Extreme-Scale Readiness of ParFlow

GPU Scaling, Architectural Portability, and In Situ Workflows

CONNECTING THE DOTS

ISC High Performance

26.06.2026 I **Muhammad Fahad**[1], Daniel Caviedes-Voulliéme[1], Andrés Bermeo Marinelli[2], Hugo Strappazzon[2], Bruno Raffin[2], Stefan Kollet[1]

[1]Institute of Bio-Geosciences (IBG-3, Agrosphere), Research Centre Jülich, Jülich, Germany

[2]Université Grenoble Alpes, Inria, CNRS, Grenoble, France

ISC Workshop: Readiness of HPC Extreme-Scaling Applications (3rd Edition)

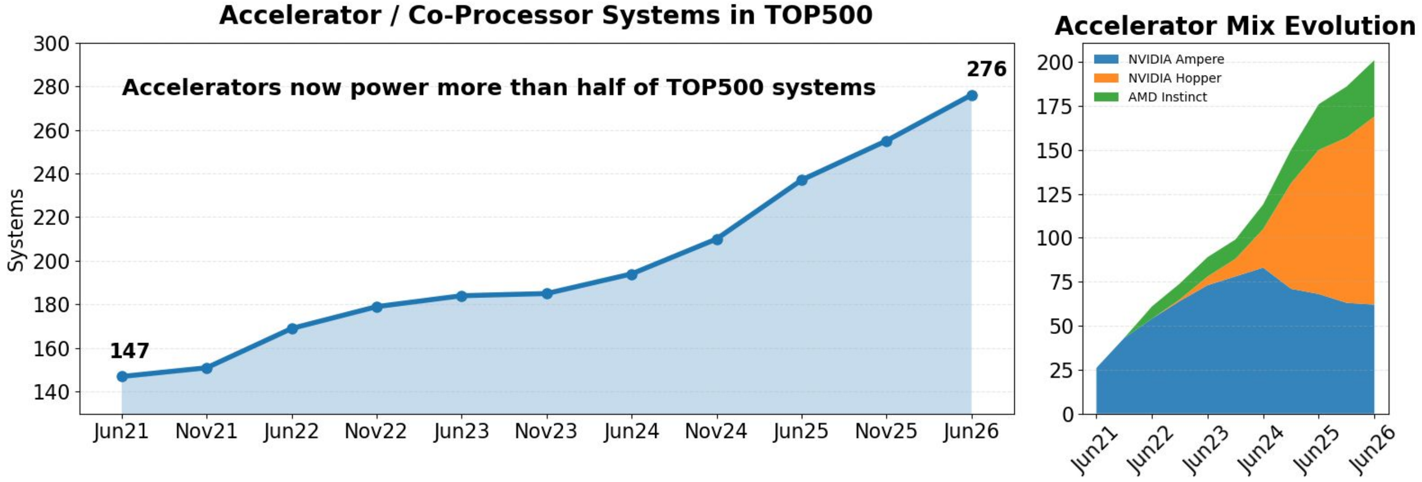
The Hardware Shift: Acceleration Dominates HPC

Accelerator Adoption Continues to Reshape the TOP500

276
Accelerator-based systems
(June 2026)

+129
Net increase
since June 2021

+88%
5-year growth


Accelerator / Co-Processor Systems in TOP500

Accelerator Mix Evolution

Key trend: rapid transition from Ampere to Hopper while AMD Instinct adoption continues to grow.

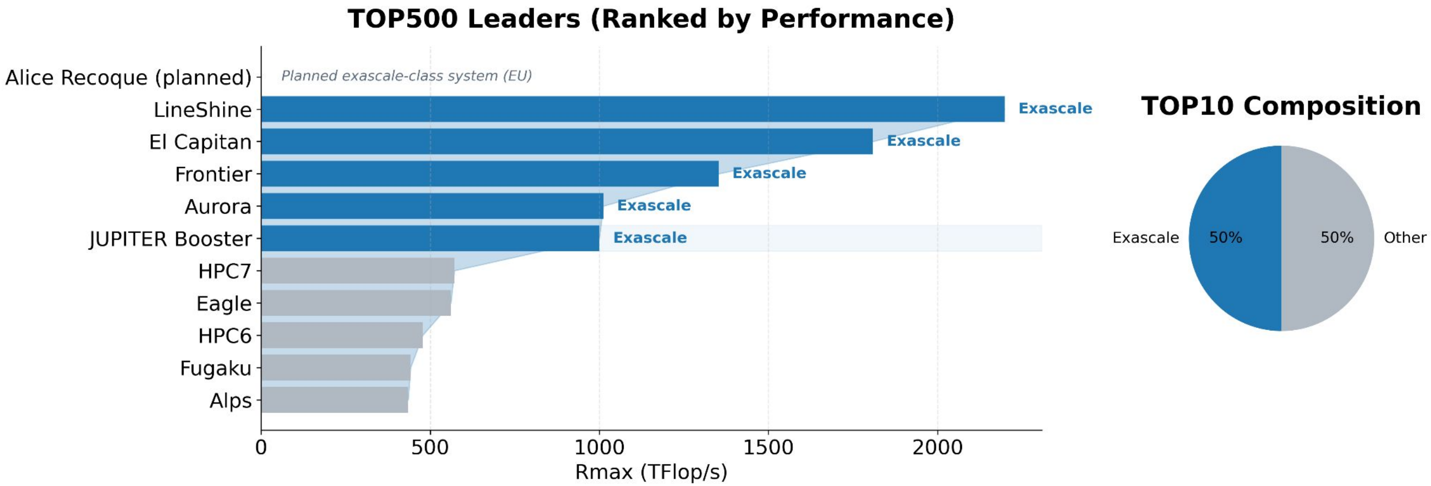
Exascale Is Here!

Exascale Systems Are Already Embedded in the TOP500 (June 2026)

5
Exascale systems
in TOP10

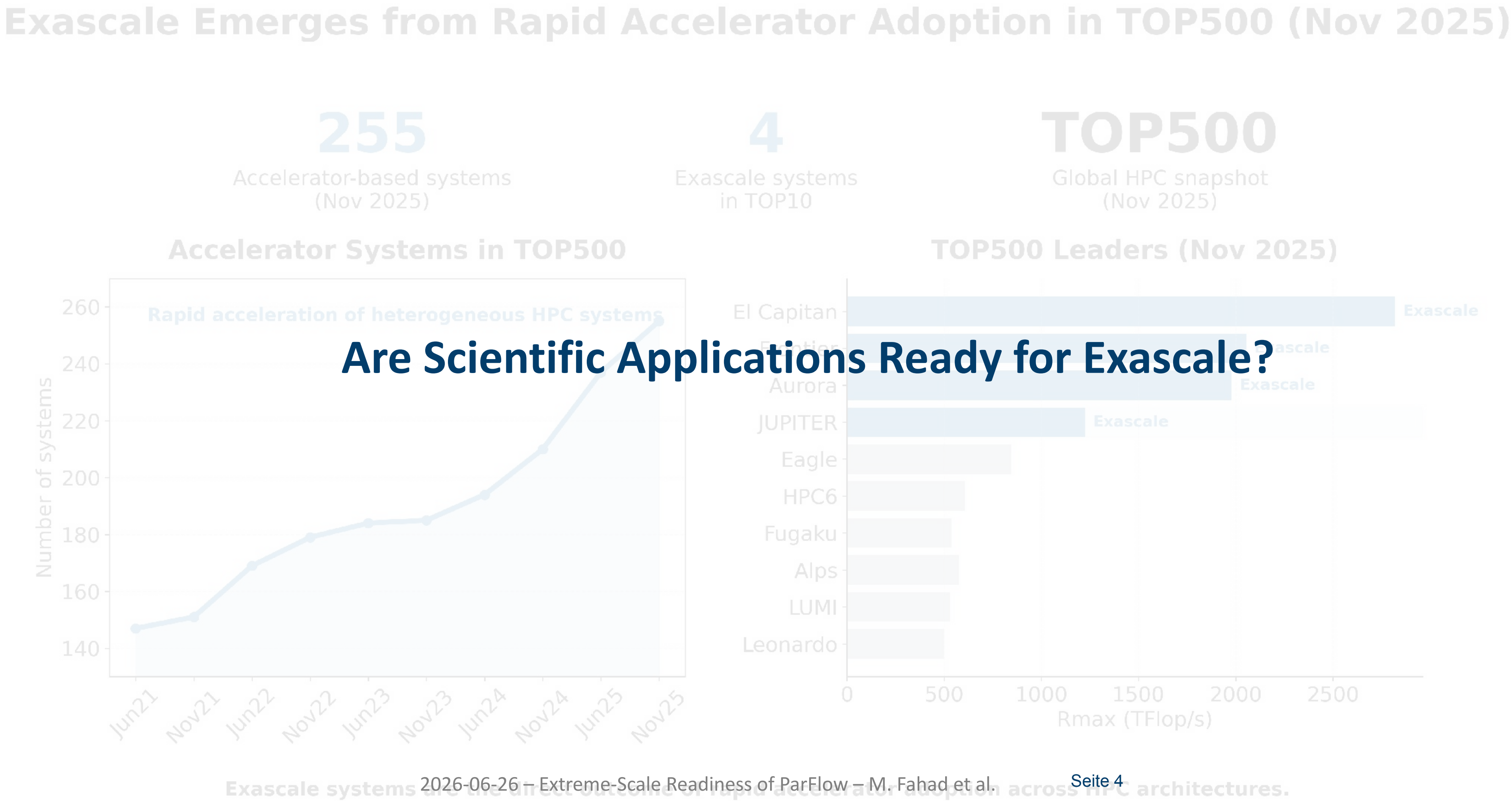
TOP500
Global HPC ranking
(June 2026)

5/10
Exascale share
in TOP10



Exascale capability is already part of the operational TOP500 leadership landscape.

Are Scientific Applications Ready for Exascale?



Why This Is Hard for Scientific Applications?

