GA Gauß-Allianz

FaST – Dynamische Topologien in höchstskalierenden Umgebungen

FAST FOR EXASCALE

Abschlussbericht im Rahmen der HPC-Statustagung in Hamburg
28. November 2016

Präsentation der Gauß-Allianz

Prof. Dr.-Ing. André Brinkmann

Projektpartner

JOHANNES GUTENBERG UNIVERSITÄT MAINZ

JG|U

RWTH AACHEN UNIVERSITY

Fraunhofer SCAI

MEGWARE®

PARTEC CLUSTER COMPETENCE CENTER

TUM Technische Universität München

GA Gauß-Allianz

Herausforderung Exascale Computing

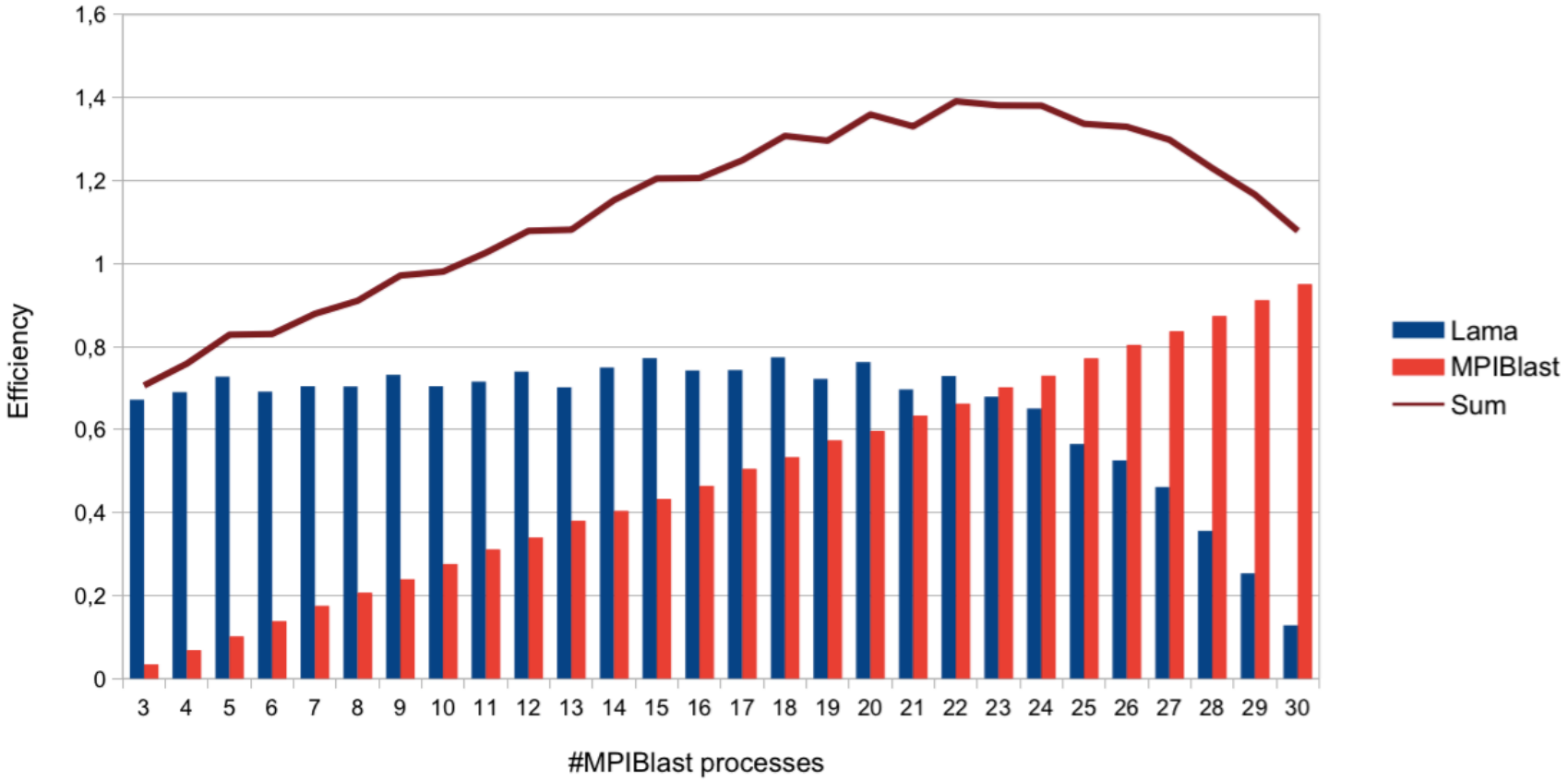
- Die meisten Anwendungen sind nicht in der Lage, alle Ressourcen eines Knotens effizient zu nutzen

- “System designs which balance the resource usage are absolutely essential”
  - Scheduler müssen die Prozesse gemäß ihres Ressourcenverbrauchs verteilen
  - Korrekturen zur Laufzeit: Migrationen der Prozesse zwischen den Knoten

Zitat aus: „HPC Trends, Multicore Computing and the Path to Exascale“
von James Sexton, IBM Research; 2009

GA Gauß-Allianz

Co-Scheduling



Gauß-Allianz

Herausforderung Exascale Computing

- Charakterisierung der Anwendungen
  - Statische Parameter, dynamisches Laufzeitverhalten
  - Daten einholen für Scheduling-Entscheidungen
- Migration von Prozessen
  - Anpassung der MPI-Laufzeitumgebung
  - Minimieren notwendiger Datentransfers
- Scheduling der Prozesse
  - Dynamisches Scheduling nach Ressourcenverbrauch und Laufzeitverhalten
  - Anpassung des Batchsystems
  - Durchsatz erhöhen und / oder Energie einsparen

GA Gauß-Allianz

Herausforderung Exascale Computing

- **Charakterisierung der Anwendungen**
  - **Statische Parameter, dynamisches Laufzeitverhalten**
  - **Daten einholen für Scheduling-Entscheidungen**

- Migration von Prozessen
  - Anpassung der MPI-Laufzeitumgebung
  - Minimieren notwendiger Datentransfers

- Scheduling der Prozesse
  - Dynamisches Scheduling nach Ressourcenverbrauch und Laufzeitverhalten
  - Anpassung des Batchsystems
  - Durchsatz erhöhen und / oder Energie einsparen

GA Gauß-Allianz

Auslastung der Ressource "Speicherhierarchie"

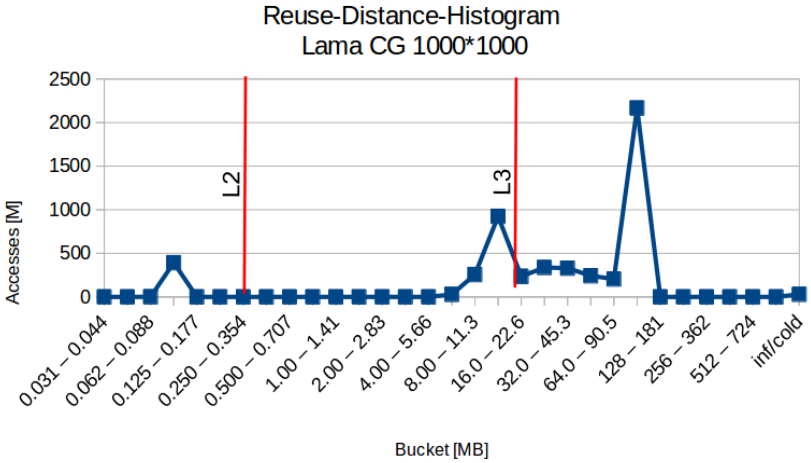
Ziele

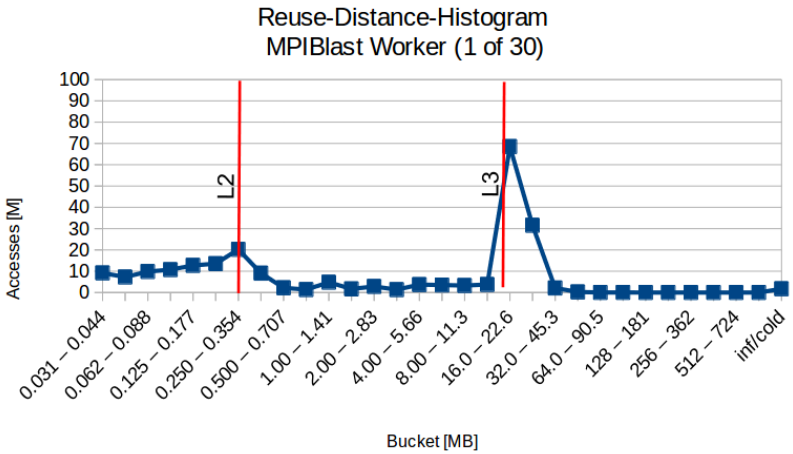
- Vermeidung (starker) gegenseitiger Ausbremsung von Anwendungen im Co-Scheduling aufgrund Nutzung der Speicherhierarchie
  - gemeinsam genutzt: L3, DRAM + verfügbare Bandbreiten
- Charakterisierung von Anwendungen bzgl. ihrer Nutzung der Speicherhierarchie
  - Offline: Analyse, um sich gegenseitig stark bremsende Anwendungen von vornherein auszuschließen
  - Online: dynamische Erfassung der aktuellen Auslastung
- Verfahren mit geringer Abhängigkeit von Prozessor-Eigenschaften (verfügbare Performance-Counter)

