JUWELS BOOSTER EARLY EXPERIENCES

CERN COMPUTE ACCELERATOR FORUM

9 June 2021 | Andreas Herten | Jülich Supercomputing Centre, Forschungszentrum Jülich

JÜLICH Forschungszentrum | JÜLICH SUPERCOMPUTING CENTRE

Overview

Jülich Supercomputing Centre

Forschungszentrum Jülich Germany, near Cologne, interdisciplinary research, 6400 employees

Jülich Supercomputing Centre

- Operation of supercomputers
- Education, training
- Application Support, Domain Science Support
- Research & Development

Accelerating Devices Lab Support, research, education for GPUs et al.; NVIDIA Application Lab at Jülich

Supercomputers

Production JUWELS, JURECA DC, JUSUF

Prototypes JUMAX, DEEP, …

Jülich Supercomputing Centre

Forschungszentrum Jülich Germany, near Cologne, interdisciplinary research, 6400 employees

Jülich Supercomputing Centre

- Operation of supercomputers
- Education, training
- Application Support, Domain Science Support
- Research & Development

Accelerating Devices Lab Support, research, education for GPUs et al.; NVIDIA Application Lab at Jülich

Supercomputers

Production **JUWELS**, JURECA DC, JUSUF

Prototypes JUMAX, DEEP, …

JUWELS Overall Architecture

JUWELS Cluster (2018)

- 2511 compute nodes (2× Skylake)
- 48 GPU nodes (4× V100 w/ NVLink2)
- Mellanox EDR 100 Gbit/s network, fat-tree topology (1:2@L1)
- 12 PFLOP/s

JÜLICH Forschungszentrum | JÜLICH SUPERCOMPUTING CENTRE

JUWELS Overall Architecture

JUWELS Cluster (2018)

- 2511 compute nodes (2× Skylake)
- 48 GPU nodes (4× V100 w/ NVLink2)
- Mellanox EDR 100 Gbit/s network, fat-tree topology (1:2@L1)
- 12 PFLOP/s

JUWELS Booster (2020)

- 936 compute nodes (2× AMD Rome, 4× A100 w/ NVLink3)
- Mellanox HDR 200 Gbit/s network, DragonFly+ topology
- 73 PFLOP/s

JUWELS Overall Architecture

JUWELS Cluster (2018)

Top500 Nov-2020:

- #1 Europe
- #7 World
- #1* Green500

JUWELS Booster (2020)

- 936 compute nodes (2× AMD Rome, 4× A100 w/ NVLink3)
- Mellanox HDR 200 Gbit/s network, DragonFly+ topology
- 73 PFLOP/s

JUWELS Booster Overview

Node Configuration

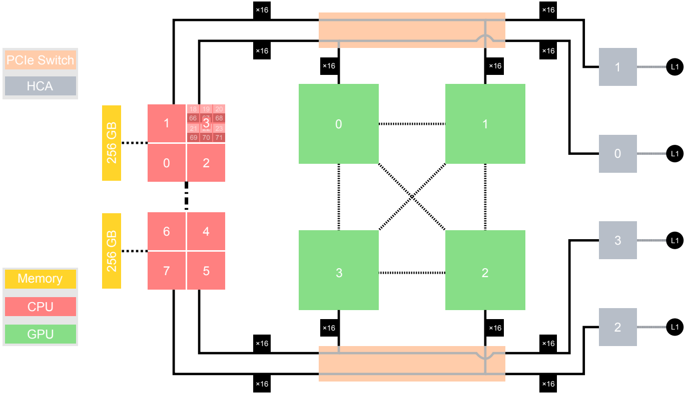
Arch Atos Bull Sequana XH2000

CPU 2 × AMD EPYC 7402:
$2_{\text{Socket}} \times 24_{\text{Core}} \times 2_{\text{SMT}}$,
2 × 256 GB DDR4-3200 RAM;
NPS-4

GPU 4 × NVIDIA A100 40 GB, NVLink3
73 PFLOP/s, 1.16 EFLOP/s_{FP16TC},
18.7 EOP/s_{BinTC}

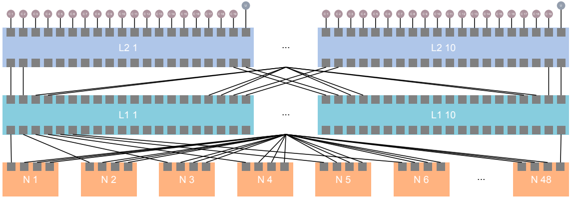
HCA 4 × Mellanox HDR200 (200 Gbit/s)
InfiniBand ConnectX 6

