

CPU-Update

**best OpenSystems Day
Herbst 2006**

Unterföhring

Wolfgang Stief
stief@best.de

Senior Systemingenieur
best Systeme GmbH
GUUG Board Member



■ Welcome to 64 Bit!

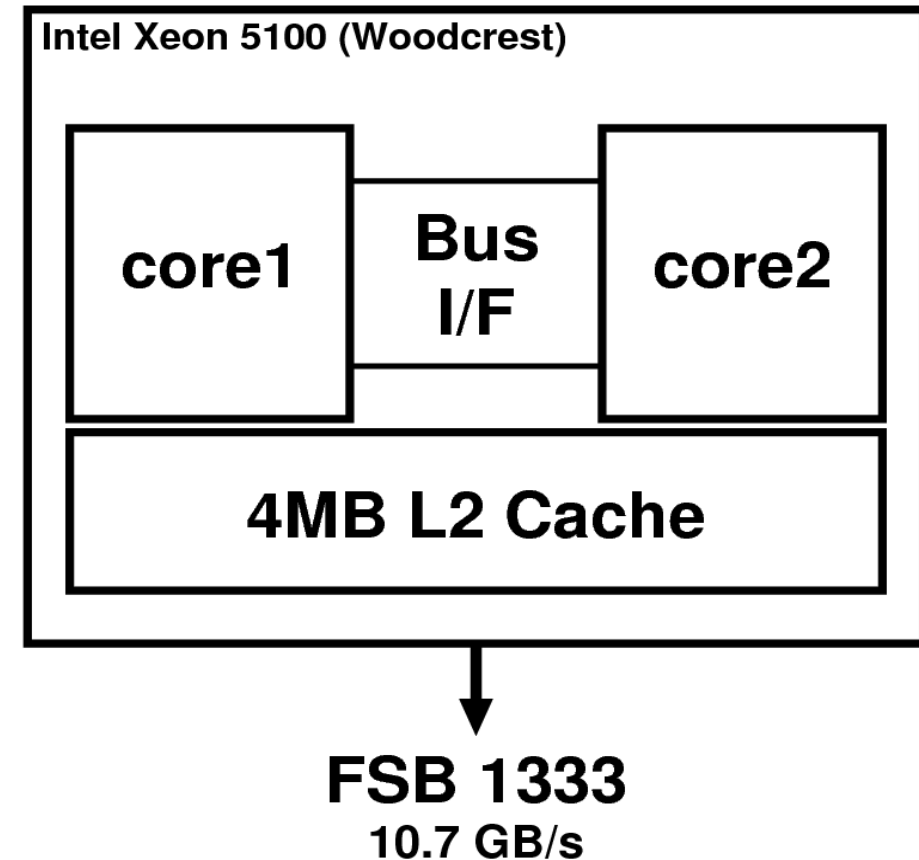
- Sämtliche Server-CPU's sind mittlerweile 64bit
- Alle wichtigen Server-Betriebssysteme sind in aktueller Version 64bit: Linux, {Net,Free,Open}BSD, Windows, Solaris, AIX, HP-UX, IRIX

■ CPU-Trends allgemein

- Multithreaded: 2...8 Threads
- Multicore: 2...8 Cores
- zunehmend Special Purpose Units (SPU) on Chip: Hypervisor, Crossbar, Netzwerk I/F, PCIe etc.
- No more „*One size fits all.*“

- **Intel Woodcrest + Intel Clovertown**
(Dualcore / Quadcore)
- **AMD Opteron + AMD Barcelona**
(Dualcore / Quadcore)
- **Sun UltraSPARC T1 und T2 (Niagara-2)**
- **Sun UltraSPARC IV+, Fujitsu SPARC64 VI und VI+**
- **Benchmark-Abenteuer**

- Dualcore, 65nm Prozess
- 4MB shared L2 Cache
- dual 1333 MHz FSB Interface (max)
- 1.6GHz / 1.86GHz @ 1066MHz FSB (8.5GB/s), 2.0GHz...3.0GHz @ 1333MHz FSB (10.7GB/s)
- 65W-85W Stromaufnahme
40W @ 2.33 GHz geplant



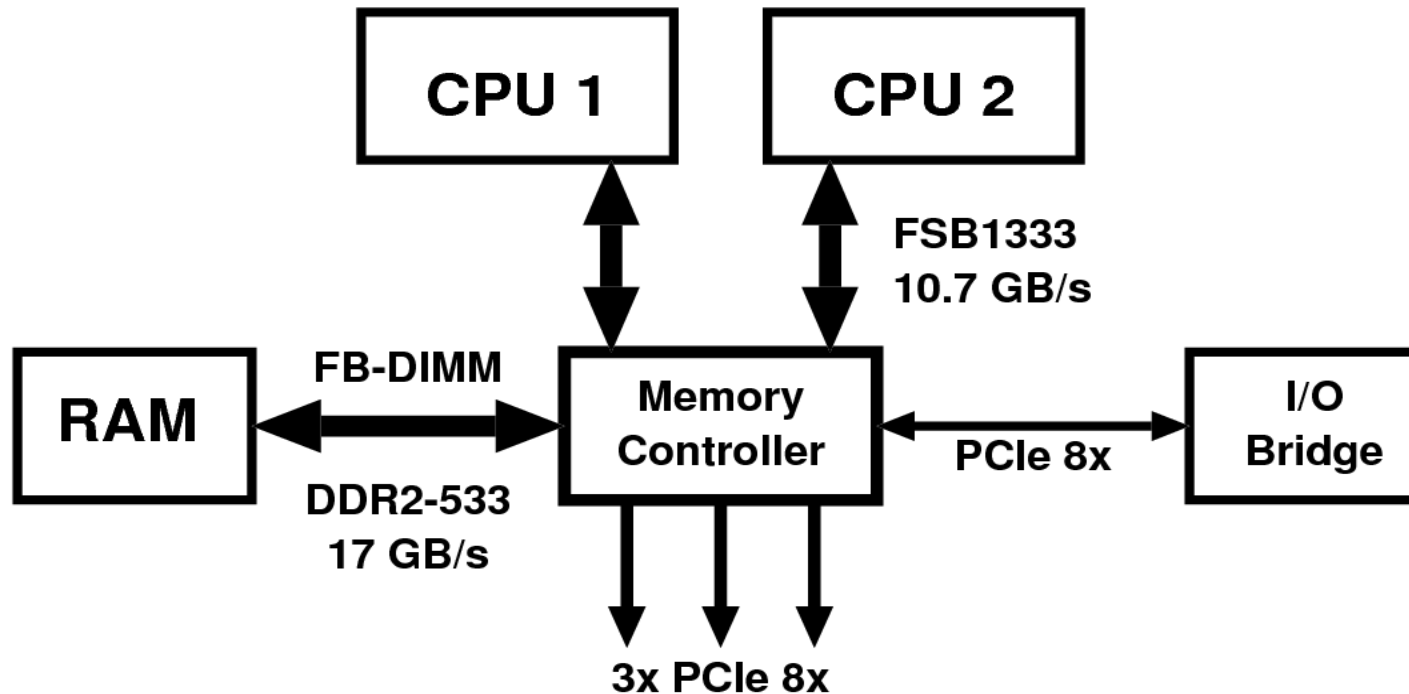
■ Smart Cache

- Data Sharing im Cache
- mehr Cache Hit Rate
- keine Cache Data Replication
- insgesamt bessere Cache-Ausnutzung

