TECHNISCHE UNIVERSITÄT DRESDEN

Center for Information Services and High Performance Computing (ZIH)

# Theorie und Einsatz von Verbindungseinrichtungen in parallelen Rechnersystemen

*Leistungsbewertung von Verbindungsnetzwerken*

08. Juli 2011

Andy Georgi

INF 1046
Nöthnitzer Straße 46
01187 Dresden

0351 - 463 38783

ZIH
Center for Information Services & High Performance Computing

## Foliensatz

**Verfügbarkeit der Folien**

Vorlesungswebseite:
`http://tu-dresden.de/die_tu_dresden/zentrale_einrichtungen/zih/lehre/ss2011/tevpr`



## Agenda





## Agenda

# Latenzzeit I

- Kommunikationszeit allgemein:

$$T_{Comm} = T_{Latency} + T_{Transfer}$$

- Übertragungszeit:

$$T_{Transfer} = \frac{D_{User} + D_{System}}{DR}$$



# Latenzzeit I

- Kommunikationszeit allgemein:

$$T_{Comm} = T_{Latency} + T_{Transfer}$$

- Übertragungszeit:

$$T_{Transfer} = \frac{D_{User} + D_{System}}{DR}$$



# Latenzzeit II

**Definition**

Die Latenz einer Verbindungseinrichtung entspricht der Zeitspanne zwischen der Einleitung eines Kommunikationsvorgangs bis zu dessen Abschluss, ohne Berücksichtigung der Übertragungszeit. Entsprechend ist die Latenz mit der Kommunikationszeit gleichzusetzen, die für die Übertragung einer Null-Byte-Nachricht benötigt wird.



# Entstehung von Latenz I


Abbildung: Open Systems Interconnection Reference Model

# Entstehung von Latenz II

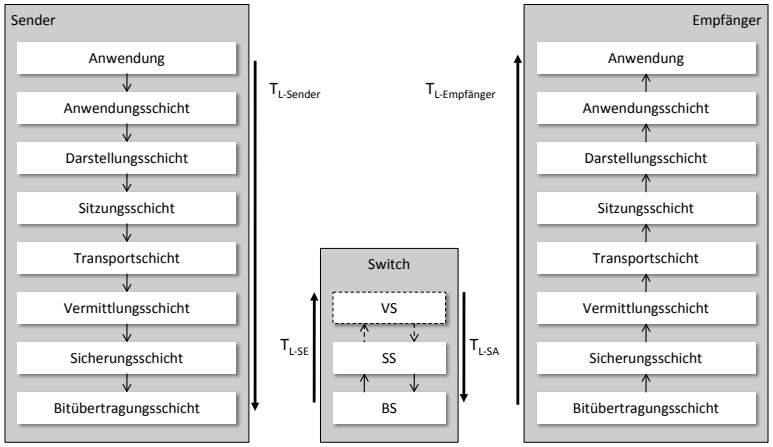


Abbildung: Zusammensetzung der Latenzzeit innerhalb einer Fabric



# Messung der Latenzzeit

- Bestimmung der Round-Trip-Time für eine Null-Byte-Nachricht
- Halbierung des gemessenen Wertes
- Probleme:
  - Abbildung der Prozesse auf physische Prozessoren
  - Ebene auf der gemessen wird
  - Großer Einfluss von Störungen durch andere Prozesse



