Paralleles Höchstleistungsrechnen

—

Parallele Rechnerarchitektur III

Stefan Lang

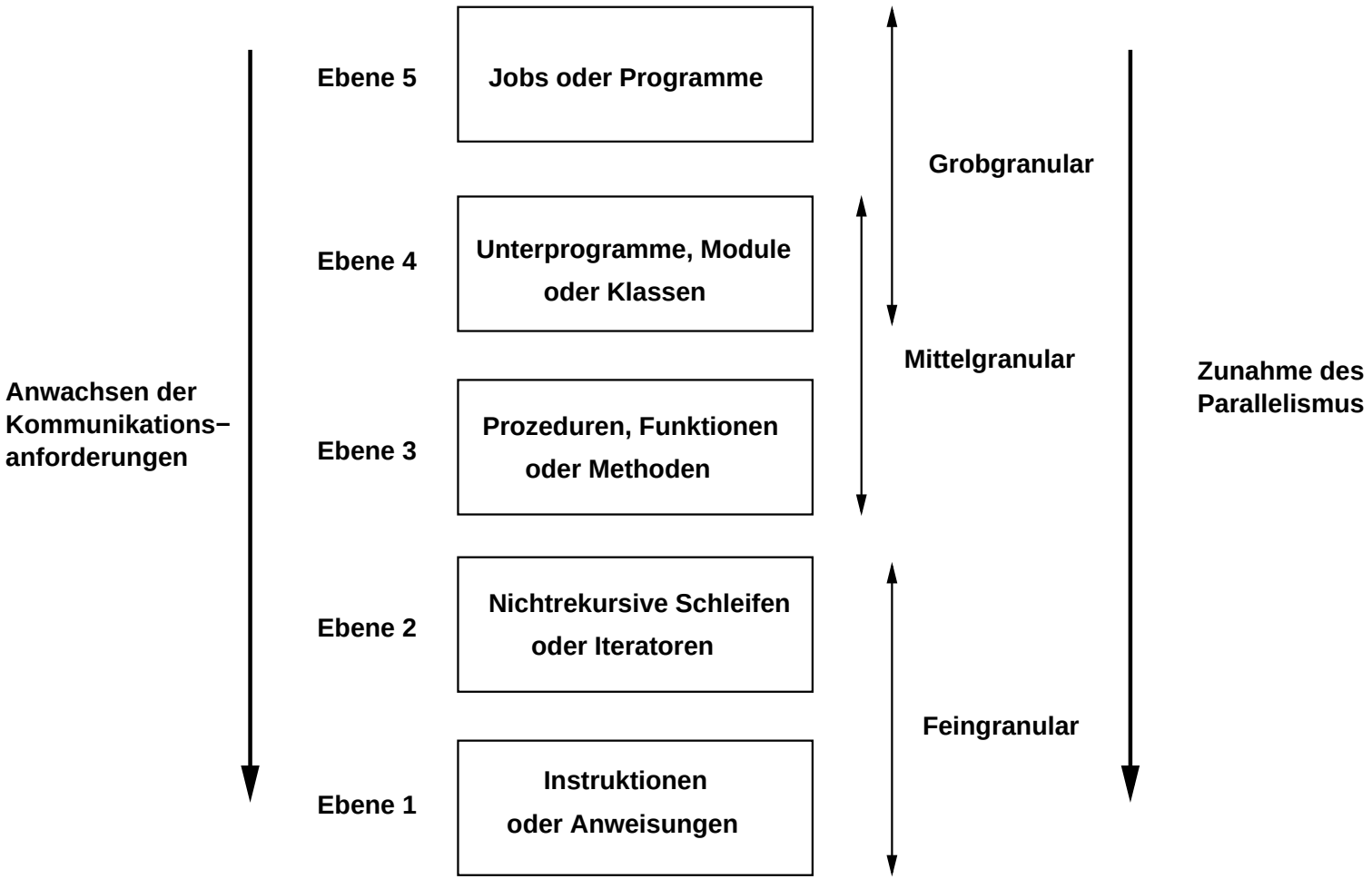
Interdisziplinäres Zentrum für Wissenschaftliches Rechnen
Universität Heidelberg
INF 368, Raum 425
D-69120 Heidelberg
phone: 06221/54-8264
email: Stefan.Lang@iwr.uni-heidelberg.de