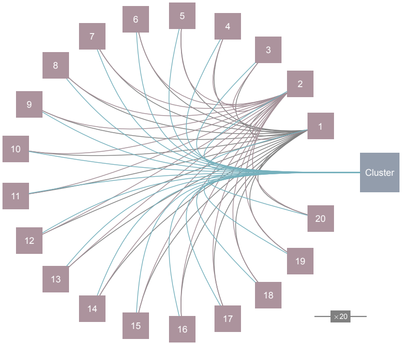
etc 2 × PCIe Gen 4 switch



JUWELS Booster Overview

Network Configuration: DragonFly+ Network





In-Cell (48 nodes): Full fat-tree in 2 levels

Inter-Cell (20 cells): 10 links between each pair of cells

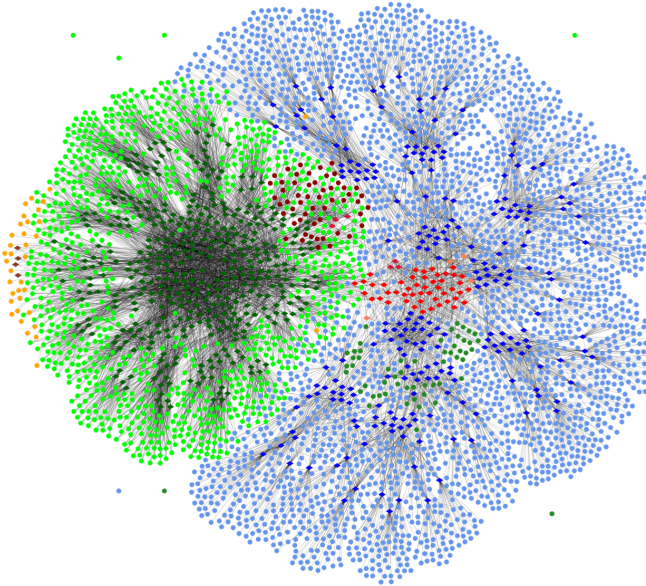
JUWELS

Cluster Booster Integration

Fully integrated system: **JUWELS** with Cluster and Booster modules

- File system: GPFS
- Network: InfiniBand
- Workload management: Slurm
- Resource management: ParaStation / ParaStation Slurm

Picture: Booster Cluster



JUWELS Software Stack

- Software
    - Software management: EasyBuild, LMod
    - Compilers: GCC, Intel, NVHPC
    - GPU-aware MPIs (ParaStationMPI, OpenMPI; via UCX)
  → https://apps.fz-juelich.de/jsc/llview/juwels_modules_booster/
- Operation
    - Operation System: CentOS 8
    - Provisioning: Ansible

Early Experiences

JUWELS Timeline

2018	JUWELS Cluster production start
2019	JUWELS Booster kick-off
2020 Apr	JUWELS Booster installation start
2020 May	JUWELS Booster Early Access Program first job
2020 Nov	JUWELS Booster production start, first compute-time period
2021 May	JUWELS Booster second compute-time period

JUWELS Timeline

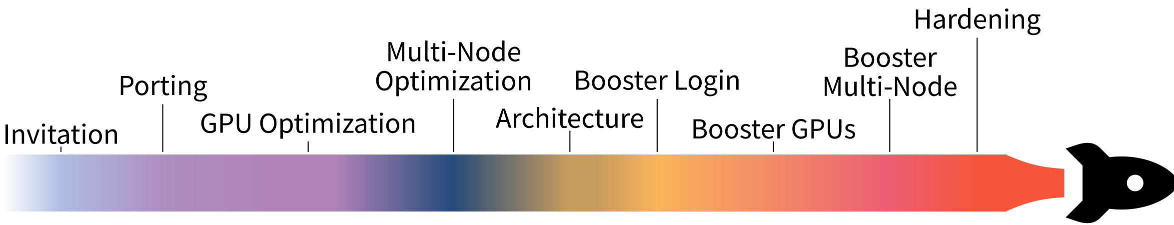
2018	JUWELS Cluster production start
2019	JUWELS Booster kick-off
2020 Apr	JUWELS Booster installation start
2020 May	JUWELS Booster Early Access Program first job
2020 Nov	JUWELS Booster production start, first compute-time period
2021 May	JUWELS Booster second compute-time period

Early Access Program

- Started in early 2020
- Invited 14 applications from various scientific domains
  - Aimed for applications that could use JUWELS Booster at scale
  - Some teams already use JUWELS Cluster, others new
- **Offer**: Use JUWELS Booster before general access; **Request**: Help improve system, compute-time allocation
- Endeavor of many parts in **JSC** and beyond
  - NVIDIA Application Lab: Steering, GPU optimization, application support, system support
  - Application support, Simulation Labs
  - Performance Optimisation and Productivity team
  - System operations team
  - Vendors: NVIDIA, ParTec, Atos

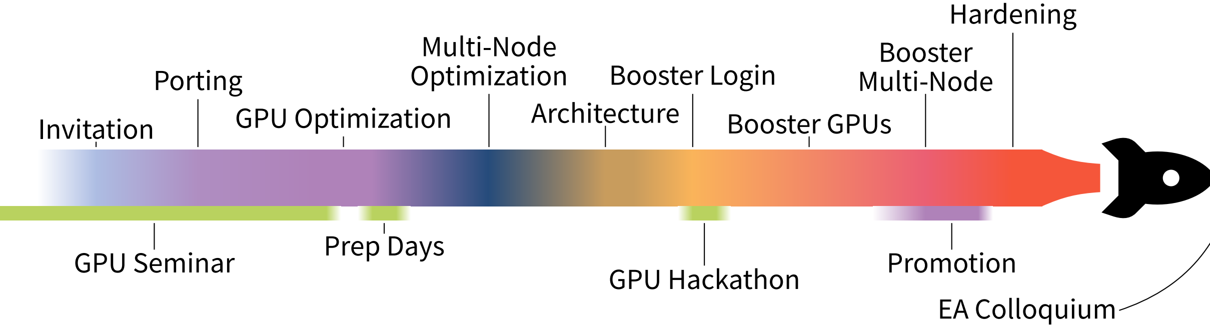
Timeline to Booster

- Preparation Timeline



Timeline to Booster

- Preparation Timeline
- Additionally: events



JÜLICH Forschungszentrum | JÜLICH SUPERCOMPUTING CENTRE

Applications I

Climate/Meteo/Hydro (*ESM*)

