Einführung in z/OS
Enterprise Computing

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

WS 2008/2009

Teil 2

System z Hardware

System z Hardware

Modern Mainframes are represented by the z9 EC and z10 EC systems. EC stands for “Enterprise Computer”. Next to the EC systems IBM also offers a BC (Business Computer) version, which is less expensive and less powerful than the EC version.

This Module describes the Hardware of z9 EC and z10 EC systems. It consists of 3 parts.

1. **Chips and chip packaging on a MultiLayer Ceramic module (MLC)**
2. **Packaging the components into a frame**
3. **Infiniband technology**
4. **Hardware Interconnection technologies**

Chip and Module Technology

In all aspects a System z mainframe hardware differs greatly from other platforms. The differences are driven by the requirements for superior reliability, availability, I/O performance, and interconnection performance.

Lets start with chips and chip packaging. PCs and most other machines are build using printed circuit boards (PCB). A PCB uses a synthetic, laminated, insulating material to which copper tracks have been added. There are three core materials used in the formation of the bare board:

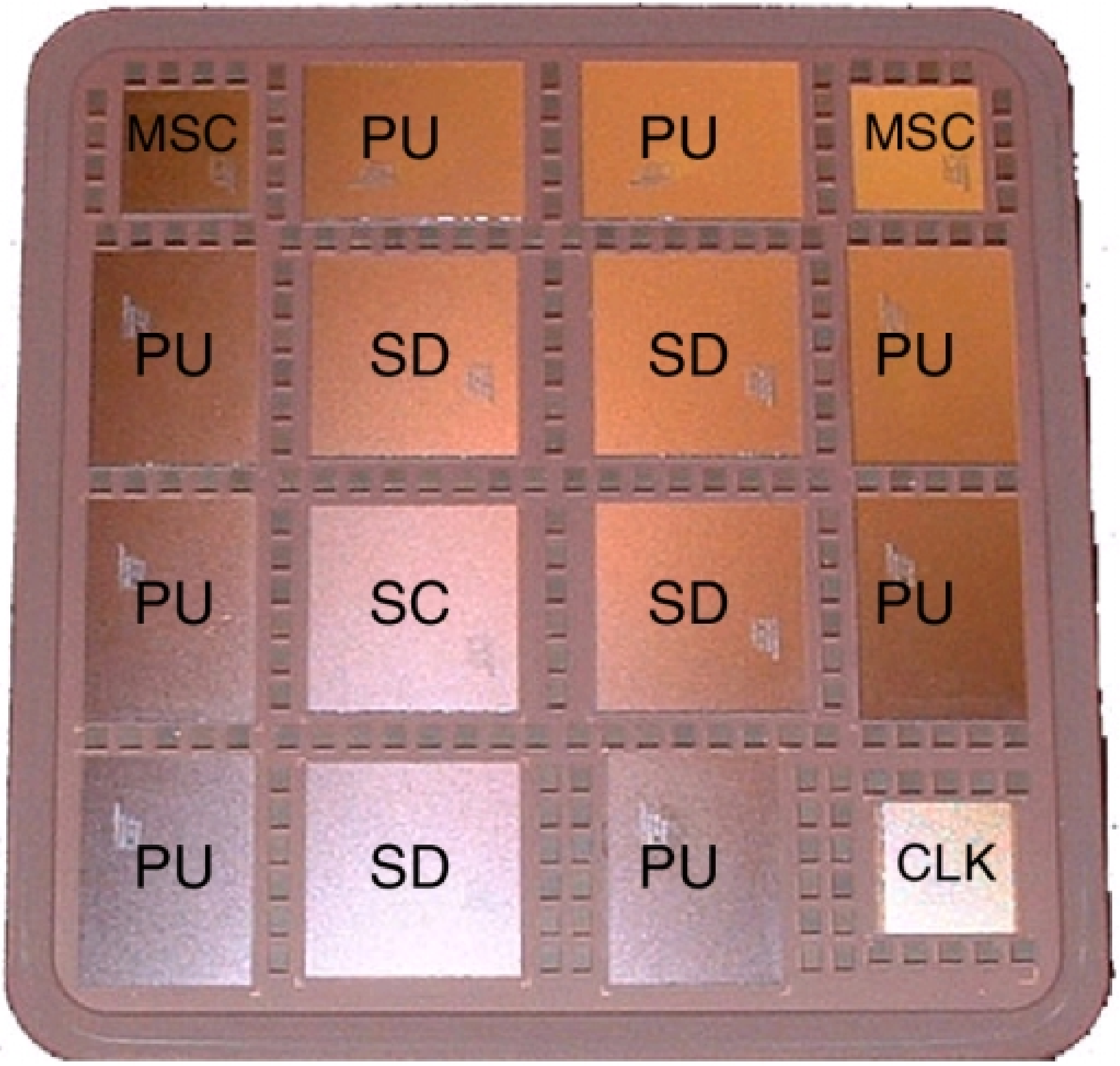
1. resin, normally epoxy
2. reinforcement, normally a woven glassfabric
3. copper foil.

System z machines use MultiLayer Ceramic (MLC) modules instead. These are manufactured by injection molding a mixture of a fine particulate such as alumina and a binder into a mold containing predesigned ridges and pins. The product is a green body layer of ceramic (green sheet) containing grooves and vias that is thereafter metallized with a conductive paste and laminated to other like layers of ceramic. Solvent extraction of the binder and sintering of the MLC forms the module.

MLC modules have been manufactured by IBM since 1979. IBM also calls them Multi-Chip Modules (MCM) and uses both terms interchangeably.

Using MLC instead of PCB as chip carriers results in tighter chip packaging, faster chip interconnect, and improved reliability. IBM claims there has not been a single electrical failure among all the MLC modules shipped since 1979.

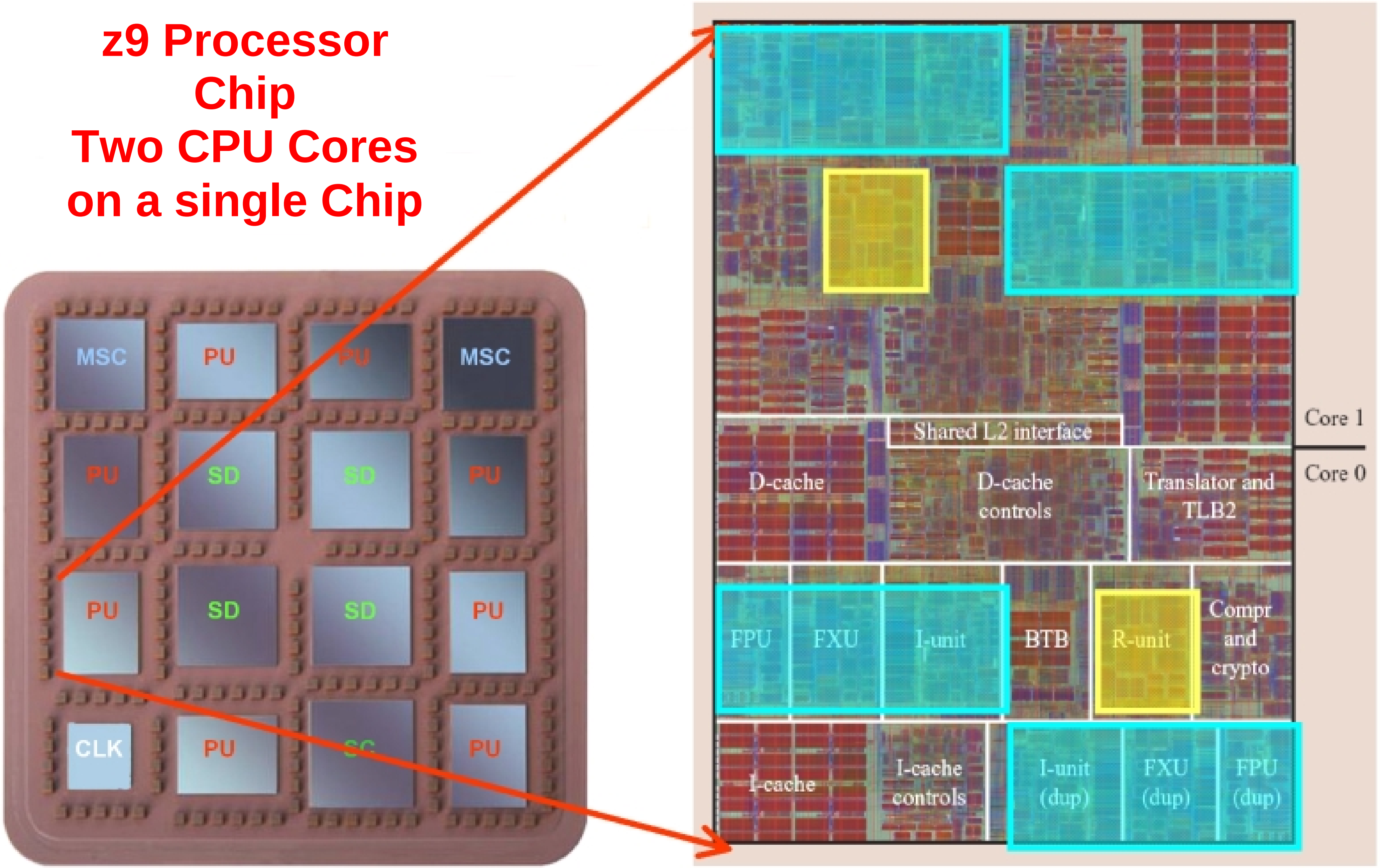
z9 MLC module

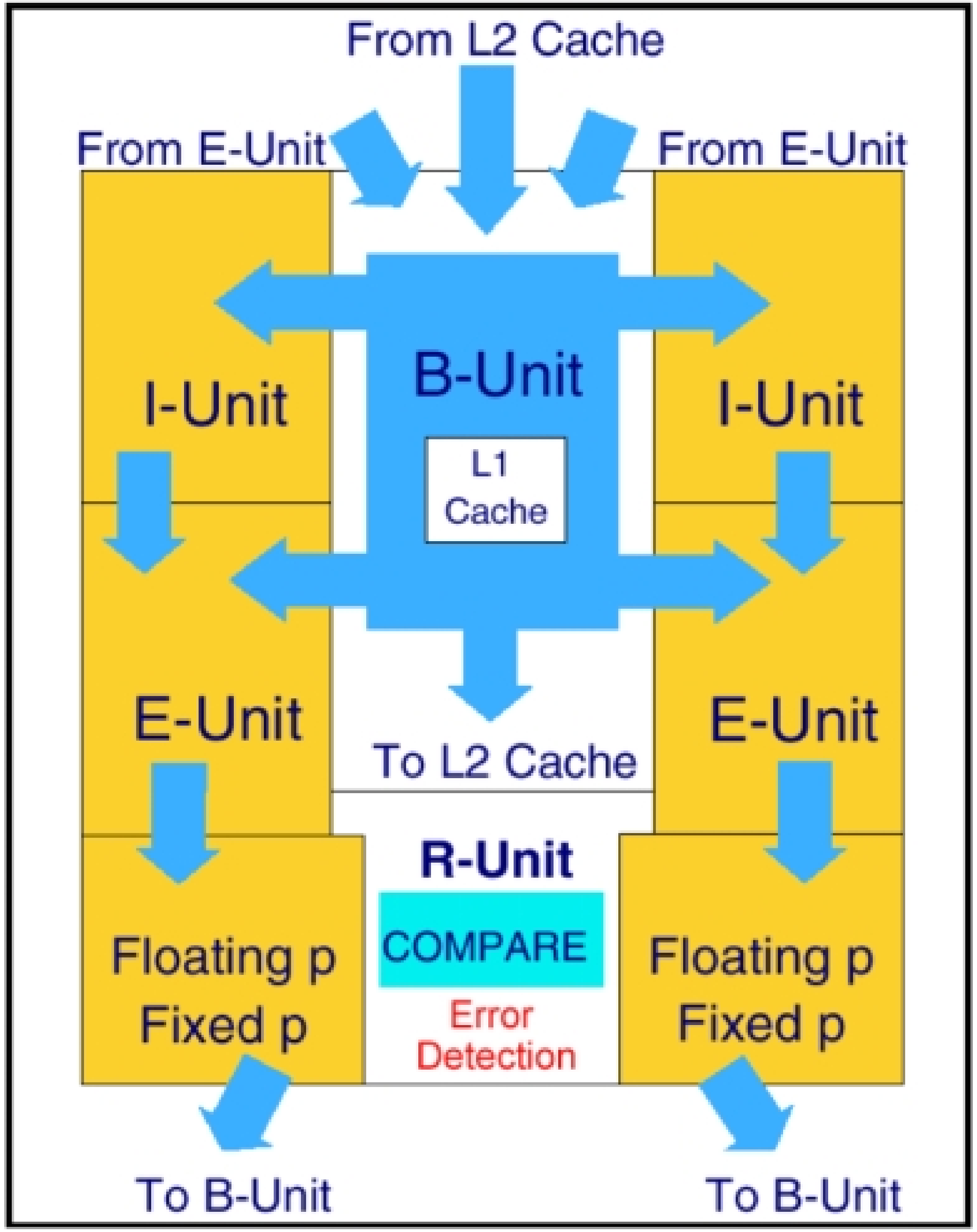


The z9 Multi Chip Moduke (MCM) is also known by the name Multilayer Ceramic Module (MLC). The module mounts:

- **8 dual core CPU chips (labeled PU) for a total of 16 CPUs,**
- **4 L2 cache chips labeled SD,**
- **1 L2 cache controller chip labeled SC,**
- **2 main store controller chips labeled MSC,**
- **and a single clock chip (CLK).**

z9 Processor Chip
Two CPU Cores on a single Chip



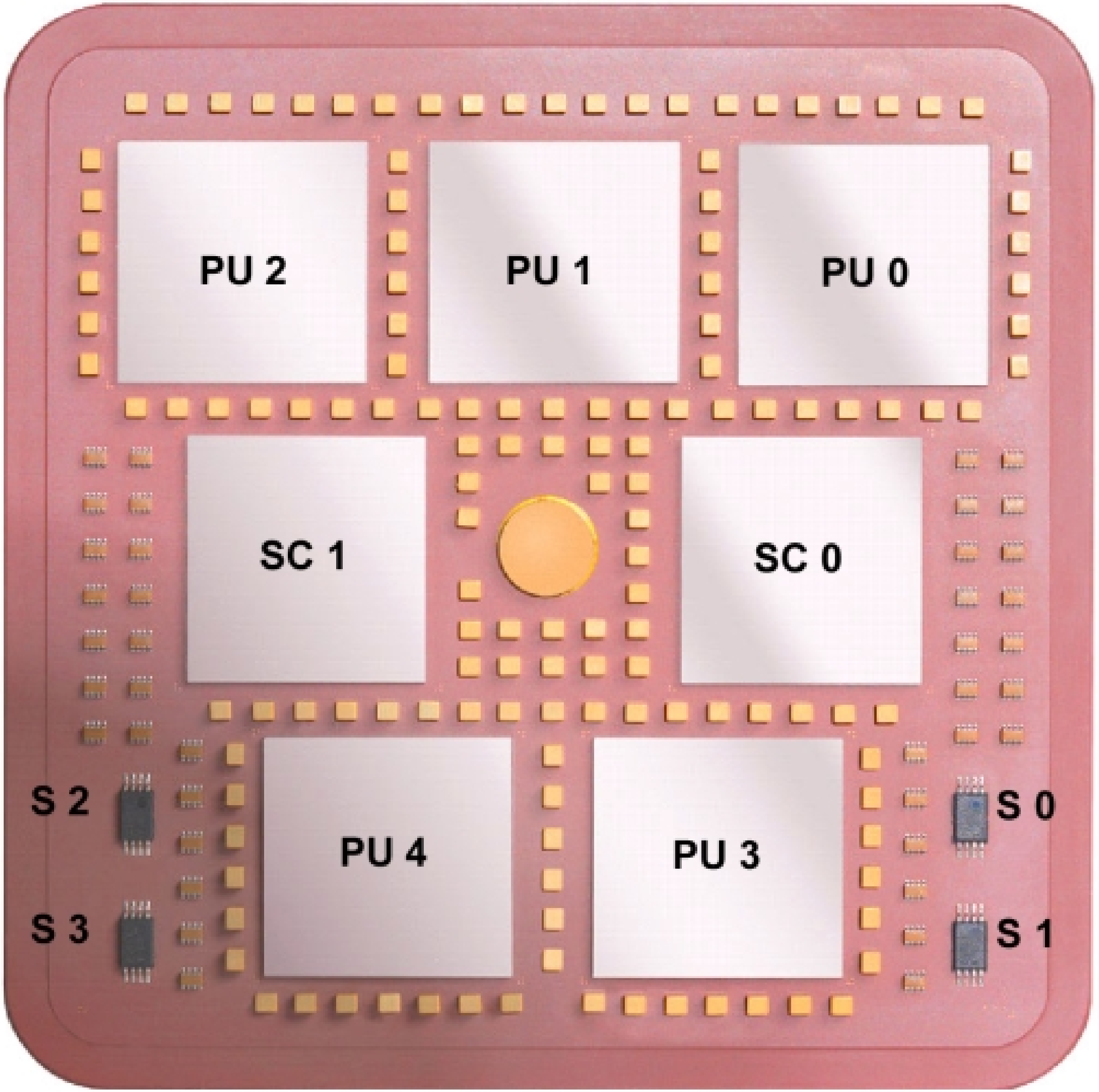


Layout of a z9 Core

Each CPU (core) duplicates (has 2 copies) of the instruction unit (I-Unit) and the execution units (E-Unit), for fixed (FXU) and for floating (FPU) operations. The instructions are executed independently and asymmetrically (nearly but not exactly in parallel) on each of the two copies of the I-, FXU- and FPU units. The results are compared after processing. In case of a mismatch, an error condition is created.

Both copies access a common L1 cache, and through it, an L2 cache which is common to all CPUs of a z9 system.

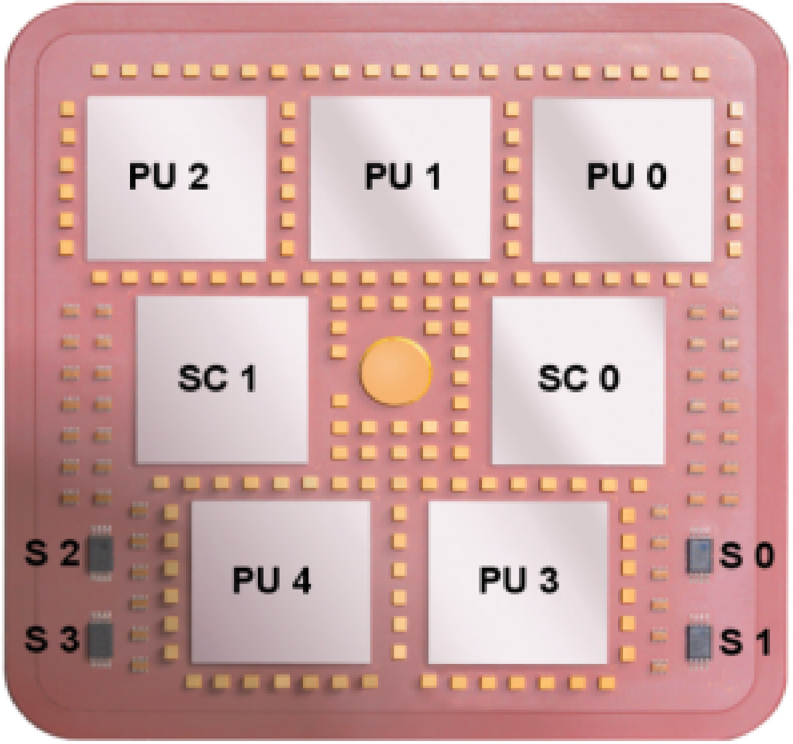
The z9 also contains error-checking circuits for data flow parity checking, address path parity checking, and L1 cache parity checking.



z10 Multi-Chip Module

The z10 Multi-Chip Module looks simpler than the z9 Multi-Chip Module. It contains five Quad core processor chips (labeled PU) for a total of 20 CPUs per MCM, as well as two Storage Control Chips (labeled SC) storing 24 MByte each, for a total of 48 MByte L2 cache for each MCM.

- **96mm x 96mm MCM**
  - 103 Glass Ceramic layers
  - 7 chip sites
  - 7356 LGA connections
  - 17 and 20 way MCMs



- **CMOS 11s chip Technology**
  - PU, SC, S chips, 65 nm
  - 5 PU chips/MCM – Each up to 4 cores
    - One memory control (MC) per PU chip
    - 21.97 mm x 21.17 mm
    - 994 million transistors/chip
    - L1 cache/PU
      - 64 KB I-cache
      - 128 KB D-cache
    - L1.5 cache/PU
      - 3 MB
    - 4.4 GHz

    2 Storage Control (SC) chip
    - 21.11 mm x 21.71 mm
    - 1.6 billion transistors/chip
    - L2 Cache 24 MB per SC chip (48 MB/Book)
    - L2 access to/from other MCMs
  - 4 SEEPROM (S) chips
    - 2 x active and 2 x redundant
    - Product data for MCM, chips and other engineering information
  - Clock Functions – distributed across PU and SC chips
    - Master Time-of-Day (TOD) and 9037 (ETR) functions are on the SC

z10 Multi-Chip Module (MCM)

The z10 EC MCM chips use CMOS 11S chip technology, based on ten-layer Copper Interconnections and Silicon-On Insulator technologies. The chip lithography line width is 0.065 micron (65 nm). The chip contains close to 1 billion transistors in a 450 mm^2 die.

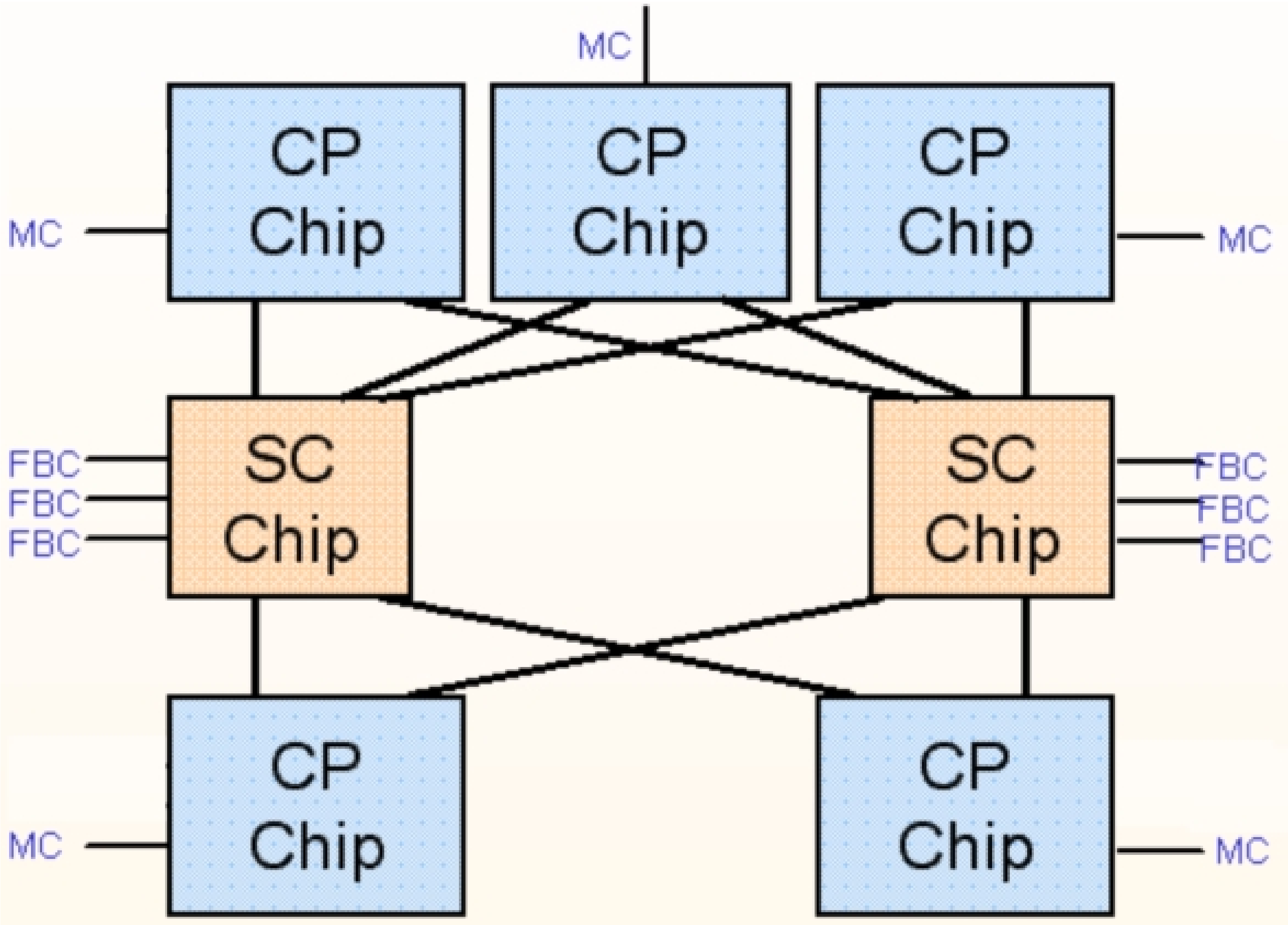
The z10 server has five processor (PU) chips per MCM and each PU chip has up to four PUs (cores).

The MCM also has two Storage Control (SC) chips. Each SC chip packs 24MB of SRAM cache, interface logic for 20 cores, and SMP fabric logic into 450 mm^2. The two SC chips are configured to provide a single 48MB cache shared by all 20 cores on the module.

There are four SEEPROM (S) chips, of which two are active and two are redundant, that contain product data for the MCM, chips and other engineering information.

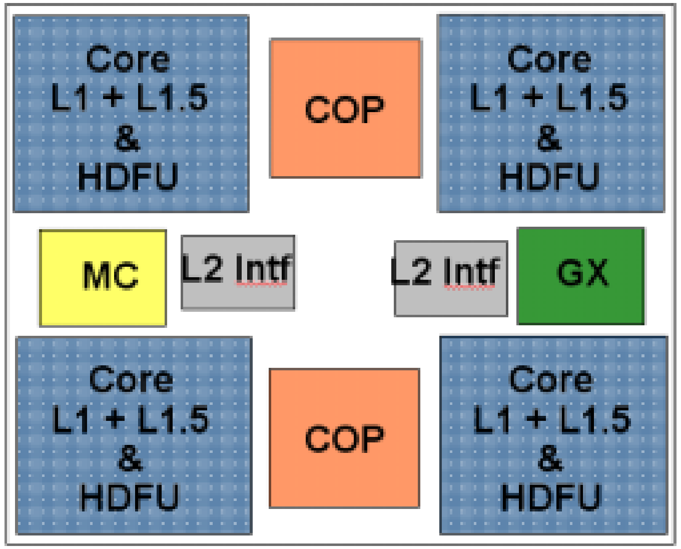
894 instructions (668 implemented in hardware)

z10 adds 50 instructions to improve compiled code efficiency



Each of the 5 Processor chips on an MCM connects directly to the two L2 cache chips (SC). The memory controller (MC) on each processor chip connects directly to main memory. Allother outside communication, especially I/O, go via the FCB busses of the L2 chips.

Enterprise quad-core z10 chip



Core L1 + L1.5 & HDFU

COP

Core L1 + L1.5 & HDFU

MC

L2 Intf

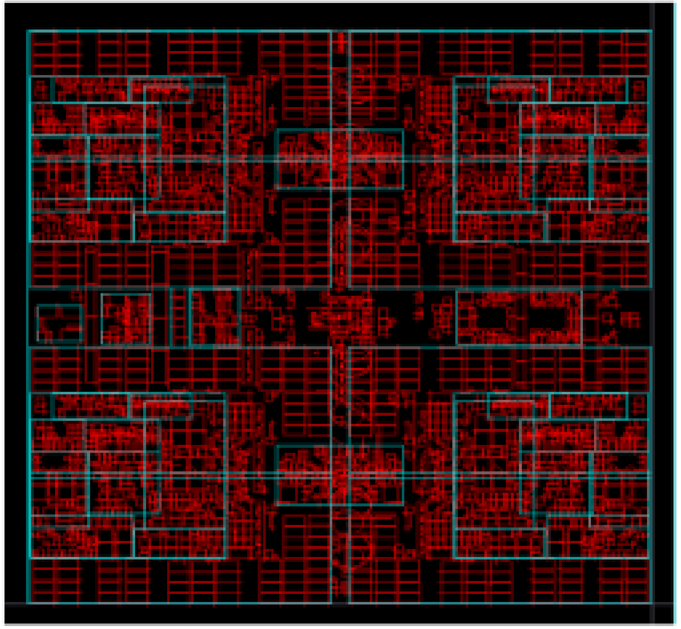
L2 Intf

GX

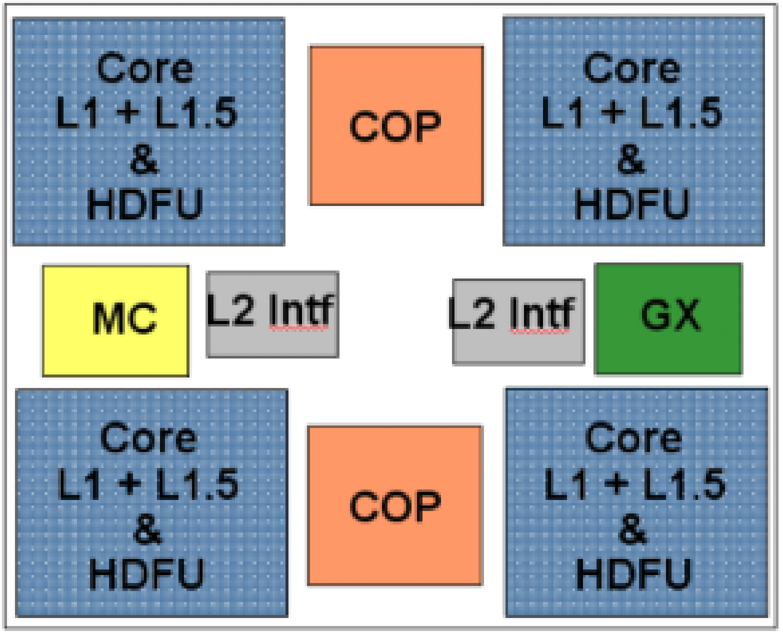
Core L1 + L1.5 & HDFU

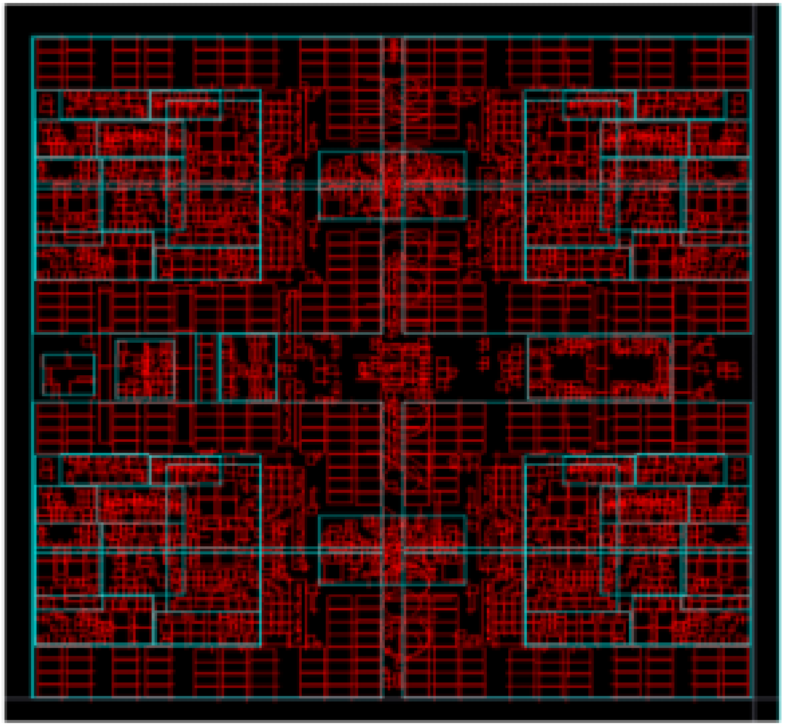
COP

Core L1 + L1.5 & HDFU



- **Four cores (PUs)**
  - 4.4 GHz
  - L1 cache/PU, 64 KB I-cache, 128 KB D-cache
  - 3MB L1.5 cache/PU
  - Each core with its own Hardware Decimal Floating Point Unit (HDFU)
- **Two Co-processors (COP)**
  - **Accelerator engines**
    - Data compression
    - Cryptographic functions
  - Includes 16KB cache
  - Shared by two cores
- **L2 Cache interface**
  - Shared by all four cores
- **I/O Bus Controller (GX)**
  - Interface to Host Channel (HCA)
  - Compatible with z9 MBA
- **Memory Controller (MC)**
  - Interface to controller on memory DIMMs





z10 Quad-core Processor Chip

Shown to the left are the major building blocks of the z10 chip (top) and the actual chip layout (bottom).

