ENABLING APPLICATIONS FOR JUWELS BOOSTER

GTC DIGITAL 2020

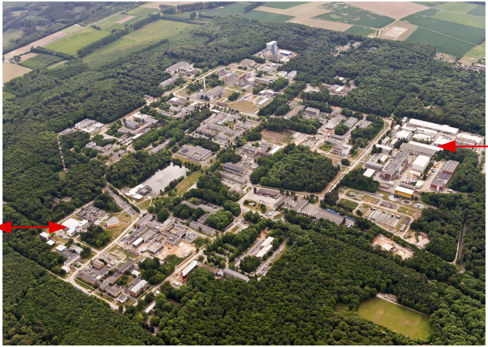
October 2020 | Dirk Pleiter, Andreas Herten | Jülich Supercomputing Centre, Forschungszentrum Jülich

Outline

Introduction

About Forschungszentrum Jülich

JSC

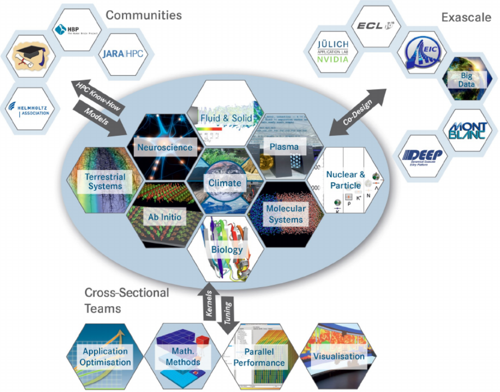


- One of Europe’s largest interdisciplinary research centres; about 6,400 employees
- Special expertise in physics, materials science, nanotechnology, neuroscience and medicine, and information technology
- Leader in various European HPC projects, including PRACE

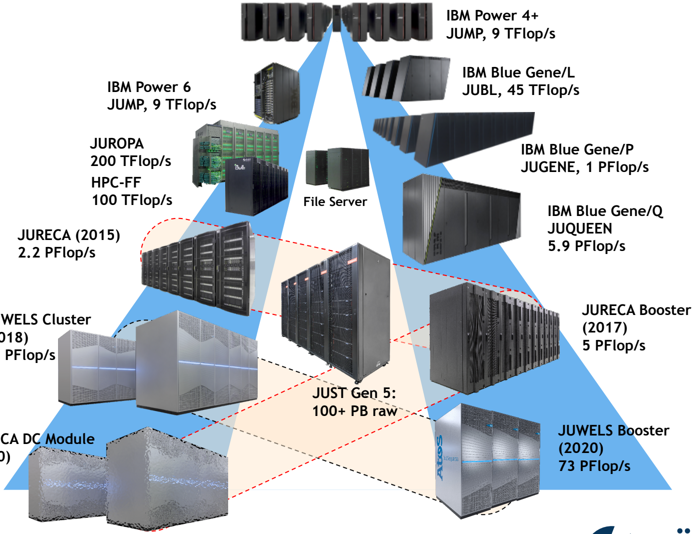
JÜLICH Forschungszentrum | JÜLICH SUPERCOMPUTING CENTRE

About Jülich Supercomputing Centre

- Supercomputer operation for
  - Centre – FZJ,
  - Regional – JARA
  - Helmholtz & National – NIC, GCS
  - Europe – PRACE, EU projects
- Education and Training
- Application support
  - User support
  - Peer review support and coordination
- Research and development
  - Computational science: SimLab
  - Algorithms, performance analysis and tools
  - HPC architectures and technologies: Exascale Laboratories, Community data management service



JSC's HPC Infrastructure

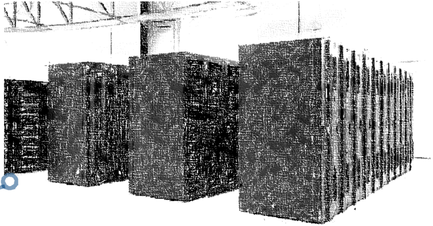


JÜLICH Forschungszentrum | JÜLICH SUPERCOMPUTING CENTRE

JUWELS Overall Architecture

JUWELS Cluster (2018)