# Beispiel Latenzzeitmessung I

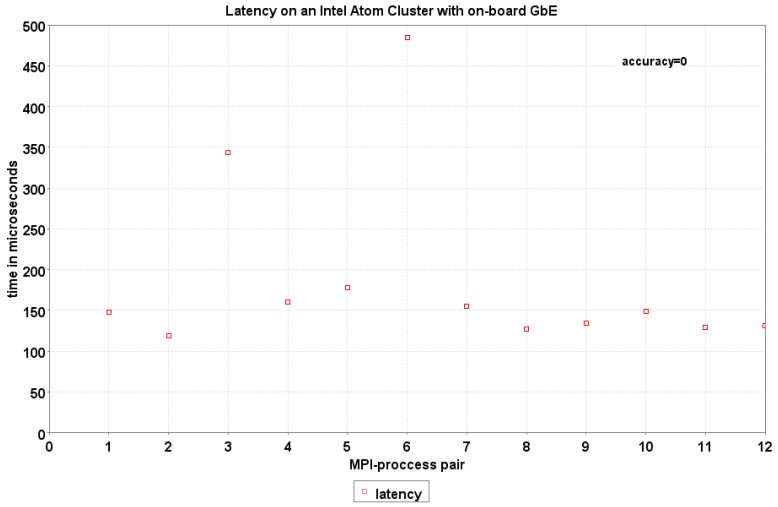


Abbildung: Einfache Latenzzeitmessung mit vier Prozessen



# Beispiel Latenzzeitmessung II

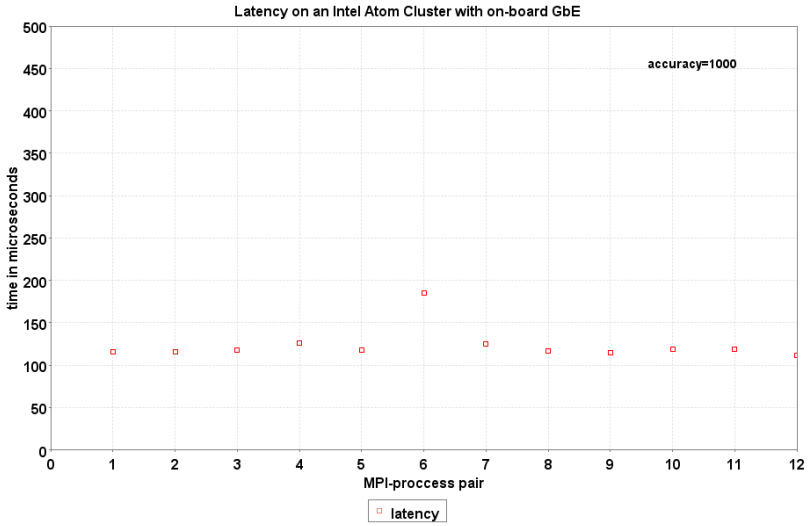


Abbildung: Latenzzeitmessung mit erhöhter Messgenauigkeit

## Agenda

2. Bandbreite, Symbolrate, Datenübertragungsrate und Datendurchsatz



## Bandbreite

**Definition**

Innerhalb eines konkreten Datenübertragungssystems entspricht die Bandbreite der Maximalfrequenz mit der ein rekonstruierbarer Signalwechsel stattfinden kann. Dadurch wird der Frequenzbereich definiert, in dem eine Signalübertragung möglich ist. Die Angabe erfolgt in Hertz [Hz].



## Symbolrate

**Definition**

Die Symbolrate - auch als Schrittgeschwindigkeit oder Baudrate bezeichnet - gibt die Anzahl der definierten Signaländerungen innerhalb eines Zeitintervalls an, welche gemessen werden können und wird in Baud [Bd] angegeben. Die theoretische obere Grenze wird dabei durch das Shannon-Hartley-Gesetz definiert:

$$C_N = 2 * B * ld(L)$$

Die so ermittelte maximale Symbolrate bezieht sich auf einen störungsfreien Übertragungskanal.



## Datenübertragungsrate

**Definition**

Die Datenübertragungsrate - häufig auch als Datentransferrate oder Datenrate bezeichnet - beschreibt die gesamte digitale Datenmenge, die auf einem Kanal übertragen werden kann und berechnet sich aus dem Produkt von Symbolrate und der Anzahl der möglichen Zustände pro Übertragungsschritt. Die Angabe erfolgt in Bit pro Sekunde [bit/s].

# Datendurchsatz

**Definition**

Der Datendurchsatz einer Verbindungseinrichtung beschreibt die Menge an Nutzdaten, die pro Zeiteinheit übertragen werden können in Bit pro Sekunde [bit/s]. Im Gegensatz zur Datenübertragungsrate bleiben Steuerinformation dabei unberücksichtigt.



