



Universität Hamburg

DER FORSCHUNG | DER LEHRE | DER BILDUNG

Praktikum: Paralleles Programmieren für Geowissenschaftler

Prof. Thomas Ludwig, Hermann Lenhart & Michael Blesel



Dr. Hermann-J. Lenhart

hermann.lenhart@informatik.uni-hamburg.de



Einführung zum „Umfeld“ vom Paralleles Programmieren:

- Hardware Voraussetzung zur Parallelen Programmierung
- Softwareaspekte zum Parallelen Programmieren
- Modellstruktur zum Parallelen Programmieren
(mit Blick auf MPI)

HPC Top500 Liste

Top500 List - November 2017 | X +

https://www.top500.org/list/2017/11/?page=1

Stand Nov. 2017

... ♥ ☆

38	Japan Aerospace eXploration Agency Japan	SORA-MA - Fujitsu PRIMEHPC FX100, SPARC64 Xlfx 32C 1.98GHz, Tofu interconnect 2 Fujitsu	110,160	3,157.0	3,481.1	1,652
39	Government United States	Cray XC30, Intel Xeon E5-2697v2 12C 2.7GHz, Aries interconnect Cray Inc.	225,984	3,143.5	4,881.3	6,328
40	Air Force Research Laboratory United States	Thunder - SGI ICE X, Xeon E5-2699v3/E5-2697 v3, Infiniband FDR, NVIDIA Tesla K40, Intel Xeon Phi 7120P HPE	152,692	3,126.2	5,610.5	4,820
41	Academic Center for Computing and Media Studies (ACCMS), Kyoto University Japan	Camphor 2 - Cray XC40, Intel Xeon Phi 7250 68C 1.4GHz, Aries interconnect Cray Inc.	122,400	3,057.3	5,483.5	748
42	DKRZ - Deutsches Klimarechenzentrum Germany	Mistral - bullx DLC 720, Xeon E5-2680v3 12C 2.5GHz/E5-2695V4 18C 2.1GHz, Infiniband FDR Bull, Atos Group	99,072	3,010.7	3,962.9	1,116
43	Information Technology Center, Nagoya University Japan	Fujitsu PRIMEHPC FX100, SPARC64 Xlfx 32C 2.2GHz, Tofu interconnect 2 Fujitsu	92,160	2,910.0	3,244.0	1,382
44	Leibniz Rechenzentrum Germany	SuperMUC - iDataPlex DX360M4, Xeon E5-2680 8C 2.70GHz, Infiniband FDR IBM/Lenovo	147,456	2,897.0	3,185.1	3,423
45	Leibniz Rechenzentrum Germany	SuperMUC Phase 2 - NeXtScale nx360M5, Xeon E5-2697v3 14C 2.6GHz, Infiniband FDR14 Lenovo/IBM	86,016	2,813.6	3,578.3	1,481

Rank Site

System

Cores

(TFlop/s)

(TFlop/s)

(kW)

Quelle: www.top500.org

HPC Top500 Liste

Top500 List - November 2017 | X +

https://www.top500.org/list/2017/11/?page=1

Quelle: www.top500.org

38 Japan Aerospace eXploration Agency
Japan
SORA-MA - Fujitsu PRIMEHPC FX100, SPARC64 Xlfx 32C
1.98GHz, Tofu interconnect 2

39 Government
United States

40 Air Force Research Laboratory
United States

Stand Nov. 2017

41 Academic Center for Computing and
Media Studies (ACCMS), Kyoto
University
Japan

42 DKRZ - Deutsches
Klimarechenzentrum
Germany

43 Information Technology Center, Nagoya
University
Japan

44 Leibniz Rechenzentrum
Germany

45 Leibniz Rechenzentrum
Germany

https://www.top500.org/list/2018/11/?page=1

Stand Nov. 2018

60 Academic Center for Computing and
Media Studies (ACCMS), Kyoto
University
Japan
Camphor 2 - Cray XC40, Intel
Xeon Phi 7250 68C 1.4GHz,
Aries interconnect
Cray Inc.

