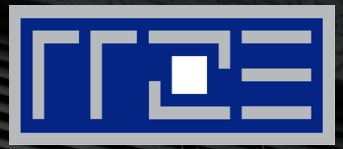
REGIONALES RECHENZENTRUM ERLANGEN [RRZE]



High Performance Computing

Systemausbildung − Grundlagen und Aspekte von Betriebssystemen und systemnahen Diensten, 21.06.2017

HPC-Gruppe, RRZE

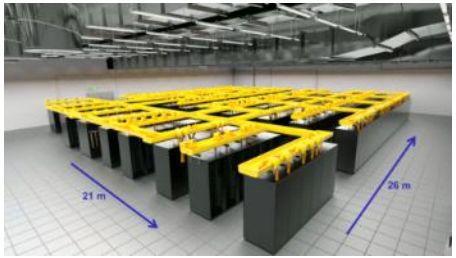
FRIEDRICH-ALEXANDER
UNIVERSITÄT
ERLANGEN-NÜRNBERG

Agenda

- Was bedeutet HPC?
- Aus welchen Komponenten besteht ein typisches HPC-System?
- Mit dem HPC-System läuft $GTA5/$Photoshop/$Videoschnittprogramm doch superschnell, oder?
- Shared-Memory-Parallelisierung vs. Message Passing
- Was ist dieses Queuing-System und wozu ist es gut?

Was bedeutet HPC?

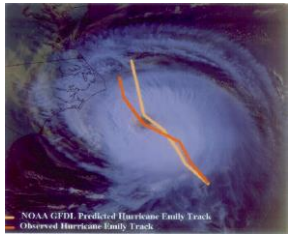
- Englisch: High Performance Computing
- Deutsch: **Hochleistungsrechnen**
- “Hochleistungsrechnen ist ein Bereich des computergestützten Rechnens. Er umfasst alle Rechenarbeiten, deren Bearbeitung einer hohen Rechenleistung oder Speicherkapazität bedarf.“ (Wikipedia)

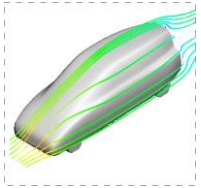


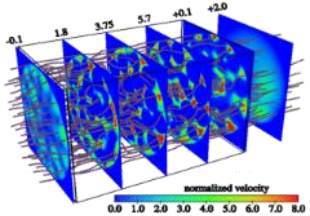
- genaue Abgrenzung ist schwierig

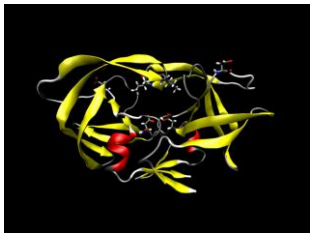
Einsatzbereiche von HPC-Systemen (Beispiele)

- Ingenieurwesen (Automotive)
  - Crashsimulation
  - Aerodynamik
  - Akustik
- Meteorologie
  - Wettervorhersage
  - Katastrophenvorhersage
- Biologie/Medizin
  - “Drug Design”
  - Aufklärung von Reaktionsprozessen









Einsatzbereiche von HPC-Systemen (Beispiele)

- Werkstoffwissenschaften
- Physik
  - Aufklärung von fundamentalen Wechselwirkungen
  - Struktur der Materie
- Filmindustrie
  - Rendering
  - CGI

- Zunehmend auch außerhalb der klassischen Natur-/Ingenieurs-wissenschaften, z.B. in Wirtschafts- oder Sprachwissenschaften

Ein Wort zur Warnung

Fricke / Lehmann: Der Rechenschieber
VEB Fachbuchverlag Leipzig 1961
7. Auflage

Man kann sich aber auch zu sehr an den Gebrauch des Rechenschiebers gewöhnen. Dann werden auch die einfachsten Berechnungen, die sonst spielend, fast ohne nachzudenken, im Kopf erledigt werden, unbedingt mit dem „Schieber" erledigt. Bei diesen Leuten ist 3 mal 3 auf dem Schieber etwa 8,99. Es erübrigt sich wohl, erst zu sagen, daß diese Methode nicht ganz die richtige ist. Grundsätzlich soll der Rechenschieber das Kopfrechnen *nicht ersetzen*; denn dieses erhält den Geist beweglich und ist außerdem beim Gebrauch des Rechenschiebers unbedingt erforderlich.

```
s/Rechenschieber/Supercomputer/g

s/Kopfrechnen/Common Sense/g
```

HPC-Systeme

- “Hochleistungsrechner sind Rechnersysteme, die geeignet sind, Aufgaben des Hochleistungsrechnens zu bearbeiten.”