27. Oktober 2009

Parallele Rechnerarchitektur III

- Parallelität und Granularität
- Detailstudie Hypertransport

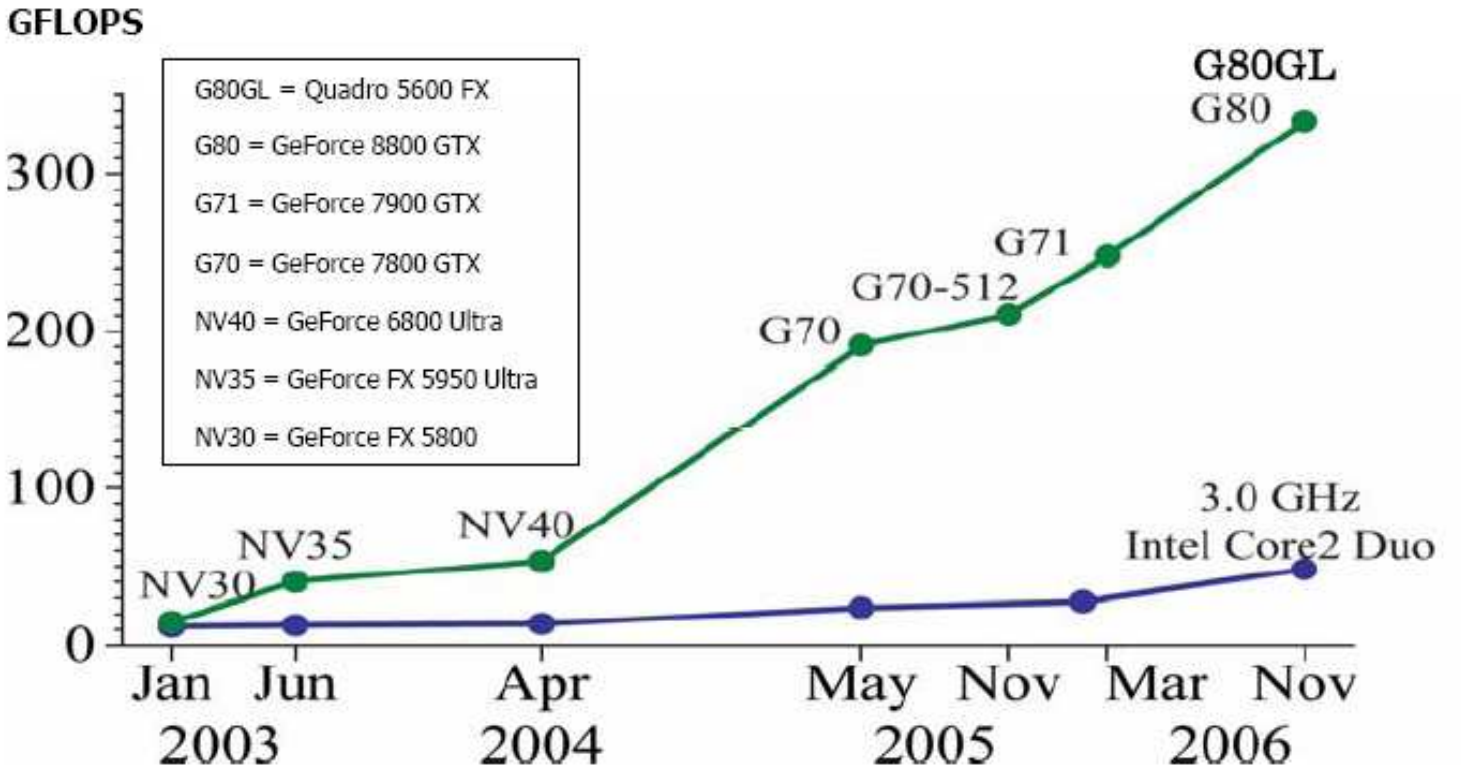
Parallelität und Granularität



Graphikkarten

- GPU = Graphics Processing Unit
- CUDA = Compute Unified Device Architecture
  - Tookit von NVIDIA zur direkten GPU Programmierung
  - Programmierung einer GPU ohne graphische API
  - GPGPU 0
- weitaus höhere Rechenleistung, Speicherbandbreite als CPU
- GPUs sind billig und breit etabliert