DeepACF High-resolution Weather Forecast Based on Deep Learning </> Lib:DL
JSC: Bing Gong, Michael Langguth, Amirpasha Mozaffari, Martin Schultz, Scarlet Stadtler

ICON Next-Generation Physical Weather and Climate Models </> OpenACC
MPI Met: Luis Kornblueh; NVIDIA: Dmitry Alexeev

MPTRAC Massive Parallel Trajectory Calculations of Volcanic Emissions </> OpenACC
JSC: Sabine Grießbach, Lars Hoffmann

ParFlow Surface, Soil, Ground Water Flow </> CUDA C
IBG-3: Jaro Hokkanen, Stefan Kollet

Biological Matter

Amber Drug Binding over Biologically Relevant Timescales (MD) </> Lib
JSC/HHU: Holger Gohlke, Christopher Pfleger, Michele Bonus

SOMA Kinetics of Nanomaterial Formation (Soft Matter) </> OpenACC
U Göttingen: Ludwig Schneider, Niklas Blagojevic

Applications II

PIConGPU Plasma Simulations for Next Generation Particle Accelerators (Plasma) </> CUDA C++
HZDR: Alexander Debus, Anton Lebedev, Rene Widera, Michael Bussmann

JUQCS-G Simulating Universal Quantum Computer (Quantum) </> CUDA Fortarn
JSC: Hans De Raedt, Kristel Michielsen, Dennis Willsch

E-train Understanding Learning Processes in Brain (Neuro) </> Lib:DL
U Graz: Franz Scherr, Wolfgang Maass; U Sussex: James Knight; INM-6: Sacha van Albada

NBODY6++GPU Dense Star Clusters and Gravitational Waves (Astro) </> CUDA Fortran
U Heidelberg: Rainer Spurzem

Lattice QCD

Bonn Flavour Singlet Structure of Hadrons </> Lib:QUDA
U Bonn: Simone Bacchio, Bartosz Kostrzewa, Carsten Urbach

Wuppertal SignQCD – Studying the Hottest Man-made Liquid </> Lib:QUDA
U Wuppertal: Szabolcs Borsányi, Kalman Szabo

Bielefeld HotQCD – Studying Extreme States of Matter </> CUDA C++
U Bielefeld: Christian Schmit, Dennis Bollweg, Frithjof Karsch

Regensburg Baryons with Charm </> Lib:Grid
Peter Boyle, Christoph Lehner, Gunnar Bali, Sara Collins

Applications II

PIConGPU Plasma Simulations for Next Generation Particle Accelerators (Plasma) </> CUDA C++
HZDR: Alexander Debus, Anton Lebedev, Rene Widera, Michael Bussmann

JUQCS-G Simulating Universal Quantum Computer (Quantum) </> CUDA Fortarn
JSC: Hans De Raedt, K
Dennis Willsch

E-train Understanding Learning Processes in Brain (Neuro) </> Lib:DL
U Graz: Franz Scherr, Wolfgang Maass; U Sussex: James Knight; INM-6: Sacha van Albada

NBODY6++GPU Dense Star Clusters and Gravitational Waves (Astro) </> CUDA Fortran
U Heidelberg: Rainer Spurzem

Lattice QCD

Bonn Flavour Singlet Structure of Hadrons
</> Lib:QUDA
U Bonn: Simone Bacchio, Bartosz Kostrzewa, Carsten Urbach

Wuppertal SignQCD – Studying the Hottest
-made Liquid </> Lib:QUDA
Wuppertal: Szabolcs Borsányi,
Kalman Szabo

Bielefeld HotQCD – Studying Extreme States of Matter </> CUDA C++
U Bielefeld: Christian Schmit, Dennis Bollweg, Frithjof Karsch

Regensburg Baryons with Charm </> Lib:Grid
Peter Boyle, Christoph Lehner, Gunnar Bali, Sara Collins

→ Details on each app online

Feedback to JSC

Issues

- Performance fluctuations (GPU, node, network)
- OpenMPI segmentation violations
- NCCL hangs
- NVHPC Fortran compiler bugs
- UCX configuration (caches)
- PCIe switch bi-directional bandwidth
- PCIe device crashes
- I/O subsystem maturity

Feedback to JSC

Issues

- Performance fluctuations (GPU, node, network)
- OpenMPI segmentation violations
- NCCL hangs
- NVHPC Fortran compiler bugs
- UCX configuration (caches)
- PCIe switch bi-directional bandwidth
- PCIe device crashes
- I/O subsystem maturity

Peculiarities

- AMD CPUs / NUMA domains
- PCIe switch
- GPU device affinity
- Network design (DragonFly+)

Feedback to JSC

Issues

- Performance fluctuations (GPU, node, network)
- OpenMPI segmentation violations
- NCCL hangs
- NVHPC Fortran compiler bugs
- UCX configuration (caches)
- PCIe switch bi-directional bandwidth
- PCIe device crashes
- I/O subsystem maturity

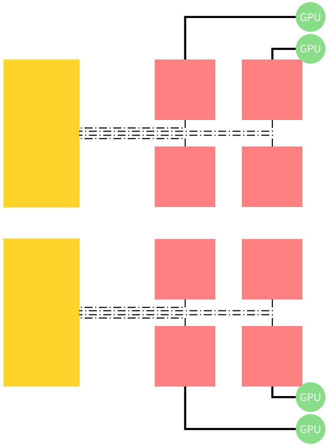
Peculiarities

- AMD CPUs / NUMA domains
- PCIe switch
- GPU device affinity
- Network design (DragonFly+)

