



bw | HPC – C5

Benchmark-Anforderungen im Beschaffungsprozess der bwForCluster

Bernd Wiebelt, Rechenzentrum, Universität Freiburg
Jürgen Salk, kiz, Universität Ulm



UNIVERSITÄT
HEIDELBERG
ZUKUNFT
SEIT 1386

Hochschule
für Technik
Stuttgart



Hochschule Esslingen
University of Applied Sciences

Universität
Konstanz



UNIVERSITÄT
MANNHEIM



Universität Stuttgart

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN



KIT
Karlsruher Institut für Technologie



ulm university universität
uulm

Funding:



Baden-Württemberg

MINISTERIUM FÜR WISSENSCHAFT, FORSCHUNG UND KUNST

www.bwhpc-c5.de

bwHPC Leistungsebenen

Europäische Höchstleistungsrechenzentren
(Tier 0) Gauss Center for Supercomputing



Nationale Höchstleistungsrechenzentren
(Tier 1) HLRS@GCS

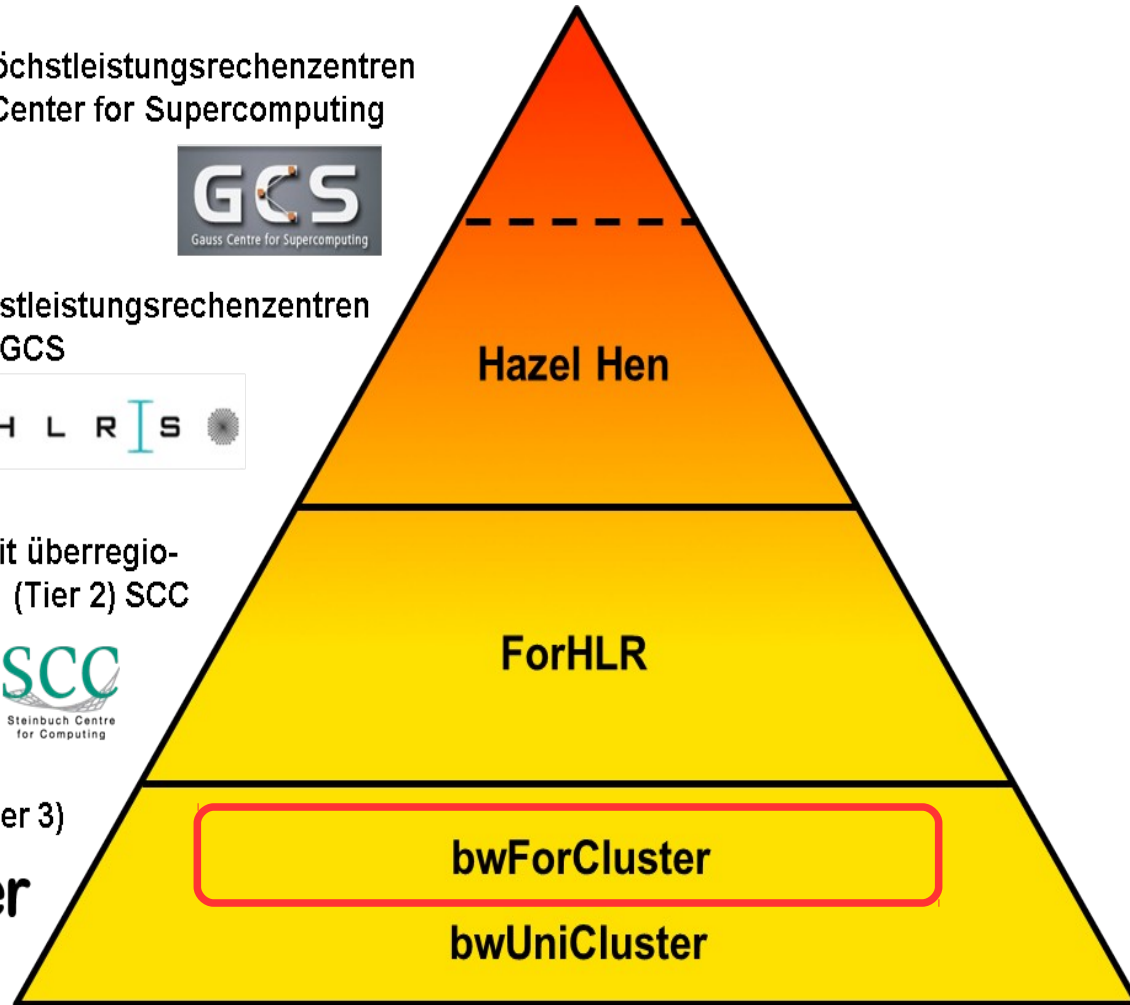


HPC-Zentren mit überregionalen Aufgaben (Tier 2) SCC

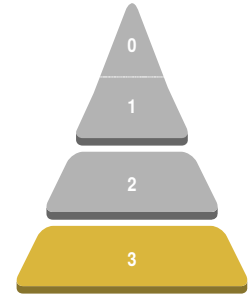


HPC Cluster (Tier 3)

