Enterprise Computing
Einführung in das Betriebssystem z/OS

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

WS 2009/2010

Teil 7

Parallel Rechner Implementierungen

Mehrfachrechner (Multiprocessor)
Parallelrecher (Parallel Processor)

Mehrfachrechner:

Auf mehreren CPUs laufen mehrere Prozesse (Threads) gleichzeitig.

Parallelrechner:

Auf mehreren CPUs läuft ein einziger Prozess.

Die Unterschiede zwischen Mehrfachrechnern und Parallelrechnern sind fließend. Meistens können Rechner mit mehreren CPUs sowohl als Mehrfachrechner als auch als Parallelrechner eingesetzt werden.

zSeries und S/390 Rechner sind für den Einsatz als Mehrfachrechner optimiert. IBM AIX, HP Superdome, SUN Sunfire25k bzw M9000, sowie Blade Server werden häufig auch als Parallelrechner eingesetzt.

Vorteile von Mehrfachrechnern und Parallelrechnern

Erhöhung der Verarbeitungskapazität

Gemeinsame Datennutzung

Bessere Nutzung von Betriebsmitteln

Verfügbarkeit

Gliederung von Mehrfachrechnern und Parallelrechnern

SISD

Single Instruction Stream - Single Data Stream (normaler von-Neumann Rechner)

SIMD

Single Instruction Stream - Multiple Data Stream (Data Level Parallelism)

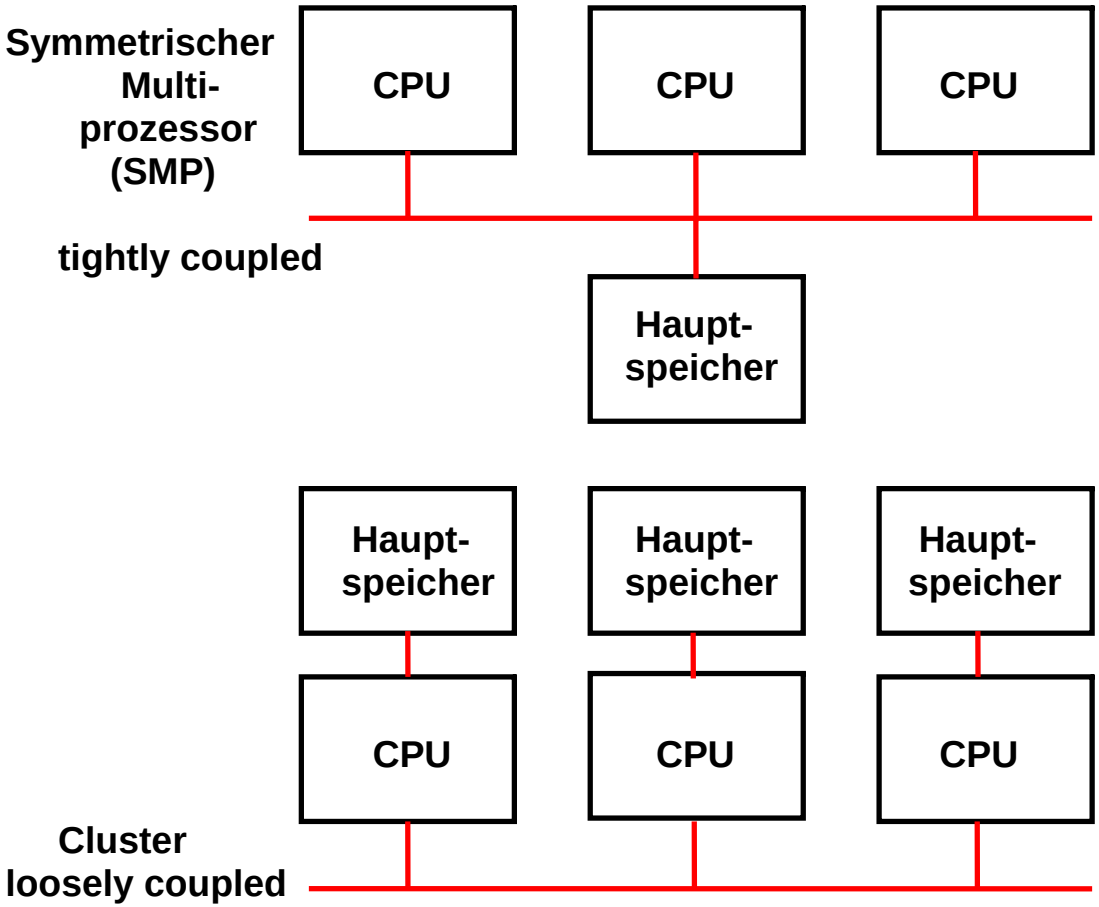
Alle Prozessoren arbeiten im Gleichschritt. Sie führen den gleichen Maschinenbefehl zur gleichen Zeit, aber mit unterschiedlichen Daten aus. Ideal für Situationen, in denen die Programmabläufe datenunabhängig sind.

MIMD

Multiple Instruction Stream - Multiple Data Stream (Programming Level Parallelism)

Vielfache unabhängige Sequenzen von Maschinenbefehlen verarbeiten vielfache, unabhängige Sequenzen von Daten

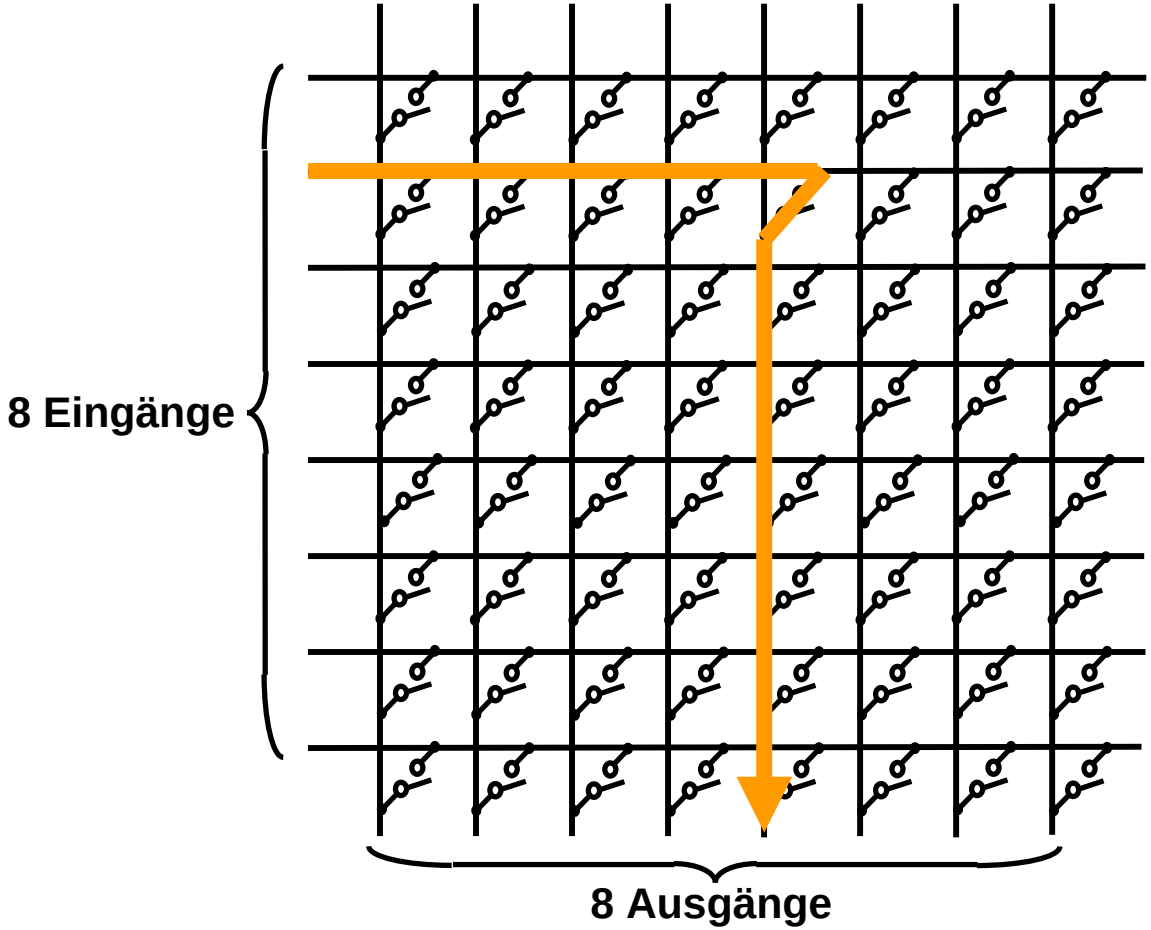
Exotische Architekturen
Systolic Arrays
Data Flow Maschinen
Neuronale Netze



Taxonomie von Mehrfach/Parallelrechnern

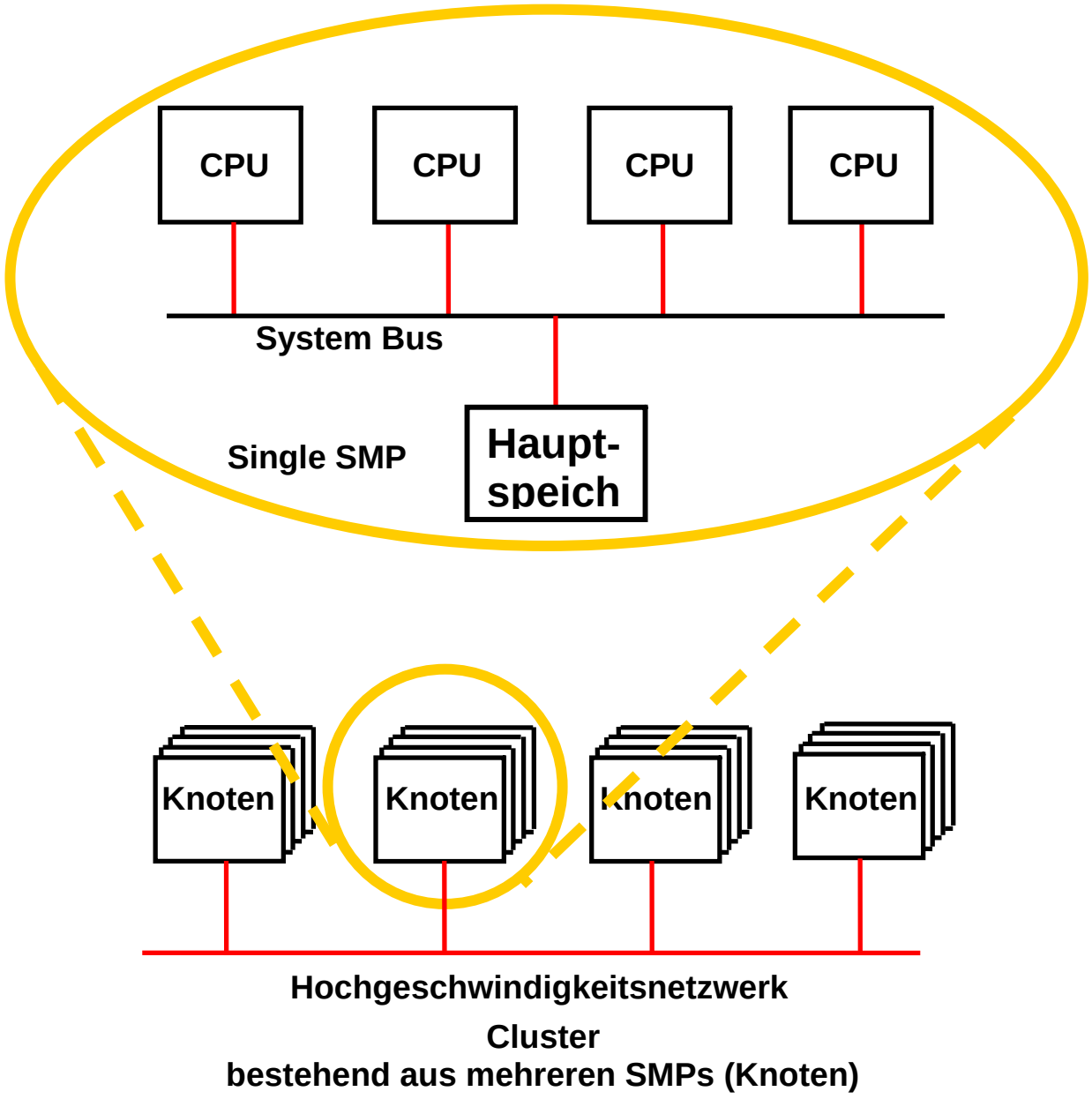
Mainframe Systeme erfordern meistens die Leistung von mehr als einer CPU. Bei Mehrfachrechnern unterscheiden wir 2 Grundtypen:

- **Ein Symmetrischer Multiprozessor (SMP) besteht aus mehreren CPUs (bis zu 64 bei einer z10 EC), die alle auf einen gemeinsamen Hauprspeicher zugreifen. In dem Hauptspeicher befindet sich eine einzige Instanz des Betriebssystems**
- **Ein Cluster besteht aus mehreren Rechnern, die jeder ihren eigenen Hauptspeicher und ihre eigene Instanz eines Betriebssystems haben.**



8 x 8 Crossbar Switch

Die CPUs eines Parallelrechners sind über ein Verbindungsnetzwerk miteinander verbunden. Dies kann ein leistungsfähiger Bus sein, z.B. der PCI Bus. Ein Bus hat aber nur eine begrenzte Datenrate. Deshalb setzen viele Implementierungen als Verbindungsnetzwerk statt dessen einen Kreuzschienenverteiler (Crossbar Switch, Crossbar Matrix Switch) ein, der die gleichzeitige Verbindung mehrerer Eingänge mit mehreren ausgängen ermöglicht. Mit derartigen Switches können fast beliebige Datenraten erreicht werden. Gezeigt ist als Beispiel ein Switch mit 8 Eingängen und 8 Ausgängen, der gleichzeitig 8 parallele Verbindungen ermöglicht.



Cluster
bestehend aus mehreren SMPs (Knoten)

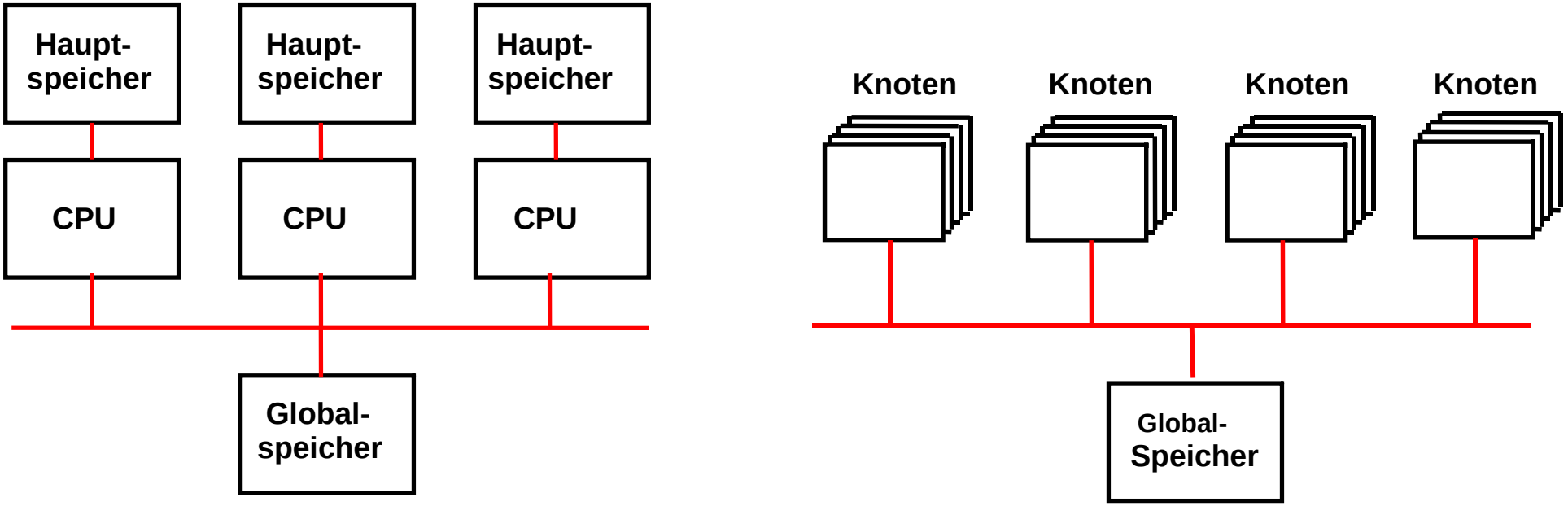
Häufig sind die Elemente eines Clusters nicht einzelne CPUs, sondern SMPs, die als Knoten (und in der IBM Terminologie als „Systeme“) bezeichnet werden.

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

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

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

Ein SMP wird als eng gekoppelter, ein Cluster als loose gekoppelter Multi-prozessor bezeichnet.



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

Frage: Warum kann mann nicht einen SMP mit 64 oder mehr Prozessoren bauen ?

Antwort: Ein z10 EC Rechner kann tatsächlich einen SMP mit bis zu 64 CPUs betreiben. Aber ...

Skalierung eines Symmetrischen Multiprozessors

Ein Multiprocessor mit zwei statt einer CPU sollte idealerweise die zweifache Leistung haben. Dies ist aus mehreren Gründen nicht der Fall.

Das größte Problem besteht darin, dass die CPUs eines SMP gleichzeitig einen Dienst des Betriebssystem-Kernels in Anspruch nehmen wollen, der nur seriell durchgeführt werden kann. Ein typisches Beispiel ist der Scheduler. Zu diesem Zweck werden Komponenten des Kernels mit Locks (Sperren) versehen. Gelegentlich muss eine CPU darauf warten. dass eine andere CPU eine Resource des Kenrnels freigibt.

Mit wachsender Anzahl der CPUs verstärkt sich das Problem.

Wie groß die Beeinträchtigung ist, hängt sehr stark von der Art der laufenden Anwendungen ab. Anwendungen, die Dienste des Kernels nur selten in Anspruch nehmen werden nur wenig beeinträchtigt.

Unglücklicherweise sind Transaktions- und Datenbankanwendungen besonders kritisch. Gleichzeitig sind dies die wichtigsten Aufgaben in einer betrieblichen IT-Infrastruktur.

