UNIVERSITÄT LEIPZIG

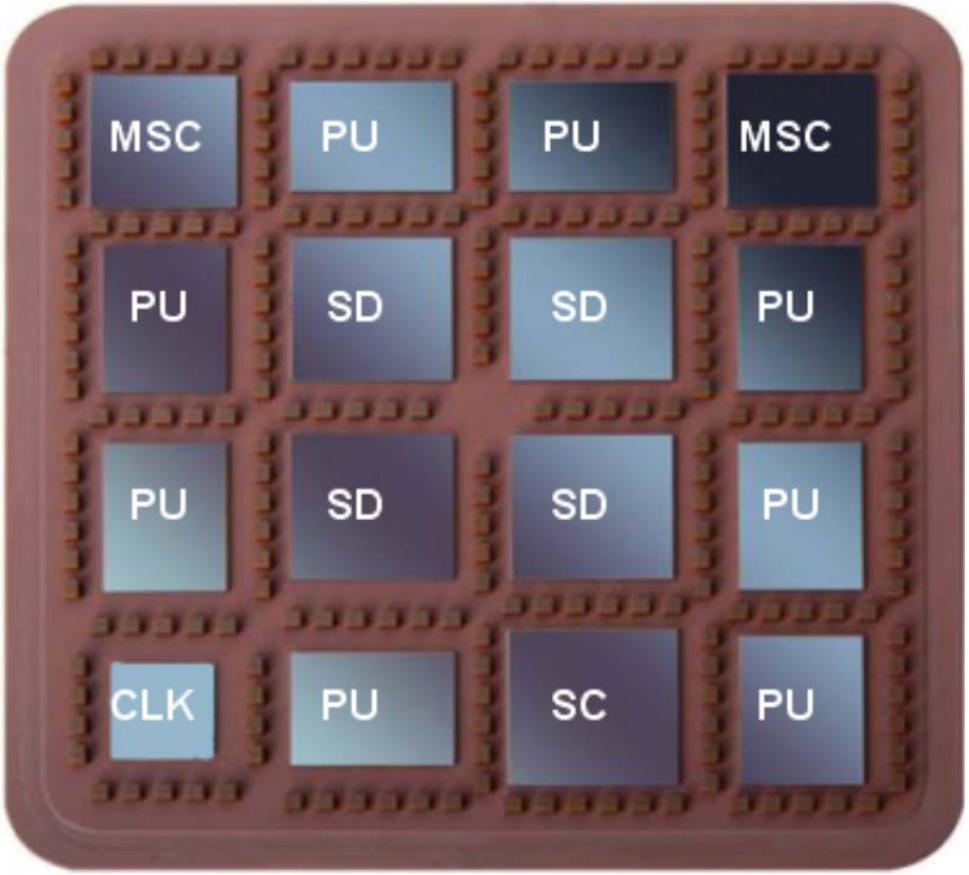
Enterprise Computing
Einführung in das Betriebssystem z/OS

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

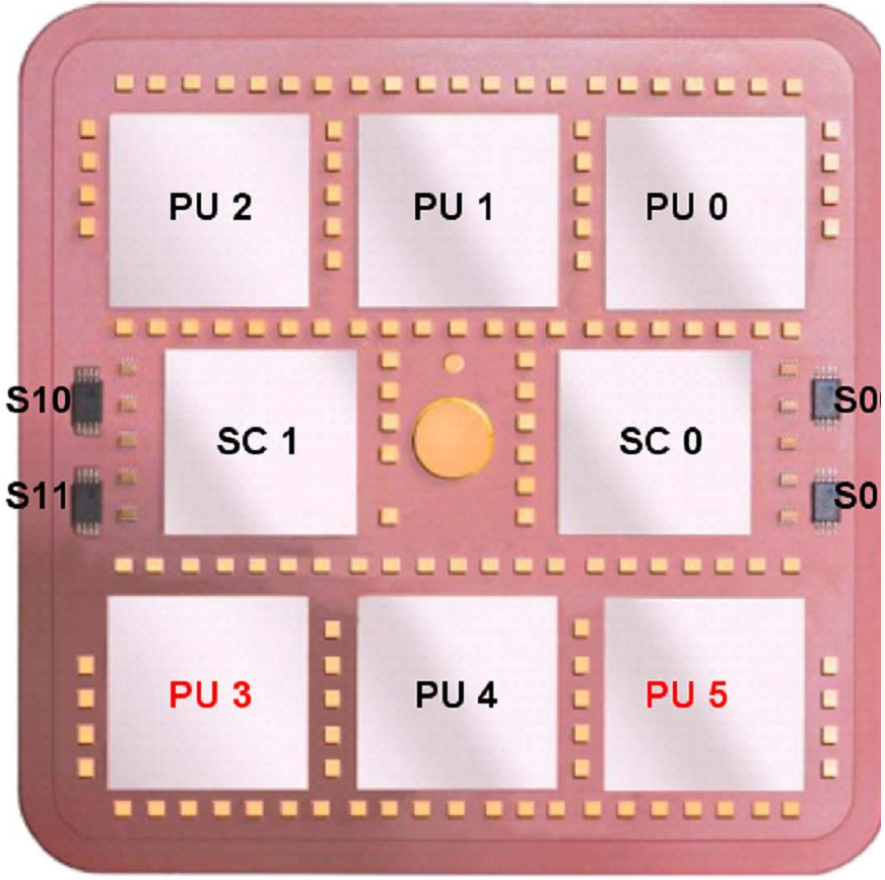
WS 2012/13

System z/Hardware Teil 2

Multichip Module



z9



z196

Vergleich der z9 und z196 Multichip Mulitilagen 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. 4 (z196) Cores, und der gemeinsame Cache (SD/SC) wuchs von 40 (z9) auf 196 (z196) MByte.