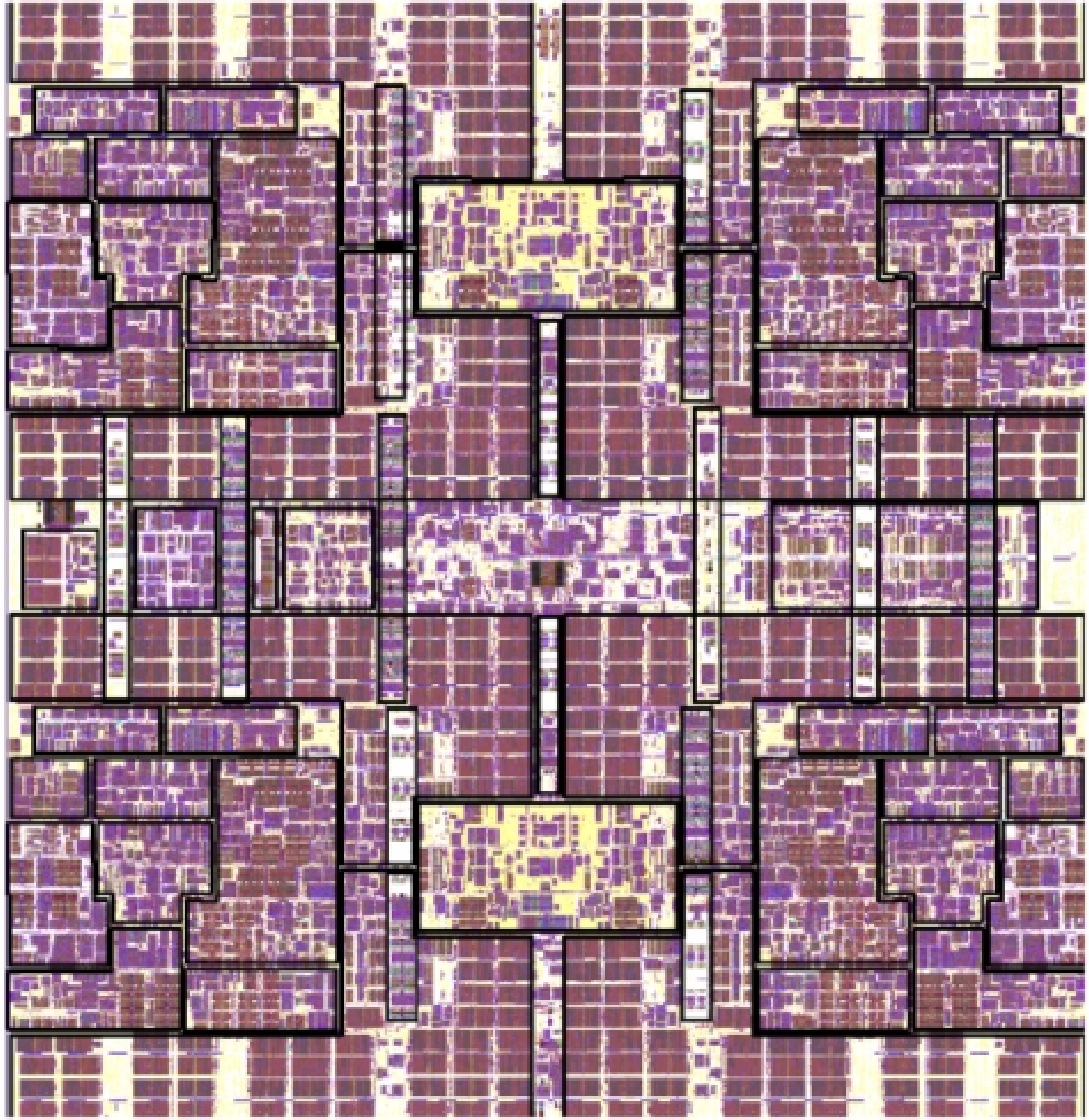
Each Chip contains 4 CPUs (cores).

The CPUs operate at 4.4 GHz. Each CPU has its own L1 cache, consisting of a 64 KByte I-cache and a 128 KByte D-cache, and an additional 3 MByte l1.5 cache.

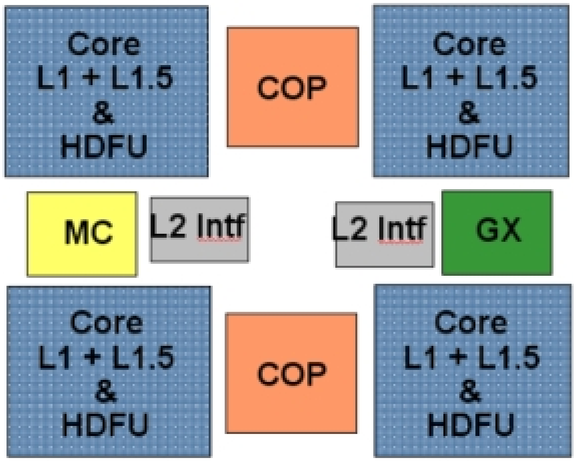
All CPUs contains the new hardware decimal floating point function (HDFU).

Each Chip has 2 Co-processors, serving as acceletators for data compression and cryptographic functions. Each Co-Processor is shared by two cores.

All 4 cores share a common Memory Controller (MC), a common L2 cache interface, and a common I/O Bus Controller (GX), which in turn interfaces to the Host Channel (HCA). The HCA is the replacement to the z9 Memory Bus adapter (MBA).



z10 Quadcore Processor chip



MC	**Memory Controller**
COP	**Crypto and Compression Coprocessors**
GX	**Interface to I/O**
HDFU	**Hardware Decimal Floating Point Unit**

Charles F. Webb: IBM z10: The Next-Generation
Mainframe Microprocessor
IEEE Micro. March 2008

IBM z10 EC Hardware Decimal Floating Point Unit (HDFU)

Meets requirements of business and human-centric applications

- **Performance, Precision, Function**
- **Avoids rounding and other problems with binary/decimal conversions**
- **Improved numeric functionality over legacy Binary Coded Decimal (BCD) operations**
- **Much of commercial computing is dominated by decimal data and decimal operations**

Growing industry support for DFP standardization

- **Java BigDecimal, C#, XML, XL C/C++, DB2 9 , Enterprise PL/1, Assembler Endorsed by key software vendors including Microsoft and SAP**
- **Open standard definition led by IBM**

The z10 processor chip also supports IBM hexadecimal and IEEE 754 floating point operations

Decimal floating-point

Binary floating-point numbers can only approximate common decimal numbers. The value 0.1, for example, would need an infinitely recurring binary fraction. This causes subtle problems; for instance, consider the calculation of adding a 5 percent sales tax to a $0.70 telephone call, rounded to the nearest cent. Using binary floating-point numbers, the result of 0.70 x 1.05 before rounding is just less than the correct result (0.735) and hence would be rounded down to $0.73. With decimal floating-point numbers, the intermediate result would be exactly 0.735, which would then round up correctly to $0.74.

For this and other reasons, binary floating-point computation cannot be used safely for financial calculations, or indeed for any calculations where the results achieved are required to match those which might be calculated by hand.

IBM’s mainframe processors have always had binary-coded decimal (BCD) instructions, but these are hard to use for anything other than fixed-point calculations. However, in recent years, decimal calculations are more common (interest rates change daily, for instance) and more complicated (more analysis is done on currency transactions, for example).

The decimal floating-point unit in the z10 processor allows all calculations, including mathematical and statistical, to be done in the new decimal formats, so no conversions to binary are needed, and exact decimal results are given where expected. Conversion to and from BCD or strings is easy too.

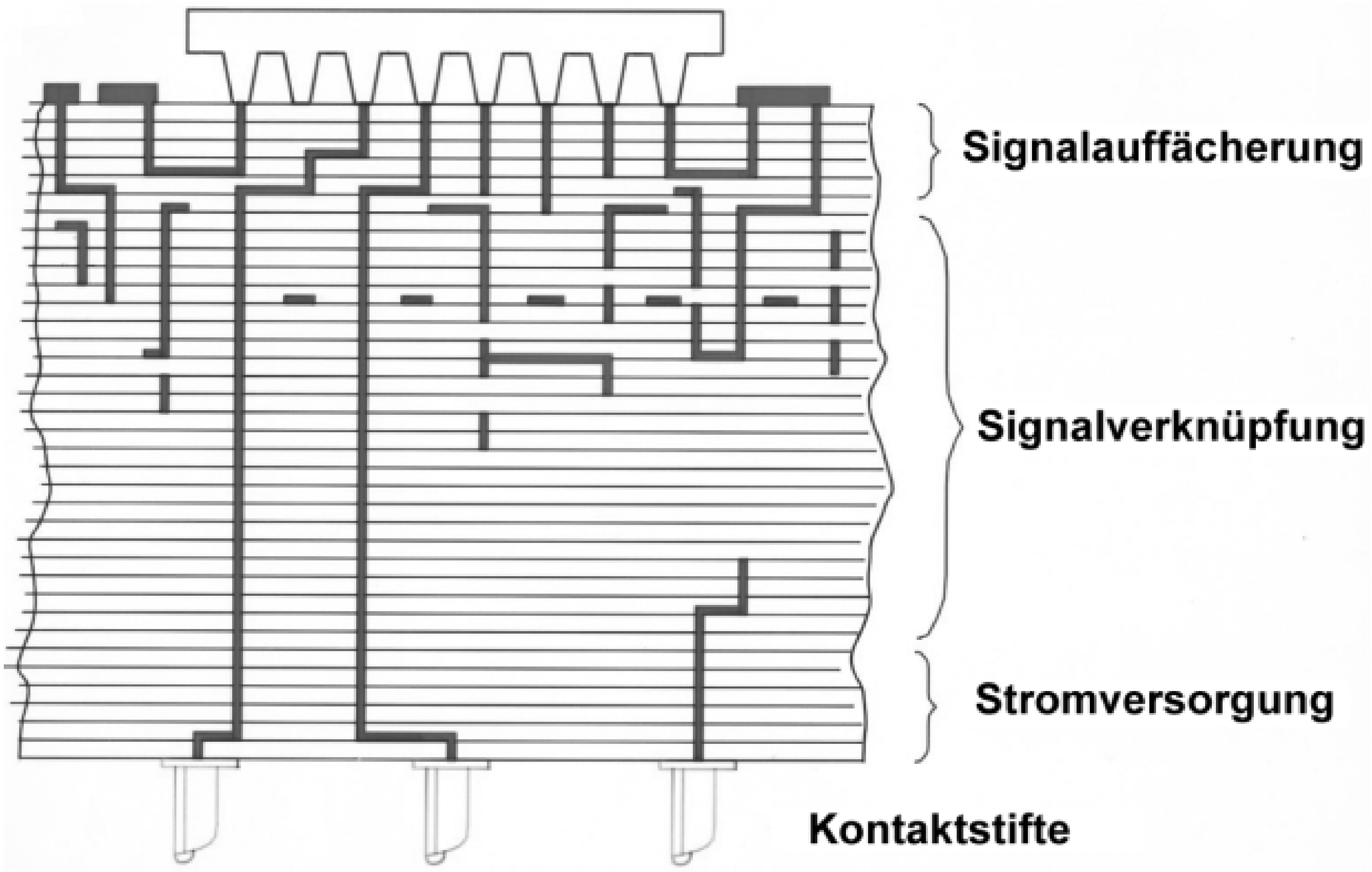
For more examples and information, see http://www2.hursley.ibm.com/decimal.

Multilayer Ceramic Module (MLC).

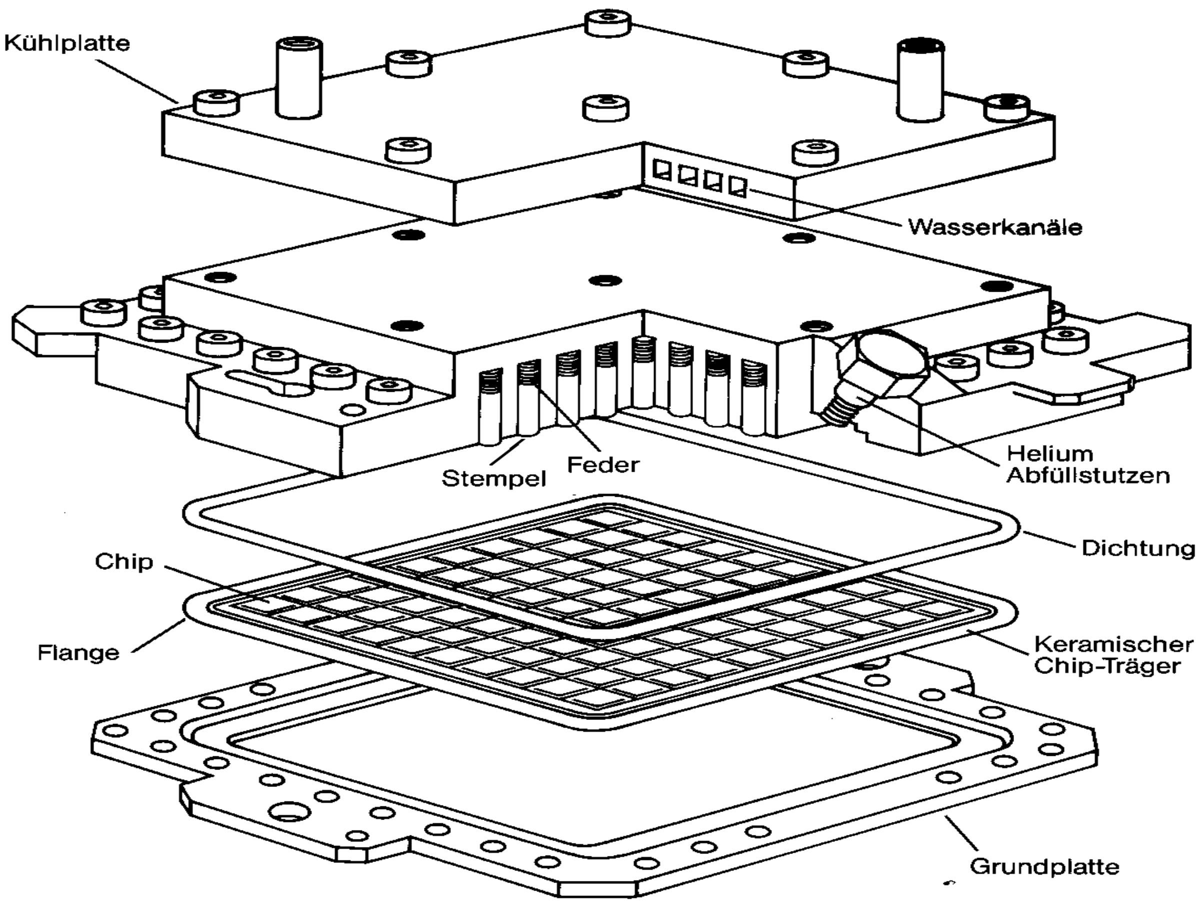
Multilayer Ceramic Module (MLC) technology is used to implement a Multi Chip Module (MCM). There are in principle other technologies to implement a MCM, but MLC is the only one currently used for complex chip carriers. IBM uses both terms interchangeably.

Some of the earlier Pentium modeles produced by Intel used MLC technology. The (then) high speed Pentium Pro consisted of two chips, a CPU chip and a separate cache chip, mounted on a single MLC substrate. This provided superior performance compared to mounting both chips on seperate modules and interconnecting them via prited circuit board technology. Intel has since discontinued this approach.

IBM has manufactured MLC modules since the early 1980s. It is claimed that of all the MLC modules delivered since then not a single one has failed due to electrical reasons.



Querschnitt durch ein MCM. Das z900-MCM benutzt einen Glas-Keramik-Träger mit 101 Glas-Keramik- und 6 Polyimide Dünnfilm-Verdrahtungslagen. In dem 127 x 127 mm-Modul sind insgesamt 1 km Draht untergebracht. Innerhalb der verschiedenen Schichten entstehen komplexe Verdrahtungsmuster. Die senkrechten Verbindungen zwischen den Schichten bestehen aus leitenden Bohrungen, die wiederum innerhalb einer Schicht in horizontalen Leiterbahnen weitergeführt werden und an einer Bohrung zu einer darunter- oder darüberliegenden Schicht enden usw. Mehr als 4000 Kontaktstifte sorgen für die Verbindung des MCM mit dem darunter liegendem Board.

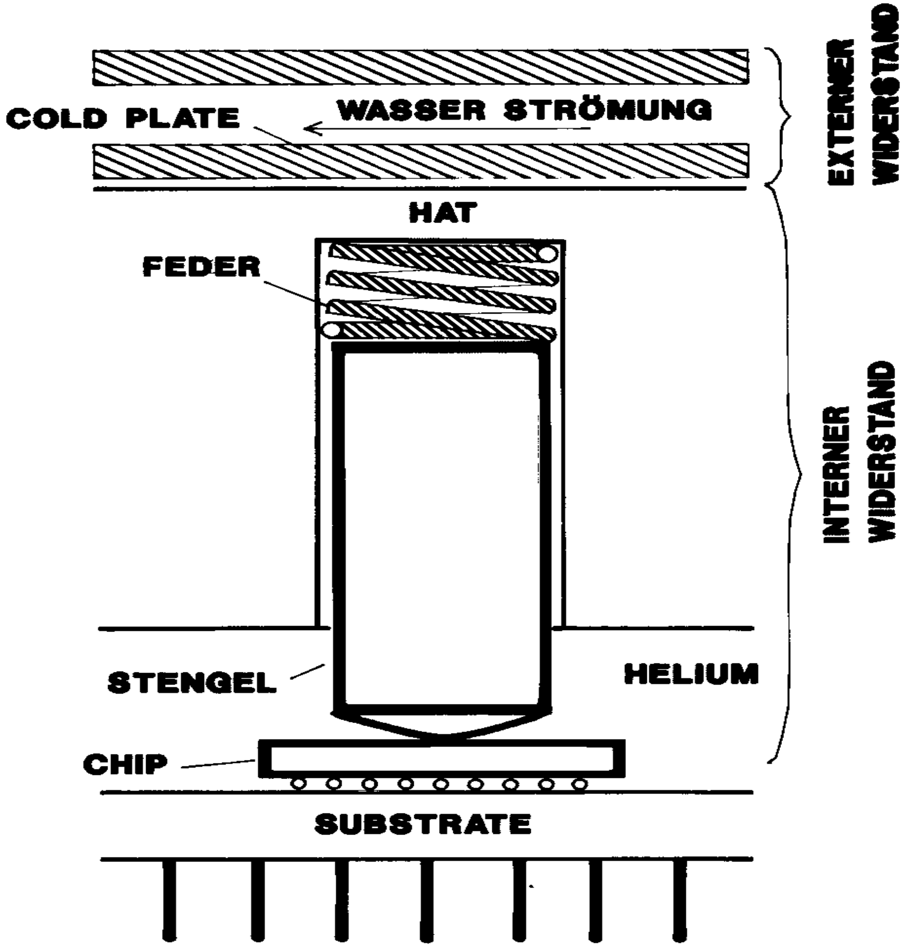


Zu Kühlungszwecken sitzt auf jedem Chip ein Aluminium-Stempel, der die Verlustwärme ableitet. Eine Spiralfeder sorgt für einen guten Kontakt des Stempels mit der Chip-Oberfläche. Die Aluminium-Stempel werden in einer Bohrung geführt und geben die Verlustwärme an die Umgebungsplatte weiter. Eine darüber liegende Kühlplatte wird mit Wasser gekühlt. Hierfür existiert ein geschlossener Kreislauf, in dem das Wasser seine Wärmeenergie an einen Radiator weitergibt, ähnlich wie in einem Automobil. Bei den z9 und z10 Rechnern befindet sich der Radiator im Rechnergehäuse.

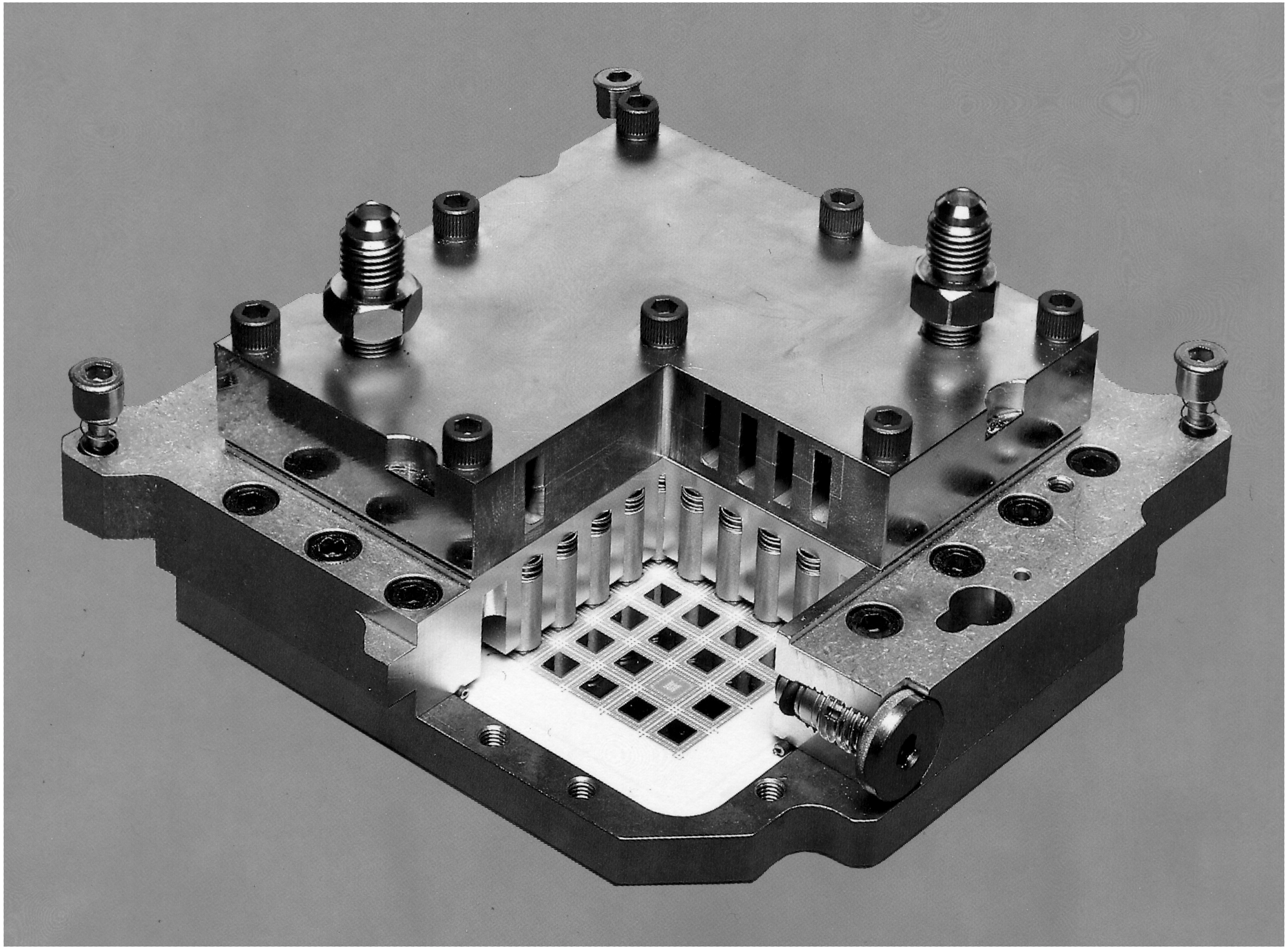
Bei dem hier vorgestellten Verfahren werden alle Chips selbst mit Luft gekühlt. Es sind in der Vergangenheit viele Versuche unternommen worden, Chips direkt mit einer Flüssigkeit zu kühlen. Diese Methode hat sich in der Praxis jedoch nicht bewährt.

Aufbau eines „Thermal Conduction Module“

WÄRMEABLEITUNG IM TCM

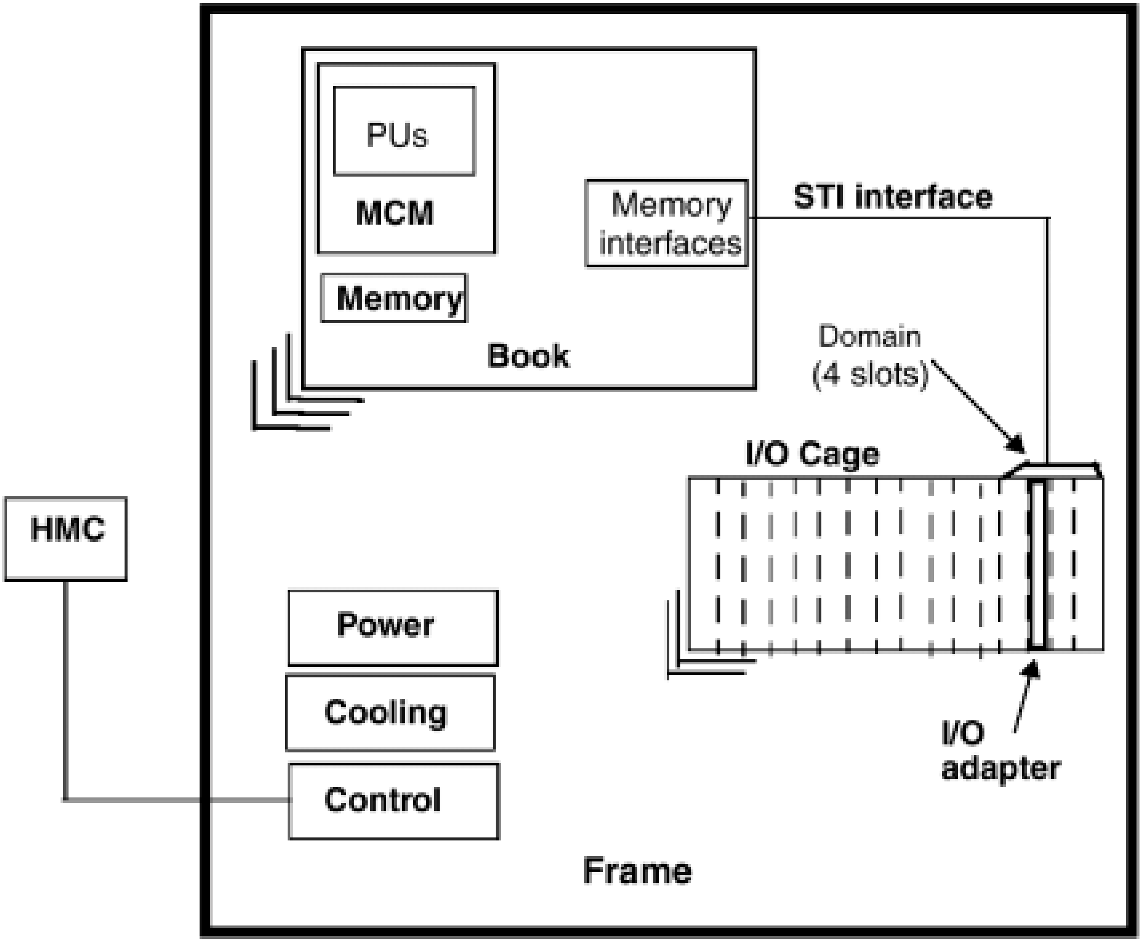


TCM Wärmeübergang



Thermal Conduction Module (TCM)
Multichip Module (MCM) Package

. A TCM or may mount as few as 4 and as many as 121 chips.



The diagram shows is the internal structure of a z9 or z10 mainframe system. CPUs and L2 cache are packaged on a Multi Chip Module (MCM). A single MCM, together with its memory (main store) and its memory interface to the outside world, form a unit called a “Book”.

The memory interface connects to I/O adapter cards housed in an “I/O cage”. Connection is performed via an I/O cable, which implements the STI protocol (z9) or Infiniband protocol. Consider these as the equivalents to the PCI bus on your PC. The I/O cage houses a number of I/O adapter cards for connecting disks, tapes, and other I/O devices.

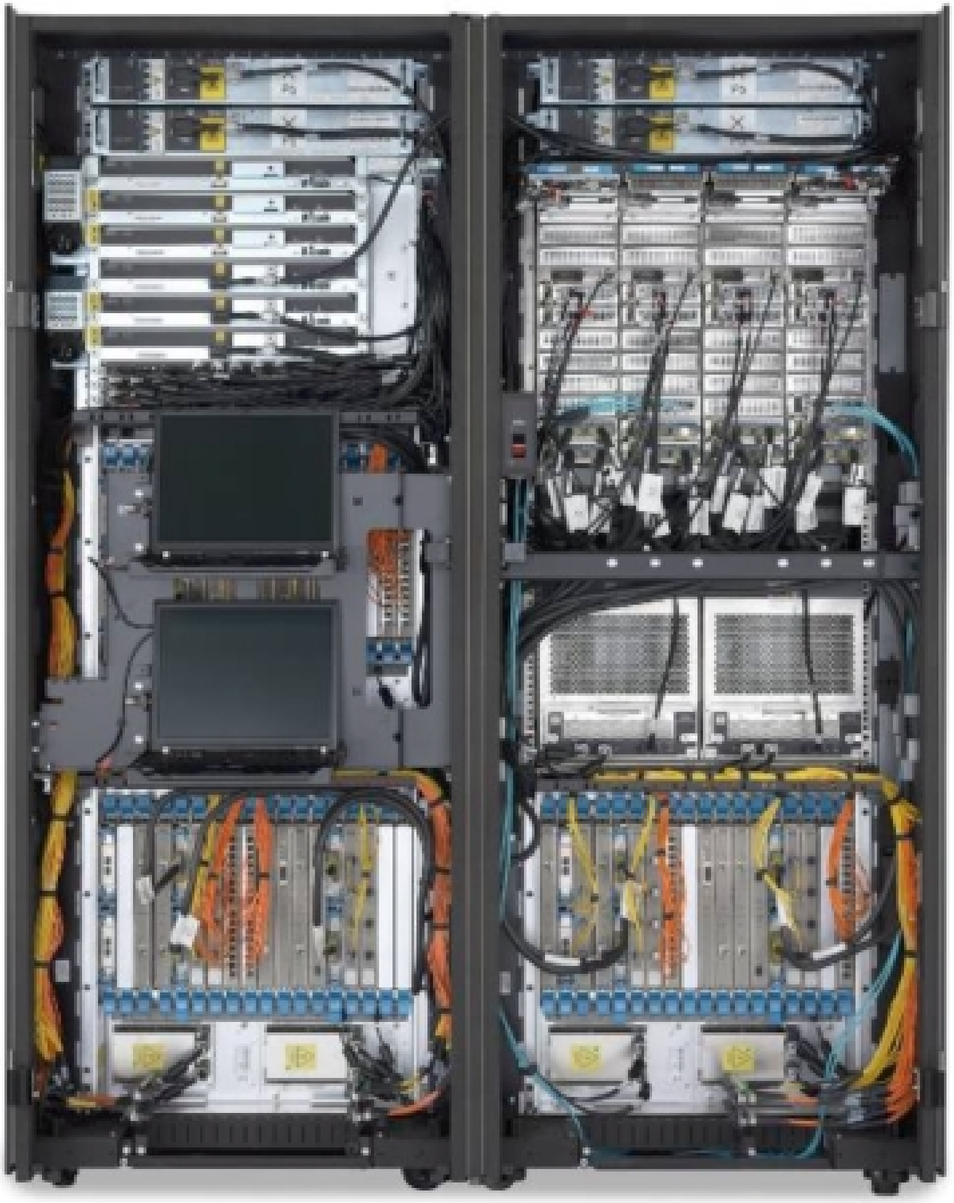
Power supplies, cooling, and controls for connection to a “Hardware Management Console” (HMC) are the remaining compnents of a mainframe system.