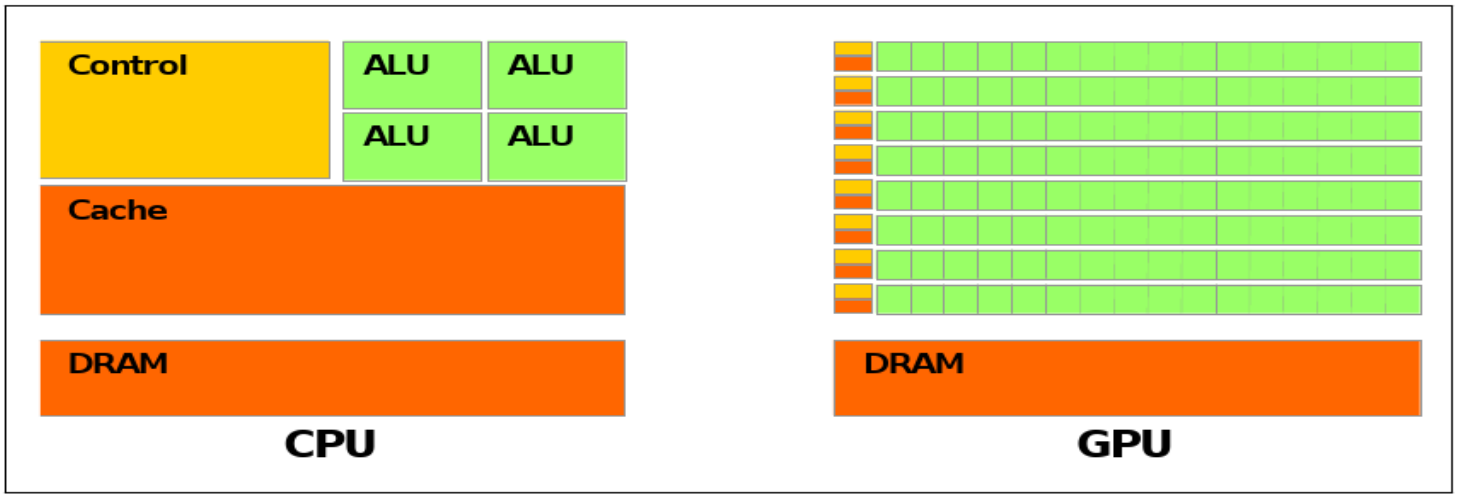
Leistungsentwicklung: CPU vs. GPU



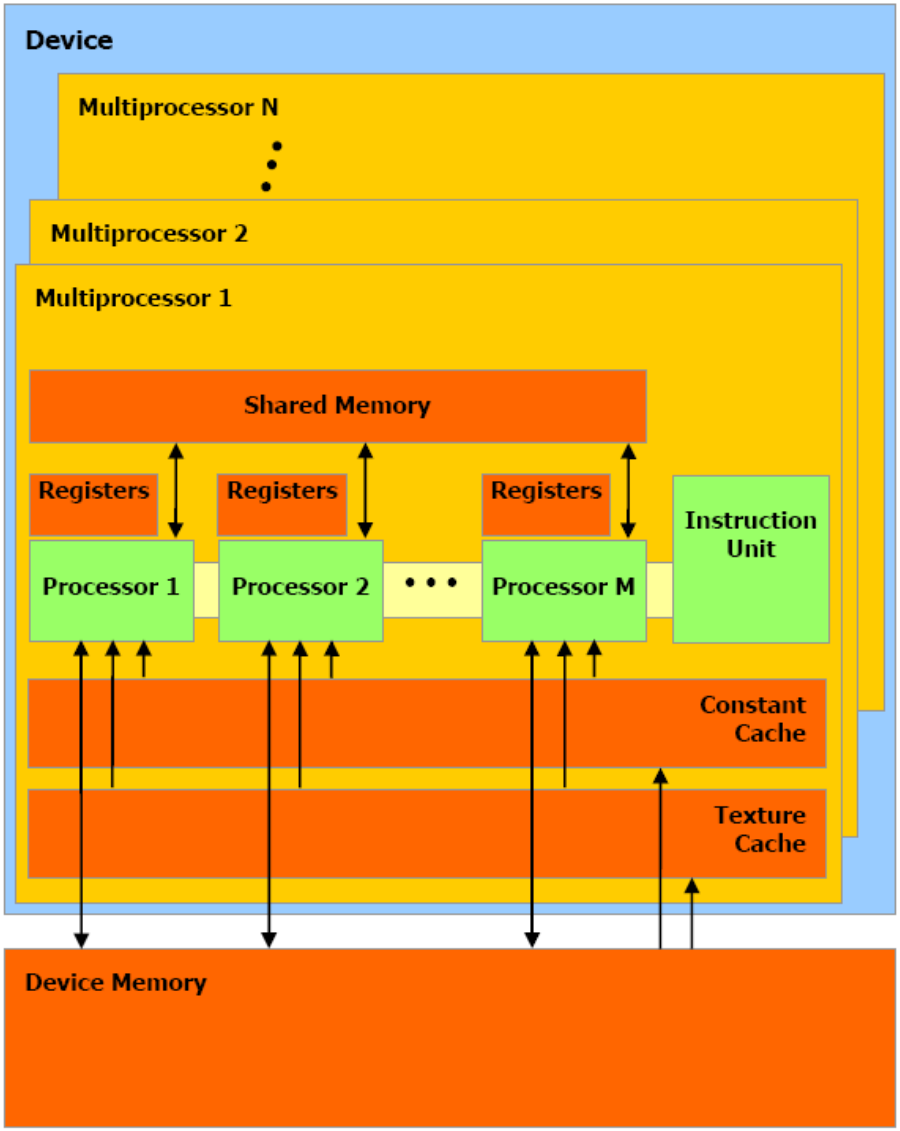
Graphikkarte: Hardwaredaten

GeForce GTX 285	
Fab (nm)	55
Transistors (million)	1400
Memory (MB)	1024
Multiprocessors	30
Streaming Processors	240
Shader Clock (MHz)	1476
Memory Bandwidth (GB/s)	159
Memory Bus width (bit)	512
SP GFLOPs (MADD + MUL)	1063
DP GFLOPs (MADD)	89
TDP (Watt)	183
Price (EUR)	~350