# Beispiel - 1000BASE-T I

- Bandbreite eines Adernpaares: $B = 62.5MHz$
- Symbolrate: $C_N = 2 * B * ld(L) = 2 * 62.5MHz * ld(2) = 125MBaud$
- Datenrate:
  - Einsatz eines Pulsamplitudenmodulationsverfahrens mit fünf Amplitudenstufen:

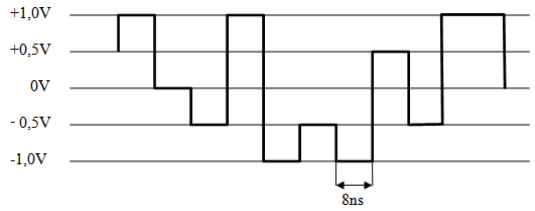


Abbildung: Eindimensionale 5-PAM Modulation



# Beispiel - 1000BASE-T II

- Berechnung der Datenrate:
  - Informationsgehalt: $ld(5) \approx 2,32 \frac{bits}{Symbol}$
  - Datenrate pro Adernpaar: $DR_{AP} = 2 \frac{bits}{Symbol} * 125MBaud = 250 \frac{Mbits}{s}$
  - Datenrate pro Kanal: $DR_{Kanal} = 4 * 250 \frac{Mbits}{s} = 1000 \frac{Mbits}{s}$



# Beispiel - 1000BASE-T III

- Berechnung des Datendurchsatzes:
  - Ethernet Type II Frame:
    - Nutzdaten: 1500 Bytes
    - Header: max. 22 Bytes
    - Präambel: 8 Bytes
    - Gap: 12 Bytes
  - Aufwendung von ca. 2.7% der Datenrate für Steuerinformationen
  - Datendurchsatz:

## Messung des Datendurchsatzes

- Messung der Zeit für eine Ping-Pong-Kommunikation unter Verwendung ausreichend großer Nachrichten
- Berechnung:

$$DD = \frac{D_{User}}{2 * T_{Comm}}$$

- Ergebnis strebt für $D_{User} \rightarrow \infty$ gegen den maximal erreichbaren Datendurchsatz



## Beispiel Durchsatzmessung

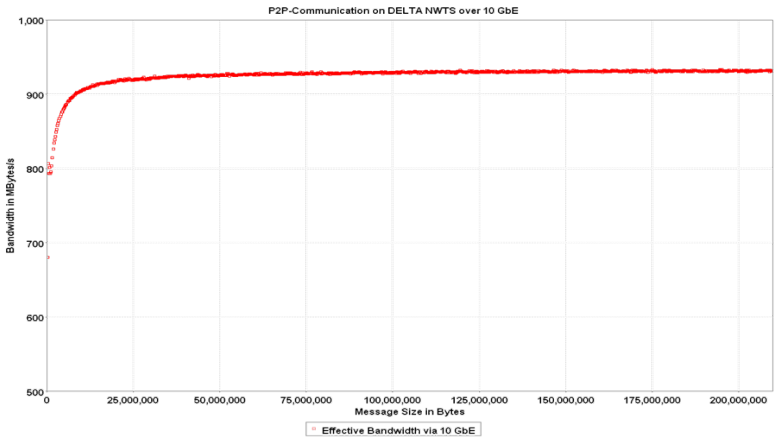


Abbildung: Datendurchsatz mit MPI über 10-Gigabit-Ethernet



## Agenda





## Overhead

- Im Zusammenhang mit Verbindungsnetzwerken existieren zwei Interpretationen:
  1. Die Menge an Informationen, welche zusätzlich zu den Nutzdaten übertragen werden müssen.
  2. Der von der CPU geleistete Aufwand, der notwendig ist um die Steuerinformationen zu berechnen.