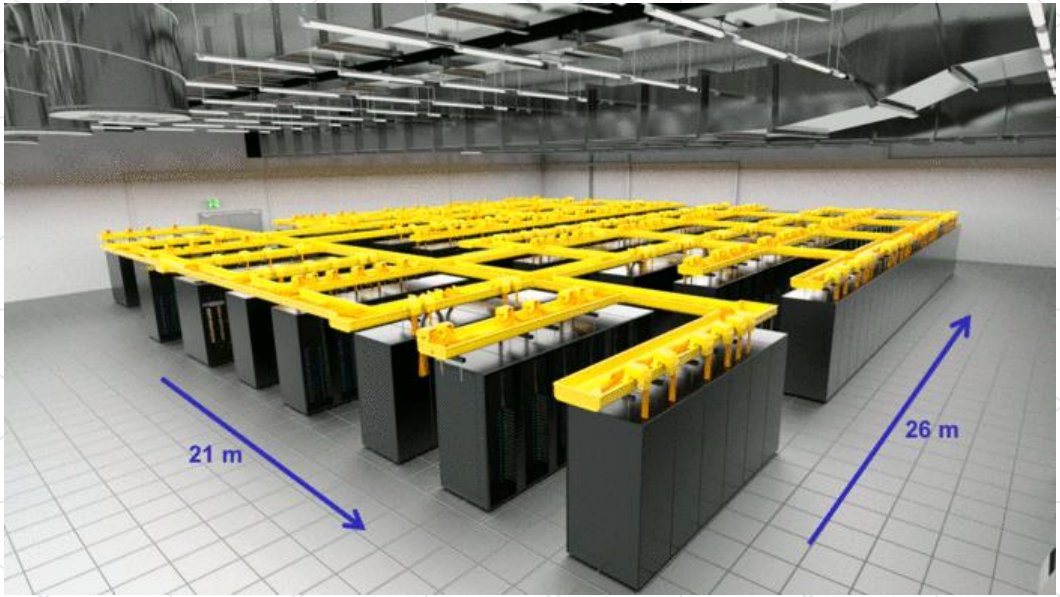
- Früher: Spezialsysteme – Spezialhardware + -software
- **Heute: Fast ausschließlich Cluster aus Standard-Hardware**, meist mit schnellem Interconnect (Netzwerk)
  - Die einzelnen Clusterknoten sind nicht schneller als ein gut ausgestatteter Standardserver!

- Vektorrechner: Nischenprodukt

Beispiel: SuperMUC Phase 1 (2012) @ LRZ

- **9216 compute nodes**
  - 2x Intel Xeon E5-2680 CPU (8 Kerne, 2.7 GHz, „Sandy Bridge“)
  - 32 GB RAM
  - Ergibt in Summe 147456 Kerne und 288 TB RAM
- Infiniband-Netz
  - 18 Inseln a 512 Knoten
    - Fully non-blocking innerhalb der Inseln
    - ausgedünnt 4:1 zwischen den Inseln
  - FDR-10 (knapp 40 Gbit/s)
- 15 PB paralleles Filesystem @250 GByte/s

“SuperMUC” Phase 1 (rendering)