Peculiarities

CPU

- AMD EPYC 7402: 24 core processor (SMT-2) $\times$ 2 sockets
- Each socket built as Multi-Chip Module (*chiplets*)
  - 3 cores: Core-Complex (CCX), shared L3 *(max 4 cores)*
  - 2 CCXs: Core Complex Die (CCD)
  - 1 CCDs: 1 Quadrant *(max 2 CCD per Quadrant)*
  - NPS-4: 1 Quadrant $=$ 1 NUMA domain
- Total: 8 memory channels, 2 per quadrant / NUMA domain
- PCIe lanes: 2 $\times$ 16, each 16 connected to 1 quadrant $\rightarrow$ *true* GPU affinity only by half of chiplets



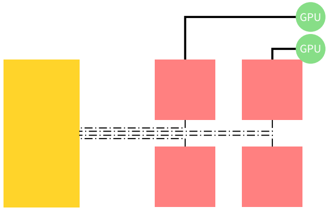
Peculiarities

CPU

- AMD EPYC 7402: 24 core processor (SMT-2) $\times$ 2 sockets
- Each socket built as Multi-Chip Module

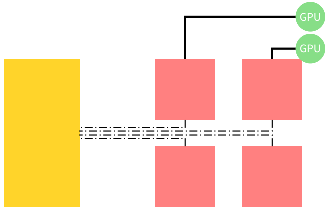
Affinity Not all device have affinity to each other

Rank	NUMA Domain	GPU ID	HCA ID
0	3	0	0
1	1	1	1
2	7	2	2
3	5	3	3



Peculiarities

CPU
- AMD EPYC 7402: 24 core processor (SMT-2) $\times$ 2 sockets
- Each socket built as Multi-Chip Module



Affinity Not all device have affinity to each other

Rank	NUMA Domain	GPU ID	HCA ID
0	3	0	0
1	1	1	1
2	7	2	2
3	5	3	3

$\rightarrow$ New Slurm defaults

```
$ srun --cpu-bind=verbose -n 2 bash -c "" |& sort
cpu_bind=THREADS - jwb0001, task  0  0 [17070]: mask 0x40000 set
cpu_bind=THREADS - jwb0001, task  1  1 [17072]: mask 0x40 set
```

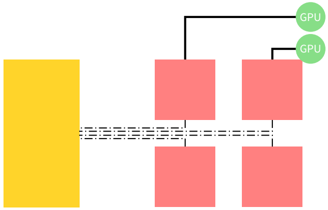
hex2bin
```
000000 000000 000000 100000 000000 000000 000000 000000  # GPU 0
000000 100000 000000 000000 000000 000000 000000 000000  # GPU 1
```

Peculiarities

CPU
- AMD EPYC 7402: 24 core processor (SMT-2) $\times$ 2 sockets
- Each socket built as Multi-Chip Module

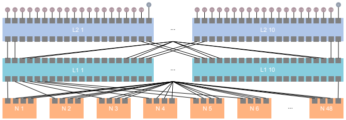
Affinity Not all device have affinity to each other

Rank	NUMA Domain	GPU ID	HCA ID
0	3	0	0
1	1	1	1
2	7	2	2
3	5	3	3



Network 20 cells: many links inside *(right)*, some between (10)
Bi-section bandwidth between N cells:

$$\mathcal{B}(N) = \lfloor (N/2)^2 \rfloor \times (10 \times \mathrm{bw}_1)$$



Peculiarities

CPU
- AMD EPYC 7402: 24 core processor (SMT-2) × 2 sockets
- Each socket built as Multi-Chip Module

Affinity Not all device have affinity to each other

Rank	NUMA Domain	GPU ID	HCA ID
0	3		
			3

Documented online

apps.fz-juelich.de/jsc/hps/juwels/booster-overview.html

Net … many links inside *(right)*, some between (10)

Bi-section bandwidth between N cells:

$$\mathcal{B}(N) = \lfloor (N/2)^2 \rfloor \times (10 \times \text{bw}_1)$$

Early Results

Early Results

Overview

- Some first results by users
- Mainly EA participants
- Most results preliminary
- Results partly on machine under construction

Early Results

Soft Matter: SOMA

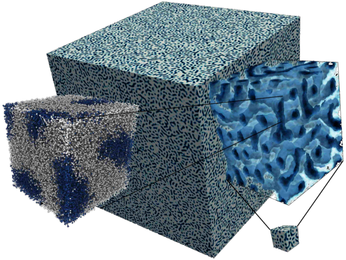
Soft Matter: SOMA

- **SOMA**: Soft, coarse-grained Monte-Carlo Acceleration
  L. Schneider and M. Müller, Comput. Phys. Commun. 235C 463–476 (2019) and GPU Seminar Talk
- Kinetics of nanomaterial formation; multi-component polymer systems (battery materials, membranes, …)
- Unique: Resolve details of polymer, but study lengths relevant to engineering