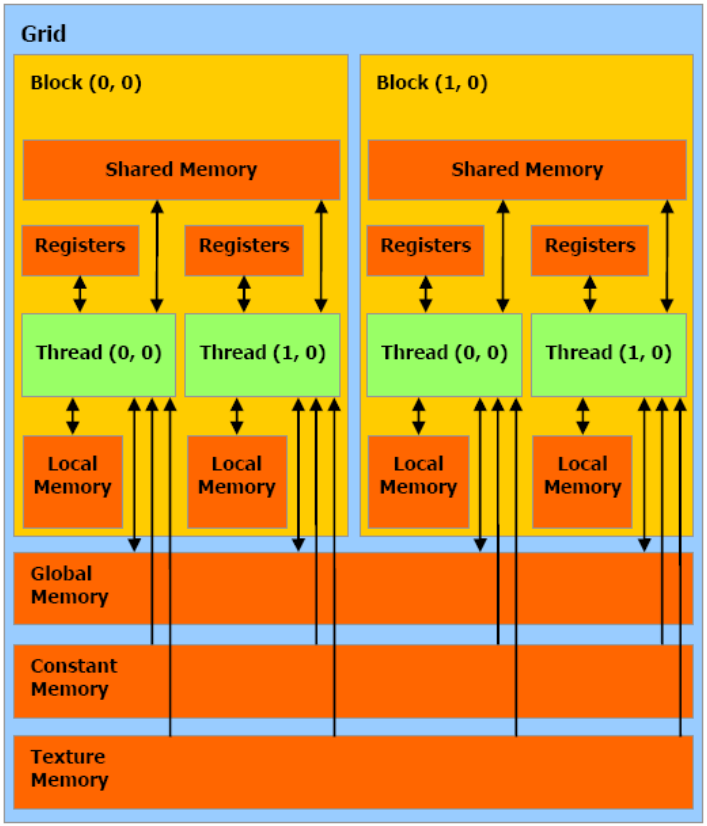
Chiparchitektur: CPU vs. GPU



Graphikkarte: Hardwareaufbau

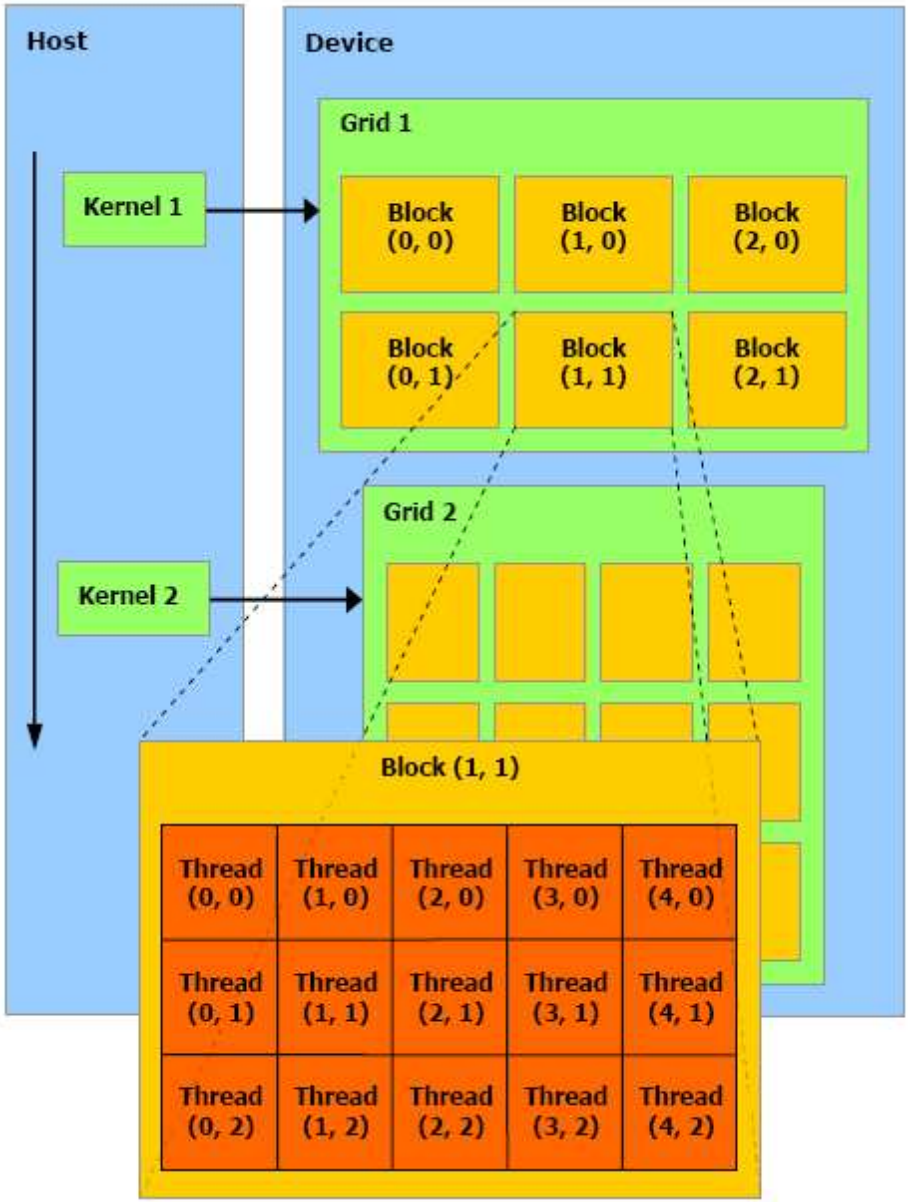


Graphikkarte: Speicheraufbau

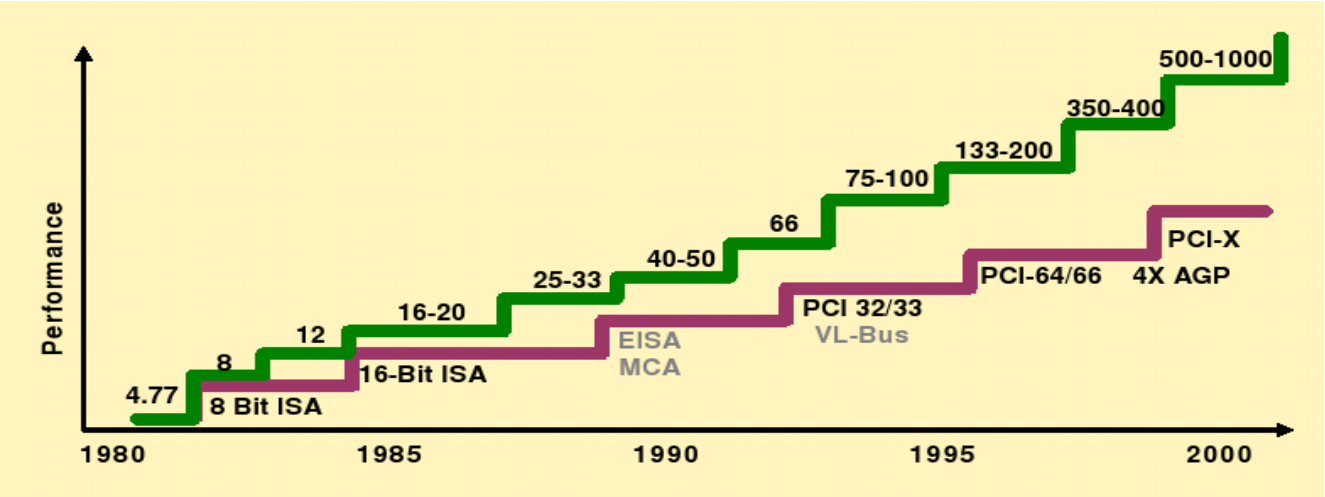


- 8192 Register (32-bit), ingesamt 32KB pro Multiprozessor
- 16KB schnelles shared memory pro Multiprozessor
- grosses globales memory (Hunderte von MB, e.g. 512MB-4GB)
- Globales memory ungecacht, Latenzzeit 400-600 Taktzyklen
- Lokales memory ist eigentlich Teil vom globalen Speicher
- Read-only konstanter Speicher
- Read-only Textspeicher
- Register und shared memory werden zwischen Blöcken, welche auf einem Multiprozessor ausgeführt werden, aufgeteilt
- Globaler Speicher, konstanter Speicher und Texturspeicher sind von der CPU zugreifbar