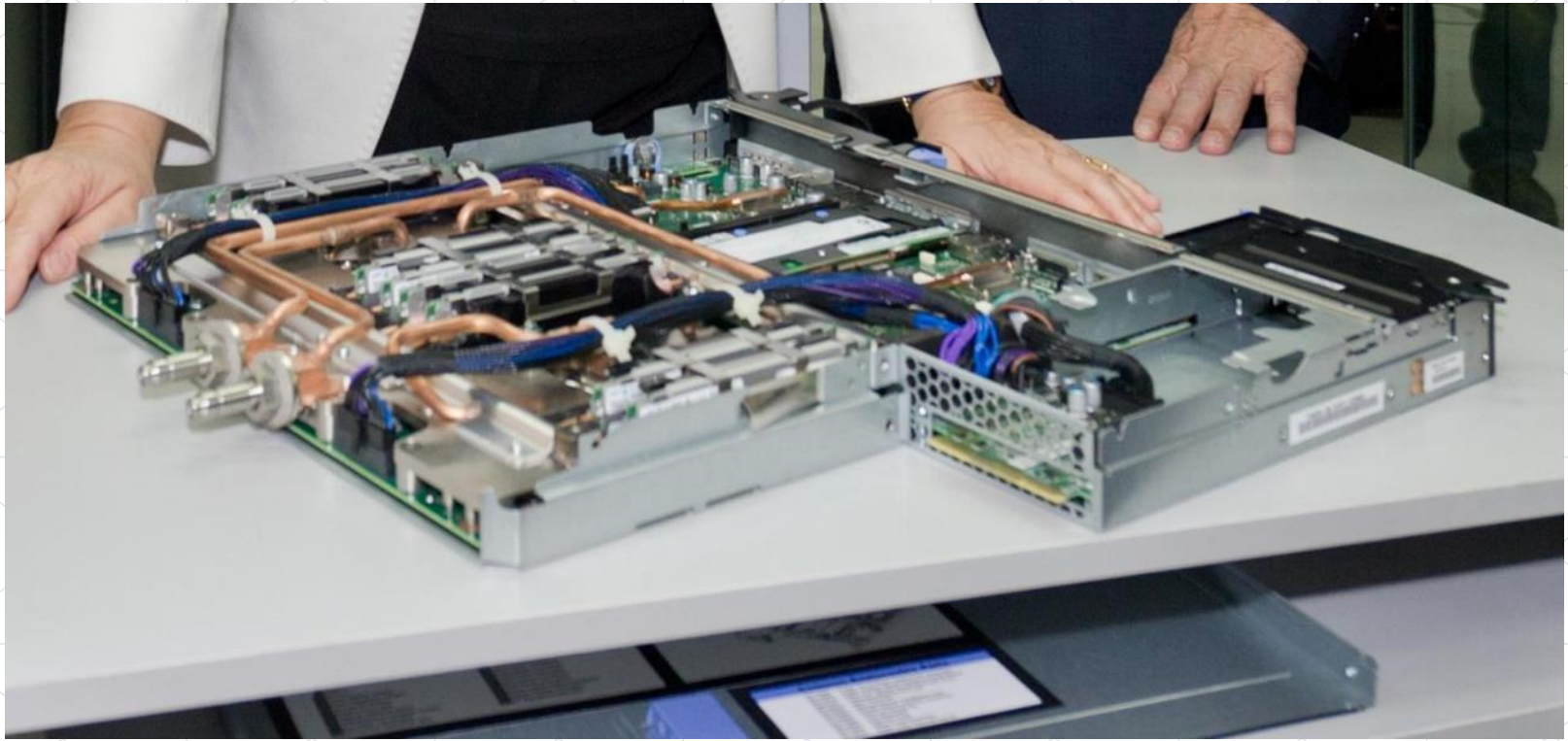


Quelle: www.lrz.de

SuperMUC 2012 (realer Kabelsalat)

Quelle: https://en.wikipedia.org/wiki/File:SuperMUC.jpg, Wikipedia-User Mdw77, CC-BY-SA 4.0

SuperMUC (Rechenknoten mit Wasserkühlung)



