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 1

Microprocessor Technologie

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 und Windows/Linux 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

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 verrsucht man Verbesserungen zu erreichen, die dann aber ins Geld gehen.**

- **Ein/Ausgabe Einrichtungen (Input/Output, I/O) sind z.B. für den Ansschluss von 5 Plattenspeichern optimiert, nicht aber für 5 000 oder 50 000 Plattenspeicher**

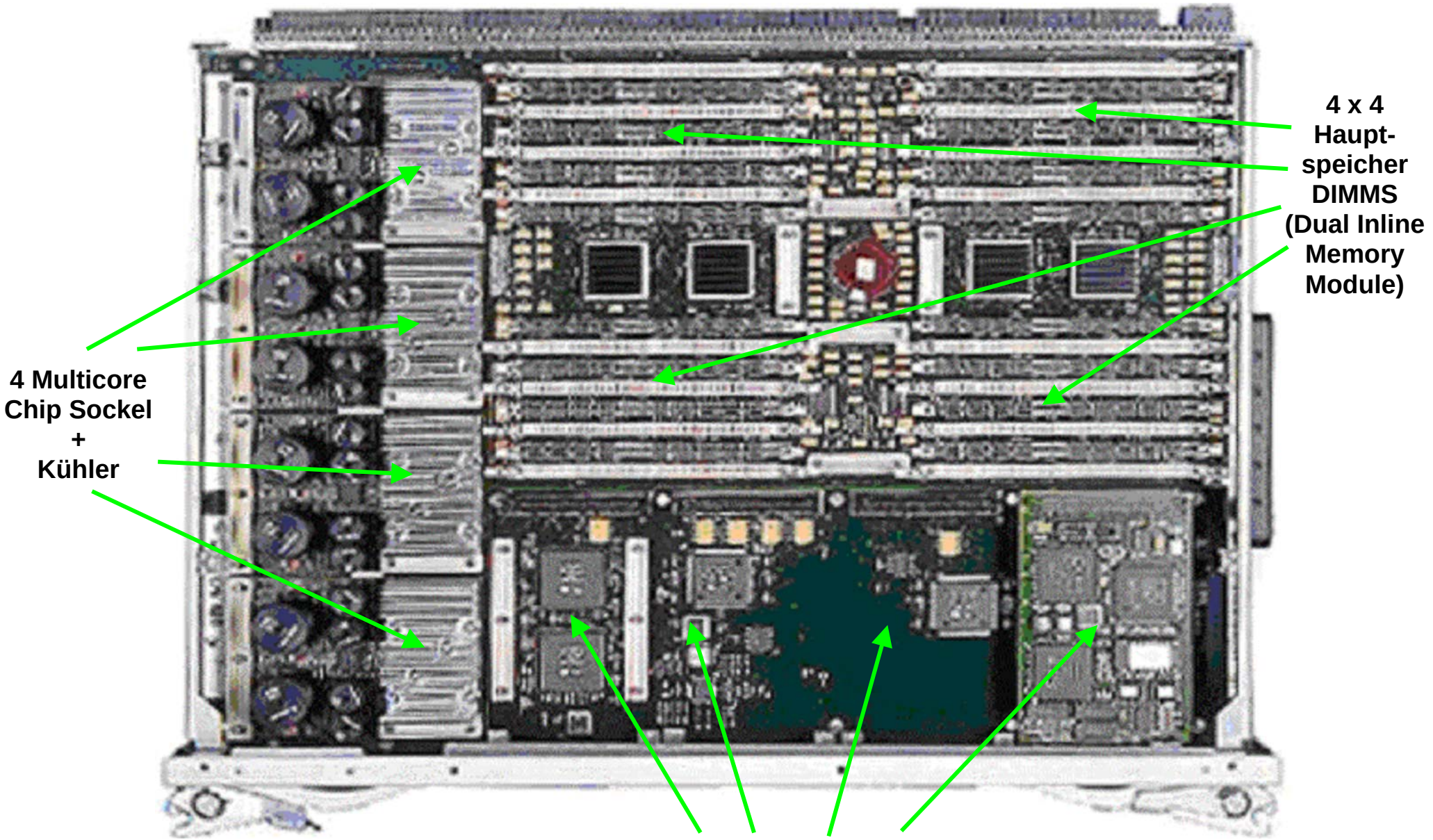
Dennoch sind betriebswirtschaftlichen Großrechner alles andere als billig, wie das folgende Beispiel eines führenden Herstellers, der Firma Sun, zeigt:

Das kleinste Modell der Sun 25k Serie hatte 4 „System Boards“ mit je 4 Dual Core SPARC CPUs, oder insgesamt 16 CPUs. 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 System Boards der Sun 25k Serie werden auch in anderen Produkten der Firma Sun eingesetzt, z.B. in einer Low End Workstation mit einem einzigen System Board.

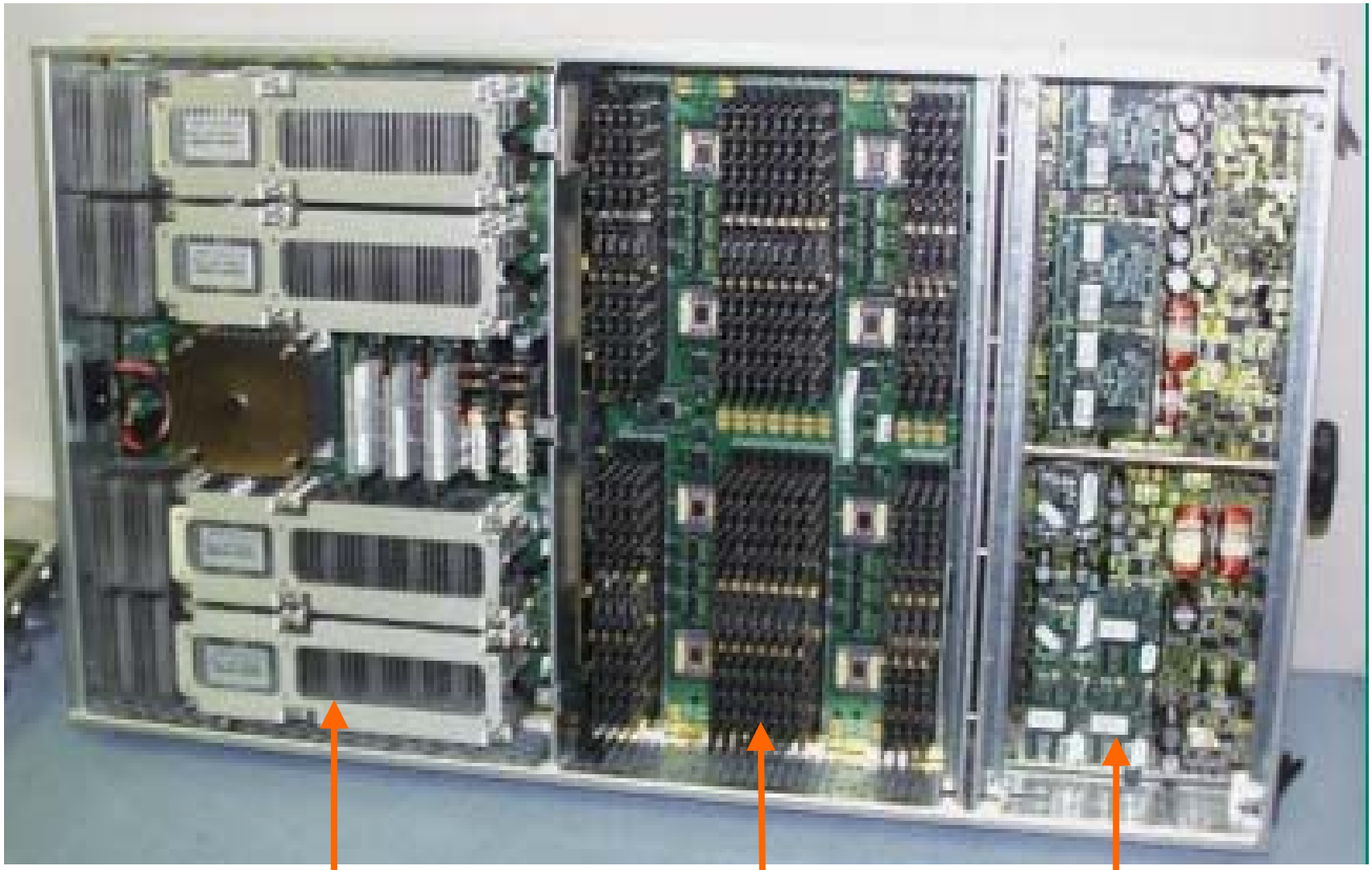
Die Firma Sun hat ihre Produkte kontinuierlich weiterentwickelt ohne dass sich am Konzept viel verändert hat. Die neuesten Modelle werden als M9000 bezeichnet, und gemeinsam von Sun und von Fujitsu/Siemens vertrieben.