Graphikkarte: Programmiermodell



Leistungsentwicklung im Bereich des I/O

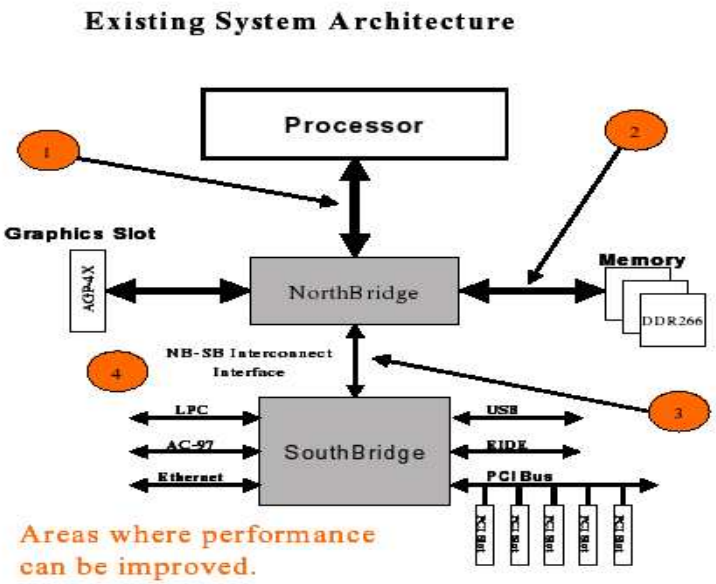


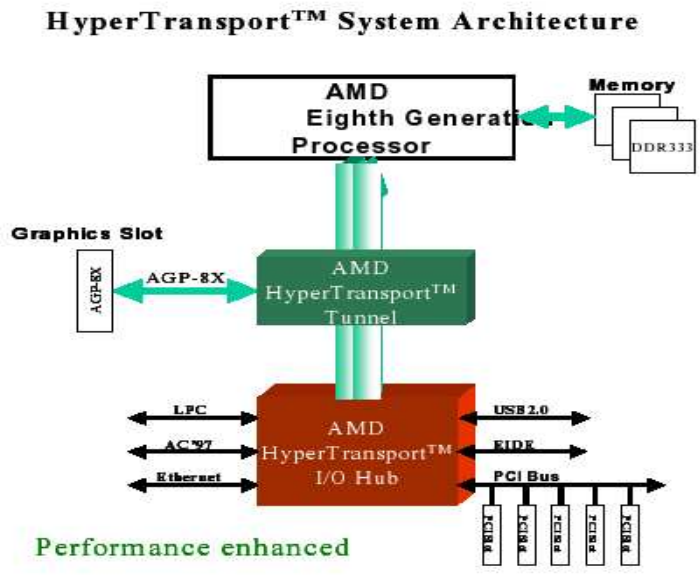
- Mikroprozessor Leistung verdoppelt sich in ca. 18 Monaten
- Leistung der I/O Architektur verdoppelt sich in ca. 36. Monaten
- Server und Workstations benötigen verschiedene Hochgeschwindigkeitsbusse (PCI-X, PCI-E, AGP)

→ hohe Komplexität und unnötige Vielfalt bei eingeschränkter Leistung

- Anforderungen an Bandbreite wachsen: Hochauflösende 3D-Graphik und Video (CPU – Graphikprozessor), Interprozessor (CPU – CPU), Hochleistungsnetzwerke wie Gigabit Ethernet und Infiniband (CPU – I/O)

Engpäße im I/O-Bereich





Mögliche Engpäße sind:

- Front-Side Bus
- Speicherinterface
- Chip-to-Chip Verbindung
- I/O zu anderen Bussystemen

Standardisierung von I/O Anforderungen

Entwicklung der Hypertransport (HT) I/O Verbindungsarchitektur (seit 1997)

- HyperTransport ist In-The-Box Lösung (Intraconnect Technologie)
- komplementäre Technik zu Netzwerkprotokollen wie Infiniband und 10 Gb/ 100 Gb Ethernet als Box-to-box Lösungen (Interconnect Technologie)
- HT ist offener Industriestandard, kein Produkt (keine Lizenzgebühren)
- Weiterentwicklung und Standardisierung durch Konsortium von Industriepartnern, z.B. AMD, Nvidia, IBM, Apple
- früherer Kodename: Lightning Data Transfer (LDT)

Funktionsüberblick

Feature/Function	HyperTransport Technology
Bus Type	Dual, unidirectional, point-to-point links
Link Width	2, 4, 8, 16, or 32 bits
Protocol	Packet-based, with all packets multiples of four bytes (32 bits). Packet types include Request, Response, and Broadcast, any of which can include commands, addresses, or data.
Bandwidth (Each Direction)	100 to 6400 Mbytes/s
Data Signaling Speeds	400 MHz to 1.6 GHz
Operating Frequencies	400, 600, 800, 1000, 1200, and 1600 Megatransfers/second
Duplex	Full
Max Packet Payload or Burst Length	64-byte packet
Power Management	ACPI-compatible
Signaling	1.2-V Low-Voltage Differential Signaling (LVDS) with a 100-ohm differential impedance
Multiprocessing Support	Yes
Environment	Inside the box
Memory model	Coherent and noncoherent

Designziele bei der Gestaltung von HyperTransport

- Verbesserung der Systemleistung
  - höhere I/O Bandbreite (high bandwidth)
  - Vermeidung von Flaschenhälsen durch langsame Geräte in kritischen Informationspfaden
  - geringere Anzahl von Systembussen
  - niedrige Antwortzeiten (low latency)
  - reduzierter Energieverbrauch
- Vereinfachung des Systemaufbaus
  - Einheitliches Protokoll für In-box Verbindungen
  - Nutzung kleiner Pinzahlen um hohe Packungsdichte und niedrige Kosten sicherzustellen
- Höhere I/O-Flexibilität
  - Modulare Brückenarchitektur
  - Unterschiedliche Bandbreiten in Upstream/Downstream Richtung
- Kompatibilität mit bestehendem Systemen
  - Ergänzung zu standardisierten, externen Bussen
  - geringe Auswirkungen auf bestehende Betriebssysteme und Treiber
- Erweiterbarkeit zu Systemnetzwerk Architekturen (SNA Busse)
- Hohe Skalierbarkeit in Mehrprozessorsystemen

Flexible I/O Architektur

Hypertransport Architektur ist in 5 Schichten unterteilt
Struktur orientiert sich am **Open-Sytem-Interconnection** (OSI) Referenzmodell

- **Bitübertragungsschicht:** Physikalische und elektrische Eigenschaften von Daten, Kontroll, Taktleitungen
- **Datenverbindungsschicht:** Initialisierung und Konfiguration von Verbindungen, periodisch zyklische Redundanz (CRC), Trennung/Wiederverbindung, Paketierung Kontrollfluß und Fehlermanagement,
- **Protokollschicht:** Virtuelle Kanäle und Kommandos
- **Transaktionsschicht:** Schreibe- und Leseaktionen unter Nutzung der Datenverbindungsschicht
- **Sitzungsschicht:** Power-, Interrupt- und Systemmanagement

