

VERNETZUNG

- ▶ Vernetzung von Rechnern und Eingabe/Ausgabe
- ▶ Allgemeine Betrachtungen
- ▶ Designaspekte effizienter Kommunikation
- ▶ Leistungsentwicklung
- ▶ Leistungsmaße
- ▶ Netztopologien
- ▶ Einbindung in TCP/IP
- ▶ InfiniBand-Vernetzung

Vernetzung

Die zehn wichtigsten Fragen

- ▶ Welche Aufgaben hat die Vernetzung?
- ▶ Welche Komponenten weist eine Vernetzung auf?
- ▶ Welche Fragen finden wir bei der Pufferverwaltung?
- ▶ Was bedeutet Überlagerung von Berechnung und Kommunikation?
- ▶ Was versteht man unter Hardware-Realisierung von Software?
- ▶ Welche Bandbreiten finden wir bei aktuellen Netztechnologien?
- ▶ Welche Charakteristiken sollte die Netzhardware aufweisen?
- ▶ Was versteht man unter Bisektionsbandbreite?
- ▶ Wie umgeht man die Engstelle TCP/IP
- ▶ Welche Protokolle finden wir bei InfiniBand?

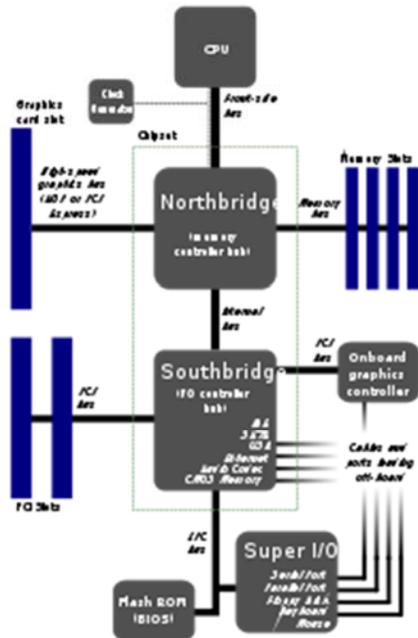
Aufgabenstellung

Wozu Vernetzung?

- ▶ Die Vernetzung verbindet unabhängige Rechnerknoten
- ▶ Ermöglicht die Interprozeßkommunikation
 - ▶ Prozesse auf verschiedenen Knoten
- ▶ Ermöglicht Prozeß-Eingabe/Ausgabe
 - ▶ E/A-Hardware meist nicht knotenlokal angeschlossen

Die Vernetzung (Hardware und Software) unterstützt die Interprozeßkommunikation und die Prozeß-E/A.

Rechnerinterne Vernetzung



- ▶ Front-Side-Bus
 - ▶ Z.B. 64 Bit mit 100 MHz und 4 Übertragungen/Takt ergibt 3.200 MByte/s
- ▶ Bussystem an der Southbridge schafft Verbindung zu Peripherie
- ▶ Prozeß-zu-Prozeß
 - ▶ Gemeinsamer Speicher
 - ▶ Socketkommunikation
- ▶ Prozeß-zu-Datei
 - ▶ Datei-E/A

Quelle: http://en.wikipedia.org/wiki/File:Motherboard_diagram.svg

Im Rechner finden wir verschiedene Bussysteme, die die Komponenten in Verbindung bringen. Softwareseitig verwenden wir zur Interprozeßkommunikation Socketprogrammierung basierend auf TCP/IP. Alternativ können gemeinsame Speicherbereiche verwandt werden. Hier gibt es aber keine genormte Programmierschnittstelle.

