Hochleistungsrechner

Exascale Computing & Resilienz

VON PIERRE BOUCHARD

Inhalt

1. **Wieso müssen Supercomputer immer leistungsfähiger werden?**
2. Was heißt Exascale Computing?
3. Existierende und geplante Hochleistungsrechner
4. Funktionsweise
5. Wie soll die Exascale-Marke geknackt werden?
6. Resilienz
7. Quellen

Wieso müssen Supercomputer immer leistungsfähiger werden?

Naturkatastrophen müssen besser vorhersagbar sein um:

- Menschenleben zu retten
- Schaden zu begrenzen
- Wiederaufbaukosten zu minimieren [1]

Nachrichtendienste hacken zum Beispiel:

- Banken
- Forschungseinrichtungen
- andere Regierungen [2]

Auch für folgende Forschungsgebiete werden größere Rechenleistungen benötigt:

- Moleküldynamik
- Astrophysik [3]

[1] : HTTPS://WWW.GOLEM.DE/NEWS/SUPERCOMPUTER-WIE-DIE-EXAFLOP-MARKE-GEKNACKT-WERDEN-SOLL-1610-122823.HTML
[2] : HTTP://WWW.SPIEGEL.DE/NETZWELT/NETZPOLITIK/NSA-WILL-SUPER-COMPUTER-ZUM-AUSSPAEHEN-ENTWICKELN-A-941604.HTML
[3] : HTTPS://SCIENCE.ENERGY.GOV/~/MEDIA/ASCR/ASCAC/PDF/REPORTS/EXASCALE_SUBCOMMITTEE_REPORT.PDF

Inhalt

1. Wieso müssen Supercomputer immer leistungsfähiger werden?
2. **Was heißt Exascale Computing?**
3. Existierende und geplante Hochleistungsrechner
4. Funktionsweise
5. Wie soll die Exascale-Marke geknackt werden?
6. Resilienz
7. Quellen

Was heißt Exascale Computing?

Exa steht für 10^{18}

Maß für die Leistungsfähigkeit eines Systems sind Flops

- **Flo**ating **P**oint Operations Per **S**econd [4]
- Floating Point = Gleitkommazahlen
  - Zahlen mit Ziffern vor und nach dem Komma [5]

Bisher existiert kein Exascale System [4]

Einheit	Fließkommaoperationen pro Sekunde
Flops	1
MegaFlops	1.000.000
GigaFlops	1.000.000.000
TeraFlops	1.000.000.000.000
PetaFlops	1.000.000.000.000.000
ExaFlops	1.000.000.000.000.000.000

[4] : HTTPS://DE.WIKIPEDIA.ORG/WIKI/FLOATING_POINT_OPERATIONS_PER_SECOND
[5] : HTTPS://WWW.ELEKTRONIK-KOMPENDIUM.DE/SITES/DIG/1807231.HTM

Inhalt

1. Wieso müssen Supercomputer immer leistungsfähiger werden?
2. Was heißt Exascale Computing
3. **Existierende und geplante Hochleistungssysteme**
   - Sunway-taihu light
   - Aurora
   - Tianhe-3
   - Planung der EU
4. Funktionsweise
5. Wie soll die Exascale-Marke geknackt werden?
6. Resilienz
7. Quellen

Sunway-taihu light

Chinesisches System ohne amerikanische Prozessoren [6]

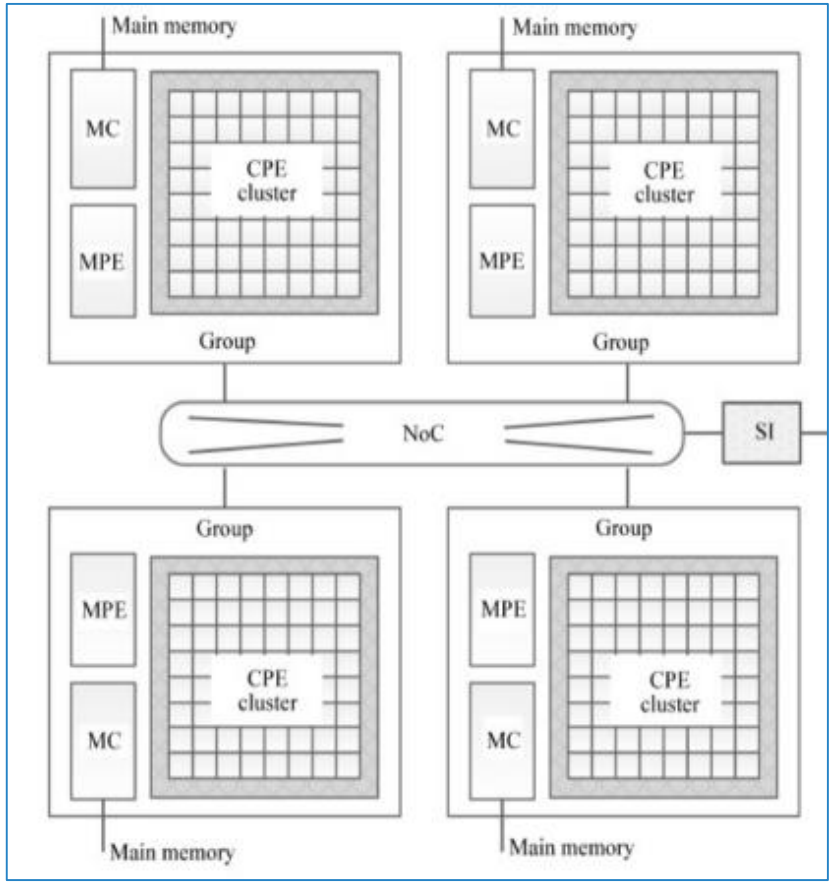
Entwicklungsstandort	National Supercomputing Center in Wuxi
Hersteller	NRCPC
Rechnerarchitektur	Hybrid Distributed-Shared Memory Computing [8]
Tatsächliche Rechenleistung	93,015 PetaFlops
Theoretische Spitzenleistung	125,439 PetaFlops
Energieverbrauch	15,371 MegaWatt
Flops pro Watt	6,02 GigaFlops pro Watt
Speichergröße	1,31 Petabytes