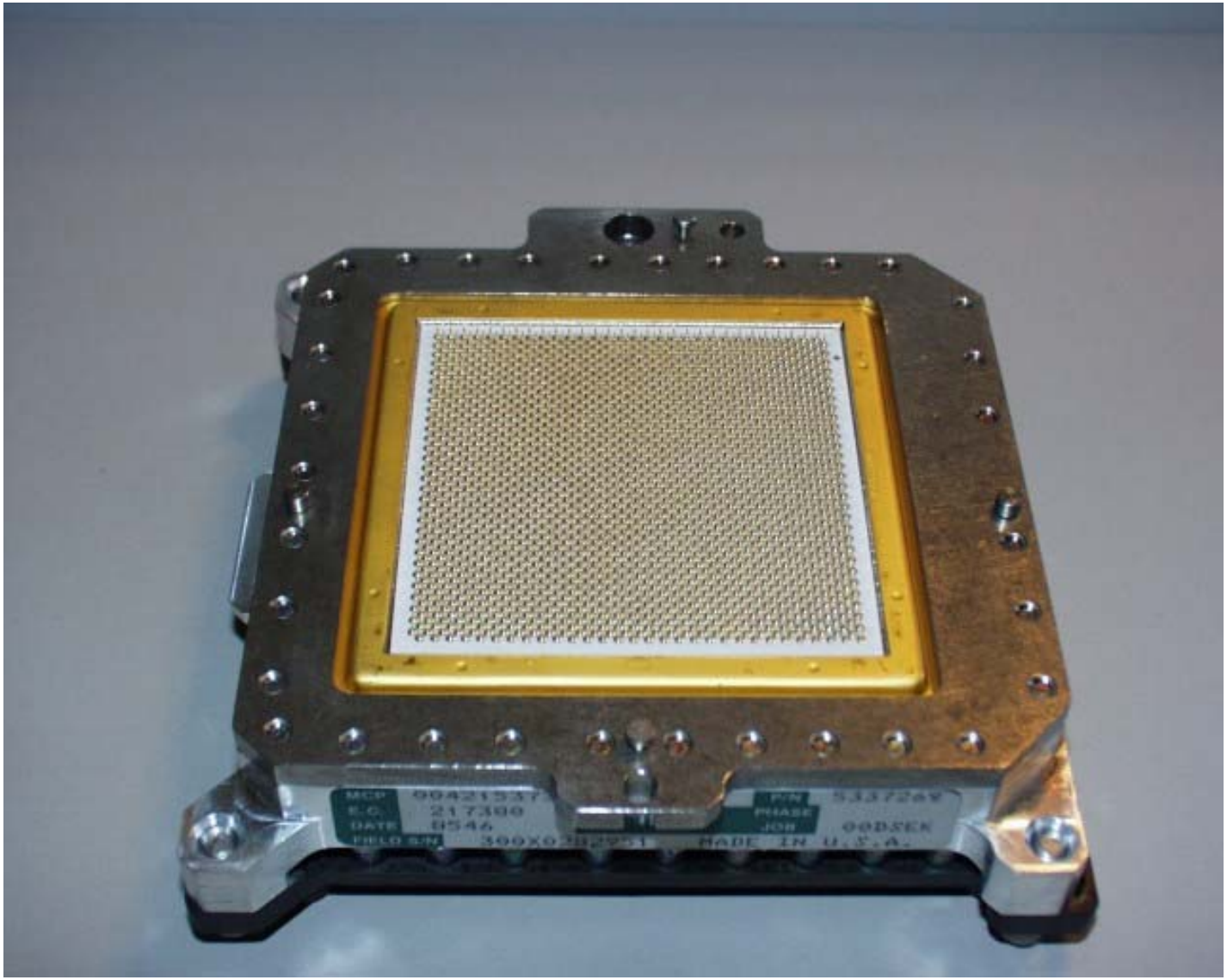
IBM bringt etwa 2 ½ Jahre verbesserte Mainframe Modelle heraus. So erschien das Modell z9 im Juli 2005, Modell z10 im Februar 2008, Modell z196 im Juli 2010 und Modell zEC12 im Juli 2012. Bei unserem Mainframe Rechner an der Uni Leipzig handelt es sich um ein Modell z9 .

z9 Multi-Chip Module

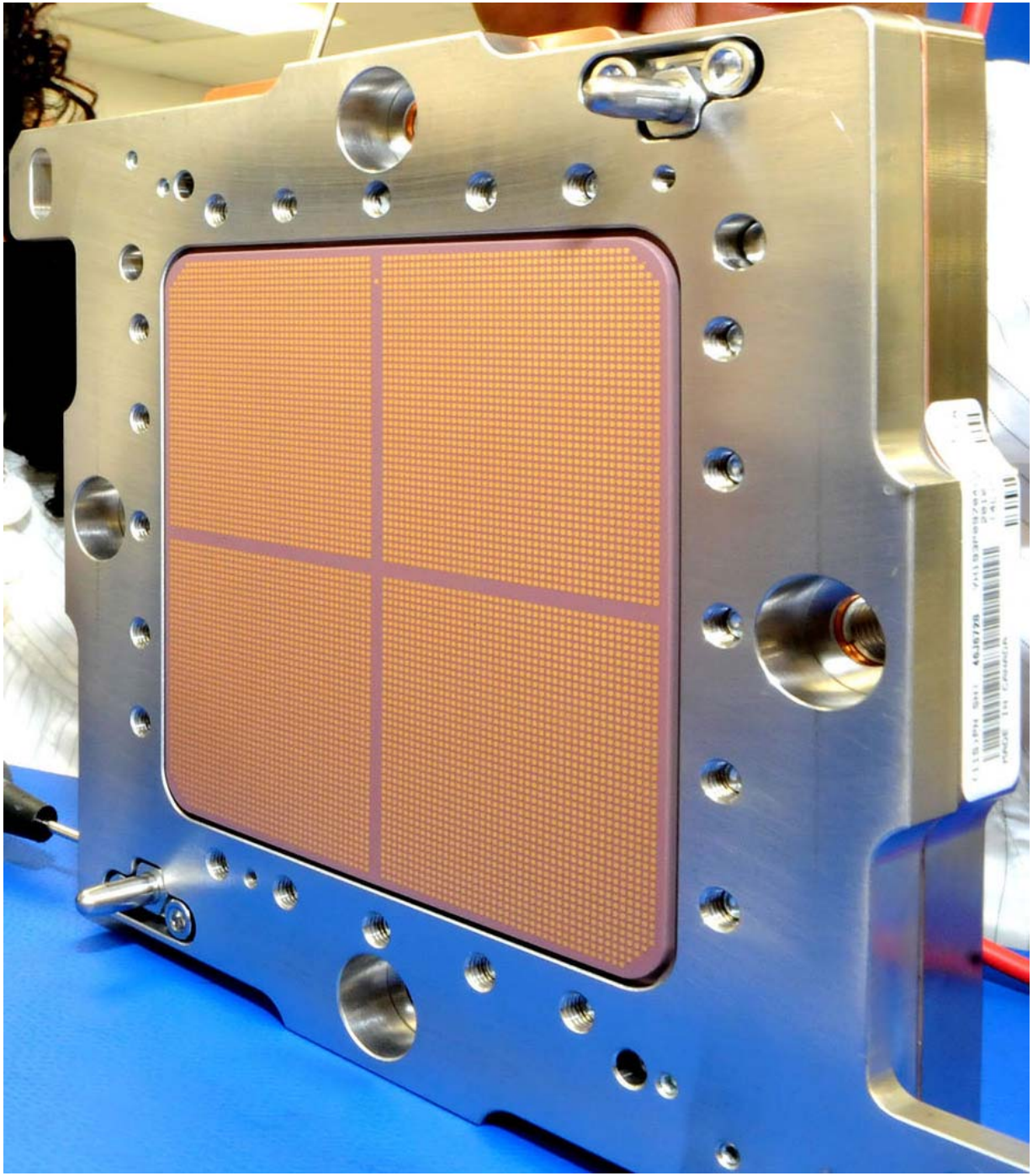
Multichip Module (MCM) benutzen die Multilagen Ceramic (MLC) Technologie. Ein MLC Module ihat etwa 100 Verdrahrungslagen und ist wenige mm dick. Gezeigt ist das Einpassen eines z9 MCM in eine Halterung.

