TECHNISCHE UNIVERSITÄT DRESDEN
Fakultät Informatik, Institut für Technische Informatik, Professur Rechnerarchitektur

Einführung in die Technische Informatik

Parallelrechner

Zellescher Weg 12
Willers-Bau A 205
Tel. +49 351 - 463 - 35450

Nöthnitzer Straße 46
Raum 1044
Tel. +49 351 - 463 - 38246

Wolfgang E. Nagel (wolfgang.nagel@tu-dresden.de)

ZIH
Center for Information Services & High Performance Computing

Gliederung

- Definitionen und Klassifizierungen
  - Klassifizierung nach FLYNN
  - Einordnung nach der Organisation des Speichers
    - Physikalische Organisation des Speichers
    - Sicht des Programmierers auf den Speicher
  - Parallelrechner-Kategorien in der TOP 500 Liste
  - Charakterisierungsansatz nach Dongarra, Sterling, Simon, Strohmaier
- Beispiele von Parallelrechnern
  - SGI Altix 4700
  - SGI UV
  - Megware-Cluster Atlas
  - Bull HPC Cluster Taurus
  - BlueGene
- Hochgeschwindigkeitsnetzwerke

Definitionen und Klassifizierungen

Definition:

Ein Parallelrechner ist eine Ansammlung von Berechnungseinheiten (Prozessoren), die durch koordinierte Zusammenarbeit große Probleme schnell lösen können.

Definition ist sehr allgemein gehalten, um Vielzahl der Parallelrechner zu erfassen. Offen bleibt z.B.:

- Anzahl und Komplexität der Berechnungseinheiten
- Struktur des Verbindungsnetzwerkes
- Koordination der Arbeit der Berechnungseinheiten
- Eigenschaften der zu lösenden Probleme

➔ Weitergehende Klassifikationen notwendig

Flynnsche Klassifizierung (siehe RAII-Vorlesung)

- **SISD** - single instruction stream, single data stream
- **MISD** - multiple instruction streams, single data stream
- **SIMD** - single instruction stream, multiple data streams
- **MIMD** - multiple instruction streams, multiple data streams

Fast alle heutigen Parallelrechner arbeiten nach dem MIMD-Prinzip, deshalb ist eine weitere Unterteilung der MIMD-Klasse notwendig.

MIMD – Klassifizierung nach der Org. des Speichers

Organisation des Speichers wird unterschieden nach:

- Physikalische Organisation des Speichers
  - Rechner mit physikalisch verteiltem Speicher (Multicomputer)
  - Rechner mit physikalisch gemeinsamen Speicher (Multiprozessor)
  - Rechner mit virtuell gemeinsamen Speicher (Multiprozessor)

  Multiprozessor
  Prozessoren des MIMD-Rechners nutzen einen gemeinsamen Adressraum

  Multicomputer
  Prozessoren des MIMD-Rechners nutzen eigene Adressräume

MIMD – Klassifizierung nach der Org. des Speichers

- Sicht des Programmierers auf den Speicher
  - Rechner mit verteiltem (lokalem) Adressraum
  - Rechner mit gemeinsamen (globalen) Adressraum

- Sicht des Programmierers kann sich unterscheiden von der physikalischen Organisation des Speichers
  - MPI auf einem Multiprozessorsystem
  - Cluster OpenMP auf einem Cluster

MIMD – Rechner mit physikalisch verteiltem Speicher

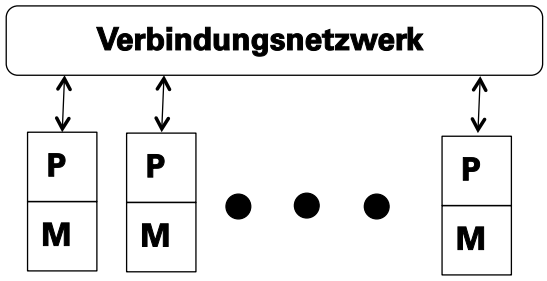
Bestandteile

- Mehreren Verarbeitungseinheiten (Knoten)
  - Ein Knoten ist eine selbständige Einheit aus:
    - Prozessor
    - Lokalem Speicher
    - Evtl. I/O-Anschlüssen
- Verbindungsnetzwerk
  - Verbindet Knoten durch physikalische Leitungen

Kommunikation zwischen Knoten

- Nachrichtenaustausch

Rechner mit verteiltem Speicher- Abstrakte Struktur



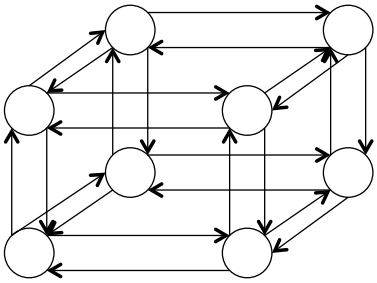
P = Prozessor

M = lokaler Speicher

Frühe Rechner mit verteiltem Speicher (Multicomputer)

Verwendung von Punkt-zu-Punkt-Verbindungen zwischen den Knoten

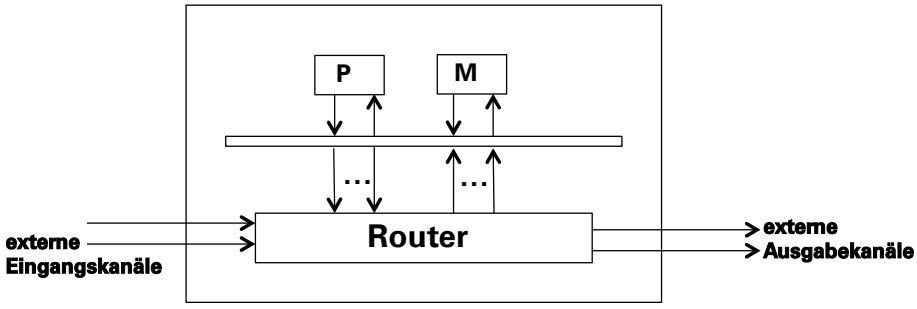
- Knoten können nur mit direkten Nachbarn kommunizieren
  - Kommunikation durch Gestalt des Netzwerkes vorgegeben
    z.B. Hypercube



- Kommunikation nur, wenn benachbarte Knoten gleichzeitig auf den verbindenden Link schreiben bzw. von ihm lesen
  - Bedingt durch kleine Puffer, die die Zwischenspeicherung von großen Nachrichten nicht zulassen

Moderne Multicomputer

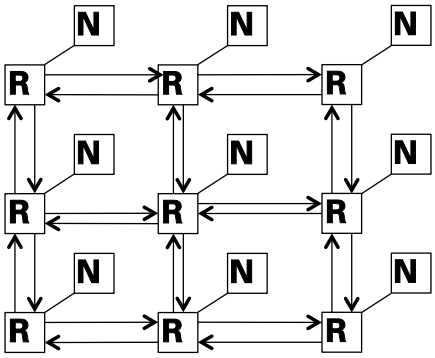
- Verwendung von Prozessor-Speicher-Knoten mit Hardware-Router



- Puffer am Ende von Eingabe- und Ausgabekanälen zum Zwischenspeichern von Nachrichten

Moderne Multicomputer

- Router bilden das eigentliche Netzwerk, das hardwaremäßig die Kommunikation mit allen auch weiter entfernten Knoten übernimmt



R = Router
N = Knoten bestehend aus Prozessor und lokalem Speicher

- Verringerung der Kommunikationszeit
- Verringerung der Kommunikationszeit-Differenz zwischen benachbarten und weit entfernten Knoten

Multicomputer

Multicomputer werden üblicherweise unterteilt in:

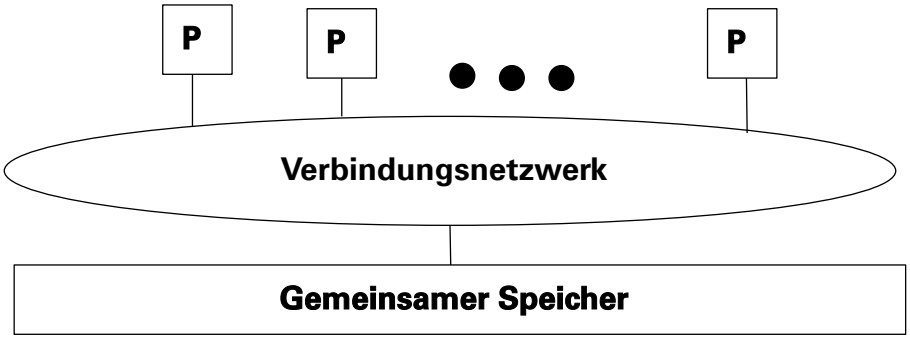
- **Cluster** (COW - Cluster of Workstations, NOW – Network of Workstations)
  - PCs oder Arbeitsstationen, die gegebenenfalls in Gestellrahmen montiert und über handelsübliche Zwischenverbindungen zusammengeschlossen sind

- **MPP** (Massively Parallel Processors)
  - Teure Supercomputer mit vielen CPUs, die über ein proprietäres Hochgeschwindigkeitsnetz eng gekoppelt sind

MIMD – Multiprozessoren

Rechner mit physikalisch gemeinsamen Speicher

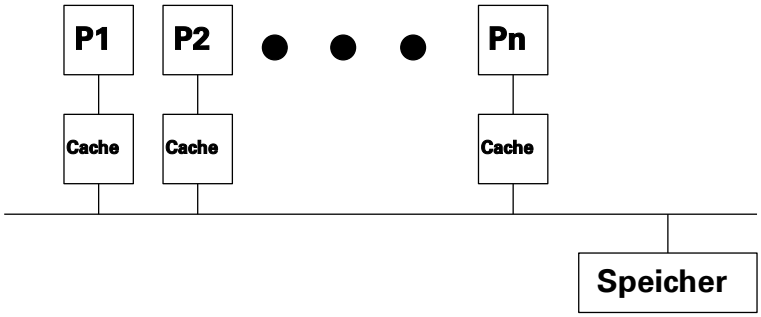
- Bestehen aus:
  - Mehreren Prozessoren oder Prozessorkernen
  - Einen gemeinsamen oder globalen Speicher
  - Verbindungsnetzwerk
    - Verbindet Prozessoren oder Prozessorkerne mit dem gemeinsamen Speicher



SMP – symmetrische Multiprozessoren

SMP (symmetric multiprocessor)

- Spezielle Variante von Rechnern mit gemeinsamen Speicher
- Bestehen aus:
  - Einer kleinen Anzahl von Prozessoren oder Prozessorkernen
  - Verbindungsnetzwerk ist oft ein zentraler Bus
  - Keine zusätzlichen privaten Speicher für Prozessoren
    - Lokale Caches für Prozessoren sind üblich