GA Gauß-Allianz

Analyse des Speicherzugriffsverhaltens

“Reuse”-Distanzen



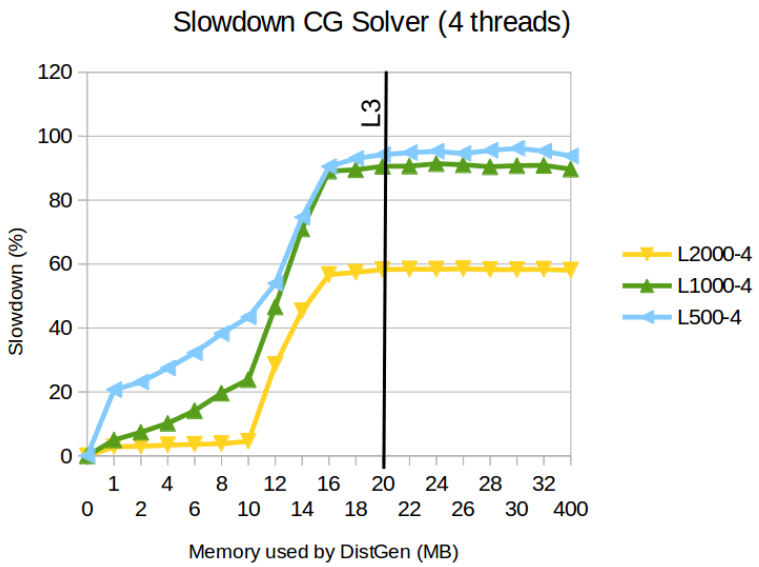


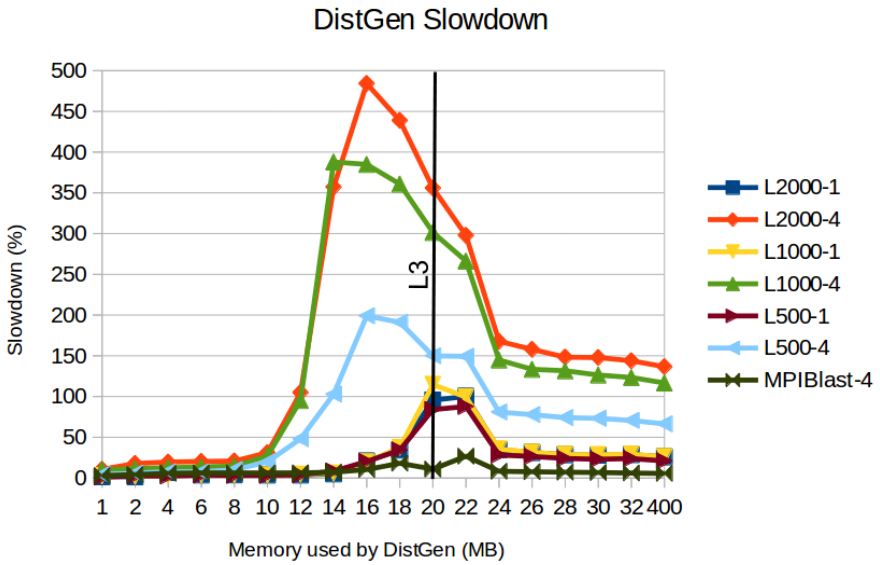
- Aussagen über Nutzung von Cache-Ebenen / Speicher
- schwer nutzbar für Vorhersagbarkeit von Bremseffekten (keine zeitlichen Informationen zur Berechnung von Korrelationen)

GA Gauß-Allianz

Analyse des Speicherzugriffsverhaltens

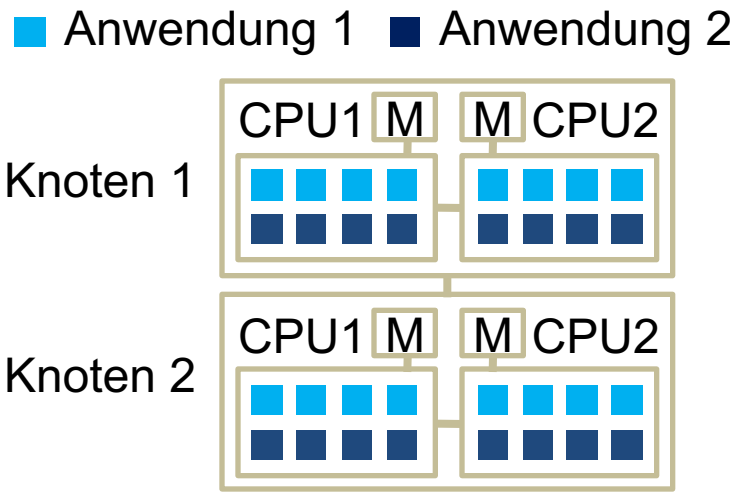
- Bremseffekte gegenüber Benchmark (Distgen) mit kontrollierbarem Verhalten

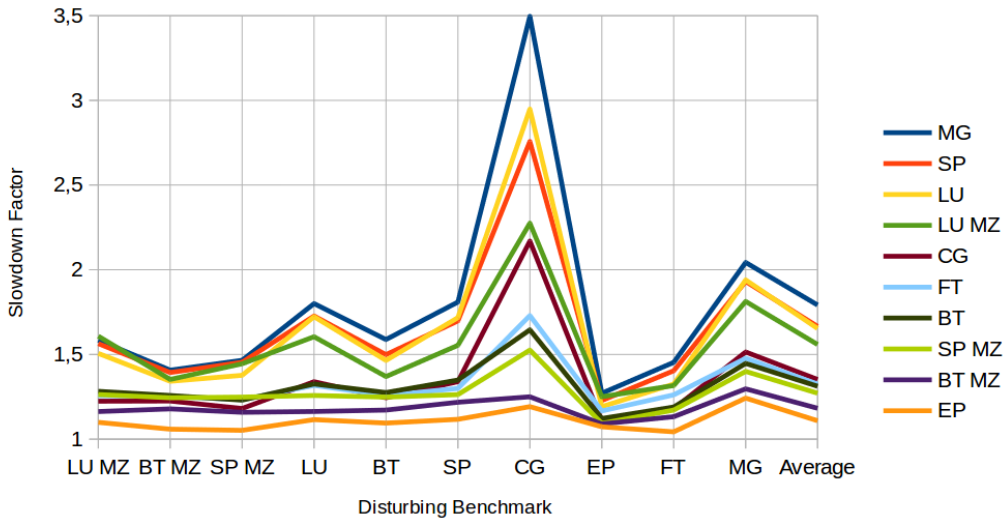




Gauß-Allianz

Gegenseitiges Ausbremsverhalten NPB (Class D)



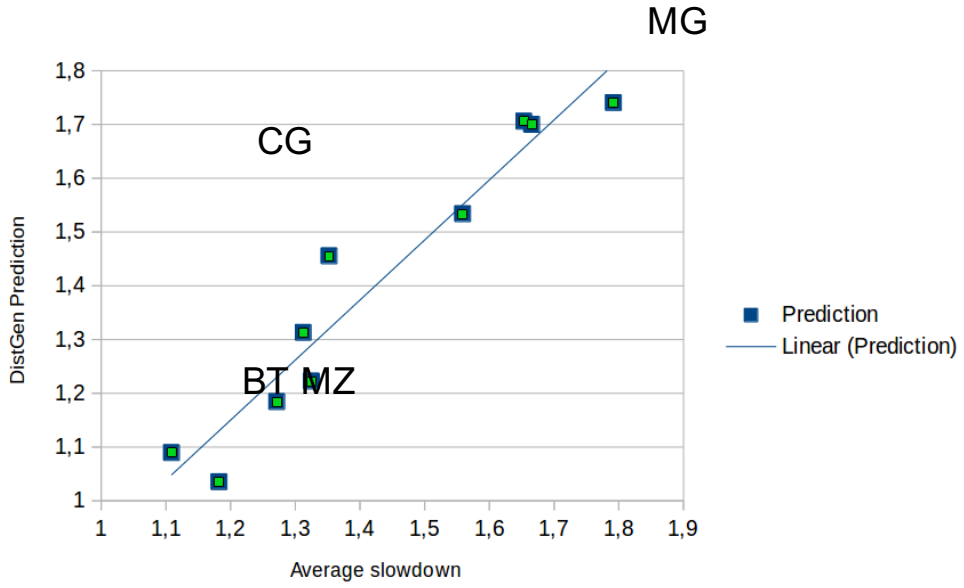


GA Gauß-Allianz

Vorhersage Bremsverhalten (NBP)