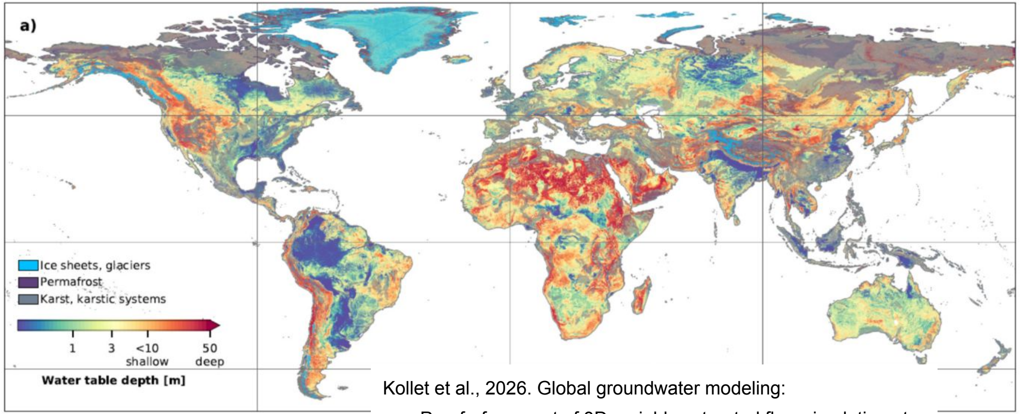
Scientific Workflows Face Bottlenecks at Scale

- Compute is no longer the only bottleneck
- Key challenges include:
  - Memory hierarchy complexity
  - Communication overhead at scale
  - I/O bottlenecks dominating runtime
  - Workflow inefficiencies
  - Diverse architectures and Programming models: Performance Portability ?
- Traditional model breaks:
  - File-based post-processing → expensive
  - Data movement becomes dominant cost

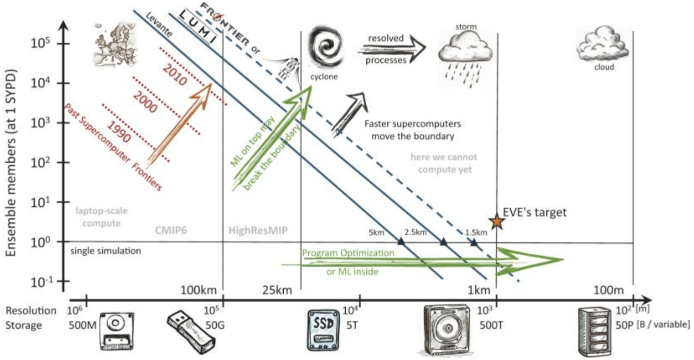
Hydrology Models

Key Challenges

- Coupled physics:
  - Surface + subsurface flow
- Large structured grids
- High-performance implicit solvers
- Strong demand for predictive Earth system simulations
  - Increasing spatial resolution and coupling complexity
- Scaling challenges:
  - Fine-resolution simulations (sub-Kilometer scale)
  - Tight coupling → synchronization costs
  - Strong memory and communication demands
  - Compute demands: Exponential growth in Degree Of Freedoms (DOFs) and communication pressure



Kollet et al., 2026. Global groundwater modeling: Proof-of-concept of 3D variably saturated flow simulation at kilometer resolution. Journal of Hydrology X 30, 100213.

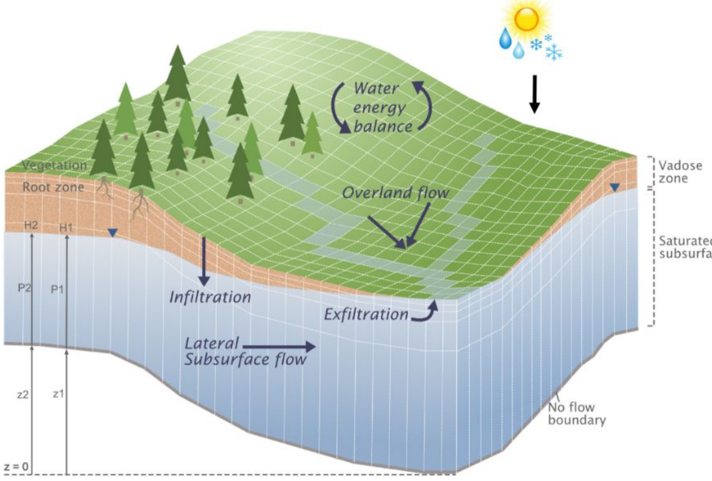


Hoefler et al.,., 2023. Earth Virtualization Engines: A Technical Perspective. Comput. Sci. Eng. 25, 50–59.

Extreme-scale Readiness

ParFlow

- Open-source[1], massively parallel Integrated Hydrological Model
- Fully coupled, 2D/3D dynamical representation of the hydrological processes in the variably saturated zone, including groundwater and overland flow
- Flagship application of Energy Oriented Center of Excellence (EoCoE) III
- **C**, C++, Python
- GPU-enabled backends via internal Performance Portability Layer
  - CUDA
  - Kokkos (HIP, CUDA)
- Supports advanced memory management with optimized allocators for enhanced performance
  - RMM, and UMPIRE



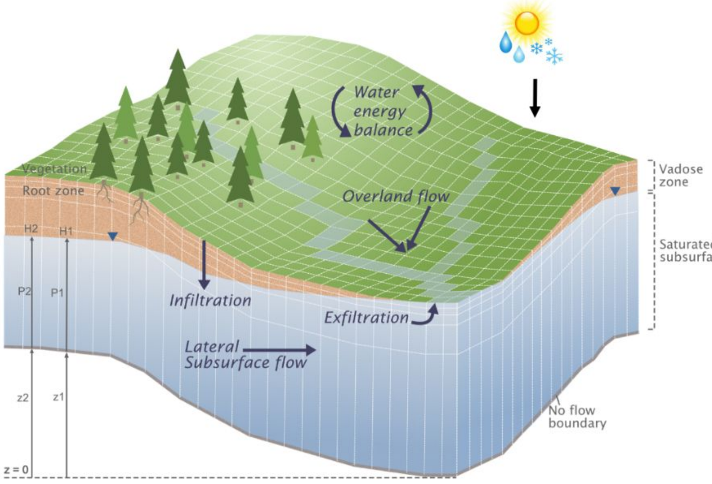
---

[1] https://github.com/parflow

Extreme-scale Readiness

ParFlow

- Open-source[1], massively parallel Integrated Hydrological Model
- Fully coupled, 2D/3D dynamical representation of the hydrological processes in the variably saturated zone, including groundwater and overland flow
- Flagship application of Energy Oriented Center of Excellence (EoCoE) III
- **C**, C++, Python
- GPU-enabled backends via internal Performance Portability Layer
  - **CUDA**
  - ~~Kokkos (HIP, CUDA)~~
- Supports advanced memory management with optimized allocators for enhanced performance
  - **RMM**, ~~and UMPIRE~~



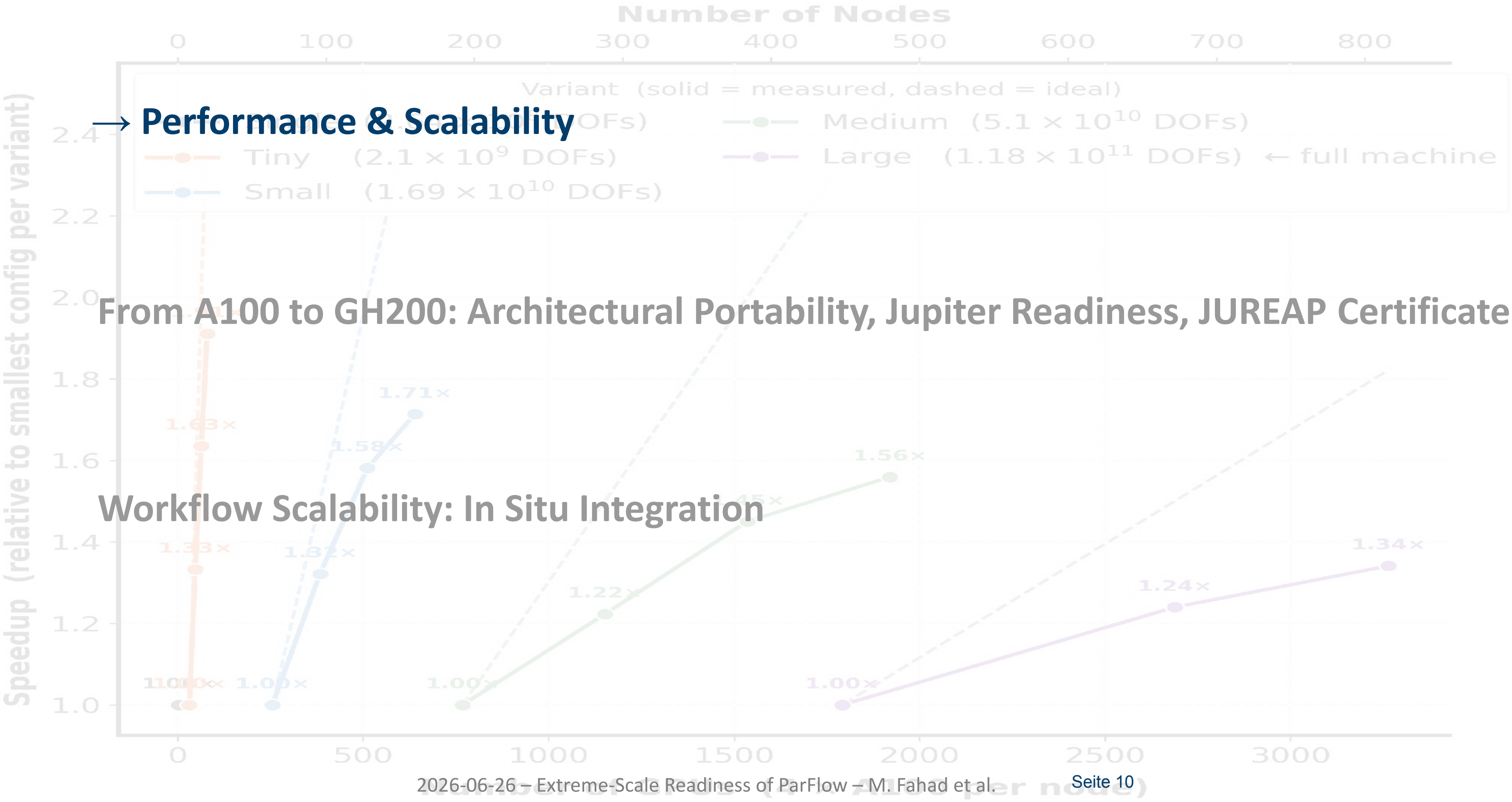
---

[1] https://github.com/parflow

ParFlow: Extreme-scale Readiness

Key Challenges

- **Performance & Scalability**
  - Scaling to thousands of GPUs
  - Maintaining efficiency at increasing node counts
- **Architectural & Performance Portability**
  - Supporting multiple architectures and vendors
  - Minimizing code modifications
  - Achieving high efficiency across platforms
- **Workflow Scalability**
  - Mitigating I/O bottlenecks
  - Managing storage and data movement
  - Enabling scalable analysis workflows



→ Performance & Scalability

From A100 to GH200: Architectural Portability, Jupiter Readiness, JUREAP Certificate

Workflow Scalability: In Situ Integration

Hardware Platforms

JUWELS BOOSTER

- CPU: AMD EPYC 7402 processor; 2 sockets, 24 cores per socket
- Memory: 512 GB DDR4-3200 RAM; 256 GB per socket
- GPU: 4 × **NVIDIA A100** Tensor Core GPU with 40 GB
- Network: 4 × Mellanox HDR200 InfiniBand ConnectX 6

Copyright: Forschungszentrum Jülich

https://www.fz-juelich.de/en/jsc/systems/supercomputers/juwels

Hardware Platforms

JUPITER Exascale Development Instrument (JEDI)

