CLAIX – Vorstellung und Technik

Christian Terboven

IT Center | RWTH AACHEN UNIVERSITY

Inhalte

- CLAIX – Phase I
  - Zwei Rack-Reihen + 2 Schränke
- Testbetrieb mit Projekten seit November 2016
  - Trier-2 HPC-System
- Abnahme im Januar 2017
- TOP500 im November 2016:
  - 506.87 TFLOPS (528 TFLOPS Peak)
  - Platz 309 in TOP500, Platz 55 in Green500: 96% Effizienz
- 15000+ Intel Xeon Cores, 80+ Main Memory
- 100 Gbit/s Intel Omni-Path HPC Netzwerk, 3.3 PB Lustre Storage

\Orchestrating a brighter world

NEC

wassergekühlt
(geschlossen)

luftgekühlt
(offen)

IT Center | RWTH AACHEN UNIVERSITY

Technische Daten: MPI Knoten

- 609 Standard MPI Knoten
  - 2x Intel Xeon Broadwell-EP (E5-2650v4): je 12 Cores, 2.2 GHz
    - Turbo: bis zu 2.9 GHz (2.5 GHz für AVX-Instruktionen)
  - 128 GB DDR4-2400 Main Memory
  - Intel Omni-Path x16 Netzwerkkarte
- 4 Server je 2 Höheneinheiten
- Knoten sind luftgekühlt in wassergekühlten, geschlossenen Schränken

Intel Xeon Broadwell

- Produktname: E5-v4
  - Intel: the latest and greatest beast since the previous one ;-)
  - 14 nm Produktionstechnologie, Nachfolger des Haswell

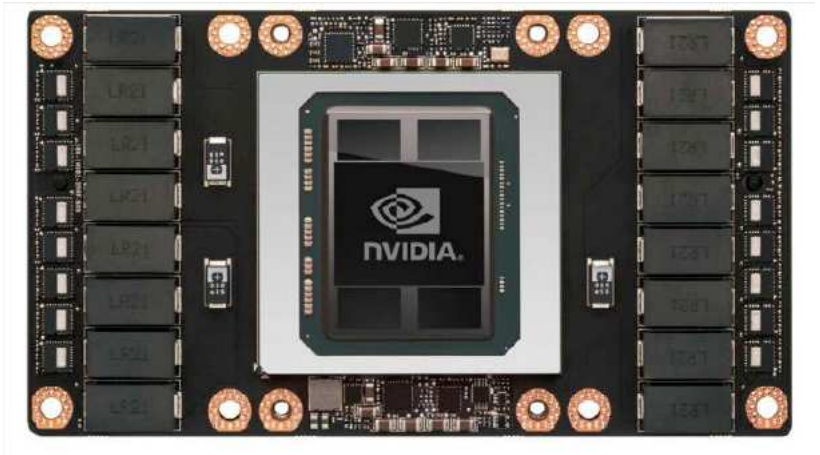
- Unsere Auswahl:
  - “Advanced” Segment: volle Speicherbandbreite und schnelle, große Caches
  - DDR4-2400 Speicher, 9.6 GT/s QPI
  - Modell mit 12 Cores: Optimierung der Preis-Leistung im Jobmix

- Spezielle Funktionen:
  - Leistungsfähige Vektoreinheiten, hohe Speicherbandbreite (ca. 60 GB/s)
  - Funktionierendes Transactional Memory
  - Schulungsangebote, z.B. http://www.itc.rwth-aachen.de/aixcelerate

Technische Daten: GPU Knoten

- 10 GPU Knoten, geplante Erweiterung auf 16
  - Technische Daten wie Standard MPI Knoten, plus:
  - 2x NVIDIA P100 (Pascal Architektur)

- Verbunden via NVLink



Technische Daten: SMP Knoten

8 Standard MPI Knoten
- 8x Intel Xeon Broadwell-EX (E7-8860v4): je 18 Cores, 2.2 GHz
- 1 TB Main Memory
- 2x 2 TB NVMe SSD
  (siehe auch NVMe Server)
- Intel Omni-Path x16 Netzwerkkart

