Energieableitung mittels Stempel und Kühlplatte. Diese Baugruppe wird als Thermal Conduction Module (TCM) bezeichnet und ist in der obigen Abbildung dargestellt.

Zu Kühlungszwecken sitzt auf jedem Chip ein Aluminium-Stempel, der die Verlustwärme ableitet . Eine Spiralfeder drückt den Stempel an die Chip Oberfläche an. Die das Chip berührende Oberfläche des Stempels ist konisch ausgebildet mit einem Konus Radius im Bereich vieler 100 Meter. Dies sorgt für einen guten Kontakt des Stempels mit der Chip-Oberfläche und für einen minimalem Luftspalt, auch wenn der Stempel geringfügig verkantet (siehe folgende Abbildung).

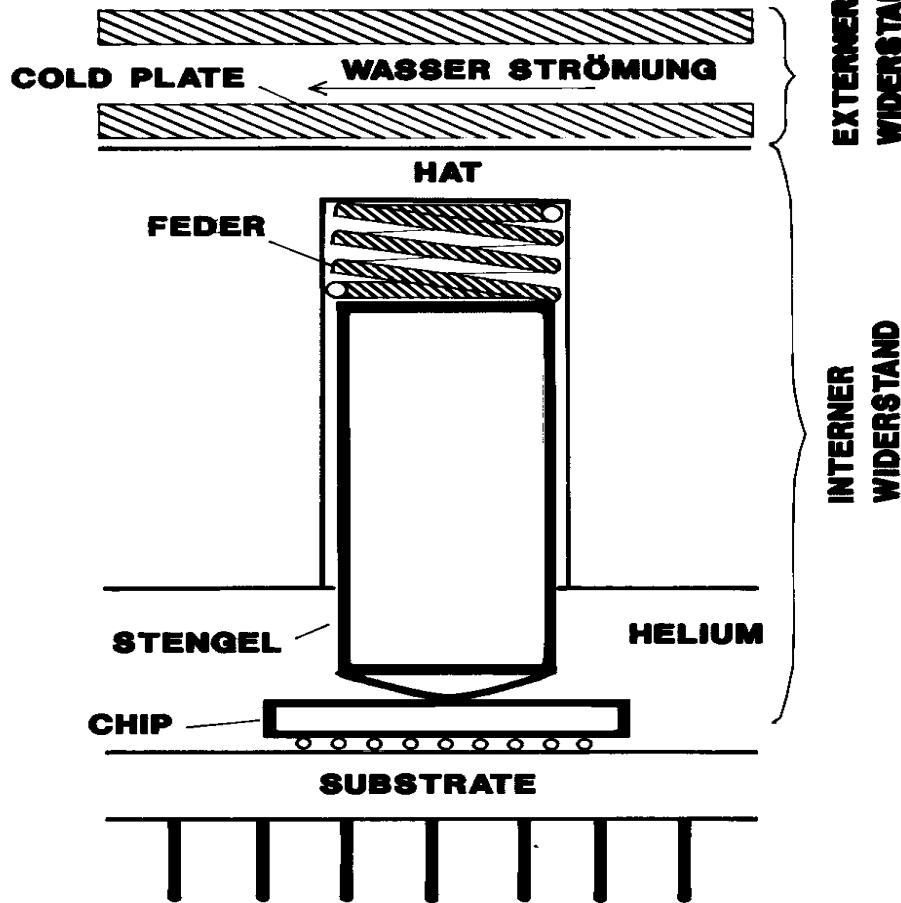
Die TCMs sind vielfach mit Helium an Stelle von Luft oder Stickstoff gefüllt. Die Wärmeleitfähigkeit von Helium ist größer als die jeder anderen bekannten Substanz.

Die Aluminium-Stempel werden in einer Bohrung geführt und geben die Verlustwärme an die Umgebungsplatte weiter. Eine darüber liegende Kühlplatte wird entweder mit Luft oder mit Wasser gekühlt. Im letzteren Fall existiert ein geschlossener Kreislauf, in dem das Wasser seine Wärmeenergie an einen Radiator innerhalb des System z Frames weitergibt, ähnlich wie in einem Automobil. Es existieren auch System z Modelle, wo die Wasserkühlung extern erfolgt.

Bei dem hier vorgestellten Verfahren werden alle Chips selbst mit Luft (Helium) gekühlt. Es sind in der Vergangenheit viele Versuche unternommen worden, Chips direkt mit einer Flüssigkeit zu kühlen. Diese Verfahren haben sich in der Praxis nicht bewährt.

# TCM Wärmeübergang

## WÄRMEABLEITUNG IM TCM



w034

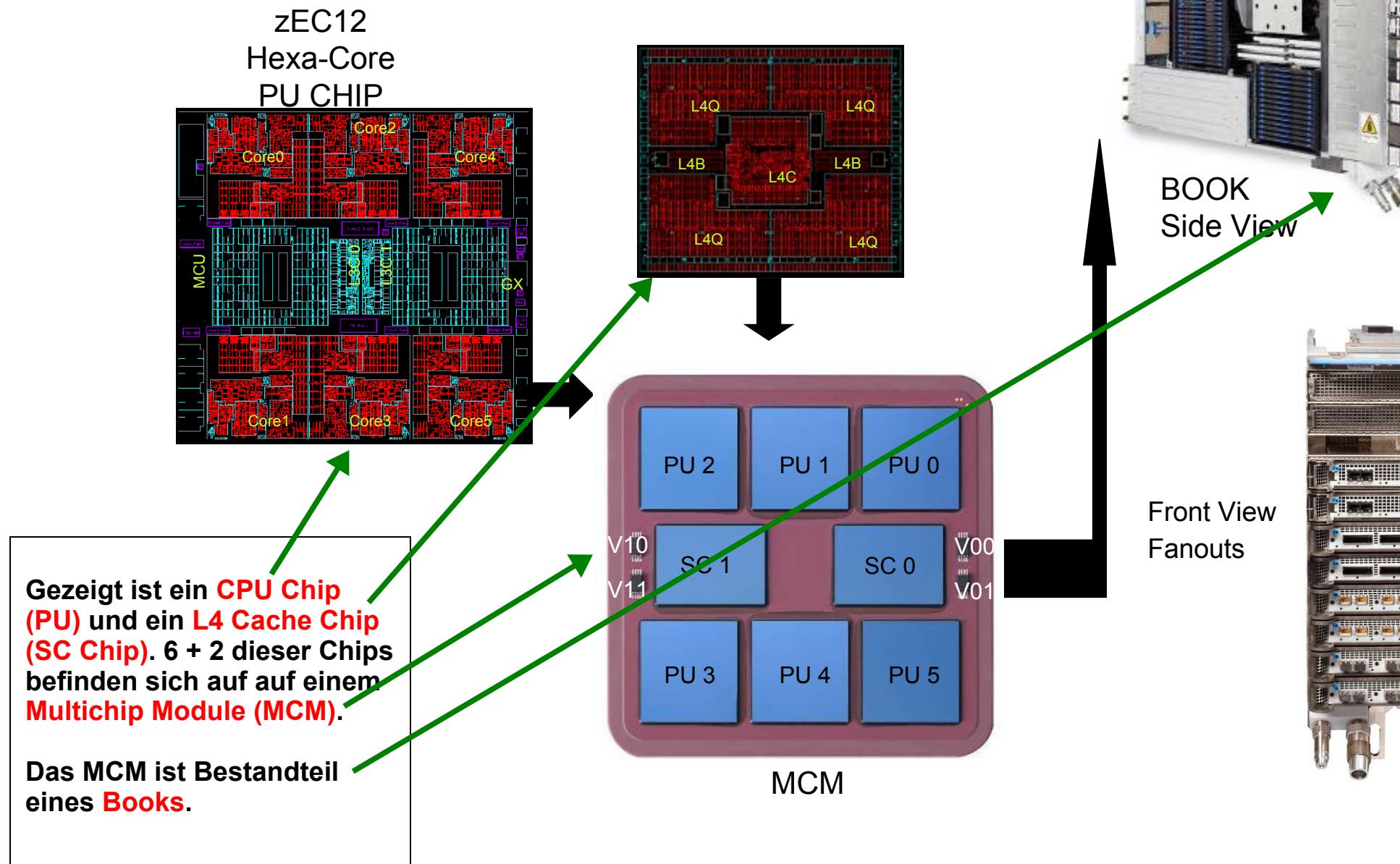
Zu Kühlungszwecken sitzt auf jedem Chip ein Aluminium-Stempel, der die Verlustwärme ableitet. Eine Spiralfeder drückt den Stempel an die Chip Oberfläche an. Die das Chip berührende Oberfläche des Stempels ist konisch ausgebildet mit einem Konus Radius im Bereich vieler 100 Meter. Dies sorgt für einen guten Kontakt des Stempels mit der Chip-Oberfläche und für einen minimalem Luftspalt, auch wenn der Stempel geringfügig verkantet.