Quelle: www.lrz.de

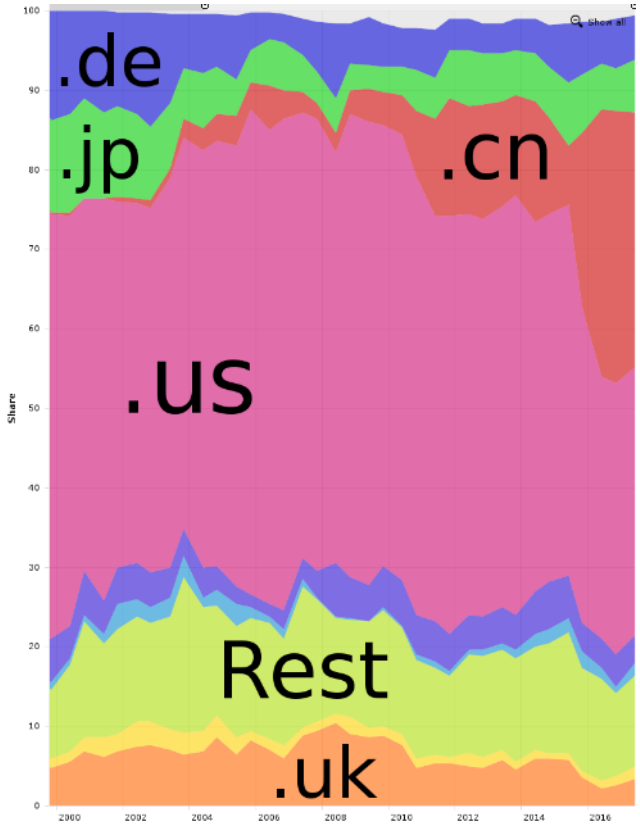
LiMa Rechenknoten

Mit dem HPC-System läuft $GTA5 doch super, oder?

- HPC-Systeme führen nicht auf magische Weise beliebige Software schneller aus
- Software muss explizit dafür programmiert werden, dass sie das HPC-System nutzen kann → Parallelisierung!
- → Selbst wenn das HPC-Cluster unter Windows laufen würde, könnte $GTA5 nicht mehr als einen Knoten eines Clusters benutzen.

Top500 http://www.top500.org

- Liste der **500 schnellsten HPC-Systeme**
  - …die öffentlich bekannt sein sollen
  - Gemessen mit “**Linpack**”-Benchmark (lösen eines großen linearen Systems)
  - Erscheint 2x pro Jahr (Nov. + Juni)
- **#1 Juni 2017: Sunway TaihuLight, China**
  - **10 649 600 cores, chinesische CPU**
- **#2: Tianhe-2, China**
  - 3 120 000 Intel Xeon Phi cores



Trends Top500-Liste Juni 2017

- 43% mit Infiniband- oder OmniPath-Vernetzung
- **>99% mit Linux** (Windows: zuletzt 07/2015 1 System)
- 33% der Systeme in den USA, 32% in China (.de: 5%);
  - Top 10: 2 .cn, 5 USA, 2 .jp, 1 .ch
  - Zum Vergleich 11/2014: 46% USA, 12% .cn.
    Top10: 1 .cn, 6 USA, 1 .jp, 1 .ch, 1 .de
- **90% mit Intel Xeon CPUs** (11/2014: 85%)

Infiniband / Omni-Path

- **De-facto Standard** für HPC-Cluster
  - Außer Ethernet praktisch keine Konkurrenz mehr
- Schnell und mit sehr geringer Latenz
  - **Switched** fabric, oft fully non-blocking (bis 648 Ports)
  - QDR: 40 Gbit roh, 32 Gbit Nutz, 1.3 µs Latenz (seit 2007)
  - FDR: 56 Gbit roh, 54 Gbit Nutz, 0.7 µs Latenz (seit 2011)
  - Omni-Path: im Prinzip Infiniband @100 GBit, nur Intel-proprietär und mit Infiniband inkompatibel :-/

Infiniband - Probleme

- Grauenhafte (Nicht-)API
  - Vom Endanwender unbemerkt
- Geringe Zuverlässigkeit
- Besorgniserregende Marktkonsolidierung
  - Nur noch zwei Hersteller für QDR, nur einer für FDR
  - Stark verlangsamte Entwicklung, steigende Preise
  - Wird durch Omni-Path nicht besser, im Gegenteil…