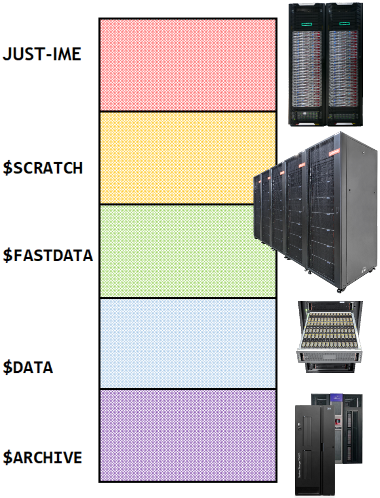
- 2511 compute nodes (2× Skylake)
- 48 GPU nodes (4× V100 w/ NVLink2)
- Mellanox EDR 100 Gbit/s network, fat-tree topology (1:2@L1)
- 12 PFlop/s



JUWELS Booster (2020)

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

JSC’s Storage Infrastructure



Bandwidth optimized, capacity limited storage (NVM based)

Capacity and bandwidth balancing HPC storage (temp.)

Capacity and bandwidth balancing HPC storage (pers.)

High capacity, low bandwidth storage (campaign use cases)

Archival storage based on high latency media

JUWELS Booster

JUWELS Booster in a Nutshell (1/2)

- 936 compute nodes
  - 2× 24-core AMD EPYC Rome CPUs
    - $C_{\text{mem}}^{(\text{CPU})} = 2 \times 256$ GByte DDR4-3200 memory
  - 4× Nvidia A100 GPUs, each GPU features
    - $B_{\text{fp}}^{(\text{GPU})} = 9.7$ TFlop/s peak performance

      With tensor cores: $B_{\text{fp}}^{(\text{GPU})} = 19.5$ TFlop/s
    - $C_{\text{mem}}^{(\text{CPU})} = 40$ GByte HBM2 memory
    - $B_{\text{mem}}^{(\text{GPU})} = 1.5$ TByte/s memory bandwidth
    - NVLink3
  - 1× HDR200 InfiniBand port per GPU
- DragonFly+ network topology with 20 cells
  - All links with 200 Gbit/s HDR200
  - 40 Tbit/s connection to Cluster

JÜLICH Forschungszentrum JÜLICH SUPERCOMPUTING CENTRE

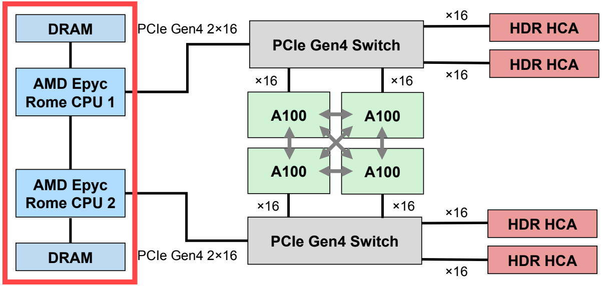
JUWELS Booster in a Nutshell (2/2)

- High I/O performance
  - $>$ 400 GByte/s bandwidth to JUST-DSS
  - Up to 1 TByte/s bandwidth to JUST-IME
- Bull Sequana XH2000 system with warm-water cooling
  - 37 °C inlet temperature

JUWELS Booster: Node Design (1/3)

2× **AMD Rome 7402 CPUs**

- DDR4-3200 memory DIMMs
- 2×8 memory channels
  $\Rightarrow B_{\text{mem}}^{(\text{CPU})} = 410\,\text{GByte/s}$
- $C_{\text{mem}}^{(\text{CPU})} = 2 \times 256\,\text{GiByte}$
- Total of 96 PCIe Gen4 lanes

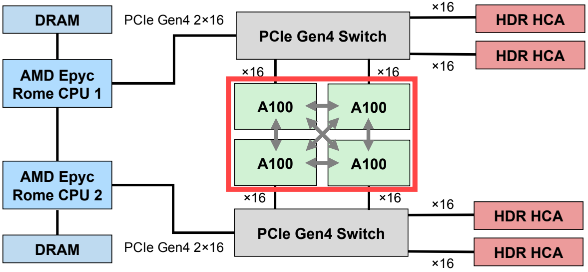


JÜLICH Forschungszentrum | JÜLICH SUPERCOMPUTING CENTRE

JUWELS Booster: Node Design (2/3)