[7]

[6] : HTTP://WWW.SPIEGEL.DE/NETZWELT/GADGETS/CHINAS-SUNWAY-TAIHULIGHT-IST-SCHNELLSTER-SUPERCOMPUTER-A-1098599.HTML
[7] : HTTPS://WWW.TOP500.ORG/SYSTEM/178764
[8] : HTTPS://WWW.NEXTPLATFORM.COM/2016/06/20/LOOK-INSIDE-CHINAS-CHART-TOPPING-NEW-SUPERCOMPUTER/

Sunway-taihu light - Rechnerarchitektur

Ein Knoten:



[8] : Rechenarchitektur eines Knotens

Sunway-taihu light

Chinesisches System ohne amerikanische Prozessoren [6]

Entwicklungsstandort	National Supercomputing Center in Wuxi
Hersteller	NRCPC
Rechnerarchitektur	Hybrid Distributed-Shared Memory Computing [8]
Tatsächliche Rechenleistung	93,015 PetaFlops
Theoretische Spitzenleistung	125,439 PetaFlops
Energieverbrauch	15,371 MegaWatt
Flops pro Watt	6,02 GigaFlops pro Watt
Speichergröße	1,31 Petabytes

[7]

[6] : HTTP://WWW.SPIEGEL.DE/NETZWELT/GADGETS/CHINAS-SUNWAY-TAIHULIGHT-IST-SCHNELLSTER-SUPERCOMPUTER-A-1098599.HTML
[7] : HTTPS://WWW.TOP500.ORG/SYSTEM/178764
[8] : HTTPS://WWW.NEXTPLATFORM.COM/2016/06/20/LOOK-INSIDE-CHINAS-CHART-TOPPING-NEW-SUPERCOMPUTER/

Aurora

Amerikanisches System in der Entwicklung

Entwicklungsstandort	Argonne National Laboratory in den USA
Hersteller	Intel
Rechnerarchitektur	Hybrid Distributed-Shared Memory Computing
Tatsächliche Rechenleistung	n.a.
Theoretische Spitzenleistung	180 - 450 PetaFlops
Energieverbrauch	13 MegaWatt
Flops pro Watt	13 GigaFlops pro Watt
Speichergröße	7 Petabytes

[9]

[9] : HTTP://WWW.INTEL.COM/NEWSROOM/ASSETS/INTEL_AURORA_FACTSHEET.PDF

Tianhe-3

Chinesisches System in Planung, daher sind noch keine genauen Fakten bekannt

Prototyp wird 2018 auf dem Markt sein

2020 soll mit dem Rechner die Exascale-Marke geknackt werden

Allgemein verfügbar für Forschungseinrichtungen

Beinhaltet nur chinesische Bauteile [10]

[10] : HTTP://WWW.CHINADAILY.COM.CN/CHINA/2017-02/20/CONTENT_28259294.HTM

Planung der EU

Bau eines Exascale Computers bis 2022/2023

Planung soll bis Ende 2017 abgeschlossen sein

5 Milliarden Euro Investitionen [11]

[B1] : Europäische Flagge

[11] : HTTPS://M.HEISE.DE/NEWSTICKER/MELDUNG/EU-WILL-BEI-EXASCALE-COMPUTING-AUFHOLEN-3710281.HTML?WT_REF=ANDROID-APP%3A%2F%2FCOM.GOOGLE.ANDROID.GOOGLEQUICKSEARCHBOX%2FHTTPS%2FWWW.GOOGLE.COM&WT_T=1494495705396

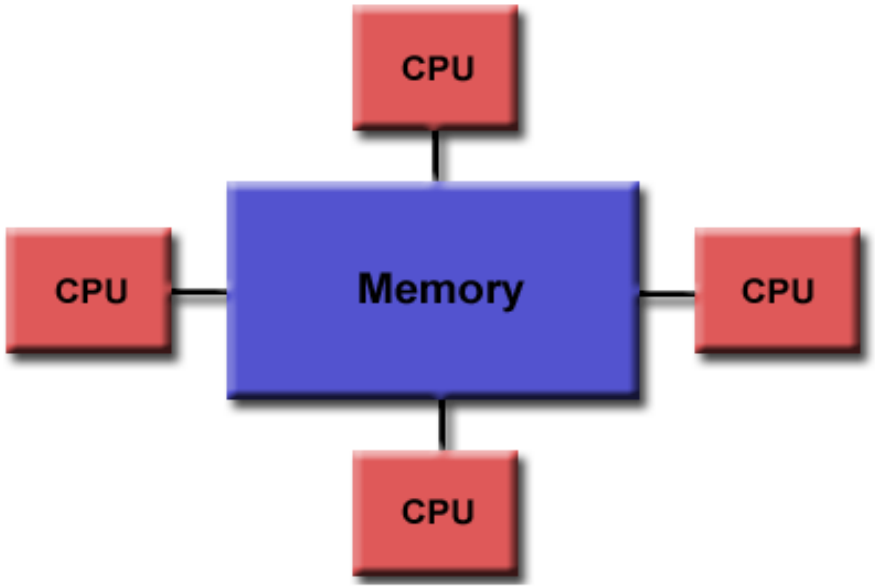
Inhalt

1. Wieso müssen Supercomputer immer leistungsfähiger werden?
2. Was heißt Exascale Computing?
3. Existierende und geplante Hochleistungssysteme
4. **Funktionsweise**
   - Shared Memory Computing
   - Distributed Memory Computing
   - Hybrid Distributed-Shared Memory Computing
5. Wie soll die Exascale-Marke geknackt werden?
6. Resilienz
7. Quellen

Funktionsweise

1. **Shared Memory Computing**
   - Einleitung
   - OpenMP
   - Vorteile
   - Probleme
2. Distributed Memory Computing
3. Hybrid Distributed-Shared Memory Computing