Vernetzung von Rechnern und E/A

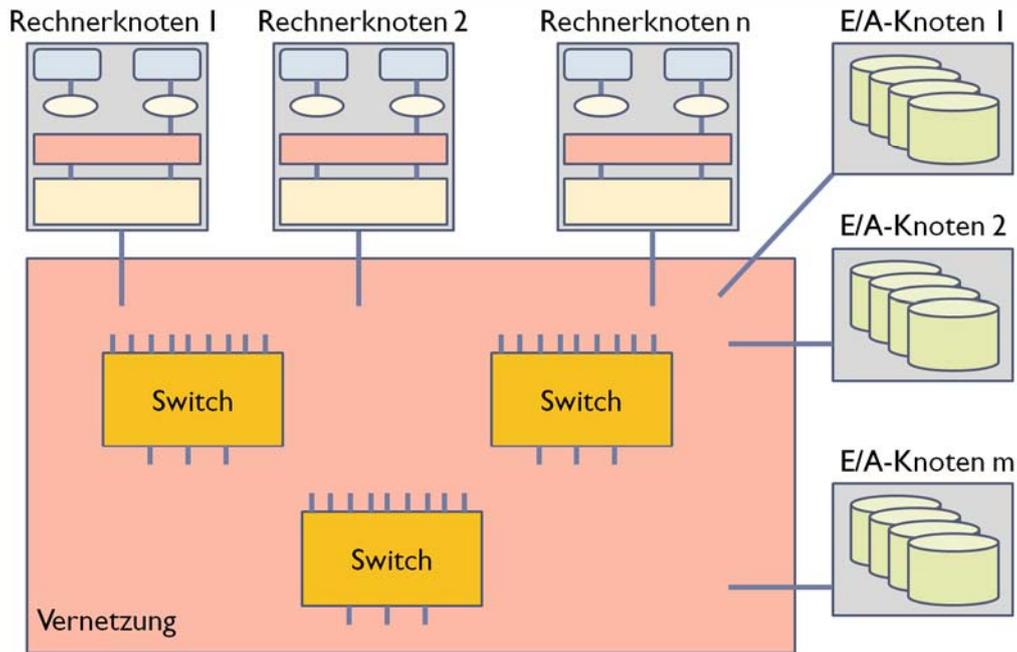
Bestandteile der Vernetzung

- ▶ NIC (network interface card) – mindestens eine pro Rechner
- ▶ Verbindungsleitungen: Kupfer, Glasfaser
- ▶ Switches – zur wechselseitigen Verbindung von Komponenten (Rechnern, E/A-Knoten)

Gegebenenfalls getrennte Vernetzung für

- ▶ Prozeßkommunikation
- ▶ Eingabe/Ausgabe zu Dateisystemen und Bandarchiv
- ▶ Wartung und Kontrolle der Rechner

Vernetzung von Rechnern und E/A



▶ 106

Hochleistungsrechnen - © Thomas Ludwig

19.04.2010

Die Rechner und die E/A-Komponenten werden über Netze verbunden. Dies können getrennte oder gemeinsam genutzte Netzwerke sein. Die Verknüpfung erfolgt über Switches.

Allgemeine Betrachtungen zur Software

Was interessiert uns an der Vernetzung?

- ▶ Programmierschnittstelle hoher Abstraktion
 - ▶ Portabilität der Programme wichtig
- ▶ Geringe Zusatzlast
- ▶ Hardware voll ausnutzbar
- ▶ Anpaßbar an unterschiedliche Realisierungen der Vernetzungshardware
- ▶ Software zum Netzmanagement

Allgemeine Betrachtungen zur Hardware

Was interessiert uns an der Vernetzung?

- ▶ Datenraten
- ▶ Latenzzeiten
- ▶ Bestandteile der Vernetzung
 - ▶ Kabel: Kupfer, Glasfaser
 - ▶ Netzkomponenten: Karten, Switches
- ▶ Ausfallsicherheit
- ▶ Adaptive Wegewahl
 - ▶ Bei Überlast / bei Fehlern
- ▶ Anbindung an TCP/IP

Designaspekte effizienter Kommunikation

Überblick über wichtige Designaspekte

- ▶ Pufferverwaltung
- ▶ Überlappung von Berechnung und Kommunikation
- ▶ Realisierung in Hardware
- ▶ Datentransport

Pufferverwaltung (1/2)

Warum ist Pufferverwaltung relevant?