61 NVIDIA Corporation
United States
Circe - NVIDIA DGX-2H POD,
Xeon Platinum 8174 24C
3.1GHz, Mellanox InfiniBand
EDR, NVIDIA Tesla V100 SXM2
Nvidia

62 DKRZ - Deutsches
Klimarechenzentrum
Germany
Mistral - bullx DLC 720, Xeon
E5-2680v3 12C
2.5GHz/E5-2695V4 18C 2.1Ghz,
Infiniband FDR
Bull, Atos Group

63 Information Technology Center, Nagoya
University
Japan
Fujitsu PRIMEHPC FX100,
SPARC64 Xlfx 32C 2.2GHz, Tofu
interconnect 2
Fujitsu

64 Leibniz Rechenzentrum
Germany
SuperMUC - iDataPlex
DX360M4, Xeon E5-2680 8C
2.70GHz, Infiniband FDR
IBM/Lenovo

65 Leibniz Rechenzentrum
Germany
SuperMUC Phase 2 - NeXtScale
nx360M5, Xeon E5-2697v3 14C
2.6GHz, Infiniband FDR14
Lenovo/IBM

Rank	Site	System	Cores	(TFlop/s)	(TFlop/s)	(kW)
------	------	--------	-------	-----------	-----------	------

HPC Top500 Liste - Stand November 2018

  <https://www.top500.org/lists/2018/11/>   

Rank	System	Cores	Rmax (TFlop/s)	Rpeak (TFlop/s)	Power (kW)
1	Summit - IBM Power System AC922, IBM POWER9 22C 3.07GHz, NVIDIA Volta GV100, Dual-rail Mellanox EDR Infiniband , IBM DOE/SC/Oak Ridge National Laboratory United States	2,397,824	143,500.0	200,794.9	9,783
2	Sierra - IBM Power System S922LC, IBM POWER9 22C 3.1GHz, NVIDIA Volta GV100, Dual-rail Mellanox EDR Infiniband , IBM / NVIDIA / Mellanox DOE/NNSA/LLNL United States	1,572,480	94,640.0	125,712.0	7,438
3	Sunway TaihuLight - Sunway MPP, Sunway SW26010 260C 1.45GHz, Sunway , NRCPC National Supercomputing Center in Wuxi China	10,649,600	93,014.6	125,435.9	15,371
4	Tianhe-2A - TH-IVB-FEP Cluster, Intel Xeon E5-2692v2 12C 2.2GHz, TH Express-2, Matrix-2000 , NUDT National Super Computer Center in Guangzhou China	4,981,760	61,444.5	100,678.7	18,482
5	Piz Daint - Cray XC50, Xeon E5-2690v3 12C 2.6GHz, Aries interconnect , NVIDIA Tesla P100 , Cray Inc. Swiss National Supercomputing Centre (CSCS) Switzerland	387,872	21,230.0	27,154.3	2,384
6	Trinity - Cray XC40, Xeon E5-2698v3 16C 2.3GHz, Intel Xeon Phi 7250 68C 1.4GHz, Aries interconnect , Cray Inc. DOE/NNSA/LANL/SNL United States	979,072	20,158.7	41,461.2	7,578

Germany:

8 Super-MUC-NG

26 Jülich

30 Jülich

Quelle: www.top500.org

HPC Top500 Liste - Stand November 2018



HPC Top500 Liste - Stand November 2018



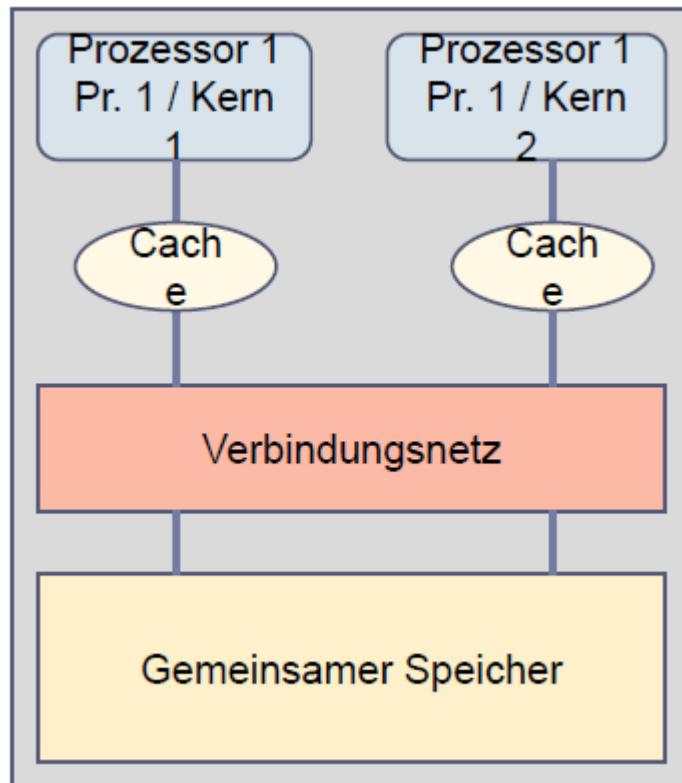
Laptop heute
ohne GPU



Möglichkeiten der Parallelen Programmierung :

- **OpenMP** - Möglich bei der Nutzung von gemeinsamem Speicher (shared memory directives)
- **MPI** (Message-Passing Interface)
 - bei Rechnerarchitektur mit verteiltem Speicher
 - derzeit einziger Standard mit Portabilität auf allen Plattformen
- **Hybride** Programmierung: Kombination von MPI und OpenMP

OpenMP - Gemeinsamer Speicherzugriff mittels SMP



SMP: Symmetrisches Multiprozessersystem
(symmetric multiprocessing)

[Multiprozessor-Architektur](#),

bei der zwei oder mehr [identische Prozessoren](#) einen gemeinsamen [Adressraum](#) besitzen.

Eine SMP-Architektur erlaubt es, die laufenden [Prozesse](#) dynamisch auf alle verfügbaren [Prozessoren](#) zu verteilen.

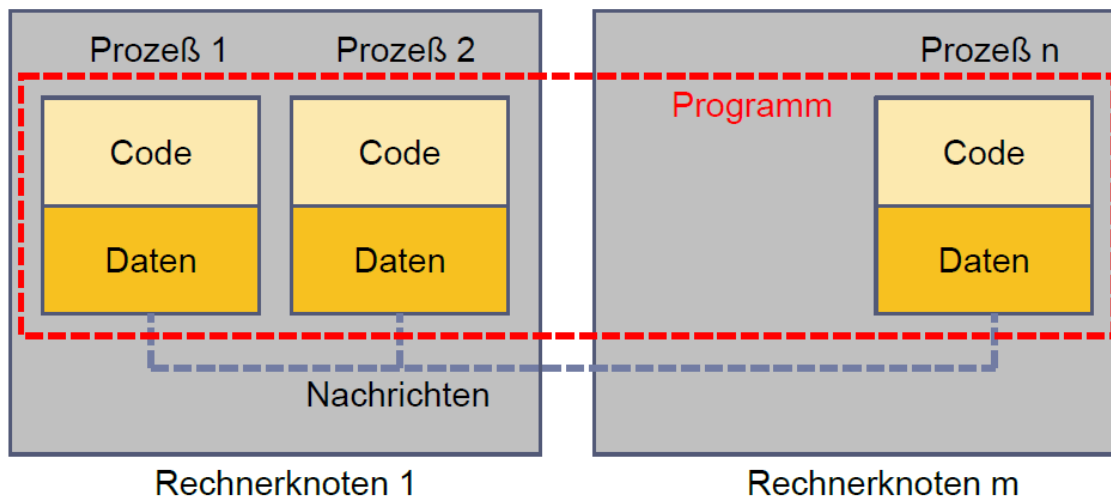
Dagegen muss beim [asymmetrischen Multiprocessing](#) jeder CPU eine Aufgabe fest zugewiesen werden (z. B. führt CPU0 Betriebssystemaufrufe und CPU1 Benutzerprozesse aus).