Shared Memory Computing Einleitung



[B2] : Schematischer Aufbau

Ein Speicher für alle Prozessoren

Shared memory access ermöglicht eine leichte Programmierung

Threading

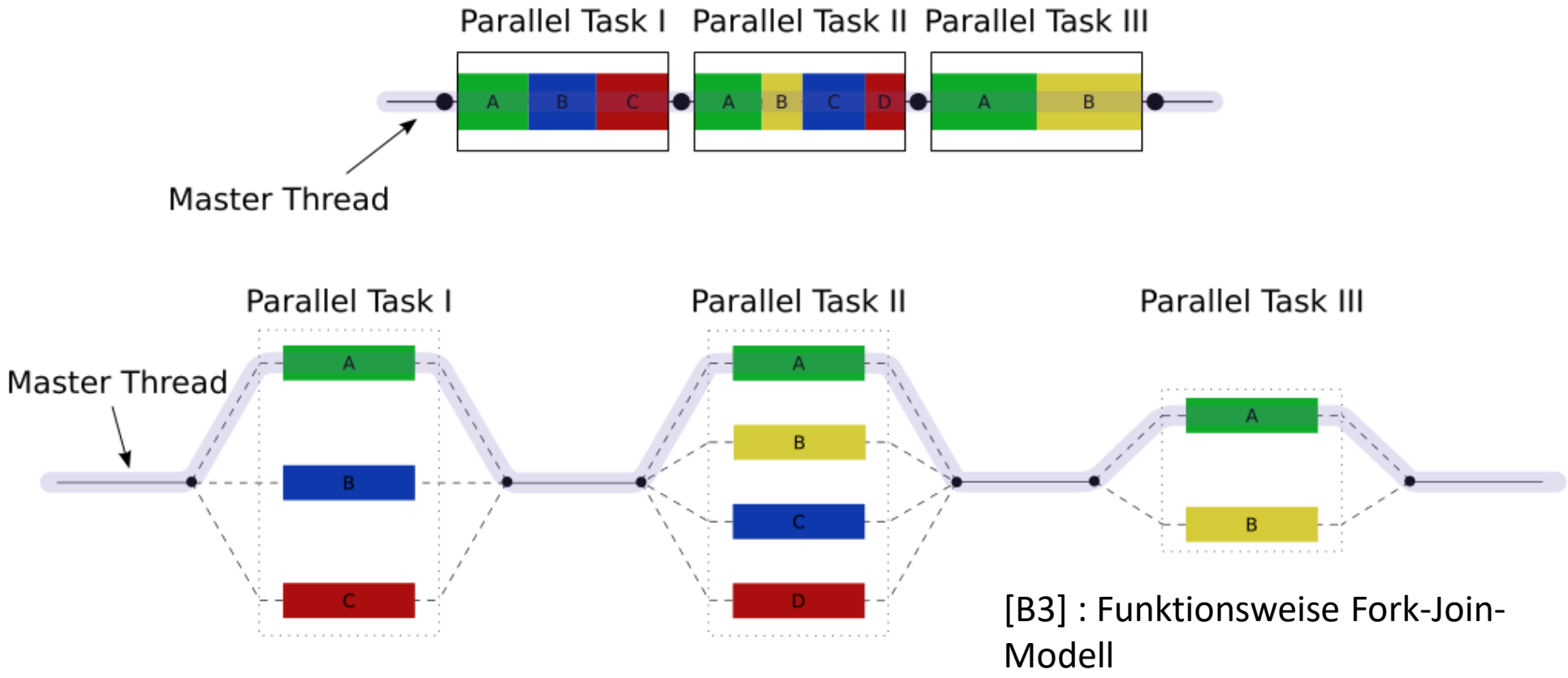
- Aufteilen eines Prozesses
- Thread hat seine eigene Adresse und seinen eigenen Stack

Programmierschnittstelle

- OpenMP [12]

[12] : HTTPS://SOFTWARE.INTEL.COM/EN-US/ARTICLES/HYBRID-PARALLELISM-PARALLEL-DISTRIBUTED-MEMORY-AND-SHARED-MEMORY-COMPUTING

OpenMP - Fork-Join-Modell



[B3] : Funktionsweise Fork-Join-Modell

Vorteile

OpenMP

- Ressourcenschonend im Bezug auf die Threads
- Automatisches Zerlegen der Aufgaben in Threads

Intels Threading Building Blocks

- Nur C++
- Viele vorprogrammierte Features
- Leicht verständliche Benutzeroberfläche [12]

[B4] : Logo OpenMP

[B4] : Logo TBB

[12] : HTTPS://SOFTWARE.INTEL.COM/EN-US/ARTICLES/HYBRID-PARALLELISM-PARALLEL-DISTRIBUTED-MEMORY-AND-SHARED-MEMORY-COMPUTING