Shown is the external view of a z10 EC Mainframe.

The system consists of two frames labeled A frame and Z frame. The frames are 40 EIA units tall; the two EIA units may be removed if necessary. The specific dimensions are:
_ 158 cm deep (62 inches), including the covers
_ 194 cm high (76 inches), including the casters
_ 154 cm wide (61 inches), with both frames together

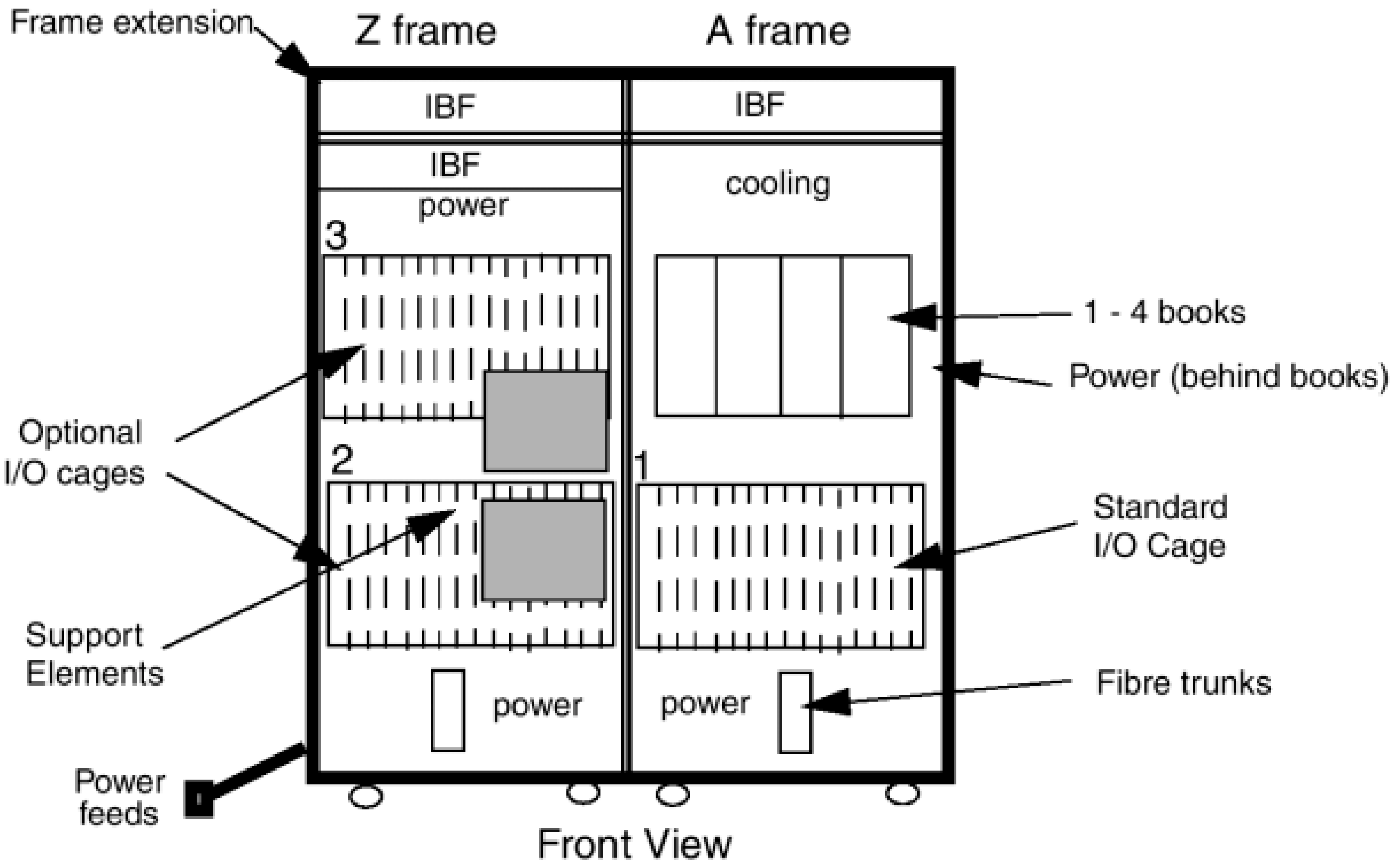
Weight (fully equipped):
_ 1025 kg (2257 pounds) for the A frame (with the internal battery feature)
_ 999 kg (2197 pounds) for the Z frame (with the internal battery feature)

Exterior of a z10 EC system



Interior of a z10 EC system

Z frame A frame

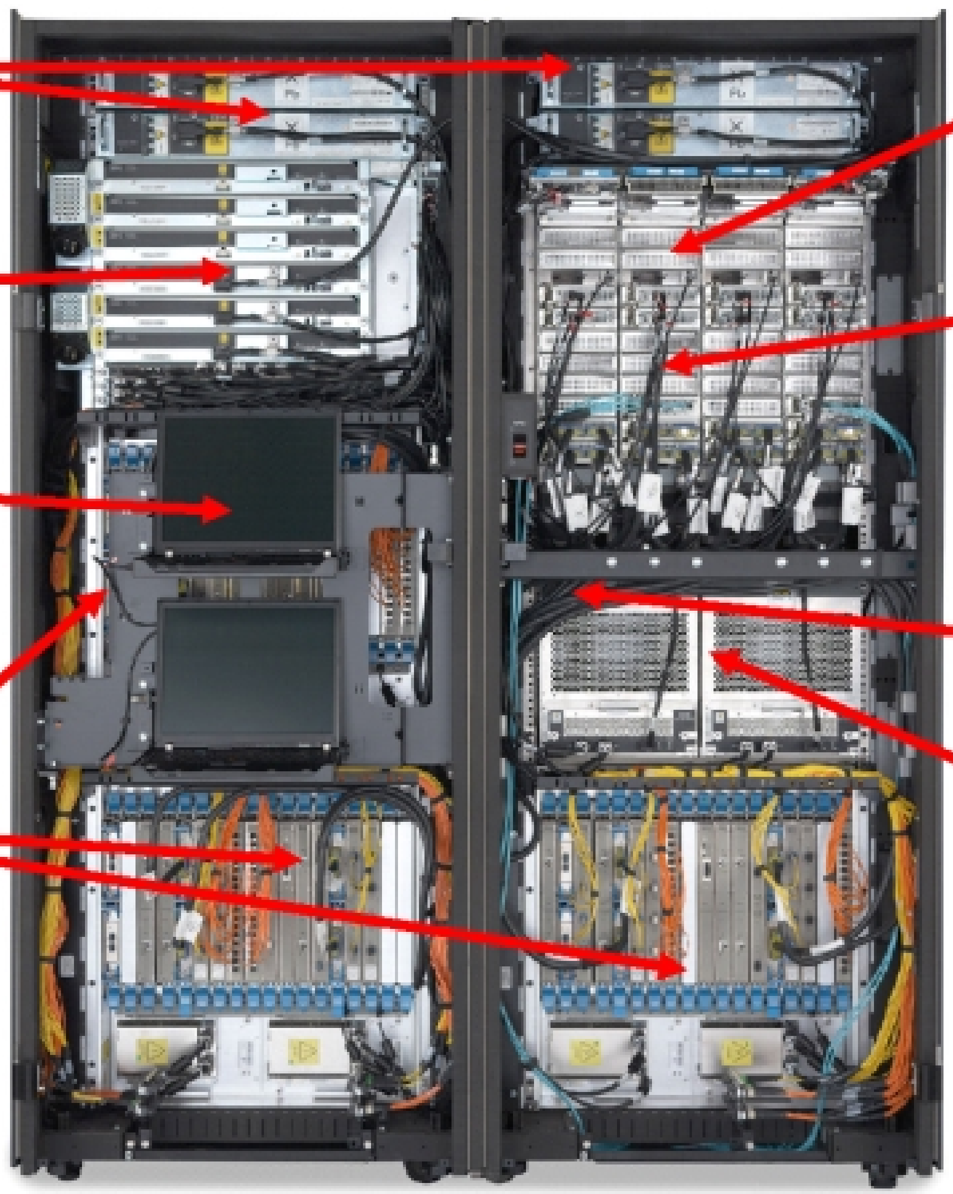


A z9 or z10 system has a minimum of 1 book and a maximum of 4 books. A minimum z9 system has a single A frame and a single I/O cage. The Z frame houses 2 additional I/O cages and a "Support Element" to control the system. There is actually a second Support Element just in case the first one fails. A regular Thinkpad Laptop implements each Support Element.

z10 EC internal front view

Z frame

A frame



Internal Batteries (optional)

Power Supplies

2 x Support Elements

3x I/O cages

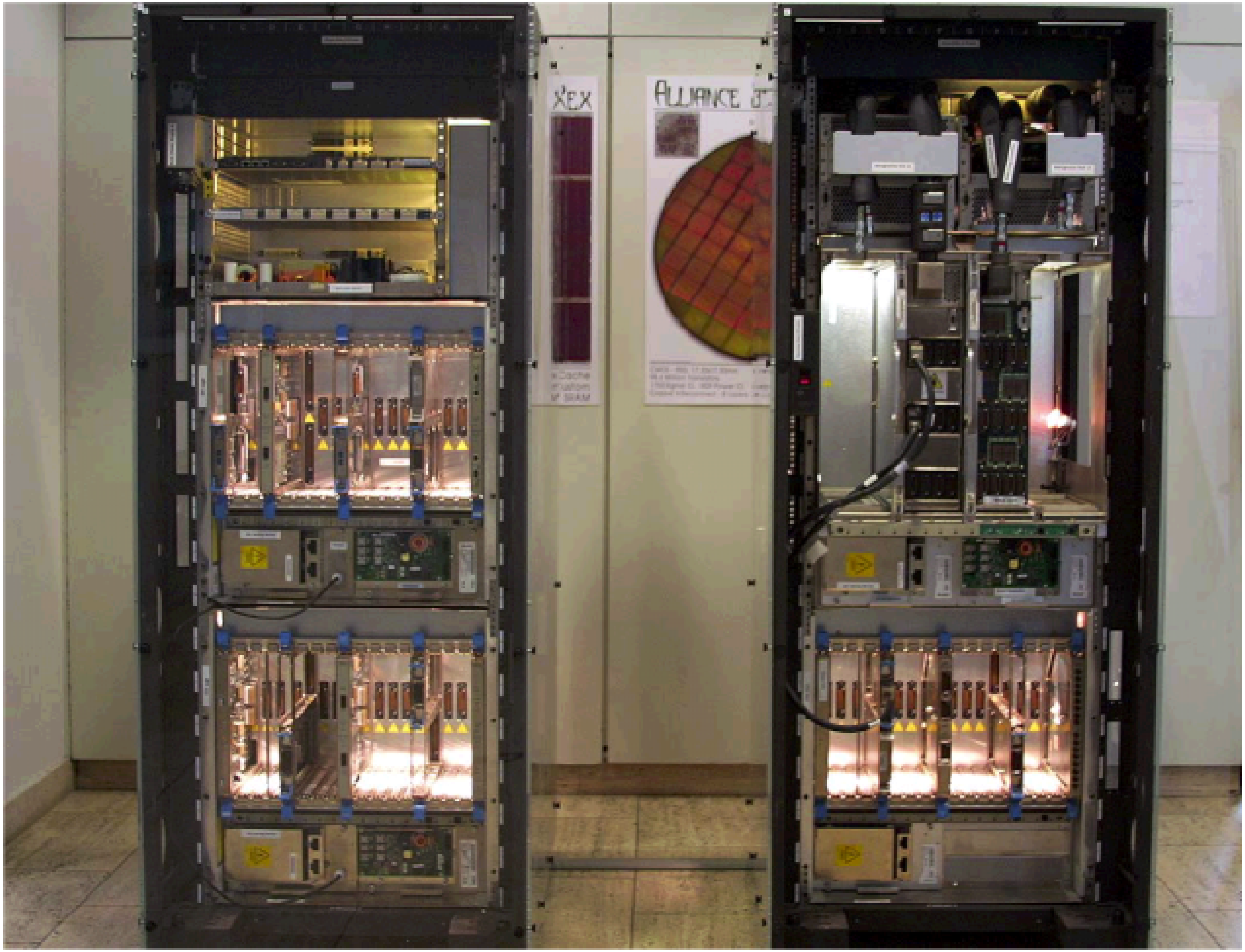
Processor Books, Memory, MBA and HCA cards

Ethernet cables for internal System LAN connecting Flexible Service Processor (FSP) cage controller cards

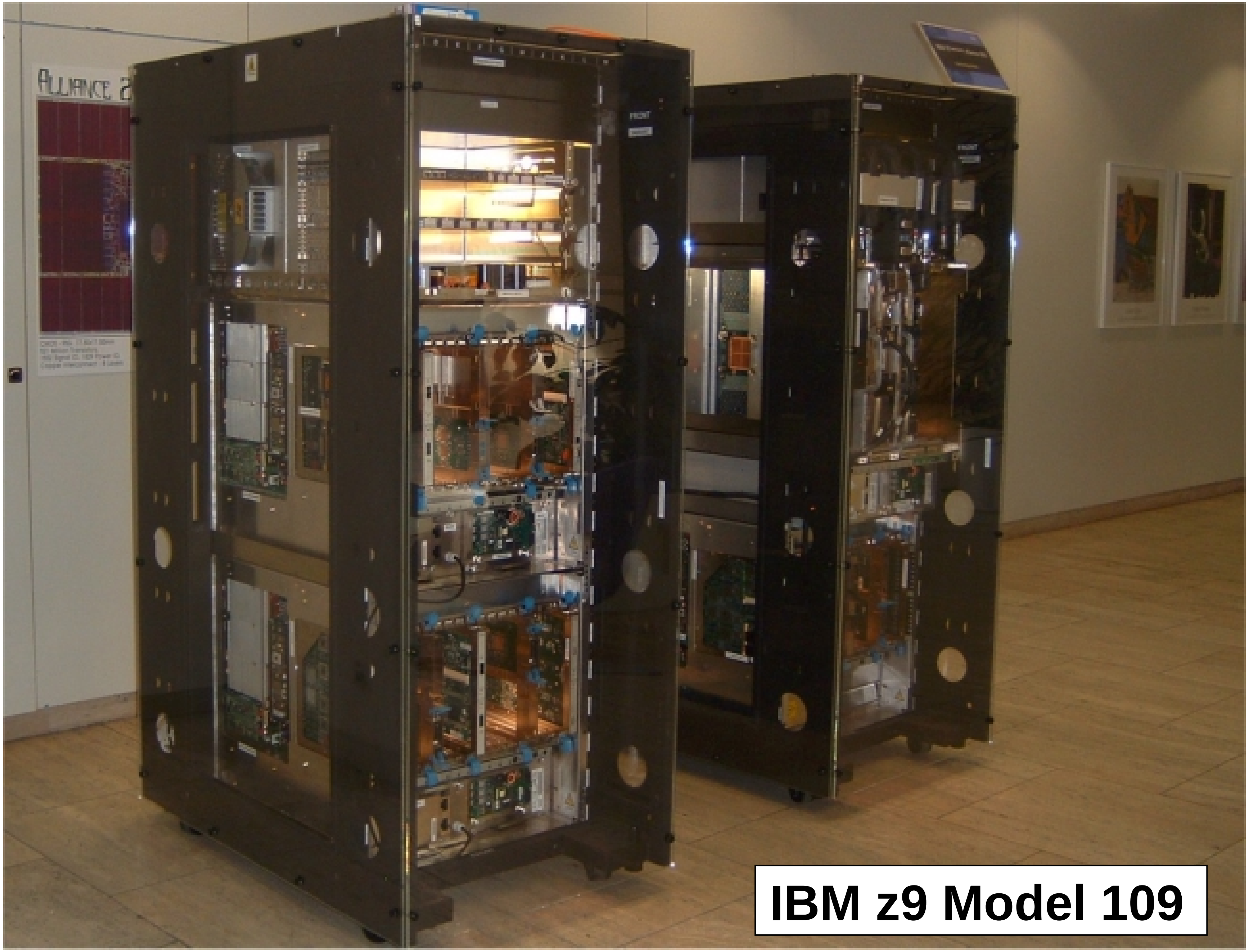
InfiniBand I/O Interconnects

2 x Cooling Units

This shows the internal components of a z10 system. The 4 books are in the right upper area.

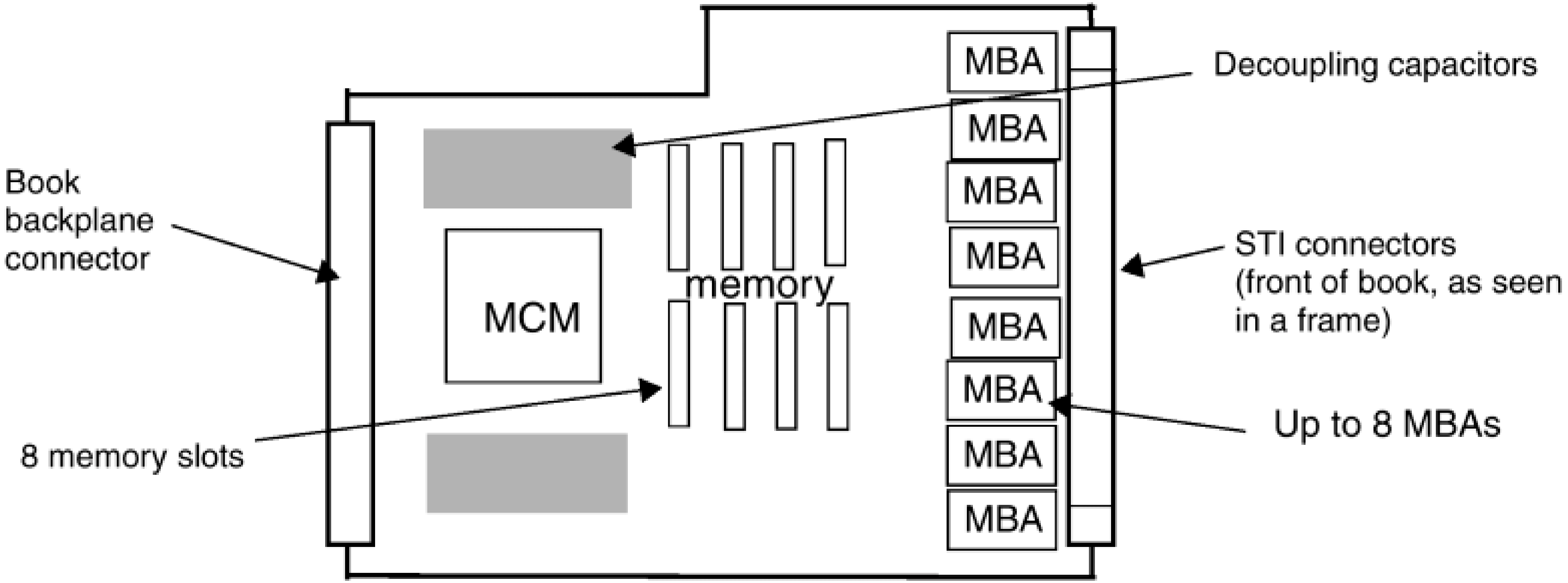


This picture hows the Z frame and A frame of a Z990 Mainframe



and here is a slightly different z9 model 109.

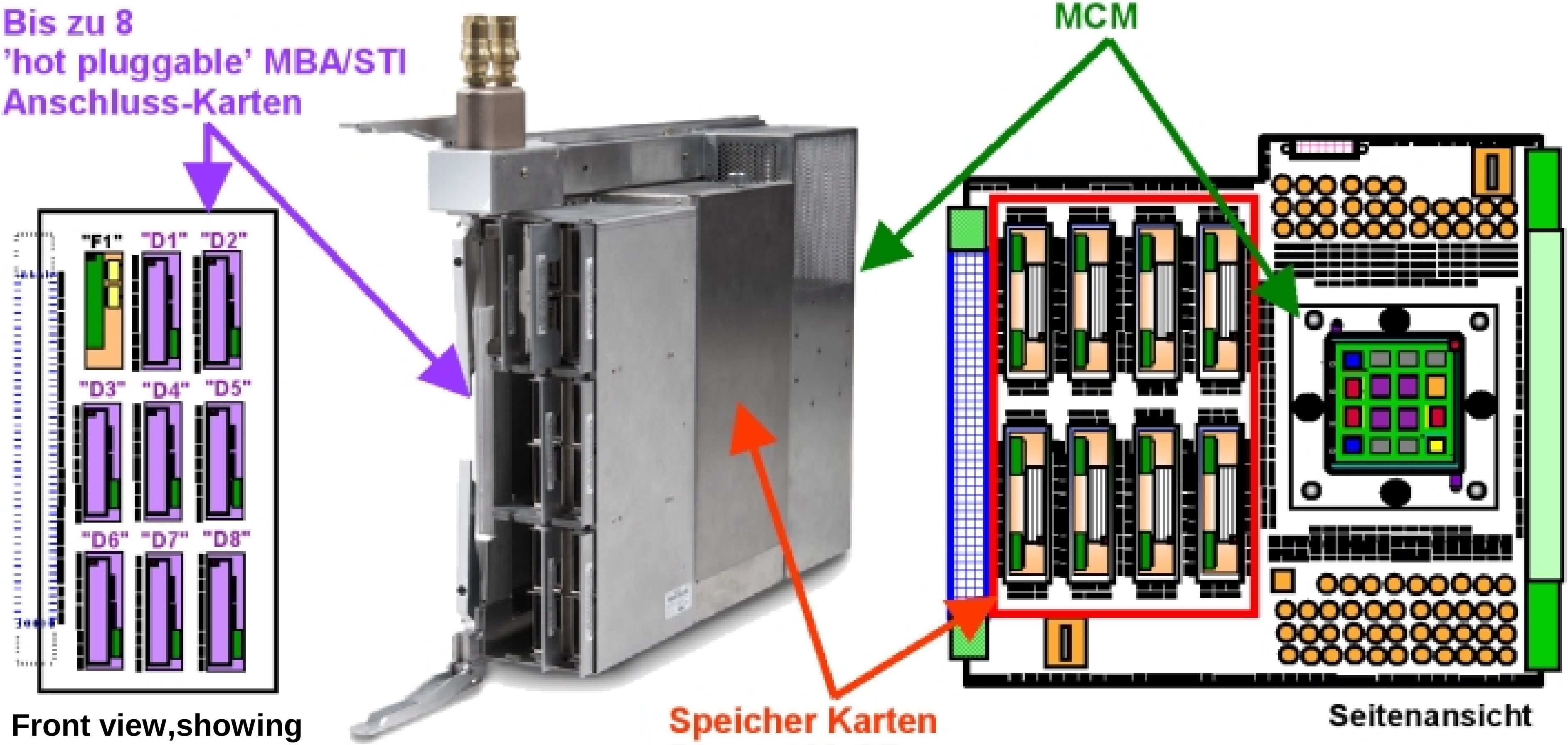
Shown is an engineer removing a book from a z9 system.



A z9 book consists of a single MCM, slots for memory cards, and a number of Memory Bus Adapter (MBA) chips/cards. Each MBA implements a Direct Memory Access (DMA) connection to the memory, which is somewhat similar in function to the DMA chip in your PC. MBAs connect via an STI (z9) or Infiniband (z10) cable to an I/O cage. The I/O cage backplane implements (similar to the PCI bus) an STI or Infiniband bus to accept the various I/O adapter cards.

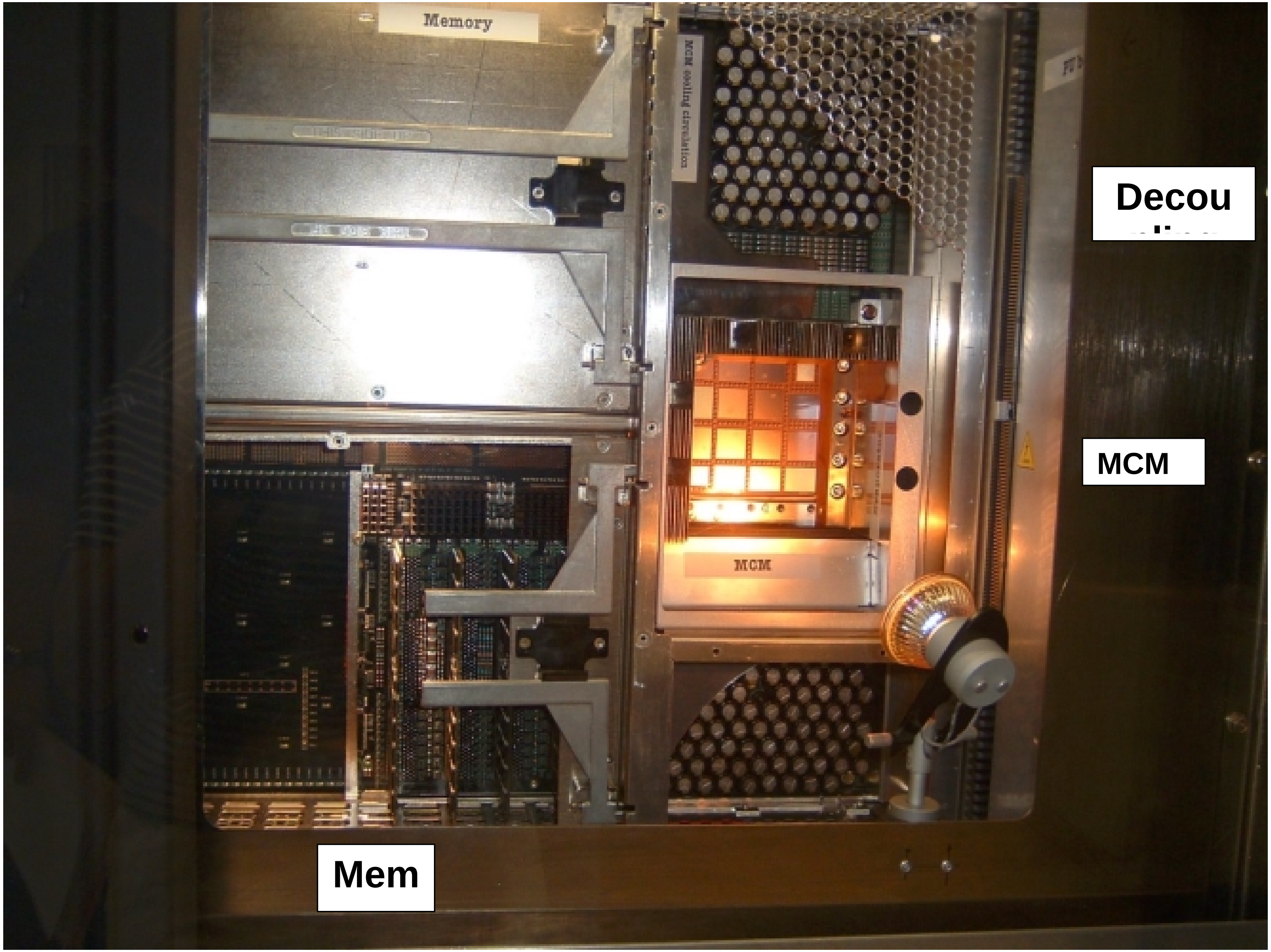
The 4 books form the “Central Electronic Complex” (CEC) , sometimes also called the “Central Processing Complex” (CPC). They are mounted in a common CEC cage and are connected to a CEC backplane via a “Book Backplane Connector”.

z9-109 Prozessor Book Layout



Front view,showing
8 MBA/STI connectors

Each MBA fanout card connects to up to two STI cables. There are up to 8 MBA fanout cards per book, each driving two STIs, resulting in 16 STIs per book. All 16 STIs in a book have a data rate of 2.7 GByte/s each.



Decou

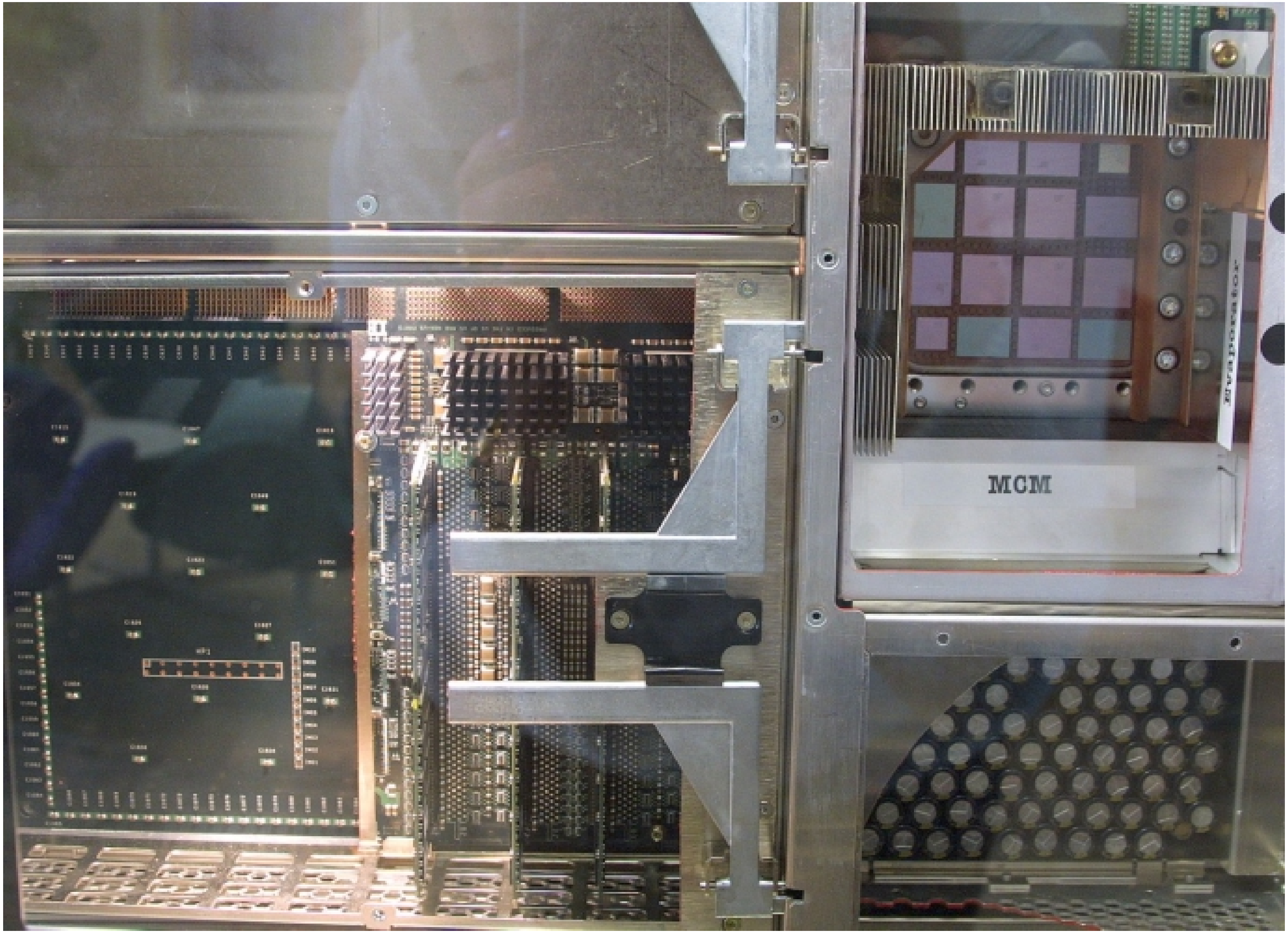
MCM

Mem

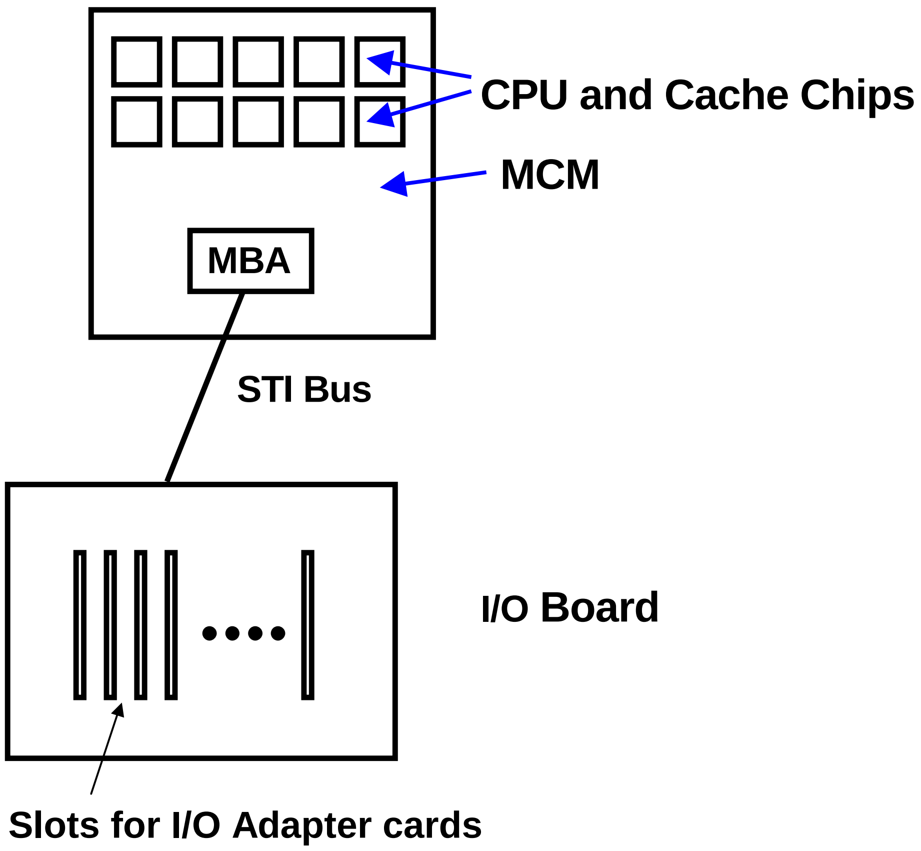
View of an opened book

The picture shows the MCM ob the right side, center, of the book. Above and below the MCM are a large number of decoupling capacitors. Short time load variations are satified from the decoupling capacitors, which in turn are charged from the central power supply. In a z10 system. the decoupling capacitors are replaced by a small auxiliary power supply contained in the book to handle short time load variations.

