Projekt „Parallele Systeme“

Anna Fuchs, Jannek Squar

Arbeitsbereich Wissenschaftliches Rechnen
Fachbereich Informatik
Fakultät für Mathematik, Informatik und Naturwissenschaften
Universität Hamburg

17.10.2024

HPC

- High Performance Computing
  - Hochleistungsrechnen (HLR)
- Wissenschaftliche Simulationen
  - Große Rechenlast, große Datenmengen
  - Serielle (sequentielle) Abarbeitung zu langsam

HPC

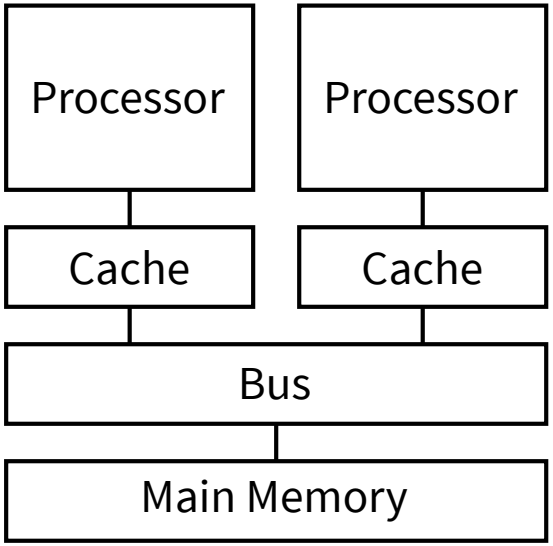
- High Performance Computing
  - Hochleistungsrechnen (HLR)
- Wissenschaftliche Simulationen
  - Große Rechenlast, große Datenmengen
  - Serielle (sequentielle) Abarbeitung zu langsam
- Parallele Programme
  - Parallele Berechnung
  - Parallele Ein-/Ausgabe
  - Parallele Visualisierung
  - Parallele Architektur
  - etc.

HPC

- High Performance Computing
  - Hochleistungsrechnen (HLR)
- Wissenschaftliche Simulationen
  - Große Rechenlast, große Datenmengen
  - Serielle (sequentielle) Abarbeitung zu langsam
- Parallele Programme
  - Parallele Berechnung
  - Parallele Ein-/Ausgabe
  - Parallele Visualisierung
  - Parallele Architektur
  - etc.
- Große Rechner
  - Top500-Liste
  - Top1: 8.699.904 Kerne mit 1.194,00 PFlop/s

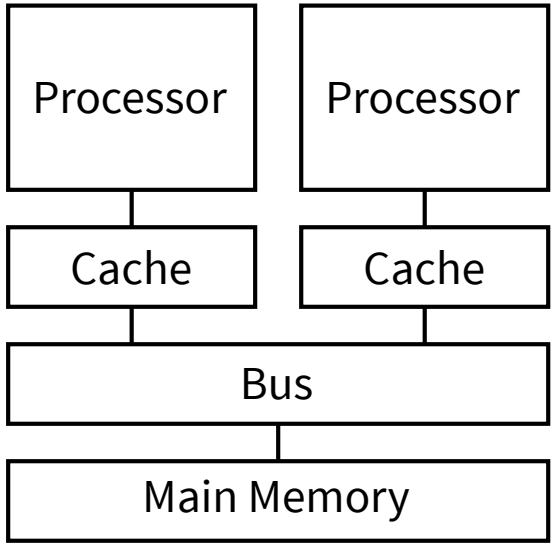
Gemeinsamer Speicher

- Knoten
  - Mehrere Prozessoren
  - Einzelner Rechner



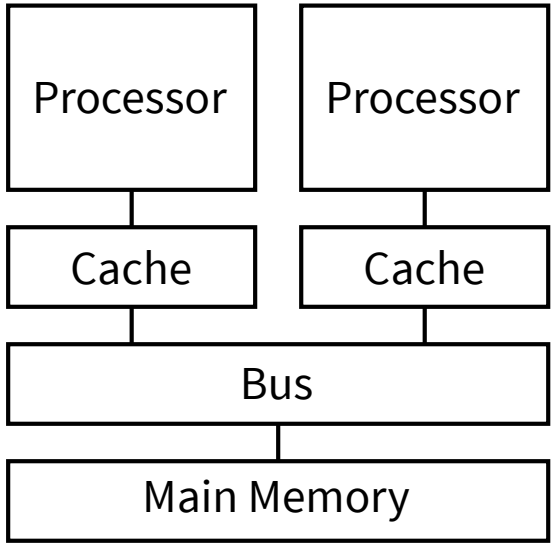
Gemeinsamer Speicher

- Knoten
  - Mehrere Prozessoren
  - Einzelner Rechner
- Mehrere Kerne pro Prozessor
  - Jeder Kern mit eigenen Caches
  - Gemeinsamer Speicher mit Zugriff per Bus
  - Limitierte Skalierung
  - Threads, OpenMP