Gründe für den Leistungsabfall beim Symmetric Multiprocessors

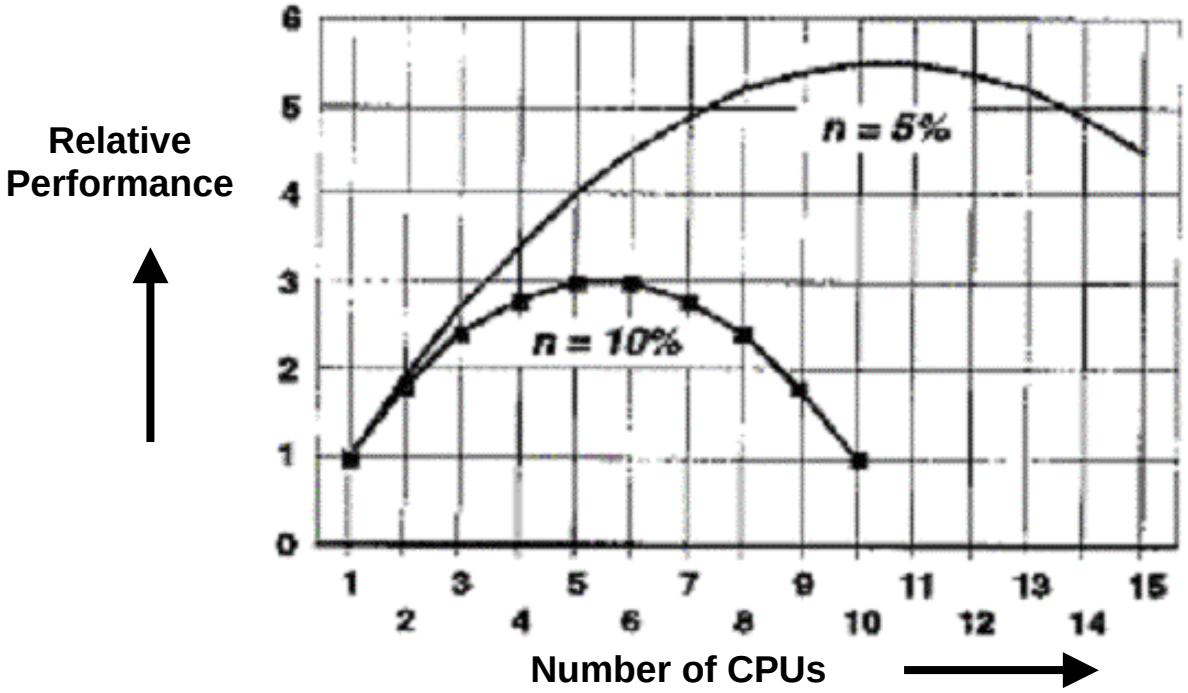
1. Hardware Konflikte. Wenn zwei CPUs gleichzeitig auf den Hauptspeicher zugreifen, muss eine CPU warten.

Andere Beispiele: gleichzeitige Nutzung von Bussen, Kanälen, Steuereinheiten, Plattenspeichern (z. B. SYSRES, Paging Disk).

Mit entsprechendem Aufwand lassen sich Hardware Konfkikte reduzieren.

2. Software Konflikte (meistens Betriebssystem/Kernel). Wenn zwei CPUs gleichzeitig den Scheduler/Dispatcher aufrufen, muss eine CPU warten.

Kernel Konflikte lassen sich nur begrenzt vermeiden.



Die beiden dargestellten Kurven repräsentieren das typische Leistungsverhalten eines SMP als Funktion der anzahl der eingesetztwen CPUs. Die untere Kurve nimmt an, dass bei 2 CPUd jede CPU 10 % ihrer Leistung verliert; bei der oberen Kurve sind es 5 %. Mit wachender Anzahl von CPUs wird der Leistungsgewinn immer kleiner; ab einer Grenze wird der Leistungsgewinn negativ.

Dis bedeutet, dass SMPs nur mit einer begrenzen maximalen CPUs sinnvoll betrieben werden können. Diese Grenze ist sehr stark von der Art der Anwendungen abhängig.

Angenommen, ein Zweifach Prozessor leistet das Zweifache minus n % eines Einfach Prozessors. Für n = 10 % ist es kaum sinnvoll, mehr als 4 Prozessoren einzusetzen. Für n = 5 % sind es 8 Prozessoren.

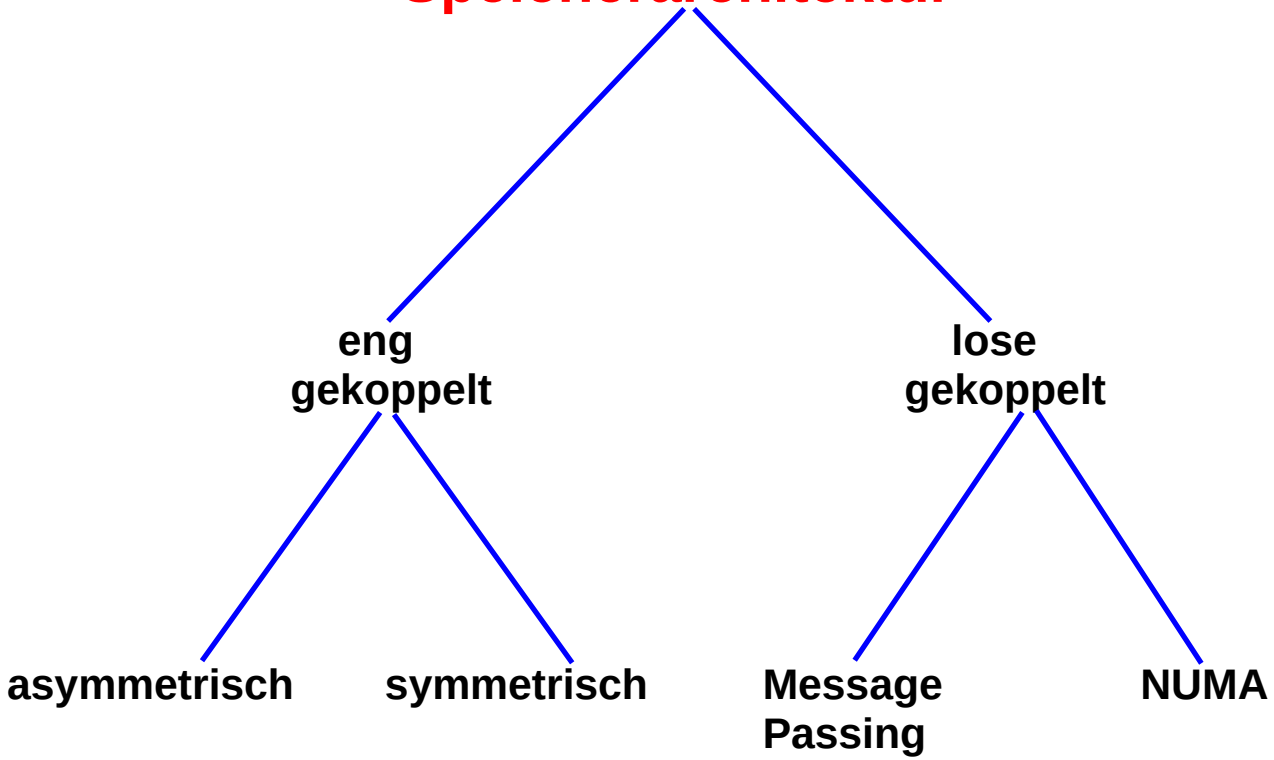
Bei einem z10 Rechner mit z/OS ist n << 2 %. Es kann sinnvoll sein, einen SMP mit bis zu 32 Prozessoren einzusetzen.

Begrenzte Anzahl CPUs in einem SMP

Die Gründe für den Leistungsabfall sind Zugriffskonflikte bei der Hardware und Zugriffskonflikte auf Komponenten des Überwachers. Die Überwacherkonflikte überwiegen.

Durch Feintuning und Abstimmung der einzelnen Kernel-Komponenten kann die Skalierung von SMP Rechnern verbessert werden. z/OS hebt sich an dieser Stelle deutlich von allen anderen Betriebssystemen ab. Mann nimmt an, dass unter z/OS in einem SMP maximal etwa 2 x soviele CPUs betrieben werden können als wie bei jedem anderen Betriebssystem.

Mehrfachrechner und Parallelrechner
Speicherarchitektur



Message Passing

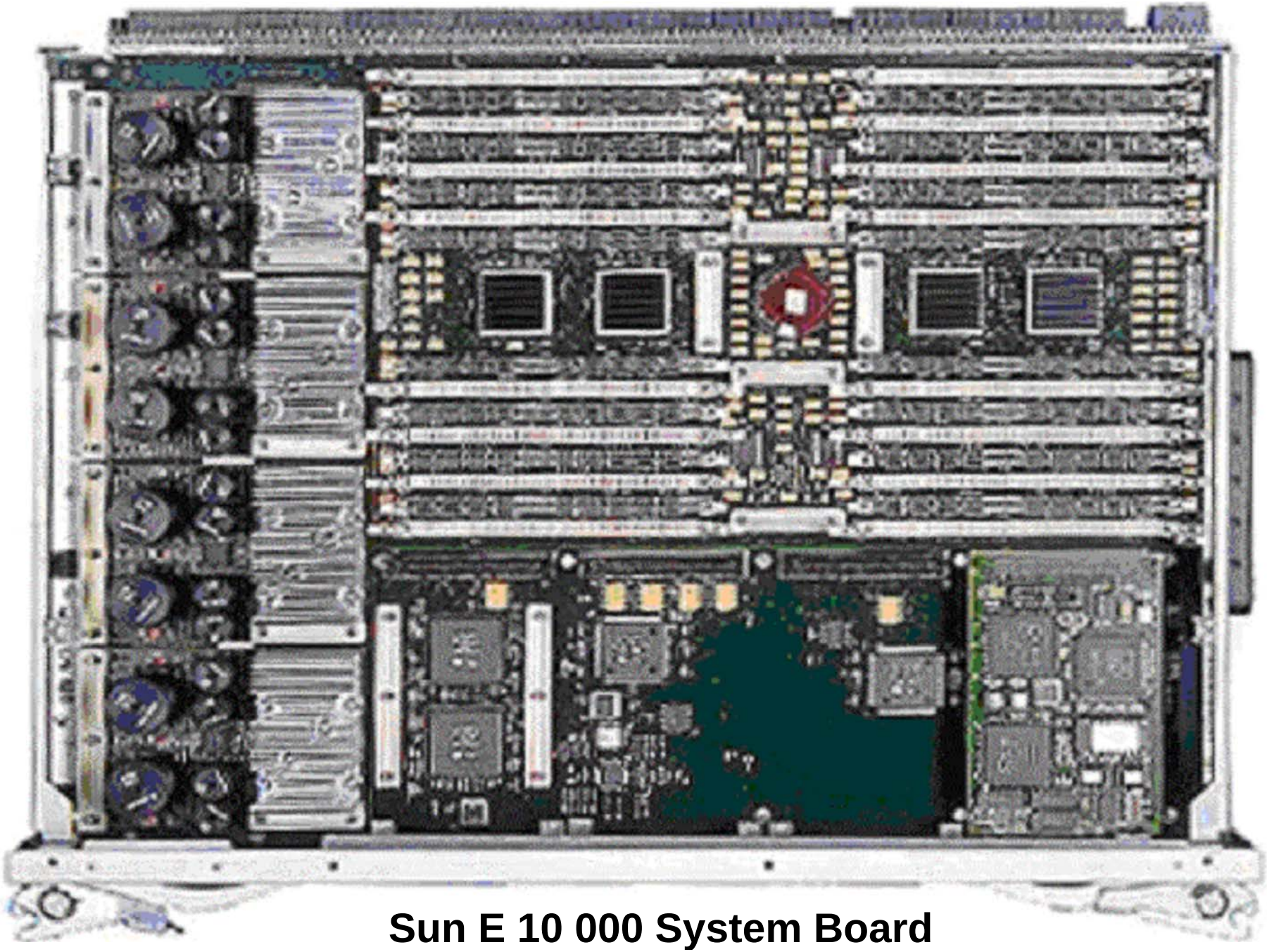
Zugriff auf entfernte Objekte wird durch Nachrichtentransport simuliert

Non Uniform Memory Access

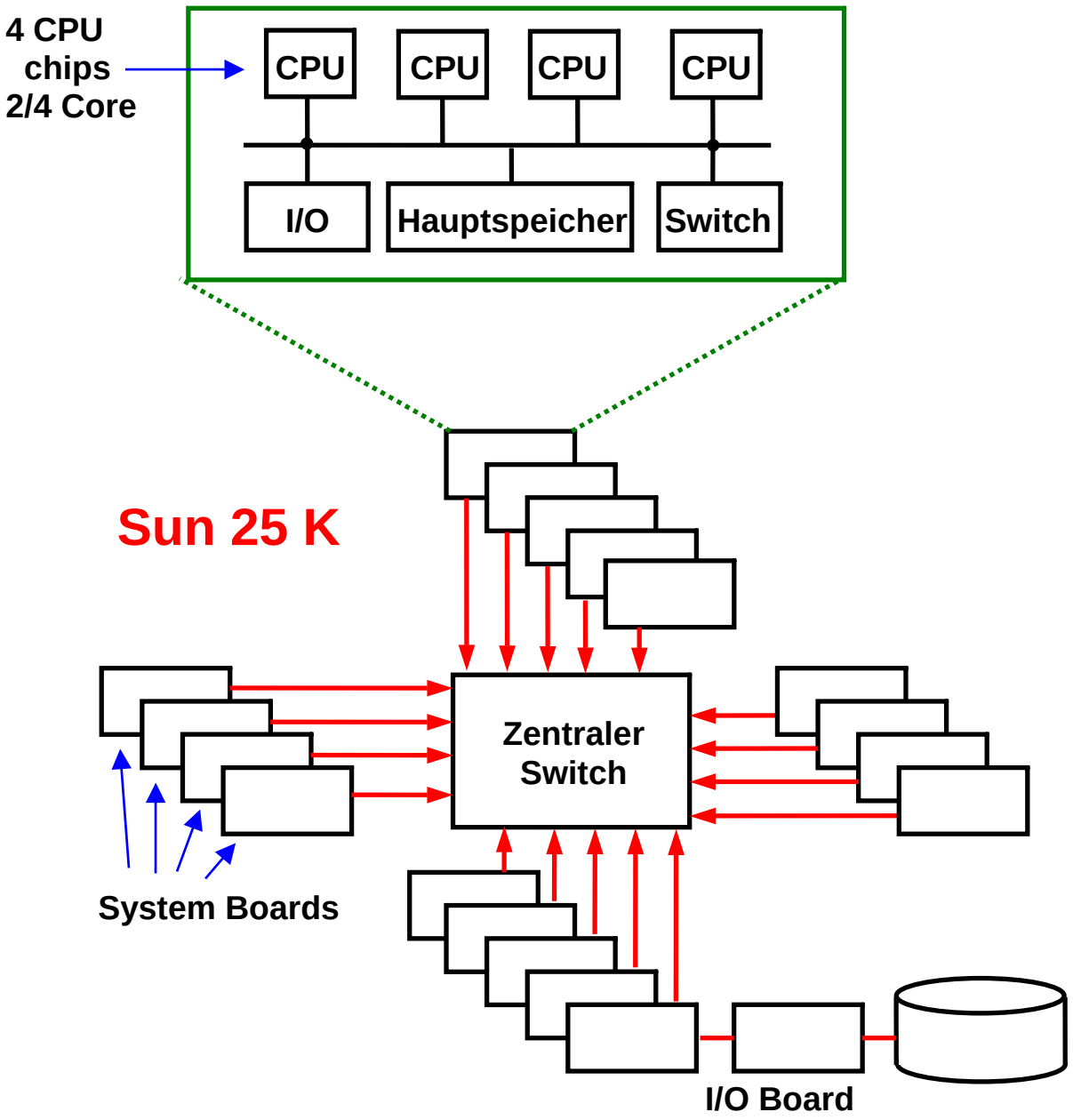
Hardwareunterstützung für den Zugriff auf den Inhalt entfernter Speicher

Kommerzielle Großrechner

Message Passing Parallel Rechner



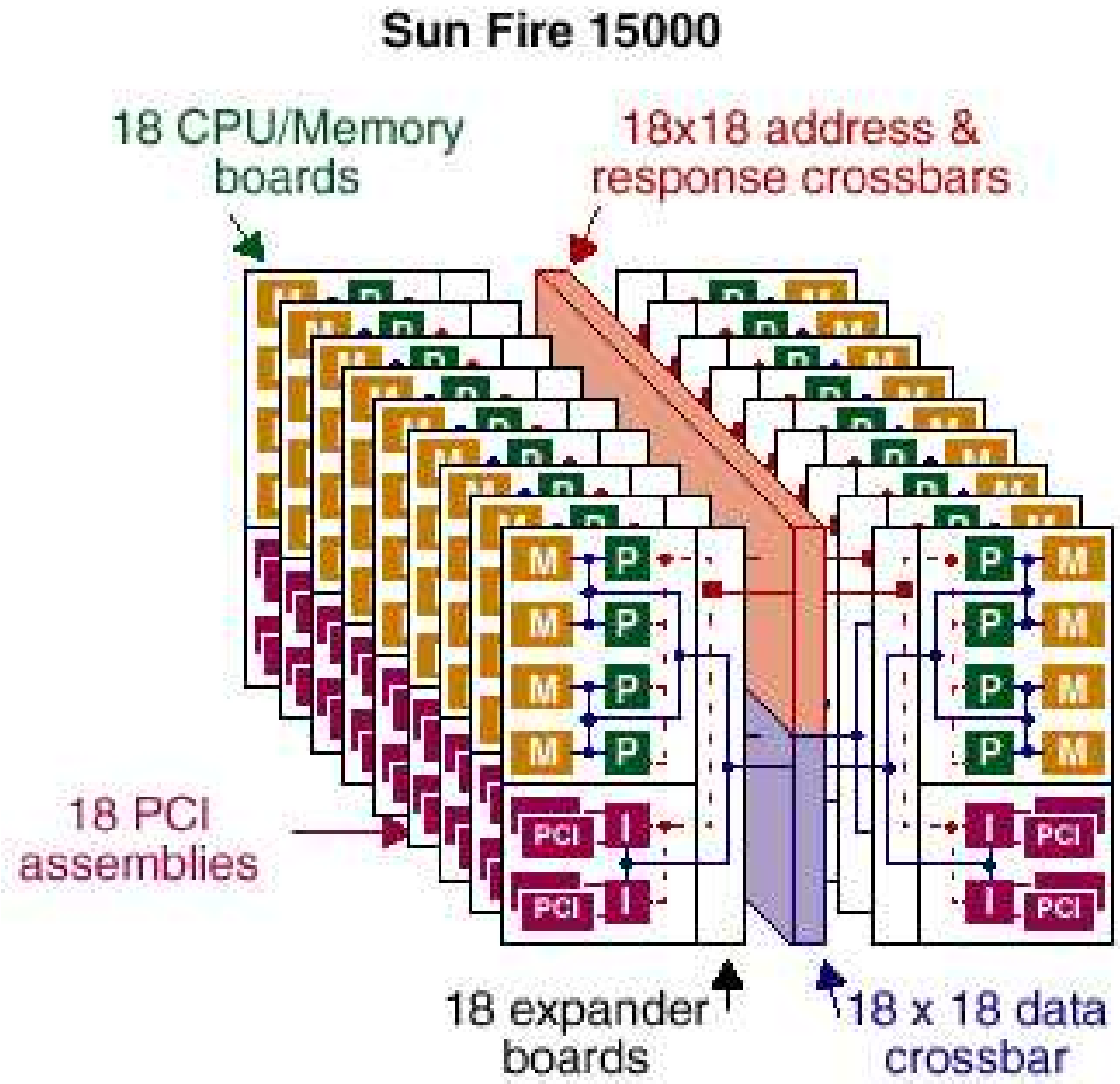
Sun E 10 000 System Board



Ein Sun 25K oder HP Superdome Großrechner enthält 16 oder 18 System Boards, jedes mit 4 CPU Chips und bis zu 128 GByte Hauptspeicher, einem Anschluss an einen zentralen Switch sowie I/O Controller auf jedem System Board.