Team: L. Schneider, N. Blagojevic, L. Pigard, M. Müller, et al

→ `gitlab.com/InnocentBug/SOMA/`

- C, OpenACC, MPI
- Frequent JUWELS user



GEORG-AUGUST-UNIVERSITÄT GÖTTINGEN

SOMA yes, soft matters

JÜLICH Forschungszentrum | JÜLICH SUPERCOMPUTING CENTRE

Soft Matter: SOMA

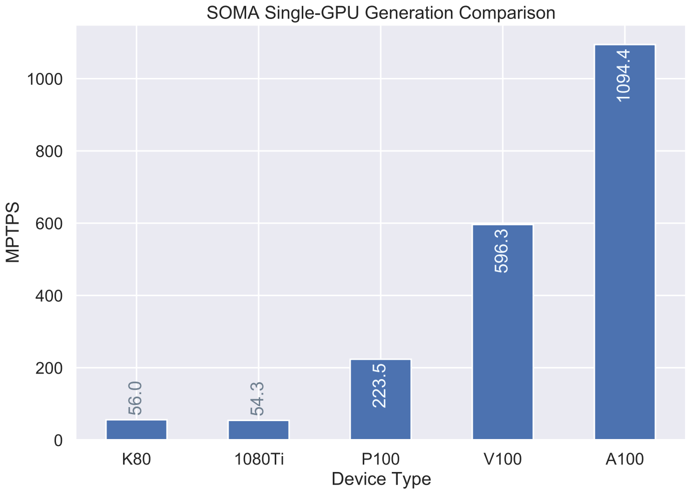
Comparison of GPU Generations

- Long experience with various GPU architectures

→ **Update to new generations early!**

- Some algorithmic changes between generations; also feature additions

- *PTPS: Particle Timesteps Per Second*



Data provided by SOMA / L. Schneider

Soft Matter: SOMA

Kernel Comparison: Memory Chart

- Many random accesses

→ Benefit from larger L1, L2 caches

→ More FP64 throughput

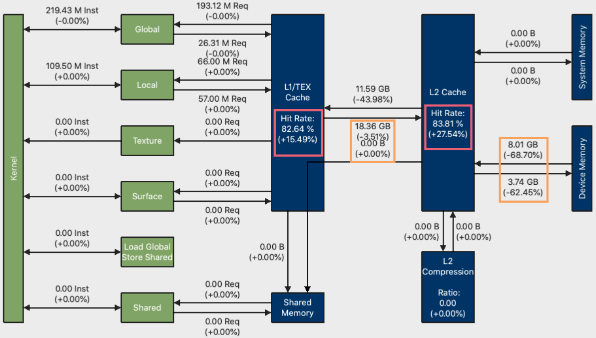
- Knock-on effect: less memory traffic
- **Kernel runtime:**

V100	25.8 ms
A100	21.5 ms
A100*	18.9 ms

exhausting the available communication bandwidth between those units (Max Bandwidth), or by reaching the maximum throughput of issuing memory instructions (Mem Pipes Busy). Detailed chart of the memory units. Detailed tables with data for each memory unit.

Memory Throughput [Gbyte/second]	488.76	(-1.21%)	Mem Busy [%]	45.27	(+99.83%)
L1/TEX Hit Rate [%]	82.64	(+15.49%)	Max Bandwidth [%]	55.82	(+1.26%)
L2 Hit Rate [%]	83.81	(+27.54%)	Mem Pipes Busy [%]	10.80	(+127.96%)
L2 Compression Success Rate [%]	0		L2 Compression Ratio	0	

Memory Chart

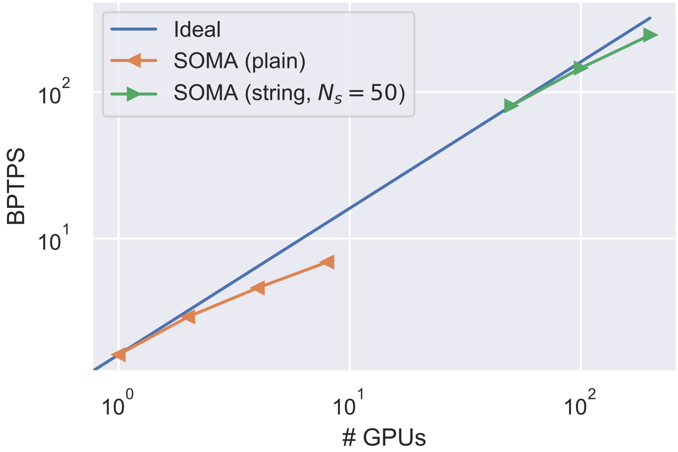


Nsight Compute profile provided by SOMA

Soft Matter: SOMA

New Method for Scaling

- **Scale of Booster: New algorithms, implementations with more scalability!**
- New project for Booster: *String* Method
- String-coupled SOMA ensemble simulation
- Master thesis of N. Blagojevic



Data provided by SOMA / N. Blagojevic

Early Results

Earth-system modelling: ParFlow

Earth-system modelling: ParFlow

- **ParFlow**: Numerical model for groundwater and surface water flow

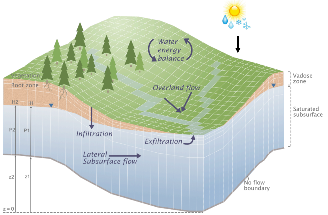
  J. Hokkanen, S. Kollet, et al, EGU General Assembly 2020, 4–8 May 2020, EGU2020-12904, and GPU Seminar Talk

- Model hydrologic processes, hill-slope to continental scale; forecasting, water cycle research, climate change; since 1990s
- Finite-difference scheme with implicit time integration