Die TCMs sind vielfach mit Helium an Stelle von Luft oder Stickstoff gefüllt. Die Wärmeleitfähigkeit von Helium ist größer als die jeder anderen bekannten Substanz.

# **System z Hardware Teil 4**

## **Book und System Frame**

# CPU Chip – MCM – Book



# Book

Das MCM ist Bestandteil einer als „Book“ bezeichneten Baugruppe, welche neben dem MCM noch Steckplätze für 30 Hauptspeicher DIMMs (Dual Inline Memory Module) sowie „Fan Out Adapter“ Karten enthält. Fan Out Cards werden auch als Host Connector Adapter (HCA) Cards bezeichnet. Sie erfüllen die gleiche Funktion wie die I/O Adapter Cards der Sun oder Hewlett Packard System Boards.

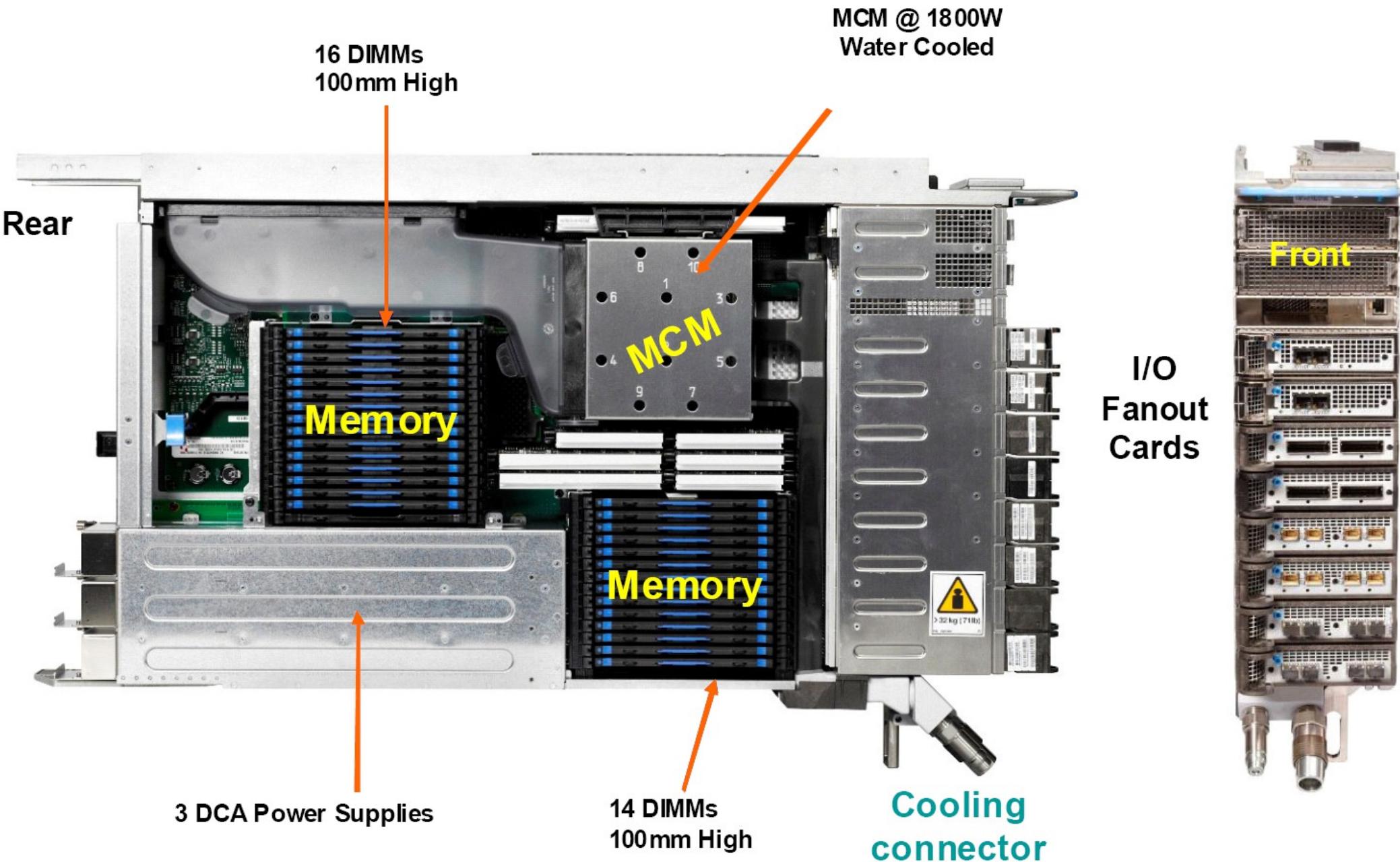
Hauptspeicher DIMMs haben eine Kapacität von je 4 GByte, 16 GByte oder 32 GByte. Die maximale physische Hauptspeicherkapazität beträgt somit 960 GByte, von denen nach Abzug für Fehlerkorrektur und Redundanz 768 GByte verfügbar sind.

Die Fanout Cards führen zu Steckkontakte auf der Vorderseite (Front View) des Books. Es existieren 8 Steckkontakte , die mit jeweils 2 Kabeln eine Verbindung zu einem „I/O Cage“ aufnehmen, in denen Steckkarten für I/O Anschlüsse untergebracht sind. Zwei weitere Steckkontakte (FSP) werden zur Verbindung zu zwei „Support Elementen“ verwendet, die später diskutiert werden.

Es existiert eine zweistufige Stromversorgung. Die primäre Stromversorgung (Bulk Supply) versorgt einen ganzen Rechner, und ist aus Zuverlässigkeitsgründen doppelt vorhanden. In jedem Book befindet sich eine zusätzliche sekundäre Stromversorgung (Distributed Converter Assembly, DCA) , welche kurzfristige Versorgungsschwankungen innerhalb eines Books ausgleicht.

Die folgende Abbildung zeigt die Seitenansicht eines (geöffneten) Books, in der das Multichip Module (MCM), die Hauptspeicher DIMMs, die (sekundäre) Stromversorgung und (hinter der Abdeckung verborgen) die Fanout Cards zu sehen sind.