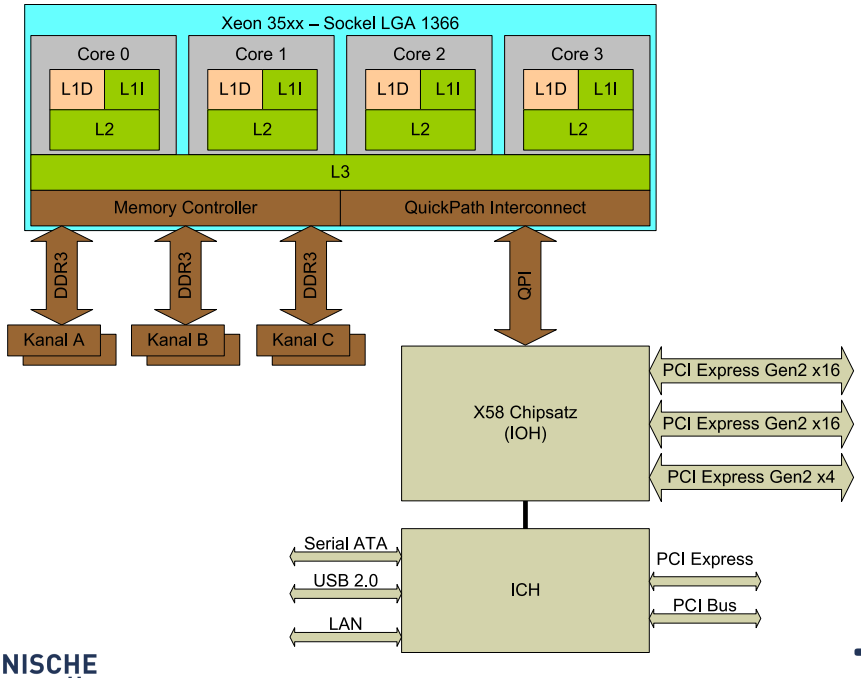


SMP – symmetrische Multiprozessoren

- „Symmetrisch“ bezieht sich auf die Prozessoren bzw. Prozessorkerne und bedeutet:
  - Prozessoren haben gleiche Funktionalität
  - Prozessoren haben gleiche Sicht auf das Gesamtsystem
    - Dauer eines Zugriffs auf den gemeinsamen Speicher dauert für jeden Prozessor gleich lange (UMA – Uniform Memory Access))
- Aktuelles Beispiel eines SMP-Systems
  - Multicore-Prozessor als 1-Sockel-System

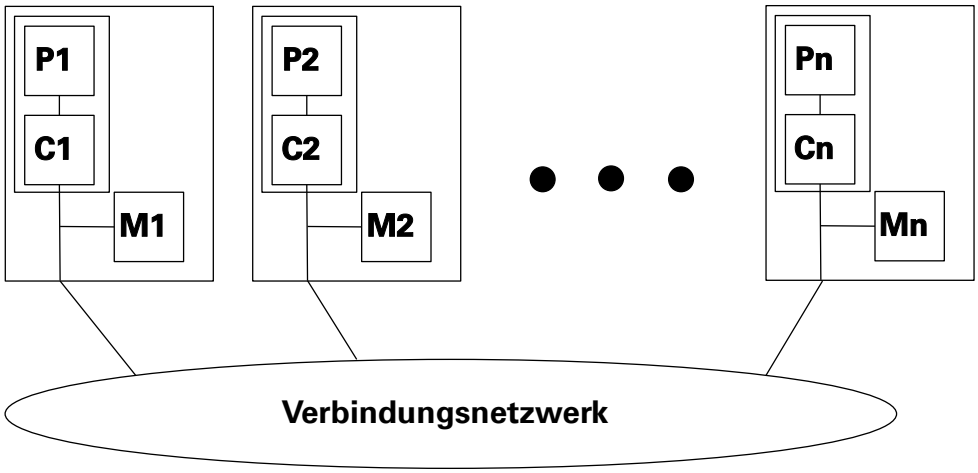
SMP – symmetrische Multiprozessoren

- 1-Sockel-System auf Basis des Xeon 3500 Prozessors



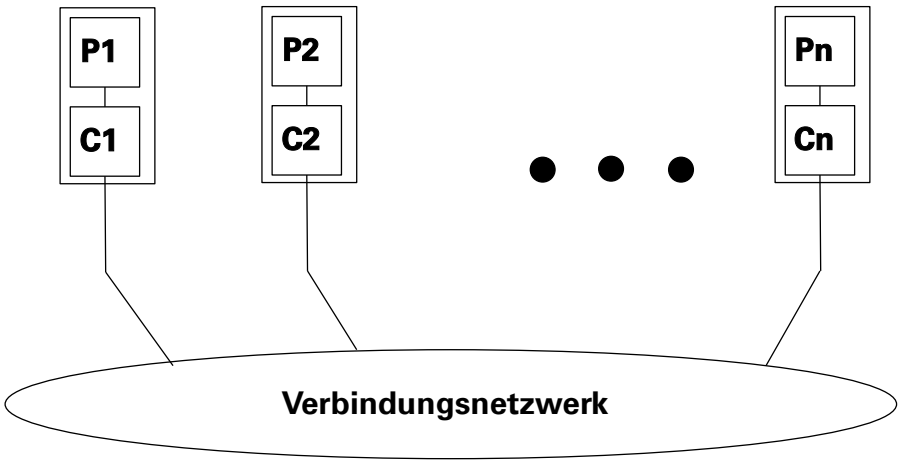
NUMA (Non-Uniform Memory Access) - Multiprozessoren

- Unterschiedliche Zugriffszeiten zu den Speichermodulen
- Zu den CPUs gehören oftmals lokale bzw. gruppenlokale Speichermodule, auf die die CPUs wesentlich schneller zugreifen können als auf entfernte Module
- Platzierung von Programmen und Daten beeinflusst die Performance
- Unterscheidung in:
  - Architektur mit Cache (CC-NUMA)
    - CC steht für Cache Coherent
  - Architektur ohne Cache (NC-NUMA)
    - NC steht für No Cache



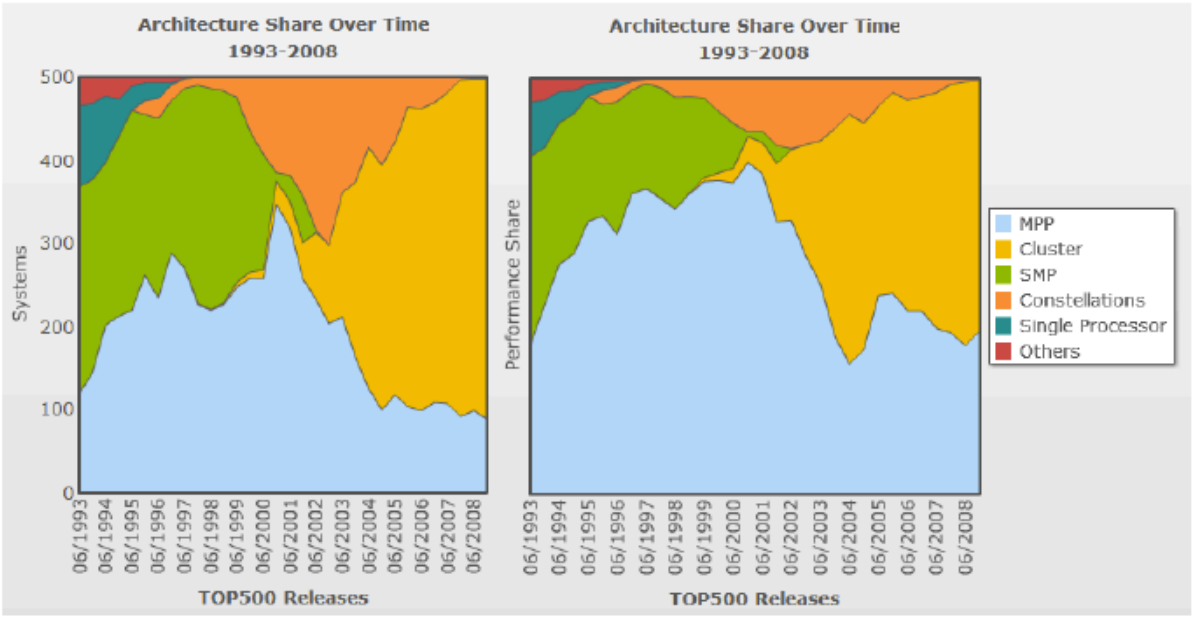
COMA (Cache Only Memory Access) - Multiprozessoren

- Physikalischer Adressraum wird in Cache-Zeilen aufgeteilt
- Cache-Zeilen wandern auf Anforderung im System umher



Parallelrechner-Kategorien in der TOP 500 Liste

- Neben MPP, Cluster, SMP wird noch der Begriff „Constellations“ verwendet



Parallelrechner-Kategorien in der TOP 500 Liste

- Unter „Constallation“ versteht man ein Cluster von SMP-Rechnern

- Definition „Constellation“ nach Tom Sterling (Beowulf-Pionier):

  A constellation is a cluster of large SMP nodes scaled such that the number of processors per node is greater than the number of nodes.

Charakterisierungsansatz nach Dongarra, Sterling, Simon, Strohmaier

Jack Dongarra (University of Tennessee)
- Entwickelte LINPACK
- Maßgeblich an TOP 500 Liste beteiligt

Thomas Sterling (California Institute of Technology)

Horst Simon (Lawrence Berkley National Laboratory)

Erich Strohmaier (Lawrence Berkley National Laboratory)

Für die Charakterisierung von Parallelrechnern werden folgende Basis-Kategorien vorgeschlagen:
- Clustering
- Name space
- Parallelism
- Latency/locality management

Daraus ergibt sich für die Charakterisierung des Parallelrechners dieses Schema
- Clustering/Naming/Parallelism/Latency

Charakterisierungsansatz nach Dongarra, Sterling, Simon, Strohmaier

Für jede Kategorie gibt es eine Anzahl von Attributen

- Clustering
  - c für commodity cluster
  - m für monolithic system
- Naming
  - d für distributed
  - s für shared
  - c für cache coherent
- Parallelism
  - t für multithreading
  - v für vector
  - c für communicating sequential processes oder message passing
  - s für systolic
  - w für VLIW
  - h für producer/consumer
  - p für parallel processes

Charakterisierungsansatz nach Dongarra, Sterling, Simon, Strohmaier

- Latency
  - c für caches
  - v für vectors
  - t für multithreaded
  - m für processor in memory
  - p für parcel or message driven split-transaction
  - f für prefetching
  - a für explicit allocation

Beispiele für ausgewählte Parallelrechner

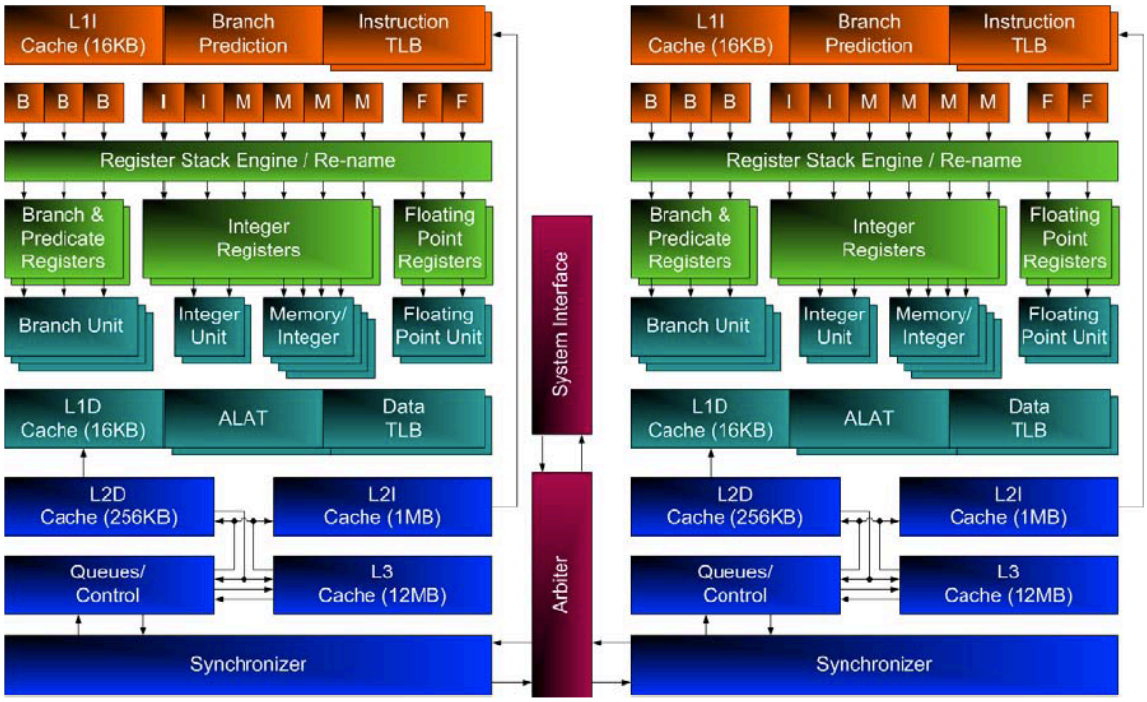
- SGI Origin: m/c/p/c
- Red Storm: m/d/c/a
- Earth Simulator: m/s/v/v

Die Community ist aufgerufen, an der Verbesserung dieses Systems mitzuarbeiten.

