TECHNISCHE UNIVERSITÄT CHEMNITZ

MEGWARE COMPUTER CLUSTER

High Performance Computing

Jürgen Gretzschel

MEGWARE

1990 gegründet

Bis zu 40 ComputerShops in Deutschland

Produktionskapazität von über 2000 PC pro Tag

Enge Beziehungen zu Komponentenherstellern

Sehr erfolgreich bis Ende der 90er Jahre

Während all dieser Jahre auch tätig als Lieferant für die Öffentliche Hand, Universitäten, Forschungsinstitute, Schulen und die kleine und mittelständige Industrie

Seit 1997 steigende Konzentration auf Serversysteme und vernetzte Computer

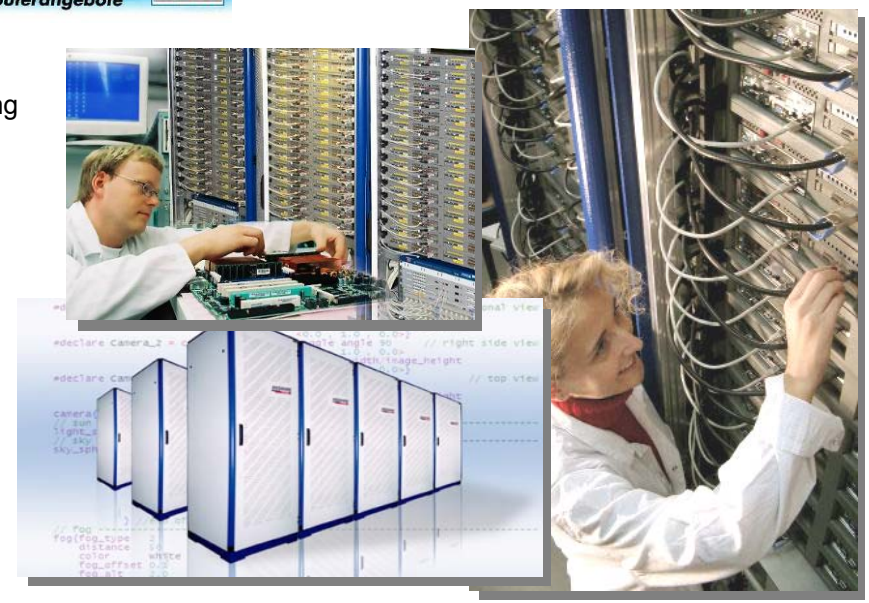
2000 erstes Cluster (CLiC)

Hauptgeschäftsfeld seit 2000

LINUX / Windows Clusterslösungen für das High Performance Computing

Schlüsselfertige Lösungen

Eigene Hard- und Softwareentwicklungen

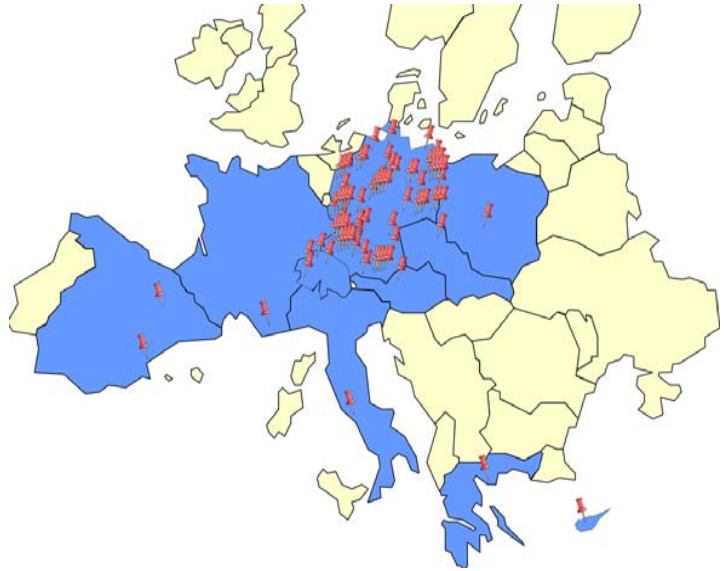


MEGWARE

AEI – Potsdam / Golm

Helics
Ruprecht-Karls-Universität Heidelberg

flight case



CLiC
Technische Universität Chemnitz

Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen

Logistikcenter in Chemnitz seit 1998

Agenda

- **Was ist High Performance Computing (HPC)**
- **Funktionsprinzipien des parallele Rechnens**
- **Anwendungsgebiete des parallelen Rechnens**
- **technologische Besonderheiten im Hochleistungsrechnen**
- **Forschung und Entwicklung von HPC-Systemen in Chemnitz**
- **Lösungen und Anwendungen für das High Performance Computing Ausgewählte Anwendungsmöglichkeiten**
  - **Erdsystemforschung – „Klimaforschung“**
  - **Neandertaler und hierarchische Matrizen**
- **Technologien im Supercomputing – das Handwerkszeug eines HPC Ingenieurs**
- **Widrigkeiten und offene Probleme**
- **Berufliche Zukunft in Chemnitz**

Was ist High Performance Computing ?

High Performance Computing (HPC) ist das computergestützte Hochleistungsrechnen.

Typische Merkmale von Hochleistungsrechnern:

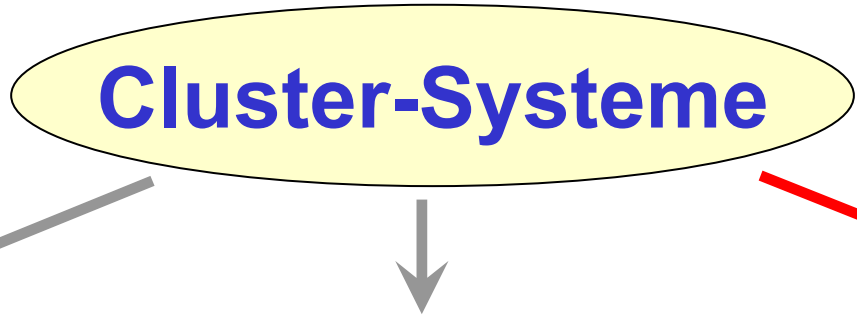
- große Anzahl Prozessoren
- parallele Verarbeitung von Rechenalgorithmen
- schnelle Netzwerke, spezielle Topologien
- u.U. gemeinsamer Zugriff auf Peripheriegeräte
- shared / distributed memory Systeme
- hohe Packungsdichte → Kühlung

Earth Simulator (Japan)

Gliederung formal in drei Bereiche:

1. HPC (High Performance Computing)
   so viele FLOPs wie möglich über eine kurze Zeit (bspw. Seconds)
2. HTC (High Troughput Computing)
   so viele FLOPs wie möglich über eine lange Zeit (bspw. Monate, Jahre)
   Beispiel: Pipelining Henry Ford
3. HAC (High Availabillity Computing)

Hochleistungsrechner / Supercomputer



Hochverfügbarkeitscluster
(High Availability)

- hohe Verfügbarkeit
- sehr geringe Ausfallzeiten
- redundante Nodes
- Gegenseitige Ersetzbarkeit
- Anwendung in fehler-kritischen Umgebungen, max. Ausfallzeit wenige Minuten/Jahr (z.B. Banken)

Load Balancing Cluster
(Load Balance)

- Lastverteilung auf mehrere Nodes
- redundante Hardware
- Anwendung bei hoher Anfragedichte der Clients (z.B. Suchmaschinen)
- Virtualisierung

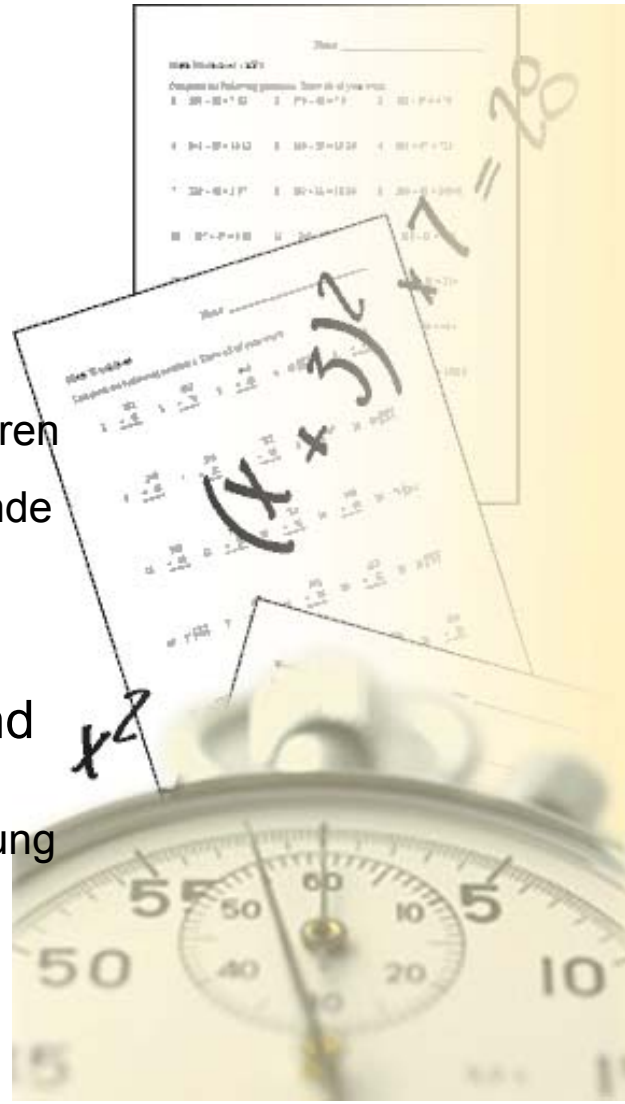
Hochleistungscluster
(High Performance)

- hohe Rechenleistung
- bis zu mehreren tausend Nodes
- parallele / verteilte Abarbeitung einzelner Jobs
- Anwendung in Forschung und Industrie

Supercomputer und Rechenleistung

Wie wird die Rechenleistung gemessen?

- Benchmark-Software
- im HPC-Bereich mit Linpack (Lösen von linearen Gleichungssystemen)
- Schwerpunkt ist reine die Rechenleistung der Prozessoren
- Ergebnis wird in Gleitkommazahloperationen pro Sekunde angegeben FLOPS



FLOPS = FLoating Point OPerations Per Second

- TOP500 listet Supercomputer nach ihrer Linpack-Leistung
- *www.top500.org* listet seit Superrechner nach Ihrer Linpack Leistung
- *Linpack* ist ein Benchmarkprogramm mit Beispielaufgaben der Linearen Algebra (http://www.netlib.org/benchmark/hpl)

Supercomputer und Rechenleistung

Top500-Liste

- Rangliste der 500 weltweit leistungsfähigsten Supercomputer
- gegründet 1986 von Prof. Dr. Hans-Werner Meuer (Uni Mannheim)
- wird 2x jährlich veröffentlicht, Juni (ISC D'land) und November (SC USA)
- Leistung der Supercomputer wird mit Linpack-Benchmark ermittelt
- Linpack-Benchmark beruht auf der Lösung linearer Gleichungssysteme
- Ergebnis wird in Gleitkommaoperationen pro Sekunde angegeben (FLOPS)

www.top500.org

Entwicklung der HPC-Rechenleistung

Historie der FLOPs

Year	Supercomputer	Peak speed	Location
1942	Atanasoff-Berry Computer (ABC)	30 OPS	Iowa State University, Ames, Iowa, USA
	TRE Heath Robinson	200 OPS	Bletchley Park
1944	Flowers Colossus	5 kOPS	Post Office Research Station, Dollis Hill
1946	UPenn ENIAC	100 kOPS	Aberdeen Proving Ground, Maryland, USA
1954	IBM NORC	67 kOPS	U.S. Naval Proving Ground, Dahlgren, Virginia, USA
1956	MIT TX-0	83 kOPS	Massachusetts Inst. of Technology, Lexington, Massachusetts, USA
1958	IBM AN/FSQ-7	400 kOPS25	U.S. Air Force sites across the continental USA and 1 site in Canada
1960	UNIVAC LARC	250 kFLOPS	Lawrence Livermore National Laboratory, California, USA
1961	IBM 7030 "Stretch"	1.2 MFLOPS	Los Alamos National Laboratory, New Mexico, USA
1964	CDC 6600	3 MFLOPS	Lawrence Livermore National Laboratory, California, USA
1975	Burroughs ILLIAC IV	150 MFLOPS	NASA Ames Research Center, California, USA
1976	Cray-1	250 MFLOPS	Los Alamos National Laboratory, New Mexico, USA (80+ sold worldwide)
1981	CDC Cyber 205	400 MFLOPS	(numerous sites worldwide)
1983	Cray X-MP/4	941 MFLOPS	Los Alamos National Laboratory
1984	M-13	2.4 GFLOPS	Scientific Research Institute of Computer Complexes, Moscow, USSR
1985	Cray-2/8	3.9 GFLOPS	Lawrence Livermore National Laboratory, California, USA
1989	ETA10-G/8	10.3 GFLOPS	Florida State University, Florida, USA
1990	NEC SX-3/44R	23.2 GFLOPS	NEC Fuchu Plant, Fuchu, Japan
1993	Thinking Machines CM-5/1024	65.5 GFLOPS	Los Alamos National Laboratory; National Security Agency
	Intel Paragon XP/S 140	143.40 GFLOPS	Sandia National Laboratories, New Mexico, USA
1994	Fujitsu Numerical Wind Tunnel	170.40 GFLOPS	National Aerospace Laboratory, Tokyo, Japan
1996	Hitachi SR2201/1024	220.4 GFLOPS	University of Tokyo, Japan
	Hitachi/Tsukuba CP-PACS/2048	368.2 GFLOPS	Center for Computational Physics, University of Tsukuba, Tsukuba, Japan
1997	Intel ASCI Red/9152	1.338 TFLOPS	Sandia National Laboratories, New Mexico, USA
2000	IBM ASCI White	7.226 TFLOPS	Lawrence Livermore National Laboratory, California, USA
2002	NEC Earth Simulator	35.86 TFLOPS	Earth Simulator Center, Yokohama-shi, Japan
2004	IBM Blue Gene/L	70.72 TFLOP	SU.S. Department of Energy/IBM, USA
2005		136.8 TFLOPS	U.S. Department of Energy/U.S. National Nuclear Security Administration,
		280.6 TFLOPS	Lawrence Livermore National Laboratory, California, USA
2008	IBM Roadrunner	1 PFLOPS	National Laboratory Los Alamos

System mit 2 CPUs
INTEL XEON, 3,2 GHz, 2x(quad core)
→ 8 cores

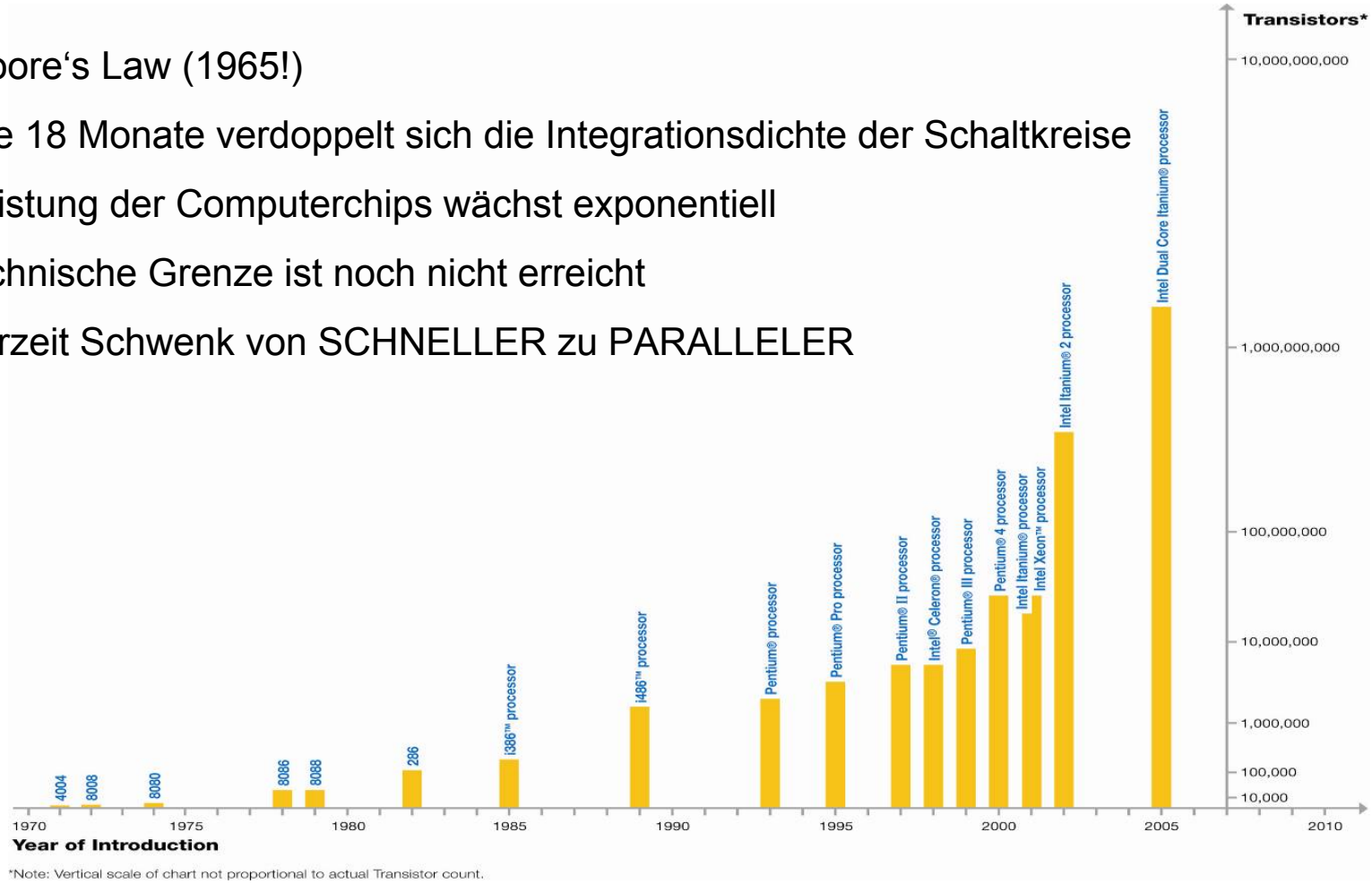
⇨ 102,4 GFLOPS

Entwicklung der HPC-Rechenleistung

Entwicklung der Rechenleistung *(Moore's Law):* Die Leistung von Computerchips wächst exponentiell, Verdopplung aller 18 Monate der Funktionen / Chip

Entwicklung der Schaltkreistechnik

- Moore‘s Law (1965!)
- alle 18 Monate verdoppelt sich die Integrationsdichte der Schaltkreise
- Leistung der Computerchips wächst exponentiell
- technische Grenze ist noch nicht erreicht
- derzeit Schwenk von SCHNELLER zu PARALLELER

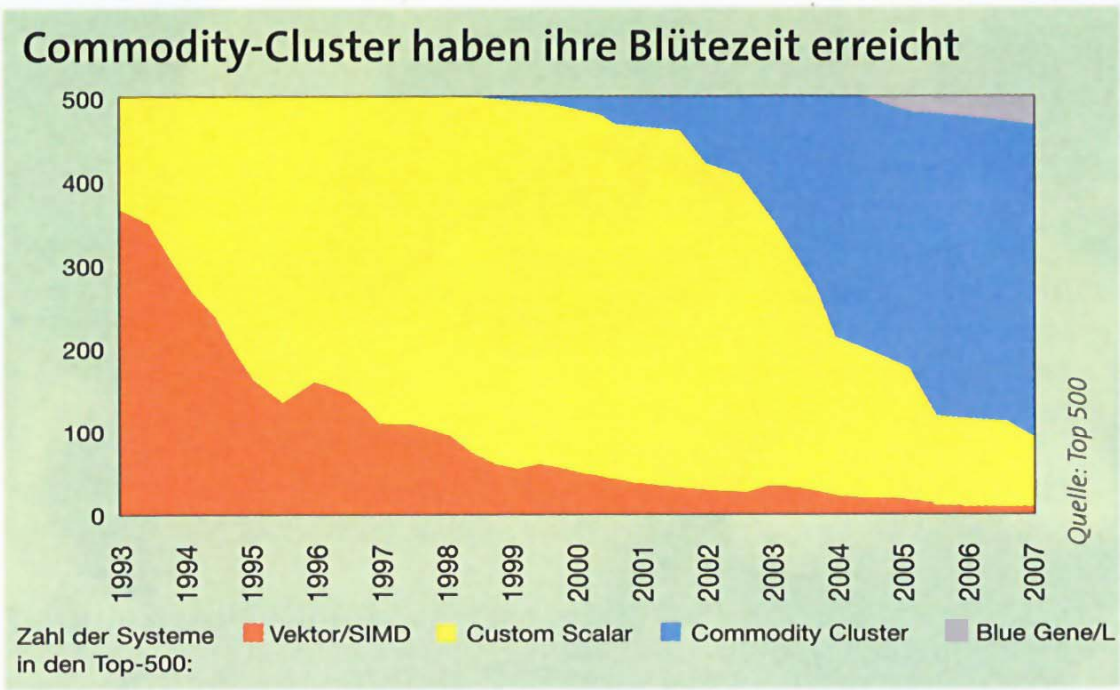


Quelle: http://download.intel.com/pressroom/images/events/moores_law_40th/Microprocessor_Chart.eps

Entwicklung hinsichtlich der Architekturen

Bell's Law

- beschreibt wie sich Computerklassen bilden, entwickeln und vergehen:
- etwa alle 10 Jahre neue Klasse



Ab Mitte der 90er Jahre begannen sich Commodity-Cluster langsam im Top-500-Ranking der Supercomputer auszubreiten – jetzt beherrschen sie drei Viertel der Liste. Mit Blick auf Bells Gesetz von der Bildung der Computerklassen könnte IBMs Blue-Gene-Architektur eine ähnliche Erfolgsgeschichte einläuten.

Quelle: Computerzeitung 38. Jahrgang Nr. 26 25.Juni2007, Seite4

Architektur der Supercomputer

… wurden bis Mitte der 1990er Jahre überwiegend als Vektorrechner konzipiert
z.B. CRAY-1, CRAY-2, Illiac-IV, CDC Star-100, TI ASC, NEC Earth Simulator

Vektorrechner:

- spezielle Prozessoren
- SIMD (*Single Instruction Multiple Data*) Prinzip (Taxonomie nach Flynn)
- führen gleichzeitig viele Berechnungen in einem Array-Prozessor aus
- jeder Prozessor hat mehrere Vektor-Register
- In den „alten Systemen“ z.B. bis zu 64 Werte gleichzeitig bearbeitbar
- können Vektoren direkt verarbeiten, ADD, MUL,…

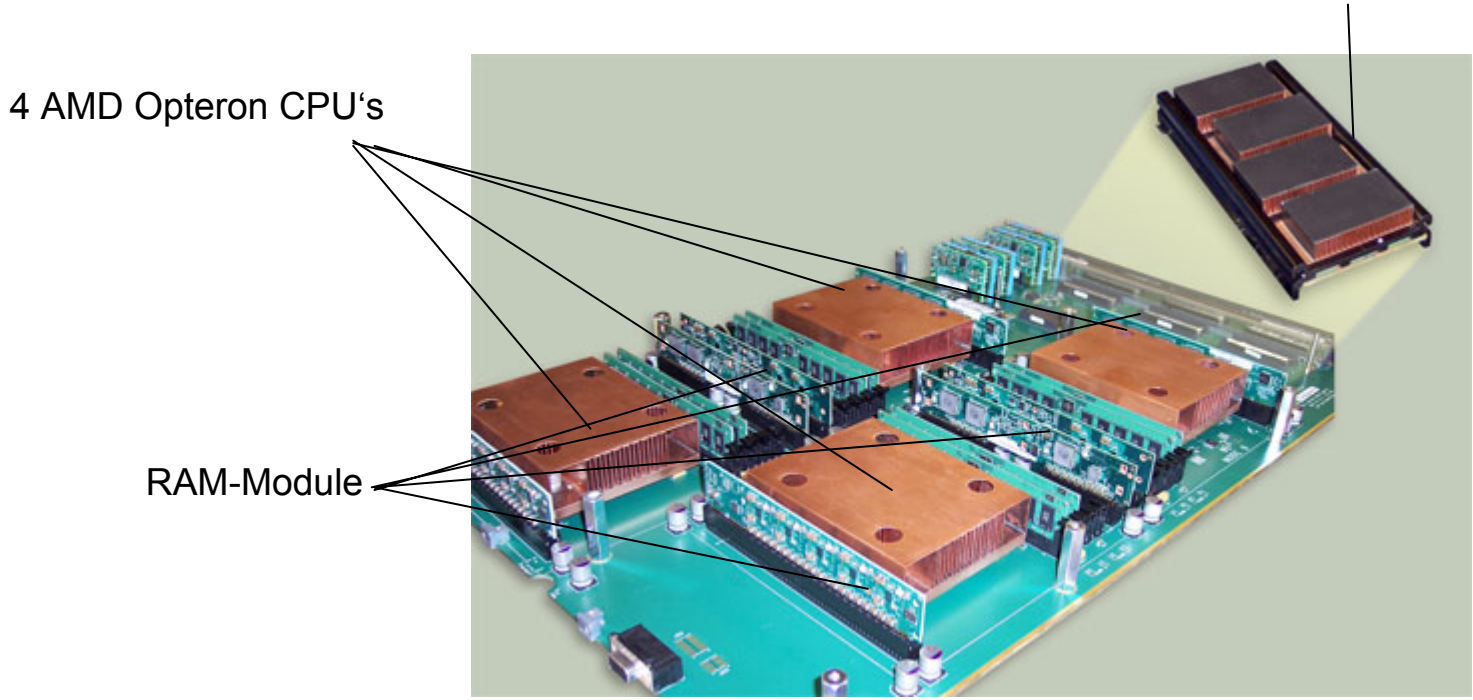
Das Prinzip findet sich aber auch in modernen CPUs wieder:

→ SSE (*Streaming SIMD Extensions)*

Cray-2 (ca.1985)

Architektur der Supercomputer

Beispiel: CRAY XT3 Prozessorelement (2004)



CRAY XT3 kann mehrere Tausend Prozessorelemente beinhalten

Architektur der Supercomputer

… ab Ende 1990er Jahre werden zunehmend Clustersysteme eingesetzt
Anbieter: Hewlett-Packard, IBM, MEGWARE u.v.a.