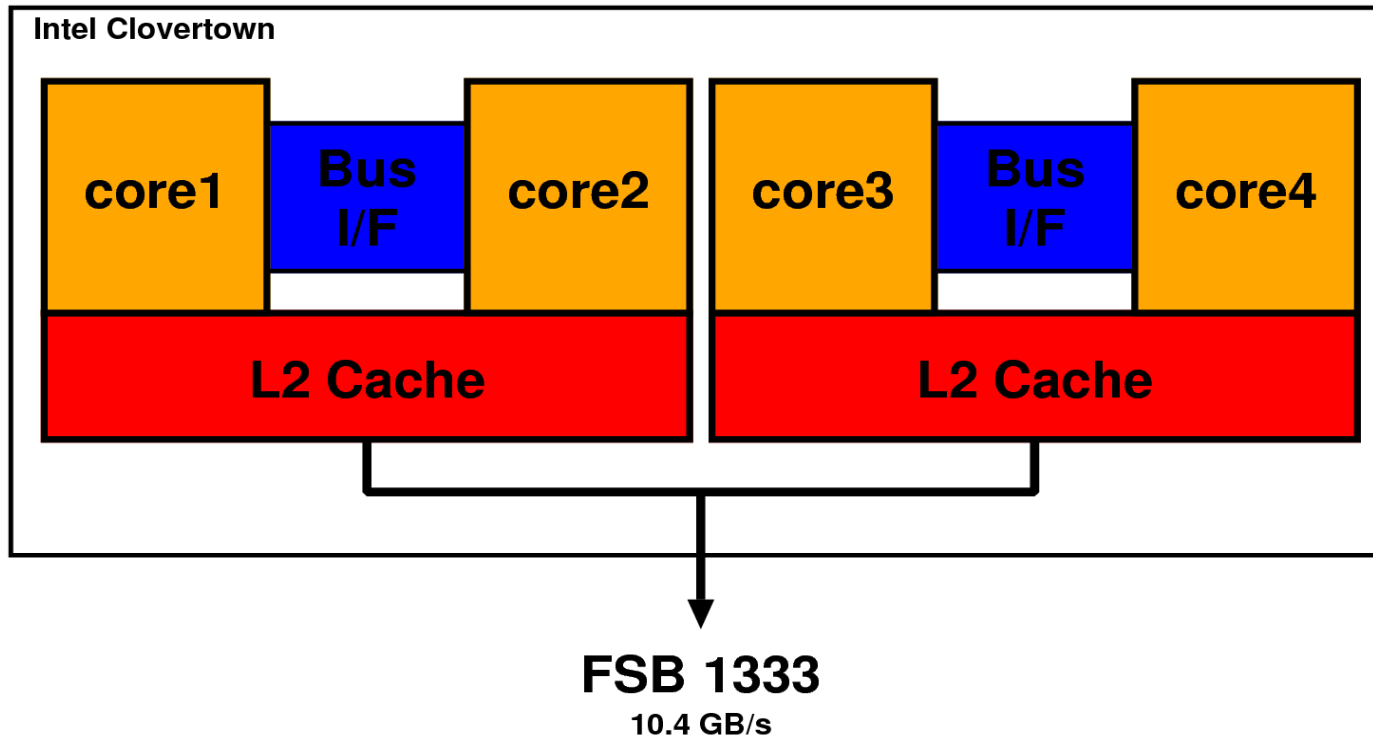
Haken bei Intel:

„In all these designs, neither core realizes the other lies in close physical proximity“

(Nathan Brookwood, Insight 64)



- switched architecture, Intel 5000P Chipset (Blackford)
- Memory Controller (= Northbridge) kann Engpass werden
- Je nach Benchmark ist Woodcrest od. Opteron vorne



- Quadcore, 65nm Prozess
- Ad hoc Design: 2x Dualcore zusammengepappt
- Engpass: FSB ↔ Memory

Insight64 zum Design:

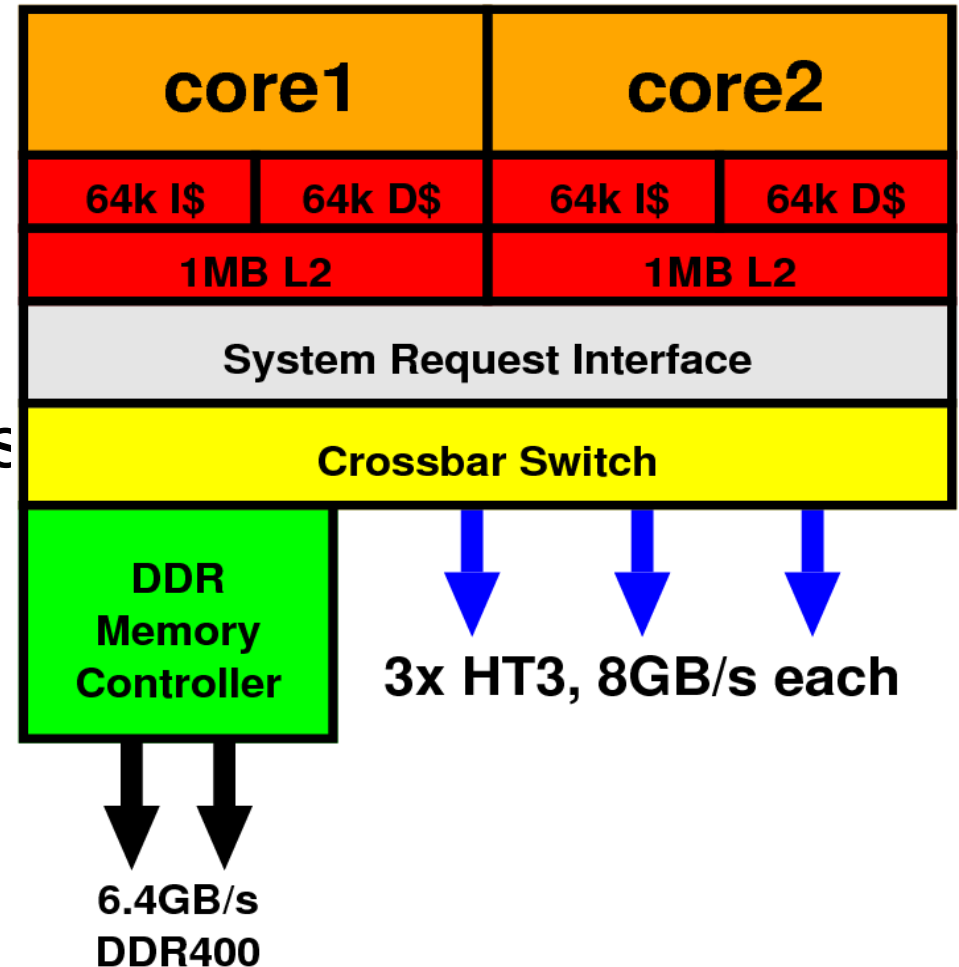
„Based on the timing of the recent drumbeats, we believe that Clovertown will consist of two woodcrest dice crammed into a single package, as illustrated. We've seen this movie before, and it didn't have a ending then, either.“