- ▶ Verwaltung von Speicherplatz ist sehr teuer
- ▶ Umkopiervorgänge sind sehr teuer

Aufgabenstellung

- ▶ Nachricht steht sendebereit im Adreßraum des sendenden Prozesses
- ▶ Nachricht durch alle Softwareschichten und das Netz in den Speicher des Empfängers befördern
- ▶ Nachricht im Adreßraum des Empfängers zur Verfügung stellen

Pufferverwaltung (2/2)

- ▶ **Zero-Copy-Mechanismen**
 - ▶ Am besten aus einem Adreßraum sofort in den Netzadapter übertragen – schwierig!
- ▶ **Speicherregistrierung**
 - ▶ Moderne Vernetzungen gestatten Remote Direct Memory Access (RDMA)
 - ▶ Setzt die Registrierung von Speicherbereichen voraus – zeitaufwendig
- ▶ **Unerwartete Nachrichten**
 - ▶ Der Empfänger hat keine Kenntnis, daß eine Nachricht eintreffen wird
 - ▶ Entsprechend sind keine Puffer allokiert und der Ablauf verlangsamt sich

Überlappung von Berechnung und Kommunikation

Welche Phasen sehen wir bei der Kommunikation?

- ▶ Daten aus der Anwendung zum Sendenetzadapter
- ▶ Daten vom Sendenetzadapter zum Empfangsnetzadapter übertragen
- ▶ Daten vom Empfangsnetzadapter zur Anwendung

Was soll überlappend stattfinden?

- ▶ Am besten alle drei Phasen!

Was kann aktuelle Hardware

- ▶ Verschieden gute Varianten der optimalen Lösung

Noch problematisch:

- ▶ Kann die Software das ausnutzen?

Realisierung in Hardware

Moderne Netztechnologien gestatten die Abarbeitung eines Teils der Software-Schichten in der Adapterhardware (genannt offloading)

Diese Hardware nennt man für TCP/IP:

- ▶ TCP/IP offload Engine, kurz ToE

Erhöht gegebenenfalls die Überlappung von Berechnung und Kommunikation

Datentransport

- ▶ **Zuverlässigkeit: weniger kritisch als in WANs**
- ▶ **Paketgröße und MTU**
 - ▶ IP-Paket: 64 KB, Ethernet Standard: 1.500 Byte
 - ▶ Ethernet verwendet sog. Jumbo-Frames
- ▶ **DMA-basiert**
 - ▶ Entlastet Prozessor
- ▶ **Unterbrechungen und Polling (Abfrage)**
 - ▶ Unterbrechungen sind schwergewichtig
 - ▶ Polling benötigt Zeit vom Prozessor
 - ▶ Wir finden beide Realisierungen

Leistungsentwicklung

In den Jahren von 1990-2010 sehen wir verschiedene Generationen von internen Bussen und externen Netztechnologien

Die Geschwindigkeitssteigerungen sind viel geringer als bei den Rechnern!

Bussysteme

| | | |
|--------------------------------|--|---|
| PCI | 1990 | 33MHz/32bit: 1.05Gbps (shared bidirectional) |
| PCI-X | 1998 (v1.0) 2003 (v2.0) | 133MHz/64bit: 8.5Gbps (shared bidirectional) 266-533MHz/64bit: 17Gbps (shared bidirectional) |
| HyperTransport (HT) by AMD | 2001 (v1.0), 2004 (v2.0) 2006 (v3.0), 2008 (v3.1) | 102.4Gbps (v1.0), 179.2Gbps (v2.0) 332.8Gbps (v3.0), 409.6Gbps (v3.1) |
| PCI-Express (PCIe) by Intel | 2003 (Gen1), 2007 (Gen2) 2009 (Gen3 standard) | Gen1: 4X (8Gbps), 8X (16Gbps), 16X (32Gbps) Gen2: 4X (16Gbps), 8X (32Gbps), 16X (64Gbps) Gen3: 4X (~32Gbps), 8X (~64Gbps), 16X (~128Gbps) |
| Intel QuickPath | 2009 | 153.6-204.8Gbps per link |

Quelle: Tutorial von Panda, Balaji, Koop: InfiniBand and 10-Gigabit Ethernet for Dummies, Supercomputing 2009.

Von 1998 bis 2009 in 12 Jahren ein Faktor von nicht mehr als 25. Bei Rechnern mehr als Faktor 1000.