Memory cards are shown on the left side of the book.



View of the memory cards

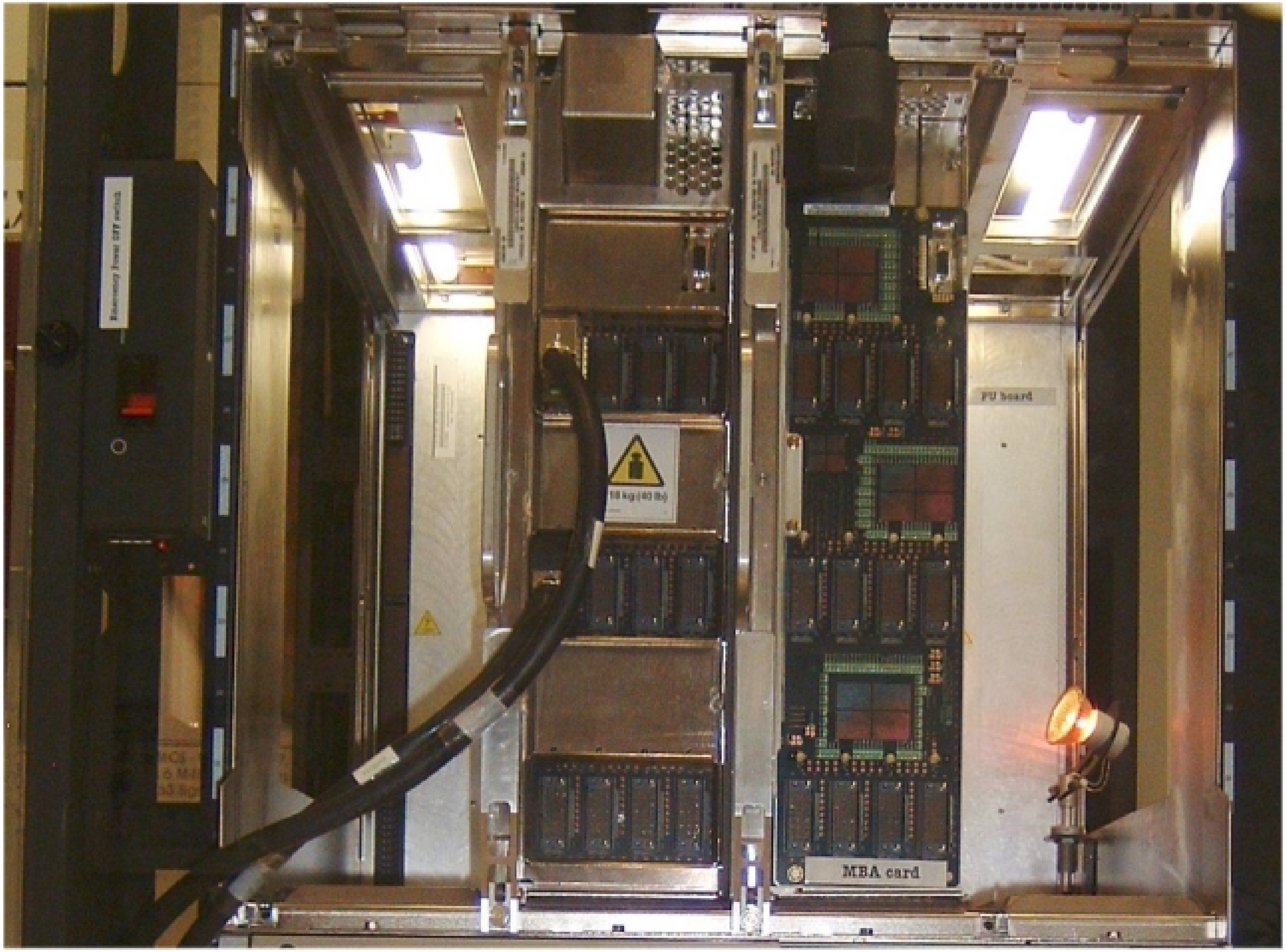


I/O Board

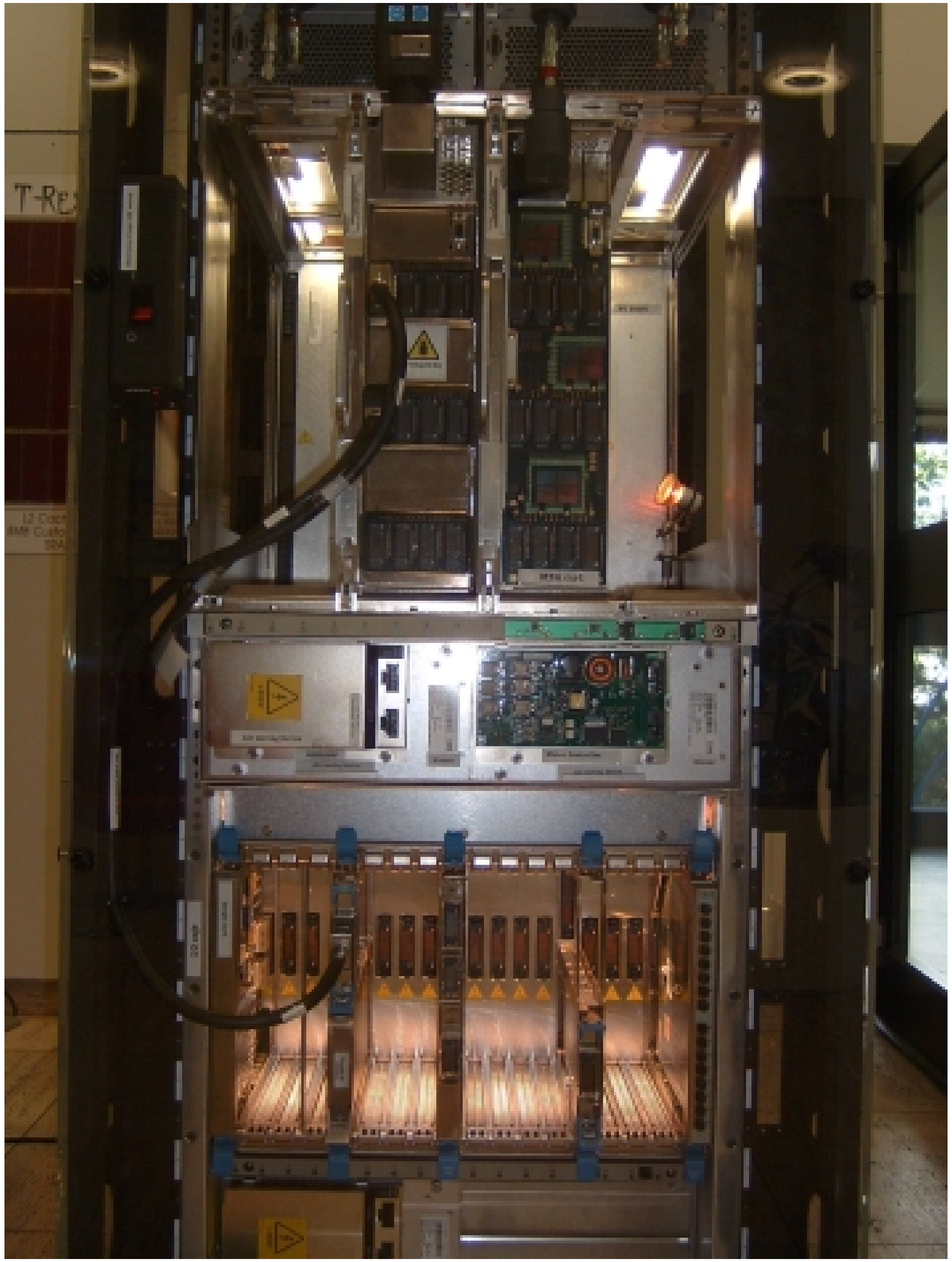
The MCM connects to 8 Memory Bus Adapter (MBA) chips, which have a function comparable to the Southbridge chip in a PC. Each MBA connects to two STI cables, comparable to the PCI Bus in a PC. Die STI Busse connect to I/O cages, that have STI Slots to accept I/O cards.

I/O adapter cards have connections for

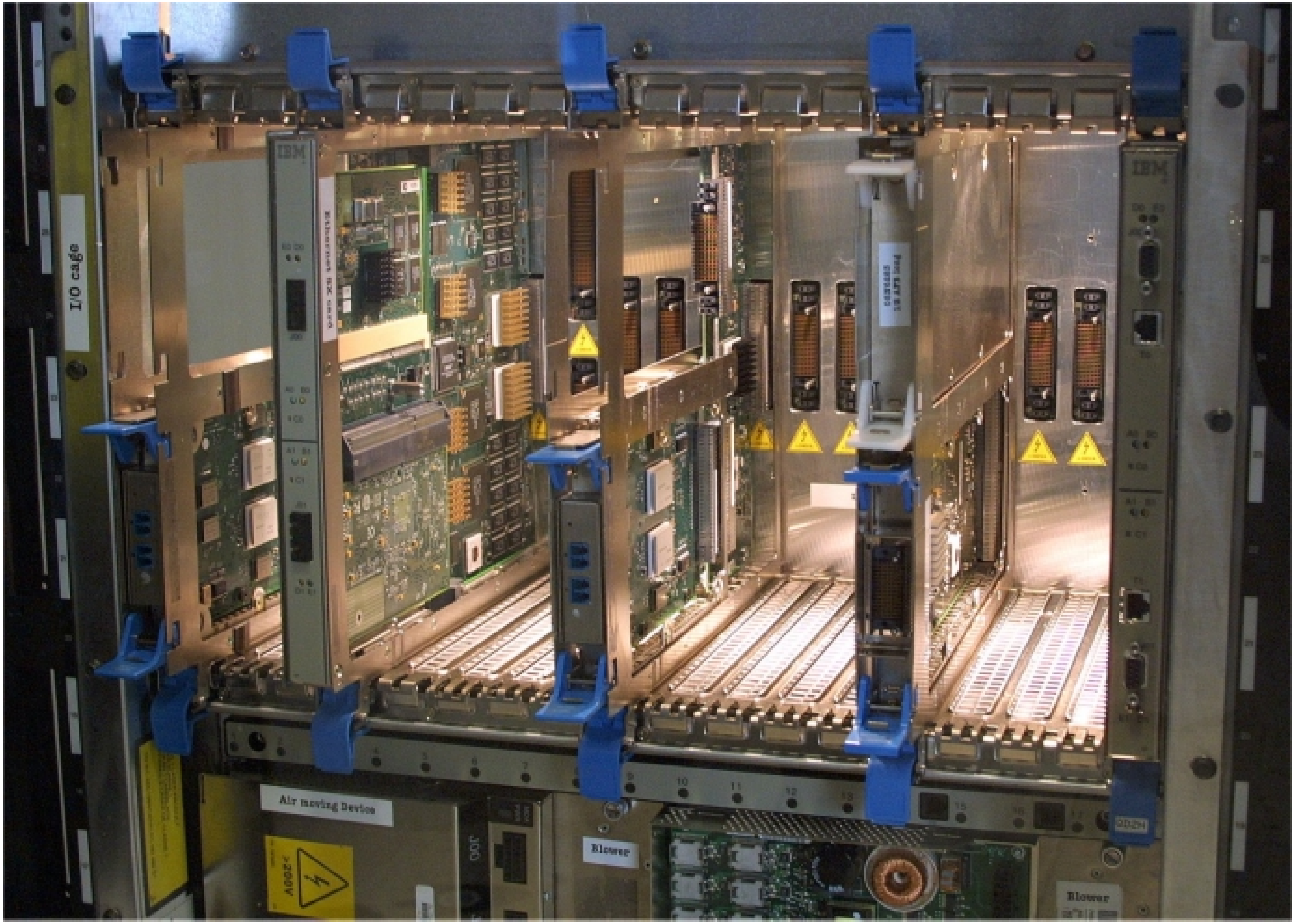
ESCON
FICON
OSA Adapter for Ethernet, Token Ring, ATM, ...



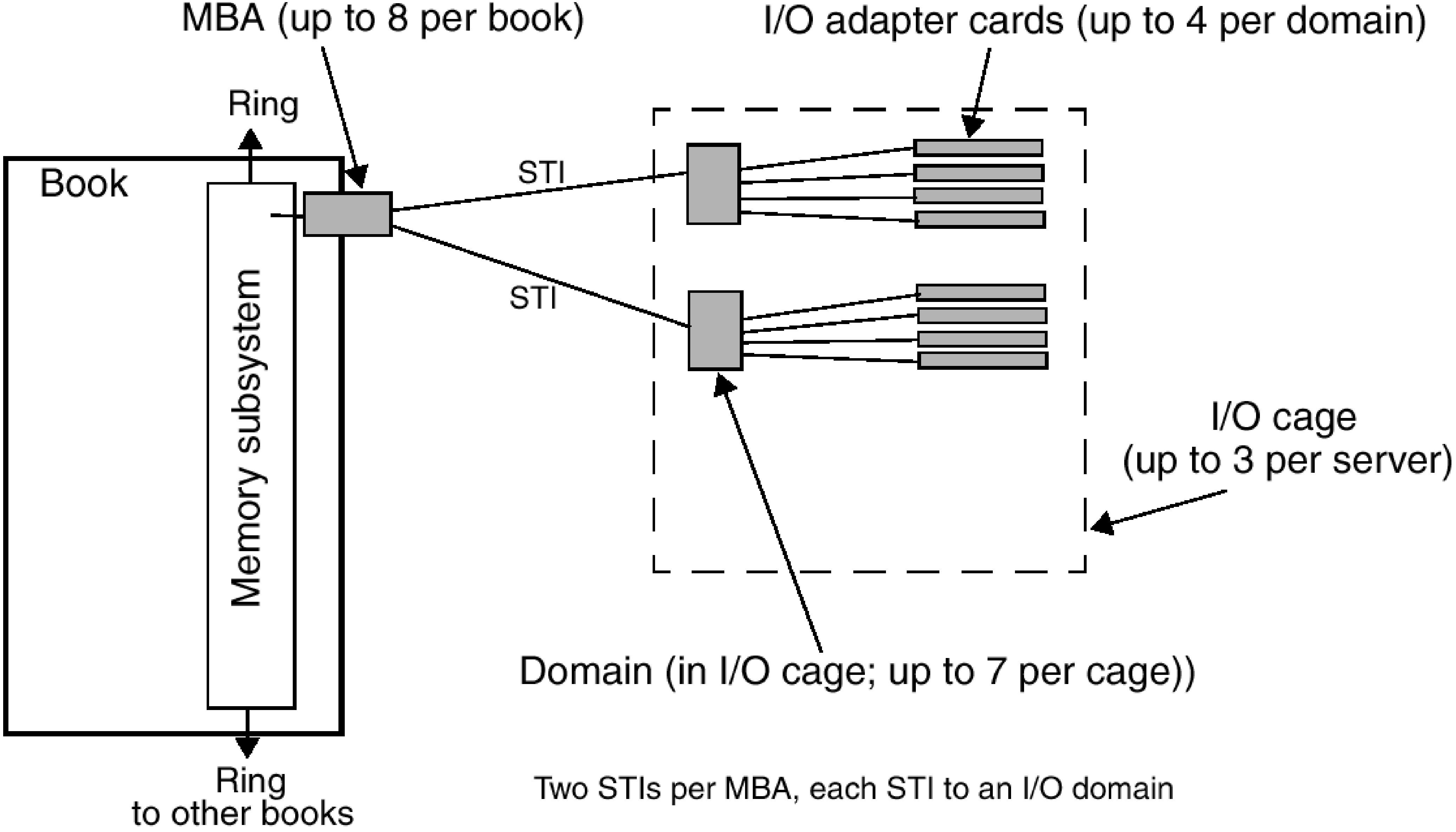
**Front side view of a CEC, populated with 2 Books (out of a possible of 4).
Two STI cables connect to an MBA adapter.**



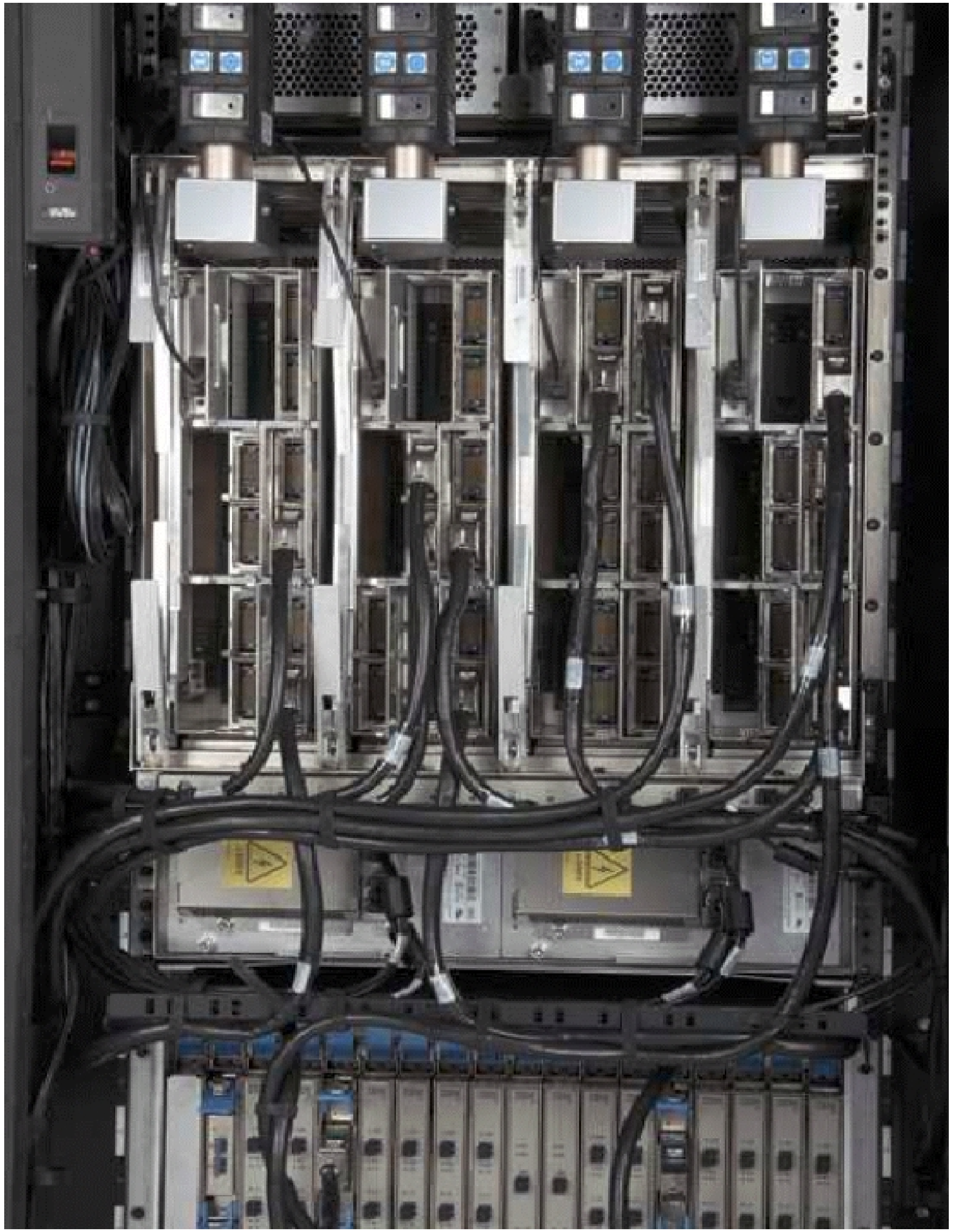
STI cable connectiong ton an I/O cage



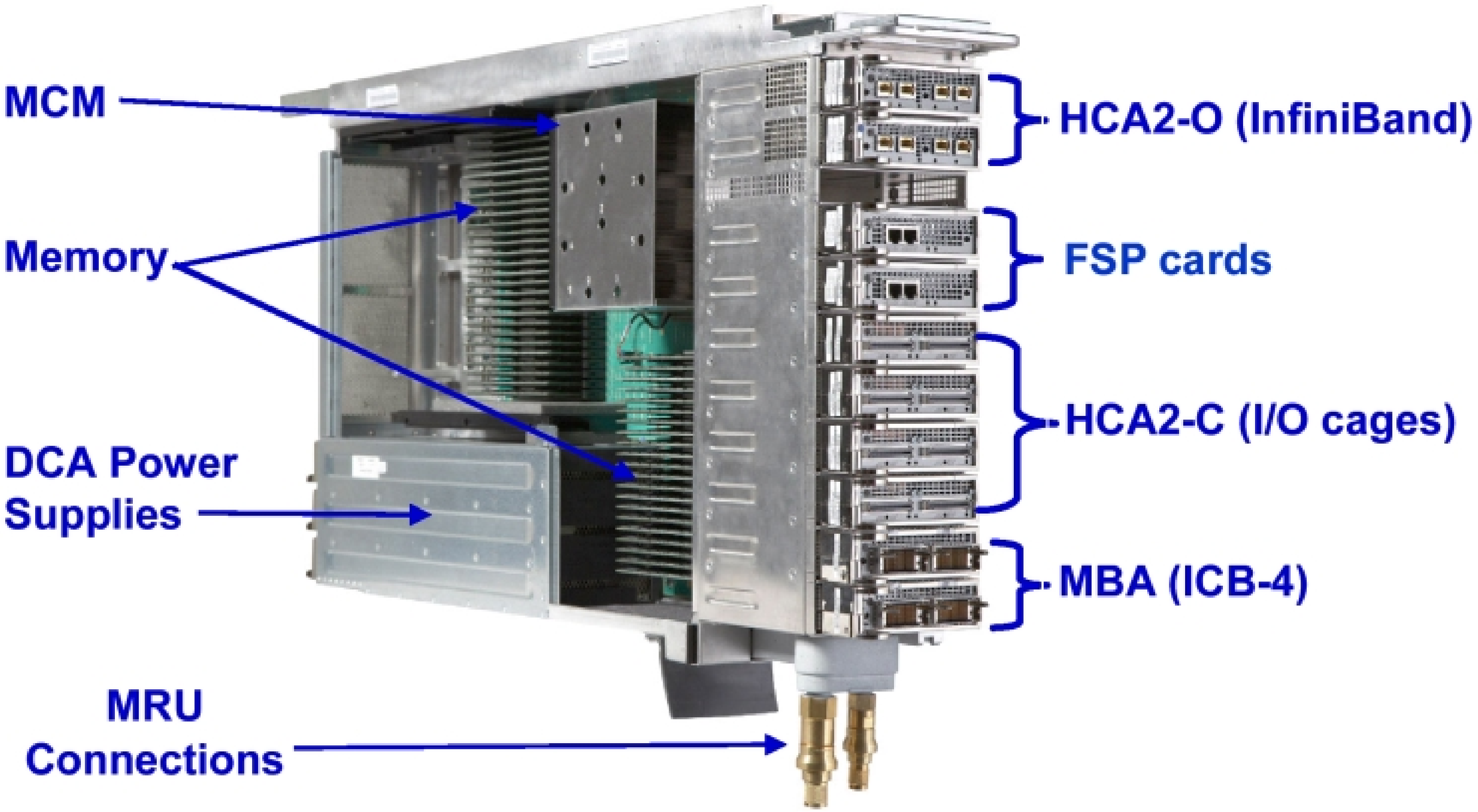
I/O cage



Two STIs per MBA, each STI to an I/O domain



A fully populated CEC with 4 books and cables connecting to a fully populated I/O cage.

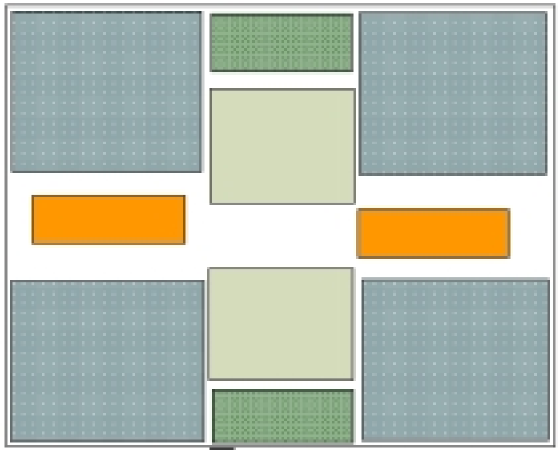


z10 EC book structure and components

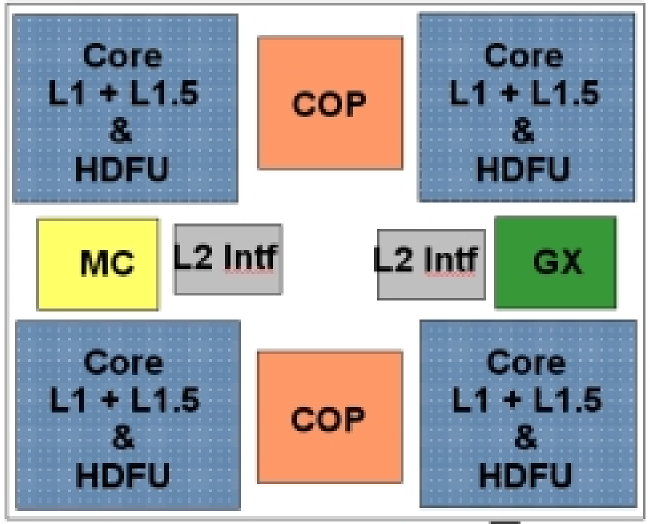
The z10 book lokks somewhat different than the z9 book. The large bank of capaciters has been replaced by a special power supply. Most important, the MBA connecters on the front side of the book habe been replaced by Infiniband HCA connectors, in ether an electrical form (HCA2-C) or an optical form (HCA2-O).

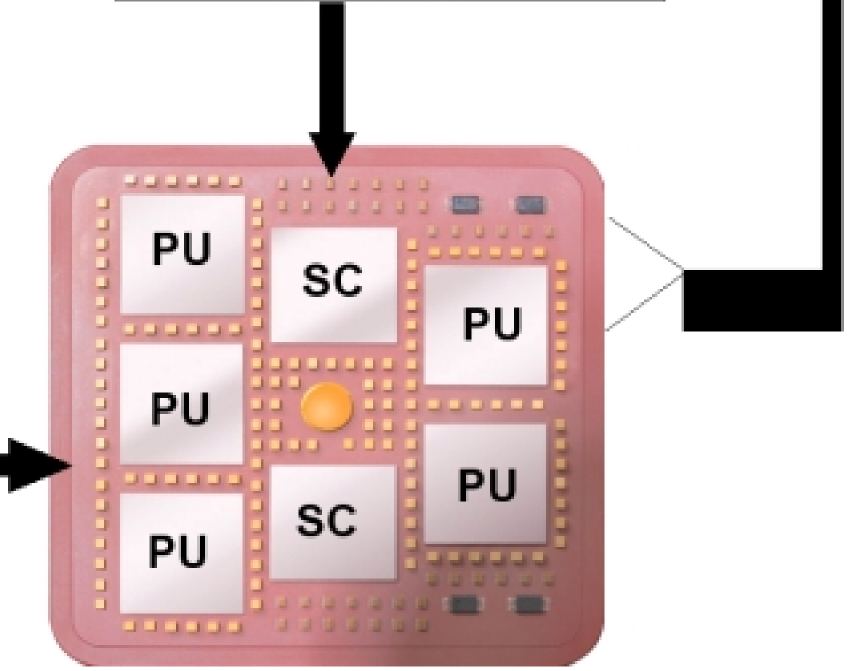
z10EC Processor/Memory/HCA and Book

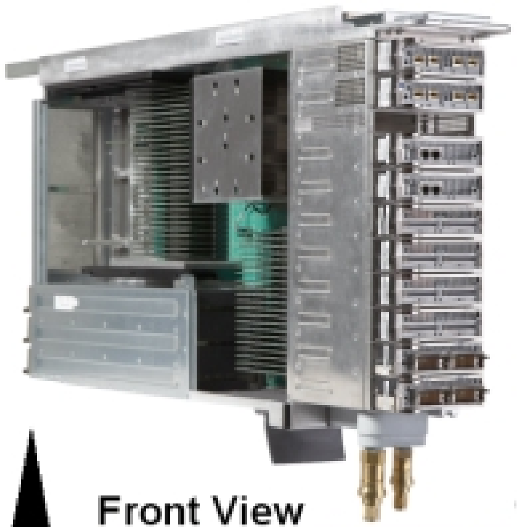
SC CHIP



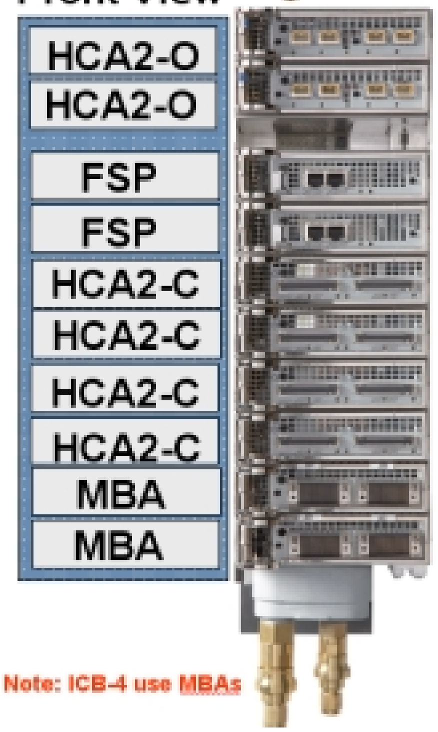
PU CHIP







Front View



Note: ICB-4 use MBAs

- Up to 8 Hot pluggable HCA fanout cards
- Plugging rules apply and dependant on Model

Infiniband (IB)

One of the significant differences between the z9 and the z10 models is the use of Infiniband technology in the z10.