bwCluster



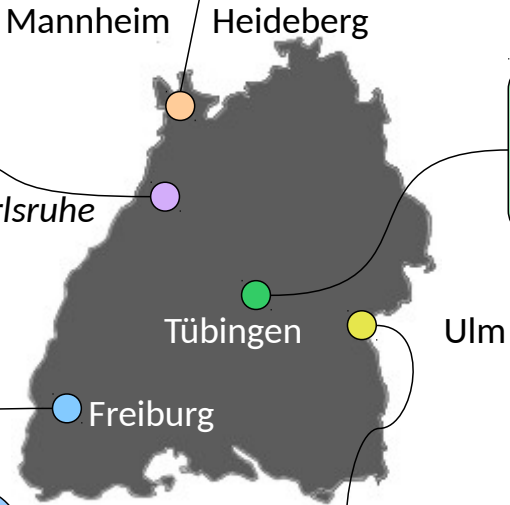
bwCluster @ Tier 3



bwForCluster MLS&WISO (10/2015):
Wirtschafts- und Sozialwissenschaften,
Molekulare Lebenswissenschaften

bwUniCluster (02/2014):
Allgemeine Versorgung,
Lehre und Training

bwForCluster BinAC (2016):
Bioinformatik,
Astrophysik

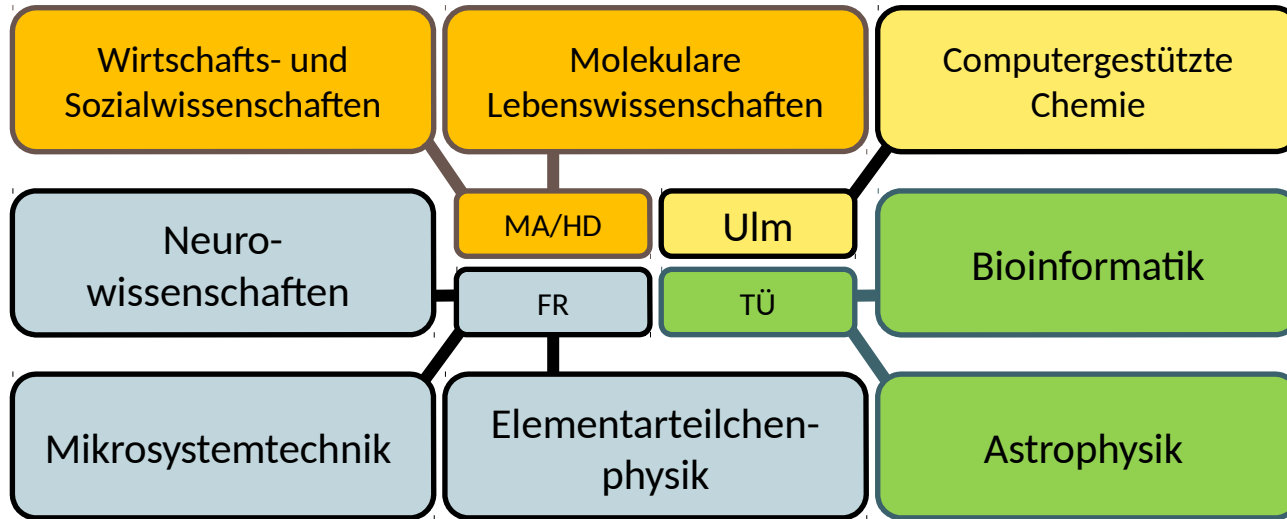


bwForCluster NEMO (Q2/2016):
Neurowissenschaften,
Elementarteilchenphysik,
Mikrosystemtechnik

bwForCluster JUSTUS (12/2014):
Computergestützte Chemie



bwForCluster

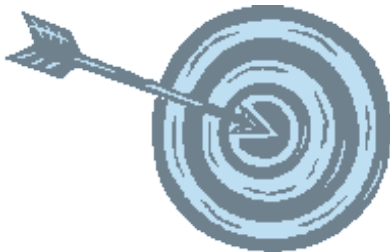


- bwForCluster sind auf spezifische Wissenschaftsbereiche zugeschnitten
- Beschaffung von Hardware und Software in Abstimmung mit den designierten Benutzern
- Begleitprojekt bwHPC-C5 (Kompetenzzentren)