Probleme

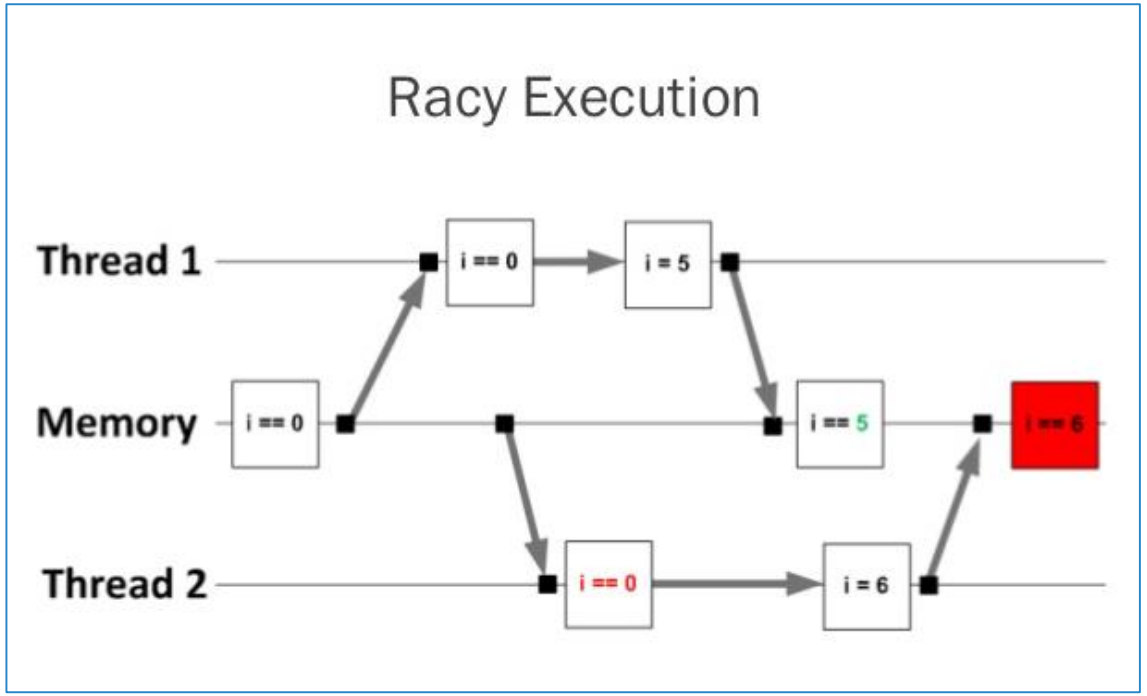
Datarace

Mittels Barrieren kann

das Problem eingedämmt werden,

dann entsteht jedoch

das nächstes Problem: Deadlocks [12]



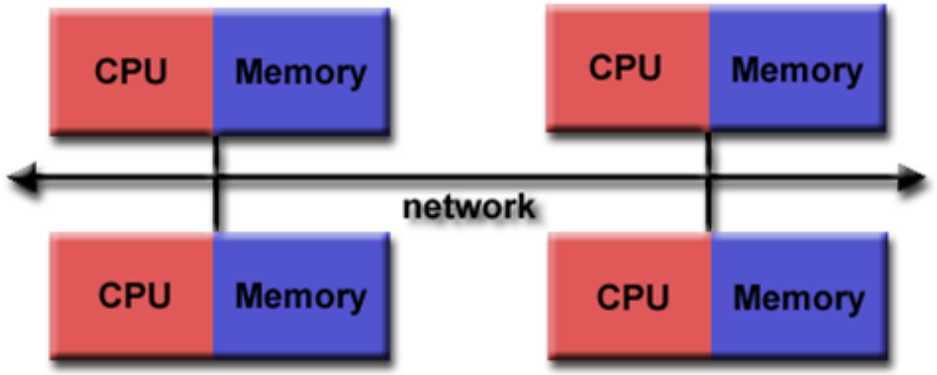
[B5] : Veranschaulichung von Datarace

[12] : HTTPS://SOFTWARE.INTEL.COM/EN-US/ARTICLES/HYBRID-PARALLELISM-PARALLEL-DISTRIBUTED-MEMORY-AND-SHARED-MEMORY-COMPUTING

Funktionsweise

1. Shared Memory Computing
2. **Distributed Memory Computing**
   - Einleitung
   - MPI
   - Probleme
3. Hybrid Distributed-Shared Memory Computing

Distributed Memory Computing Einleitung



[B2] : Schematischer Aufbau

Gliederung in Einheiten

- Pro Prozessor ein Speicher

Kommunikation zwischen Einheiten

- MPI
- Aktives Versenden und Nutzen von Daten [12]

Schneller Zugriff auf den lokalen Speicher [13]

[12] : HTTPS://SOFTWARE.INTEL.COM/EN-US/ARTICLES/HYBRID-PARALLELISM-PARALLEL-DISTRIBUTED-MEMORY-AND-SHARED-MEMORY-COMPUTING

[13] : HTTPS://COMPUTING.LLNL.GOV/TUTORIALS/PARALLEL_COMP/#WHATIS

MPI – Message Passing Interface

Dekomposition durch Programmierer

One-side data movement

- Datentransfer ohne Bestätigung

Remote direct memory access

- Direkter Zugriff auf einen anderen Speicher
- Durch mehr Vorarbeit wird der Rechenprozess beschleunigt [12]

[B4] : Logo MPI

[12] : HTTPS://SOFTWARE.INTEL.COM/EN-US/ARTICLES/HYBRID-PARALLELISM-PARALLEL-DISTRIBUTED-MEMORY-AND-SHARED-MEMORY-COMPUTING

Probleme

Mehr Vorarbeit ist nötig [12]

Größerer Speicheraufwand

Eigene Darstellung

Komplizierte Speicherorganisation

Keine einheitliche Zugriffsdauer auf den Speicher [13]

[12] : HTTPS://SOFTWARE.INTEL.COM/EN-US/ARTICLES/HYBRID-PARALLELISM-PARALLEL-DISTRIBUTED-MEMORY-AND-SHARED-MEMORY-COMPUTING

[13] : HTTPS://COMPUTING.LLNL.GOV/TUTORIALS/PARALLEL_COMP/#WHATIS

Funktionsweise

1. Shared Memory Computing
2. Distributed Memory Computing
3. **Hybrid Distributed-Shared Memory Computing**

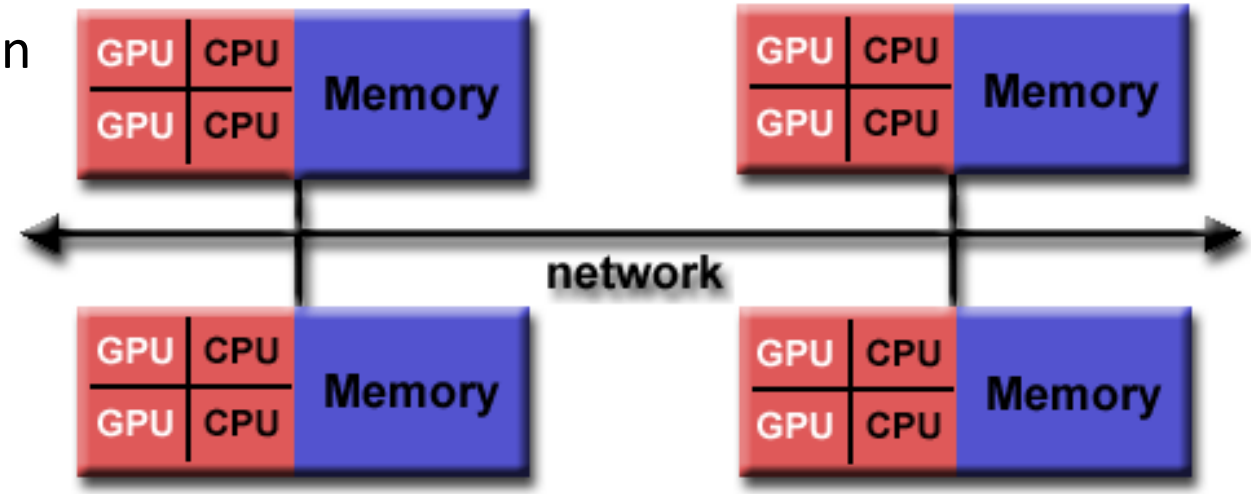
Hybrid Distributed-Shared Memory Computing