JUPITER Exascale Development Instrument (JEDI)

- 4× NVIDIA **GH200 Grace-Hopper Superchip**
  - CPU: NVIDIA Grace (Arm Neoverse-V2), 72 cores; 120 GB LPDDR5X memory
  - GPU: NVIDIA Hopper, 96 GB HBM3 memory at 4 TB/s
- Network: 4 x InfiniBand NDR200 (Connect-X7)

The JUPITER Exascale Development Instrument (left, orange) at the the Jülich Supercomputing Centre. Copyright: Forschungszentrum Jülich / Ralf-Uwe Limbach

Copyright: Forschungszentrum Jülich / Ralf-Uwe Limbach

https://www.fz-juelich.de/en/news/archive/press-release/2024/european-exascale-supercomputer-jupiter-sets-new-energy-efficiency-standards-w n-1-ranking-on-green500

JUREAP Framework

From System Capability to Application Readiness

- JUPITER Research and Early Access Program (JUREAP)
  - Defines readiness criteria for JUPITER Supercomputer
  - Supports the smooth build-up during the assembly phase of JUPITER
  - Two phases:
    - SPEP, the Scalability and Performance Evaluation Phase
      - Strong/Weak scaling on JUWELS BOOSTER
      - Full-machine (more than 800 nodes) run on JUWELS BOOSTER
    - Performance portability to JUPITER
      - Systematic cross-platform comparison of JUWELS BOOSTER and JUPITER

JUREAP Framework

From System Capability to Application Readiness

- JUPITER Research and Early Access Program (JUREAP)
  - Defines readiness criteria for JUPITER Supercomputer
  - Supports the smooth build-up during the assembly phase of JUPITER
  - Two phases:
    - SPEP, the Scalability and Performance Evaluation Phase
      - Strong/Weak scaling on JUWELS BOOSTER
      - Full-machine (more than 800 nodes) run on JUWELS BOOSTER
    - Performance portability to JUPITER
      - Systematic cross-platform comparison of JUWELS BOOSTER and JUPITER

No hardware performance normalization or roofline analysis is performed in this study, the results should be interpreted as direct application-level runtime comparisons rather than peak efficiency estimates.

JUREAP Framework

From System Capability to Application Readiness

- JUPITER Research and Early Access Program (JUREAP)
  - Defines readiness criteria for JUPITER Supercomputer
  - Supports the smooth build-up during the assembly phase of JUPITER
  - Two phases:
    - SPEP, the Scalability and Performance Evaluation Phase
      - Strong/Weak scaling on JUWELS BOOSTER
      - Full-machine (more than 800 nodes) run on JUWELS BOOSTER
    - Performance portability to JUPITER
      - Systematic cross-platform comparison of JUWELS BOOSTER and JUPITER

> **OUTCOME**: **JUREAP Certificate** acknowledging the suitability to run efficiently on JUPITER

ParFlow: Extreme-scale Readiness

Benchmarks

- **ClayL:** a variably saturated infiltration problem; part of benchmarking suite of JUPITER[1]
- **Terrain Sine:** a 3D simulation of groundwater-surface water flow with sinusoidal topography; part of vendor benchmarks at JSC; represents all linear physics present in scientific production runs such as Global 1km[2]

1. A. Herten et al., "Application-Driven Exascale: The JUPITER Benchmark Suite," SC24: International Conference for High Performance Computing, Networking, Storage and Analysis, Atlanta, GA, USA, 2024, pp. 1-45, doi: 10.1109/SC41406.2024.00038
2. Kollet et al., 2026. Global groundwater modeling: Proof-of-concept of 3D variably saturated flow simulation at kilometer resolution. Journal of Hydrology X 30, 100213, https://doi.org/10.1016/j.hydroa.2025.100213

ParFlow: Extreme-scale Readiness

Strong scaling on JUWELS BOOSTER

- **Setup**
  - JUREAP's structured node selection guidelines
    - Problem sizes (DOFs) to ensure per-GPU memory footprint approximately constant for **base node** counts across various scaling scenarios

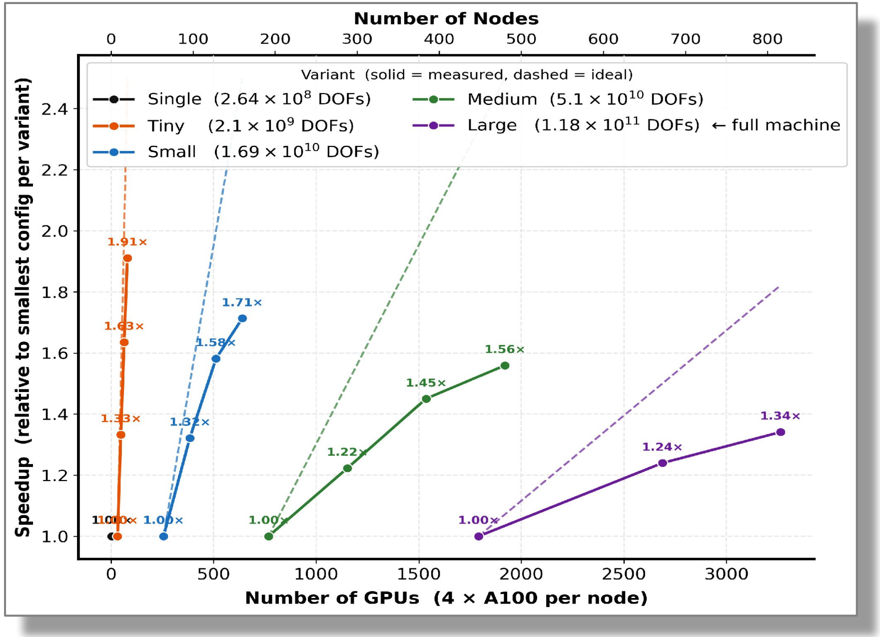
Variant	Problem Size (DOFs)	Node count
Single	2.64×10^8	1
Tiny	2.1×10^9	**8**, 12, 16, 20
Small	1.69×10^{10}	**64**, 96, 128, 160
Medium	5.1×10^{10}	**192**, 288, 384, 480
Large	1.18×10^{11}	**448**, 672, 816

ParFlow: Extreme-scale Readiness

Strong scaling on JUWELS BOOSTER

- Terrain Sine
- Max speedup achieved is **1.9x** w.r.t corresponding ideal baseline for Tiny variant

- **Scalability:** Full machine run over **816** nodes (**3264 A100 GPUs**)



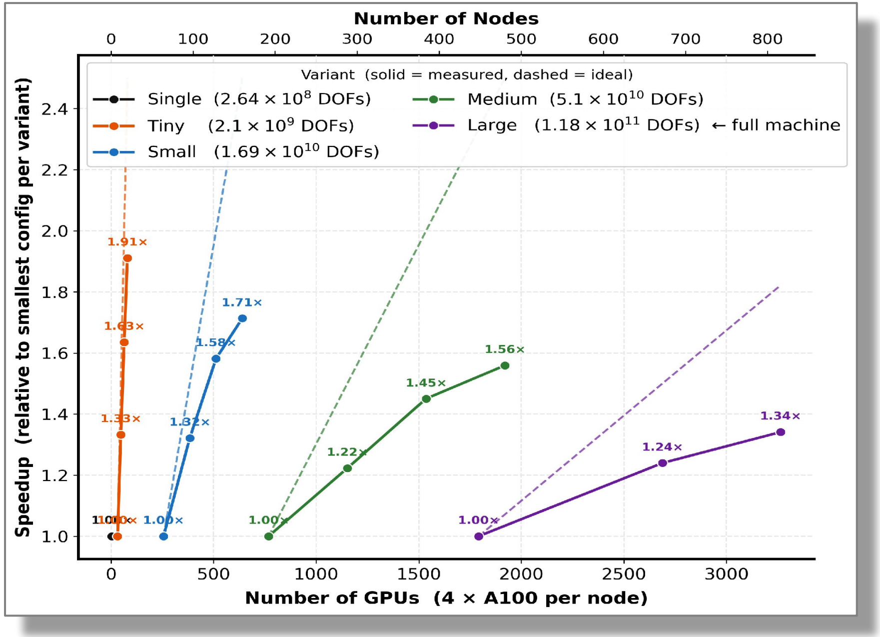
ParFlow: Extreme-scale Readiness

Strong scaling on JUWELS BOOSTER

- Terrain Sine
- Max speedup achieved is **1.9x** w.r.t corresponding ideal baseline for Tiny variant

- **Scalability:** Full machine run over **816** nodes (**3264 A100 GPUs**)

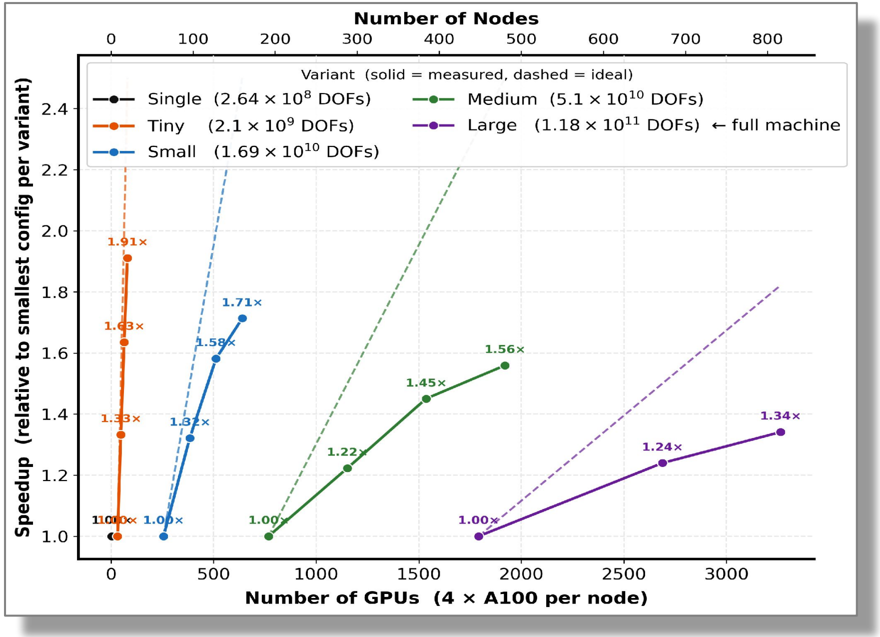
> exceeds the JUREAP scalability target of demonstrating efficient execution beyond 800 nodes (3,200 GPUs)



ParFlow: Extreme-scale Readiness

Strong scaling on JUWELS BOOSTER

Efficiency degrades at high decomposition including the two effects identified from interconnect counters via JSC monitoring infrastructure LLView[1]