- Nutzung der Messung des Ausbremsverhaltens von Distgen (50 MB, 4 Cores)

➔ hohe Korrelation mit tatsächlichem Ausbremsen



Gauß-Allianz

Energieverbrauch für Scheduling-Entscheidungen

- Ziel: Neben Peformance-Charakterisierung auch Energieverbrauch für Scheduling-Entscheidungen verwenden

- Voraussetzungen:
  - Messwerte mit ausreichender Granularität & Genauigkeit
  - Passende Schnittstellen

- Umsetzung:
  - RAPL-Unterstützung zur Ermittlung des Energieverbrauchs von CPU und Hauptspeicher
  - Erhöhung der zeitlichen Granularität der Leistungsmessung von **MEGWARE** ClustSafe PDUs zur Messung des Verbrauchs ganzer Compute Nodes auf >1Hz

GA Gauß-Allianz

Herausforderung Exascale Computing

- Charakterisierung der Anwendungen
  - Statische Parameter, dynamisches Laufzeitverhalten
  - Daten einholen für Scheduling-Entscheidungen

- **Migration von Prozessen**
  - **Anpassung der MPI-Laufzeitumgebung**
  - **Minimieren notwendiger Datentransfers**

- Scheduling der Prozesse
  - Dynamisches Scheduling nach Ressourcenverbrauch und Laufzeitverhalten
  - Anpassung des Batchsystems
  - Durchsatz erhöhen und / oder Energie einsparen

GA Gauß-Allianz

Anwendungstransparenter Shutdown/Reconnect

Ziele

- Minimierung der Migrationskosten
  - Minimale Beeinflussung der eigentlichen Migration
- Vermeidung von Laufzeitoverhead
  - Keine Beeinflussung von Anwendungen die nicht migriert werden
- Hardwareunabhängigkeit
  - Der Lösungsansatz sollte für verschiedene Netze adaptierbar sein
- Anwendungstransparenz
  - Migration von MPI Anwendungen ohne Codeanpassungen

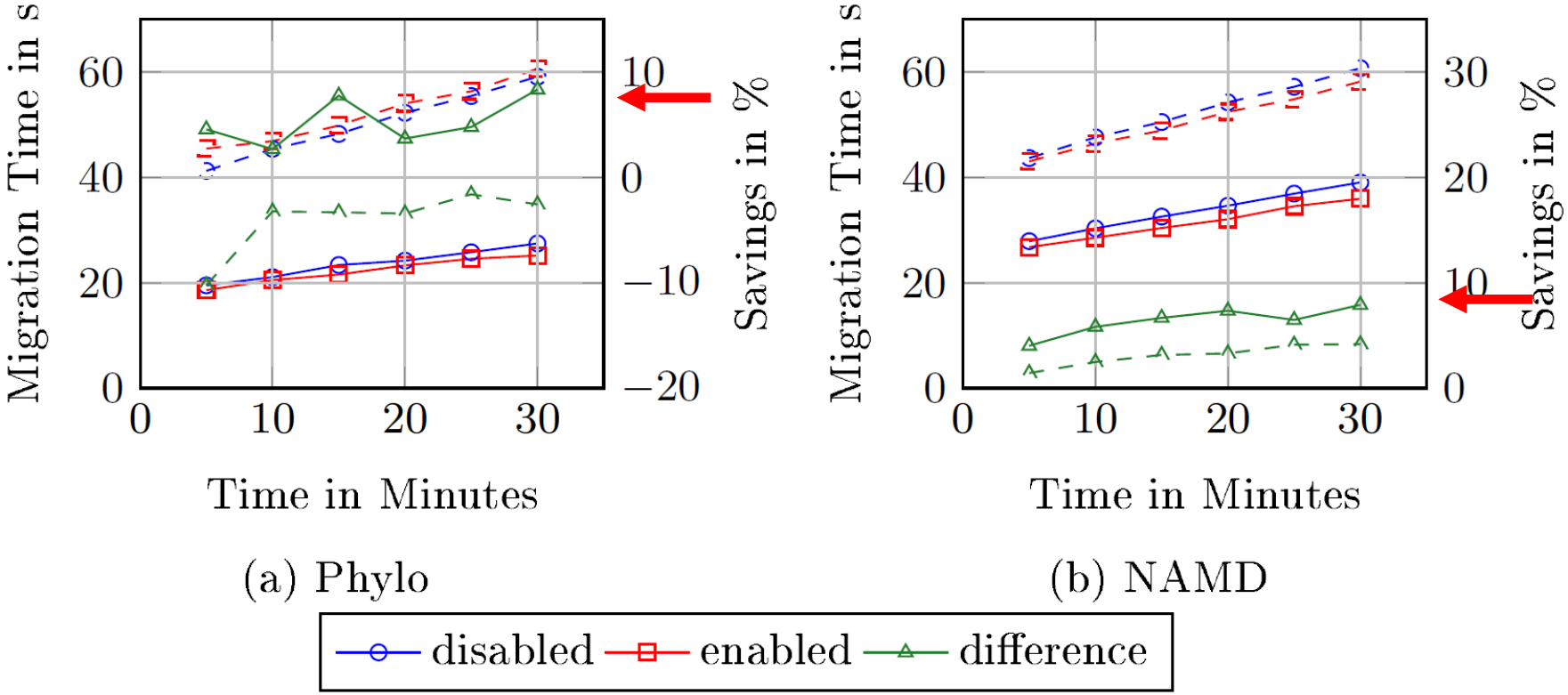
GA Gauß-Allianz

Minimieren notwendiger Datentransfers I

- Nicht genutzte Speicherbereiche einer Anwendung bleiben Bestandteil einer VM
- Zeroing-Ansatz: Überschreibe diese Bereiche mit „Nullen“
- Der Hypervisor erkennt anschließend Zero-Pages und komprimiert diese
- Die Größe des VM-Images wird hierdurch reduziert und die Migration beschleunigt

GA Gauß-Allianz

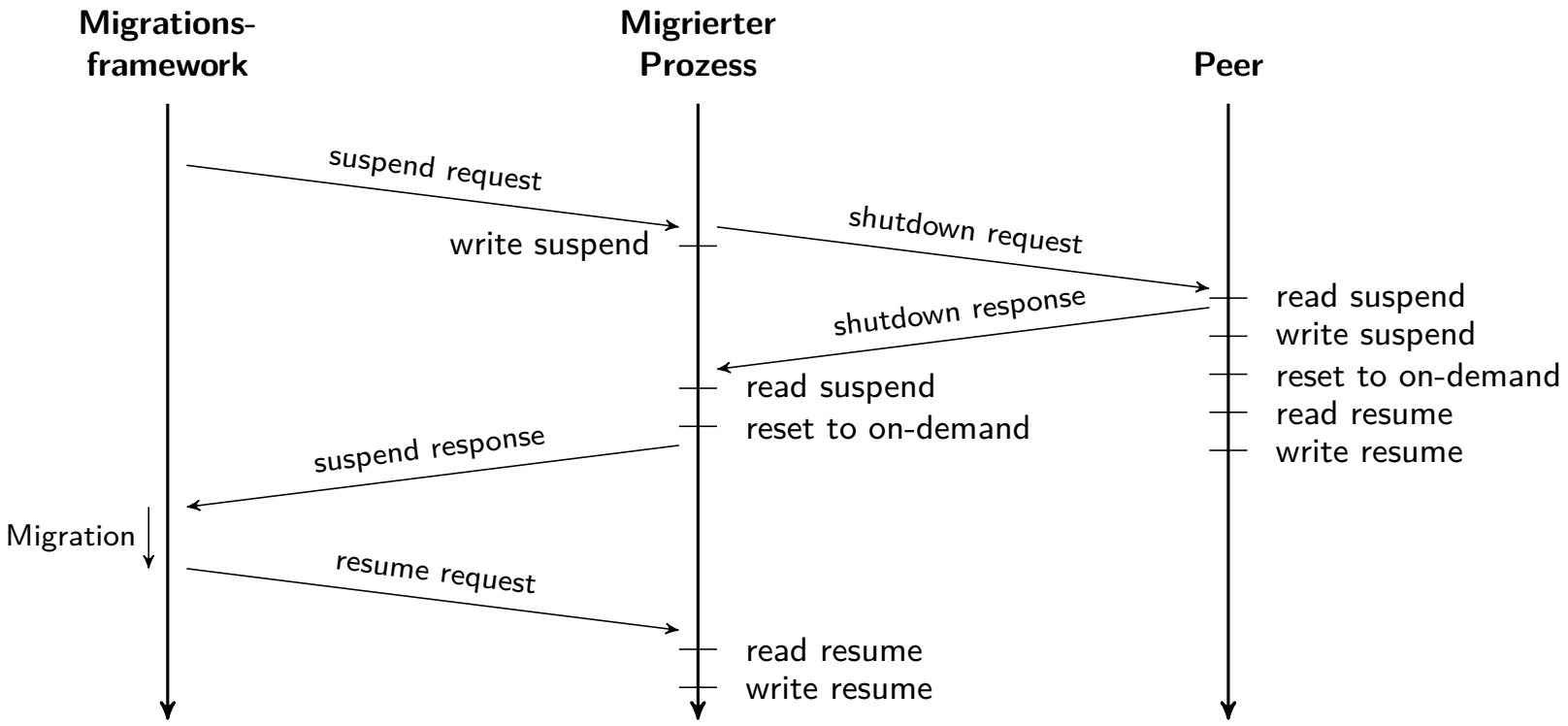
Minimieren notwendiger Datentransfers II