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

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



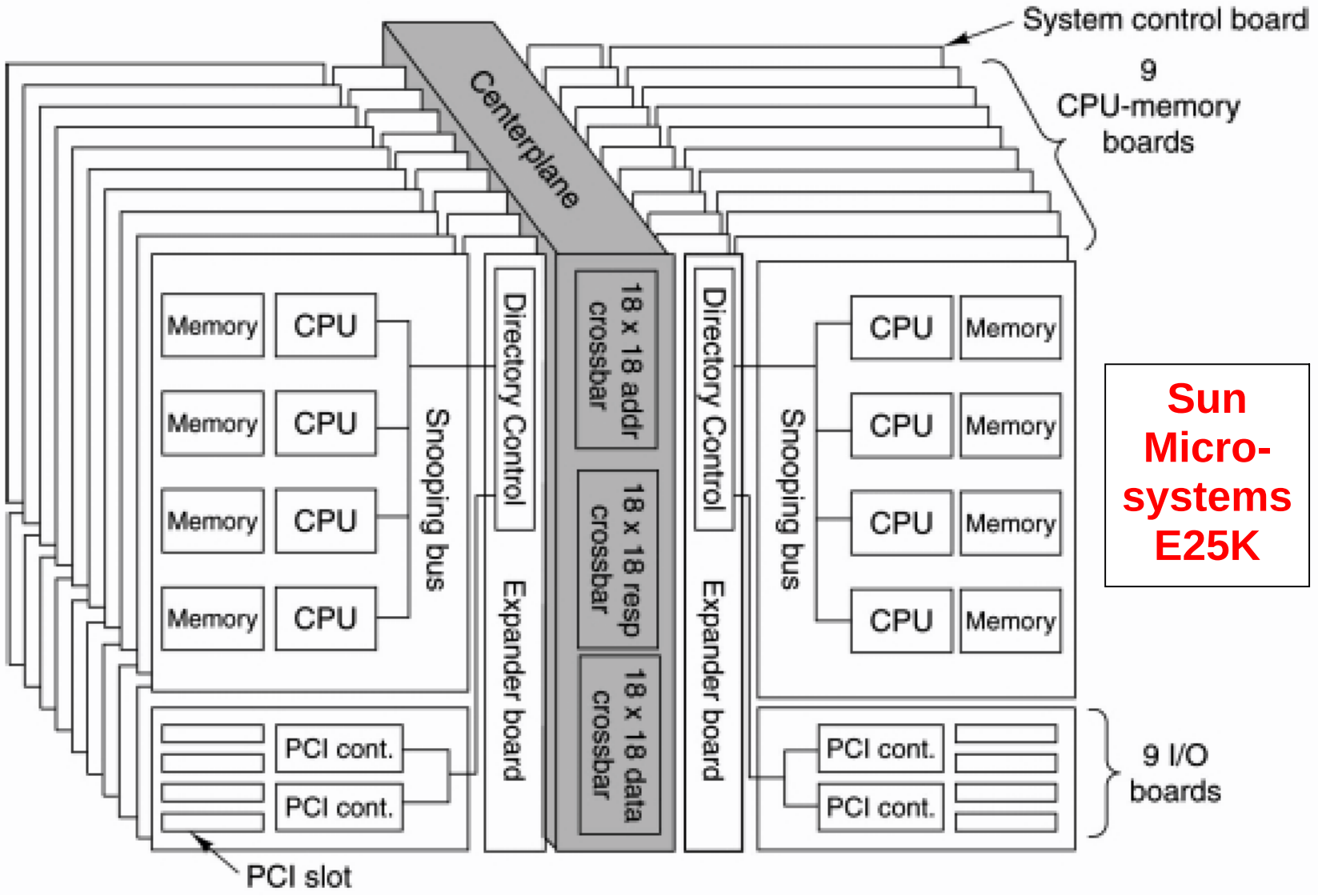
Die 16 oder 18 System Boards befinden sich in Steckplätzen auf einem Crossbar Board. Dieses enthält eine Reihe von Crossbar Chips, welche alle System Boards miteinander verbindet. Dies sind spezifisch ein 16 x 16 Daten Crossbar Chip, ein 16 x 16 Response Crossbar Chip und ein 16 x 16 Adressen Crossbar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

Auch die I/O Karten in den I/O Cages sind über diese Switche mit den System boards verbunden.

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

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

Eine solche Aufteilung wird als „Harte Partitionnierung“ bezeichnet.



Sun 25 K
Central Electronic Complex

Keine Konkurrenz für Aldi-PCs . Ab 1 Mill. $ ist man dabei. Die Konkurrenzprodukte von HP und IBM sind eher noch teurer.

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.

Sun SPARC Enterprise M9000 server

- **Backplane Interconnect mit 368 GByte/second,**
- **Verbindet alle chips miteinander**
- **Unterstützt bis zu 24 “Dynamic Domains”.**
- **Granularität der Domain Partitionierung besteht aus einem Processor Chip, 4 DIMMs und zwei I/O Slots.**

Sun SPARC Enterprise M9000 Server

64 dual core SPARC64 VI oder 64 quad core SPARC64 VII Prozessoren
Bis zu 2 TByte Hauptspeicher
Bis zu 24 Dynamic Domains
100% binäre Kompatibilität mit UltraSPARC Prozessoren

Sun –Fujitsu/Siemens

Die drei Highend-Systeme "M8000", "M9000-32" und "M9000-64" können mit maximal 16, 32 und 64 Sparc64-VI-CPUs ausgestattet werden. Die größtmögliche Hauptspeicherkapazität beträgt je nach Modell 512 GB oder 1 beziehungsweise 2 TB. Das Modell M9000-64 kann 128 Prozessorkerne am Laufen halten

Ebenfalls je nach Rechnertyp lassen sich bis 16, 32 oder 64 2,5-Zoll-SAS-Festplatten einbauen. Maximal können bei diesen Hochleistungsservern bis zu 24 Hardware-Partitionen eingerichtet werden.

Die Rechner laufen unter dem Betriebssystem Solaris 10. Die Unternehmen garantieren eine binäre Rückwärtskompatibilität bis einschließlich Solaris 2.6.

Solaris 10 unterstützt Haus-eigene Virtual-Machines- und Betriebssystem-Virtualisierungskonzepte.

“The I/O system interface features hot-swappable I/O Units (IOUs), each of which provides eight PCIe slots and four 2.5-inch SAS disk drive bays. Fully configured, these servers support up to 32 or 64 internal SAS boot disks, 64 or 128 internal PCIe slots, and up to 1024 GB or 2048 GB of memory. With the optional external I/O expansion unit, you can increase the number of supported I/O slots to a maximum of 288 slots”.

April 2007. Seit 2008 sind auch Sparc64-VII-CPUs möglich.

Sun SPARC Enterprise M9000 Server

SPARC V9 Architektur, ECC protected

SPARC64VII Cache per processor (Level 1): 64-KB D-cache and 64-KB I-Cache

Cache per processor (Level 2): 6-MB on-chip

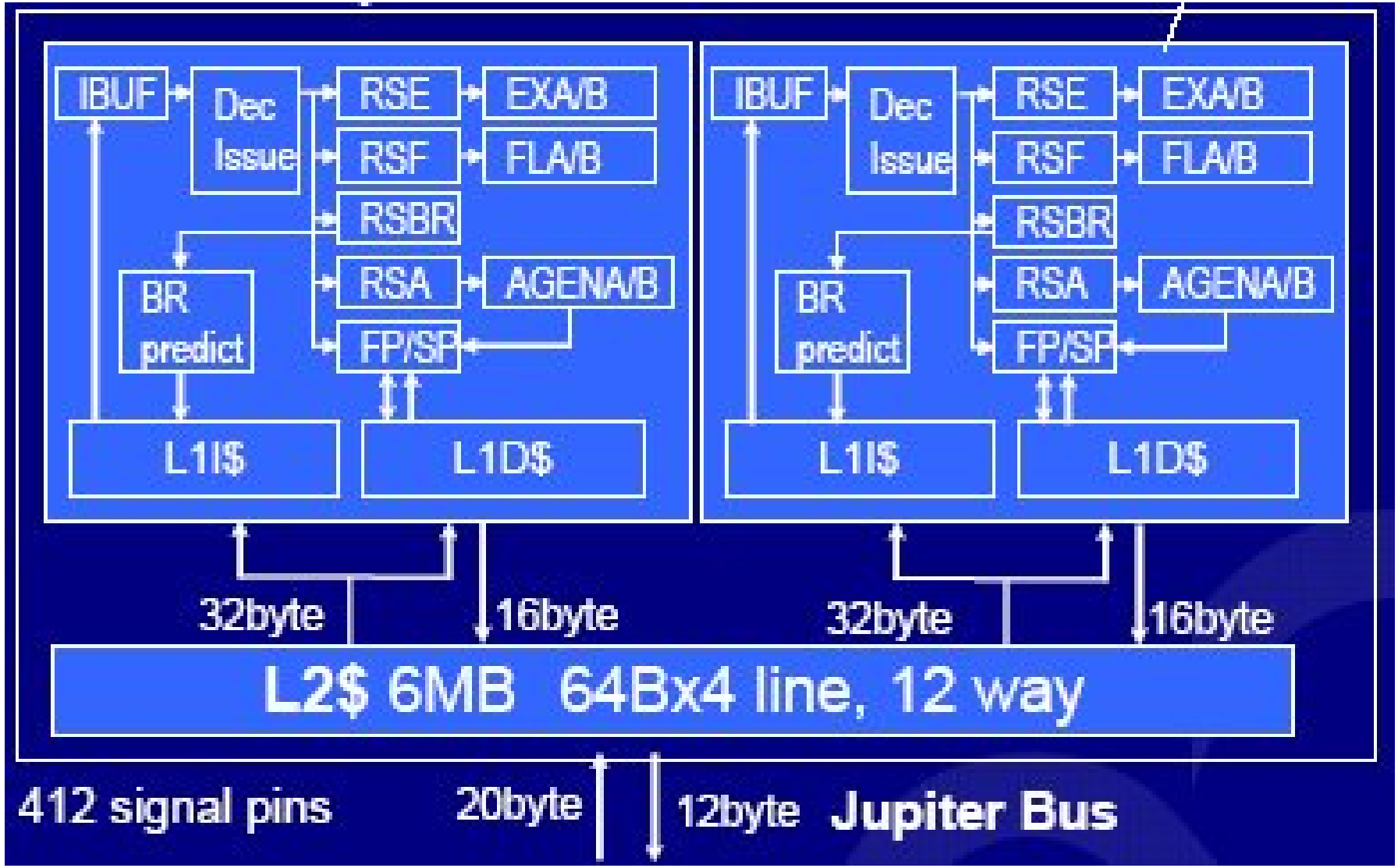
Chip Multithreading (CMT) Technologie

Bis zu 4TByte Hauptspeicher

Bis zu 288 PCIe or PCI-X slots with the optional External I/O Expansion Unit

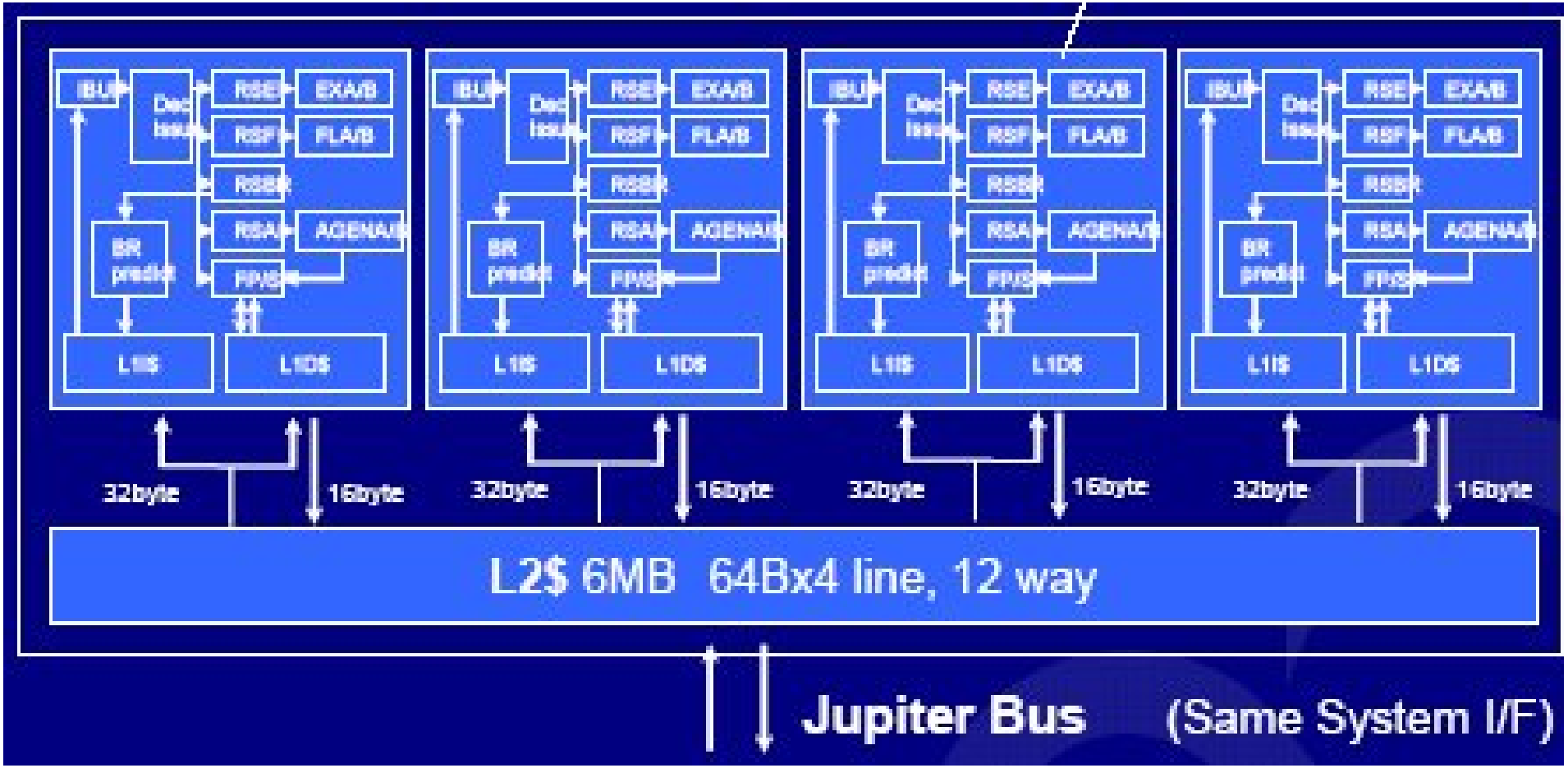
Bis zu 24 dynamic domains

2009



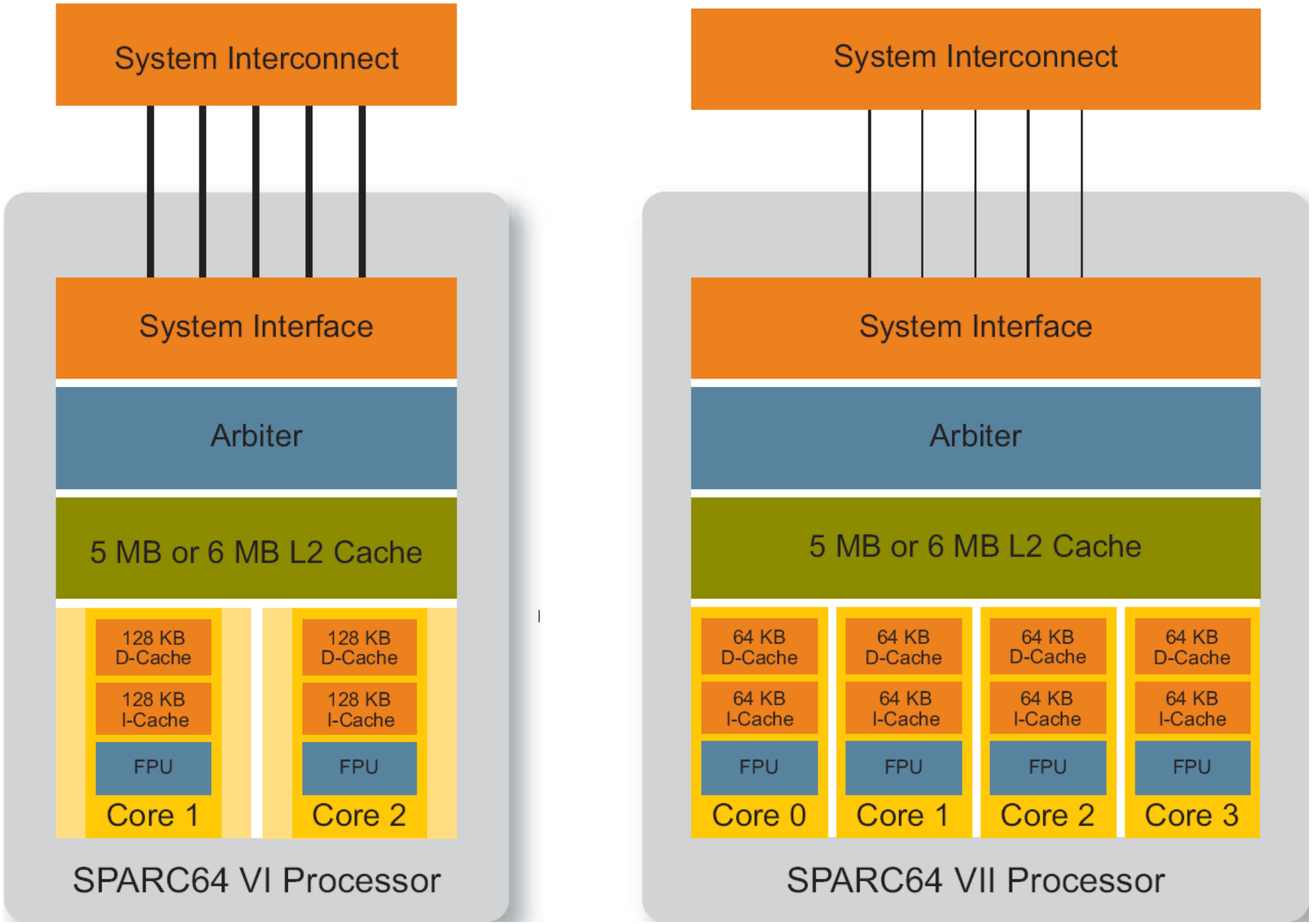
Sun-Fujitsu Kooperation SPARC64 VI, 3Q 2007

Jeder Kern kann 2 Threads parallel ausführen und hat einen eigenen L1-Cache. Den 6 MByte großen L2-Cache teilen sich die beiden Kerne.



Seit April 2008 sind SPARC64-VII-Chips mit vier Kernen und 2,7 GHz (65 nm) verfügbar. Diese nutzen die selben Kerne wie das VI-Modell. Allerdings teilen sich dann vier Kerne die 6 MByte L2-Cache. Die VII-Modelle werden Bus-kompatibel zu den VI-Modellen sein.

A fully-loaded, SPARC-Enterprise M9000-64 with 64 CPU chips would have 256 cores, capable of running 512 concurrent threads in a single Solaris image. With 512 DIMMs, and assuming 4GB DIMMs are available, the system would max-out at 2TB of RAM.