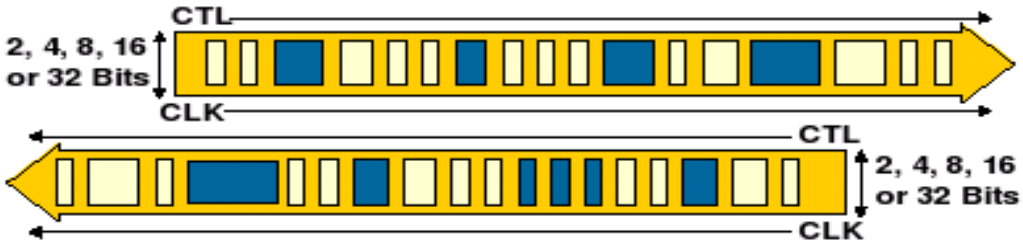
Geräteklassen und -konfigurationen

3 Geräteklassen werden hinsichtlich Funktion und Lage innerhalb der HT-Kette unterschieden: Cave, Tunnel und Bridge

- HT Bridge: Vermittler zwischen primärer Seite (CPU bzw. Speicher) und sekundärer Seite (HT-Geräten) einer Kette
- HT Tunnel: besitzt zwei Seiten mit jeweils einer Empfangs- und einer Sendeeinheit, z.B. Netzwerkkarte oder Bridge zu weiteren Protokoll
- HT Cave: markiert Ende der Kette und besitzt nur eine Kommunikationsseite. Durch die Verschaltung von mindestens einer HT Bridge und einem HT Cave kann eine einfache HT-Kette aufgebaut werden.

Bitübertragungsschicht I

Aufbau einer HT-Verbindung



- zwei unidirektionalen Punkt-zu-Punkt Datenpfade
- Datenpfadbreiten: 2,4 8 und 16 Bit je nach Geräteanforderungen
- Kommandos, Adressen und Daten (CAD) verwenden die selben Signalleitungen
  $\longrightarrow$ geringere Leitungsanzahl, kleinere Pakete, geringerer Energieverbrauch, bessere thermische Eigenschaften
- Pakete enthalten CAD und sind Vielfache von 4 Byte (32 bit)
- Verbindungen mit weniger als 32 bit verwenden aufeinanderfolgende Takte um Pakete vollständig zu übertragen

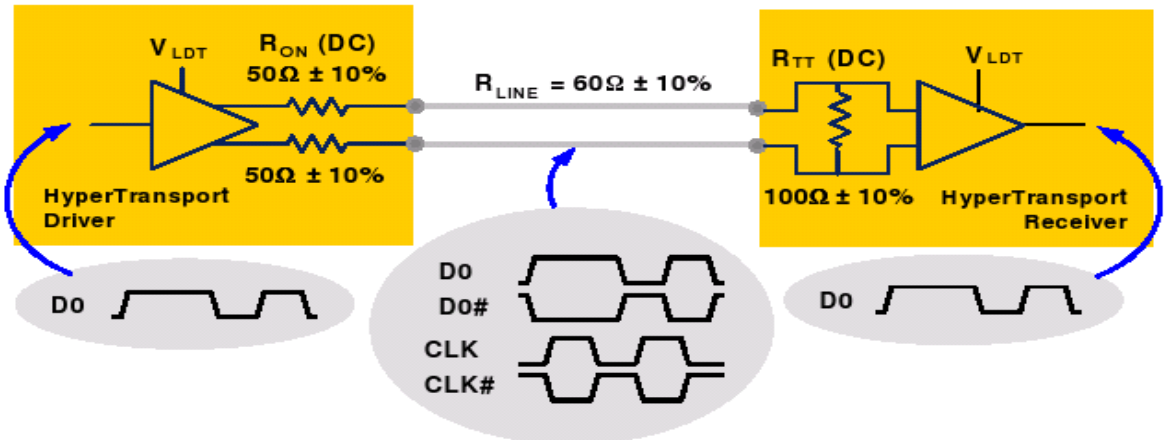
Hohe Performanz und Skalierbarkeit

- hohe Datenraten aufgrund von low voltage differential signaling (LVDS, 1,2V $\pm$ 5%)
- skalierbare Bandbreite mittels Skalierung von Übertragungsfrequenz und Linkbreite

Bitübertragungsschicht II

Hypertransport Low Voltage Differential Signaling

- Differenzielle Signalübertragung über zwei Leitungen: Spannungsdifferenz ($\pm$ 0,3 V) entspricht Logikzustand Logikwechsel durch Umpolen der Leitungen (symmetrische Signalübertragung, double-ended)
- HT Signalübertragung entspricht erweitertem IEEE LVDS Standard (1,2 V statt 2,5 V)
- 60 cm maximale Leitungslänge bei 800 Mbit/s
- Übertrager sind in Controllchips integriert, 100 Ω Impedanz verhindert Reflexionen
- einfach realisierbar auf Standard 4-Schicht PCBs (Printed Circuit Boards) und zukunftsfähig



Bitübertragungsschicht III

Elektrische Signale

Signal Name	Description	Comment
CAD	**Commands, Addresses and Data:** Carries command, address, or data information.	CAD width can be different in each direction.
CTL	**Control:** Used to distinguish control packets from data packets.	
CLK	**Clock:** Forwarded clock signal.	Each byte of CAD has a separate clock signal. Data is transferred on each clock edge.
PWROK	**Power OK:** Power and clocks are stable.	Single-ended.
RESET#	**HyperTransport Technology Reset:** Resets the chain.	Single-ended.
LDTSTOP#	**HyperTransport Technology Stop:** Enables and disables links during system state transitions.	Used in systems requiring power management. Single-ended.
LDTREQ#	**HyperTransport Technology Request:** Requests re-enabling links for normal operation.	Used in systems requiring power management. Single-ended.