Cluster *(engl.)* – Gruppe, Schwarm, Haufen
Compute-Cluster – bezeichnet eine Anzahl vernetzter Computer, die eine Aufgabe gemeinsam lösen können

Compute-Cluster:

- sehr große Anzahl Prozessoren
- Verbund von mehreren bis sehr vielen Rechnern
- **!!! sehr gutes Preis-/Leistungsverhältnis !!!**

MEGWARE (2000)

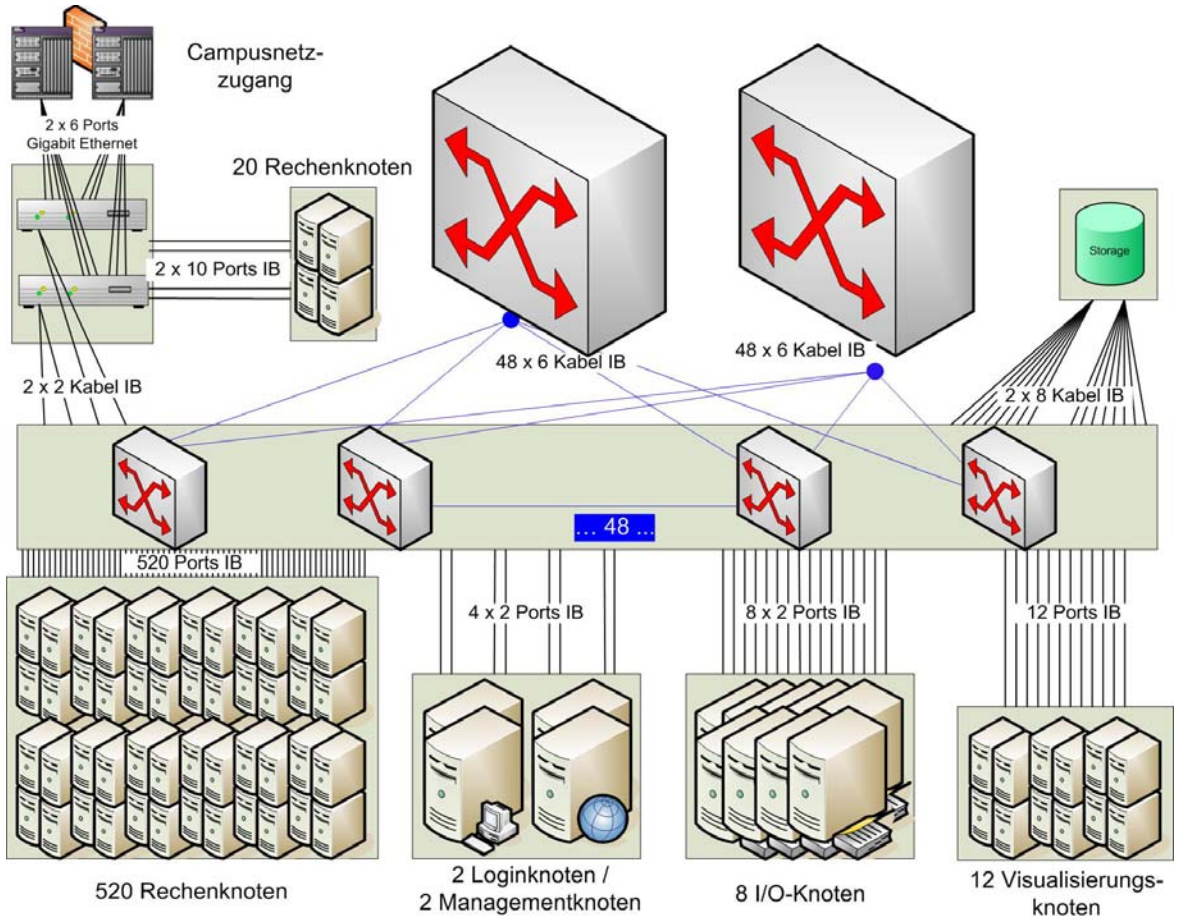
Architektur der Supercomputer

Compute-Cluster:

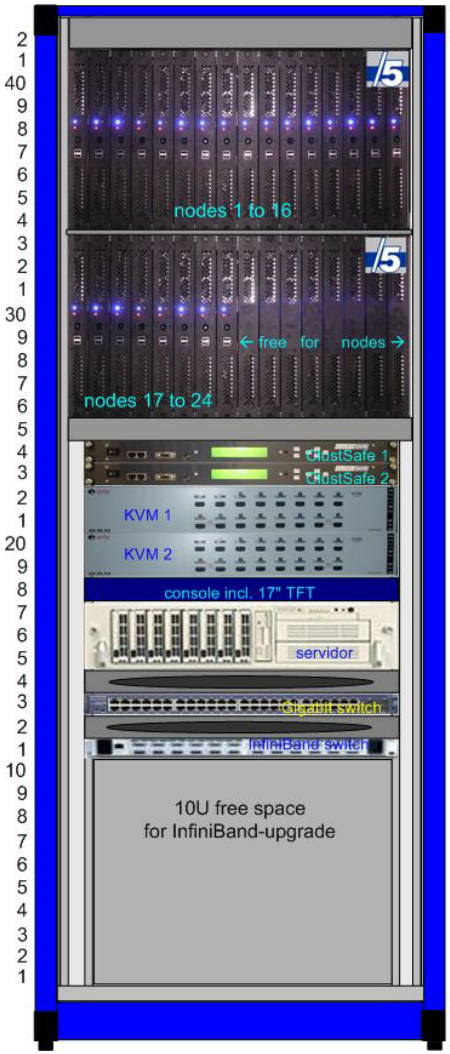
- Verbund von Rechnern mit handelsüblichen Hardware-Komponenten (PC)
- i.d.R Service- und seperates Interprozess-Netzwerk (Ethernet, InfiniBand, SCI, Myrinet)
- Programme und Bibliotheken für parallele Verarbeitung (MPI, PVM, SGE, PBS, Atlas …)

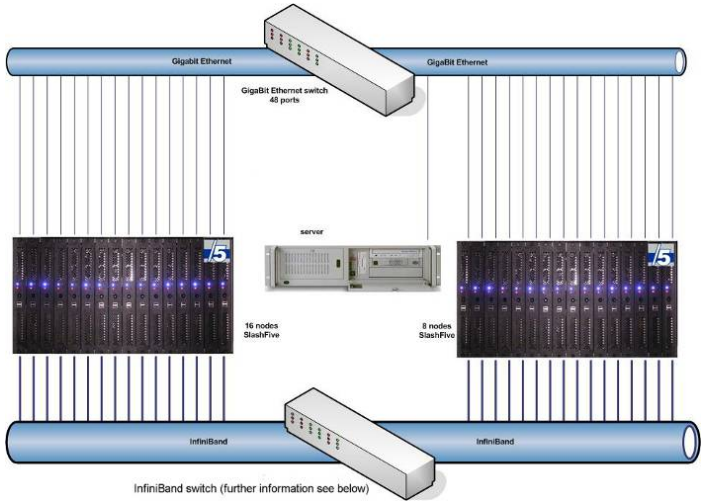
Architektur der Supercomputer

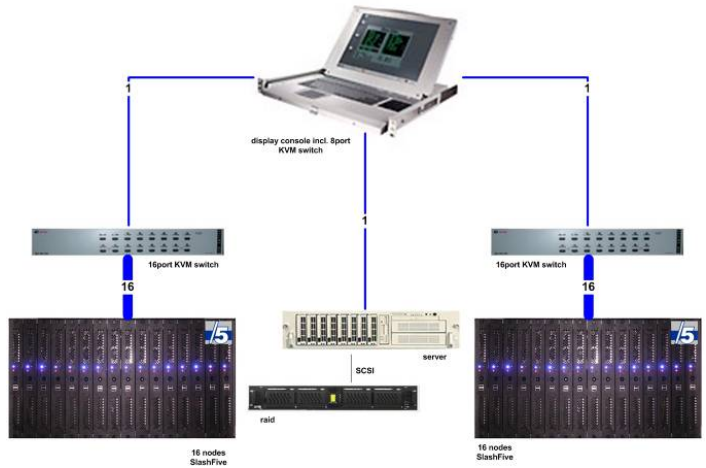
- Beispiel:
- Heterogenes Cluster
- DDR Infiniband Voltaire
- Diskless Node
- Booten über IB
- Paralleles IO mit Lustre
- 12 Visualisierung Node
- Ethernet Management optional
- Realisierung mit Scientific Linux

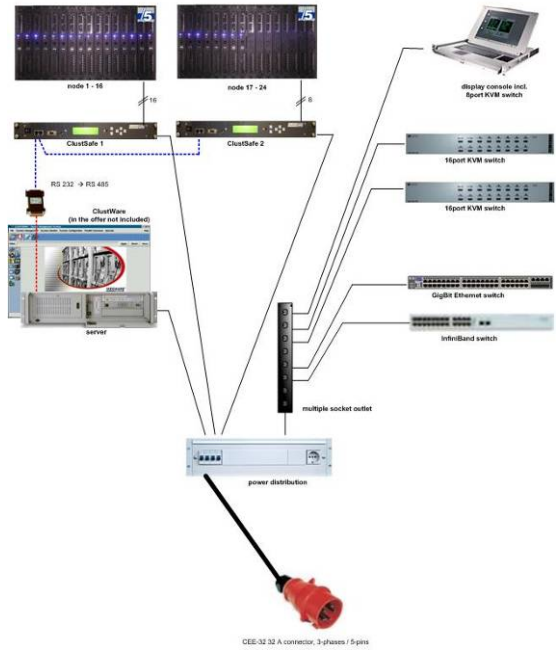


Architektur der Supercomputer









Beispiel: Compute-Cluster für Madrid (MEGWARE)

Architektur der Supercomputer

Proprietäre Bauformen oder eine neue Hauptgruppe:
z.B. von IBM, SiCortex

Bestehen aus wiederholbaren Bausteinen:

- System On Chip (SoC):
  - *CPU*
  - *PCI-Controller*
  - *Netzwerk Interface*
  - *kein RAM*
- Board mit mehreren solchen SoCs und Speicher
- Backplane mit mehreren solchen Boards

Typische Vertreter:

- IBM BlueGene/L, SiCortex SC5832

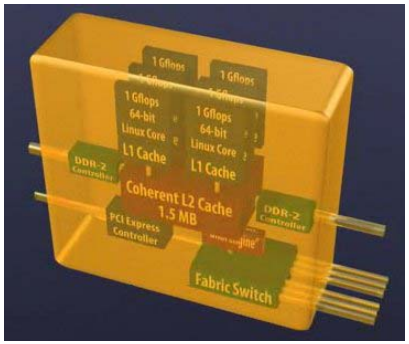
Architektur der Supercomputer

Beispiel: SiCortex SC5832

- 972 Six-Way SMP-Compute-Nodes
- 5832 CPUs
- kompletter Cluster-Node auf einem Chip
- nur 18 KW Leistungsaufnahme
- stromsparend, umweltfreundlich
- seit Herbst 2007 Vertrieb über MEGWARE

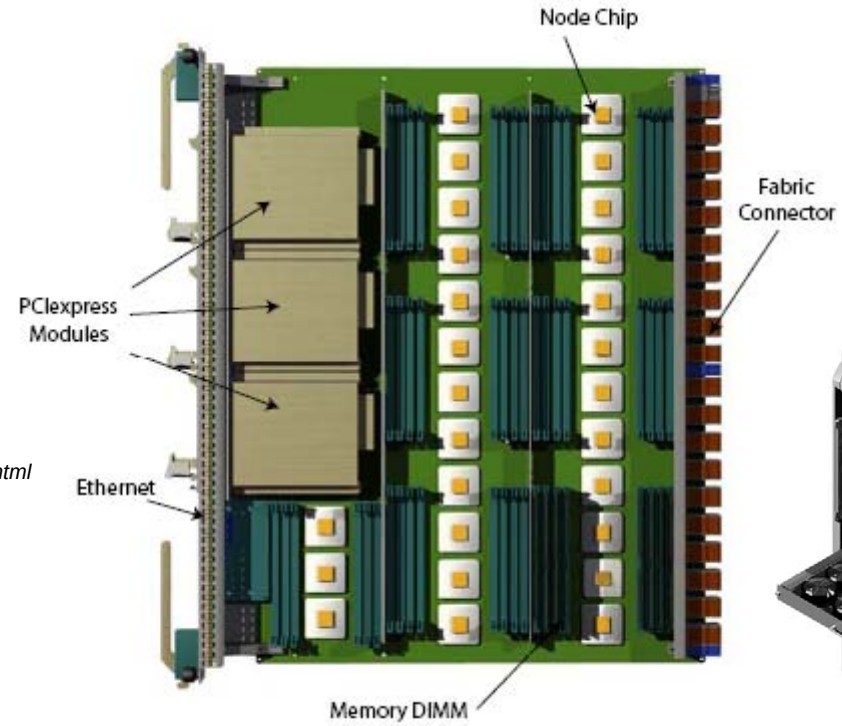
http://www.thealarmclock.com/mt/archives/2006/11/sicortex_superc.html

Architektur der Supercomputer



Quelle: http://www.linuxdevices.com/news/NS3651965718.html

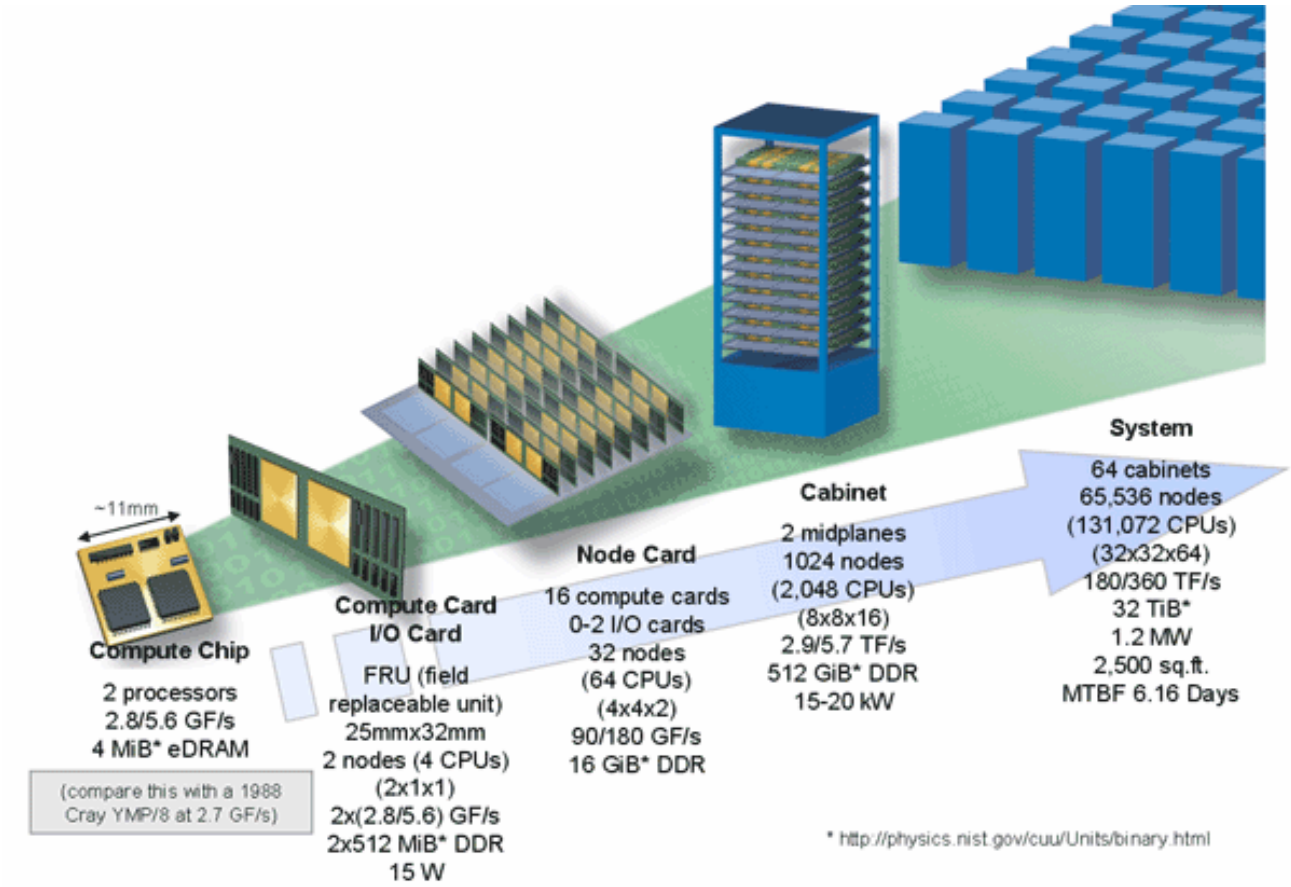
Node-Board (SiCortex SC648 / SC5832)



Quelle: http://sicortex.com/products/white_papers/sicortex_technical_summary

Quelle: http://sicortex.com/products/white_papers/sicortex_technical_summary

Blue Gene

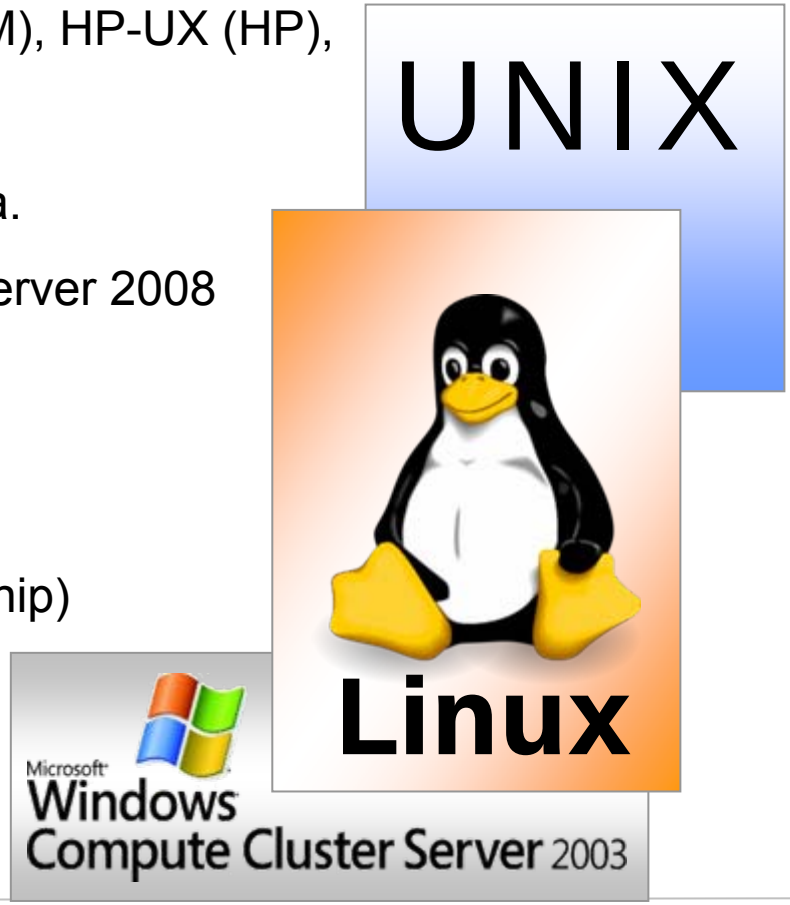


http://www.llnl.gov/asci/platforms/bluegenel/images/bgl_slide2.gif

Betriebssysteme für Supercomputer

Im Wesentlichen sind derzeit 3 Betriebssysteme von Bedeutung.

- UNIX und deren Derivate, z.B. AIX (IBM), HP-UX (HP), Solaris (Sun) u.a.
- Linux, z.B. Debian, Red Hat, SUSE u.a.
- Microsoft Windows Compute Cluster Server 2008



Linux ist im HPC dominierend:

- geringe Kosten (Total Cost of Ownership)
- gute Skalierung und Performance
- relativ einfache Administration

Betriebssysteme in der TOP500

Operating System	Count	Share %	Rmax Sum (GF)	Rpeak Sum (GF)	Processor Sum
Linux	347	69,40%	2459803	3924833	501876
AIX	36	7,20%	386440	495647	72156
CNK/SLES	9	34 6,80%	870513	1100176	391168
SuSE Linux Enterprise Server	9	18 3,60%	311738	431090	57188
HP Unix (HP-UX)	15	3,00%	69939	118451	26356
SUSE Linux Enterprise Server 10	9	1,80%	163750	187339	27288
UNICOS/SUSE Linux	6	1,20%	277242	338060	66744
RedHat Enterprise 4	5	1,00%	110760	167098	15420
SLES10 + SGI ProPack 5	4	0,80%	19823	21299	3328
Solaris	4	0,80%	22675	45659	7488
Super-UX	4	0,80%	52899	59186	5952
UNICOS	3	0,60%	35250	42305	3186
MacOS X	3	0,60%	32989	53008	6296
Redhat Linux	3	0,60%	20774	39369	5152
SUSE Linux	2	0,40%	24197	28672	3584
UNICOS/Linux	2	0,40%	46718	57928	11140
Tru64 UNIX	2	0,40%	18343	26512	11208
Windows Compute Cluster Server 2003	2	0,40%	15518	36357	3808
SuSE Linux Enterprise Server 8	1	0,20%	7215	10259	1776
Totals	**500**	**100,00%**	**4946586,05**	**7183245,39**	**1221114**

http://www.top500.org/stats/list/29/os

Betriebssysteme in der TOP500

Operating System Family	Count	Share %	Rmax Sum (GF)	Rpeak Sum (GF)	Processor Sum
Linux	389	77,80%	3118060	4809959	615612
Unix	60	12,00%	532647	728573	120394
Mixed	42	8,40%	1194473	1496163	469052
BSD Based	4	0,80%	52899	59186	5952
Mac OS	3	0,60%	32989	53008	6296
Windows	2	0,40%	15518	36357	3808
Totals	**500**	**100,00%**	**4946586,05**	**7183245,39**	**1221114**

http://www.top500.org/stats/list/29/osfam

- ca. 63% der Performance mit Linux
- ca. 24% der Performance aus Mischlösungen (Unix und Linux)
- ca. 10% der Performance mit Unix
- ca. 0.3% mit Windows

Gründe für hohen Linux/Unix Anteil:

- Geringe Kosten (Total Cost of Ownership)
  TCO beinhaltet Anschaffungs- sowie Betriebskosten (Also auch Support usw.) (HW+SW)
- Gute Skalierung und Performance
- Relativ einfache Administration

Agenda

- Was ist High Performance Computing (HPC)
- Funktionsprinzipien des parallelen Rechnens
- Anwendungsgebiete des parallelen Rechnens
- technologische Besonderheiten im Hochleistungsrechnen
- Forschung und Entwicklung von HPC-Systemen in Chemnitz
- Lösungen und Anwendungen für das High Performance Computing Ausgewählte Anwendungsmöglichkeiten
  - Erdsystemforschung – „Klimaforschung“
  - Neandertaler und hierarchische Matrizen
- Technologien im Supercomputing – das Handwerkszeug eines HPC Ingenieurs
- Widrigkeiten und offene Probleme
- Berufliche Zukunft in Chemnitz

Funktionsprinzipien des parallelen Rechnens

Ideal: Jeder beteiligte Prozessor löst ein Teil des Gesamtproblems
=> SpeedUp

Beispiel: Wettervorhersage
Maschenweite 7 km in Mitteleuropa, 35 vertikale Schichten kann am DWD in zwei Stunden auf 200 Prozessoren gerechnet werden.