Design bwForCluster

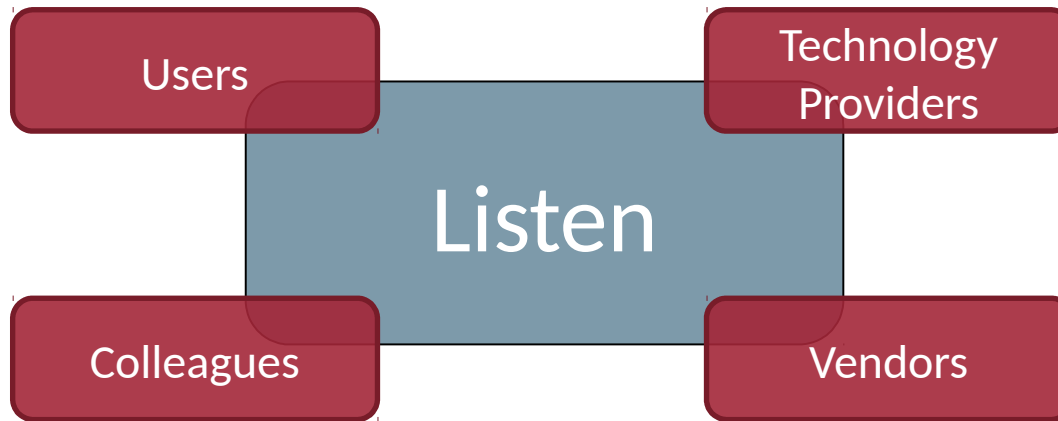
■ Generelle Zielsetzung:

- Schaffung von Forschungsclustern, die speziell auf die Anforderungen der Nutzer im Bereich der jeweiligen Fachrichtungen zugeschnitten ist.
- Anpassung der Kapazität und Architektur an die Zielgruppen aus 9 Universitäten Baden-Württembergs
- Volle Integration in bwHPC-Konzept



Design bwForCluster

- Generelle Frage am Anfang: Wie findet man das “richtige” System?



Co-Design

This process believes that by **encouraging the trained designer and the user to create solutions together**, the final result will be more appropriate and acceptable to the user.

Source: Wikipedia (<http://en.wikipedia.org/wiki/Co-design>) and Albinsson, L., M. Lind, et al. (2007). Co-Design: An approach to border crossing, Network Innovation. eChallenges 2007, The Hague, The Netherlands

Benchmarks als Designmerkmal

- **Wesentliches** Bewertungsmerkmal bei Beschaffung
 - Beschreibt letztlich die “Leistungsfähigkeit” des Gesamtsystems
- Aber viele Optimierungsvektoren:
 - CPU-Architektur
 - Anzahl Nodes, Anzahl Cores pro Node, Taktfrequenz
 - Memory-Bandbreite, Cachegröße
 - I/O
 - Spezial-Hardware
 - ...
- Leistet das System, was sich die **Betreiber** davon versprechen?
- Leistet das System, was sich die **Benutzer** davon versprechen?

Welche Benchmarks werden ausgewählt?
Wie werden sie gewichtet?

Standard-Benchmarks

- Robust und umfangreich getestet
- Single-Component
 - Beispiel: Stream (Arbeitsspeicher), IOR (Filesystem)
 - Gut geeignet, um minimale Qualitätsstandards zu definieren
 - Bei **korrektem** Einsatz eindeutige Qualitätsaussagen
- Multi-Component
 - Beispiel: Spec, HPL
 - Abschätzung der Gesamtsystemleistung in speziellen Szenarien
 - Bei **korrektem** Einsatz eindeutige quantitative Aussagen

Frage ist nicht ob, sondern welche
Standard-Benchmarks genutzt werden



Applikations-Benchmarks

- In Kooperation mit Benutzern
 - Umfragen, fachspezifische Konsultationen, Benchmark-Workshops
 - Auswahl repräsentativer Applikationen
 - Definition geeigneter Testläufe
- Bei **korrektem Einsatz** hohe Aussagekraft für Wert des Gesamtsystems
- Probleme/Fragen:
 - Wie kommt man an geeignete Benchmark-Daten?
 - Was wird eigentlich getestet?
 - Wie robust ist der Benchmark im Eigenbau gegenüber verschiedenen Hardwarekonfigurationen (z.B. Scale-Up)?
 - Könnte der Benchmark durch einen Standard-Benchmark ersetzt werden?