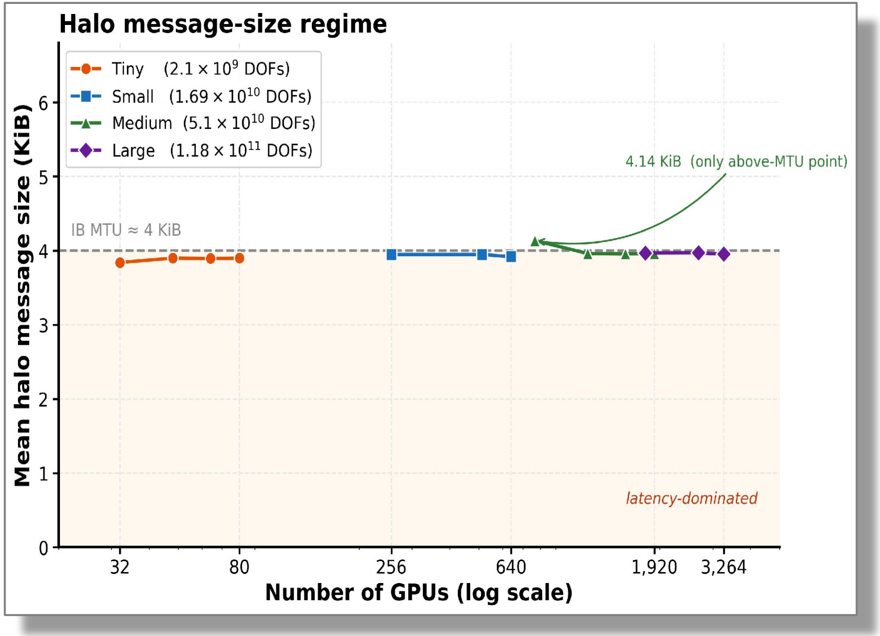
---

1. https://www.fz-juelich.de/en/jsc/services/user-support/software-tools/llview

ParFlow: Extreme-scale Readiness

Strong scaling on JUWELS BOOSTER

- **Message size:**
  - Message size remains ~4 KiB across all scales
  - Communication granularity remains nearly constant.
  - Strong scaling reduces workload per GPU, not communication size.
  - Communication overhead therefore becomes increasingly visible.

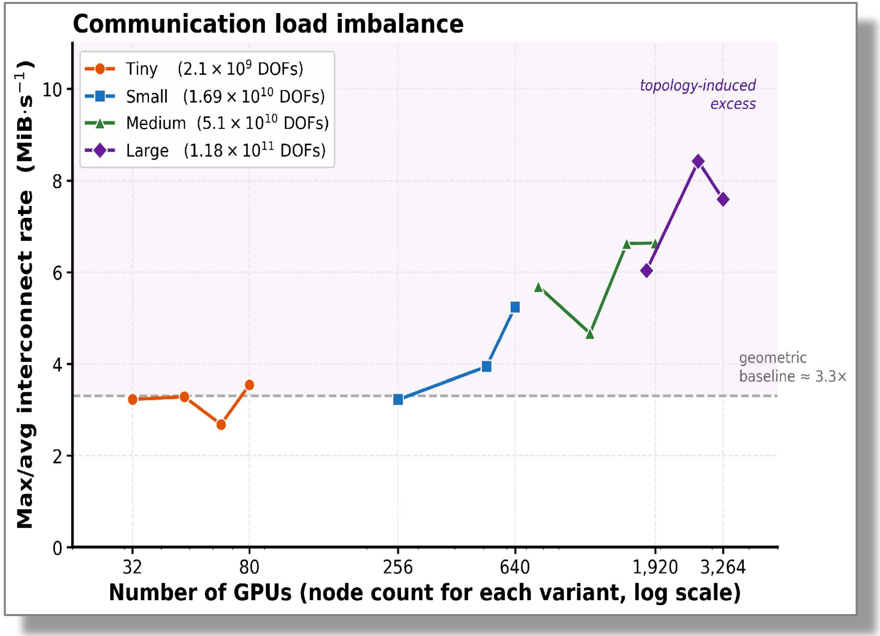


1. https://www.fz-juelich.de/en/jsc/services/user-support/software-tools/llview

ParFlow: Extreme-scale Readiness

Strong scaling on JUWELS BOOSTER

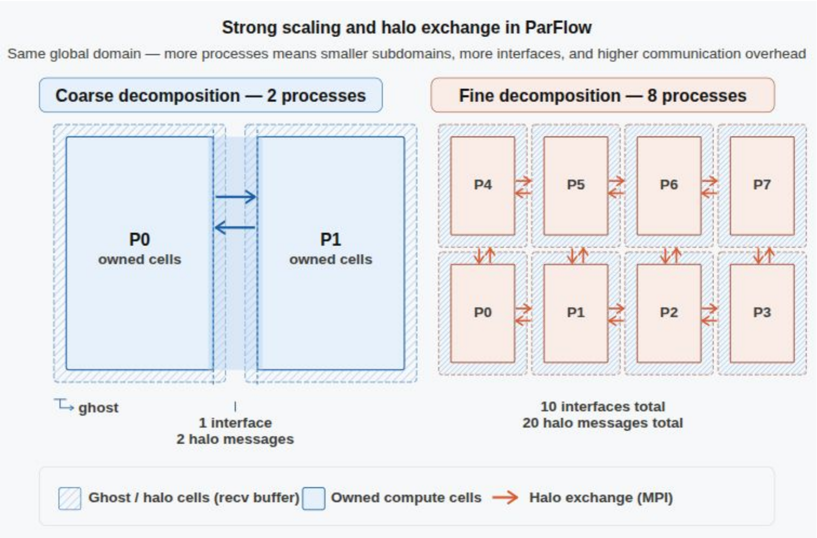
- **Communication load imbalance**: Interconnect traffic becomes increasingly imbalanced
  - Max/avg traffic ratio grows at scale
  - Larger allocations exhibit greater communication variability
  - Consistent with increased network complexity at scale



ParFlow: Extreme-scale Readiness

Strong scaling on JUWELS BOOSTER

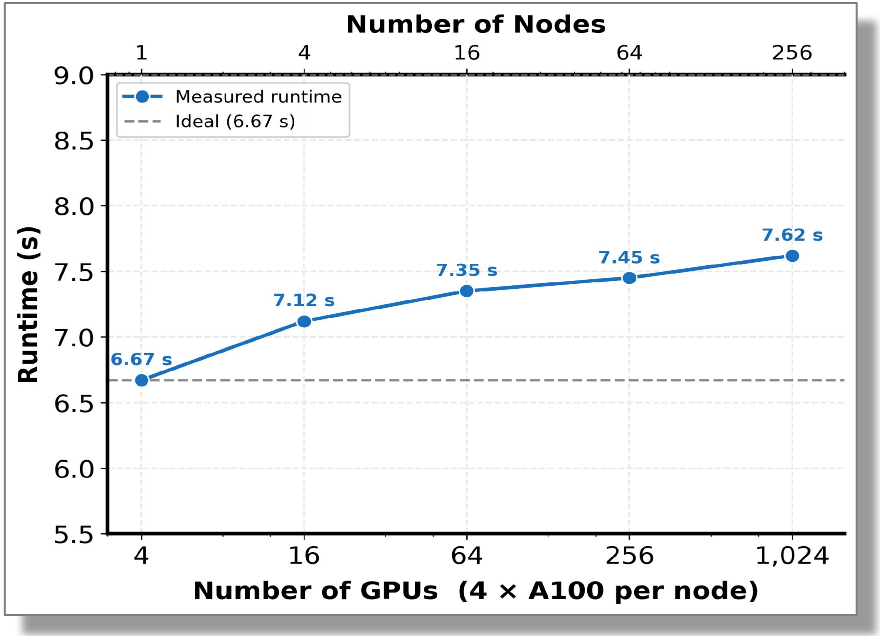
- Domain Decomposition: Core Idea
  - 3-D computational domain split across processes
  - Each process owns a subgrid + ghost (halo) layers
  - Ghost layers required for stencil-based updates (neighbor communication)
- Increased Halo Exchange Cost
  - Surface-to-volume ratio increases as subdomains shrink
  - Relative amount of boundary cells grows
  - More frequent / proportionally larger ghost layer updates
  - Communication becomes dominant:
    - Higher MPI message count
    - Increased latency sensitivity
    - Reduced compute/communication overlap efficiency



ParFlow: Extreme-scale Readiness

Weak scaling on JUWELS BOOSTER

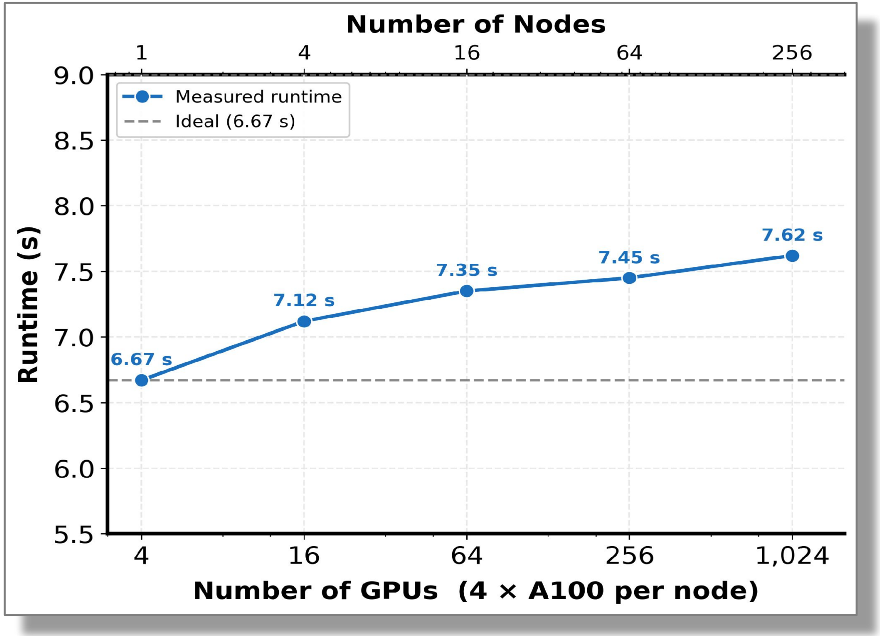
- **Setup**
  - ClayL benchmark
  - Fixed problem size: 2.4x10^8 cells for each node to fill its GPU memory capacity (40 GB per GPU)
  - Scalability: 1-256 JUWELS BOOSTER nodes