Team: J. Hokkanen, S. Kollet

→ `parflow.org`

- C, C++, CUDA, MPI
- Fresh GPU port in prepartion for Booster

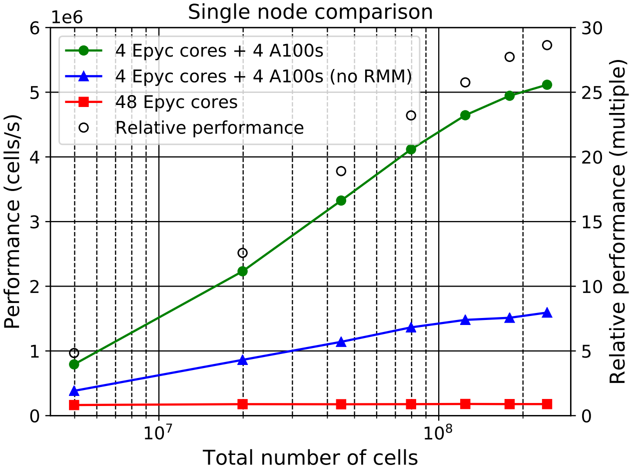


JÜLICH Forschungszentrum | JÜLICH SUPERCOMPUTING CENTRE

Earth-system modelling: ParFlow

Single-Node Performance

- Comparing CPU of Booster node with GPUs
- **Good speed-up, max. 29×**
- Memory pool (*RMM*) gives extra boost
- Larger problem sizes solvable per node

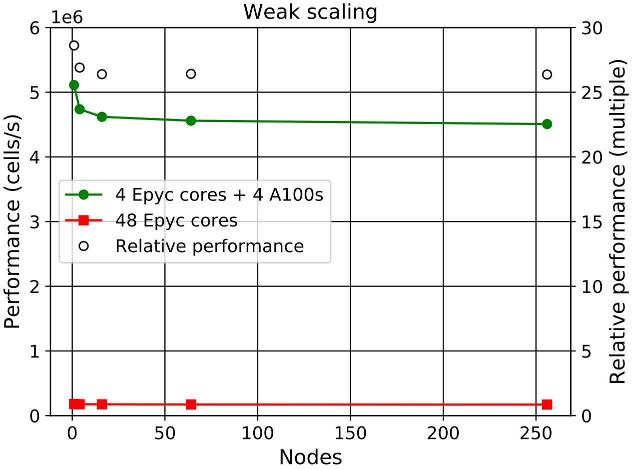


Plot provided by ParFlow / J. Hokkanen

Earth-system modelling: ParFlow

Weak Scaling

- Fixed problem size per node
- **26× speed-up achieved over $\mathcal{O}(100)$ nodes**



Plot provided by ParFlow / J. Hokkanen

Early Results

JUQCS

Quantum Computing: JUQCS

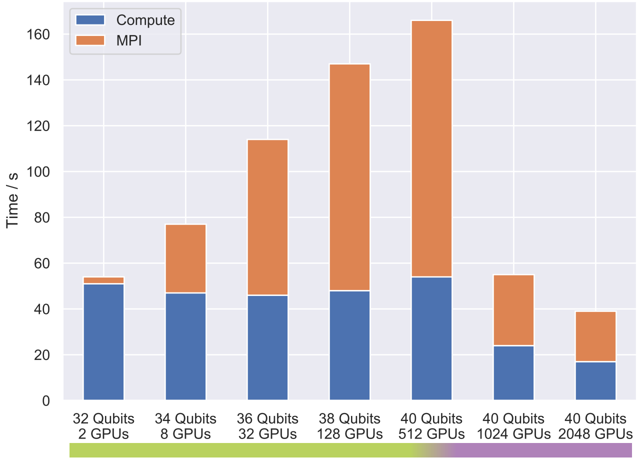
- **JUQCS**: Jülich Universal Quantum Computer Simulator
  De Raedt et al., Comp. Phys. Comm. 237 47–61 (2019)
- Universal quantum computing on digital computer
- Network-, memory-intensive computations
- Team: Research group Quantum Information Processing

- Fortran, CUDA Fortran
- Frequent JUWELS user

JÜLICH SUPERCOMPUTING CENTRE

JUQCS

- **40 qubits:**
  - $>$ 16 TiB memory needed $\rightarrow$ 512 A100s
  - Each quantum operation: Update states, 8 TB transfer
- Weak scaling: Compute constant, MPI as expected
- Strong scaling: Still investigate DragonFly+ topology

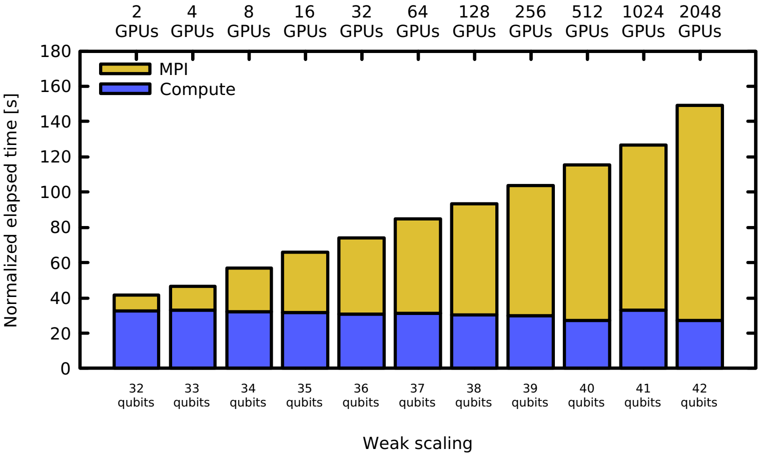


Data provided by JuQCS / H. De Raedt

JUQCS

More Weak Scaling

- Weak scaling to 2048 GPUs / 42 qubits
- Good behavior, but MPI still limiter



Early Results

LQCD: Bonn

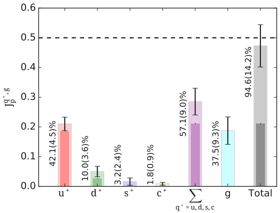
LQCD: Bonn

- **ETMC**: Extended Twisted Mass Collaboration
  C. Alexandrou and S. Bacchio et al, Phys. Rev. D 101 094513 (2020)
- Study of the Flavour Singlet Structure of Hadrons