Aber: Nicht jedes Problem ist einfach parallelisierbar

Es gibt verschiedene Techniken, um ein Problem parallel zu lösen

- Divide and Conquer (teile und herrsche)
- Pipelining (vergl. Richard-Cole-Sort)
- … und viele weitere

Funktionsprinzipien des parallelen Rechnens

Beispiel: Matrix – Vektor Multiplikation

- Problem: $A \cdot x = y$ mit $A \in N^{n \times n}$ und $x, y \in N^{n \times 1}$ haben p Prozessoren
- Lösung vertikal:
  - teilen A vertikal in p Teil-Matrizen: $A_i \in N^{n \times \frac{n}{p}}$ und x in $z_i \in N^{\frac{n}{p} \times 1}$

$$A \cdot x = \left(\underbrace{\begin{matrix} a_{1,1} & \cdots & a_{1,r} \\ \vdots & & \vdots \\ a_{n,1} & \cdots & a_{n,r} \end{matrix}}_{A_1} \underbrace{\begin{matrix} a_{1,r+1} & \cdots & a_{1,2r} \\ \vdots & & \vdots \\ a_{n,r+1} & \cdots & a_{n,2r} \end{matrix}}_{A_2} \cdots \underbrace{\begin{matrix} a_{1,n-r+1} & \cdots & a_{1n} \\ \vdots & & \vdots \\ a_{n,n-r+1} & \cdots & a_{nn} \end{matrix}}_{A_n} \right) \cdot \left(\begin{matrix} x_1 \\ \vdots \\ x_r \end{matrix} \right] z_1 \quad \vdots \quad \left[\begin{matrix} x_{n-r} \\ \vdots \\ x_n \end{matrix} \right] z_n \Bigg) = y$$

- Teilergebnisse $A_i \cdot z_i = w_i$ wieder einsammeln und Gesammtaufgabe lösen
- $y = \sum_i w_i$ (bspw. als Baumreduktion)
- Implementierung bspw. mittels Bibliotheken wie MPI, PVM,...

Funktionsprinzipien des parallelen Rechnens

Beispiel: MPI Programm zu Summe zufälliger Zahlen

```
#define _MPI_CPP_BINDINGS
#include <mpi.h>
#include <iostream>
#include <time.h>

int main(int argc, char* argv[])
{
  MPI::Init(argc, argv);

  int rank = MPI::COMM_WORLD.Get_rank();
  int size = MPI::COMM_WORLD.Get_size();

  srand(rank + time(NULL));
  int data = rand()%100;
  int result = 0;

  std::cout << "Hello World! I am:\t" << rank << " of:\t" << size;
  std::cout << " having:\t" << data << std::endl;

  MPI::COMM_WORLD.Reduce(&data, &result, 1, MPI::INT, MPI::SUM, 0);

  if (rank == 0) std::cout << "sum:\t" << result << std::endl;

  MPI::Finalize();
}
```

MPI – Reduce:

- baumartige Reduktion, so dass zum Schluss Ergebnis bei der Wurzel (rank 0) liegt
- andere Knoten symbolisieren Verknüpfung von zwei Datensätzen mittels MPI::SUM

Beispiel: MPI Programm zu Summe zufälliger Zahlen
→ Outputs (3 Knoten und 13 Knoten)

```
math@math-desktop:~/presentations/gastprofessur/MatrixVectorMul$/usr/local/openmpi-1.2.4/bin/mpirun -np 3
./hello
Hello World! I am:      0 of:   3 having:       15
Hello World! I am:      1 of:   3 having:       44
Hello World! I am:      2 of:   3 having:       26
sum:    85
math@math-desktop:~/presentations/gastprofessur/MatrixVectorMul$/usr/local/openmpi-1.2.4/bin/mpirun -np 13
./hello
Hello World! I am:      3 of:   13 having:      30
Hello World! I am:      0 of:   13 having:      56
Hello World! I am:      1 of:   13 having:      10
Hello World! I am:      2 of:   13 having:      30
Hello World! I am:      4 of:   13 having:      72
Hello World! I am:      5 of:   13 having:      18
Hello World! I am:      6 of:   13 having:      53
Hello World! I am:      7 of:   13 having:      99
Hello World! I am:      8 of:   13 having:      50
Hello World! I am:      9 of:   13 having:      76
Hello World! I am:      12 of:  13 having:      29
Hello World! I am:      10 of:  13 having:      16
Hello World! I am:      11 of:  13 having:      26
sum:    565
```

Gesetze Parallelen Programmierens:

Theoretischer maximaler SpeedUp Amdahls Law:

- Wenn nicht alle Teile eines Programms parallelisierbar sind
- SpeedUp eines Programms bei mehreren Rechnern:

$$S = \frac{\text{Execution time without enhancement}}{\text{Enhanced execution time}}$$

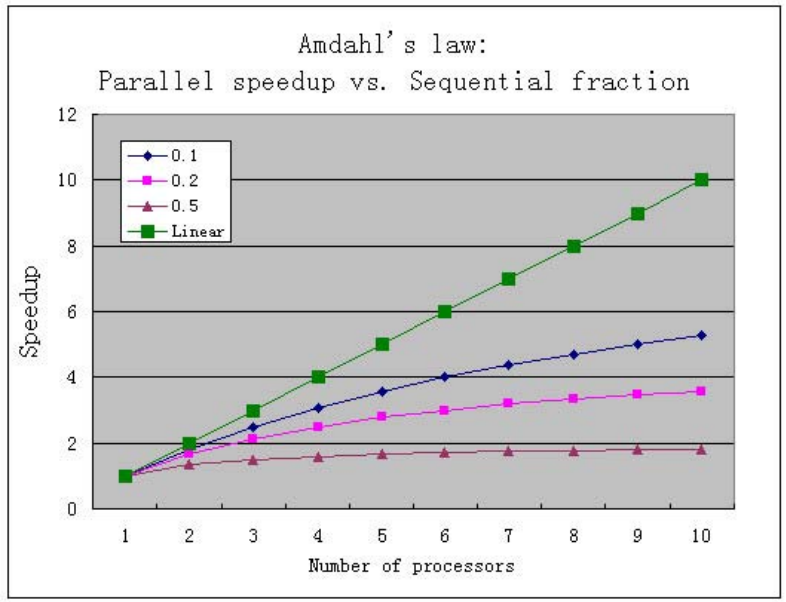
Beispiel: $S = T_1 / T_p$

T_1 sequentielle Ausführungszeit,
T_p Ausführungszeit bei p *Prozessoren*

- Maximaler SpeedUp hängt von 2 Faktoren ab:
  - Anteil des Programms welcher optimierbar ist: P (max. 100%)
  - SpeedUp des optimierten Anteils: $S_{partial}$

$$S_{overall} = \frac{1}{(1-P) + \frac{P}{S_{partial}}}$$

- Problem: Serieller Anteil ist konstant modelliert
  → In Praxis auch abhängig von #CPU
- Lösung: Zusammen betrachten mit *Gustavson Barsis' Law*



Quelle: http://en.wikipedia.org/wiki/Image:Amdahl-law.jpg

Gesetze Parallelen Programmierens:

- Amdahls Gesetz sorgte 21 Jahre dafür, dass massiv paralleles Rechnen in einer unbedeutenden Nische verbannt war.
- Erst Gustavson durchbrach diese Schranke, indem er erkannte die parallelen Probleme müssen nur groß genug sein um einen Speedup auch bei 1000 Prozessen zu erhalten.
- Es gelten immer beide Gesetze bei der Untersuchung eines theoretischen Speedup.
- Speedup Werte können bis zum Wert P gehen, bekannt sind auch Superlineare Werte. Letztere entstehen wenn ein Prozessor durch die Parallelisierung die Cache sehr gut nutzen kann und damit den sequentiellen Anteil mehr als ausgleicht.

Agenda

- Was ist High Performance Computing (HPC)
- Funktionsprinzipien des parallele Rechnens
- Anwendungsgebiete des parallelen Rechnens
- technologische Besonderheiten im Hochleistungsrechnen
- Forschung und Entwicklung von HPC-Systemen in Chemnitz
- Lösungen und Anwendungen für das High Performance Computing Ausgewählte Anwendungsmöglichkeiten
  - Erdsystemforschung – „Klimaforschung“
  - Neandertaler und hierarchische Matrizen
- Technologien im Supercomputing – das Handwerkszeug eines HPC Ingenieurs
- Widrigkeiten und offene Probleme
- Berufliche Zukunft in Chemnitz

Anwendungsgebiete des parallelen Rechnens

- Einsatzfelder überall dort, wo
- großer Bedarf an hohen Rechenleistungen besteht
- die Leistung von wenigen Prozessoren nicht mehr ausreicht

Wissenschaft

Forschung & Lehre

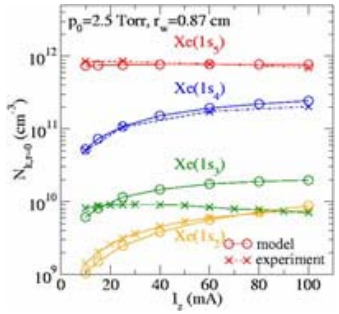
Industrie

Bedarf verdoppelt sich jährlich !

Anwendungsgebiete des parallelen Rechnens

Wissenschaft

Institut für Niedertemperatur-Plasmaphysik Greifswald



Beispiel Plasmamodellierung

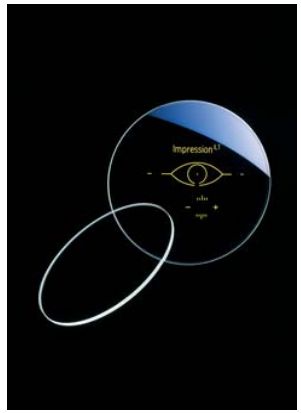
Entwicklung von Modellen und Simulationen für anisotherme und thermische Plasmen
Analyse wissenschaftlich und technologisch relevanter Plasmen in enger Kopplung mit Experimenten und Anwendungen

Behandlung plasmaspezifischer Problemstellungen wie
- Kinetik geladener Spezies
- Plasmachemie und Transportprozesse
- Strahlungstransport und Spektrenanalyse
- Wechselwirkung von Plasmen mit Wänden und Elektroden
- Mehrflüssigkeitsbeschreibung und Strömungssimulation

Hauseigene numerische Verfahren und kommerzielle Codes

z.B. Nanostrukturphysik – Entwicklung neuer Werkstoffe und Oberflächen

- Veredlung von Kunststoffen mit funktionellen Oberflächenschichten z.B. Brillengläser
- Entwicklung von Plasmatechnologien für die Halbleiterindustrie, z.B. Microchipfertigung



Quelle: http://www.inp-greifswald.de/web.nsf/sfdm-projekte

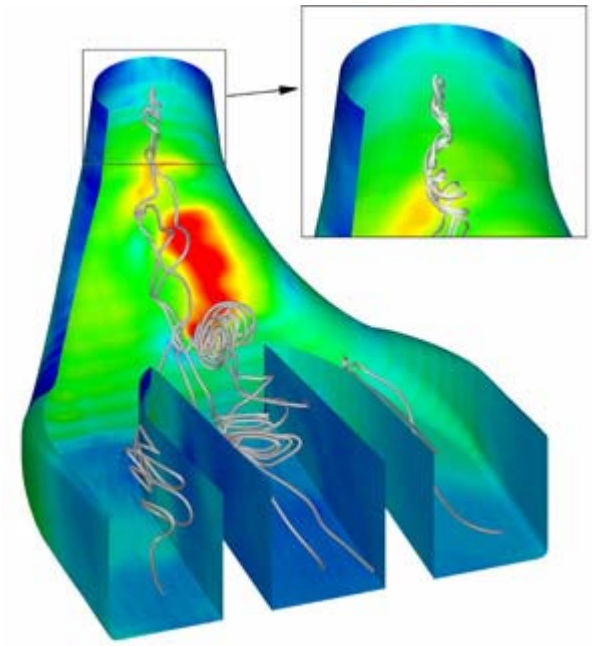
Anwendungsgebiete des parallelen Rechnens

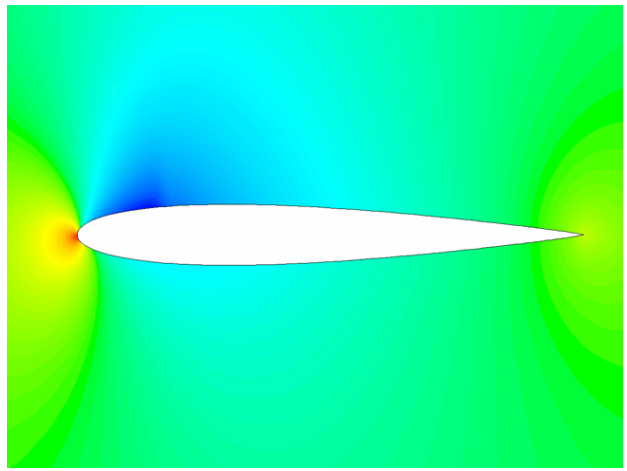
Forschung & Lehre

in fast allen Bereichen der Naturwissenschaft und Technik

z.B. Strömungsmechanik – Simulation von Strömungsvorgängen

- Luft- und Raumfahrtforschung
- Energiegewinnung
- Umwelttechnik
- uvm.





Anwendungsgebiete des parallelen Rechnens

Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen

•Anzahl CPU: 300 (600 Core)

•Anzahl Nodes: 150

•Intel Xeon 5160

•Interconnect Infiniband

•Racksystem SlashFive

•Anwendung: Forschungsaufgaben in verschiedenen Wissenschaftsbereichen, wie der Astronomie, der biophysikalischen und anorganischen Chemie. Mit diesem Cluster werden z.B. die zeitliche Entwicklung von Instabilitäten in unserem Sonnensystem simuliert, die mechanischen Eigenschaften von DNA-Molekülen und die Faltungszustände von Proteinen untersucht.

•Expertenmeinung: Dr. Ulrich Schwardmann (GWDG) „Mit der neuen Intel-Woodcrest-Architektur erwarten wir eine erhebliche Leistungssteigerung gegenüber vorangegangenen Cluster-Architekturen. Von der Dual-Core-Technologie werden Programme mit SMP-Skalierbarkeit besonders bevorzugt.“

Anwendungsgebiete des parallelen Rechnens

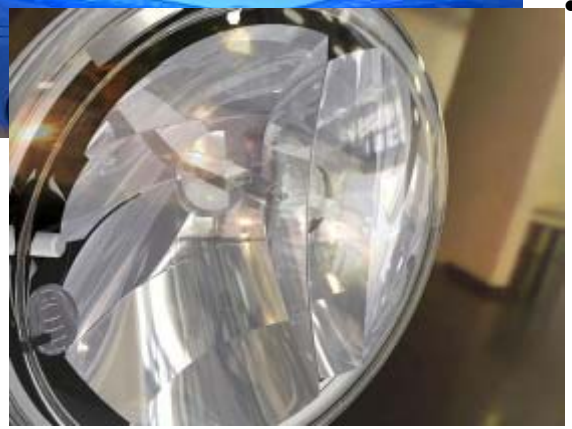
Damiana – für das Albert Einstein Institut für Gravitationsphysik (AEI) in Potsdam Golm

Anwendungsgebiete des parallelen Rechnens

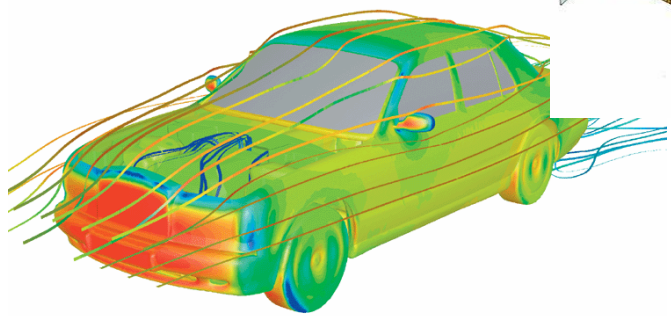
Industrie

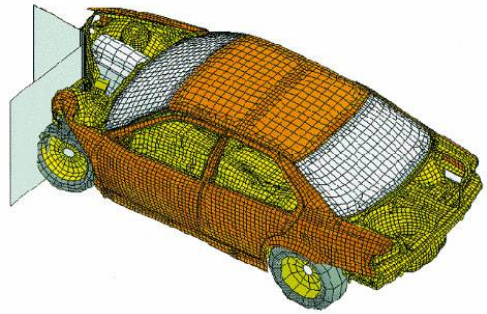
hier vor allem in der Automobilindustrie

z.B. Produktentwicklung und Formdesign



- Karosserieoptimierungen (Strömungstests)
- Materialkontrolle (Crashtests)
- Minimierung von Geräuschen und Vibrationen
- Oberflächendesign (Spiegelungen, Reflexionen)
- Fahrverhalten
- uvm.



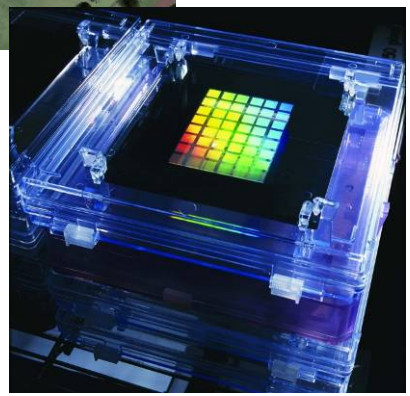


Anwendungsgebiete des parallelen Rechnens

Industrie

AMTC (Advanced Mask Technology Center)

- Joint Venture von AMD und Infineon
- Sitz in Dresden, direkt neben AMD (Fab 30 und 36)
- gemeinschaftliches Forschungsprojekt mit MEGWARE



- Herstellung von Photomasken für die Halbleiterindustrie (z.B. AMD, Infineon)
- z.B. die Maskensätze für den Opteron
- riesige Datenmengen → Speicherkapazität
- sehr hohe Genauigkeit → 64Bit
- schnelle Verarbeitung → hohe Rechenleistung
- Datensicherheit → weit reichende Redundanz

Quelle: http://www.amtc-dresden.com/homepage/content/index.php?js=1

Agenda

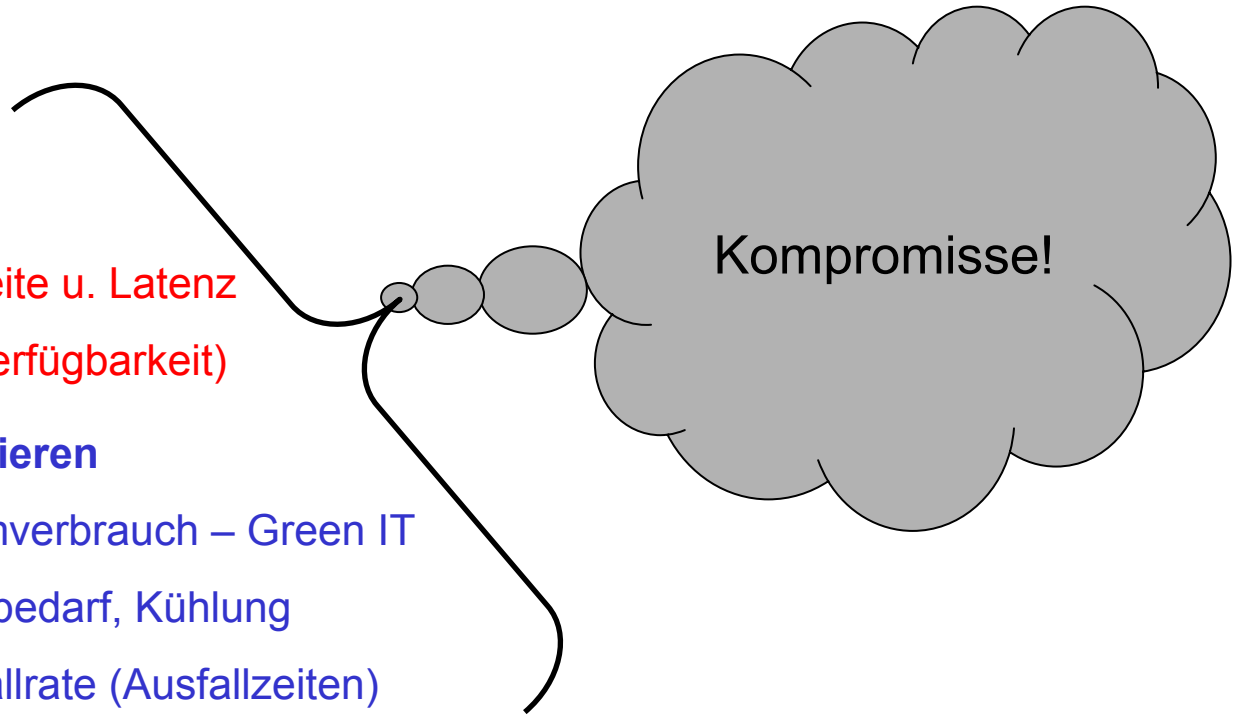
- Was ist High Performance Computing (HPC)
- Funktionsprinzipien des parallele Rechnens
- Anwendungsgebiete des parallelen Rechnens
- technologische Besonderheiten im Hochleistungsrechnen
- Forschung und Entwicklung von HPC-Systemen in Chemnitz
- Lösungen und Anwendungen für das High Performance Computing Ausgewählte Anwendungsmöglichkeiten
  - Erdsystemforschung – „Klimaforschung“
  - Neandertaler und hierarchische Matrizen
- Technologien im Supercomputing – das Handwerkszeug eines HPC Ingenieurs
- Widrigkeiten und offene Probleme
- Berufliche Zukunft in Chemnitz

Technologieentwicklung Supercomputer

Der Wunschzettel:

Erhöhen:

- Rechenleistung
- Netzwerke – Bandbreite u. Latenz
- Betriebssicherheit (Verfügbarkeit)



Reduzieren

- Stromverbrauch – Green IT
- Platzbedarf, Kühlung
- Ausfallrate (Ausfallzeiten)

Verbessern / Vereinfachen

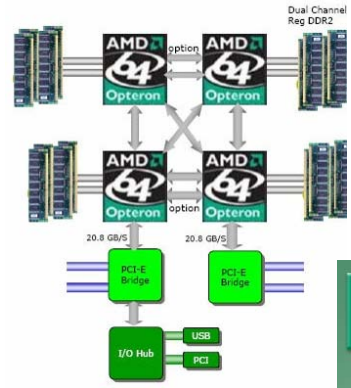
- Monitoring und Management
- Administration
- Effizienz der Applikationen

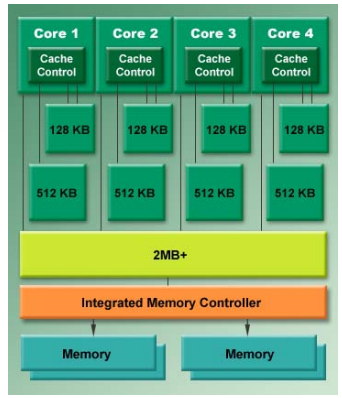
Technologieentwicklung Supercomputer

Erhöhen der Rechenleistung

Schneller durch paralleler

- weitere Steigerung der Taktfrequenzen der CPUs ist problematisch,
- alternativer Weg ist die Parallelisierung
- Motherboards mit mehreren CPUs pro Board
- nun auch Prozessoren mit mehreren Kernen → Multicore





Quelle: http://multicore.amd.com/de-de/AMD-Multi-Core/Vorteile-von-Quad-Core/Leistung.aspx

Schnellere Netzwerke

- Fast Ethernet → **Gigabit Ethernet** → *10G Ethernet*
- Myrinet, **InfiniBand** u.a.
- spezielle Topologien (z.B. 3D-Torus)

Technologieentwicklung Supercomputer

Reduzieren des Stromverbrauch und Platzbedarfs

Mehr Rechenleistung pro Watt
- die Gehäuse werden immer mehr strömungsoptimiert, um Lüfterleistung zu sparen,
- es werden Netzteile mit sehr hohem Wirkungsgrad eingesetzt,
- Netzteile am Anfang des Kühlluftstroms → Temperatur niedriger (besserer Wirkungsgrad),
- Komponenten mit geringem Stromverbrauch (z.B. keine Boards mit unnötigen SCSI-Interfaces o.ä.),
- Prozessorauswahl mit Blick auf einen geringeren Stromverbrauch (Beispiel HE-CPUs von AMD)
- z.B. „CoolCore“- Technologie im Opteron

Verringerung des Platzbedarfs einzelner Rechenknoten → z.B. Blades → z.B. Blue Gene / SiCortex
- eine gewisse Abkehr von dem Grundprinzip auf Standardkomponenten zurückzugreifen,
- eine Reduzierung der Universalität der Lösung hinsichtlich der Nutzbarkeit z.B. unterschiedlichster Interconnects,
- die Boards werden auf die notwendigen Komponenten reduziert,
- dadurch werden sie stark verkleinert und im Stromverbrauch reduziert,
- geringere Taktfrequenzen senken ebenfalls den Stromverbrauch *(!!! Lizenzmodelle !!!)*
- Verkabelungsaufwand durch die Backplane stark verringert

Die Verringerung des Stromverbrauchs senkt die Betriebstemperaturen.