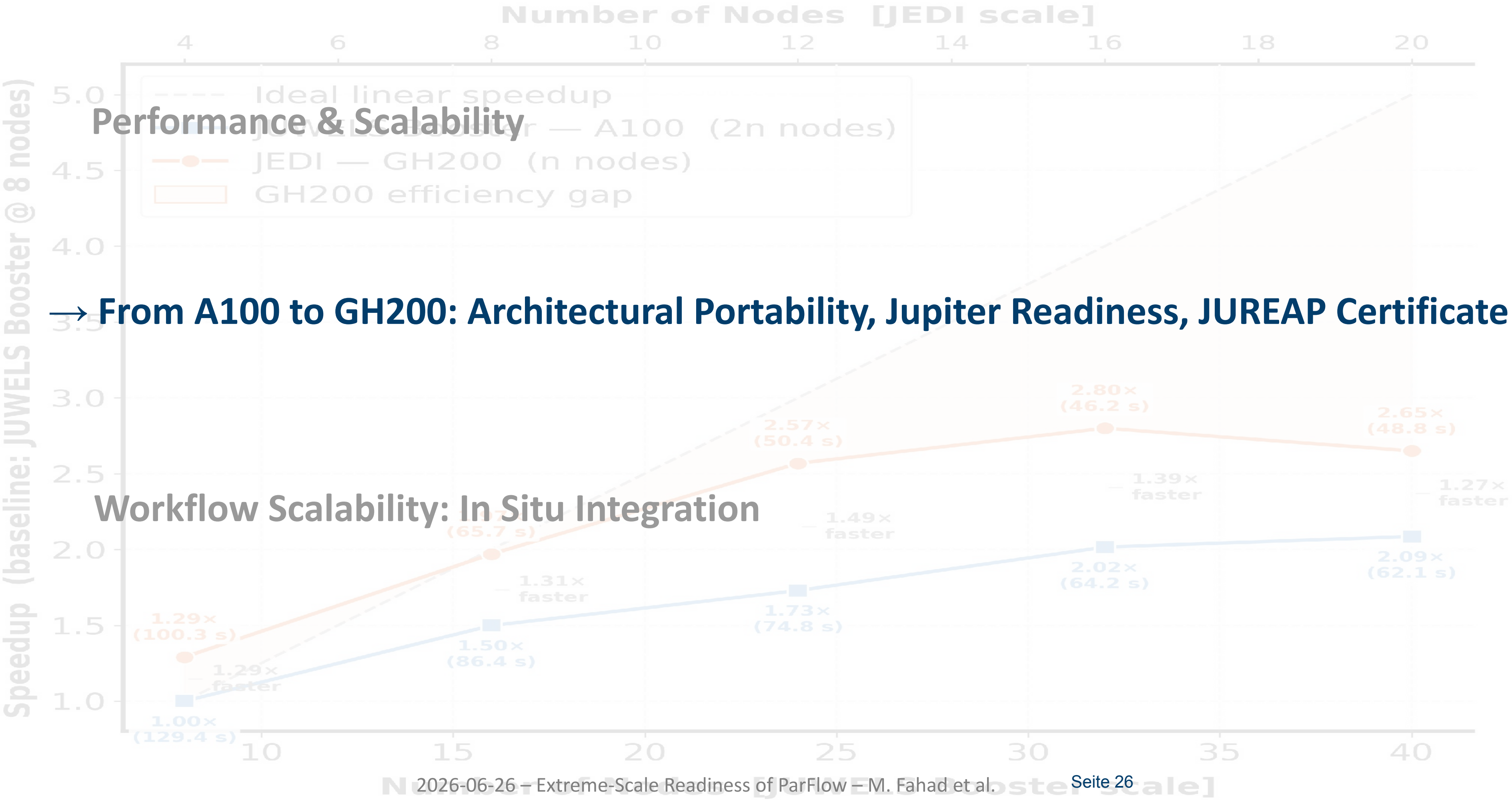
ParFlow: Extreme-scale Readiness

Weak scaling on JUWELS BOOSTER

- **Setup**
  - ClayL benchmark
  - Fixed problem size: 2.4x10^8 cells for each node to fill its GPU memory capacity (40 GB per GPU)
  - Scalability: 1-256 JUWELS BOOSTER nodes



> A largely flat weak-scaling profile over the evaluated range, with only a modest increase at higher node counts



Performance & Scalability

→ From A100 to GH200: Architectural Portability, Jupiter Readiness, JUREAP Certificate

Workflow Scalability: In Situ Integration

ParFlow: Extreme-scale Readiness

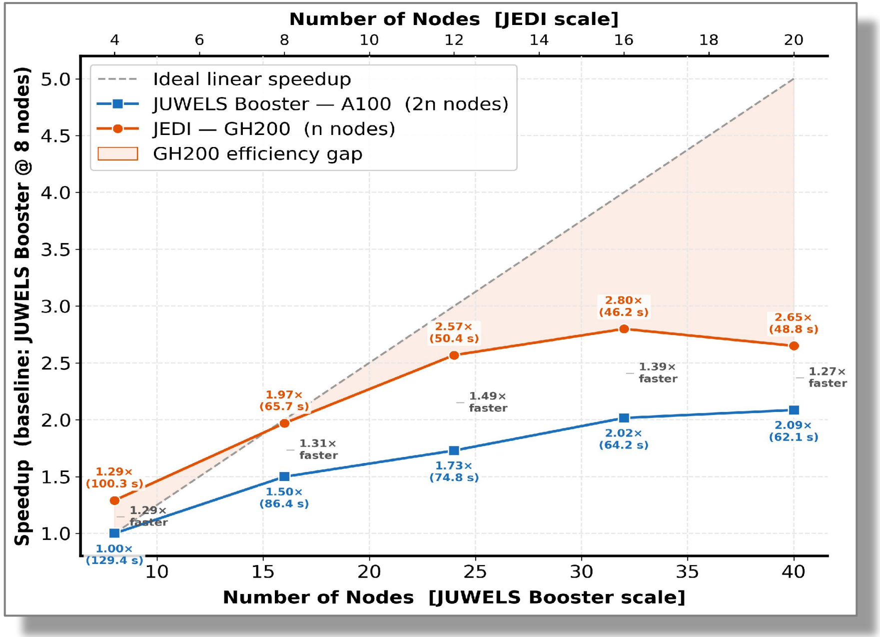
Architectural Portability

- **Setup**
  - JUPITER Exascale Development Instrument (JEDI)
  - Terrain sine benchmark
  - JUREAP's evaluation protocol
    - Identical per-node workload while respecting the memory constraints of JUWELS Booster
    - Fixed problem sizes (DOFs) executed on ***n*** JEDI nodes and ***2n*** JUWELS BOOSTER nodes

Problem Size (DOFs)	JEDI node count	JUWELS BOOSTER node count
1.06×10^9	4, 8, 12, 16, 20	8, 16, 24, 32, 40

ParFlow: Extreme-scale Readiness

Architectural Portability



> The minimum runtime of 46.24 seconds is observed on 16 JEDI nodes, corresponding to a 2.8× speedup over the JUWELS Booster baseline, using half the node counts.

ParFlow: Extreme-scale Readiness

Summary and Highlights

- JUPITER Research and Early Access Program (JUREAP)
  - Strong scaling: Max speedup achieved is **1.9x** w.r.t corresponding corresponding ideal baseline
  - Weak scaling: sustained throughput and near-constant execution time across a 256-fold increase in problem size and node count
  - Scalability: Full machine run over **816** nodes (**3264 A100 GPUs**)
  - Architectural Portability: Without any code modifications, JEDI achieves a maximum speed up of **2.8x** over JUWELS BOOSTER using half the node counts

ParFlow: Extreme-scale Readiness

Summary and Highlights

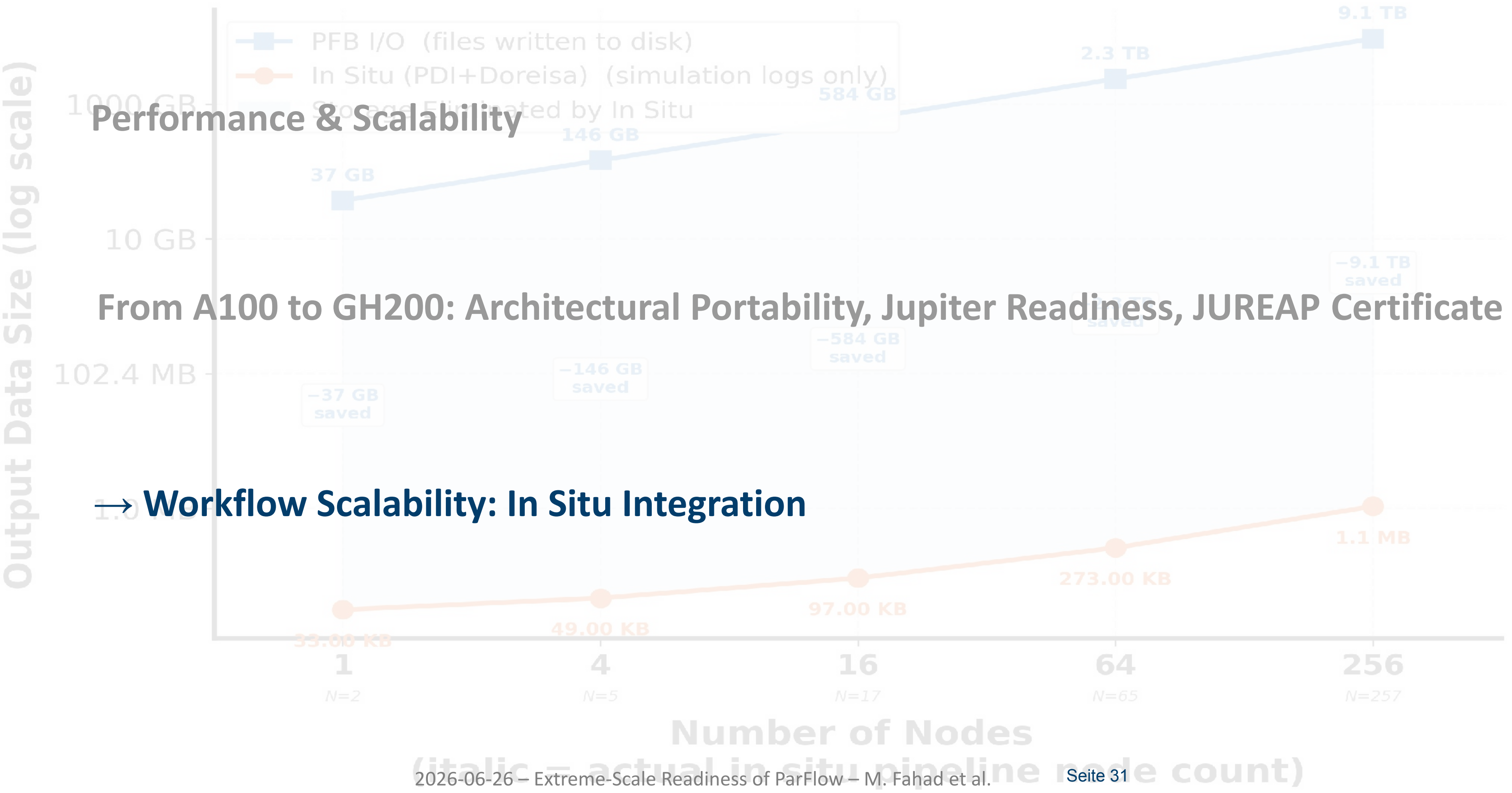
- JUPITER Research and Early Access Program (JUREAP)
  - Strong scaling: Max speedup achieved is **1.9x** w.r.t corresponding corresponding ideal baseline
  - Weak scaling: sustained throughput and near-constant execution time across a 256-fold increase in problem size and node count
  - Scalability: Full machine run over **816** nodes (**3264 A100 GPUs**)
  - Architectural Portability: Without any code modifications, JEDI achieves a maximum speed up of **2.8x** over JUWELS BOOSTER using half the node counts