Voltage Transformation Modules (VTM) bewirken Power Conversion in dem Book und benutzen Triple Redundancy. Die Redundanz schützt die Processoren vor einem Spannungsverlust als Folge des Versagens einer VTM Card. Triple redundancy wird auch für die Humidity (Luftfeuchtigkeit) Sensoren zur Verbesserung der Zuverlässigkeit eingesetzt.

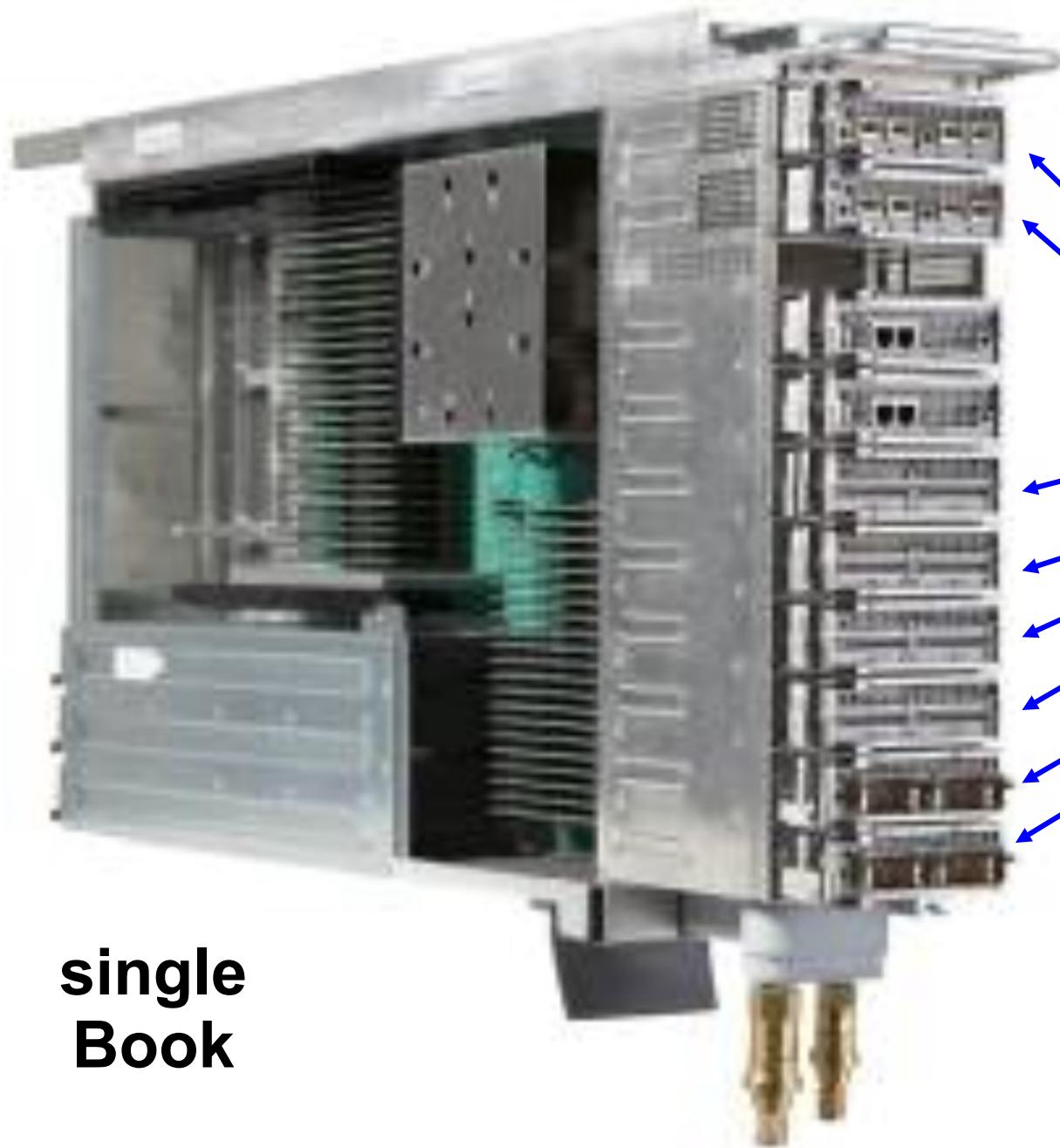


Note: Unlike the z196, zEC12 Books are the same for the Radiator based Air and Water cooled Systems

## **zEC12 Book**

Jedes zEC12 Book enthält die folgenden Komponenten:

- Ein Multi-Chip-Modul (MCM) mit sechs Hex-Core-Mikroprozessor-Chips, und zwei Storage Control Chips mit 384 MByte Level 4 Cache.
- Speicher DIMMs in 30 verfügbaren Slots. Dies ermöglicht von 60 GByte bis 960 GByte physischen Speicher pro Book.
- Eine Kombination von bis zu acht Host Channel Adapter (HCA) oder PCIe Fanout-Karten.
- PCIe Fanouts werden für 8Gbit/s Links zu den PCIe I/O-Karten verwendet. Die HCA-Optical Fanouts verbinden zu externen Coupling Links.
- Drei verteilte Wandler-Baugruppen (DCAs), die das Book mit Strom versorgen. Bei Verlust einer DCA ist genügend Power vorhanden ( $n + 1$  Redundanz), um den Strombedarf des Books zu befriedigen. Die DCAs können während des laufenden Betriebes gewartet und ausgetauscht werden.
- Zwei „Flexible Service-Prozessor“ (FSP)-Karten für die Systemsteuerung.