2 Systeme mit 1x NVIDIA P100
- PCIe Karte

- Je Server: 8 CPUs in
  7 Höheneinheiten

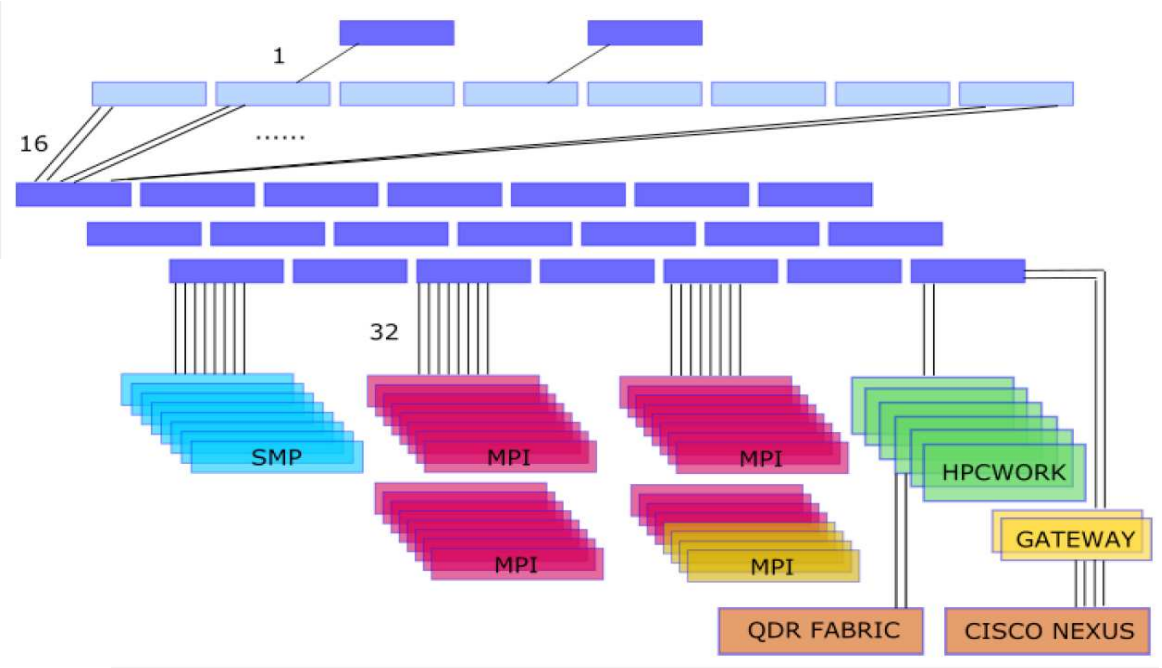
- Knoten sind luftgekühlt

Technische Daten: NVMe Knoten

- 8 NVMe Knoten
  - Technische Daten wie Standard MPI Knoten, plus:
  - 2 TB NVMe SSD
- Derzeit: Evaluation von Betriebsmodi
  - Block-Dateisystem
  - Ad-hoc (Block-)Dateisystem
  - Mapped Memory
  - Allocated Memory
- Evaluation von weiteren Anwendungsszenarien
  - In-situ Analyse
  - I/O-Nodes

HPC Netzwerk

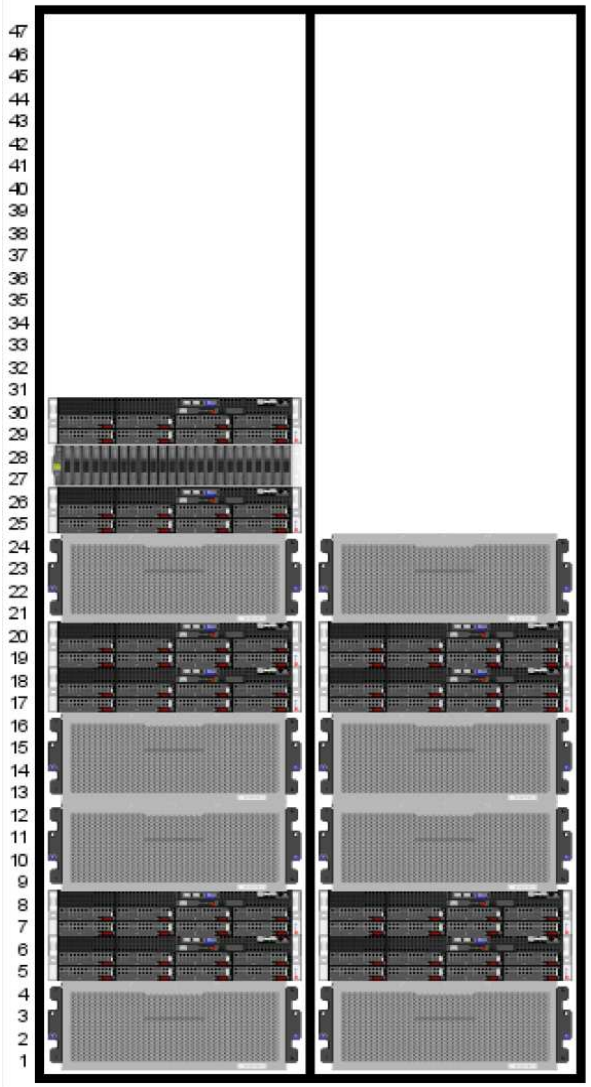
- Intel Omni-Path Netzwerk
  - 1:2 Blocking
  - x16 PCIe Karten (100 Gbit/s)
- 2 Omni-Path zu Eth Gateways
  - 2x 40 Gbit/s pro Gateway