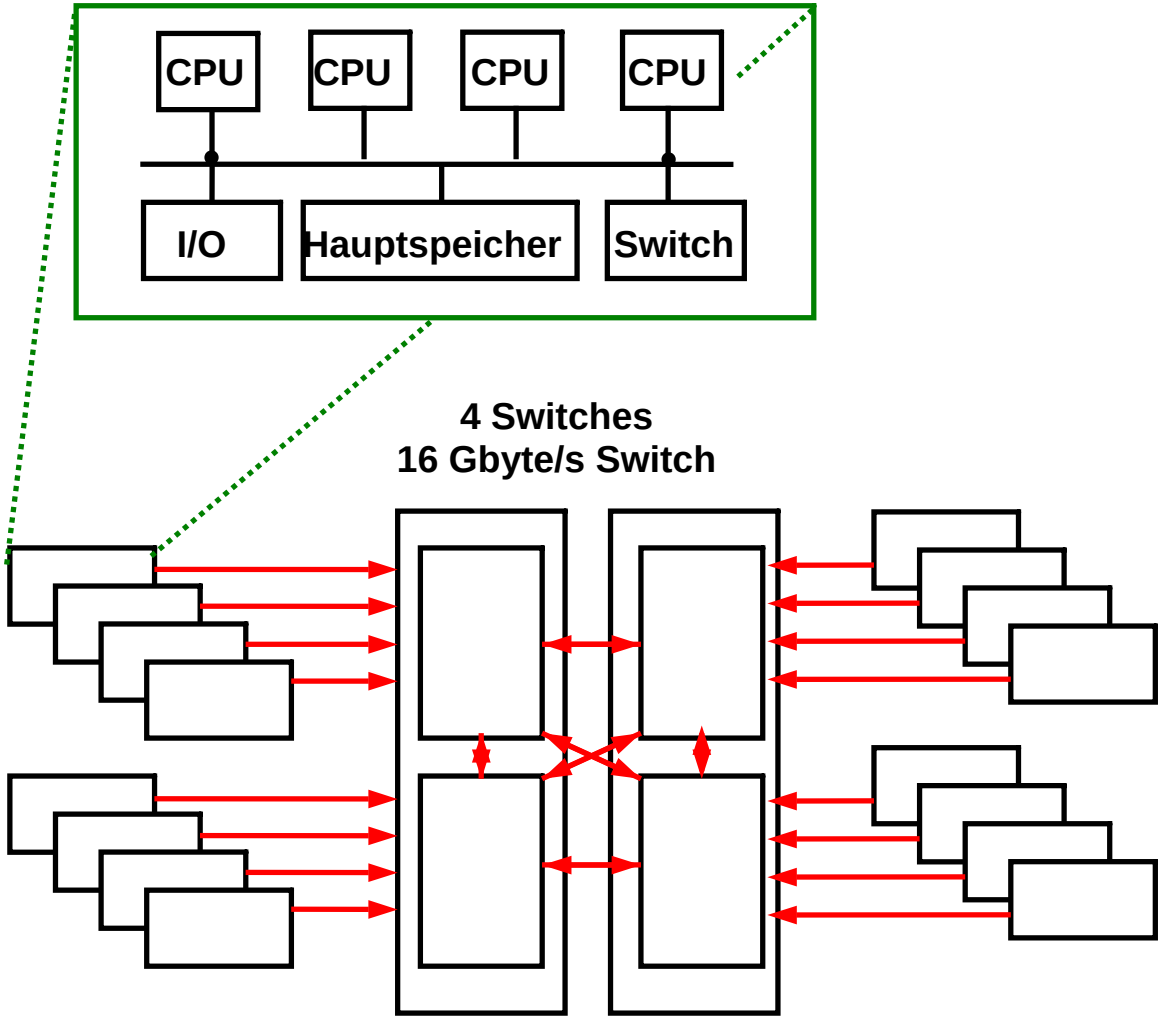

System Interconnect
System Interface
Arbiter
5 MB or 6 MB L2 Cache
128 KB D-Cache
128 KB I-Cache
FPU
Core 1
128 KB D-Cache
128 KB I-Cache
FPU
Core 2
SPARC64 VI Processor
System Interconnect
System Interface
Arbiter
5 MB or 6 MB L2 Cache
64 KB D-Cache
64 KB I-Cache
FPU
Core 0
64 KB D-Cache
64 KB I-Cache
FPU
Core 1
64 KB D-Cache
64 KB I-Cache
FPU
Core 2
64 KB D-Cache
64 KB I-Cache
FPU
Core 3
SPARC64 VII Processor

Hewlett-Packard Superdome Cell Board

4 Itanium 2 CPUs 1,5 GHz

64 Gbyte Hauptspeicher

E/A Bus Anschlüsse

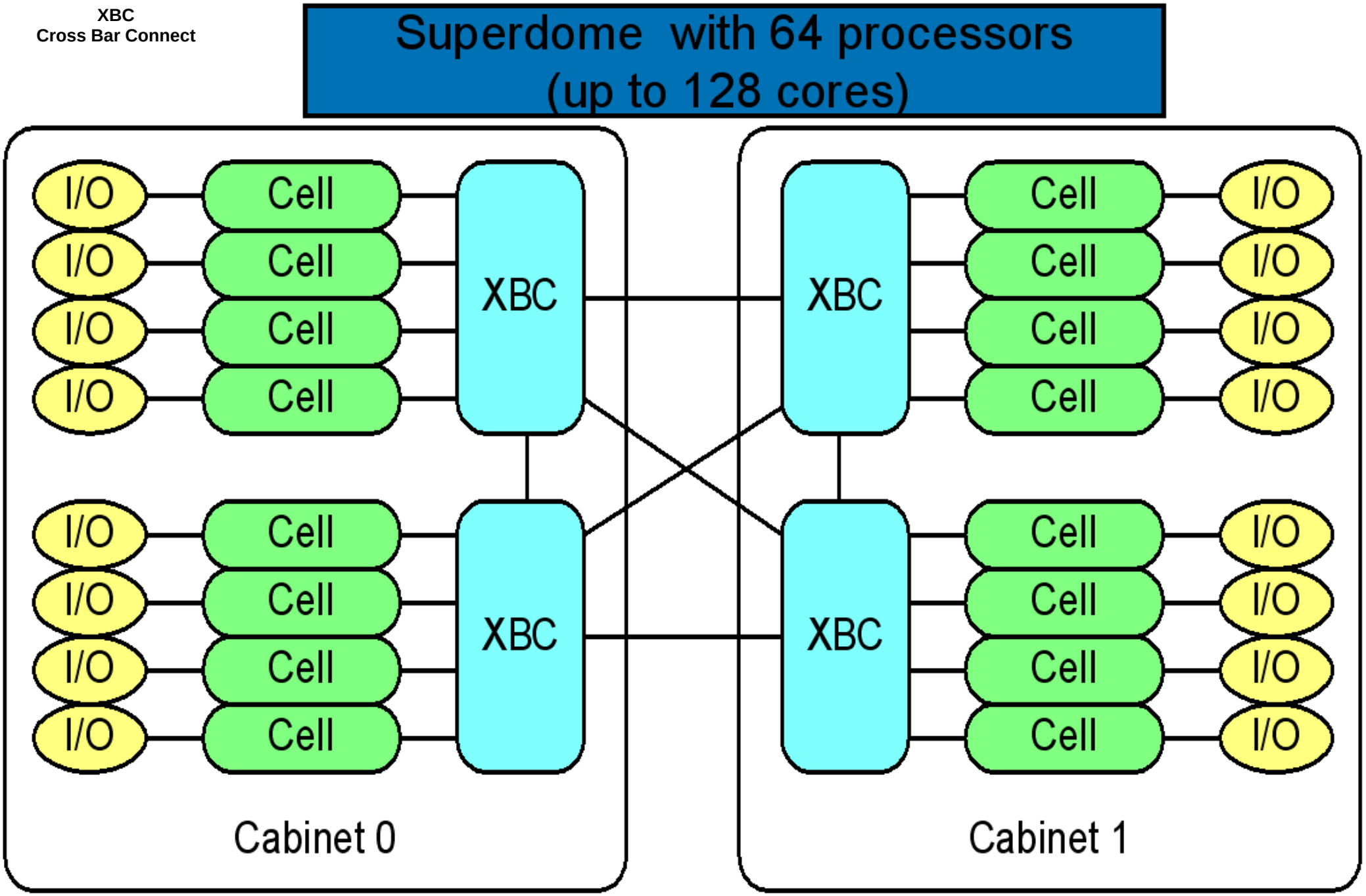


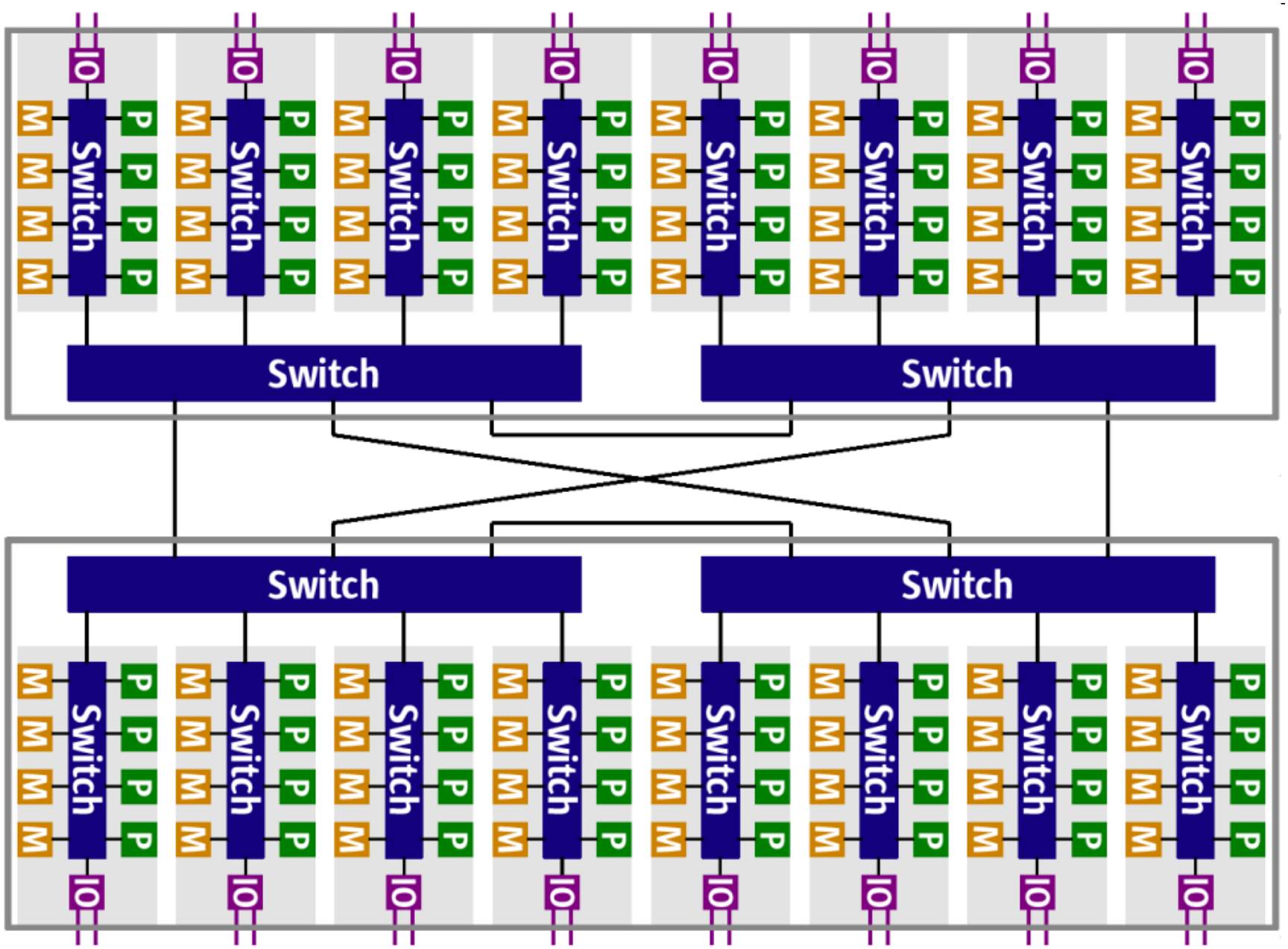
128 (64) CPU´s, 16 Knoten (Cell Boards), je 8 CPU/Knoten
I/O Anschluß auf jedem Cell Board

HP Superdome Cluster

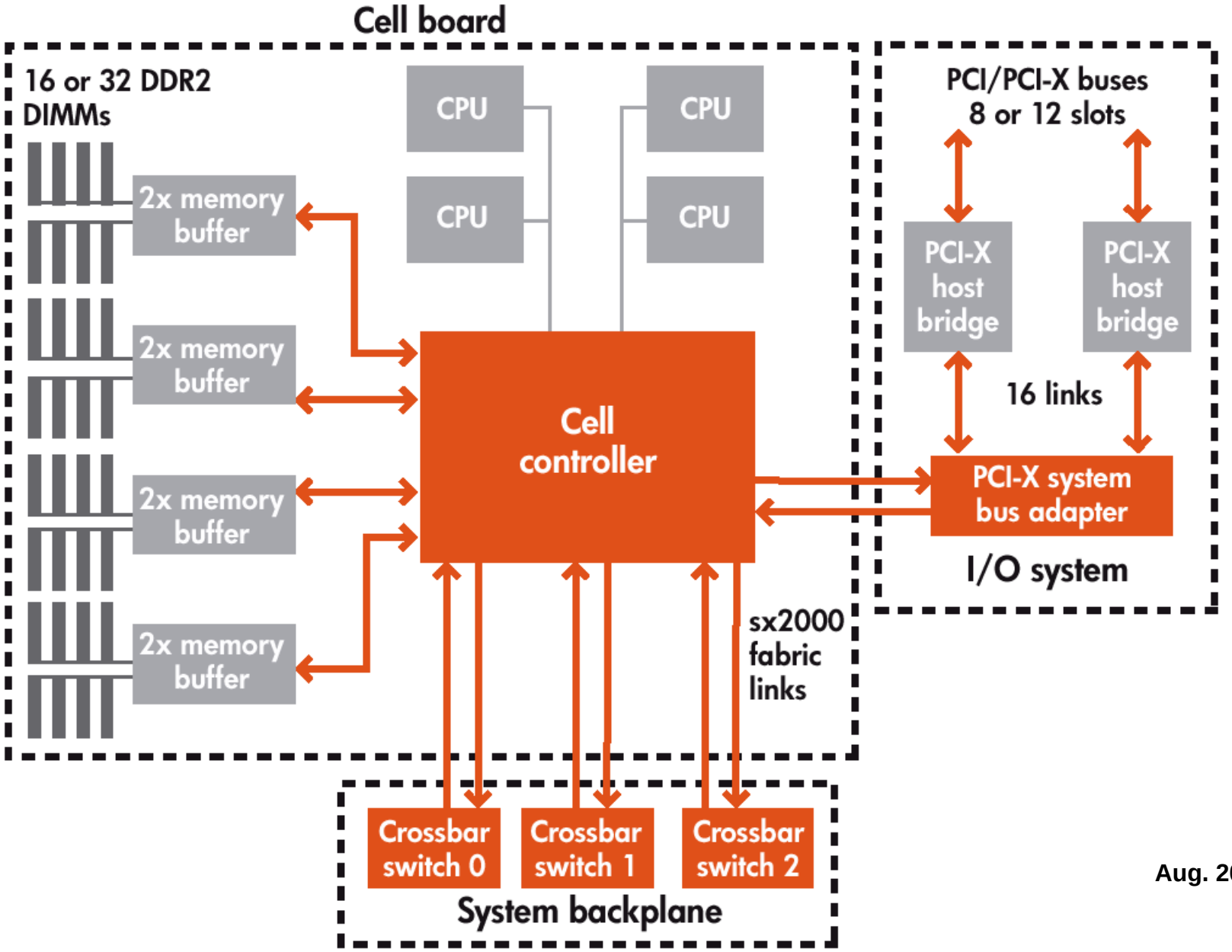
http://www.serverworldmagazine.com/webpapers/2001/05_hpsuperdome.shtml

XBC
Cross Bar Connect





HP Superdome Cluster

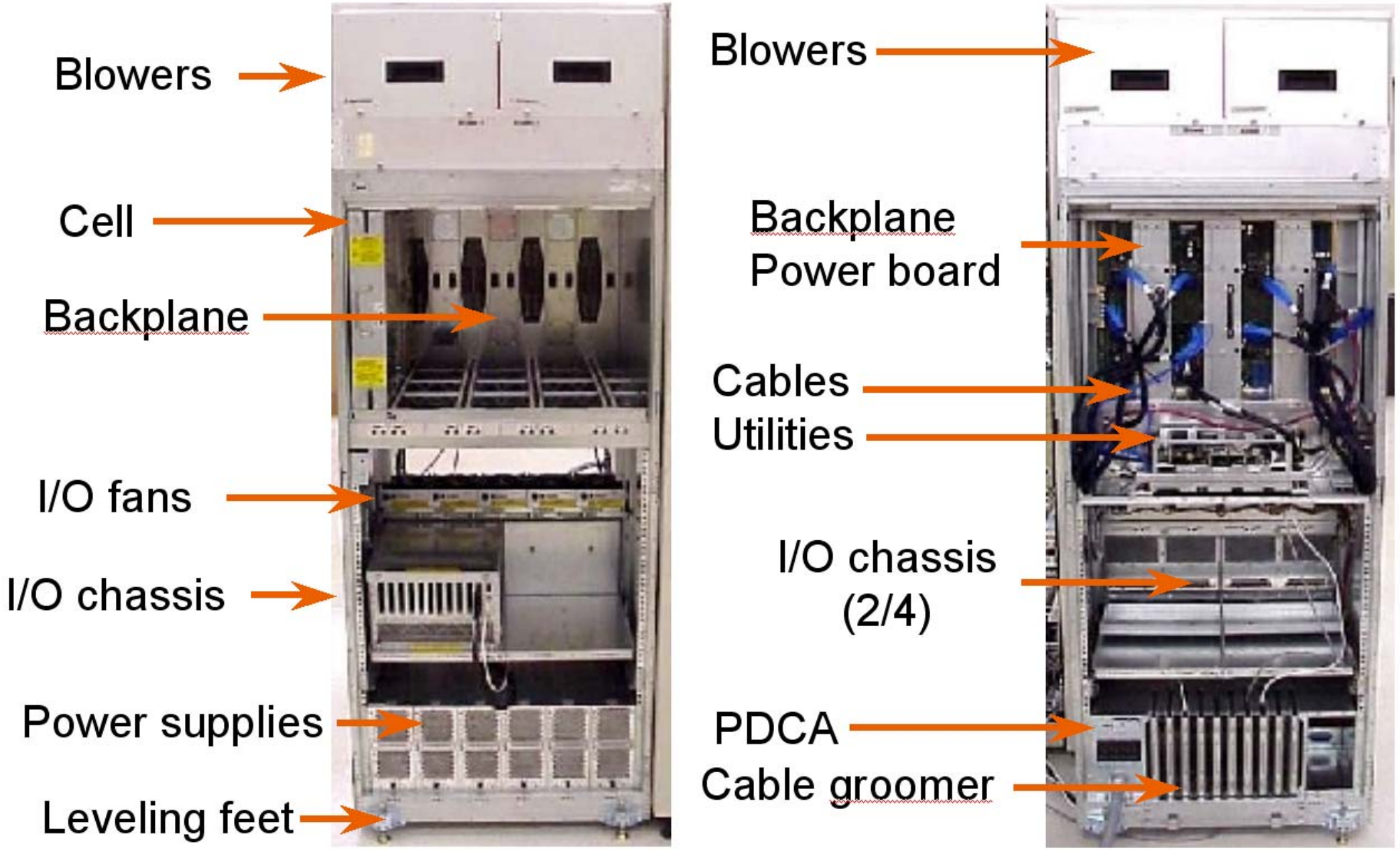

Cell board
16 or 32 DDR2 DIMMs
CPU
CPU
CPU
CPU
2x memory buffer
2x memory buffer
2x memory buffer
2x memory buffer
Cell controller
PCI/PCI-X buses 8 or 12 slots
PCI-X host bridge
PCI-X host bridge
16 links
PCI-X system bus adapter
I/O system
sx2000 fabric links
Crossbar switch 0
Crossbar switch 1
Crossbar switch 2
System backplane

Aug. 2009

HP Superdome

64 dual core Itanium 9M , 128 cores **Aug. 2009**

16 Cell Boards, 512 DIMM slots, 1 TByte (2TByte with 4 GB DIMMS, 96 (192) PCI-X slots



Layout of an HP 9000 Superdome 32-processor server cabinet.