Erster grober Design-Ansatz bwForCluster JUSTUS

- 3 Knoten-Typen für verschiedene Anwendungsklassen
- Grobe Verteilung der Typen: 100:100:10

Node Type 1:

- Very high I/O performance
- Applications with very low scalability
- Problem sizes demand 128GB RAM
- Turbomole, Gaussian, Molpro, Schrödinger-Jaguar, Orca, Cfour and Gamess.

Node Type 3:

- Very high I/O performance
- Single node
- Problem sizes demand 512GB RAM
- Same App. mix as Seg. 1

Node Type 2:

- I/O performance less relevant
- Applications with low to medium scalability
- Problem sizes demand 128GB RAM
- VASP, ADF, AIMS, DACAPO, ABINIT, QuantumEspresso, CPMD, Amber, Gromacs




Benchmarks bwForCluster JUSTUS

■ Applikations-Benchmarks:

- Auswahl **repräsentativer Applikationen** für jeden Knotentyp:
 - Typ 1/3: Molpro
 - Typ 2: VASP
- Herausforderung: Definition von Test-Cases mit **annehmbaren Laufzeiten für Benchmarks** (sehr viel **kürzer als normale Jobs**), die aber trotzdem die Vor- und Nachteile verschiedener Lösungen aufdecken.
- Erfordert hohes Maß an Expertenwissen zu den ausgewählten Applikationen
- Beispielwerte aus Molpro Benchmarks:

Bench	128GB + 2*SATA	128GB + 4*SSD
DFT j1c8	300s	301s
CCF12 j1c8	6000s	5618s
LCCSD j1c8	9000s	4992s

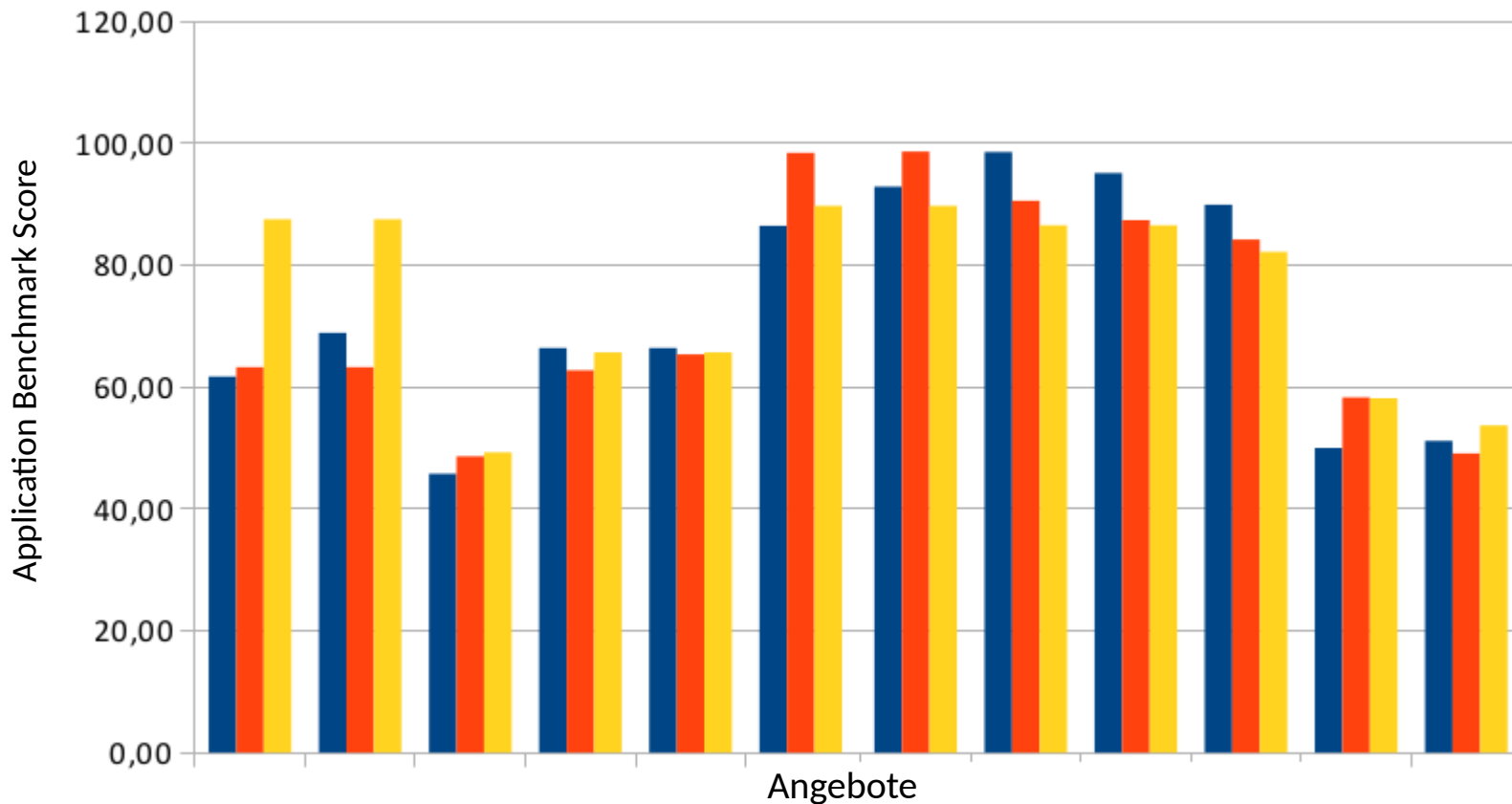
Große Laufzeitunterschiede!
Bei mehreren Jobs pro Knoten
(=Emulation größerer Jobs) ist der
Effekt noch ausgeprägter.