The z9 uses the STI technology to interconnect the books of the CEC to the I/O cards in an I/O cage. STI is somewhat superior but similar to the PCI bus employed in a PC. Large systems from other suppliers, e.g. the Superdome from Hewlett Packard, use PCI bus cables to interconnect their CEC to the I/O cages.

InfiniBand is a switched fabric communications link primarily used in high-performance computing. Its features include quality of service and failover, and it is designed to be scalable. The InfiniBand architecture specification defines a connection between processor nodes and high performance I/O nodes such as storage devices.

Like Fibre Channel, PCI Express, Serial ATA, and many other modern interconnects, InfiniBand is a point-to-point bidirectional serial link intended for the connection of processors with high speed peripherals such as disks. It supports several signalling rates and, as with PCI Express, links can be bonded together for additional bandwidth.

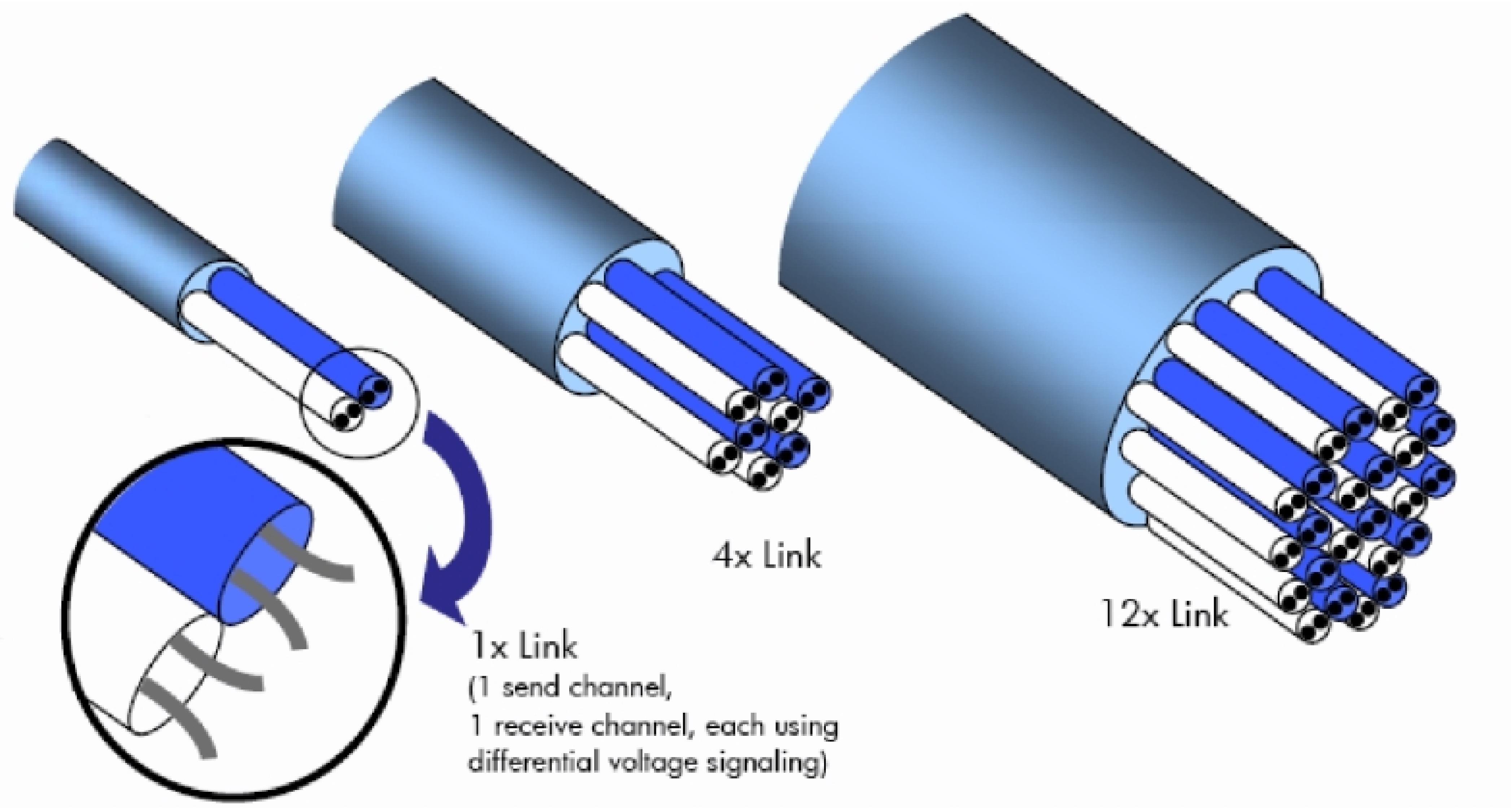
Infiniband physical lanes

Each individual InfiniBand link is based on

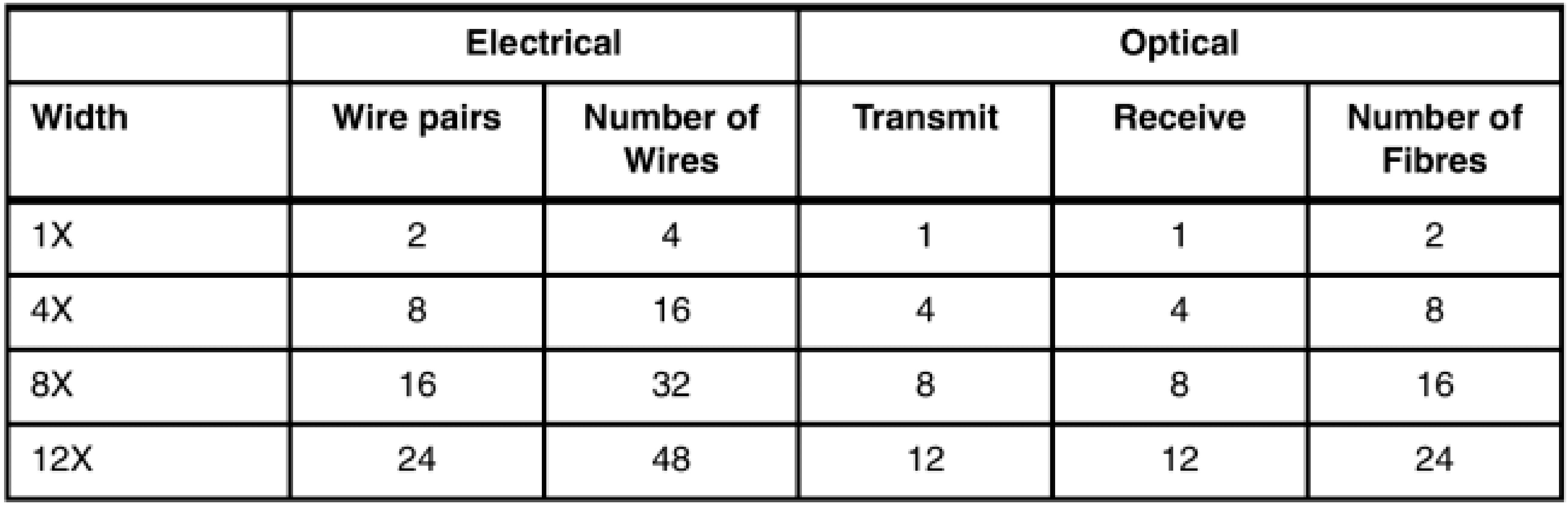
- **a two-fiber 2.5 Gbit/ss bidirectional connection for an optical (fiber cable) implementation, or**
- **a four wire 2.5 Gbit/s bidirectional connection for an electrical (copper cable) implementation.**

An optical two-fiber 2.5 Gbps bidirectional connection or an electrical four wire 2.5 Gbps bidirectional connection is called a physical lane. Each lane supports multiple transport services for reliability and multiple prioritized virtual communication channels. Physical lanes are grouped together in support of

- **one physical lane or 1 link (1X),**
- **4 physical lanes or 4 link (4X),**
- **8 physical lanes or 8 link (8X),**
- **12 physical lanes or 12 link (12X).**


4x Link
12x Link
1x Link
(1 send channel,
1 receive channel, each using
differential voltage signaling)

The one physical lane (one link) Infiniband copper cable implementation has 2 copper wires for the send channel and 2 copper wires for the receive channel, for a total of 4 wires. The 4 link version has 16 wires, and the 12 link version (used in the z10) has 48 wires.



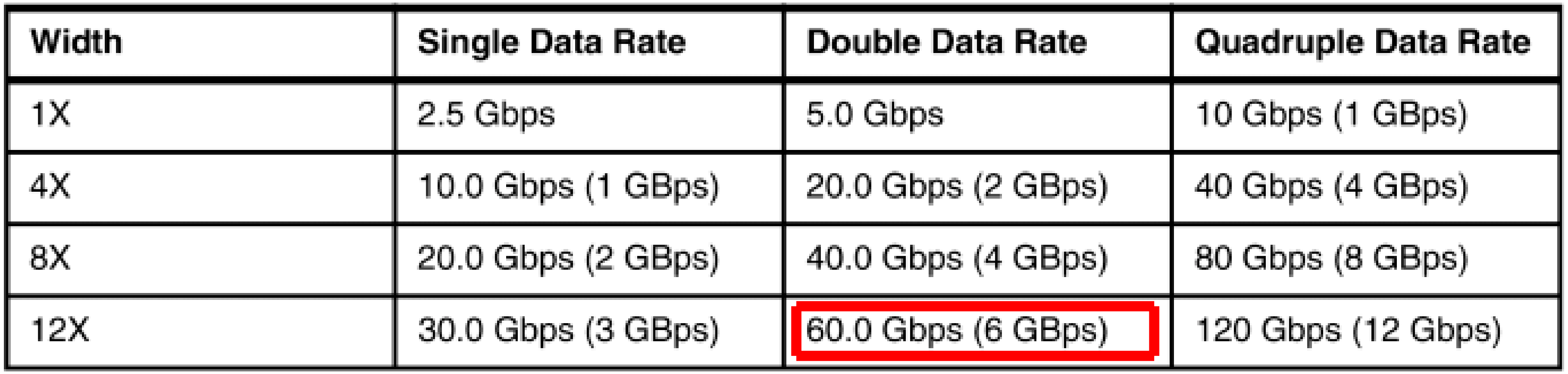
	Electrical		Optical		
Width	Wire pairs	Number of Wires	Transmit	Receive	Number of Fibres
1X	2	4	1	1	2
4X	8	16	4	4	8
8X	16	32	8	8	16
12X	24	48	12	12	24

Infiniband interface widths and electrical and optical wiring

For example, the z10 uses a width of 12X

- **in an electrical configuration to interconnect the CEC with an I/O cage**
- **in an optical configuration to interconnect the CEC with the CEC of another z10 or z9 system**

In a z10, an electrical Infiniband cable has 48 wires, and an optical cable has 24 fibres.



Width	Single Data Rate	Double Data Rate	Quadruple Data Rate
1X	2.5 Gbps	5.0 Gbps	10 Gbps (1 GBps)
4X	10.0 Gbps (1 GBps)	20.0 Gbps (2 GBps)	40 Gbps (4 GBps)
8X	20.0 Gbps (2 GBps)	40.0 Gbps (4 GBps)	80 Gbps (8 GBps)
12X	30.0 Gbps (3 GBps)	60.0 Gbps (6 GBps)	120 Gbps (12 Gbps)

Infiniband interface width and link speeds

InfiniBand defines several link speeds at the physical layer. It negotiates the use of:

- **Single Data Rate (SDR) delivering 2.5 Gbit/s per physical lane.**
- **Double Data Rate (DDR) delivering 5.0 Gbit/s per physical lane.**
- **Quadruple Data Rate (QDR) delivering 10.0 Gbit/s per physical lane.**

Links use 8B/10B encoding (every ten bits sent carry eight bits of data), so that the useful data transmission rate is four-fifths the signalling or link rate (signalling and link rate equal the raw bit rate). 1X single, 4x double, and 8x quad rates carry 2 Gbit/s, 4 Gbit/s, or 8 Gbit/s of useful data respectively.

The z10 uses double data rate.

4B/5B Coding Scheme
5 Bit “Symbol” or “Code Group”

HEX Data	FDDI or Infiniband Symbol
0	11110
1	01001
2	10100
3	10101
4	01010
5	01011
6	01110
7	01111
8	10010
9	10011
A	10110
B	10111
C	11010
D	11011
E	11100
F	11101

32 Combinations

16 are Data-Symbols (Nibbles), Hex 0..F
8 are Control-Symbols (e.g. Start, End)
8 are “Violation”-Symbols
(e.g. run-length violation)

If you string an arbitrary sequence of 5 bit data symbols together, you will never have more than three consecutive zero’s.

InfiniBand Technology on System z Servers

Used as follows:

- An I/O interface which includes the InfiniBand Double Data Rate (IB-DDR) infrastructure replacing the Self-Timed Interconnect (STI) features found in prior System z servers.

- Parallel Sysplex InfiniBand (PSIFB) links have a bandwidth of up to 6 GByte/s between two System z10 servers and of up to 3 GByte/s between System z10 and System z9 servers. Since it supports point-to-point to up to 150 m, it replaces Inter-system Cluster Bus-3 (ICB-3) and Inter-system Cluster Bus-4 (ICB-4) for distances greater than seven meters.

- Server Time Protocol timekeeping signals are supported on Parallel Sysplex InfiniBand (PSIFB) links.

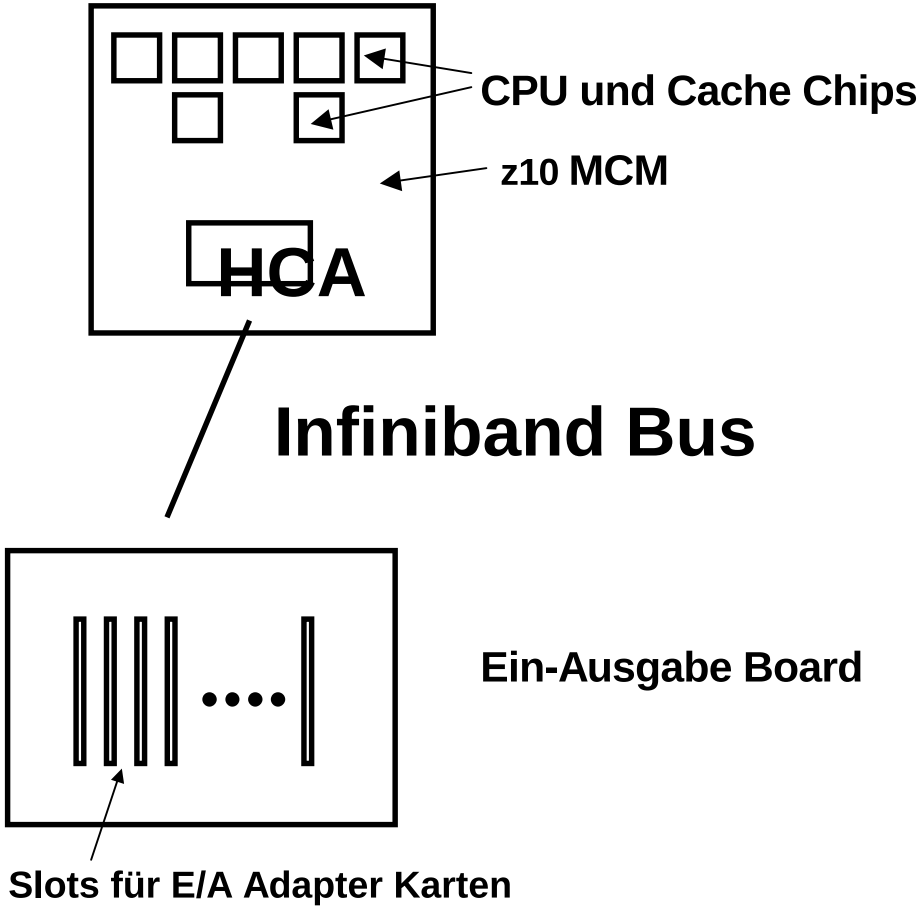
Host Channel Adapter

Host Channel Adapters (HCAs) are physical devices in servers and I/O equipment that create and receive packets of information. The host channel adapter is a programmable Direct Memory Access (DMA) engine that is able to initiate local memory operations. The DMA engine offloads costly memory operations from the processor since it can access system memory directly for reading and writing independently from the central processor thus enabling transfer of data with much less CPU overhead. The CPU initiates the transfer and switches to other operations while the transfer is in progress. Eventually the CPU receives an interrupt once the transfer operation has been completed.

A host channel adapter has one or more ports each having its own set of transmit and receive buffers called Virtual Lanes (VLs) enabling the port to support multiple simultaneously send and receive operations. For example, the host channel adapter ports provide multiple communication interfaces by providing send and receive queues for each interface.

A host channel adapter provides an interface to a host device (I/O device, for example a DASD) and supports “verbs” defined to InfiniBand. Verbs describe the service interface between a host channel adapter and the software supporting it. Verbs allow the device driver and the hardware to work together.

Ein-/Ausgabe Board



CPU und Cache Chips

z10 MCM

HCA

Infiniband Bus

Ein-Ausgabe Board

Slots für E/A Adapter Karten

Ein/Ausgabegeräte werden über I/O Adapter Karten mit dem L2 Cache verbunden. Ein Beispiel ist die FICON Express Karte, die 4 FICON Ports mit je 4 Gbit/s enthält.

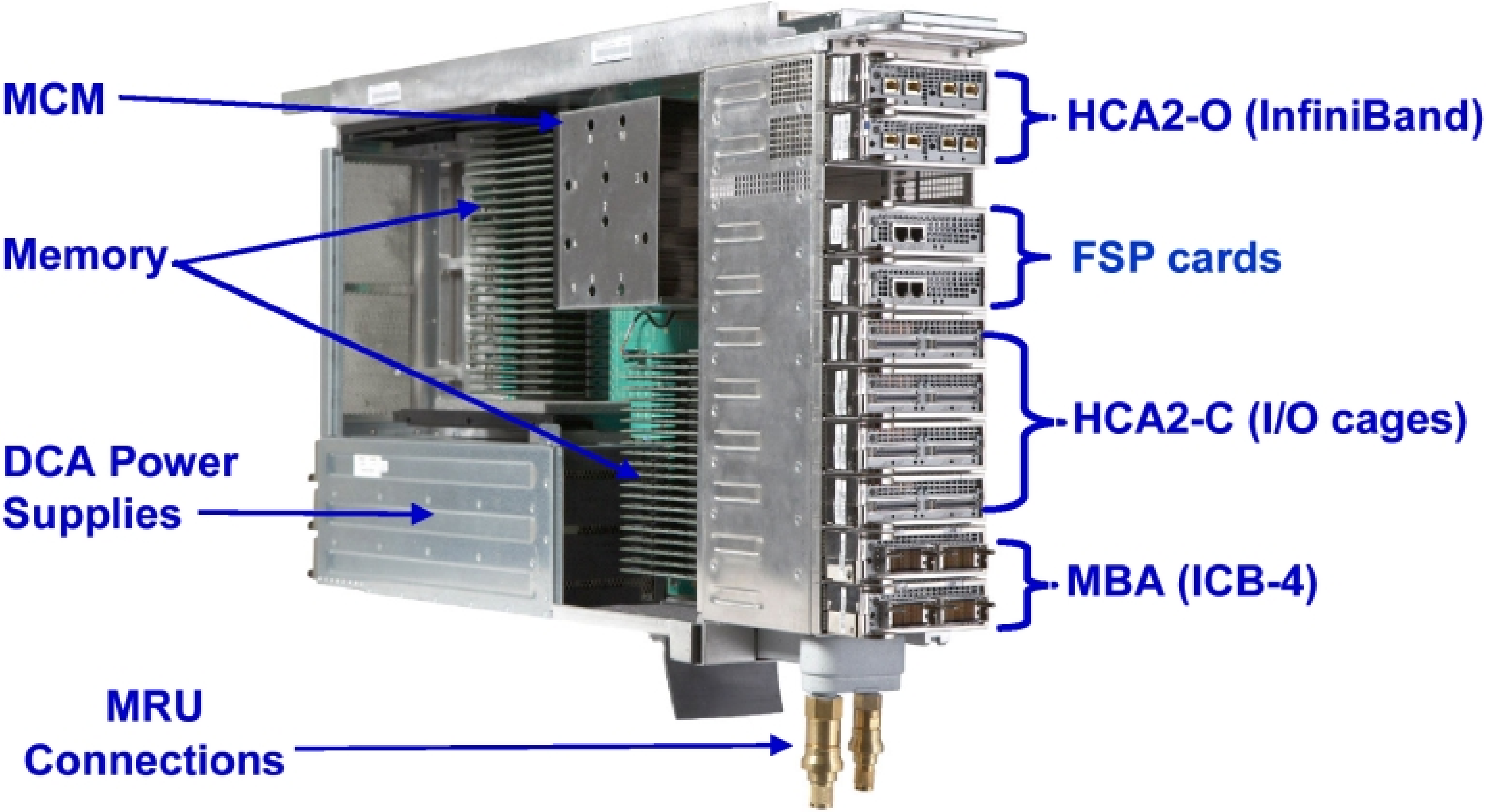
I/O Adapter Karten sind über Infiniband und 6 MByte/s mit dem Host Channel Adapter (HCA) verbunden. Der HCA ist eine programmierbare Direct Memory Access (DMA) Engine.

Die wichtigsten E/A Adapter Board Typen haben Anschlüsse für

FICON Kanal
OSA Adapter für Ethernet, Token Ring, ATM, ...

z10 uses two types of Host Channel Adapters:

HCA2-C for copper interconnections between the CEC and an I/O cage
HCA2-O for optical interconnections between the CEC and another z10 or z9 system.



z10 EC book structure and components

Host Channel Adapter Types on System z10 and System z9

The I/O interface includes the InfiniBand Double Data Rate (IB-DDR) infrastructure replacing the Self-Timed Interconnect features found in prior System z servers.

On System z10 and System z9 servers, host channel adapters are implemented for a variety of applications:

- **A host channel adapter identified as HCA2-C provides a copper InfiniBand connection from a book to I/O cages on a System z10. HCA2-C is used in combination with 12X IB-DDR link rating to provide a link rate of up to 6 GByte/s.**
- **A host channel adapter identified as HCA2-O provides an optical InfiniBand connection for coupling links from System z10 to System z10, or from System z10 to System z9. HCA2-O is used in combination with 12X IB-DDR link rating to provide a link rate of up to 6 GByte/s.**
- **A host channel adapter identified as HCA1-O provides an optical InfiniBand connection on System z9. HCA1-O is used in combination with the 12X IB-SDR link rating to provide a link rate of up to 3 GByte/s.**

In z 9 systems, an STI or eSTI connection with data rates of 2 GBps and 2.7 GBps respectively was used. The connection was made between the MBA, which is installed in a fanout slot in the front of the book, to the STI-MP (Self-Timed Interconnect - Multiplex) cards installed within the I/O cages.

With the System z10 an HCA2-C fanout in the front of the book connects via an IFB (Infiniband) copper cable to an IFB-MP (InfiniBand - Multiplex) card, installed in the I/O cages

Each z10 EC book supports a combination of up to eight I/O hubs (HCA-C, HCA-O or the new MBA), implemented via a combination of eight InfiniBand Host Channel Adapter (HCA2-Optical or HCA2-Copper) fanout cards, or Memory Bus Adapter fanout cards (MBAs).

Each of the cards have two ports thereby supporting up to 16 connections. HCA2-Copper connections are for links to the I/O cages in the server and the HCA2-Optical and MBA connections are to external servers (coupling links). MBA cards are used for ICB-4 links connecting to System z mainframes without Infiniband capability.

Sysplex InfiniBand technology

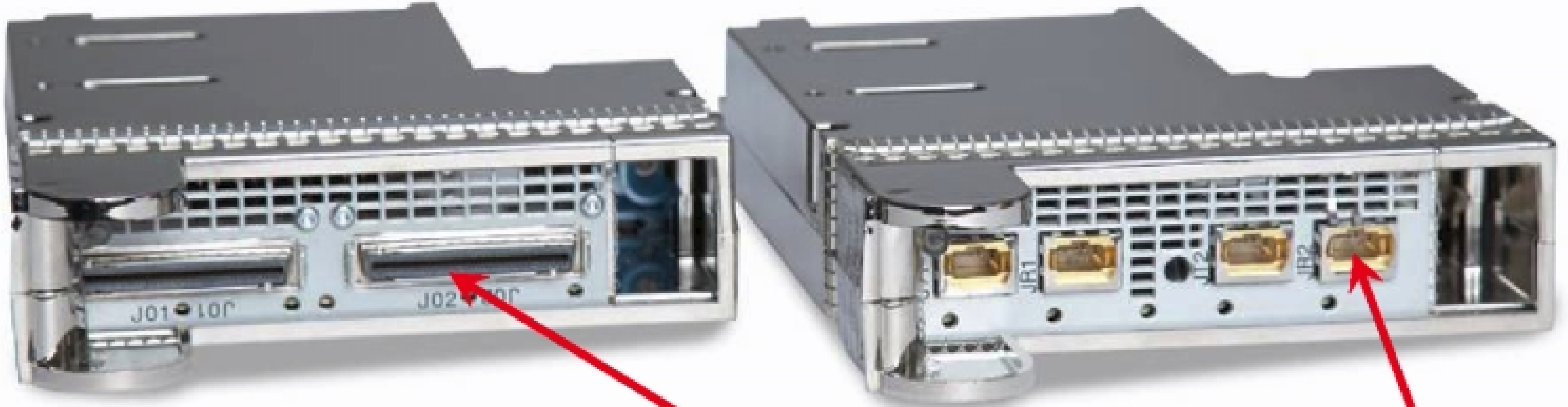
A “Sysplex” (Parallel Sysplex is the correct name) is a cluster of interconnected z/OS systems. Infiniband links can be used to interconnect z10 and z9 systems in a sysplex.

- Parallel Sysplex InfiniBand (PSIFB) links have a bandwidth of up to 6 GBps between two System z10 servers and of up to 3 GBps between System z10 and System z9 servers. Since it supports point-to-point to up to 150m, it replaces Inter-system Cluster Bus-3 (ICB-3) and Inter-system Cluster Bus-4 (ICB-4) for distances greater than seven meters.

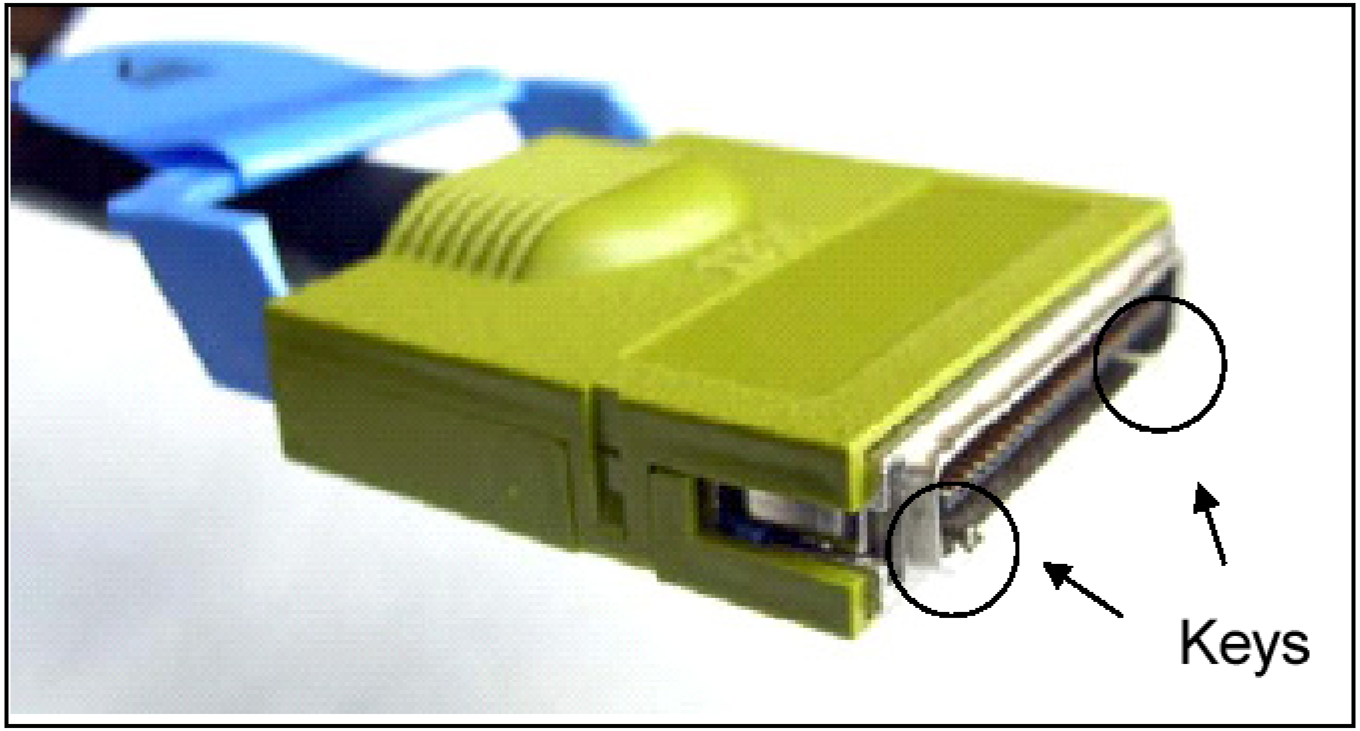
- Server Time Protocol timekeeping signals are supported on Parallel Sysplex InfiniBand (PSIFB) links.

Prior to the z10, sysplex interconnections were implemented by a variety of means, including the ICB-3 and ICB-4 links. MBA cards on a z10 are used for ICB-4 links connecting to System z mainframes without Infiniband capability.