Source: Grafik Ludwig WS12/13, Text Wikipedia



MPI – Hardware Voraussetzung

Quelle: Ludwig WS12/13



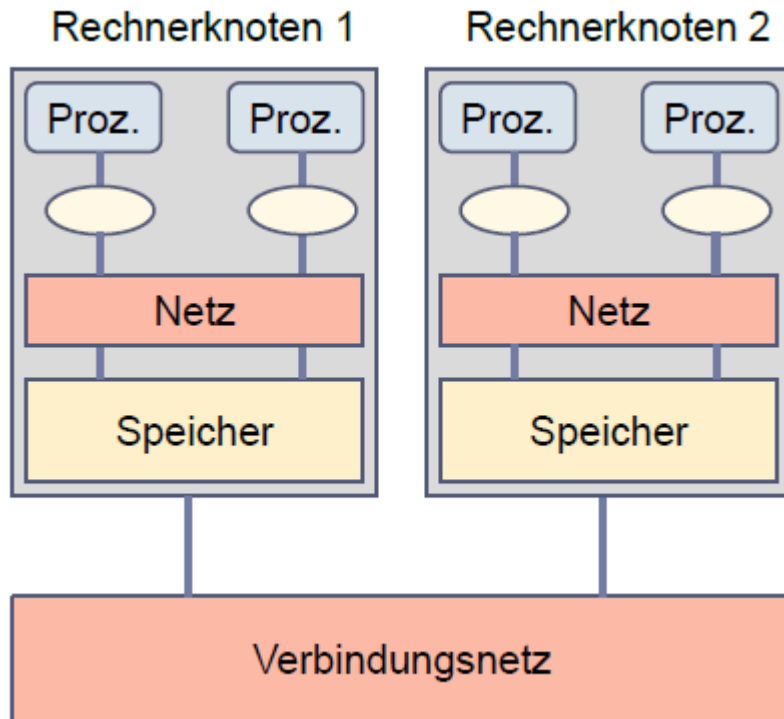
- Keinen direkten Zugriff auf Memory (Daten) von anderen Prozessen.
- Datenverfügbarkeit über expliziten Datenaustausch (Senden/Empfangen)

mit anderen Prozessen!

!! Vorteil: MPI Prozesse lassen sich skalieren !!



Hybride Programmierung



Existierende HPC Rechner sind heute meist eine Kombination aus **Rechnerknoten mit gemeinsamem Speicher**, von den man viele verwendet und **über ein Verbindungsnetz** verbindet.

Hybride Programmierung ermöglicht die Kombination von MPI und OpenMP, aber bedarf mehr Struktur der Zugriffsrechte.