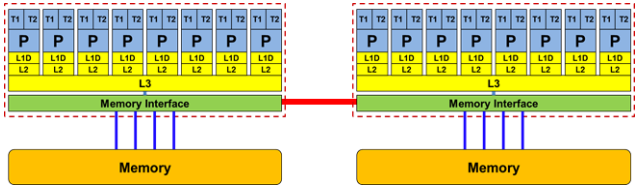
Ein Teil des Infiniband-Netzes von LiMa…

Acceleratoren

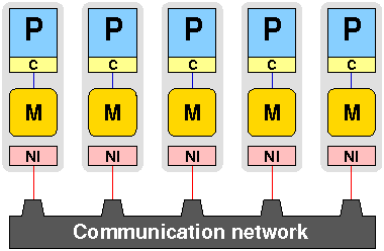
- Einsteckkarten, auf die Berechnungen ausgelagert werden können
  - GPUs
  - Intel Xeon Phi
- Noch schwieriger effizient zu programmieren als normale shared-memory-Systeme oder Cluster.
  Programmiermodelle
  - NVIDIA CUDA, OpenCL, OpenACC (GPUs)
  - OpenMP (Intel Xeon Phi)
- **Wunder darf man auch im Idealfall nicht erwarten!**

Parallelisierungsarten

- Shared Memory
  - **OpenMP**: Compiler-Erweiterung
  - Pthreads: POSIX library
  - TBB, OmpSs, Cilk+,…
  - Automatische Parallelisierung: nicht praxisrelevant



- Distributed Memory
  - **MPI** (Message Passing Interface)
    - Version 1.0: Juni 1994; Aktuelle Version: MPI 3.1 (06/2015)



- Sprachbasierte Parallelität
  - X10, Chapel, …

OpenMP

- Spracherweiterungen, die von einem OpenMP-fähigen Compiler ausgewertet und umgesetzt werden
- Trivial-Beispiel aus Wikipedia (in C):

```
int i, a[100000];

#pragma omp parallel for
for (i = 0; i < N; i++)
    a[i] = 2 * i;
```

- Jeder Thread bekommt einen Teil der Schleife zugewiesen und arbeitet ihn ab (Shared Memory!)

MPI

- Alt aber konstant weiterentwickelt
- **Library-basiert** (Funktionsaufrufe)
  - Programmiersprache ist weiterhin seriell
- Praktisch jede Software die parallel auf mehr als einem Rechner laufen kann setzt darauf auf

- Alle hochskalierbaren Produktionscodes auf Supercomputern sind MPI-basiert.

MPI

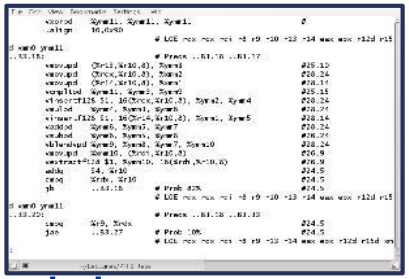
- MPI-Standard definiert ein **Set von Funktionen** für die Programmierung von Parallelrechnern
  - Senden / Empfangen von Nachrichten
  - Parallel I/O
  - Synchronisation
    - warten bis alle Prozesse eine Barriere erreicht haben

“Hello World” in MPI (C)

```
int main(int argc, char ** argv) {
 int numprocs, myid;
 MPI_Init(&argc, &argv);
 MPI_Comm_size(MPI_COMM_WORLD, &numprocs);
 MPI_Comm_rank(MPI_COMM_WORLD, &myid);
 printf(“Hi, ich bin Prozess %d von %d\n",
        myid, numprocs);
 MPI_Finalize();
 return 0;
}
```

Nutzung von HPC-Clustern in der Praxis: Batchbetrieb

- Fast wie in der guten alten Zeit: Lochkarten mit Programmen darauf abgeben, auf Verarbeitung warten, Ergebnis abholen
  - Nur heutzutage natürlich ohne Lochkarten…
- Man kann sich nicht einfach direkt auf beliebigen Knoten des Clusters einloggen
- Stattdessen: Einloggen auf **Frontend**, Jobs submittieren
  - Frontends sind interaktiv zugänglich (meist SSH)
  - Dort können Jobs vorbereitet und abgeschickt werden
  - Jobs müssen komplett ohne Benutzereingaben auskommen!