> **Awarded JUREAP Certificate** recognising the **Exascale** potential with **strong** recommendation for execution on **JUPITER** Booster

Performance & Scalability

From A100 to GH200: Architectural Portability, Jupiter Readiness, JUREAP Certificate

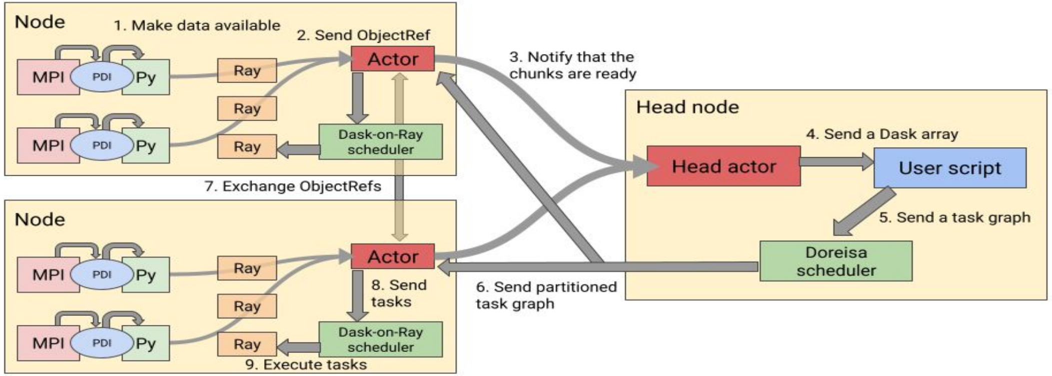
→ Workflow Scalability: In Situ Integration



ParFlow: Workflow Scalability

In Situ Integration: Pipeline Architecture

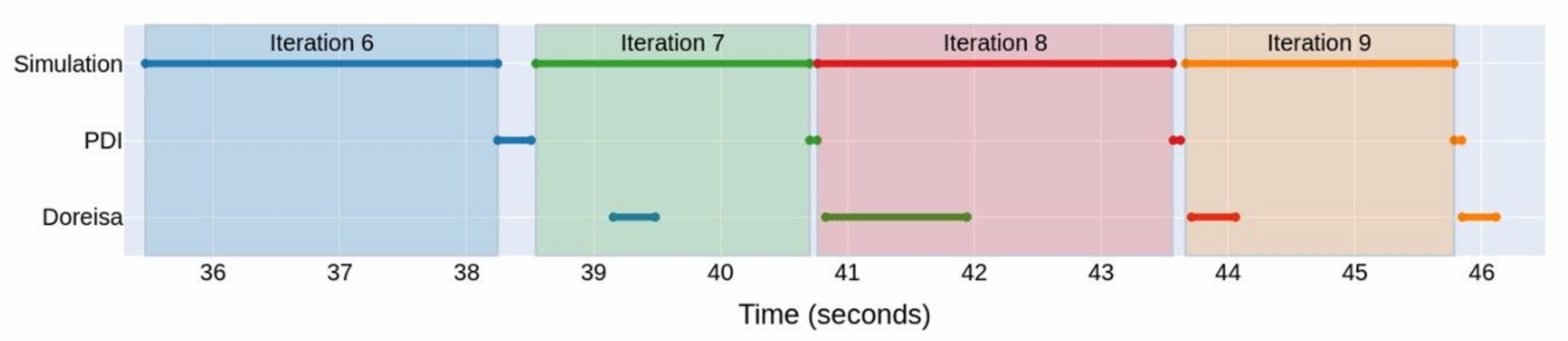
- **PDI (PDI Data Interface):** a lightweight middle-ware library that decouples the simulation from I/O handling through a declarative API
- **Doreisa:** a Ray extension of the Deisa distributed analytics framework
- **Data interception & distribution:** PDI captures simulation output buffers and, via the PDI Pycall plugin, forwards them to distributed Ray workers running across compute nodes.
- **Distributed diagnostics:** Ray workers execute user-defined diagnostics (e.g., spatial aggregation) expressed as NumPy operations and coordinated by a Ray head node.



ParFlow: Workflow Scalability

In Situ Integration: Pipeline Architecture

- Non-intrusive execution
- Data exchange happens after each timestep, allowing diagnostics to run concurrently with bookkeeping without affecting solver convergence or the numerical solution.



ParFlow: Workflow Scalability

In Situ Integration: Results

- Setup
  - JUWELS BOOSTER
  - ClayL benchmark
  - Typical PFB outputs (write only) vs In situ analytics
  - Weak scaling

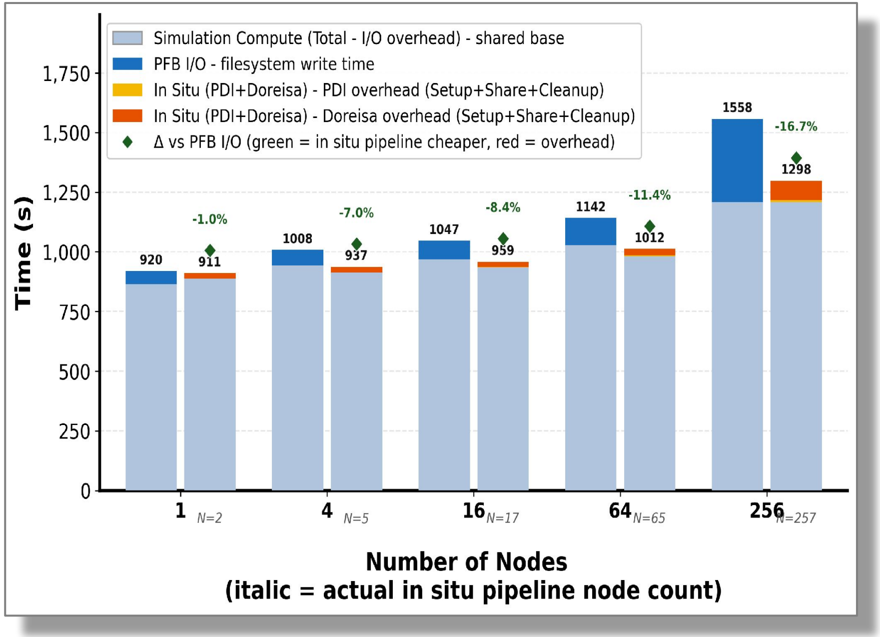
Problem Size (DOFs)	Node count
2.48×10^8	1, 4, 16, 64, 256

- ParFlow (PFB) Setup: N simulation nodes
- In situ Setup: N simulation nodes + 1 Ray head node

ParFlow: Extreme-scale Readiness

In Situ Integration: Runtime Efficiency

- 3.9× lower data-movement cost through in-memory coupling (89.1 s vs. 349.6 s at 256 nodes).
- ≥7.8× end-to-end advantage when snapshot read-back is included (9.1 TB of data movement avoided).



ParFlow: Extreme-scale Readiness

In Situ Integration: Runtime Efficiency

- 3.9× lower data-movement cost through in-memory coupling (89.1 s vs. 349.6 s at 256 nodes).
- ≥7.8× end-to-end advantage when snapshot read-back is included (9.1 TB of data movement avoided).

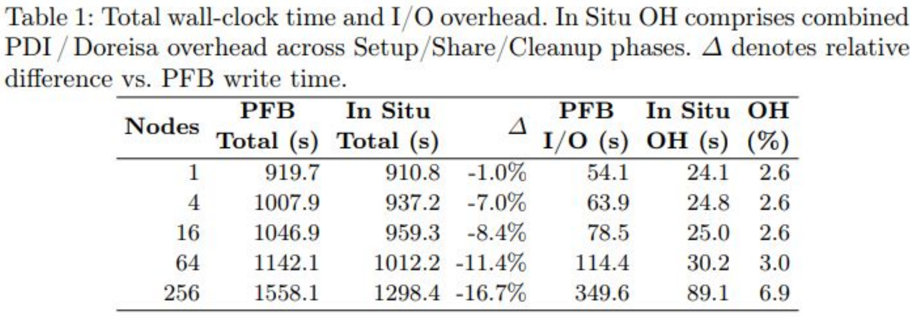


Table 1: Total wall-clock time and I/O overhead. In Situ OH comprises combined PDI / Doreisa overhead across Setup/Share/Cleanup phases. Δ denotes relative difference vs. PFB write time.

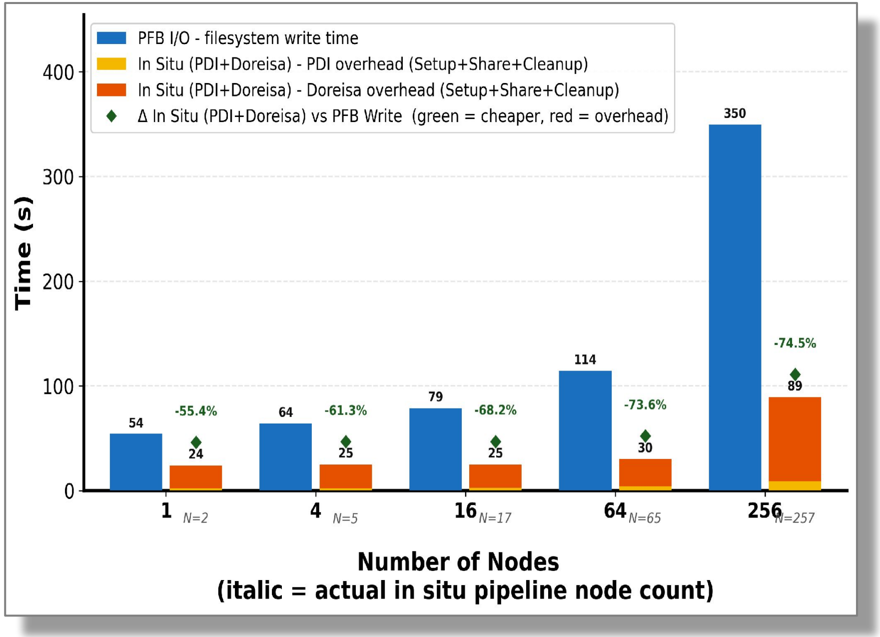
Nodes	**PFB Total (s)**	**In Situ Total (s)**	Δ	**PFB I/O (s)**	**In Situ OH (s)**	**OH (%)**
1	919.7	910.8	-1.0%	54.1	24.1	2.6
4	1007.9	937.2	-7.0%	63.9	24.8	2.6
16	1046.9	959.3	-8.4%	78.5	25.0	2.6
64	1142.1	1012.2	-11.4%	114.4	30.2	3.0
256	1558.1	1298.4	-16.7%	349.6	89.1	6.9

> Better scalability: diagnostics are computed in memory, eliminating expensive filesystem I/O and post-processing bottlenecks

ParFlow: Extreme-scale Readiness

In Situ Integration: I/O Overhead

- In situ overhead scales only 3.7× across a 256× increase in node count, compared to 6.5× growth for PFB I/O overhead.
- The performance advantage widens with scale, increasing from 2.2× faster at 1 node to 3.9× faster at 256 nodes.