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 .



Die Rückseite des MLC Modules enthielt früher vergoldete Kontaktstifte, welche 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 z196 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 (land) 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 Lotperlen gleich mit. Alle drei Varianten sind hauptsächlich für ICs mit Hunderten bis über Tausend Anschlüssen gedacht.

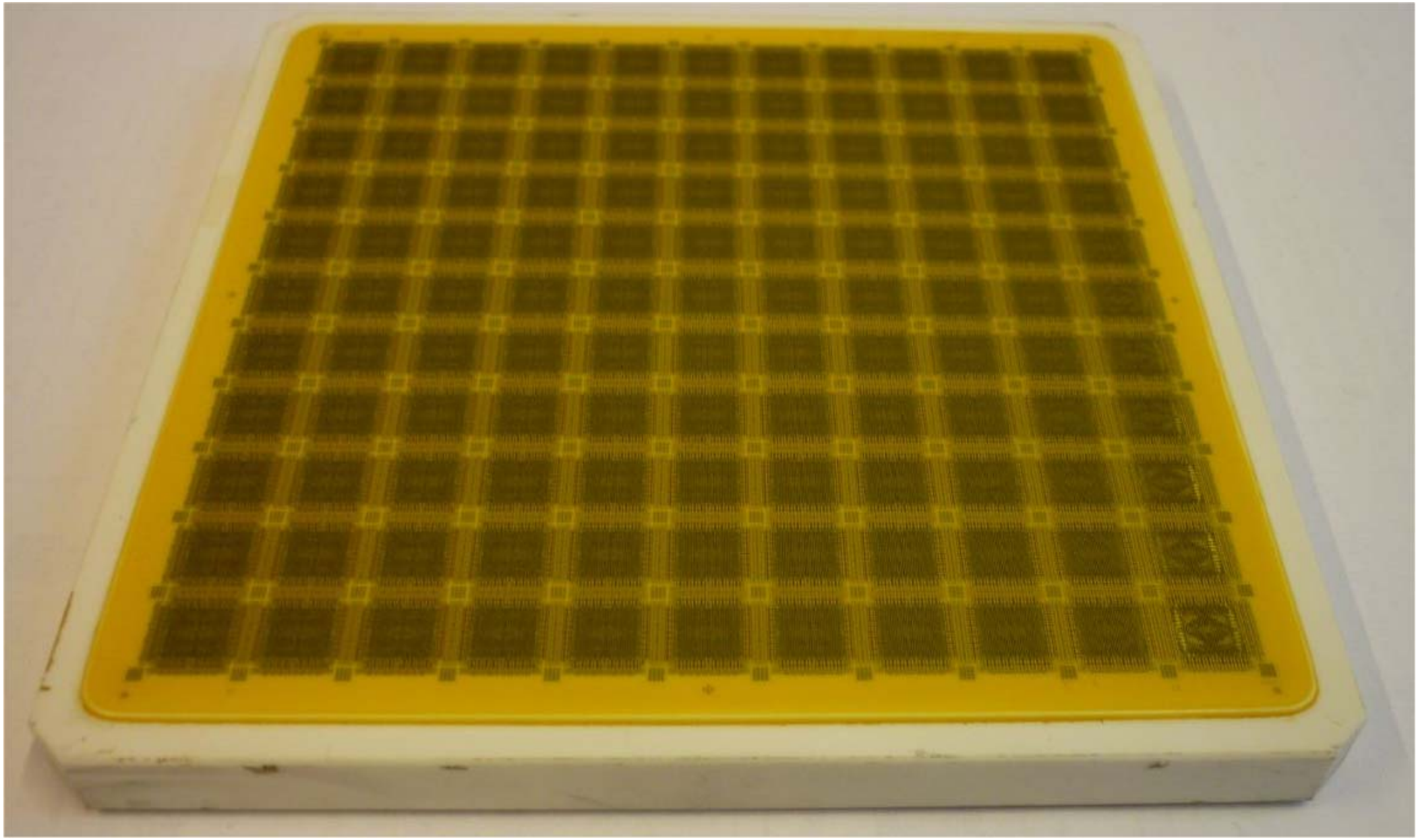
Das Land Grid Array ist im Gegensatz zum Pin Grid Array für höhere Frequenzen geeignet und günstiger zu produzieren.