Beispiel-Batchscript

```
#!/bin/bash -l
#
#PBS -M michael.meier@fau.de
#PBS -m abe
#PBS -N LINPACK
#PBS -l walltime=01:45:00
#PBS -l nodes=16:ppn=24
#

cd $PBS_O_WORKDIR
module load intelmpi
mpirun_rrze -npernode 12 -pinexpr "S0:0-5@S1:0-5" ./xhpl
```

Spezifikation von Ressourcen und anderen Parametern

Arbeit!

Batchbetrieb

- **Scheduler** legt fest, wann der Job ausgeführt wird und berücksichtigt dabei Dinge wie…
  - **Priorität des Nutzers** und seiner Gruppe
  - wieviel Rechenzeit hat der Nutzer und seine Gruppe in der jüngeren Vergangenheit schon bekommen („Fair Share“)
  - Andere laufende Jobs
  - Reservierungen (z.B. für Nutzer oder Wartungsarbeiten)
  - Verfügbarkeit von Lizenzen die der Job braucht

ORGANISATORISCHES

- Die Vorträge im Überblick
- Andere Vortragsreihen des RRZE
- Ablageorte Vortragsfolien
- RRZE-Veranstaltungskalender / Mailingliste abonnieren
- Themenvorschläge & Anregungen

Weitere Vorträge zur „Systemausbildung“

26.04.2017 − Geschichte der Betriebssysteme
03.05.2017 − Unixoide Betriebssysteme (Unix, Linux, OS X)
10.05.2017 − Systemüberwachung / Monitoring
17.05.2017 − Storage & Filesysteme
31.05.2017 − Windows-Betriebssysteme
21.06.2017 − High Performance Computing
28.06.2017 − Benutzerverwaltung: MS Active Directory
05.07.2017 − Virtualisierung
12.07.2017 − Backup / Archiv
19.07.2017 − Kerberos
26.07.2017 − IT-Sicherheit

- Immer mittwochs (ab 14 c.t.),
- Raum 2.049 im RRZE

Andere Vortragsreihen des RRZE

Campustreffen

- immer donnerstags ab 15 Uhr c.t.
- vermittelt Informationen zu den Dienstleistungen des RRZE
- befasst sich mit neuer Hard- & Software, Update-Verfahren sowie Lizenzfragen
- ermöglicht den Erfahrungsaustausch mit Spezialisten

Netzwerkausbildung „Praxis der Datenkommunikation“

- immer mittwochs in den Wintersemestern, ab 14 Uhr c.t.
- Vorlesungsreihe, die in die Grundlagen der Netztechnik einführt
- stellt die zahlreichen aktuellen Entwicklungen auf dem Gebiet der (universitären) Kommunikationssysteme dar

Vortragsfolien

Die Vortragsfolien werden nach der Veranstaltung auf der Webseite des RRZE abgelegt:

http://www.rrze.fau.de/news/systemausbildung.shtml

RRZE-Veranstaltungskalender & Mailinglisten

- Kalender abonnieren oder bookmarken
  - Alle Infos hierzu stehen auf der Webseite des RRZE unter: http://www.rrze.fau.de/news/kalender.shtml

- Mailingliste abonnieren
  - Wöchentliche Terminhinweise werden zusätzlich an die Mailingliste RRZE-Aktuelles gesendet.
  - Auch diese Liste kann man abonnieren: https://lists.fau.de/mailman/listinfo/rrze-aktuelles

Themenvorschläge & Anregungen

Themenvorschläge und Anregungen nehmen wir gerne entgegen!

Bitte schreiben Sie uns einfach eine E-Mail an:
rrze-zentrale@fau.de (Betreff: Systemausbildung)

Vielen Dank für Ihre Aufmerksamkeit!

Regionales RechenZentrum Erlangen [RRZE]
Martensstraße 1, 91058 Erlangen
http://www.rrze.fau.de