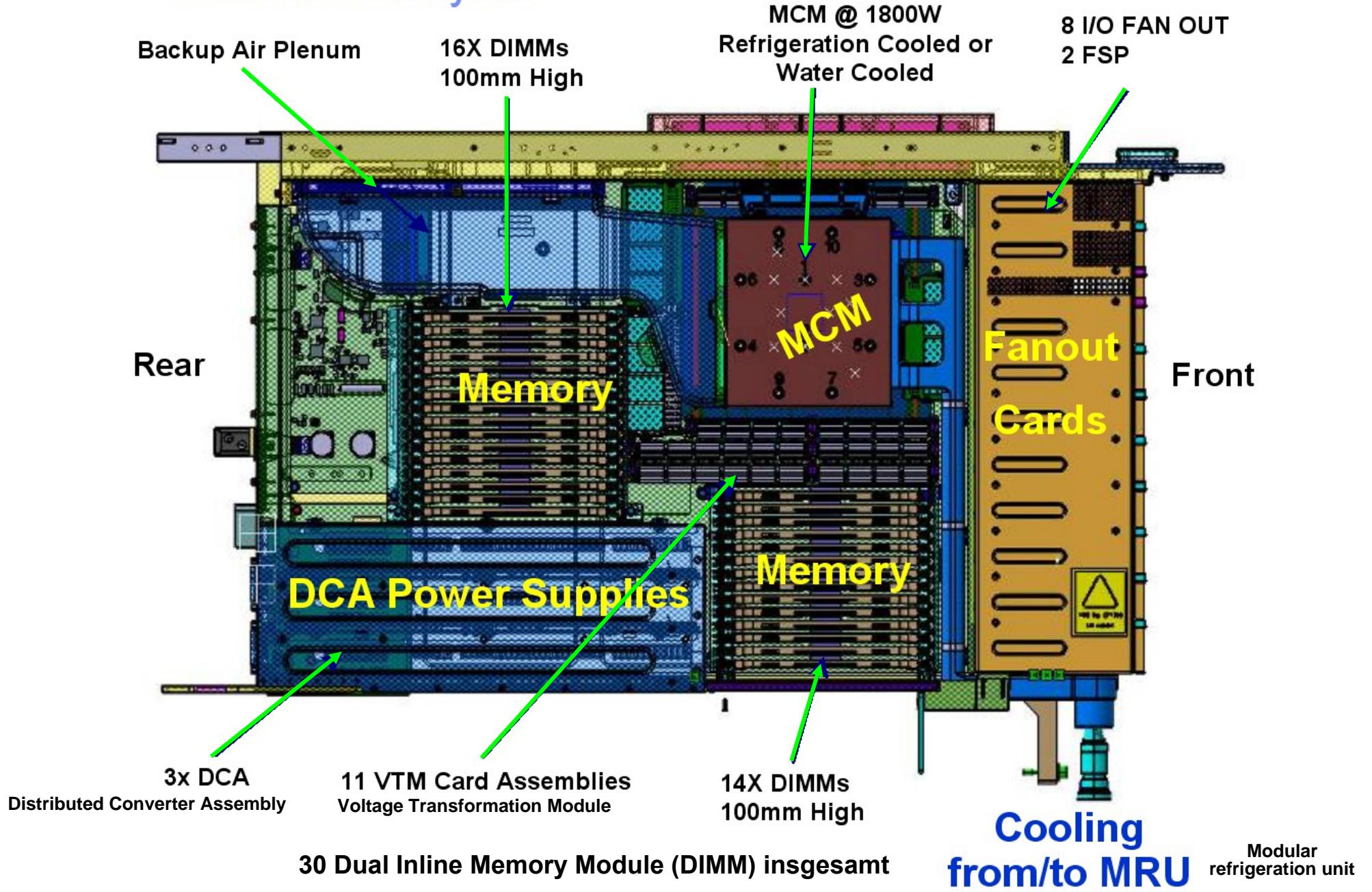
**single  
Book**

**8 Fanout  
Card Anschlüsse**

**2 Anschlüsse pro  
Fanout Card**

Fanout Cards werden auch als „Host Channel Adapter“ (HCA) Cards bezeichnet. Pro Fanout Card sind 2 nebeneinander-liegende Steckkontakte für Kabel-verbindungen mit einem I/O Cage vorhanden.

# z196 Book Layout



## Verbindung MCM - Hauptspeicher

Die folgende Abbildung zeigt den Hauptspeicheranschluss.

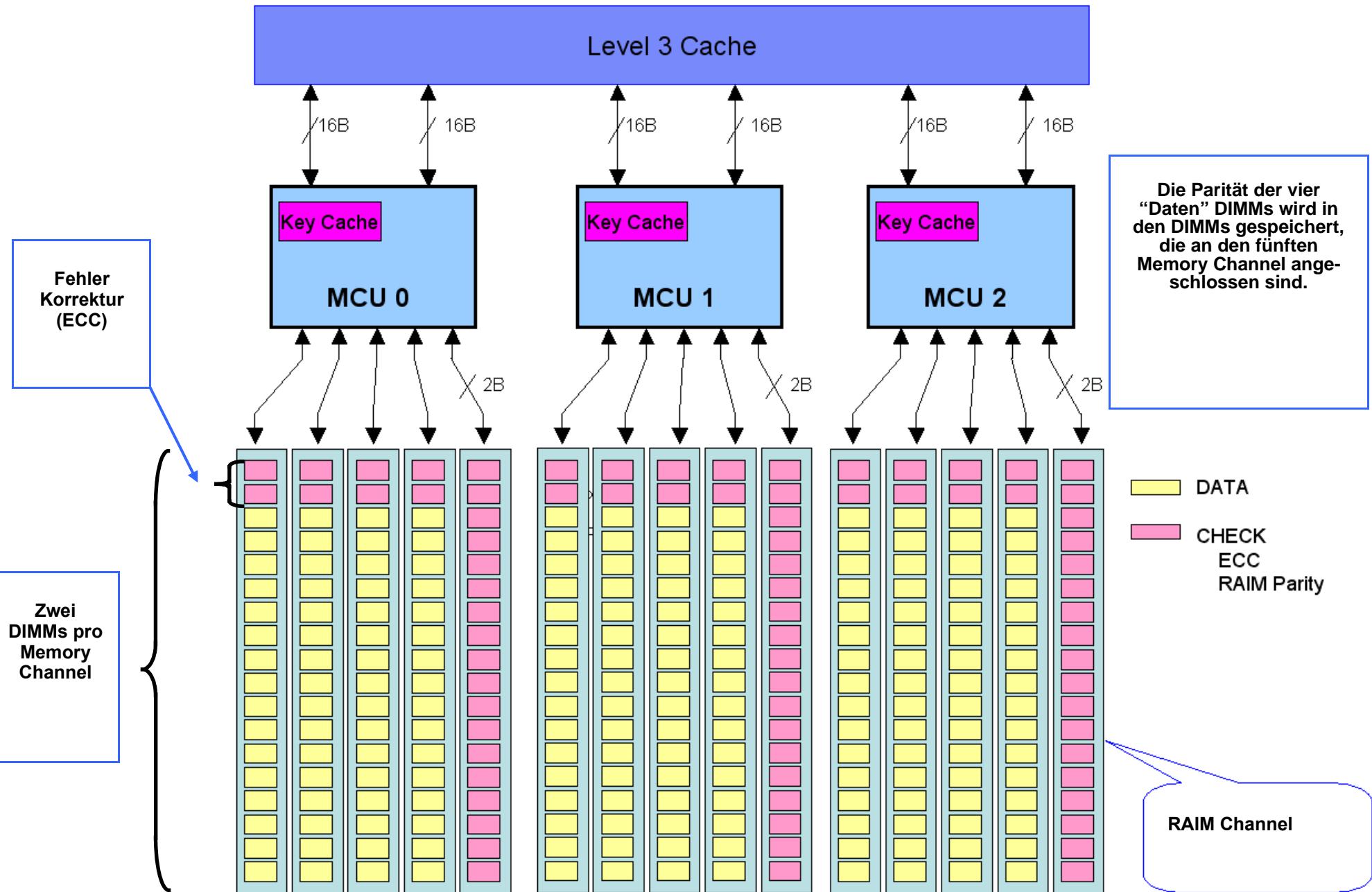
Dieser erfolgt mit Hilfe von drei „Memory Control Units“ (MCU), welche über zwei je 16 Byte (2 x 128 Bit) breite Busse mit den L3 Caches der einzelnen CPU Chips verbunden sind. Die Memory Control Units enthalten zusätzlich einen Cache, der die am häufigsten gebrauchten Storage Protection Keys (siehe Abschnitt „Storage Protection“ in Einführung, Teil 3 dieser Vorlesung) enthält.

Der Hauptspeicher selbst enthält 4 „Channels“ (Columns) von Hauptspeicher Quad High DIMMs (Dual Inline Memory Module), vom Aussehen her ähnlich wie die DIMMs in Ihrem PC. Jeder Channel verwendet eine spezielle Version des ReedSolomon Fehlerkorrektur Codes an Stelle des sonst für Hauptspeicher üblichen Hamming Fehlerkorrektur Codes.