Die nächste Abbildung zeigt ein Sun System Board.



Sun E 10 000 System Board

Hewlett-Packard Superdome Cell Board

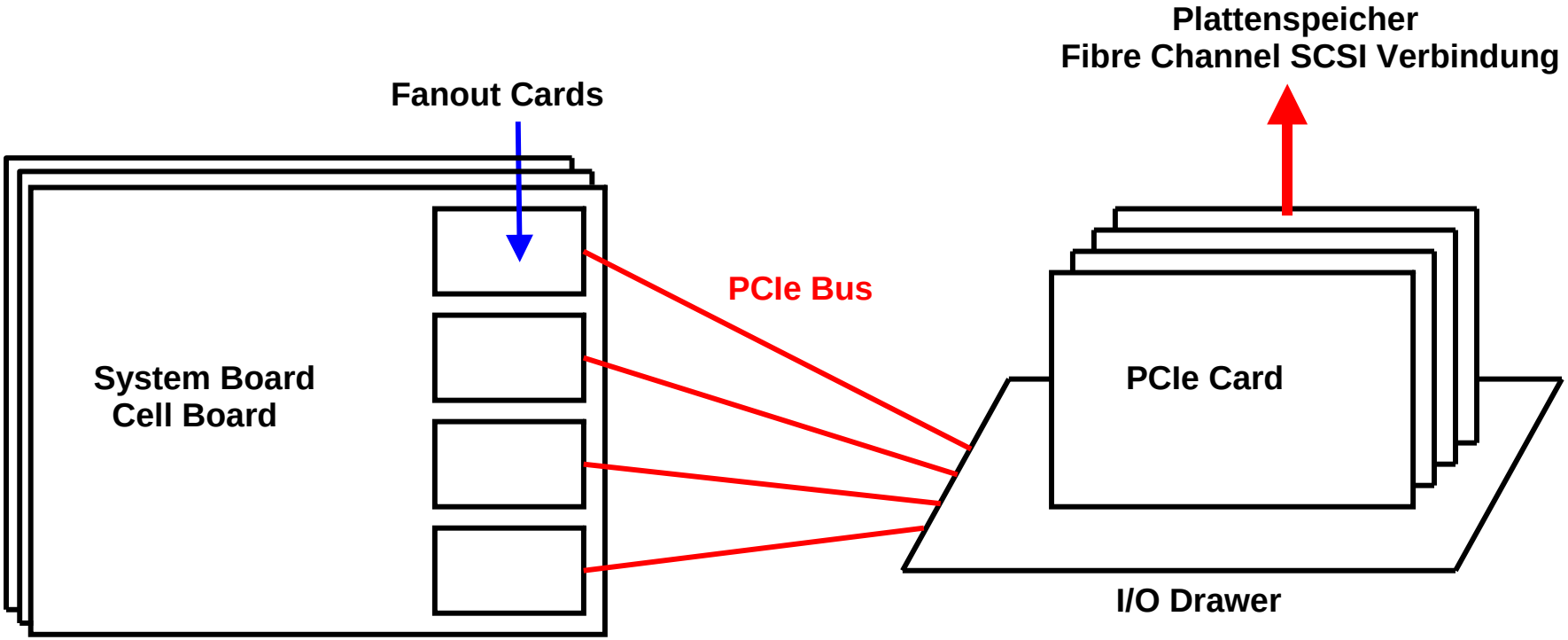


4 Itanium 2 CPU Chips

Hauptspeicher
2 x 16 DIMMs

6 Fanout Cards

Das Cell Board des Hewlett Packard Superdome Rechners haben 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 Plattenspeichern, von denen evtl. hunderte oder mehr angeschlossen werden müssen. Statt dessen existieren Fanout Cards, die mittels DMA auf den Hauptspeicher zugreiffen 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 Plattenspeicher Anschlüsse mittels Fibre Channel SCSI Kabeln anzutreffen. Auf diese Art ist es möglich, hunderte oder mehr Plattenspeicher an einen SUN oder HP Großrechner anzuschließen.

Bei einem Mainframe werden die Fanout Cards auch als „Host Channel Adapter“ (HCA) bezeichnet.

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.

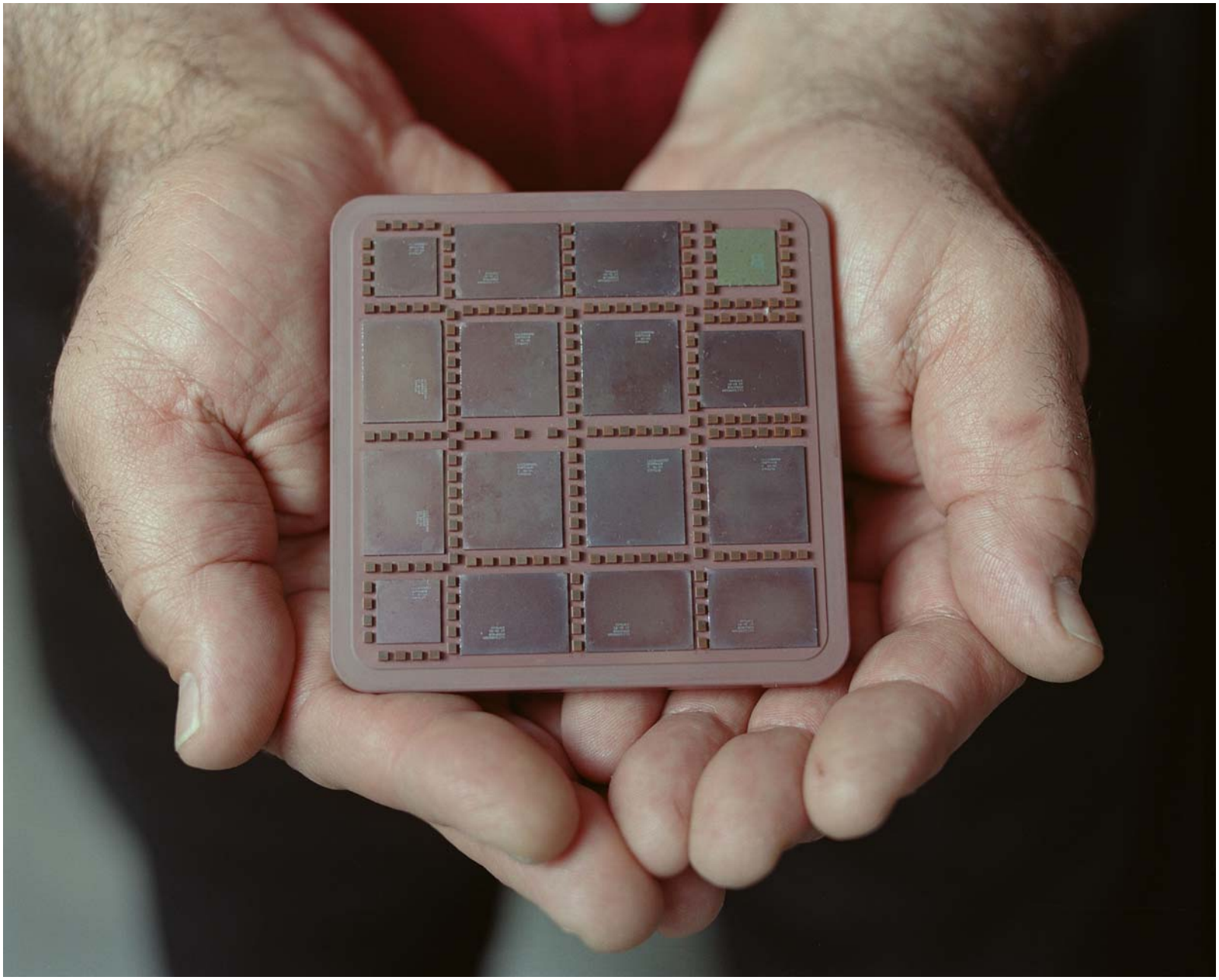
System z verwendet an Stelle von Prozessor Boards sog. „Books“. Das Äquivalent zur Fanout Adapter Card wird als HCA (Host Channel Adapter) Card bezeichnet.