Eine Reduzierung auf die nötigsten Komponenten bietet geringere „Ausfallchancen“.

→ Reduzieren der Ausfallrate (Ausfallzeiten)
→ MTBF, MTTR, Verfügbarkeit (availability)

Quelle: http://www.thealarmclock.com/mt/archives/2006/11/sicortex_superc.html

Technologieentwicklung Supercomputer

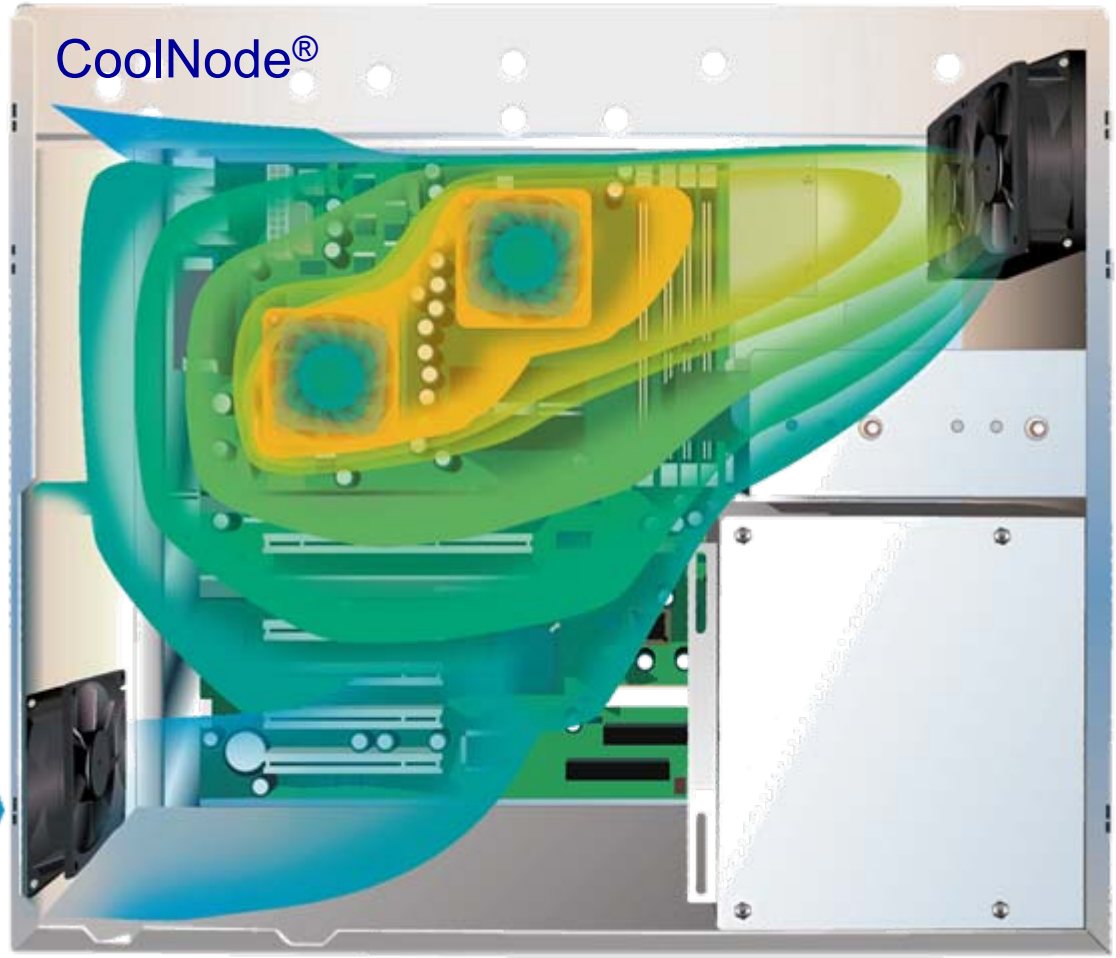
Einbeziehung weiterer FPUs (FPGA,Grafikkarten)

- spezielle Chips mit einer Struktur, die ähnlich den Vektorprozessoren (SIMD) ist,
- das können z.B. Prozessoren sein, wie sie für Grafikkarten eingesetzt werden,
- aber auch FPGAs (**F**ield **P**rogrammable **G**ate **A**rray),

- führen Operationen in einem oder wenigen Schritten aus, für die die CPU viele Schritte und Schleifen braucht,
- Idee ist die Auslagerung solcher Berechnungen in eine geeignete externe Recheneinheit,
- technische Realisierung z.B. per Erweiterungskarten, die solche Chips tragen,
- derzeit laufen aber auch Versuche mit kombinierten Motherboards (CPU(s) + FPGS(s))
- Beispiel: NVIDIA Tesla C870 – 500 GigaFlop (GPU)

Ein Nachteil solcher Systeme liegt in der Notwendigkeit der Programmierung des FPGA auf die konkrete Operation. Diese Systeme sind eher in hoch spezialisierten Einsatzfällen anzutreffen. Für Cluster im RZ erscheint diese Lösung momentan häufig noch ungeeignet.

Technologieentwicklung Supercomputer



warm

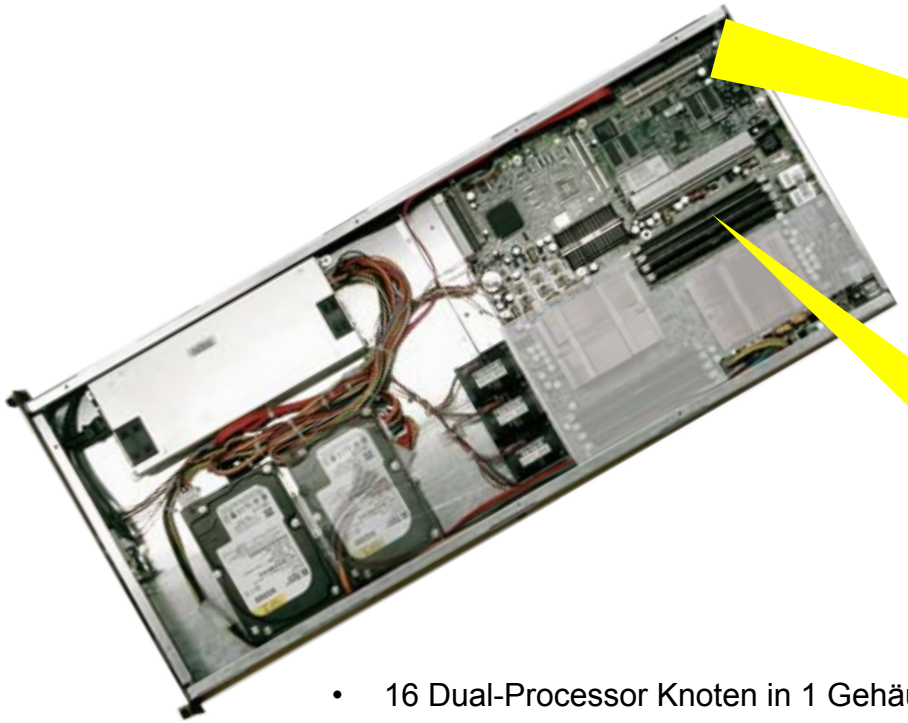
kalt

Technologieentwicklung Supercomputer

SlashTwo®

- bis zu 80 CPUs (160 Cores) pro Schrank
- alle low profile PCI-X slots des Mainboards nutzbar
- 2HE für 2 Knoten entsprechen 1HE-Knoten, aber geringere Kosten und besser Kühung
- Flüssigkeitskühlung möglich

Technologieentwicklung Supercomputer



SlashFive®

- 16 Dual-Processor Knoten in 1 Gehäuse (nur 8HE - entspricht 0,5HE pro Knoten)
- Standardkomponenten
- optimierte Kühlung
- bis zu 160 CPUs (640 Cores) pro Schrank
- Flüssigkeitskühlung möglich

Technologieentwicklung Supercomputer

Wasser vs. Luft

$Q = c \cdot m \cdot \Delta T$ (Q... Wärme, c… Wärmekoeffizient, m… Masse, ΔT… Temperaturerhöhung)
$Q = P \cdot t$ (P... Heizleistung, t... Zeit)
(ΔT) – Die Temperaturerhöhung ist direkt proportional zu P - der Heizleistung
wieviel, hängt von c und m ab:

c…	Wärmekoeffizient von Luft ca. 1000 kJ/kg	
	Wärmekoeffizient von Wasser ca. 4200 kJ/kg	Wasser ca. 4,2 mal besser
m…	Dichte der Luft = 0.00118 kg/L	
	Dichte von Wasser = 1kg/L	Wasser ca. 850 mal besser

Schlussfolgerungen:

Wasser kühlt ca. 3500 mal (4,2 x 850) besser als Luft.

Das bedeutet, wir können den gleichen Kühleffekt mit einem Kühlwasservolumen erreichen, das 3500 mal kleiner ist als das Kühlluftvolumen.
Mit einem Kühlwasservolumen, das 1000 mal kleiner ist als das Kühlluftvolumen, erreichen wir in einem direkt wassergekühlten System einen Temperaturanstieg, der 3.5 mal geringer ist als der in einem luftgekühlten System.

Beispiel:	**Raumtemperatur**	**= 20°C**	
	Arbeitstemperatur einer luftgekühlten CPU	**= 70°C**	**ΔT = 50°C**
	Arbeitstemperatur einer wassergekühlten CPU	**=** (50/3,5=14,29) **= 35°C** (35°C geringer)	**ΔT = 15°C**

Technologieentwicklung Supercomputer

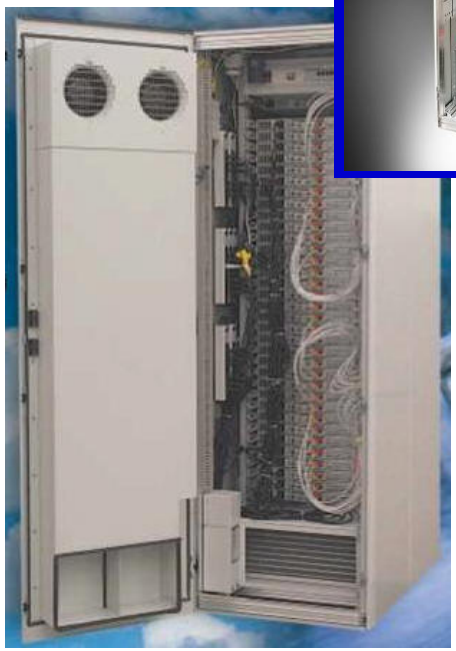
Gesamtsystem per Luft gekühlt

bedarf einer ausreichenden Raumklimatisierung !!!

Intern Luftkühlung

Schrank ist geschlossenes System, enthält Luft-Wasser Wärmetauscher.

Schrank wird per Wasser gekühlt.

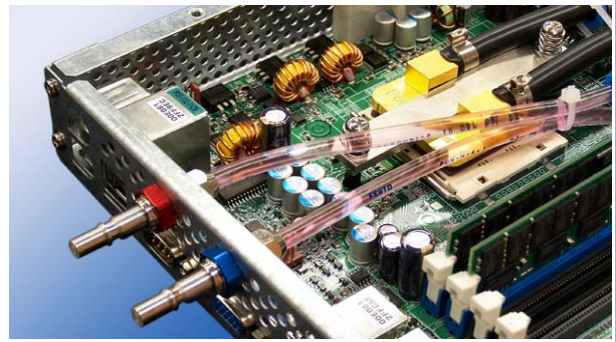


benötigt Kühlwasser oder einer Rückkühlanlage !!!

CPUs wassergekühlt (ca. 66% der Gesamtwärmeverlustleistung).

Rest luftgekühlt.

Gesamtsystem arbeitet bei deutlich geringerer Temperatur → höhere Zuverlässigkeit



Technologieentwicklung Supercomputer

Angepasste Schranklösungen / -konzepte

ClustRack ®

Universelles Rack

Spezielle Lösung für ein portables HPC- System zur mobilen Datenerfassung und Verarbeitung (z.B. zur Entwicklung des A380)

flight case

Personal **S**uper**C**omputer z.B. für den Engineering - Bereich

Compact cluster PSC4000

Technologieentwicklung Supercomputer

Model – MEG F1200

- bis zu 12 CPUs Intel® Xeon® 54xx (48 Core)
- max. 192GB RAM pro Chassis
- ScaleMP® VersatileSMP Architektur
- 7 Gigabit Ethernet Ports
- 6 Festplatten bis 4.5 TB (SATA)
- Betriebssystem: alle Standard-Linux-Distributionen
- 6U 19“ oder Workstation
- erweiterbar auf bis zu 48 CPUs und 768GB RAM (Kombination von 4 Systemen)

Perfekt für folgende Anforderungen:

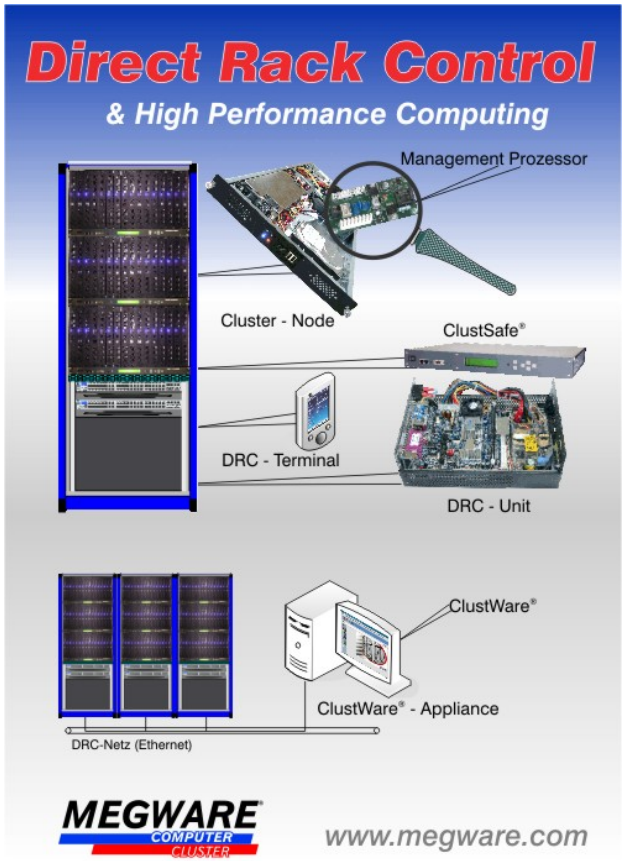
- CFD, FEM, Chemie (Fluent, Abaqus, Gaussian)
- Pre- Postprocessing mit hoher Speicheranforderung
- OpenMP und Threads auf Basis von SHMEM
- hoher Speicherdurchsatz (Stream 30 GByte/s)
- einfachstes Management durch SingleSystemImage

SMP – Server als Tower

Technologieentwicklung Supercomputer

Monitoring und Management vereinfachen
- Handhabung eines HPC- Systems immer weiter vereinfachen,
- die Anwender in der Industrie sollen sich auf ihre Kernaufgaben konzentrieren können

Direct Rack Control (DRC) und ClusterWare Appliance



- Rack-orientiertes Management
- in jedem Node befindet sich ein MP
- MP liest Sensoren, CPU- und Netzlast aus
- Verbunden mit Reset- und Power- Button
- MP gibt alle Daten über USB-Bus an DRC
- DRC sammelt alle MP-Daten eines Racks
- ID- LED für jeden Knoten (vorn und hinten)

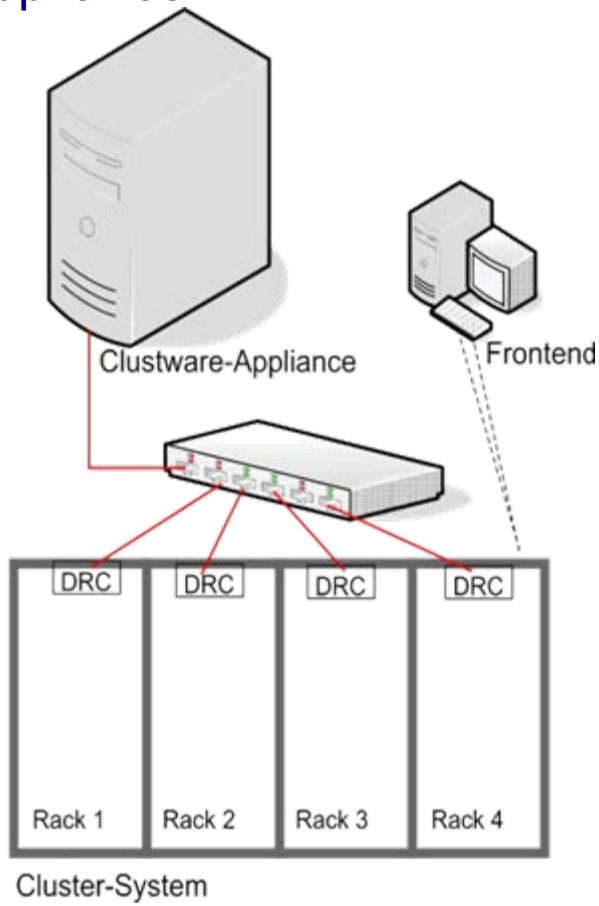
Technologieentwicklung Supercomputer

Monitoring und Management vereinfachen
- Handhabung eines HPC- Systems immer weiter vereinfachen,
- die Anwender in der Industrie sollen sich auf ihre Kernaufgaben konzentrieren können

Direct Rack Control (DRC) und ClusterWare Appliance

ClustWare – Appliance

- Management – Server
- Cluster – Managementsoftware
- Überwachung des kompletten Cluster
  Auslastung, Archivierung, Statistiken
- Backup Cluster-Software + Konfiguration
- Automatische Remoteinstallation aller Server + Nodes
- serielle Konsole zu allen angeschlossenen Nodes und Servern
- automatische Biosupdates auf allen Boards mgl.



Agenda

- Was ist High Performance Computing (HPC)
- Funktionsprinzipien des parallele Rechnens
- Anwendungsgebiete des parallelen Rechnens
- technologische Besonderheiten im Hochleistungsrechnen
- Forschung und Entwicklung von HPC-Systemen in Chemnitz
- Lösungen und Anwendungen für das High Performance Computing Ausgewählte Anwendungsmöglichkeiten
  - Erdsystemforschung – „Klimaforschung“
  - Neandertaler und hierarchische Matrizen
- Technologien im Supercomputing – das Handwerkszeug eines HPC Ingenieurs
- Widrigkeiten und offene Probleme
- Berufliche Zukunft in Chemnitz

High Performance Computing in Chemnitz

Lehre und Forschung

- TU Chemnitz
- Studiengang Master: Parallele und Verteilte Systeme
- Forschungsbereiche in verschiedenen Fachgebieten

High Performance Computing in Chemnitz

Unternehmen

- MEGWARE Computer GmbH
- Entwicklung von innovativen HPC-Systemlösungen
- Installation und Support von Supercomputern im In- und Ausland

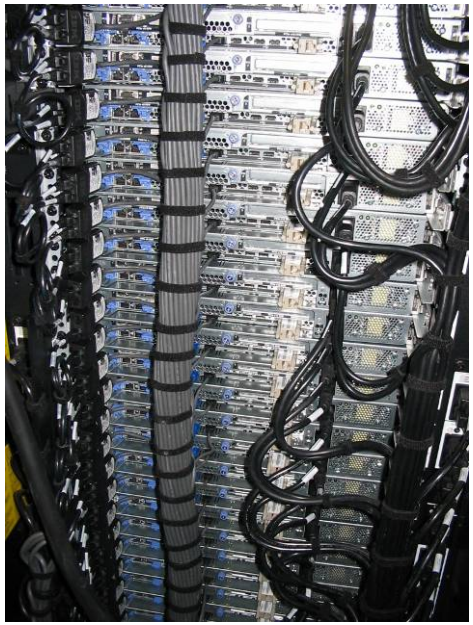
CLIC (Chemnitzer Linux Cluster)

- 2000 installiert, 530 CPU's:
  Intel IA-32 Pentium 3 800 MHz (0.8 Gflops)
- Interconnect: Fast Ethernet
- November 2000:
  Rang 126
  Rmax 143.3 GFlops
- November 2001:
  Rang 137
  Rmax 221.6 GFlops

Anwendungen:

- Simulation von Temperaturflüssen
- Simulation von Deformationsvorgängen
- http://www.clug.de/vortraege/CLIC/slides.html

http://www.clug.de/vortraege/CLIC/images/gang2.jpg



CHIC (Chemnitz High Performance Linux Cluster)

- 2007 installiert, 2152 CPU's:
  AMD x86_64 Opteron Dual Core (2218 Step2) 2600 MHz (5.2 GFlops)
- Interconnect: Infiniband
- Juni 2007:
  Rang 117
  Rmax 8210 GFlops
- http://www.tu-chemnitz.de/chic/

http://www.tu-chemnitz.de/chic/CHICClustercomputer/Bildergalerie/IMG_0792.jpg

Partnerschaft zwischen TU und MEGWARE

- langjährige erfolgreiche Zusammenarbeit
- ganz besonders zur Professur Rechnerarchitektur (Prof. Dr. Rehm)

2000: Chemnitzer Linux Cluster (CLiC)

- 528 Rechner Nodes
- 528 CPUs: AMD Pentium III 800 MHz
- 44 Racks in 4 Reihen
- 143 Gigaflops (Linpack)

TOP500-Liste – Platz 126

- zweitschnellster Compute-Cluster in Europa
- weltweit bestes Preis-/Leistungsverhältnis

Aussage: Prof. Dr. Meuer

Partnerschaft zwischen TU und MEGWARE

2007: Chemnitzer Hochleistungs-Linux Cluster (CHiC)

- gemeinsames Projekt IBM und MEGWARE
- 538 Rechner Nodes
- 1.076 CPUs: AMD Opteron Dual Core
- 60 Terabyte Storage-System
- 18 wassergekühlte Racks in 2 Reihen
- 8210 Gigaflops (Linpack)

Vergleich zum CLiC

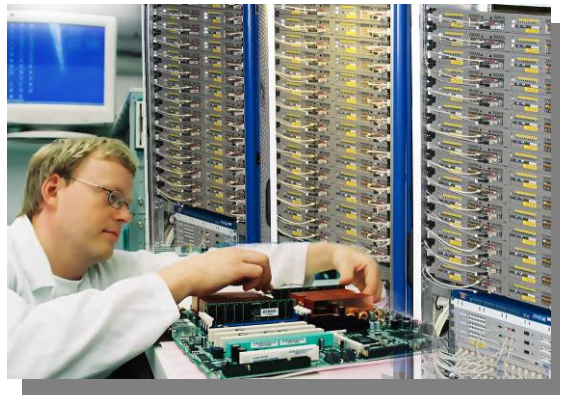
- ca. 4x mehr CPU-Cores
- ca. 60% weniger Racks
- ca. 57x höhere Rechenleistung

TOP500
117

Partnerschaft zwischen TU und MEGWARE

Wir suchen: aus der Fakultät Informatik oder artverwandten Fachbereichen

- Werksstudenten
- Praktikanten
- Diplomanden



Wir bieten: interessante und anspruchsvolle Entwicklungsaufgaben

z.B.