Aug. 2009, PDCA - Power distribution control assembly

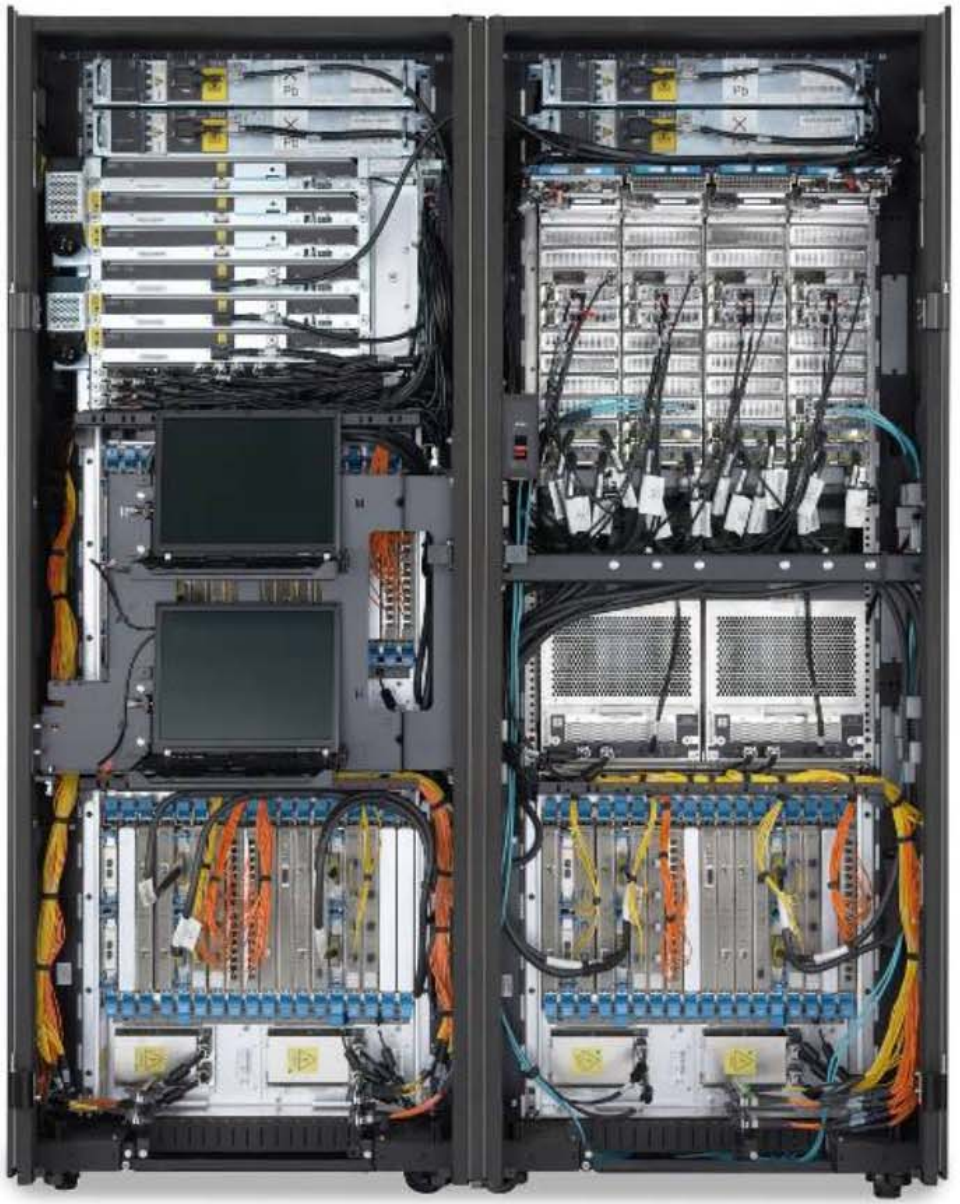
Layout of an HP 9000 Superdome 32-processor server cabinet.

At the top of the cabinet, hot-swappable main fans are installed after the cabinet arrives at your site. Below these fans is a cage for the eight cell boards on which processors and memory DIMMs reside. In a future release of HP-UX 11i, these cell boards will support hot-swap capability, so they can be replaced without bringing down the system.

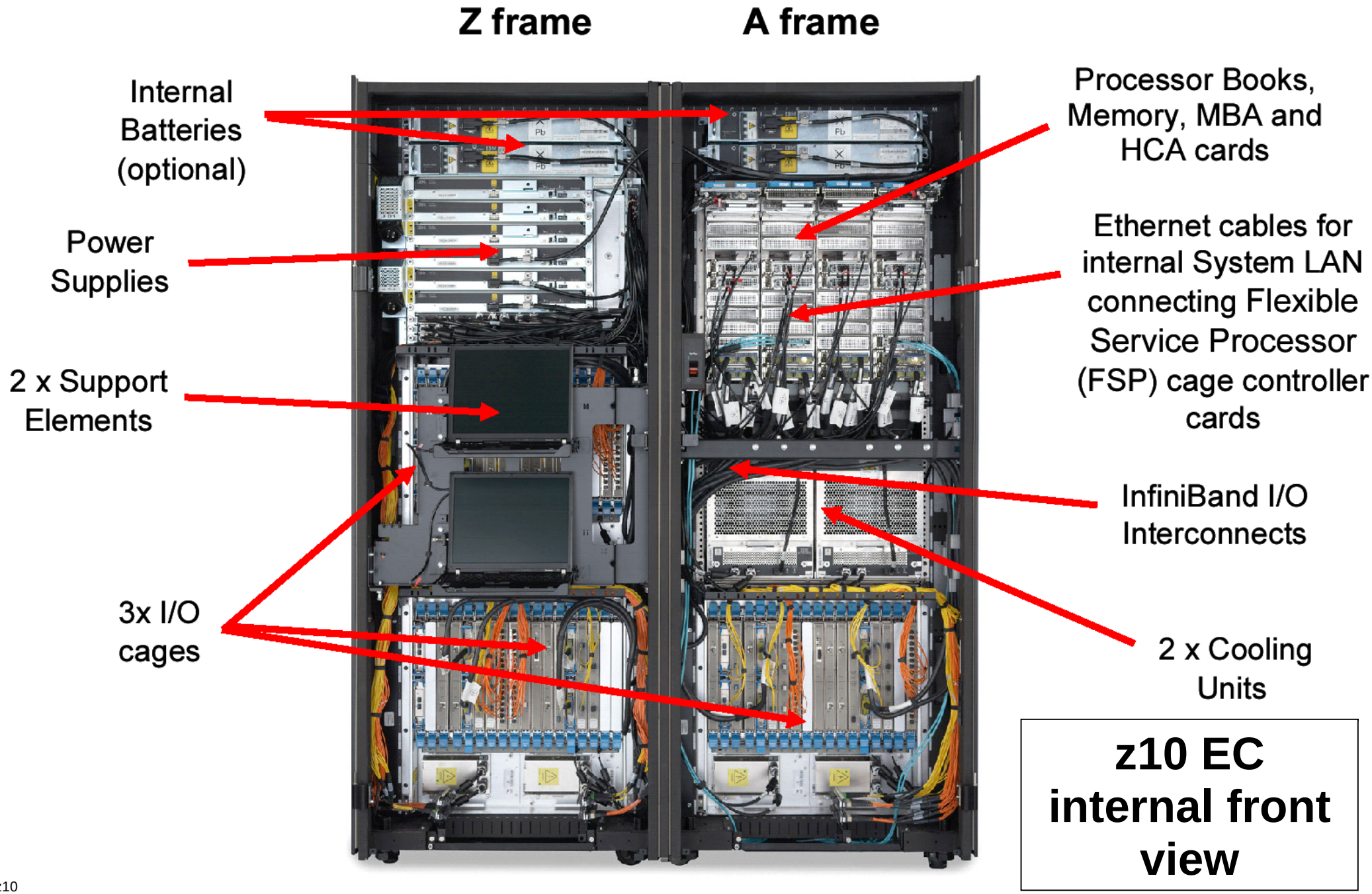
Directly below the cell boards is the main air intake, and below that are two I/O chassis. Each I/O chassis holds 12 PCI-X I/O cards.

Redundant power supplies are at the bottom of the cabinet. The HP 9000 Superdome Server does not use electrical plugs; instead, 48 V dc power is hard-wired to each cabinet. There are two redundant power inputs, so the system can be powered up by two different power grids.

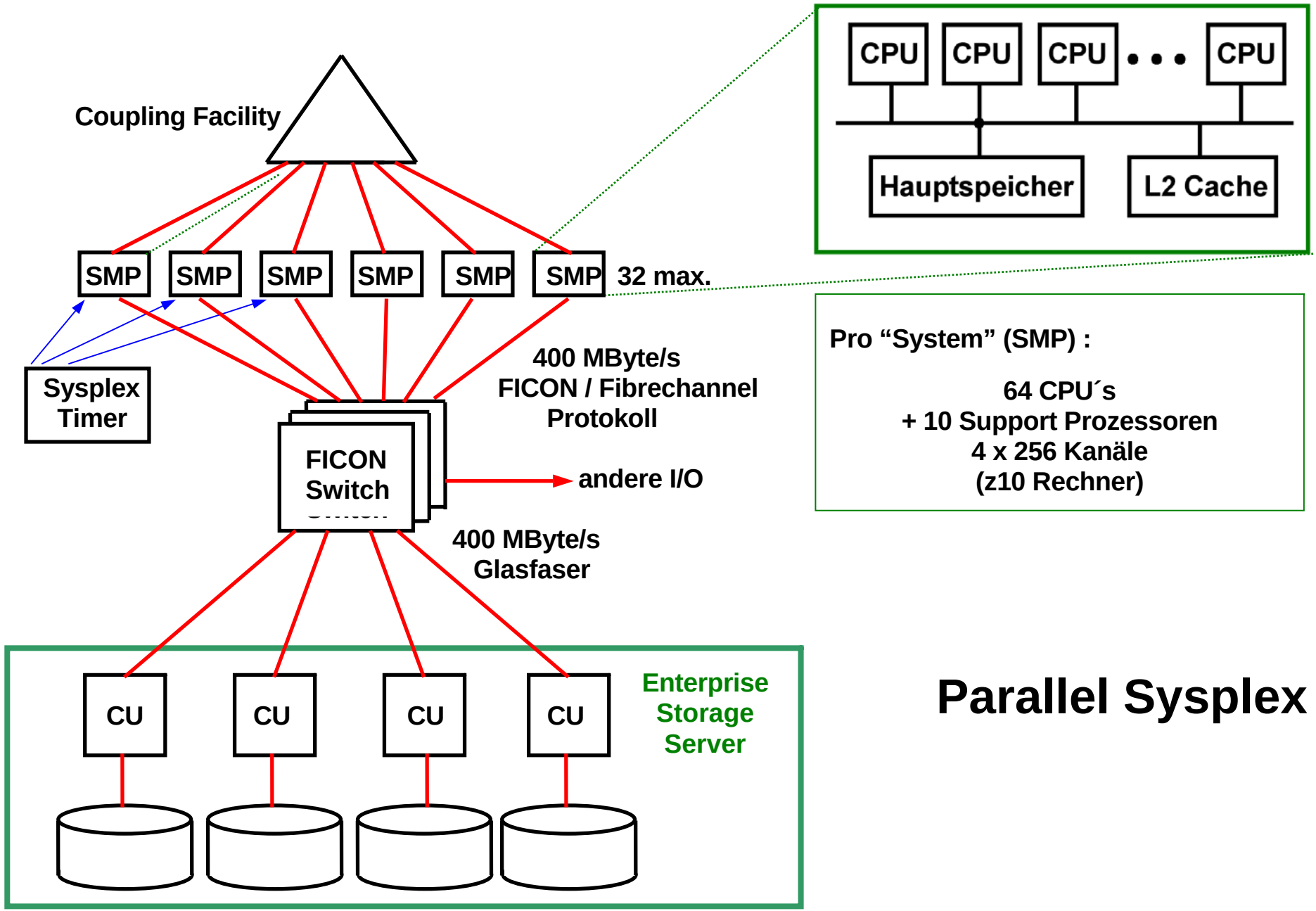
An opening in the side of the cabinet allows two HP 9000 Superdome 32-processor cabinets to be cabled together as an HP 9000 Superdome system with 64 processors.