NVIDIA HGX A100 (“Redstone”) board

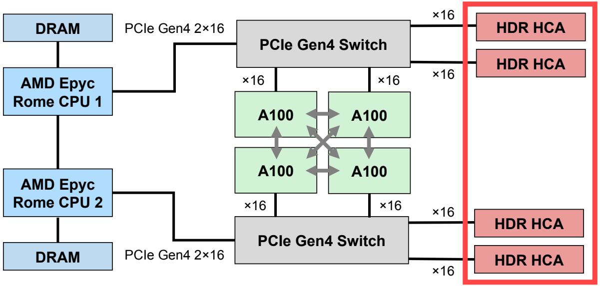
- $4\times$ A100
- NVLink3 full mesh
  - $4\times$ GPU-to-GPU links $\Rightarrow$ 100 GByte/s per direction
- $\times 16$ PCIe Gen4 links to CPUs
  - 63 GByte/s between CPUs and GPUs per direction



JUWELS Booster: Node Design (3/3)

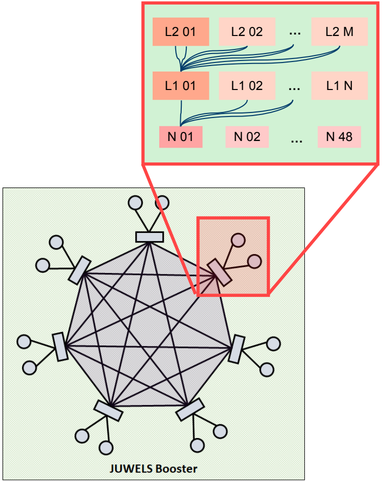
4× HCA Mezzanine cards

- Mellanox ConnectX-6 cards
- 200 Gbit/s per card and direction
- GPUdirect RDMA support



JUWELS Booster: Network Design

- DragonFly+ topology [A. Shpiner et al., 2017]
  - Maximally 5 hops between two nodes (or more with dynamic routing)
- 20$\times$ switch groups (“cells”)
  - 48 nodes $\Rightarrow$ 192 up-links
  - 10 leaf + 10 spine routers
  - Full fat-tree topology within switch group $\Rightarrow$ 40 Tbit/s bi-section bandwidth
- 10 links connecting each pair of switch groups
  - 4 Tbit/s bi-section bandwidth between switch groups
  - 400 Tbit/s global bi-section bandwidth



JÜLICH Forschungszentrum | JÜLICH SUPERCOMPUTING CENTRE

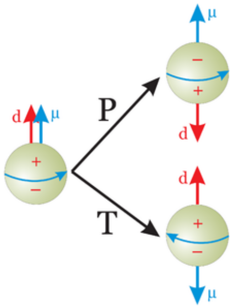
JUWELS Booster: Software Stack

- Fully integrated Cluster and Booster module
  - ParaStation as core enabler
    - Resource Management
    - MPI Implementation (MPICH-based)
    - Extensions to support multi-GPU nodes
- Slurm as Workload Manager for JUWELS
- Red Hat Enterprise Linux and CentOS 8

Scientific Grand Challenges (Selection)

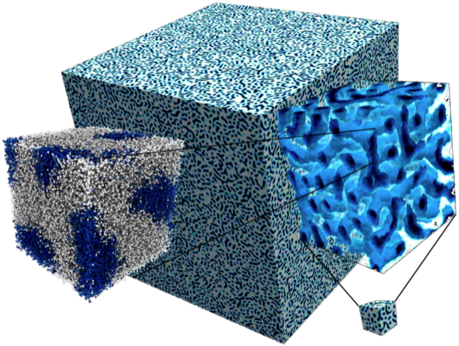
Particle Physics

- **Scientific challenge:** Exploring CP symmetry violations through high-precision determination of the neutron electric dipole moment (nEDM)
  - Search for physics beyond the Standard Model
- **Approach:** Simulation of Quantum Chromodynamics on the lattice
- **Computational challenge(s):**
  - Computation of very long trajectories that requires strong scaling
  - Simulations at physical quark masses using fine lattices



Soft Matter

- **Scientific challenge:** Enable optimisation of polymeric materials for Lithium-Ion batteries polymeric electrodes by simulating their transport properties
- **Approach**: Simulation of large systems comprising polymers
- **Computational challenge(s):**
  - Simulation of particle-based models using a very large number of particles

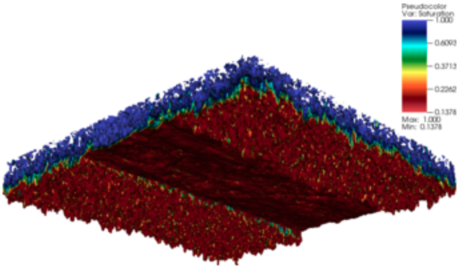


[L. Schneider, 2020]