Gliederung

- Beispiele von Parallelrechnern
  - SGI Altix 4700
  - SGI UV
  - Megware-Cluster Atlas
  - Bull HPC Cluster Taurus
  - BlueGene
- Hochgeschwindigkeitsnetzwerke

SGI Altix 4700 - Prozessorbasis Intel Itanium II Montecito



TECHNISCHE UNIVERSITÄT DRESDEN
Wolfgang E. Nagel
ZIH Center for Information Services & High Performance Computing

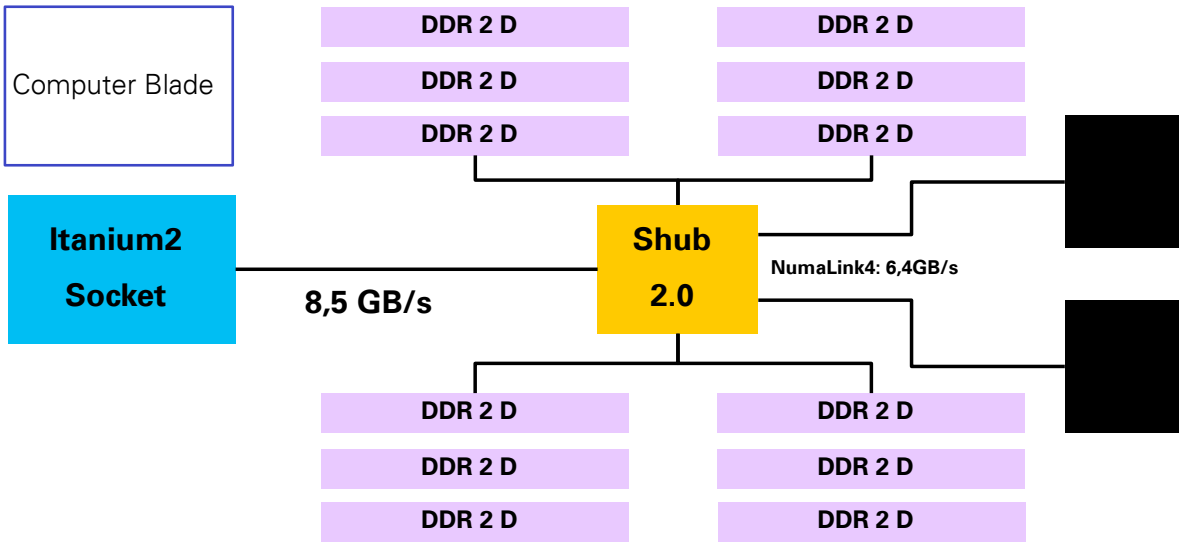
Intel Itanium Montecito

- EPIC „Explicit Parallel Instruction Computing“
- 128 Bit Befehlsbündel (VLIW)

Instruction 2 (41 Bit)	Instruction 1 (41 Bit)	Instruction 0 (41 Bit)	Template (5 Bit)

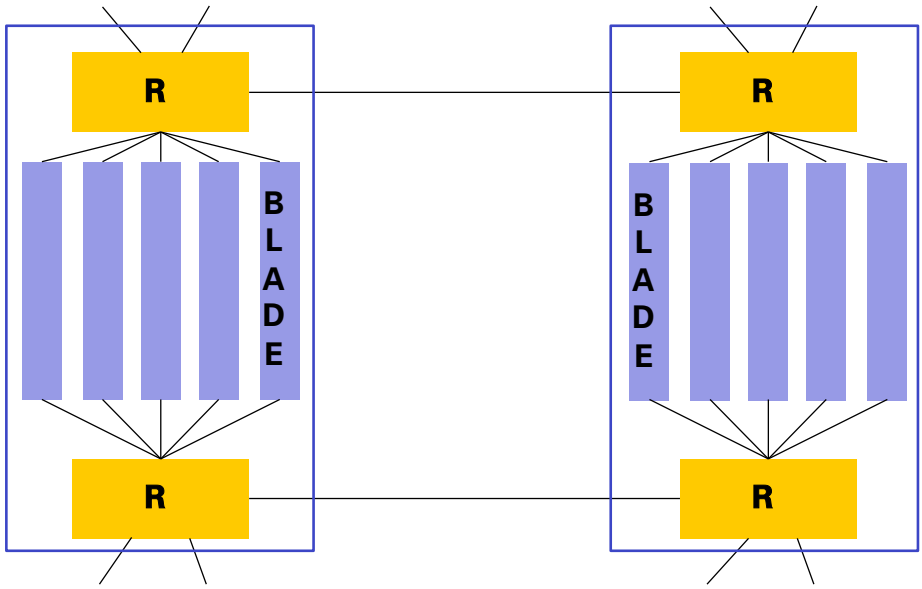
- Template liefert Informationen zur parallelen Ausführung von Befehlen
- Taktfrequenz: 1,6 GHz
- Theoretische Floating Point Peak Performance pro Core: 6,4 Gflop/s

SGI Altix 4700: Systemknoten



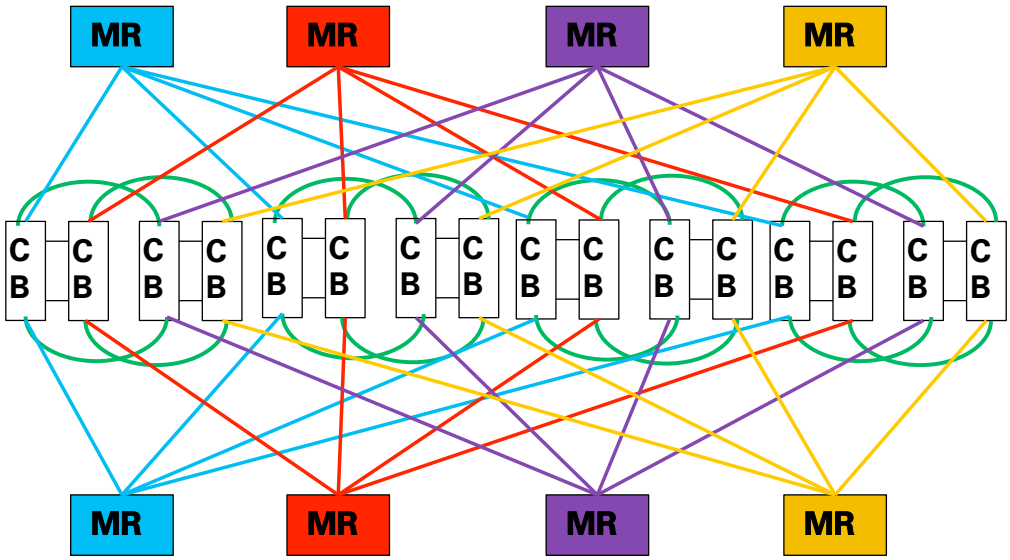
Schematische Darstellung eines Systemknotens

SGI Altix 4700: 8-Port-Router

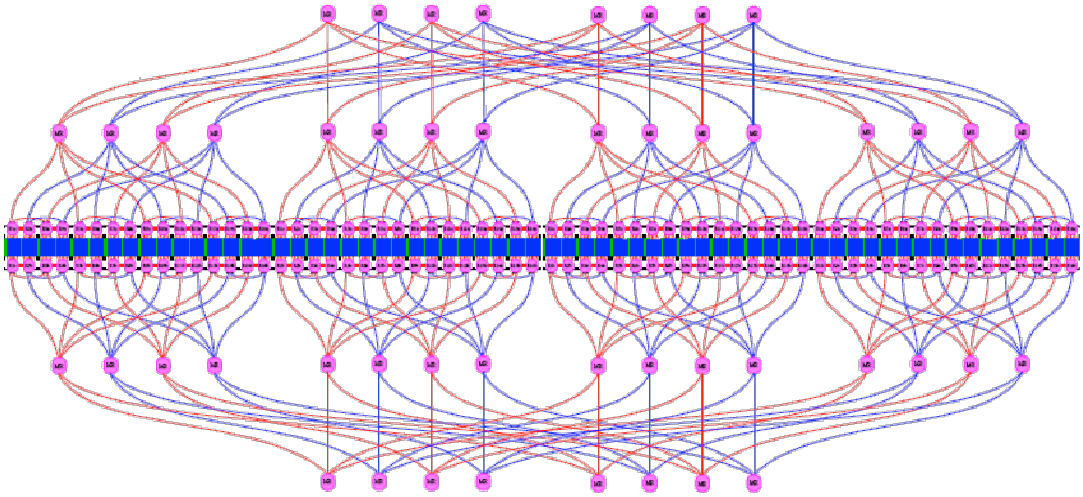


8-Port-Router verbindet jeweils vier Prozessoren-Blades mit maximal einem I/O-Blade zu einem Compute Brick (CB)

SGI Altix 4700: 128 Cores (MR –Metarouter)



SGI Altix 4700: 512 Cores



SGI Altix 4700 am ZIH

- 2048 Cores, aufgeteilt in fünf Partitionen
  - Login-Partition: Mars
    - 4 Cores Boot-CPU-Set
    - 32 Cores für Login und interaktiven Betrieb
    - 348 Cores im Batchbetrieb
    - 1 GB Hauptspeicher je Core
  - Compute-Partitionen (Jupiter, Saturn, Uranus)
    - 4 Cores Boot-CPU-Set
    - 2 Cores für das Batchsystem LSF selbst (zur Verwaltung)
    - 506 Cores im Batchbetrieb
    - 4 GB Hauptspeicher je Core
  - Interaktive Partition: (Neptun)
    - 128 Cores im interaktiven Betrieb
    - Grafik und FPGA
    - 1 GB Hauptspeicher je Core