Vortrag von Dr. Panagiotis Adamidis als DKRZ Beitrag am 4. Juli 2019

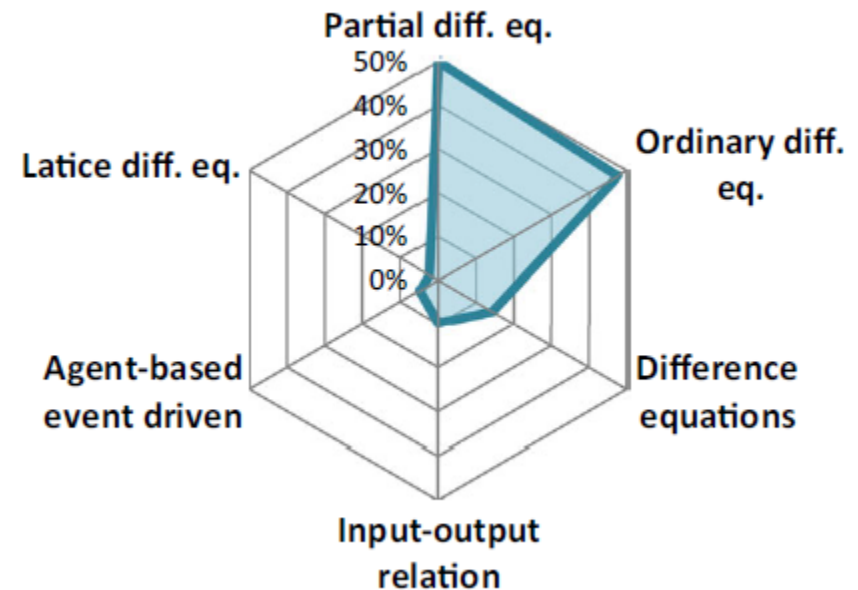
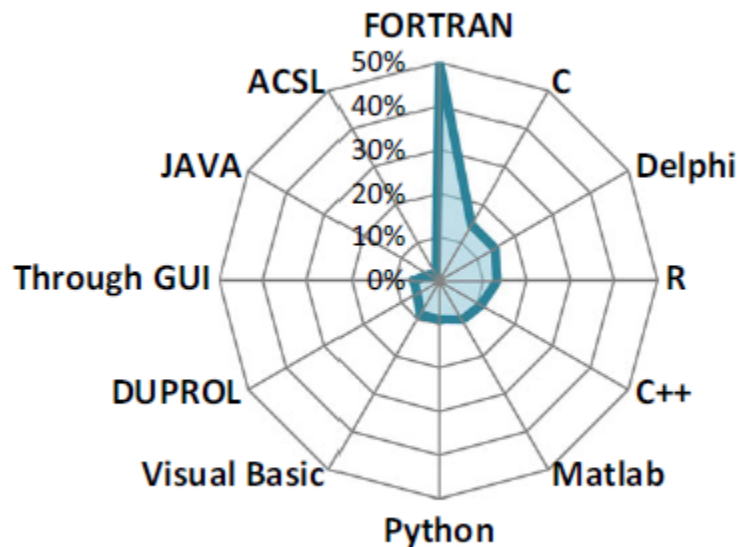
Quelle: Ludwig WS12/13



Software-Anwendungen in Erdsystemmodellen

Paper: Exploring, exploiting and evolving diversity of
aquatic ecosystem models: a community perspective.

Janssen et al., 2015 in Aquatic Ecology





Praktische-Aspekte für Parallele Programmierung

Problem beim Parallelen Programmieren:

Reihenfolge der Ausführung nicht steuerbar, bzw. nicht deterministisch.

Ein Beispiel für einen unkontrollierten Zustand des Speichers sind sogenannte

Wettlaufsituation (Race Condition)

Wettlaufsituationen bezeichnen in der Programmierung eine Konstellation, in der das Ergebnis einer Operation vom zeitlichen Verhalten bestimmter Einzeloperationen abhängt.

D.h. die Reihenfolge in der die Berechnungen und Speicherung ablaufen

ist nicht vertauschbar!



Software-Aspekte für Parallele Programmierung I

Neben Hardware gibt es auch einige Anforderungen an Softwarekomponenten, z.B. für die Einbindung von Programm-Bibliotheken.

Thread-Sicherheit (Thread safety)

Softwarekomponente können gleichzeitig von verschiedenen Programmbereichen mehrfach ausgeführt werden, ohne dass diese sich gegenseitig behindern.

Zum Zweck der Mehrfachausführung bieten Betriebssysteme das Konzept von sogenannte **Threads**, die als „leichtgewichtige Prozesse“ gelten.



Universität Hamburg

DER FORSCHUNG | DER LEHRE | DER BILDUNG



Software-Aspekte für Parallele Programmierung II

Jeder Thread arbeitet dabei unabhängig von den anderen einen Programmteil ab.

Häufig muss das Programm dabei gleichzeitig auf einen gemeinsamen Speicherbereich des Computers zugreifen.

Änderungen im Speicher durch verschiedene Threads müssen koordiniert werden, um einen chaotischen Zustand des Speichers zu verhindern.



Programmstruktur in der Modellierung

Der Ablauf eines Modelles wird üblicherweise in 3 Phasen eingeteilt:

- 1) Initialisierung
- 2) Berechnung des Modells
- 3) Finalisierung

Diese Einteilung wird z.B. von Kopplern wie ESMF gefordert.