Dynamischer [12] Hybridcode aus OpenMP (4.0, oder neuer [14]) und MPI

Aufteilung in Cluster

GPUs und CPUs können nur auf ihren eigenen Speicher zugreifen

- Ein Kommunikationssystem ist notwendig [13]



[B2] : Schematischer Aufbau

[12] : HTTPS://SOFTWARE.INTEL.COM/EN-US/ARTICLES/HYBRID-PARALLELISM-PARALLEL-DISTRIBUTED-MEMORY-AND-SHARED-MEMORY-COMPUTING [13] : HTTPS://COMPUTING.LLNL.GOV/TUTORIALS/PARALLEL_COMP/#WHATIS
[14] : HTTPS://EN.WIKIPEDIA.ORG/WIKI/OPENACC

Inhalt

1. Wieso müssen Supercomputer immer leistungsfähiger werden?
2. Was heißt Exascale Computing?
3. Existierende und geplante Hochleistungsrechner
4. Funktionsweise
5. **Wie soll die Exascale-Marke geknackt werden?**
6. Resilienz
7. Quellen

Wie soll die Exascale-Marke geknackt werden?

1. **Lösungsansatz**
2. Resultierende Probleme
3. Bestehende Problemlösungsansätze

Lösungsansatz

Prozessoren beschleunigen

- Mehr Kerne
- Größere Caches

Verwendung von GPU‘s

Breitere Vektoreinheiten [1]

[1] : HTTPS://WWW.GOLEM.DE/NEWS/SUPERCOMPUTER-WIE-DIE-EXAFLOP-MARKE-GEKNACKT-WERDEN-SOLL-1610-122823.HTML

Wie soll die Exascale-Marke geknackt werden?

1. Lösungsansatz
2. **Resultierende Probleme**
3. Bestehende Problemlösungsansätze

Resultierende Probleme

Kein Platz

Hoher Energieverbrauch

- Stärkere Wärmeentwicklung [15]
- Energieverschwendung [16]

Heterogene Systeme

- Benötigen komplexere Programme
- Keine Ausnutzung der gesamten Rechenleistung [12]

Ungeeignete Systemsoftware [1]

[12] : HTTPS://SOFTWARE.INTEL.COM/EN-US/ARTICLES/HYBRID-PARALLELISM-PARALLEL-DISTRIBUTED-MEMORY-AND-SHARED-MEMORY-COMPUTING

REFERENZEN AUF PROBLEMLÖSUNGSSEITEN

Wie soll die Exascale-Marke geknackt werden?

1. Lösungsansatz
2. Resultierende Probleme
3. **Bestehende Problemlösungsansätze**

Bestehende Problemlösungsansätze

1. Kein Platz
2. Hoher Energieverbrauch
   - Stärkere Wärmeentwicklung
   - Energieverschwendung
3. Ungeeignete Systemsoftware

Dreidimensionale Positionierung der Transistoren

Größter Stromverbrauch durch
- Elektrischen Widerstand der Leiterbahnen
  - 10 mal größer als beim Schalten der Transistoren
  - 99% des Energieverlustes

Verkürzen der Leiterbahnen
- In der 2D Ebene geht das jedoch nicht mehr viel kürzer

3D Bauweise
- Reduzierung des Energieverbrauchs um das 100-fache und das Computervolumen um das 1000-fache

Geordnete fraktale Struktur
- Gehirn als Vorlage
- Zusätzlich Stromverbrauch / 30, Volumen / 1000 [15]

[15] : HTTP://WWW.SPEKTRUM.DE/NEWS/KAMPF-GEGEN-DIE-HITZE/1180336

Resultierende Probleme

1. Kein Platz
2. Hoher Energieverbrauch
   - Stärkere Wärmeentwicklung
   - Energieverschwendung
3. Ungeeignete Systemsoftware

Lösung für das Wärmeproblem

Flüssigkeitskühlung direkt auf dem Chip

- Nichtwässriges, nichtreaktives und flüssiges Kühlmittel
- Fluorkarbone im superMUC (IBM)

Wasser als Elektrolyt-Treibstoff

- Energierückgewinnung durch redox-flow-Zellen
- Betrieb von Mikroprozessoren [15]

[15] : HTTP://WWW.SPEKTRUM.DE/NEWS/KAMPF-GEGEN-DIE-HITZE/1180336

Resultierende Probleme

1. Kein Platz
2. Hoher Energieverbrauch
   - Stärkere Wärmeentwicklung
   - Energieverschwendung
3. Ungeeignete Systemsoftware

Energieverschwendung

Maximal 20 Megawatt dürfte das Exascale System verbrauchen

- 50 GigaFlops/Watt

Stand heute

- 6 GigaFlops/Watt im sunway-taihu light [1]

Steigender Verbrauch von Kühlsystemen

- Großteil der Energie wird als Wärme abgegeben
- Beheizen der Umgebung - Block Heizkraftwerk [16]

[1] : HTTPS://WWW.GOLEM.DE/NEWS/SUPERCOMPUTER-WIE-DIE-EXAFLOP-MARKE-GEKNACKT-WERDEN-SOLL-1610-122823.HTML
[16] : HTTP://WWW.CMCC.IT/WP-CONTENT/UPLOADS/2012/05/RP0121-SCO-12-2011.PDF

Woher weiß man, was energiesparender werden muss?