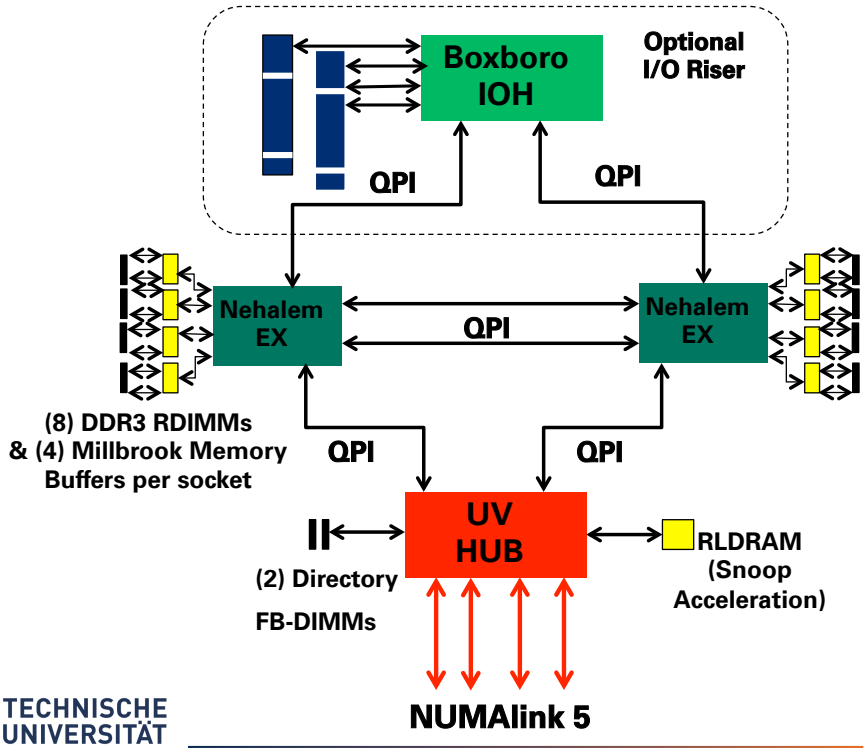
SGI UV 1000

- November 2009 auf der SC09 vorgestellt.
- Prozessorbasis ist Intel Nehalem-EX
  - Acht Kerne per Socket
  - Jeder Kern zwei 128-Bit-SSE-Einheiten
    - Für eine Taktfrequenz von 2,27 GHz → 9,08 Gflop/s pro Kern (DP)
- Zwei Sockets per Blade
  - 16 Kerne/Blade x 9,08 Gflop/s pro Kern = 145,28 Gflop/s pro Blade (DP)
- 32 Blades pro Rack
  - 4,65 Tflop/s pro Rack
- Bis zu 512 Rack lassen sich koppeln
  - 512 Racks x 32 Blades/Rack x 16 Kerne/Blade = 262 144 Kerne
  - 2,38 Pflop/s

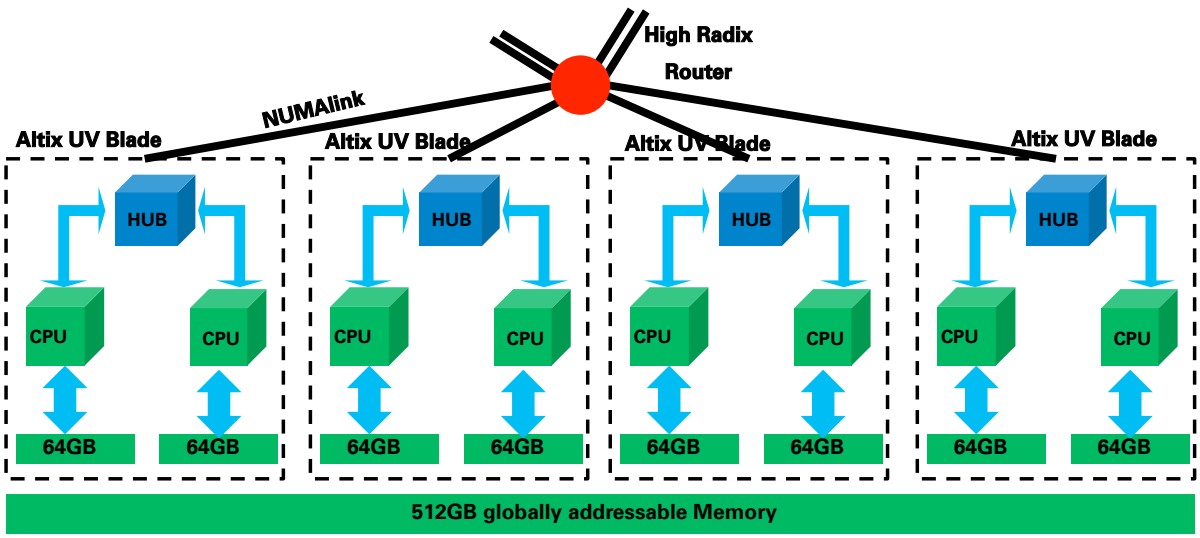
SGI UV 1000

- Skalierbar bis 2048 Kerne und 16 TB als Shared Memory
- Verbindungsnetzwerk
  - Kopplung der Blades über UV Hub
  - Verbindung zwischen UV HUBs basiert auf auf NUMALink 5
    - 15 GB/s pro Link x 4 Links = 60GB/s per Blade
- Betriebssystem Linux

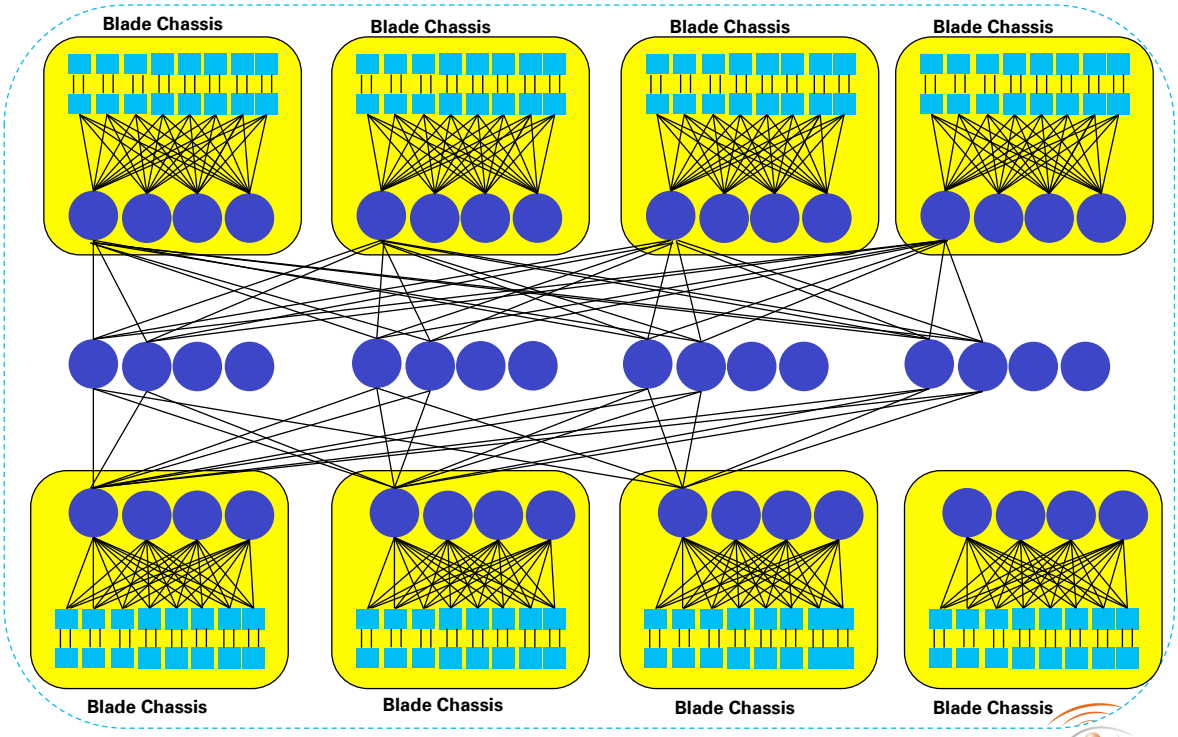
SGI UV 1000: Compute Blade



SGI UV 1000: Interconnect with Global Addressing

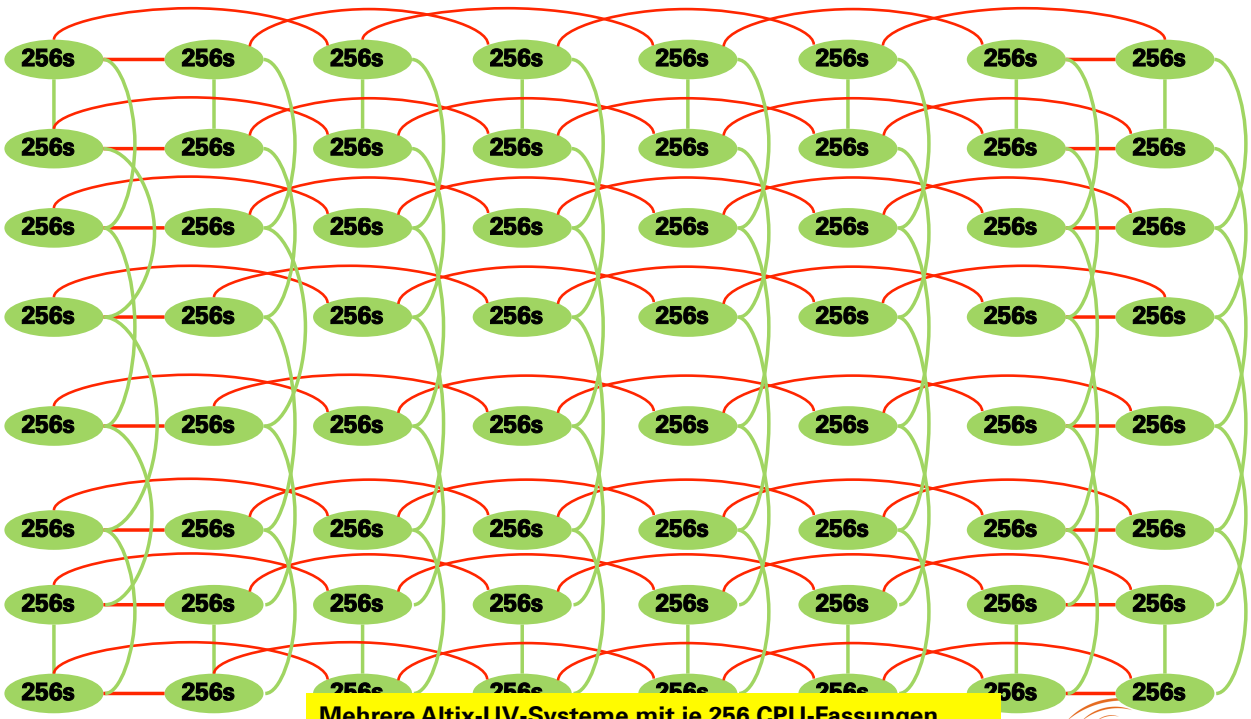


SGI UV 1000: bis 256-CPU-Fassungen (2048 Kerne) als Shared Memory



NUMALink5 koppelt 256 CPU-Fassungen auf 128 Blades in vier Racks; hier ist nur ein Viertel aller Kabel abgebildet

SGI UV 1000: 131072-Cores

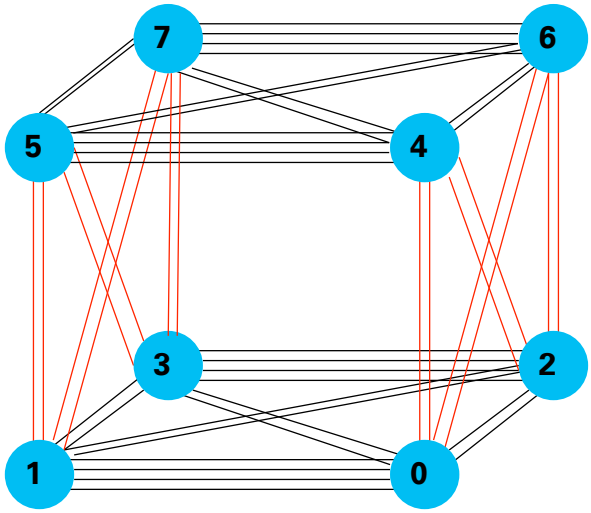


Mehrere Altix-UV-Systeme mit je 256 CPU-Fassungen lassen sich wiederum koppeln, hier zu einem System mit 16.384 Prozessoren und 131.072Kernen.