Hypervisor Technologie:

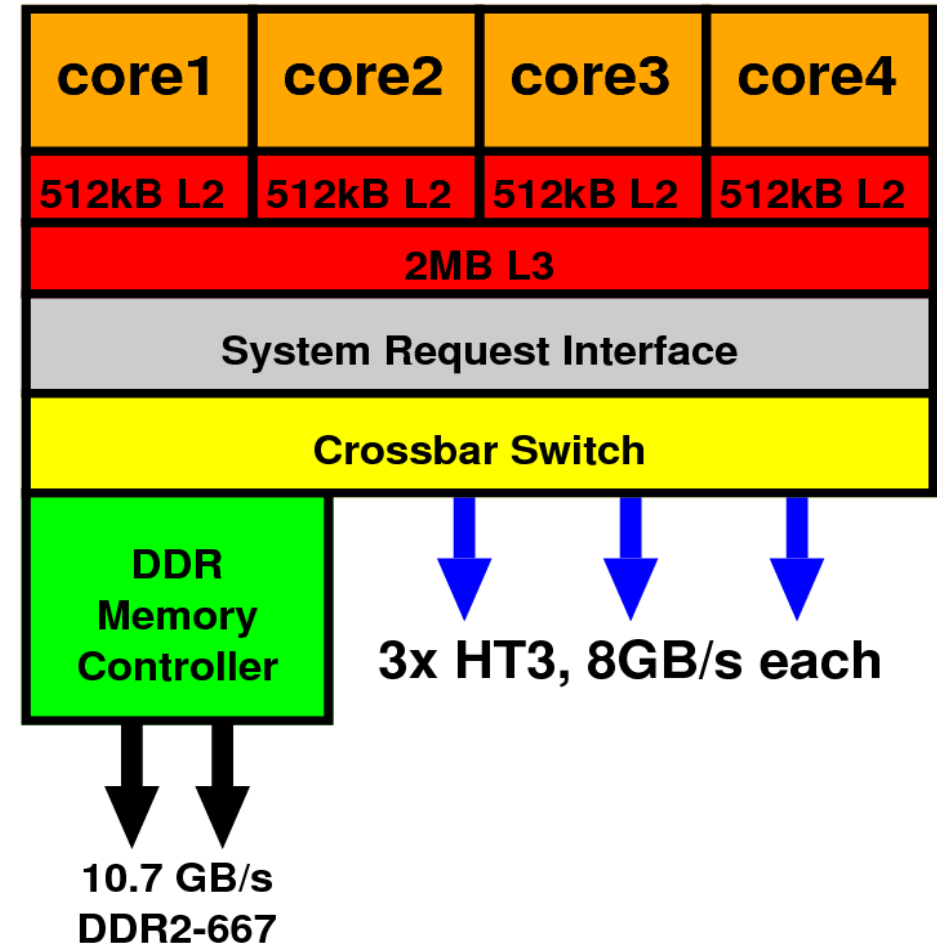
- Vanderpool

- Intel Woodcrest + Intel Clovertown
(Dualcore / Quadcore)
- **AMD Opteron + AMD Barcelona**
(Dualcore / Quadcore)
- Sun UltraSPARC T1 und T2 (Niagara-2)
- Sun UltraSPARC IV+, Fujitsu SPARC64 VI und VI+
- Benchmark-Abenteuer

- Dualcore, 90nm Prozess
- 1MB L2 Cache je Core
- Crossbar on Chip
- MMU on Chip: DDR400, 6.4GB/s memory bandwidth (dual core Rev. E)
- Stromaufnahme max. 90W bzw. 120W, wird von AMD garantiert (design goal)
- Neighbours + I/O: Hypertransport, 3x 8GB/s



- Quadcore, 65nm Prozess
- 2MB shared L3 Cache (?)
- Crossbar on Chip
- MMU on Chip: DDR2-667, 10.7GB/s memory bandwidth (dual core Rev. F + quad core)
- Stromaufnahme max. 90W bzw. 120W, wird von AMD garantiert (design goal)
- Neighbours + I/O: Hypertransport, 3x 8GB/s

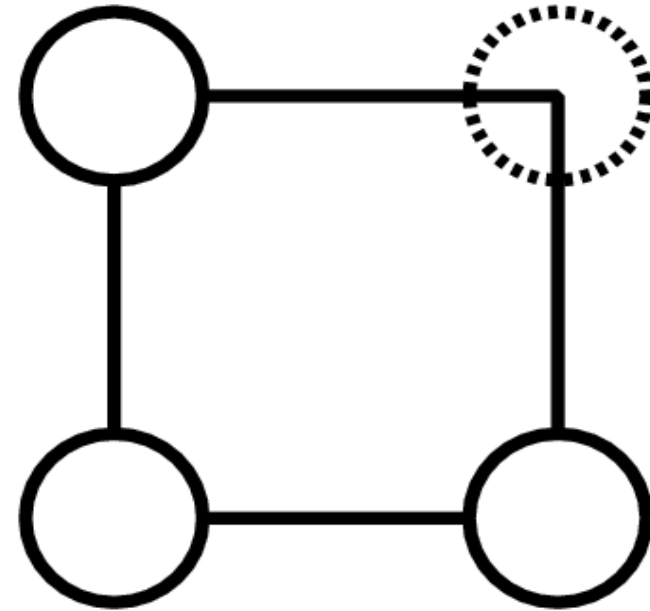
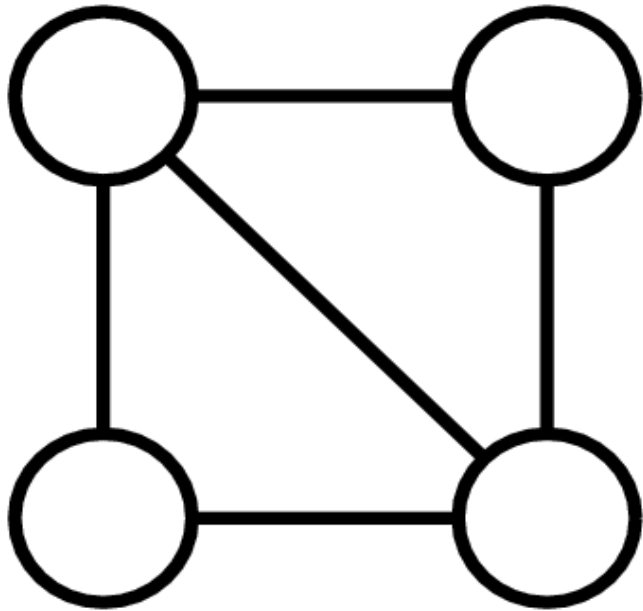


- Vorteil: Memory-Bandbreite zu lokalem RAM
- Nachteil: NUMA (Latenz bei Memory-Zugriff, Cache Coherency)