Siehe auch: http://en.wikipedia.org/wiki/List_of_device_bandwidths - für verschiedenste Gerätetypen

Vernetzungstechnologien

| | |
|------------------------------|-----------------------|
| Ethernet (1979 -) | 10 Mbit/sec |
| Fast Ethernet (1993 -) | 100 Mbit/sec |
| Gigabit Ethernet (1995 -) | 1000 Mbit/sec |
| ATM (1995 -) | 155/622/1024 Mbit/sec |
| Myrinet (1993 -) | 1 Gbit/sec |
| Fibre Channel (1994 -) | 1 Gbit/sec |
| InfiniBand (2001 -) | 2 Gbit/sec (1X SDR) |
| 10-Gigabit Ethernet (2001 -) | 10 Gbit/sec |
| InfiniBand (2003 -) | 8 Gbit/sec (4X SDR) |
| InfiniBand (2005 -) | 16 Gbit/sec (4X DDR) |
| | 24 Gbit/sec (12X SDR) |
| InfiniBand (2007 -) | 32 Gbit/sec (4X QDR) |
| InfiniBand (2011-) | 64 Gbit/sec (4X EDR) |

Quelle: Tutorial von Panda, Balaji, Koop: InfiniBand and 10-Gigabit Ethernet for Dummies, Supercomputing 2009.

In den 9 Jahren von 2001-2009 sehen wir bei Infiniband einen Faktor von 16. Bei den Rechnern grob einen Faktor von 300.

Leistungsmaße

Wünschenswerte Charakteristika

- ▶ Niedrige Latenz (Aufsetzzeit der Nachrichten)
- ▶ Hohe Bandbreite
- ▶ Geringe Belastung der CPU
- ▶ Hohe Bisektionsbandbreite des Netzes

Vorgriff auf Programmierkonzepte

- ▶ Möglichst wenig Kommunikation verwendet
- ▶ Wenn überhaupt, dann lieber wenige große als viele kleine Pakete

Netztopologien

Wünschenswerte Eigenschaften

- ▶ Identische Datenraten zwischen beliebigen Knoten
- ▶ Vermeidung von Engpässen, Überlastungen, Ausfällen
- ▶ Erweiterbarkeit
- ▶ Skalierbarkeit
- ▶ Gutes Preis/Leistungsverhältnis

Bisektion des Netzes

Bisektionsbreite

Anzahl der Verbindungen, die man trennen muß, damit das Netz in zwei isolierte Teile mit gleicher Knotenzahl zerfällt

- ▶ Je mehr, desto besser

Bisektionsbandbreite

Aggregierte Bandbreite der durchtrennten Leitungen
(=Fluß an der engsten Stelle)

Bisektion des Netzes

Volle Bisektionsbandbreite

Ein Netz mit n Knoten hat volle Bisektionsbandbreite, wenn die Summe aller Bandbreiten von Verbindungen zwischen zwei beliebigen Hälften des Netzes $n/2$ -mal die Bandbreite einer einzelnen Verbindung ist

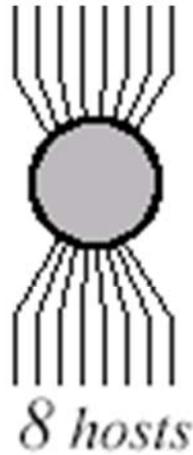
Warum interessiert uns das?

- ▶ Wir benötigen ausreichend Switches und Wege, um alle Komponenten leistungsfähig zu vernetzen

Wir betrachten einen Switch mit 64 Ports und 1 Gbit/s pro Port. Wenn 32 Sender mit 32 Empfängern kommunizieren wollen, benötigen wir 32 Gbit/s an Bandbreite durch den Switch hindurch. Bidirektional das doppelte. Die interne Topologie des Switches sollte also diese Gesamtleistung ermöglichen. In der Praxis ist sie meist geringer.