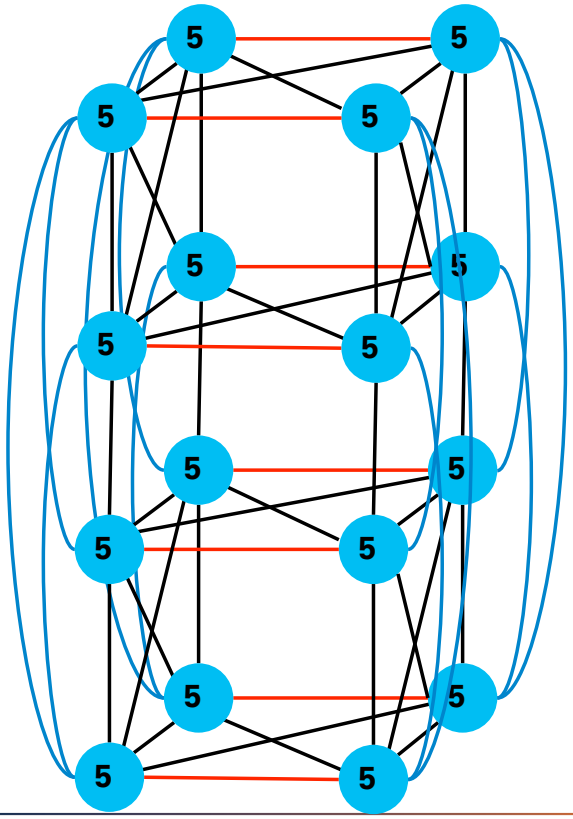
SGI UV 2000

- Juni 2012 auf der ISC12 in Hamburg vorgestellt.
- Prozessorbasis ist Intel SandyBridge
  - Acht Kerne per Socket
  - Zwei Sockets pro Blade
- Verbindungsnetzwerk
  - Kleine Systeme (16 Sockets) gibt es auch Backplane-gekoppelt
  - Mittlere Systeme können direkt über NUMALink 6 verkabelt werden
  - Größere Systeme werden über externe Router verbunden
    - Mehrere 16 Socket Systeme werden über die 8 unteren Ports der Router verbunden

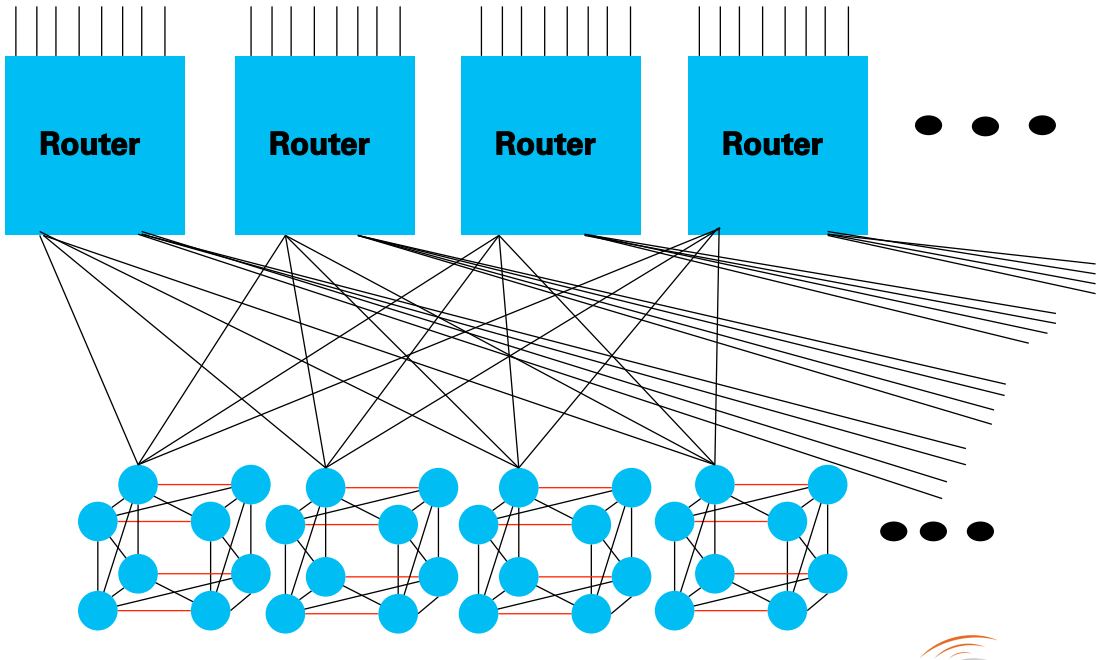
SGI UV 2000: 16 Socket Topologie



SGI UV 2000: 32 Socket Topologie



SGI UV 2000: Topologie mit externen Routern



SGI UV 2000

- Bilding-Block-Konfiguration
  - Acht 16-Socket-Systeme werden mit 32 Routern gekoppelt
  - Jeder Knoten (Blade) jedes 16-Socket-Systems ist mit vier Routern verbunden
  - Sehr große Systeme nutzen obere acht Ports der Router und verkoppeln die Bilding-Block-Konfigurationen zu verschiedenen Topologien:
    - Variante von all-to-all
    - Fat-Tree
    - Torus
- Skalierbar bis 4096 Kerne und 64 TB als Shared Memory

SGI UV 2000 am ZIH

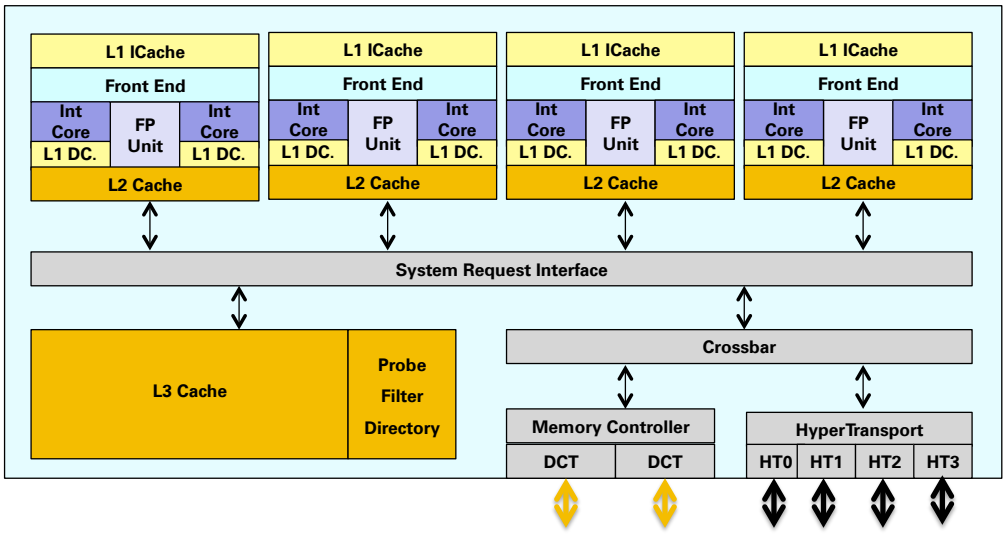
- 512 Kerne
  - 64 x 8 Kerne SandyBridge (2,6 GHz)
  - 10,6 Tflop/s Peak
- 8 TB RAM
  - 32 Blades je 256 GB
  - QPI (QuickPath Interconnect) 8 GT/s (Gigatransfers pro Sekunde)
  - DIMMs 1600 MT/s
- Enhanced HyperCube
- Ein wassergekühltes Rack
- Leistungsaufnahmen abhängig von der Konfiguration – ca. 16-20 kW
  - Turbomode 3.1 GHz
  - Power Saving Potential

Megware-Cluster „Atlas“ am ZIH

- Prozessorbasis ist: AMD Interlagos Opteron 6274
- Codename für die Mikroarchitektur des AMD Interlagos Opteron 6274 Prozessors ist „Bulldozer“
- Bulldozer beschreibt den inneren Aufbau der Module der Prozessoren
- AMD Interlagos Opteron 6274 Prozessor
  - Ein Sockel mit zwei Nodes
  - Jeder Node hat 4 Module
    - 8 (Integer-)Cores
    - 4 Floating Point Units
    - Node ist CC-NUMA
- Verbindung der zwei Nodes über HyperTransport (HT) Links
  - Max. 25,6 GB pro Sekunde über einen HT Link

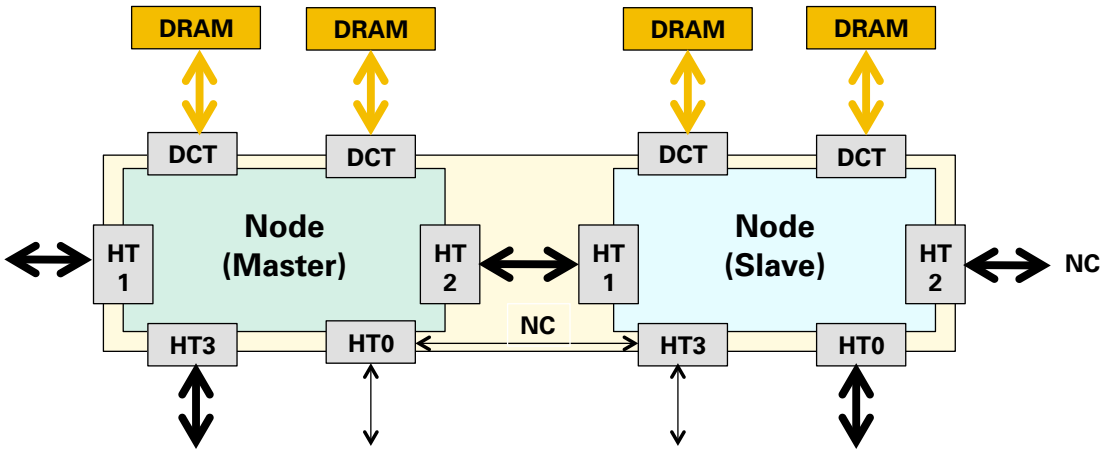
Megware-Cluster Atlas am ZIH

- Ein Node des AMD Interlagos Opteron 6274 Prozessor



Megware-Cluster Atlas am ZIH

- AMD Interlagos Opteron 6274 Prozessor als Dual-Node Package



Megware-Cluster Atlas am ZIH

- Compute-Knoten mit vier AMD Interlagos Opteron 6274 Prozessoren (563,2 Gflop/s)

Nutzerbetrieb ab März 2012

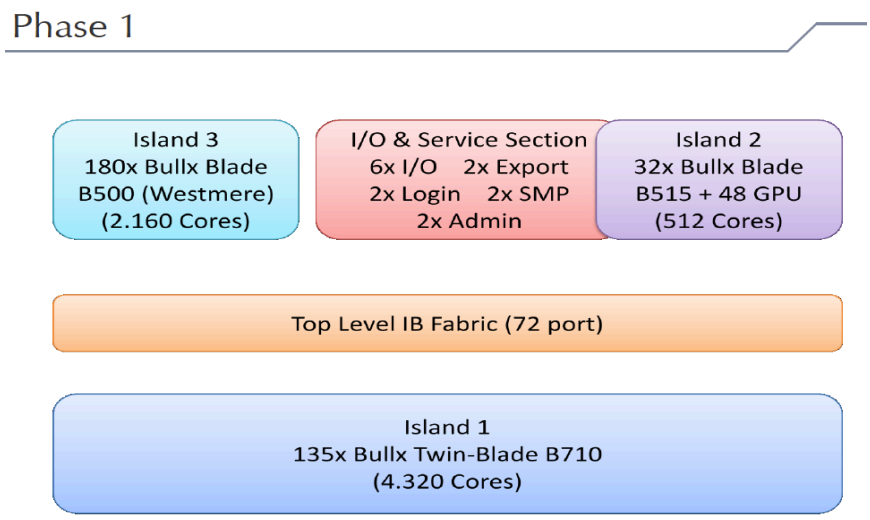
- Megware-Cluster Atlas besteht insgesamt aus
  - 92 Compute-Knoten (mit Speicher von 64 GB bis 512 GB)
    - 51,8 Tflop/s
  - 2 Login-Knoten
  - 2 Master-Knoten
- Kopplung der Knoten über Infiniband
  - Fünf QDR Leaves
  - Zwei QDR Spines