- Gestrichelte Kurven repräsentieren Ergebnisse ohne Kompression
- Einsparungen von ca. 8%

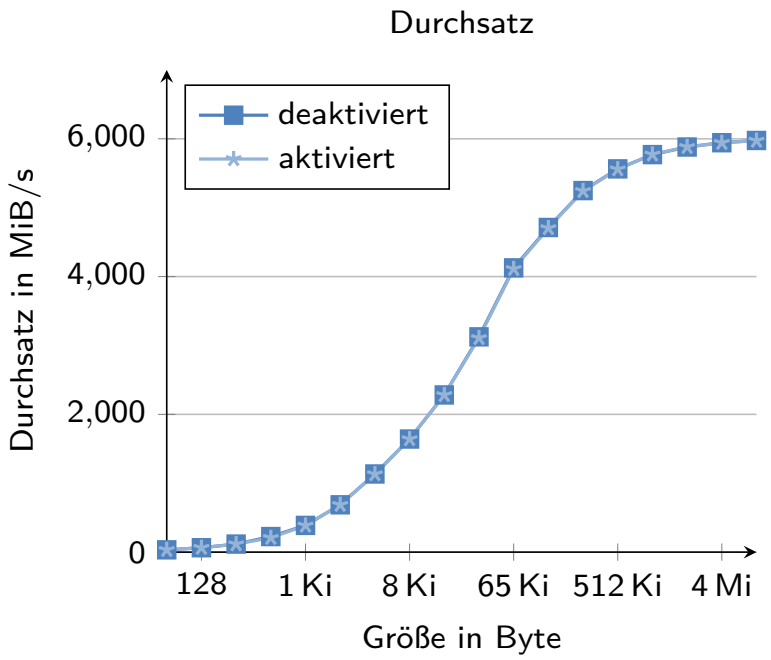
GA Gauß-Allianz

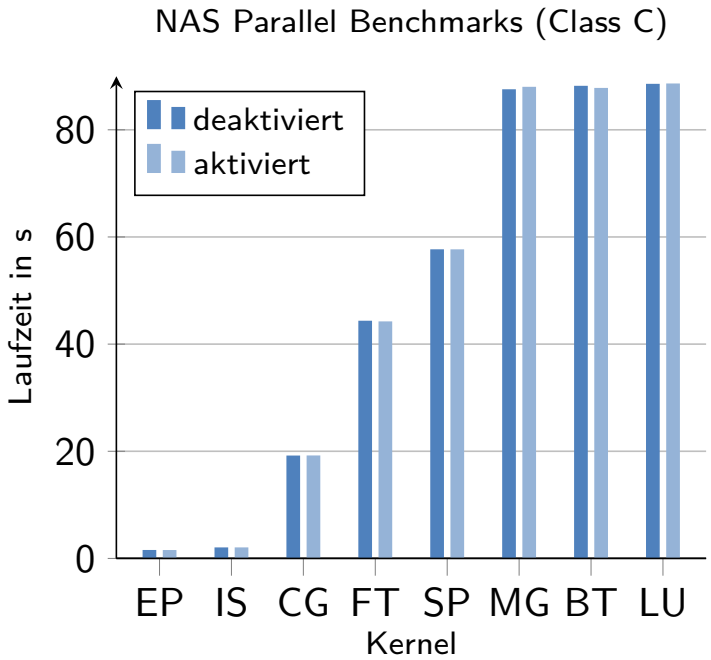
Das Shutdown/Reconnect Protokoll



Gauß-Allianz

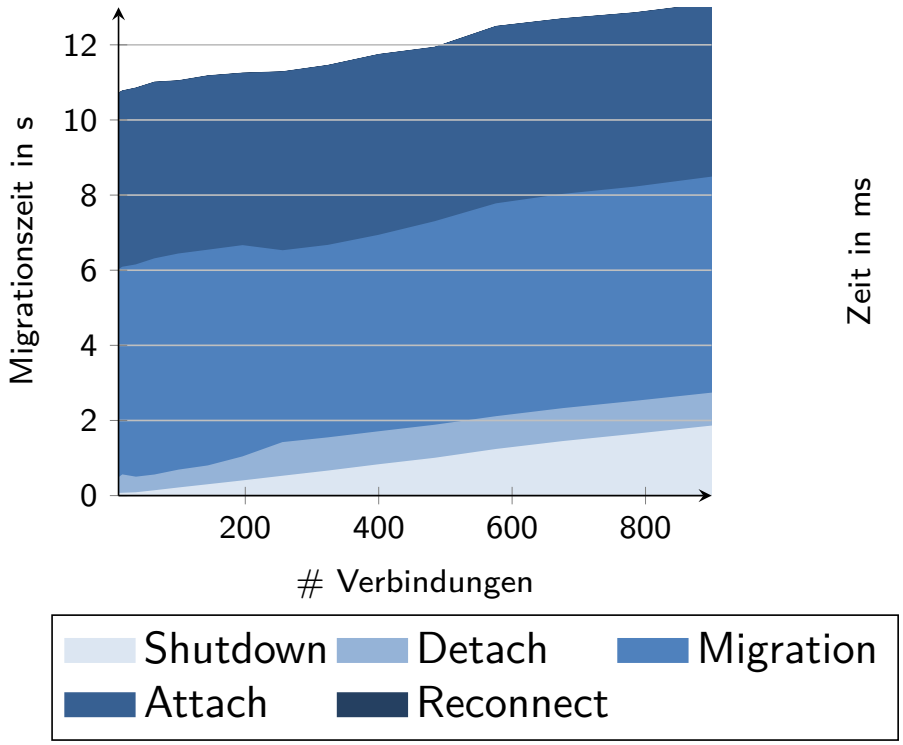
Laufzeitverhalten

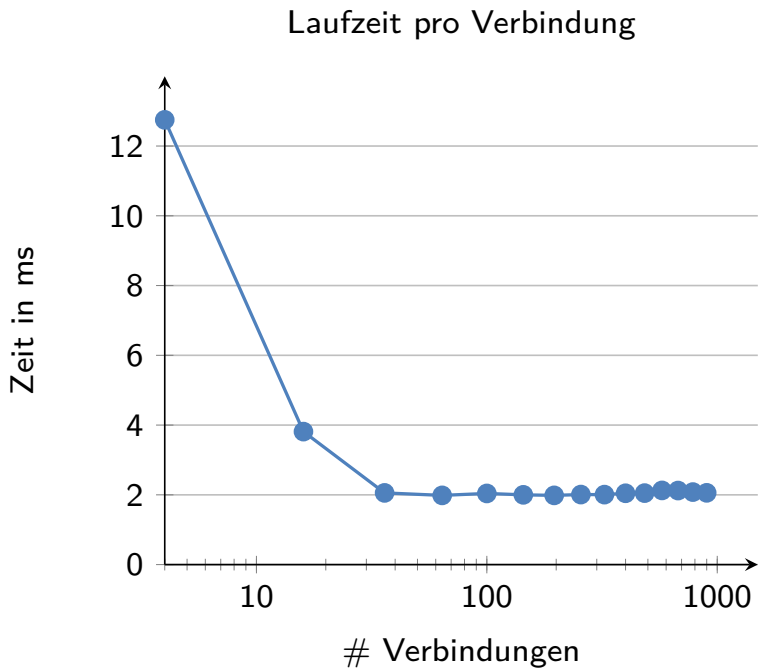




GA Gauß-Allianz

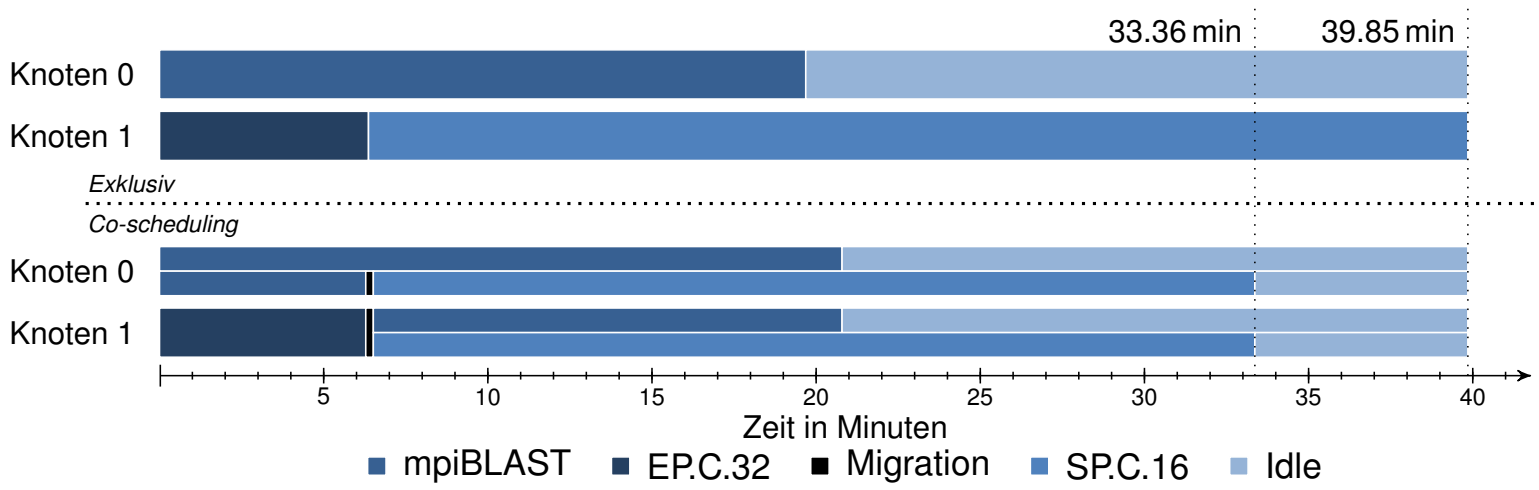
Skalierungsverhalten





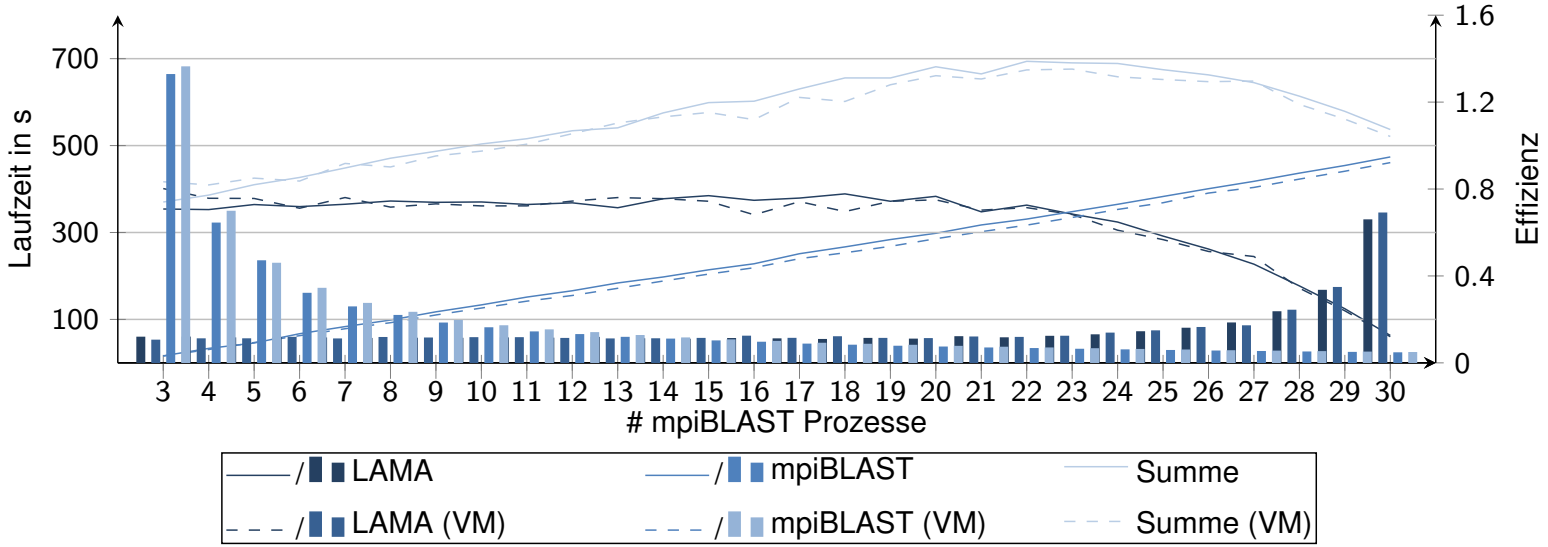
GA Gauß-Allianz

Co-scheduling virtueller Maschinen I



Gauß-Allianz

Co-scheduling virtueller Maschinen II



Gauß-Allianz

Herausforderung Exascale Computing