Benchmarks bwForCluster JUSTUS

■ Applikations-Benchmarks:

- Bei der Bewertung der Benchmarks wird Performance des Gesamtsystems beurteilt
- Es zählt nicht der einzelne Benchmark, sondern Jobs pro Zeiteinheit (Throughput)
- Knotenzahl spielt eine Rolle



Design-Ansatz bwForCluster NEMO

- Neurowissenschaft
 - Simulation von biologischen neuronalen Netzwerken (Standard-Software NEST)
 - Compute-Inseln (500-800 Cores) mit non-blocking Hochleistungsnetzwerk
 - Datenanalyse mit Matlab oder Python
 - Viel Speicher (mind. 128 GB) oder viel Durchsatz (Cores)
- Elementarteilchenphysik
 - Einsatz des Software-Stack des CERN (cvmfs)
 - Viele Cores, mindestens 4 GB pro Core
 - Mathematica (Theoretische Physik)
 - Viel Speicher (mind. 128 GB)
 - Virtuelle Forschungsumgebungen
 - Möglichst geringer Verlust beim Einsatz von virtuellen Maschinen
- Mikrosystemtechnik
 - Viele Cores, Hochleistungsnetzwerk (MPI)



Benchmarks bwForCluster NEMO

- Ausschreibung: Uniformer Node-Typ mit 128 GB Speicher
- Applikationsbenchmarks Neurowissenschaft:
 - NEST (CPU + Memory + Hochleistungsnetzwerk)
 - ~~Beitrag (Matlab Code) eliminiert, im Wesentlichen CPU Benchmark~~
- Applikationsbenchmarks Elementarteilchenphysik
 - HEP-SPEC06 (CPU)
- Applikationsbenchmarks Mikrosystemtechnik
 - ~~Beitrag (C Quellcode) eliminiert, im Wesentlichen CPU Benchmark~~

Außer NEST und HEP-SPEC06 keine
Applikationsbenchmarks



Benchmark-Ausführungsregeln

- Hardware benchmarken, nicht die Skills der Mitarbeiter des Anbieters
 - Praxisnahe Compiler und Compile-Optionen vorgeben
 - Programmaufrufe vorgeben (z.B. ohne CPU-Pinning)
 - Exakte Versionsvorgaben machen oder Quellcode als Download bereitstellen
- Ergebnisse vergleichbar machen, Features abstellen
 - Simultaneous Multithreading
 - Turbo Mode
 - Cluster-on-Die
- Ausführung auf der realen Hardware favorisieren
 - Wer extrapoliert, muss spätestens bei der Abnahme des Systems auch die versprochenen Ergebnisse liefern



Alle Benchmarks (NEMO)