JÜLICH Forschungszentrum | JÜLICH SUPERCOMPUTING CENTRE

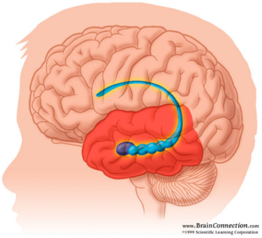
Earth System Modelling

- **Scientific challenge:** Reliable prediction of hydrometeorological extremes
- **Approach:**
  - Create seamless land-ocean-hydro-meteorological prediction system at pan-European scale
  - Enable forecasting and projecting weather-driven extremes over days up to multiple decades
- **Computational challenge(s):**
  - Very large number of free parameters when going to finer resolution
  - Coupled applications of different characteristics



Brain Research

- **Scientific challenge**: Create understanding of higher brain functions (learning, memory, spatial navigation) as well as dysfunctions causing mental diseases including Alzheimer
- **Approach**: Simulation of the brain at different scales
  - Large-scale models based on biologically realistic networks
  - Detailed neuron/synapse models
  - Effective brain-level models
- **Computational challenge(s)**:
  - Coupled applications
  - Large memory footprint



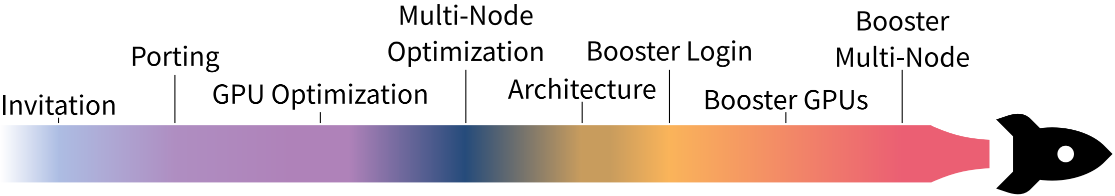
Early Access Program

Overview

- Started in early 2020
- Selected 14 applications from various scientific domains (access closed)
  - Aimed for applications that could use JUWELS Booster at scale
  - Some teams already use JUWELS Cluster, others are new
- **Offer**: Use JUWELS Booster before general access
- Involved many groups at **JSC**
  - 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

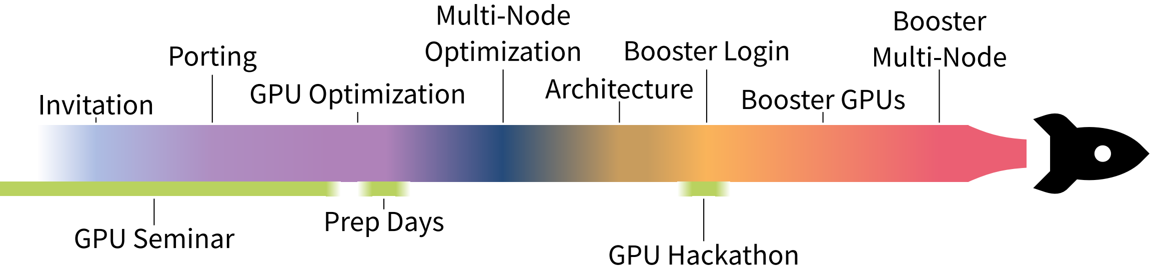
Timeline to Booster

- Possible timeline of an application preparing for JUWELS Booster
- Not all applications need all steps, or start at beginning



Timeline to Booster

- Possible timeline of an application preparing for JUWELS Booster
- Not all applications need all steps, or start at beginning
- Additionally: events



Early Experiences, Lessons Learned

- EA Program tailored individually around each application
  - Very different statuses of GPU acceleration in application
  - Different ways of working
  - Diverse response times
- Fresh system, fresh software stack: Update as early as possible
- One can never start early enough
- Knowledge dissemination programs well-received (talks, newsletter, overview documentations, chat)
- Challenging to schedule EA runs and low-level system tests at same time