- Charakterisierung der Anwendungen
  - Statische Parameter, dynamisches Laufzeitverhalten
  - Daten einholen für Scheduling-Entscheidungen

- Migration von Prozessen
  - Anpassung der MPI-Laufzeitumgebung
  - Minimieren notwendiger Datentransfers

- **Scheduling der Prozesse**
  - **Dynamisches Scheduling nach Ressourcenverbrauch und Laufzeitverhalten**
  - **Anpassung des Batchsystems**
  - **Durchsatz erhöhen und / oder Energie einsparen**

GA Gauß-Allianz

Scheduling

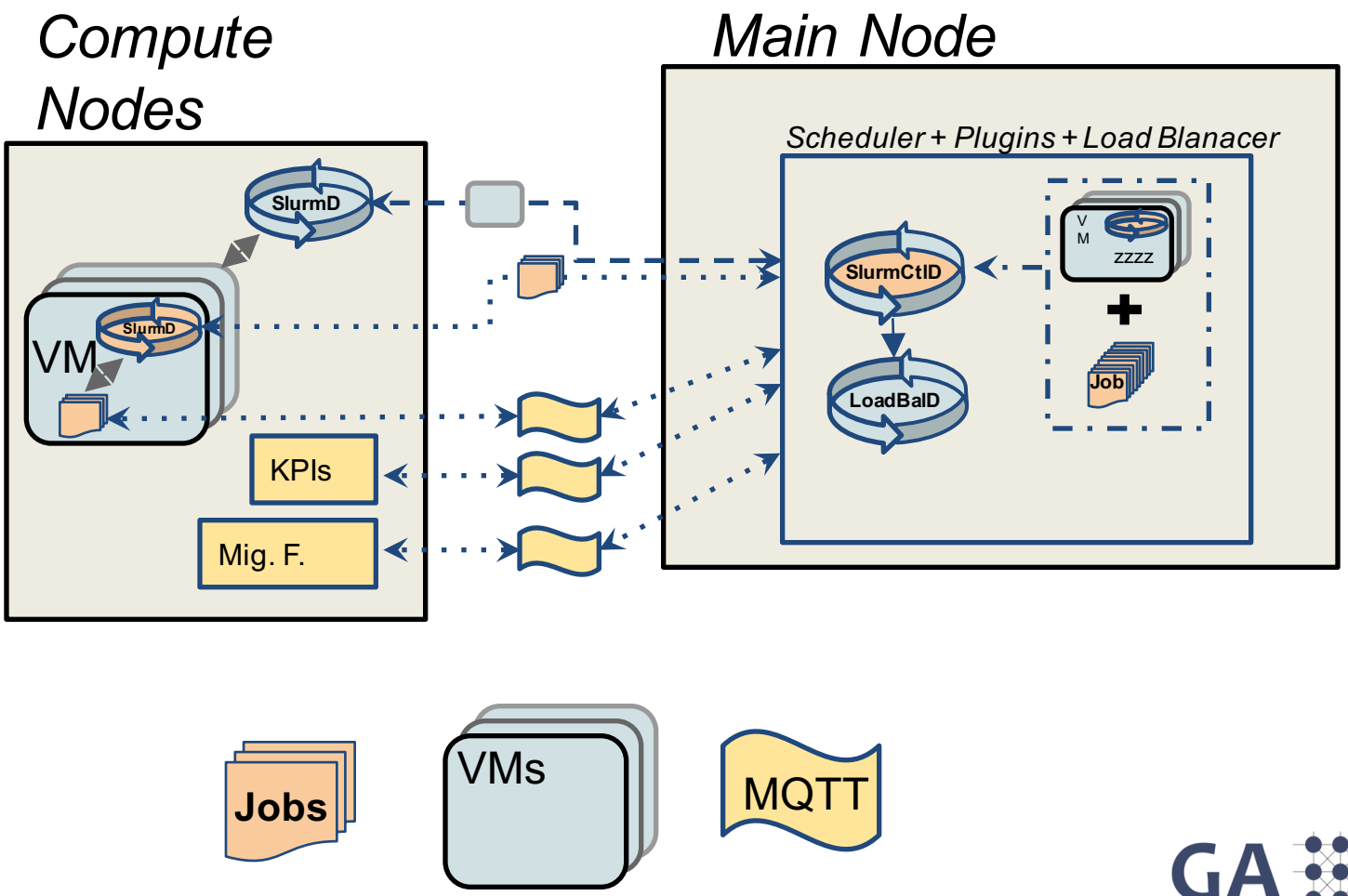
Entwicklung neuer Scheduling-Algorithmen für anforderungsorientiere Zuordnung / Gruppierung von Prozessen