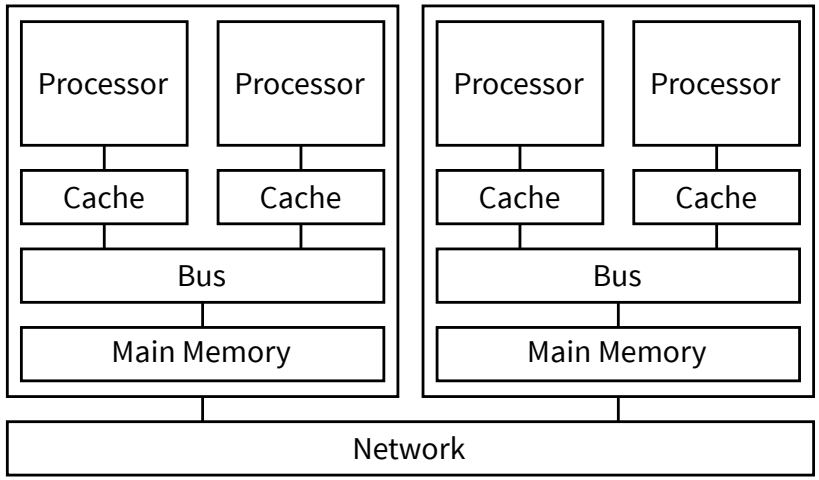
Gemeinsamer Speicher

- Knoten
  - Mehrere Prozessoren
  - Einzelner Rechner
- Mehrere Kerne pro Prozessor
  - Jeder Kern mit eigenen Caches
  - Gemeinsamer Speicher mit Zugriff per Bus
  - Limitierte Skalierung
  - Threads, OpenMP
- Beschleuniger
  - Meist GPUs
  - Andere Spracherweiterungen
  - Eigener Speicher/Adressraum
  - CUDA, OpenCL



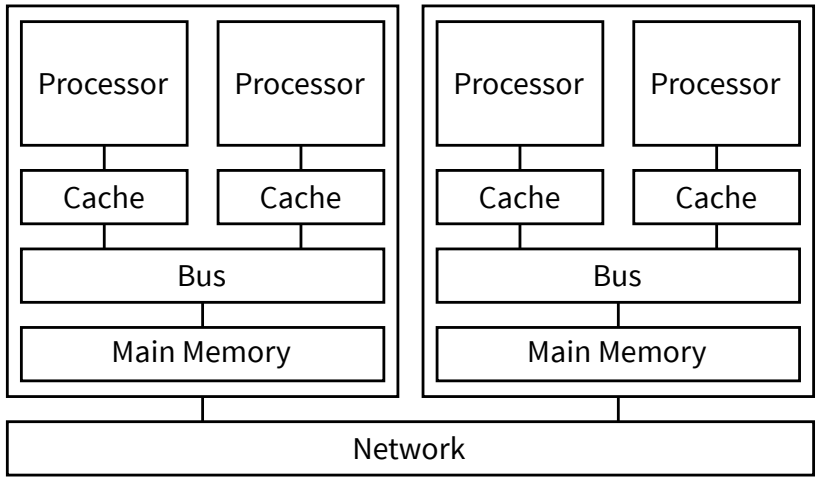
Verteilter Speicher

- Cluster von Knoten
- Berechnungen werden verteilt
  - Braucht Kommunikation über Knoten hinweg
  - Zusätzliche Software-Bibs
  - MPI - Senden, Empfangen, Synchronisieren