- Evaluierung von Cluster-Storagesystemen
- Tests und Gegenüberstellung von Queuing- bzw. Batchsystemen
- Implementierung eines Command-Line Interface für das Cluster-Management
- Aufbau und Programmierung universeller Mikrocontroller-Baugruppen

Einige Entwicklungsschwerpunkte

Anforderungen an Forschung und Entwicklung eines HPC-System Herstellers

Auszug:

- Erhöhung der Packungsdichte in HPC-Systemen
- Erhöhung der Zuverlässigkeit und Verfügbarkeit von HPC-Systemen
- finden von angepassten Kühlungslösungen
- das bedeutet auch das Preis/Leistungs-Verhältnis im Auge zu halten
- Vereinfachung des Managements und der Administration des HPC-System
- bessere Werkzeuge zur Planung einer optimierten Wartung
- u.v.m.

Gehäuselösungen

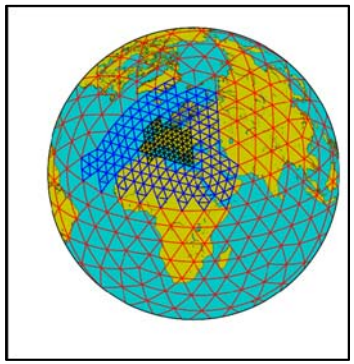
Kühlungslösungen

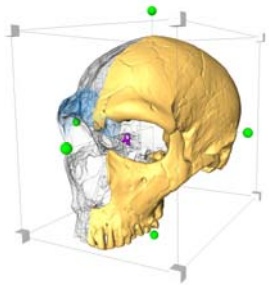
Lösungen zur Clusteradministration und für das Powermanagement (Software + Hardware)

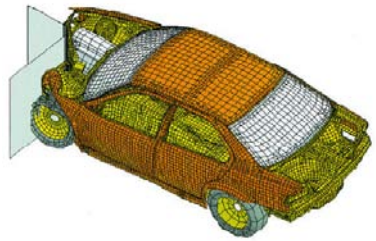
Agenda

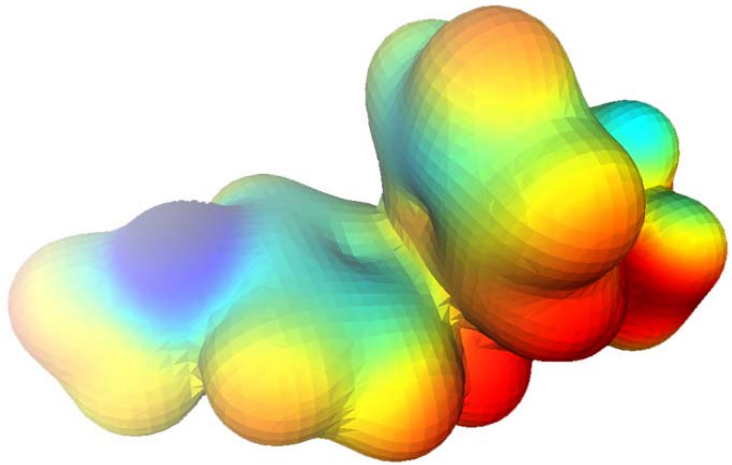
- Was ist High Performance Computing (HPC)
- Funktionsprinzipien des parallele Rechnens
- Anwendungsgebiete des parallelen Rechnens
- technologische Besonderheiten im Hochleistungsrechnen
- Forschung und Entwicklung von HPC-Systemen in Chemnitz
- Lösungen und Anwendungen für das High Performance Computing Ausgewählte Anwendungsmöglichkeiten
  - Erdsystemforschung – „Klimaforschung“
  - Neandertaler und hierarchische Matrizen
- Technologien im Supercomputing – das Handwerkszeug eines HPC Ingenieurs
- Widrigkeiten und offene Probleme
- Berufliche Zukunft in Chemnitz

Lösungen und Anwendungen für das High Performance Computing







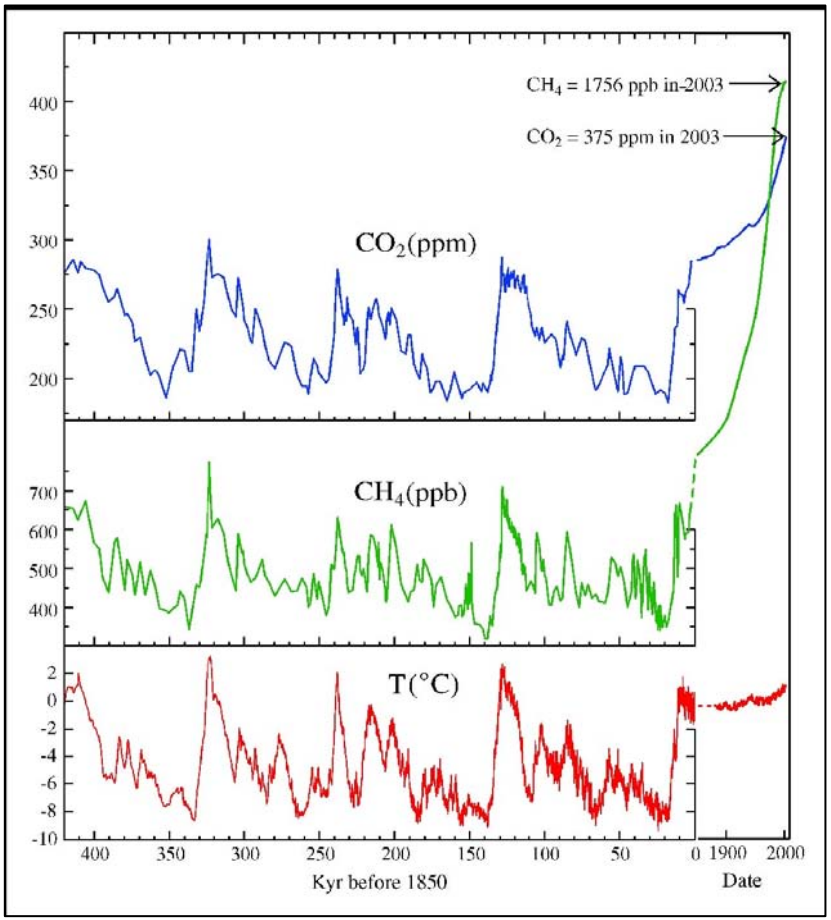


Febreze??

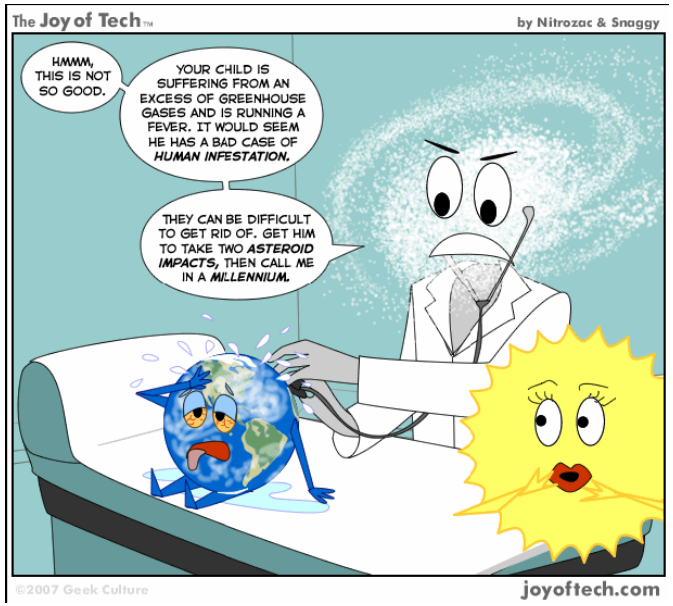
Agenda

- Was ist High Performance Computing (HPC)
- Funktionsprinzipien des parallele Rechnens
- Anwendungsgebiete des parallelen Rechnens
- technologische Besonderheiten im Hochleistungsrechnen
- Forschung und Entwicklung von HPC-Systemen in Chemnitz
- Lösungen und Anwendungen für das High Performance Computing Ausgewählte Anwendungsmöglichkeiten
  - Erdsystemforschung – „Klimaforschung“
  - Neandertaler und hierarchische Matrizen
- Technologien im Supercomputing – das Handwerkszeug eines HPC Ingenieurs
- Widrigkeiten und offene Probleme
- Berufliche Zukunft in Chemnitz

Erdsystemforschung – „Klimaforschung“



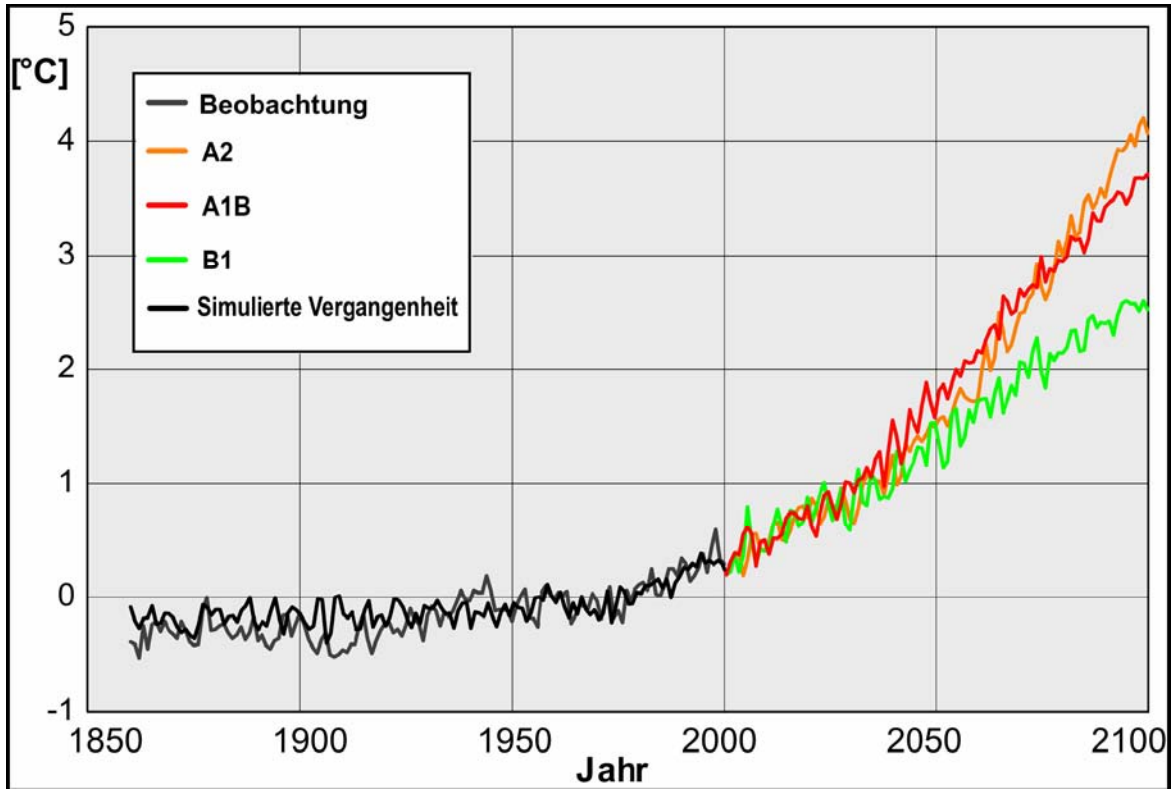
In den letzten 150 Jahren haben wir einige klimarelevante Faktoren massiv verändert.
Wie wird das System reagieren???



Quelle: Rainer Weigle – MPI für Meteorologie

Erdsystemforschung – „Klimaforschung“

IPCC* AR4 Resultate:

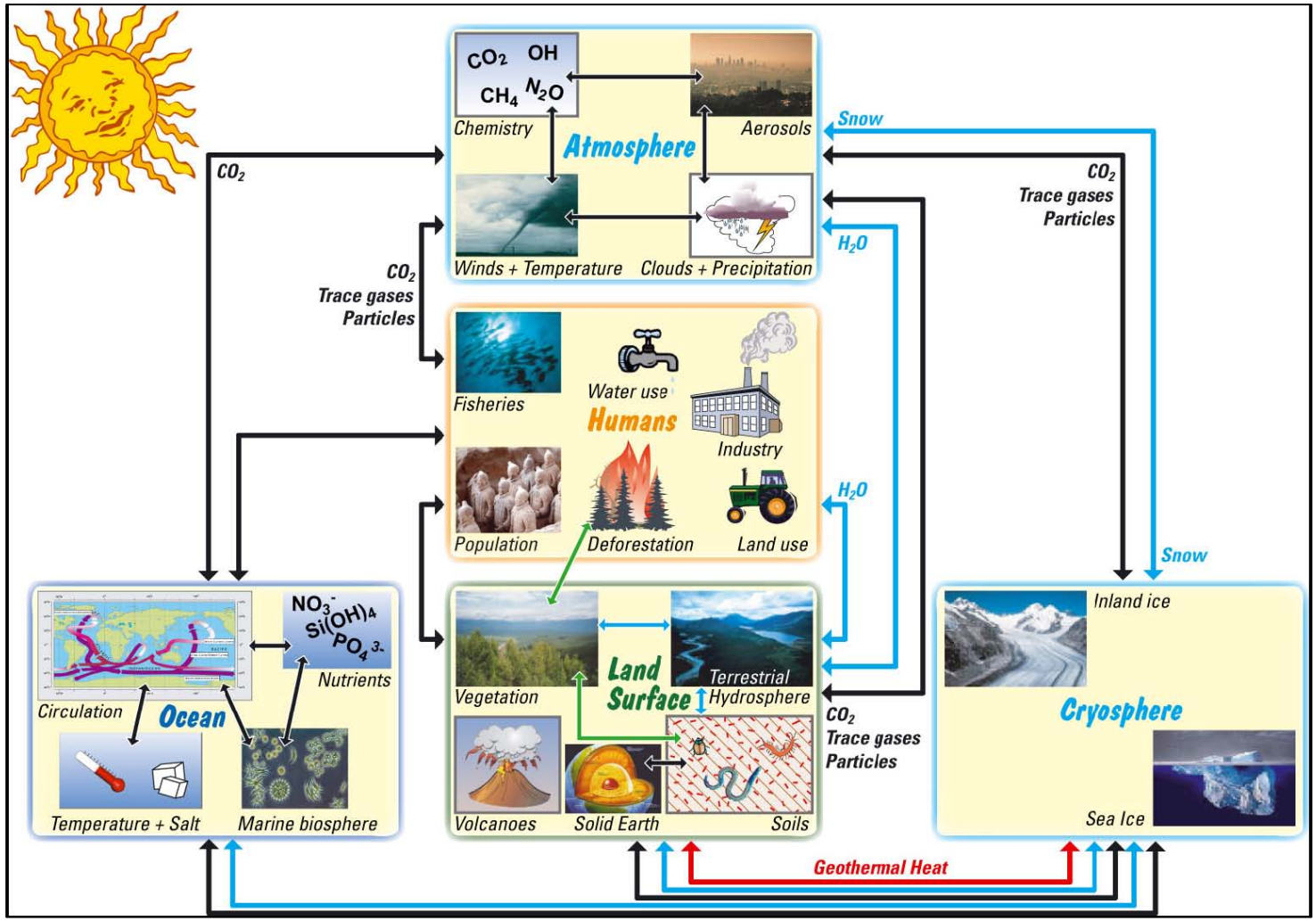


*Zwischenstaatlichen Ausschusses für Klimawandel der Vereinten Nationen - Intergovernmental Panel on Climate Change

Entwicklung der Globalen mittleren Temperatur im Vergleich zu dem Mittelwert des Zeitraums 1961 – 1990.

Quelle: Rainer Weigle – MPI für Meteorologie

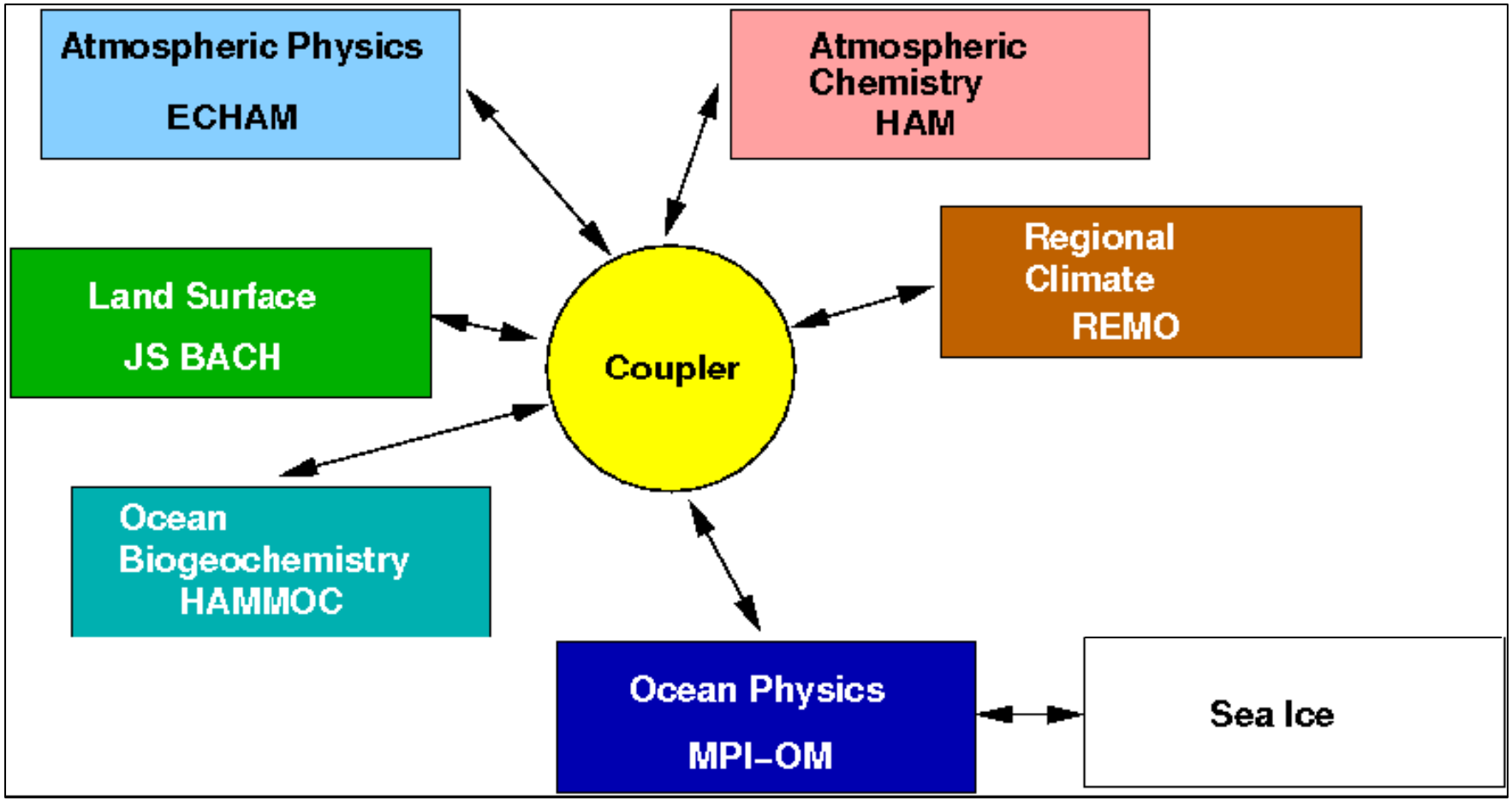
Erdsystemforschung – „Klimaforschung“



Quelle: Rainer Weigle – MPI für Meteorologie

Erdsystemforschung – „Klimaforschung“

Modelkomponenten



Quelle: Rainer Weigle – MPI für Meteorologie

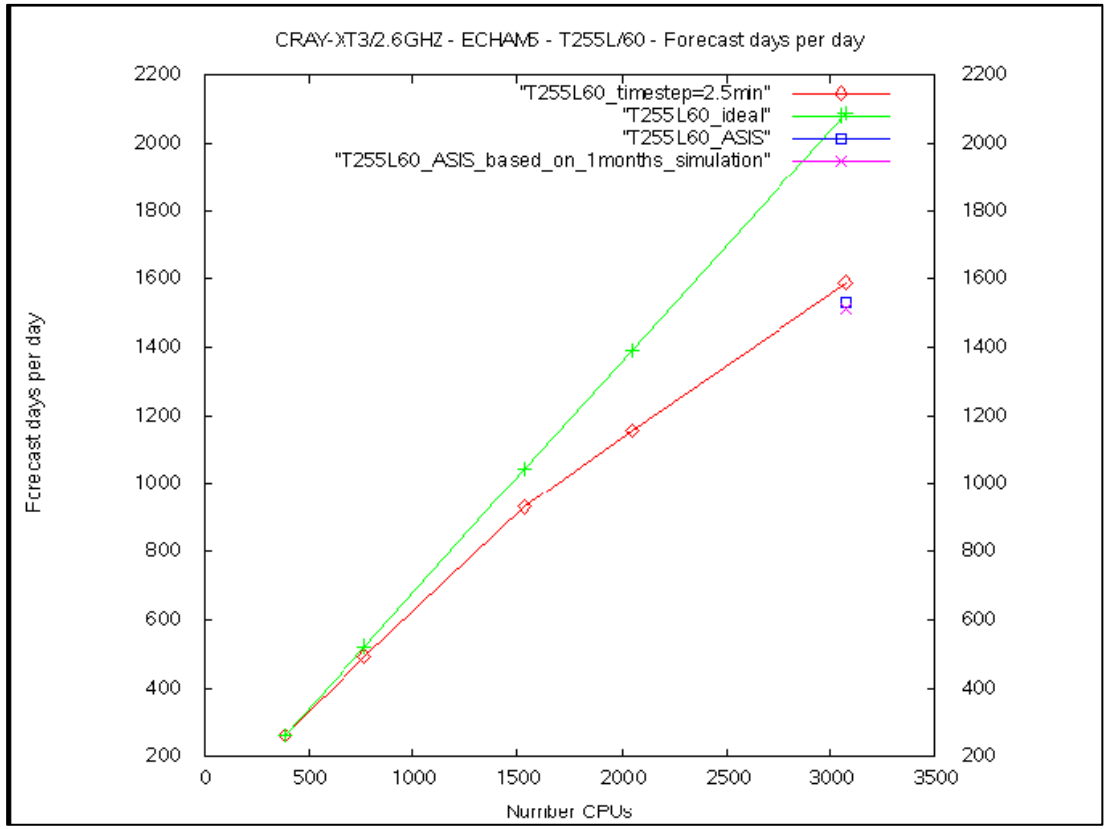
Erdsystemforschung – „Klimaforschung“

Rechenressourcen und Skalierungen am Beispiel des ECHAM5

- Skalierung auf > 3000 Prozesse
- Parallele Effizienz von 75 %
- 1.4 TFlops/s sustained
- ca. 15 % der Peak Performance

aber ...