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.

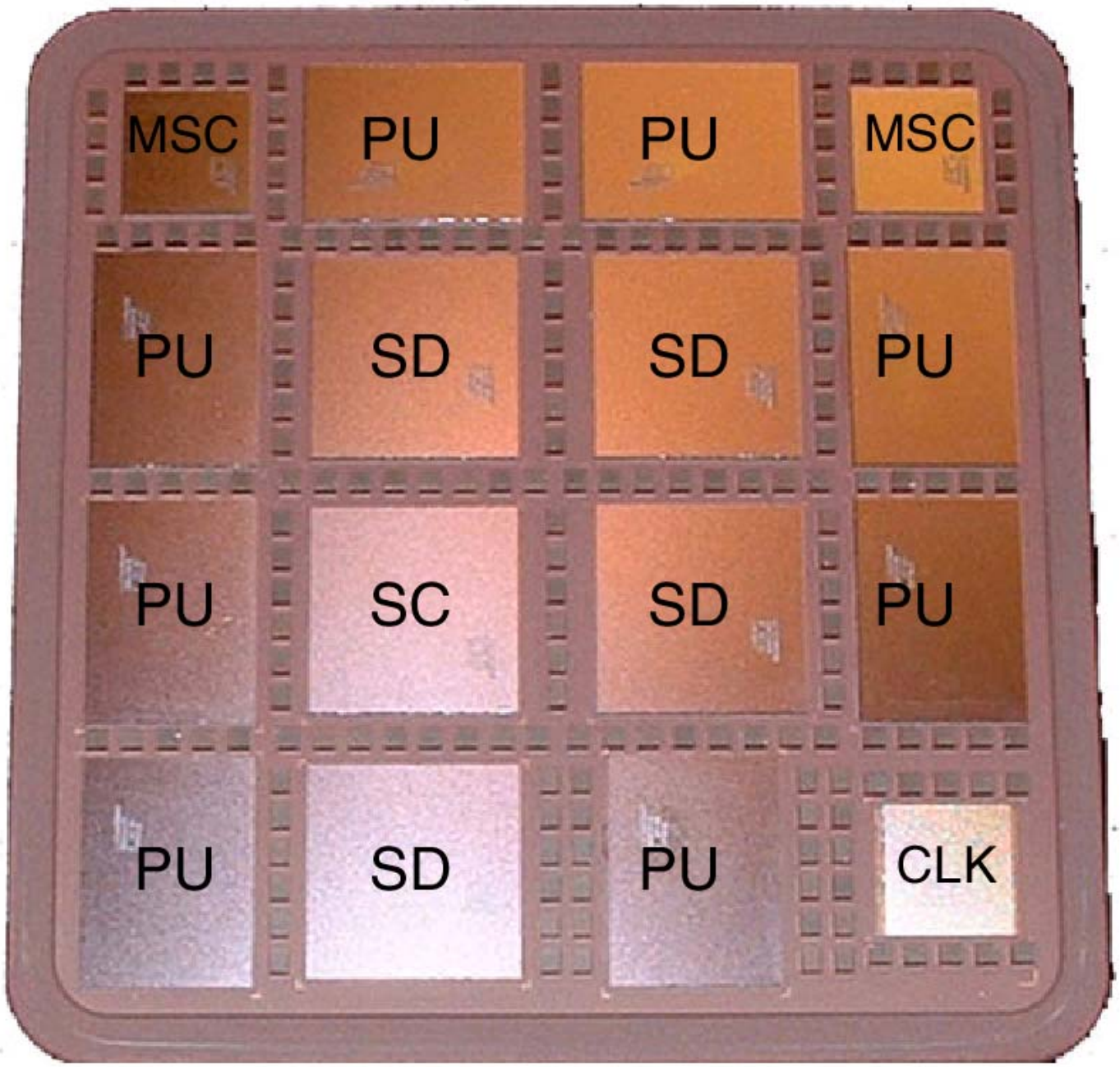
Alle diese Hersteller außer IBM setzen aus der PC-Welt abgeleitete Technologien für ihre Processor Boards ein. Besonders verwenden die Prozessor Boards die gleich Printed Circuit Board (PCB) Technologie, die auch für die Mainboards der PCs benutzt wird.

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) aufgegelötet. Multilayer Boards können aus bis zu 48 Schichten bestehen. PC Mainboards haben in der Regel unter 10 Schichten.

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



Multi-Chip Module eines z9 Rechners



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 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 Mulitilayer 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:

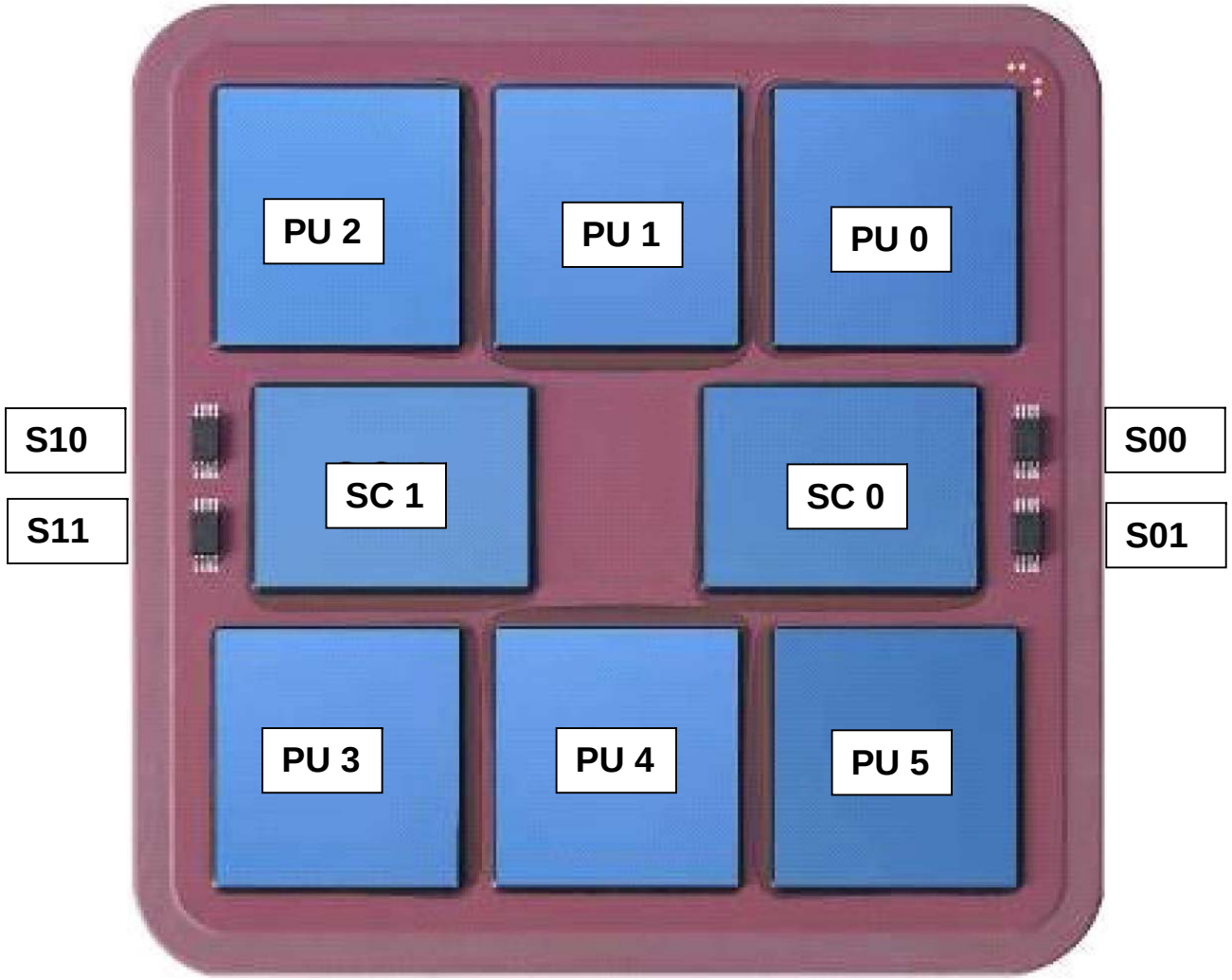
- **kleinere Abstände zwischen den Chips**
- **günstigere Dielektrizitätskonstante von MLC im Vergleich zu zur Printed Circuit Technologie**

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 seint 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.