Exa2Green

- EU-Projekt
- Gerät/Tool zur Messung des Energieverbrauchs von einzelnen Komponenten
- Lokalisierung der energiefressenden Komponenten

Gezielte Verbesserung der betroffenen Komponenten

- Verbesserung des Anwendungscodes [17]

[17] : HTTPS://WWW.STEINBEIS-EUROPA.DE/FILES/TRANSFER3-2015_EXA2_GREEN.PDF

Resultierende Probleme

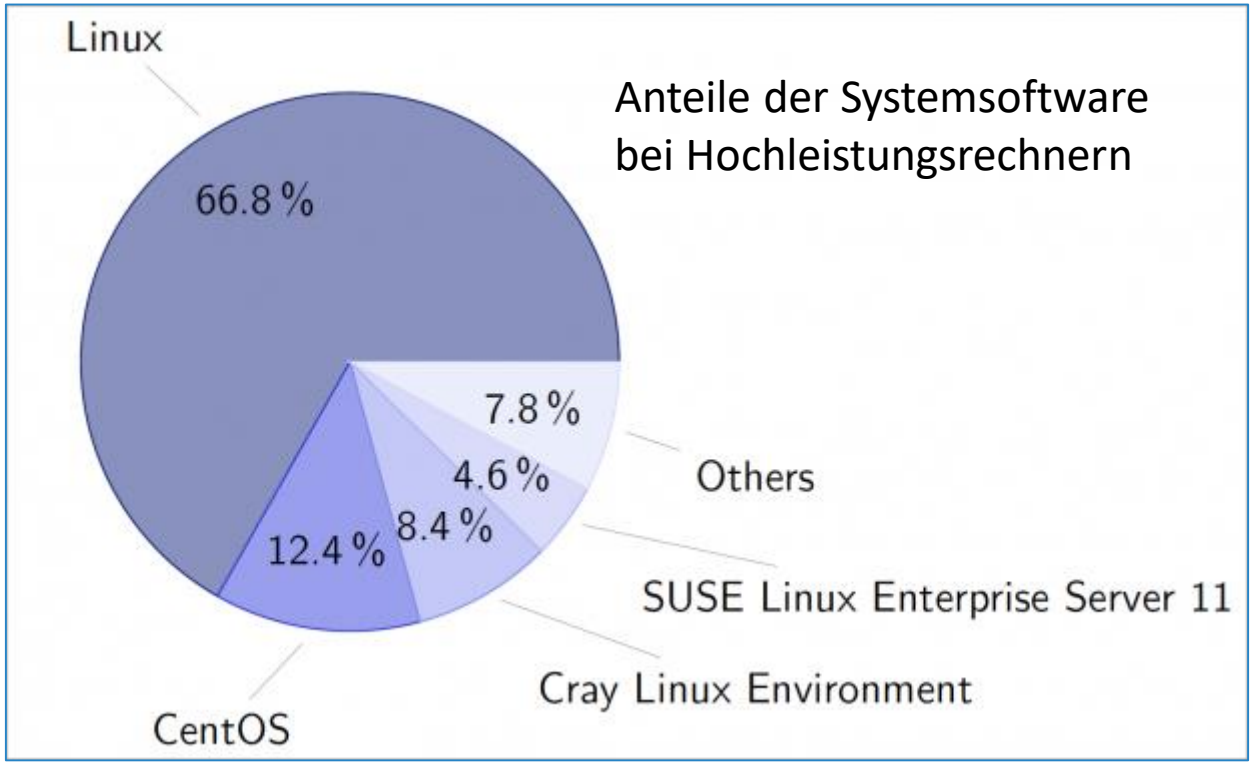
1. Kein Platz
2. Hoher Energieverbrauch
   - Stärkere Wärmeentwicklung
   - Energieverschwendung
3. Ungeeignete Systemsoftware

Systemsoftware

Linux:
- Sehr komplex
- Anfällig für Fehler

Lightweight Kernel
- Vermeidung von Seiteneffekten
- Hoher Programmieraufwand

Lösung
- Zeptos-OS (abgespecktes Linux)
- Multi-Kernel [1]



[B6] : Diagramm zur prozentualen Nutzung von Systemsoftware

[1] : HTTPS://WWW.GOLEM.DE/NEWS/SUPERCOMPUTER-WIE-DIE-EXAFLOP-MARKE-GEKNACKT-WERDEN-SOLL-1610-122823.HTML

Inhalt

1. Wieso müssen Supercomputer immer leistungsfähiger werden?
2. Was heißt Exascale Computing?
3. Existierende und geplante Hochleistungsrechner
4. Funktionsweise
5. Wie soll die Exascale-Marke geknackt werden?
6. **Resilienz**
7. Quellen