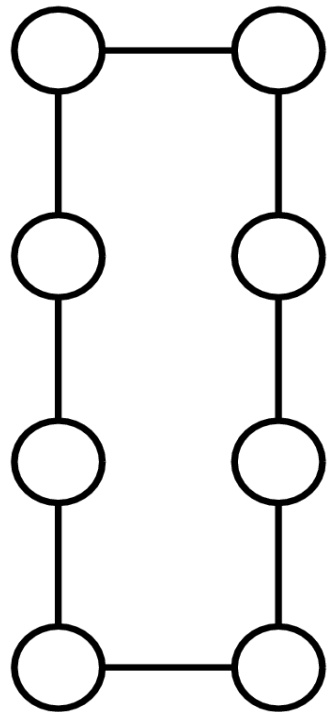
Insight 64:

„AMD system partners accomplished the move from single to dual-core systems by merely „dropping“ dual-core Opterons into the socket previously designated for single-core processors, a far simpler move than the gymnastics Intel's system OEMs had to undertake in their move to dualcore Xeon systems“

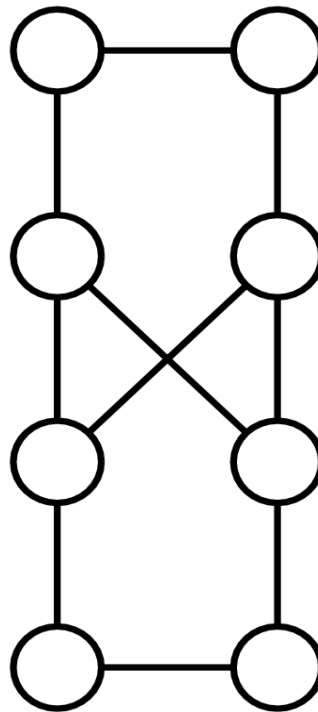
- Dual-Core und Quad-Core haben den selben Sockel F und das selbe Chipset, also einfaches Drop-In Replacement



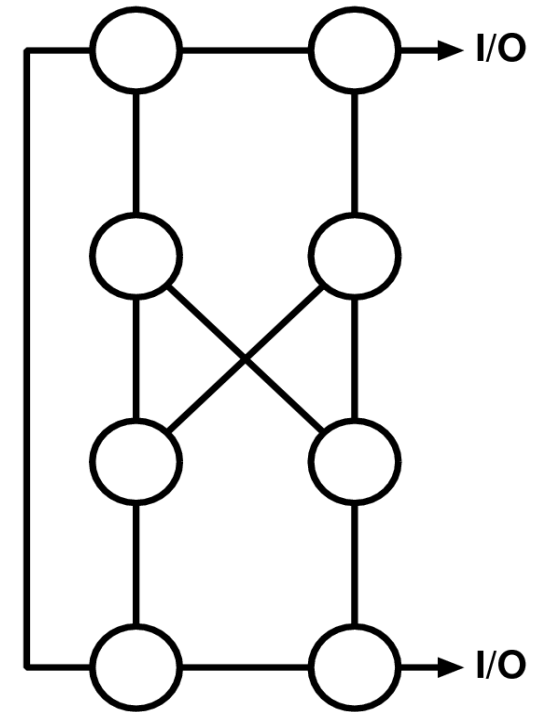
- u. U. Performance-Gewinn bei 4 → 3 CPUs wg. Verkürzung der Memory-Latenzen



ladder



twisted ladder

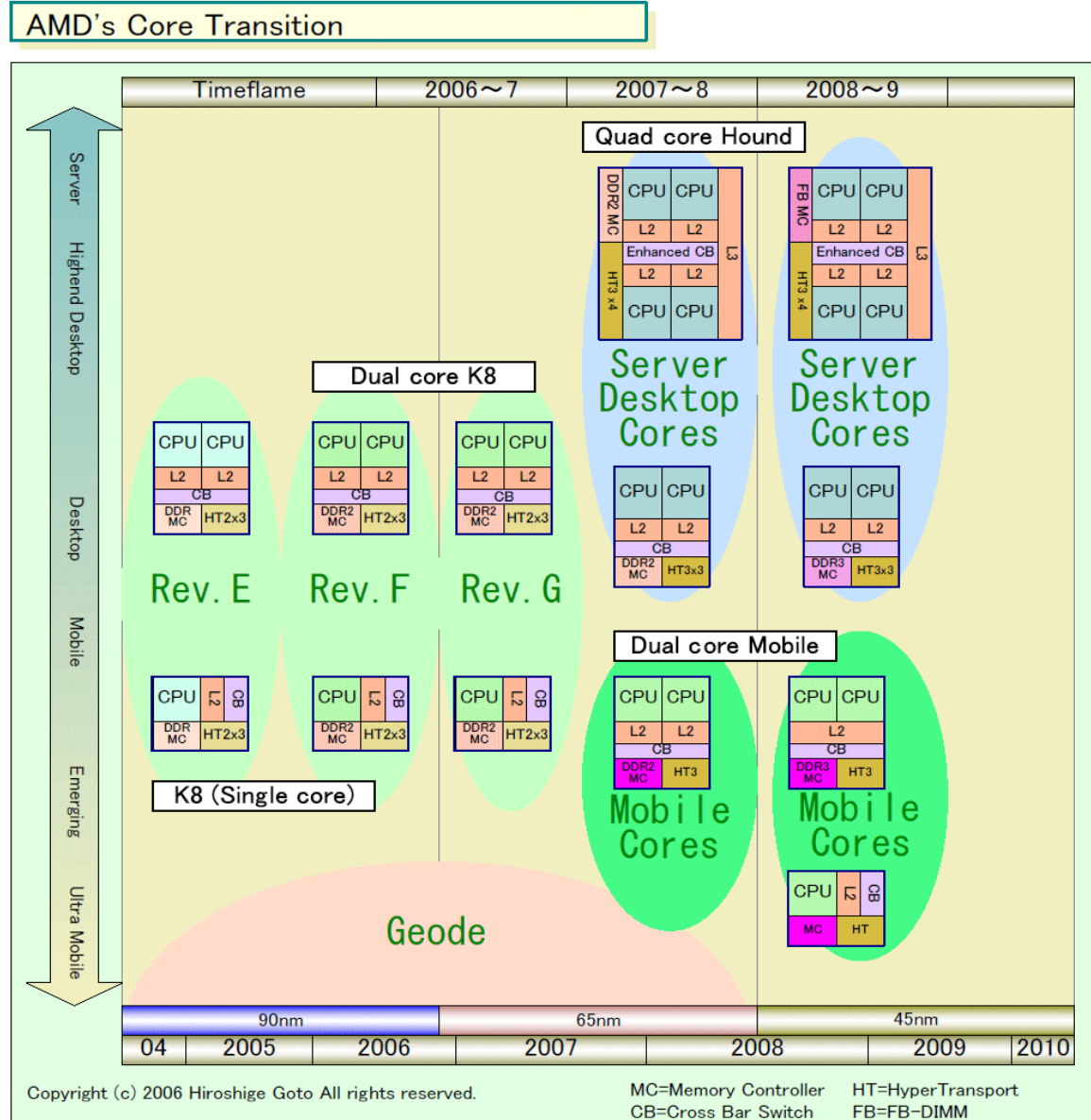


extended twisted ladder

- entscheidend: Anzahl Hops ins nicht lokale RAM
- Ausfall einer CPU = Ausfall des Systems, kein Blacklisting
- Sun X4600: extended twisted ladder