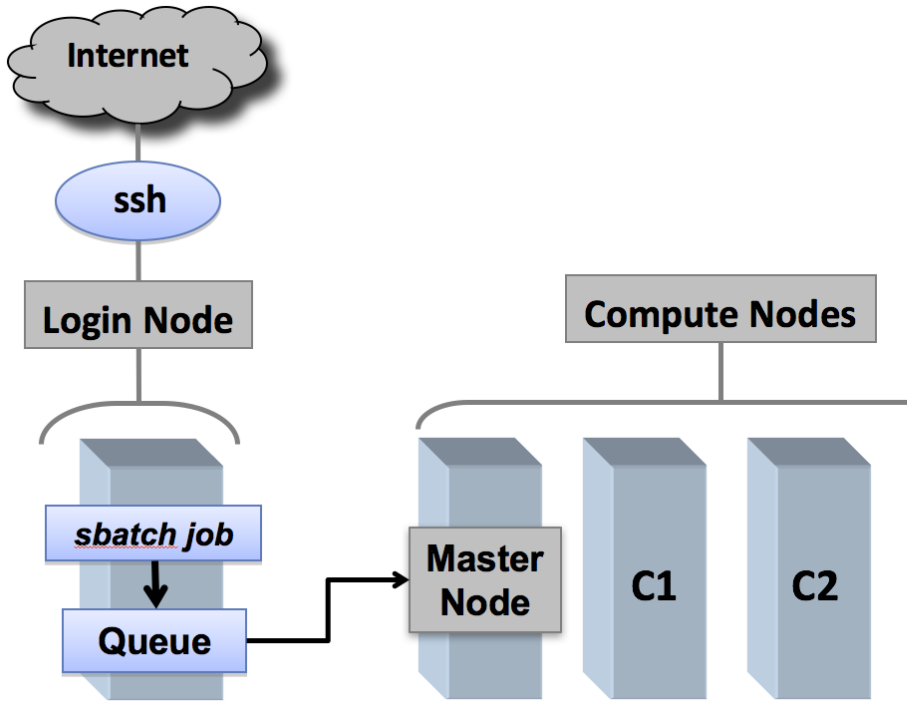
Verteilter Speicher

- Cluster von Knoten
- Berechnungen werden verteilt
  - Braucht Kommunikation über Knoten hinweg
  - Zusätzliche Software-Bibs
  - MPI - Senden, Empfangen, Synchronisieren
- Ausgabe von mehreren Knoten in z.B. eine Datei
  - Verteilte Dateisysteme - Lustre
  - Neue Bibliotheken für Synchronisation - MPI