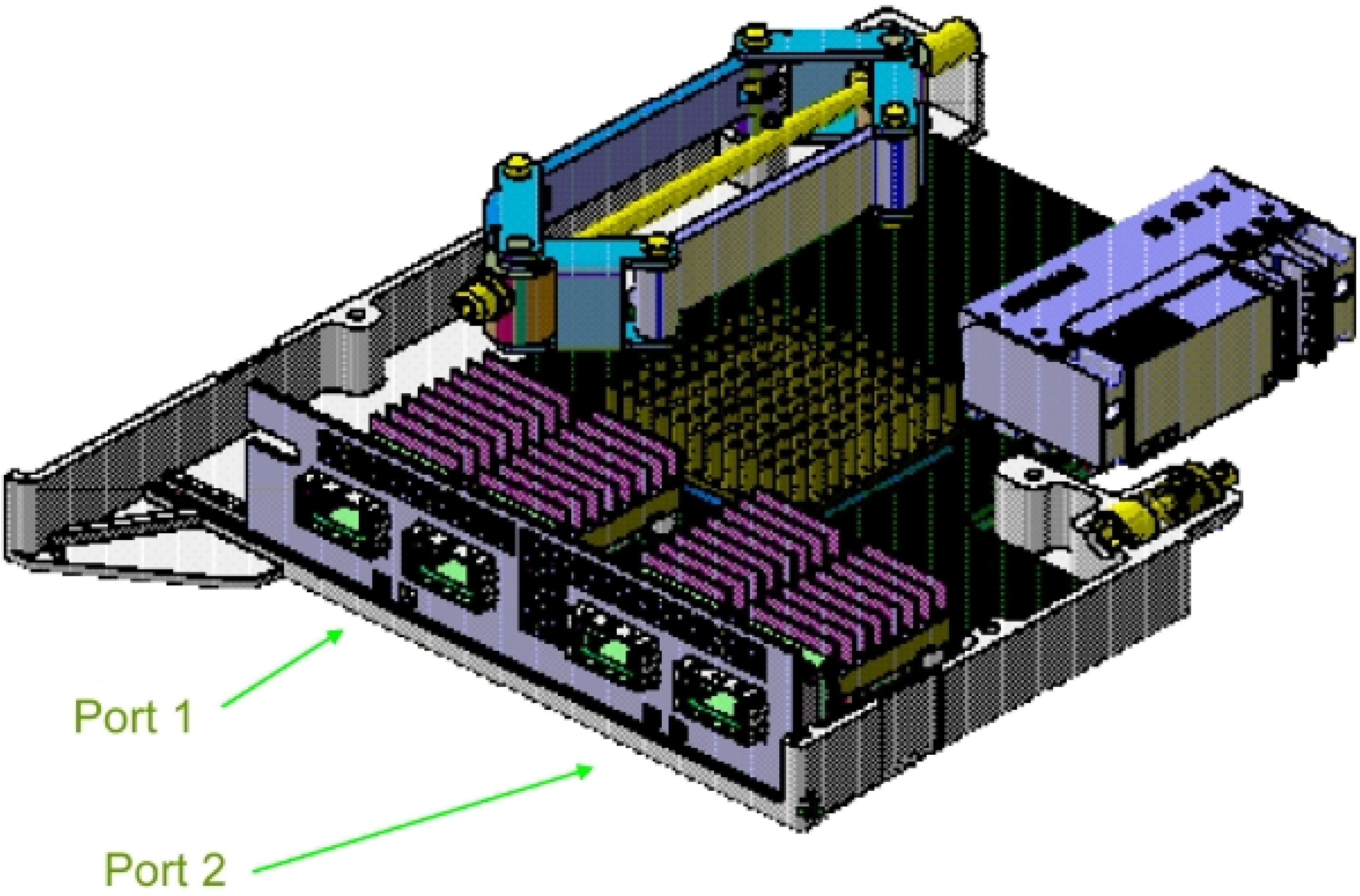
HCA2-C **HCA2-O**



- ***12x IB-DDR interfaces, 6 GBps***
  - ***HCA2-C (copper) = to I/O cages***
  - ***HCA2-O (optical) = External coupling***

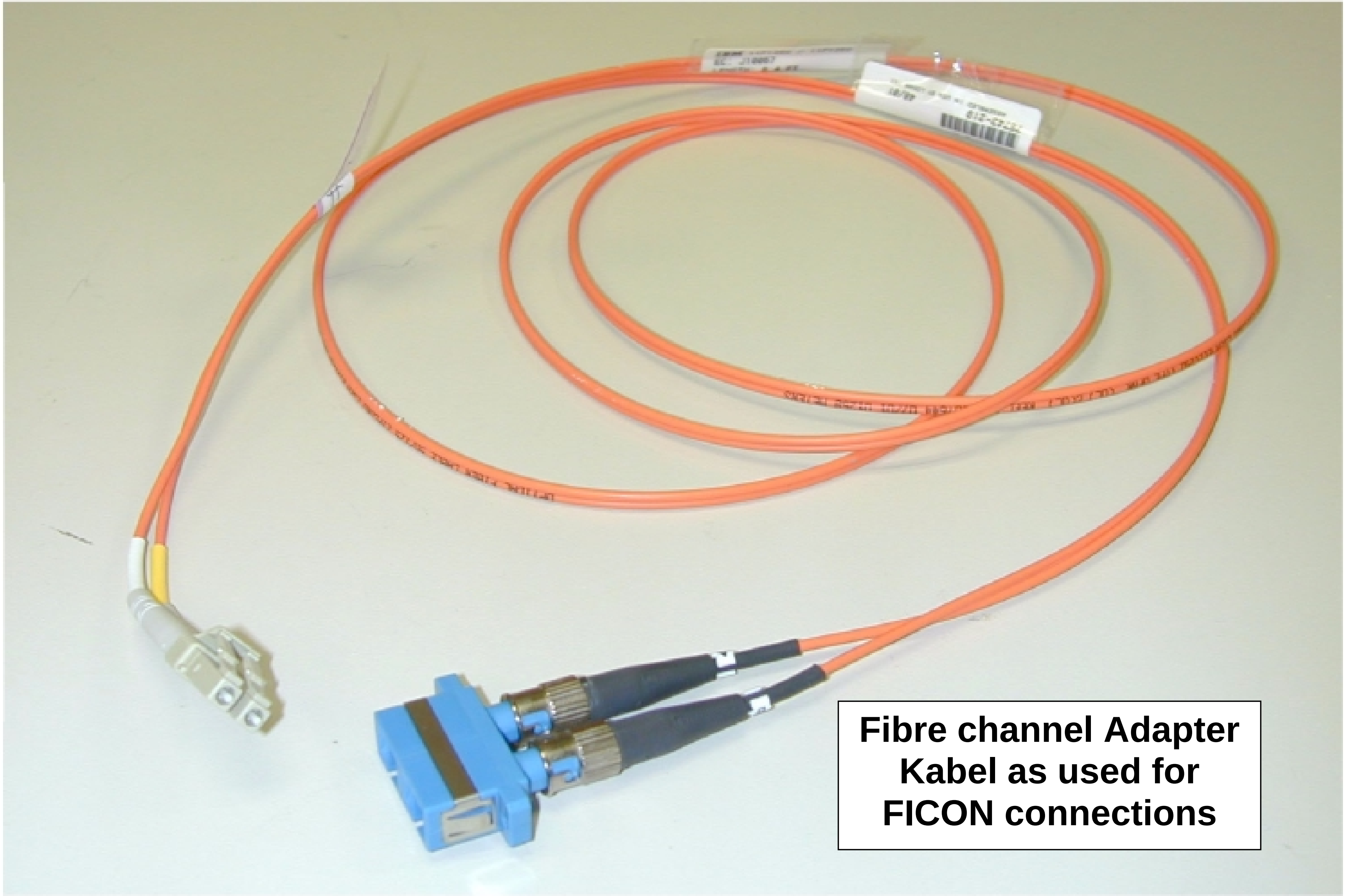


Infiniband Copper Cable Connector

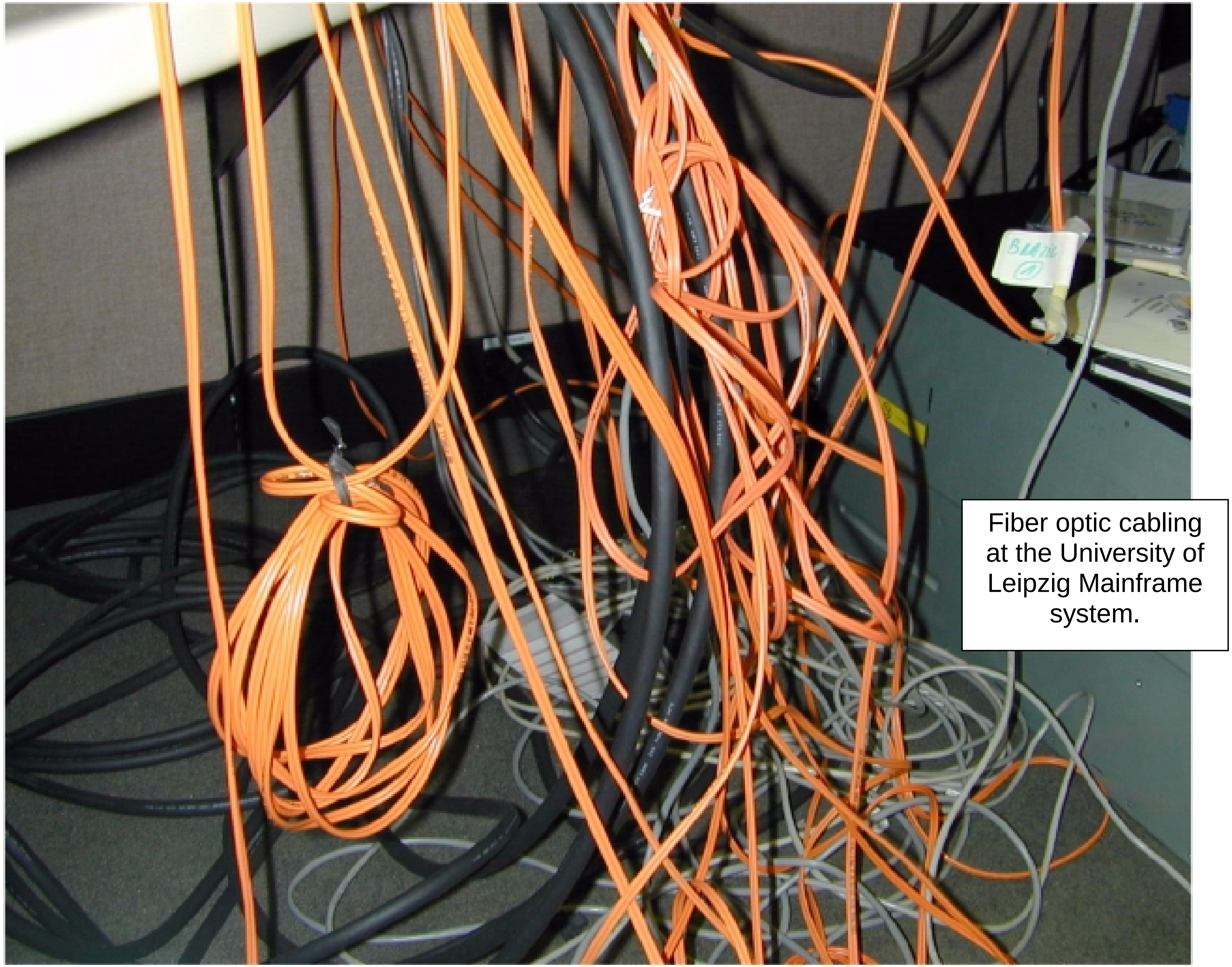


Optical interface

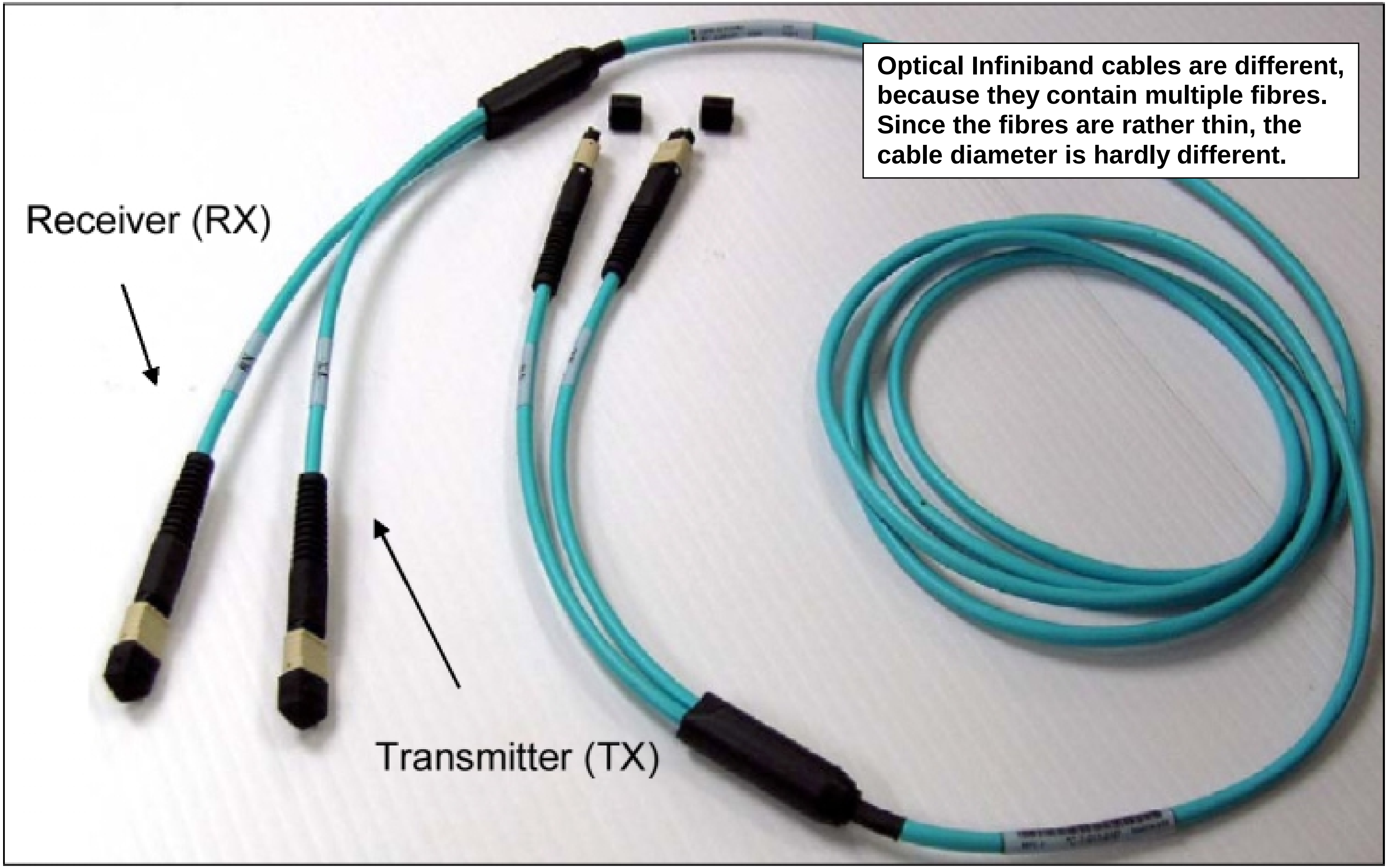
Each HCA2-O fanout card has two ports to connect either an optical cable or a copper cable. The optical cable is used for coupling link connectivity, while the copper cable is used for I/O cage connectivity.



Fibre channel Adapter Kabel as used for FICON connections

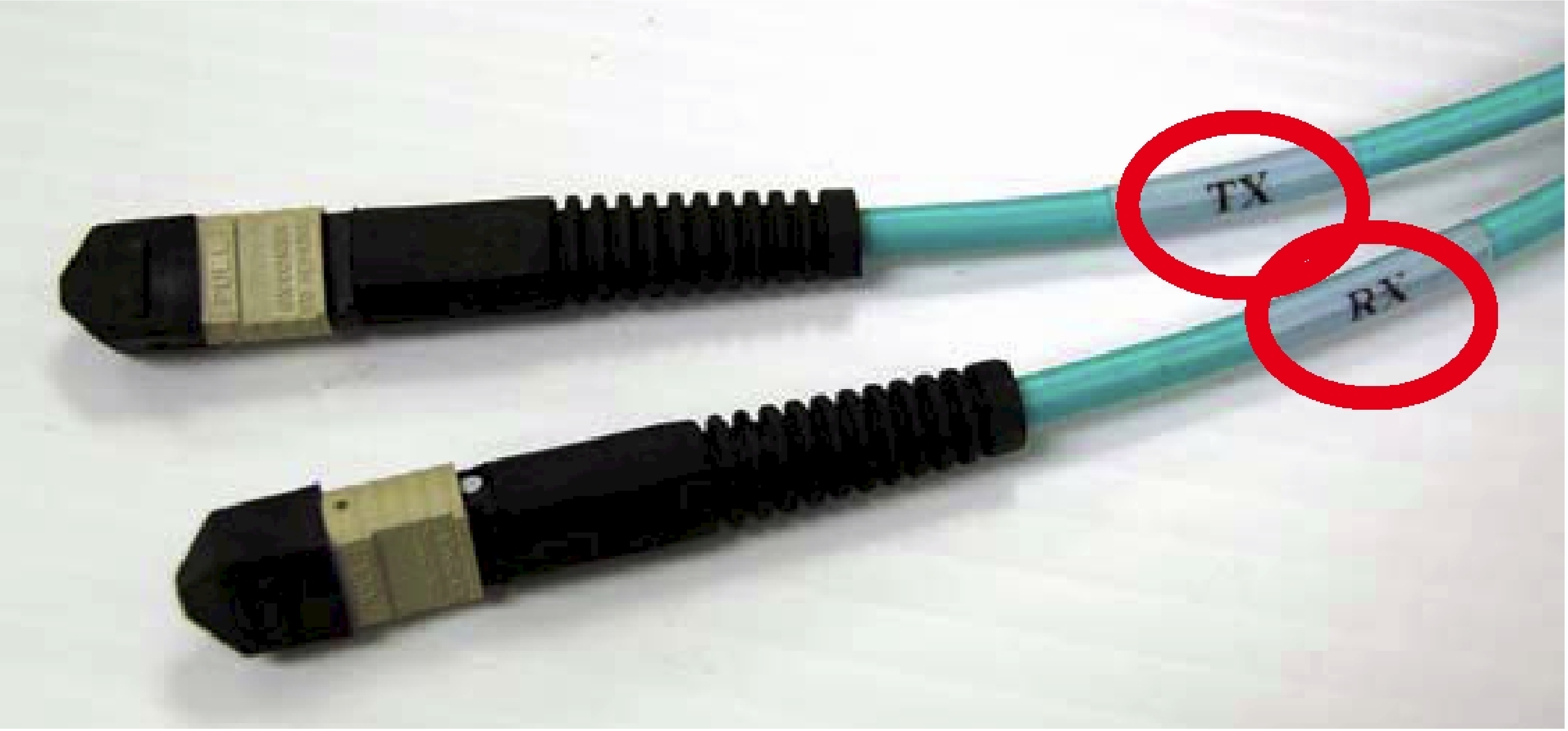


Fiber optic cabling at the University of Leipzig Mainframe system.

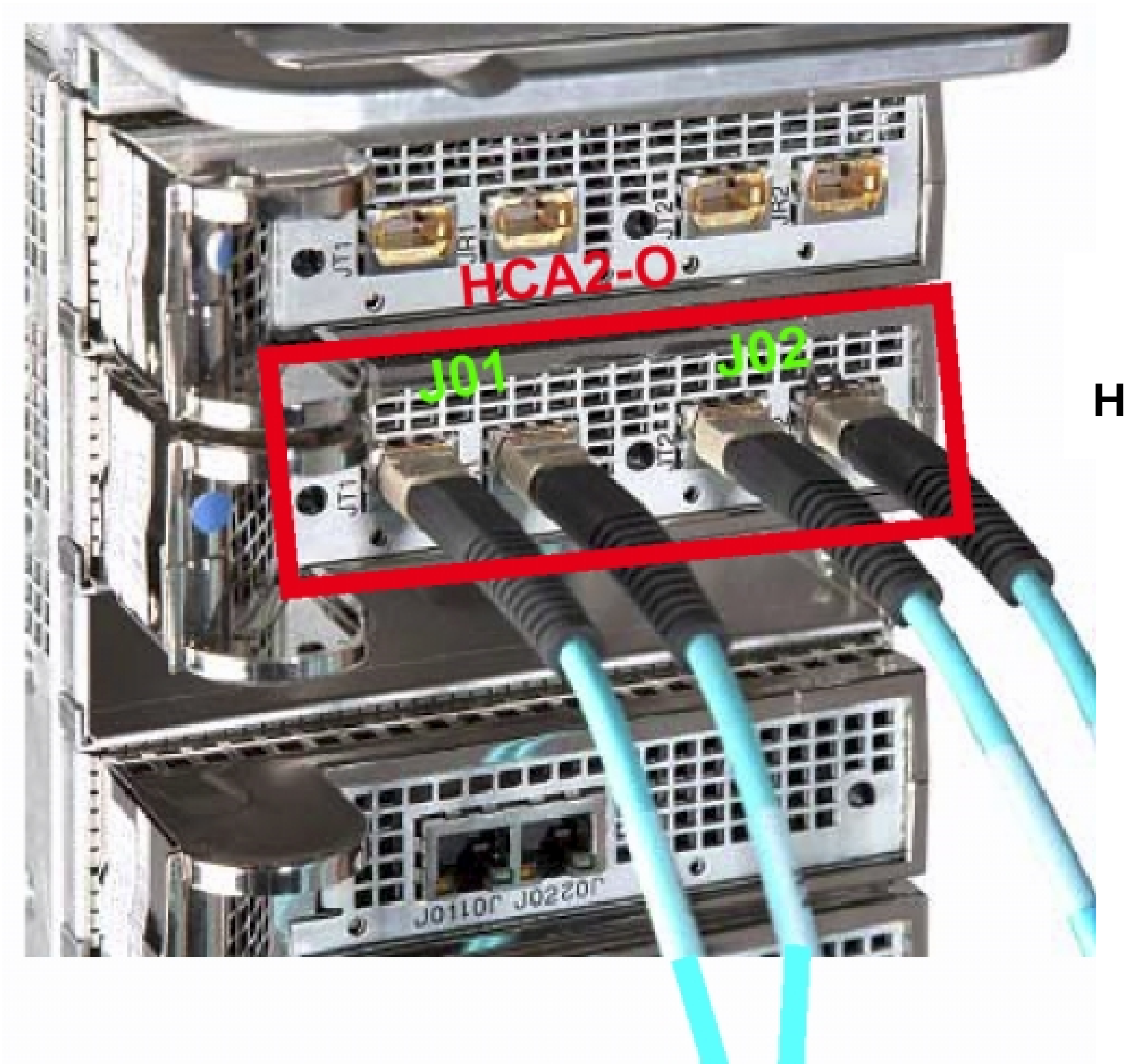


Optical Infiniband cables are different, because they contain multiple fibres. Since the fibres are rather thin, the cable diameter is hardly different.

Optical Infiniband cable

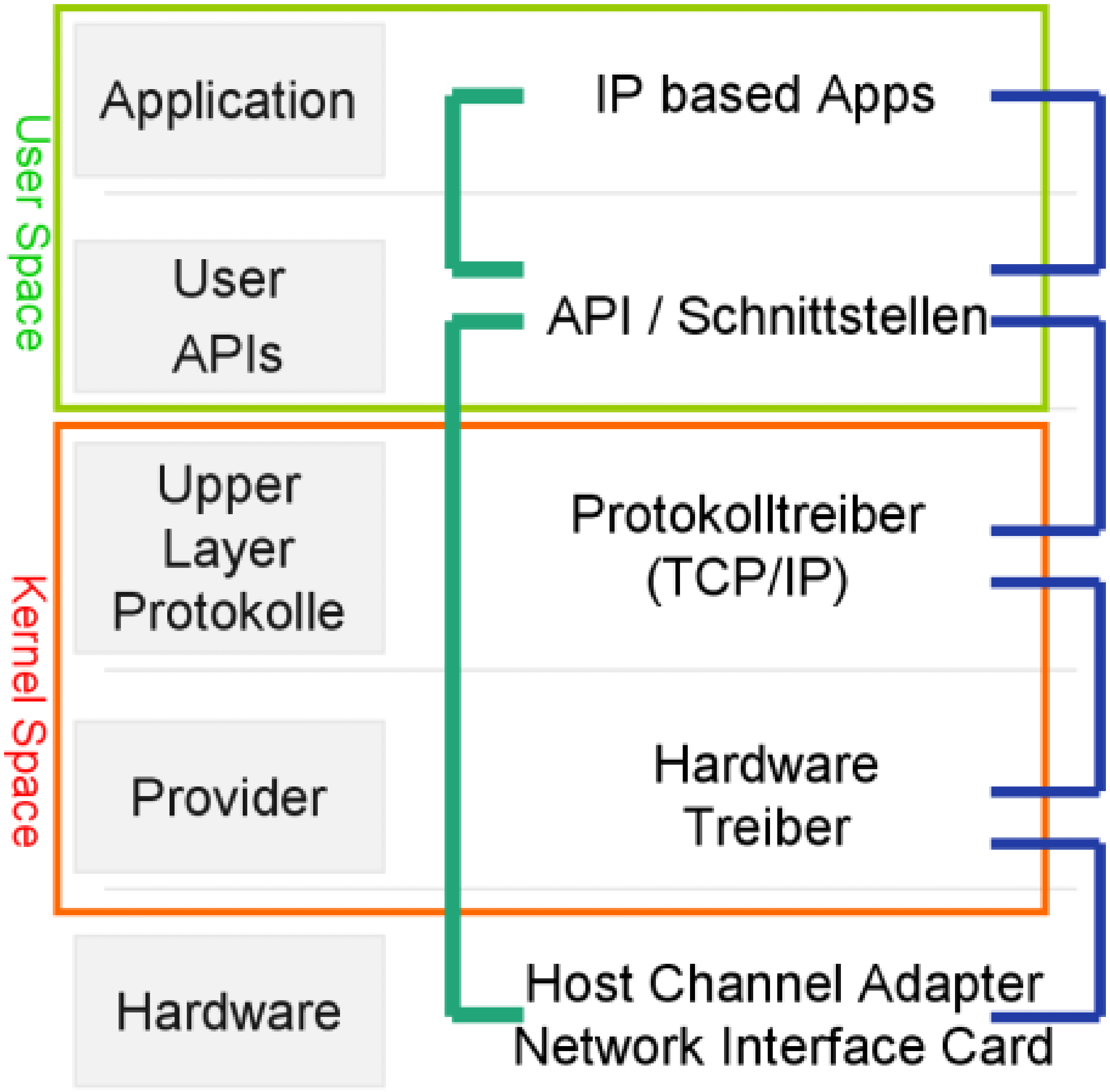


The transmitter cable is labeled TX, the receiver cable RX



Infiniband cable connection to an HCA2-O fanout card

	Gigabit-Ethernet	InfiniBand
Datenrate	1 Gbit/s (theoretisch)	2.5 Gbit/s (theoretisch) *Single Data Rate*
Latenz	Ca. 60 µs	Ca. 1,5 µs
Programmierung	Sockets (TCP/UDP)	Connection, Datagram, Remote DMA
Vorteile	Billig Etabliert	Bidirektionale Übertragung Kanalbündelung (1x, 4x, 12x, …) Protokollstack in **Hardware**
Nachteile	Protokollstack in **Software**	Teurer



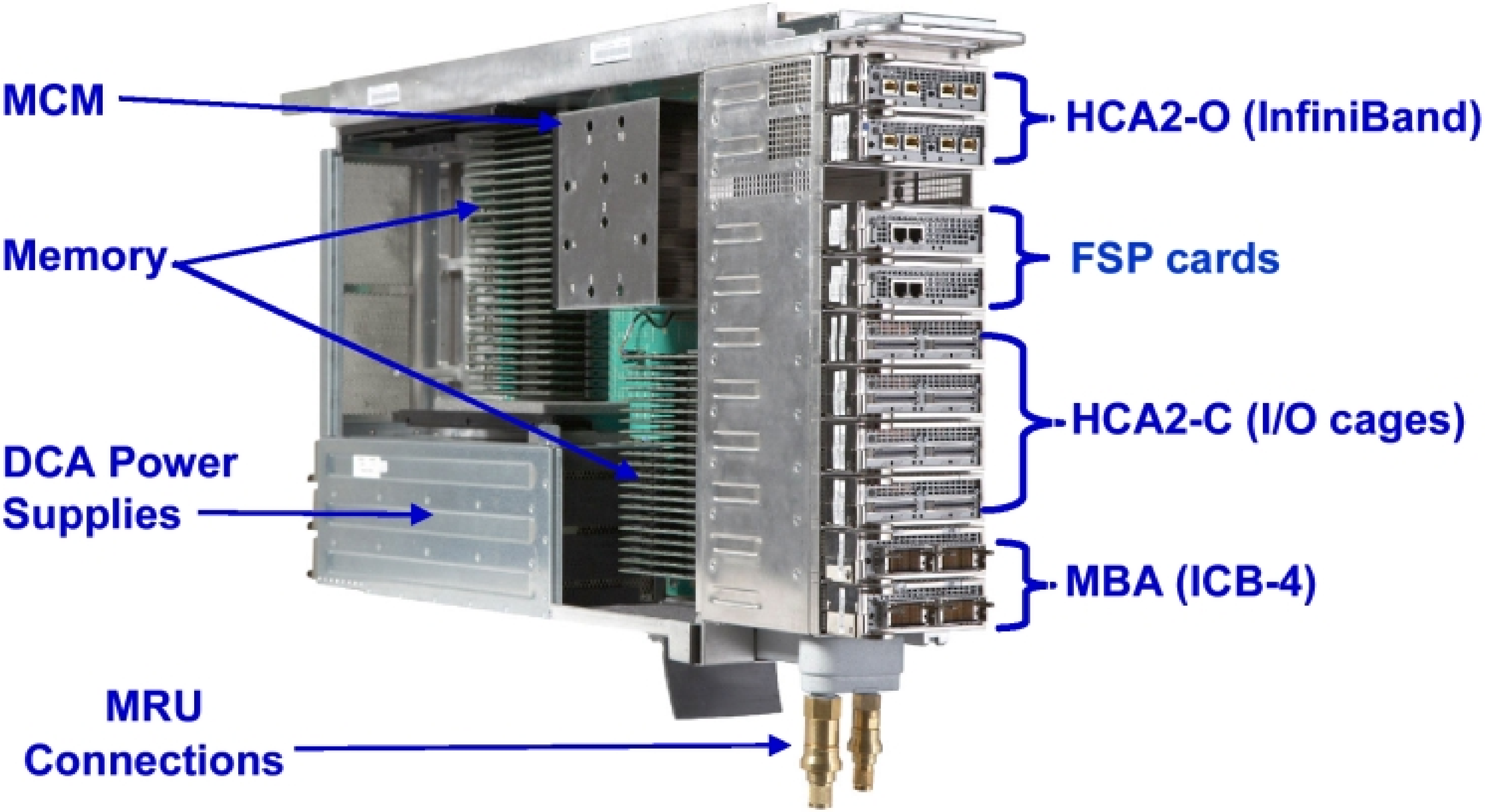
Ansprüche an das Netzwerk:

- **Schneller Datentransfer**
- **Entlastung der CPU**
- **Schlankes Protokoll**
- **Hardware-Management des Datentransfers**

Geringe Latenz

— Ethernet: TCP/IP

— InfiniBand: Kernel Bypass



z10 EC book structure and components (1)

HCA - Host Channel Adapter

This is an example of how and where different fanout cards are installed. The quantities installed will depend on the actual I/O configuration

z10 EC book structure and components (2)

The z10 EC has fanout cards with associated connectors residing on the front of the book package. There are three kinds of fanout cards:

- An InfiniBand HCA2-C (copper) fanout card supporting ESCON, FICON, OSA, ISC-3, and Crypto Express2 cards in the I/O cages.
- An InfiniBand HCA2-O (optical) fanout card supporting up to 6 GByte/s z10 EC to z10 EC and up to 3 GByte/s z10 EC to System z9 Parallel Sysplex connections.
- A MBA fanout card used for ICB-4 connections only.

The z10 EC supports up to eight I/O fanout card (HCA-C, HCA-O or the new MBA) for each book, with a maximum of 24 for a four book system. Each fanout card comes with two ports giving a maximum of 48 ports for I/O connectivity.

An additional two FSB cards attach to a local ethernet contained within a z10 EC system and are used for administrative and diagnostic purposes only.