Team: S. Bacchio, B. Kostrzewa, et al; Uni Bonn, Uni Cyprus, Cyprus Institute, Uni Rome, …

→ `github.com/etmc`, PLEGMA, QUDA, tmLQCD

- C/C++, CUDA, MPI, OpenMP
- Frequent JUWELS user



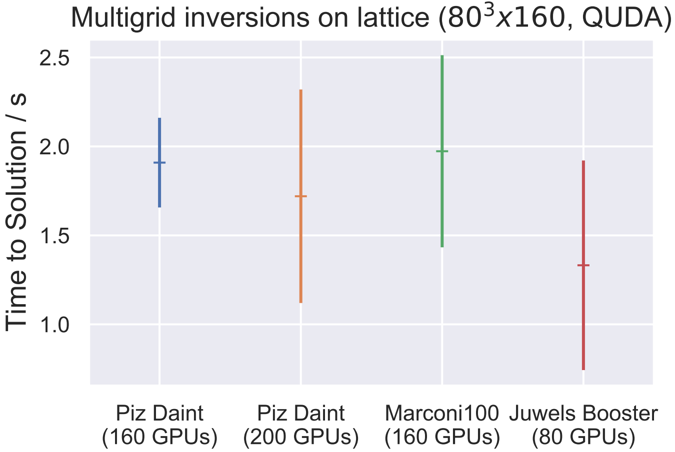
LQCD: Bonn

Comparison of GPU HPC Machines

- Multigrid inversion
- Mean time-to-solution, spread
- Systems

  Piz Daint Haswell, P100; DragonFly

  Marconi100 POWER9, V100; DragonFly+
- JUWELS Booster: Low time to solution; but large spread (being investigated)



Data provided by ETMC / S. Bacchio

Early Results

PIConGPU

PICon GPU

- **PIConGPU**: Plasma simulation

  H. Burau et al, IEEE Transactions on Plasma Science 38 10 (2010)

- Particle-in-cell simulation for Exascale-level GPUs

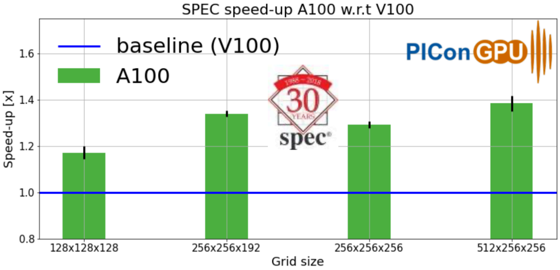
Team: A. Lebedev, A. Debus, M. Bussmann, et. al

→ `github.com/ComputationalRadiationPhysics/picongpu`

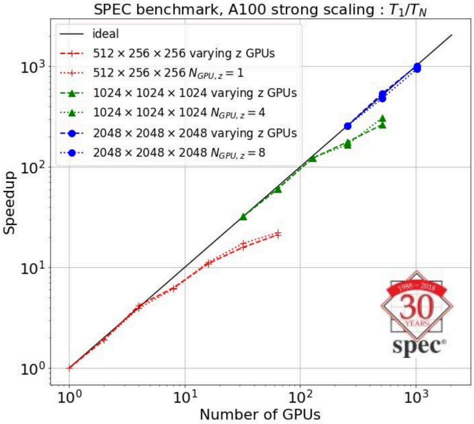
- C/C++, CUDA, MPI, Alpaka

PIConGPU

Results



Comparison A100 vs. V100



Strong scaling for different grid sizes

Graphs by PIConGPU / M. Bussmann

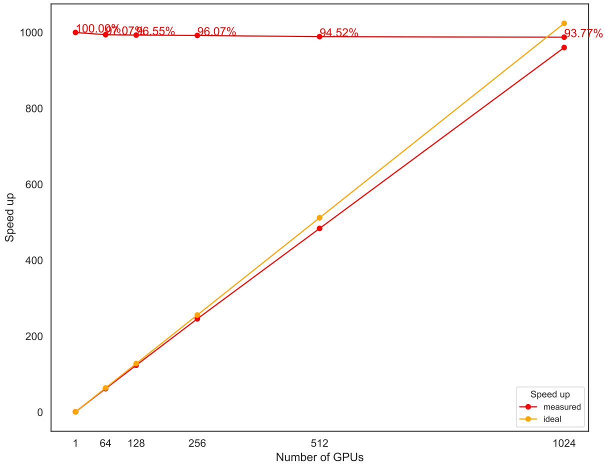
Early Results

Others

Large-Scale Pre-Training on Transfer Learning for Images

Deep-Learning

- **Publication**: Effect of large-scale pre-training on full and few-shot transfer learning for natural and medical images
- Authors: Mehdi Cherti, Jenia Jitsev; JSC
- Status: Preprint (under review) arXiv:2106.00116 [cs.LG]