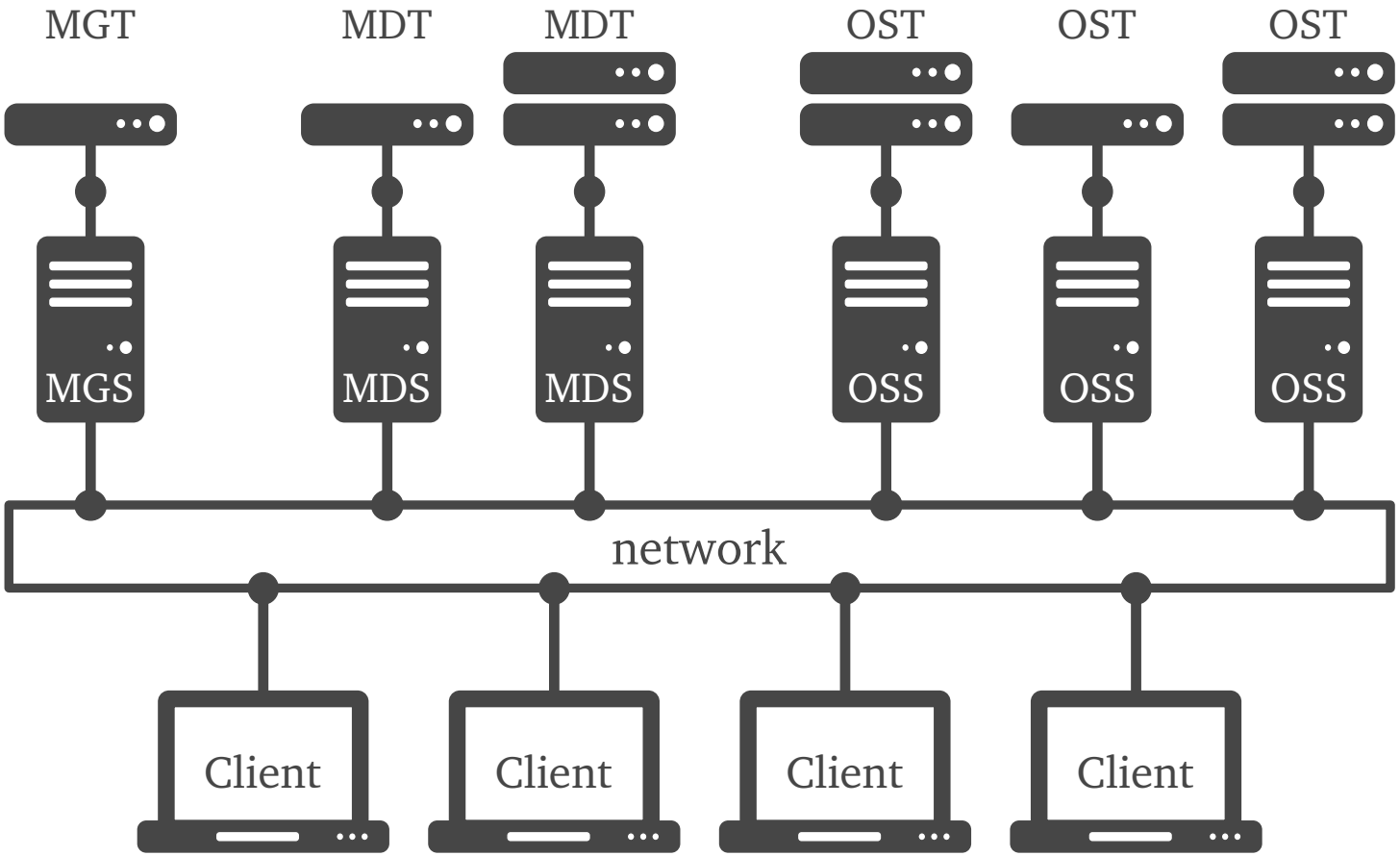


Soft

- Workload manager
- Viele Nutzer gleichzeitig
- Limitierte Ressourcen
- SLURM, flux
- Maximale Auslastung, keine Idle-Zeilen, Energieeffizienz

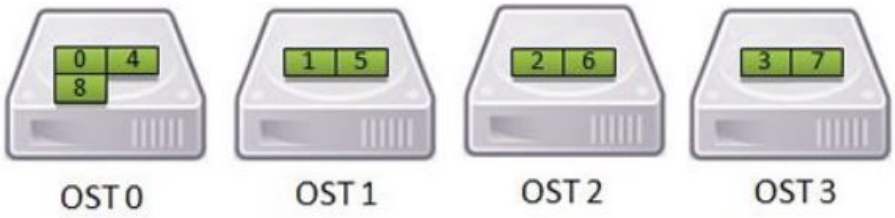


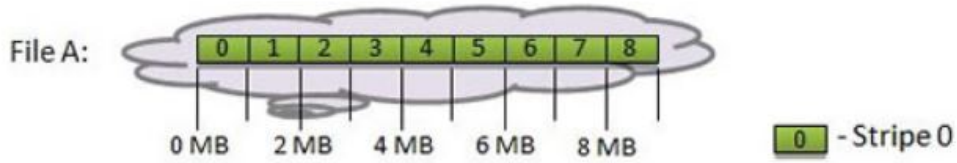
Dateisystem

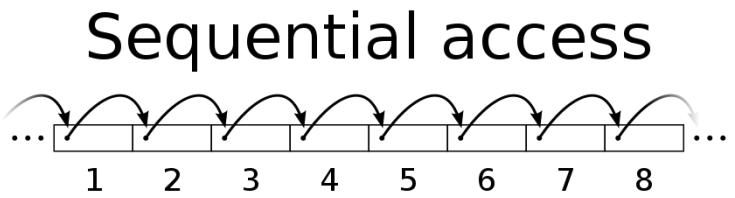


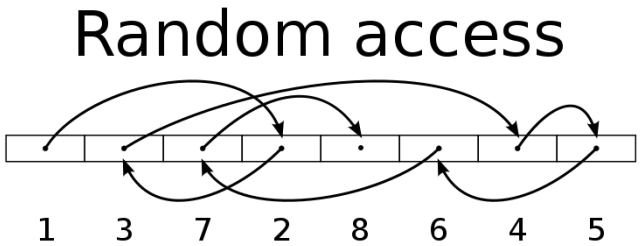
I/O

- Zugriffe
  - Schreiben, überschreiben, lesen
  - Muster erkennen
- Striping
- Small random access vs. große zusammenhängende Blöcke