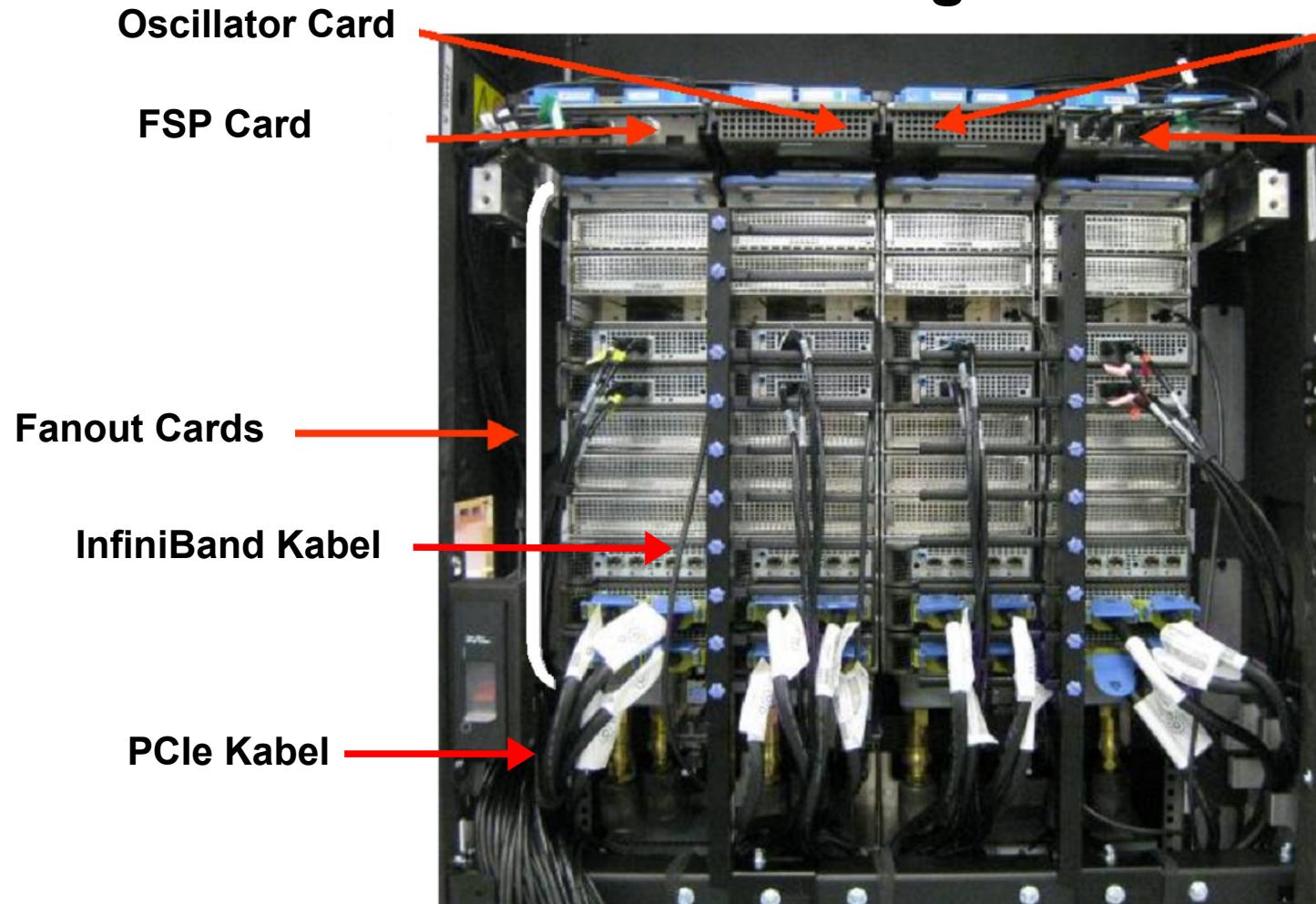
Als weitere Fehlerkorrekturmaßnahme existiert ein 5. RAIM (Redundant Array of Independent Memory) Channel. RAIM benutzt das gleiche Verfahren wie RAID für Plattspeicher. Wenn eine der 4 Columns ausfällt, kann der Hauptspeicherinhalt mit Hilfe der 5. Column wieder hergestellt werden. RAIM wurde erstmalig mit der z196 für kommerziell erhältliche Rechner eingeführt.

An jedem Channel hängen 2 Dual In-Line Memory Modules (DIMMs); jede einzelne MCU bedient 10 DIMMs. Jedes DIMM Module speichert 32 GByte. Jedes Book verfügt über DIMMs, für eine maximale Speicherkapazität von 960 GByte pro Book, oder 3 480 GByte für ein 4 Book System.

Da die RAIM DIMMs 20 % der Speicherkapazität in Anspruch nehmen, stehen dem Benutzer lediglich 768 GByte pro Book, oder 3 072 GByte pro System zur Verfügung. Davon werden 32 GByte für die „Hardware System Area“ (HSA) benötigt. Die HSA speichert Firmware und wird später erläutert.



# CPU Cage



Oscillator Card  
STP: Server Time Protocol  
FSP Card  
FSP: Flexible Service Processor

PCIe I/O Infrastruktur mit einer Bandbreite von 8 Gbit/s

InfiniBand I/O Infrastructure mit einer Bandbreite von 6 GByte/s

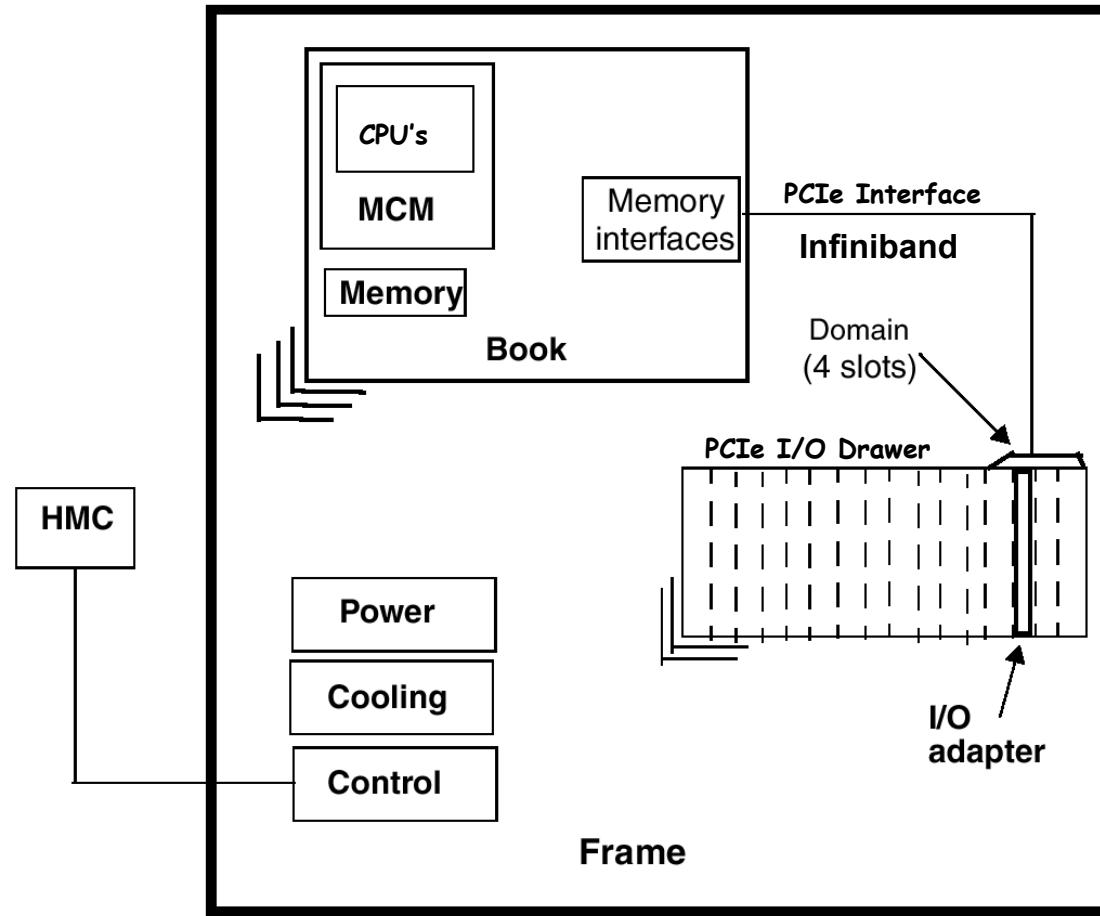
Die obige Abbildung zeigt die Vorderseite(n) von 4 nebeneinander stehenden Books. Zu sehen sind die Steckkontakte, über die mit Infiniband oder PCIe Kabeln die Verbindung zu einer (max 3) I/O Drawer hergestellt werden, welcher I/O Adapter Karten (z.B Plattspeicher Anschlüsse) aufnimmt, die eine Verbindung zur Außenwelt übernehmen.