Thermal Conduction Module

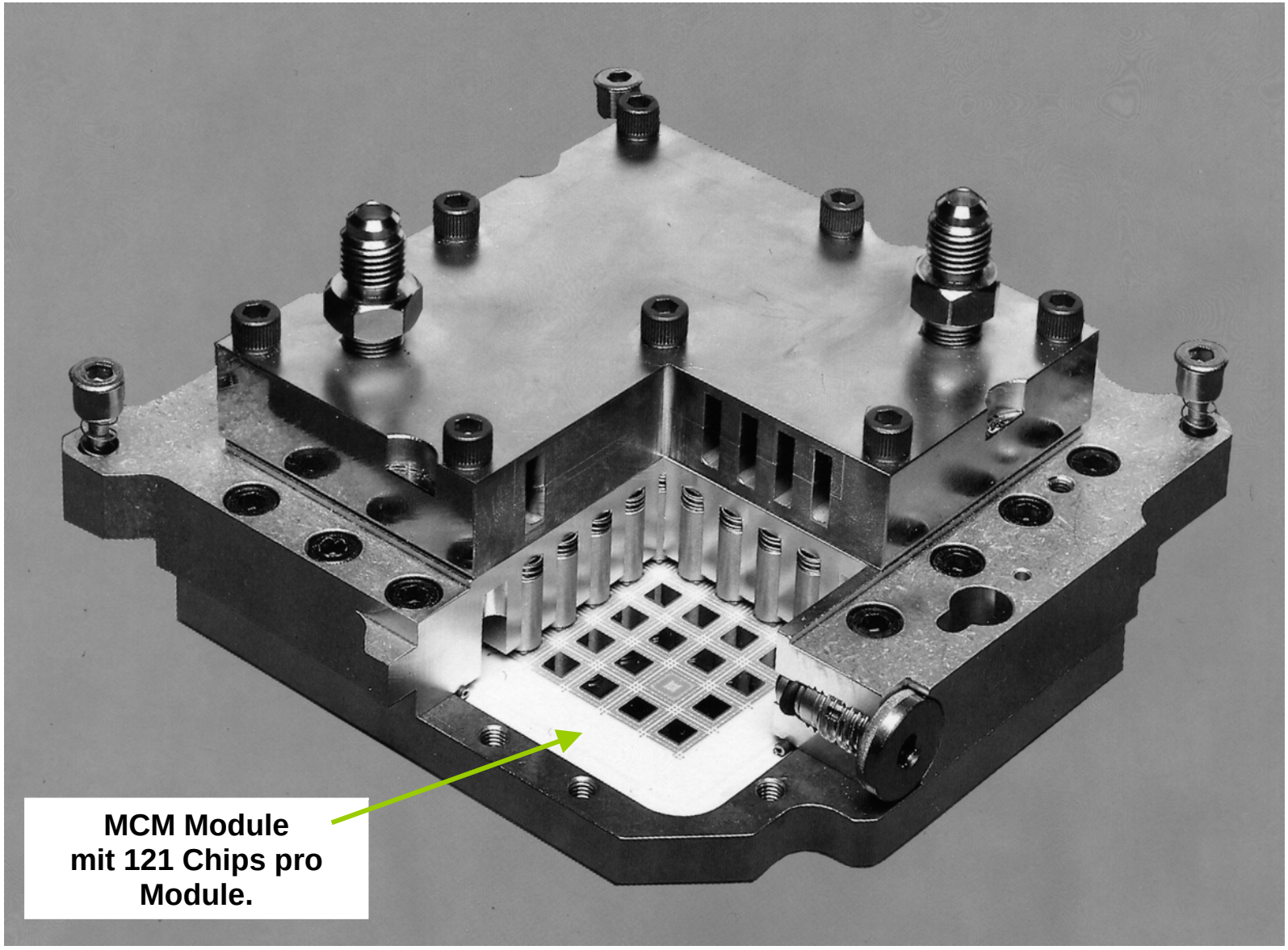
Ein Thermal Conduction Module (TCM) ist eine Baugruppe, welche ein Multilayer Ceramic (MLC) Multichip Module (MCM) enthält, und die für die Energieabfuhr (Kühlung) erforderliche Hardware enthält.

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

Die allermeisten Mainframe Modelle benutzen seitsem 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.

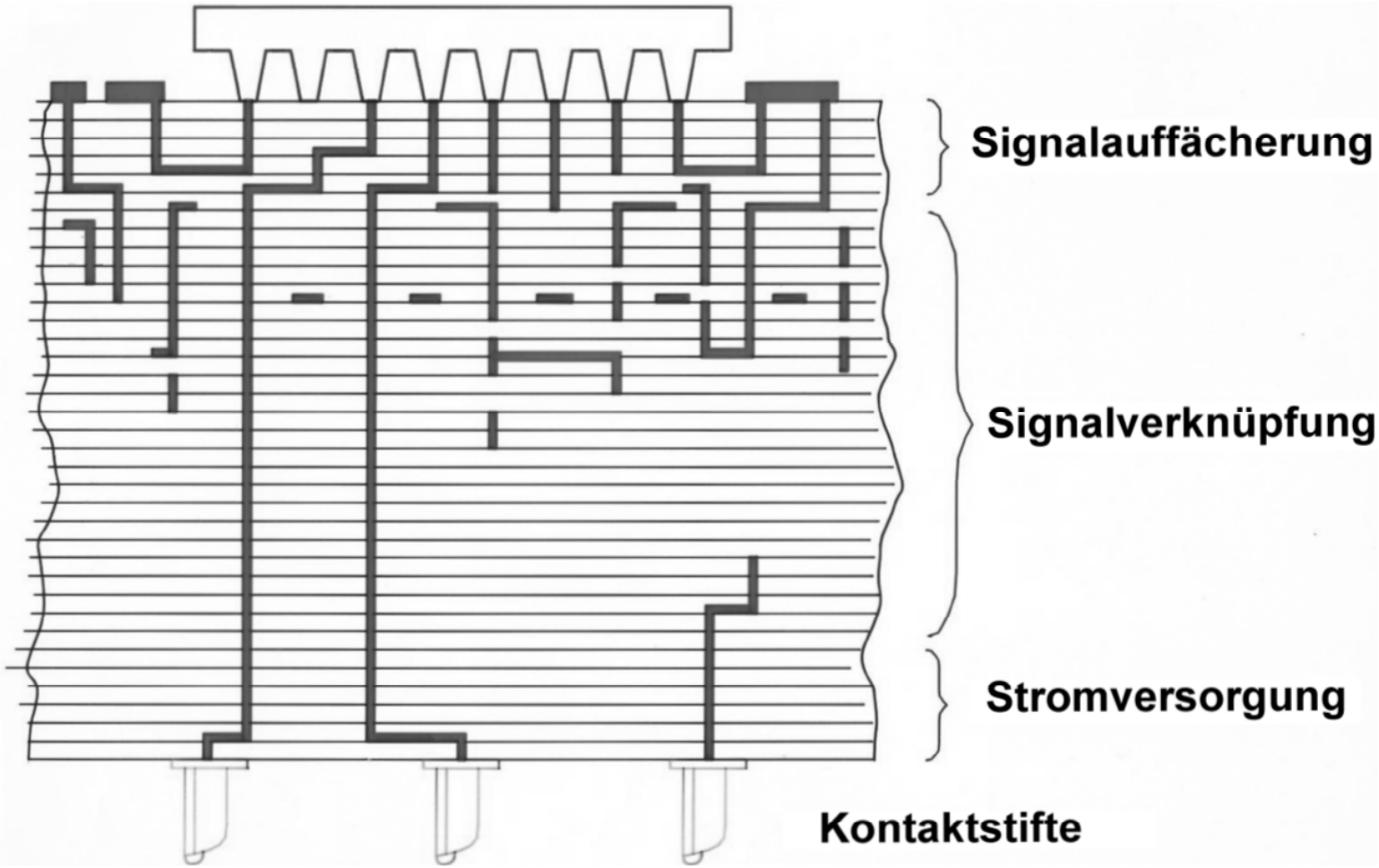


Mainframe Multichip Multilayer Ceramic Module mit 121 Chip Sites aus den 80er Jahren



Thermal Conduction Module (TCM)

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



Multilayer Ceramic Technologie

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

Das zEC12 - MCM 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 (AL2O3)-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.