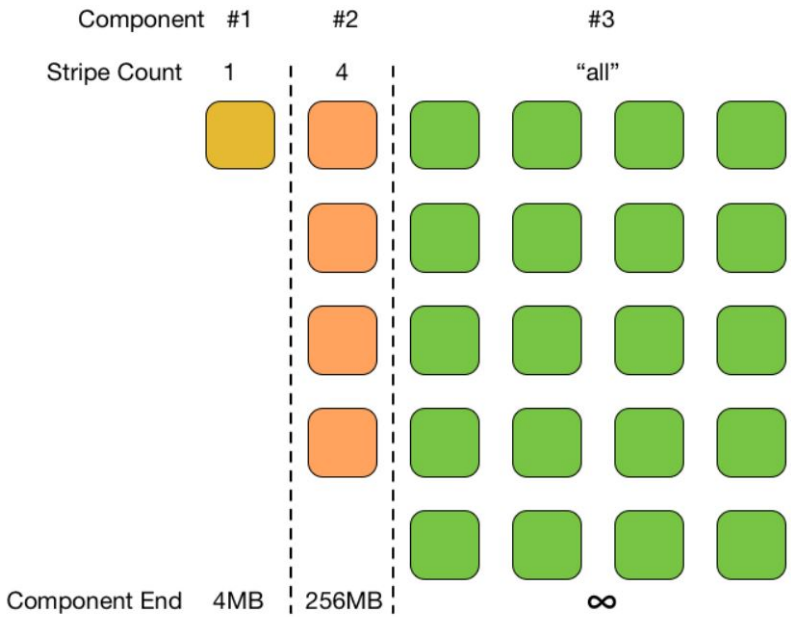


Herausforderungen

- Skalierung
  - Hardware ausnutzen
- Optimierung
  - Compute/memory/x bound application
  - Compiler
- Automatisierung
  - Händischer Workflow nicht mehr machbar
- Analyse
  - Parallelität bringt eine Extraklasse an Fehlern
  - Neue Komplexität an Zusammenhängen
- Hardwareunterstützung
  - NVMe SSD, 200Gb Netzwerk, 1Mio.-Kerne-System
  - Topologie
- Energieeffizienz
- Machine Learning

Thema: Progressive File Layouts (PFL) auf Levante

- Bewertung der Performance von PFL auf Levante evaluieren
- Unterschiedliche Lesezugriffe landen auf unterschiedlichen Komponenten
- Pro Komponenten sind verschiedene Hardwareeinheiten
- Wie, wann und ob sich das auf die I/O-Laufzeiten auswirkt



Thema: ISC SCC 2025 – online Wettbewerb

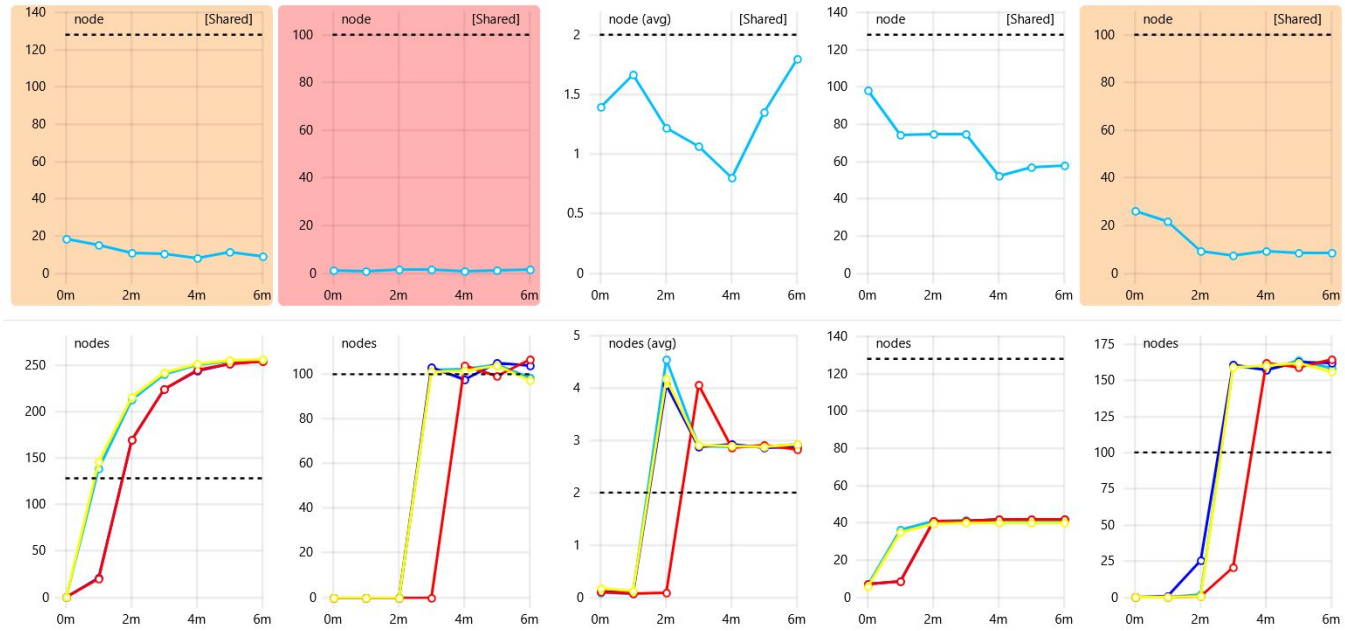
- Studentischer Wettbewertf auf der International Supercomputing Ceonference 2025 in Hamburg
- Kooperation mit Universitäten Magdeburg und Dresden
- Wissenschaftliche Anwendungen auf gegebenen Systemen bauen und laufen lassen
- Stromlimits beachten
- Compilen, Benchmarken, Optimieren, Testen