- MPI Latenz (Knoten zu Knoten): 0.93 usec
- MPI Bandbreite (Knoten zu Knoten): 93 Gbit/s

Lustre Storage

- Mount: $HPCWORK
- Kapazität: 3300 TB
- Leistung (Bandbreite)
  - 55+ GB/s schreibend
  - 50+ GB/s lesend
- Erste Installation von LXFS mit ZFS in Europa
  - Ziel: End-to-end Integrity
- Besonderheit: Hadoop-Adapter für Lustre



Ausblick I

- Big Data:
  - Lustre Storage unterstützt prinzipiell Hadoop und ähnliche Paradigmen
  - CentOS 7.x (Linux Betriebssystem) unterstützt prinzipiell Container
  - NVMe-Technologie verfügbar in einigen Knoten sowie über Intel OPA
  - Vision: Integration von “Big Data”-Workloads in das HPC-System
- Evaluation von Architekturen für 2018
  - Anschaffung eines Teilclusters mit Intel Xeon Phi Architektur (Intel KNL)
    - Many-Core Technologie: 240+ Cores je Prozessor
  - Kooperation mit NEC: Evaluation der Vektorarchitektur AURORA
  - Kooperation mit FZ Jülich: Evaluation von ARM- und POWER-basierten Systemen

Ausblick II

- Jobgenaues Energiemonitoring und -controlling
  - Detaillierte Erfassung und Auswertung der Energiedaten
  - Korrelation mit Informationen vom Batchsystem und Anwendungsanalyse
  - Implementierung der Möglichkeit des Power Capping
  - Ziel: Weitere Reduktion der Betriebskosten
- Bis 2018:
  - Ausbau des Angebots des Integrativen Hosting
  - Ausbau der Tier-3 Angebote für die RWTH Aachen
  - Beschaffung und Inbetriebnahme von CLAIX – Phase 2
- 2018: Realisierung einer freien Kühlung

Vielen Dank
für Ihre Aufmerksamkeit

Christian Terboven

E-Mail: Terboven@itc.rwth-aachen.de

HPC Rechenzeitprojekt

- Tier-2 System: Zugang für Projekte nach Antrag und Begutachtung
- Anforderungen an Antrag und Begutachtung:
  - < 2000 core-h pro Monat: persönliches Kontingent jedes MA
  - < 20000 core-h pro Monat: technische Begutachtung am IT Center
    - Projekte aus Forschung und Lehre, Abschlussarbeiten, etc.
  - < 1.2 Mio core-h pro Jahr: technische und wissenschaftliche Begutachtung am IT Center und in der RWTH Aachen
  - > 1.2 Mio core-h pro Jahr: Projektanträge in de JARA-HPC Partition
  - > 10 Mio. core-h pro Jahr: Empfehlung auf ein Tier-1 Center auszuweichen
- CLAIX:
  - 80 % der Rechenzeit stehen in JARA zur Verfügung
  - Bis zu 20 % der Rechenzeit stehen zukünftig bundesweit zur Verfügung

HPC Rechenzeitprojekt

Bis auf Weiteres noch auf dem Bull-Cluster

- Freie Nutzung: Wissenschaftler max 0,024 Mio Coreh/Jahr
- Freie Nutzung: Studenten max 0,006 Mio Coreh/Jahr
- Kleine Projekte: max 0,24 Mio Coreh/Jahr
  kurze Projektbeschreibung, technische Begutachtung
- Mittlere Projekte: max 1,2 Mio Coreh/Jahr
  Projektbeschreibung, techische und wissenschaftliche Begutachtung

Claix

- Große Projekte: über 1,2 Mio Coreh/Jahr
  Projektanträge zweimal jährlich (nächste Abgabefrist: Ende Februar)
  ausführliche technische und wissenschaftliche Begutachtung
- Geplant: 20% der Ressourcen für externe Projekte
  technische und wissenschaftliche Begutachtung
- “Schnupperkontingente“ für technische Evaluierungen