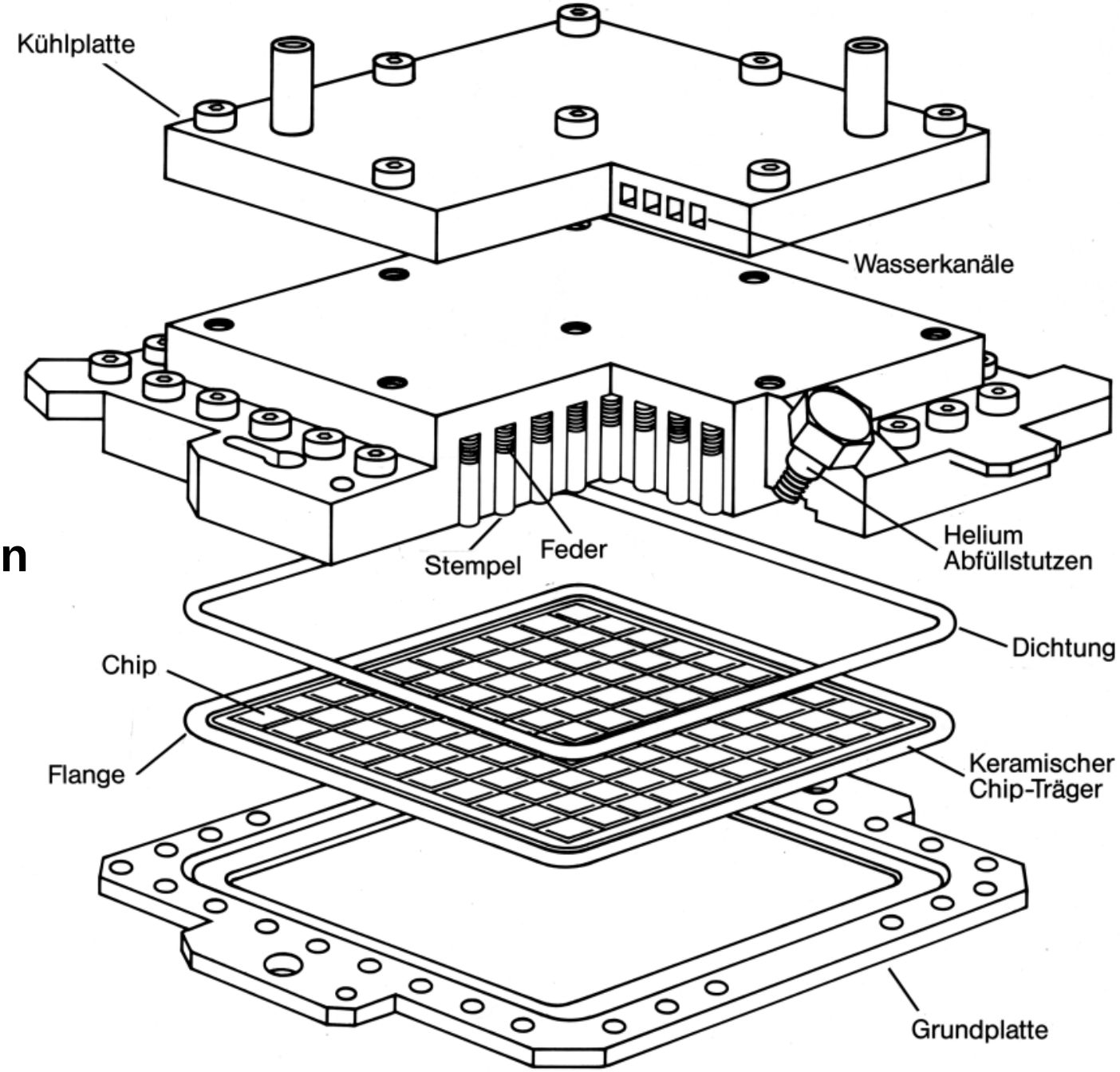
MLC Literatur

**[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.

Thermal Conduction Module



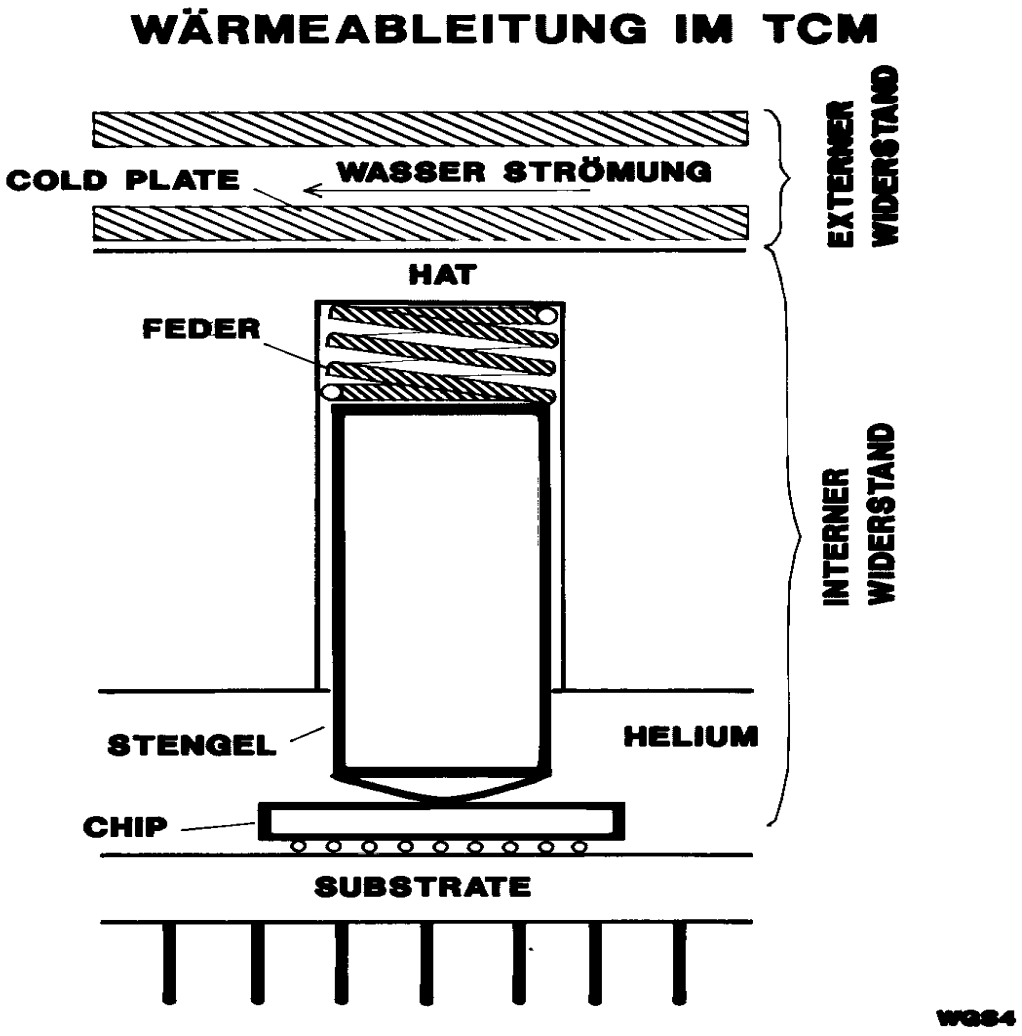
Eine MCM-Baugruppe enthält neben dem keramischen Chip-Träger die Mechanik für die Energieableitung mittels Stempel und Kühlplatte. Diese Baugruppe wird als Thermal Conduction Module (TCM) bezeichnet und ist in der obigen Abbildung dargestellt.