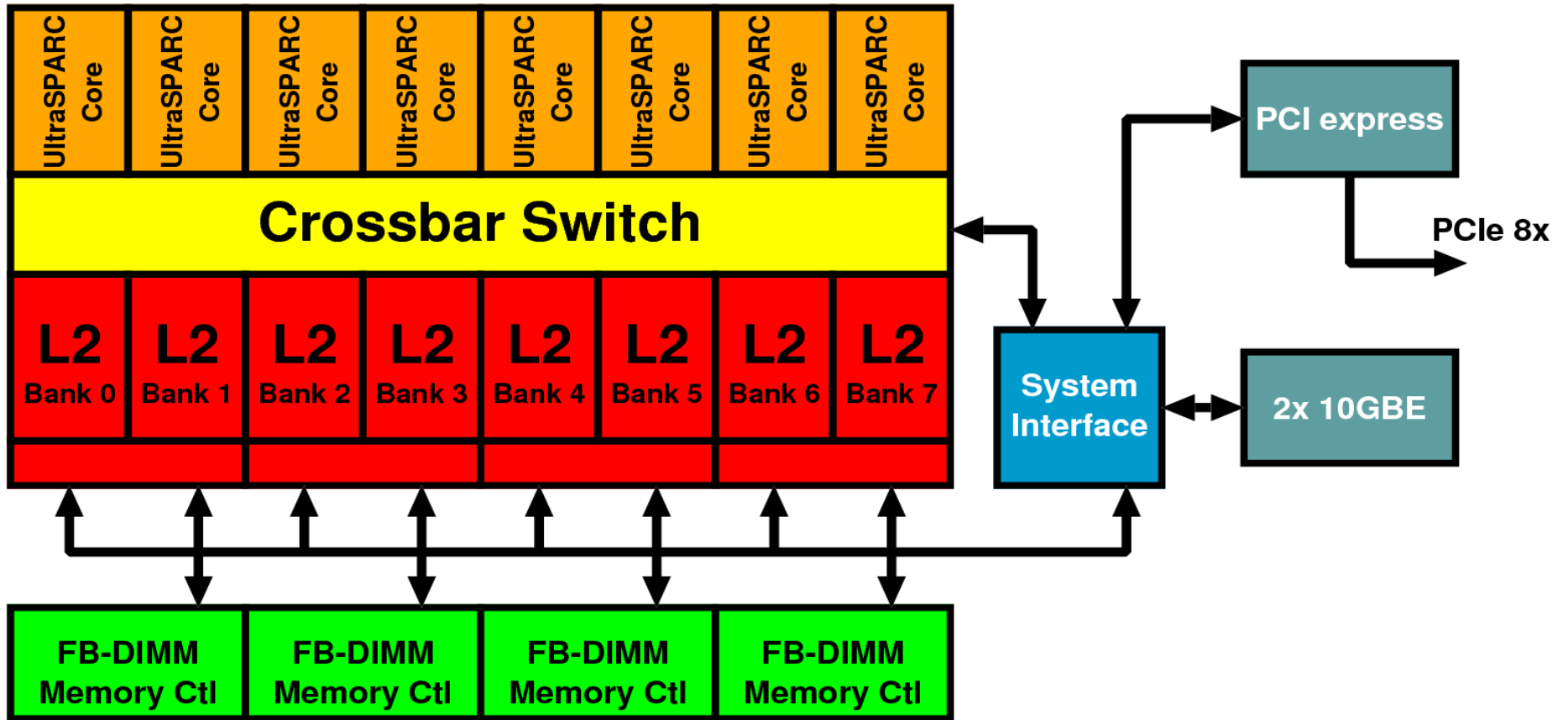
- Hypervisor: Pacifica
 - nicht kompatibel mit Intel Vanderpool (wg. MMU)
 - kommt mit Sockel F (= Rev. F)
 - Drop-In Replacement

- Vor Kurzem: Übernahme ATI durch AMD: Grafik-Subsystem auf dem Die? (SPU)



- Intel Woodcrest + Intel Clovertown
(Dualcore / Quadcore)
- AMD Opteron + AMD Barcelona
(Dualcore / Quadcore)
- Sun UltraSPARC T1 und T2 (Niagara-2)
- Sun UltraSPARC IV+, Fujitsu SPARC64 VI und VI+
- Benchmark-Abenteuer

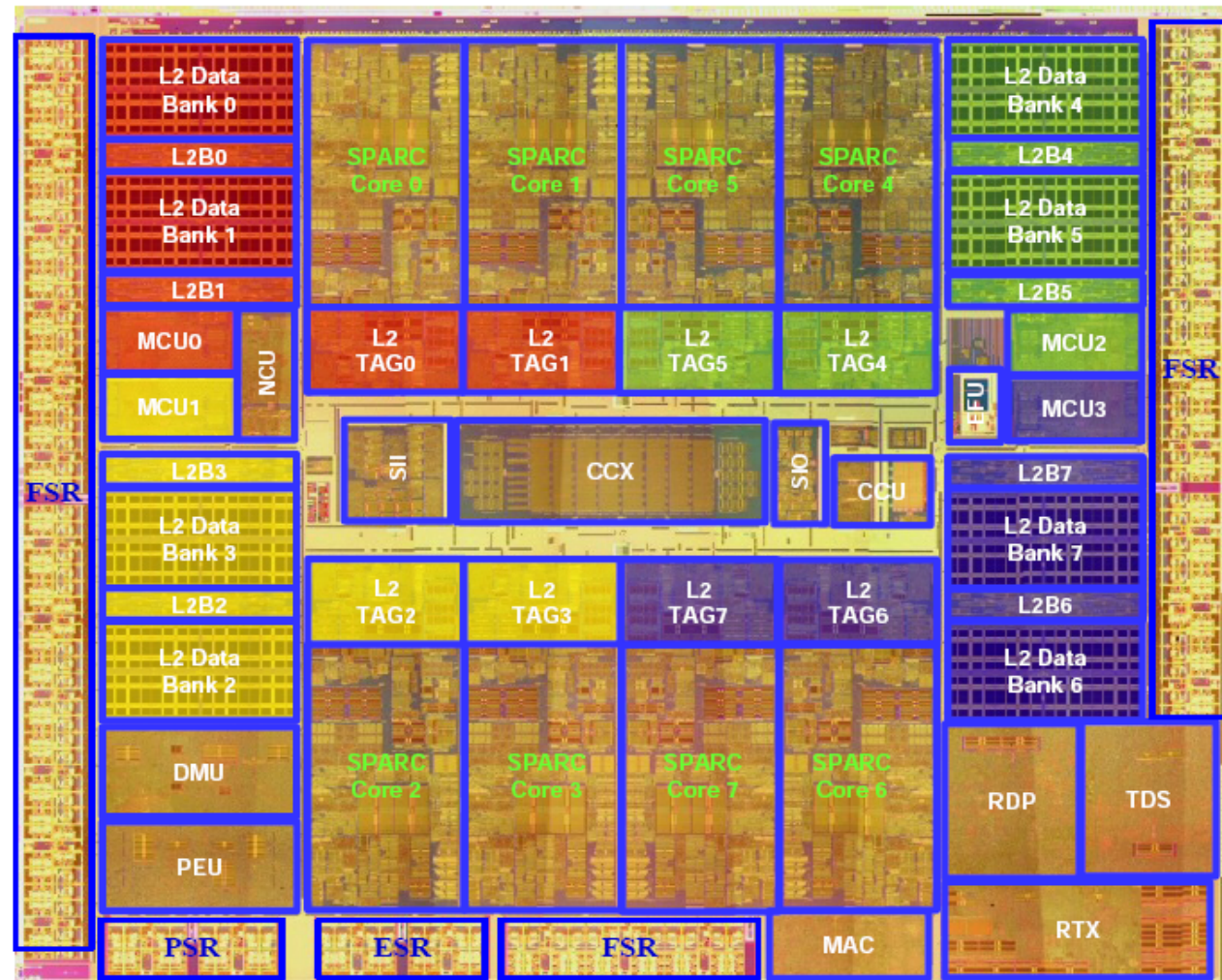
Sun Niagara-2 (UltraSPARC T2?)