z10 EC

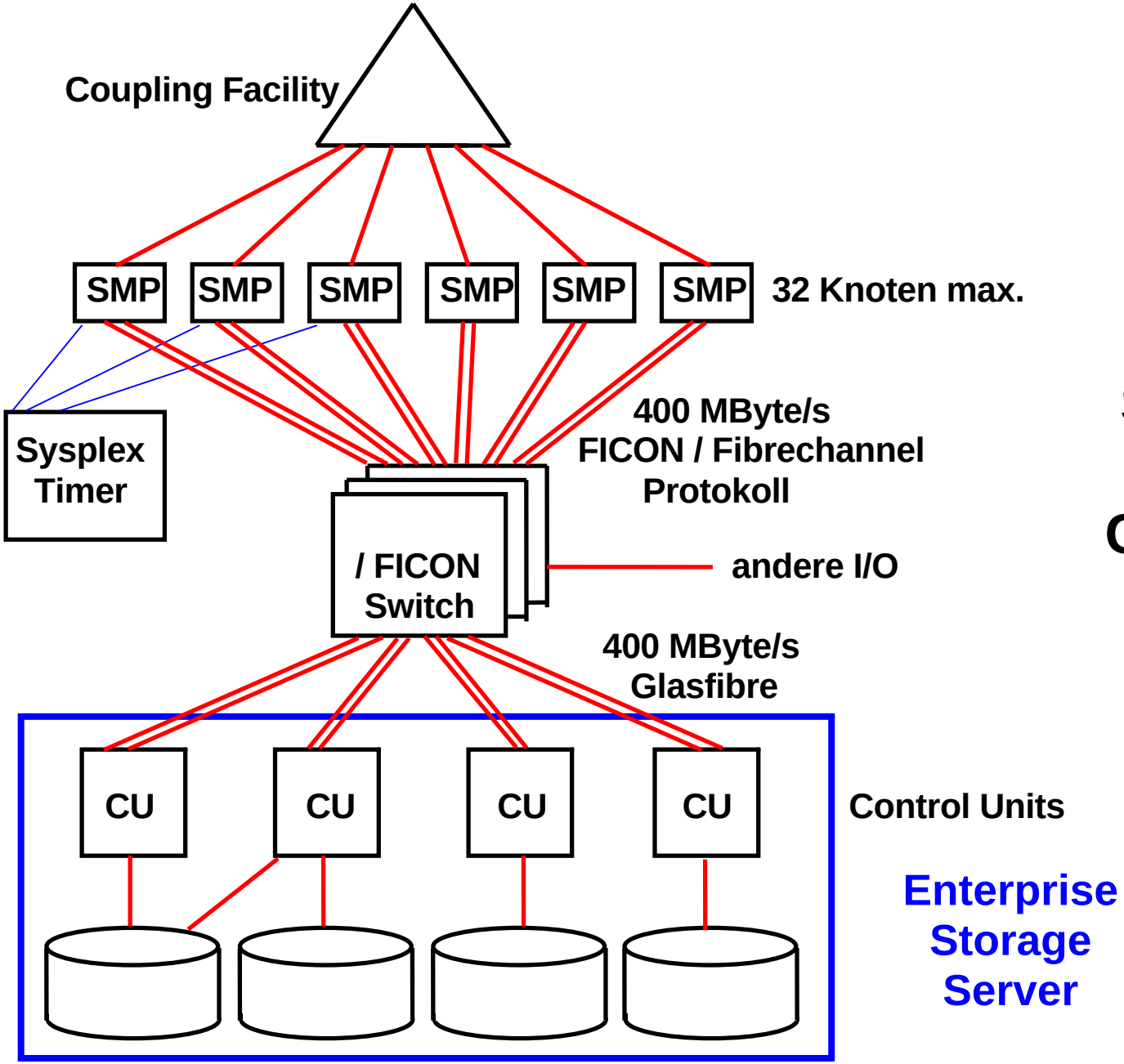


z10 EC internal front view



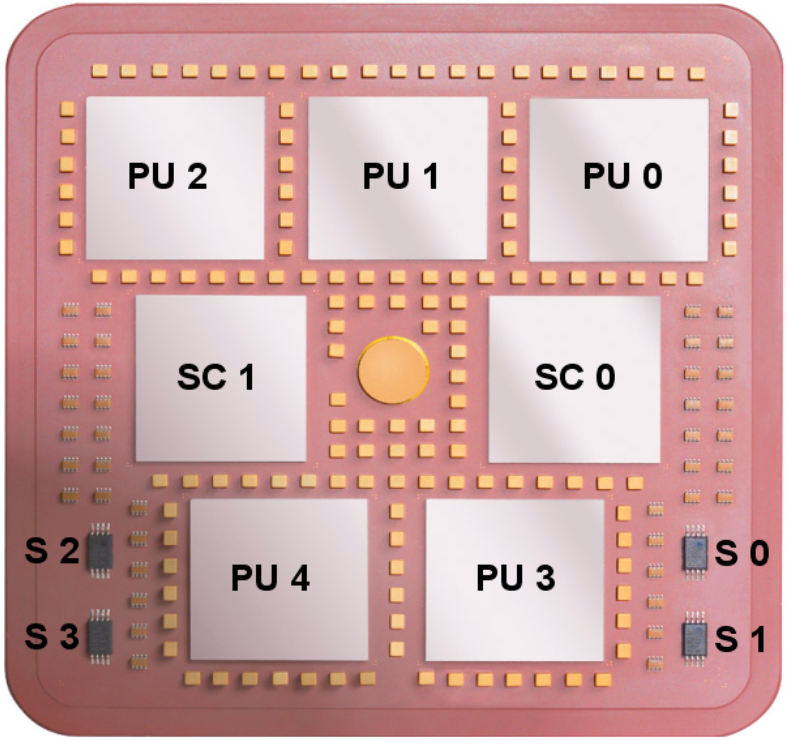
Parallel Sysplex

Sysplex mit Coupling Facility



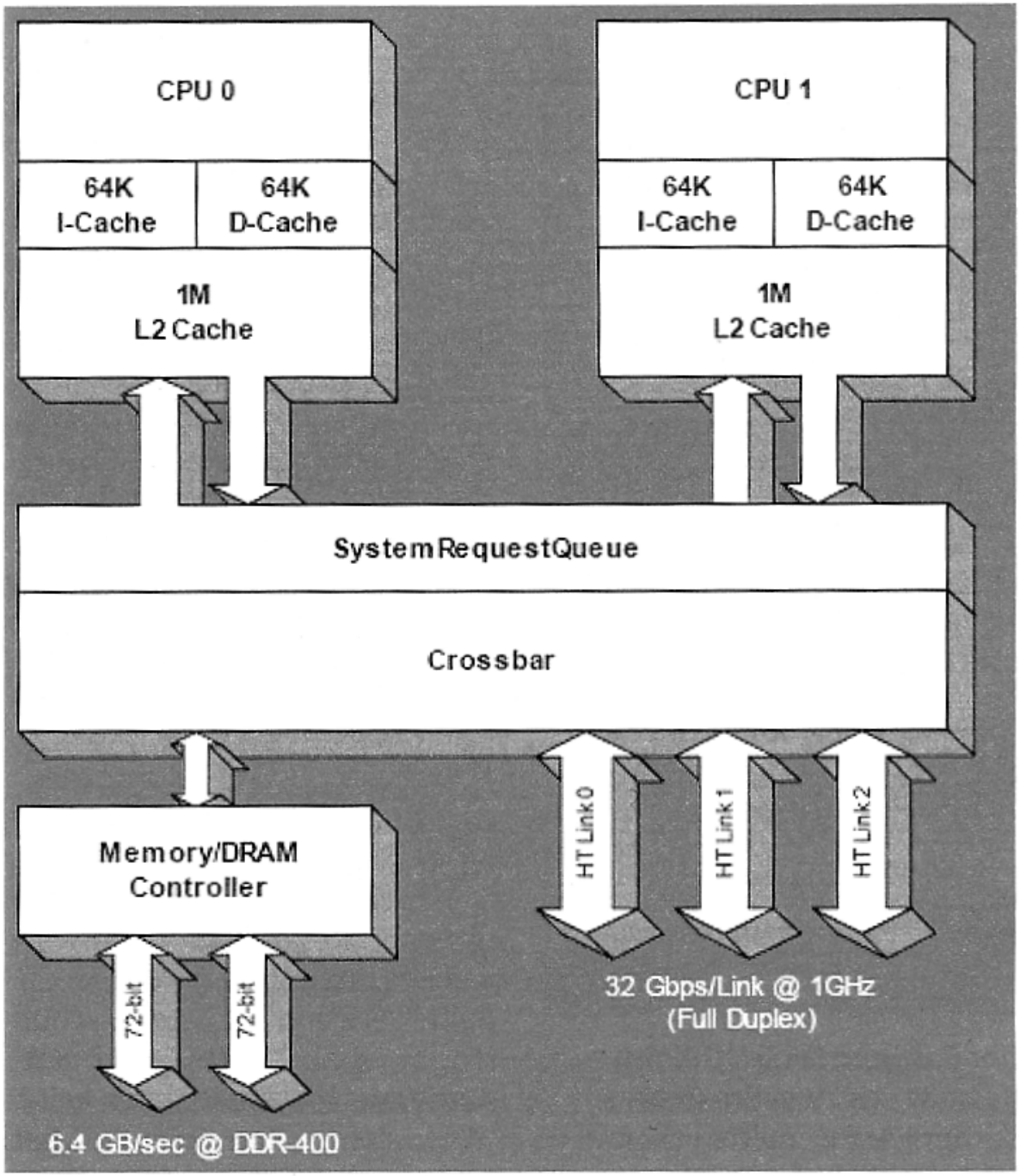
z10 EC Multi-Chip Module (MCM)

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

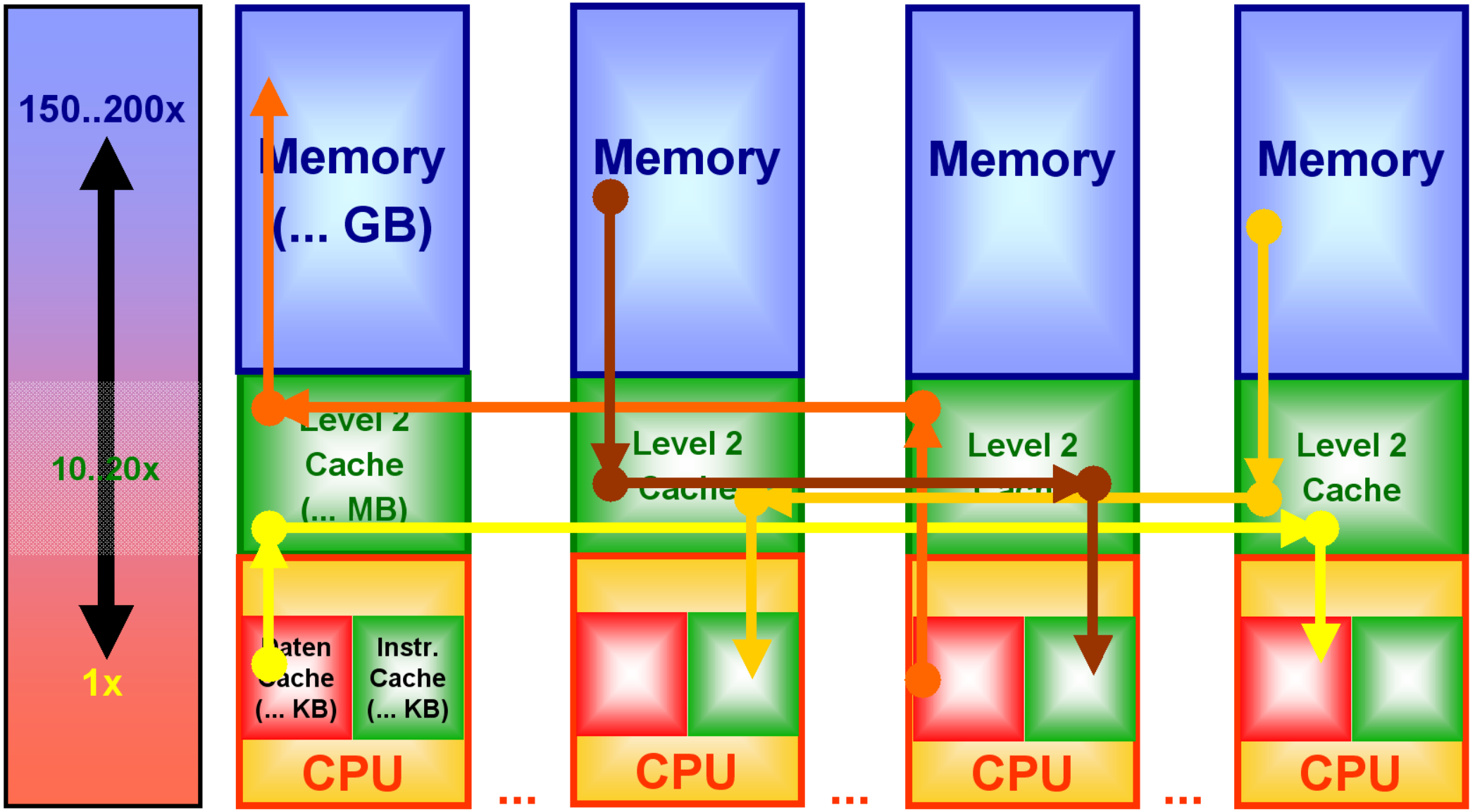


- **CMOS 11s chip Technology**
  - PU, SC, S chips, 65 nm
  - 5 PU chips/MCM – Each up to 4 cores
    - One memory control (MC) per PU chip
    - 21.97 mm x 21.17 mm
    - 994 million transistors/chip
    - L1 cache/PU
      - 64 KB I-cache
      - 128 KB D-cache
    - L1.5 cache/PU
      - 3 MB
   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

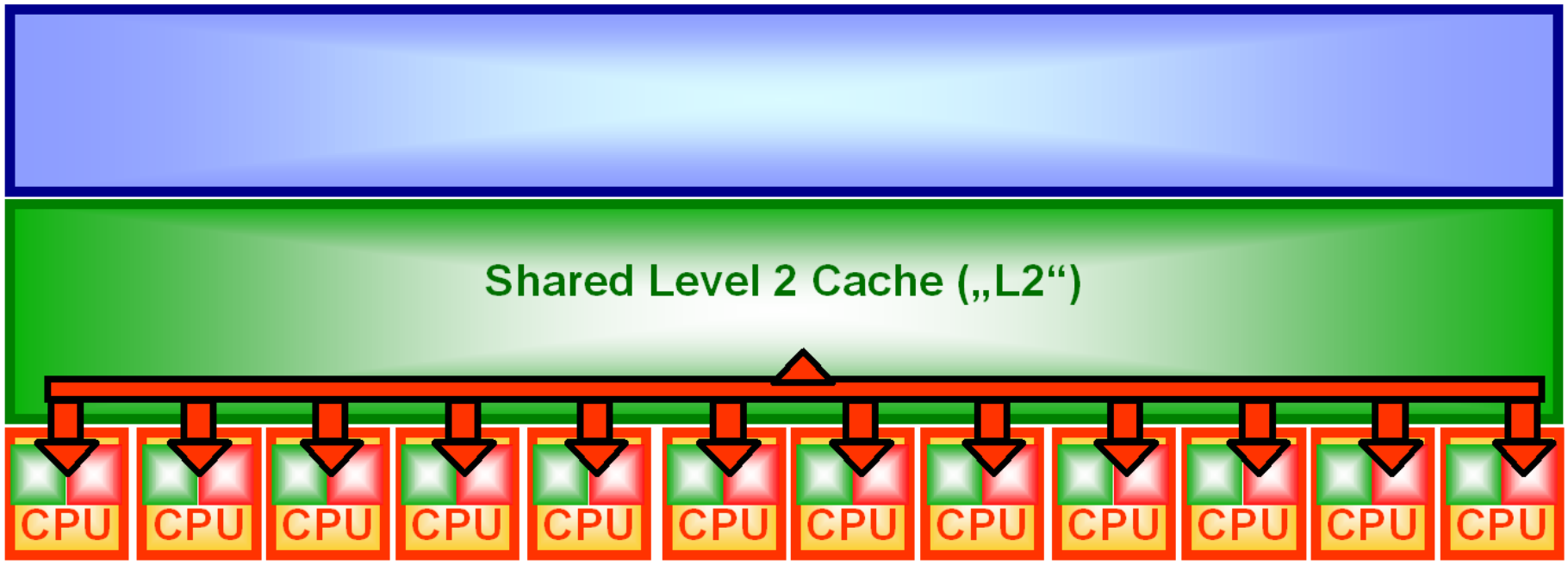
    2 Storage Control (SC) chip
    - 21.11 mm x 21.71 mm
    - 1.6 billion transistors/chip
    - L2 Cache 24 MB per SC chip (48 MB/Book)
    - L2 access to/from other MCMs
  - 4 SEEPROM (S) chips
    - 2 x active and 2 x redundant
    - Product data for MCM, chips and other engineering information
  -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



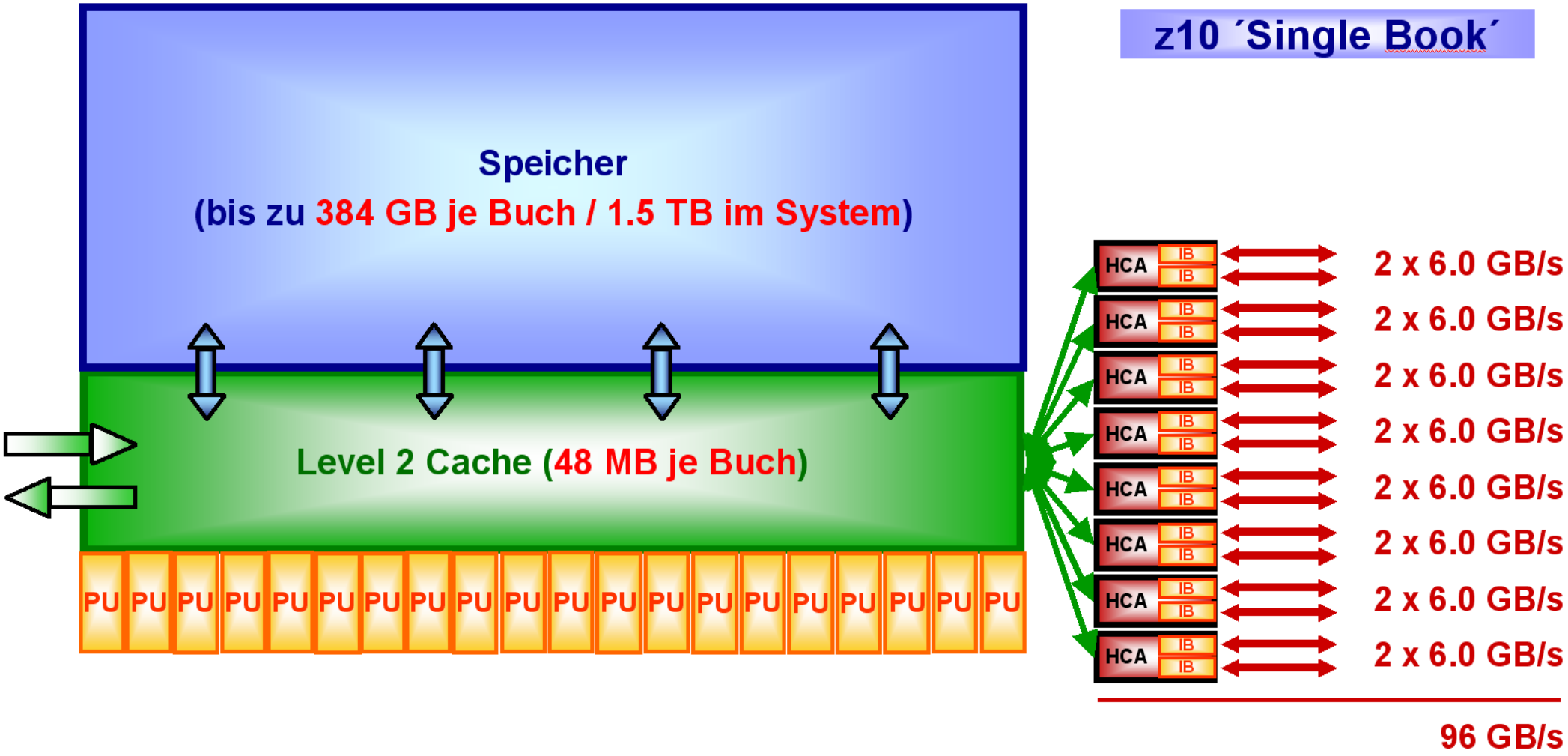
Beim Athlon 64 X2 und dem Opteron hängen beide Kerne an einem schnellen Crossbar-Switch. Über diesen greifen sie auf den Speicher und die Peripherie zu.