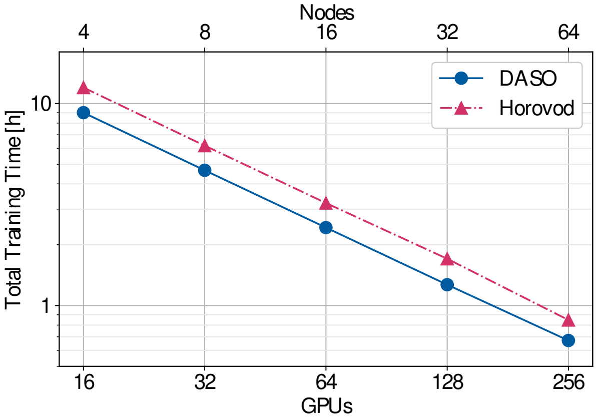


Graph from publication arXiv:2106.00116 [cs.LG]

Distributed Training with DASO

Deep-Learning

- **Publication:** Accelerating Neural Network Training with Distributed Asynchronous and Selective Optimization (DASO)
- Authors: D. Coquelin et. al; KIT, DLR
- Unique: 25 % improvement over Horovod
- Status: Preprint arXiv:2104.05588 [cs.LG]

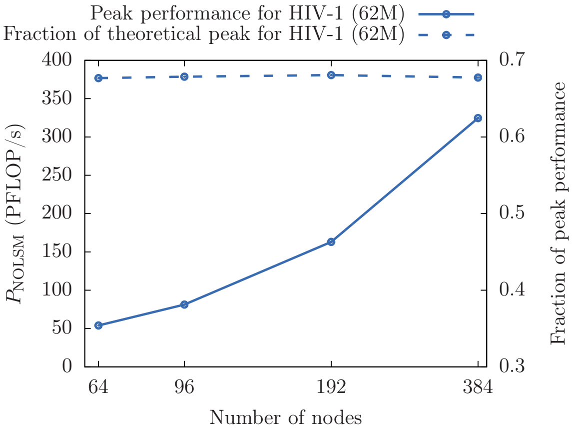


Graph from publication arXiv:2104.05588 [cs.LG]

Large-Scale Ab-Initio Molecular Dynamics

Molecular Dynamics

- **Publication**: Enabling Electronic Structure-Based Ab-Initio Molecular Dynamics Simulations with Hundreds of Millions of Atoms
- Authors: R. Schade et. al; Paderborn University
- Unique: FP16/FP32 mixed precision, 1536 GPUs, 324 PFLOP/s
- Status: Preprint arXiv:2104.08245 [physics.comp-ph]



Graph from publication arXiv:2104.08245 [physics.comp-ph]

Summary and Conclusions

Summary

- JUWELS Booster: European flagship system based on A100 GPUs and HDR200 InfiniBand network
- Highly scalable system design with > 70 PFLOP/s$_{\mathrm{FP64}}$ compute performance and 749 Tbit/s acc. injection bandwidth
- In production since end of November, some applications earlier through Early Access Program
- First results incoming; second allocation period started

Summary

- JUWELS Booster: European flagship system based on A100 GPUs and HDR200 InfiniBand network
- Highly scalable system design with $>$ 70 PFLOP/s_{FP64} compute performance and 749 Tbit/s acc. injection bandwidth
- In production since end of November, some applications earlier through Early Access Program
- First results incoming; second allocation period started

Thank you for your attention!

a.herten@fz-juelich.de

Appendix

Appendix

Network Performance

References

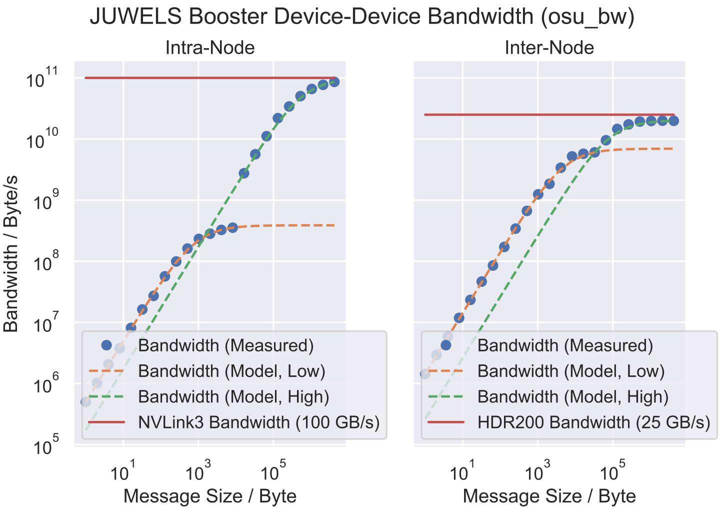
Appendix

Network Performance

Network Performance

OSU Micro-Benchmarks: Bandwidth

- OSU Microbenchmarks: device-device bandwidth (`osu_bw D D`)
- Good results, expected limiters
- Intra-node: NVLink3 bandwidth
- Inter-node: HDR200 bandwidth
- Model fits show 2 regimes (---/ ---)



Appendix

References

References: Images, Graphics I

[1] Forschungszentrum Jülich GmbH (Ralf-Uwe Limbach). *JUWELS Cluster*.

[2] Forschungszentrum Jülich GmbH (Ralf-Uwe Limbach). *JUWELS Booster*.