ParFlow: Extreme-scale Readiness

In Situ Integration: I/O Overhead

- **Overhead Breakdown**
    - PDI middleware costs are negligible (<10 s at all scales).
    - Data sharing cost remains flat (~7–8 s), indicating effective overlap with simulation execution.
    - In Situ Setup dominates at scale, increasing from 14.4 s ⟶ 71.5 s and accounting for 80% of total overhead at 256 nodes.

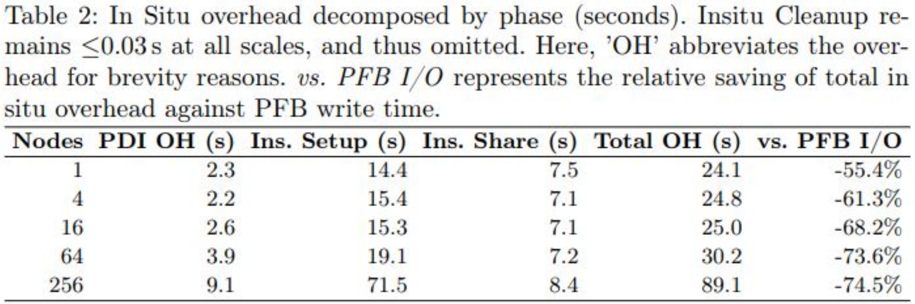


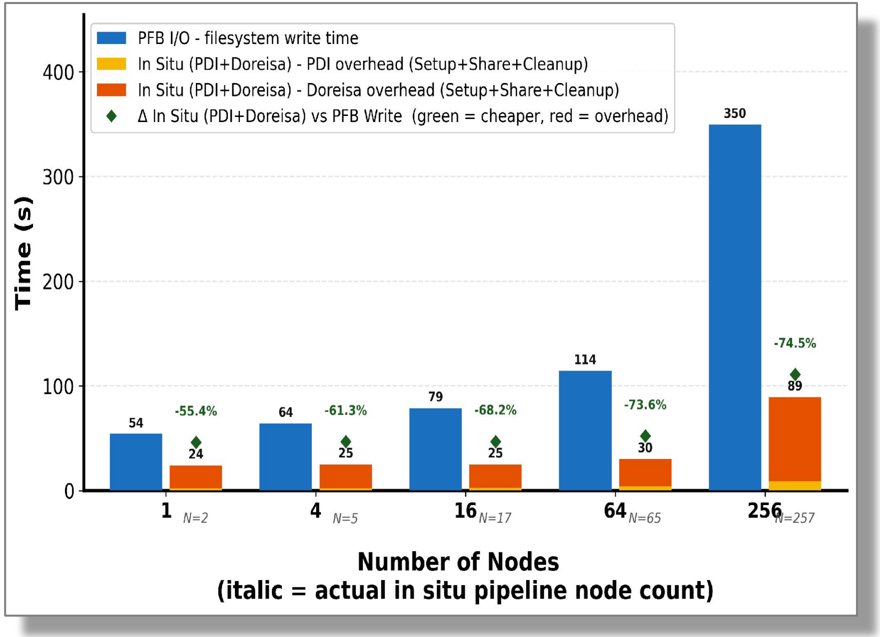
Table 2: In Situ overhead decomposed by phase (seconds). Insitu Cleanup remains ≤ 0.03 s at all scales, and thus omitted. Here, 'OH' abbreviates the overhead for brevity reasons. *vs. PFB I/O* represents the relative saving of total in situ overhead against PFB write time.

Nodes	PDI OH (s)	Ins. Setup (s)	Ins. Share (s)	Total OH (s)	vs. PFB I/O
1	2.3	14.4	7.5	24.1	-55.4%
4	2.2	15.4	7.1	24.8	-61.3%
16	2.6	15.3	7.1	25.0	-68.2%
64	3.9	19.1	7.2	30.2	-73.6%
256	9.1	71.5	8.4	89.1	-74.5%

ParFlow: Extreme-scale Readiness

In Situ Integration: I/O Overhead

- **Overhead Breakdown**
  - PDI middleware costs are negligible (<10 s at all scales).
  - Data sharing cost remains flat (~7–8 s), indicating effective overlap with simulation execution.
  - In Situ Setup dominates at scale, increasing from 14.4 s → 71.5 s and accounting for 80% of total overhead at 256 nodes.

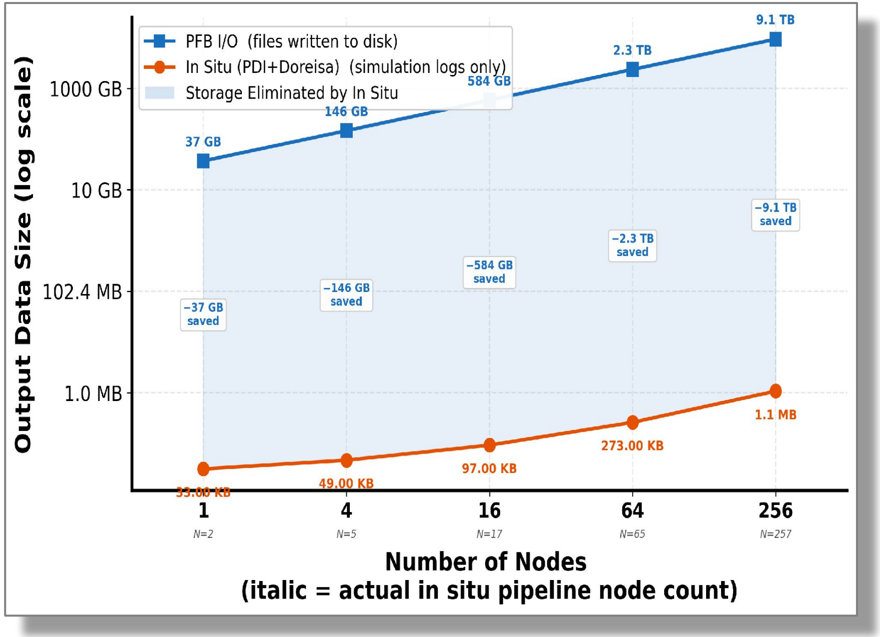


> Runtime savings are increasingly dominated by eliminating I/O, while setup initialization is the primary remaining optimization target.

ParFlow: Extreme-scale Readiness

In Situ Integration: Eliminating the Storage Bottleneck

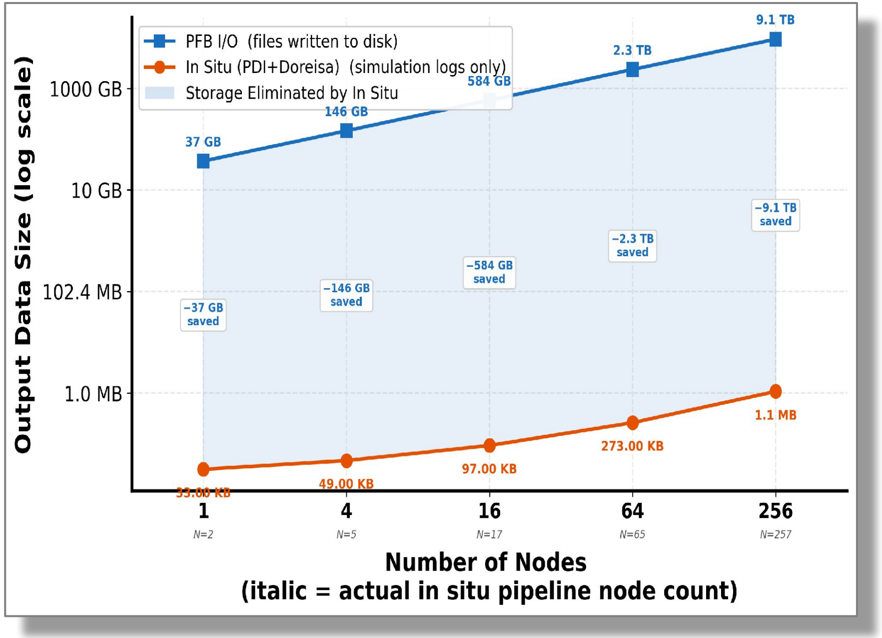
- PFB output grows linearly with scale, reaching 9.1 TB for only 10 timesteps at 256 nodes.
- The in situ workflow generates no simulation snapshots, producing only logs and metadata (<1.1 MB at 256 nodes).
- Extrapolated to a production campaign (10,000 timesteps, 256 nodes), PFB output would reach ~91 PB, exceeding typical Tier-0 storage allocations.



ParFlow: Extreme-scale Readiness

In Situ Integration: Eliminating the Storage Bottleneck

- PFB output grows linearly with scale, reaching 9.1 TB for only 10 timesteps at 256 nodes.
- The in situ workflow generates no simulation snapshots, producing only logs and metadata (<1.1 MB at 256 nodes).
- Extrapolated to a production campaign (10,000 timesteps, 256 nodes), PFB output would reach ~91 PB, exceeding typical Tier-0 storage allocations.



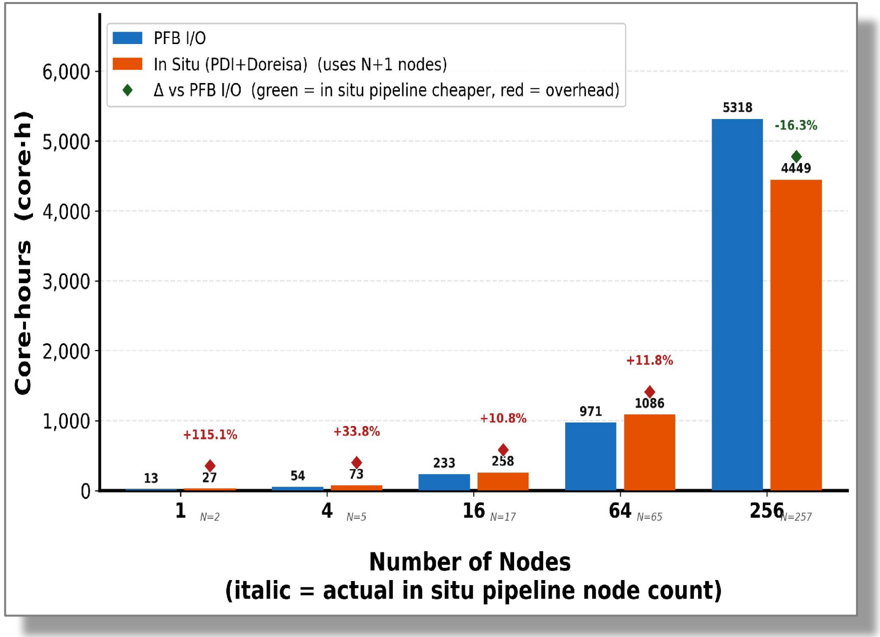
> In situ analysis transforms output from petabyte-scale datasets to megabyte-scale metadata, removing a major HPC workflow bottleneck.