Für je 8 Bit Datenbreite gibt es eine Taktleitung, welche vom Sender zum Empfänger verläuft, mit der die Daten auf den 8 Datenleitungen beim Empfänger abgetastet werden (quell-synchrone Taktung)
→ Abweichungen vom Sendertakt werden minimiert

Bitübertragungsschicht IV

Bandbreitenskalierung

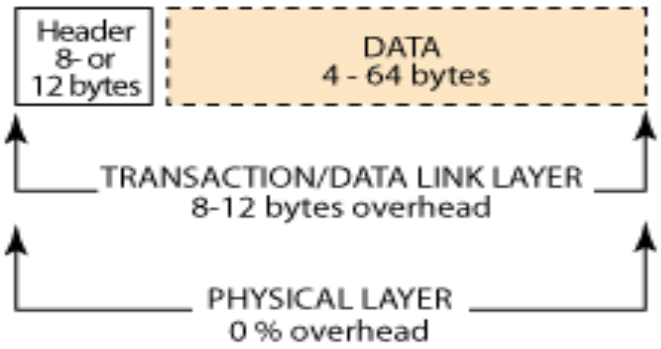
- Datenübertragung von CAD bei steigender und fallender Flanke des Taktsignals (DDR)
  $\rightarrow$ Verbindungstakt von 800 MHz entspricht 1600 MHz Datentakt
- Anzahl der Leitungen ermöglicht Anpassung an Bandbreitenerfordernisse
  $\rightarrow$ 2 x 16 CAD bits + 800 GHz clock = 2 x 3,2 GByte Bandbreite (103 Pins)
  $\rightarrow$ 2 x 2 CAD bits + 400 MHz clock = 2 x 200 MByte Bandbreite (24 Pins)

Link Width (Each Way)	2	4	8	16	32
Data Pins (total)	8	16	32	64	128
Clock Pins (total)	4	4	4	8	16
Control Pins (total)	4	4	4	4	4
Subtotal (High Speed)	**16**	**24**	**40**	**76**	**148**
V_{LOT}	2	2	3	6	10
GND	4	6	10	19	37
PWROK	1	1	1	1	1
RESET#	1	1	1	1	1
Total Pins	**24**	**34**	**55**	**103**	**197**

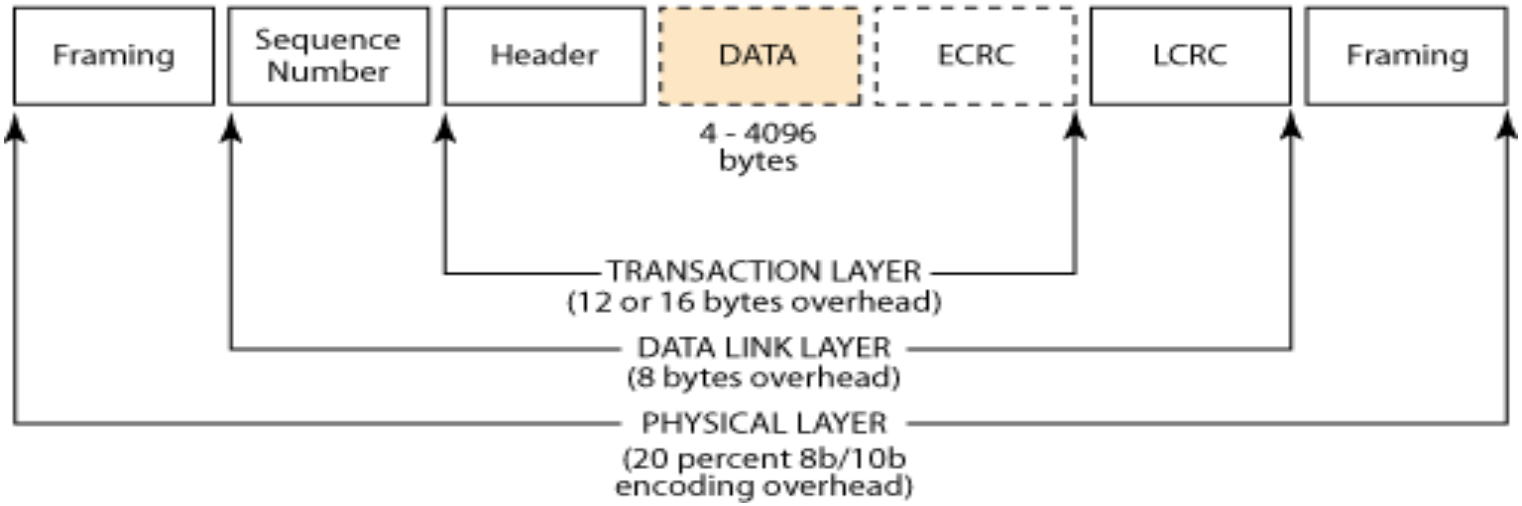
Verbindungsschicht

Packetaufbau von HT im Vergleich zu PCI-E

HyperTransport Packet Format



PCI Express Packet Format



Protokoll- und Transaktionsschicht

- Protokollschicht: Kommandos, virtuelle Kanäle und Flußkontrolle
- Transaktionsschicht: Durchführung von Aktionen wie Leseanforderungen und Antworten

Kommandoübersicht

Virtual Channel	Command	Comment
Posted	Posted Write	Followed by data packet(s).
	Broadcast	Issued by host bridge downstream to communicate information to all devices.
	Fence	All posted requests in a stream cannot pass it.
Non-Posted	Non-Posted Write	
	Read	Designates whether response can pass posted requests or not.
	Flush	Forces all posted requests to complete.
	Atomic Read-Modify-Write	Generated by I/O devices or bridges and directed to system memory controlled by the host.
Responses	Read Response	Response to read command, is followed by data packet(s).
	Target Done	A transaction not requiring returned data has completed at its target.

Paketstruktur

Bit-Time	7	6	5	4	3	2	1	0
0	SeqID[3:2]		Cmd[5:0]					
1	PassPW	SeqID[1:0]	UnitID[4:0]					
2	*Command-Specific*							
3	*Command-Specific*							
4	Addr[15:8]							
5	Addr[23:16]							
6	Addr[31:24]							
7	Addr[39:32]							

Paketweiterleitung

Weiterleitungsmethoden der HT-3.0 Spezifikation

- Store-and-Forward Routing: Ein Paket wird vollständig empfangen und zwischengespeichert, die Checksumme (CRC) berechnet und mit der im Paket verglichen, der Paketheader wird zur Empfängerermittlung dekodiert, dann wird das Paket entweder selbst verarbeitet oder über die Verbindung in Richtung des Empfängers weitergeleitet (1 hop)
- Cut-Through Routing: Weiterleitung des Pakets sobald er Paketheader empfangen und dekodiert ist, das Paket wird ohne Zwischenspeicherung durchgeleitet
  Problem: weitergeleitete Pakete mit CRC-Fehler Lösung: Abbruch der Weiterleitung, Empfänger ermittelt Paketfehler und verwirft das Paket
- Spekulative Weiterleitung: Spekulation auf Korrektheit und Empfängerport, nur noch Feststellen ob korrektes Kommando im Paket vorliegt (1 Byte)
- Spekulative Weiterleitung mit aggr. Implementierung: Paket wird sofort, also mit dem ersten Leitungstakt und ohne irgendeine Dekodierung weitergeleitet

Alle Methoden können ohne Veränderung der Verbindungsspezifikation implementiert werden!