Thema: Lustre Monitoring Tools evaluieren

- OpenSource Parallels Dateisystem, auf WR-Cluster und Levante
- I/O Zugriffsmuster von Anwendung != Client pattern != Server pattern
- Viele OpenSource Frameworks im Web, die bereits was auslesen können
- Durchprobieren, vergleichen, analysieren, benchmarken
- Zusammenführen und für ClusterCockpit auf Levante aufbereiten

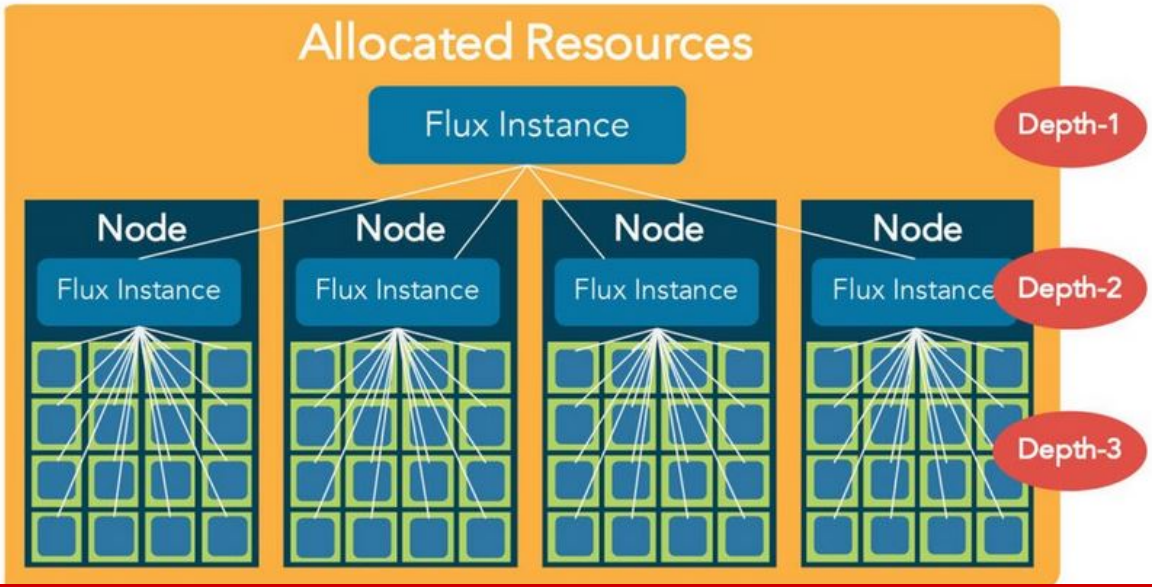
Thema: Anomaly detection in ClusterCockpit monitoring

- Monitoring-Framework auf Levante
- Einige Jobs laufen offensichtlich ineffizient, einige Werte lassen es vermuten
- Wie kann man die Jobs automatisiert erkennen, labeln und informieren?
- Im nächsten Schritt ggf. Maßnahmen ergreifen
- Gruppieren und sortieren