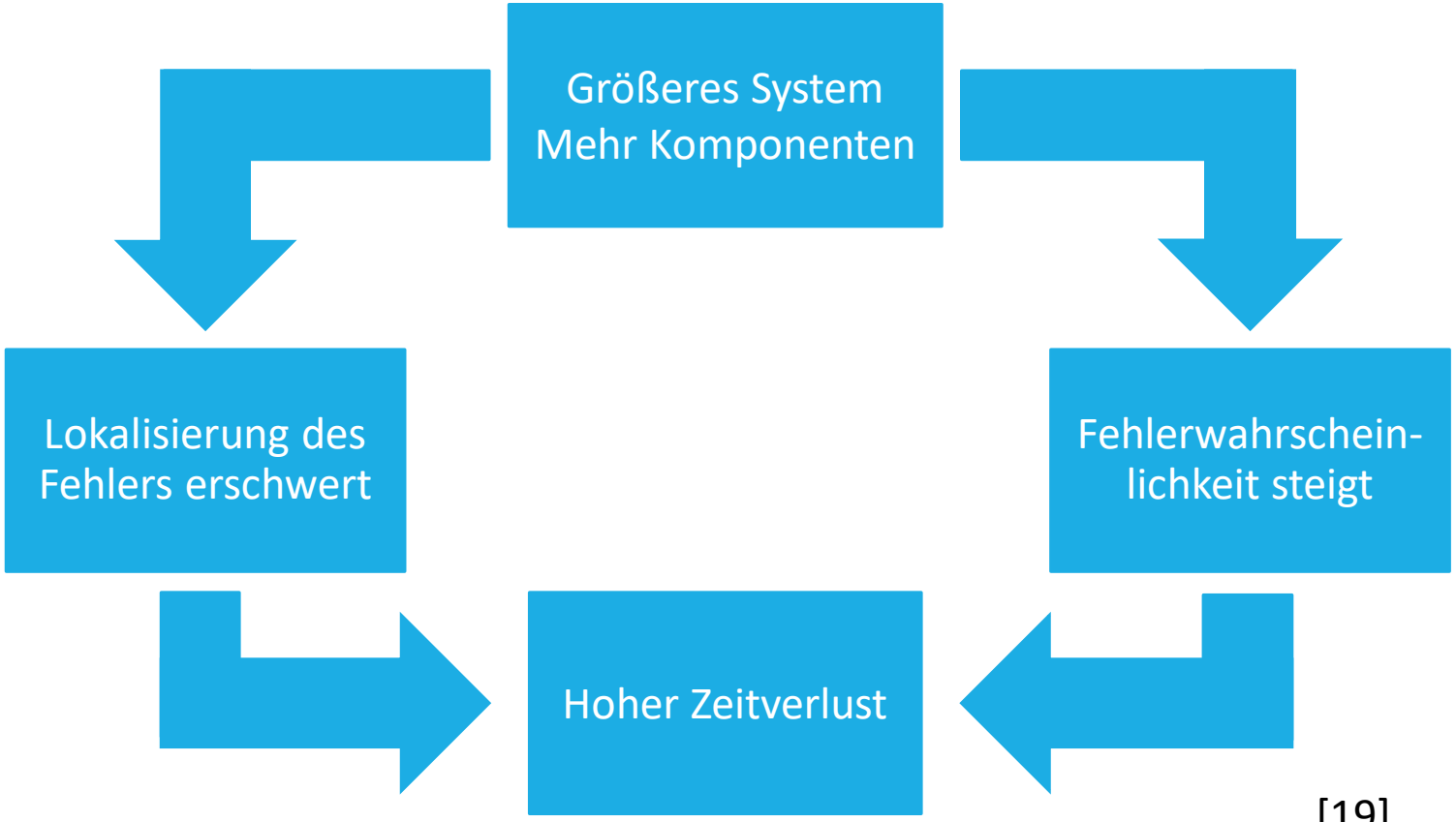
Resilienz

Widerstandsfähigkeit und die Fähigkeit,

mit Problemsituationen umgehen zu können [18]

[18] : HTTP://RESILIENZ-FREIBURG.DE/INDEX.PHP/WAS-IST-RESILIENZ/DEFINITION-UND-MERKMALE

Resilienz



[19]

[19] : HTTP://RESILIENZ-FREIBURG.DE/INDEX.PHP/WAS-IST-RESILIENZ/DEFINITION-UND-MERKMALE

FeToL - eine Fehler-Tolerante Umgebung für peta-scale MPI Löser

Erhöhung der Ausfallsicherheit

Funktionsweise:

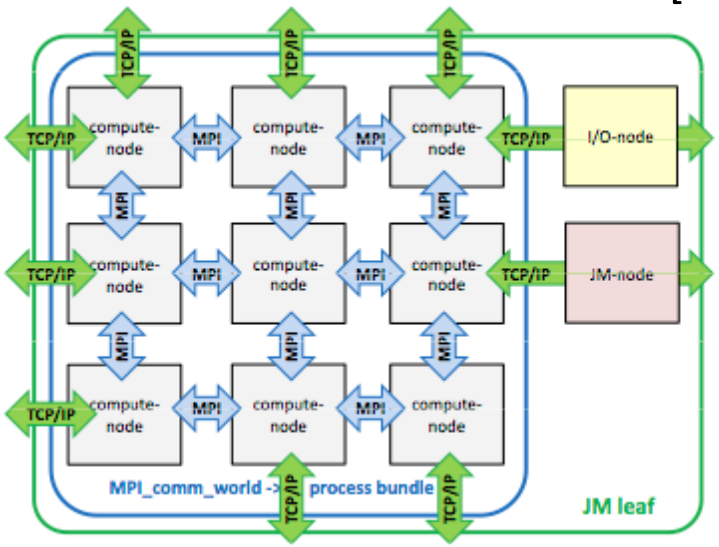
- Divide-and-conquer-Prinzip
- Gruppierung von Prozessen in Prozess-Bündel (PB)
- Prozesse innerhalb eines PB kommunizieren untereinander über ein MPI
- Prozesse in unterschiedlichen PB kommunizieren über ein Multi-Agent-System, namens BOND [19]