Early Performance Results

Disclaimer

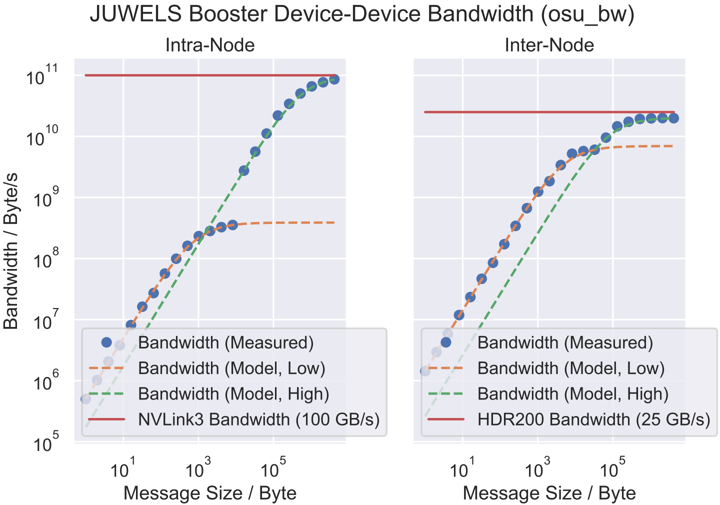
- Following results obtained on very, very fresh JUWELS Booster
- …while system integration work was done at same time
- Only few nodes available
- System will be tuned and improved
- …also due to results obtained by EA applications!

- Software used
  - GCC 9.3.0
  - CUDA 11.0 (with CUDA Driver 450.51.06)
  - NVHPC 20.7
  - ParaStationMPI 5.4.7 (with UCX 1.8.1)

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 (---/ ---)



Soft Matter: SOMA

- **SOMA**: Soft, coarse-grained Monte-Carlo Acceleration

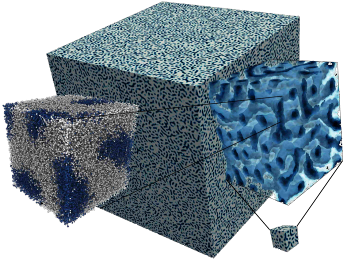
  L. Schneider and M. Müller, Comput. Phys. Commun. 235C 463–476 (2019) and GPU Seminar Talk
- Kinetics of nanomaterial formation; multi-component polymer systems (battery materials, membranes, ...)
- Unique: Resolve details of polymer, but study lengths relevant to engineering
- Team: L. Schneider, N. Blagojevic, L. Pigard, M. Müller, et al