> Best suited when required diagnostics can be computed during runtime.

ParFlow: Extreme-scale Readiness

In Situ Integration: Resource Efficiency at Scale

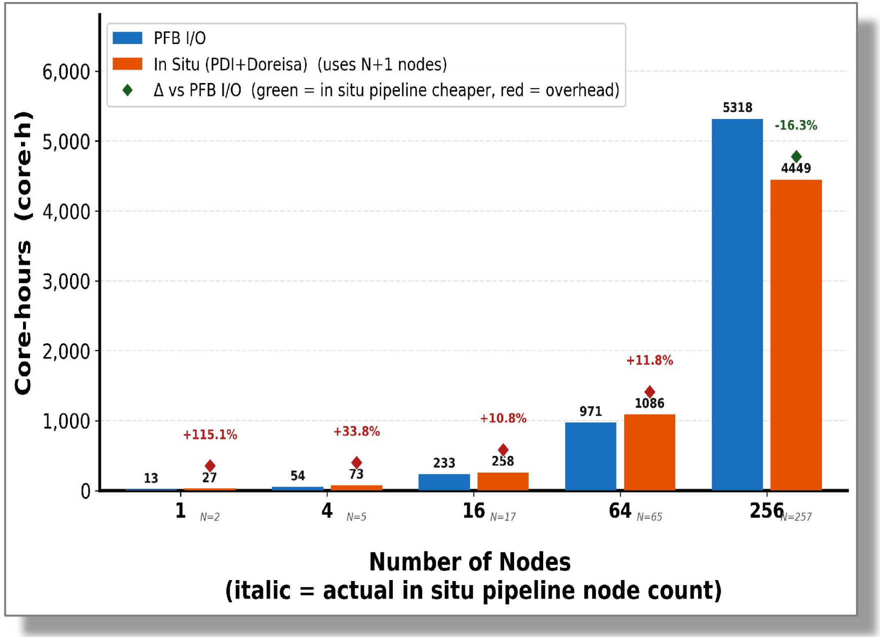
- The In situ workflow requires one additional node for the Doreisa Ray head (N+1 configuration).
- At small scales (1–16 nodes), the extra node introduces a measurable core-hour cost.
- From 64–256 nodes, reduced runtime outweighs the additional resource requirement.
    - At 256 nodes, the in situ pipeline saves 259.7 seconds (~16% reduction) despite using an extra node.



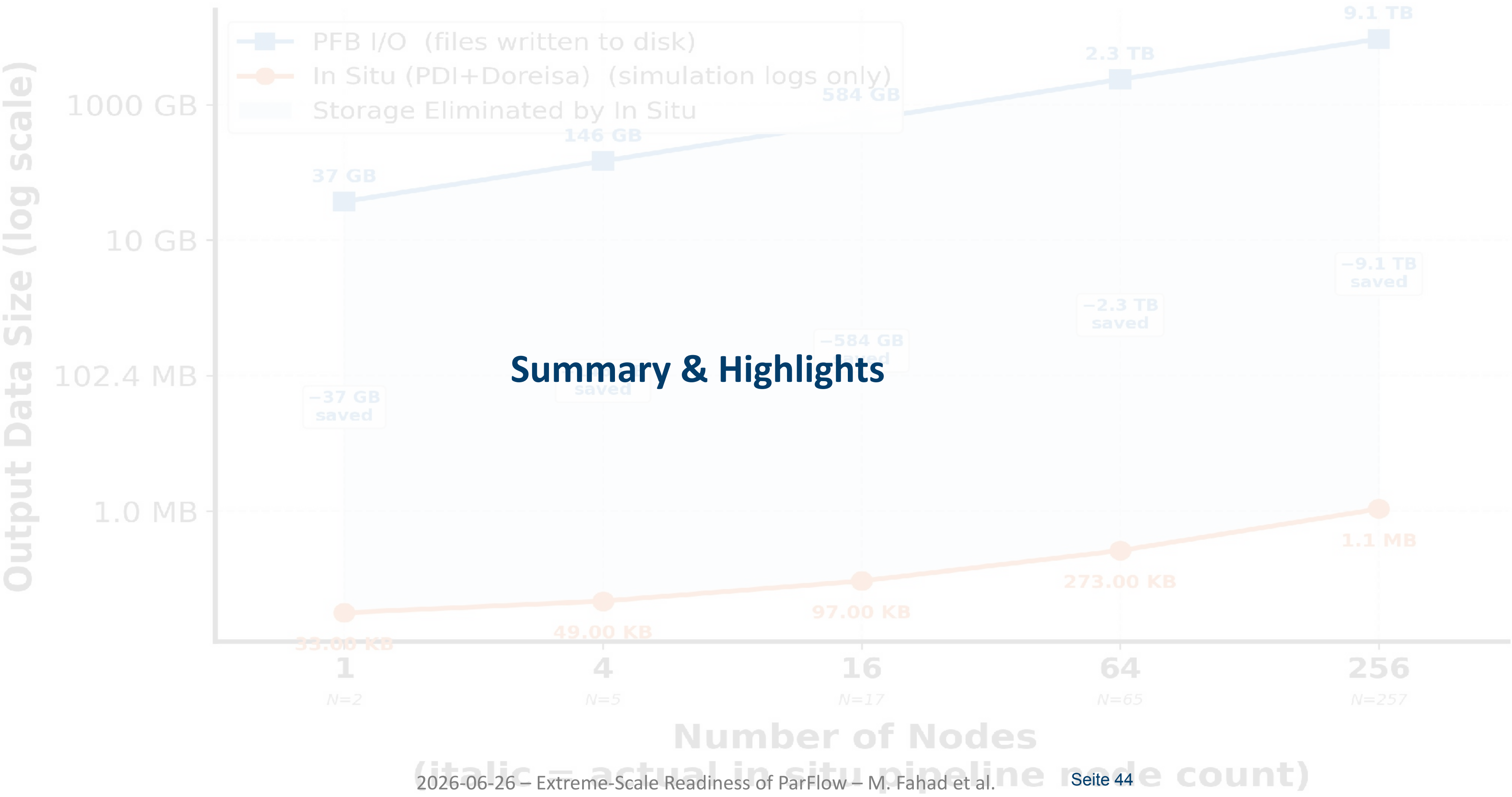
ParFlow: Extreme-scale Readiness

In Situ Integration: Resource Efficiency at Scale

- The workflow requires one additional node for the Doreisa Ray head (N+1 configuration).
- At small scales (1–16 nodes), the extra node introduces a measurable core-hour cost.
- From 64–256 nodes, reduced runtime outweighs the additional resource requirement.
  - At 256 nodes, the in situ pipeline saves 259.7 seconds (~16% reduction) despite using an extra node.



> The additional orchestration node becomes negligible at scale, while the reduction in I/O overhead delivers a net gain in resource efficiency and time-to-insight.



Summary & Highlights

ParFlow: Extreme-scale Readiness

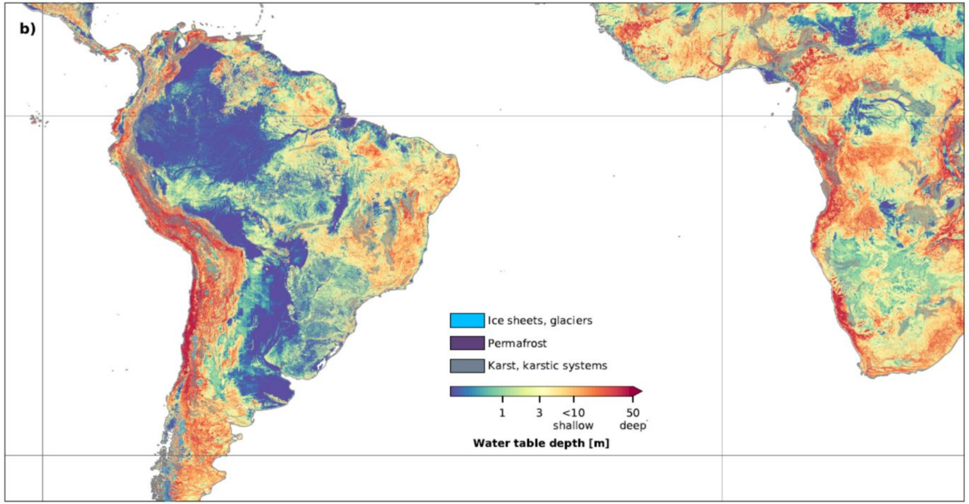
Summary and Highlights

Aspect	Indicator	Evidence	Observed Limitation
Strong scaling	Achieved	Full-machine run with 1.18 × 10^{11} DOFs; JUREAP certificate confirming exascale suitability and strong node-level performance	Sub-linear efficiency beyond ~400 nodes due to latency-dominated halo exchange and network traffic imbalance
Weak scaling	High efficiency	Largely flat profile	Residual loss from halo exchange and global synchronisation
Architectural Portability to JUPITER	Verified	2.80× speedup on GH200 relative to A100 at equal problem size and half the node count; no code modifications required	Differences in scaling behavior under fixed problem size, consistent with architectural differences between GH200 and A100 systems; no hardware-level profiling performed
Workflow scalability	Demonstrated	In situ pipeline reduces runtime by up to 16.7%; write-only I/O cost reduced 3.9× (excluding post-processing read-back and analysis); output volume reduced from 9.1 TB to 1.1 MB at 256 nodes	Insitu Setup exhibits super-linear growth at large scale

ParFlow: Extreme-scale Readiness

Future Outlook

- Optimize *In situ Setup* phase
  - Ray worker initialisation and connection management
- In situ - hybrid CPU-GPU setup (on JUPITER)
  - N+1 nodes → N nodes
- Energy efficiency studies
- Node-level optimizations
- Sub-kilometer spatial resolution



Kollet et al., 2026. Global groundwater modeling: Proof-of-concept of 3D variably saturated flow simulation at kilometer resolution. Journal of Hydrology X 30, 100213.

https://www.parflow.org

https://github.com/parflow/parflow

Acknowledgement

The authors gratefully acknowledge the Earth System Modelling Project (ESM) for funding this work by providing computing time on the ESM partition of the supercomputer JUWELS at the Jülich Supercomputing Centre (JSC). This project has received funding from the European High Performance Computing Joint Undertaking under grant agreement no. 101144014 (EoCoE-III). The authors acknowledge the JUPITER Research and Early Access Program (JUREAP) for providing access to JEDI, the JUPITER Exascale Development Instrument. JUPITER is funded by the EuroHPC Joint Undertaking, the German Federal Ministry of Research, Technology and Space, and the Ministry of Culture and Science of the German state of North Rhine-Westphalia.