- 65nm, vorauss. 1.4GHz Core-Takt
- 8 Cores mit je 8 Strands \Rightarrow 64 „CPUs“ on die!
Verdoppelung der Strands ist flächeneffizienter als Verdoppelung der Cores.
- Crossbar on Chip (SPU)

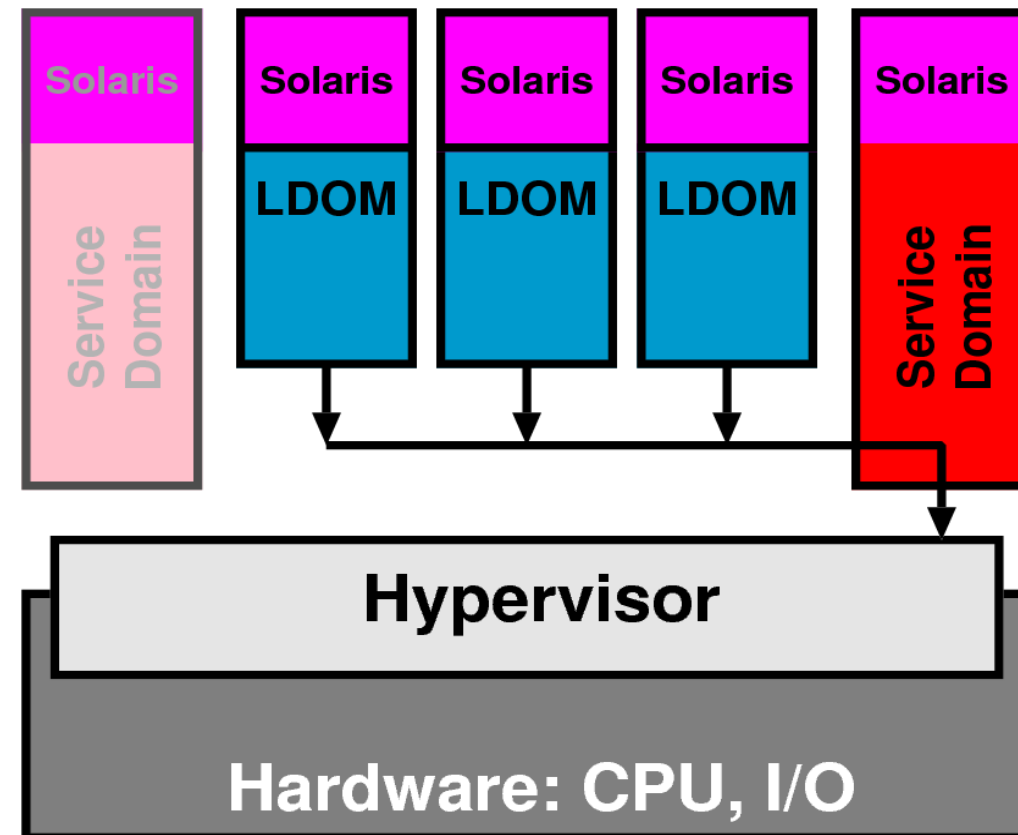
Sun Niagara-2 (UltraSPARC T2?)

- PCIe 8x + 2x 10GBE on Chip
- Bandbreite der Datenpfade noch nicht öffentlich
- je Core 1 FPU
- je Core 1 Crypto-Unit: RSA, Polynomial Elliptic Curve, DES/3DES, AES128, AES192, AES256, SHA-1, SHA-256, RC4, MD5
- Design-Ziel: 2x Throughput T1
- Design-Ziel: Crypto in Wirespeed auf beiden 10GBE
- Integer Pipeline: 8 Stages
- Float Pipeline: 12 Stages (divide/sqrt ist länger)
- DMA Engine teilt sich den Crossbar-Port mit dem jeweiligen Core



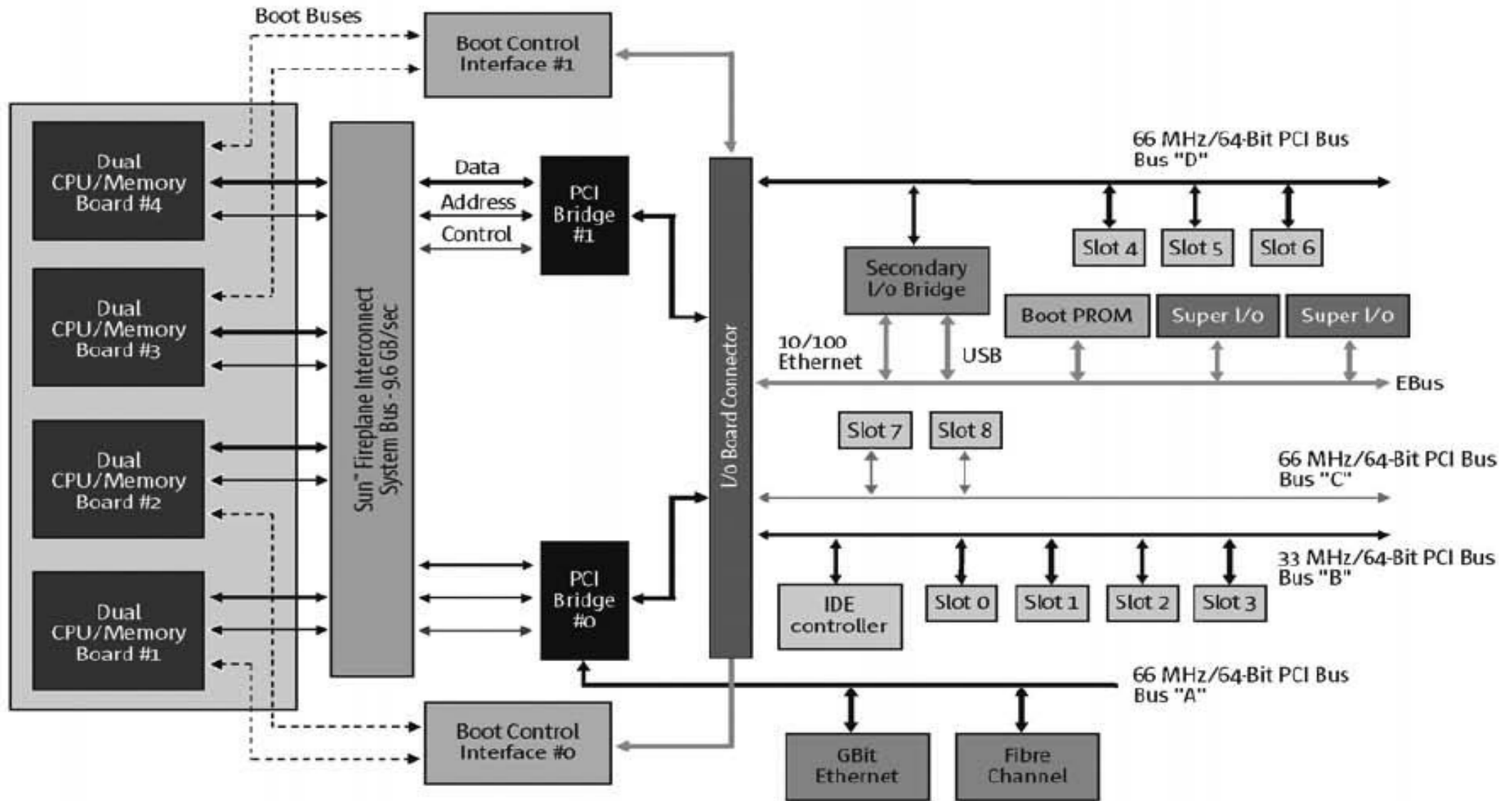
Sun Niagara-2 (UltraSPARC T2?)