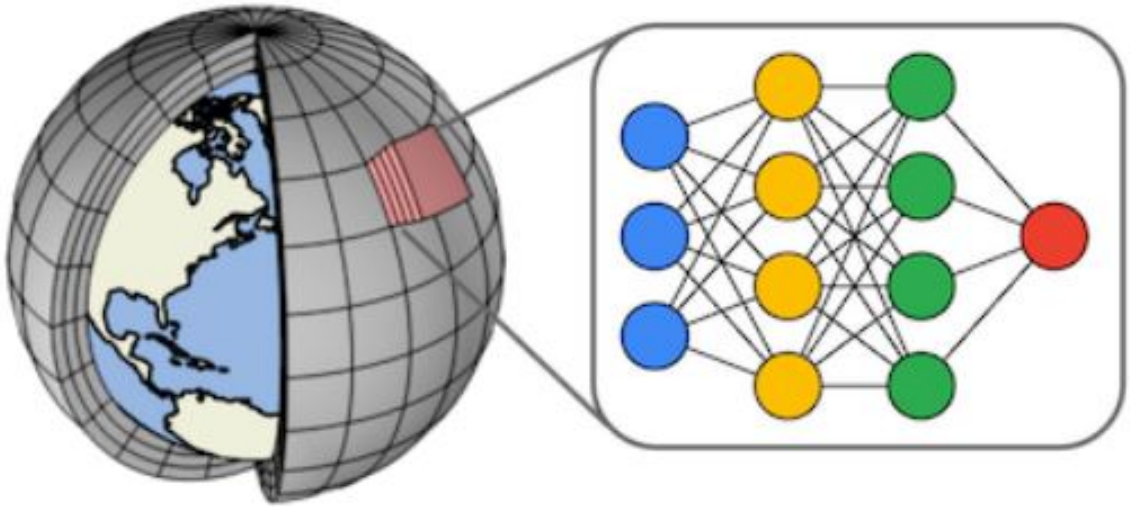
Thema: I/O aware Flux

- De Facto Standard Scheduler ist Slurm
- Kann keinerlei Verhalten zu I/O oder Storage-Ressourcen berücksichtigen
- Plugins brauchen zwingend root Rechte
- Flux soll flexibler sein
- Mit Slurm vergleichen und eine Erweiterung einbauen
  - Bei angenommen bekanntem IO Verhalten von Anwendungen dieses berücksichtigen



Thema: NeuralGCM benchmarken

- Bauen und laufen lassen, Ressourcenbedarf ermitteln
- Metriken einsammeln
- IO-Bedarf erörtern und evaluieren
- Unterschiede zu pytorch erarbeiten



NeuralGCM

Thema: Vergleich von Energieverbrauchsmessungen

- System Counter vs. externe Messung
- System Counter von Intel, AMD und ARM als gängige interne Energiemessmethoden
- Externe Messungen über PDUs liefern präzise Messdaten
- Versuche zeigen teils erhebliche Abweichungen zwischen internen und externen Messungen
- Einsatz automatisierter Datenextraktion aus PDUs
- Python-basierte Datenanalyse und Auswertung

Thema: IO in WRF benchmarken

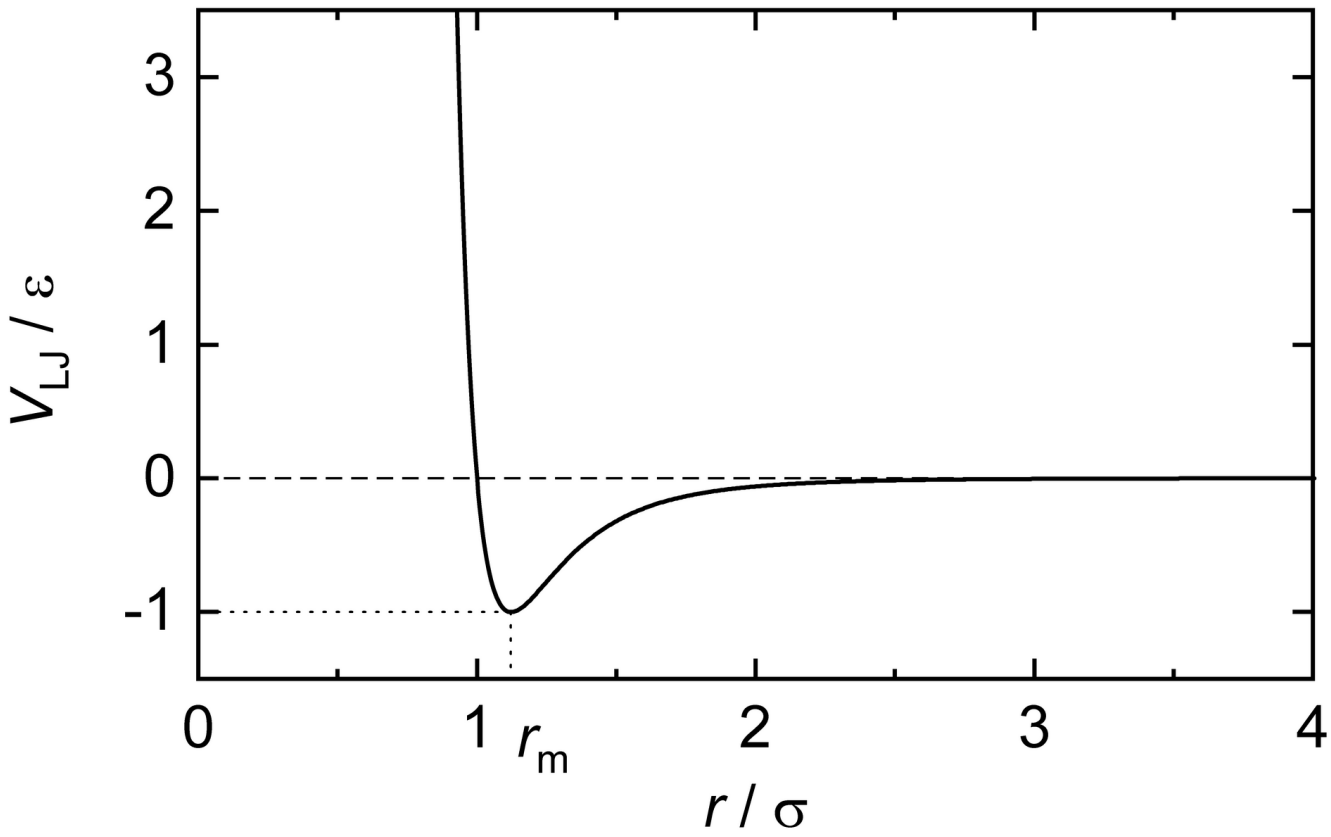
- 2 Paper mit Grundlagen
- Bauen, Installieren, Experimente laufen lassen
- Verschiedene Backends nutzen
- Evaluieren auf verschiedenen Systemen (CPU/GPU/HDD/SSD)
- Parameter in Bezug zur Leistung setzen
- Setups finden, die man CPU, Netzwerk oder IO-Bound hinbekommt