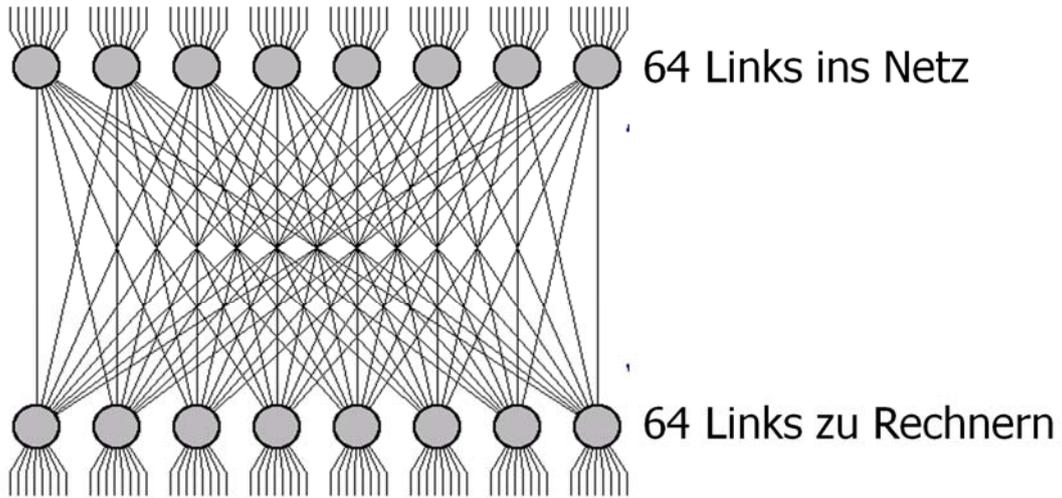
Bisektion am Beispiel Myrinet

- ▶ Basisbaustein: Kreuzschiene mit 16 Anschlüssen
- ▶ Topologie benannt nach Charles Clos (1952)