- Skalierung ist von der Gitterweite abhängig
- I/O ist nicht berücksichtigt!
- ECHAM ist die fortschrittlichste Modellkomponente
- MPI-OM ist noch nicht für cache-basierte Maschine optimiert



Quelle: Rainer Weigle – MPI für Meteorologie

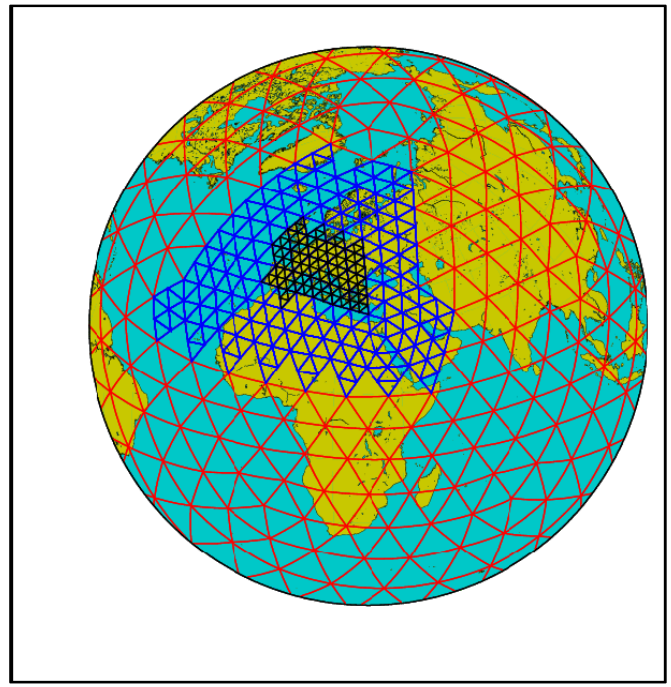
Erdsystemforschung – „Klimaforschung“

Petaflop Herausforderungen:

- Modelle sind nicht granular genug für Tausende von CPUs
- Die CPU Performance pro Core stagniert
- Balancierung zwischen den Modellkomponenten ist schwierig
- Skalierung des I/O
- Anforderungen an das Postprocessing

Petaflop Lösungen:

- Mehr hervorragende (Fortran-) Programmierer
- Entwicklung besserer Algorithmen
- Ensemble Rechnungen (mehrere Prognoseläufe)
- Enger Kontakt zu den Hardwareherstellern
- Auf bessere Technologie warten ;-)

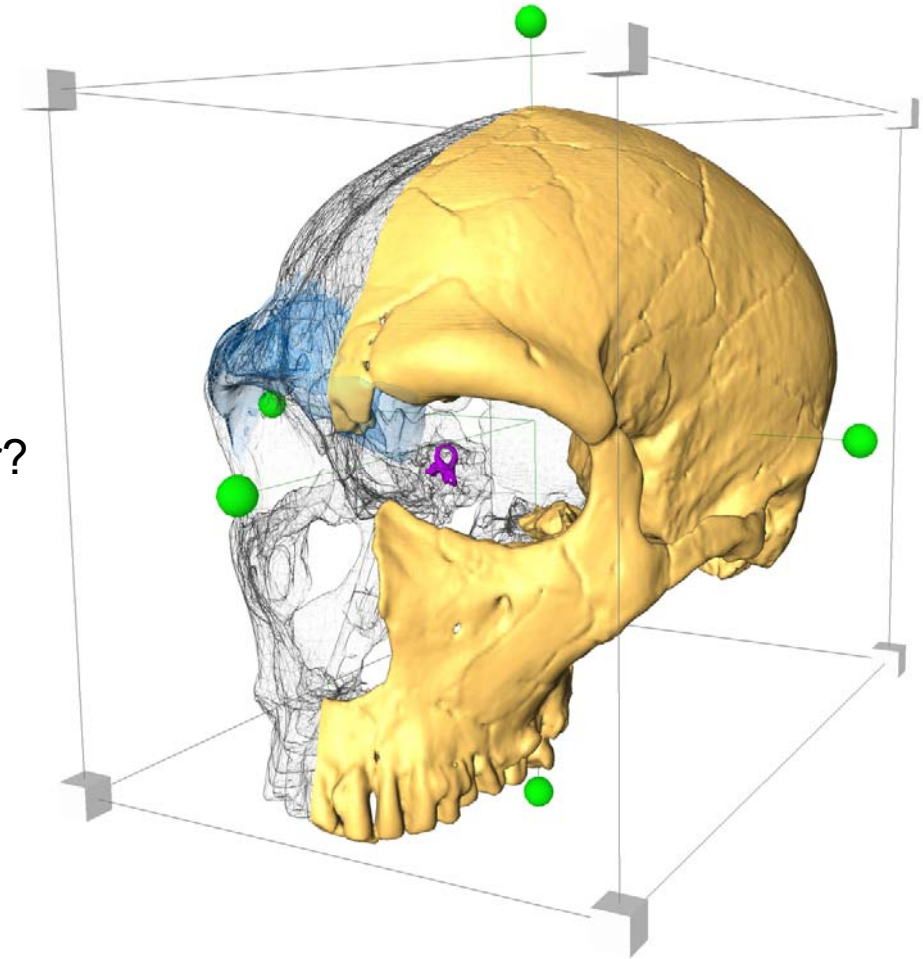


Quelle: Rainer Weigle – MPI für Meteorologie

Agenda

- Was ist High Performance Computing (HPC)
- Funktionsprinzipien des parallele Rechnens
- Anwendungsgebiete des parallelen Rechnens
- technologische Besonderheiten im Hochleistungsrechnen
- Forschung und Entwicklung von HPC-Systemen in Chemnitz
- Lösungen und Anwendungen für das High Performance Computing Ausgewählte Anwendungsmöglichkeiten
  - Erdsystemforschung – „Klimaforschung“
  - Neandertaler und hierarchische Matrizen
- Technologien im Supercomputing – das Handwerkszeug eines HPC Ingenieurs
- Widrigkeiten und offene Probleme
- Berufliche Zukunft in Chemnitz

Warum nicht wir?

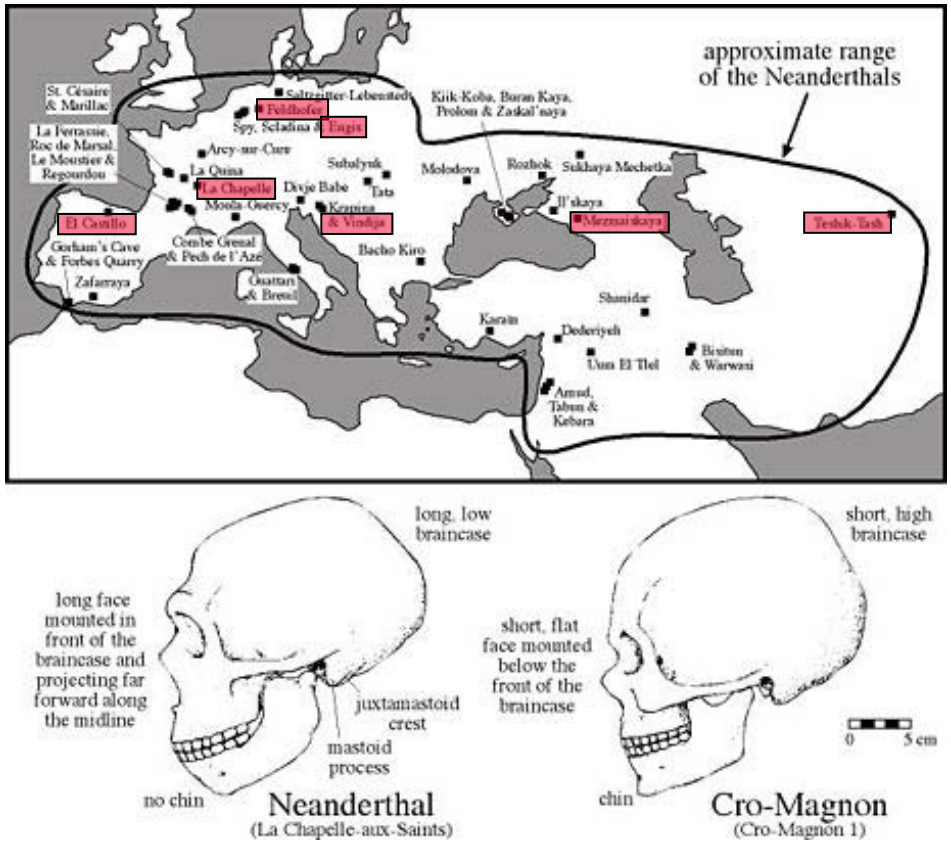


Neandertaler und hierarchische Matrizen

Quelle: Max-Planck-Institut für evolutionäre Anthropologie

Neandertaler und hierarchische Matrizen

Entschlüsselung des Neandertaler-Genoms



Verbreitungsgebiet der Fundstätten

Quelle: Max-Planck-Institut für evolutionäre Anthropologie

Neandertaler und hierarchische Matrizen

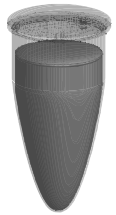
Entschlüsselung des Neandertaler-Genoms

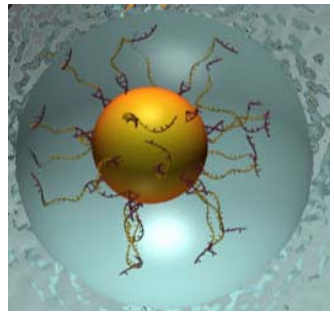
Entnahme von DNA - Resten im Reinraum

Quelle: Max-Planck-Institut für evolutionäre Anthropologie

Neandertaler und hierarchische Matrizen

Entschlüsselung des Neandertaler-Genoms auf einem Linux Cluster





DNA extract

454™ direct sequencing

NCBI

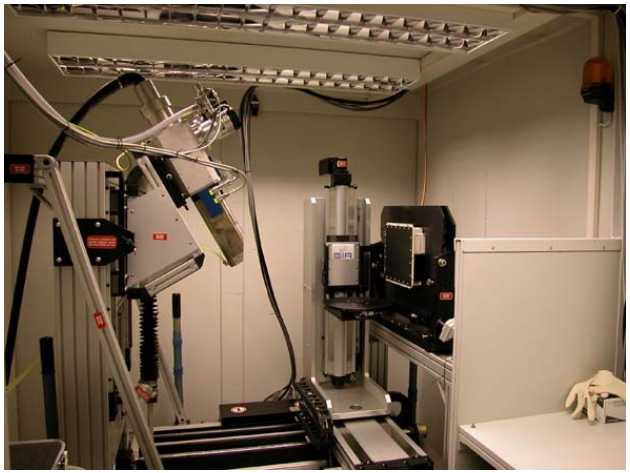
Sequence Database (70,000-250,000 sequences)

Alignment database ~20% of input sequences

Quelle: Max-Planck-Institut für evolutionäre Anthropologie

Neandertaler und hierarchische Matrizen

Portabler CT-Scanner

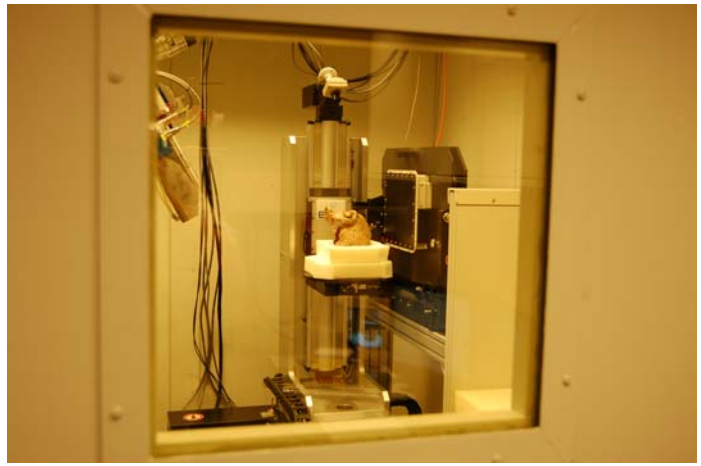


Quelle: Max-Planck-Institut für evolutionäre Anthropologie

Neandertaler und hierarchische Matrizen

Vor Ort Scans in aller Welt

Naturhistorisches Museum Zagreb



Archäologisches Museum Rabat

3D Visualisierung aus allen Fundstücken

Quelle: Max-Planck-Institut für evolutionäre Anthropologie

Neandertaler und hierarchische Matrizen

3D Print in Gips und Visualisierung

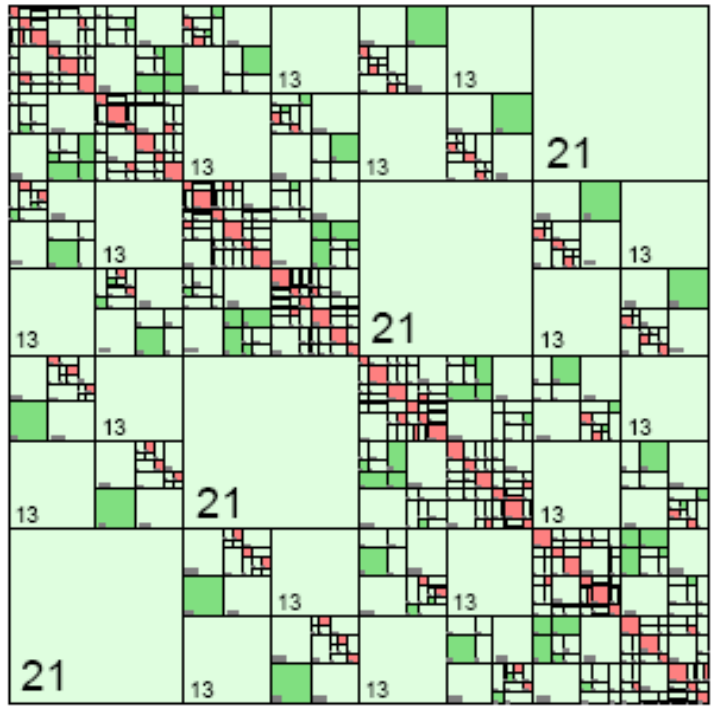
Quelle: Max-Planck-Institut für evolutionäre Anthropologie

Neandertaler und hierarchische Matrizen

Und hierarchische Matrizen? – Mathematik am:
Max-Planck-Institut für Mathematik in den Naturwissenschaften, Leipzig

- Im Wesentlichen erfolgt Grundlagenforschung, typischerweise auf dem Gebiet der hierarchischen Matrizen,
- selten auch anwendungsbezogene Forschung Hierarchische Matrizen

Beispiel
2000 × 2000 $\mathcal{H}$-Matrix für eine Integralgleichung:
Die Zahlen geben den Rang des jeweiligen Matrixblockes an, z.B. unten links Rang 21 anstelle von Rang 500.
Gesamtkompression: 92 % (41 MB statt 512 MB)
Beobachtung: Eine H-Matrix hat sehr viele Einzelblöcke mit unterschiedlichen Datenmengen (Rang).



(grün: datenschwach, rot: vollbesetzt)

Quelle: Max-Planck-Institut für evolutionäre Anthropologie

Neandertaler und hierarchische Matrizen

Hardware:

- 72 Knoten, je 2 Opteron 250 (2,4 GHZ), 4GB RAM
- 34 Knoten, je 2 Opteron 254 (2,8 GHZ), 16GB RAM
- Single Core Prozessoren
- Infiniband für 32 Knoten (Mellanox Gazelle 9600)
- 2 x Frontends
- 2 x Fileserver, 4 TB + 3 TB
- 2 x Gigabit-Netze (HP Procurve 5308XL)
- Admin-Netz (3com)

Software:

- Redhat Enterprise Linux V4 AS U3
- SUN Gridengine 6.0u8
- zwei Nutzergruppen: EVA, MIS

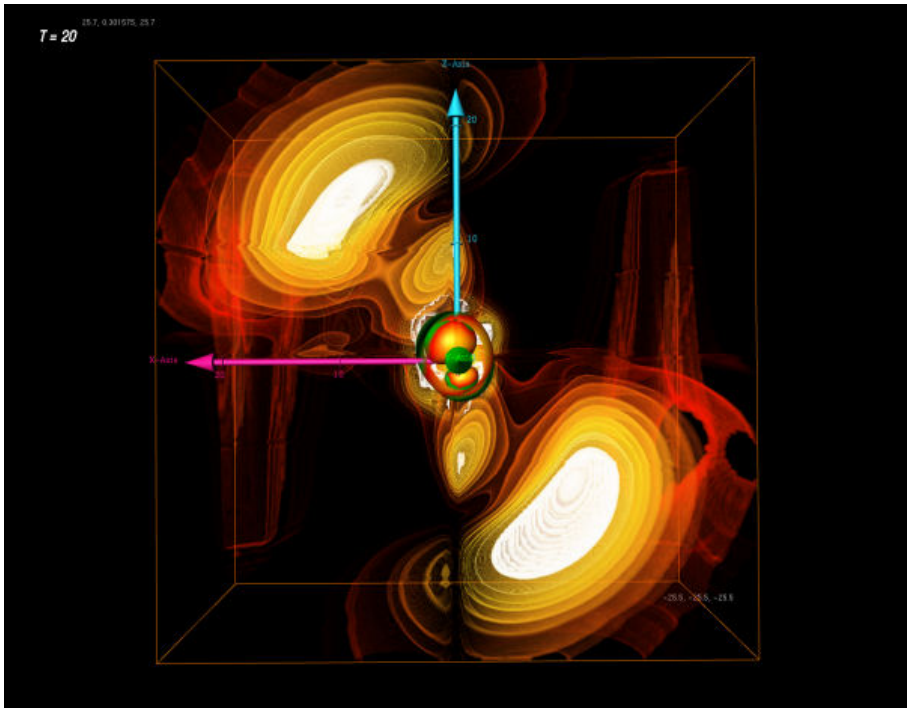
Cluster der beiden Nutzergruppen in Leipzig

Quelle: Max-Planck-Institut für evolutionäre Anthropologie

Weitere Beispiele

Visualisierung astrophysikalischer Prozesse

Albert-Einstein-Institut Golm (Potsdam)



Hochleistungsrechner DAMIANA

TOP500
192
2007

- Was passiert beim Zusammentreffen von zwei Schwarzen Löchern?
- Wie verschmilzt ein Neutronenstern mit einem Schwarzen Loch?

Weitere Beispiele

Schachspiel – Kombinatorik

- Hydra Maschine (16 / 32 CPUs; 16 x FPGA)
- sehr erfolgreich in internationalen Meisterschaften für Computer-Schach
- Mensch gegen Maschine / Maschine gegen Maschine
- Hydra rechnet ca. 8 Mio. mal schneller als menschliche Nervenzelle
- berechnet mehrere Mio. Zugkombination innerhalb einer Sekunde voraus
- **'Hydra is the Kasparov of computers'**

PAL Group - Abu Dhabi (VAE)

Agenda

- Was ist High Performance Computing (HPC)
- Funktionsprinzipien des parallele Rechnens
- Anwendungsgebiete des parallelen Rechnens
- technologische Besonderheiten im Hochleistungsrechnen
- Forschung und Entwicklung von HPC-Systemen in Chemnitz
- Lösungen und Anwendungen für das High Performance Computing Ausgewählte Anwendungsmöglichkeiten
  - Erdsystemforschung – „Klimaforschung“
  - Neandertaler und hierarchische Matrizen
- Technologien im Supercomputing – das Handwerkszeug eines HPC Ingenieurs
- Widrigkeiten und offene Probleme
- Berufliche Zukunft in Chemnitz

Technologien im Supercomputing

"Ich denke, dass es einen Weltmarkt für vielleicht fünf Computer gibt."
Thomas Watson, CEO von IBM, 1943

"Computer der Zukunft werden nicht mehr als 1,5 Tonnen wiegen."
US-Zeitschrift Popular Mechanics, 1949

"Es gibt keinen Grund dafür, dass jemand einen Computer zu Hause haben wollte."
Ken Olson, Präsident von Digital Equipment Corp., 1977

"640KByte sollten genug für jeden sein."
Bill Gates, Microsoft-Gründer,1981

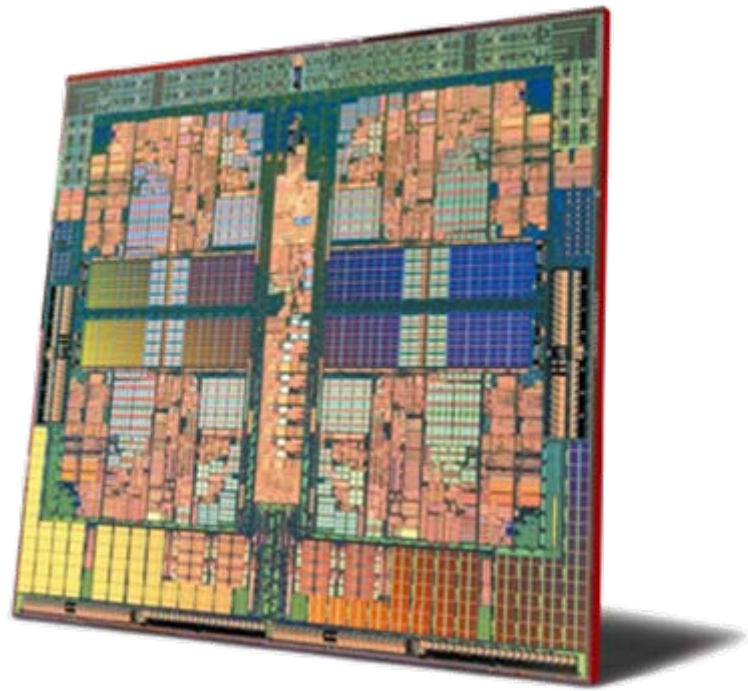
- Vorhersagen sind sehr schwierig
- die Grenzen des Machbaren sind ständig in Bewegung
- ABER: Miniaturisierung ist nicht beliebig erweiterbar (Quanteneffekte)

Technologien im Supercomputing

Einige ausgewählte Kriterien

Mehrfache Prozessorkerne (Multicore):

- klarer Trend zu vielen Kernen je CPU
- Verbindung der Kerne großes Problem:
  - Cache Kohärenz
  - hoher Energieverbrauch
- internationale Fachkreise prognostizieren:
  - Heterogene Sammlung von Cores:
  - Scalar-, Vector-, Multithreaded-,... Kerne in einem Chip



- ab dreistelliger Anzahl von Kernen => Verbindung der Kerne großes Problem

Technologien im Supercomputing

Einige ausgewählte Kriterien

Betriebssystem Virtualisierung:

- mehrere Betriebsysteme (Gäste) laufen („gleichzeitig“) auf einem Host
- wird von CPUs unterstützt: AMD-Virtualization, INTEL-VirtualizationTechnology
- Gäste werden vom Hypervisor (auf SW oder HW aufsetzend) gemanaged
- Unterscheidung in:
  - Paravirtualisierte (angepasste) Gäste
  - Unmodifizierte Gäste

- Beispiel: Virtual Appliance: JeOS (Just enough Operating System) [pronounced „Juice“]
  - Ubuntu basiertes OS
  - Auf das nötigste beschränkt => geringer Overhead
  - => einfaches Aufsetzen eines kleinen Webservers

Technologien im Supercomputing

Einige ausgewählte Kriterien

Grid-Computing:

- Analogie zum Stromnetz (Power-Grid) um die Jahrhundertwende
- viele Firmen (auch Haushalte) hatten eigenen Generator (heute: Cluster)
- man wollte aber Strom (Rechenpower):
  - überall verfügbar
  - einheitlicher Zugriff
  - relativ günstig
- Anbieter haben sich etabliert

zukünftig:

- viele Fachbereiche einer Uni haben Cluster, Superrechner, Pools (in d. Nacht)
- ==> Campus Grid
  - größere Aufgaben möglich
  - Resourcenauslastung, …
- Sun bietet On-Demand-Grid: http://network.com
  - „eine Stunde Rechenzeit“ kostet z.B. einen Dollar/CPU

Technologien im Supercomputing

Einige ausgewählte Kriterien

Green-Computing:

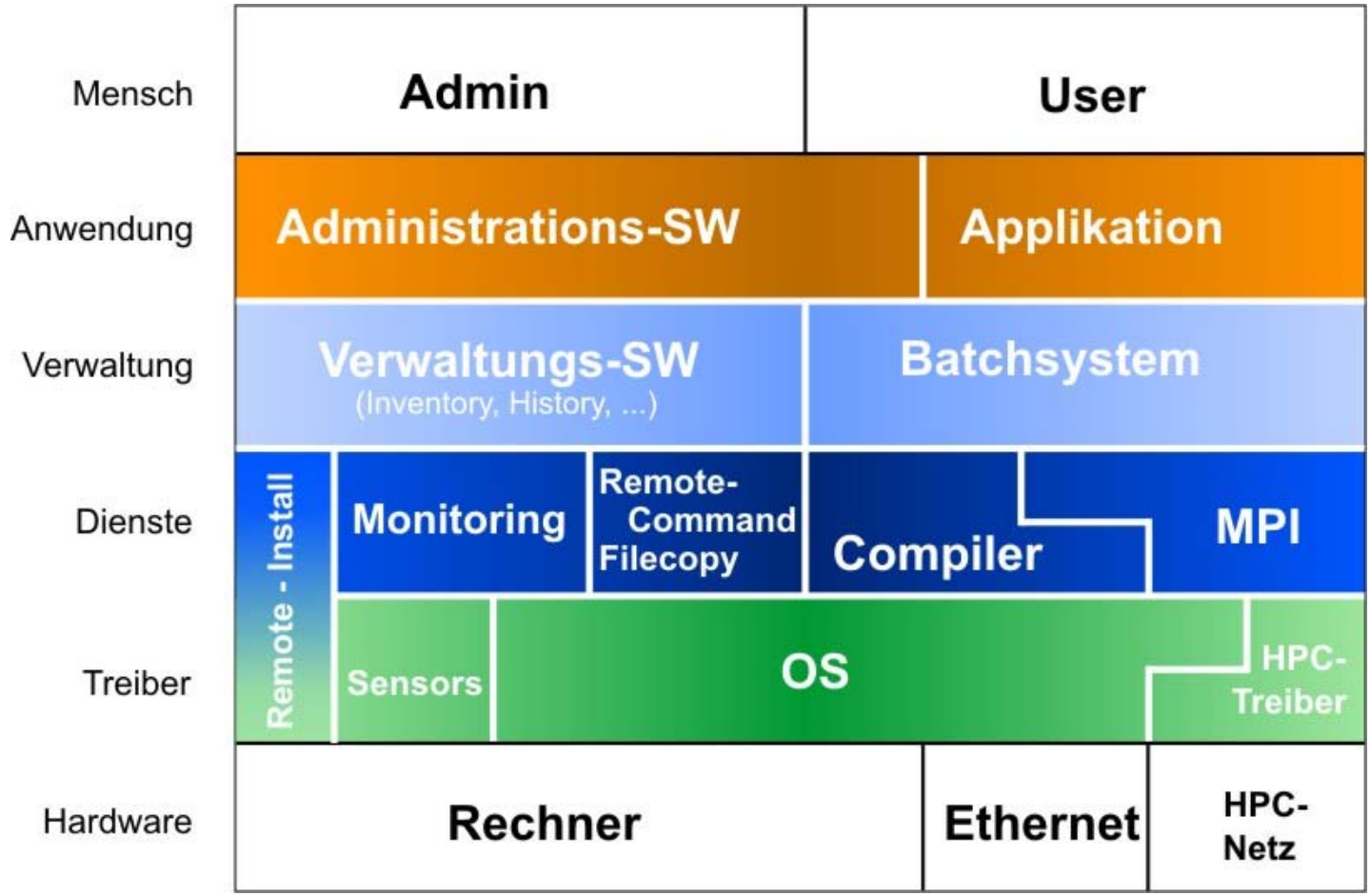
- Klimawandel wird immer dramatischer
- HPC verschlingt immer noch gewaltige Energiemengen
- Starke Nachfrage nach *High Efficiency/Low Voltage* Prozessoren