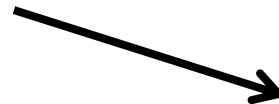
MPI Nachrichtenaustausch in der Modellierung I

1) Initialisierung: Aufteilung der Rechengebiete und Initialisierung

Initialisierungsdaten



Modellmatrix



Teilmatrizen pro Prozessor

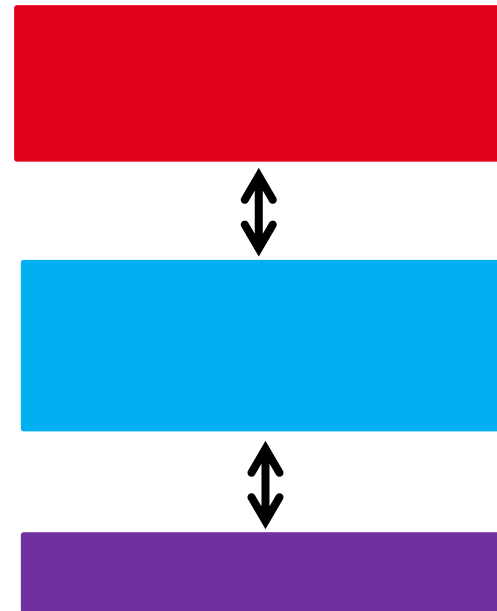


MPI Nachrichtenaustausch in der Modellierung II

2) Berechnung des Modells auf Teilmatrizen: Austausch zwischen Teilmatrizen



Modellmatrix