[19] : HTTP://RESILIENZ-FREIBURG.DE/INDEX.PHP/WAS-IST-RESILIENZ/DEFINITION-UND-MERKMALE

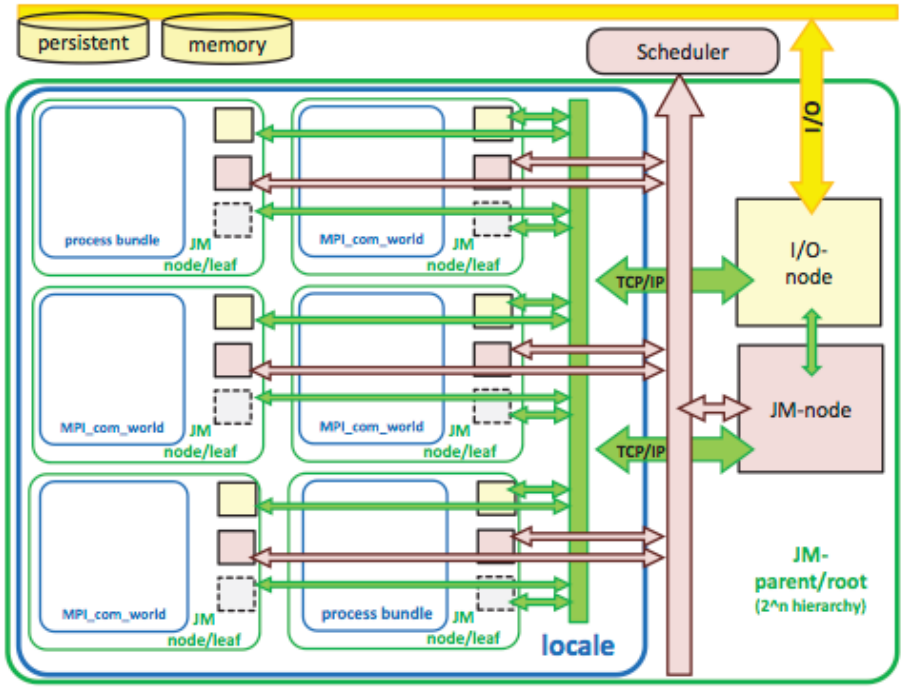
FeToL - eine Fehler-Tolerante Umgebung für peta-scale MPI Löser

Ausfall eines oder mehrerer Prozesse

- Neustart eines Prozessbündels
- Basierend auf
  - Checkpoint Daten
  - Daten der Prozessnachbarn [19]



[B7] : Funktions-schema FeToL



[B7] : Funktionsschema FeToL

[19] : HTTP://RESILIENZ-FREIBURG.DE/INDEX.PHP/WAS-IST-RESILIENZ/DEFINITION-UND-MERKMALE

Inhalt

1. Wieso müssen Supercomputer immer leistungsfähiger werden?
2. Was heißt Exascale Computing?
3. Existierende und geplante Hochleistungsrechner
4. Funktionsweise
5. Wie soll die Exascale-Marke geknackt werden?
6. Resilienz
7. **Quellen**

Quellen

[1] : https://www.golem.de/news/supercomputer-wie-die-exaflop-marke-geknackt-werden-soll-1610-122823.html

[2] : http://www.spiegel.de/netzwelt/netzpolitik/nsa-will-super-computer-zum-ausspaehen-entwickeln-a-941604.html

[3] : https://science.energy.gov/~/media/ascr/ascac/pdf/reports/Exascale_subcommittee_report.pdf

[4] : https://de.wikipedia.org/wiki/Floating_Point_Operations_Per_Second

[5] : https://www.elektronik-kompendium.de/sites/dig/1807231.htm

[6] : http://www.spiegel.de/netzwelt/gadgets/chinas-sunway-taihulight-ist-schnellster-supercomputer-a-1098599.html

[7] : https://www.top500.org/system/178764

Quellen

[8] : https://www.nextplatform.com/2016/06/20/look-inside-chinas-chart-topping-new-supercomputer/

[9] : http://www.intel.com/newsroom/assets/Intel_Aurora_factsheet.pdf

[10] : http://www.chinadaily.com.cn/china/2017-02/20/content_28259294.htm

[11] : https://m.heise.de/newsticker/meldung/EU-will-bei-Exascale-Computing-aufholen-3710281.html?wt_ref=android-app%3A%2F%2Fcom.google.android.googlequicksearchbox%2Fhttps%2Fwww.google.com&wt_t=1494495705396

[12] : https://software.intel.com/en-us/articles/hybrid-parallelism-parallel-distributed-memory-and-shared-memory-computing

[13] : https://computing.llnl.gov/tutorials/parallel_comp/#Whatis

[14] : https://en.wikipedia.org/wiki/OpenACC

[15] : http://www.spektrum.de/news/kampf-gegen-die-hitze/1180336

Quellen

[16] : http://www.cmcc.it/wp-content/uploads/2012/05/rp0121-sco-12-2011.pdf

[17] : https://www.steinbeis-europa.de/files/transfer3-2015_exa2_green.pdf

[18] : http://resilienz-freiburg.de/index.php/was-ist-resilienz/definition-und-merkmale

[19] : http://resilienz-freiburg.de/index.php/was-ist-resilienz/definition-und-merkmale

Bildquellen

[1] : http://storyworkshopmw.org/eu.php

[2] : https://computing.llnl.gov/tutorials/parallel_comp/#Whatis

[3] : https://en.wikipedia.org/wiki/Fork%E2%80%93join_model#/media/File:Fork_join.svg

[4] : https://www.google.de/url?sa=i&rct=j&q=&esrc=s&source=images&cd=&cad=rja&uact=8&ved=0ahUKEwiJ9vWIlu3TAhXLOxQKHfPtDI8QjRwIBw&url=https%3A%2F%2Fwww.florian-rappl.de%2FNews%2FPage%2F331%2Fconstruction-of-a-supercomputer&psig=AFQjCNFbanLVlfy0C3hWXHIMz2O8-hb8aQ&ust=1494775298738972

[5] : https://www.google.de/imgres?imgurl=https%3A%2F%2Fimage.slidesharecdn.com%2Fdrdsecrfinal3-121106080213-phpapp01%2F95%2Fdynamic-data-race-detection-in-concurrent-java-programs-5-638.jpg%3Fcb%3D1352191909&imgrefurl=https%3A%2F%2Fwww.slideshare.net%2FDevexperts%2Fdynamic-data-race-detection-in-concurrent-java-programs&docid=SSwmwgSXulLQKM&tbnid=HhCf5uZ9AjNkuM%3A&vet=10ahUKEwjvm5jfjsTTAhVGAxoKHYOvDGsQMwgpKAQwBA..i&w=638&h=479&bih=702&biw=1536&q=data%20race&ved=0ahUKEwjvm5jfjsTTAhVGAxoKHYOvDGsQMwgpKAQwBA&iact=mrc&uact=8#spf=1

Bildquellen

[6] : https://www.golem.de/news/supercomputer-wie-die-exaflop-marke-geknackt-werden-soll-1610-122823-3.html

[7] : https://www.tu-braunschweig.de/irmb/forschung/abgeschlosseneprojekte/fetol