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 EPROM chips labeled S00, S01, S10 und S11. Sie dienen der Personalisierung (characterization) jedes einzelnen MCMs.**

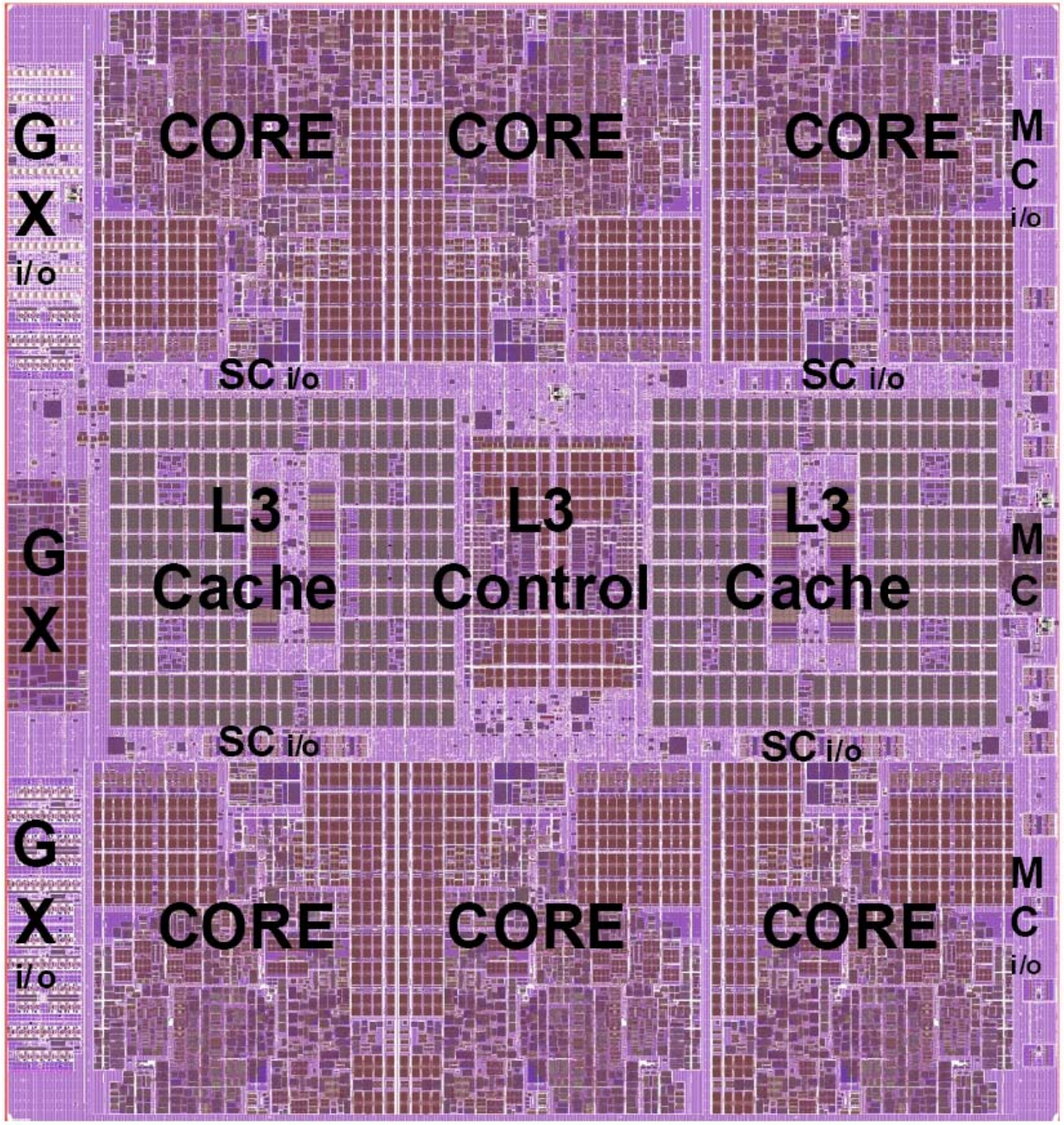
zEC12 MCM

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

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

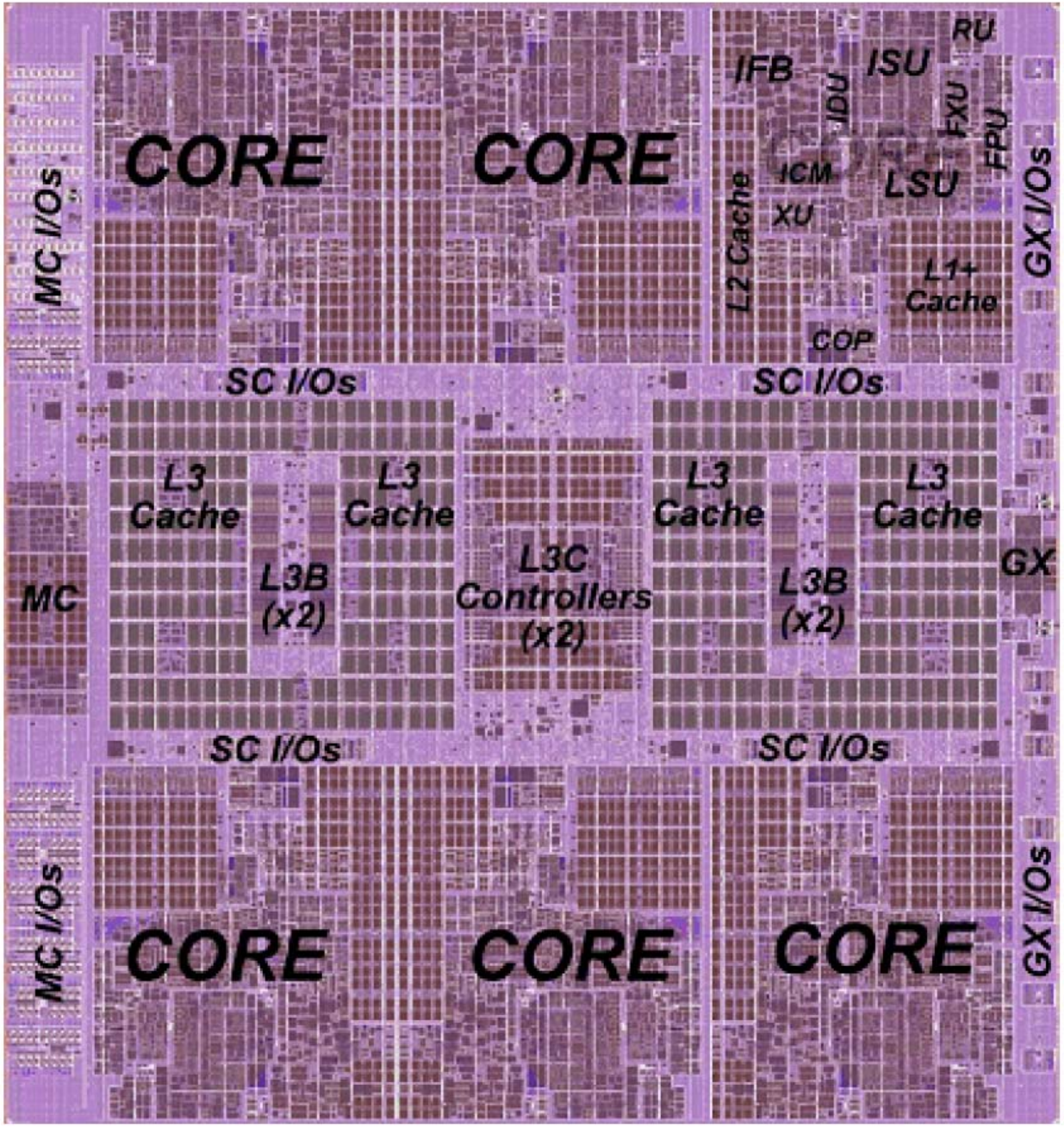
Modell z9	**verfügbar seit Juli 2005**
Modell z10	**verfügbar seit Februar 2008**
Modell z196	**verfügbar seit Juli 2010**
Modell zEC12	**verfügbar seit August 2012**

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.