- Parity Protection: I\$, D\$ Tags, D\$ Data, ITLB, DTLB, Arithmetic Memory, Store Buffer Address
- „Server-on-a-Chip“: alle wichtigen Funktionen auf dem Chip enthalten
- verbesserte Stromsparfunktionen
- eingebauter Hypervisor (seit Niagara-1), Logical Domains (LDMs) wird vorauss. mit Solaris 10 U4 (ca. Anfang 2008) unterstützt
- CPU ist OpenSource:
<http://www.opensparc.net/>
 - Verilog RTL Sources
 - Dokumentation
 - Tools / Application Stack (Apache, PHP, MySQL etc.)
 - GCC for SPARC Systems
 - bereits seit UltraSPARC T1 (= Niagara)



- **Intel Woodcrest + Intel Clovertown
(Dualcore / Quadcore)**
- **AMD Opteron + AMD Barcelona
(Dualcore / Quadcore)**
- **Sun UltraSPARC T1 und T2 (Niagara-2)**
- **Sun UltraSPARC IV+, Fujitsu SPARC64 VI und VI+**
- **Benchmark-Abenteuer**

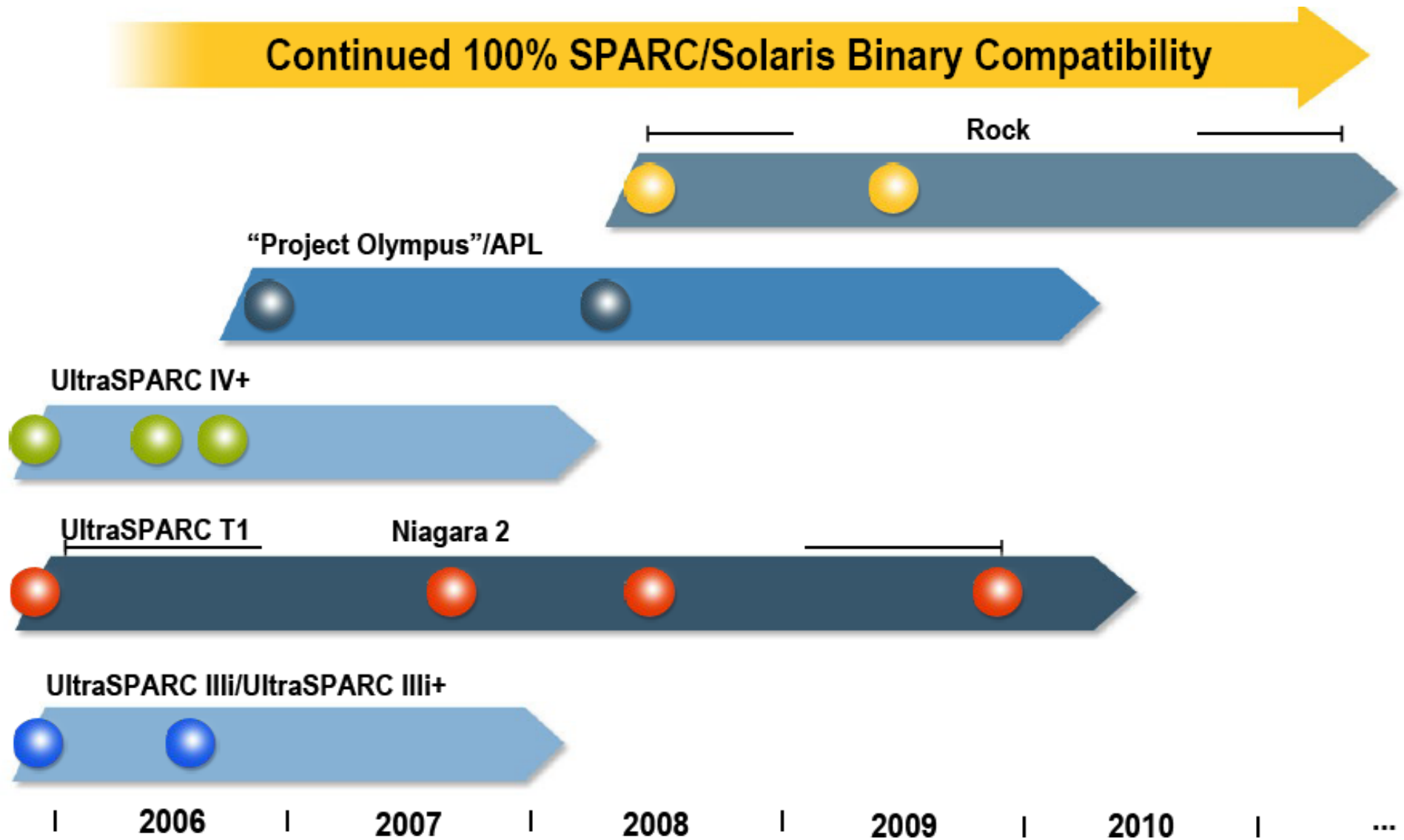
- 2x UltraSPARC III Core auf einem Die
- 90nm Prozess (Texas Instruments)
- bis 1.8GHz, ca. 90W
- 2MB L2 Cache on Chip
- 32MB L3 Cache, Tags on Chip, Data off Chip
- Mixed CPU-Betrieb möglich, je nach System mit UltraSPARC III, IV und IV+
- NUMA-Architektur, Crossbar-Switch auf Backplane (SunFire Interconnect)



- 2 Cores + 2 Strands je Core
- 90nm Prozess
- 2.4GHz @ 120W
- 6MB L2 Cache on Chip
(ggf. wesentlicher Performance-Vorteil)
- verbesserter SPARC64 V, keine wesentlichen
Änderungen (Modellpflege)

- 4 Cores + 2 Strands je Core
- 65nm Prozess
- > 2.7GHz
- wahrscheinlich > 6MB L2 Cache on Chip
- wird sehr wahrscheinlich in Sun APL verbaut werden (Advanced Product Line), ≈ Anfang/Mitte 2007