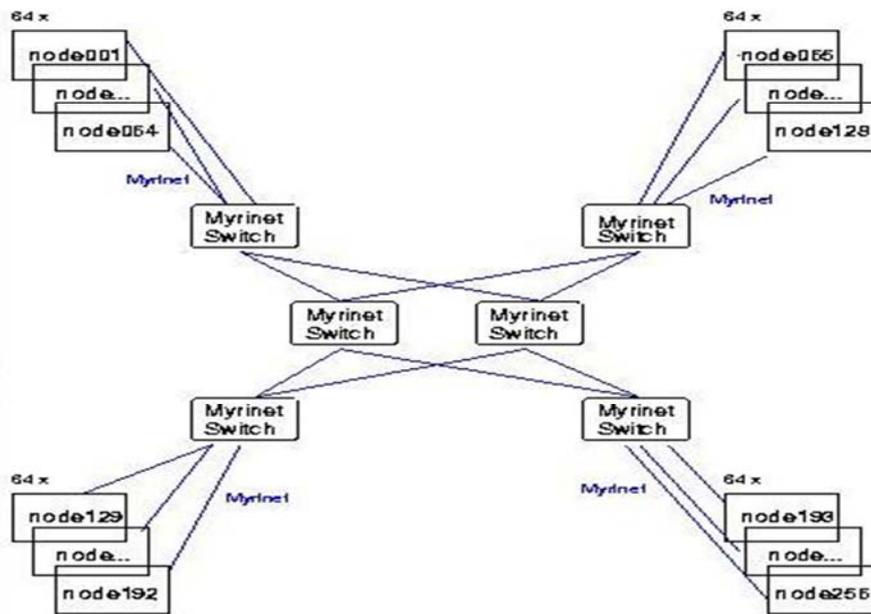


Myrinet-Switch

Neue Basiskomponente aus 16 des bisherigen Typs

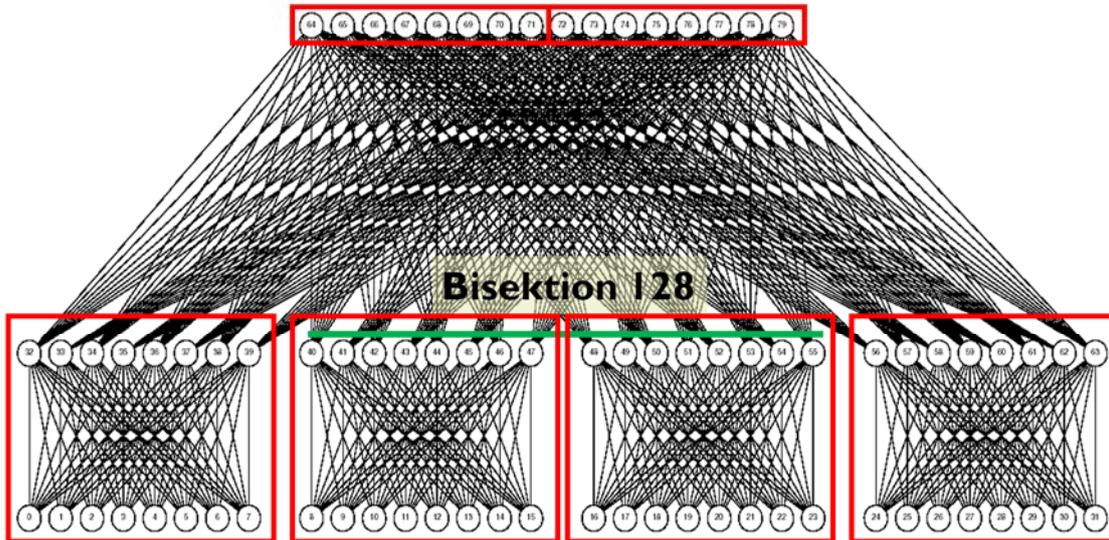


Heidelberger Helics-Cluster mit Myrinet



Myrinet im Detail...

4 Clos64-Bausteine und 2 Switche für 256 Rechner



... und in der Praxis



▶ 126

Hochleistungsrechnen - © Thomas Ludwig

19.04.2010

Einbindung in TCP/IP

- ▶ **Aus Sicht des Programms**
 - ▶ Socket-Programmierung
 - ▶ TCP/IP-Stack im Betriebssystem
- ▶ **Aus Sicht der Hardware**
 - ▶ Varianten von Ethernet sehr populär

- ▶ **Im Rechner-Cluster praktisch immer auch TCP/IP involviert**
- ▶ **Problem: TCP/IP ist schwerfällig**

Einbindung in TCP/IP

- ▶ Probleme mit TCP/IP
 - ▶ Viele Protokollschichten
 - ▶ Nicht geringe Prozessorbelastung
 - ▶ Integration in das Betriebssystem
 - ▶ Kopiervorgänge
 - ▶ Anzahl der Unterbrechungen

In den zahlreichen Protokollschichten werden Aktivitäten ausgeführt, die in einem lokalen Cluster ohne Bedeutung sind, aber Software-Aufwand hervorrufen. Z.B. die Überlaststeuerung zwischen entfernten Knoten.

Die umfangreichen Software-Schichten bedeuten auch eine nicht vernachlässigbare Belastung des Prozessor, der diese ja abarbeiten muß.

Die Integration in das Betriebssystem ist nachteilig, da die Anwendung ja immer erst durch dieses hindurch auf die NIC zugreifen muß, anstatt einen eigenen Weg zu haben.

Beim Durchlaufen der Schichten und beim Übergang zwischen Betriebssystem und Anwenderprogramm kommt es immer wieder zu zeitaufwendigen Kopiervorgängen.

Aufgrund der maximalen MTU (maximum transmission unit) von 1500 Byte kommt es auch bei GigaBit-Ethernet zu hohen Raten von Unterbrechungen. Größere Frames (Jumbo-Frames) schaffen hier ein wenig Abhilfe.

Einbindung in TCP/IP

▶ Lösungsansätze

- ▶ TCP-Bypass
 - ▶ Definition eines eigenen Paketformats
 - ▶ Evtl. Problem: nur cluster-lokal verwendbar
- ▶ TCP Offload Engine
 - ▶ Z.B. eine GigE Verbindung kann einen Pentium IV mit 2,4 GHz auslasten
 - ▶ Deshalb extra Prozessor für TCP/IP-Protokollstapel
 - ▶ Normalerweise auf der NIC untergebracht
- ▶ Kernel Bypass
 - ▶ Die NIC kommuniziert direkt mit dem Programm
 - ▶ Dies findet man bei allen nicht Ethernet-Vernetzungen!

Einbindung in TCP/IP

▶ Lösungsansätze...