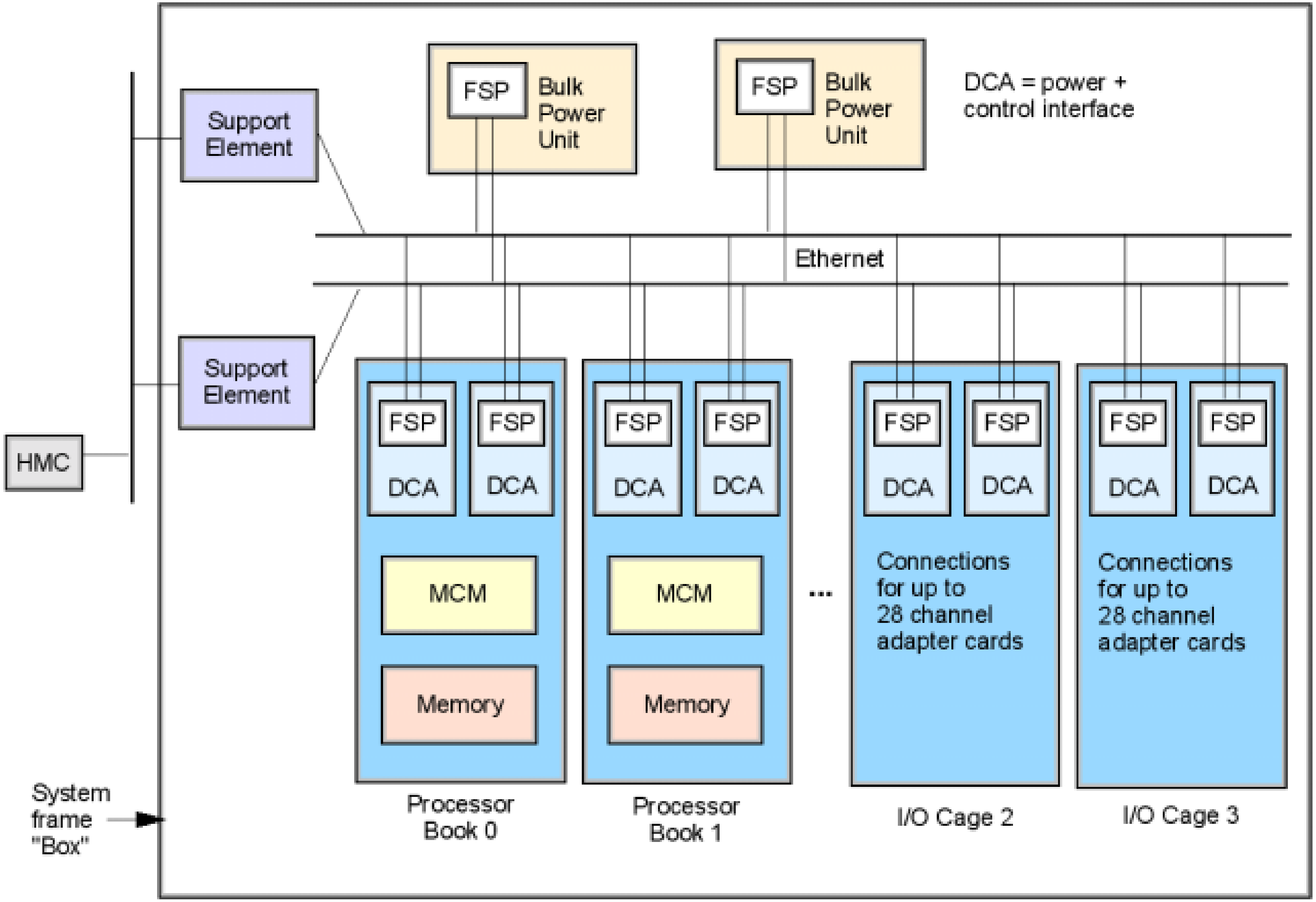
z10 EC book structure and components (3)

The z10 EC exploits InfiniBand (IFB) connections to I/O cages driven from the Host Channel Adapter (HCA2-C) fanout cards that are located on the front of the book. The HCA2-C fanout card is designated to connect to an I/O cage by a copper cable. The two ports on the fanout card are dedicated to I/O. This is different from the z9 EC which uses Self Timed Interface (STI) connections driven from the Memory Bus Adapters (MBAs) to connect to the I/O cages.

The HCA2-C fanout card is designated to connect to another z9 or z10 system.

For the z10 EC server there are up to eight fanout cards per book, each driving two IFB cables (two ports), resulting in up to 16 IFB connections per book (16 STI connections with the z9 EC server). All 16 InfiniBand (IFB) connections work with a data rate of 6 GByte/s. In a system configured for maximum availability, alternate paths will maintain access to critical I/O devices, such as disks, networks, and so on.

There is a maximum of 24 fanout cards and 48 ports for a four book system.



Flexible Service Processor (FSP)

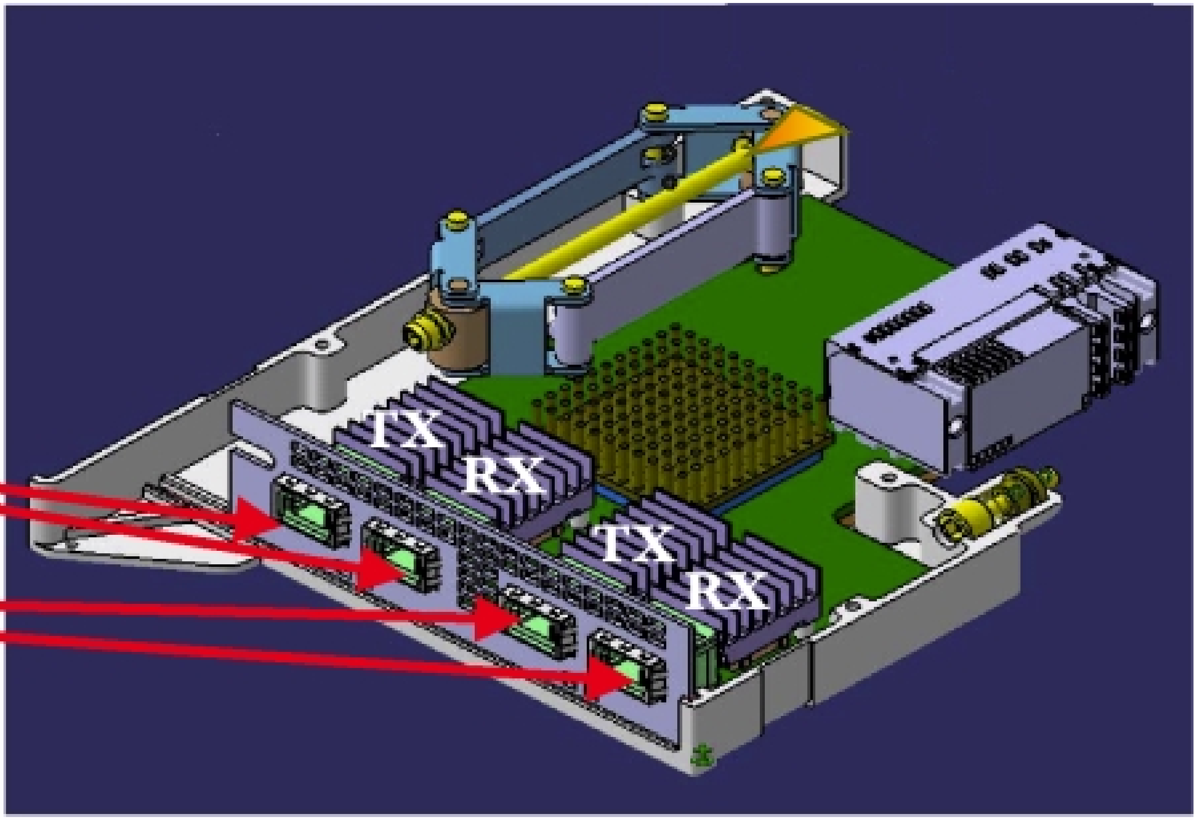
The Flexible Service Processor (FSP) cards use Ethernet connectors for internal System services. This is firmware that provides diagnostics, initialization, configuration, run-time error detection and correction. FSP connects via the Support Element (Thinkpad within the z10 frame) to the Hardware Management Console (HMC).

The FSB Ethernet is contained within the z10 EC frame and interconnects all books, all I/O cages, and the two Hardware Support Elements (Thinkpads). For reliability and availability reasons all connections are available in duplicate. The support elements connect via another Ethernet do the Hardware Management Console (HMC). The HMC is used by the system administrator to manage the system (e.g. perform IPL). The Ethernet connection between the Support Element and the HMC is usually dedicated to this single purpose.

The FSB is used for an ethernet network that connects elements within the z10 EC system only, including the Thinkpad Service Element, and is duplicated for reliability reasons.

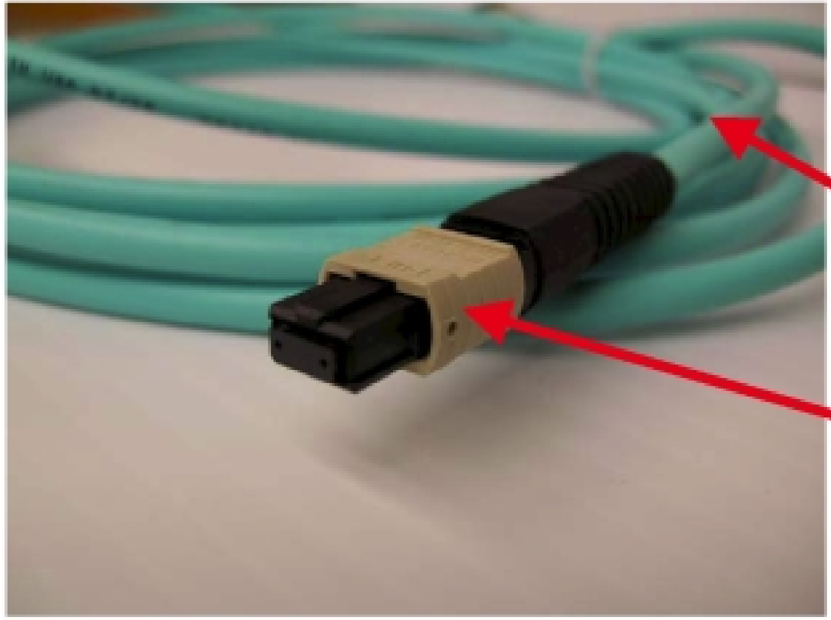
The Hardware Management Console (HMC) is not directly connected th this internal ethernet. Instead it is connected via a separate ethernt to the Thinkpad service element. The HMC can be routed at great distance from the z10 system.

HCA2-O fanout



Port 0

Port 1



OM3 cable (aqua cladding)
2000 MHz-km, 50 micron multimode fiber

MPO connector
150 meters point-to-point

I/O cage

The z9 and z10 EC has a a minimum of one and a maximum of three I/O cages

Each I/O cage can house up to seven I/O domains. Each I/O domain may have up to 4 I/O slots, making a total of 28 I/O slots per cage for the z9 and z10. Each slot can accept a single I/O card, for example a FICON Express card.

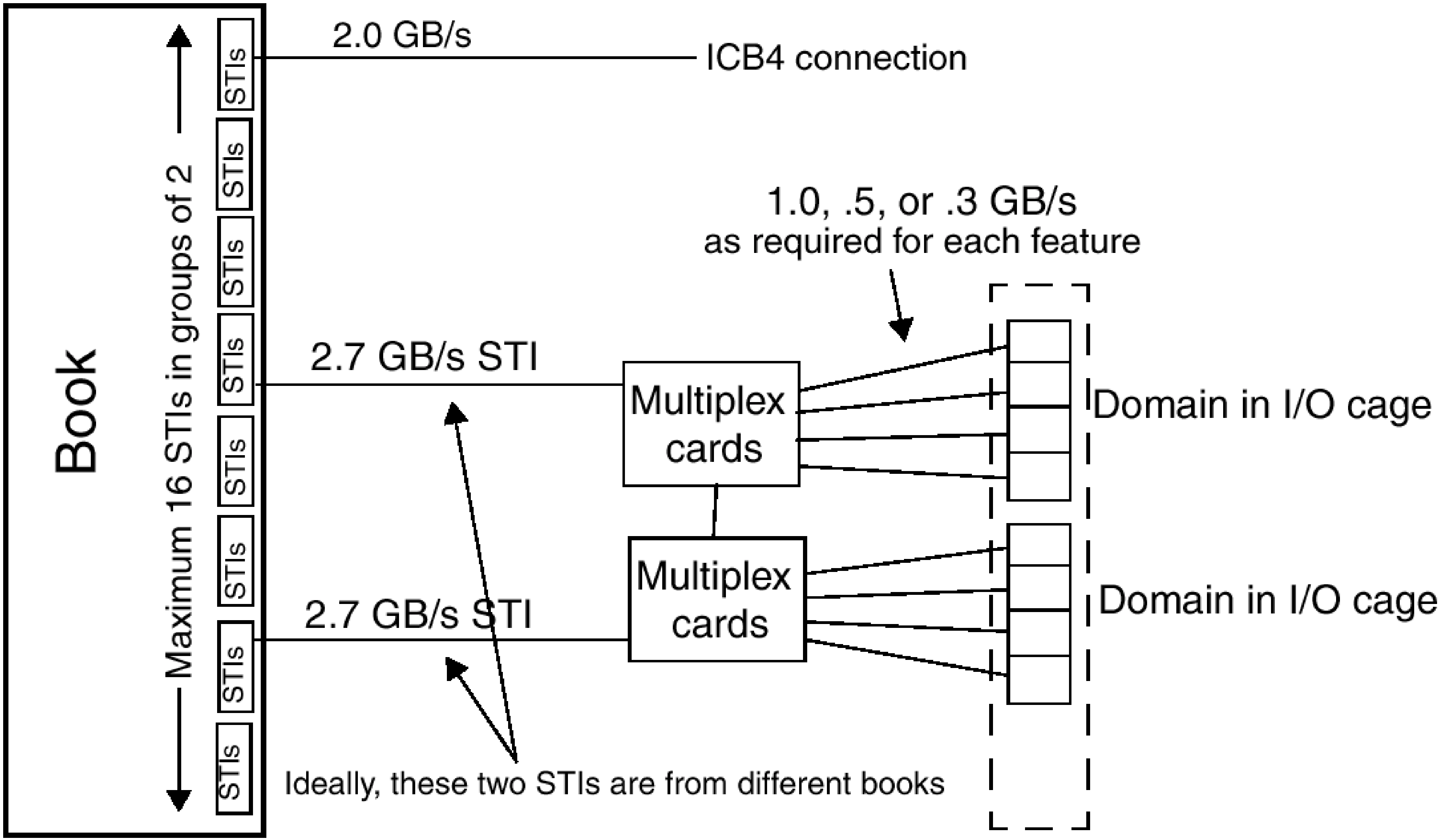
It is possible to populate the 28 I/O slots in an I/O cage with any mix of these cards:

- **ESCON**
- **FICON Express4, FICON Express2, or FICON Express**
- **OSA-Express3 and OSA-Express2**
- **Crypto Express2**
- **Coupling links**

Each I/O domain is connected to the CEC

- **on the z9 via a Self-Timed Interface – Multiplexer (STI-M) card and a 2.5 GByte/s STI cable. Seven STI-M cards and STI cables are needed to support a full I/O cage.**

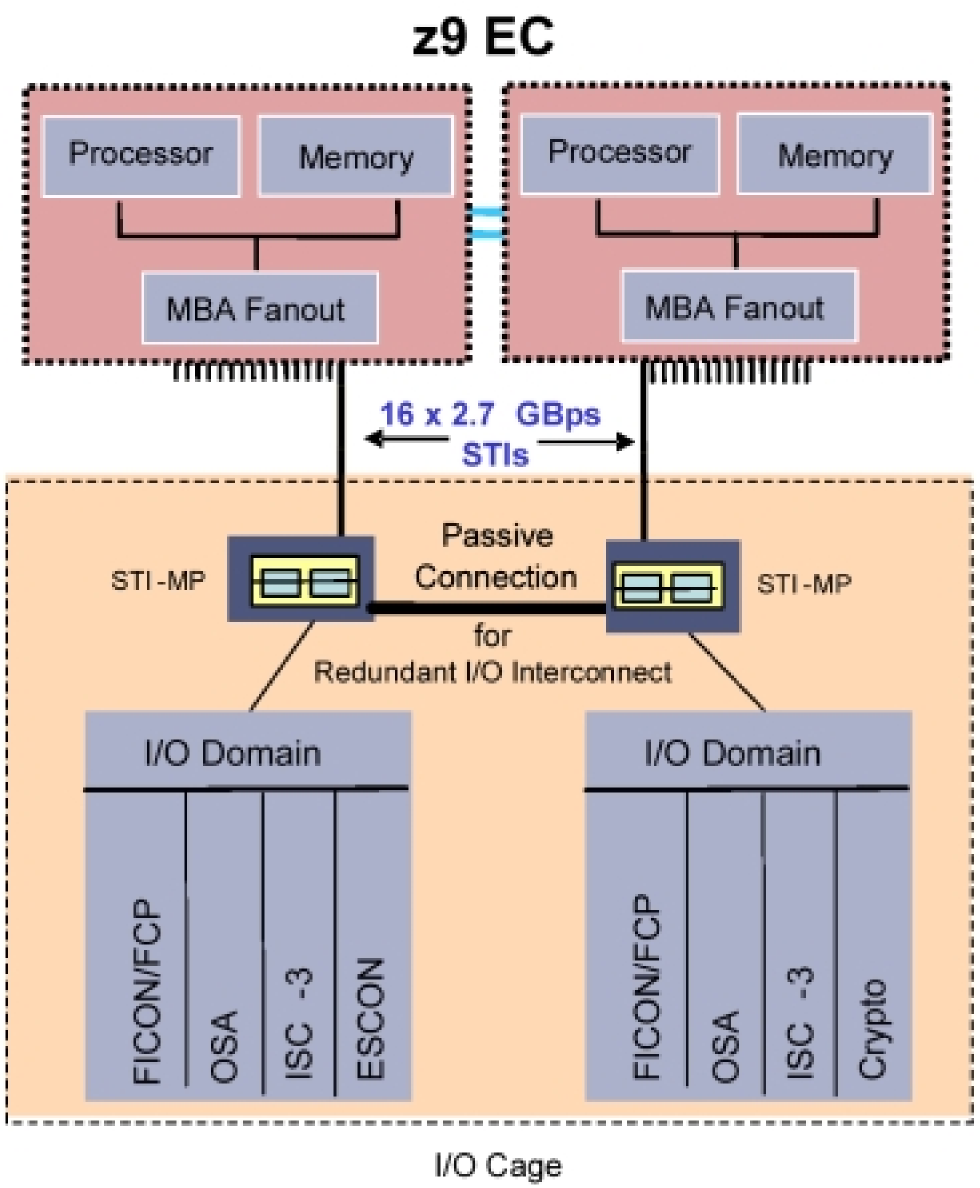
- **on the z10 via a Infiniband I/O Interconnect (IFB-MP) card and a 6.0 GByte/s Copper-Infiniband cable. Seven STI-MP cards and Infiniband cables are needed to support a full I/O cage.**

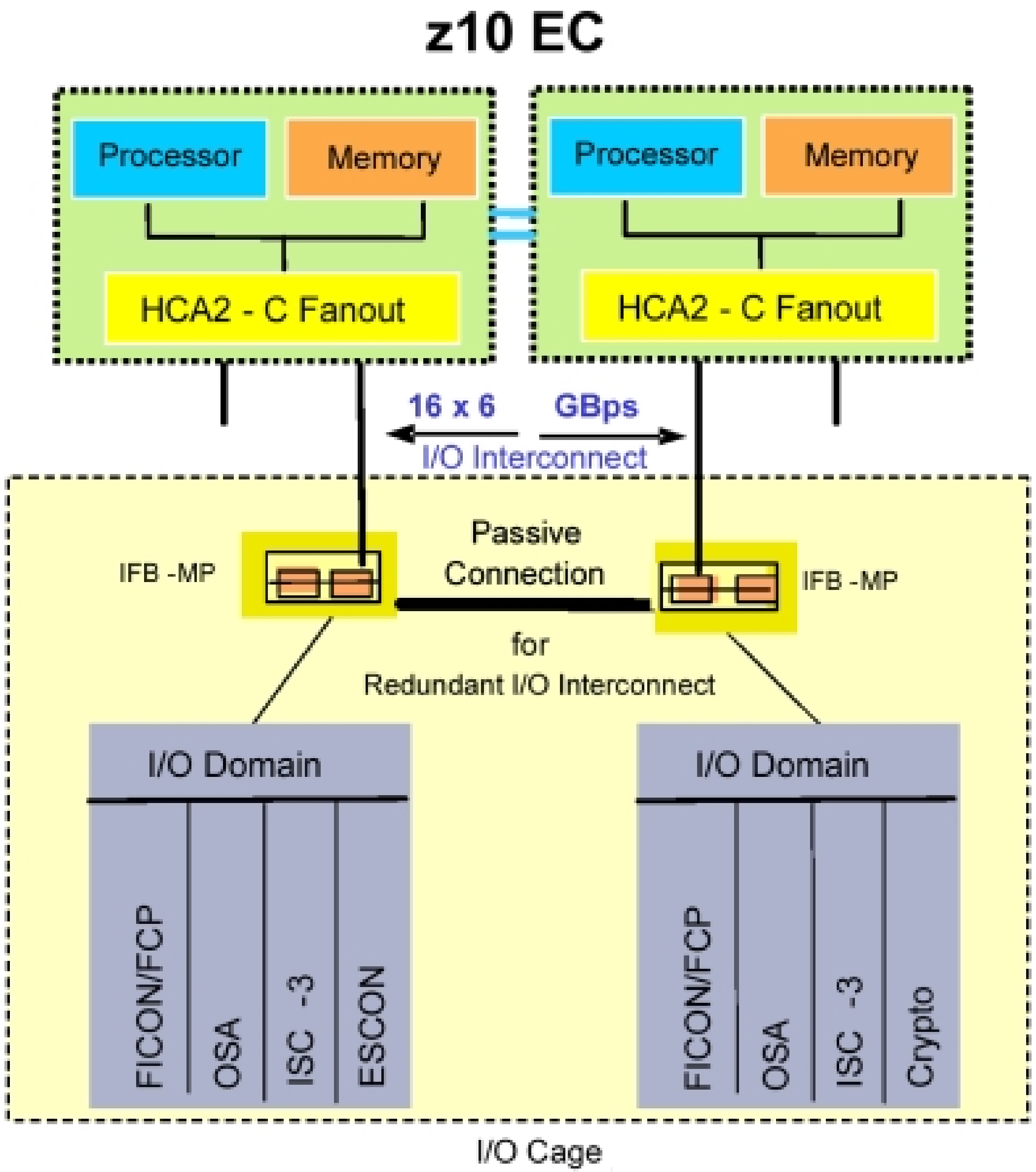

Book
Maximum 16 STIs in groups of 2
STIs
STIs
STIs
STIs
STIs
STIs
STIs
STIs
2.0 GB/s
ICB4 connection
1.0, .5, or .3 GB/s
as required for each feature
2.7 GB/s STI
Multiplex cards
Domain in I/O cage
2.7 GB/s STI
Multiplex cards
Domain in I/O cage
Ideally, these two STIs are from different books

The figure above shows the interconnection of a book via 2 STI cables to an I/O cage in a z9 system. The 28 card slots in an I/O cage are subdivided into 7 domains of 4 cards each. An additional 4 slots are available for STI Multiplexor cards; each STI cable connects to an STI multiplexor (STI-MP) card.

As shown, two multiplexor cards are interconnected to facilitate a redundant I/O interconnect.

Ideally the two STI cables shown above would originate from different books.


z9 EC
Processor
Memory
Processor
Memory
MBA Fanout
MBA Fanout
16 x 2.7 GBps
STIs
STI -MP
Passive
Connection
for
Redundant I/O Interconnect
STI -MP
I/O Domain
FICON/FCP
OSA
ISC -3
ESCON
I/O Domain
FICON/FCP
OSA
ISC -3
Crypto
I/O Cage


z10 EC
Processor
Memory
Processor
Memory
HCA2 - C Fanout
HCA2 - C Fanout
16 x 6 GBps
I/O Interconnect
IFB -MP
Passive
Connection
for
Redundant I/O Interconnect
IFB -MP
I/O Domain
FICON/FCP
OSA
ISC -3
ESCON
I/O Domain
FICON/FCP
OSA
ISC -3
Crypto
I/O Cage

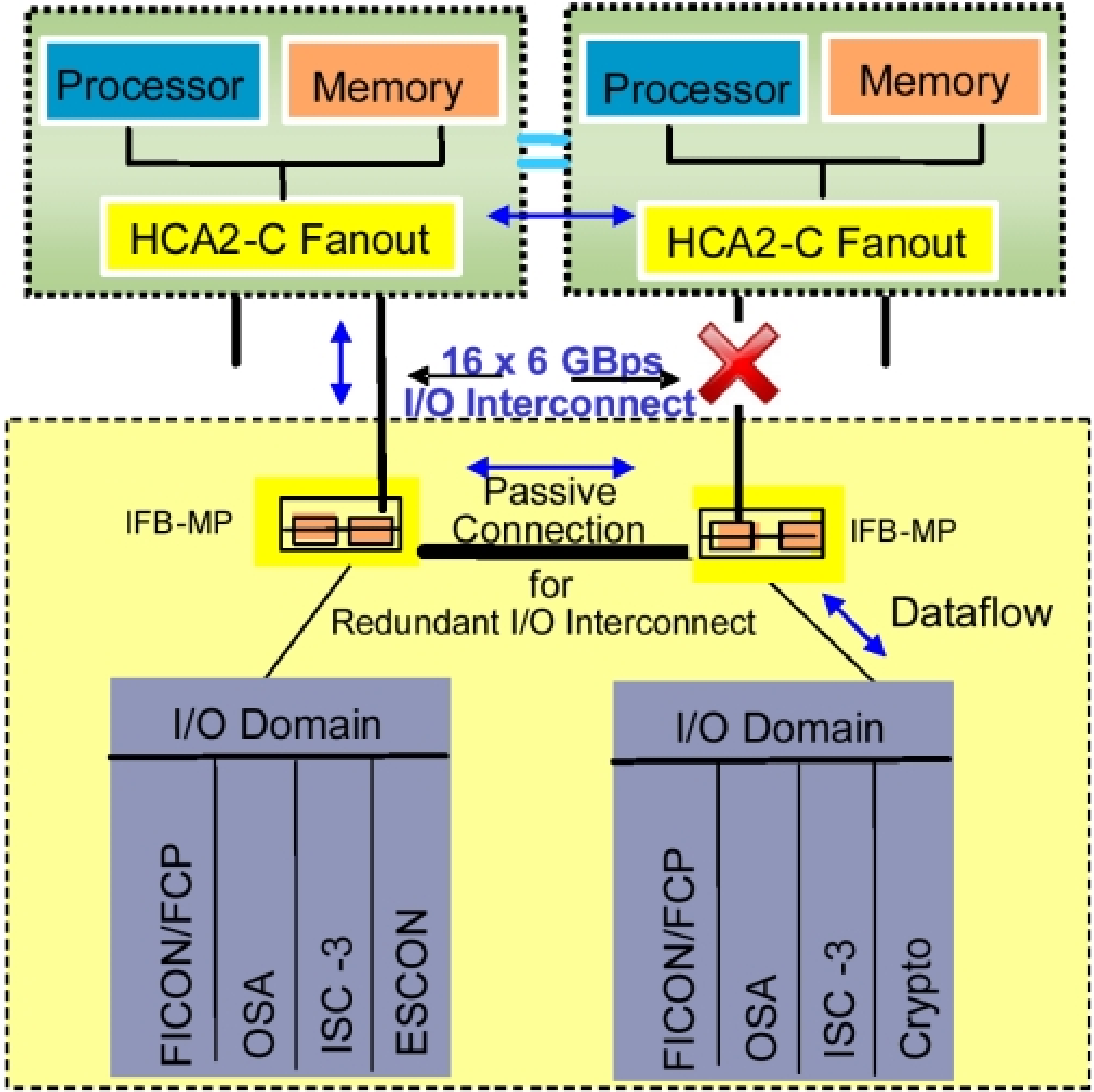
The figure above compares the connection difference between a z9 and a z10 system. The MBA adapter is replaced by a HCA adapter. The STI cable is replaced by a copper Infiniband cable, and the STI multiplexor (STI-MP) is replaced by an Infiniband multiplexor (IFB-MP).

In previous systems, an STI or eSTI connection with data rates of 2 GBps and 2.7 GBps respectively was used. The connection was made between the MBA, which is installed in a fanout slot in the front of the book, to the STI-MP (Self-Timed Interconnect - Multiplex) cards installed within the I/O cages.

As I/O cards continue to support higher data transfer rates to the devices, the connection between the I/O cards and the CEC cage needs to provide a higher data rate as well. The connectivity to the I/O cages (I/O domains) in the System z10 is supported by InfiniBand technology, which provides a data rate of 6 GBps.

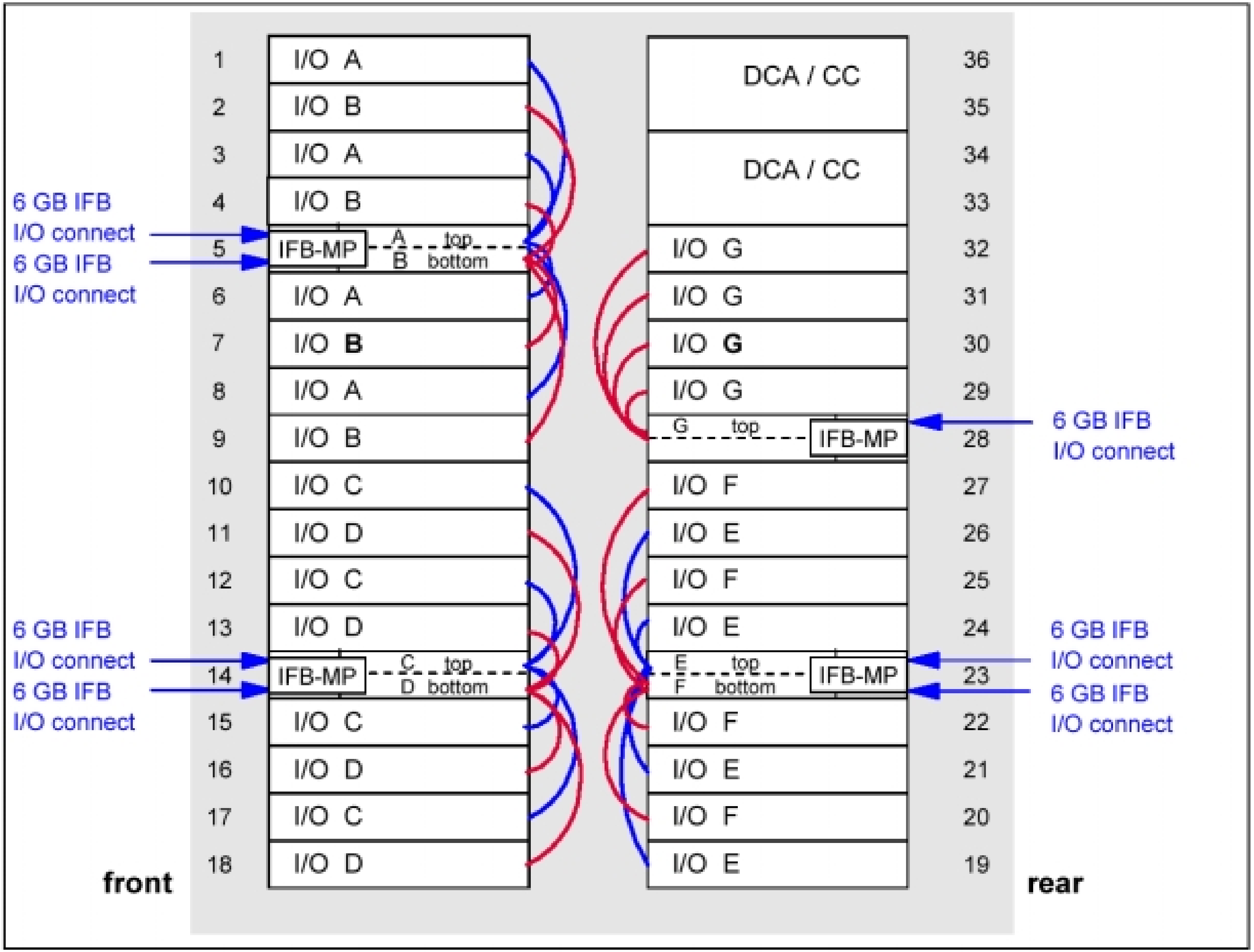
With the System z10 an HCA2-C fanout in the front of the book connects via an IFB copper cable to an IFB-MP (InfiniBand - Multiplex) card, installed in the I/O cages

On System z9, two STI-MP cards are installed in one STI mother card. For the System z10 the STI mother cards are still used and are populated with the IFB-MP cards.