→ `gitlab.com/InnocentBug/SOMA/`

- C, OpenACC, MPI
- Frequent JUWELS user

SOMA yes, soft matters

GEORG-AUGUST-UNIVERSITÄT GÖTTINGEN



SOMA Performance Results

Kernel Comparison: Memory Chart

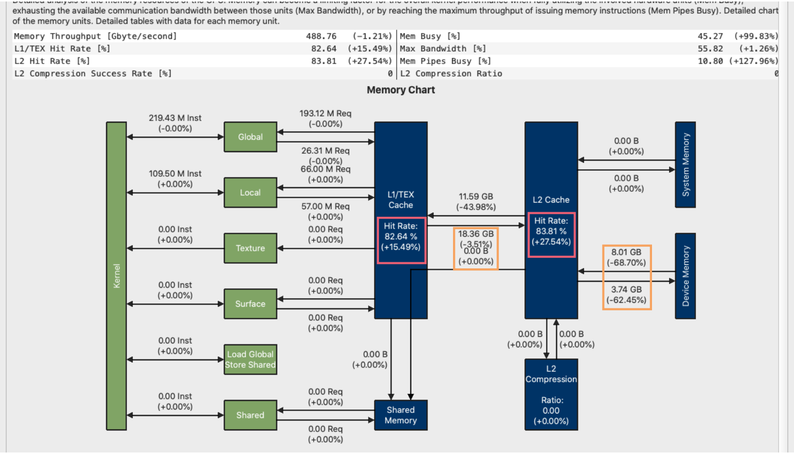
- Many random accesses

→ 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

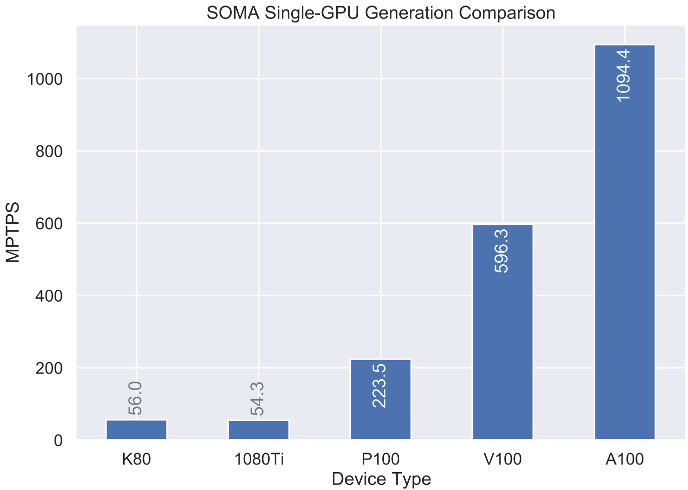


Nsight Compute profile provided by SOMA

SOMA Performance Results

Comparison of GPU Generations

- Long experience with various GPU architectures
- Good performance increase with each generation
- Some algorithmic changes between generations; also feature additions
- *PTPS: Particle Timesteps Per Second*



Data provided by SOMA / L. Schneider

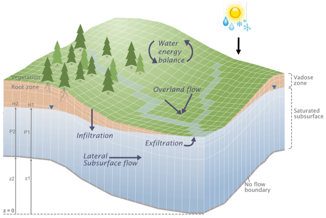
Earth-system modelling: ParFlow

- **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`

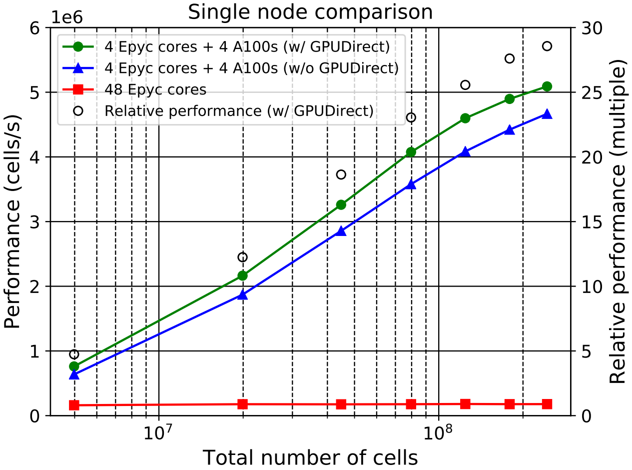
- C, C++, CUDA, MPI
- Fresh GPU port in prepartion for Booster



ParFlow Performance Results

Single-Node Performance

- Comparing CPU of Booster node with GPUs
- Good speed-up, max. 29×
- GPUDirect gives extra boost
- Larger problem sizes solvable per node

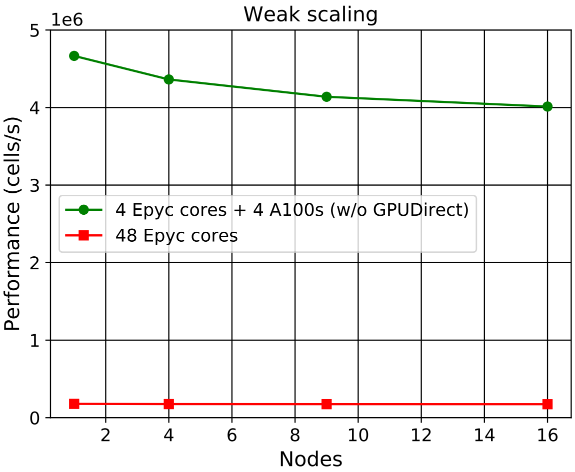


Plot provided by ParFlow / J. Hokkanen

ParFlow Performance Results

Weak Scaling

- Fixed problem size per node
- *w/ GPUDirect* currently under investigation



Plot provided by ParFlow / J. Hokkanen

Summary and Conclusions

Summary

- JUWELS Booster: European flagship system based on A100 GPUs
  - Science instrument for various scientific grand challenges
- Planned to go into production in November 2020
  - Applications are prepared through an Early Access Program
- Very early performance results are encouraging

Acknowledgements

- JSC High Performance Systems: Dorian Krause, Damian Alvarez, Benedikt von St. Vieth
- NVIDIA Collaborators: Markus Hrywniak, Jiri Kraus, Mathias Wagner
- Participants of Early Access Program, especially

SOMA	Ludwig Schneider, Louis Pigard, Niklas Blagojevic
ParFlow	Jaro Hokkanen
LQCD Bonn	Bartosz Kostrzewa