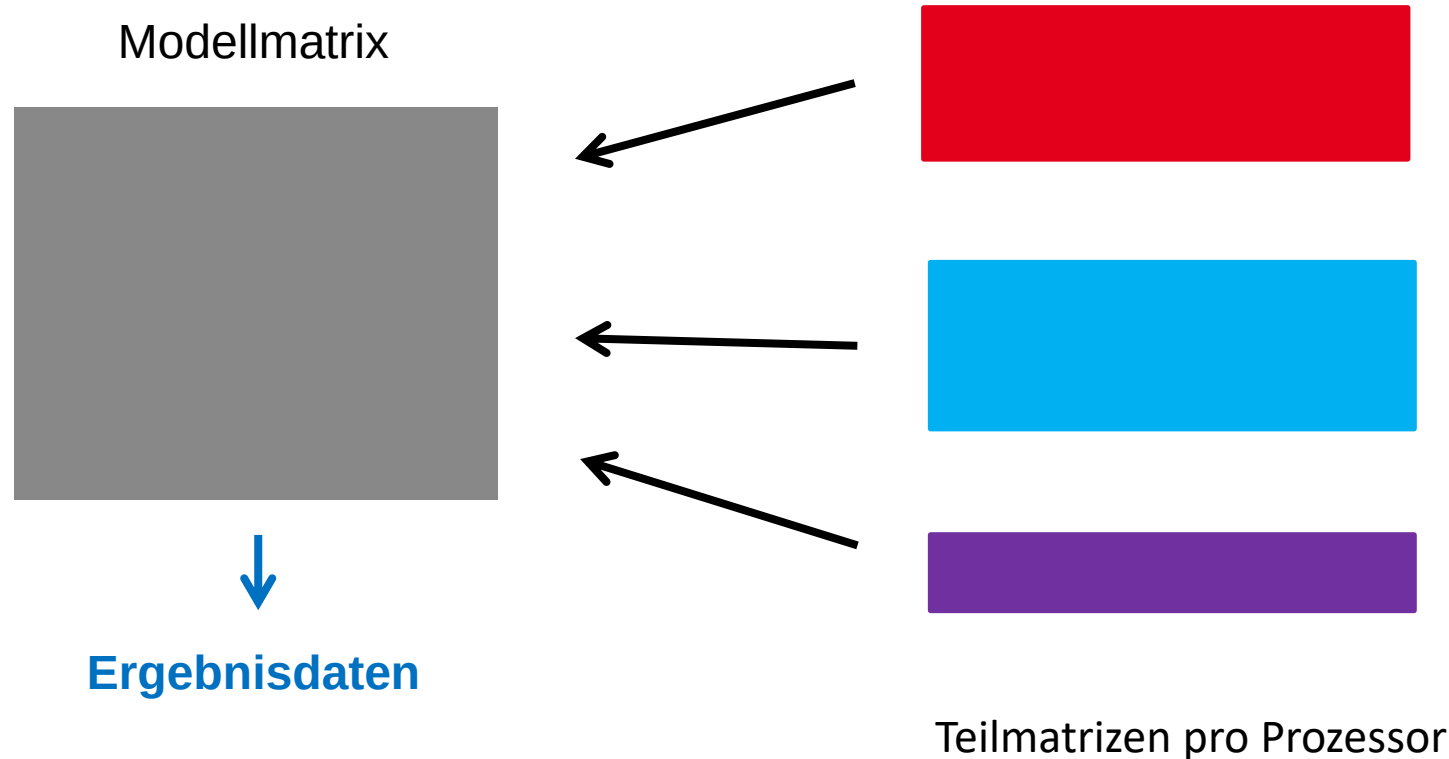


Teilmatrizen pro Prozessor

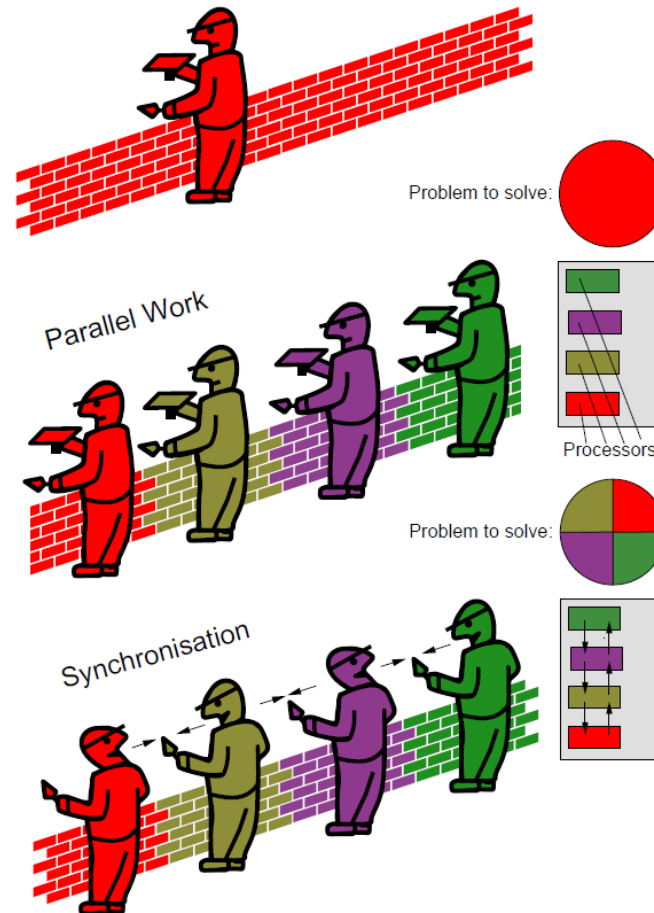


MPI Nachrichtenaustausch in der Modellierung III

3) **Finalisierung**: Zusammenfügen der Rechenergebnisse der Teilmatrizen



MPI Nachrichtenaustausch: **Zu Beachten!!**



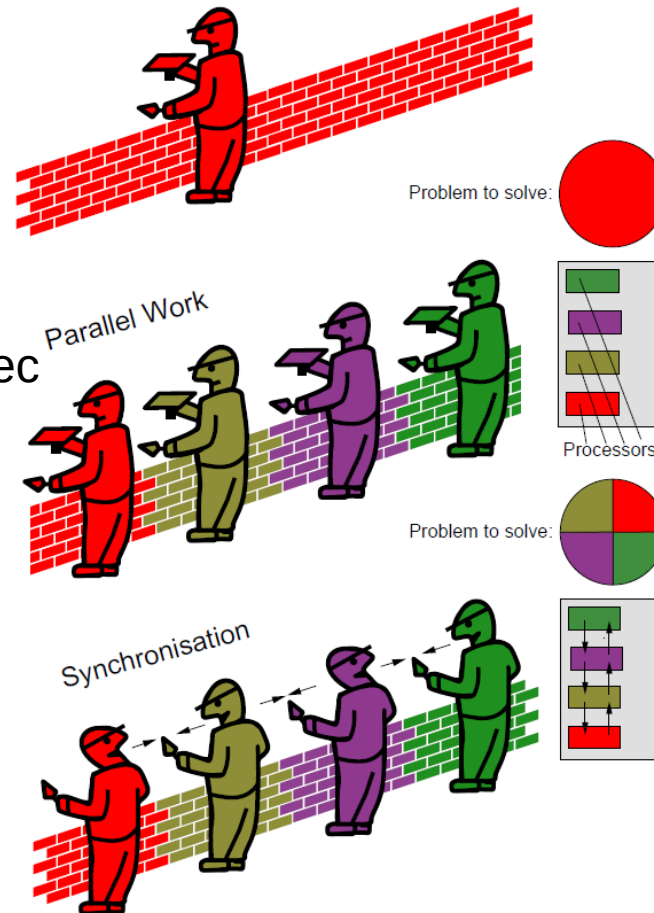
MPI Nachrichtenaustausch: **Zu Beachten!!**

Als Info für das Verhältnis