1. Zentrales Re-scheduling mit SLURM
   - Key Performance Indicators (KPI): Ressourcenauslastung, Stromverbrauch, Informationen von den Anwendungen über den zukünftigen Ressourcenverbrauch
   - Erweiterung von SLURM, so dass es die KPIs berücksichtigen kann

2. Dezentrales Re-scheduling:
   - Initiales Scheduling mit (erweitertem) SLURM
   - Nachoptimierung des Schedules, die von den Knoten angestoßen wird
   - Peer-to-peer-Verfahren, die Prozesse migrieren
   - Ergebnisse der Nachoptimierung werden an SLURM zurückgegeben, um das initiale Scheduling zu optimieren

GA Gauß-Allianz

FAST SLURM's inner workings



Gauß-Allianz

Anwendungsintegration

- Runtime-System:
  - Virtual Machines
  - Scheduler

- Schnittstelle zur Anwendung: MQTT-Messages (durch fastlib)
  - liefert Key-Performance-Indikatoren über den Zustand der Anwendung
    - IO-Intensity
    - Compute-Intensity
    - Communication-Intensity

- Input-Geber:
  - Agent
  - Anwendung

GA Gauß-Allianz

Anwendungspartner Fraunhofer SCAI: LAMA

LAMA High Performance Library für Heterogene Rechen-Cluster

- Besonderheit: Anwendung ist eine Bibliothek, die die Instrumentierung für die Interaktion mit dem Runtime-System übernimmt

- Vorteil: End-Anwender muss nicht selbst instrumentieren
- Nachteil: Library hat kein Modelwissen über die Anwendung

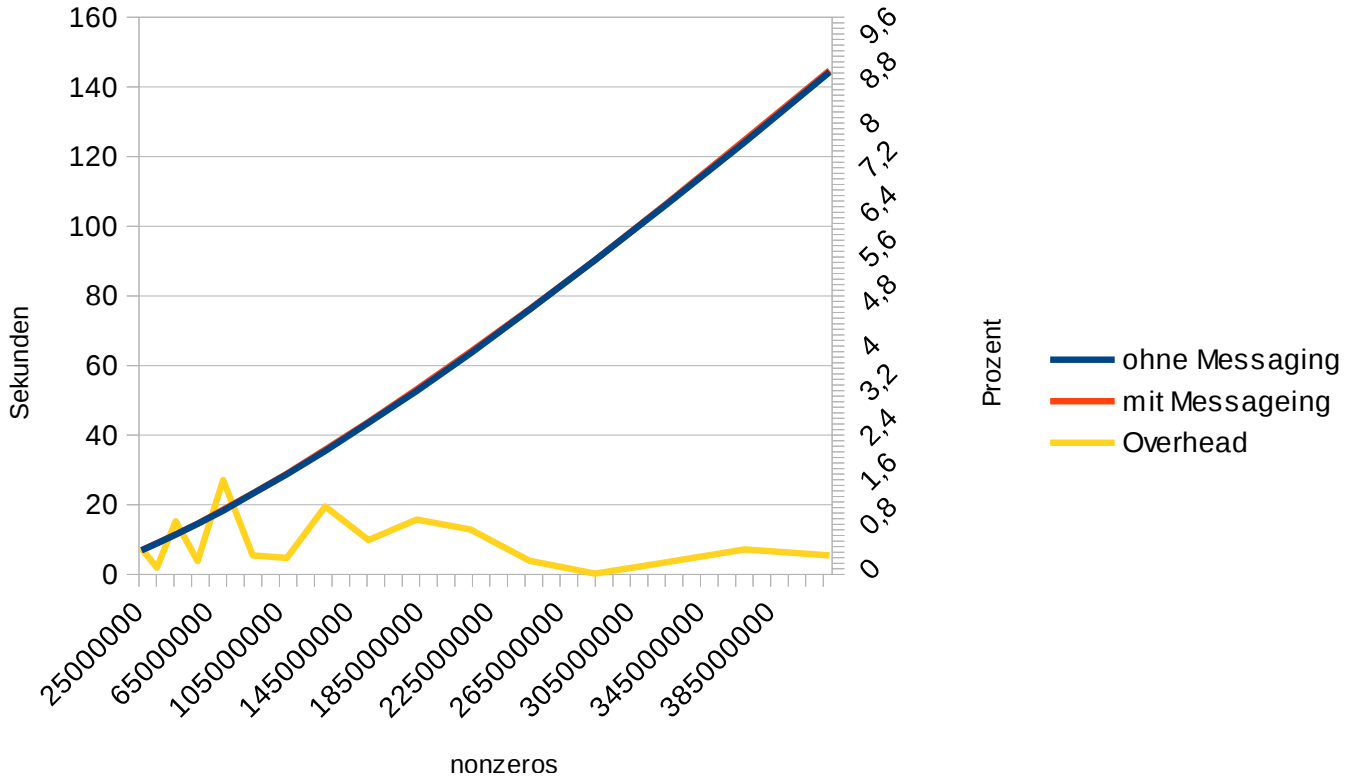
- Deswegen: Instrumentierung erfolgt auf Basis von Low-Level-Routinen

  → erzeugt eine große Menge an Input für den Scheduler

GA Gauß-Allianz

Messaging-Overhead

(Beispiel: Lösen eines großen dünn besetzten linearen Gleichungssystems mittels CG)



Overhead liegt nur bei wenigen (0,01 – 1,7) Prozent der Gesamtlaufzeit

GA Gauß-Allianz

Weiter-/Neuentwicklungen an LAMA

- Aufdeckung von Optimierungsmöglichkeiten durch das Agentensystem
- Einführung von parallelem I/O für ein einheitlicheres & balancierteres Verhalten in den I/O-Phasen (innerhalb dessen meist nicht migriert werden sollte – Anfang/Ende der Anwendung)
- Neue Lastbalancierungsstrategien durch Graph-Partitionierung
  - für balancierteres Verhalten der parallelen Anwendung
  - führt zu besseren Prognosen des Laufzeitverhalten nach Repartitionierung

WAVE

GA Gauß-Allianz

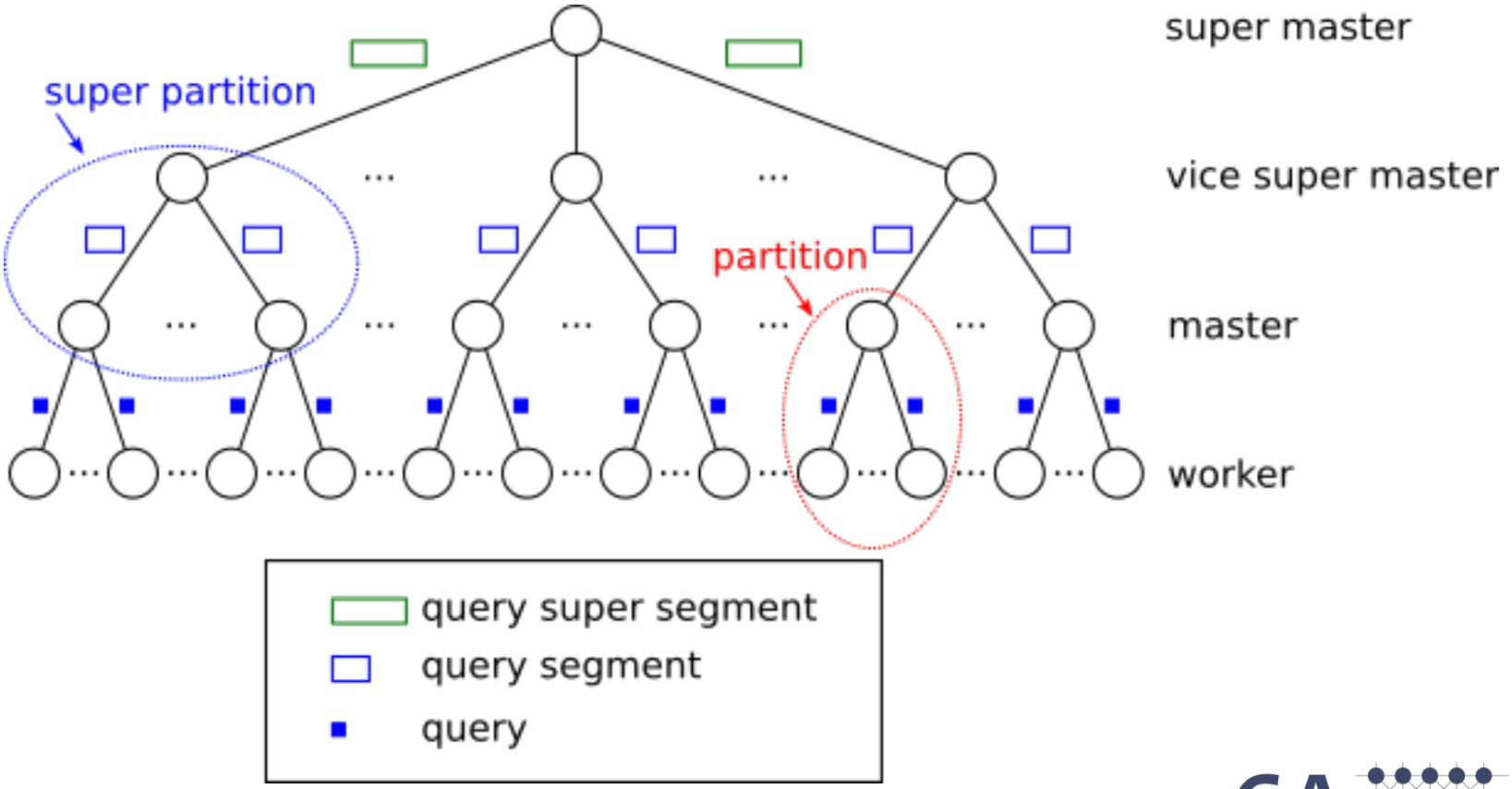
Evaluation von FaST-Mechanismen durch mpiBLAST