- Standard-Benchmarks
 - HPL (Flops)
 - Stream (Arbeitsspeicher)
 - IOR (paralleler Storage)
- Applikations-Benchmarks
 - NEST (Hochleistungsnetzwerk, CPU, Memory)
 - HEP-SPEC06 (CPU)

Strategie:
Wenige, robuste Benchmarks



Weitere Benchmarks bwForCluster JUSTUS

■ IO Benchmarks:

■ Unterschiedliche Benchmarks für verschiedene Storage-Bereiche des Clusters:

- Lokaler Scratch
- Home Filesystem
- Globaler Scratch (paralleles Filesystem)

■ Herausforderung bei allen IO-Benchmarks: **Page-Caching des Kernels beachten**

- Erfordert hinreichend große Datensätze und/oder spezielle Flags bei Benchmark-Tools
- Ausschnitt aus IO Benchmark-Guide für JUSTUS:

“The sum of all working file sizes on each compute node [...] must be greater than 2 times the available memory on the individual compute nodes involved in the test.”



Benchmarks bwForCluster JUSTUS

■ Verwendete IO Benchmark Tools:

■ FIO (<http://freecode.com/projects/fio>)

- “fio is an I/O tool meant to be used both for benchmark and stress/hardware verification.”

■ Auszug aus IO Benchmarking Guide für JUSTUS:

*“The storage system will be subjected to different file access patterns, including pure sequential and random read/write operations **as well as mixed workloads based on IO patterns expected for quantum chemistry applications.**”*

■ IOR (<https://sourceforge.net/projects/ior-sio>)

- “The IOR software is used for benchmarking parallel file systems using POSIX, MPIIO, or HDF5 interfaces.”
- Single-Node und Multi-Node Benchmarks gefordert (nur POSIX IO)

■ MDTEST (<https://sourceforge.net/projects/mdtest>)

- “mdtest is an MPI-coordinated metadata benchmark test that performs open/stat/close operations on files and directories and then reports the performance.”
- Single-Node und Multi-Node Benchmarks gefordert



Benchmarks bwForCluster JUSTUS

■ IO Benchmark Matrix:

Cluster segment	Storage Entity	fio	IOR	mdtest
Segment 1	local	X		X
	home		X	X
	global		X	X
Segment 2	local			
	home		X	X
	global		X	X
Segment 3	local	X		X
	home		X	X
	global		X	X

Benchmarks bwForCluster JUSTUS

- Weitere verwendete Benchmarks: Memory Bandwidth und Messaging
 - **STREAM** (<https://www.cs.virginia.edu/stream>)
 - “The STREAM benchmark is a simple synthetic benchmark program that measures sustainable memory bandwidth (in MB/s) and the corresponding computation rate for simple vector kernels.”
 - Nicht verwendet, aber evtl. gute Alternative zu STREAM: Intel MLC Benchmark
 - <https://software.intel.com/en-us/articles/intelr-memory-latency-checker>
 - **Intel IMB** (<http://software.intel.com/en-us/articles/intel-mpi-benchmarks>)



Benchmarks bwForCluster JUSTUS

■ Was ist rausgekommen?



bwHPC@Ulm

In Produktionsbetrieb seit Dez. 2014

Benchmarks bwForCluster JUSTUS

■ Was ist rausgekommen?

Segment 1: 204 nodes

- Dual Socket E5-2630v3 2.4 GHz
- NEC "Green Gem" Platform
- 128 GB DDR4-RAM
- 4x240 GB SSD (RAID0)
- 64 GB (max.) Ramdisk

Segment 3: 38 nodes

- Dual Socket E5-2630v3 2.4 GHz
- NEC "Green Gem" Platform
- 256 - 512 GB DDR4-RAM
- 4x480 GB SSD (RAID0)
- 128 - 256 GB (max.) Ramdisk

Segment 2: 202 nodes

- Dual Socket E5-2630v3 2.4 GHz
- NEC "Green Gem" Platform
- 128 GB DDR4-RAM
- Diskless
- 64 GB (max.) Ramdisk

Visualization: 2 nodes

- Dual Socket E5-2630v3 2.4 GHz
- NEC "Green Gem" Platform
- Nvidia K6000 Graphic Card
- VirtualGL based remote vis

Benchmarks bwForCluster NEMO

■ Was ist rausgekommen?



- ...besuchen Sie den bwHPC-C5 Stand auf der ISC 2016

Danke für Ihre Aufmerksamkeit!