Zwei weitere Steckkontakte (FSP) werden zur Verbindung zu zwei „Service Elementen“ verwendet, die später diskutiert werden. Zwei Oscillator Karten stellen Clock Signale zur Verfügung und stellen eine Synchronisation aller CPUs mittels des Server Time Protocols (STP) sicher.



Hier entfernt ein Entwicklungsingenieur ein Book aus einem z9 Rechner.

# Struktur eines zEC12 Mainframe Systems.



Die Abbildung zeigt die interne Struktur eines z9, z10 z196 Mainframe oder zEC12 Systems. CPUs und Caches befinden sich auf einem Multi Chip Module (MCM). Ein einziges MCM, zusammen mit dem Hauptspeicher (main store) bilden eine als "Book" bezeichnete Baugruppe. Bis zu 4 Books sind möglich.

Über PCIe oder Infiniband I/O Drawer können I/O Adapter angeschlossen werden, die sich in separaten Gehäusen (I/O Drawer) befinden.

# Hardware Management Console (HMC)

Ein PC verfügt über ein BIOS. Letzteres implementiert Maschinencode, der in einem Teil des Hauptspeichers liegt, auf den ein Benutzerprozess nicht zugreifen kann.

Eine äquivalente Mainframe-Funktion heißt Firmware, häufig auch als LIC (Licensed Internal Code) bezeichnet. Der Firmware-Bereich ist außerhalb des Adressbereichs der Maschinenbefehle im Hauptspeicher untergebracht. Er wird beim Hochfahren eines Rechners durch einen als IML (Initial Microcode Load) bezeichneten Vorgang in den Speicher geladen.

Firmware hat zahlreiche Funktionen. Er implementiert z.B. manche komplexe Maschineninstruktionen oder umfangreiche Diagnostik- und Fehlerbehandlungs-Funktionen. Die später beschriebene PR/SM-Einrichtung wird ebenfalls mittels Firmware implementiert.

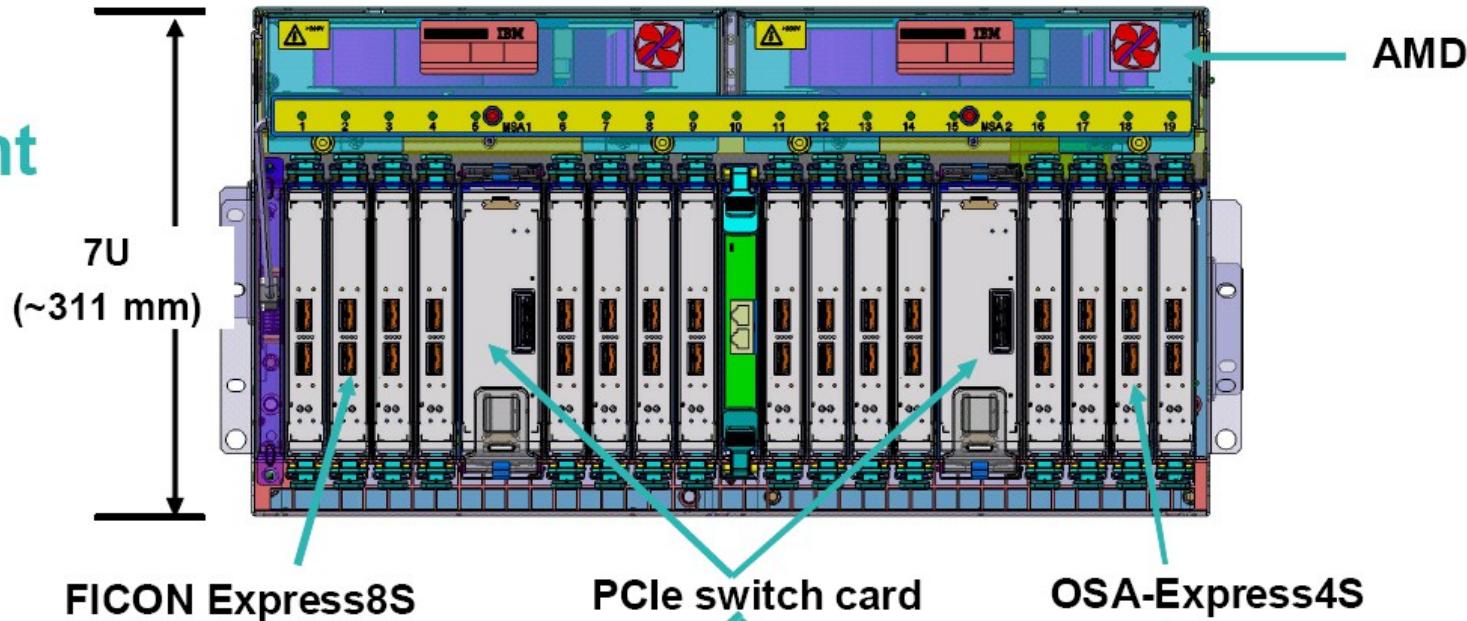
Wenn man beim Hochfahren eines PC eine Funktionstaste drückt, kann man BIOS-Funktionen aufrufen. Beim Mainframe ist diese Funktionalität sehr viel umfangreicher, während des laufenden Betriebs verfügbar und wird als „Operator Facilities“ bezeichnet. Ein System-Administrator kommuniziert mit Hilfe der Operator Facilities mit dem Rechner. Er erledigt damit Aufgaben wie das Setzen von Datum und Zeit, das Reset von Subsystemen, Architectural Mode Selection, das Eingreifen in einen ausführenden Programmablauf oder ein Reagieren auf Maschinenfehler-Unterbrechungen. Ebenfalls dazu gehört die Funktion eines Boot-Managers, beim Mainframe als Initial Program Load (IPL) bezeichnet. Mainframe-Rechner verfügen über ein als „Hardware Management Console“ (HMC) bezeichnetes Bildschirm-Gerät, das ausschließlich der Nutzung der Operator Facilities durch den System-Administrator dient.