Layout des zEC12 CPU Chips

There is one L3 cache, with 48 MB. This 48 MB L3 cache is a store-in shared cache across all cores in the PU chip. It has 192 x 512Kb eDRAM macros, dual address-sliced and dual store pipe support, an integrated on-chip coherency manager, cache and cross-bar switch. The L3 directory filters queries from local L4. Both L3 slices can deliver up to 160 GB/s bandwidth to each core simultaneously. The L3 cache interconnects the six cores, GX I/O buses, and memory controllers (MCs) with storage control (SC) chips. The memory controller (MC) function controls access to memory. The GX I/O bus controls the interface to the host channel adapters (HCAs) accessing the I/O. The chip controls traffic between the cores, memory, I/O, and the L4 cache on the SC chips.



Layout des zEC12 CPU Chips

Jedes CPU chip hat 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 (näheres dazu später).

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

~300 watts/ PU Chip

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

eDRAM

Ein Static Random Access Memory (SRAM) benötibt zwischen 4 – 10 Transistoren in einer Flip-Flop Schaltung für jedes gespeicherte Bit. Ein Dynamic Random Access Memory (DRAM) benötigt 1 Transistor un einen kleinen Kondensator für jedes gespeicherte Bit. Ein DRAM Speicher klann 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).

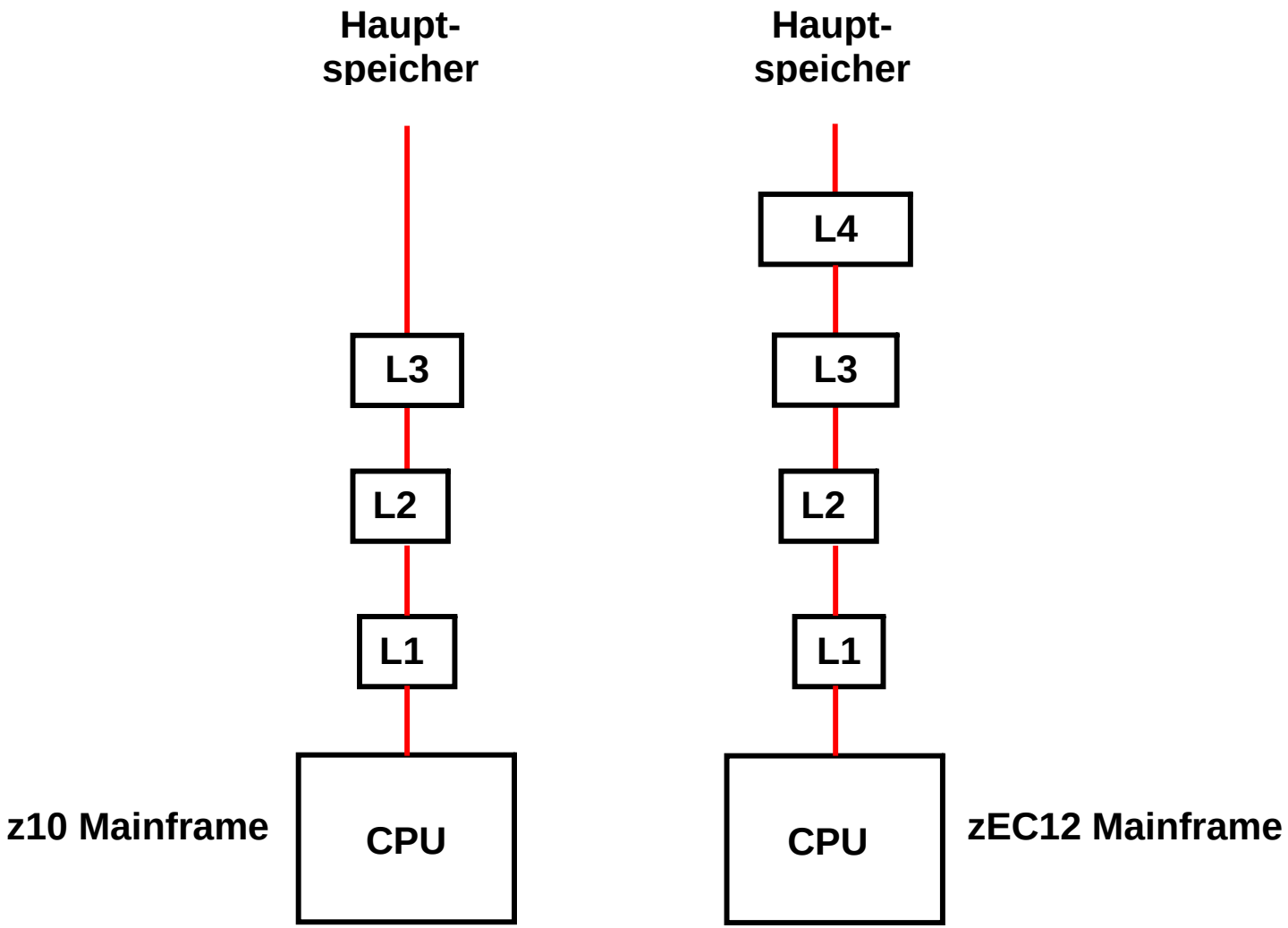
Die Hauptspeicher nahezu aller Rechner verwenden fast immer DRAMS, häufig in der Form von SIMM oder DIMM Steckkarten. Cache Speicher werden fast immer in SRAM Technologie implementiert.

Die L3 und L4 Caches eines z196 Rechners werden in einer neuartigen eDRAM (embedded DRAM) Technologie implementiert. eDRAM ist nahez so schnell wie SRAM, benötigt aber viel weniger Platz.

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

Das Einbetten des Speichers ermöglicht gegenüber externen Speichermodulen die Nutzung größerer Busse und höhere Arbeitsgeschwindigkeiten. Durch den geringeren Platzbedarf ermöglicht DRAM verglichen mit SRAM 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 Spielkonsolen genutzt, z.B. Xbox 360 von Microsoft, Wii von Nintendo und PlayStation 3 von Sony.



Mainframe Cache Hierarchien

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

Vier oder acht Cores pro Chip ?

Ein Blick auf das Layout des zEC12 CPU Chips zeigt, dass die 4 L2 Caches mit 4 x 1,5 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 neuste 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

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.

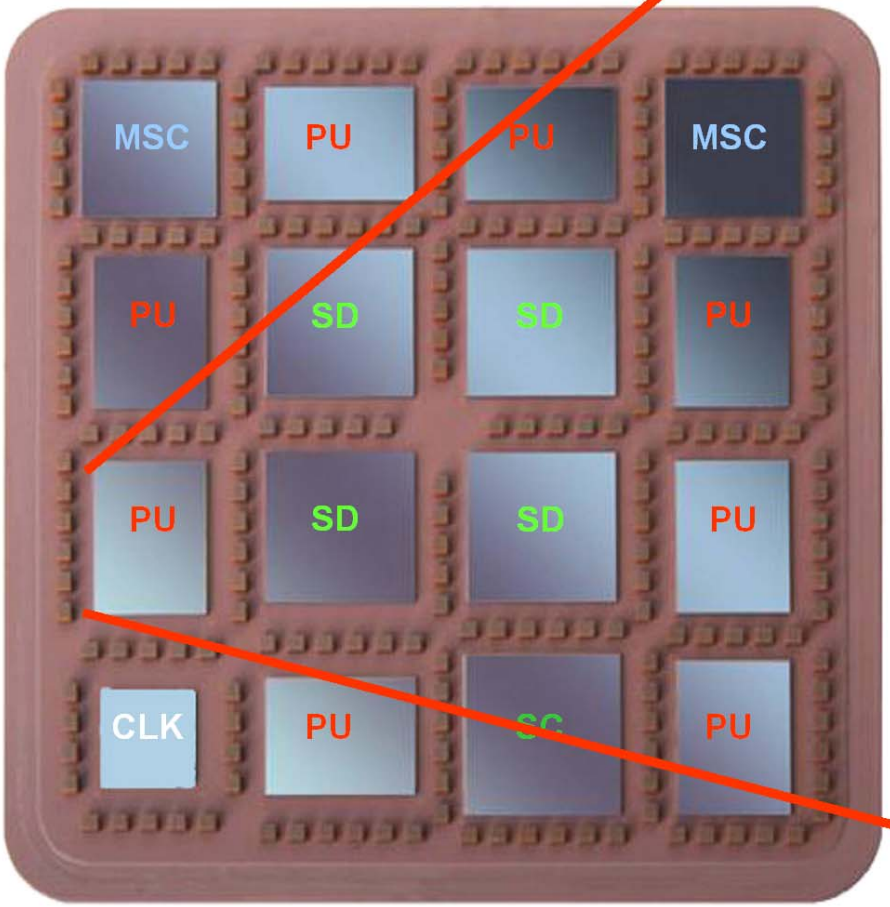
Das T5 Chip ist für High CPU Performance Server vorgesehen, während das M5 Chip für betriebswirtschaftliche Großrechner vorgesehen ist, die evtl. mit Mainframes konkurrieren sollen.