## Messung von Overhead

- Bestimmung der Menge an hinzugefügten Steuerdaten mit Hilfe von sog. ,,Sniffern"
- Beispiel - Ethernet mit TCP/IP-Stack:

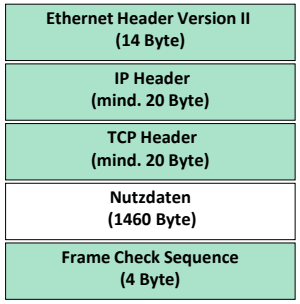


Abbildung: Protokoll-Overhead von Ethernet mit TCP/IP



## Auswirkung auf die Rechenleistung

- Hardware-Zugriff in herkömmlichen Netzwerken lediglich aus dem Kernel-Space
- Kontextwechsel verursachen zusätzlichen Aufwand
- Stetige Unterbrechung des laufenden Prozesses beeinträchtigt die erreichbare Rechenleistung

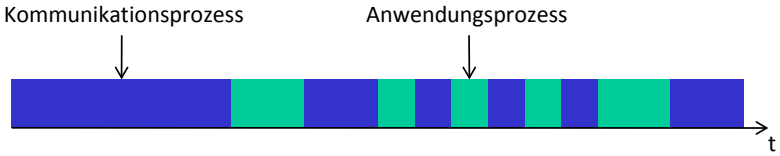


Abbildung: Kontextwechsel bei Kommunikation ohne Hardwareunterstützung



## Kommunikationsprozessoren I

- Integration vollständiger Chipsätze auf aktuellen Netzwerkadaptern
- Aufgaben:
  - Ausführung der Firmware
  - Berechnung der Steuerinformationen
  - Direct Memory Access
- Beispiele: Lanai [Myri], Terminator ASIC [Chel], InfiniHost [Mel]



## Kommunikationsprozessoren II

- Einsparung von Kontextwechseln und Verlagerung des Rechenaufwands führt zu höherer Rechenleistung
- Ermöglicht die ,,Maskierung" von Kommunikationsvorgängen

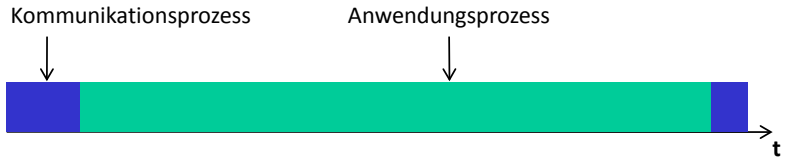


Abbildung: Kommunikationsverlauf beim Einsatz von Kommunikationsprozessoren

## Zusammenhänge

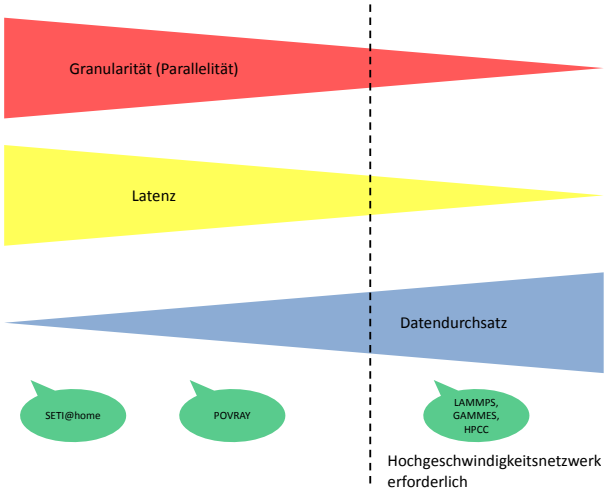


Abbildung: Zusammenhänge zwischen Granularität, Latenz und Datendurchsatz



## Agenda





## Literaturverzeichnis

[Myri] Myricom
Custom-VLSI Chips, Nov 2008
http://www.myri.com/vlsi/

[Chel] Chelsio Communications
T4 ASIC, 2010
http://www.chelsio.com/t4_asic.html

[Mel] Mellanox Technologies
InfiniBand Adapter Silicon, Nov 2010
http://www.mellanox.com/content/pages.php?pg=products_dyn&product_family=6&menu_section=32