Sun und HP Cluster – Kommunikation zwischen den Prozessoren

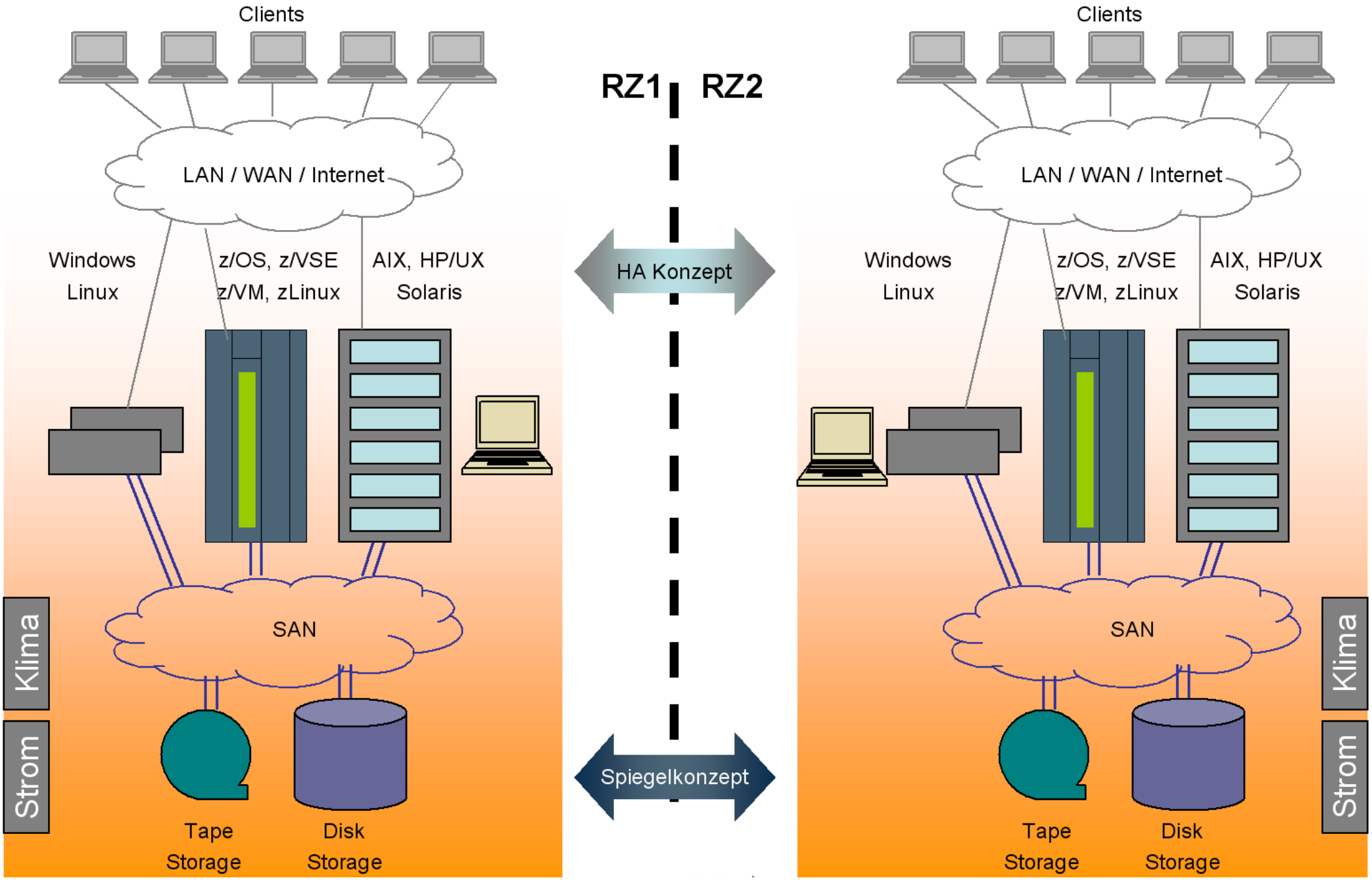


zSeries - Zentraler Switch mit gleichzeitigem Zugriff aller Prozessoren



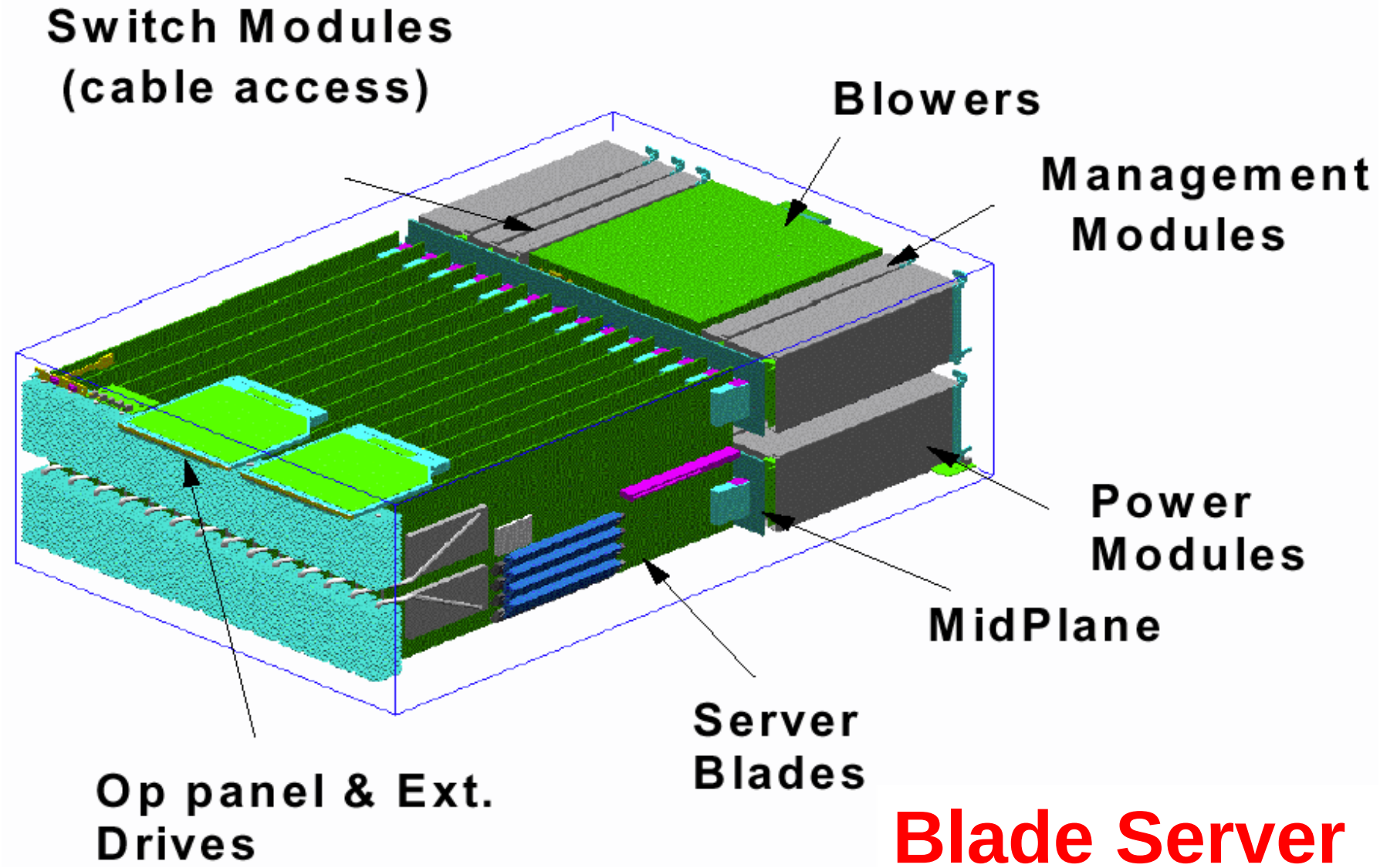
This is another unique z10 (and z9) feature. In all non-IBM systems, I/O adapter cards attach to main memory. In a z10, the Host Channel Adapter (HCA) attaches to the L2 cache, supporting a much higher bandwidth.

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.

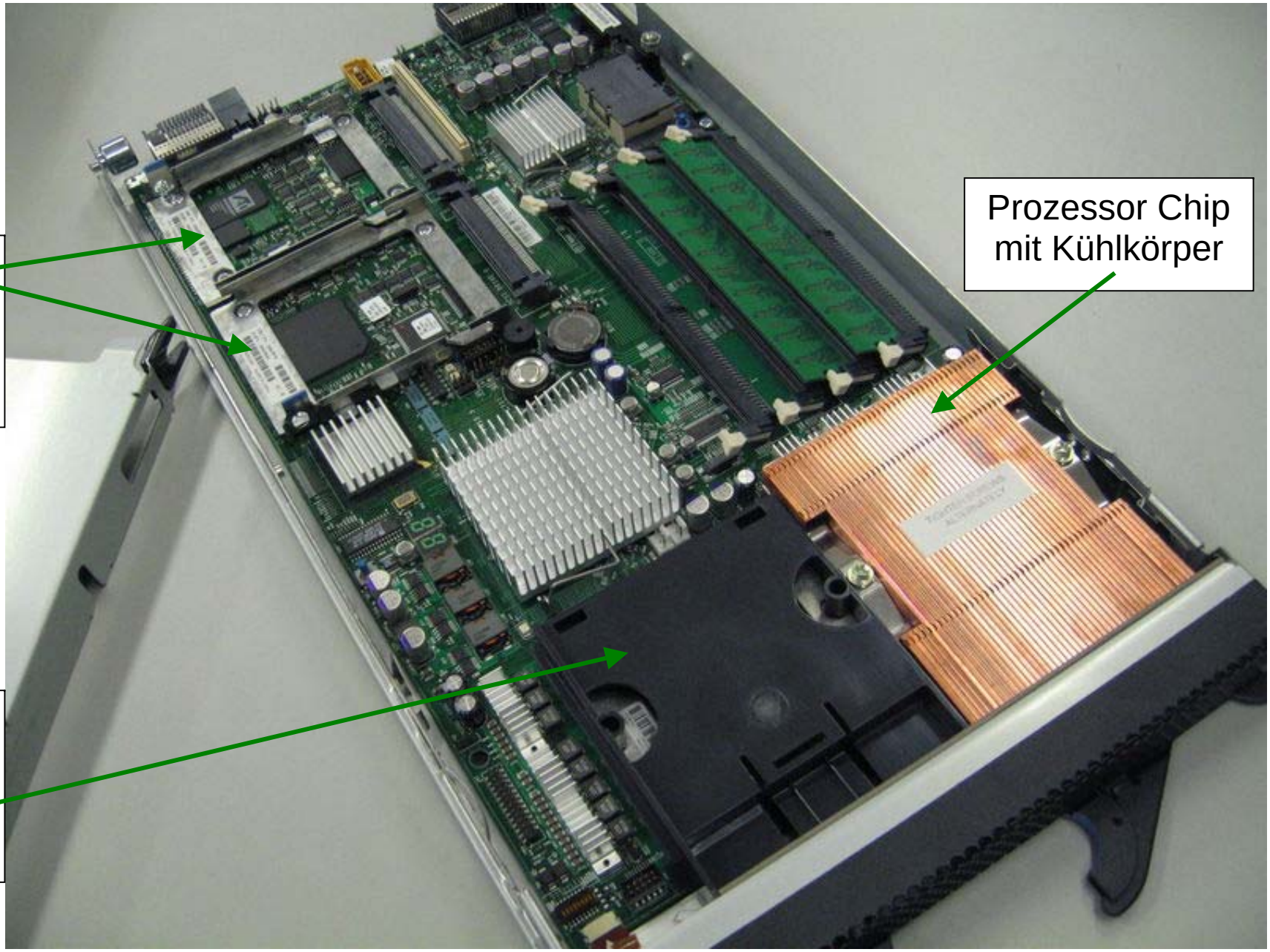

Clients
LAN / WAN / Internet
Windows
Linux
z/OS, z/VSE
z/VM, zLinux
AIX, HP/UX
Solaris
SAN
Klima
Strom
Tape
Storage
Disk
Storage
RZ1
RZ2
HA Konzept
Spiegelkonzept
Clients
LAN / WAN / Internet
Windows
Linux
z/OS, z/VSE
z/VM, zLinux
AIX, HP/UX
Solaris
SAN
Klima
Strom
Tape
Storage
Disk
Storage

Kommerzielle Großrechner

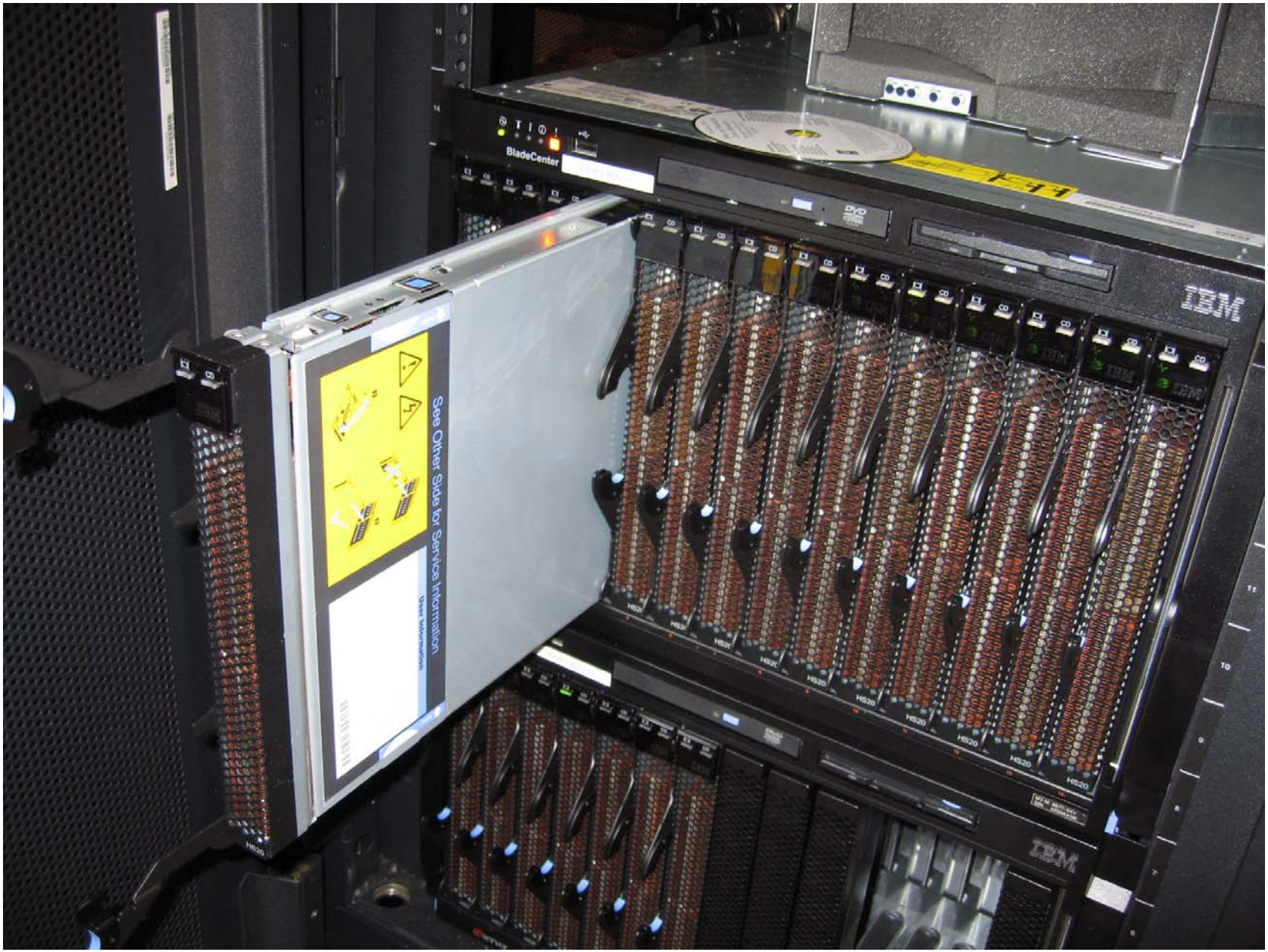
Message Passing Parallel Rechner



Blade Server



IBM Blade for Intel-, PowerPC- und Cell



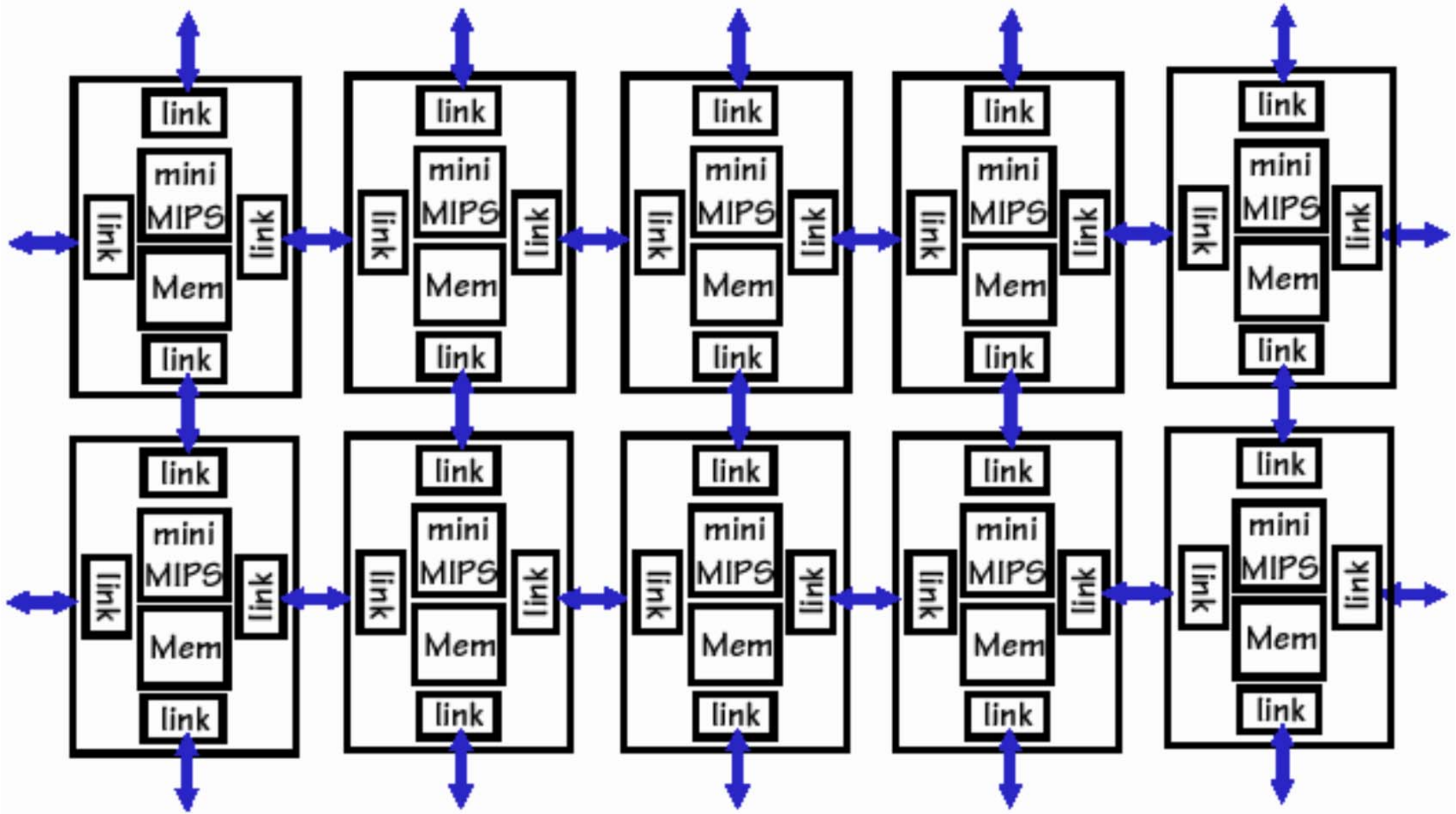
IBM BladeCenter

HP ProLiant BladeSystem

- **Unterstützt bis zu 16 voll ausgestattete Blade-Server.**
- **Bis zu 6 hocheffiziente Netzgeräte für höchste Flexibilität und schwankungsfreien Wechselstromversorgung.**
- **Bis zu 4 redundante I/O-Fabrics.**
- **Integrierte Verwaltung für einfache und unkomplizierte Einrichtung.**
- **Einfache Einrichtung und Fehlerdiagnose direkt an der Vorderseite des Systems.**
- **Einstellbare Gebläsegeschwindigkeit zur Steuerung von Stromverbrauch, Luftbewegung und Geräuschentwicklung.**
- **Optionales zweites Modul für uneingeschränkte Redundanz.**
- **Umschaltbare Stromanschlüsse für Gleichstrom oder Wechselstrom.**

HP ProLiant BladeSystem

- **Unterstützt bis zu 16 voll ausgestattete Blade-Server.**
- **Bis zu 4 redundante I/O-Fabrics.**
- **Integrierte Verwaltung für einfache und unkomplizierte Einrichtung.**
- **Optionales zweites Modul für uneingeschränkte Redundanz.**



Message Passing

Homogeneous CPUs, can leverage existingCPU designs and development tolls. Hardware focusses on communication (2-D Mesh, N-Cube). SW focusses on partittioning of data and algorithms of a single process.