- Optimierungen von mpiBLAST
  - Sleep-Länge/-Frequenz zwischen Anfragen
  - I/O: Caching des I/O Zugriffs
  - Dynamische Anpassung der Segment-granularität
  - Erweiterung der hierarchischen Kommunikationsstruktur für Exascale Architekturen

- Kommunikation zwischen mpiBLAST und FaST mittels MQTT Nachrichten

GA Gauß-Allianz

Optimierung hierarchischen Kommunikationsstruktur



Gauß-Allianz

Status/KPI Nachrichten von mpiBLAST

Status nach Start	KPI: Distribution Phase	KPI: Compute Phase
task: status source: SLURM_JOB_ID migration: yes supported_KPI: - comm. intensity (network) - IO intensity - compute intensity	task: KPI source: SLURM_JOB_ID KPIS: - comm. Intensity: high - IO intensity: high - compute intensity: low	task: KPI source: SLURM_JOB_ID KPIS: - comm. Intensity: medium - IO intensity: low - compute intensity: high

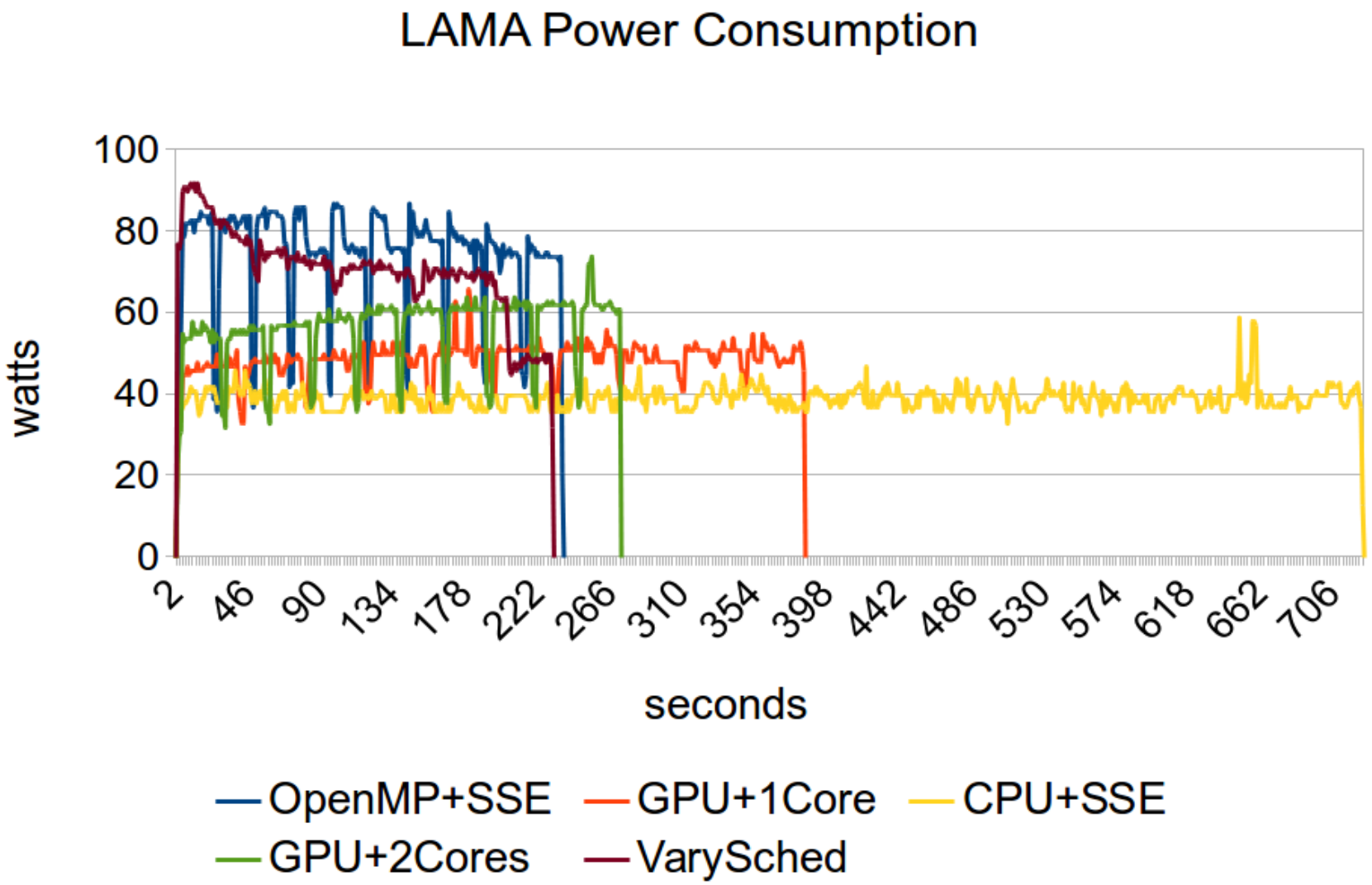
- Verschicken von MQTT Nachrichten zu Beginn der Anwendung und der beiden Phasen ermöglicht:
  - Co-Scheduling
  - Ressourcenoptimierung

GA Gauß-Allianz

VarySched

- **Ausnutzen aller Ressourcen**: Scheduling von Rechenjobs mit dem Ziel, möglichst alle Ressourcen der Knoten zu nutzen: CPUs, GPUs, ...
- **Aufspalten der Programme in Kernel**: Die Programme, deren Code vorliegen muss, werden in Kernel (Subjobs) aufgespalten, die sich viel flexibler auf die Ressourcen verteilen lassen
- **Automatische Code-Generierung**: Codes für die verschiedenen Ressourcen können manuell oder mit Hilfe von Werkzeugen wie PluTo-SICA automatisch erzeugt werden
- **Scheduling unter Berücksichtigung von Präferenzen**: VarySched verteilt die Kernel auf die Ressourcen, wobei der Scheduler berücksichtigt, auf welchen Ressourcen sie am besten laufen
- **Auswahl der Schedulingstrategie**: Zur Laufzeit kann die Strategie jederzeit geändert werden, um z. B. Energieverbrauch oder Durchsatz zu optimieren

VarySched: Laufzeit & Energieverbrauch im Vergleich



ParaStation Integration in FAST

Implementierung eines Mechanismus zum dynamischen Ab- und Wiederaufbau von Verbindungen

Erfolgsbericht der prototypischen Implementierung für InfiniBand:

- leicht auf andere Netzwerktechnologien zu übertragen · verwendet skalierbare In-Band-Signalisierung
- von externen Instanzen durch MQTT triggerbar
- vollkommen transparent für die MPI-Anwendungen
- minimaler Overhead bei Verwendung
- keinerlei Laufzeitverluste bei Nichtverwendung
- nachweislich hoch skalierbar
- wiederverwendbar für Checkpoint/Restart

GA Gauß-Allianz

ParaStation Integration in FAST

Integration und Anbindung von Job-Scheduler und Migrationsrahmenwerk mit dem ParaStation Prozessmanager

Geplante Einbettung der ParaStation Management Dämonen:

- unproblematische Realisierung dank der einfachen Konfigurationsmöglichkeiten der ParaStation-Umgebung
- lediglich eine geringfügige Anpassung von Skripten erforderlich
- erfolgreich auf dem Testcluster eingerichtet und getestet
- Grundlage für zukünftige Projekte und Produkte

GA Gauß-Allianz

Integration in Cluster Management Systeme

ClustWare-Appliance ist MEGWAREs Cluster Management Software

- Web-basierte Oberfläche für
  - Monitoring (Erfassung, Überwachung, Archivierung und Visualisierung aller Hardwareparameter)
  - Management (Ferngesteuertes Ein-/Ausschalten sowie Reset von Knoten, Steuerung von Infrastruktur-Komponenten wie PDUs, Klima, o.ä.)
- Neu in FAST: Integration von Batch Scheduling Systemen (SLURM)
  - Visualisierung von Jobs, Node-Zuordnungen, etc.
  - Archivierung der (Co-)Scheduling-Daten
  - Automatische Erweiterung bestehender ClustWare-Funktionalitäten auf die Scheduler-Anbindung (z.B. automatisches Herausnehmen von Knoten aus dem Scheduler bei Wartungsfällen)

GA Gauß-Allianz

Dissemination

Veröffentlichungen 2016

- Josef Weidendorfer, Jens Breitbart: **Inclusive Cost Attribution for Cache Use Profiling**, International Conference on Computational Science (ICCS), San Diego, California, June 2016.
- Simon Pickartz, Stefan Lankes, Antonello Monti, Carsten Clauss, Jens Breitbart: **Application Migration in HPC -- A Driver of the Exascale Era?** International Conference on High Performance Computing Simulation (HPCS), July 2016.
- Petra Berenbrink, Tom Friedetzky, Peter Kling, Frederik Mallmann-Trenn, Lars Nagel, Christopher Wastell: **Self-stabilizing Balls & Bins in Batches: The Power of Leaky Bins**. Proceedings of the 2016 ACM Symposium on Principles of Distributed Computing (PODC), July 2016.
- Jens Breitbart, Simon Pickartz, Josef Weidendorfer, Antonello Monti: **Viability of Virtual Machines in HPC**. ROME Workshop at Euro-Par, Grenoble, France, August 2016.