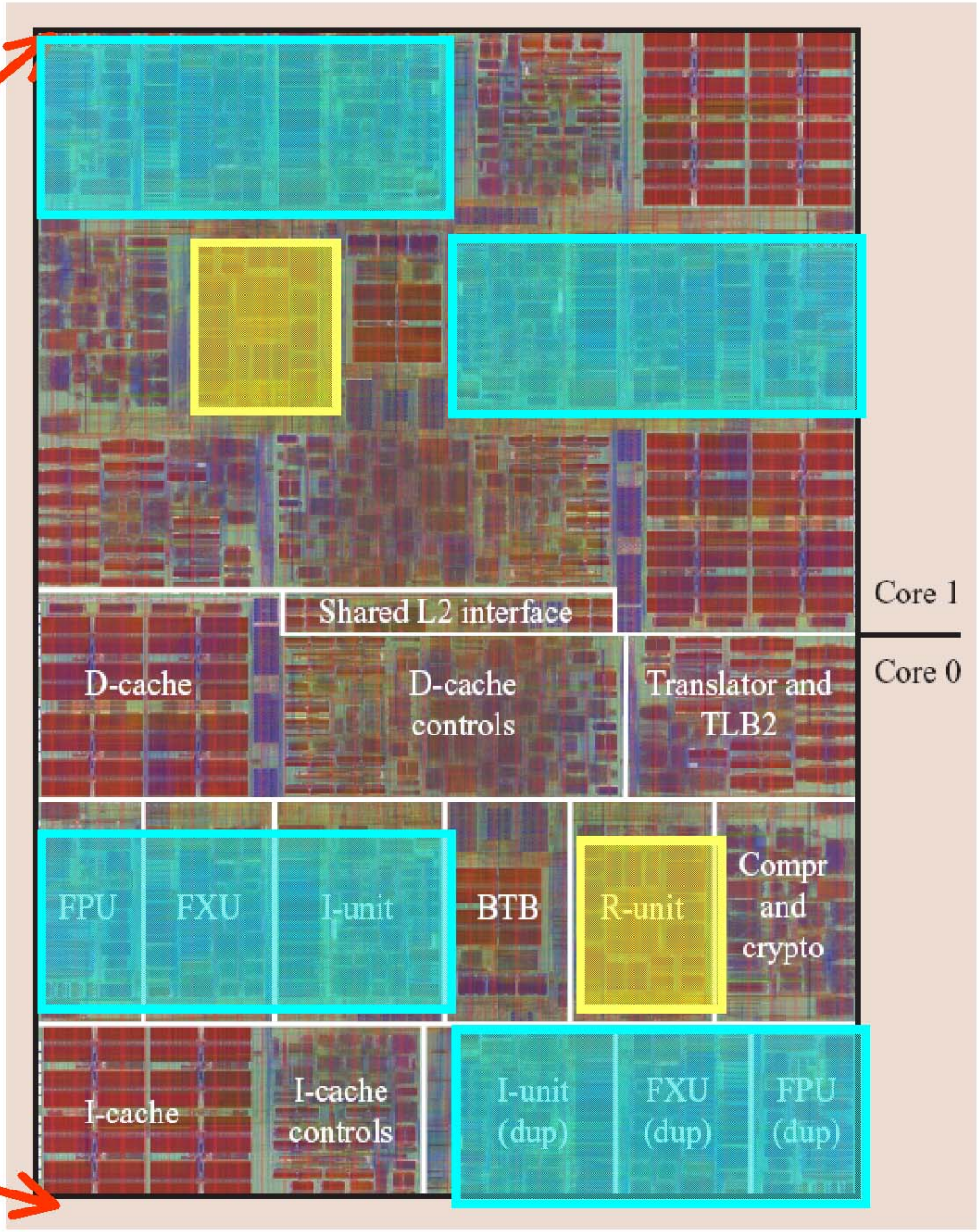
Das M5 Chip wird in dem derzeitigen Spitzenprodukt, dem Sparc M5-32 Server eingesetzt, Nachfolger des bisherigen M9000 Servers. Ein M5-32 Server hat 32 M5 CPU Chips, oder bis zu 192 CPU Cores.

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



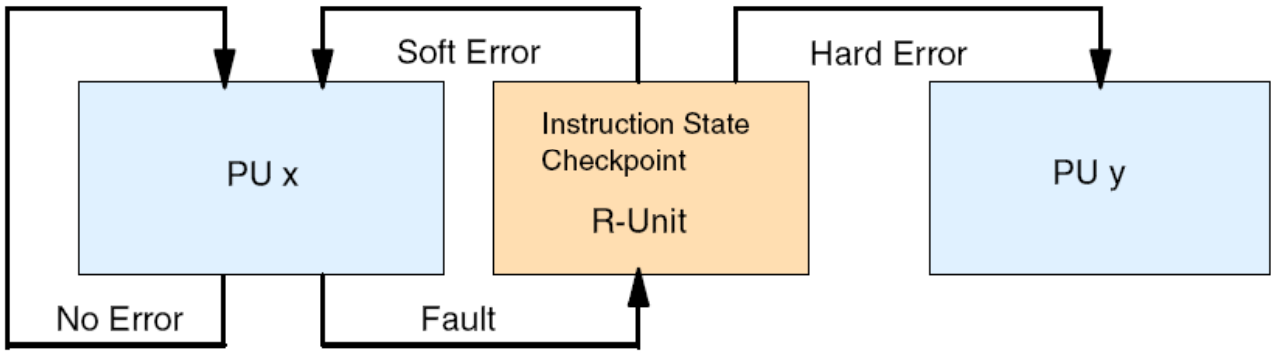


Zusätzliche Eigenschaften des System z CPU Chips

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 ein 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.