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.

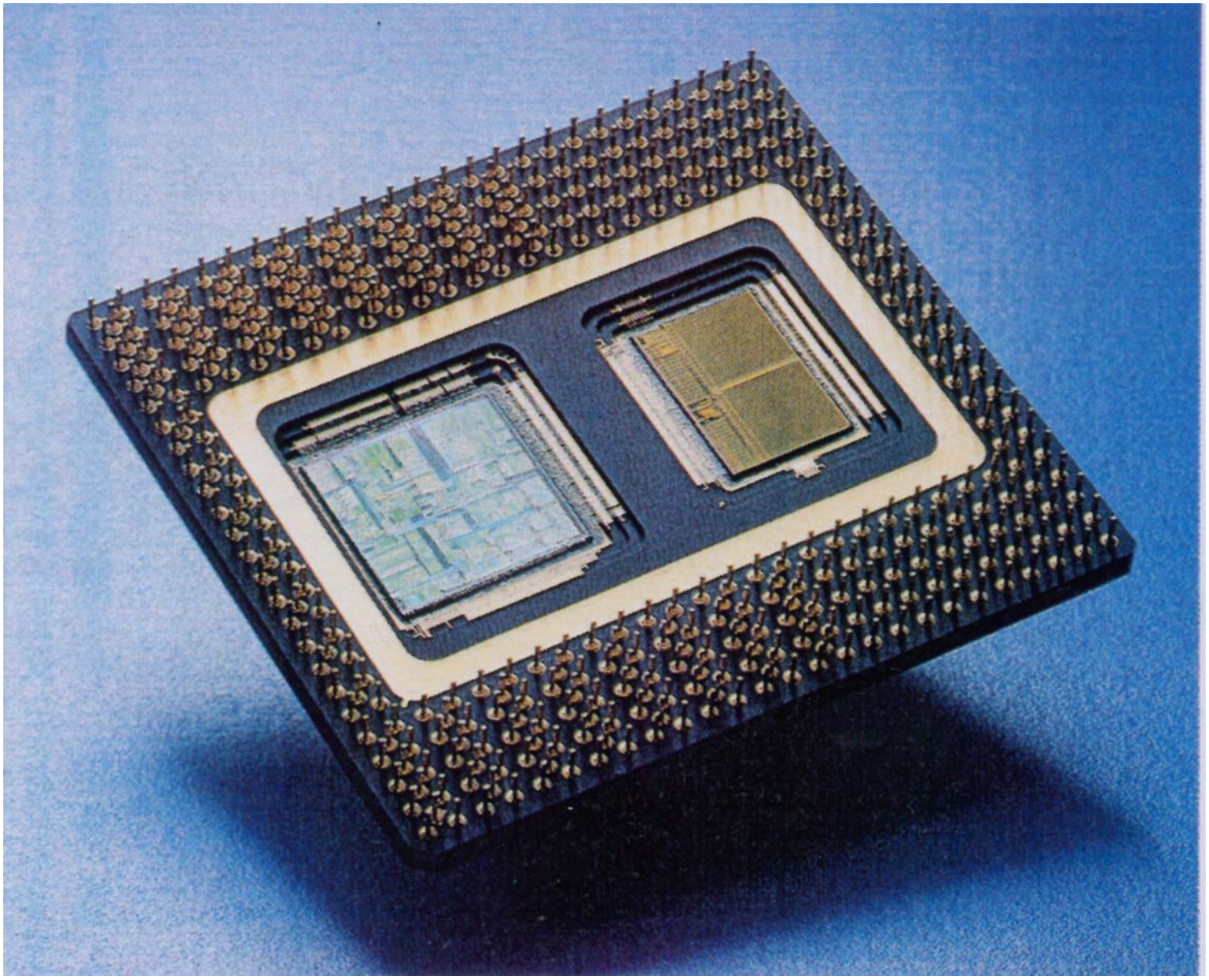
Bei dem hier vorgestellten Verfahren werden alle Chips selbst mit Luft gekühlt. Es sind in der Vergangenheit viele Versuche unternommen worden, Chips direkt mit einer Flüssigkeit zu kühlen. Diese Verfahren haben sich sich in der Praxis nicht bewährt.



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.