Roadrunner

Roadrunner

- **Roadrunner will primarily be used to ensure the safety and reliability of the nation's nuclear weapons stockpile. It will also be used for research into astronomy, energy, human genome science and climate change.**
- **Roadrunner is the world's first hybrid supercomputer. In a first-of-a-kind design, the Cell Broadband Engine -- originally designed for video game platforms such as the Sony Playstation 3® -- will work in conjunction with x86 processors from AMD.**
- **Made from Commercial Parts. In total, Roadrunner connects 6,948 dual-core AMD Opteron chips (on IBM Model LS21 blade servers) as well as 12,960 Cell engines (on IBM Model QS22 blade servers). The Roadrunner system has 80 terabytes of memory, and is housed in 288 refrigerator-sized, IBM BladeCenter racks occupying 6,000 square feet. Its 10,000 connections – both Infiniband and Gigabit Ethernet -- require 57 miles of fiber optic cable. Roadrunner weighs 500,000 lbs.**
- **Custom Configuration. Two IBM QS22 blade servers and one IBM LS21 blade server are combined into a specialized "tri-blade" configuration for Roadrunner. The machine is composed of a total of 3,456 tri-blades. Standard processing (e.g., file system I/O) is handled by the Opteron processors. Mathematically and CPU-intensive elements are directed to the Cell processors. Each tri-blade unit can run at 400 billion operations per second (400 Gigaflops).**
- **Roadrunner operates on open-source Linux software from Red Hat.**
- **Energy Miser. Compared to most traditional supercomputer designs, Roadrunner's hybrid format sips power (3.9 megawatts) and delivers world-leading efficiency – 376 million calculations per watt. Roadrunner is a top energy-efficient systems.**
- **Roadrunner's massive software effort targets commercial applications for hybrid supercomputing. With corporate and academic partners, IBM is developing an open-source ecosystem that will bring hybrid supercomputing to financial services, energy exploration and medical imaging industries among others. Applications for Cell-based hybrid supercomputing include: calculating cause and effect in capital markets in real-time, supercomputers in financial services can instantly predict the ripple effect of a stock market change throughout the markets. In medicine, complex 3-D renderings of tissues and bone structures will happen in real-time, as patients are being examined.**

Roadrunner costs about $100 million

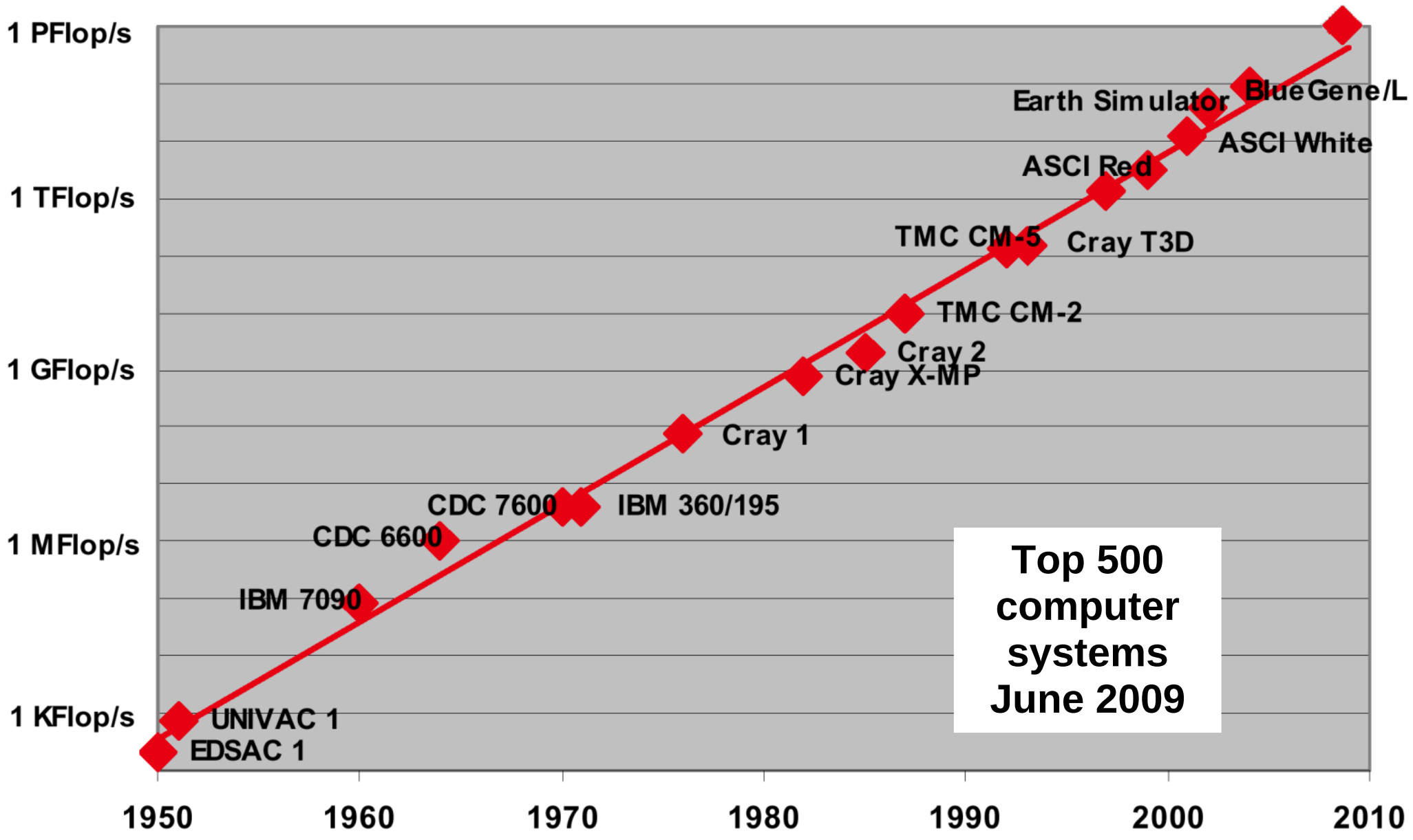
Rang 2009 (Nov. 2008)	Standort	Computer/Hersteller	Cores	Leistung
1 (1)	Los Alamos Lab, USA	Roadrunner (PowerXCell und Opteron, Infiniband), 2008,IBM	129600	1105.00
2 (2)	Oak Ridge National Lab, USA	Jaguar (Cray XT5), 2008, Cray	150152	1059.00
3 (11)	Forschungszentrum Jülich, Deutschland	Jugene (Blue Gene/P), 2009, IBM	294912	825.50
4 (3)	NASA Ames Research, USA	Pleiades (SGI Altix ICE 8200EX), 2008,SGI	51200	487.01
5 (4)	Lawrence Livermore National Lab, USA	BlueGene/L (eServer Blue Gene), 2007, IBM	212992	478.20
6 (15)	University of Tennessee, USA	Kraken XT5 (Cray XT5), 2008, Cray	66000	463.30
7 (5)	Argonne National Lab, USA	Blue Gene/P, 2007, IBM	163840	458.61
8 (6)	University of Texas, USA	Ranger (SunBlade, Opteron, Infiniband), 2008,Sun	62976	433.20
9 (-)	Lawrence Livermore National Lab, USA	Dawn (Blue Gene/P), 2009, IBM	147456	415.70
10 (-)	Forschungszentrum Jülich, Deutschland	Juropa (NovaScale R422-E2/Sun Blade 6048, Intel Xeon X5570, Infiniband, Parastation/Partec-MPI), 2009, Bull SA0	26304	274.8

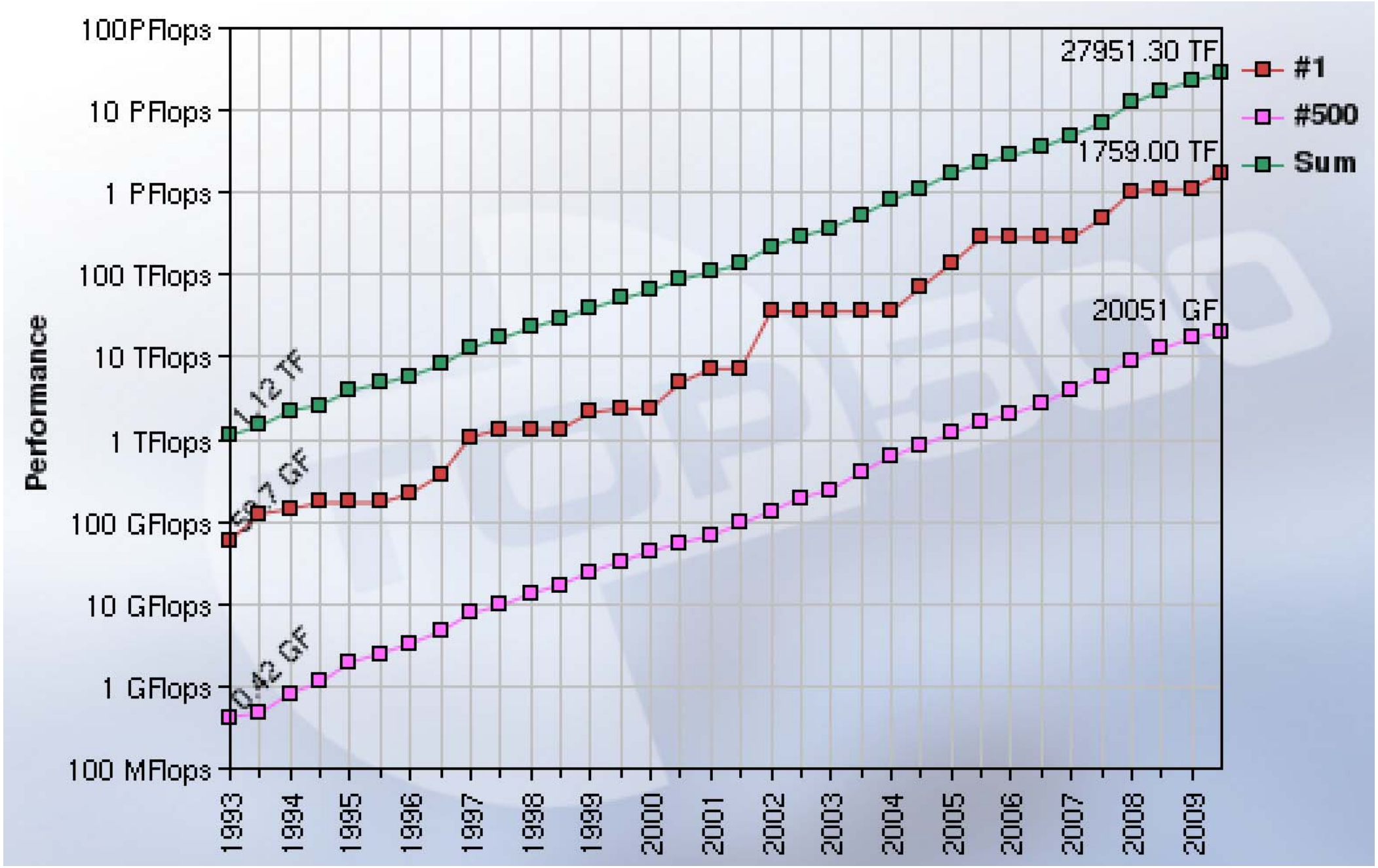
Top 10 Computer Systems June 2009

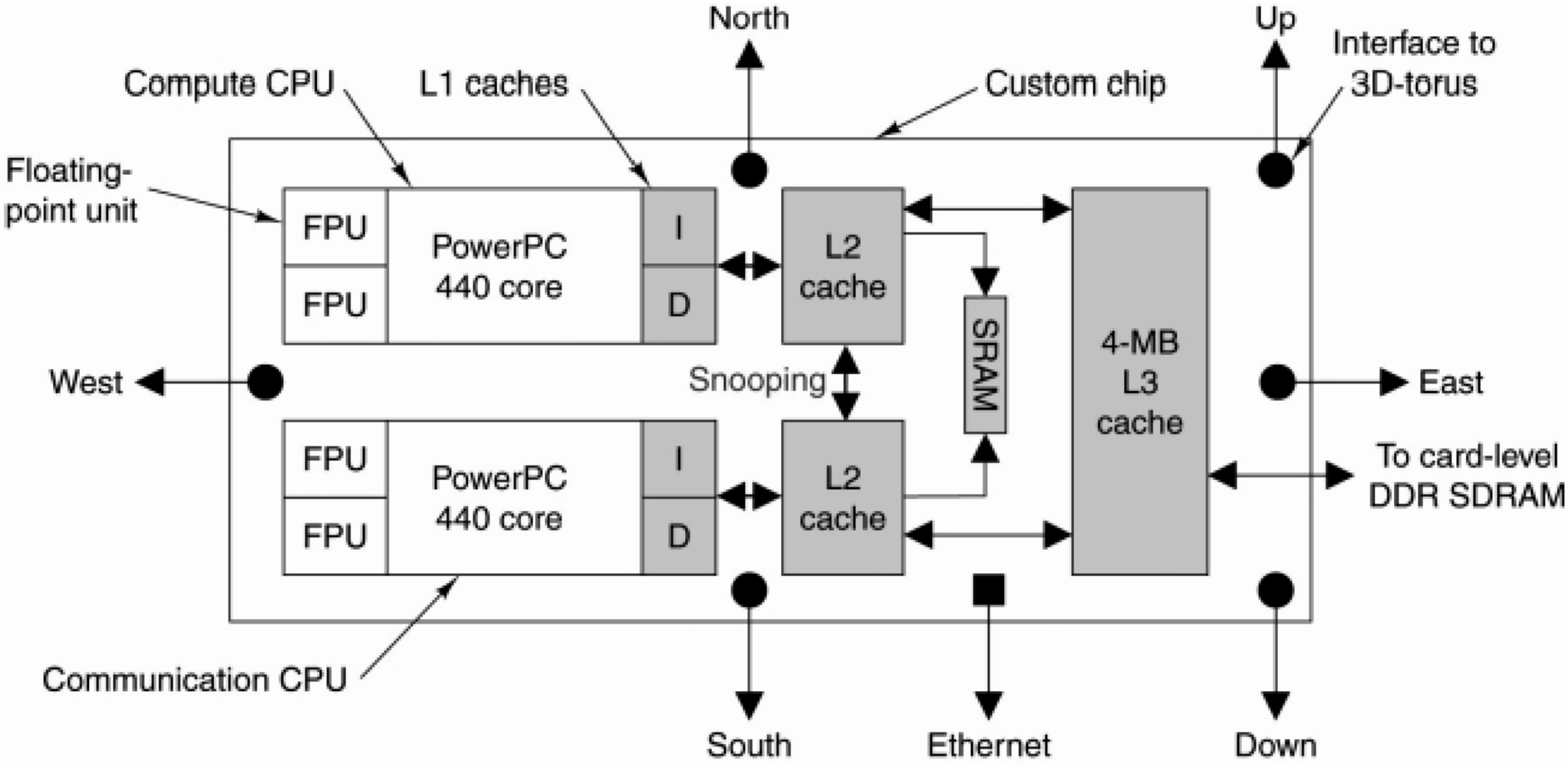
Rank	Site	Computer
1	Oak Ridge National Laboratory United States	Jaguar - Cray XT5-HE Opteron Six Core 2.6 GHz Cray Inc.
2	DOE/NNSA/LANL United States	Roadrunner - BladeCenter QS22/LS21 Cluster, PowerXCell 8i 3.2 Ghz / Opteron DC 1.8 GHz, Voltaire Infiniband IBM
3	National Institute for Computational Sciences/University of Tennessee United States	Kraken XT5 - Cray XT5-HE Opteron Six Core 2.6 GHz Cray Inc.
4	Forschungszentrum Juelich (FZJ) Germany	JUGENE - Blue Gene/P Solution IBM
5	National SuperComputer Center in Tianjin/NUDT China	Tianhe-1 - NUDT TH-1 Cluster, Xeon E5540/E5450, ATI Radeon HD 4870 2, Infiniband NUDT
6	NASA/Ames Research Center/NAS United States	Pleiades - SGI Altix ICE 8200EX, Xeon QC 3.0 GHz/Nehalem EP 2.93 Ghz SGI
7	DOE/NNSA/LLNL United States	BlueGene/L - eServer Blue Gene Solution IBM
8	Argonne National Laboratory United States	Blue Gene/P Solution IBM
9	Texas Advanced Computing Center/Univ. of Texas United States	Ranger - SunBlade x6420, Opteron QC 2.3 Ghz, Infiniband Sun Microsystems
10	Sandia National Laboratories / National Renewable Energy Laboratory United States	Red Sky - Sun Blade x6275, Xeon X55xx 2.93 Ghz, Infiniband Sun Microsystems

Top 10 Computer Systems November 2009

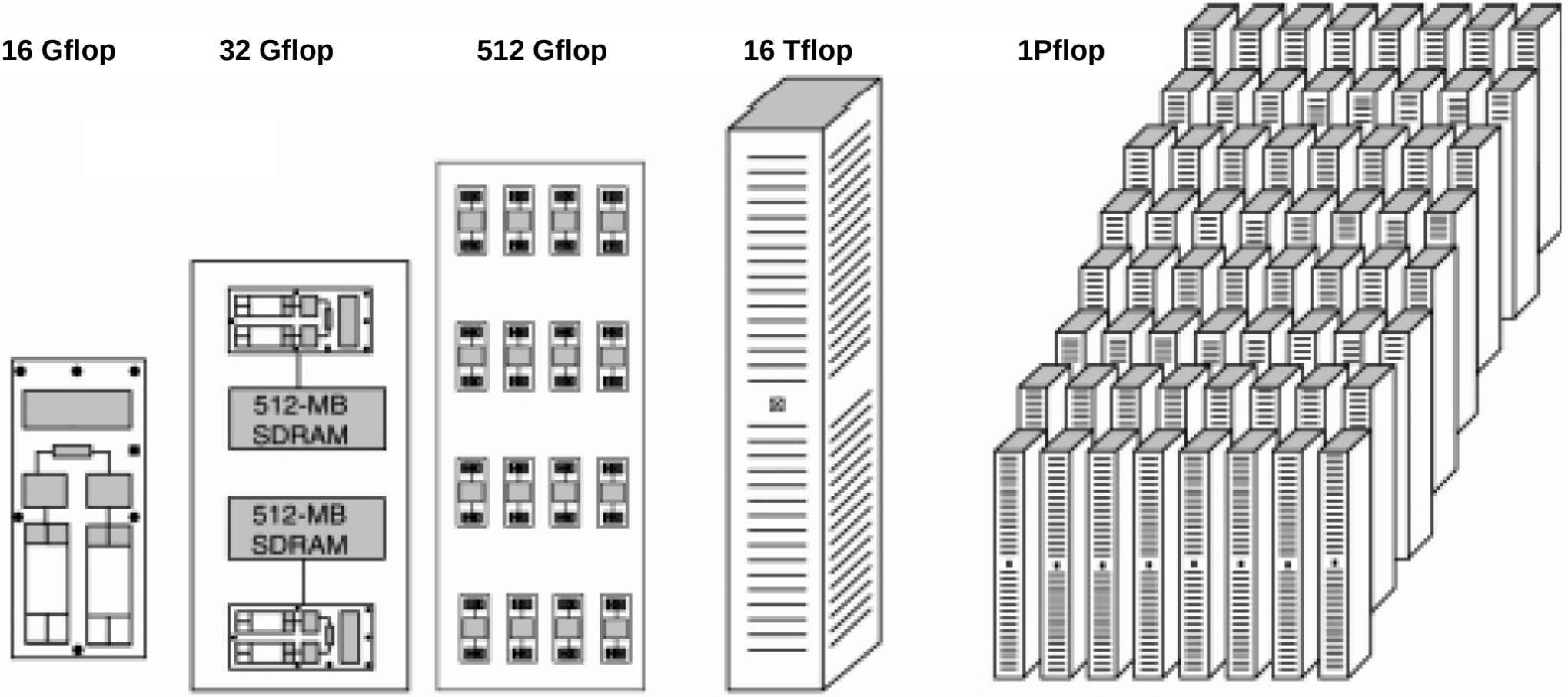

52 80
JAGUAR


1 PFlop/s
1 TFlop/s
1 GFlop/s
1 MFlop/s
1 KFlop/s
1950
1960
1970
1980
1990
2000
2010
Earth Simulator
BlueGene/L
ASCI White
ASCI Red
TMC CM-5
Cray T3D
TMC CM-2
Cray 2
Cray X-MP
Cray 1
CDC 7600
IBM 360/195
CDC 6600
IBM 7090
UNIVAC 1
EDSAC 1
Top 500 computer systems June 2009


Performance
100 PFlops
10 PFlops
1 PFlops
100 TFlops
10 TFlops
1 TFlops
100 GFlops
10 GFlops
1 GFlops
100 MFlops
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
27951.30 TF
1759.00 TF
20051 GF
1.12 TF
59.7 GF
0.42 GF
#1
#500
Sum



BlueGene/L custom processor chip



Chip:	Card:	Board	Cabinet	System
	2 Chips	16 Cards	32 Boards	64 Cabinets
	1 GB	32 Chips	512 Cards	2048 Boards
		16 GB	1024 Chips	32,768 Cards
			512 GB	65,536 Chips
				32 TB

BlueGene System