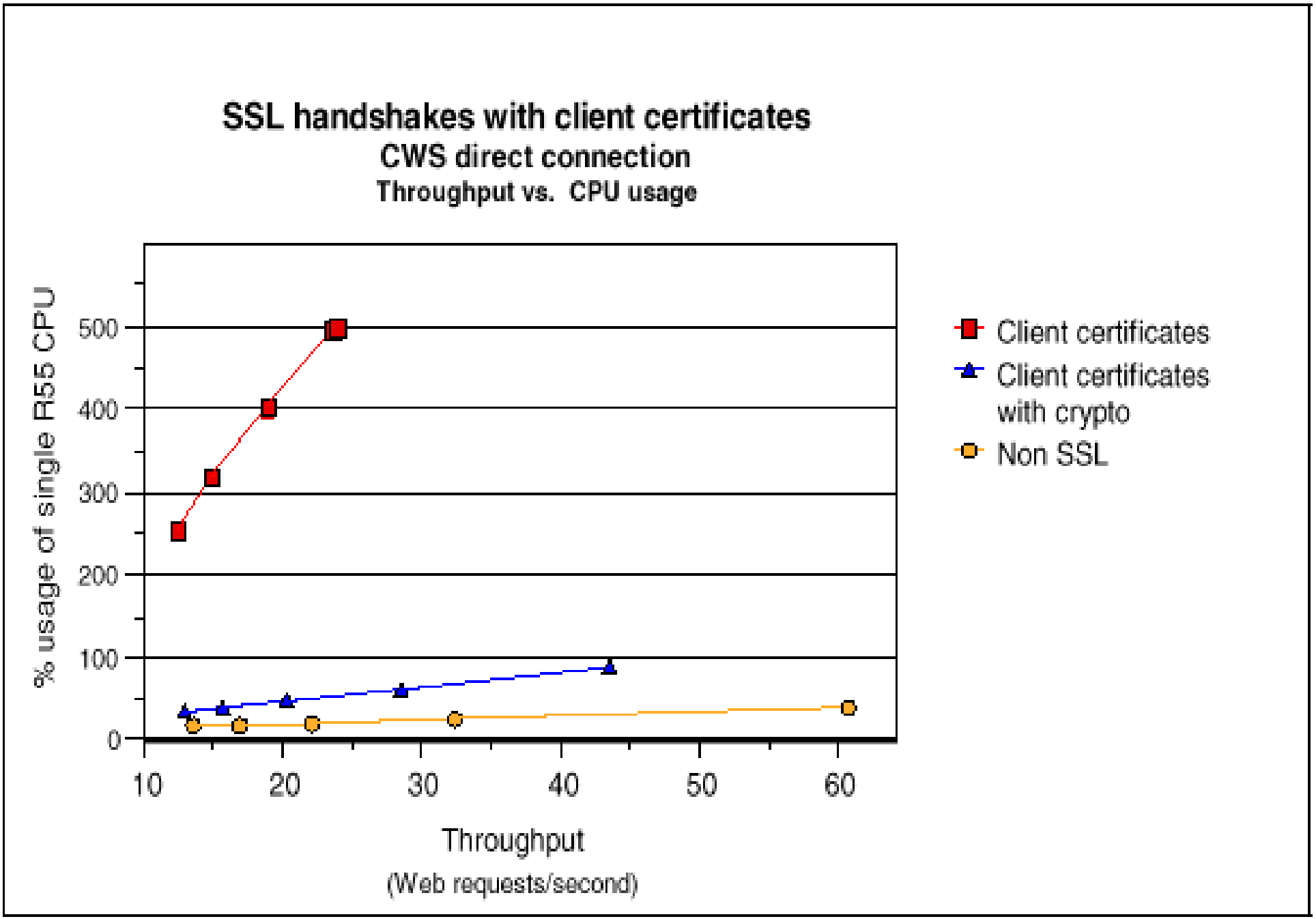
Die wichtigsten Elemente in jedem Core sind die Instruction Unit (I-unit), Fixed Point Execution Unit (FXU) und Floating Point Execution Unit (FPU). Jedes Core enthält alle drei Units in zweifacher Ausführung. Maschinenbefehle werden unabhängig und (nahezu aber nicht exakt in parallel) auf beiden Kopien der I-, FXU- and 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 Maschinenbefehlswiederholung (instruction retry). Dies geschieht unbemerkt vom Betriebssystem oder Benutzerprogramm.



Recovery Unit

Von besonderem Interesse ist in diesem Zusammenhang die „Recovery Unit“ (RU). Wenn eine Maschinen-befehlswiederholung 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 Augeblick den vollständigen Architekturstatus der CPU in ihrer R-Unit zwischenspeichert wird, wobei diese Zwischenspeicherung wiederum über Hamming Fehlerkorrekturcodes abgesichert ist.

The PU uses a process called transient recovery as an error recovery mechanism. When an error is detected, the instruction unit retries the instruction and attempts to recover the error. If the retry is not successful (that is, a permanent fault exists), a relocation process is started that restores the full capacity by moving work to another PU. Relocation under hardware control is possible because the R-unit has the full architected state in its buffer.


SSL handshakes with client certificates
CWS direct connection
Throughput vs. CPU usage
% usage of single R55 CPU
500
400
300
200
100
0
10
20
30
40
50
60
Throughput
(Web requests/second)
Client certificates
Client certificates with crypto
Non SSL

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

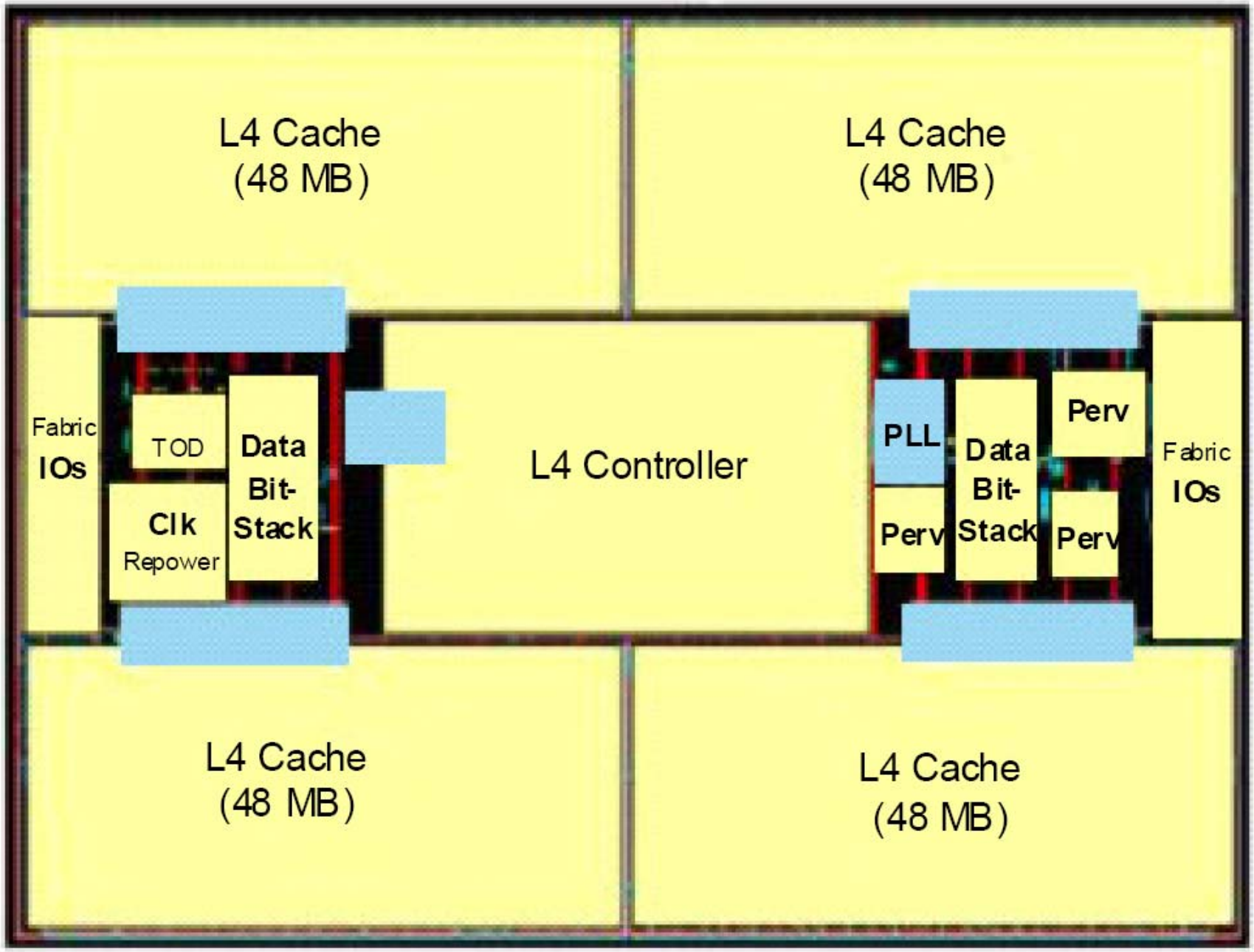
Die obige Abbildung demonstriert den Nutzen der Crypto Unit. Angenommen ist eine 6-Prozessor Einheit, die eine theoretische Leistung von 600 % verglichen mit einem einzelnen Prozessor erbringen kann.