Thema: FS OnDemand

- Einige Anwendung lesen und schreiben viele kleine Daten/Dateien
- Das sorgt für hohe IOPs, viel Netzwerktraffic, unoptimierte Zugriffe
- Eine lokale schnelle Flash-Platte könnte diese abfangen und erstmal sammeln
- Dann wird in Wellen synchronisiert
- BeeOND ?
- Aufsetzen, ausprobieren, in Kombination mit Lustre testen

Thema: Teilchensimulation mit Lennard-Jones-Potential

- Teilchen-Simulation
- Dynamische Lastverteilung, Cutoff-Radius, Erhaltungsgrößen



Thema: Cloud Survey

- Übersicht verfügbarer Cloud-Anbieter
- Übersicht angebotener Dienste
  - Fokus auf Hardware
  - Fokus auf Buchungsmodalitäten
- Erarbeitung einer Kostenfunktion
  - Buchungskosten
  - Varianten (z.B. spot)
  - Berücksichtigung der Verfügbarkeit (wie viele "9er"?)

Thema: (IO) pattern recognition

- "Was tut mein Programm?"
- Statische Code Analyse?
- Infromationsbeschaffung
  - Sampling
  - Non-invasives tracing
  - Invasives tracing
- Erkennung von Mustern

Thema: Anwendung von KI zur Auswertung der U.S. SEC Enforcement- und Litigation-Datenbank

- Durchsetzung des Bundeswertpapiergesetze
- Mustererkennung auf DB
  - Art der Verstöße
  - Trends
  - Maßnahmen und deren Wirkung bzw Veränderungen
  - Umfang der Maßnahmen

Thema: Data LifeCycle tracking

- Dateien aus dem warmen Speicher (Lustre auf HDD/SSD) werden manchmal ins Archiv migriert
- Dann wieder zurpckgelesen
- Unterschiedliche Systeme mit unterschiedlichen Identifizierungsmechanismen für ein logisches Datenobjekt
- Deren Lebensweg nachverfolgen ist sehr wichtig
- DB-Anwendung bauen, die das verfolgt und erfasst
- Skalierbar (wie viele Dateien haben wir so?)

Thema: Retraction Watch Database

- Automatisierte Datenbeschaffung
- Automatisierte Analyse auf DB
- Mustererkennung
  - Autoren, Institut, Verbindungen
  - Auffälligkeiten im Review-Prozess
    - Journal, Editoren
    - Geschwindigkeit bzgl Datum

Thema: PyDarshan

- Erweiterung der Datenerfassung
  - Kompression
  - ?
- Erweiterung des PDF-Reports
- Kombination mit otf2 Traces erkunden

Thema: OpenMP Offloading

- Paper Johannes Doerfert
- OpenMP Offloading vs OpenACC
- Offloading auf GPU
- Offloading auf CPU
- OpenMP Offloading vs MPI

Thema: Datenreduktion

- Deduplikation
- Kompression
- Einteilung in Äquivalenzklassen
  - Stellvertreter-Dateien