Rechnen / Nachrichtenaustausch
 soll folgende Abschätzung dienen:

Ein moderner Parallelrechner schafft
 ca. 3 Mrd. floating point operationen / Sec

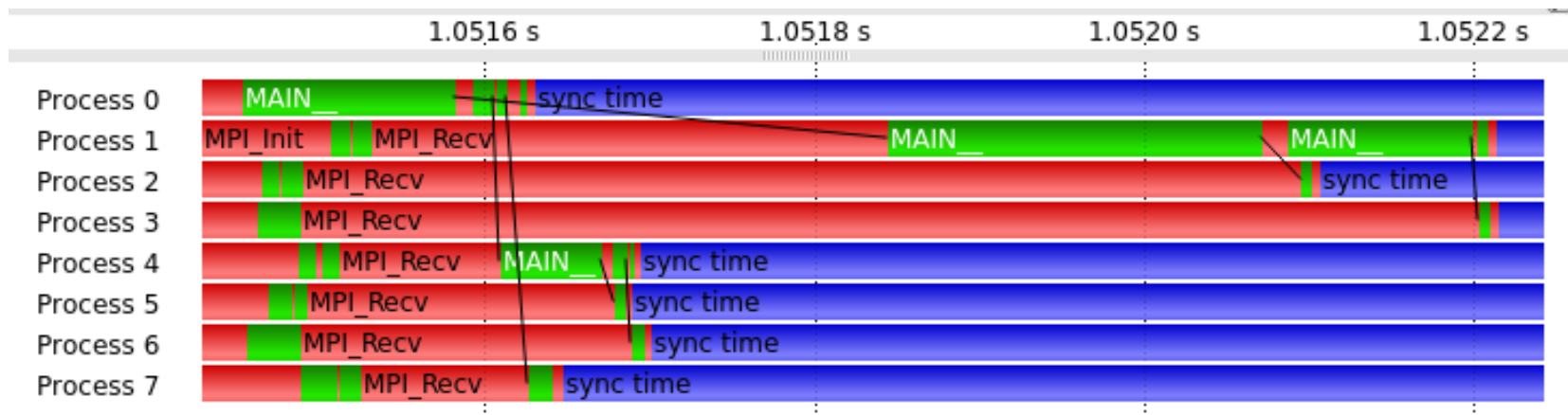
Der Nachrichtenaustausch
 aber nur 10 Mio. Wörter / Sec

Faktor 300 !!





Visualisierung des Programmablaufes mit Vampire:





Universität Hamburg

DER FORSCHUNG | DER LEHRE | DER BILDUNG



**Danke,
gibt es noch Fragen?**