TCM Wärmeübergang



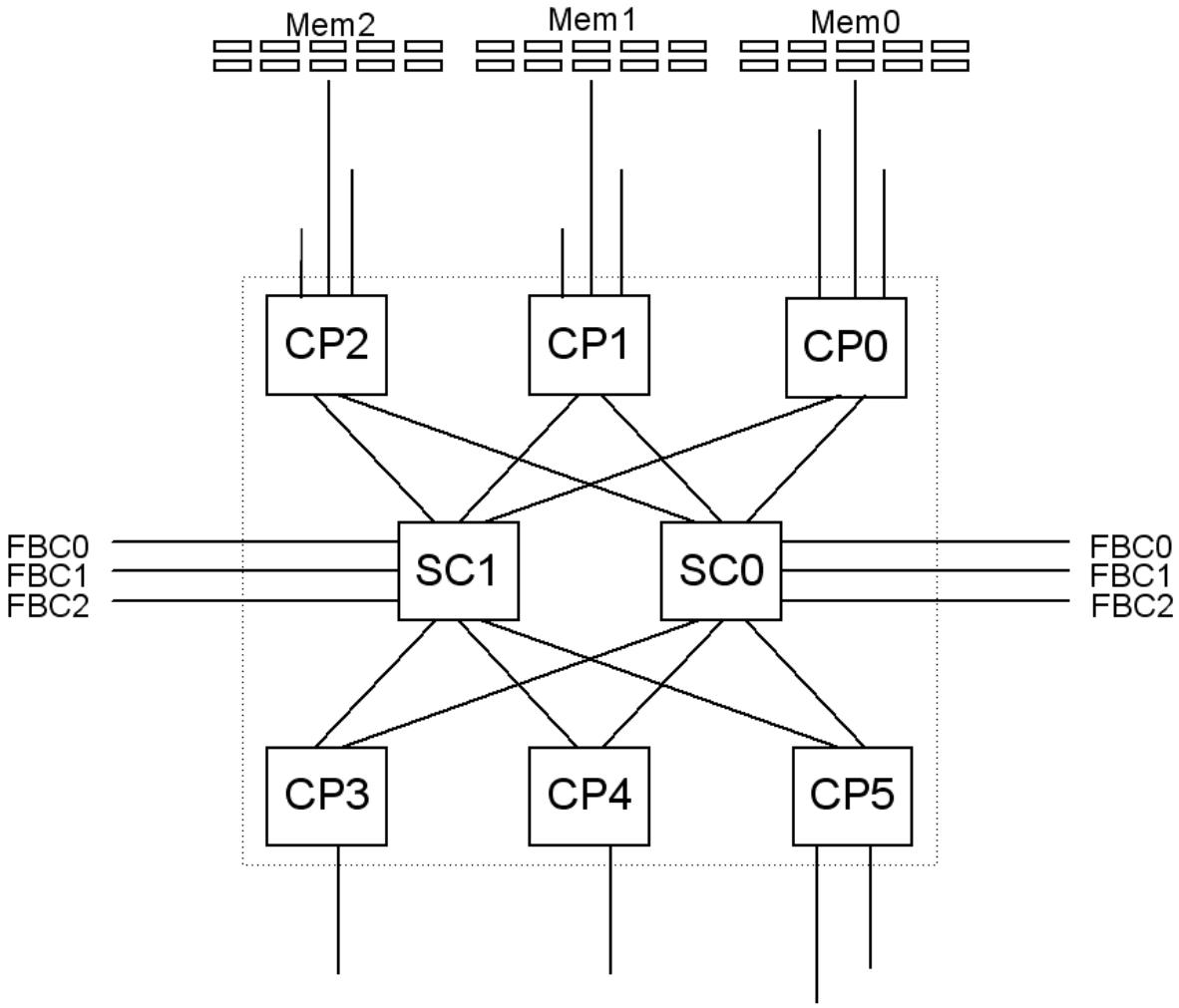
Der 1995 von Intel herausgebrachte Pentium Pro Microprozessor verwendete ein 387 Pin Multi Layer Ceramic (MLC) Multi Chip Carrier (MCM) Module und hatte eine überdurchschnittliche Leistung.

Intel Pentium Pro

Bei der ursprünglichen Einführung des Pentium hatte Intel eine ähnliche MCM-Technologie für den Pentium Pro eingesetzt. Der Pentium Pro Microprocessor wurde von Intel im November 1995 eingeführt. Er wurde für Server und für High-End Desktop Processor en 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.

Die MCM-Technologie ermöglicht besonders günstige Signallaufzeiten zwischen den Chips.

Der Pentium Pro benutzte ein 387 Pin Multi Layer Ceramic (MLC) Multi Chip Carrier (MCM) Module. Das Module enthielt zwei Chips, ein CPU chip und ein getrenntes L2 Cache Chip, beide auf dem gleichen MLC Substrate.