- ▶ Remote Direct Memory Access (RDMA)
 - ▶ Die NIC kann direkt in den Speicher eines entfernten Rechners schreiben
- ▶ Zero-Copy Networking
 - ▶ Vermeide Kopien zwischen Betriebssystem und Anwendungsprogramm
 - ▶ Verwende stattdessen DMA und MMU
- ▶ Interrupt Mitigation
 - ▶ Fasse mehrere Unterbrechungen zu einer zusammen

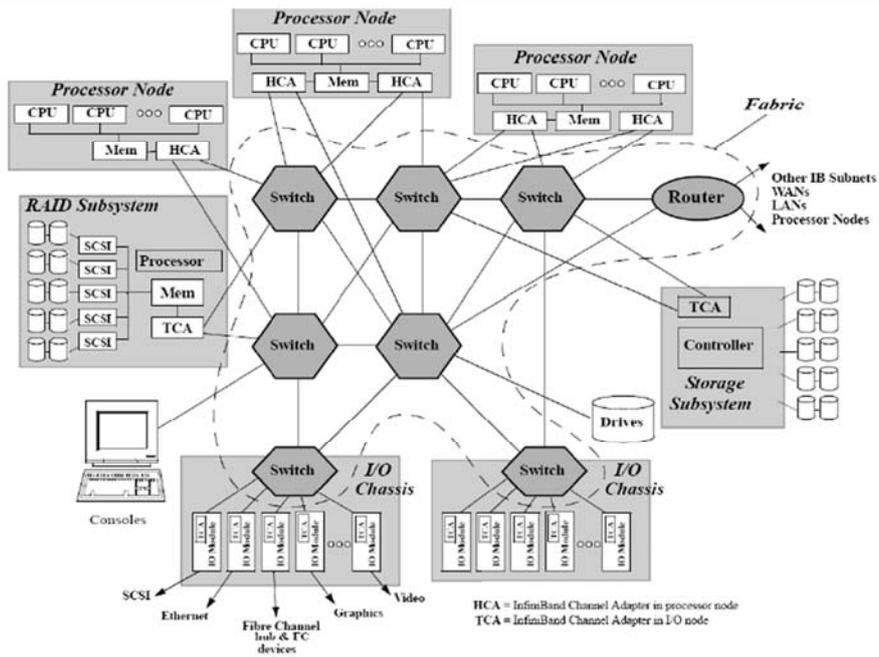
InfiniBand-Vernetzung

- ▶ InfiBand ist ein Industriestandard einer Gruppe von Herstellern genannt Open Fabrics Alliance
- ▶ Definiert einen Standard für ein Systemnetz
 - ▶ Früher auch als Spezifikation eines rechnerinternen Busses bekannt – dieser Ansatz wurde nicht weiter verfolgt

- ▶ Unterscheidung zwischen Rechnern und E/A-Geräten
 - ▶ Für Rechner verwendet: Host Channel Adapter (HCA)
 - ▶ Für E/A-Geräte verwendet: Target Channel Adapter (TCA)

Siehe: <http://de.wikipedia.org/wiki/InfiniBand>

InfiniBand-Netz: ein Überblick



InfiniBand-Software

- ▶ **IP over InfiniBand (IPoIB)**
 - ▶ Ermöglicht IP-basierten Protokollen eine Kommunikation über InfiniBand
 - ▶ Durchsatz von 300 MByte/s ist höher als der von Gigabit-Ethernet (GE)
 - ▶ Latenz von 20 μ s vergleichbar zu GE
- ▶ **Sockets direct protocol (SDP)**
 - ▶ Durchsätze von 900 MByte/s
 - ▶ Latenzen von 12 μ s

Vernetzung

Zusammenfassung

- ▶ Die Vernetzung bringt Rechner miteinander und mit den E/A-Geräten in Verbindung
- ▶ Vernetzungshardware sind Netzadapter, Kabel, Switches
- ▶ Wir möchten hohe Datenraten und geringe Latenzen bei gleichzeitiger guter Skalierbarkeit
- ▶ Effiziente Kommunikation umfaßt viele Einzelaspekte
- ▶ In den letzten 10 Jahren haben sich Rechengeschwindigkeiten 10x schneller gesteigert als Netzgeschwindigkeiten
- ▶ Ein wichtiges Maß der Topologie ist die Bisektionsbandbreite
- ▶ TCP/IP ist schwerfällig und wird häufig ersetzt
- ▶ Infiniband ist die aktuelle Hochleistungsvernetzung beim Hochleistungsrechnen