Bull HPC Cluster Taurus am ZIH

- Besteht aus drei Inseln und zentraler Komponenten
  - Insel 1: 4320 Cores Intel E5-2690 (Sandy Bridge) 2.90GHz
  - Insel 2: 704 Cores Intel E5-2450 (Sandy Bridge) 2.10GHz
    + 88 NVIDIA Tesla K20x GPUs
  - Insel 3: 2160 Cores Intel X5660 (Westmere) 2.80GHz
  - Zentrale Komponenten
    - Zwei SMP-Knoten
      - Jeder Knoten 32 Cores Intel E5-4650L (Sandy Bridge) 2.60 GHz
    - Zwei Login-Knoten
      - Jeder Knoten 8 Cores Intel E5-2670 (Sandy Bridge) 2.60GHz
    - Transfer-Knoten
      - Zwei Server

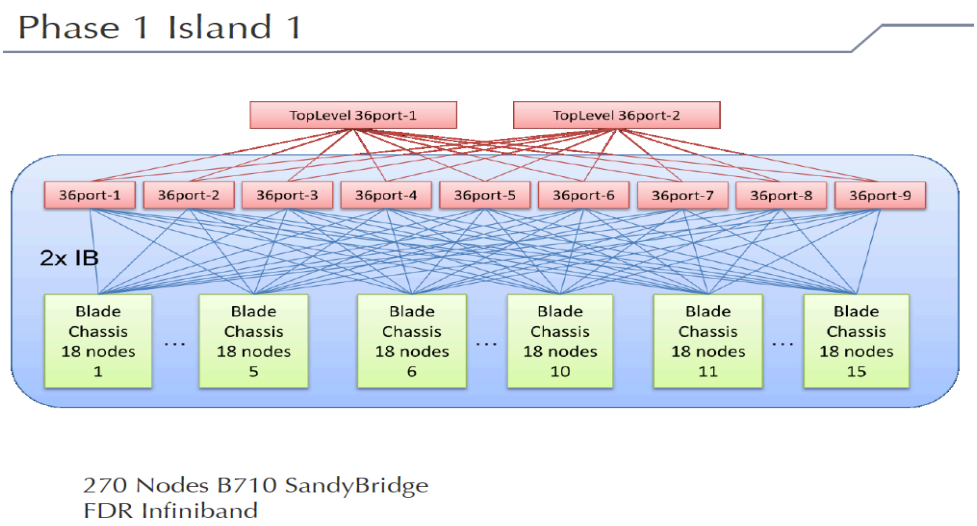
Bull HPC Cluster Taurus am ZIH

- Verbindung der Hauptgruppen über Infiniband



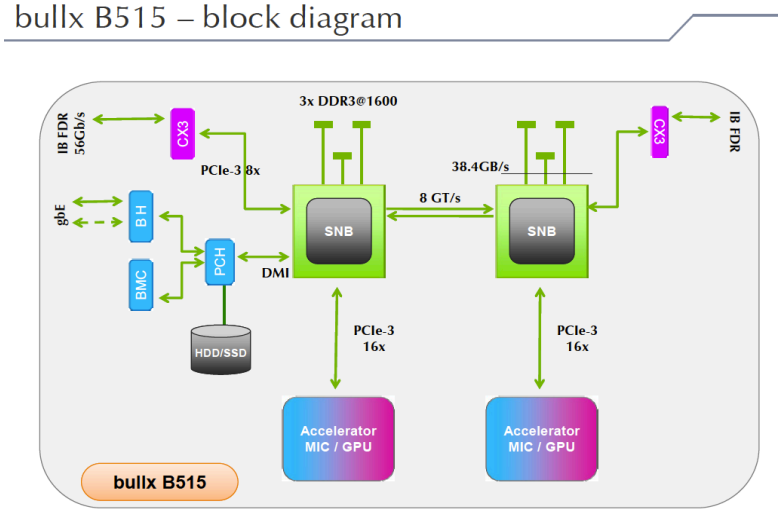
Bull HPC Cluster Taurus am ZIH

- Interne Kopplung der einzelnen Inseln über Infiniband (z.B. Insel 1)



Bull HPC Cluster Taurus am ZIH

- Basis der Inseln 1 und 2 sind Compute Blades mit zwei Prozessoren (Sandy Bridge)
- Compute Blades der Insel 2 haben Accelerator-Anschlüsse



Bull HPC Cluster Taurus am ZIH

- Floating Point Peak Performance (DP): 137 TFLOPS
- Batchsystem SLURM (Simple Linux Utility for Resource Management)
  - Realisiert Zugriff zu den Ressourcen (compute nodes)
  - Liefert ein Framework zum Starten, Ausführen und Monitoren der Arbeit auf den allokierten Cores
  - Regelt den Zugriff auf die Ressourcen durch Warteschlangen
- NVIDIA Tesla K20x Karte
  - Basis ist GPU GK110
    - 15 SMX (Streaming Multiprozessors)
    - 192 Cores pro SMX
    - 64 DP Units pro SMX
    - Weitere Details zur GPU GK110 in der ETI-Vorlesung „GPU Programmierung“

IBM BlueGene Projekt

Das BlueGene Projekt

- Projekt von IBM, das seit 1999 läuft
- Ziele
  - Entwicklung von Supercomputern im Petaflop/s-Bereich
    - MPP (Massively Parallel Processors)
  - Minimierung des Energieverbrauchs
- Haupteinsatzgebiet Biogenetik und darin speziell die Proteinfaltung
  - Davon wurde Namensgebung abgeleitet
- Architekturen des BlueGene Projektes unterscheiden sich vor allem in der Prozessorbasis
  - Erste Architektur war BlueGene/L
    - Belegte November 2004 Platz 1 der Top 500 Liste und hielt Platz 1 bis November 2007

IBM BlueGene Projekt

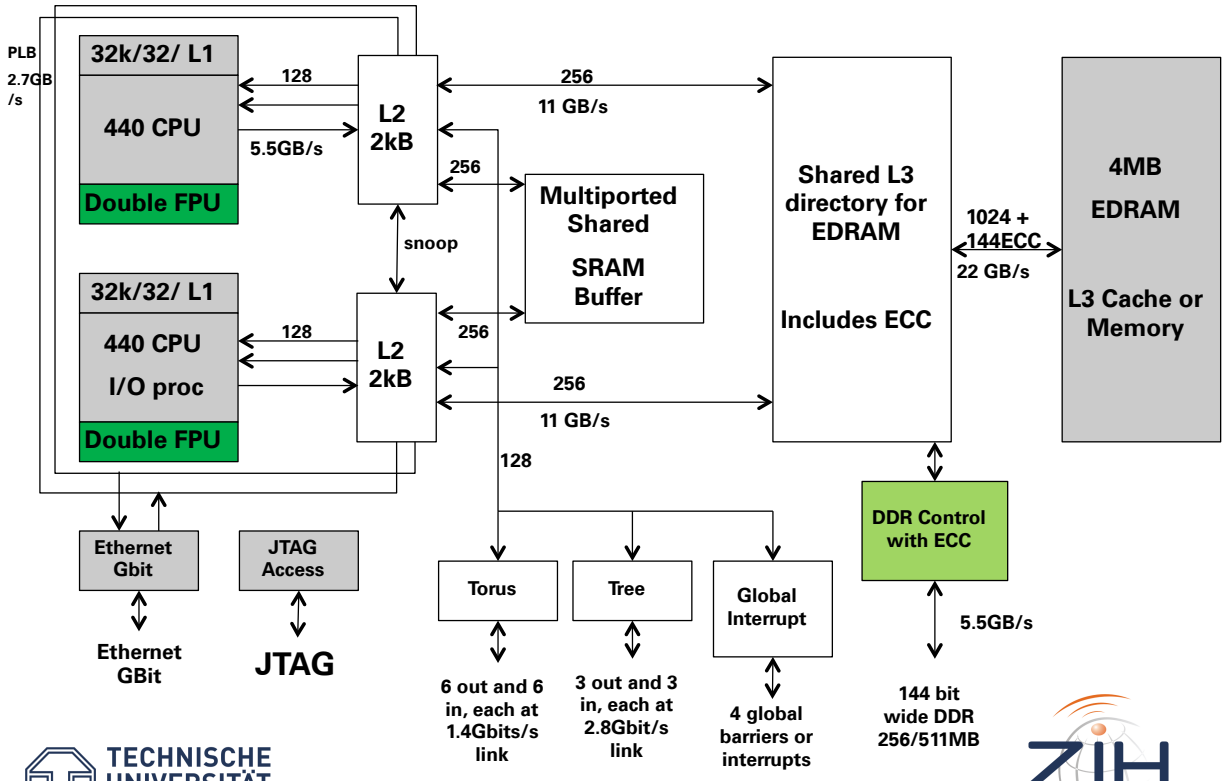
- Zweite Architektur war BlueGene/P
  - Juni 2007 auf ISC in Dresden
  - Eines der ersten Auslieferungen 2008 an das Forschungszentrum Jülich (JUGENE)
    - 180 Tflop/s, November 2008 TOP 500 Platz 11
    - Mai 2009 Ausbau auf 1 Petaflop/s
- Dritte Architektur ist BlueGene/Q
  - TOP 500 Juni 2012 Platz 1; November 2012 Platz 2
    - Rmax=16,3 Petaflop/s und
    - Rpeak=20,1 Petaflop/s

IBM BlueGene/L

Compute Chip (Knoten)

- Basis ist IBM Prozessor PowerPC 440d
  - 32-Bit-Architektur
  - 700 MHz
    - Reduzierung der Leistungsaufnahme
    - Verminderung der Lücke zwischen Prozessorperformance und Speicherperformance
  - Doppelte 64-Bit-Floating-Point-Einheit
- Zwei Prozessoren pro Compute Chip
  - Betriebsart 1: beide Prozessoren für Rechenaufgaben → 5,6 Gflop/s
  - Betriebsart 2: ein Prozessor für Rechenaufgaben → 2,8 Gflop/s
    ein Prozessor als Kommunikationsprozessor

BlueGene/L: Compute Chip (Knoten)



IBM BlueGene/L: Aufbau

Einzelne Compute Chips (Knoten) werden in mehreren Schritten in 3D-Torus Topologie zum Gesamtsystem verbunden

- Compute Card
  - Zwei Knoten mit 4 Prozessoren und zwei 512 MByte DDR-Speichern
  - Max. 11,2 Gflop/s bei 15 Watt, 1x2x1 Torus
  - I/O-Card hat gleichen Aufbau wie Compute Card
    - Gigabit-Ethernet Schnittstelle zur I/O-Kommunikation in jedem Knoten vorhanden aber nur im I/O-Knoten genutzt
- Node Card
  - 16 Compute Cards, 32 Knoten, 4x4x2 Torus
- Rack
  - Erste eigenständige abgeschlossene Einheit
  - 32 Node Cards, 1.024 Knoten, 2.048 Prozessoren
  - 8x8x16 Torus