# PCIe I/O drawer

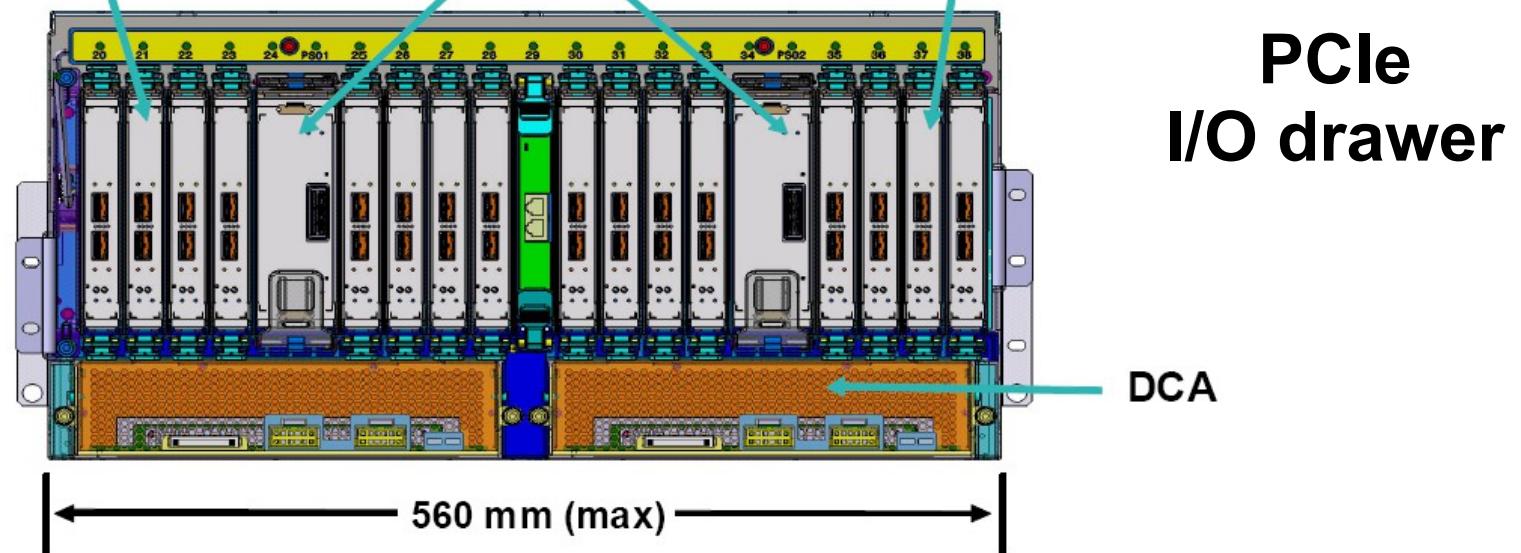
Die Hauptspeicherschnittstelle (memory interface) verbindet ein Book mit „I/O Adaptern“ (I/O Cards) die in einem „I/O Drawer“ (auch als I/O Cage bezeichnet) untergebracht sind. Die Verbindung zwischen Book und I/O Drawer wird durch I/O Kabel hergestellt, welche das PCIe Protokoll implementieren, äquivalent zu dem PCIe Protokoll in einem PC. Der PCIe Drawer ist ein zweiseitige Drawer (I/O-Adapter auf beiden Seiten). Ein Drawer enthält 32 I/O-Steckplätze, und kann bis zu 32 I/O Adapter Cards aufnehmen, für Verbindungen zu Plattspeichern, Magnetbändern und anderen I/O Geräten. Letztere sind grundsätzlich in getrennten Gehäusen untergebracht.

Der PCIe I/O Drawer nutzt PCIe als Infrastruktur. Die PCIe I/O-Bus-Infrastruktur Datenrate beträgt 8 Gbit/s. Bis zu 128 Kanäle (64 PCIe I/O-Features) sind in einem I/O Drawer möglich. Der Drawer enthält 4 Switch Cards (zwei vorne, 2 hinten), und zwei DCAs für die redundante Stromversorgung.

**Front**



**Rear**





**Z Frame**

**A Frame**

Ein z9, z10, z196 oder zEC12 Rechner besteht aus 2 meistens nebeneinander aufgestellten Rahmen, etwa 1,8 Meter hoch, welche von IBM als „Z Frame“ und „A Frame“ bezeichnet werden. Die Türen zu den beiden Frames sind künstlerisch geformt und enthalten viel leere Luft.

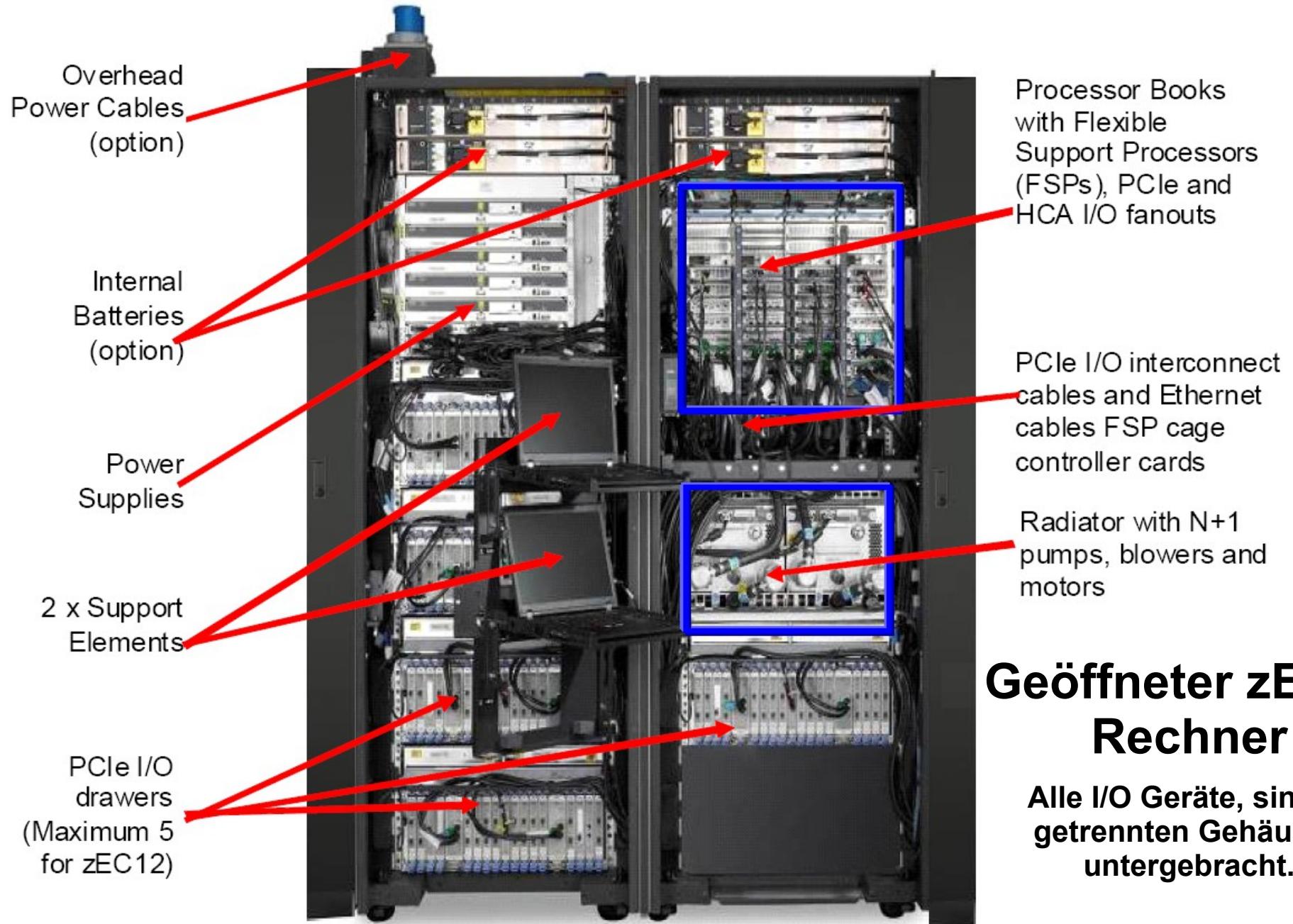
Ein zEC 12 Rechner enthält bis zu 4 Books mit je 6 x 6 CPU Chips und  $4 \times 6 \times 6 = 144$  Prozessoren. Von diesen können 101 als CPUs eingesetzt werden, 16 arbeiten als „**System Assist Prozessoren**“ (SAP) und die restlichen dienen als Reserve (Spares), die aktiviert werden können, wenn ein anderer Processor ausfällt.