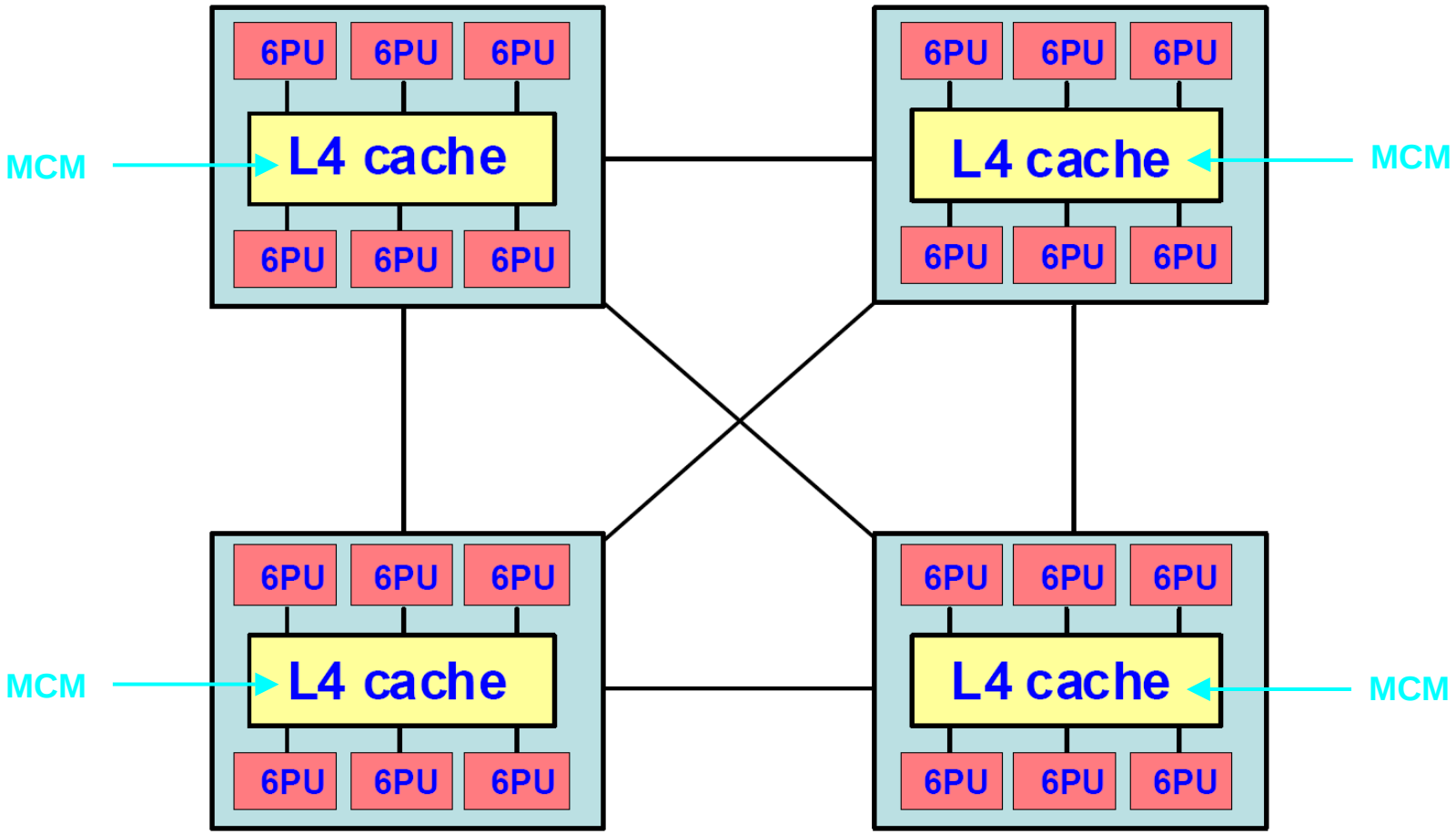
z196 Multichip Module

Auf dem EC12 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.

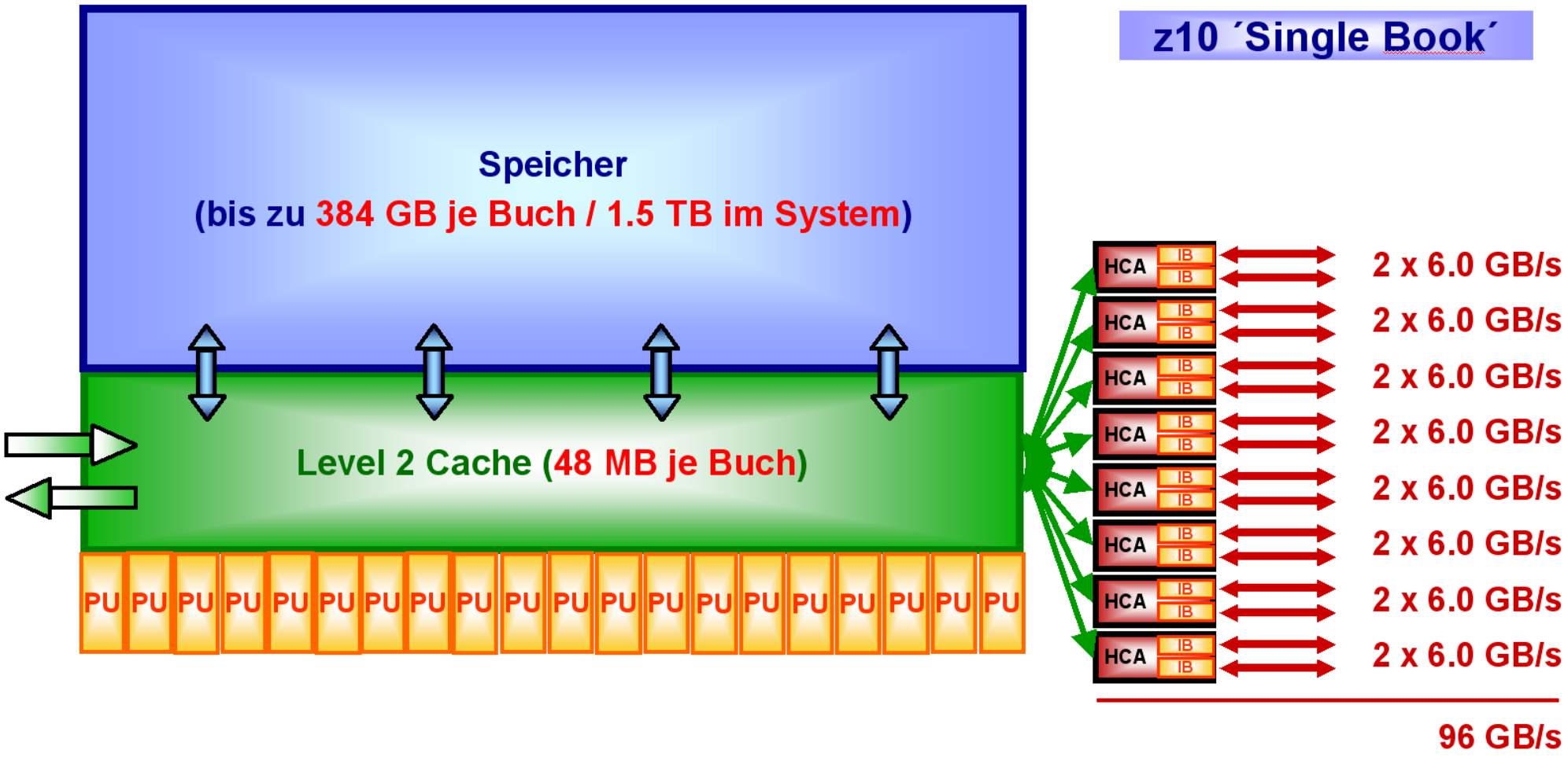
Ein z 196 System kann vier als „Books“ bezeichnete Baugruppen enthalten. Jedes Book beinhaltet ein Multichip Module.

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



Die vier L4 Caches (je 2 L4 Cache Chips) der vier z196 Books mit je 24 CPU Cores sind mittels einer Punkt zu Punkt Topologie über „Fabric Book Connectivity (FBC)“ Leitungen direkt miteinander verbunden. Direkter Datenaustausch zwischen den vier L4 Caches.

Alle 4 x 24 = 96 Cores greifen auf die L4 Caches aller Books direkt zu. Die vier L4 Caches implementieren einen NUMA (Non Uniform Memory Architecture) Cache.



Dies ist ein weiteres Mainframe Alleinstellungsmerkmal: In allen nicht-Mainframe Servern (und allen PCs) bewirken die Fanout Adapter Karten einen Datentransfer zwischen Plattenspeicher und Hauptspeicher. In den Mainframe Rechnern erfolgt der Datentransfer zwischen Plattenspeicher 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.