IBM BlueGene/L: Aufbau

- Mindesten 8 und maximal 64 I/O-Cards
- Theoretische Floating Point Peak Performance 5,6 Tflop/s, 512 GBytes Speicher
- 15 bis 20 kW im Betrieb, maximal 27,6 kW

- Maximalausbau
  - 64 Racks, 65.536 Knoten, 131.072 Prozessoren
  - 64x32x32 Torus
  - Theoretische Floating Point Peak Performance 360 Tflop/s

IBM BlueGene/L: Verbindungsnetzwerk

Netzwerk der BlueGene/L setzt sich aus fünf unabhängigen Einzelnetzwerken zusammen:

- 3D-Torus Netzwerk
- Kollektives Netzwerk
- Barriere Netzwerk
- Gigabit-Ethernet
- Kontrollnetzwerk

Netzwerklogik und die Schnittstellen vollständig auf Compute Chip (Knoten) integriert

- Kommunikation müssen die beiden Prozessoren selbst durchführen
- Für kommunikationsintensive Anwendungen kann ein Prozessor des Knotens direkt als Koprozessor für Kommunikation genutzt werden

IBM BlueGene/L: Verbindungsnetzwerk

3D-Torus

- Pro Rack zwei 8x8x8 Toren (zweimal 512 Knoten)
- Jeder Torus mit Midplane verbunden
  - 24 Link-Chips pro Midplane
    - Verstärken Signale
    - Verbinden Racks miteinander
- Zwischen benachbarten Knoten
  - Hardwarewarelatenz 69 ns, Gesamtlatenz 100ns
  - Worst-Case für Maximalkonfiguration 64x32x32 Torus
    - Nachricht muss 64 Knoten (32+16+16) durchlaufen
    - Latenz 6,4 μ s
- Bandbreite 175 MByte/s in jede Richtung zu allen 6 Nachbarn

  = 2,1 GByte/s pro Knoten
- Virtual cut-through Routing (Paketvermittlung)

IBM BlueGene/L: Verbindungsnetzwerk

Kollektives Netzwerk

- Baumstruktur
  - 10 Ebenen für Maximalausbau mit 65.536 Knoten
    - Effektiver Broadcast von einem Knoten an alle anderen Knoten 20 μ s
  - Punkt-zu-Punkt Verbindungen möglich, die es Knoten gestattet mit seinem I/O-Knoten zu kommunizieren

Barriere Netzwerk

- Besteht hauptsächlich aus vier unabhängigen Interruptkanälen
  - Verlaufen parallel zum Global Tree
  - Komplett getrennt vom Global Tree und allen anderen Netzwerken
  - Für Maximalausbau mit 65.536 Knoten werden alle Knoten in 1,5 μ s erreicht

IBM BlueGene/L: Verbindungsnetzwerk

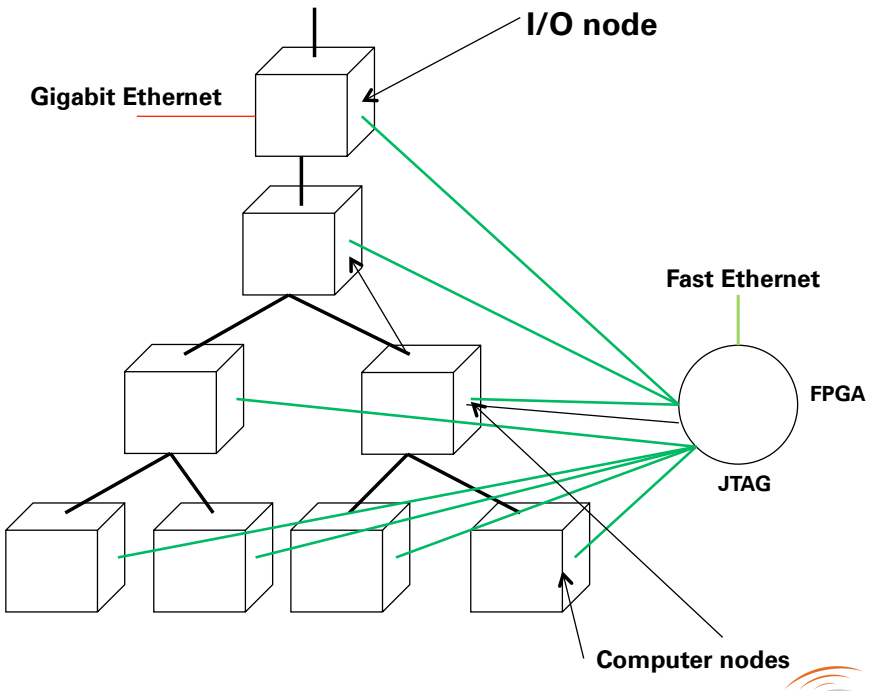
Gigabit-Ethernet

- Gigabit-Ethernet Schnittstelle in jedem Knoten vorhanden aber nur vom I/O-Knoten genutzt
  - Verbindung zu externen Host und externen Filesystem
- Normalerweise greifen 64 Rechen-Knoten über einen I/O-Knoten auf das Dateisystem zu
- Anzahl der I/O-Knoten kann so erhöht werden, dass 8 Rechenknoten pro I/O-Knoten zugeordnet werden können

Kontrollnetzwerk

- Eigenständiges Netzwerk zur Überwachung der Systemkomponenten
- Besonderheit ist JTAG-Interface in jedem Knoten
- JTAG (Joint Test Action Group) ist ein Standard, der entwickelt wurde, um von außen auf integrierte Schaltkreise zugreifen zu können
- Service-Knoten greift über Gigabit Ethernet auf einem FPGA zu, der die Ethernet-Pakete in JTAG-Signale konvertiert
  - Service-Knoten kann auf jeden Knoten des Systems zugreifen

BlueGene/L: Kontrollnetzwerk



IBM BlueGene/P

- Grundlegender Aufbau von BlueGene /L übernommen
  - Basis ist IBM Prozessor PowerPC 450
    - Leicht erhöhte Taktfrequenz von 850 MHz
    - Vier Prozessoren (statt zwei) auf einem Compute Chip (Knoten) →13,6 Gflop/s
  - Nur ein (statt zwei) Compute Chip auf einer Compute Card
    - Preiswerterer Austausch von fehlerhaften Komponenten
  - 32 Compute Cards (statt 16) auf einer Node Card
  - 32 Node Cards pro Rack (wie bei BluGene/L)
  - Maximalausbau 72 Racks, 294.912 Prozessoren, 1Petaflop/s

IBM BlueGene/Q

- Aktuellste BlueGene-Architektur
- Neuer Prozessor „PowerA2“
  - 18 Kerne
    - 16 Kerne zum Rechnen
    - ein Kern für Kontroll- oder I/O-Aufgaben
    - ein Reservekern, um Ausbeute zu erhöhen
  - Theoretische Floating Point Peak Performance pro Kern
    - Vier doppeltgenaue Fused-Multiply-Add-Befehle →8 Flop pro Takt
    - 1,6 GHz → 12,8 Gflop/s
  - Theoretische Floating Point Peak Performance pro Prozessor
    - 16 Rechenkerne x 12,8 Gflop/s = 204,8 Gflop/s
- Compute Card enthält einen Prozessor
- Node Card enthält 32 Compute Cards → 6,55 Teraflop/s
- Rack enthält 32 Node Cards → 209,7 Teraflop/s
- 96 Racks → 20 Petaflop/s

Gliederung

- Hochgeschwindigkeitsnetzwerke

Kennwerte

Bestimmung der Kommunikationsleistung über

- Kommunikationslatenz
- Bandbreite, Symbolrate, Datenübertragungsrate oder Datendurchsatz
- Overhead