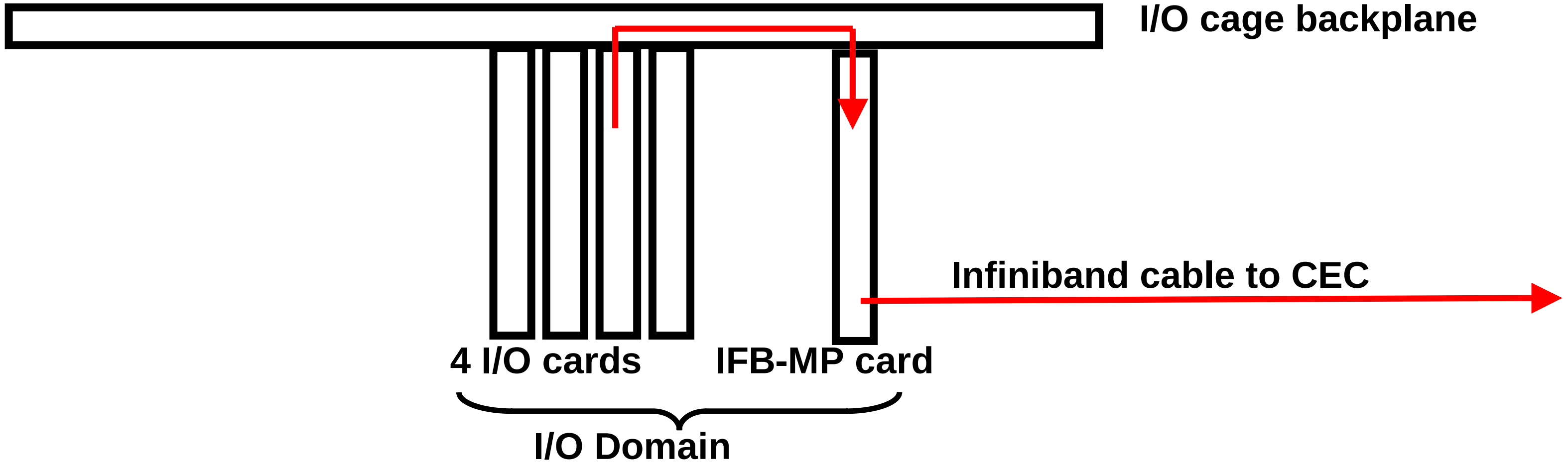

Processor
Memory
HCA2-C Fanout
Processor
Memory
HCA2-C Fanout
16 x 6 GBps
I/O Interconnect
Passive
Connection
for
Redundant I/O Interconnect
IFB-MP
IFB-MP
Dataflow
I/O Domain
FICON/FCP
OSA
ISC -3
ESCON
I/O Domain
FICON/FCP
OSA
ISC -3
Crypto

There is a passive connection in the Infiniband multiplexor (IFB-MP) card to provide the redundancy for the I/O interface. This allows for concurrent repairs against the cabling or the HCA-2C fanout.

Typically the two IFB cables originate from different books to improve redundancy.

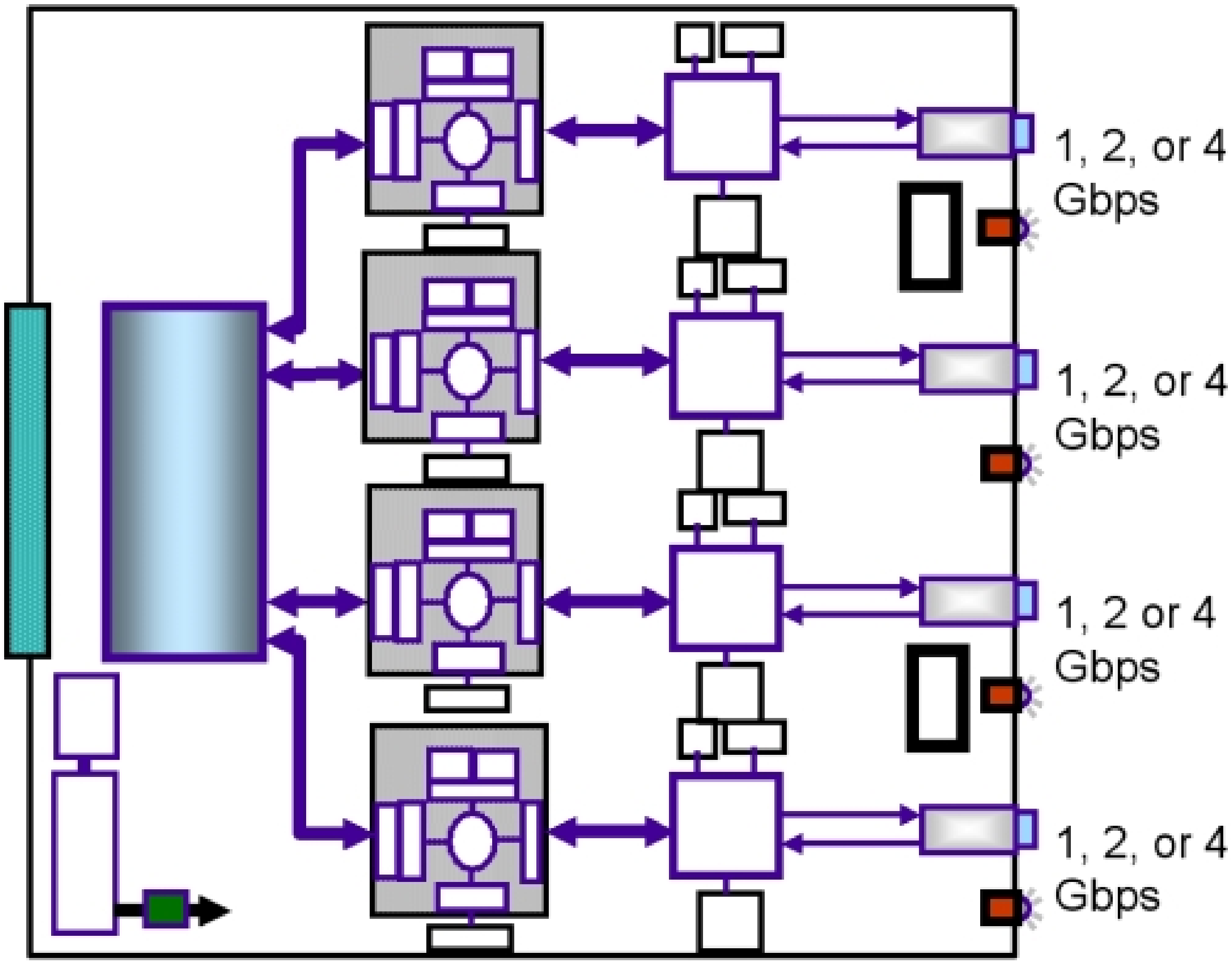


The 7 different I/O domains (A,B,C,D,E,F,G) and the InfiniBand MultiPlexer (IFB-MP) of a z10 EC I/O cage are shown. Each domain has 4 I/O card slot positions, for a total of 28 slots. An additional 4 card positions accept IFM-MP cards. They connect both to the CEC cage and to the I/O cards in the I/O cage.

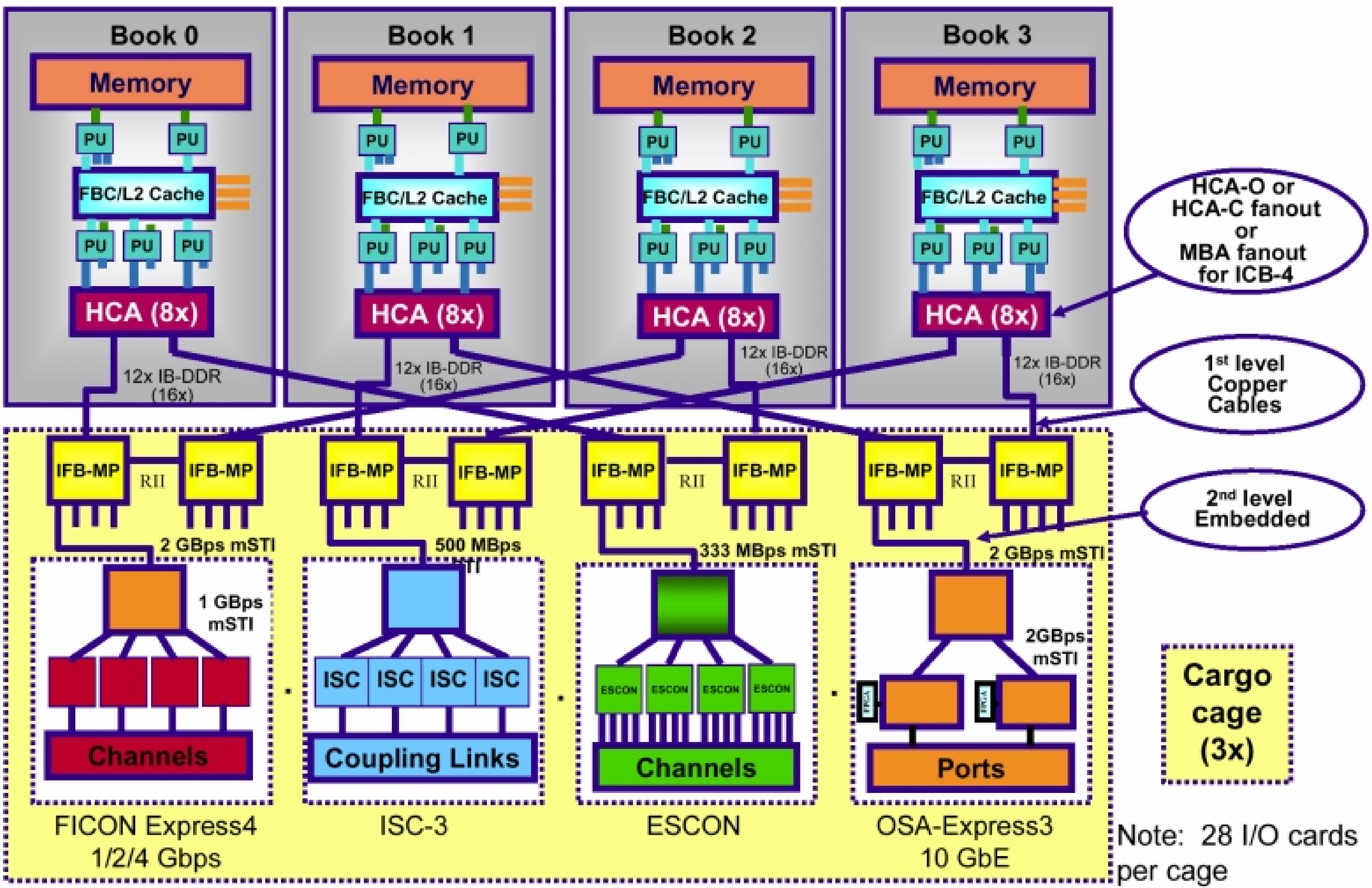


The IFB-MP (InfiniBand - Multiplex) card shown above provides the basic connection between the CEC cage and the I/O cages. Data received on any I/O card (e.g. a “FICON Espress 4” I/O card as shown below) travels through the I/O cage backplane to the IFB-MP card.

The IFB-MP card is connected to the HCA2-C fanout card in a book of the central the processor complex (CEC) via an Infiniband internal copper cable. The link between the IFB-MP card and the HCA2-C fanout is operated with a link speed of 6 GByte/s.



As an example of an I/O card, the configuration of a “FICON Espress 4” I/O card is shown. It attaches up to 4 duplex FICON cables.



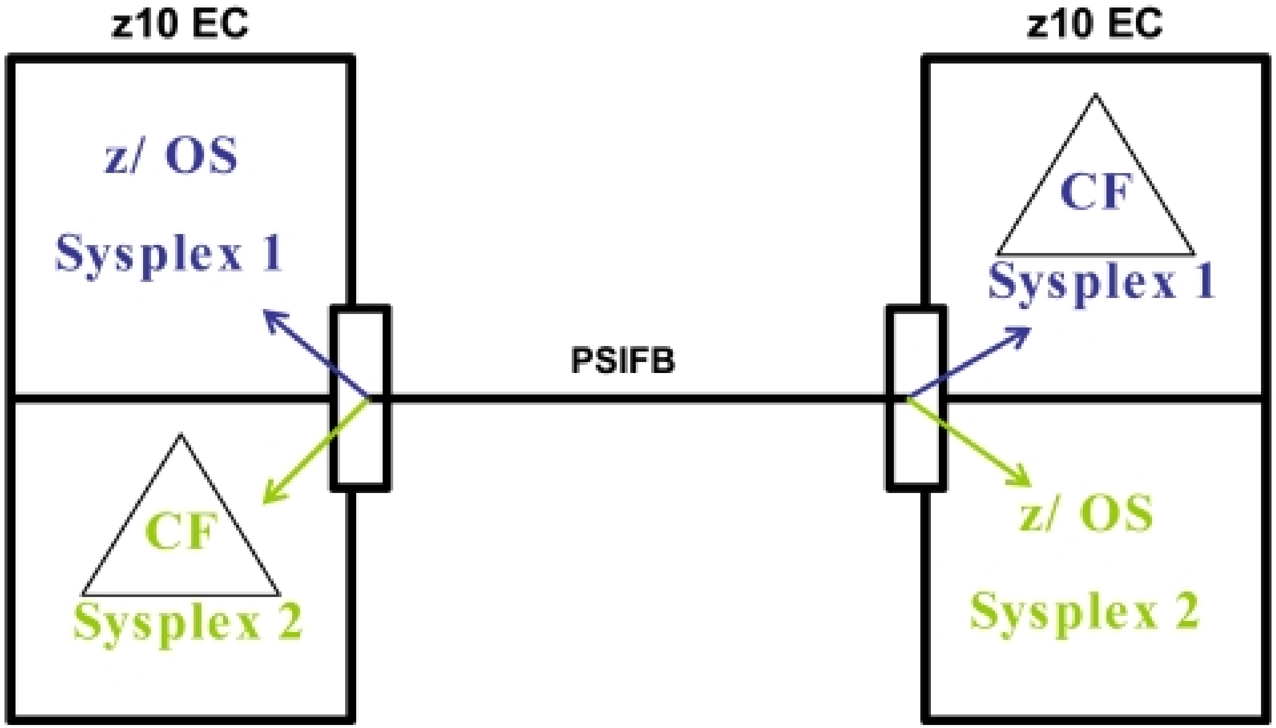
Note: 28 I/O cards per cage

Shown above are the interconnections of the 4 books of a z10 EC system to one of the I/O cages (also called a cargo cage). Each book has up to 8 HCA adapter Each HCA adapter accepts 2 Infiniband copper cables, for a total of 16 cables. Each copper cable has 12 Infiniband physical lanes with 24 wire pairs. Each wire pair operates at 5 Gbit/s DDR (Double Data Rate) for a total of 60 Gbit/s per cable.

Pairs of Infiniband multiplexors (IFB-MP) are coupled with a Remote I/O Interconnect(RII).

A number of differen I/O card types are available. The most important types are:

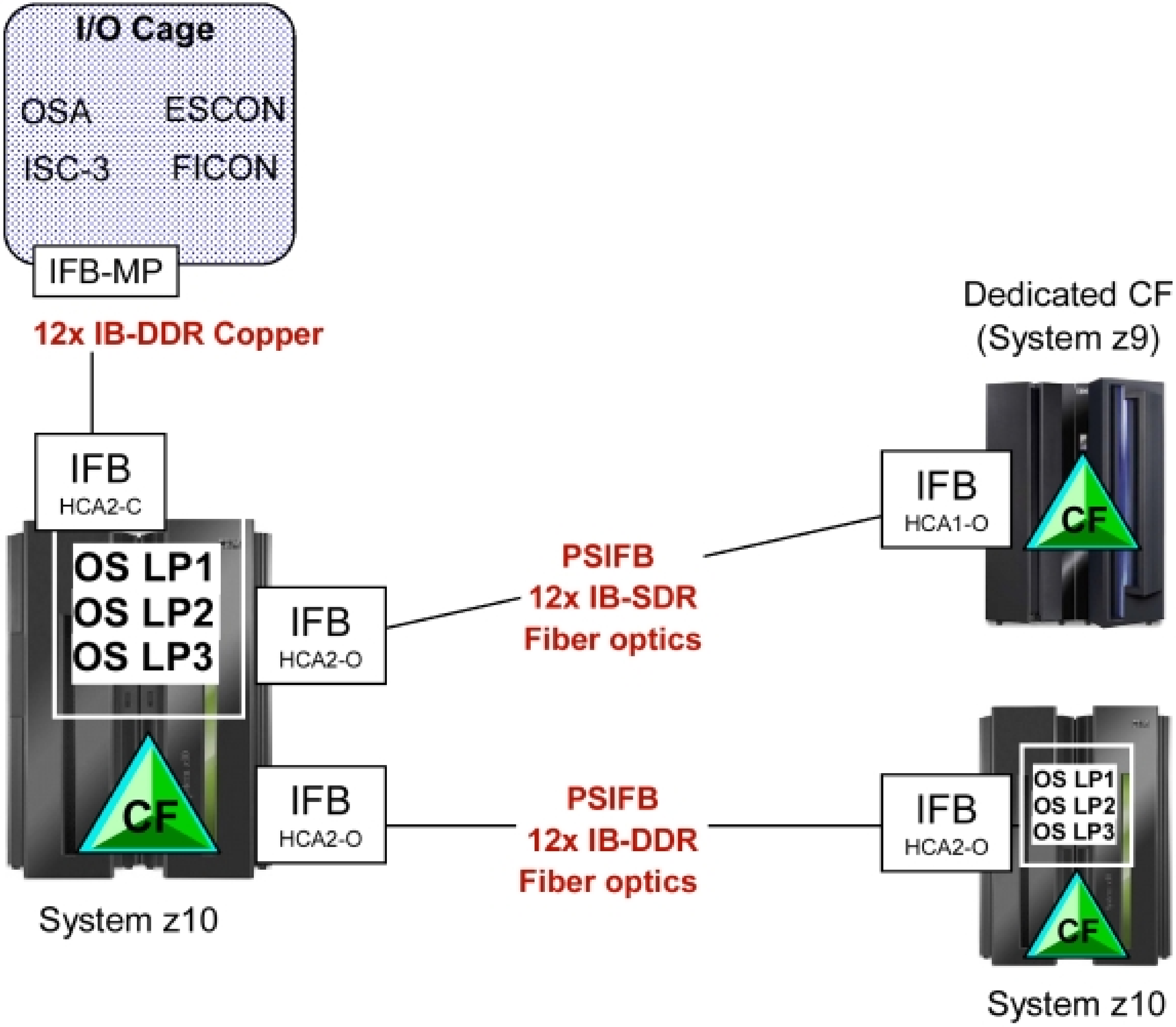
- **FICON Express for the attachment of up to 4 FICON fibre optical links with a data rate of up to 4 Gbit/s each**
- **ESCON for the attachment of the older ESCON fibre optical links .**
- **ISC-3 fibre optic connections to another system z.**
- **Several types of OSA Express for multiple Ethernet attachments up to 10 Gbit/s data rate.**



Interconnecting several Mainframe systems

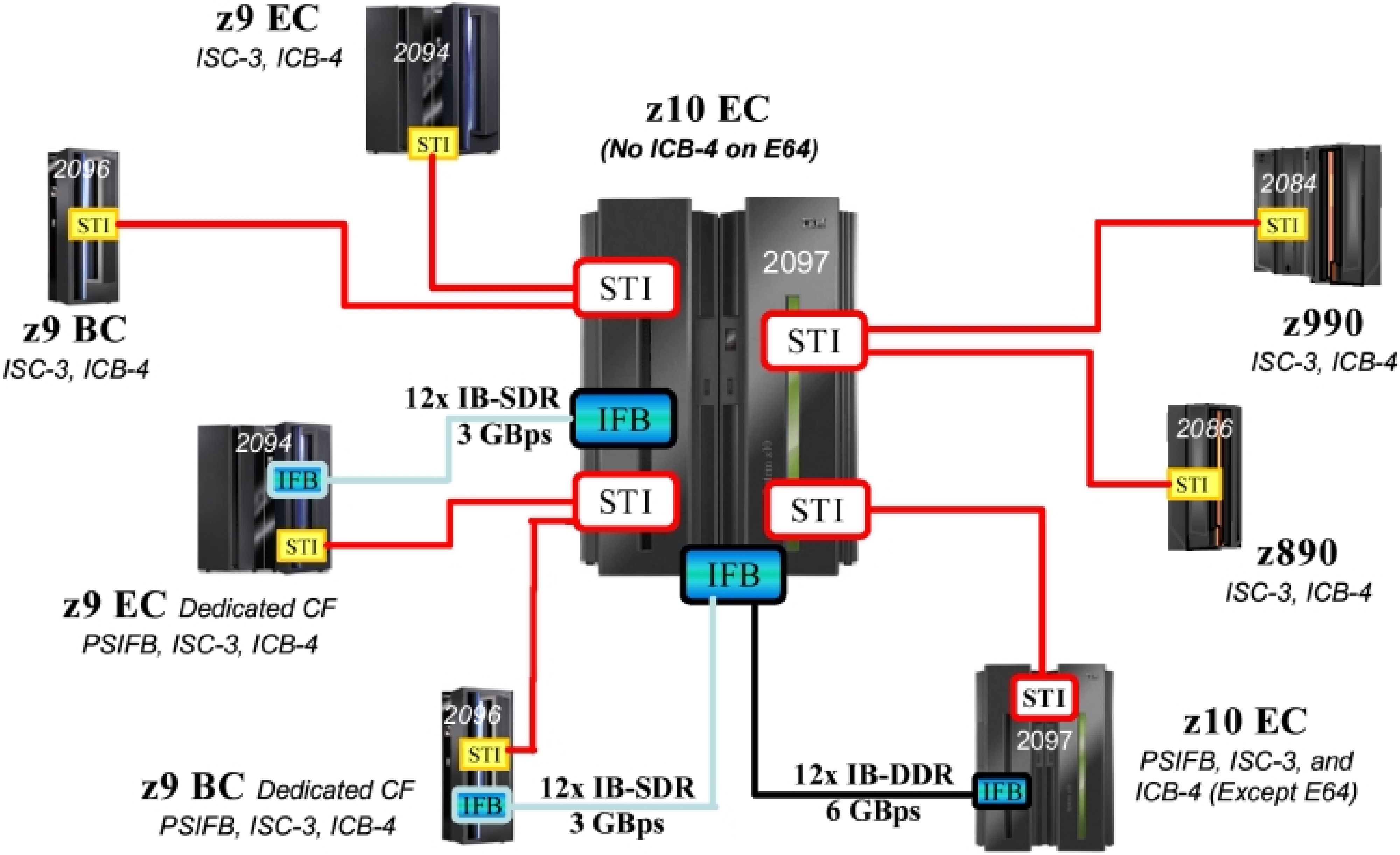
Traditionally there have been several alternatives to interconnect multiple mainframe systems. These were especially the ISC-3 and ICB-4 cabling approaches. Interconnected multiple mainframe systems are called a Sysplex; a Coupling Facility (CF) is ahighly specialised mainframe system..

With the z10 it is possible to use Parallel Sysplex Infiniband (PSIFB).The PSIFB optical cable for 12x IB-DDR is a 12 fiber pair cable (total of 24 optical fibers) of 50 micron multimode optical fiber. The maximum cable length for PSIFB is 150 meters. Single mode cables are used for larger distances.


I/O Cage
OSA
ESCON
ISC-3
FICON
IFB-MP
12x IB-DDR Copper
IFB
HCA2-C
OS LP1
OS LP2
OS LP3
IFB
HCA2-O
CF
IFB
HCA2-O
System z10
PSIFB
12x IB-SDR
Fiber optics
PSIFB
12x IB-DDR
Fiber optics
Dedicated CF
(System z9)
IFB
HCA1-O
CF
IFB
HCA2-O
OS LP1
OS LP2
OS LP3
CF
System z10

Shown above are InfiniBand connection types supported on System z10 and System z9 servers. While the data rate for interconnecting two z10 systems is 6 GByte/s, interconnecting a z10 and a z9 permits only 3 GByte/s.

The same Infiniband protocol is used to alternatively interconnect a book with an I/O cage and to interconnect two systems. The difference is in the cabling: Book to I/O cage uses a copper cab le and the HCA2-C adapter while the inter-sytem connection uses an optical cable and the HCA2-O adapter.

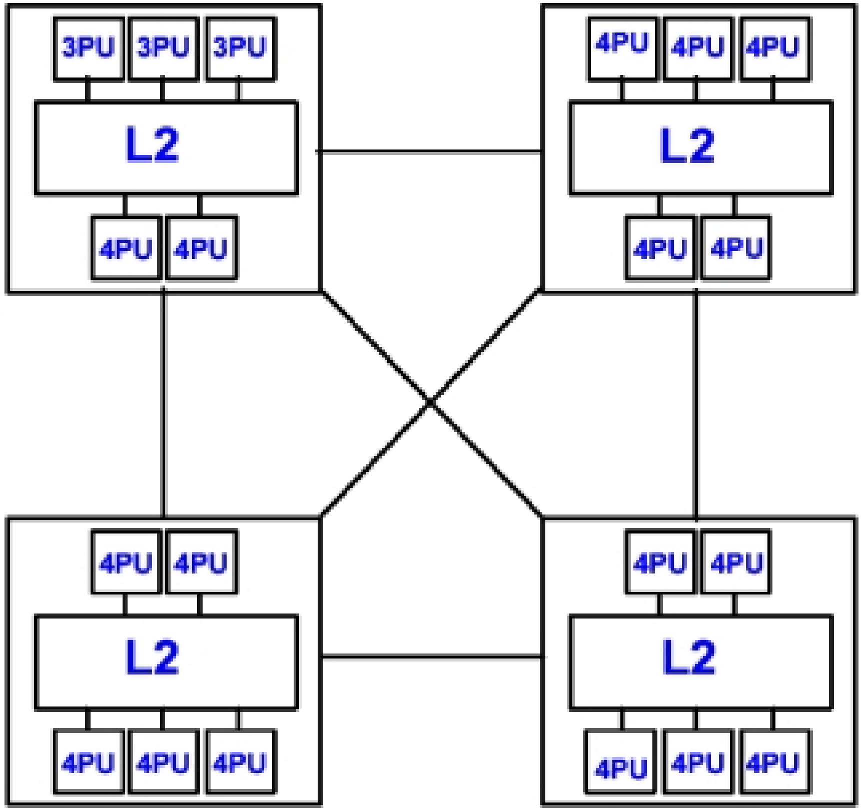

z9 EC
ISC-3, ICB-4
2094
STI
z10 EC
(No ICB-4 on E64)
2096
STI
z9 BC
ISC-3, ICB-4
STI
2097
STI
2084
STI
z990
ISC-3, ICB-4
12x IB-SDR
3 GBps
IFB
2094
IFB
STI
STI
STI
2086
STI
z890
ISC-3, ICB-4
IFB
z9 EC Dedicated CF
PSIFB, ISC-3, ICB-4
2096
STI
IFB
z9 BC Dedicated CF
PSIFB, ISC-3, ICB-4
12x IB-SDR
3 GBps
STI
2097
12x IB-DDR
6 GBps
IFB
z10 EC
PSIFB, ISC-3, and
ICB-4 (Except E64)

Coupling link configuration options for System z10

Interconnections between z10 and z9 systems may use Infiniband. Interconnections to older systems require the traditional ISC-3 and ICB-4 links.

The ICB-4 cable connects to an MBA card. It interconnects two systems with a copper cable and a data rate of 2 MByte/s over a distance of up to 10 meters.

The ISC-3 cable connects to an ISC-3 I/Ocard housed in an I/O cage. It interconnects two systems with an optical cable and a data rate of 2 Mbit/s or 200MByte/s over a distance of up to 100 km.

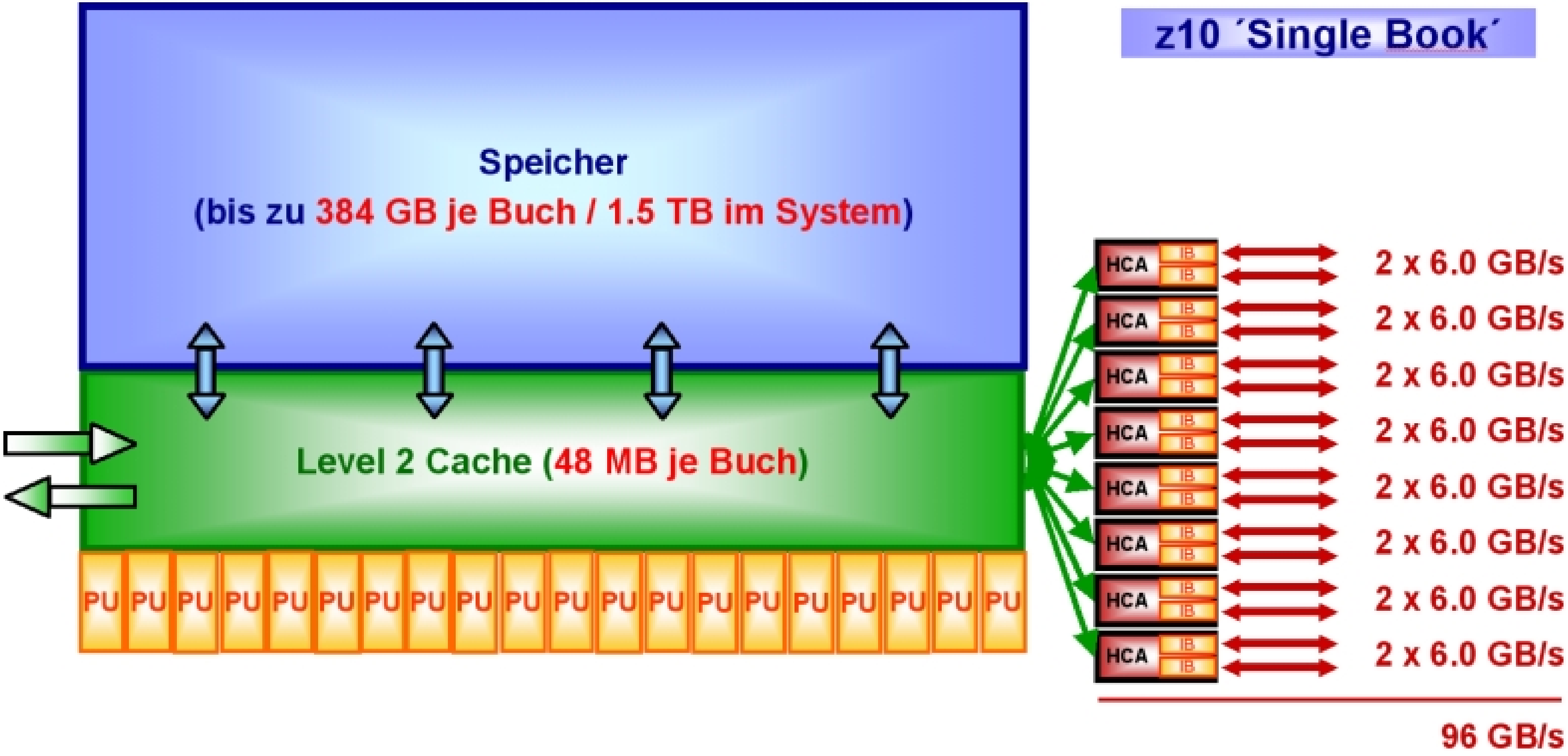


z10 configuration of the L2 cache

A maximum z10 system can have a total of 77 processing units (PU) in 4 books. Of these, 64 may be configured as CPUs. 3 of the 4 books may have 20 PUs each, the 4th book has 17 PUs.

The four L2 caches of the four books are interconnected by point-to-point links and form a single common and shared L2 cache that is used by the 77 PUs in all four books.

This is a unique System z feature. In other large systems, e.g from HP or Sun, the L2 cache is not shared, or is shared by a few CPUs at best.



This is another unique z10 (and z9) feature. In all non-IBM systems, I/O adapter cards attach to main memory. In a z10, the Host Channel Adapter (HCA) attaches to the L2 cache, supporting a much higher bandwidth.

Each book has a maximum of 8 HCA adapter cards, and each HCA has 2 ports, each attaching a 6 GByte/s Infiniband link, for a total of 96 GByte/s I/O data rate per book.