Es sind 786 GByte Hauptspeicher pro Book möglich, insgesamt also 3 TByte für ein System mit 4 Books.

Es kann ein Rechner mit 1, 2, 3 oder 4 Books ausgeliefert werden. Unser z9 Rechner hat nur 1 Book.



**Beim Entwurf der Türen für die zEC12 Rahmen durften sich die künstlerisch motivierten Designer austoben.**



## Geöffneter zEC12 Rechner

Alle I/O Geräte, sind in getrennten Gehäusen untergebracht.

## **Geöffneter zEC12 Rechner**

Gezeigt ist ein geöffneter zEC12 Rechner (ohne Türen). Rechts oben sind 4 Books zu sehen mit Anschluss Steckern auf der Vorderseite. Hinter den Steckern sitzen Host Connector Adapter (HCA) Cards, welche die Verbindung zu dem MCM herstellen. Diese nehmen entweder PCIe Bus Kabel für die Verbindung zu den I/O Drawers oder Infiniband Kabel für die Verbindung zu anderen Rechnern auf.

Zu beachten ist: Alle I/O Geräte, besonders auch Plattenspeicher, sind in getrennten Gehäusen untergebracht.

Ein Support Element (SE) ist ein herausklappbarer Laptop Computer, der innerhalb des linken Frames herausklappbar und fest verschraubt untergebracht ist. Er wird u.A. für die Initialisierung des z/OS Systems benutzt.

# **System z Hardware Teil 5**

## **Mainframe Emulation**

# Mainframe Emulation

Das IBM System z Personal Development Tool (zPDT) besteht aus einem Dongle (USB Stick). Dieser kann mit einem regulären x86 (Intel, AMD) Rechner betrieben werden. Hiermit ist es möglich, z/OS auf einem x86-Rechner laufen zu lassen, allerdings mit stark verringelter Leistung.

<http://www.redbooks.ibm.com/redbooks/pdfs/sg247721.pdf>.

Ähnliches leistet der Hercules Emulator, der aber von IBM (im Gegensatz zu zPDT) nicht unterstützt wird.

<http://www.hercules-390.eu/>

Die Fujitsu Siemens SX Serie Systeme emulieren auf SPARC oder Intel Prozessoren das hauseigene BS2000 Betriebssystem.

# **System z Hardware Teil 6**

## **Weiterführende Information**

## Hauptspeicher DIMMs



**512 MByte DIMM**

Die Speicherkapazität eines Speichermoduls ergibt sich normalerweise als Produkt aus der Speicherkapazität der meist gleichartigen Speicherchips und deren Anzahl.

Als Beispiel sei hier ein Speichermodul genannt, das mit 16 Chips des Typs GM72V16821CT10K bestückt ist. Aus dem Datenblatt erfährt man, dass dieser Chip in zwei Bänken mit je 524.288 (=  $2^{19}$ ) Wörtern mit einer Wortbreite von jeweils 16 Bit organisiert ist ( $2 \times 2^{19} \times 16$ ). Daraus ergibt sich eine Speicherkapazität pro Chip von  $2 \times 2^{19} \times 16$  bit =  $2^{24}$  bit = 16.777.216 bit. Mit 16 dieser Chips ergibt sich eine Speicherkapazität des Speichermoduls von  $2^{28}$  bit = 268.435.456 bit oder – mit 8 Bits pro Byte –  $2^{25}$  Byte = 33.554.432 Byte = 32 MiB.

Manche Speichermodule besitzen ein oder zwei zusätzliche Chips (gleichen oder anderen Typs), die für Fehlerkorrektur- bzw. Paritätsfunktionen zuständig sind. Hier werden für ein Byte häufig 9 Bits verwendet (8 Datenbits und 1 Prüfbit).

Der heute übliche DDR/DDR2-Speicher besitzt 64 Daten-Signalleitungen (beziehungsweise 72 bei ECC). Die einzelnen SDRAM-Chips sind so verschaltet, dass sie die gesamte Breite des Datenbusses belegen. Jeder Chip ist für bestimmte Datenleitungen zuständig. Ein Chip mit einer „ $\times n$ “-Organisation kann  $n$  Datenleitungen versorgen. Für einen Datenbus mit 64 Leitungen sind folglich  $64/n$  Chips mit der Organisation „ $\times n$ “ erforderlich. Bei Modulen mit mehreren Bänken (siehe unten) sind mehrere Chips (2 oder 4) an den Datenleitungen parallel geschaltet. Folglich enthält ein Modul mit  $k$  Bänken  $64/n \times k$  Chips mit der Organisation „ $\times n$ “.

Zusätzliche Eingangsleitungen regeln die Auswahl des Speicherbausteins (Chip Select) und die Schreib- bzw. Leserichtung (R/W) der Date

<http://de.wikipedia.org/wiki/Speichermodul>

Der zEC12 Rechner verwendet in der Maximalausrüstung DIMMs mit einer Kapazität von 32 GByte. Das sind 2GByte pro Speicherchip, oder 16 Gbit pro Speicherchip.

## **Multi-Layer Ceramics (MLC) Literatur**

[ BUR ] W. G. Burger, C. W. Weigel: Multi-Layer Ceramics Manufacturing. IBM Journal of Research and Development, Volume: 27 No.1, Jan. 1983, p. 11 - 19

[ KAT ] G. A. Katopis et al. : *MCM technology and design for the S/390 G5 system.* IBM Journal of Research and Development, Vol. 43, Nos. 5/6, 1999, p. 621.

A. J. Blodgett, D. R. Barbour: Thermal Conduction Module: A High-Performance Multilayer Ceramic Package. Volume 26, Number 1, 1982, Page 30.

## Mainframe Earthquake Test

Die hohe Verfügbarkeit der Mainframes beruht auf einer sehr großen Anzahl einzelner Maßnahmen und Eigenschaften der Hardware, Software oder einer Kombination von beiden.

Sehr sichtbar ist dies, wenn man sich die Hardware anschaut. Alles ist grundsolide, kein Aufwand ist zu groß.

Ein Beispiel ist der „Erdbebentest“. Gefordert ist, dass bei einem Erdbeben mit dem Wert 9,0 auf der Richterskala der Mainframe Rechner dies ungestört überlebt, und dass spezifisch alle Software unbeeinträchtigt und ohne Absturz, (während des Bebens und hinterher) weiterläuft.

Das Video

<http://www.informatik.uni-leipzig.de/cs/Literature/esiisup/earthquake.html>

kann mit dem Microsoft Windows Media Player wiedergegeben werden. Es zeigt einen „Erdbebentest“, durchgeführt in der IBM Fabrik in Poughkeepsie, N.Y. Während des Testes laufen z/OS und alle anderen Subsysteme und Anwendungsprogramme ungestört weiter.

Wenn Sie einen System z Rechner öffnen, fällt auf, wie solide der ganze mechanische Aufbau ist.