Sun SPARC Roadmap (nicht brandaktuell)



Note: Roadmap reflects estimated system delivery dates

- **Intel Woodcrest + Intel Clovertown
(Dualcore / Quadcore)**
- **AMD Opteron + AMD Barcelona
(Dualcore / Quadcore)**
- **Sun UltraSPARC T1 und T2 (Niagara-2)**
- **Sun UltraSPARC IV+, Fujitsu SPARC64 VI und
VI+**
- **Benchmark-Abenteuer**

- Benchmarks are Voodoo:
 - Understand, **what** to benchmark!
 - Understand, **how** to benchmark!
 - **Never** trust the numbers alone!
- SPECint und SPECfp, 2000 und 2006:
 - Kombination CPU / Cache / RAM
 - kein Realworld-Benchmark
 - Suite passt sich alle paar Jahre an die CPU-/Systementwicklung an, Ergebnisse nur schwer oder gar nicht vergleichbar
- TPC-C und TPC-D
 - I/O-Subsystem, lassen sich durch geschickte Plattenkombination stark beeinflussen!
 - kein Realworld-Benchmark
 - veraltet, nicht adäquat für aktuelle Systeme

Frage: Welche Metrik? Wie werden Systeme vergleichbar?

Ansatz: Performance / Watt

- Leistungsaufnahme ~ erforderlichen Kühlleistung
- Leistungsdichte ist in RZs maßgebliche Größe: Kühlleistung je Quadratmeter
- weniger Leistungsaufnahme \Rightarrow mehr Server auf gleichem Raum

- aktueller „Spitzenreiter“ ist Niagara bzw. Niagara-2:
 - ca. 60W – 70W bei 100% Load
(in Summe auf **allen** Cores mit **allen** Strands)
- AMD Opteron und Intel Woodcrest ca. gleich für CPU:
 - ca. 90W – 120W bei 100% Load
(lt. Datenblatt)

Aber:

- auch RAM braucht Strom!
- FBDIMM deutlich mehr als DDR, dafür auch deutlich schneller
- Auch Peripherie braucht Strom (Bootplatten, NICs etc.)

Forderung: Bei Benchmarks auch genau angeben, welche Konfiguration gemessen wurde!



Beispiel:

	Woodcrest 2-Way			T2000	T1000
RAM	8GB	16GB	32GB	32GB	16GB
Disk	1x150GB SATA 7k2	1x73GB SAS 15k	2x SATA 7k2	4x73GB SAS 10k	??
Leistungsaufnahme	330W	430W	510W	330W	185W

⇒ **1 Rack mit 32GB Woodcrest:**

≈ 510W x 20 Server = 10.2kW / Rack

⇒ ca. 20kW Gesamtleistung (incl. Cooling)

Skalierung über Taktfrequenz:

- SPEC CPU2000 Benchmark (<http://www.spec.org/>)
- Fujitsu Siemens Celsius R540

CPU / GHz	Config	SPECfp_rate 2000
Intel 5160 / 3.0	4 Core / 2 Socket	80.6
Intel 5160 / 2.6	4 Core / 2 Socket	77.4
Intel 5160 / 2.0	4 Core / 2 Socket	68.4

- 2GHz → 3GHz $\hat{=}$ 50% Steigerung der Taktrate
- Steigerung SPECfp_rate bei 17.8%
- Memory Latency @ FSB?

Skalierung über Anzahl CPUs:

- Fluent 6.2 Benchmark (<http://www.fluent.com/>)
- Intel S5000 XAL, 3.0GHz Xeon Woodcrest 5160

	FL5L1 (scaling)	FL5L2 (scaling)
4 Core / 2 Socket	631.8 (3.3)	400.0 (3.0)
2 Core / 1 Socket	372.8 (1.9)	226.0 (1.7)
1 Core / 1 Socket	194.0 (1.0)	133.0 (1.0)

- mit 4 Cores nur 3.0- bis 3.3-fache Leistung
- Memory Latency @ FSB?
- Sun insgesamt etwas schlechter, skaliert aber nahezu linear, insgesamt wenig Streuung bei allen gemessenen Systemem

Opteron vs. Woodcrest

- Intel S5000 XAL, 3.0GHz Xeon Woodcrest 5160
- Sun X4100 M2, 2.8GHz Opteron DC 2200
- Fluent 6.2 Benchmark (<http://www.fluent.com/>)

	# Cores	FL5M3 (scaling)	FL5L2 (scaling)
Intel	4 Core	827.0 (2.8)	400.0 (2.9)
	2 Core	553.7 (1.9)	226.0 (1.6)
	1 Core	297.3 (1.0)	138.0 (1.0)
Sun/AMD	4 Core	979.9 (3.6)	486.6 (4.1)
	2 Core	516.1 (1.9)	241.8 (2.1)
	1 Core	273.5 (1.0)	117.6 (1.0)

Server haben doch Stromsparmodi?

■ **Richtig!** Aber:

■ Server 1: 400W @ 100%, spart 20W je 10% weniger Auslastung

■ Server 2: 300W @ 100%, spart 20W je 10% weniger Auslastung

Utilization		100%	90%	80%	70%	60%	50%	40%	30%	20%	10%
Server 1	Watts @ Utilization	400	380	360	340	320	300	280	260	240	220
	Watts / Work	400	422	450	486	533	600	700	867	1200	2200
Server 2	Watts @ Utilization	300	280	260	240	220	200	180	160	140	120
	Watts / Work	300	311	326	343	367	400	450	533	700	1200

Und was sagt uns das jetzt?

Annahme 1: 5 Server @ 10% Load

⇒ $5 \times 220\text{W} = 1100\text{W}$

Aber! 1 single server mit $5 \times 10\% = 50\%$ Load

⇒ 300W

Annahme 2: 4 Server @ 20% Load

⇒ $5 \times 140\text{W} = 560\text{W}$

Aber! 1 single server mit $4 \times 20\% = 80\%$ Load

⇒ 260W

Ziel: vorhandene Ressourcen besser ausnutzen!

- *Greg Grohoski*
Niagara-2: A highly threaded Server-on-a-Chip
Sun Microsystems, August 2006

- *David Kanter*
Niagara II: The Hydra Returns
<http://www.realworldtech.com/>

- *Nathan Brookwood*
The role of intelligent design in the evolution of multiple processors
Insight 64, 2006

- *Jason Clark, Ross Whitehead*
Intel Woodcrest: the birth of a new king
<http://www.anandtech.com/printarticle.aspx?i=2793>

- **Weblog: BM Seer**
<http://blogs.sun.com/bmseer>

■ CoolTools

<http://cooltools.sunsource.net/>

Compiler, Binary Tools, Testwerkzeuge für UltraSPARC-Systeme

■ OpenSparc

<http://www.opensparc.net/>

Dokumentation, Foren, Sourcen zu Sun UltraSPARC T1

Danke für die Aufmerksamkeit.

Fragen?

**best OpenSystems Day
Herbst 2006**

Unterführung

**Wolfgang Stief
stief@best.de**

Senior Systemingenieur
best Systeme GmbH
GUUG Board Member