Paketweiterleitung

Paket mit n bytes (inkl. CRC), Linkbreite w bit, Pfadlänge d hops
t_{wire} Übertragungszeit, t_{proc} Zeit für CRC-Check + Empfängerermittlung
h zu empfangende Bytes bis zum Forwarding
Latenzzeit der verschiedenen Weiterleitungsmethoden

- Store-and-Forward Switching

$$L = d \cdot (n \cdot 8/w + t_{wire} + t_{proc})$$

- Cut-Through Switching

$$L = d \cdot (max(1, 8 * h/w) + t_{wire} + t_{proc}) + 8/w \cdot (n - max(h, w/8))$$

- Spekulative Weiterleitung

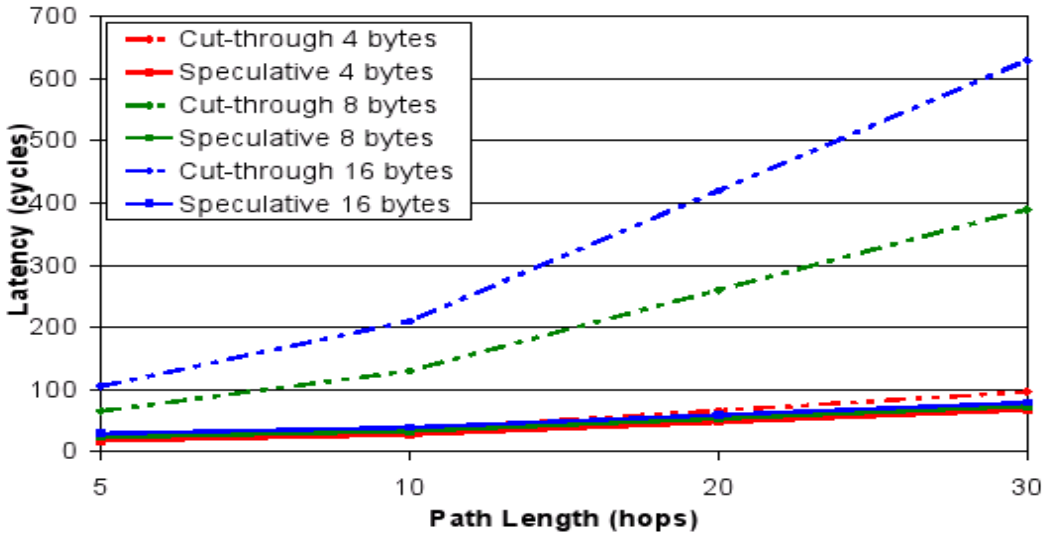
$$L = d \cdot (max(1, 8/w) + t_{wire} + t_{proc}) + 8/w \cdot (n - max(1, w/8))$$

- Spekulative Weiterleitung mit aggr. Implementierung

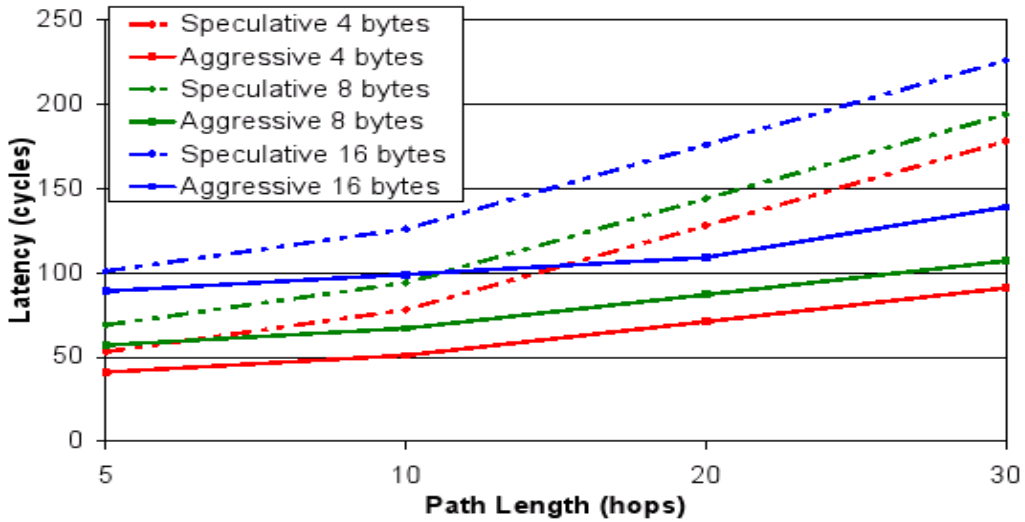
$$L = d \cdot (1 + t_{wire} + t_{proc}) + 8/w \cdot (n - 1)$$

Paketweiterleitung

Latenzzeit bei standard bzw. aggressiver spekulativer Weiterleitung



8-Bit Verbindungsbreite



Literatur

Literatur zu Paralleler Rechnerarchitektur

- Hennessy J., Patterson D.: Computer Architecture – A Quantitative Approach, 3. Ausgabe, Morgan Kaufmann, 2003
- Hennessy J., Patterson D.: Computer Architecture – A Quantitative Approach, 4. Ausgabe, Morgan Kaufmann, 2007
- Culler D., Singh J., Gupta A.: Parallel Computer Architecture – A Hardware/Software Approach, Morgan Kaufmann, 1999
- HyperTransport Konsortium: www.hypertransport.org
- Hwang K.: Advanced Computer Architecture: Parallelism, Scalability, Programmability, McGraw-Hill, 1993
- Grama A., Gupta A., Karypis G., Kumar V.: Introduction to Parallel Computing, 2. Ausgabe, Benjamin/Cummings, 2003