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

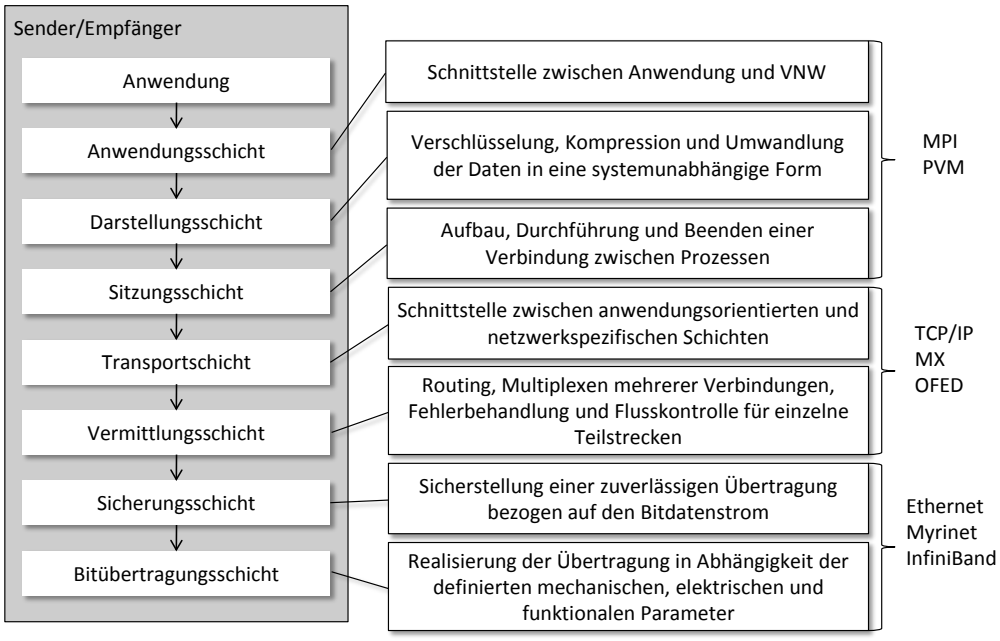


Abbildung: Open Systems Interconnection Reference Model

Entstehung von Latenz II

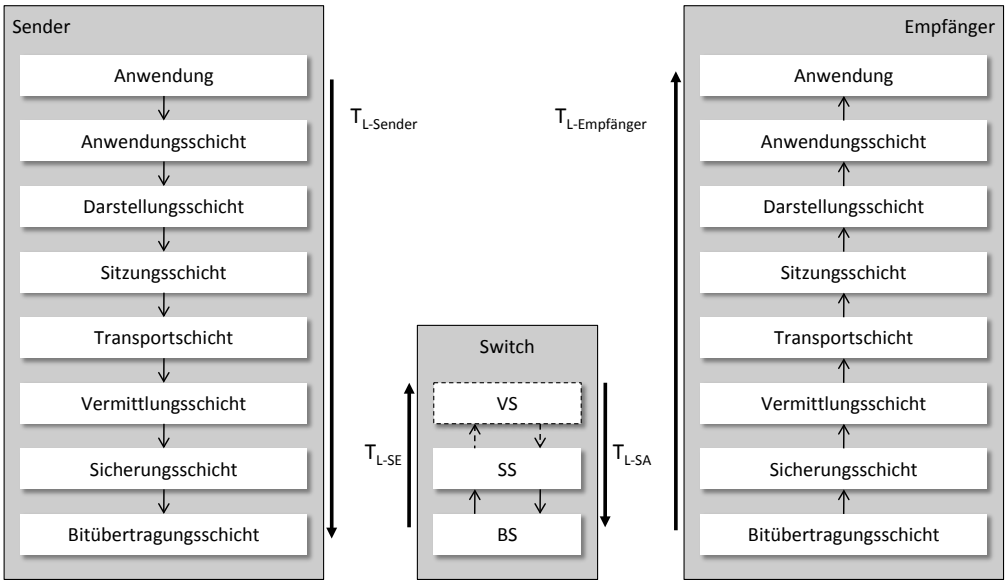


Abbildung: Zusammensetzung der Latenzzeit innerhalb einer Fabric

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

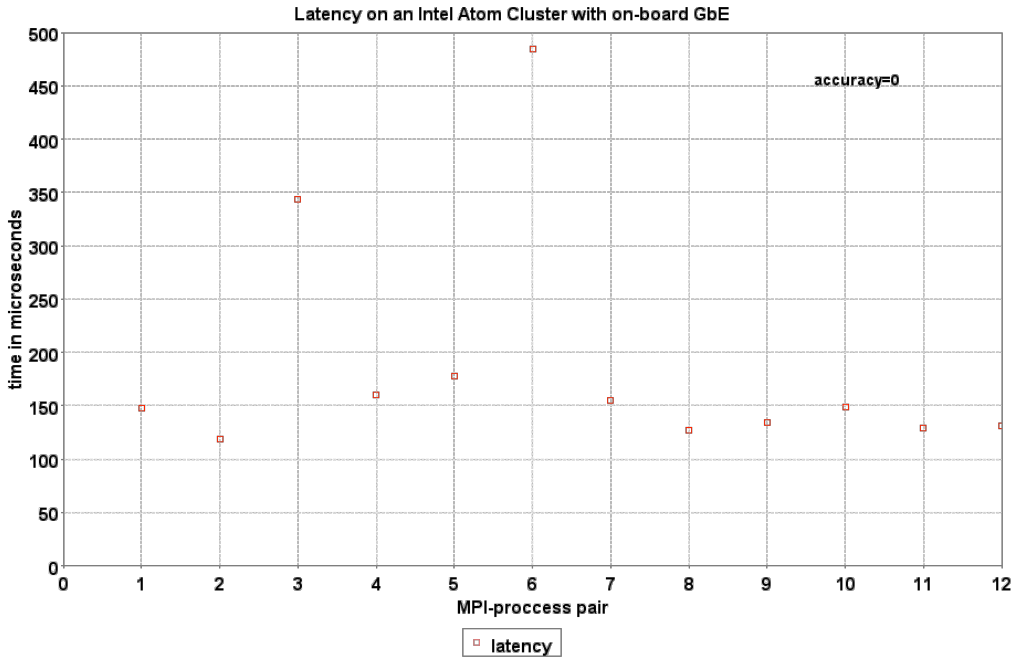


Abbildung: Einfache Latenzzeitmessung mit vier Prozessen

Beispiel Latenzzeitmessung II

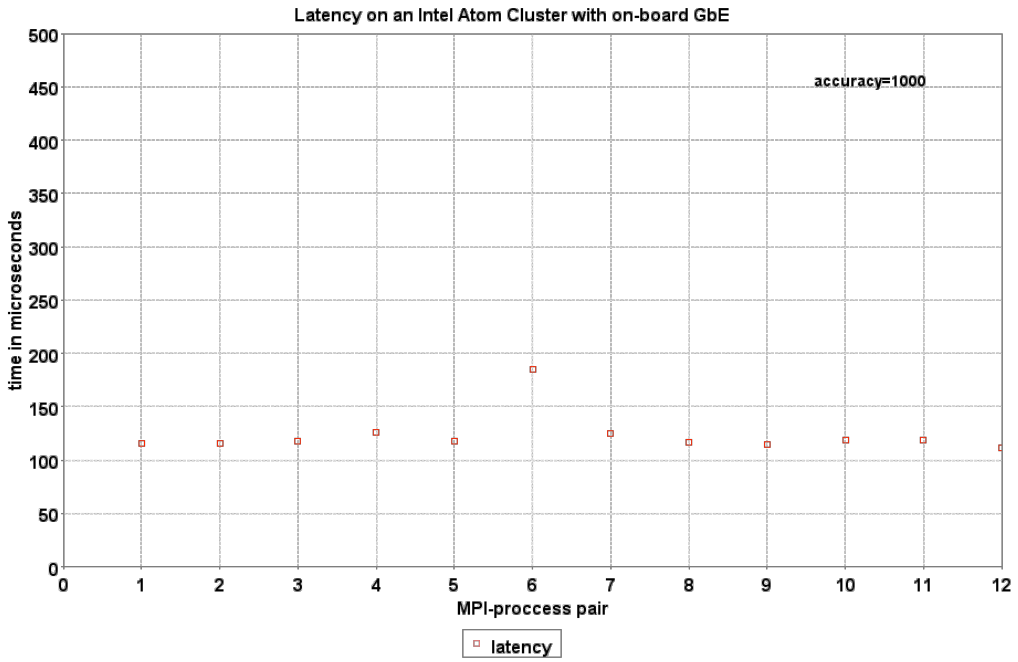


Abbildung: Erhöhung der Qualität durch Mehrfachmessungen

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 der Symbolrate und dem Informationsgehalt eines Symbols. 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.

Messung des Datendurchsatzes

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

$$DD = D_{user} / (2 * T_{comm})$$

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

Beispiel Datendurchsatzmessung

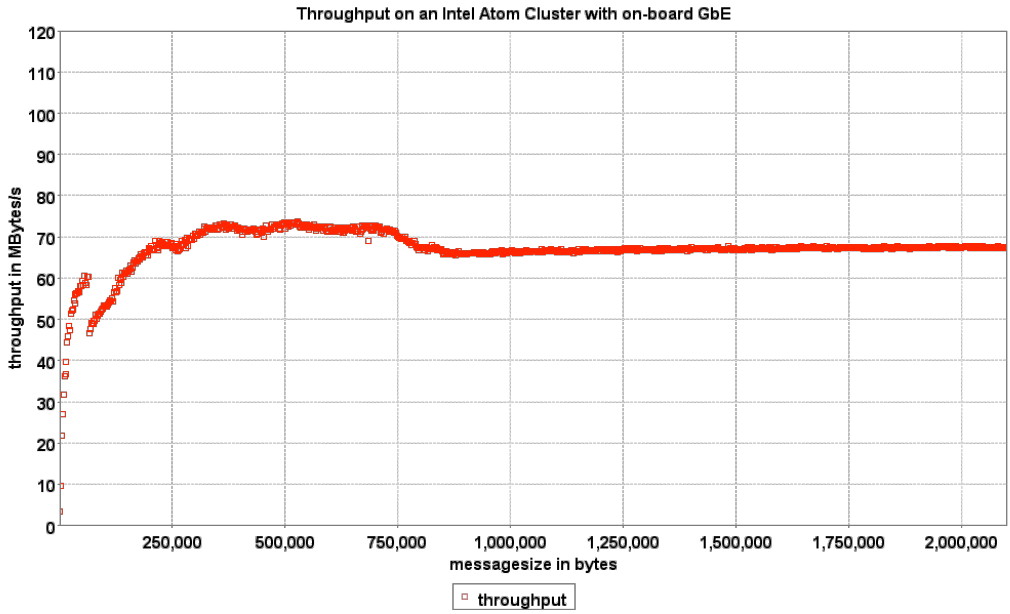


Abbildung: Datendurchsatz über MPI auf Gigabit Ethernet

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. ,,Packet Sniffern"
- Beispiel - Ethernet mit TCP/IP-Stack:

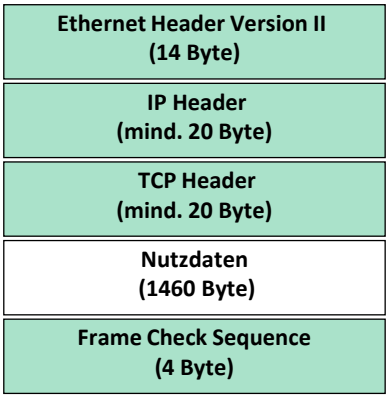


Abbildung: Protokoll-Overhead von Ethernet mit TCP/IP

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

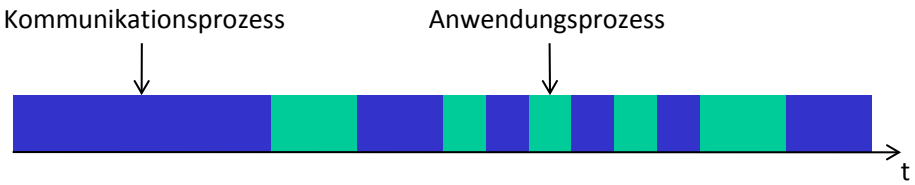


Abbildung: Kontextwechsel bei Kommunikation ohne Hardwareunterstützung

Kommunikationsprozessoren I

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

Kommunikationsprozessoren II

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

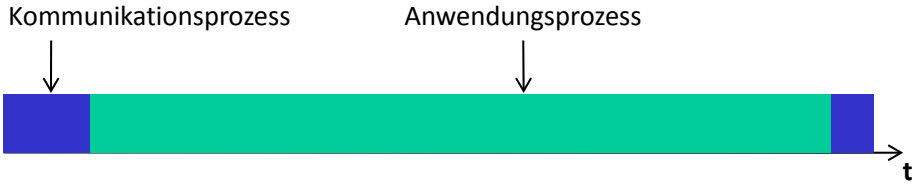


Abbildung: Kommunikationsverlauf beim Einsatz von Kommunikationsprozessoren

Zusammenhänge

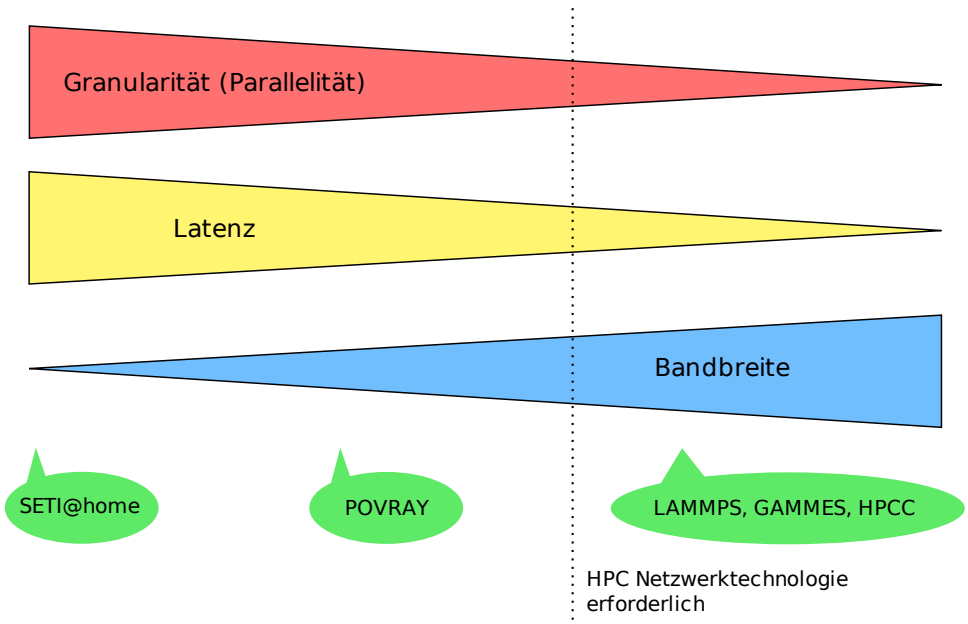


Abbildung: Zusammenhänge zwischen Granularität, Latenz und Bandbreite