Beispiel: Earth Simulator (Japan)
- Rechenleistung ca. 36 TFLOP
- hat eigenes Kraftwerk
- verbraucht 8 Megawatt

Lösungansätze: Blue Gene /L (USA)
- Blue Gene /L: 280,6 TFLOP
  - verbraucht ‚nur' 500 Kilowatt
  - ca. 60 mal effizienter als Earth Simulator

Technologien im Supercomputing

Handwerkzeug eines HPC Ingenieurs



Typische Installation eines HPC Cluster

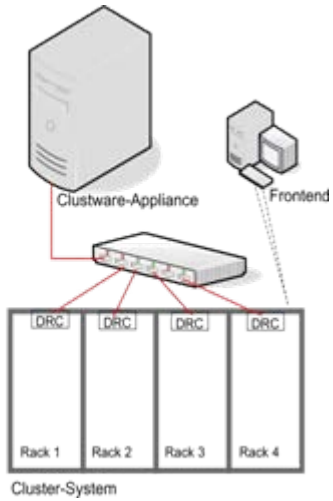
Technologien im Supercomputing

Handwerkzeug eines HPC Ingenieurs

- **64 bit CPU Architektur**
  - Intel Xeon MP u. DP, Itanium – Single bis Quad Core bis 4 Sockel
  - AMD Opteron – Single - Quad Core bis 8 Sockel
- **Software**

  Management/Batchsysteme
  - ClustWare Appliance und DRC von MEGWARE
  - Ganglia
  - SCALI Manage
  - LSF HPC – Workload Management von Platform
  - Torque mit Maui
  - SUN N1 Grid Engine 6 u. Cluster Tools
  - ParaStation Management 4 v. Partec
  - INTEL Cluster Tool Kit

MPI Bibliotheken:
MPICH, MPICH2, LAM-MPI, MICH G2, SCALI MPI, MPI/Pro, Intel MPI Library



Technologien im Supercomputing

Handwerkzeug eines HPC Ingenieurs

Compiler
- IntelCompiler 10.1 x64, IA64, EM64T, F95, C, C++, Cluster Tools
- PathScale Fortran 77/90/95, C,C++
- PGI Fortran 77/90 HPF, C,C++

Interprozessnetz
- Dolphin SCI – 10 Gbit/s
- Ethernet – 10 Gbit/s
- Infiniband – 20 Gbit/s
- Myrinet 10G – 10 Gbit/s
- Quadrics Elan II + III 20Gbit/s

The Portland Group™

PathScale™

ABAQUS
Superior Finite Element Analysis

FLUENT

CD-adapco

Technologien im Supercomputing

Handwerkzeug eines HPC Ingenieurs

Vierte-Generation von Myricom Produkten, eine Konvergenz von 10-Gigabit/s Ethernet und Myrinet

- Basierend auf 10-Gigabit Ethernet PHYs (layer 1)
- Standard 10-Gigabit Ethernet Kabel, (CX4) Kupfer und Fiber
- Myri-10G NICs Support für Ethernet und Myrinet Netzwerkprotokoll auf layer 2
- und, wie immer bei Myricom entlasten Firmware und Prozessor der Karten die CPU durch Kernelbypass
- Myri-10G Swiches haben volle Bisektionsbandbreite
- ein Mischbetrieb zwischen 10-Gigabit Myrinet und 10-Gigabit Ethernet ist möglich

Technologien im Supercomputing

MareNostrum Cluster in Barcelona. The central Myrinet-2000 switch has 2560 host ports. Photo courtesy of IBM.

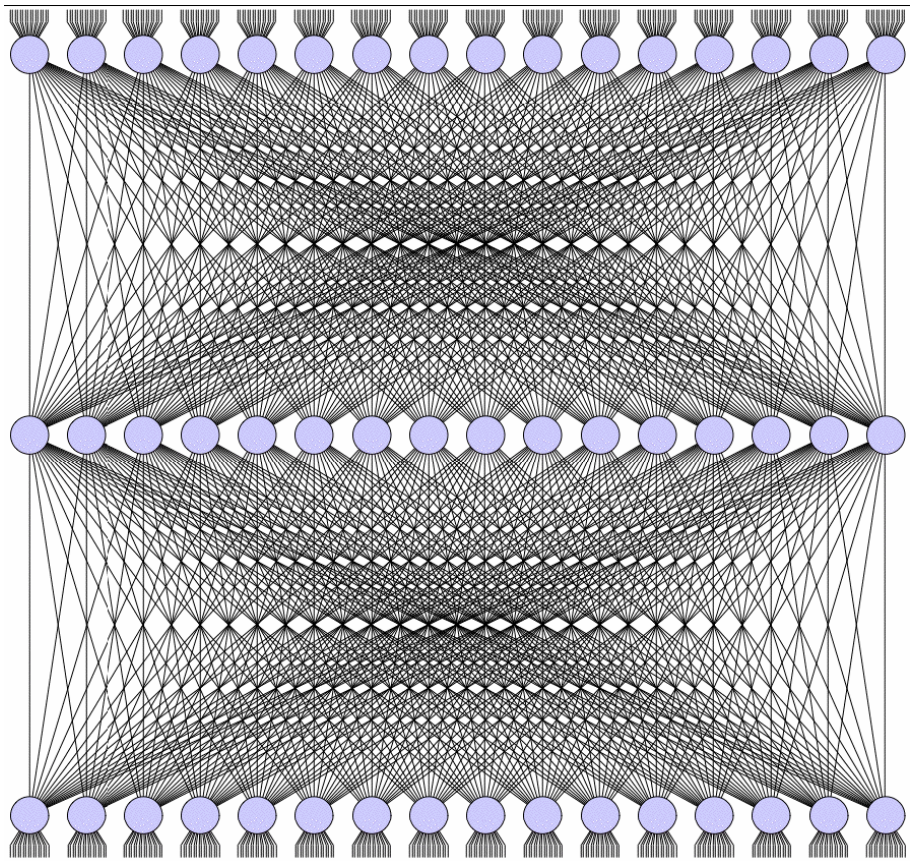
Technologien im Supercomputing

Handwerkzeug eines HPC Ingenieurs

Topology of the 512-Port Switch Network

Diameter = 3
Clos network
of 32-port
crossbarswitch
chips

A total of 48
10G_XBar32
chips



16 leaf switches
on 16 line cards

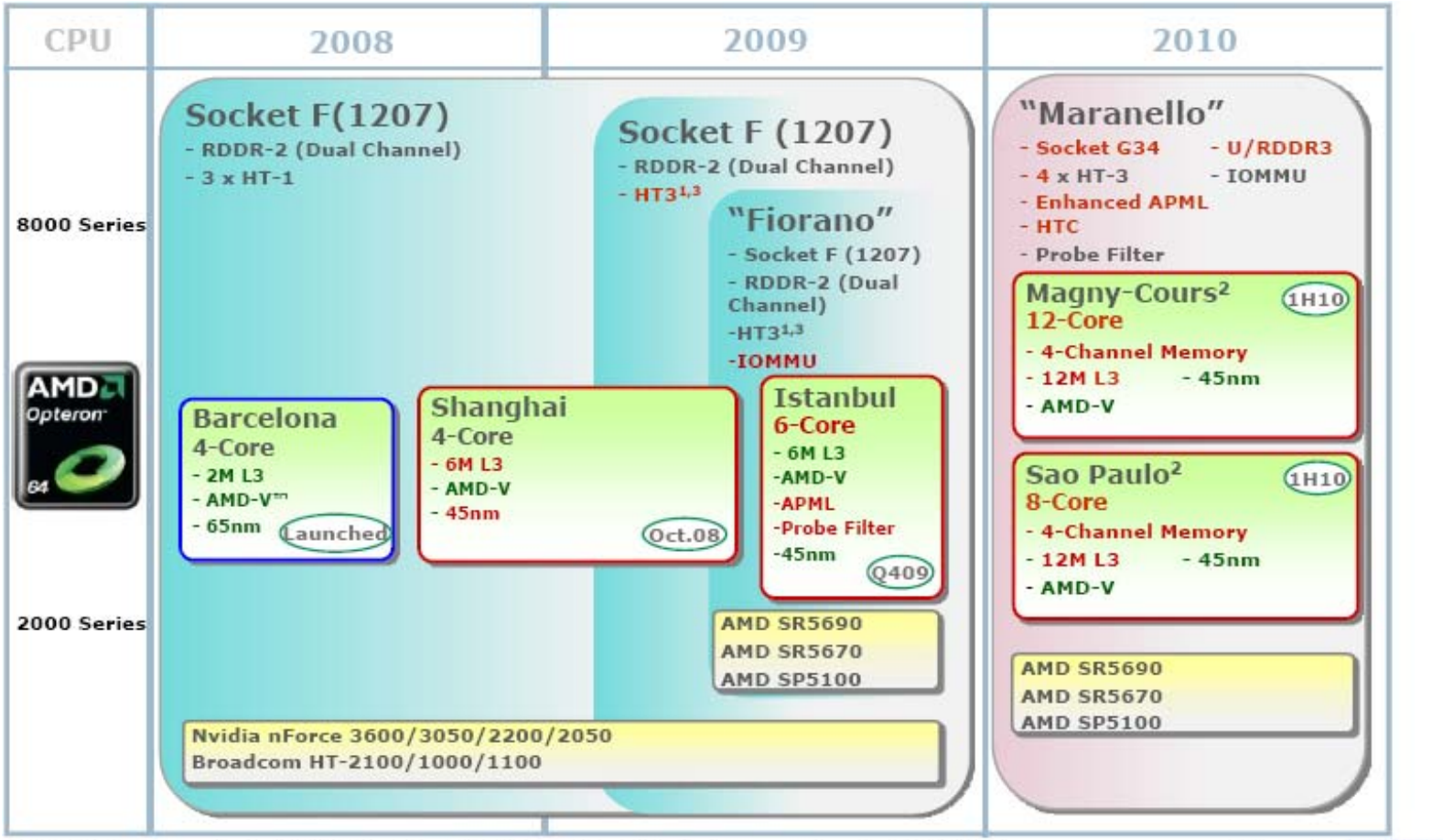
Clos spreader
network in the
backplanes

16 spine switches

Clos spreader
network in the
backplanes

16 leaf switches
on 16 line cards

Technologien im Supercomputing



[1] 2300 Series supports 1 cHT3 link (or 2 in dual-link mode). 8300 Series supports up to 3 cHT3 links. ncHT3 support dependant upon chipset.

[2] Up to 4 Socket Support Only [3] HT3 will be available on Shanghai processors in 1st half of 2009

Chipset 65 nm 45nm

Red type denotes new features

Technologien im Supercomputing

Handwerkzeug eines HPC Ingenieurs

Dedicated L1

- AMD's 64KB/64KB vs. Intel's 32KB/32KB
- Allows 2 loads per cycle

Handle Data Quickly and Efficiently.

Dedicated L2

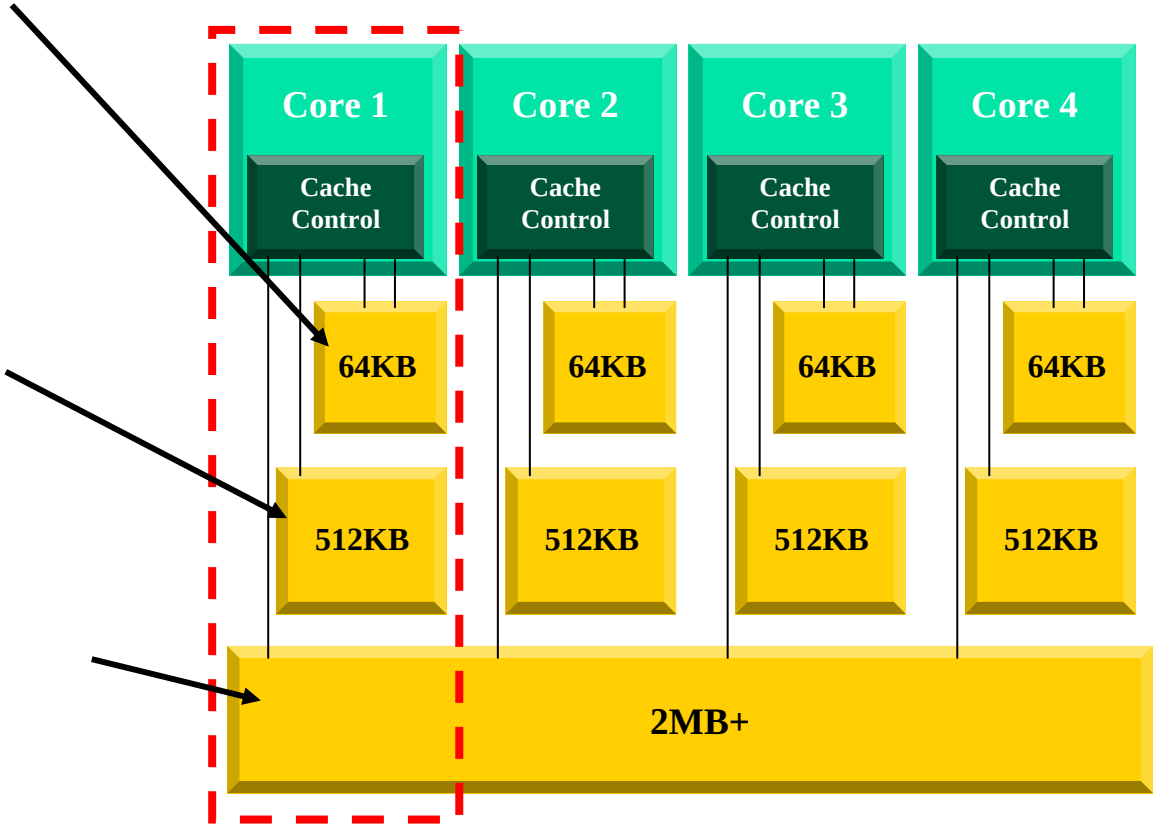
- Dedicated cache to eliminate conflicts of shared caches
- Designed for true working data sets

Avoid Thrashing. Minimize Latency.

Shared L3 - New

- Designed for optimum memory use and allocation for multi-core
- Ready for expansion at the right time for customers

Efficient memory handling reduces need for "brute force" cache sizes



Technologien im Supercomputing

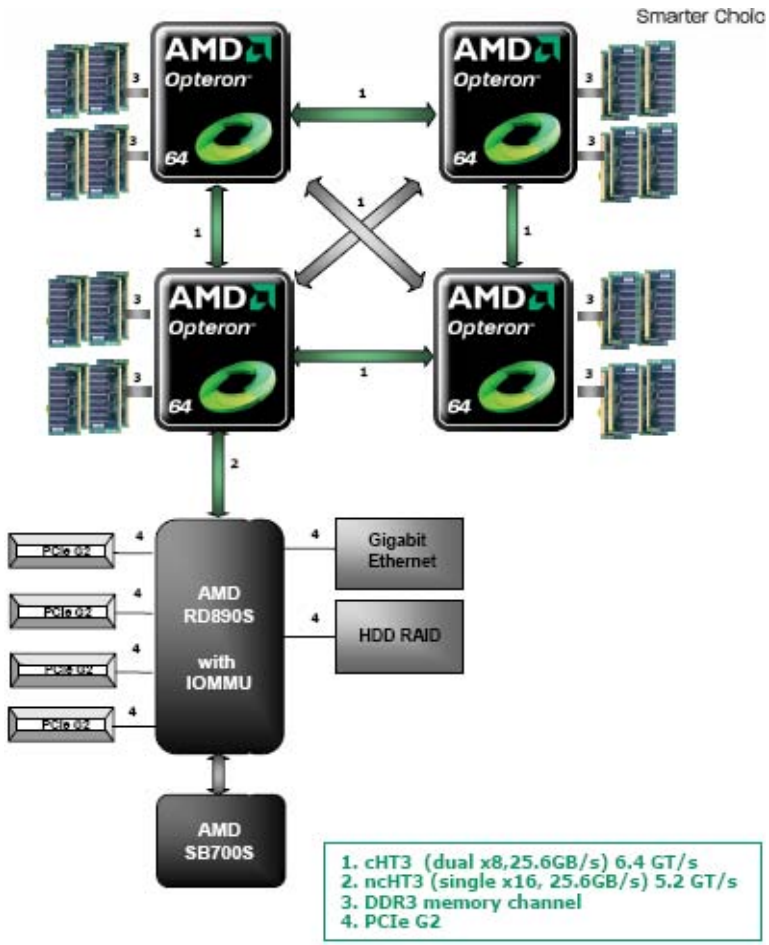
Handwerkzeug eines HPC Ingenieurs

1Q10: AMD Opteron™ - “Maranello”

Quad-core and Octal-core–Enhanced
- Cache Architecture (1M L2, 6M L3)

Memory
- Next generation DDR-3
- G3 Memory eXtender(G3MX)
- High bandwidth/low latency
- Dual channel DDR-3 (directly connected)
- Quad Channel Buffered DDR3

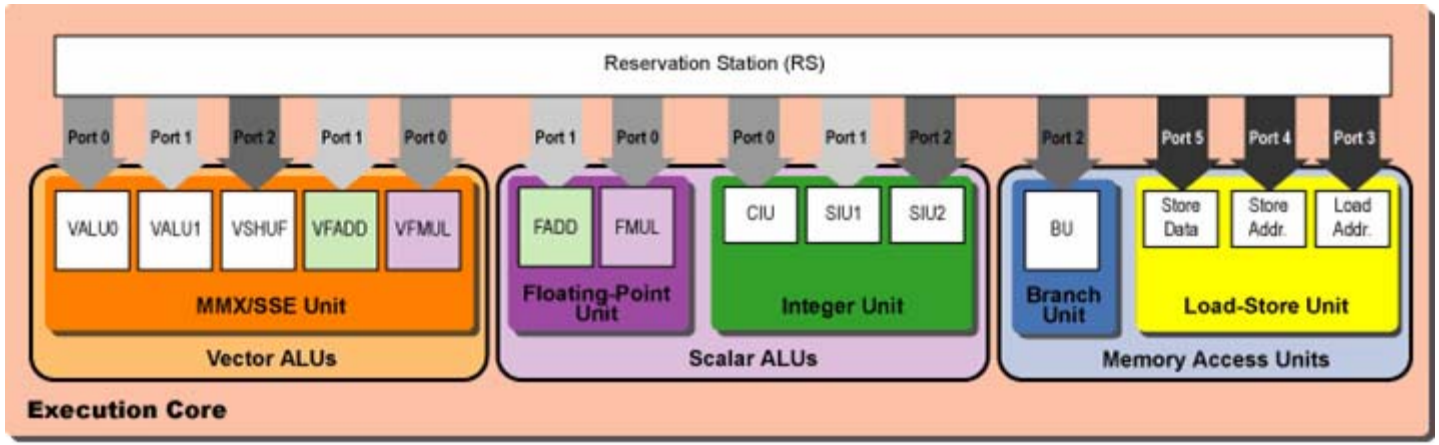
Interconnect
- HT-3 up to 25.6 GB/s (6.4 GT/s)
- 4 x 16-bit HT Links:
- Can be split to 8 x 8-bit links
- Fully connected 4P or 8P Configurations
- Either 2 directly connected channels
- Or4 buffered channels (G3MX)



Technologien im Supercomputing

Handwerkzeug eines HPC Ingenieurs

INTEL Architecture

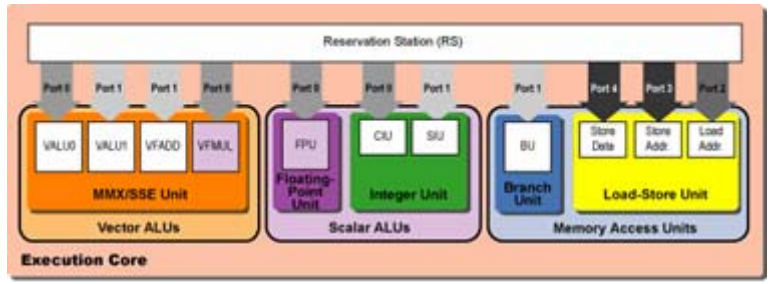


VALU0	VALU1	VSHUF	VFADD	VFMUL
-VALU	-VALU	-Vshuff	-VFadd	-VFmul
-Vmul	-Vshift	-Vrecip/ rsqrt	-Vmov	-VFdiv
		-Vmov		-Vsqrt
				-Vmov

FADD	FMUL
-Fadd	-Fmul
	-Fdiv

CIU	SIU1	SIU2
-ALU	-ALU	-ALU
-mul		
-shift		

Core



NetBurst

VALU0	VALU1	VFADD	VFMUL
-VALU	-VALU	-VFadd	-VFmul
-Vmul	-Vshift	-Vrecip/ rsqrt	-VFdiv
		-Vshuff	-Vsqrt
		-Vmov	-Vmov

FPU
-Fadd
-Fmul
-Fdiv

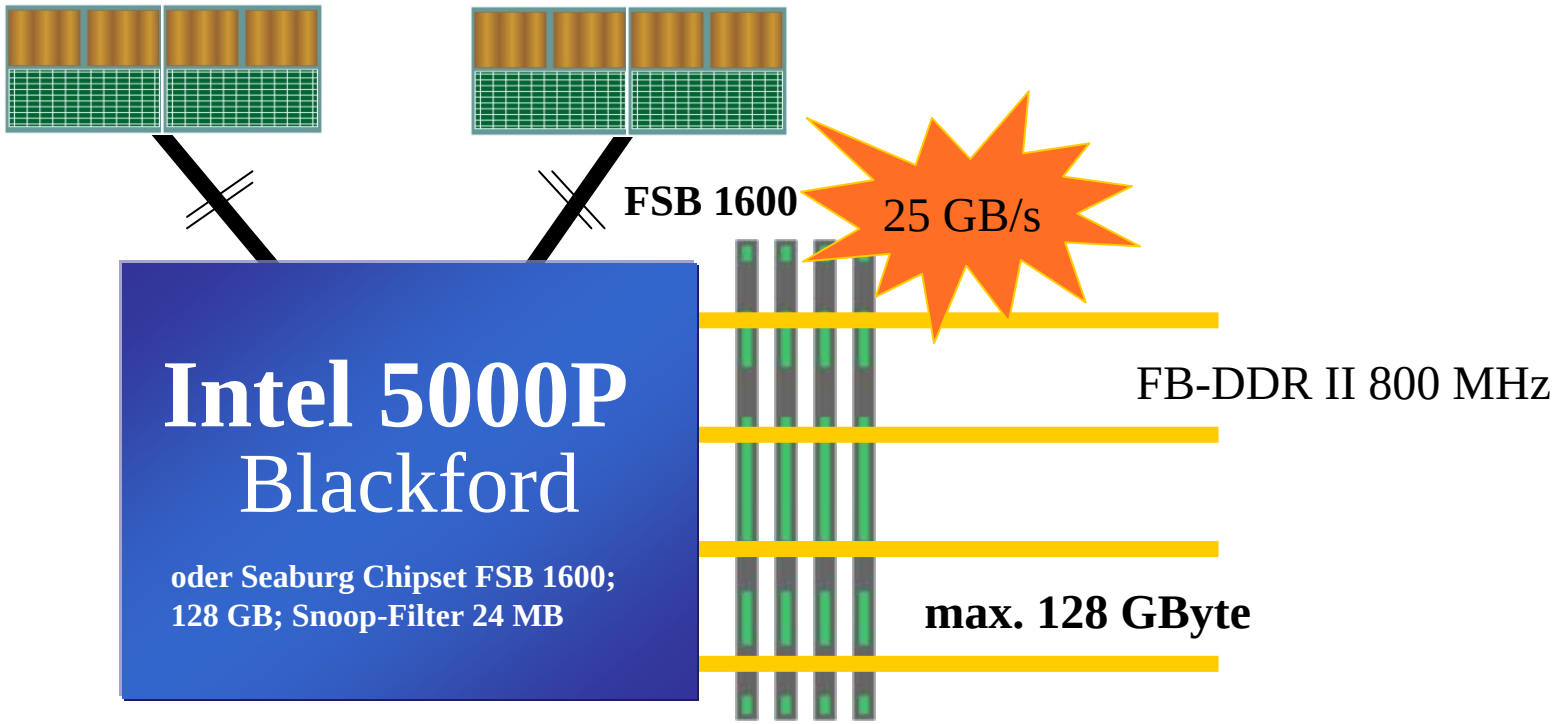
CIU	SIU
-ALU	-ALU
-mul	
-shift	

Technologien im Supercomputing

Handwerkzeug eines HPC Ingenieurs

INTEL Architecture

2 x Intel Xeon 5xx0, bis 3,2 GHz, 2 x 6 MB Cache



Technologien im Supercomputing

Handwerkzeug eines HPC Ingenieurs

Infiniband

Industriestandard in der vierten Generation

- Bandbreiten bis zu 40 Gbit/s
- Connect-X Latenzen ca. 1,5 µs
- neuste HCA sind Ethernet kompatibel
- universeller Interconnect mit Vielzahl an Protokollen:

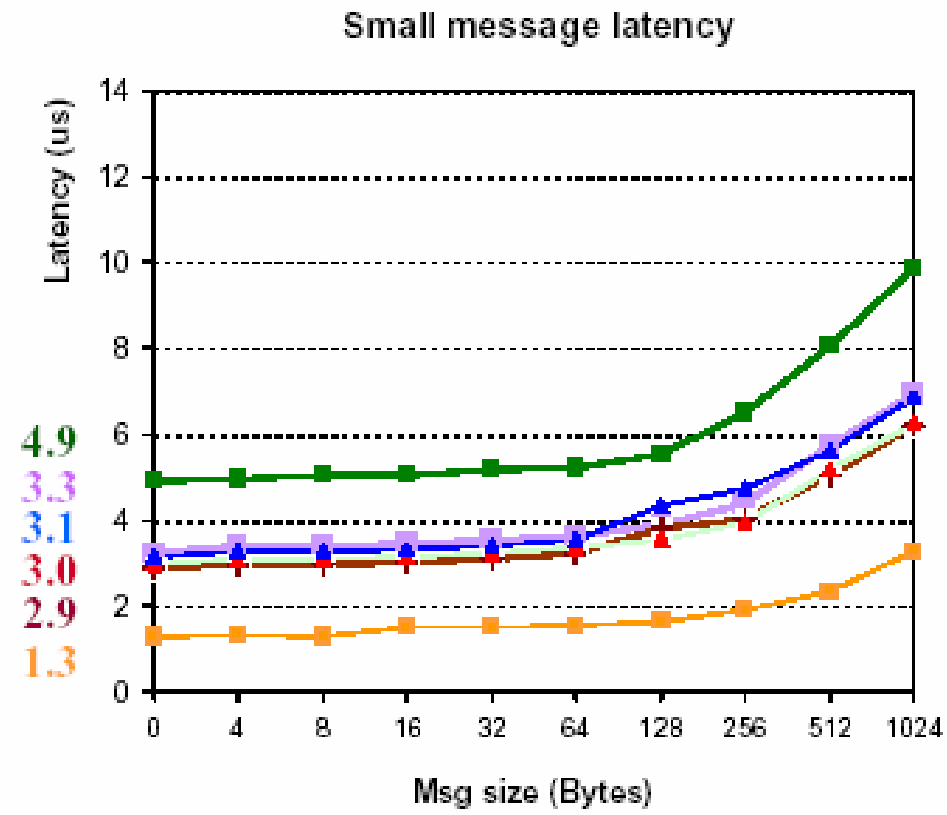
  MPI, IPoIB, iSER, SRP,SDP,DAPL

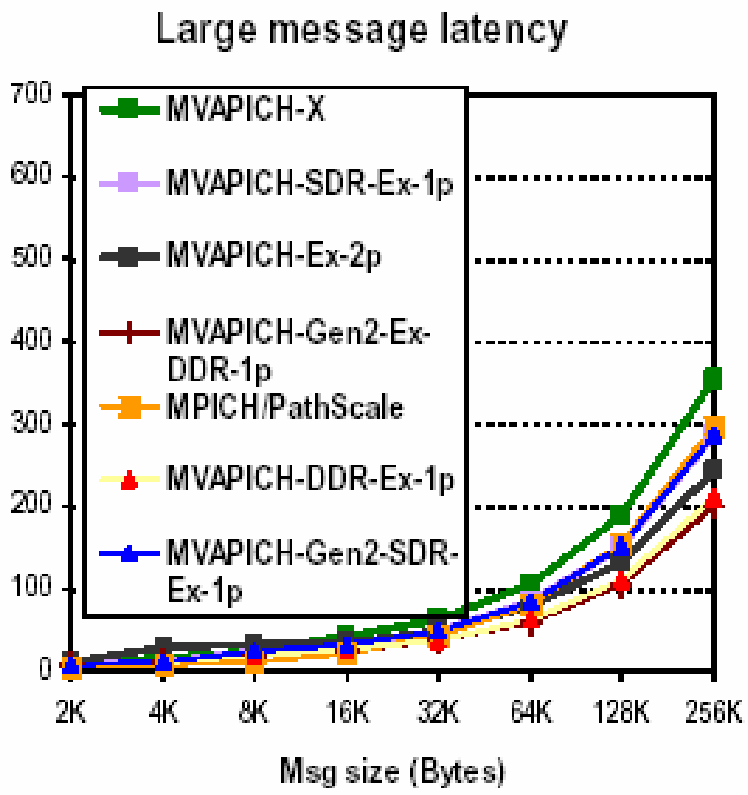
Myrinet & 10G Ethernet

- Myrinet 10G 10-Gigabit Ethernet und 10-Gigabit Myrinet Protokoll MPI
- Latenzen von ca. 1,8 µs

Technologien im Supercomputing

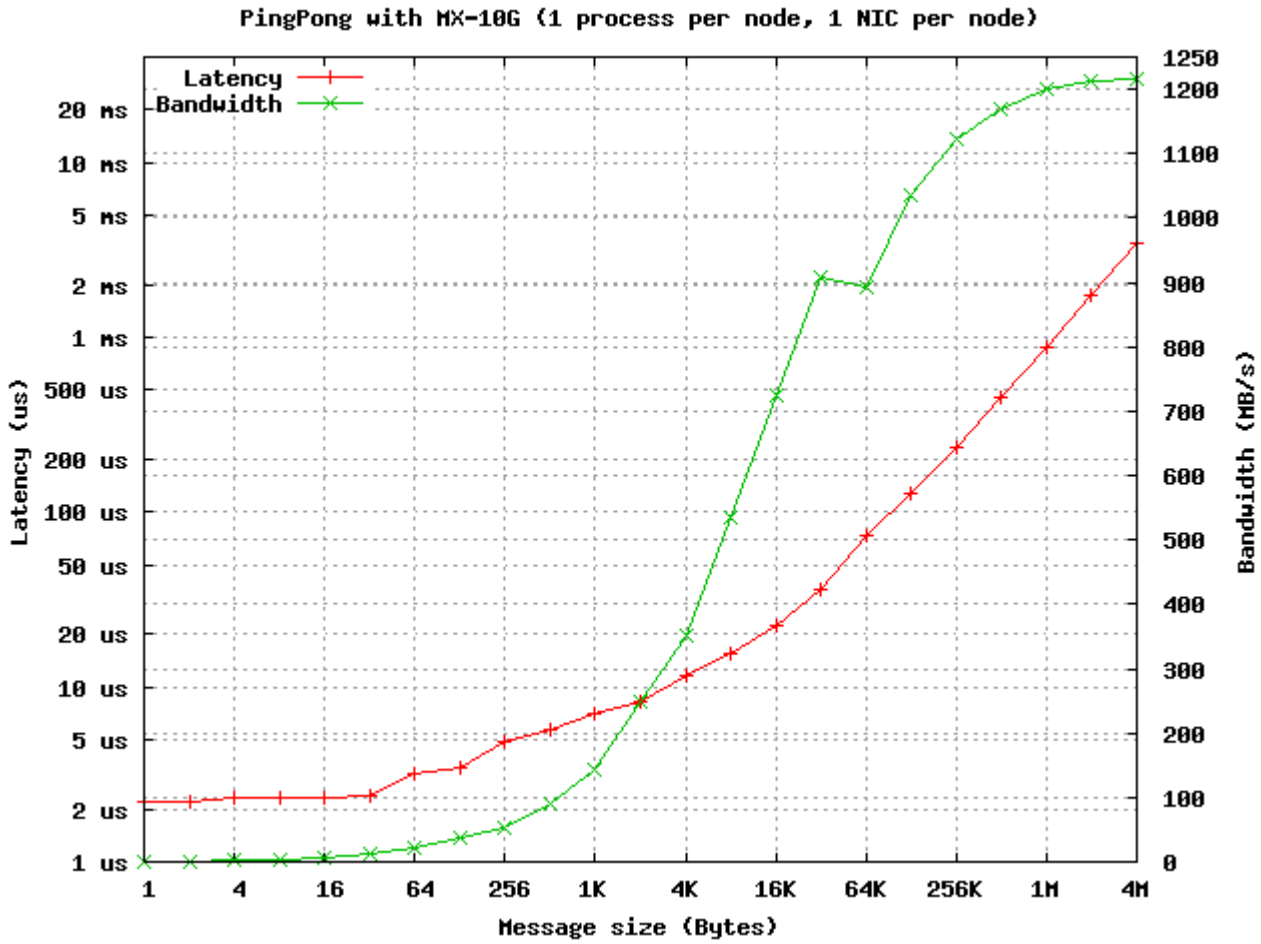
Handwerkzeug eines HPC Ingenieurs





Technologien im Supercomputing

Handwerkzeug eines HPC Ingenieurs



Agenda

- Was ist High Performance Computing (HPC)
- Funktionsprinzipien des parallele Rechnens
- Anwendungsgebiete des parallelen Rechnens
- technologische Besonderheiten im Hochleistungsrechnen
- Forschung und Entwicklung von HPC-Systemen in Chemnitz
- Lösungen und Anwendungen für das High Performance Computing Ausgewählte Anwendungsmöglichkeiten
  - Erdsystemforschung – „Klimaforschung“
  - Neandertaler und hierarchische Matrizen
- Technologien im Supercomputing – das Handwerkszeug eines HPC Ingenieurs
- Widrigkeiten und offene Probleme
- Berufliche Zukunft in Chemnitz

Widrigkeiten und offene Probleme

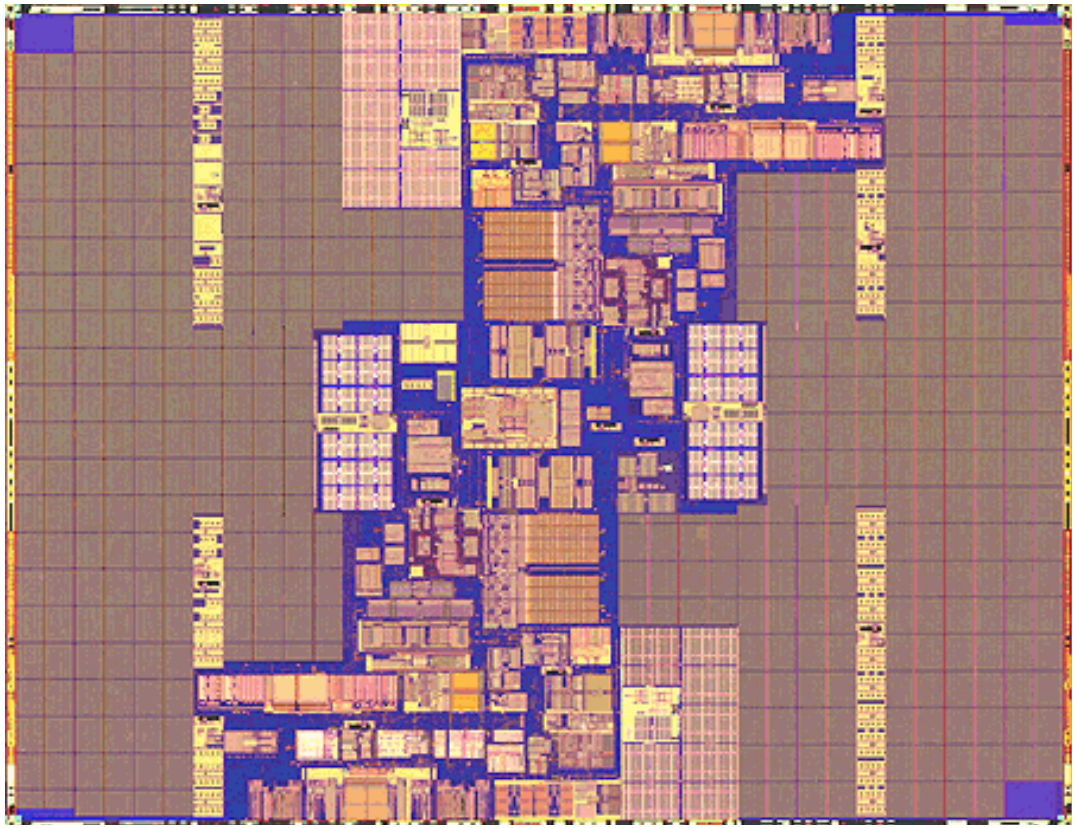
Management von mehreren Hundert bis Tausende CPUs

Management im GRID:
Starten von Jobs auf einer bestimmten Anzahl von CPUs (Batchsystem)
Installieren benötigter Applikationen/Pakete auf bestimmten Rechenknoten
Verwalten der Rechenknoten („Gesundheitszustand“)
Einordnen von Wartungen
Neuinstallation von einzelnen oder mehreren Knoten
Konfiguration der Knoten
Accounting
Verarbeiten von Monitoring-Daten
- Lüfterdrehzahlen,
- CPU-Temperaturen,
- Spannungen
- Lasten (Netz, CPU…)
- Laufende Prozesse
- Uvm.

→ Anzeige, Auswertung auch bei großen Mengen von Sensorwerten (viele Knoten…)
→ automatische Analyse dieser Werte mit Vorausschau für notwendige Wartungen (z.B. zu empfehlender Lüftertausch)
→ geeignetes System der Information an z.B. Admin o.ä.
Wird mit steigender Knoten-, CPU- und Core- Anzahl immer komplexer und komplizierter
Größte Herausforderung sehen wir in der effektiven Nutzung vom Multicore Architekturen

Widrigkeiten und offene Probleme

Größte Herausforderung effektive Nutzung vom Multicore Architekturen



Dualcore Itanium

Widrigkeiten und offene Probleme

Größte Herausforderung effektive Nutzung von Multicore Architekturen

Parallelrechner
in jedem Haushalt

Widrigkeiten und offene Probleme

Größte Herausforderung effektive Nutzung von Multicore Architekturen

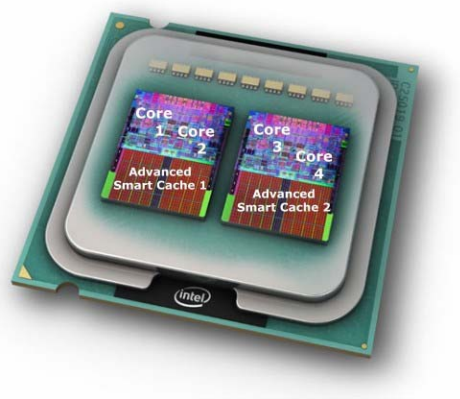
- Hersteller gehen dazu über, Chips mit mehreren integrierten Prozessorkernen zu entwickeln ==> Chip Multiprocessors (CMP).
- Prognose: In einigen Jahren hunderte Prozessorkerne auf einem Chip. ==> Rechenleistung wird wie bisher ansteigen.
- Rechenleistung verdoppelt sich alle 18 Monate.

- „Computer“ wandern in Chip!
- Neue Speicherhierarchien ==> Caches!
- Neue Verbindungstechnologien
- Drei Stufen Parallelismus:
  - On-chip
  - On-board
  - Cluster

Widrigkeiten und offene Probleme

Größte Herausforderung effektive Nutzung von Multicore Architekturen

- Neue Anforderungen an "Durchschnittsprogrammierer":
  - Fähigkeit, parallelen Code zu schreiben.
  - Wissen über:
    - Parallele Algorithmen,
    - Parallele Programmierkonzepte,
    - Parallele Programmiersprachen und -modelle,
    - Erfahrung mit Debuggern / Analysewerkzeugen für **parallelen Code**.

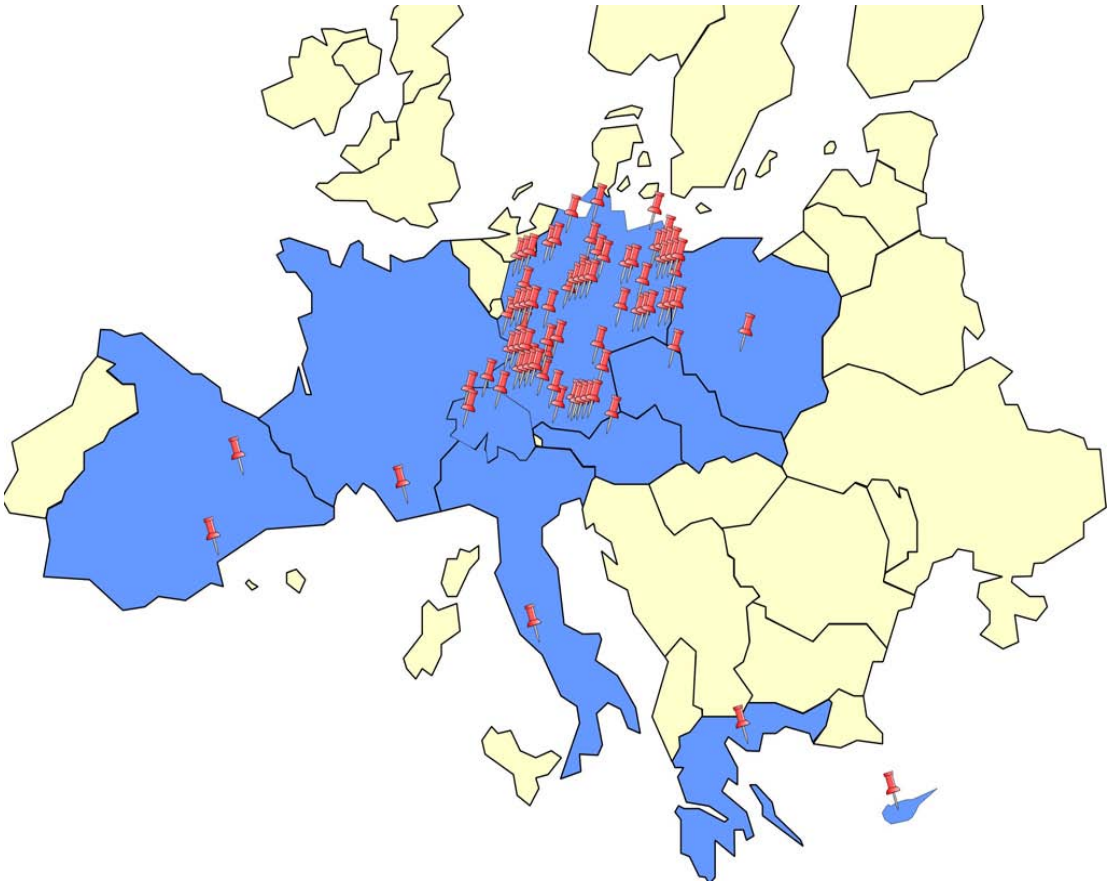


Haben Sie Interesse am High Performance Computing

?

Berufliche Zukunft bei MEGWARE

In 11 europäischen Ländern rechnet man mit MEGWARE.



MEGWARE Cluster
an der Universität Madrid

Berufliche Zukunft bei MEGWARE

Aus dem Arbeitsleben unserer HPC Ingenieure:
Hier Eindrücke der letzten Monate

UNI - Zaragoza 05.06

Cluster - Installation in Madrid

Cluster 2006 in Barcelona

Albert Einstein Institut Golm

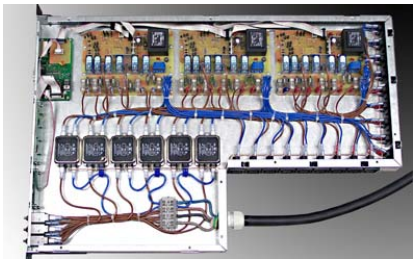
GeoForschungszentrum Potsdam

Fa. Hoffmann-LA Roche

TU Chemnitz, CHIC – Cluster mit IBM

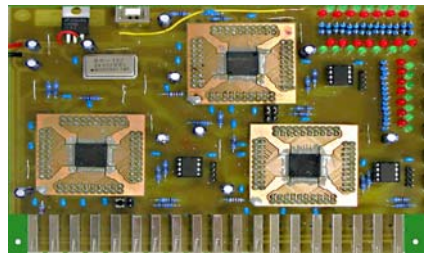
Berufliche Zukunft bei MEGWARE

MEGWARE eigene Entwicklungen im Sinne der Trends moderner Cluster – Architekturen, für mittlere und große Cluster



Formfaktor 0,5 HE - unser SlashFive

CoolNode: für heiße Tage

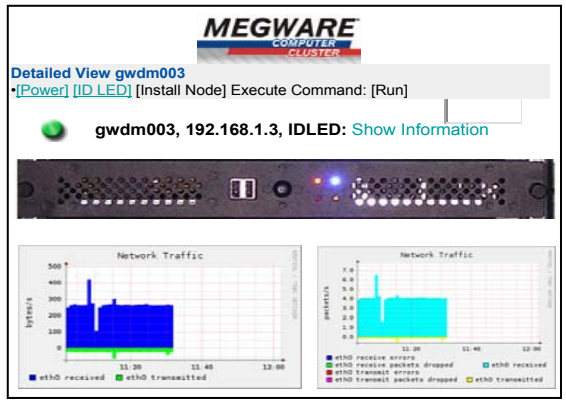


Erster Versuchsaufbau eines Switch für DRC

ClustSafe, I²C, Display, Bedienung per Tasten

SlashTwo zwei Höheneinheiten zwei Server

Unsere erste Management Lösung ClustWare bis V 3.0

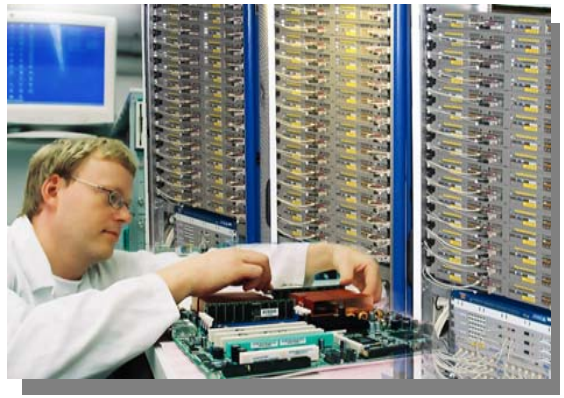


Ein neues Management mit Cluster Appliance und DirectRackControl

Berufliche Zukunft bei MEGWARE

Wir suchen: aus der Fakultät Informatik oder artverwandten Fachbereichen

- Werksstudenten
- Praktikanten
- Diplomanden



Wir bieten: interessante und anspruchsvolle Entwicklungsaufgaben

+

den Einstieg in eine sichere berufliche Zukunft

Lernen Sie die Welt kennen…

und

bleiben Sie …

… in Chemnitz ☺

Vielen Dank für Ihre Aufmerksamkeit

Jürgen Gretzschel

MEGWARE Computer GmbH
Vertrieb und Service
Tel 03722 528 85
Fax 03722 528 15

E-Mail juergen.gretzschel@megware.com

http://www.megware.com