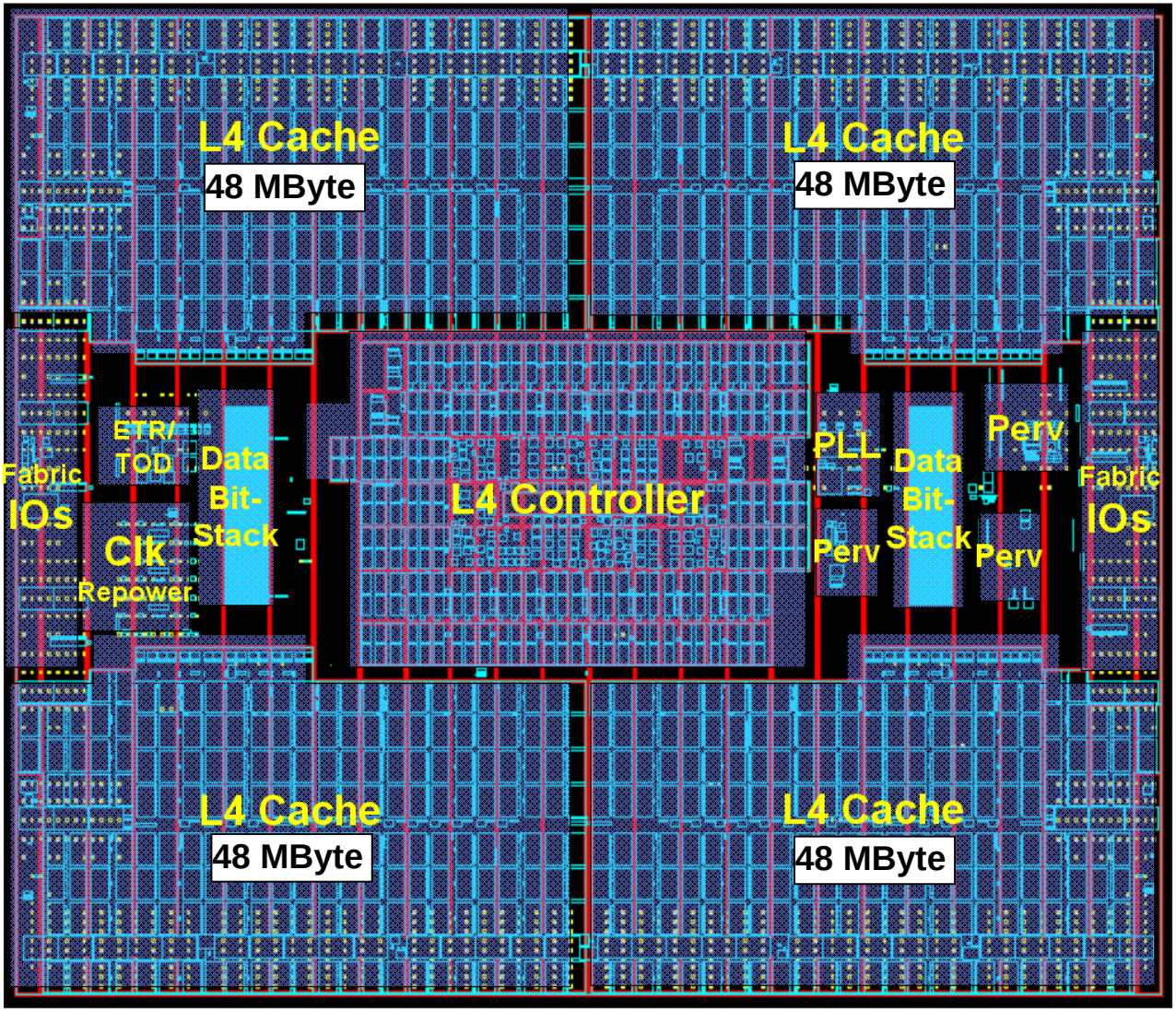
Die gelbe Kurve stellt die CPU Auslastung für eine bestimmte Art von Web Requests pro Sekunde dar. Ohne SSL beträgt die Auslastung bei 60 Transaktionen/s weniger als 50 % der Leistung einer einzigen CPU.

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

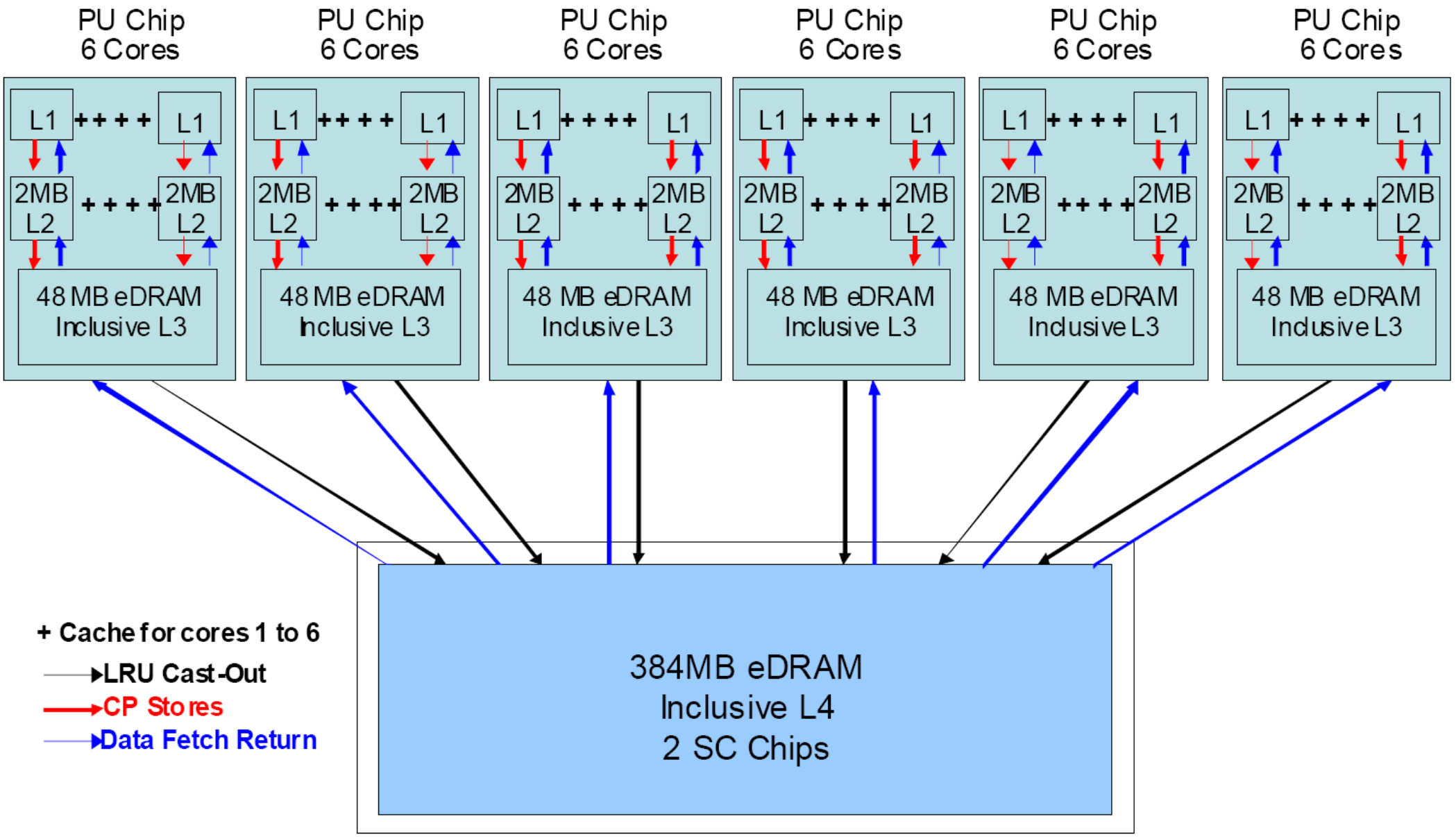
Wird die Crypto Unit für die Verschlüsselung eingesetzt, entsteht die blaue Kurve. Die CPU Auslastung ist zwar deutlich höher als ohne SSL aber deutlich besser als SSL ohne Crypto.

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


L4 Cache
(48 MB)
L4 Cache
(48 MB)
Fabric
IOs
TOD
Data
Bit-
Stack
Clk
Repower
L4 Controller
PLL
Perv
Data
Bit-
Stack
Perv
Perv
Fabric
IOs
L4 Cache
(48 MB)
L4 Cache
(48 MB)



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.



Cache Struktur eines zEC12 Multichip Modules

Cache Struktur eines zEC12 Multichip Modules

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

Auf dem CPU Chip befinden sich 4 Cores. Jeder Core hat seine eigenen privaten L1 und L2 Cache. L1 und L2 verwenden unterschiedliche Technologien und haben unterschiedliche Zugriffszeiten.

Alle 4 Cores des CPU Chips verwenden einen gemeinsam geutzten (shared) L3 Cache.

Die 6 CPU Cips 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 96 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.