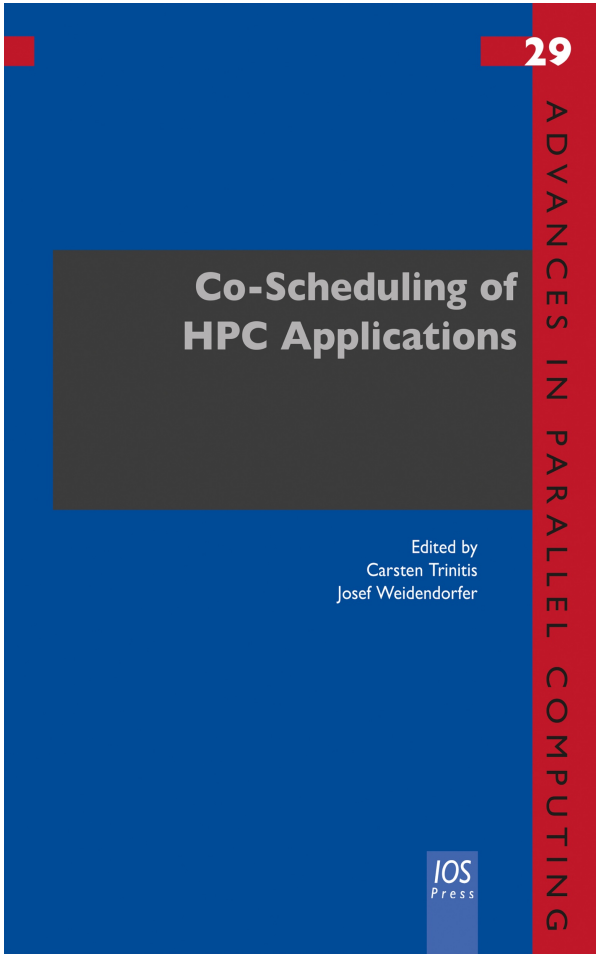
GA Gauß-Allianz

Dissemination

- Andreas Kohne, Damian Pasternak, Lars Nagel, Olaf Spinczyk: **Evaluation of SLA-based decision strategies for VM scheduling in cloud data centers**. In Proceedings of the 3rd Workshop on CrossCloud Infrastructures & Platforms, CrossCloud 2016, London, April 2016.
- Simon Pickartz, Carsten Clauss, Stefan Lankes, Stephan Krempel, Thomas Moschny, Antonello Monti: **Non-Intrusive Migration of MPI Processes in OS-bypass Networks**. Proceedings of the 2016 IEEE International Symposium on Parallel and Distributed Processing Workshops and PhD Forum, May 2016.
- Jens Breitbart, Josef Weidendorfer, Carsten Trinitis: **Automatic Co-Scheduling Based on Main Memory Bandwidth Usage**. 20th Workshop on Job Scheduling Strategies for Parallel Processing (JSSPP) at IPDPS, Chicago, IL, May 2016.
- Simon Pickartz, Niklas Eiling, Stefan Lankes, Lukas Razik, Antonello Monti: **Migrating LinuX Containers Using CRIU**. In Proceedings of the 11th Workshop on Virtualization in High-Performance Cloud Computing (VHPC), June 2016.
- Ramy Gad, Simon Pickartz, Tim Süß, Lars Nagel, Stefan Lankes, André Brinkmann: **Accelerating Application Migration in HPC**. In Proceedings of the 11th Workshop on Virtualization in High-Performance Cloud Computing (VHPC), June 2016.

GA
Gauß-Allianz

Dissemination



GA Gauß-Allianz

Dissemination

- Carsten Clauss, Thomas Moschny, Norbert Eicker: **Dynamic Process Management with Allocation-internal Co-Scheduling Towards Interactive Supercomputing**. 1st Workshop on Co-Scheduling of HPC Applications (COSH 2016). Prague, Czech Republic, 2016.
- Josef Weidendorfer, Jens Breitbart: **Detailed Characterization of HPC Applications for Co-Scheduling**. 1st Workshop on Co-Scheduling of HPC Applications (COSH 2016). Prague, Czech Republic, 2016.
- Simon Pickartz, Jens Breitbart, Stefan Lankes: **Implications of Process Migration in Virtualized Environments**. 1st Workshop on Co-Scheduling of HPC Applications (COSH 2016). Prague, Czech Republic, 2016.
- Tim Süß, Nils Döring, Ramy Gad, Lars Nagel, André Brinkmann, Dustin Feld, Eric Schricker, Thomas Soddemann: **Impact of the Scheduling Strategy in Heterogeneous Systems That Provide Co-Scheduling**. 1st Workshop on Co-Scheduling of HPC Applications (COSH 2016). Prague, Czech Republic, 2016.
- Josef Weidendorfer, Carsten Trinitis: **Proceedings of the First Workshop on Co-Scheduling of HPC Applications**. 1st Workshop on Co-Scheduling of HPC Applications (COSH 2016). Prague, Czech Republic, 2016.

GA
Gauß-Allianz

Dissemination

- T. Süß, N. Döring, R. Gad, L. Nagel, A. Brinkmann, D. Feld, T. Soddemann, S. Lankes: **VarySched: A Framework for Variable Scheduling in Heterogeneous Environments**, IEEE Cluster 2016
- J. Kaiser, R. Gad, T. Süß, L. Nagel, A. Brinkmann: **Deduplication Potential of HPC Applications' Checkpoints**, IEEE Cluster 2016

GA Gauß-Allianz

Dissemination

Veröffentlichungen 2015

- Matthias Grawinkel, Michael Mardaus, Tim Süß, André Brinkmann: **Evaluation of a Hash-Compress-Encrypt for Storage System Applications Pipeline**. A poster at 10th IEEE International Conference on Networking, Architecture, and Storage (NAS), August 6-7, Boston, USA
- Stefan Lankes: **Revisiting Co-scheduling for Upcoming ExaScale Systems**. International Conference on High Performance Computing & Simulation (HPCS 2015), Amsterdam, Netherlands, 2015.
- Jens Breitbart, Carsten Trinitis, Josef Weidendorfer: **Case Study on Co-Scheduling for HPC Applications**. International Workshop on Scheduling and Resource Management for Parallel and Distributed Systems (SRMPDS 2015). Beijing, China, 2015.

GA Gauß-Allianz

Dissemination

Veröffentlichungen 2014

- Petra Berenbrink, André Brinkmann, Tom Friedetzky, Lars Nagel: **Balls into non-uniform bins**. J. Parallel Distrib. Comput. 74(2): 2065-2076 (2014)
- André Brinkmann, Peter Kling, Friedhelm Meyer auf der Heide, Lars Nagel, Sören Riechers, Tim Süß: **Scheduling shared continuous resources on many-cores**. In Proceedings of the 26th ACM Symposium on Parallel Algorithms and Architectures (SPAA), Prague, Czech Republic 2014: 128-137
- Ramy Gad, Tim Süß, André Brinkmannn: **Compiler Driven Automatic Kernel Context Migration for Heterogeneous Computing**. In Proceedings of the 34th International Conference on Distributed Computing Systems (ICDCS), 2014, Madrid, Spain
- Simon Pickartz, Ramy Gad, Stefan Lankes, Lars Nagel, Tim Süß, André Brinkmann, Stephan Krempel: **Migration Techniques in HPC Environments**. In Proceedings of the 9th Workshop on Virtualization in High-Performance Cloud Computing (VHPC) held in conjunction with Euro-Par 2014, August 25-29, Porto, Portugal

GA Gauß-Allianz

Weitere Informationen erhalten Sie unter:

Prof. Dr.-Ing. André Brinkmann

Johannes Gutenberg-Universität Mainz
Zentrum für Datenverarbeitung
D-55128 Mainz

Telefon: +49 6131 3926390
Fax: +49 6131 3926407

E-mail: brinkman@uni-mainz.de
WWW: research.zdv.uni-mainz.de
WWW: fast-project.de

GA
Gauß-Allianz